Class 11

Emily Chase (PID: 14656894)

We saw last time that the PDB (the main experimentally resolved structure db) has 209,886 entries (Oct/Nov 2025). UniProtKB (i.e. protein sequence database) has 199,579,901 entries … PDB 0.1% coverage of UniProt.

Enter Alpha fold database (AFDB). https://alphafold.ebi.ac.uk/ It attempts to provide computed models for all sequences in UniProt. According to AFDB website: "AlphaFold DB provides open access to over 200 million protein structure predictions to accelerate scientific research."

AFDB relies on MSAs and people have dug into the black box: coevolution events are predictive of contacts, then refines with physics-based approaches.

Alpha Fold

AlphaFold DB has 3 main outputs:

- the predicted coordinates (PDB files)
- a local quality score called **pLDDT** (one for each amino acid)
- a second quality score called **PAE** (packing score; predicted aligned error; for each pair of aa)

We can run alphafold ourselves if we are unhappy with AFDB (ie no coverage or poor model) on colabfold with mmer2.

- .a3m –> all found sequences
- .pdb –> structure
- .json –> confidence scores (plDDT)

We are now looking on molstar at the pdb files. We took the most confident structure (#1) and the answer (1hsg). We selected two chains and hit "superpose" and now we can see how similar the predicted vs actual structure is.

Interpreting/analyzing AF results in R

```
results_dir <- "HIVPR_dimer_23119/"

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

basename(pdb_files)
```

```
[1] "HIVPR_dimer_23119_unrelaxed_rank_001_alphafold2_multimer_v3_model_4_seed_000.pdb"
[2] "HIVPR_dimer_23119_unrelaxed_rank_002_alphafold2_multimer_v3_model_1_seed_000.pdb"
[3] "HIVPR_dimer_23119_unrelaxed_rank_003_alphafold2_multimer_v3_model_5_seed_000.pdb"
[4] "HIVPR_dimer_23119_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000.pdb"
[5] "HIVPR_dimer_23119_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000.pdb"
```

```
library(bio3d)

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

```
Reading PDB files:
HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_001_alphafold2_multimer_v3_model_4_seed_
HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_002_alphafold2_multimer_v3_model_1_seed_
HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_003_alphafold2_multimer_v3_model_5_seed_
HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_
HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_
.....

Extracting sequences

pdb/seq: 1   name: HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_001_alphafold2_multime
pdb/seq: 2   name: HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_002_alphafold2_multime
pdb/seq: 3   name: HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_003_alphafold2_multime
pdb/seq: 4   name: HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_004_alphafold2_multime
pdb/seq: 5   name: HIVPR_dimer_23119//HIVPR_dimer_23119_unrelaxed_rank_005_alphafold2_multime
```

Alignment file

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

```
[1] "HIVPR_dimer_23119//HIVPR_dimer_23119.a3m"
```

```
pdb <- read.pdb("1hsg")
```

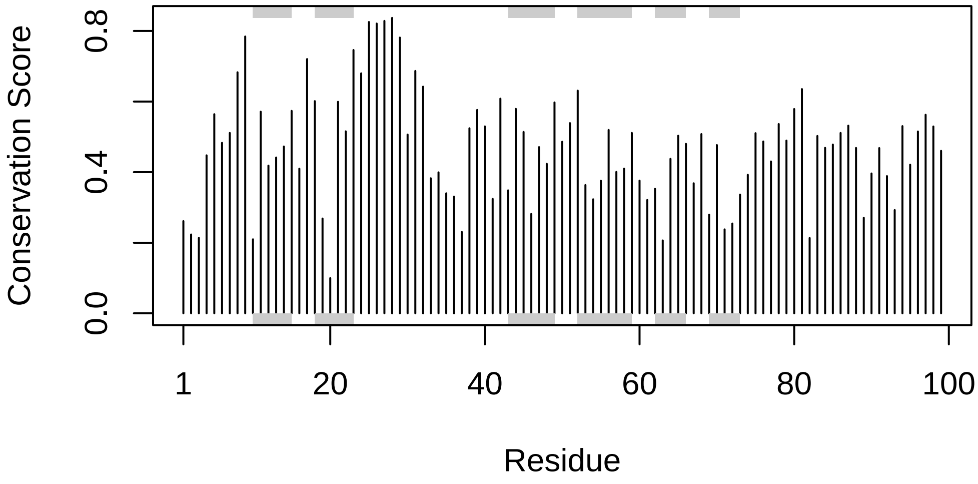
```
  Note: Accessing on-line PDB file
```

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

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

```
sim <- conserv(aln)

plotb3(sim[1:99], sse=trim.pdb(pdb, chain="A"),
       ylab="Conservation Score")
```


Conservation Score
Residue