Identifying Disordered Proteins using AlphaFold2 and ESMFold

Amrit Rajbhandari (arajbhan@uno.edu), Md Wasi Ul Kabir (mkabir3@uno.edu), Md Tamjidul Hoque (thoque@uno.edu)
Department of Computer Science, University of New Orleans, New Orleans, LA, USA

Abstract

Disordered proteins are a class of proteins that lack a well-defined three-dimensional structure under physiological conditions. Unlike structured proteins, which typically have a specific shape that is critical to their function, disordered proteins are highly flexible and can adopt a wide range of conformations. They are also sometimes referred to as intrinsically disordered proteins (IDPs) or intrinsically unstructured proteins (IUPs). Disordered proteins are present in all domains of life, and they play critical roles in various cellular processes, including signaling, transcription, translation, and regulation of protein-protein interactions. They are also involved in a number of diseases, including cancer, neurodegenerative disorders, and infectious diseases. Recently, the development of AlphaFold2 and ESMFold marked a paradigm shift in the structural biology community. AlphaFold2 and ESMFold are neural network-based models that predict protein structures only from amino acid sequences. ESMfold adapts the same AlphaFold2 architecture but removes AlphaFold2 dependence on MSAs with a protein language model. Compared to AlphaFold2, ESMFold predicts protein structure significantly faster. Successes of these two state-of-the-art methods (AlphaFold2 and ESMfold) motivates us to assess their ability for disordered protein prediction.

Structure and disordered proteins

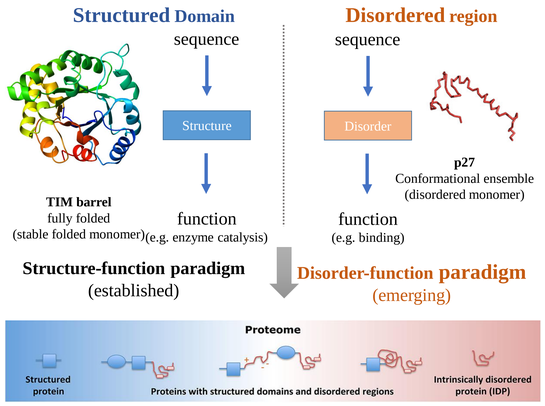
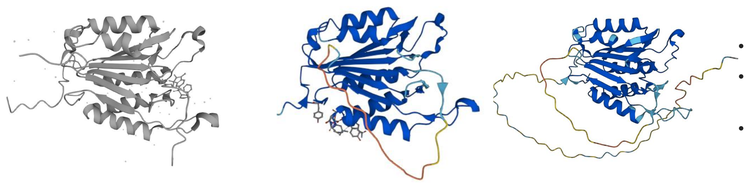


Image Source: https://pubs.acs.org/doi/10.1021/cr400525m

Experimental Vs. Prediction of Alphafold2 Vs. ESMFold

Caspase-3: Thiol protease that acts as a major effector caspase involved in the execution phase of apoptosis. Following cleavage and activation by initiator caspases (CASP8, CASP9 and/or CASP10), mediates execution of apoptosis by catalyzing cleavage of many proteins.



X-ray crystallography-based structural model | Alphafold2 prediction | ESMfold prediction

https://www.uniprot.org/uniprotkb/P42574/feature-viewer
https://alphafold.ebi.ac.uk/entry/P42574
https://esmatlas.com/resources/detail/MGYP002476601952

Discussion

- In this study, we propose a method for disordered region prediction using the predicted protein structure of two state-of-the-art methods, EMSFold and AlphaFold2.
- The EMSFold and AlphaFold2 prediction accuracy is very good, and these methods not only predict the structure but also provide confidence scores for each residue.
- The confidence score correlates with the disordered regions.
- We will extract features from the protein structures and use the confidence score for disordered prediction.
- As the protein sequence is similar to Natural Language Processing (NLP) tasks, we plan to implement an attention-based transformer model to predict the disordered regions.

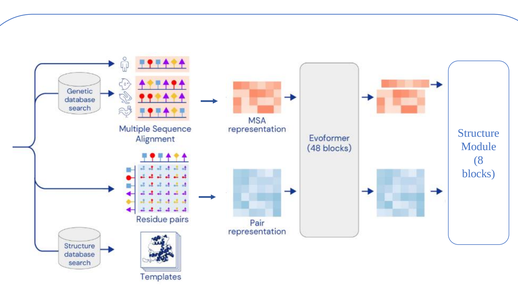
Proposed Method

Protein sequence

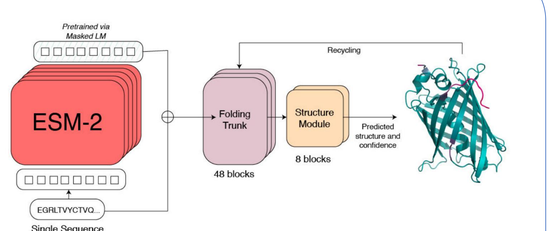
MTDVET…..

Input

AlphaFold2 Model



ESMFold Model



Output

Features

The AlphaFold network directly predicts the 3D coordinates of all heavy atoms for a given protein using the primary amino acid sequence and aligned sequences of homologues as inputs

	Atom coordinates			
M	$M_1(x_1,y_1,z_1)$	$M_2(x_2,y_2,z_2)$	--	$M_{37}(x_{37},y_{37},z_{37})$
T	$T_1(x_1,y_1,z_1)$	$T_2(x_2,y_2,z_2)$	--	$T_{37}(x_{37},y_{37},z_{37})$
--	--	--	--	--
--	--	--	--	--

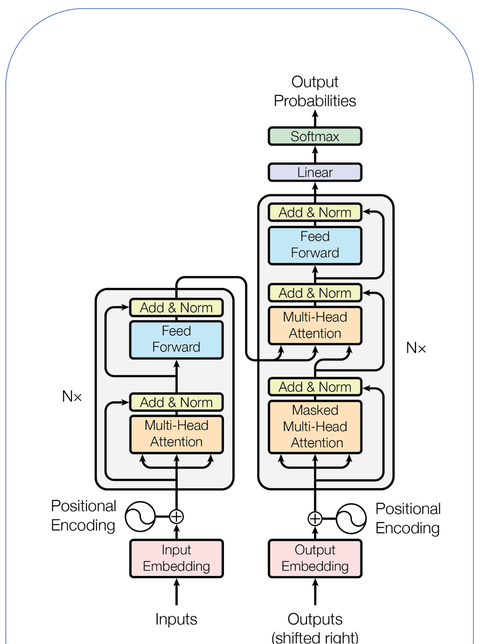
Per-residue confidence metric called predicted local distance difference test (pLDDT) on a scale from 0 to 100. pLDDT estimates how well the prediction would agree with an experimental structure based on the local distance difference test Cα (lDDT-Cα).

In CASP9, the local Distance Difference Test (lDDT) score was introduced, assessing how well local atomic interactions in the reference protein structure are reproduced in the prediction.

(The Critical Assessment of protein Structure Prediction (CASP) is a community-wide prediction competition where research groups are required to predict 3-D structures from protein sequences). [source: https://predictioncenter.org/]

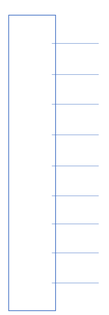
Input

A transformer is a deep learning model that adopts the mechanism of self-attention, differentially weighting the significance of each part of the input data. It is used primarily in the fields of natural language processing (NLP) and computer vision (CV).



Output

Propensity Scores



1. Lin, Zeming, et al. Language Models of Protein Sequences at the Scale of Evolution Enable Accurate Structure Prediction. bioRxiv, 21 July 2022. bioRxiv, https://doi.org/10.1101/2022.07.20.500902.
2. Jumper, John, et al. "Highly Accurate Protein Structure Prediction with AlphaFold." *Nature*, vol. 596, no. 7873, Aug. 2021, pp. 583–89. *www.nature.com*, https://doi.org/10.1038/s41586-021-03819-2.
3. Vaswani, Ashish, et al. *Attention Is All You Need*. arXiv, 5 Dec. 2017. *arXiv.org*, https://doi.org/10.48550/arXiv.1706.03762.