# **Structural Bioinformatics pt 2**

Emily Hendrickson (PID: A69034780)

First, finish last week's stuff. library(bio3d) Warning: package 'bio3d' was built under R version 4.3.3 id <- "1ake\_A" aa <- get.seq(id)</pre> Warning in get.seq(id): Removing existing file: seqs.fasta Fetching... Please wait. Done. 60 pdb|1AKE|A MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT 61 120 DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI pdb|1AKE|A 180 pdb|1AKE|A VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG 180 181 214

214

YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG

pdb|1AKE|A

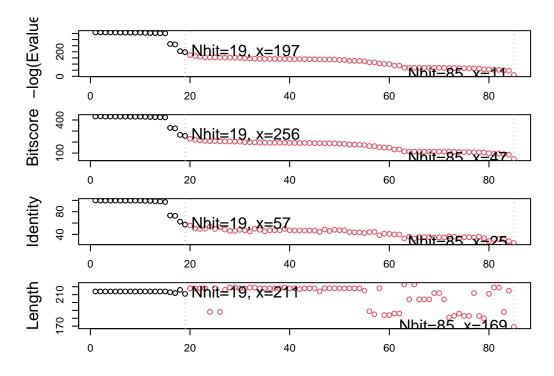
181

\* Possible cutoff values: 197 11

Yielding Nhits: 19 85

\* Chosen cutoff value of: 197

Yielding Nhits: 19



files <- get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE)</pre>

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/1AKE.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/8BQF.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/4X8M.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/6S36.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/8Q2B.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/8RJ9.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/6RZE.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/4X8H.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/3HPR.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/1E4V.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/5EJE.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/1E4Y.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE):
pdbs/3X2S.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/6HAP.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/6HAM.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/4K46.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/4NP6.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/3GMT.pdb.gz exists. Skipping download

Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE):
pdbs/4PZL.pdb.gz exists. Skipping download

1		
	I	0%
  ==== 	I	5%
  ======	I	11%
  =========	I	16%
  ===================================	I	21%
  ===================================	I	26%
  ===================================	I	32%
  ===================================	I	37%
 	I	42%
 	I	47%
 	I	53%
 	I	58%
 	I	63%
 	I	68%
 	I	74%
 	I	79%
 	ı	84%
 	ı	89%
 	I	95%
 	==	100%

I have now downloaded the all the ADK structures in the PDB database but viewing them is

challenging. They need to be aligned and superposed. I already had BiocManager installed and installed the pdbaln package using BiocManager::install()

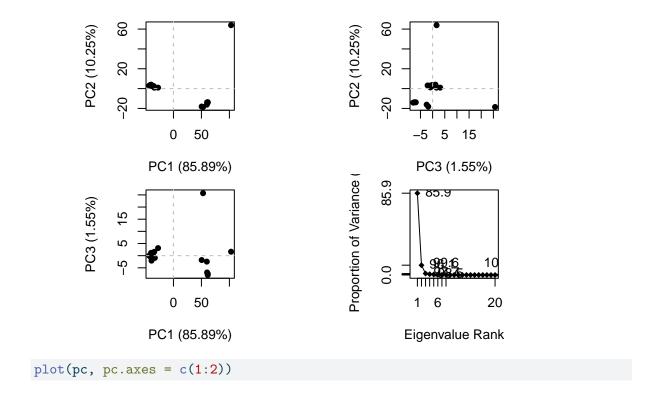
```
pdbs <- pdbaln(files, fit = TRUE, exefile="msa")</pre>
```

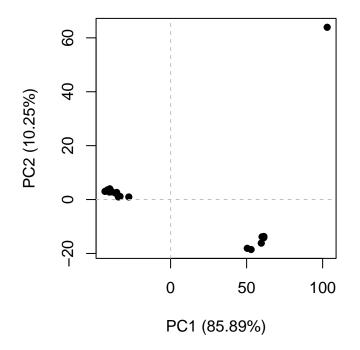
```
Reading PDB files:
pdbs/split chain/1AKE A.pdb
pdbs/split_chain/8BQF_A.pdb
pdbs/split_chain/4X8M_A.pdb
pdbs/split_chain/6S36_A.pdb
pdbs/split_chain/8Q2B_A.pdb
pdbs/split_chain/8RJ9_A.pdb
pdbs/split_chain/6RZE_A.pdb
pdbs/split_chain/4X8H_A.pdb
pdbs/split_chain/3HPR_A.pdb
pdbs/split_chain/1E4V_A.pdb
pdbs/split_chain/5EJE_A.pdb
pdbs/split_chain/1E4Y_A.pdb
pdbs/split_chain/3X2S_A.pdb
pdbs/split_chain/6HAP_A.pdb
pdbs/split_chain/6HAM_A.pdb
pdbs/split_chain/4K46_A.pdb
pdbs/split_chain/4NP6_A.pdb
pdbs/split_chain/3GMT_A.pdb
pdbs/split_chain/4PZL_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
    PDB has ALT records, taking A only, rm.alt=TRUE
     PDB has ALT records, taking A only, rm.alt=TRUE
    PDB has ALT records, taking A only, rm.alt=TRUE
    PDB has ALT records, taking A only, rm.alt=TRUE
    PDB has ALT records, taking A only, rm.alt=TRUE
     PDB has ALT records, taking A only, rm.alt=TRUE
     PDB has ALT records, taking A only, rm.alt=TRUE
       PDB has ALT records, taking A only, rm.alt=TRUE
    PDB has ALT records, taking A only, rm.alt=TRUE
. . . .
Extracting sequences
             name: pdbs/split_chain/1AKE_A.pdb
pdb/seq: 1
   PDB has ALT records, taking A only, rm.alt=TRUE
             name: pdbs/split_chain/8BQF_A.pdb
pdb/seq: 2
```

```
PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 3
             name: pdbs/split_chain/4X8M_A.pdb
pdb/seq: 4
             name: pdbs/split_chain/6S36_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
             name: pdbs/split_chain/8Q2B_A.pdb
pdb/seq: 5
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 6
             name: pdbs/split_chain/8RJ9_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
             name: pdbs/split_chain/6RZE_A.pdb
pdb/seq: 7
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 8
             name: pdbs/split_chain/4X8H_A.pdb
             name: pdbs/split_chain/3HPR_A.pdb
pdb/seq: 9
   PDB has ALT records, taking A only, rm.alt=TRUE
              name: pdbs/split_chain/1E4V_A.pdb
pdb/seq: 10
              name: pdbs/split_chain/5EJE_A.pdb
pdb/seq: 11
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 12
              name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 13
              name: pdbs/split_chain/3X2S_A.pdb
pdb/seq: 14
              name: pdbs/split_chain/6HAP_A.pdb
pdb/seq: 15
              name: pdbs/split_chain/6HAM_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 16
              name: pdbs/split_chain/4K46_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 17
              name: pdbs/split_chain/4NP6_A.pdb
pdb/seq: 18
              name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 19
              name: pdbs/split_chain/4PZL_A.pdb
```

##Principal Component Analysis

```
pc <- pca(pdbs)
plot(pc)</pre>
```





visualize the structural components of the variation in xray structures that is described by PC1.

```
pc1 <- mktrj(pc, pc=1, file="pc_1.pdb")</pre>
```

## AlphaFold

I used the predicted protein sequence from my find a gene project.

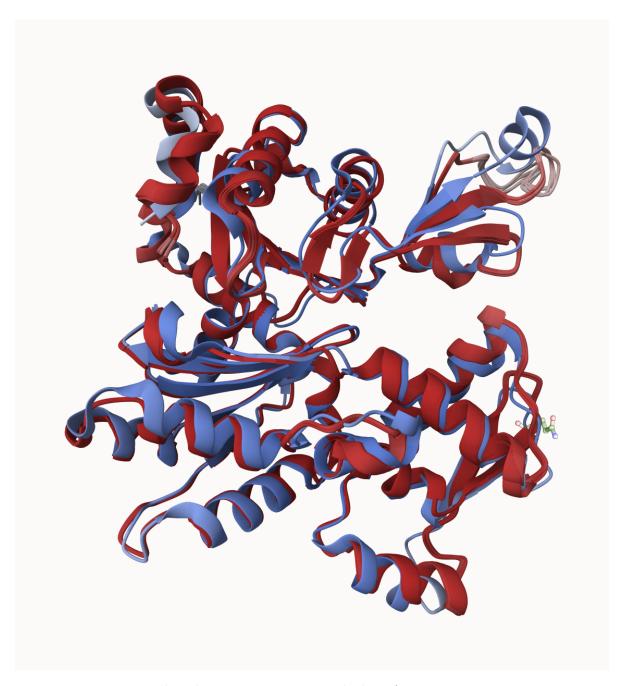


Figure 1: My predicted protein structures with the reference actin PDB structure

```
full.names = TRUE)
basename(pdb_files)
[1] "test_3910f_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb"
[2] "test_3910f_unrelaxed_rank_002_alphafold2_ptm_model_1_seed_000.pdb"
[3] "test_3910f_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000.pdb"
[4] "test_3910f_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb"
[5] "test_3910f_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb"
library(bio3d)
pdbs <- pdbaln(pdb_files, fit=TRUE, exefile="msa")</pre>
Reading PDB files:
test_3910f//test_3910f_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000.pdb
test_3910f//test_3910f_unrelaxed_rank_002_alphafold2_ptm_model_1_seed_000.pdb
test 3910f//test 3910f unrelaxed rank 003_alphafold2_ptm model_5_seed_000.pdb
test_3910f//test_3910f_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000.pdb
test_3910f//test_3910f_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000.pdb
Extracting sequences
             name: test_3910f//test_3910f_unrelaxed_rank_001_alphafold2_ptm_model_4_seed_000
pdb/seq: 1
pdb/seq: 2
             name: test_3910f//test_3910f_unrelaxed_rank_002_alphafold2_ptm_model_1_seed_000
             name: test_3910f//test_3910f_unrelaxed_rank_003_alphafold2_ptm_model_5_seed_000
pdb/seq: 3
             name: test_3910f//test_3910f_unrelaxed_rank_004_alphafold2_ptm_model_3_seed_000
pdb/seq: 4
             name: test_3910f//test_3910f_unrelaxed_rank_005_alphafold2_ptm_model_2_seed_000
pdb/seq: 5
pdbs
                                                                               50
[Truncated Name:1]test 3910f
                               CDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQK
[Truncated_Name:2]test_3910f
                               CDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQK
[Truncated Name:3]test 3910f
                               CDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQK
[Truncated_Name:4]test_3910f
                               CDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQK
[Truncated_Name:5]test_3910f
                               CDEEVAALVVDNGSGMCKAGFAGDDAPRAVFPSIVGRPRHQGVMVGMGQK
                               **************
```

51

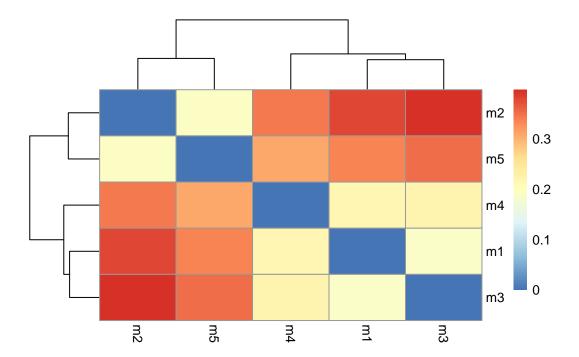
50

100

[Truncated_Name:1]test_3910f [Truncated_Name:2]test_3910f [Truncated_Name:3]test_3910f [Truncated_Name:4]test_3910f [Truncated_Name:5]test_3910f	DSYVGDEA DSYVGDEA DSYVGDEA DSYVGDEA	AQSKRGILTL AQSKRGILTL AQSKRGILTL AQSKRGILTL	KYPIEHGIV KYPIEHGIV KYPIEHGIV KYPIEHGIV	TNWDDMEKIWH TNWDDMEKIWH TNWDDMEKIWH TNWDDMEKIWH TNWDDMEKIWH ********	HTFYNELRVA HTFYNELRVA HTFYNELRVA HTFYNELRVA	PEE PEE PEE PEE
[Truncated_Name:1]test_3910f [Truncated_Name:2]test_3910f [Truncated_Name:3]test_3910f [Truncated_Name:4]test_3910f [Truncated_Name:5]test_3910f	HPVLLTEA HPVLLTEA HPVLLTEA HPVLLTEA	APLNPKANRE APLNPKANRE APLNPKANRE APLNPKANRE	KMTQIMFET KMTQIMFET KMTQIMFET KMTQIMFET	. FNTPAMYVAIG FNTPAMYVAIG FNTPAMYVAIG FNTPAMYVAIG FNTPAMYVAIG FNTPAMYVAIG ********	QAVLSLYASGR QAVLSLYASGR QAVLSLYASGR QAVLSLYASGR	TTG TTG TTG TTG
[Truncated_Name:1]test_3910f [Truncated_Name:2]test_3910f [Truncated_Name:3]test_3910f [Truncated_Name:4]test_3910f [Truncated_Name:5]test_3910f	IVLDSGDO IVLDSGDO IVLDSGDO	SVSHTVPIYE SVSHTVPIYE SVSHTVPIYE SVSHTVPIYE	GYALPHAIL GYALPHAIL GYALPHAIL GYALPHAIL	. RLDLAGRDLTE RLDLAGRDLTE RLDLAGRDLTE RLDLAGRDLTE RLDLAGRDLTE RLDLAGRDLTE ************************************	OYLMKILTERG OYLMKILTERG OYLMKILTERG OYLMKILTERG	YSF YSF YSF YSF
[Truncated_Name:1]test_3910f [Truncated_Name:2]test_3910f [Truncated_Name:3]test_3910f [Truncated_Name:4]test_3910f [Truncated_Name:5]test_3910f	TTTAEREI TTTAEREI TTTAEREI TTTAEREI	VRDIKEKLO VRDIKEKLO VRDIKEKLO VRDIKEKLO	YVALDFEQE YVALDFEQE YVALDFEQE YVALDFEQE	. MATAASSSSLE MATAASSSSLE MATAASSSSLE MATAASSSSLE MATAASSSSLE ********	EKSYELPDGQV EKSYELPDGQV EKSYELPDGQV EKSYELPDGQV	TITI TITI TITI TITI
[Truncated_Name:1]test_3910f [Truncated_Name:2]test_3910f [Truncated_Name:3]test_3910f [Truncated_Name:4]test_3910f [Truncated_Name:5]test_3910f	GNERFRCF GNERFRCF GNERFRCF GNERFRCF	PEALFQPSFL PEALFQPSFL PEALFQPSFL PEALFQPSFL	GMEACGIHE GMEACGIHE GMEACGIHE GMEACGIHE	. TTYNSIMKCDV TTYNSIMKCDV TTYNSIMKCDV TTYNSIMKCDV TTYNSIMKCDV *******	/DIRKDLYANT /DIRKDLYANT /DIRKDLYANT /DIRKDLYANT	VLS VLS VLS VLS

```
301
                                                                              350
[Truncated_Name:1]test_3910f
                              GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLS
[Truncated_Name:2]test_3910f
                              GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLS
[Truncated Name:3]test 3910f
                              GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLS
[Truncated_Name:4]test_3910f
                              GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLS
[Truncated Name:5]test 3910f
                              GGTTMYPGIADRMQKEITALAPSTMKIKIIAPPERKYSVWIGGSILASLS
                              ***************
                            301
                                                                              350
                            351
                                                      375
[Truncated_Name:1]test_3910f
                              TFQQMWISKQEYDESGPSIVHRKCF
[Truncated_Name:2]test_3910f
                              TFQQMWISKQEYDESGPSIVHRKCF
[Truncated_Name:3]test_3910f
                              TFQQMWISKQEYDESGPSIVHRKCF
[Truncated_Name:4]test_3910f
                              TFQQMWISKQEYDESGPSIVHRKCF
[Truncated_Name:5]test_3910f
                              TFQQMWISKQEYDESGPSIVHRKCF
                              ********
                            351
                                                      375
Call:
  pdbaln(files = pdb_files, fit = TRUE, exefile = "msa")
Class:
  pdbs, fasta
Alignment dimensions:
  5 sequence rows; 375 position columns (375 non-gap, 0 gap)
+ attr: xyz, resno, b, chain, id, ali, resid, sse, call
rd <- rmsd(pdbs, fit=T)</pre>
Warning in rmsd(pdbs, fit = T): No indices provided, using the 375 non NA positions
range(rd)
[1] 0.000 0.398
library(pheatmap)
```

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



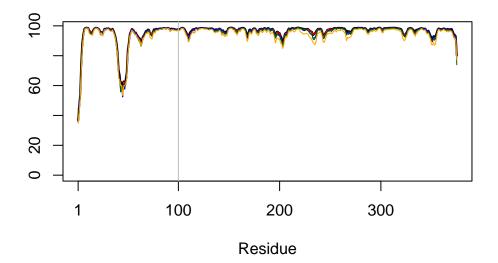
```
pdb <- read.pdb("1j6z")</pre>
```

Note: Accessing on-line PDB file PDB has ALT records, taking A only, rm.alt=TRUE

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



#### core <- core.find(pdbs)</pre>

```
core size 374 of 375
                       vol = 0.973
core size 373 of 375
                       vol = 0.841
core size 372 of 375
                       vol = 0.774
core size 371 of 375
                       vol = 0.72
core size 370 of 375
                       vol = 0.664
core size 369 of 375
                       vol = 0.629
core size 368 of 375
                       vol = 0.592
core size 367 of 375
                       vol = 0.567
core size 366 of 375
                       vol = 0.55
core size 365 of 375
                       vol = 0.536
core size 364 of 375
                       vol = 0.519
core size 363 of 375
                       vol = 0.506
core size 362 \text{ of } 375 \text{ vol} = 0.493
FINISHED: Min vol (0.5) reached
```

### core.inds <- print(core, vol=0.5)</pre>

# 363 positions (cumulative volume <= 0.5 Angstrom^3)
start end length</pre>

```
1 3 41 39
2 49 194 146
3 198 375 178
```

xyz <- pdbfit(pdbs, core.inds, outpath="corefit\_structures")</pre>

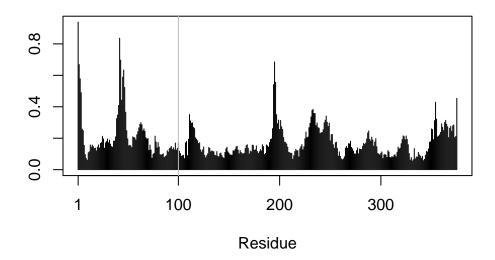


Figure 2: Core superposed structures colored by B-factor

```
rf <- rmsf(xyz)
plotb3(rf, sse=pdb)</pre>
```

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)
```

Warning: package 'jsonlite' was built under R version 4.3.3

```
$names
```

[1] "plddt" "max\_pae" "pae" "ptm"

#### head(pae1\$plddt)

[1] 36.53 46.22 54.09 74.19 88.12 93.88

Max pae scores

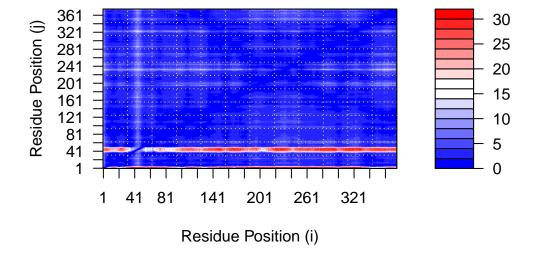
#### pae1\$max\_pae

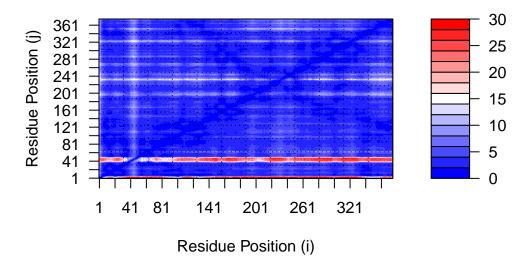
[1] 31.09375

#### pae5\$max\_pae

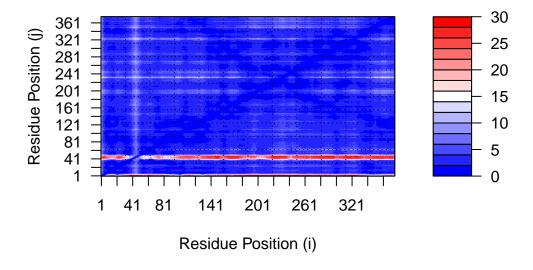
[1] 30.9375

Plotted residue-by-residue pae scores for model 1 and 5  $\,$ 





model 1 with range of 0-30



#### Residue conservation from alignment file

[1] "test\_3910f//test\_3910f.a3m"

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

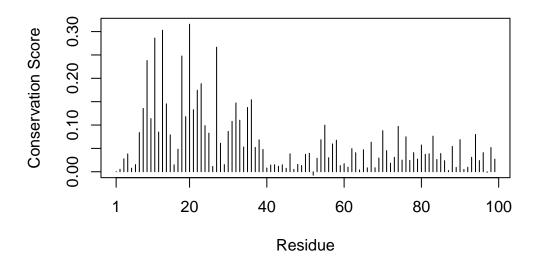
[1] " \*\* Duplicated sequence id's: 101 \*\*"

```
# number of sequences
dim(aln$ali)
```

[1] 21157 757

Socring residue conservation

Warning in plotb3(sim[1:99], sse = trim.pdb(pdb, chain = "A"), ylab =
"Conservation Score"): Length of input 'sse' does not equal the length of input
'x'; Ignoring 'sse'



```
con <- consensus(aln, cutoff = 0.25)
con$seq</pre>
```

```
[181] "A" "G" "R" "D" "L" "T" "-" "Y" "L" "-" "-" "L" "L" "L" "-" "E" "-" "G" "-"
[199] "-" "F" "-" "T" "-" "A" "E" "-" "E" "I" "V" "R" "D" "I" "K" "E" "K" "-"
[235] "-" "-" "-" "-" "-" "Y" "-" "L" "P" "D" "G" "-" "-" "-" "-" "-" "-" "G" "-"
[253] "E" "R" "F" "-" "-" "P" "E" "-" "L" "F" "-" "P" "-" "-" "-" "G" "-" "-"
[289] "-" "-" "-" "-" "-" "-" "-" "-" "-" "L" "-" "G" "G" "-" "-" "-" "-" "-"
[757] "-"
```