

# class10

Tin Nguyen

```
read.csv("Data Export Summary.csv")
```

	Molecular.Type	X.ray	EM	NMR	Multiple.methods	Neutron	Other
1	Protein (only)	172,081	18,135	12,627	210	84	32
2	Protein/Oligosaccharide	10,027	2,974	34	10	2	0
3	Protein/NA	8,849	5,386	286	7	0	0
4	Nucleic acid (only)	2,947	185	1,536	15	3	1
5	Other	170	10	33	0	0	0
6	Oligosaccharide (only)	11	0	6	1	0	4
	Total						
1	203,169						
2	13,047						
3	14,528						
4	4,687						
5	213						
6	22						

```
pdb.data <- read.csv("Data Export Summary.csv")
```

Q1. What percentage of structures in the PDB are solved by X-Ray and electron Microscopy

```
xray <- as.numeric(gsub(",", "", pdb.data$X.ray))
em <- as.numeric(gsub(",", "", pdb.data$EM))
total <- as.numeric(gsub(",", "", pdb.data$Total))

total_sum <- sum(total)
xray_percent <- 100 * sum(xray) / total_sum
em_percent <- 100 * sum(em) / total_sum

xray_percent
```

```
[1] 82.35596
```

```
em_percent
```

```
[1] 11.32535
```

Q2. What proportion of structures in the PDB are protein

```
protein_index <- which(pdb.data$Molecular.Type == "Protein (only)")  
protein_total <- total[protein_index]  
protein_percent <- 100 * protein_total / total_sum  
protein_percent
```

```
[1] 86.21057
```

Q3. Type HIV in the PDB website search box on the home page and determine how many HIV-1 protease structures are in the current PDB

around 1090 structures

## Introduction to Bio3D in R

```
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

`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>	O	<NA>										
5	<NA>	C	<NA>										
6	<NA>	C	<NA>										

```
adk <- read.pdb("6s36")
```

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

```
adk
```

```
Call: read.pdb(file = "6s36")
```

```
Total Models#: 1
```

```
Total Atoms#: 1898, XYZs#: 5694 Chains#: 1 (values: A)
```

```
Protein Atoms#: 1654 (residues/Calpha atoms#: 214)
```

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

```
Non-protein/nucleic Atoms#: 244 (residues: 244)
```

```
Non-protein/nucleic resid values: [ CL (3), HOH (238), MG (2), NA (1) ]
```

```
Protein sequence:
```

```
MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMRLRAAVKSGSELGKQAKDIMDAGKLV  
DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDKI  
VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG  
YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG
```

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

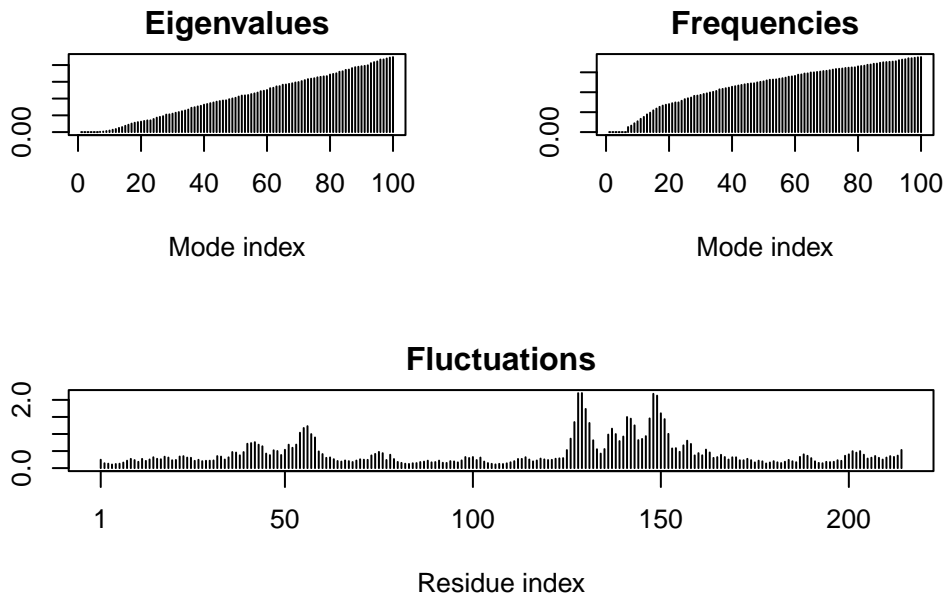
```
# Perform flexibility prediction
```

```
m <- nma(adk)
```

```
Building Hessian... Done in 0.014 seconds.
```

```
Diagonalizing Hessian... Done in 0.288 seconds.
```

```
plot(m)
```



```
mktrj(m, file="adk_m7.pdb")
```

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

Warning in get.seq("lake\_A"): Removing existing file: seqs.fasta

Fetching... Please wait. Done.

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

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

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/6S36.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/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/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

	0%
=====	8%
=====	15%
=====	23%



```
# Align related PDBs
pdbs <- pdbaln(files, fit = TRUE, exefile="msa")
```

Reading PDB files:

```
pdbs/split_chain/1AKE_A.pdb
pdbs/split_chain/6S36_A.pdb
pdbs/split_chain/6RZE_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/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
...

```

#### Extracting sequences

```

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

```

```

# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbc$id)

# Draw schematic alignment
#plot(pdbc, labels=ids)

```

```

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

```

```

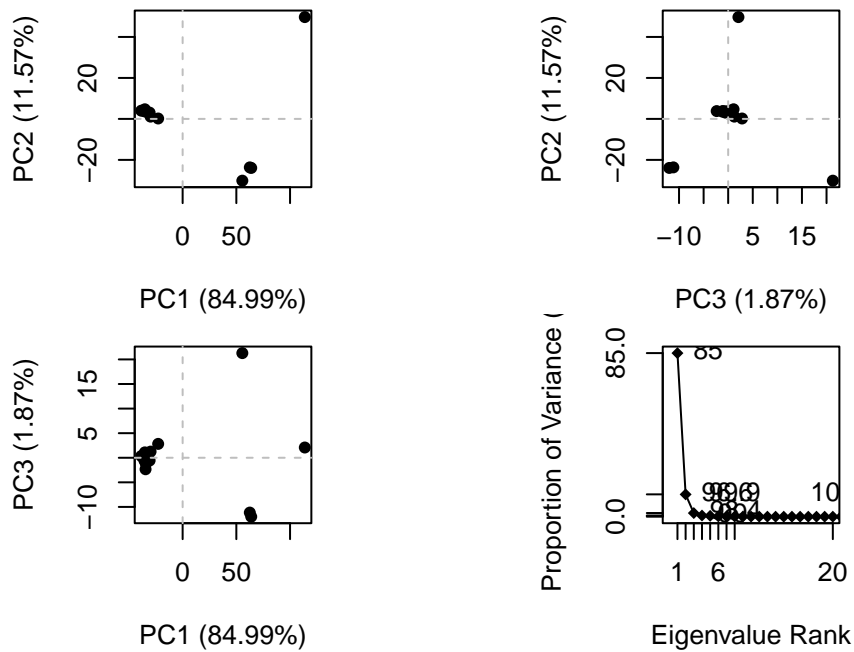
[1] "Escherichia coli"
[2] "Escherichia coli K-12"
[3] "Escherichia coli 0139:H28 str. E24377A"
[4] "Escherichia coli str. K-12 substr. MDS42"
[5] "Photobacterium profundum"

```



```
[6] "Burkholderia pseudomallei 1710b"
[7] "Francisella tularensis subsp. tularensis SCHU S4"
```

```
# Perform PCA
pc.xray <- pca(pdbbs)
plot(pc.xray)
```

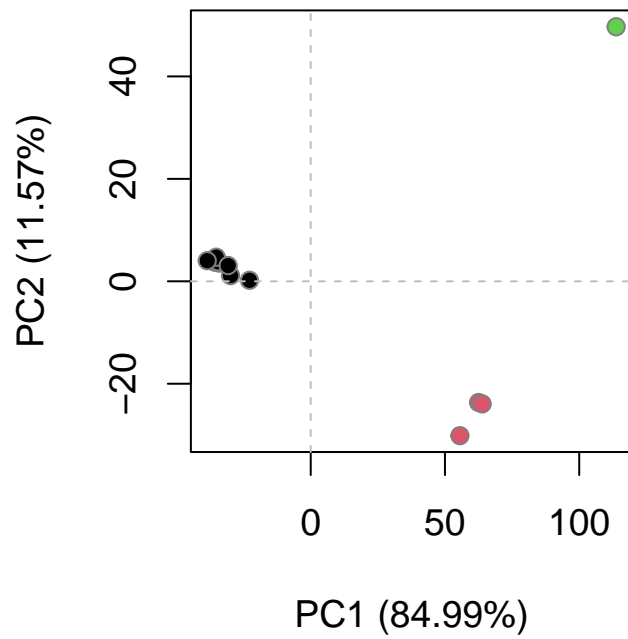


```
# Calculate RMSD
rd <- rmsd(pdbbs)
```

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

```
# Structure-based clustering
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)
```



```
#Plotting results with ggplot2
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
```

