HomologousProteinSeeker tool: User Manual

(HPS)

Table of contents:

* Introduction
  + About this pipeline
  + Workflow diagram
* Package introduction
  + Esm-fold
  + Pymol
  + Foldseek
  + Swissport
* Installation guideline
* Running Pipeline
  + Required input
  + Usage
  + Command Line
  + Parameters
* Getting Started with the Pipeline
* Pitfalls and Limitation

**Introduction:**

Describing the pipeline:

The pipeline is for taking in amino acid sequences and using a combination of bioinformatic tools in order to compare homologous protein structures visually and annotatively as the goal. Esm-fold is a way for us to get structure predictions of the amino acid sequences from Fasta files to Pdb files. Through this pipeline you can take and compare multiple predicted structures through Pymol and search for similar proteins in various protein databases via Foldseek. A python script is made for each fasta file and output to a sbatch script which then would be submitted to GPU-shared clusters in Bridges to run Esm-fold. Once the predictions are given as pdb files, Pymol can visualize and compare the structures and get RMSD values as txt files and png(s) of the structures. Foldseek is also used on the pdb files to search databases to get similar proteins and do structural annotation to understand structure conservation, understand functional domains, and for distant homology detection.

**Workflow:**

The workflow as a visualization and annotation portion to get qualitative and quantitative comparison of the isoform predictions to themselves and to known proteins in 2 databases (Alpha Fold/Proteome and Alpha Fold/Swiss Prot).

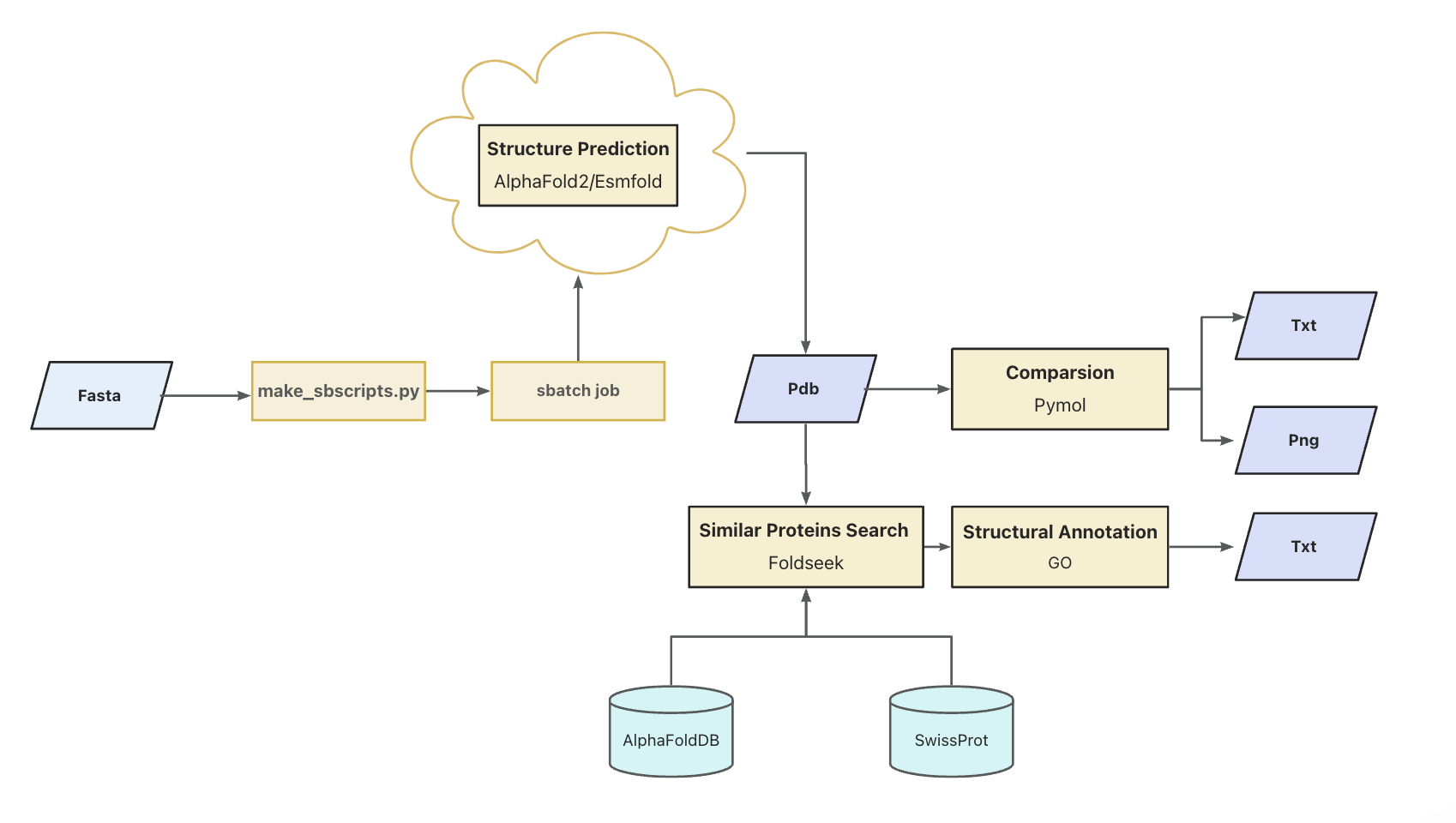


Figure 1: The workflow of our pipeline

**Packages:**

**ESM-fold:**

Takes fasta files as input and is an AI tool that predicts the 3d structure of proteins as PDB files to be then visualized.

**More details:**

<https://esmatlas.com/about>

**Pymol:**

It can visualize the protein structures in a 3d manner and also compare the structures in depth to between the isoforms/homologous proteins through aligning and getting the RMSD values.

More details:

<https://pymol.org/dokuwiki/doku.php?id=welcome>

Download from:

<https://pymol.org/>

**Foldseek:**

Compare the isoform predictions to structure databases and find proteins that share similar three dimensional structure. We used databases Alpha Fold/Proteome and Alpha Fold/Swiss Prot.

More details:

<https://github.com/steineggerlab/foldseek?tab=readme-ov-file>

**Package installation:**

**Esm-fold:**

$ git clone https://github.com/facebookresearch/esm.git

# the .yml file is inside the repo folder

$ conda env create -f environment.yml

# Activate esm fold environment

$ conda activate esmfold

$ conda install -c conda-forge einops=0.6.1

$ pip install git+https://github.com/facebookresearch/esm.git

$ pip install "fair-esm[esmfold]"

$ pip install 'dllogger @ git+https://github.com/NVIDIA/dllogger.git'

$ pip install 'openfold @ git+https://github.com/aqlaboratory/openfold.git@4b41059694619831a7db195b7e0988fc4ff3a307'

**Foldseek:**

# Linux AVX2 build (check using: cat /proc/cpuinfo | grep avx2)

wget https://mmseqs.com/foldseek/foldseek-linux-avx2.tar.gz; tar xvzf foldseek-linux-avx2.tar.gz; export PATH=$(pwd)/foldseek/bin/:$PATH

# Linux SSE2 build (check using: cat /proc/cpuinfo | grep sse2)

wget https://mmseqs.com/foldseek/foldseek-linux-sse2.tar.gz; tar xvzf foldseek-linux-sse2.tar.gz; export PATH=$(pwd)/foldseek/bin/:$PATH

# Linux ARM64 build

wget https://mmseqs.com/foldseek/foldseek-linux-arm64.tar.gz; tar xvzf foldseek-linux-arm64.tar.gz; export PATH=$(pwd)/foldseek/bin/:$PATH

# MacOS

wget https://mmseqs.com/foldseek/foldseek-osx-universal.tar.gz; tar xvzf foldseek-osx-universal.tar.gz; export PATH=$(pwd)/foldseek/bin/:$PATH

# Conda installer (Linux and macOS)

conda install -c conda-forge -c bioconda foldseek



Database required for foldseek could be downloaded using foldseek command. For example, Alpha Fold/Proteome could be downloaded to current through this command:

# alphafold db

foldseek databases Alphafold/Proteome afdb tmp

Foldseek currently supports the following databases: (When running foldseek, the path to the database will be required.)

 Name Type Taxonomy Url

- Alphafold/UniProt Aminoacid yes https://alphafold.ebi.ac.uk/

- Alphafold/UniProt50 Aminoacid yes https://alphafold.ebi.ac.uk/

- Alphafold/Proteome Aminoacid yes https://alphafold.ebi.ac.uk/

- Alphafold/Swiss-Prot Aminoacid yes https://alphafold.ebi.ac.uk/

- ESMAtlas30 Aminoacid - https://esmatlas.com

- PDB Aminoacid yes https://www.rcsb.org



**PyMOL:**

# Conda installer (Linux and macOS)

conda install -c schrodinger pymol



**Running Pipeline:**

**Required Input Files:**

1. Amino acid sequence Fasta files of homologous proteins/ isoforms
2. Pre-generated databases for Foldseek

**Usage:**

All the scripts under ./scripts should not be moved. The relative path to the binaries is used.

Our intermediate files are the PDB files produced by ESM-fold and the pipeline has to wait for all PDB files to proceed. This will take some time and this is the message that the user will see"Wait for batch jobs done...". During the wait, the program will constantly check if all the PDB files are produced, so there is **no need for users to panic** when they see “No such directory” warnings during the waiting process since it is just because the check does not find all the PDB files needed. Once the output of PDB files equals the input file count the user will get the message “Structure Prediction done”.

**Command line:**

Example:

**$ bash HPS\_pipeline.sh -i <Data/path> -o <Output/path> -d <foldseek/Database/path>**

****

**Parameters:**

Required:

-i <data dir> input fasta directory path

-o <output dir> output directory path

-d <database path> your foldseek db path, e.g. -d ./foldseek/database

**Getting Started with the Pipeline:**

**Esm-fold:**

It predicts the three-dimensional structure of proteins based on their amino acid sequences. It leverages a deep learning technique called evolutionary scale modeling (ESM) to understand patterns in biological sequences. The input files should be fasta files for amino acid sequences. The output files are in PDB file format.

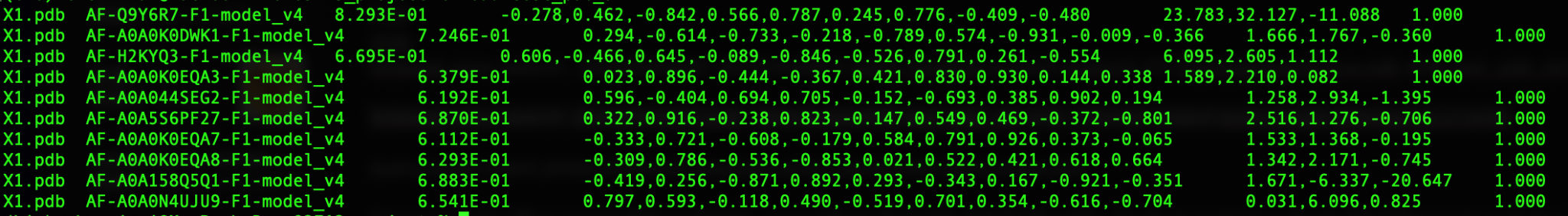
**Foldseek:**

It will perform searches in the databases using the input files (predicted protein structure) and output the comparison result with protein in pre-generated databases.The 10 output files will be stored under ./2\_foldseek

Based on the foldseek result, the annotation process would extract the uniport ID of top 10 closest protein and search annotation information from <https://www.uniprot.org> and output to local .txt fileThe annotation file will be stored under ./3\_annotation

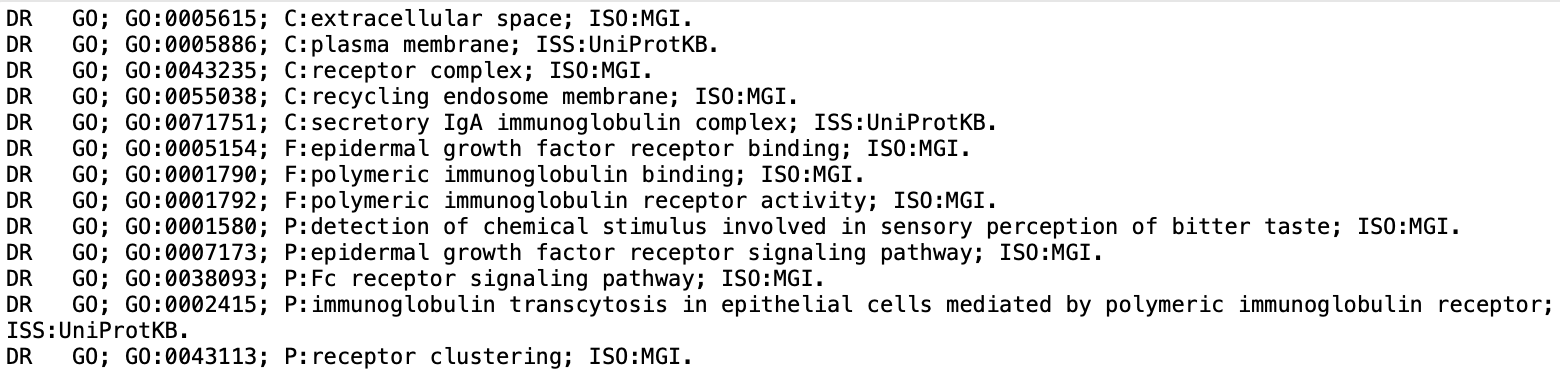
Sample Output:

The result of Foldseek would contain information about [query file, search target, TM-score of the alignment, Rotation matrix, Translation vector,Probability]



Structure comparison result example

The results of the annotations are from the search results in Gene Ontology database using the protein IDs from Foldseek as queries. The annotations suggest the functions of those query proteins.



Annotation result example

**Pymol:**

It takes the PDB files produced by ESMfold as inputs. It will output the images as png files for the visualized protein structures of inputting PDB files. It will output the RMSD scores between every pair of protein structures as a txt file

Sample Output:

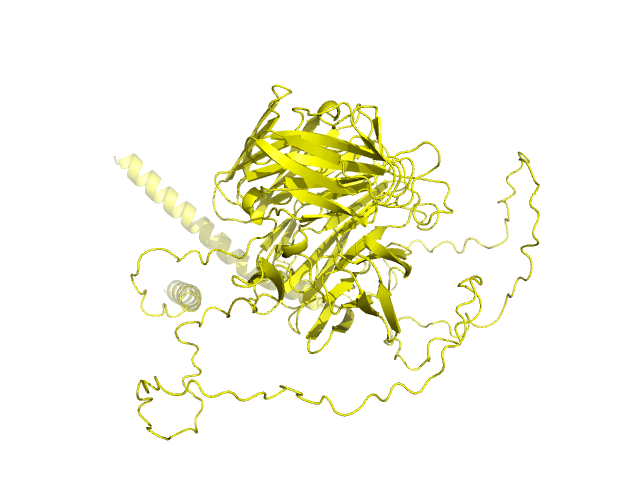


Fig. 1 Single molecule visualization image

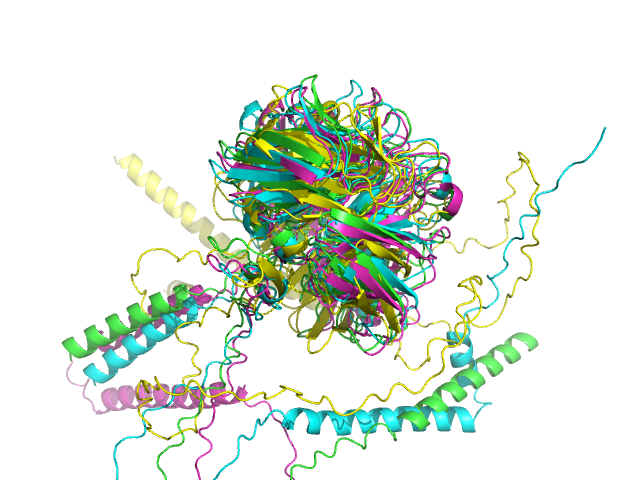


Fig. 2 Multiple molecules visualization image

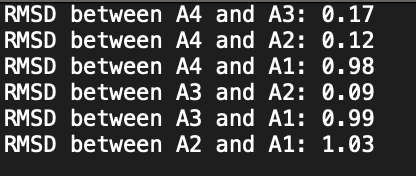


Fig.3 RMSD scores for molecules pairs in the output txt file

**Pitfalls and Limitations:**

The ESM-fold part of our pipeline now requires access to Bridges and may need to wait for available computational resources. Limited test cases is also an issue since our proteins tested are relatively small, so more complex proteins can cause issues with the pipeline. For foldseek analysis you need to locally download the two databases which take large storage space.

The function annotations returned at the end of the pipeline are based on Gene Ontology annotations of the similar proteins. It could be that the functions of similar proteins identified by the Foldseek are not the same with the query protein. Also, if there is no annotation of the proteins in the Gene Ontology database, the pipeline will not be able to return any annotations. Also, if there are lots of PDB files that Pymol has to visualize, it will take a significant amount of time.