

Class 9 - Structural Bioinformatics

Jenny

The RCSB Protein Data Bank (PDB)

Protein structures by X-ray crystallography dominate this database. We are going to use Molstar Skipping questions 1-3 since the website was too slow for us

- Computer doesn't allow me to open WebGL since my browser seems to be outdated so I can't do this step... Without having access to this website I'm not able to answer questions 4-7

3. Introduction to Bio3D in R

Bio3D is an R package for structural bioinformatics. To use it we need to call it up with the 'library()' function (just like any package)

```
library("bio3d")
```

```
pdb <- read.pdb("1hsg")
```

Note: Accessing on-line PDB file

```
pdb
```

```
Call: read.pdb(file = "1hsg")
```

```
Total Models#: 1
```

```
Total Atoms#: 1686, XYZs#: 5058 Chains#: 2 (values: A B)
```

```
Protein Atoms#: 1514 (residues/Calpha atoms#: 198)
```

Nucleic acid Atoms#: 0 (residues/phosphate atoms#: 0)

Non-protein/nucleic Atoms#: 172 (residues: 128)

Non-protein/nucleic resid values: [HOH (127), MK1 (1)]

Protein sequence:

PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYD
QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPQITLWQRPLVTIKIGGQLKE
ALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTP
VNIIGRNLLTQIGCTLNF

+ attr: atom, xyz, seqres, helix, sheet,
calpha, remark, call

Q7. How many amino acid residues are there in this pdb object ? 198

Q8. Name one of the two non-protein residues ? mk1

Q9. How many protein chains are in this structure ? 2

`attributes(pdb)`

\$names

[1] "atom" "xyz" "seqres" "helix" "sheet" "calpha" "remark" "call"

\$class

[1] "pdb" "sse"

`head(pdb$atom)`

	type	eleno	elety	alt	resid	chain	resno	insert	x	y	z	o	b
1	ATOM	1	N	<NA>	PRO	A	1	<NA>	29.361	39.686	5.862	1	38.10
2	ATOM	2	CA	<NA>	PRO	A	1	<NA>	30.307	38.663	5.319	1	40.62
3	ATOM	3	C	<NA>	PRO	A	1	<NA>	29.760	38.071	4.022	1	42.64
4	ATOM	4	O	<NA>	PRO	A	1	<NA>	28.600	38.302	3.676	1	43.40
5	ATOM	5	CB	<NA>	PRO	A	1	<NA>	30.508	37.541	6.342	1	37.87
6	ATOM	6	CG	<NA>	PRO	A	1	<NA>	29.296	37.591	7.162	1	38.40

	segid	elesy	charge
1	<NA>	N	<NA>
2	<NA>	C	<NA>
3	<NA>	C	<NA>

```
4 <NA>      0 <NA>
5 <NA>      C <NA>
6 <NA>      C <NA>
```

Comparative structure analysis of Adenylate Kinase

Adenylate kinase structure (PDB code: 1AKE)

Make sure to install the packages first

Q10. Which of the packages above is found only on BioConductor and not CRAN ? msa

Q11. Which of the above packages is not found on BioConductor or CRAN ? none

Q12. True or False? Functions from the devtools package can be used to install packages from GitHub and BitBucket ? TRUE

First we get it's primary sequence :

```
library(bio3d)
aa <- get.seq("1ake_A")
```

Warning in get.seq("1ake_A"): Removing existing file: seqs.fasta

Fetching... Please wait. Done.

```
aa
```

```

      1      .      .      .      .      .      .      60
pdb|1AKE|A MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMRLAAVKSSELGKQAKDIMDAGKLV
      1      .      .      .      .      .      .      60

      61      .      .      .      .      .      .      120
pdb|1AKE|A DELVIALVKERIAQEDCRNGFLLDGFPR TIPQADAMKEAGINVDYVLEFDVPDELIVDRI
      61      .      .      .      .      .      .      120

      121      .      .      .      .      .      .      180
pdb|1AKE|A VGRRVHAPSGRVYHV KFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQM TAPLIG
      121      .      .      .      .      .      .      180
```

```

      181          .          .          .      214
pdb|1AKE|A  YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG
      181          .          .          .      214

```

Call:

```
read.fasta(file = outfile)
```

Class:

```
fasta
```

Alignment dimensions:

```
1 sequence rows; 214 position columns (214 non-gap, 0 gap)
```

+ attr: id, ali, call

Q13.How many amino acids are in this sequence, i.e. how long is this sequence ?
214

```
# Blast or hammer search
b <- blast.pdb(aa)
```

```
Searching ... please wait (updates every 5 seconds) RID = NNSNN990016
.....
Reporting 98 hits
```

```
hits <- NULL
hits$ pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A','3X2S_A',
```

```
# List out some 'top hits'
head(hits$ pdb.id)
```

```
[1] "1AKE_A" "6S36_A" "6RZE_A" "3HPR_A" "1E4V_A" "5EJE_A"
```

```
# Download releated PDB files
files <- get.pdb(hits$ pdb.id, path="pdbc", split=TRUE, gzip=TRUE)
```

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

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

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

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

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

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

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

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

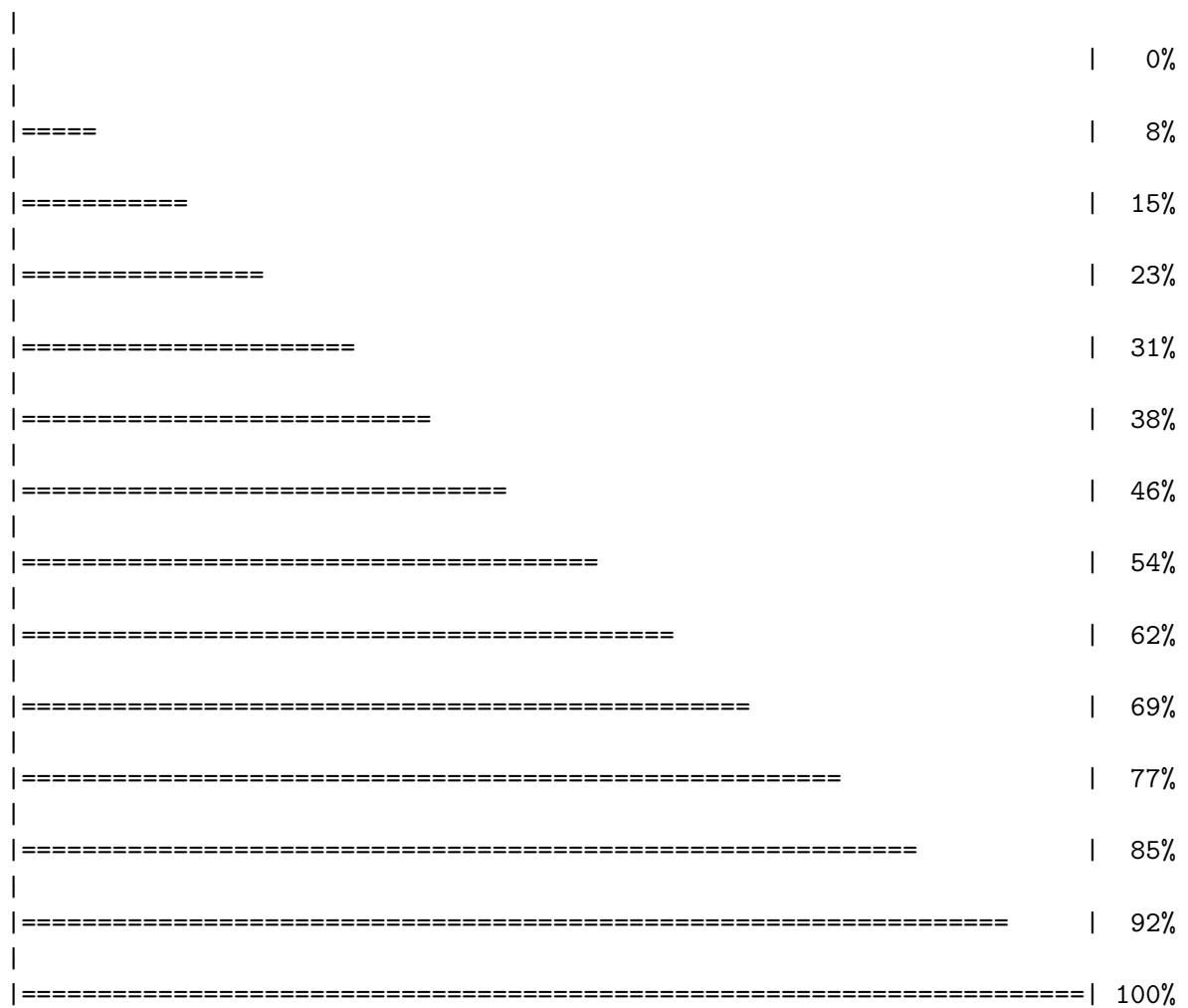
Warning in get.pdb(hits\$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
6HAP.pdb exists. Skipping download

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

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

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

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



```
pdbbs <- pdbaln(files, fit = TRUE, exefile = "msa")
```

Reading PDB files:

```
pdbbs/split_chain/1AKE_A.pdb
pdbbs/split_chain/6S36_A.pdb
pdbbs/split_chain/6RZE_A.pdb
pdbbs/split_chain/3HPR_A.pdb
pdbbs/split_chain/1E4V_A.pdb
pdbbs/split_chain/5EJE_A.pdb
pdbbs/split_chain/1E4Y_A.pdb
pdbbs/split_chain/3X2S_A.pdb
```

```

pdbh/split_chain/6HAP_A.pdb
pdbh/split_chain/6HAM_A.pdb
pdbh/split_chain/4K46_A.pdb
pdbh/split_chain/3GMT_A.pdb
pdbh/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
...

```

Extracting sequences

```

pdb/seq: 1   name: pdbh/split_chain/1AKE_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 2   name: pdbh/split_chain/6S36_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 3   name: pdbh/split_chain/6RZE_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 4   name: pdbh/split_chain/3HPR_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 5   name: pdbh/split_chain/1E4V_A.pdb
pdb/seq: 6   name: pdbh/split_chain/5EJE_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 7   name: pdbh/split_chain/1E4Y_A.pdb
pdb/seq: 8   name: pdbh/split_chain/3X2S_A.pdb
pdb/seq: 9   name: pdbh/split_chain/6HAP_A.pdb
pdb/seq: 10  name: pdbh/split_chain/6HAM_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 11  name: pdbh/split_chain/4K46_A.pdb
  PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 12  name: pdbh/split_chain/3GMT_A.pdb
pdb/seq: 13  name: pdbh/split_chain/4PZL_A.pdb

```

```

ids <- basename.pdb(pdbh$id)
dev.off()

```

```

null device
1

```

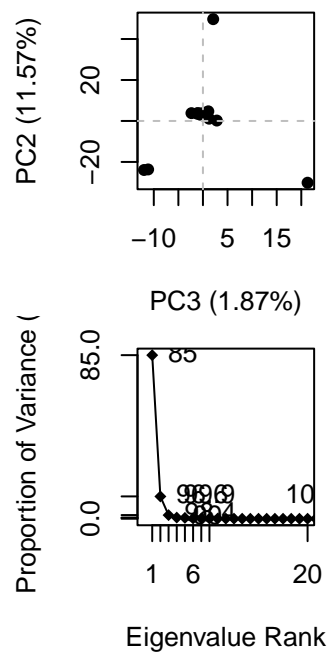
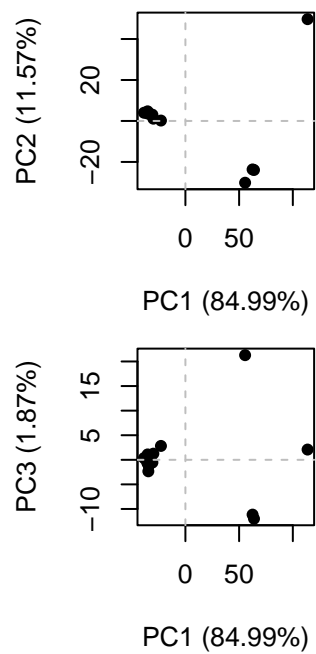
```
plot(pdb, labels=ids)
par(mar=c(1,1,1,1))
```

```
anno <- pdb.annotate(ids)
unique(anno$source)
```

```
[1] "Escherichia coli"
[2] "Escherichia coli K-12"
[3] "Escherichia coli O139:H28 str. E24377A"
[4] "Escherichia coli str. K-12 substr. MDS42"
[5] "Photobacterium profundum"
[6] "Burkholderia pseudomallei 1710b"
[7] "Francisella tularensis subsp. tularensis SCHU S4"
```

PCA

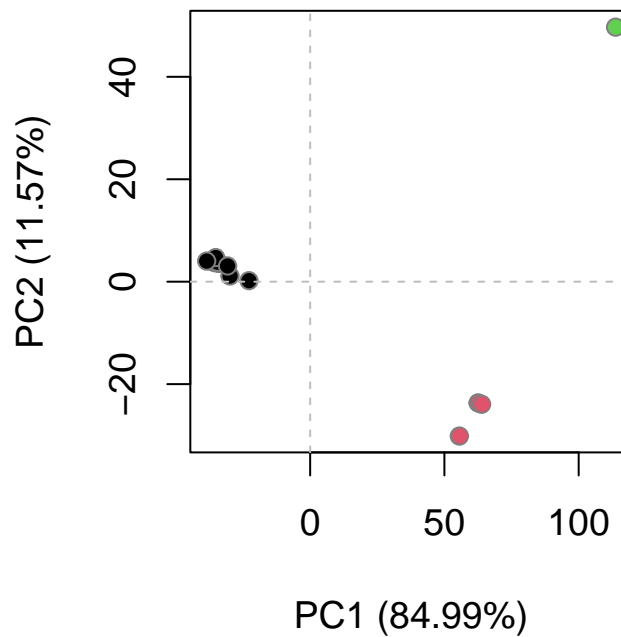
```
pc.xray <- pca(pdb)
plot(pc.xray)
```




```
rd <- rmsd(pdbbs)
```

Warning in rmsd(pdbbs): No indices provided, using the 204 non NA positions

```
hc.rd <- hclust(dist(rd))  
grps.rd <- cutree(hc.rd, k=3)  
  
plot(pc.xray, 1:2, col="grey50", bg=grps.rd, pch=21, cex=1)
```



```
library(ggplot2)  
library(ggrepel)  
  
df <- data.frame(PC1=pc.xray$z[,1],  
                 PC2=pc.xray$z[,2],  
                 col=as.factor(grps.rd),  
                 ids=ids)  
  
p <- ggplot(df) +  
  aes(PC1, PC2, col=col, label=ids) +  
  geom_point(size=2) +
```

```
geom_text_repel(max.overlaps = 20) +
theme(legend.position = "none")
p
```

