Augmenting Evolutionary Models with Structure-based Retrieval

Yining Huang†*
Biomedical Informatics
Harvard Medical School

Zuobai Zhang†*
Mila - Quebéc AI Institute
Université de Montréal

Jian Tang
Mila - Quebéc AI Institute
HEC Montréal

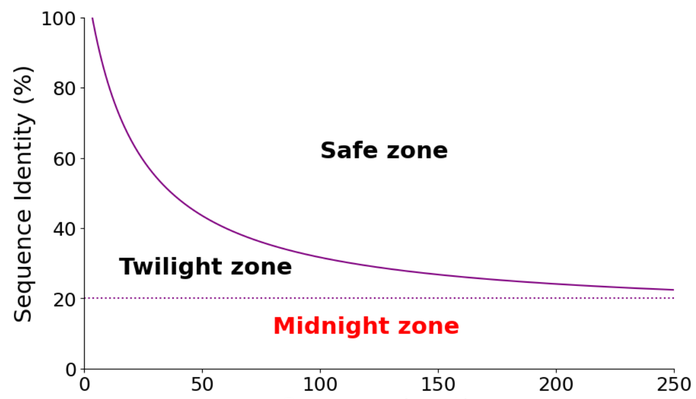
Debora S. Marks
Harvard Medical School
Broad Institute

Pascal Notin*
System Biology
Harvard Medical School

*† Equal contribution; * Correspondence: yininghuang@hms.harvard.edu, zuobai.zhang@mila.quebec, pascal_notin@hms.harvard.edu*

Background

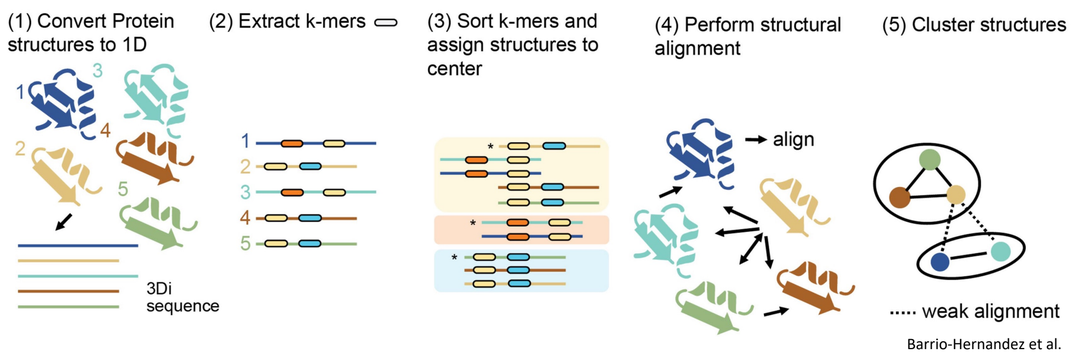
- Multiple Sequence Alignment (MSA) is a fundamental tool for identifying homologous proteins sharing a common evolutionary origin.
- Existing Multiple Sequence Alignment (MSA) search tools retrieve homologous protein sequences from protein sequence databases based on sequence similarity.
- However, many homologous proteins have high structure and functional similarity but low sequence similarity.
- Structure similarity search tools are necessary to recover homologous proteins in the 'midnight zone.'



Multiple Structure Alignment (MStructA)

Searching

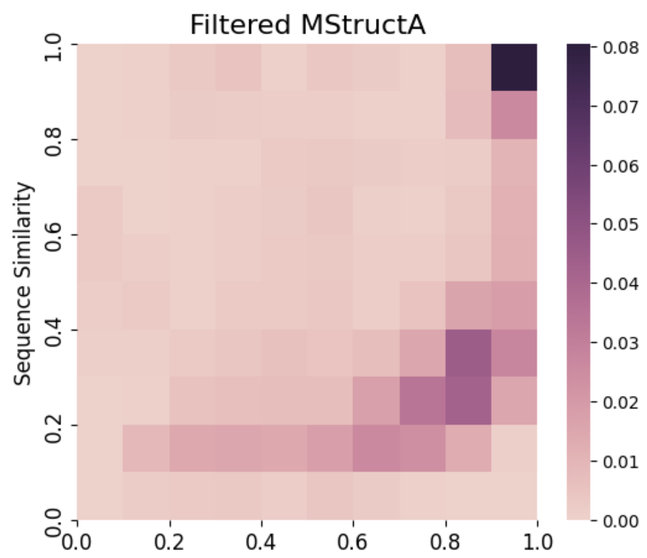
- MStructA is constructed by retrieving and aligning homologous proteins from large protein structure databases using structure similarity search tools.
- We constructed 197 MStructA for each wild-type protein in ProteinGym by using Foldseek to search for structurally similar proteins within the AlphaFold Database.



Filtering

- To ensure MStructA contains high-quality structural homologs, we filter sequences recovered by Foldseek following similar practices of previous alignment-based models.
- The final MStructA only keeps sequences with Sequence Identity > 0.1, E-value < 1e-10, and Gaps in Sequence < 50%.

Complementary Effect of Multiple Structure Alignment on MSA



- After combining MStructA with MSA, some alignments show more than a 200% depth increase.
- Most proteins identified by MStructA exhibit high structural similarity but low sequence similarity to target proteins, indicating MStructA effectively supplements the MSA by identifying previously undetected structural homologs.

Experiment

- **Task:** Protein fitness prediction in ProteinGym
- **Dataset:** We randomly selected 30 ProteinGym DMS assays of varying properties with more than 5% depth increase when MStructA is combined with MSA.
- **Model:** We choose EVE, a zero-shot alignment-based model that has shown good performance using multiple sequence alignments (MSA) for protein fitness prediction.
- **Training:** For each assay, we train an EVE model on the combined MSA and MStructA.

Performance

	Spearman	AUC
EVE (MSA)	0.434	0.737
EVE (MSA+MStructA)	**0.443**	**0.742**
% Assays Improved	60%	61%

- Our method of training EVE with combined MSA and MStructA outperforms the original EVE with only MSA on average Spearman correlation and Area Under the ROC Curve (AUC) for each assay.

Greater Improvement on Low MSA Depth and Binding Function Type Assays

Low MSA-Depth Assays

- The low-depth MSAs may not contain enough structurally informative sequences for the model to effectively capture the constraints that maintain protein fitness. The MStructAs complement these MSAs by providing additional structurally similar protein sequences.

Original MSA Depth	EVE (MSA)	EVE (MSA+MStructA)	Improvement
Low	0.497	0.534	**0.037**
Medium	0.425	0.424	-0.001
High	0.349	0.363	0.014

Binding Function Type

- Binding function inherently relies on structural information, implying that MSA built on sequence similarity alone might not fully capture the structural context essential to protein functions involving complex structural interactions.

Function Type	EVE (MSA)	EVE (MSA+MStructA)	Improvement
Activity	0.436	0.441	0.005
Binding	0.444	0.510	**0.066**
Expression	0.450	0.462	0.012
OrganismalFitness	0.426	0.431	0.005

Discussion and Future Work

MStructA Quality

- Future work will focus on optimizing the filtering pipeline to balance the quality of included sequences with the number of retrieved sequences to provide optimal structural information gain.

Benchmarking

- We will conduct comprehensive benchmarking for all assays in ProteinGym once we develop a better MStructA construction pipeline.
- We aim to develop unified protein retrieval packages to conduct combined sequence-based and structure-based search.

Reference

- Barrio-Hernandez, I., Yeo, J., Jänes, J. et al. Clustering predicted structures at the scale of the known protein universe. Nature 622, 637–645 (2023).
- Frazer, J., Notin, P., Dias, M., Gomez, A., Min, J. K., Brock, K., Gal, Y., and Marks, D. S. Disease variant prediction with deep generative models of evolutionary data. Nature, 599(7883):91–95, Nov 2021.
- Notin, P., Kollasch, A., Ritter, D., van Niekerk, L., Paul, S., Spinner, H., Rollins, N., Shaw, A., Orenbuch, R., Weitzman, R., Frazer, J., Dias, M., Franceschi, D., Gal, Y., and Marks, D. Proteingym: Large-scale benchmarks for protein fitness prediction and design. In Advances in Neural Information Processing Systems, 2023.
- van Kempen, M., Kim, S.S., Tumescheit, C. et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol 42, 243–246 (2024).