Chapter 18

Computational Structural Biology Research Team

18.1 Members

Florence Tama (Team Leader)

Miki Nakano (Postdosctoral Researcher)

Sandhya Tiwari (Postdosctoral Researcher)

Yumeno Kusuhara (Assistant)

18.2 Research Activities

18.2.1 Introduction

Biological molecular complexes of such as proteins and RNAs are of great interest in the area of molecular biology, as they are involved in cell replication, gene transcription, protein synthesis, regulation of cellular transport and other core biological functions. Those systems undergo large conformational transitions to achieve functional processes. Therefore characterization of dynamical structures of these macromolecular complexes is crucial to understand their functional mechanisms and play an important role in the development of new drugs to treat human disease.

Experimentally, X-ray crystallography has been the primary tool to study protein structures, providing high-resolution structures. In recent years, cryo electron microscopy (EM) is also becoming a vital approach, due to rapid developments in technology and data processing software. It can be used to observe large macromolecules without the necessity of crystallization and has revealed a wealth of critical information on structure and dynamics of important large biological molecules. More recently, X-ray free-electron laser (XFEL) light sources offer a new possibility to image single biological macromolecules. RIKEN/SACLA is one of a few facilities that currently exist and several more being constructed in the world. It can produce photon pulses significantly stronger than previous facilities and enables instant "single shot imaging" of biological systems. Since crystallization is not necessary for such measurements, it would be possible to investigate the structure of biomolecules under various physiological conditions or to observe elementary steps of a biochemical function. However, in the current experimental condition, it cannot achieve atomic level resolution such as obtained by X-ray crystallography.

Computationally, algorithms to process experimental data play critical roles to obtain the structural models of biological molecules, because biological molecules are exceedingly complex, consisting of thousands to millions of atoms. Even without experimental data, atomic models could be predicted using homology modeling and ab initio prediction methods. Algorithms to analyze protein/proteins interactions also have shown successes in predicting proteins complexes. While such ab initio predictions, based solely on computation, succeed in predictions for small proteins, it still remains difficult for large proteins. Therefore, in the studies of biologically important large macromolecular complexes, it is essential to integrate computational modeling approaches with experimental data.

The ultimate line of our interdisciplinary research is to bring experimental data as obtained from X-ray, cryo-EM, and XFEL with development and applications of computational tools, through K computer, to acquire knowledge on the structures of physiologically important protein complexes that are unattainable with existing experimental techniques. We have been working on the development of data analysis algorithms, with the emphasis on the integration of molecular mechanics simulation and experimental data processing, and applying the developed algorithms to experimental data through collaborations.

18.2.2 Computational tools for XFEL experimental data

The recent development of intense X-ray free-electron laser (XFEL) light sources offers a possibility to obtain new structural information of biological macromolecules. Strong X-ray pulse allows measurement of X-ray diffractions from microcrystals, and furthermore, enables the imaging of single biological macromolecules. Measurements using such a strong pulse also has advantages, enabling time-resolved measurements at femtoseconds resolution as well as the measurements at the ambient temperature. SPring-8/SACLA is one of a handful of XFEL facilities in the world. However, with current experimental conditions, atomic level resolution such as obtained by X-ray crystallography cannot yet be achieved. As XFEL experiments are very recent and still undergoing further development for routine applications to biological molecules, computational algorithms and tools to analyze experimental data also need to be developed simultaneously. One focus of our research is the development of such computational tools from multiple angles summarized below.

18.2.2.1 3D structure reconstruction from 2D diffractions data

Three-dimensional (3D) structural analysis for single particles using X-ray free electron laser (XFEL) enables us to observe hard-to-crystallize biomolecules in a state close to nature. To restore the 3D structure of the molecule from the diffraction patterns obtained by XFEL experiments, computational algorithms are necessary as one needs to estimate the angles of incident laser beams to the molecule for each of 2D diffraction patterns and assemble them into 3D volumetric structural information.

We have demonstrated the 3D structure reconstruction of a large biological molecule, ribosome, from diffraction data created by computer simulation. We proposed a sequence of algorithms that allow accurate angle estimations and 3D reconstructions. In addition, we showed the experimental conditions that are required (the number of diffraction images and the intensity of the laser beam) to restore the molecular structure at a certain resolution (Figure 18.1) (Nakano et al, J. Sync Radiat. 2018). Following this work, we are developing data processing protocols to apply such reconstruction algorithms to actual experimental data. Experimental data contains noise and uncontrollable fluctuation of samples and measurement conditions, unlike synthetic data. Therefore, we are developing algorithms for filtering high-quality diffraction patterns that contain signals useful to reconstructions

These data analysis tools are being developed using XMIPP, which is commonly used for image processing of single-particle 3D cryo EM. Since XMIPP is designed to work with 2D data in real space, some of the routines were modified to deal with 2D diffraction patterns in Fourier space. The programs for the reconstructions are ported to K-computer and being customized to improve performance.

18.2.2.2 "Idea generator" from 2D data of biological systems

Data analysis for XFEL data remains challenging. XFEL diffraction pattern is an unintuitive representation of the projection image of the sample in reciprocal space. For biological systems, the current standard approach to reconstruct a real-space image of the sample, phase recovery, often fails due to the low diffraction power of biological samples. Therefore, we are developing a new hybrid approach to interpret diffraction patterns that utilizes image analysis with database search. In this approach, for an obtained XFEL diffraction pattern, the algorithm proposes a few low-resolution 3D models that are consistent with the data using a database of known or hypothetical shapes of biological systems (Figure 18.2).

We previously demonstrated the feasibility of such an approach using cryo EM images, which are in "real space". We generated a database of shape, utilizing the existing EM database and developed a protocol to identify 3D models that match a set of inquiry 2D images (submitted). Following this, we have been adopting this approach for the application to XFEL diffraction patterns, which is in "Fourier space". From the database of shapes, expected diffraction patterns in all possible beam orientations are precomputed, and inquiry XFEL diffraction patterns are compared against this database. We are refining the similarity detection procedures to increase the accuracy of diffraction pattern comparisons.

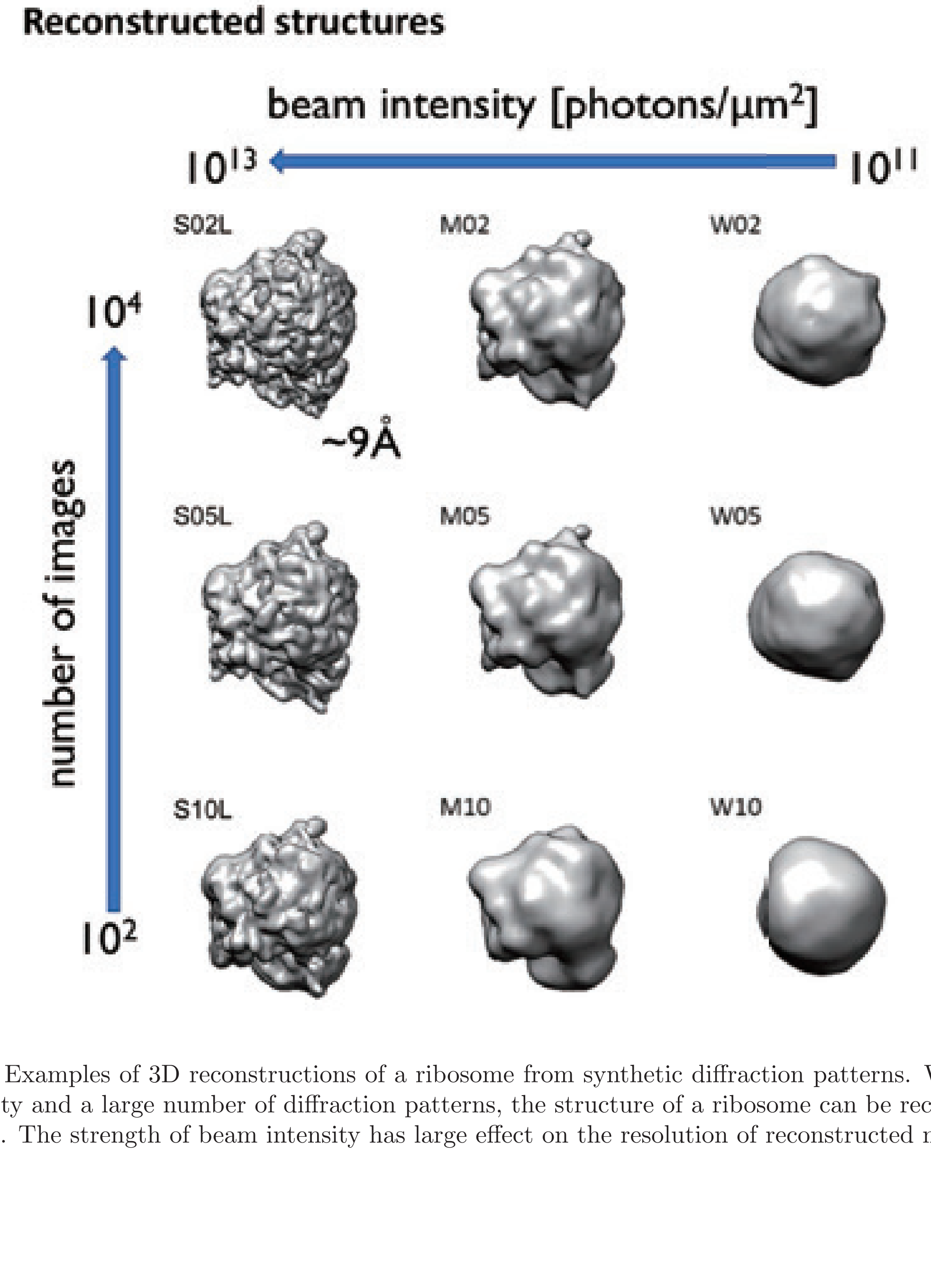

Figure 18.1: Examples of 3D reconstructions of a ribosome from synthetic diffraction patterns. With stronger beam intensity and a large number of diffraction patterns, the structure of a ribosome can be reconstructed at 9Åresolution. The strength of beam intensity has large effect on the resolution of reconstructed models.

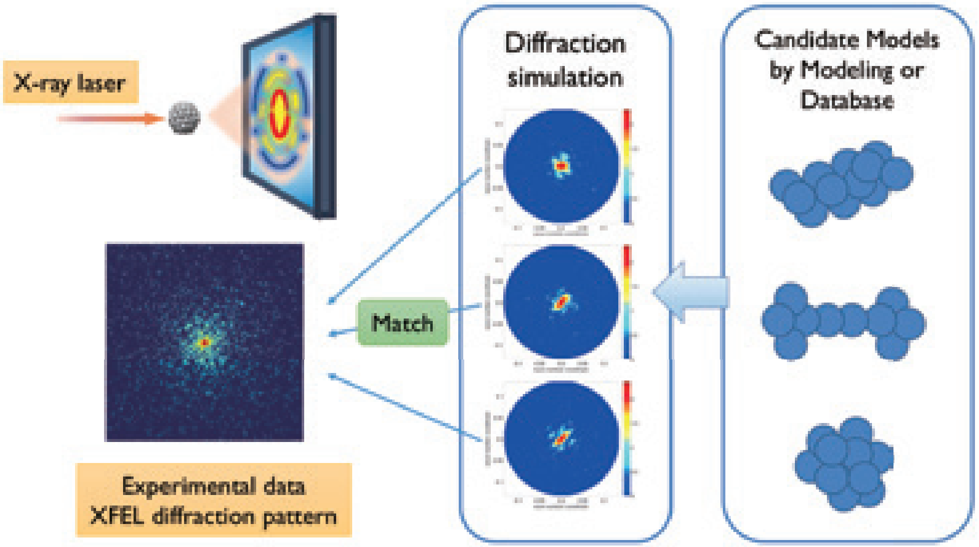
X-ray laser
Diffraction simulation
Candidate Models by Modeling or Database
Match
Experimental data XFEL diffraction pattern

Figure 18.2: Scheme of structure model reconstruction from a few experimental diffraction patterns. A large number of possible structure models are generated using the database that assembles known biological shapes, or sampling techniques using coarse-grained models. For each candidate model, expected diffraction patterns are calculated and compared against the target experimental diffraction pattern to identify the structure that is most consistent with the data.

18.2.2.3 Coarse-grained structure representation for XFEL data analysis

In parallel to the above database approach, we are also exploring ab initio approaches, where the structural models that match experimental data are generated through optimization procedures (Figure 2). We are examining the use of Gaussian mixture model (GMM), which approximates a biomolecular shape by the superposition of Gaussian distributions. As the Fourier transformation of GMM can be quickly performed, GMM can be used to efficiently simulate XFEL diffraction patterns from approximated sample structures. We have shown that with a sufficient number of GMMs, diffraction patterns that are highly consistent with the ones simulated from atomic models can be obtained using a fraction of CPU time (Figure 18.3). These results demonstrate that GMM serve as a useful coarse-grained models, in the context of hybrid modeling approach, in XFEL single particle experiments (submitted).

18.2.3 Refinement of cryo-EM structure models using X-ray structures

Recent advances in cryo-EM experiments and data processing have produced high-resolution structures of biomolecules. Nevertheless, in many cases, the obtained resolution still remains in the 6-12 Årange, which requires computational techniques to build reliable atomic models. In an approach, “flexible fitting”, known X-ray structures are optimally deformed to match experimental data using molecular mechanics simulations. In collaboration with Dr. Sugita’s laboratory in Wako, we have been implementing such type of approaches in GENESIS. Taking advantage of the generalized ensemble algorithms embedded in GENESIS to maximize conformational sampling, increase in reliability of the atomic models was achieved. Fitting subroutines are fully parallelized in the latest implementation and allow the modeling of large macromolecules.

We have been further improving the flexible fitting approaches to improve its applicability to real experimental data. 3D volume from cryo-EM reconstruction contains strong noise and there are large uncertainties in the data. In recent applications, we are examining the protocols to calibrate the parameters for flexible fitting to ensure the quality of the resulting models. The result strongly indicates that the noise in the data can easily cause “overfitting”, where the resulting model is distorted by noise. We also examined the procedure to calibrate the size parameter of pixels in the microscopy data, which may contain some error (Figure 18.4).

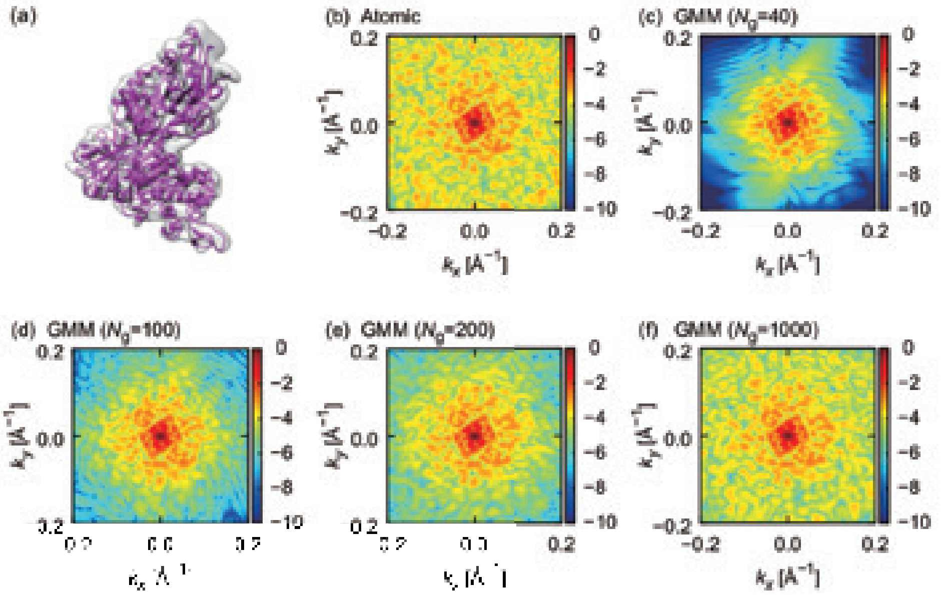

Figure 18.3: Demonstration of fast diffraction pattern simulation using GMM. (a) The original atomic model of a protein, EF2. (b) A diffraction pattern that is simulated from the atomic model. (c-f) The diffraction patterns simulated using GMM. As the number of Gaussians that are used to approximate the volume of original protein increases, the simulated diffraction patterns agrees better with the one from the atomic model. With 100 Gaussians, diffraction patterns can be simulated up to 0.1 Å^{-1}, which is sufficient for the modeling of large macromolecules.

18.3 Schedule and Future Plan

Cryo-EM experiments are quickly becoming an essential tool for studying biomolecular complexes. New XFEL facilities are getting into operation every year in the world, providing opportunities for new experiments. The amount of data from such experiments will continue to grow in numbers and analysis of such big datasets will increase the necessity of high performance computing. We aim to utilize K and post-K to break the limitation of current processing power and to obtain a new level of structural information of biological complexes from EM and XFEL data. For this goal, we plan to develop algorithms and software to analyze large dataset to obtain not only structural models but also dynamical information that can utilize computers in different sizes such as cluster and supercomputer. By sharing the software and results from structural modeling with other research institutes, we aim to contribute to the structural biology community.

XFEL facility SACLA started its operation in 2012. Other XFEL facilities are gradually established in the world and the number is still limited, but data for 亮 m systems, such as viruses and organelle, are being obtained. We have been developing algorithms to obtain structural models of biological systems by using XFEL data from two aspects.

In one approach, we aim to predict structural models from a limited amount of XFEL data. Such an approach is necessary since XFEL data collection is still a challenge and, moreover, the biological samples are intrinsically not uniform and it is difficult to obtain a large number of diffraction patterns from the samples in an identical conformation. We have developed a series of algorithms for coarse-grained representation of the structures, matching the model against experimental data. At the next step, we plan to establish sampling and optimization procedures to generate models that match experimental data.

In another approach, we aim to reconstruct structural models from a large number of XFEL data. With the advancement of experimental techniques, more than 10,000 XFEL diffraction images could be soon obtained, and furthermore, theoretical studies suggest that 1 million images are necessary to obtain high resolution models. In the near future, we will need to utilize such a large amount of dataset to reconstruct detailed 3D models from XFEL data. Thus, we have been developing algorithms to efficiently construct 3D models from XFEL diffraction patterns. We are aiming to apply the algorithm to new experimental data, and for this purpose, we

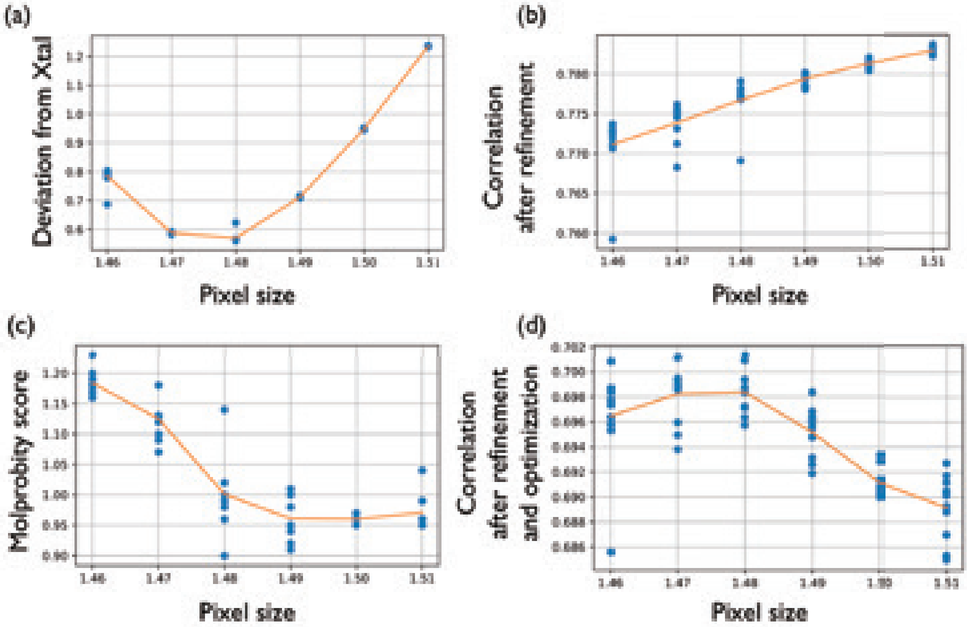

Figure 18.4: A various indicators were tested to find an indicator that can be used for calibration of the pixel size of cryo-EM density map during structure refinement procedures for a protein molecule. (a) The deviation of protein structure from the original conformation during refinement. At pixel size of 1.48, the struct is least deviated. (b) The value of CC (correlation coefficient), which is commonly used to assess the match between cryo-EM maps and protein structures. This indicator is not appropriate since it can be higher even with wrongly large pixel sizes. (c) Another commonly used indicator, Molprobity score. This indicator also cannot detect the error from wrongly large pixel size. (d) CC value after structure optimization. This indicator can identify the correct pixel value.

are developing data processing protocols to filter the data with good signals for reconstruction. These programs need to be improved further to allow the analysis of the anticipated large dataset, utilizing HPC.

We also plan to continue to improve the software for structure modeling from cryo-EM data. With the improvement of technology and algorithms, structural models with higher resolutions are being obtained. The current challenge in computation is how we ensure the quality and reliability of the resulting models. We plan to add new approaches to the current flexible fitting algorithm so that it can be applied to experimental data in a variety of qualities.

From these projects, new techniques for data analysis will become available, which will provide new structural information on biomolecules. Using the developed programs, we will work with experimental groups to obtain revolutionary structural information and contribute to the understanding of mechanisms of biological functions. We will maintain the new software and also provide usage support so that it is easily accessible to experimental groups from other institutions.

18.4 Publications

18.4.1 Journal Articles

1. M. Nakano, O. Miyashita, S. Jonic, A. Tokuhisa, F. Tama. Single-particle XFEL 3D Reconstruction of Ribosome-size Particles based on Fourier Slice Matching: Requirements to Reach Subnanometer Resolution. *J. Synchrotron Rad.* 2018 25:1010-1021

2. A. Srivastava, T. Hirota, S. Irle and F. Tama. Conformational dynamics of human Protein Kinase CK2 留 and its effect on function and inhibition. *Proteins.* 2018 86: 344-3

3. S.P. Tiwari, N. Reuter, Conservation of intrinsic dynamics in proteins-what have computational models taught us? *Curr Opin Struct Biol.* 2018 50:75-81