DAQ-Score Database: Deep-learning Based Quality Estimation of Cryo-EM Derived Protein Models

Tsukasa Nakamura[1], Xiao Wang[2], Genki Terashi[1], and Daisuke Kihara[1,2,*]

[1] Department of Biological Sciences, Purdue University, West Lafayette, Indiana, 47907, USA

[2] Department of Computer Science, Purdue University, West Lafayette, Indiana, 47907, USA

An increasing number of protein structures are determined by cryo-electron microscopy (cryo-EM) as cryo-EM has become one of the most important methods to determine structures. On the other hand, it has been noticed that errors occur in the model building process from cryo-EM maps, probably more frequently than one might think, particularly when the map resolution is not very high. Thus, establishing quality assessment methods has become a crucial and urgent task for biomolecular structure determination with cryo-EM.

We have recently developed a quality assessment method to detect protein structural model outliers using machine learning techniques. Our method, called DAQ (Deep-learning-based Amino acid-wise model Quality) score, uses deep neural network to capture local density features of amino acids and atoms in proteins and assesses the likelihood that modeled residues in a structural model are correct (Terashi *et al.*, *Nature Methods*, 2022). DAQ is also able to detect not only errors in conformations but also shifts in sequence assignment to otherwise correct main-chain conformations, which is often not easy to detect by checking density fitting.

Here, we performed a PDB-scale model analysis by DAQ. We applied DAQ to around 10,000 protein structure models in PDB that were derived from cryo-EM maps deposited in Electron Microscopy Data Bank (EMDB). We report the tendency of common errors made in the models through the large-scale analysis. When authors deposited updated structure models to PDB over an initial model, we see clear improvement of DAQ score in the updated version of the model. A common type of errors observed includes sequence shifts along alpha helices. Model assessment results with DAQ are made available in a database (https://daqdb.kiharalab.org) (Nakamura *et al.*, *Nature Methods*, 2023). The DAQ score can be computed on the Google Colab site (https://bit.ly/daq-score) or local machine (https://github.com/kiharalab/DAQ).