BITS :: Call for Abstracts 2024 - Oral communication

Type	Oral communication
Session	Bioinformatics AI, Models and Tools
Title	Alpha&ESMhFolds: a web server for the comparison of 42,942 AlphaFold2 and ESMFold models of the human reference proteome
All Authors	Matteo Manfredi (1), Castrense Savojardo (1), Pier Luigi Martelli (1), Rita Casadio (1)

Affiliation

(1) Bologna Biocomputing Group, University of Bologna

Motivation

Results from CASP15 confirm the relevance of AI-based modelling on the accuracy of protein structure prediction. The best performances are reported by methods differently based on DeepMind's AlphaFold2. As an alternative, methods such as ESMFold take advantage of Protein Language Models, allowing for faster computations. To compare the two approaches, we develop a novel database storing AlphaFold2 and ESMFold models for 42,942 proteins covering the Human Reference Proteome, alongside 2,900 experimental structures from the PDB.

Methods

Proteins adopted in this study come from the human Reference Proteome, available at UniProt. From the initial set, we exclude fragments, short peptides, sequences for which AlphaFold2 models were not available in AlphaFoldDB, and sequences for which ESMFold failed to generate a model. We end up with 42,942 protein sequences. We then extract structural data from the PDB. We exclude all structures with coverage <70% to the UniProt sequence, retaining the best PDB for 2,900 proteins.
Both AlphaFold2 and ESMFold compute for each residue a pLDDT value (ranging from 0 to 100) that provides an estimate of the quality of the prediction. We evaluate the quality of each model by computing the percentage of residues with pLDDT ≥70. Additionally, to evaluate the structural similarity, we adopt Foldseek to superimpose paired models and to superimpose each model to the associated PDB structure. This produces a TM-score (ranging from 0 to 1) for each superimposition.

Results

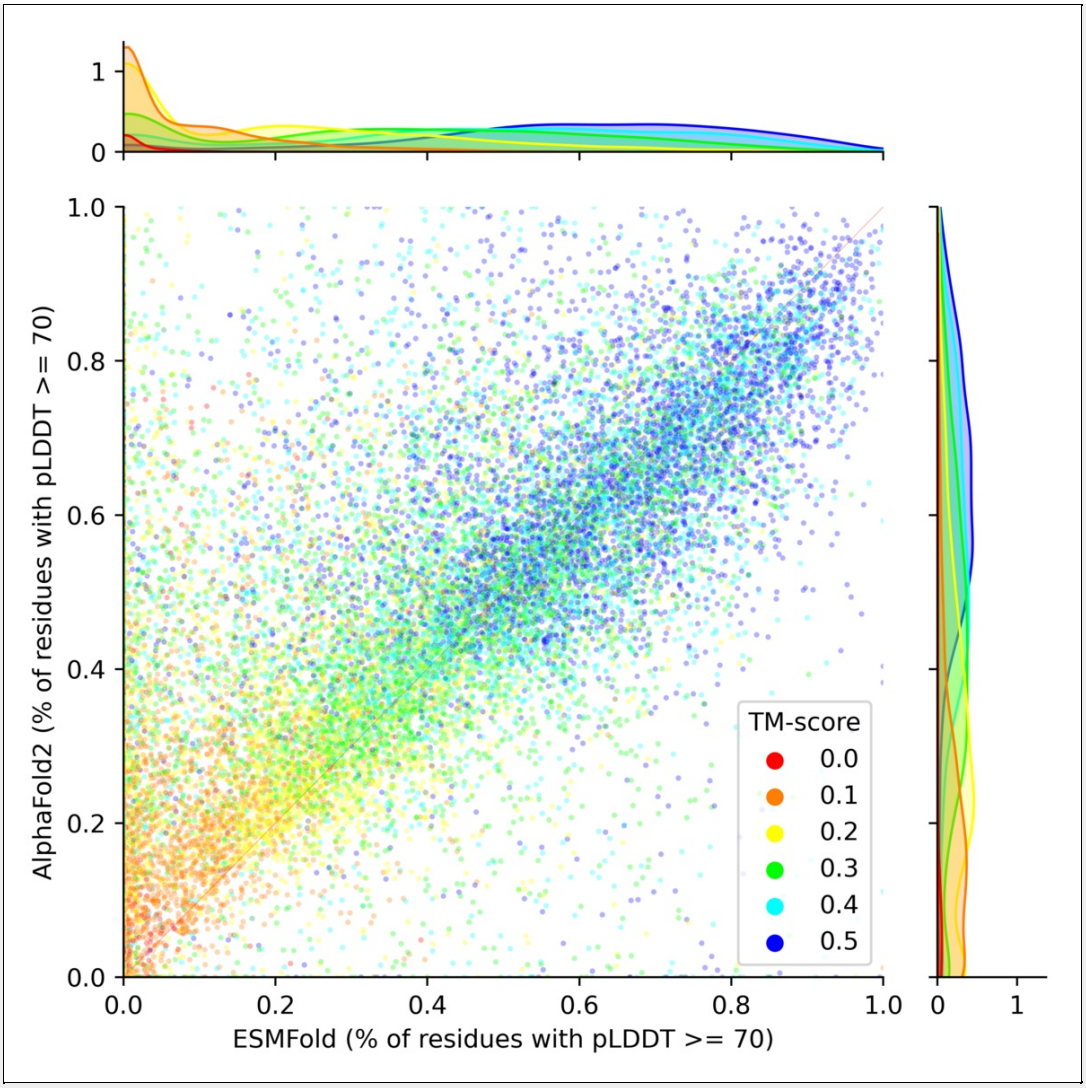
When comparing the TM-scores of AlphaFold2 and ESMFold models against the PDB structures, for 81% of the dataset the difference is <0.1, showing similar performances for the two methods. For the remaining proteins, AlphaFold2 tends to perform better than ESMFold. This is expected, as the first method retrieves known templates during the prediction phase.
Looking at the whole dataset, we find that for 45% of the proteins the TM-score between the two models is >0.6. The remaining 23,701 proteins are predicted with diverging models endowed with TM-scores <0.6. In Figure 1 we focus on the latter subset, comparing the quality of the paired models, and we indicate with a colour code six different bins of TM-scores ([0, 0.1); [0.1, 0.2); [0.2, 0.3); [0.3, 0.4); [0.4, 0.5); [0.5, 0.6)). It appears that the more the two models diverge (at decreasing TM-score), the more the confidence of each model decreases. Moreover, AlphaFold2 seems to generate better quality models than ESMFold for 62% of the proteins. Being this a self-predicted assessment of the model quality, the value is however not sufficient for an estimate of the actual model validity.
Our database is freely accessible as a web server at https://alpha-esmhfolds.biocomp.unibo.it/, allowing users to easily compare models for 42,942 human proteins. From the Home page, it is possible to use a UniProt accession to directly access the information on the selected protein. An advanced search is available for selecting filtering parameters such as the gene name, PDB ID, and different thresholds of TM-scores. Alternatively, we offer the possibility to run a BLAST search with a protein given by the user against the whole dataset. For all entries of the database, the corresponding page shows multiple information. At the top, a table highlights general protein data, including cross-links to UniProt and PDB (when a structure is available). After that, a tab shows the comparison between the AlphaFold2 and ESMFold models, including the sequence alignment, the structure superimposition, and several statistics based on the quality of each model and their similarity. When a PDB structure is available, two similar tabs show the comparison between it and each predicted model. The web server provides a Help page describing each functionality, including four examples.

Info

Figure 1 reports the scatter plot of the percentage of residues with pLDDT ≥ 70 for ESMFold (x-axis) and AlphaFold2 (y-axis). Points are coloured based on the TM-scores between the two models and fall into 6 bins: red [0.0, 0.1), orange [0.1, 0.2), yellow [0.2, 0.3), green [0.3, 0.4), light blue [0.4, 0.5) and blue [0.5, 0.6). The graphs on the side report the distribution of each axis (same colour code).

filename	Figure 1.png

Figure



Availability	https://alpha-esmhfolds.biocomp.unibo.it/

Dissemination Material

Social

-

Summary

Alpha&ESMhFolds is a novel database publicly released as a web server to compare AlphaFold2 and ESMFold models for 42,942 proteins from the Human Reference Proteome. We adopted it to conduct an extensive analysis of the methods comparison, including PDB structures for 2,900 proteins.

Corresponding Author

Name, Surname	Matteo, Manfredi
Email	matteo.manfredi4@unibo.it

Submitted on	29.04.2024