Structural Bioinformatics (pt2)

Yerin Go A16272901

AlphaFold has changed the game for protein structure prediction and allows anyone with bioinfo skills to predict the structure of virtually any protein

We ran via GoogleColab at: https://github.com/sokrypton/ColabFold

In particular we used their AlphaFold2_mmseq2 version that uses mmseq2 rather than HMMer for sequence search.

The main outputs include a set of **PDB structure** along with matching **JSON format files** that tell us how good the resulting models might be.

Lets start by loading these structures in Mol*

```
library(bio3d)

# Change this for YOUR results dir name
results_dir <- "test_94b5b"
```

```
# 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] "test_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb"
[2] "test_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb"
[3] "test_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb"
[4] "test_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb"
[5] "test_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb"
```

```
library(bio3d)

# Read all data from Models
#  and superpose/fit coords
pdbs <- pdbaln(pdb_files, fit=TRUE, exefile="msa")
```

```
Reading PDB files:
test_94b5b/test_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb
test_94b5b/test_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb
test_94b5b/test_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb
test_94b5b/test_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb
test_94b5b/test_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb
.....

Extracting sequences

pdb/seq: 1   name: test_94b5b/test_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.p
pdb/seq: 2   name: test_94b5b/test_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.p
pdb/seq: 3   name: test_94b5b/test_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.p
pdb/seq: 4   name: test_94b5b/test_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.p
pdb/seq: 5   name: test_94b5b/test_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.p
```

```
pdbs
```

```
                                 1        .         .         .         .         50
[Truncated_Name:1]test_94b5b   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:2]test_94b5b   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:3]test_94b5b   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:4]test_94b5b   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
[Truncated_Name:5]test_94b5b   PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGI
                               **************************************************
                                 1        .         .         .         .         50

                                51        .         .         .         .        99
[Truncated_Name:1]test_94b5b   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:2]test_94b5b   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:3]test_94b5b   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:4]test_94b5b   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:5]test_94b5b   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
                               **************************************************
```

```
                                51        .         .         .         .        99

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

Class:
  pdbs, fasta

Alignment dimensions:
  5 sequence rows; 99 position columns (99 non-gap, 0 gap)

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

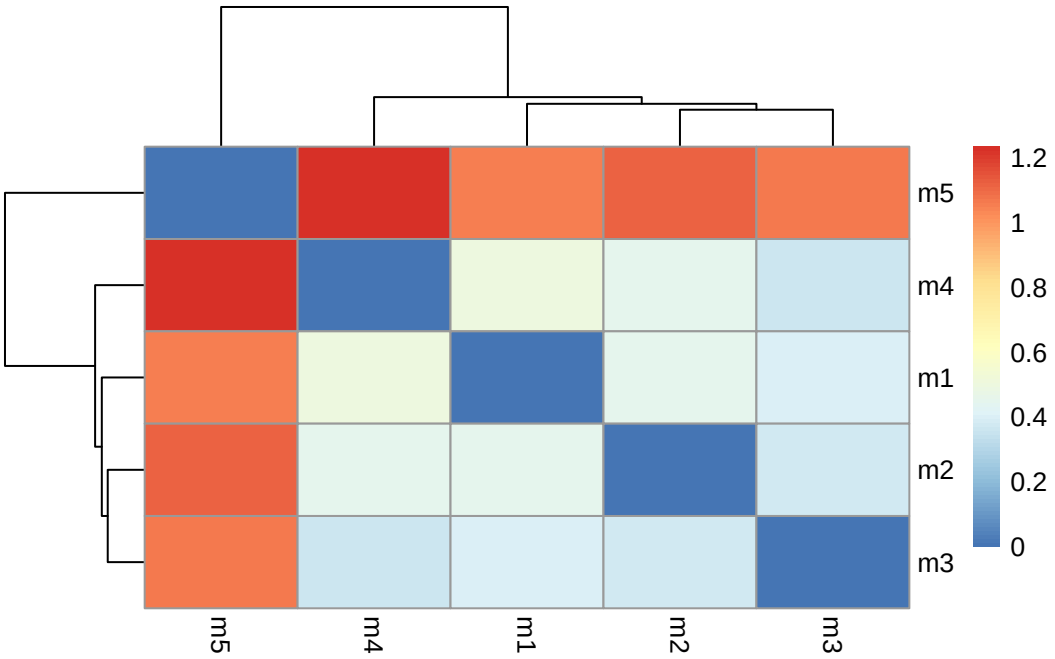
```
rd <- rmsd(pdbs, fit=T)
```

```
Warning in rmsd(pdbs, fit = T): No indices provided, using the 99 non NA positions
```

```
range(rd)
```

```
[1] 0.000 1.233
```

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



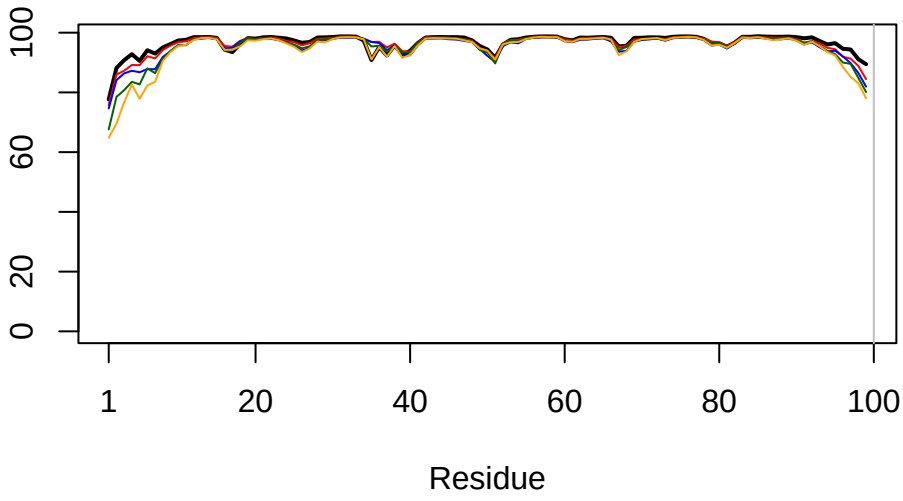

```
# Read a reference PDB structure
pdb <- read.pdb("1hsg")
```

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

```
plotb3(pdbs$b[1,], typ="l", lwd=2, sse=pdb)
```

```
Warning in plotb3(pdbs$b[1, ], typ = "l", lwd = 2, sse = pdb): Length of input
'sse' does not equal the length of input 'x'; Ignoring 'sse'
```

```
points(pdbs$b[2,], typ="l", col="red")
points(pdbs$b[3,], typ="l", col="blue")
points(pdbs$b[4,], typ="l", col="darkgreen")
points(pdbs$b[5,], typ="l", col="orange")
abline(v=100, col="gray")
```




```
core <- core.find(pdbs)
```

```
 core size 98 of 99  vol = 3.577
 core size 97 of 99  vol = 2.759
 core size 96 of 99  vol = 2.244
 core size 95 of 99  vol = 1.736
 core size 94 of 99  vol = 1.353
 core size 93 of 99  vol = 1.058
 core size 92 of 99  vol = 0.833
 core size 91 of 99  vol = 0.601
 core size 90 of 99  vol = 0.406
 FINISHED: Min vol ( 0.5 ) reached
```

```
core.inds <- print(core, vol=0.5)
```

```
# 91 positions (cumulative volume <= 0.5 Angstrom^3)
  start end length
1     2   2      1
2     7  96     90
```

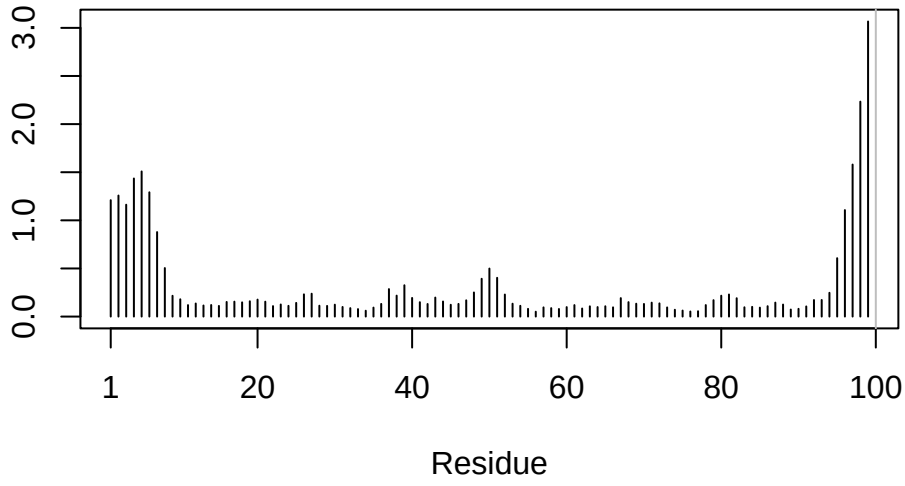
```
xyz <- pdbfit(pdbs, core.inds, outpath="corefit_structures")
```

```
rf <- rmsf(xyz)

plotb3(rf, sse=pdb)
```

```
Warning in plotb3(rf, sse = pdb): Length of input 'sse' does not equal the
length of input 'x'; Ignoring 'sse'
```

```
abline(v=100, col="gray", ylab="RMSF")
```



If the predicted model has more than one domain, each domain may have higher confidence, yet the relative positions of the domain may not. The estimated reliability of relative domain positions is in graphs of predicted aligned errors (PAE) which are included in the downloaded zip files and analyzed in R above.