Carnegie Mellon University

Midterm Presentation for the Merck Capstone Project

Protein Folding

Joon Jung, Britney Wang, Fangzhou Yuan

Agenda

Background information - Protein Folding, Tools available

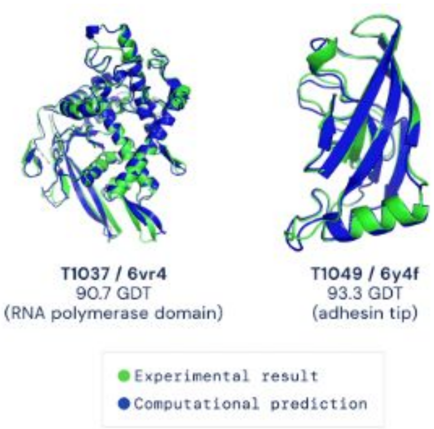
Scholarly articles research

Installation of Protein Folding Tools on Bridges-2 PSC

Summary of “AlphaFold2 can predict single-mutation effects”

What is Protein Folding?

- **Process where linear sequence of amino acids folds into structure**
- **Crucial for its functionality:** Correct folding of protein is crucial for its functionality
  - Enzymes binding to substrates, signaling molecules to receptors
  - Neurodegenerative diseases : Alzheimer’s, Parkinson’s disease
- **Important to understand and be able to predict correctly**
  - Biological function, disease mechanism, drug design



Protein Prediction Tasks

- **Structure Prediction:** prediction of how the backbone and side chains fold in 3D space.
    - Primary, Secondary, **Tertiary**, Quaternary
- **Prediction of Protein Disorder Regions:** some regions of proteins (eg: RNA) are intrinsically disordered or highly flexible, which can be critical for their function.
- **Antibody-antigen Structure Prediction**
- **Effects of Mutations Prediction:** how changes in the amino acid sequence (due to mutations) can affect the overall structure and function of the protein.

Why is it so hard?

- **Complexity of Protein folding process:** Highly complex process that involves numerous interatomic interactions
  - Hydrogen bonds, hydrophobic interactions, van der Waals force, etc
- **Limited Experimental data:** Collecting data through experiments are time-consuming and expensive.
- **Conformational Flexibility:** Proteins adopt into multiple conformations under different environmental conditions or with other molecules
- **Computational Complexity:** complexity of prediction increases exponentially as the sequence gets long and size of protein increases (degrees of freedom, local minima)

Model Comparison (Research)

- **AlphaFold2 (DeepMind):** network-based model, relies on Multiple Sequence Alignments (MSAs)
  - Winner of CASP13 (2018), CASP14 (2020)
  - High accuracy, got a GDT score above 90
- **RoseTTAFold (UW):** inspired by AF2, but different similar NN architecture
  - RMSD score is comparable to AF2's
- **ESMFold (Meta AI)**: large-scale language based model
  - Requires only a single input sequence
  - Faster prediction speed, but lower accuracy
  - Predict the structures of orphan proteins with higher accuracy than AF
- **OmegaFold (Helixon)**: DL-based method that uses only single primary sequence and no MSAs
  - better suited for proteins that have low sequence coverage

Model Evaluation Metrics

Overall Structural Accuracy

- **GDT_TS** and **GDT_HA**: overall shape and fold of a protein (Used in CASP)
- **TM-score:** overall topological similarity rather than specific atomic positions

Local Structural Accuracy

- **plDDT:** confidence in the accuracy of local structural features

Backbone Conformation

- **RMSD:** average distance between the atoms (usually the backbone atoms)

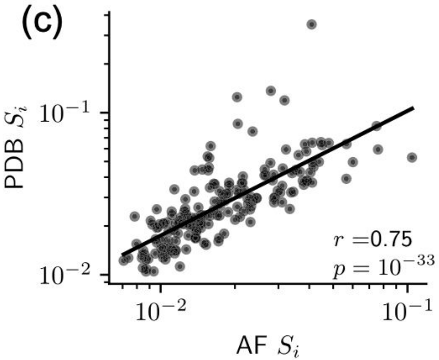
Evaluation on Mutations Effects Prediction

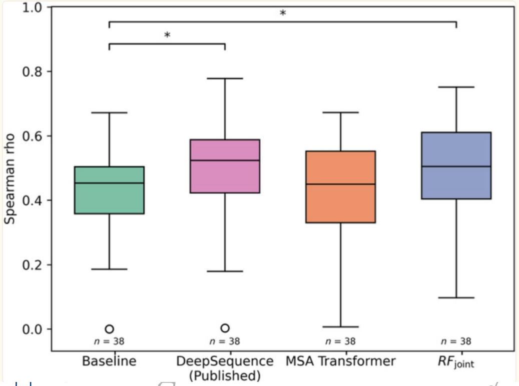
Metrics

- pLDDT – insufficient to sort folding from nonfolding proteins
- **Effective Strain(ES)**: measure local deformation – more robust measure of structural change upon mutation.
- **Log odds ratio**: how often the mutant (altered) amino acid appears at a specific position in a set of related proteins

Model Results

- AlphaFold2
- RoseTTAFold Joint (RFjoint)





sources: AlphaFold2 Can Predict Single-Mutation Effects, Accurate Mutation Effect Prediction using RoseTTAFold

Installing Folding Models

Bridges-2 (thanks to CMU and PSC)

- Storage (~1TB for our current database, up to 2.6TB for the full database)
- GPU

Installing Folding Models

AlphaFold2 (AlphaFold Non-Docker)

- AlphaFold release v2.3.1
- OpenMM patches
- Small genetic database (instead of the full database)

- Docker restrictions -> Singularity containers -> AlphaFold non-Docker
- Environment has been set up
- Errors in prediction phase
  - Reduced database/test sequence issues
  - cudatoolkit and nvidia cuda driver version compatibility issues

Installing Folding Models

OmegaFold (OmegaFold v1.1.0)

- Fully operational
- A single .fasta file -> A list of .pdb files

ESMFold (ESM-2 v1.0.3)

- Half way in environment setup
- Alternative implementations: ColabFold and ESM Metagenomic Atlas

Next step

- **AlphaFold**
  - Debugging (possible source of error: cudatoolkit compatibility issues)
  - Running test sequences
- **OmegaFold**
  - Running test sequences
- **ESMFold & RoseTTAFold2**
  - Finishing the setup

Paper summary

“AlphaFold2 can predict single-mutation effects”

- **Comparing AF predictions with curated set of proteins from PDB**
  - “AF can detect the effect of mutation on structure by identifying local deformations between protein pairs differing by 1-3 mutations.
  - Recent evidence suggests that AF learns the energy functional underlying folding
- **“AF can predict local structural change.”**
  - wild-type(WT 6BDD_A) and single mutant (6BDE_A) structures of H-NOX protein
  - Metrics used(local deformation): Effective strain (ES) per residue Si, and phenotype change (no consistent correlation with RMSD, pLDDT)
- **“AF predicts structure, not folding”**
  - They emphasize that AF is only trained to predict structures of stable proteins, and no claim whether protein will indeed fold into predicted structure

sources: AlphaFold2 Can Predict Single-Mutation Effects

Paper summary

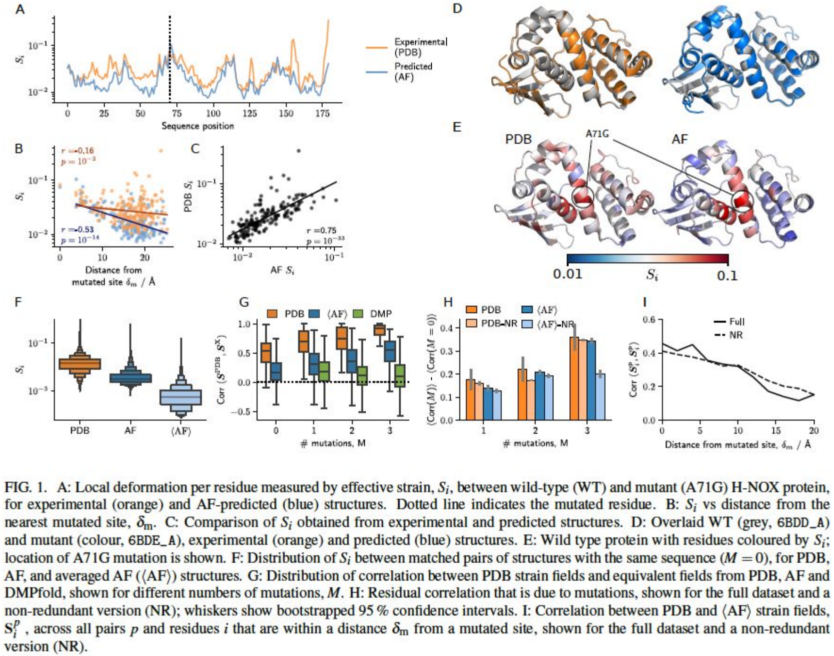


FIG. 1. A: Local deformation per residue measured by effective strain, S_i, between wild-type (WT) and mutant (A71G) H-NOX protein, for experimental (orange) and AF-predicted (blue) structures. Dotted line indicates the mutated residue. B: S_i vs distance from the nearest mutated site, δ_m. C: Comparison of S_i obtained from experimental and predicted structures. D: Overlaid WT (grey, 6BDD_A) and mutant (colour, 6BDE_A), experimental (orange) and predicted (blue) structures. E: Wild type protein with residues coloured by S_i; location of A71G mutation is shown. F: Distribution of S_i between matched pairs of structures with the same sequence ($M = 0$), for PDB, AF, and averaged AF (⟨AF⟩) structures. G: Distribution of correlation between PDB strain fields and equivalent fields from PDB, AF and DMPfold, shown for different numbers of mutations, M. H: Residual correlation that is due to mutations, shown for the full dataset and a non-redundant version (NR); whiskers show bootstrapped 95 % confidence intervals. I: Correlation between PDB and ⟨AF⟩ strain fields, S_i^p, across all pairs p and residues i that are within a distance δ_m from a mutated site, shown for the full dataset and a non-redundant version (NR).

sources: AlphaFold2 Can Predict Single-Mutation Effects

Thank you

Questions?

We want to thank CMU, PSC, and Merck.co team for the resources and time to allow us do this capstone project.