Using cryo-EM methods to uncover structure and function of bacteriophages

Ouyang, R.

8

Chapter 8

General Discussion

General Discussion

Cryo-electron microscopy (cryo-EM) has revolutionized the field of structural biology and has increasingly found applications in cell biology. However, despite its demonstrated efficacy, the technique is limited in its ability to study disordered, flexible complexes, as well as large macromolecular biological machines. Our investigation revealed that one of the primary challenges in the study of flexible complexes is their inherent structural heterogeneity. Additionally, the study of large macromolecular biological machines poses a challenge due to their sheer size and complexity, which requires specialized procedures for data acquisition, processing, and analysis. Furthermore, the presence of multiple conformations and subpopulations within such machines requires the use of advanced data processing techniques.

Our findings emphasize the need for specialized procedures and workflows to overcome the limitations of cryo-EM in studying flexible complexes and large macromolecular biological machines. Through the implementation of specialized procedures and workflows and diligent data processing, we can overcome these challenges and expand our understanding of the structures and functions of biological molecules. In my thesis work, I have investigated several such complexes and have encountered and overcome various challenges, which I will discuss in detail in this chapter.

1. Optimized classification strategies of SPA reconstruction

Optimized 2D classification strategy

An optimized 2D classification strategy was used for processing data obtained during the single-particle analysis (SPA) workflow using the Relion software. The Laplacian-of-Gaussian option allows for the auto-picking of hundreds of thousands of particles with different conformations without the need for templates. Sophisticated image classification algorithms are required to compensate for non-homogeneous samples during reconstruction. To efficiently screen a large quantity of particles, I developed an optimized strategy for 2D classification, involving multiple iterations to screen complete particles with clear edges, a key characteristic for virus-capsid particles. This strategy could eliminate most of the "bad" particles with unclear edges and blurry

surfaces, thereby providing significant advantages in screening and classifying large amounts of data. The optimized strategy of 2D classification was found to retain more good quality particles as compared to the strategy shown in the official tutorial. This optimized approach can be applied to a large amount of data obtained by cryo-EM during reconstruction for SPA.

Typically, the extracted particle data is often binned by two or four (pixel binning) to accelerate data processing. However, the binning approach reduce the characteristic structural features of the proteins. Binning is useful during the first iteration of 2D classification, enabling the quick screening and removal of most obvious wrongly-selected particles. However, unbinned data should be used during the subsequent particle extraction and classification steps to maintain high accuracy of the classification results. This optimized workflow for small particles enables the classification of millions of particles with high speed while maintaining high accuracy of the classification results.

Optimized 3D classification method

In this study, I used an optimized 3D classification method to address the challenge of limited computing resources to process large datasets that also contain different conformations of the research target. The method employs a "branch method" strategy that uses the results from the initial classification as references for further processing steps. Iterations of 3D classification are then run separately for each of these classes, resulting in reconstructed 3D models that represent the different conformations of the target. The 3D models and the proportion of classes are used to select particles from final classes to run 3D refinement separately for reconstruction. The optimized branch method of 3D classification can screen and classify hundreds of thousands of particles and is user friendly, making it suitable for use on a small computational workstation in any laboratory. Although the optimized branch method for 3D classification requires more time and patience from the user, it is highly advantageous for screening and classifying high-quality particles, particularly those with flexible proteins exhibiting different conformations. This approach is well-suited for use in large datasets with many conformations during SPA reconstruction.

2. The challenges and optimization of macromolecular reconstruction

Structural studies of complex biological macromolecules with unique morphologies present significant challenges. Despite the past preference for phages as targets for structural studies due to their high symmetry, jumbo or mega phages with irregular structures, such as the jumbo phage φKp24 and phage 7-7-1, still pose a challenge. Nevertheless, such unique morphologies are required for successful infection of a host and therefore are especially interesting to understand structurally and functionally. In my work, I used a novel combination of cryo-EM single particle analysis, tomography and machine learning approaches analyze these challenging samples. In turn, this allowed me to gain novel insights into the infection process of bacteriophages of less-well studied architectures.

More specifically, in **Chapter 4** and **Chapter 5**, I studied the jumbo phage φKp24, which targets *Klebsiella pneumoniae* as a host. To reconstruct the capsid, tail, and tail fibers of the phage, we used two different cryo-electron microscopy reconstruction methods, single-particle analysis and cryo-electron tomography, in combination with protein structure prediction software and molecular dynamics simulation methods. Furthermore, we employed machine learning to investigate the conformational changes and rearrangement process of the tail fibers of φKp24 after attachment to the host cell surface. In addition, we determined the capsular polysaccharide (CPS) type of Klebsiella pneumoniae that φKp24 is capable of infecting in a clinical setting. Our findings shed light on the challenges and optimization of macromolecular reconstruction and offer insights into the structural biology of jumbo phages.

Similarly, in **Chapter 6** and **Chapter 7**, I studied phage 7-7-1 with infects *Agrobacterium* sp. H13-3. Within this context, I focused on the unusual capsid of this phage. The elucidation of the fundamental capsid structure was achieved through SPA, synergistically integrating cutting-edge protein structure prediction tools and the efficacy of molecular dynamics simulation techniques. Additionally, the interaction between the capsid fibers and the host flagella was investigated using of cryo-ET. These studies unveiled fresh insights into the structure and operational mechanisms governing flagellitrophic phages, a class that still remains relatively underexplored in the scientific discourse.

The reconstruct challenges of the large size capsids

In the study of phage φKp24, the large size of the capsid and the heterogeneous nature of the tail fibers present significant challenges for both data collection and analysis. Addressing these challenges requires a combination of various structural techniques and an innovative approach to data processing. In **Chapter 4**, the large size of the capsid, coupled with a high volume of data, presents computational challenges during data processing. Specifically, the diameter of the ϕKp24 capsid is approximately 1450 Å. and the pixel size of the collected micrographs is 1.37 Å. As a result, a box size of 1100 is required during the extraction step if we use raw data. However, processing particles of this size with conventional methods, such as 2D or 3D classification programs, can be computationally intensive (one classification job may cost many hundreds of hours). Furthermore, it increases the likelihood of computer crashes or memory issues, even with high-end workstations such as the one equipped with sixty-four cores CPU, four 2080Ti GPU, and 256G memory that I used in this study. Of course, there are other options, such as processing the data using a powerful server or another higher-configuration workstation. However, this may be a prohibitive financial burden for research labs. Additionally, the use of powerful servers also requires competition for computing resources with other users, which may further hinder data processing efficiency.

In order to optimize computational resources and balance efficiency with final reconstruction quality, I decided to use a unconventional data processing strategy to deal with the challenge of processing data from large particles. During the extraction step, particles prepared for 2D classification were binned by a factor of four (pixel binning). Due to the high symmetry and large size of capsid particles, this binning method effectively screens particles and produces good results during 2D classification. Binning data could also be used for the initial model and 3D classification, but in the final 3D refinement the dataset needs to be binned by a smaller number to achieve high resolution results. The particles selected after 3D classification could then be binned by different numbers, and 3D refinement can be run iteratively. It is important to adjust the pixel size of the input model used for 3D refinement to the same pixel size corresponding to the different binning numbers in every round of refinement. Subsequently, the falling gradient of the FSC curve is used to determine if the dataset needs to be binned by a smaller number. If the end part of the FSC curve descends and approaches the X-axis, it indicates a good result, and that the reconstruction has reached the limit of the quality of the micrographs.

In the case of the capsid of jumbo phage ϕKp24, I used a data processing strategy that balances computational resources, efficiency, and final reconstruction quality. I began by selecting particles after 3D classification and binning them by 2x, which allowed me to run a 3D refinement and achieve a resolution of about 4.4 Å. However, I found that the reconstruction did not reach the limit of the quality of the micrographs. Therefore, I binned the dataset by 1.5x during the second 3D refinement to balance the final resolution with available computational resources. This resulted in a high-resolution density map with a resolution of 4.1 Å of ϕKp24's capsid, which is currently the highest resolution capsid of any jumbo phage deposited in the EMDB.

Prediction of atomic models without homologs

AlphaFold2 is a deep learning system developed by DeepMind that predicts protein structures with high accuracy. The system uses a neural network to predict the 3D structure of a protein from its amino acid sequence. AlphaFold2 has been shown to be highly accurate, with predictions that are often close to the true structure of the protein. The system has been used to predict the structures of thousands of proteins and has been used in a number of scientific studies. Protein structure prediction algorithms, such as AlphaFold2[63,64] and RoseTTAFold[198] have become invaluable tools for cryo-EM based macromolecular structure elucidation. In the research about phage φKp24, the genomic analysis revealed that the sequences show very limited similarities to any other known bacteriophage. This means that there were no suitable homologous atomic structures available to aid in model building. Therefore, I used AlphaFold2 to predict the protein structures of the capsid, sheath, and inner tube of phage φKp24 (**Chapter 4),** and the capsid of phage 7-7-1 (**Chapter 6**). After getting the predicted atomic models, I used the flexible fitting software ISOLDE[59] to fit the predicted atomic models into the density maps.

Building atomic models into low to medium resolution maps

ISOLDE is a flexible fitting software that enables precise adjustments of every atom in density maps. During the model building process, ISOLDE leverages molecular dynamics forcefields, such as AMBER and CHARMM, to create a realistic environment for atom fitting. These forcefields provide high-fidelity descriptions of the forces that govern macromolecules and small molecule ligands. By employing

molecular dynamics engines such as OpenMM, ISOLDE is able to efficiently solve Newton's laws of motion for these forces, even on systems containing just a few thousand atoms, thanks to the massive parallel computing capabilities of modern GPUs. Furthermore, the implementation of ISOLDE as a ChimeraX[172] plugin allows for real-time rendering of ongoing simulations through a fast and flexible API.

In **Chapter 4**, I directly build atomic models for the sheath and inner tube based on a 3 Å map, leading to the generation of high-quality models through flexible fitting. For lower to medium resolution experimental maps, such as the 4 Å map of the capsid, a combination of flexible fitting and molecular dynamic flexible fitting methods (MDFF)[69] was employed, yielding optimized and correct atomic models of the capsid for analysis of assembly and connection residues in subsequent steps.

Reconstruction and 3D structural visualization of complicated fibers

Visualizing the complex and tangled structures of fibers is still a big challenge in structural biology. One example are the highly heterogeneous tail fibers of phage φKp24 that are not suitable for single-particle analysis and sub-tomo averaging. In **Chapter 5,** I therefore used a novel combination of approaches to gain insight into the structure of these tail fibers using machine learning techniques. Specifically, a neural network was designed to accurately analyze cryo-ET datasets to trace the fibers of the imaged phages[205]. To achieve a reliable performance of the fiber tracing, I trained the program by manual segmentations of only seven phage particles. In turn, this enabled an automatic analysis of tail fibers from over 600 phages. Overall, we obtained hundreds of three-dimensional structures representing the various conformations of tail fibers in the tomograms. The resulting dataset offers a quantitative means to investigate the morphology of phage infection *in vivo* and in 3D. The results presented in **Chapter 5** demonstrate that the neural network successfully extracted and displayed the complex tail fiber structures. As an effective data processing method, it provides a new option for processing complex, complicated fiber data.

Chapter 7 describes an additional application of the neural network technology. Specifically, I examined phage 7-7-1, which features flexible fibers that extend from its capsid vertices and attach to the flagella of the bacterial host. These fibers, aided by the rotational motion of the flagellum, enable the phage to move along the flagella of

Agrobacterium sp. H13-3 and reach the host's cell envelope to initiate infection. Using cryo-ET, I could visualize the phages 7-7-1 and their interaction with the host flagella during the initial stages of the infection process. Furthermore, by leveraging the neural network algorithm developed in **Chapter 5**, I was able to automatically track and extract the capsid fibers and flagella of phage 7-7-1 and display them in different conformations as a three-dimensional structure. Taken together, these two applications of neural network technology demonstrate the potential of machine learning to facilitate the analysis of complex structures.

3. Future directions

Synergistic application of Cryo-ET and SPA

Cryo-ET and SPA are two widely used cryo-EM techniques for studying the three-dimensional structure of biological molecules. Cryo-ET involves imaging a specimen from different angles to produce a 3D image of the sample, while SPA involves imaging individual, purified particles and then process the images to produce a 3D structure. Cryo-ET is well-suited for examining large, complex structures like viruses and cells, whereas SPA is suitable for studying smaller, simpler structures such as proteins. Cryo-ET can capture dynamic processes and reveal the spatial organization of macromolecules in complex environments, while SPA provides high-resolution information of purified and homogeneous samples. Most commonly, both methods are used separately depending on the aim of the study.

In my research presented in this thesis, I show that the combination of Cryo-ET and SPA in combination presents numerous advantages for scientific research. More specifically, the simultaneous use of both techniques enabled me to overcome some of the inherent limitations of each individual method and obtain complementary information on the structure and function of macromolecular complexes. As a result, I could gain insight into structure and function of novel phages that would have been unattainable without the combination of cryoEM methods.

Furthermore, Cryo-ET and SPA can be used in conjunction to validate and refine structural models obtained from other techniques such as X-ray crystallography or NMR spectroscopy. Cryo-ET can validate the overall shape and arrangement of

components within the complex, while SPA can provide higher resolution information on individual components or subunits.

In summary, Cryo-ET and SPA exhibit distinct advantages and limitations in structural biology research. The combined use of Cryo-ET and SPA allows for a more comprehensive and detailed understanding of the structure and function of macromolecular complexes. This combined approach overcomes some of the limitations associated with each individual technique and increases the accuracy and reliability of the obtained results. The choice of technique should be based on the specific research objectives and the properties of the samples being investigated.

Future applications about AI in cryo-EM

Recent advances in cryo-EM have led significant increases of the obtainable resolution of macromolecular complex structures. However, the inherent complexity and heterogeneity of biological complexes still present significant challenges for structural analysis. To address these challenges, deep learning-based algorithms, specifically convolutional neural network (CNN)-based models, have been developed to enhance the cryo-EM image processing workflow. These algorithms have demonstrated their effectiveness in particle recognition and image classification, and they are also capable of reconstructing high-resolution 3D structures for heterogeneous samples. Moreover, recent developments in deep learning-based post-processing approaches have further improved the quality of the reconstructed 3D electron density maps, opening up new possibilities for high-resolution characterization of complex samples.

While cryo-EM has traditionally been used to analyze purified samples in vitro, recent studies have focused on characterizing molecular complexes in their native environments, known as protein communities. However, the high complexity of cell extracts makes it challenging to accurately quantify and reconstruct molecular complexes of interest. To overcome this, neural network-based approaches are being developed to detect and isolate particles of different shapes and sizes within protein communities from EM images of cell extracts. The primary challenge in this approach is determining the 3D model of each component from heterogeneous 2D projections of imaged cell extracts. The development of AI-based protein structure prediction tools has opened up new opportunities for studying native cell extracts, providing researchers

with reliable model prediction tools to gain insights into the 3D structure of molecular complexes.

Recent advances in the application of AI on cryoEM data has enabled the high-resolution structural analysis of biomolecular machines at various stages of their function, leading to unprecedented insights into molecular mechanisms. A prime example is the successful integration of cryo-EM with AI-driven multi-resolution simulations to observe the coronavirus-2 replication-transcription machinery in action[243]. By using AI-driven multi-resolution simulations, complex structures can be reconstructed from low-resolution data, thereby allowing the identification of previously undetected features and the refinement of existing models. In the future, AI methods about multi-resolution simulations could continuously learn about identification of different structural conformations of transient states for various biomolecular machines or proteins, even evolute to master the rule of conformational changes of biomolecules. As a result, this method could aid in understanding the wide range of structural changes of molecular complexes in the process of fulfilling biological functions. This promising method has the potential to revolutionize the field of structural biology, enabling the study of biological systems at an unprecedented level of detail.

Difficulties and challenges of cryo-EM

The development of cryo-EM has faced a range of challenges, including the specialized equipment and expertise required, limited access to the technology for many researchers due to its high cost and limited availability, and the difficulties in sample preparation for cryo-EM analysis. Cryo-EM data analysis is also a complex and computationally intensive process, requiring sophisticated algorithms for image processing, data analysis, and structure determination. While the integration of AI and machine learning techniques has helped in addressing some of these challenges, the development of effective and reliable algorithms remains an ongoing area of research. Furthermore, limitations in the resolution of the final structures obtained from cryo-EM due to factors such as sample heterogeneity, low signal-to-noise ratio, and image blurring require ongoing efforts to advance sample preparation techniques, data acquisition, and image processing algorithms.

Standardization of methods and protocols is necessary to ensure reproducibility and comparability of results between different laboratories. Therefore, ongoing efforts are needed to develop and implement best practices and quality control measures for cryo-EM experiments. Overcoming these challenges and obstacles will enable cryo-EM to revolutionize the field of structural biology, enabling the study of biological systems at an unprecedented level of detail.

Present and future development of cryo-EM

Cryo-EM research has been focused on a range of topics in recent years, including the study of membrane proteins, viral structures, and large macromolecular complexes. One of the prominent areas of investigation in cryo-EM has been the analysis of membrane proteins, which are critical components of many biological processes and represent important drug targets. Cryo-EM has proven to be a powerful tool in obtaining the structures and functions of these proteins, enabling researchers to better understand their roles in cellular signaling, transport, and other vital functions.

Another significant area of interest in cryo-EM research has been the investigation of virus structures and their mechanisms of action. Cryo-EM has played a key role in the development of effective vaccines and treatments for COVID-19 by enabling researchers to obtain high-resolution structures of viral proteins and complexes. These structures have provided insights into viral replication, pathogenesis, and host interactions, which are critical for the development of antiviral strategies.

In addition to these areas of research, cryo-EM has also been used to explore the structures of large macromolecular complexes, such as those involved in DNA replication and transcription. The detailed visualization of these complex structures has allowed researchers to better understand their mechanisms of action and interactions with other cellular components. This knowledge can lead to new insights into fundamental biological processes and disease mechanisms.

Moreover, significant advances have been made in cryo-EM with the development of new algorithms and methods for image processing, structure determination, and data analysis. These efforts have included the integration of artificial intelligence and machine learning techniques to improve the speed and accuracy of cryo-EM data

analysis. Cryo-EM has also been increasingly utilized in combination with other structural biology techniques, such as X-ray crystallography and NMR spectroscopy, to obtain a more comprehensive understanding of the structural features and dynamics of biological molecules and complexes.

As a rapidly advancing field with a promising future, cryo-EM has several directions of future development that hold great potential. One area of development is the push towards achieving higher resolution structures of macromolecules. While cryo-EM has already achieved atomic resolution for some macromolecules, it is still limited for many others. Future developments will focus on reducing the effects of radiation damage, improving sample preparation techniques, and developing more advanced image processing algorithms. Achieving higher resolution structures will provide deeper insight into the structural and functional mechanisms of biological systems.

Another area of future development is the study of larger and more complex samples, such as whole cells and tissues. Cryo-EM has already been used to study a range of samples, from single proteins to large macromolecular complexes. Cryo-ET, in particular, is well-suited to study large macromolecular complexes, but limited by the thickness of the investigated samples. In order to understand how macromolecular complexes interact in their native environment, it is essential to investigate thicker and more complex samples such as intact cells or tissues. To surmount this challenge, sample thinning methods have been developed, including cryo-sectioning and focused ion beam (FIB) milling. FIB milling, in particular, holds substantial potential to contribute significantly to the preparation of samples involving membranes, organelles, cytoskeleton, and macromolecular complexes[244]. This can provide insights into the molecular mechanisms of cellular processes and functions. Cryo-FIB milling can also be used to thin protein microcrystals for microcrystal electron diffraction (MicroED), which is a technique that produces high-resolution 3D molecular structures of small chemical compounds or biological macromolecules. This can enable the structural determination of proteins that are difficult to crystallize or purify, such as membrane proteins or protein complexes. With the continuous evolution of Cryo-plasma FIB technology, new electron microscopy tools, and automated methods for data collection and processing developed using AI and machine learning algorithms, the study of cells, tissues, and expansive macromolecular assemblies is poised for enhanced efficiency and deeper insights.

Another area of future development is the integration of cryo-EM with other techniques, such as X-ray crystallography, NMR spectroscopy, and fluorescence microscopy. By combining these complementary approaches, researchers can obtain a more comprehensive view of biological systems. Future developments will focus on developing more advanced methods for integrating data from multiple techniques. Additionally, the ability to study dynamic processes in real-time is an important area of development. Cryo-EM is currently limited in this regard, but future developments will focus on developing methods for studying dynamic processes using time-resolved cryo-EM or by combining cryo-EM with other techniques that can provide temporal information.

Finally, automation of cryo-EM data collection and processing is a critical area of development. Current methods are time-consuming and labor-intensive, but progress is being made by the development of automated methods, including the use of robotics and machine learning algorithms. These developments will not only enhance the efficiency of cryo-EM research but also have the potential to provide new insights into the mechanisms of biological systems.