

lab09

Nate Tran

Exploring the PDB

```
data <- read.csv("Data_Export_Summary.csv")
knitr::kable(data)
```

Molecular.Type	X.ray	EM	NMR	Multiple.methods	Neutron	Other	Total
Protein (only)	152914	9495	12121	191	72	32	174825
Protein/Oligosaccharide	9008	1663	32	7	1	0	10711
Protein/NA	8069	2949	282	6	0	0	11306
Nucleic acid (only)	2602	78	1434	12	2	1	4129
Other	163	9	31	0	0	0	203
Oligosaccharide (only)	11	0	6	1	0	4	22

```
## Creating function for removing commas and summing numbers
sum_clean <- function(x){
  sum(as.numeric(gsub(",", "", x)))
}

sum_clean(data$X.ray) ## test of sum_clean returns correct value
```

```
[1] 172767
```

Q1

85.9% of the PDB structures are solved by X-ray and 7.1% are solved by EM.

```
100*(sum_clean(data$X.ray)/sum_clean(data$Total))
```

```
[1] 85.87
```

```
100*(sum_clean(data$EM)/sum_clean(data$Total))
```

```
[1] 7.054812
```

Q2

86.9% of structures are protein only.

```
100*(data$Total[1]/sum_clean(data$Total))
```

```
[1] 86.89288
```

Q3

There are 1286 HIV-1 protease structures in the PDB.

Mol* Practice

Q4

We only see one atom per water molecule because the resolution of the structure is not high enough to capture hydrogen atoms.

Q5

Water molecule 308

Q6

Intro to Biod3D

```
library(bio3d)
```

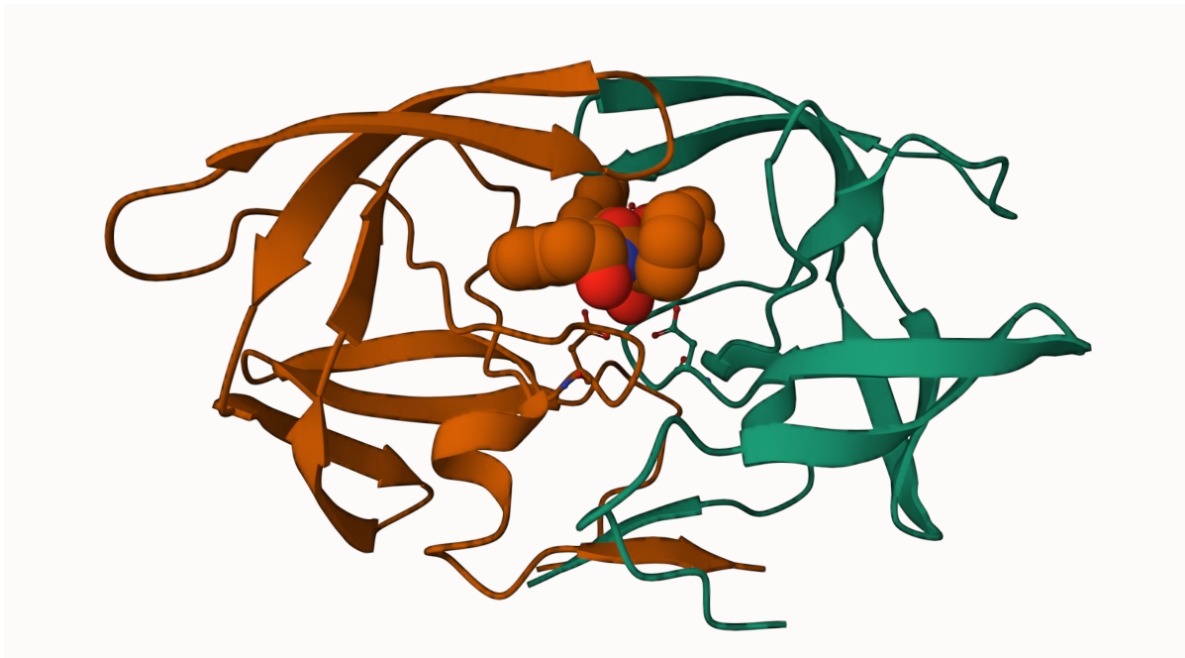


Figure 1: A snapshot of HIV-1 protease (PDB code: 1HSG) highlighting two D25 residues and water molecule 308

```
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

There are 198 residues.

Q8

MK1 is a non-protein residue.

Q9

There are 2 protein chains.

Normal Mode Analysis (NMA)

```
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:

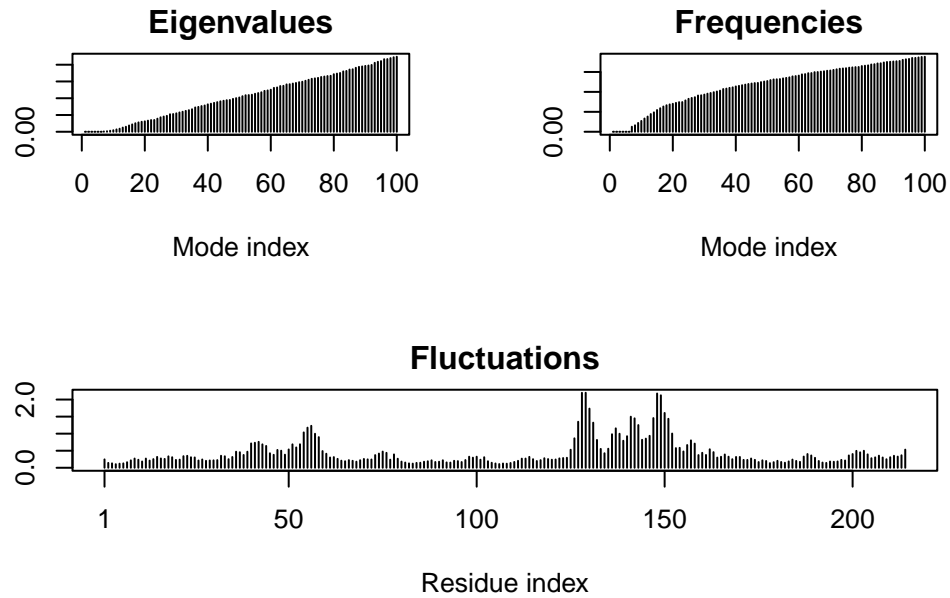
MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMRLRAAVKSGSELGKQAKDIMDAGKLV
DELVIALVKERIAQEDCRNGFLDGFPRTPQADAMKEAGINVDYVLEFDVPDELIVDKI
VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG

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

```
## performing NMA on adk  
nma_adk <- nma(adk)
```

Building Hessian... Done in 0.04 seconds.
Diagonalizing Hessian... Done in 0.29 seconds.

```
## plotting results of NMA  
plot(nma_adk)
```



```
mktrj(nma_adk, file="adk_nma.pdb")
```

Comparative Structure Analysis of ADK

Q10

“msa” is only found on BioConductor

Q11

“bio3d-view” is not found on BioConductor or CRAN

Q12

True

```
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  MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
      1      .      .      .      .      .      60

      61      .      .      .      .      .      120
pdb|1AKE|A  DELVIALVKERIAQEDCRNGFLLDGFRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
      61      .      .      .      .      .      120

     121      .      .      .      .      .      180
pdb|1AKE|A  VGRRVHAPSGRVYHVKNPPKVEGKDDVTGEELTRKDDQEETVRKRLVEYHQMTAPLIG
     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

There are 214 amino acids in this sequence.

```
##b <- blast.pdb(aa)
##b
```

```
##hits <- plot(b)
##hits
```

```
##hits
```

```
hits <- NULL
hits$pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A','3X2S_A','
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

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

		0%
=====		8%
=====		15%
=====		23%
=====		31%
=====		38%
=====		46%
=====		54%
=====		62%
=====		69%
=====		77%
=====		85%
=====		92%
=====		100%

```
pdbs <- pdbaln(files, fit=T, exefile="msa")
```

Reading PDB files:

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

```

pdb/seq: 12   name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 13   name: pdbs/split_chain/4PZL_A.pdb

```

```
ids <- basename.pdb(pdb$id)
```

```
##plot(pdb, labels=ids)
```

```

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

```

```

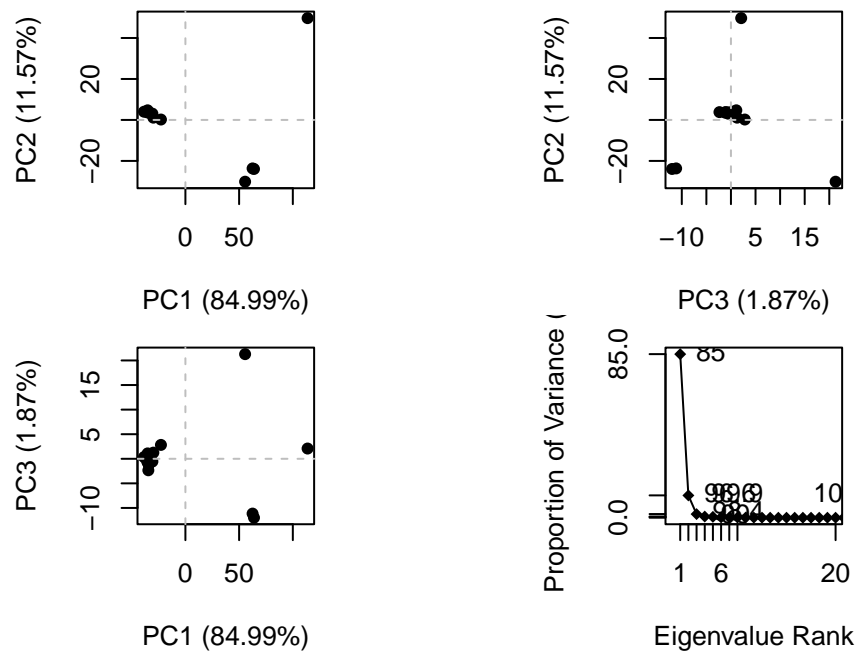
[1] "Escherichia coli"
[2] "Escherichia coli K-12"
[3] "Escherichia coli O139:H28 str. E24377A"
[4] "Escherichia coli str. K-12 substr. MDS42"
[5] "Photobacterium profundum"
[6] "Burkholderia pseudomallei 1710b"
[7] "Francisella tularensis subsp. tularensis SCHU S4"

```

```

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

```



Trajectory visualization

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