class11-find-a-gene-project

Aaron (PID A17544470)

Table of contents

HIV-Pr-Monomer

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

8a. Custom analysis of resulting models

```
# Change this for YOUR results dir name
results_dir <- "hivprmonomer_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] "hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb"
[2] "hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb"
[3] "hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb"
[4] "hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb"
[5] "hivprmonomer_94b5b_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")
```

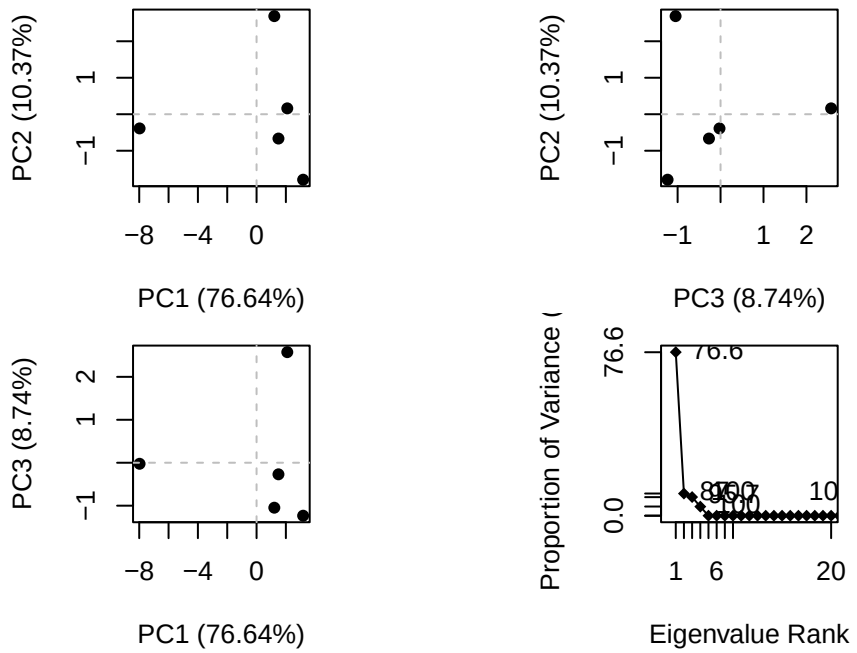
```
Reading PDB files:
hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000.pdb
hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000.pdb
hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000.pdb
hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb
hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb
.....

Extracting sequences

pdb/seq: 1   name: hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_m
pdb/seq: 2   name: hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_m
pdb/seq: 3   name: hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_m
pdb/seq: 4   name: hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_m
pdb/seq: 5   name: hivprmonomer_94b5b//hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_m
```

```
pdbs
```

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

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

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



RMSD analysis

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

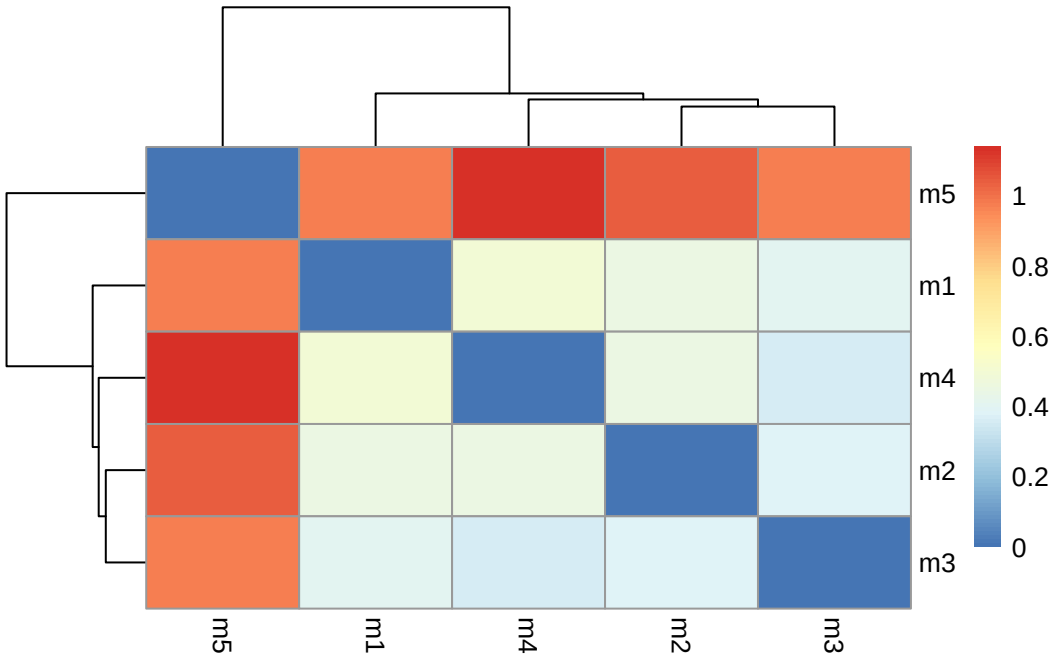
```r
rd <- rmsd(pdbs, fit = T)
```

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

```r
rd
```

```
                                                                   hivprmonomer_94b5b_unre
hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                   hivprmonomer_94b5b_unre
hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                   hivprmonomer_94b5b_unre
hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                   hivprmonomer_94b5b_unre
hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
                                                                   hivprmonomer_94b5b_unre
hivprmonomer_94b5b_unrelaxed_rank_001_alphafold2_ptm_model_5_seed_000
hivprmonomer_94b5b_unrelaxed_rank_002_alphafold2_ptm_model_4_seed_000
hivprmonomer_94b5b_unrelaxed_rank_003_alphafold2_ptm_model_1_seed_000
hivprmonomer_94b5b_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
hivprmonomer_94b5b_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
```

```
library(pheatmap)

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



HIV-Pr-Dimer

8b. Custom analysis of resulting models

```
# Change this for YOUR results dir name
results_dir1 <- "HIVPrDimer_23119_0/"
```

```
# File names for all PDB models
pdb_files1 <- list.files(path=results_dir1,
                         pattern="*.pdb",
                         full.names = TRUE)

# Print our PDB file names
basename(pdb_files1)
```

```
[1] "HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000.pdb"
[2] "HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000.pdb"
[3] "HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000.pdb"
[4] "HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000.pdb"
[5] "HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000.pdb"
```

I will use the Bio3d package for analysis

```
library(bio3d)
```

Align and superpose

```
pdbs1 <- pdbaln(pdb_files1, fit=TRUE, exefile="msa")
```

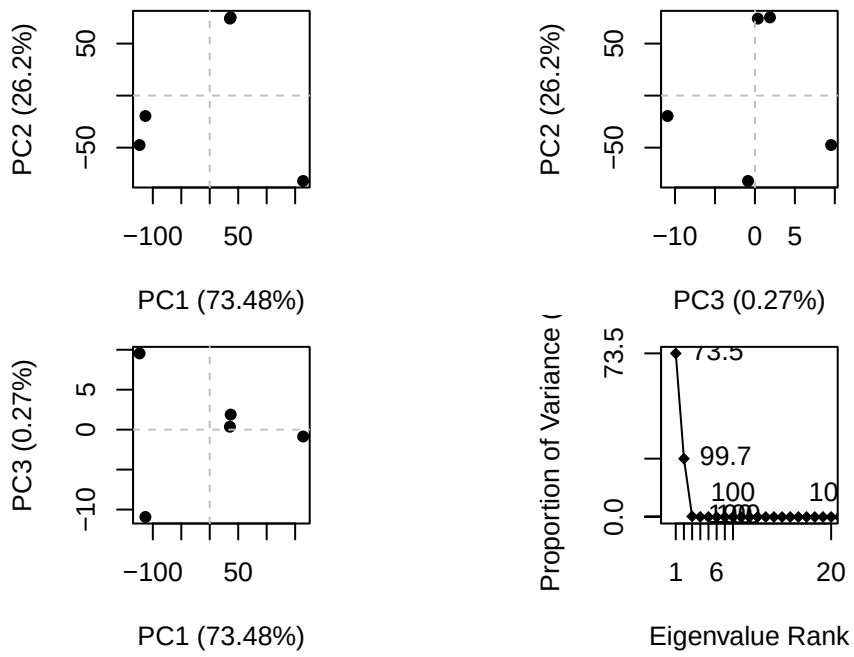
```
Reading PDB files:
HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed
HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed
HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed
HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed
HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed
.....

Extracting sequences

pdb/seq: 1   name: HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multi
pdb/seq: 2   name: HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multi
pdb/seq: 3   name: HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multi
pdb/seq: 4   name: HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multi
pdb/seq: 5   name: HIVPrDimer_23119_0//HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multi
```

```
pdbs1
```

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

```
                             51        .         .         .         .         100
[Truncated_Name:1]HIVPrDimer   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:2]HIVPrDimer   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:3]HIVPrDimer   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:4]HIVPrDimer   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
[Truncated_Name:5]HIVPrDimer   GGFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFP
                               **************************************************
                             51        .         .         .         .         100

                            101        .         .         .         .         150
[Truncated_Name:1]HIVPrDimer   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:2]HIVPrDimer   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:3]HIVPrDimer   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:4]HIVPrDimer   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
[Truncated_Name:5]HIVPrDimer   QITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIG
                               **************************************************
                            101        .         .         .         .         150

                            151        .         .         .         .       198
[Truncated_Name:1]HIVPrDimer   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:2]HIVPrDimer   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:3]HIVPrDimer   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:4]HIVPrDimer   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
[Truncated_Name:5]HIVPrDimer   GFIKVRQYDQILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNF
                               *************************************************
                            151        .         .         .         .       198

Call:
  pdbaln(files = pdb_files1, fit = TRUE, exefile = "msa")

Class:
  pdbs, fasta

Alignment dimensions:
  5 sequence rows; 198 position columns (198 non-gap, 0 gap)

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

```
pc1 <- pca(pdbs1)
pdbplot1 <- plot(pc1)
```



RMSD analysis

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

```
rd1 <- rmsd(pdbs1, fit = T)
```

```
Warning in rmsd(pdbs1, fit = T): No indices provided, using the 198 non NA positions
```

```
range(rd1)
```

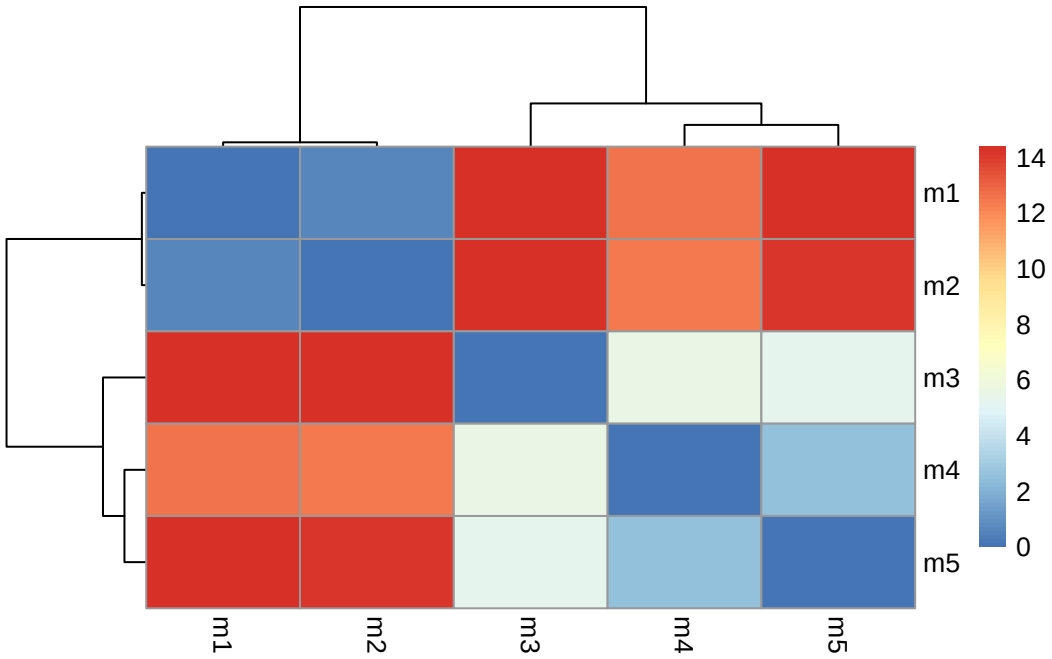
```
[1]  0.000 14.376
```

```
rd1
```

```
                                                                                HIVPrDimer_231
HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000
```

```
                                                                                  HIVPrDimer_2311
HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000
                                                                                  HIVPrDimer_2311
HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000
                                                                                  HIVPrDimer_2311
HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000
                                                                                  HIVPrDimer_2311
HIVPrDimer_23119_0_unrelaxed_rank_001_alphafold2_multimer_v3_model_1_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_002_alphafold2_multimer_v3_model_5_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_003_alphafold2_multimer_v3_model_4_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_004_alphafold2_multimer_v3_model_2_seed_000
HIVPrDimer_23119_0_unrelaxed_rank_005_alphafold2_multimer_v3_model_3_seed_000
```

```r
library(pheatmap)

colnames(rd1) <- paste0("m",1:5)
rownames(rd1) <- paste0("m",1:5)
pheatmap(rd1)
```



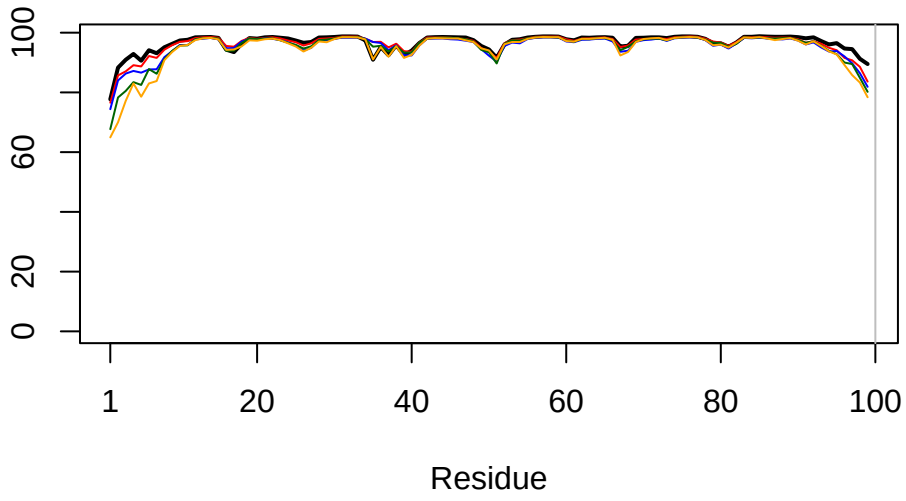

```
# Read a reference PDB structure
pdb <- read.pdb("1hsg")
```

```
  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 = 3.327
 core size 97 of 99  vol = 2.585
 core size 96 of 99  vol = 2.128
 core size 95 of 99  vol = 1.629
 core size 94 of 99  vol = 1.326
 core size 93 of 99  vol = 0.958
 core size 92 of 99  vol = 0.731
 core size 91 of 99  vol = 0.555
 core size 90 of 99  vol = 0.406
 FINISHED: Min vol ( 0.5 ) reached
```

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

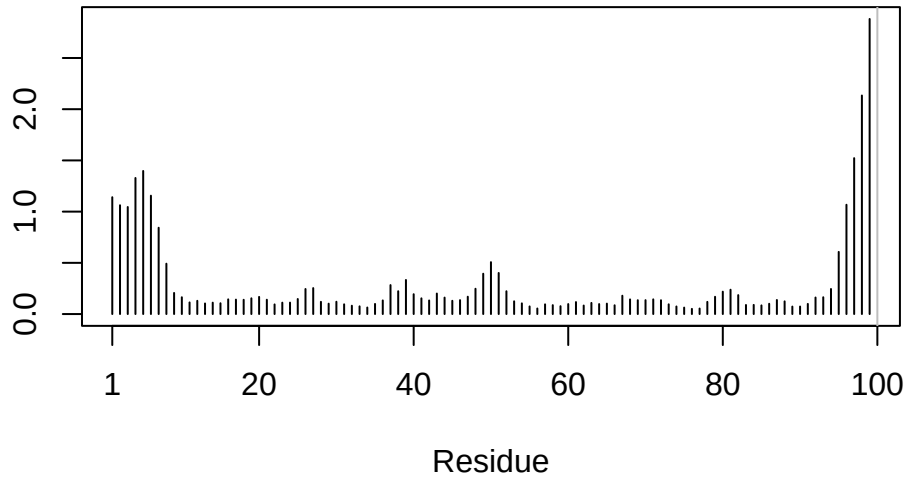
```
# 91 positions (cumulative volume <= 0.5 Angstrom^3)
  start end length
1     3   3      1
2     7  96     90
```

```
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_dir1,
                        pattern=".*model.*\\.json",
                        full.names = TRUE)
```

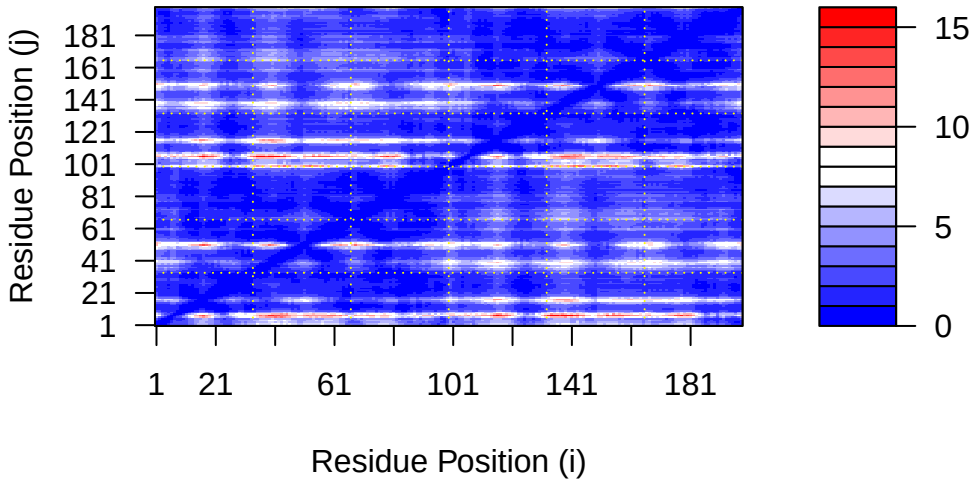
```
pae1 <- read_json(pae_files[1],simplifyVector = TRUE)
pae5 <- read_json(pae_files[5],simplifyVector = TRUE)

attributes(pae1)
```

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

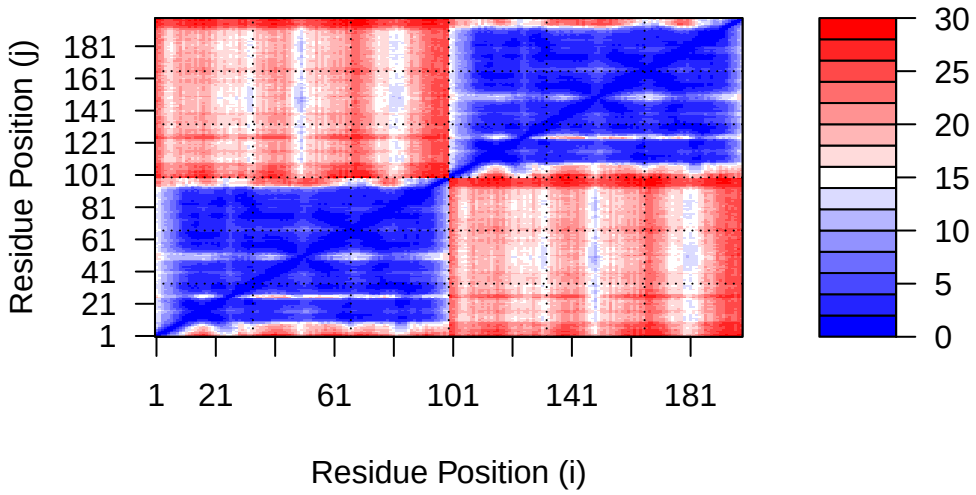
```
# Per-residue pLDDT scores
#  same as B-factor of PDB..
head(pae1$plddt)
```

```
[1] 87.69 90.81 90.38 90.88 93.44 86.06
```

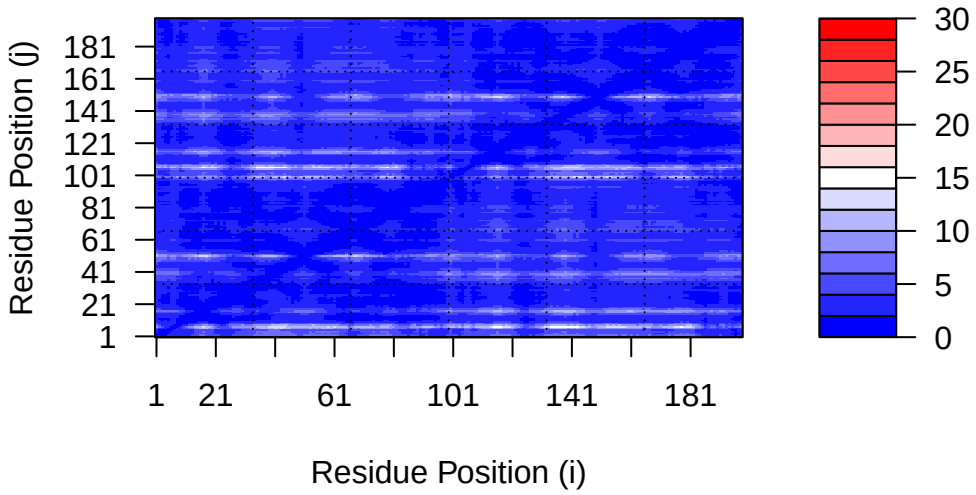
```
plot.dmat(pae1$pae,
          xlab="Residue Position (i)",
          ylab="Residue Position (j)")
```



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




```
plot.dmat(pae1$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=results_dir,
                       pattern=".a3m$",
                        full.names = TRUE)
aln_file
```

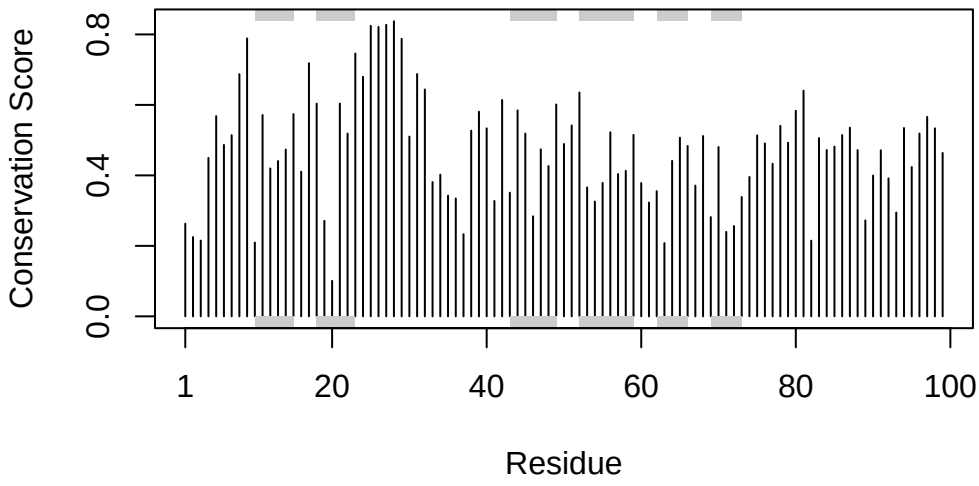
```
[1] "hivprmonomer_94b5b//hivprmonomer_94b5b.a3m"
```

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

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

```
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] "-" "-" "-" "-" "-" "-"
```

```
m1.pdb <- read.pdb(pdb_files1[1])
occ <- vec2resno(c(sim[1:99], sim[1:99]), m1.pdb$atom$resno)
write.pdb(m1.pdb, o=occ, file="m1_conserv.pdb")
```