Quality Assessment and Biomolecular Structure Modeling for Cryo-EM using Deep Learning

Genki Terashi[1], Xiao Wang[1], Tsukasa Nakamura[1], Devashish Krishna Prasad[1], Daisuke Kihara[1]
Purdue University, West Lafayette
gterashi@purdue.edu

In recent years, an increasing number of protein and nucleotide structures have been modeled from cryo-electron microscopy (cryo-EM) maps. However, even though the EM map resolution has generally improved steadily over the past years, there are still many situations where modeling errors occur in high-resolution EM maps, or modelers face difficulties in modeling biomolecular structures due to locally low resolution in the map. To address such challenges, we have applied deep learning to three tasks: model quality assessment, protein structure modeling, and DNA/RNA structure modeling in cryo-EM maps. 1: Model Quality Assessment Modeling a protein structure into a cryo-EM map is a challenging task. One of the main difficulties is assigning the correct amino acids to their corresponding positions. Moreover, even with high-quality maps, there is always a risk of human error in the modeling process. To ensure the resulting atomic model is as accurate as possible, it's essential to perform rigorous validation using appropriate methods. To validate protein structure models in cryo-EM maps, our group developed a novel method based on the Deep-learning-based Amino-acid-wise model Quality (DAQ) score. In the DAQ score, the neural network detects specific map features for protein amino acid residue types, Cα atoms, and secondary structures, and computes the likelihood that each residue assignment is correct. By quantifying the incompatibilities between the protein model and the EM map at the amino acid level, the DAQ score provides a more accurate and sensitive measure of model quality compared to other methods [1]. Overall, the DAQ score offers a powerful tool for assessing protein structure models in EM maps and advancing cryo-EM research. The DAQ score can be computed on the Google Colab site (https://bit.ly/daq-score) or local machine by installing the code from (https://github.com/kiharalab/DAQ). Our group has also recently released the DAQ-Score Database [2] (https://daqdb.kiharalab.org/), which provides precomputed quality assessment results for protein models deposited in the Protein Data Bank (PDB) and their corresponding cryo-EM maps in the Electron Microscopy Data Bank (EMDB). Currently, the DAQ-Score Database contains over 152,129 protein chain models from 9,469 PDB entries derived from cryo-EM maps. In addition, the database provides the DAQ scores for multiple previous major versions of models if they exist. An example of a database entry is shown in Figure 1, which shows the first and revised version of the model for PDB ID: 7JSN Chain B (ID: 22458_7jsn_B_v1-1 and 22458_7jsn_B_v2-0). The computed DAQ scores are presented in a color code on a model within an interactive structure viewer, coupled with a graph showing the three DAQ score types along the residue sequence number. Figure 1 shows that the first model version (top) has regions that indicate negative DAQ scores (i.e., low quality). The corresponding regions in the revised model (bottom) were updated to positive DAQ scores, indicating substantial improvement 2: DeepMainmast: Protein structure modeling Protein structure modeling from a cryo-EM map is challenging, particularly when the resolution is worse than about 3 Å. To address this problem, we have developed an integrated protein structure modeling protocol called DeepMainmast [3]. This protocol employs a new de novo protein main-chain tracing method that uses deep learning to identify positions of Cα atoms and the types of amino acids. The core process of DeepMainmast employs an effective main-chain tracing approach, the Vehicle Routing Problem solver, and Constraint Problem Solver. Additionally, the protocol can accurately assign chain identity to the structure models of homo-multimers. To enhance the performance of the protocol, we also incorporate AlphaFold2 models when applicable. These models provide valuable information that can improve the accuracy of the resulting protein structures. Overall, DeepMainmast is a powerful tool that can help researchers overcome the challenges of protein structure modeling from cryo-EM maps. Our benchmarking results demonstrate that DeepMainmast substantially outperforms existing methods on the benchmark dataset. Compared to AlphaFold2, DeepMainmast achieves higher accuracy on a larger number of maps within the dataset, which consists of 178 high-resolution maps. Figure 2 shows an example of the modeling result for EMD-6551. DeepMainmast generated the accurate model with the correct chain ID assignment. The code is available at https://github.com/kiharalab/DeepMainMast. CryoREAD: DNA/RNA structure modeling Modelling the structure of DNA/RNA from cryo-EM maps is generally more challenging than protein structure modelling due to a number of factors. For example, DNA and RNA molecules can exhibit greater flexibility and variability in their structure than proteins, and available 3D structure data for DNA/RNA is significantly less than that of proteins. As a result, most biomolecular structure modeling software is primarily designed for proteins. To overcome this challenge, we have developed CryoREAD [4], which is a novel method for automated de novo DNA/RNA structure modeling from cryo-EM maps of a resolution range, between 2.0 Å to 5.0 Å. The method uses a deep neural network to identify the potential positions of phosphate, sugar, and bases, construct the backbone structure, map the nucleic acid sequence along the backbone, and construct a full atom model. Figure 3 illustrates the input EM map (EMD-12217), outputs of deep learning, and the final structure model with the native structure (PDB-ID:7BL4). Based on our benchmarking of 68 cryo-EM maps, on average, 84.9% of the atoms were correctly placed within a 5 Å, and 52.1% of nucleotides were correctly identified. The CryoREAD is available at https://github.com/kiharalab/CryoREAD.

References

{1} Genki Terashi, Xiao Wang, Sai Raghavendra Maddhuri Venkata Subramaniya, John JG Tesmer, and Daisuke Kihara. "Residue-wise local quality estimation for protein models from cryo-EM maps." Nature Methods 19(9) (2022): 1116-1125.
{2}. Tsukasa Nakamura, Xiao Wang, Genki Terashi and Daisuke Kihara. "DAQ-Score Database: Assessment of Map-Model Compatibility for Protein Structure Models from Cryo-EM Maps." In revision (2023).
{3}. Genki Terashi, Xiao Wang, Devashish Prasad, and Daisuke Kihara. "Integrated Protocol of Protein Structure Modeling for cryo-EM with Deep Learning and Structure Prediction." In submission (2023).
{4}. Xiao Wang, Genki Terashi and Daisuke Kihara. "De novo structure modeling for nucleic acids in cryo-EM maps using deep learning." In revision (2023).

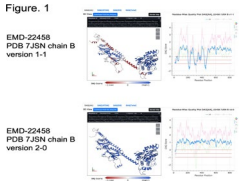

Figure 1

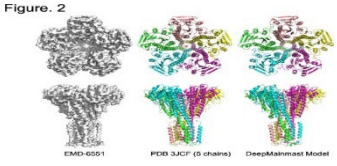

Figure 2

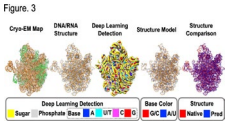

Figure 3