Protein Structure Prediction in a Post-AlphaFold2 World

Presented by *Mohammed AlQuraishi* on February 28, 2022
Hosted by UT Southwestern Medical Center via Zoom

Dr. Mohammed AlQuraishi serves as an Assistant Professor in the Department of Systems Biology at Columbia University. He is also a member of Columbia's Program for Mathematical Genomics which engages in research involving machine learning, biophysics, and systems biology. The primary research interests of his lab include the development of machine learning models for the prediction of protein structure and function, learned representations of proteins and proteomes, as well as protein-ligand interactions. Using these models, the lab investigates the organization and computational paradigms of signal transduction networks, how these networks vary in the human population, and how different diseases, particularly cancer, result in their dysregulation.

Dr. AlQuraishi's seminar primarily focused on the applications and limitations of various tools for protein structure prediction in context of the advancements made in the development of DeepMind's AlphaFold2 - an AI system that predicts a protein's 3D structure based on multiple sequence alignments of the protein. Initially, he discussed the improvement in the quality of protein structure prediction (based on the Free Modelling results from CASP - a biannual competition where different prediction algorithms are compared) through the previous decade, noting the leaps in progress made by the original AlphaFold in 2018 and subsequently by its successor, AlphaFold2 in 2020.

Next, he discussed some of the features of AlphaFold2 as well as some of its limitations. AlphaFold2 has a high level of accuracy (median accuracy of 2.1Å) when predicting the structures of single domain Apo proteins (up to 2000 residues in size). However, it appears to struggle with predictions for multi-domain proteins, potentially due to their flexibility (ability to change their conformations). Currently, it hasn't been tested with predicting mutations and synthetically designed proteins. Also, it is currently unable to handle complexes, single sequences, ligands, co-factors, etc.

Later in the seminar, he described several other algorithms for protein structure prediction, and their strengths and weaknesses, mainly focusing on the relationship between prediction accuracy and sequence information. In this discussion, he described *Rosetta* - a relatively inaccurate algorithm for predicting the structure of single sequences, and *RaptorX* - a more accurate algorithm that is able to account for co-evolution. Dr. AlQuraishi also described *RGN*, a method he developed that worked with data from Position Specific Scoring Matrices (PSSM) and utilized differentiable neural network components which allowed for greater optimization of the predictions. *RGN2* improved upon the original *RGN*, and has been found to outperform AlphaFold2 - it has faster execution times and requires fewer parameters. He concludes the presentation by discussing various use cases for AlphaFold2 in better understanding the effects of genetic variants, rapidly evolving viruses, and protein design.

In summary, the seminar discussed the application of various deep learning methods in solving the protein structure prediction problem. Recent innovations in deep learning and artificial intelligence have allowed for the development of more accurate and efficient tools for protein modeling that no longer rely on physics-based models. These tools, such as *RGN2* and *AlphaFold2* will allow for significant leaps in our understanding of molecular interactions.

Question: Could AlphaFold2 be used to predict the structure of proteins with multiple chains using a polyglycerin linker to connect those chains.

Dr. AlQuraishi's reply: People have already been using AlphaFold2 unofficially to do that, since

it was not trained for that purpose. I believe that a glycine linker wouldn't even be needed since it has a positional encoding of where the residues are in the protein. However, this encoding is fairly limited, so it only tells you how far apart two residues are, up to 32 positions. Anything past that it doesn't know, but it is still able to use this information to create the structure for a single sequence. It is able to deal with protein complexes in a very simple way though, since two proteins that are an infinite distance away is the same as them being 32 positions away.