Class 11: AlphaFold

Renny Ng (A98061553)

AlphaFold is a cool new bioinformatics method for structure prediction from sequence.

We can run AlphaFold on our own computers without installing it or we can run on GoogleCo-lab (without need to install anything) via: https://colab.research.google.com/github/sokrypton/ColabFold/blob

Here's a function to list files:

```
pth <- "hivdimer_23119/"

list.files(path=pth)
```

```
 [1] "cite.bibtex"
 [2] "config.json"
 [3] "hivdimer_23119_coverage.png"
 [4] "hivdimer_23119_env"
 [5] "hivdimer_23119_pae.png"
 [6] "hivdimer_23119_plddt.png"
 [7] "hivdimer_23119_predicted_aligned_error_v1.json"
 [8] "hivdimer_23119_scores_rank_001_alphafold2_multimer_v3_model_5_seed_000.json"
 [9] "hivdimer_23119_scores_rank_002_alphafold2_multimer_v3_model_1_seed_000.json"
[10] "hivdimer_23119_scores_rank_003_alphafold2_multimer_v3_model_4_seed_000.json"
[11] "hivdimer_23119_scores_rank_004_alphafold2_multimer_v3_model_2_seed_000.json"
[12] "hivdimer_23119_scores_rank_005_alphafold2_multimer_v3_model_3_seed_000.json"
[13] "hivdimer_23119_unrelaxed_rank_001_alphafold2_multimer_v3_model_5_seed_000.pdb"
[14] "hivdimer_23119_unrelaxed_rank_002_alphafold2_multimer_v3_model_1_seed_000.pdb"
[15] "hivdimer_23119_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000.pdb"
[16] "hivdimer_23119_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000.pdb"
[17] "hivdimer_23119_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000.pdb"
[18] "hivdimer_23119.a3m"
[19] "hivdimer_23119.csv"
[20] "hivdimer_23119.done.txt"
[21] "log.txt"
```

The multiple sequence alignment (MSA) is contained in the "a3m" file of our AlphaFold output.

Here we only want to open the a3m file, and we want the full path as well (the location in its current directory). Add additional arguments for "pattern" and for "full.names".

```
aln.file <- list.files(path=pth, pattern = ".a3m", full.names=TRUE)
```

```
library(bio3d)
aln <- read.fasta(aln.file, to.upper=TRUE)
```

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

```
attributes(aln)
```

```
$names
[1] "id"   "ali"  "call"

$class
[1] "fasta"
```

This is a large alignment (almost too big to really look at).

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

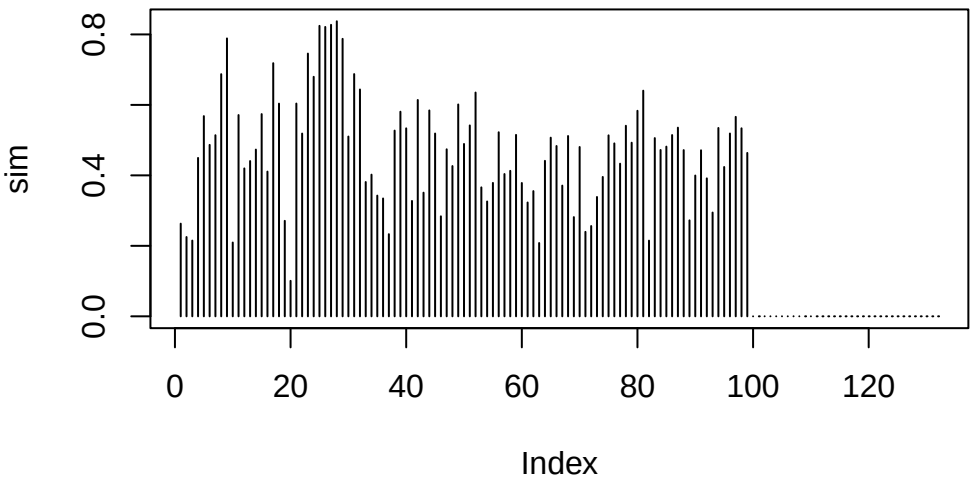
```
[1] 5378  132
```

There are 5,378 sequences, and 132 positions.

Let's calculate summary summary info such as conservation scores (if you go down these 132 columns, shows which are the most conserved position).

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

```
plot(sim, typ="h")
```



We can summarize these conserved columns (those with high scores above) via a consensus sequence. This will reveal the most conserved amino acids which are likely essential for normal function.

```
consensus(aln, cutoff = 0.9)$seq
```

```
  [1] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
 [19] "-" "-" "-" "-" "-" "-" "D" "T" "G" "A" "-" "-" "-" "-" "-" "-" "-" "-"
 [37] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
 [55] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
 [73] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
 [91] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[109] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[127] "-" "-" "-" "-" "-" "-"
```

Now we can read in our structure model to see where the D" "T" "G" "A" motif is located. Read the predicted alignment error (PAE) files into R to make sense of these different multichain models. These are stored as JSON format. Reading these require the "jsonlite" package.

```
library(jsonlite)
```

Now we must find our JSON files.

```
list.files(path=pth, pattern = "000.json", full.names=TRUE)
```

```
[1] "hivdimer_23119//hivdimer_23119_scores_rank_001_alphafold2_multimer_v3_model_5_seed_000.j
[2] "hivdimer_23119//hivdimer_23119_scores_rank_002_alphafold2_multimer_v3_model_1_seed_000.j
[3] "hivdimer_23119//hivdimer_23119_scores_rank_003_alphafold2_multimer_v3_model_4_seed_000.j
[4] "hivdimer_23119//hivdimer_23119_scores_rank_004_alphafold2_multimer_v3_model_2_seed_000.j
[5] "hivdimer_23119//hivdimer_23119_scores_rank_005_alphafold2_multimer_v3_model_3_seed_000.j
```

pae1 <- read_json(pae.files[1], simplifyVector = TRUE) pae5 <- read_json(pae.files[5], simplifyVector = TRUE)