Genome-scale structural prediction of protein sequences and complexes with deep learning

Authors: Mu Gao[1]* (mu.gao@gatech.edu), Davi N. An[1], Mark Coletti[2], Russell B. Davidson[2], Jerry M. Parks[2], Jianlin Cheng[3], Ada Sedova[2], and **Jeffrey Skolnick**[1]

Institutions: [1]Georgia Institute of Technology, Atlanta, GA; [2]Oak Ridge National Laboratory, Oak Ridge, TN; [3]University of Missouri, Columbia, MO.

Project Goals: With the advances in next generation sequencing technologies, the number of sequenced genomes is growing exponentially. This has resulted in a bottleneck for the translation of sequence information into functional hypotheses about each gene. Current gene annotation technologies are primarily based on evolutionary inference by sequence comparison; however, many proteins in a proteome remain uncharacterized. To address this challenge, this collaborative team is currently developing a suite of novel high-performance-computing (HPC), deep-learning methods that predict protein structures at unprecedented accuracy, making use of the Summit supercomputer at the DOE leadership computing facility at the Oak Ridge National Laboratory. The combination of deep learning, HPC, and structural-based analysis will help break the gene annotation bottleneck and enable rapid, accurate prediction of gene function on a genomic scale.

Abstract text: One key aspect of protein annotation is the atomic structure encoded in a protein sequence. The release of AlphaFold2 in July 2021 has provided a powerful deep learning based computational method to decode protein sequences by predicting high confidence structural models of individual proteins. Taking advantage of this advance, our team has developed a genome-scale protein structural modeling and analysis pipeline using AlphaFold2 and deployed this workflow successfully on several full proteomes on Summit. This workflow has been applied to a few bacterial and plant species of interests to DOE's Office of Biological and Environmental Science. One of them is *Smagellanicum*, whose proteome consists of about 25,227 protein sequences with 11 million total amino acids. We have modeled 25,134 (99.7%) of the proteome and about 57% of sequences have at least one high confidence model.

Moreover, interactions between proteins are vital to the understanding of their functions. Excitingly, we have developed AF2Complex, a generalization of AlphaFold2 for predicting physical interactions between different proteins via deep docking, i.e., by exploring physically favorable structural models of a putative multimeric protein complex with the same deep learning neural networks originally developed for modeling a single protein sequence. By incorporating a new evaluation metric and optimizing input data, AF2Complex can effectively predict if a set of protein sequences interact, and if so, then provide high-confidence models for the predicted protein complex. In a benchmark test on dimeric protein pairs, it achieves higher accuracy than strategies that combine AlphaFold2 and protein-protein docking. It also achieved significantly better performance than DeepMind's AF-multimer when the same set of deep learning models of AF-multimer are used. Importantly, going beyond most other approaches that focus on dimeric protein pairs, we demonstrate that a protein complex consisting of multiple proteins can be accurately

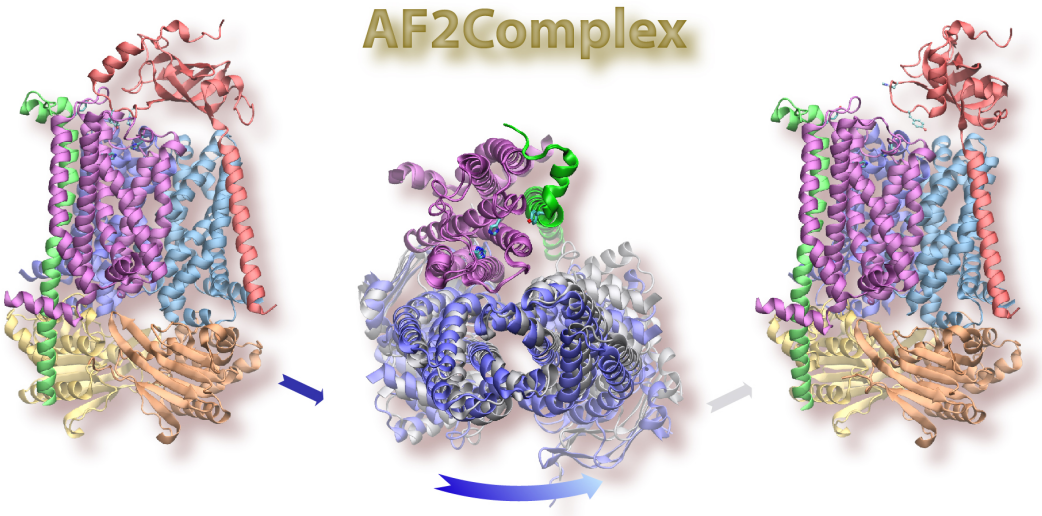

Figure 1. AF2Complex models of the cytochrome c maturation system $CcmA_2B_2CD$ from E. coli. Conformational changes relevant to the function of the system are observed in the superimposition (center) top two models (left and right). The superimposition of the two models uses one component (CcmE in purple) as the reference and is viewed from a perspective that is above the two individual models. For clarity, CcmF (red) is not shown in the superposition.

modeled using this deep learning approach. AF2Complex was successfully validated on some challenging CASP14 multimeric targets, a small but appropriate benchmark set, and the *E. coli* proteome. In a practical application, using the cytochrome *c* biogenesis system as an example, we predict high-confidence models of three sought-after assemblies formed by eight members of this system (the predicted models of one complex assembly are shown in Figure 1 above). To the best of our knowledge, this is the first time that a deep learning algorithm has been applied to model the atomic structures of a whole biomolecular system.

References/Publications

1. Gao, M., Lund-Andersen P, Morehead A, Mahmud S, Chen C, Chen X, Giri N, Roy R S, Quadir F, Effler T C, Prout R, Abraham S, Skolnick J, Cheng J, Sedova A. *High-Performance Deep Learning Toolbox for Genome-Scale Prediction of Protein Structure and Function.* In Machine Learning in HPC Environments, held in conjunction with The International Conference for High Performance Computing, Networking, Storage, and Analysis. 2021. St. Louis.

2. Gao, M, An D, Parks, P, and J. Skolnick. *Predicting direct physical interactions in multimeric proteins.* bioRxiv, 2021: p. 2021.11.09.467949.

Funding statement: This research was supported by the DOE Office of Science, Office of Biological and Environmental Research (BER), grant no. DE-SC0021303.