Class 11: Protein Structure Prediction with Alphafold2

Shazreh Hassan (PID: A13743949)

Table of contents

Background

We saw last day that the PDB has 209,886 entries (Oct/Nov 2025). UniProtKB (i.e. protein sequence database) has 199,579,901 entries.

```
209886/199579901*100
```

```
[1] 0.1051639
```

The PDB has only 0.1% coverage of the main sequence database. Enter AlphaFold database (AFDB). < https://alphafold.ebi.ac.uk/ > that attempts to provide computed models for all sequences in UniProt. "AlphaFold DB provides open access to over 200 million protein structure predictions to accel- erate scientific research."

Alphafold

AlphaFold has 3 main outputs:

1) The predicted coordinates (PDB files)
2) A local quality score plDDT (one for each amino-acid)

3) A second quality score Predicted Aligned Error PAE (for each pair of amino-acids)

We can run AlphaFold ourselves if we are not happy with AFDB (could be no coverage or poor model)

Interpreting/analyzing alphafold results in R

```
results_dir <- "HIV_pr_94b5b/"
```

```
pdb_files <- list.files(path=results_dir,
pattern="*.pdb",
full.names = TRUE)

basename(pdb_files)
```

```
[1] "HIV_pr_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb"
[2] "HIV_pr_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_5_seed_000.pdb"
[3] "HIV_pr_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_3_seed_000.pdb"
[4] "HIV_pr_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_2_seed_000.pdb"
[5] "HIV_pr_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_1_seed_000.pdb"
```

```
library(bio3d)
```

```
Warning: package 'bio3d' was built under R version 4.5.2
```

```
pdbs <- pdbaln(pdb_files, fit=TRUE, exefile="msa")
```

```
Reading PDB files:
HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb
HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_5_seed_000.pdb
HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_3_seed_000.pdb
HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_2_seed_000.pdb
HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_1_seed_000.pdb
.....

Extracting sequences

pdb/seq: 1   name: HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_
```

```
pdb/seq: 2   name: HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_5_seed_
pdb/seq: 3   name: HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_3_seed_
pdb/seq: 4   name: HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_2_seed_
pdb/seq: 5   name: HIV_pr_94b5b/HIV_pr_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_1_seed_
```

```
## library(bio3dview)
## view.pdbs(pdbs)
```

Alignment file

```
aln_file <- list.files(path=results_dir,
                       pattern=".a3m$",
                       full.names=TRUE)

aln_file
```

```
[1] "HIV_pr_94b5b/HIV_pr_94b5b.a3m"
```

```
aln <- read.fasta(aln_file[1], to.upper=TRUE)
```

```
[1] " ** Duplicated sequence id's: 101 **"
```

```
sim <- conserv(aln)
```