class10

Joshua Martin (PID: A18545389)

AlphaFold Data Base (AFDB)

The EBI maintains the largest database of AlphaFold structure prediction models at: https://alphafold.ebi.ac.uk

From last class, Class09, (before Halloween) we saw that the PDB had 244,290 (Oct 2025)

The total number of protein sequences in UniProtKB is 199,579,901

> **Key Point**: This is a tiny fraction of sequence space that has structural coverage (0.12%)

```
244290/199579901 * 100
```

```
[1] 0.1224021
```

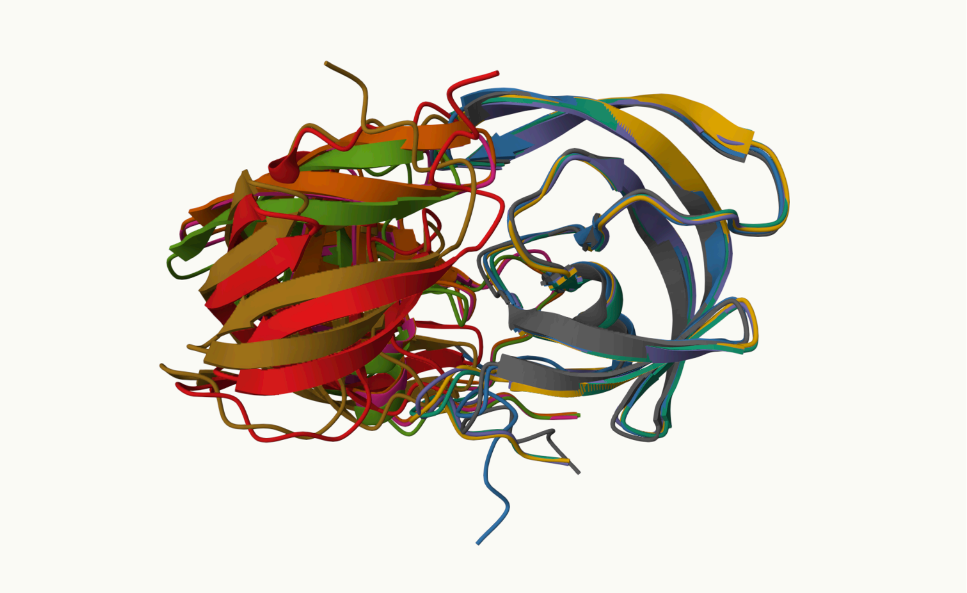
AFDB is attempting to address this gap…

There are two "Quality Scores" from AlphaFold one for residues (i.e each amino acid) called "pLDDT" socre. The other "PAE" score that measures the confidence in the relative position of two residues (i.e. a score for every pair of residues)

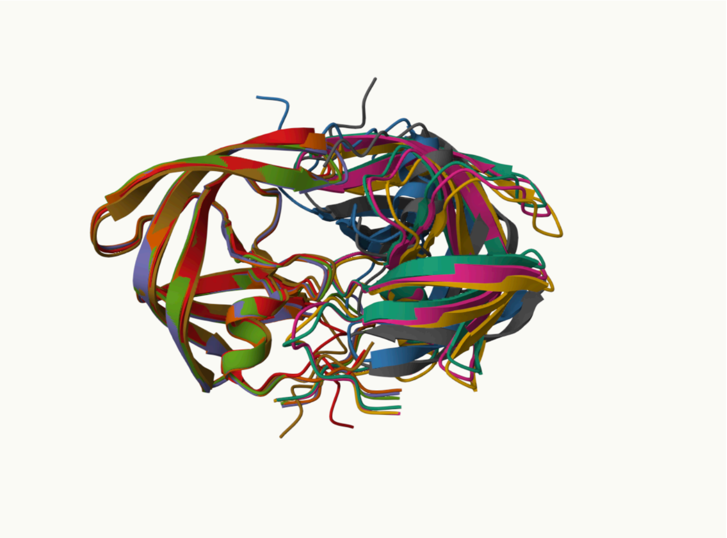
Generating your own structure predictions

image of all 5 models

```
knitr::include_graphics("5models.png")
```



```r
knitr::include_graphics("fin5modelsuperimposed.png")
```



Section 8

Custom analysis of resulting models in R

Read key result files into R. The first thing I need to know is what my results directory/folder is called (i.e. its name is different for every AlphaFold run/job)

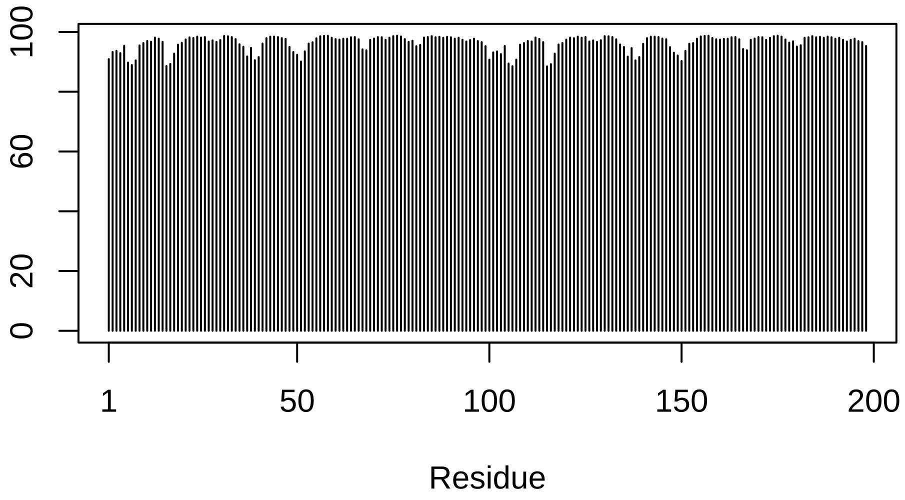
```r
results.dir <- "HIVPR_dimer_23119"

# Change this for YOUR results dir name
results_dir <- "HIVPR_dimer_23119/"

# File names for all PDB models
pdb_files <- list.files(path=results_dir,
                        pattern="*.pdb",
                        full.names = TRUE)

# Print our PDB file names
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"
```

```r
library(bio3d)

m1 <- read.pdb(pdb_files[1])
m1
```

```

 Call:  read.pdb(file = pdb_files[1])

   Total Models#: 1
     Total Atoms#: 1514,  XYZs#: 4542  Chains#: 2  (values: A B)

     Protein Atoms#: 1514  (residues/Calpha atoms#: 198)
     Nucleic acid Atoms#: 0  (residues/phosphate atoms#: 0)
     Non-protein/nucleic Atoms#: 0  (residues: 0)
     Non-protein/nucleic resid values: [ none ]

   Protein sequence:
      PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYD
      QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPQITLWQRPLVTIKIGGQLKE
      ALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTP
      VNIIGRNLLTQIGCTLNF

+ attr: atom, xyz, calpha, call
```

```
plot.bio3d(m1$atom$b[m1$calpha])
```



```
head(m1$atom)
```

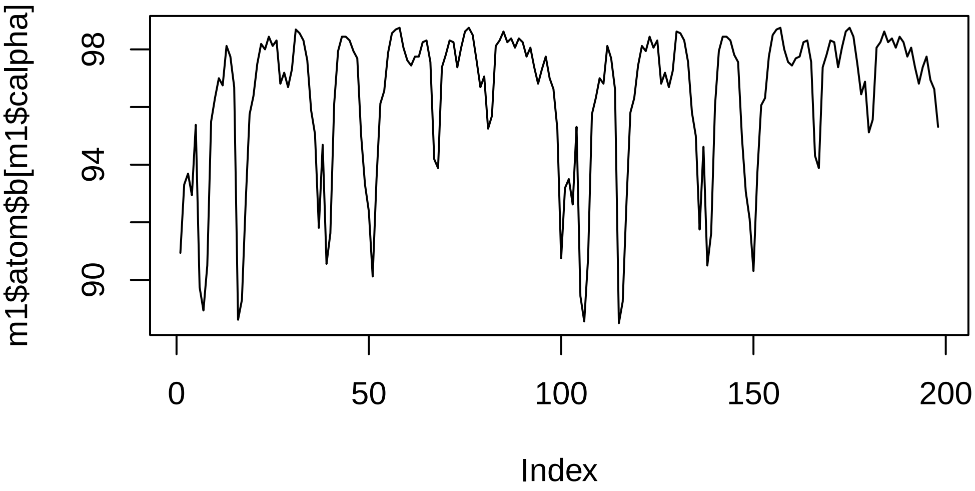
```
  type eleno elety  alt resid chain resno insert      x      y      z o     b
1 ATOM     1     N <NA>   PRO     A     1   <NA> 16.938 -3.990 -6.129 1 90.94
2 ATOM     2    CA <NA>   PRO     A     1   <NA> 16.938 -2.557 -6.430 1 90.94
3 ATOM     3     C <NA>   PRO     A     1   <NA> 16.438 -1.707 -5.266 1 90.94
4 ATOM     4    CB <NA>   PRO     A     1   <NA> 15.992 -2.449 -7.629 1 90.94
5 ATOM     5     O <NA>   PRO     A     1   <NA> 15.836 -2.232 -4.324 1 90.94
6 ATOM     6    CG <NA>   PRO     A     1   <NA> 15.070 -3.623 -7.496 1 90.94
```

```
  segid elesy charge
1  <NA>     N   <NA>
2  <NA>     C   <NA>
3  <NA>     C   <NA>
4  <NA>     C   <NA>
5  <NA>     O   <NA>
6  <NA>     C   <NA>
```

```
m1$atom$b[m1$calpha]
```

```
  [1] 90.94 93.31 93.69 92.94 95.38 89.75 88.94 90.50 95.50 96.31 97.00 96.75
 [13] 98.12 97.75 96.69 88.62 89.31 92.75 95.75 96.38 97.50 98.19 98.00 98.44
 [25] 98.12 98.31 96.81 97.19 96.69 97.31 98.69 98.56 98.31 97.62 95.88 95.06
 [37] 91.81 94.69 90.56 91.62 96.12 97.94 98.44 98.44 98.31 97.94 97.69 95.00
 [49] 93.31 92.38 90.12 93.50 96.12 96.56 97.88 98.56 98.69 98.75 98.06 97.62
 [61] 97.44 97.75 97.75 98.25 98.31 97.56 94.19 93.88 97.38 97.81 98.31 98.25
 [73] 97.38 98.06 98.62 98.75 98.50 97.62 96.69 97.06 95.25 95.69 98.12 98.31
 [85] 98.62 98.25 98.38 98.06 98.38 98.25 97.75 98.06 97.38 96.81 97.31 97.75
 [97] 97.00 96.62 95.25 90.75 93.19 93.50 92.62 95.31 89.44 88.56 90.75 95.75
[109] 96.31 97.00 96.81 98.12 97.69 96.62 88.50 89.25 92.75 95.81 96.31 97.44
[121] 98.12 97.94 98.44 98.06 98.31 96.81 97.19 96.69 97.25 98.62 98.56 98.31
[133] 97.56 95.81 95.00 91.75 94.62 90.50 91.62 96.06 97.94 98.44 98.44 98.31
[145] 97.81 97.56 94.94 93.06 92.12 90.31 93.69 96.06 96.31 97.75 98.50 98.69
[157] 98.75 98.00 97.56 97.44 97.69 97.75 98.25 98.31 97.56 94.31 93.88 97.38
[169] 97.81 98.31 98.25 97.38 98.06 98.62 98.75 98.44 97.50 96.44 96.88 95.12
[181] 95.56 98.06 98.25 98.62 98.25 98.38 98.06 98.44 98.25 97.75 98.06 97.38
[193] 96.81 97.38 97.75 96.94 96.62 95.31
```

```
plot( m1$atom$b[m1$calpha], typ="l")
```



Residue conservation from alignment file

Find the large AlphaFold alignment file

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

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

Read this int oR

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

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

How many sequences are in this alignment

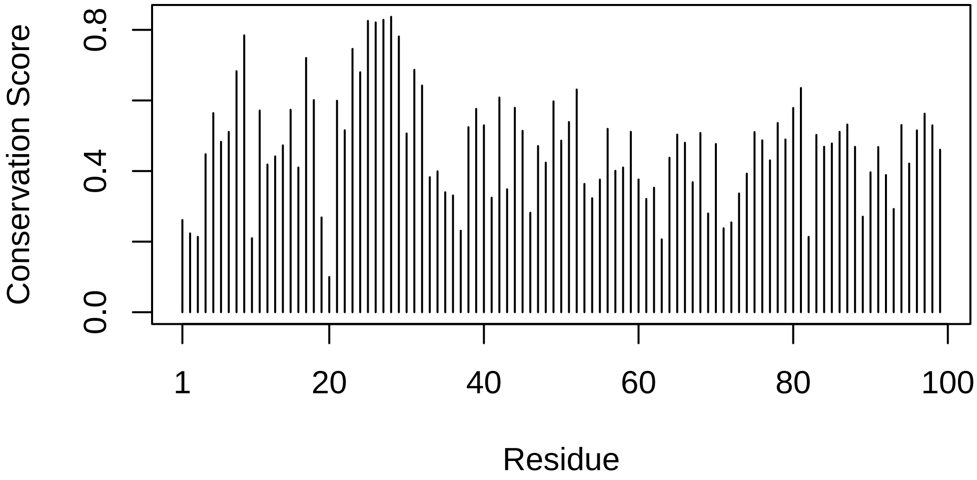
```
dim(aln$ali)
```

```
[1] 5397  132
```

We can score residue conservation in the alignment with the conserv() function.

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

```
plotb3(sim[1:99], ylab="Conservation Score")
```



Note the conserved Active Site residues D25, T26, G27, A28. These positions will stand out if we generate a consensus sequence with a high cutoff value:

```
con <- consensus(aln, cutoff = 0.9)
con$seq
```

```
 [1] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[19] "-" "-" "-" "-" "-" "-" "D" "T" "G" "A" "-" "-" "-" "-" "-" "-" "-" "-"
[37] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[55] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[73] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
```

```
 [91] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[109] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[127] "-" "-" "-" "-" "-" "-"
```

```
m1.pdb <- read.pdb(pdb_files[1])
occ <- vec2resno(c(sim[1:99], sim[1:99]), m1.pdb$atom$resno)
write.pdb(m1.pdb, o=occ, file="m1_conserv.pdb")
```