Title: Learning Protein Fitness Models from Evolutionary and Experimental Data

Authors: Chloe Hsu (chloehsu@berkeley.edu),[1]*, Hunter Nisonoff,[1] Clara Fannjiang,[1] and **Jennifer Listgarten**[1]

Institutions: [1]University of California, Berkeley

Website URL: https://sc-programs.llnl.gov/biological-and-environmental-research-at-llnl/secure-biosystems-design

Project Goals: Many protein design applications, including synthetic gene entanglement designs, rely on protein fitness models that predict protein function based on their amino acid sequence. This project aims to reduce the amount of data that the model requires to make reliable functional predictions for a protein by including sequences of evolutionarily related proteins as additional input.

Abstract Text:

There are several approaches to predict functional properties of a given protein from the protein's amino acid sequence. Existing machine learning-based models of protein fitness typically learn from either unlabeled, evolutionarily related sequences or variant sequences with experimentally measured labels. To reduce the amount of data that the model requires to make reliable functional predictions for a protein, recent work has suggested methods for combining both sources of information including evolutionary and experimental data.

Toward that goal, we propose a simple combination approach that is competitive with, and on average outperforms more sophisticated methods. Our approach uses ridge regression on site-specific amino acid features combined with a probability density feature from modeling the evolutionary data. Within this approach, we find that a variational autoencoder-based probability density model showed the best overall performance regardless which evolutionary density model was used. Moreover, our analysis highlights the importance of systematic evaluation and sufficient baseline. In addition to evolutionary and assay-labeled data, we also demonstrate that our combination approach can be extended to include protein structure information to further improve fitness prediction.

References/Publications

1. Hsu, C., Nisonoff, H., Fannjiang, C. and Listgarten, J.. " Learning protein fitness models from evolutionary and assay-labeled data." in press, Nature Biotechnology

Funding Statement:
This work is supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Lawrence Livermore National Laboratory Secure Biosystems Design SFA "From Sequence to Cell to Population: Secure and Robust Biosystems Design for Environmental Microorganisms". Support was also provided by the Chan Zuckerberg

Investigator program, C3.ai, the National Library of Medicine of the National Institutes of Health, and the National Science Foundation Graduate Research Fellowship Program.