Application of ESMFold and AlphaFold for Prediction of Protein Stability Upon Mutation

Jack Ringer

University of New Mexico, Department of Computer Science

CENTER FOR ADVANCED RESEARCH COMPUTING

Background and Overview

Background

Single-point mutations occur when a single nucleotide base pair is swapped with another in the DNA or RNA sequence of an organism's genome. Although many point mutations are benign, they can cause functional changes in proteins which can lead to disease and other serious health issues [5].

Within this work we look at using various computational methods in order to predict the change in protein stability upon point-mutation. The change in Gibbs free energy ($\Delta\Delta G$, measured in kcal/mol) between a wildtype and mutant protein ($\Delta\Delta G = \Delta G_{wildtype} - \Delta G_{mutant}$) is used to measure stability.

Why is this important?

- Determining the effects of point mutations on protein stability can provide insight into the causes of disease and inform drug development [6]
- Computational methods can dramatically reduce the expense of determining mutational effects in comparison to in vitro methods

Overview

This work investigates how predictions from ESMFold [3] compare to AlphaFold [2] in predicting protein stability change ($\Delta\Delta G$) for mutant proteins.

- DDGun [4] and ACDCN-NN [1] are employed to generate $\Delta\Delta G$ values from predicted structures.
- Use sequence-based predictions as a baseline
- Predicted structures of mutants from ESMFold are used to explore potential improvement of $\Delta\Delta G$ predictions.

Dataset

Our dataset of $\Delta\Delta G$ experiments was assembled by combining entries from FireProtDB [7] and ThermoMutDB [8]. Various methodologies were employed to clean/filter data and ensure quality, including:

- Filtering entries missing critical information
- Removing duplicate entries
- Removing entries with inconsistent mutation codes

In addition, given the computational expense of ESMFold and AlphaFold we restrict our dataset to only include proteins with 400 or less residues.

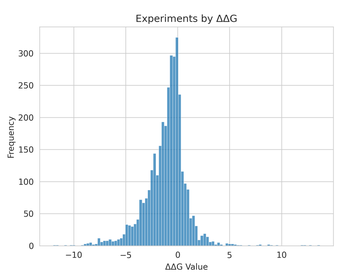

Figure 1. Histogram of experimental $\Delta\Delta G$ values in the final dataset

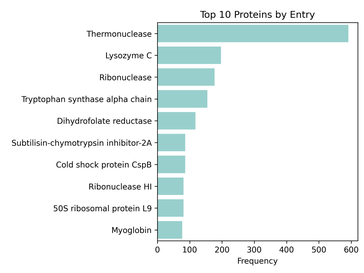

Figure 2. Top 10 most common wild-type proteins represented in the final dataset.

Methods

Structure Prediction

AlphaFold and ESMFold are applied to predict the structures for each of the 184 wild-type proteins in our dataset. In addition to predicting the structure of wild-type proteins, ESMFold is used to predict structures of mutants. ESMFold is generally considered less accurate than AlphaFold, but does not rely on computing multiple-sequence alignments (MSAs) and as a result is significantly faster.

Stability Prediction

ACDC-NN and DDGun are used to estimate $\Delta\Delta G$ values for single-point mutations.

- ACDC-NN is a data-driven approach using neural network architectures
- DDGun is an untrained method that uses evolutionary and statistical potentials
- Both of these methods are capable of taking in a protein's wild-type structure, evolutionary information, and the mutation information to generate a predicted $\Delta\Delta G$ value.

Results

An overview of results are shown in Table 1 and Figure 3. Here the Pearson correlation coefficient (r), mean-absolute error (MAE), and root-mean square error (RMSE) are used to measure the quality of predicted $\Delta\Delta G$ values against ground-truth (i.e., experimental) values.

Method	Structures From	r	MAE	RMSE
DDGun-Seq	-	0.4084	1.3178	1.9211
DDGun	AlphaFold	0.4617	1.2489	1.8631
DDGun	ESMFold	0.4611	1.2495	1.8636
ACDC-NN-Seq	-	0.4593	1.2406	1.8547
ACDC-NN	AlphaFold	0.4754	1.2299	1.8452
ACDC-NN	ESMFold	0.4765	1.2307	1.8442
ACDCN-NN-Mut	ESMFold	0.4721	1.2457	1.8606

Table 1. Assessment of each methodology on our dataset.

In general $\Delta\Delta G$ predictions were compressed around 0, with all methodologies struggling to make predictions for larger experimental values of $|\Delta\Delta G|$.

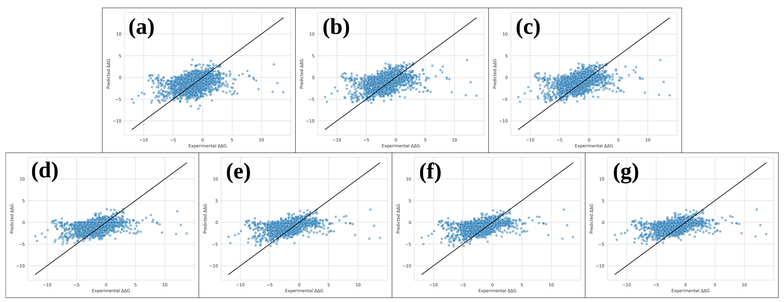

Figure 3. Parity plots showing experimental (x-axis) vs. predicted (y-axis) values of $\Delta\Delta G$. Labels for each plot correspond to methodology as follows: **(a)** DDGun-Seq **(b)** DDGun using AlphaFold wild-type structures **(c)** DDGun using ESMFold wild-type structures **(d)** ACDC-NN-Seq **(e)** ACDC-NN using AlphaFold wild-type structures **(f)** ACDCN-NN using ESMFold wild-type structures **(g)** ACDCN-NN using ESMFold wild-type and mutant structures

Discussion

- Computational structures can be applied towards protein stability prediction using existing methods
- Little difference in stability prediction accuracy when using structures from AlphaFold or ESMFold
- Inclusion of mutant structures from ESMFold did not improve performance
- Structure confidence scores (pLDDTs) were not found to correlate with errors of $\Delta\Delta G$ predictions (results not shown)
- Both DDGun and ACDC-NN biased towards predicting little change in stability
- Future work may include improving theoretical understanding of protein stability and/or constructing model architectures that are able to learn more from relatively smaller datasets

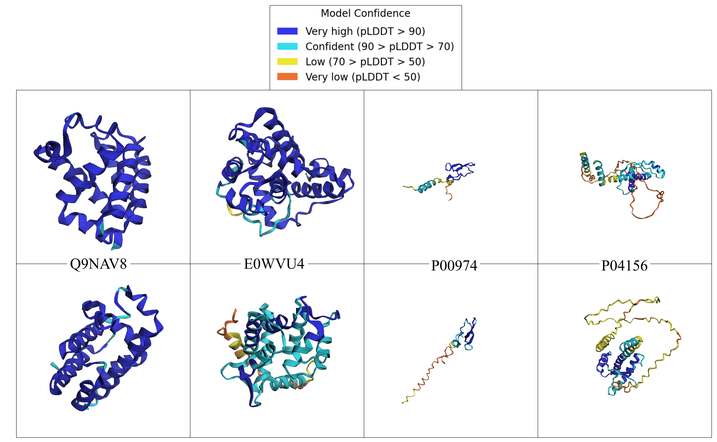

Figure 4. Sample structure predictions from AlphaFold (top row) and ESMFold (bottom row).

Acknowledgement

Much thanks to my advisor Vincent Metzger for his guidance and input over the course of this project. We'd also like to acknowledge the UNM Center for Advanced Research Computing (CARC) for allowing us to use computational resources for several aspects within this project.

References

[1] S Benevenuta, C Pancotti, P Fariselli, G Birolo, and T Sanavia.
An antisymmetric neural network to predict free energy changes in protein variants.
Journal of Physics D: Applied Physics, 54(24):245403, 2021.

[2] John M. Jumper et al.
Highly accurate protein structure prediction with alphafold.
Nature, 596:583 – 589, 2021.

[3] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Sal Candido, et al.
Language models of protein sequences at the scale of evolution enable accurate structure prediction.
bioRxiv, 2022.

[4] Ludovica Montanucci, Emidio Capriotti, Yotam Frank, Nir Ben-Tal, and Piero Fariselli.
Ddgun: an untrained method for the prediction of protein stability changes upon single and multiple point variations.
BMC Bioinformatics, 20, 2019.

[5] Medline Plus.
How can gene mutations affect health and development?: Medlineplus genetics.
`https://medlineplus.gov/genetics/understanding/mutationsanddisorders/mutationscausedisease/`, Mar 2021.

[6] Castrense Savojardo, Piero Fariselli, Pier Luigi Martelli, and Rita Casadio.
Inps-md: a web server to predict stability of protein variants from sequence and structure.
Bioinformatics, 32(16):2542–2544, Apr 2016.

[7] Jan Stourac et al.
Fireprotdb: database of manually curated protein stability data.
Nucleic Acids Research, 49:D319 – D324, 2020.

[8] Joicymara S. Xavier et al.
Thermomutdb: a thermodynamic database for missense mutations.
Nucleic Acids Research, 49:D475 – D479, 2020.