AlphaFold Analysis

Yaniv Iny (PID: A18090586)

Here we analyze our AlphaFold structure prediction models. The input directory/folder comes from the ColabFold server:

```
# Change this for YOUR results dir name
results_dir <- "CORRECTHIVMODEL_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] "CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb"
[2] "CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb"
[3] "CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb"
[4] "CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb"
[5] "CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb"
```

I will use the Bio3D package for analysis & Align and superpose

```
library(bio3d)

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

```
Reading PDB files:
CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_00
```

```
CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_00
CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_00
CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_00
CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_00
.....

Extracting sequences

pdb/seq: 1   name: CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_
pdb/seq: 2   name: CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_
pdb/seq: 3   name: CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_
pdb/seq: 4   name: CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_
pdb/seq: 5   name: CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_
```

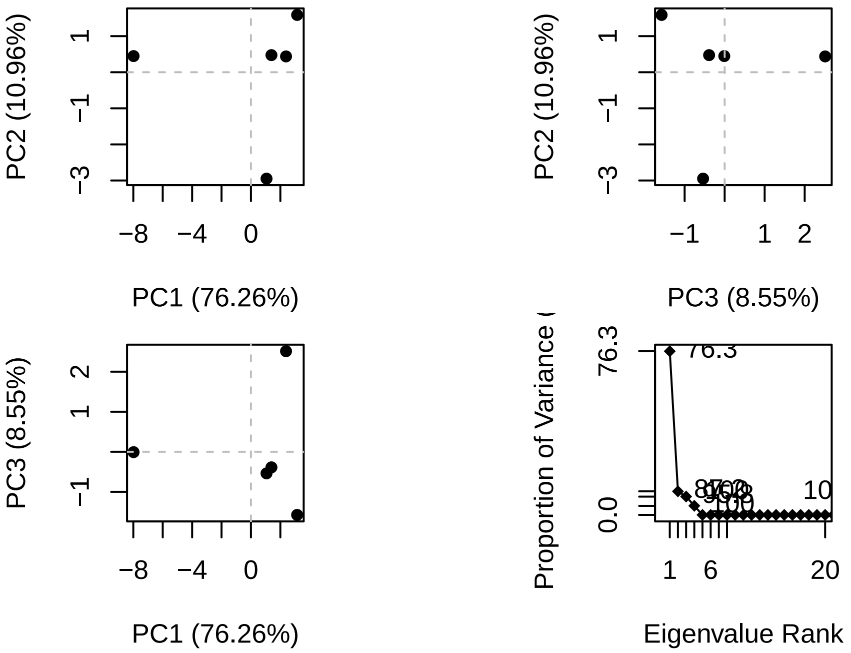
```
pdbs
```

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

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

A quick PCA plot

```r
pc <- pca(pdbs)
plot(pc)
```



RMSD Analysis

RMSD is a common measure of structural distance used in structural biology

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

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

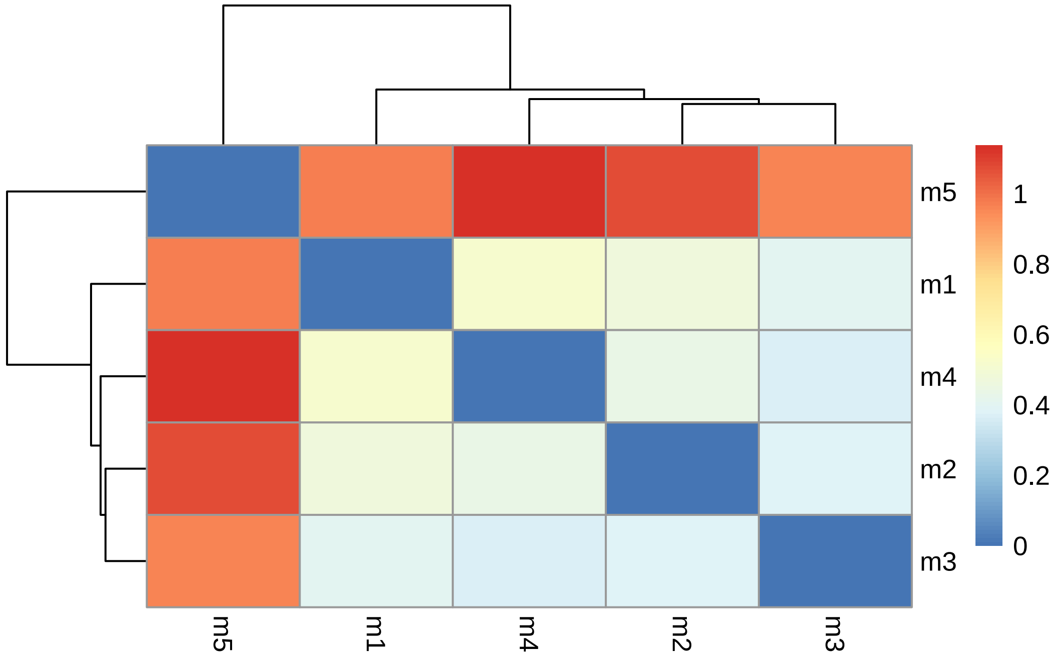
```r
rd
```

```
                                                                         CORRECTHIVMODEL_94b5
CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
```

```
CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         CORRECTHIVMODEL_94b5
CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         CORRECTHIVMODEL_94b5
CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         CORRECTHIVMODEL_94b5
CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         CORRECTHIVMODEL_94b5
CORRECTHIVMODEL_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
CORRECTHIVMODEL_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
```

,

```r
library(pheatmap)

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



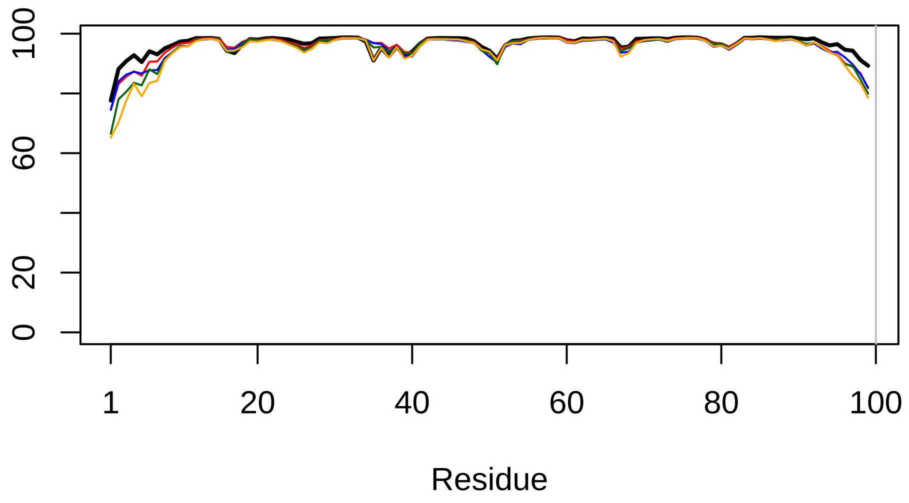
```r
# Read a reference PDB structure
pdb <- read.pdb("1hsg")
```

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

```r
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'
```

```r
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")
```



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

```
 core size 98 of 99  vol = 3.583
 core size 97 of 99  vol = 2.722
 core size 96 of 99  vol = 2.217
 core size 95 of 99  vol = 1.713
 core size 94 of 99  vol = 1.299
 core size 93 of 99  vol = 0.944
 core size 92 of 99  vol = 0.722
 core size 91 of 99  vol = 0.531
 core size 90 of 99  vol = 0.389
 FINISHED: Min vol ( 0.5 ) reached
```

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

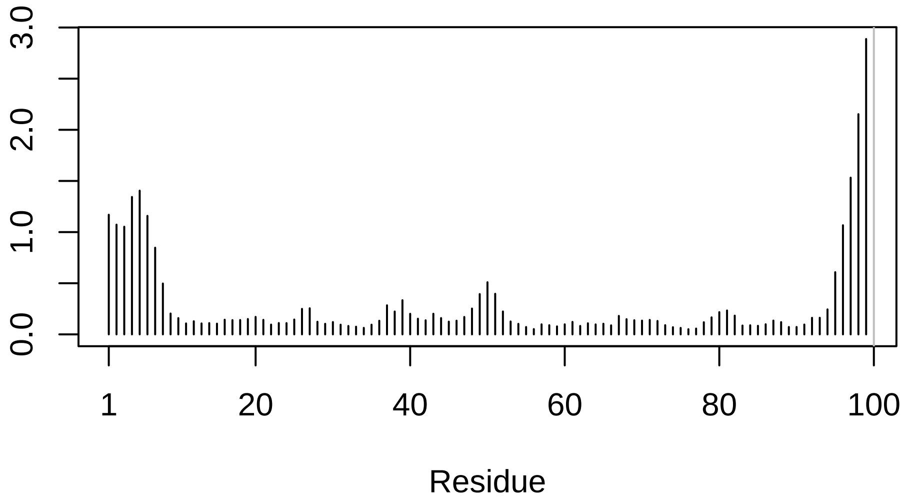
```
# 91 positions (cumulative volume <= 0.5 Angstrom^3)
  start end length
1     3   3      1
2     7  96     90
```

```
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")
```



Predicted Alignment error for domains

```
library(jsonlite)

# Listing of all PAE JSON files
pae_files <- list.files(path=results_dir,
                        pattern=".*model.*\\.json",
                        full.names = TRUE)
```

```r
pae1 <- read_json(pae_files[1],simplifyVector = TRUE)
pae5 <- read_json(pae_files[5],simplifyVector = TRUE)

attributes(pae1)
```

```
$names
[1] "plddt"   "max_pae" "pae"     "ptm"
```

```r
# Per-residue pLDDT scores
#  same as B-factor of PDB..
head(pae1$plddt)
```

```
[1] 77.62 88.19 90.81 92.81 90.50 94.12
```

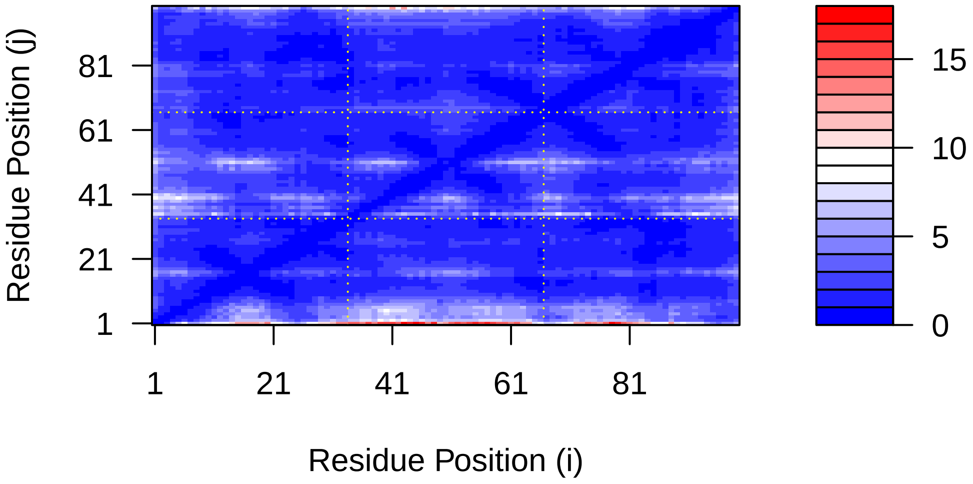
```r
pae1$max_pae
```

```
[1] 17.8125
```

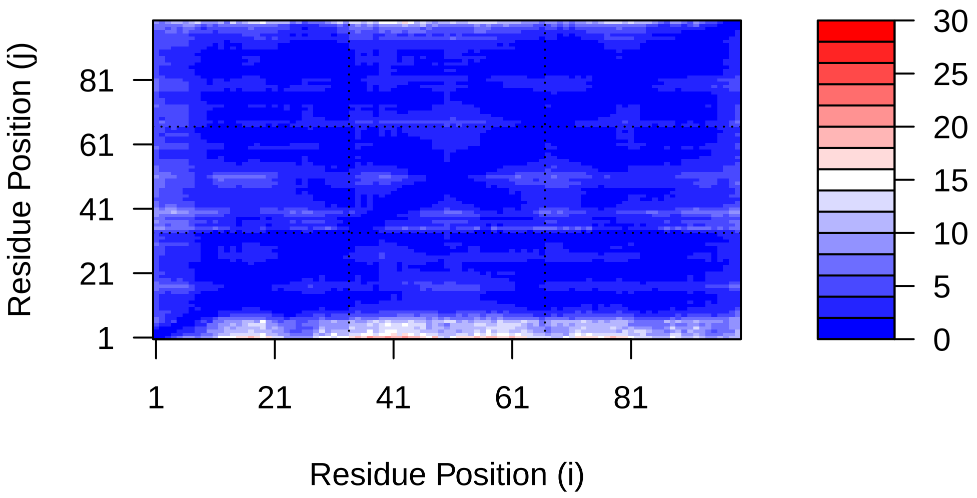
```r
pae5$max_pae
```

```
[1] 20.3125
```

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



```r
plot.dmat(pae5$pae,
          xlab="Residue Position (i)",
          ylab="Residue Position (j)",
          grid.col = "black",
          zlim=c(0,30))
```



Reisudue conservation from alignment file

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

```
[1] "CORRECTHIVMODEL_94b5b/CORRECTHIVMODEL_94b5b.a3m"
```

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

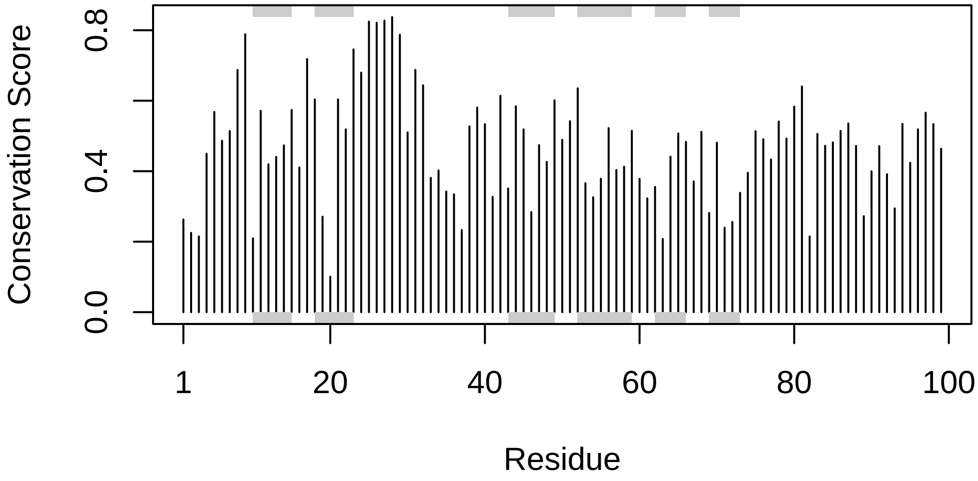
```
[1] " ** Duplicated sequence id's: 101 **"
```

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

```
[1] 5378  132
```

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

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



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

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