AlphaFold Analysis

Sion Kang (PID: A17052234)

2025-02-11

Table of contents

Custom Results of resulting models

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

```
results_dir <- "hivpr_monomer_94b5b_0/"
```

```
# 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_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb"
[2] "hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb"
[3] "hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb"
[4] "hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb"
[5] "hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb"
```

I will use the Bio3D package for analysis

```
library(bio3d)
```

Align and superpose

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

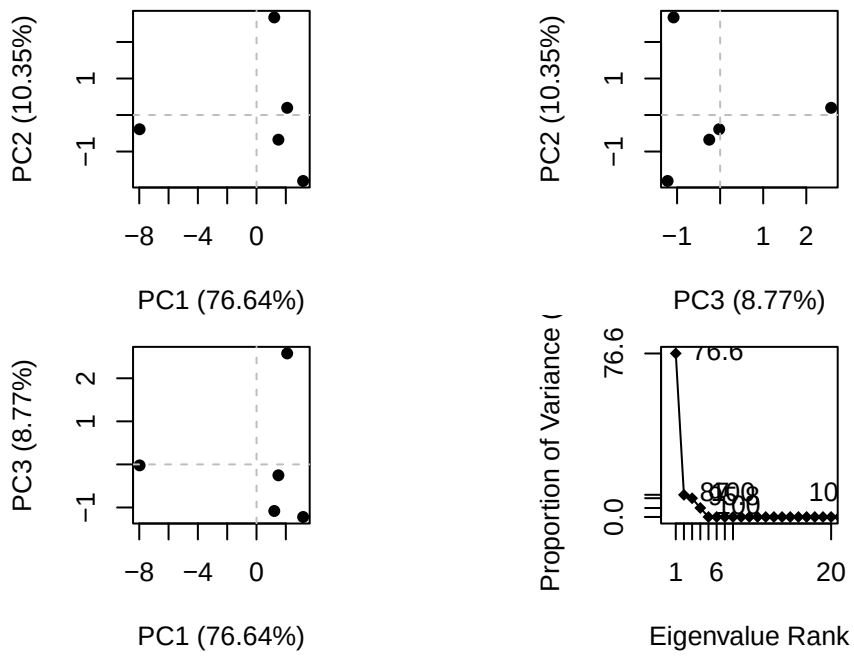
```
Reading PDB files:
hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_
hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_
hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_
hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_
hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_
.....

Extracting sequences

pdb/seq: 1   name: hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold
pdb/seq: 2   name: hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold
pdb/seq: 3   name: hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold
pdb/seq: 4   name: hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold
pdb/seq: 5   name: hivpr_monomer_94b5b_0//hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold
```

A quick PCA

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



RMSD analysis

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

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

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

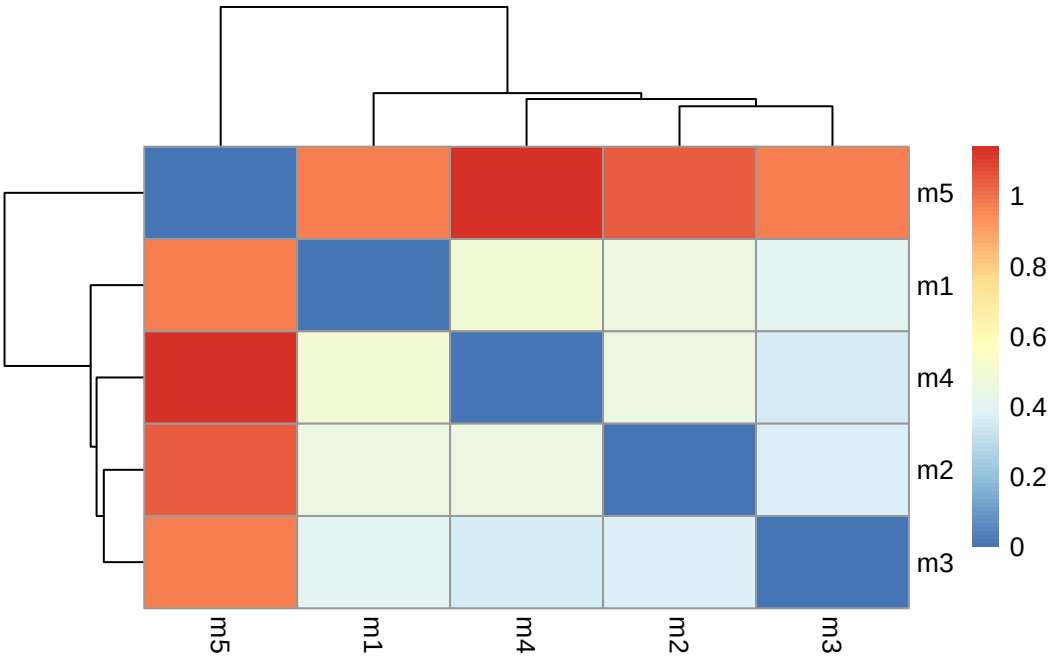
```
rd
```

```
                                                                    hivpr_monomer_94b5b_
hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                    hivpr_monomer_94b5b_
hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
```

```
hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         hivpr_monomer_94b5b_
hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         hivpr_monomer_94b5b_
hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                         hivpr_monomer_94b5b_
hivpr_monomer_94b5b_0_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivpr_monomer_94b5b_0_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivpr_monomer_94b5b_0_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)
```



Analysis of AlphaFold structure prediction models for my Find a Gene Project

Used ColabFold to generate a model for my structure of interest for the Find a Gene Project.

```
tasl_dir <- "findagene_TASL_a2aa2/"
```

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

# Print our PDB file names
basename(taslpdbfiles)
```

```
[1] "findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb"
[2] "findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000.pdb"
[3] "findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000.pdb"
[4] "findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000.pdb"
[5] "findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb"
```

Align and superpose

```
taslpdbs <- pdbaln(taslpdbfiles, fit=TRUE, exefile="msa")
```

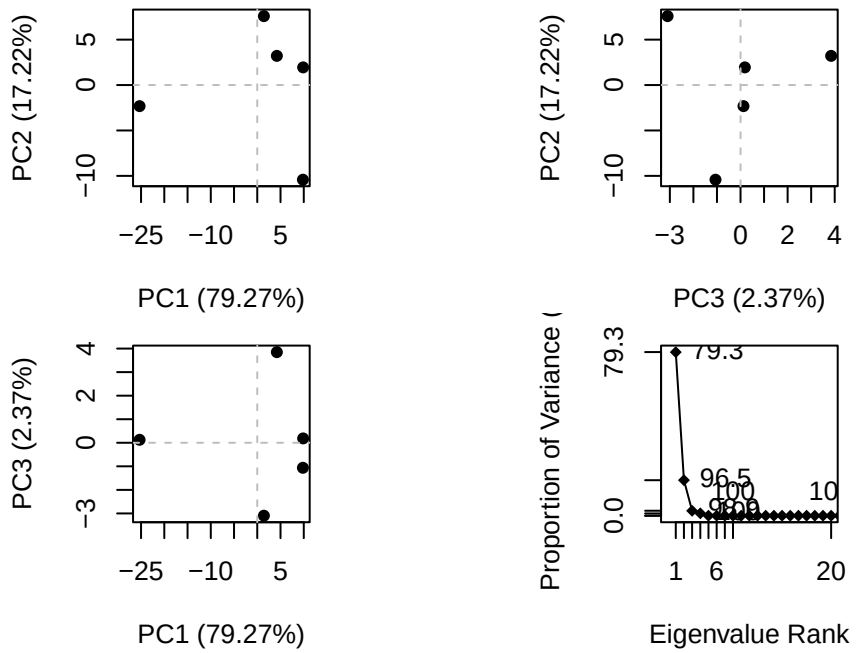
```
Reading PDB files:
findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
.....

Extracting sequences

pdb/seq: 1   name: findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_p
pdb/seq: 2   name: findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_p
pdb/seq: 3   name: findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_p
pdb/seq: 4   name: findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_p
pdb/seq: 5   name: findagene_TASL_a2aa2//findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_p
```

A quick PCA

```
tasl.pc <- pca(taslpdbs)
plot(tasl.pc)
```



RMSD analysis

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

```
taslrd <- rmsd(taslpdbs, fit=T)
```

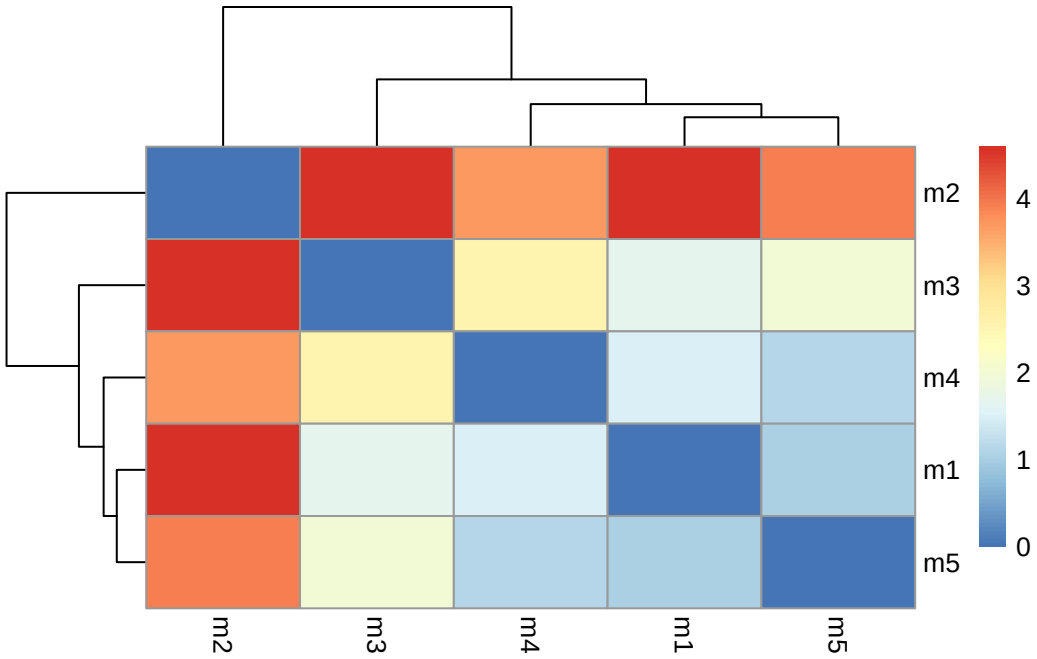
```
Warning in rmsd(taslpdbs, fit = T): No indices provided, using the 60 non NA positions
```

```
taslrd
```

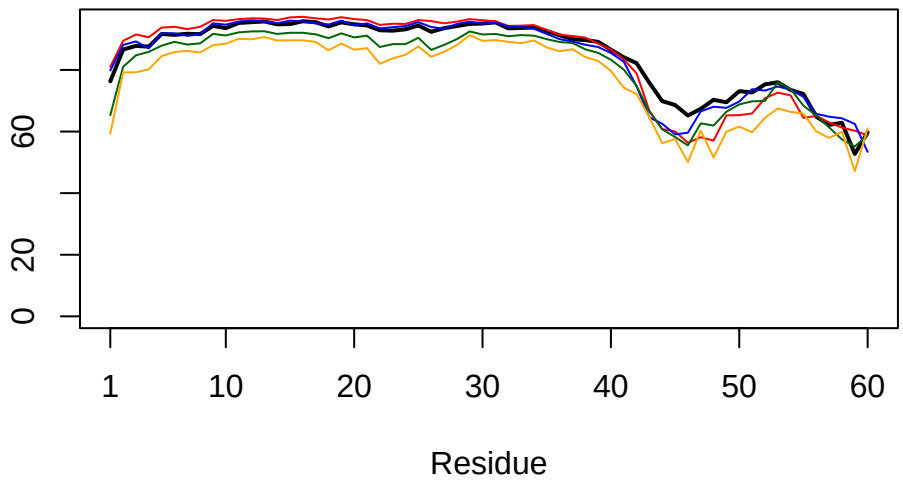
```
                                                                    findagene_TASL_a2aa2_
findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                    findagene_TASL_a2aa2_
findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                    findagene_TASL_a2aa2_
findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                    findagene_TASL_a2aa2_
findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                    findagene_TASL_a2aa2_
findagene_TASL_a2aa2_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_002_alphafold2_ptm_model_3_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_004_alphafold2_ptm_model_1_seed_000
findagene_TASL_a2aa2_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
```

```
library(pheatmap)

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




```
plotb3(taslpdbs$b[1,], typ="l", lwd=2)
points(taslpdbs$b[2,], typ="l", col="red")
points(taslpdbs$b[3,], typ="l", col="blue")
points(taslpdbs$b[4,], typ="l", col="darkgreen")
points(taslpdbs$b[5,], typ="l", col="orange")
abline(v=100, col="gray")
```




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

```
 core size 59 of 60  vol = 70.003
 core size 58 of 60  vol = 55.056
 core size 57 of 60  vol = 53.513
 core size 56 of 60  vol = 49.883
 core size 55 of 60  vol = 46.861
 core size 54 of 60  vol = 43.317
 core size 53 of 60  vol = 39.188
 core size 52 of 60  vol = 35.499
 core size 51 of 60  vol = 29.9
 core size 50 of 60  vol = 24.083
 core size 49 of 60  vol = 17.676
 core size 48 of 60  vol = 13.991
 core size 47 of 60  vol = 7.669
 core size 46 of 60  vol = 2.502
 core size 45 of 60  vol = 1.329
 core size 44 of 60  vol = 0.483
 FINISHED: Min vol ( 0.5 ) reached
```

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

```
# 45 positions (cumulative volume <= 0.5 Angstrom^3)
  start end length
1     1  45     45
```

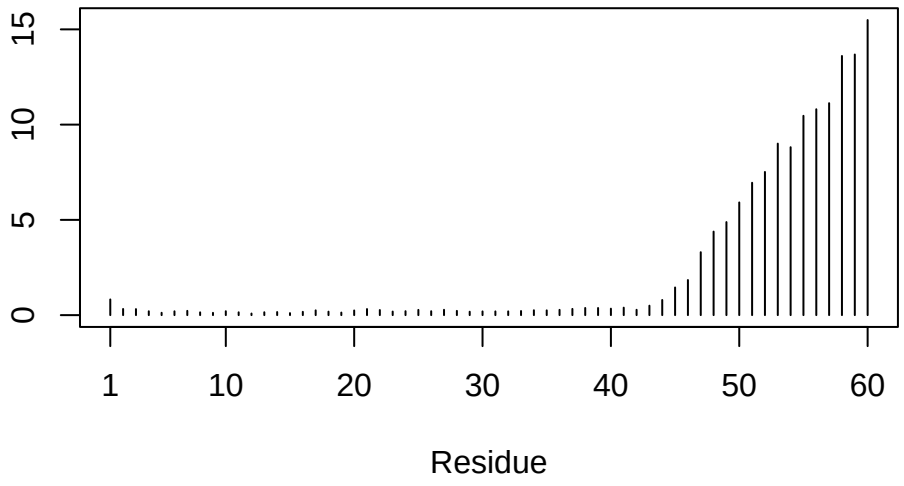
```
xyz <- pdbfit(taslpdbs, core.inds, outpath="corefit_structures")
```

```
rf <- rmsf(xyz)

plotb3(rf, sse=taslpdbs)
```

```
Warning in plotb3(rf, sse = taslpdbs): 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=tasl_dir,
                        pattern=".*model.*\\.json",
                        full.names = TRUE)
```

```
pae1 <- read_json(pae_files[1],simplifyVector = TRUE)
pae2 <- read_json(pae_files[2],simplifyVector = TRUE)
pae3 <- read_json(pae_files[3],simplifyVector = TRUE)
pae4 <- read_json(pae_files[4],simplifyVector = TRUE)
pae5 <- read_json(pae_files[5],simplifyVector = TRUE)

attributes(pae1)
```

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

```
pae1$max_pae
```

```
[1] 30.4375
```

```
pae2$max_pae
```

```
[1] 30.54688
```

```
pae3$max_pae
```

```
[1] 29.98438
```

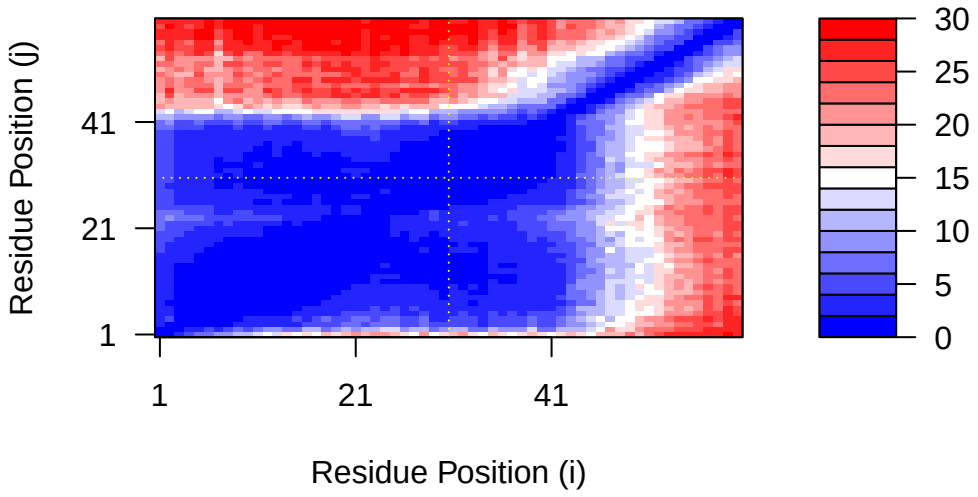
```
pae4$max_pae
```

```
[1] 30.70312
```

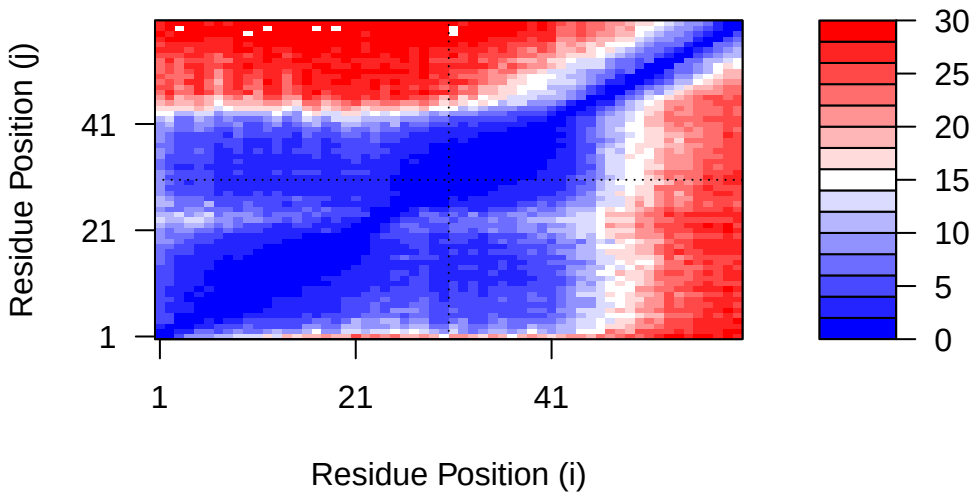
```
pae5$max_pae
```

```
[1] 30.32812
```

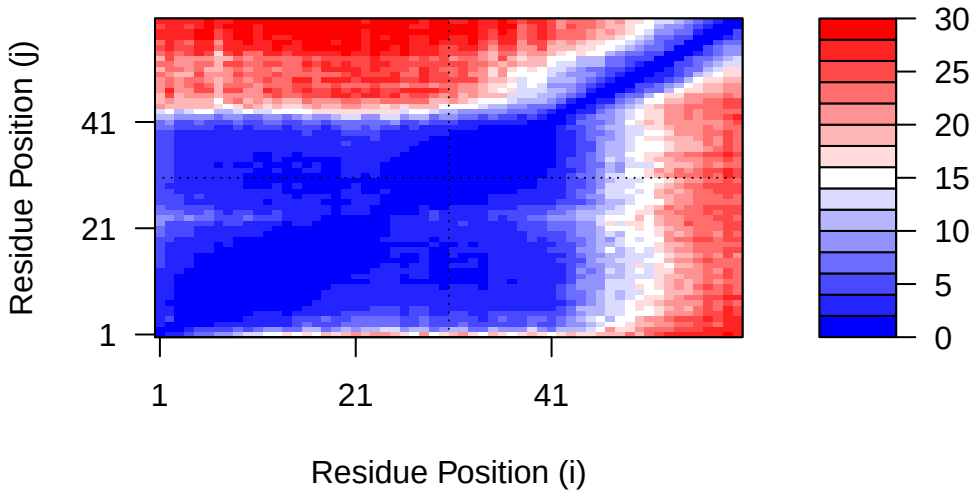
```
#plot of the N by N PAE scores of the model with the lowest max PAE score
plot.dmat(pae3$pae,
          xlab="Residue Position (i)",
          ylab="Residue Position (j)")
```




```
#plot of the N by N PAE scores of the model with the highest max PAE score
plot.dmat(pae5$pae,
          xlab="Residue Position (i)",
          ylab="Residue Position (j)",
          grid.col = "black",
          zlim=c(0,30))
```




```
plot.dmat(pae3$pae,
          xlab="Residue Position (i)",
          ylab="Residue Position (j)",
          grid.col = "black",
          zlim=c(0,30))
```



Residue conservation from alignment file

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

```
[1] "findagene_TASL_a2aa2//findagene_TASL_a2aa2.a3m"
```

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

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

How many sequences are in this alignment

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

```
[1] 487  76
```

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

```
 [1]  1.493447e-01  5.951412e-01  3.629030e-01  6.720714e-01  4.315470e-01
 [6]  5.394065e-01  5.363027e-01  3.298426e-01  4.483095e-01  5.389696e-01
[11]  5.285277e-01  5.551567e-01  8.614597e-01  7.974455e-01  5.502531e-01
[16]  3.820629e-01  7.542509e-01  5.088701e-01  9.157359e-01  5.383063e-01
[21]  9.007411e-01  3.011163e-01  7.316577e-01  4.550739e-01  7.808249e-01
[26]  4.681285e-01  2.897026e-01  4.317861e-01  8.630390e-01  6.482512e-01
[31]  6.424181e-01  3.982694e-01  5.890647e-01  5.842650e-01  3.251527e-01
[36]  3.011002e-01  7.662256e-01  4.302321e-01  2.227850e-01  2.605285e-01
[41]  3.636102e-01  3.211702e-01  1.099501e-01  6.446625e-02  1.243305e-01
[46]  5.977328e-01  2.912752e-01  2.073415e-01  1.947896e-01  5.574518e-01
[51]  3.801286e-01  2.404467e-01  1.466280e-01  2.678590e-01  2.080876e-01
[56]  2.673410e-01  4.152382e-01  6.767900e-02  7.052163e-02  7.161085e-02
[61]  7.466390e-02  3.244860e-03  1.902975e-03  3.287111e-04  7.605141e-05
[66]  1.774533e-05  4.225078e-06 -1.690031e-06  0.000000e+00  0.000000e+00
[71]  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00  0.000000e+00
[76]  0.000000e+00
```

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

```
 [1] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "Q" "-" "-" "-" "-" "-" "S"
[20] "-" "E" "-" "-" "-" "-" "-" "-" "-" "A" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[39] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
[58] "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-" "-"
```

```
m1.pdb <- read.pdb(taslpdbfiles[1])
#occ <- vec2resno(c(sim[1:60], sim[1:60]), m1.pdb$atom$resno)
#write.pdb(m1.pdb, o=occ, file="m1_conserv.pdb")

# Unable to run code. Original example using HIV-Pr monomer also does not work (see below)
```

```
aln_file1 <- list.files(path=results_dir,
                        pattern=".a3m$",
                         full.names = TRUE)
aln1 <- read.fasta(aln_file1[1], to.upper = TRUE)
```

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

```
dim(aln1$ali)
```

```
[1] 5378  132
```

```
sim1 <- conserv(aln1)
con1 <- consensus(aln1, cutoff = 0.9)
con1$seq
```

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

```
m1.pdb1 <- read.pdb(pdb_files[1])
#occ1 <- vec2resno(c(sim1[1:99], sim1[1:99]), m1.pdb1$atom$resno)
#write.pdb(m1.pdb1, o=occ1, file="m1_conserv1.pdb")

# This is the original code, which gives " Error in vec2resno(c(sim1[1:99], sim1[1:99]), m1.
```