Integrated deep-learning computational approach to proteome annotation

Authors: Mu Gao[1]* (mu.gao@gatech.edu), Jianlin Cheng[2], Jerry Parks[3], Dwayne Elias[3], Ada Sedova[3], and **Jeffrey Skolnick**[1]

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

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 infer protein structure information 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: The ability to predict the structure and function of a protein-coding gene from its sequence is a grand challenge in biology. Advances in next generation sequencing technologies have led to an exponential increase in the size of genomic datasets. Genome functional annotation, the assignment of validated molecular functions to the majority of the protein coding genes in a genome, has been challenged by these massive datasets. Experimental methods offer the gold standard for proof-of-function, but even high-throughput experiments are orders of magnitude too slow compared with the speed at which gene-sequencing big-data is generated, causing a new technology bottleneck. With growing computing power and the success of advanced machine-learning analyses, computational methods could help eliminate this bottleneck by accurately inferring protein function and thereby providing experimentally testable hypotheses.

Current computational technologies to infer function are often based on evolutionary inference through sequence comparison with known, annotated protein sequences. However, these methods fail when the sequence similarity is low, e.g., at 30% sequence identity. For many organisms, this may represent a significant portion of the genome. To tackle this important issue, our team is developing a suite of novel high-performance computing (HPC) and deep learning-based computational methods that predict structural information, and then apply it to help predict the function of proteins with low sequence similarity to any known annotated protein. Using deep learning for the prediction of the structure of these low-similarity proteins has recently achieved some dramatic breakthroughs. The inclusion of structure-related information from these predictions can help to fill in the knowledge gaps for this proteomic “dark matter.”

There are a number of ways that deep-learning inspired structural inference can help infer function. One such method is SAdLSA, which is trained to conduct sequence alignment from deep-learning protein structural alignments [1]. The implicit structural information used by deep learning reveals structural similarities not apparent from standard sequence comparison. SAdLSA has been deployed on the Summit supercomputer and can use hundreds of GPUs to search for hidden matches to experimentally deter crystal structures. We applied SAdLSA to 559 uncharacterized protein sequences in *Desulfovibrio vulgaris*, a model organism for sulfur-reducing bacteria. Scanning a large sequence library of 83,000 sequences, the pilot runs of SAdLSA on *Desulfovibrio vulgaris* found some significant hits for over 25% of these uncharacterized sequences. Preliminary analysis on just a few sequences has already revealed interesting predictions. For example, one bacterial protein's top structural match points to human PHPT1, a eukaryotic phosphohistidine phosphatase with no known prokaryotic counterpart. Moreover, the sequence alignment prediction is corroborated by the MULTICOM2 method, discussed below.

Predicting the full three-dimensional structure of a protein can provide a wealth of essential information derived from analyzing this structure using it to model binding interactions. The multi-task deep learning method DeepDist [2] is based on residual convolutional neural networks and predicts inter-residue distances from protein sequences via both regression and multi-classification. DeepDist was used to predict inter-residue distances for template-free (*ab initio*) protein structure prediction in the latest version of the comprehensive protein structure prediction system, MULTICOM2, which was ranked 7th out of 146 predictors in the tertiary structure prediction and 3rd out of 136 predictors in the inter-domain structure prediction in the CASP14 experiment. We are currently deploying several of MULTICOM2's tasks on Summit in order to train on larger datasets for even more accurate models. After a structure is predicted, fold analysis, and binding pocket analysis is performed, and molecular interactions can be predicted and modeled with simulations. This provides more information about the protein's function, and its placement in metabolic and interaction networks. Together the methods and applications we are developing on leadership computing resources will help provide a new generation of solutions to help break the genome functional annotation bottleneck.

References/Publications

1. Gao, M. and J. Skolnick, *A novel sequence alignment algorithm based on deep learning of the protein folding code.* Bioinformatics, 2020.
2. Wu, T., Z. Guo, J. Hou, and J. Cheng, *DeepDist: real-value inter-residue distance prediction with deep residual convolutional network.* bioRxiv, 2020: p. 2020.03.17.995910.

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