class11

Melanda Aboueid (PID A17473102)

Background

We saw last day that the main repositonary for biomolecular structure (the PDB database) only has about 250,000 entries.

UniProtKB (the main protien sequence database) has over 200 million entries!

In this hands-on session we will utilize AlphaFold to predict protein structure from sequence (Jumper et al. 2021).

Without the aid of such approaches, it can take years of expensive laboratory work to determine the structure of just one protein. With AlphaFold we can now accurately compute a typical protein structure in as little as ten minutes.

The EBI AlphaFold database

The EBI alphafold database contains lots of computed structure models. It is increasing likely that the structure you are intreased in is already in this database at https://alphafold.ebi.ac.uk/

There are 3 major outputs from AlphaFold

1. A model of structure in PDB format,
2. a pLDDT score: that tells us how confident the model is for a given residue in your protien (High valuses are good, above 70)
3. a **PAE score** that tells us about protien packing quality.

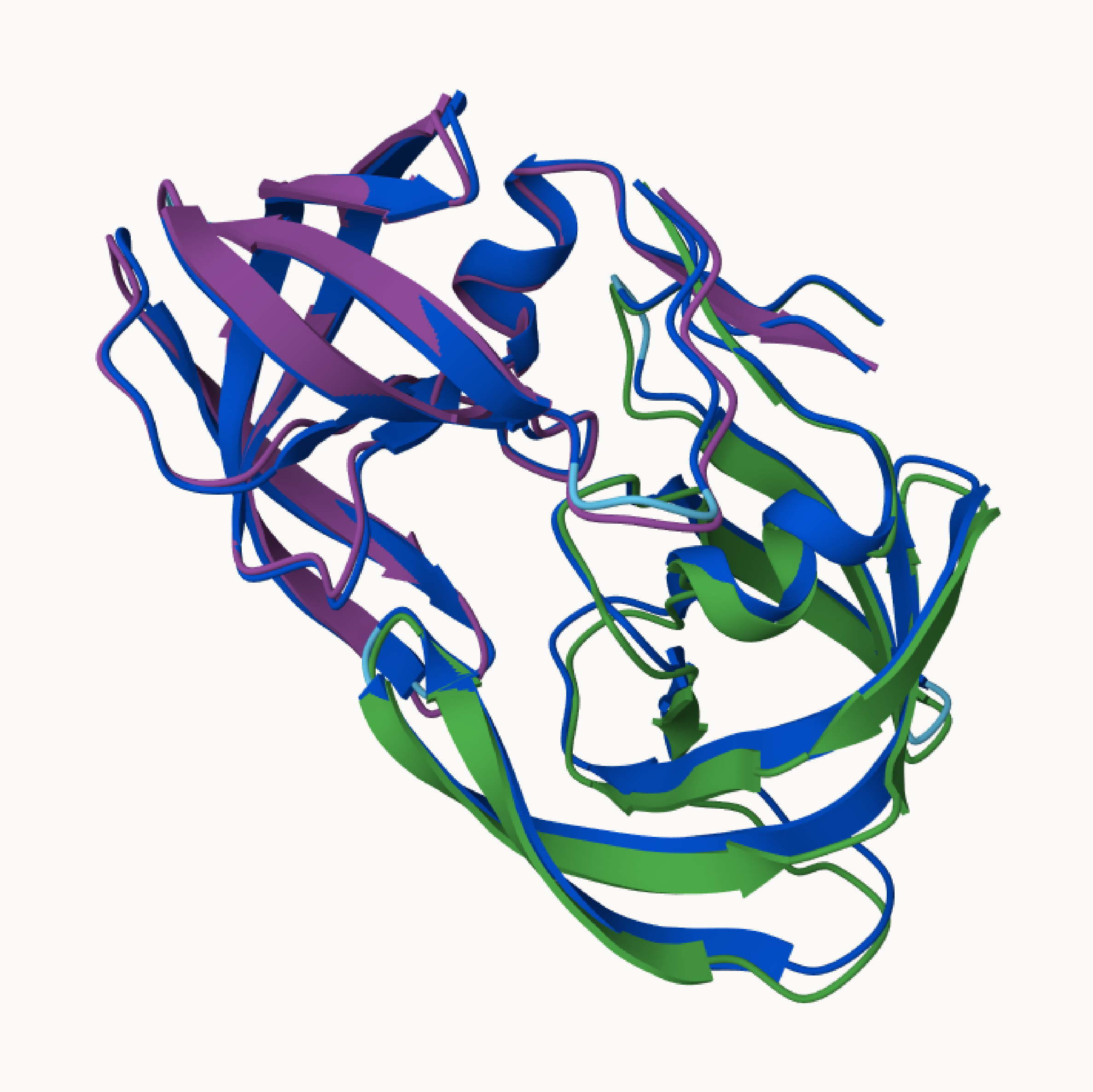
If you can't find a matching entry fir the sequence you are intreasted in AFDB you can run AlphaFold yourself…

Running AlphaFold

We will use Colab https://github.com/sokrypton/ColabFold

Figure from AlphaFold here!

```
knitr::include_graphics("Picture.png")
```



Interpreting results

coustom analysis of resulting models

we can read all AlphaFold results into R

```r
results_dir <- "hivpr_23119_0"

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

basename(pdb_files)
```

```
[1] "hivpr_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_4_seed_000.pdb"
[2] "hivpr_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_1_seed_000.pdb"
[3] "hivpr_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_5_seed_000.pdb"
[4] "hivpr_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000.pdb"
[5] "hivpr_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000.pdb"
```

```r
library(bio3d)
pdbs <- pdbaln(pdb_files, fit = TRUE, exefile = "msa")
```

```
Reading PDB files:
hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_4_seed_000.pdb
hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_1_seed_000.pdb
hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_5_seed_000.pdb
hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000.pdb
hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000.pdb
.....

Extracting sequences

pdb/seq: 1   name: hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_mode
pdb/seq: 2   name: hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_mode
pdb/seq: 3   name: hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_mode
pdb/seq: 4   name: hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_mode
pdb/seq: 5   name: hivpr_23119_0/hivpr_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_mode
```

```r
pdbs
```

```
                               1        .         .         .         .         50
[Truncated_Name:1]hivpr_2311   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:2]hivpr_2311   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:3]hivpr_2311   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:4]hivpr_2311   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:5]hivpr_2311   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
                               **************************************************
                               1        .         .         .         .         50

                              51        .         .         .         .         100
[Truncated_Name:1]hivpr_2311   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:2]hivpr_2311   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:3]hivpr_2311   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:4]hivpr_2311   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:5]hivpr_2311   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
                               **************************************************
                              51        .         .         .         .         100

                             101        .         .         .         .         150
[Truncated_Name:1]hivpr_2311   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:2]hivpr_2311   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:3]hivpr_2311   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:4]hivpr_2311   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:5]hivpr_2311   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
                               **************************************************
                             101        .         .         .         .         150

                             151        .         .         .         .       198
[Truncated_Name:1]hivpr_2311   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:2]hivpr_2311   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:3]hivpr_2311   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:4]hivpr_2311   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:5]hivpr_2311   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
                               ************************************************
                             151        .         .         .         .       198

Call:
  pdbaln(files = pdb_files, fit = TRUE, exefile = "msa")

Class:
  pdbs, fasta

Alignment dimensions:
  5 sequence rows; 198 position columns (198 non-gap, 0 gap)

+ attr: xyz, resno, b, chain, id, ali, resid, sse, call
```

```
#pdbs
```

Similarity and diffrences between the models

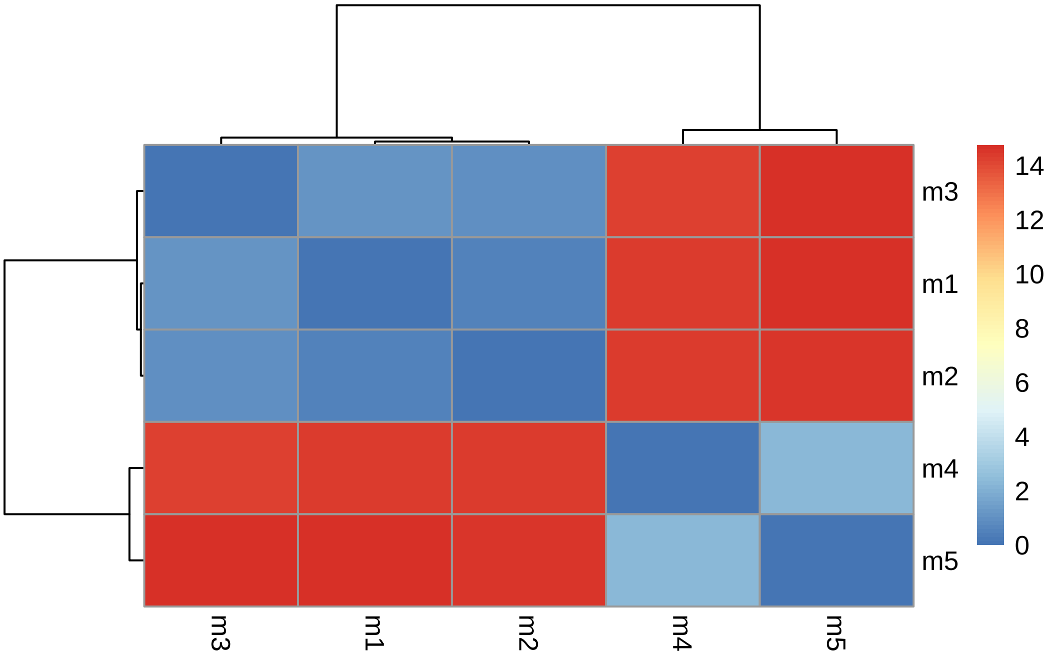
```
rd <- rmsd(pdbs, fit=TRUE)
```

```
Warning in rmsd(pdbs, fit = TRUE): No indices provided, using the 198 non NA positions
```

```
range(rd)
```

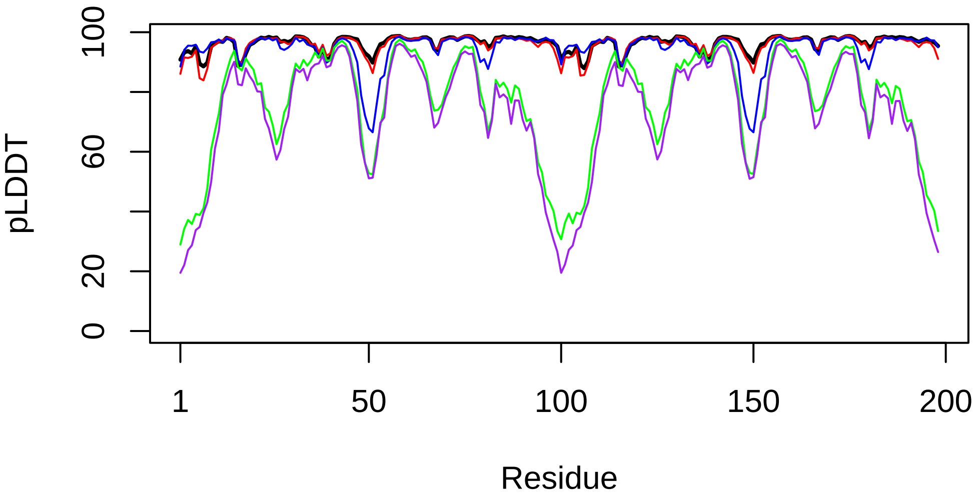
```
[1]  0.000 14.754
```

```
library(pheatmap)

colnames(rd) <- paste0("m",1:5)
rownames(rd) <- paste0("m",1:5)

pheatmap(rd)
```



Models with lower RMSD values are more structurally similar. Clustering in the heatmap indicates which predictions agree most strongly.

```
plotb3(pdbs$b[1,], typ="l", lwd=2,
       ylab="pLDDT", xlab="Residue")

points(pdbs$b[2,], col="red", typ="l")
points(pdbs$b[3,], col="blue", typ="l")
points(pdbs$b[4,], col="green", typ="l")
points(pdbs$b[5,], col="purple", typ="l")
```



Most show high pLDDT scores so they are reliable predictions. The Lower scores suggest flexible or disordered regions.

```
library(jsonlite)

pae_files <- list.files(path=results_dir,
                        pattern=".*model.*\\.json$",
                        full.names=TRUE)

pae_files
```

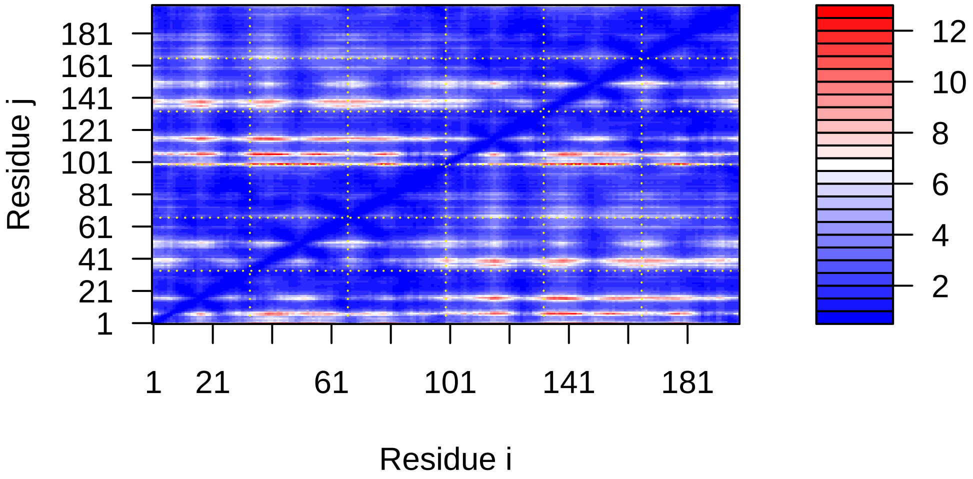
```
[1] "hivpr_23119_0/hivpr_23119_0_scores_rank_001_alphafold2_multimer_v3_model_4_seed_000.json
```

```
[2] "hivpr_23119_0/hivpr_23119_0_scores_rank_002_alphafold2_multimer_v3_model_1_seed_000.json
[3] "hivpr_23119_0/hivpr_23119_0_scores_rank_003_alphafold2_multimer_v3_model_5_seed_000.json
[4] "hivpr_23119_0/hivpr_23119_0_scores_rank_004_alphafold2_multimer_v3_model_2_seed_000.json
[5] "hivpr_23119_0/hivpr_23119_0_scores_rank_005_alphafold2_multimer_v3_model_3_seed_000.json
```

```
pae1 <- read_json(pae_files[1], simplifyVector=TRUE)

str(pae1)
```

```
List of 5
 $ plddt  : num [1:198] 90.8 93.2 93.7 92.9 95.2 ...
 $ max_pae: num 12.8
 $ pae    : num [1:198, 1:198] 0.75 0.86 1.39 1.78 2.34 4.23 7.09 5.05 3.2 2.37 ...
 $ ptm    : num 0.91
 $ iptm   : num 0.9
```

```
plot.dmat(pae1$pae,
          xlab="Residue i",
          ylab="Residue j")
```



The PAE plot is mostly low (blue), indicating high confidence in residue positioning, with a few higher-error regions showing minor uncertainty in domain packing but overall a reliable model.