```
---
title: "Class 9 Forrest Wang"
format: pdf
---

```{r}
setwd("~/Downloads")
Class9data <- read.csv("Data Export Summary.csv")
```
1. PDB Intro

Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.
A1: X-Ray is 86.3%, EM is 6.58%, total is 92.9%

Q2: What proportion of structures in the PDB are protein?
A2: Protein accounts for 171828 out of 197512, equaling to 87.0%

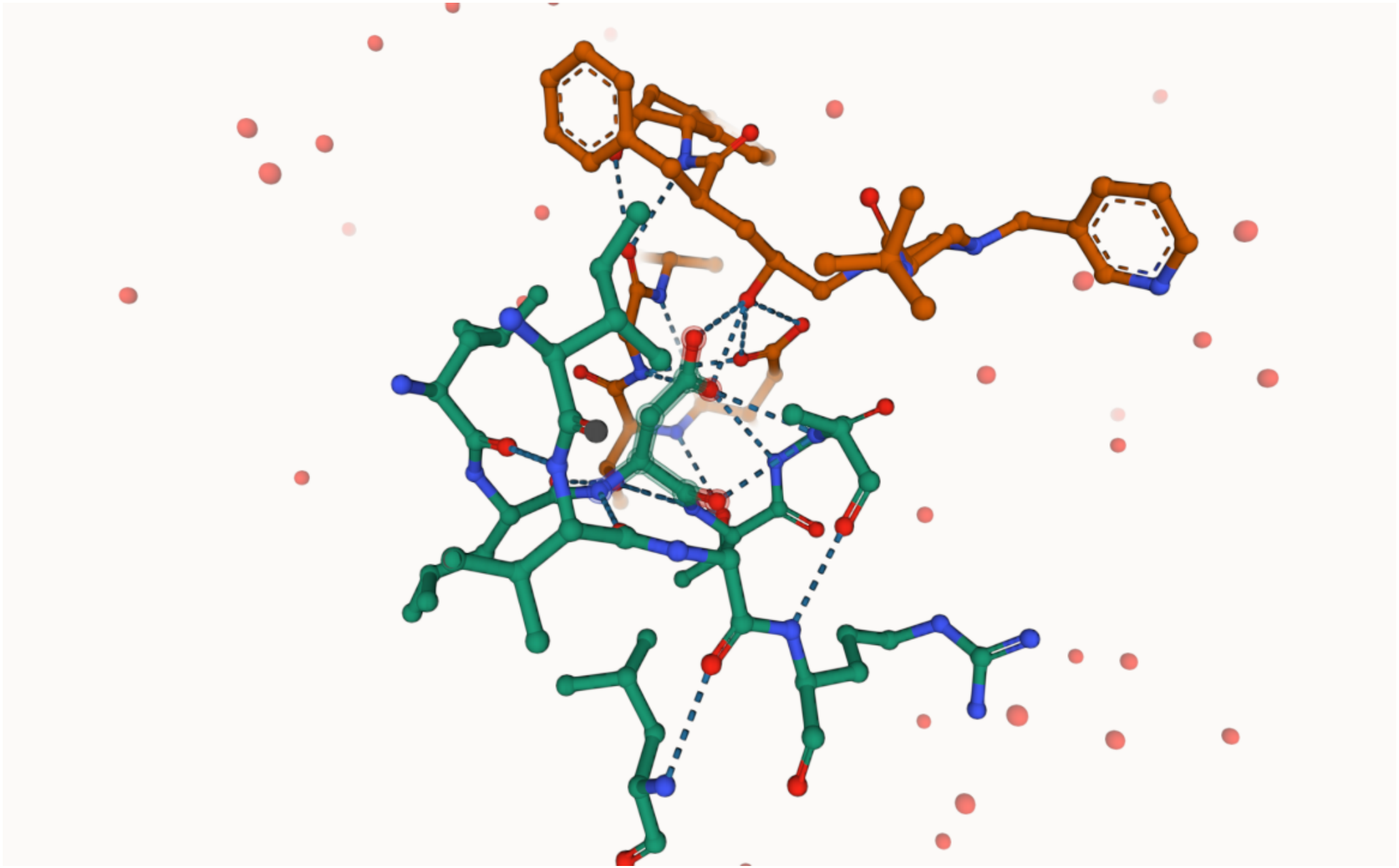
Q3: Type HIV in the PDB website search box on the home page and determine how many HIV-1 protease structures are in the current PDB?
A3: There are 240 total HIV-1 protease structures

2. PDB Visualization

Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?
A4: We only see 1 atom/h2o molecule in this structure maybe because the environment is basic and protons are being pulled off the water

Q5: There is a critical “conserved” water molecule in the binding site. Can you identify this water molecule? What residue number does this water molecule have
A5: The conserved H2O molecule closest to the ligand binding site is HOH 308

Q6: Generate and save a figure clearly showing the two distinct chains of HIV-protease along with the ligand. You might also consider showing the catalytic residues ASP 25 in each chain and the critical water (we recommend “Ball & Stick” for these side-chains). Add this figure to your Quarto document.
```



3. Intro to Bio3D

```{r}
library(bio3d)
```

Read PDB File

```{r}
pdb <- read.pdb("1hsg")
```

File Summary:

```{r}
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

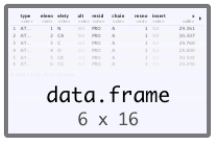
 + attr: atom, xyz, seqres, helix, sheet,
         calpha, remark, call
```

```
Q7: How many amino acid residues are there in this pdb object?
A6: 198 residues

Q8: Name one of the two non-protein residues?
A8: HOH (127) and MK1 (1)

Q9: How many protein chains are in this structure?
A9: 2 chains

Attributes
```{r}
attributes(pdb)
head(pdb$atom)
```
```



	type <chr>	eleno <int>	elety <chr>	alt <chr>	resid <chr>	chain <chr>	resno <int>	insert <chr>	x <dbl>
1	ATOM	1	N	*NA*	PRO	A	1	*NA*	29.361
2	ATOM	2	CA	*NA*	PRO	A	1	*NA*	30.307
3	ATOM	3	C	*NA*	PRO	A	1	*NA*	29.760
4	ATOM	4	O	*NA*	PRO	A	1	*NA*	28.600
5	ATOM	5	CB	*NA*	PRO	A	1	*NA*	30.508
6	ATOM	6	CG	*NA*	PRO	A	1	*NA*	29.296

6 rows | 1-10 of 16 columns

```
Predicting functional motions of a single structure
```{r}
adk <- read.pdb("6s36")
adk
```
```

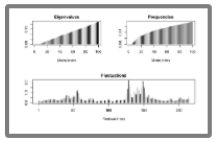
```
  Note: Accessing on-line PDB file
Warning: /var/folders/3v/y80n4ckj6yq900214dt8yy9c0000gn/T//Rtmp78IpSL/6s36.pdb  exists. Skipping download   PDB has ALT
records, taking A only, rm.alt=TRUE

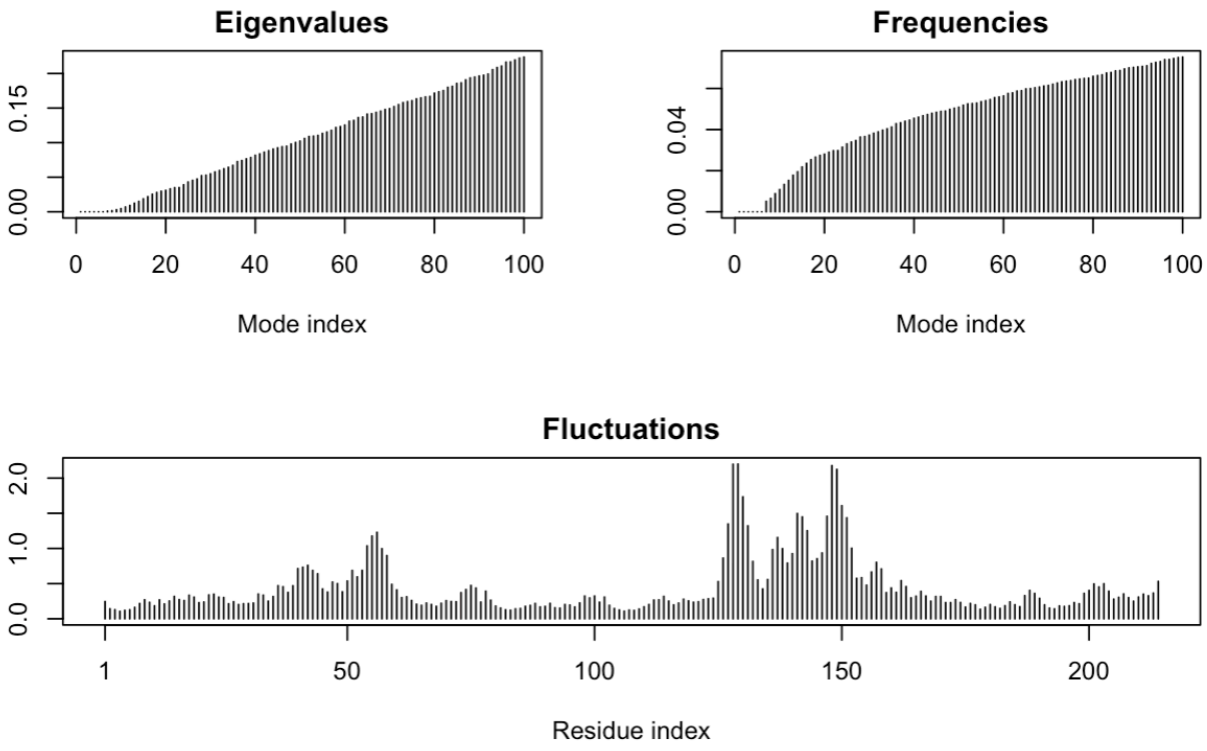
 Call:  read.pdb(file = "6s36")

   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)
```

```
Normal Mode Analysis
```{r}
m <- nma(adk)
plot(m)
```
```






```
View like movie
```{r}
mktrj(m, file="adk_m7.pdb")
```
```

4. Comparative structure analysis of Adenylate Kinase

Q10. Which of the packages above is found only on BioConductor and not CRAN?
A10: "msa"
```
```

```
Q11. Which of the above packages is not found on BioConductor or CRAN?
A11: "Grantlab/bio3d-view"

True or False? Functions from the devtools package can be used to install packages from GitHub and BitBucket?
A; True
```

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

Visualize AA

```{r}
aa
```

```
Q13. How many amino acids are in this sequence, i.e. how long is this sequence?
A13: 214 AAs

BLAST Search
```

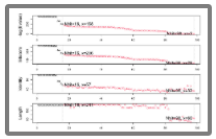
```{r}
# Blast or hmmer search
b <- blast.pdb(aa)
```

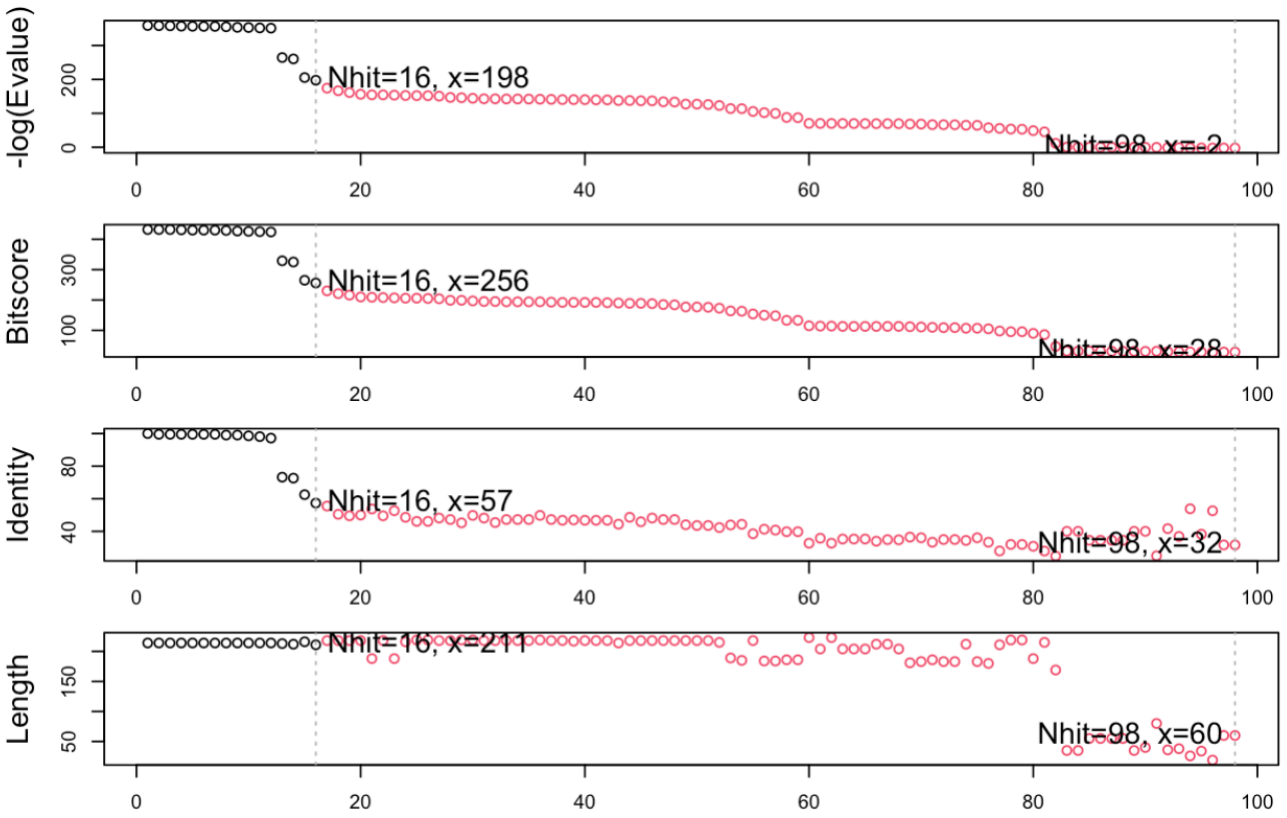
```{r}
b
```

```
$hit.tbl
       queryid subjectids identity alignmentlength mismatches gapopens q.start q.end s.start s.end    evalue bitscore
1  Query_44619     1AKE_A  100.000             214          0        0       1   214       1   214 1.32e-156      432
2  Query_44619     4X8M_A   99.533             214          1        0       1   214       1   214 2.36e-156      432
3  Query_44619     6S36_A   99.533             214          1        0       1   214       1   214 3.47e-156      432
4  Query_44619     6RZE_A   99.533             214          1        0       1   214       1   214 9.94e-156      431
5  Query_44619     4X8H_A   99.533             214          1        0       1   214       1   214 1.31e-155      430
6  Query_44619     3HPR_A   99.533             214          1        0       1   214       1   214 1.86e-155      430
7  Query_44619     1E4V_A   99.533             214          1        0       1   214       1   214 1.96e-155      430
8  Query_44619     5EJE_A   99.065             214          2        0       1   214       1   214 5.88e-155      429
9  Query_44619     1E4Y_A   99.065             214          2        0       1   214       1   214 3.40e-154      427
10 Query_44619     3X2S_A   98.598             214          3        0       1   214       1   214 5.70e-154      426
11 Query_44619     6HAP_A   98.131             214          4        0       1   214       1   214 1.69e-153      425
12 Query_44619     6HAM_A   97.196             214          6        0       1   214       1   214 3.44e-153      424
13 Query_44619     4K46_A   73.239             213         57        0       1   213       1   213 1.53e-115      329
14 Query_44619     4NP6_A   72.642             212         58        0       2   213       5   216 8.49e-114      325
15 Query_44619     3GMT_A   62.500             216         75        1       2   211      10   225  6.64e-90      265
16 Query_44619     4PZL_A   57.346             211         86        2       2   209      26   235  1.58e-86      256
17 Query_44619     5G3Y_A   55.505             218         88        2       1   214       1   213  2.34e-76      230
```

Plot!

```{r}
hits <- plot(b)
hits
```





Top Hits:

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

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

```{r}
hits <- NULL
hits$pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A','3X2S_A','6HAP_A','6HAM_A','4K46_A','3GMT_A','4P
```

```

Download related files

```{r}
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

```
  |======================================================================================================================|
100%ZE.pdb.gz  exists. Skipping downloadWarning: pdbs/3HPR.pdb.gz  exists. Skipping downloadWarning: pdbs/1E4V.pdb.gz  exists.
Skipping downloadWarning: pdbs/5EJE.pdb.gz  exists. Skipping downloadWarning: pdbs/1E4Y.pdb.gz  exists. Skipping
downloadWarning: pdbs/3X2S.pdb.gz  exists. Skipping downloadWarning: pdbs/6HAP.pdb.gz  exists. Skipping downloadWarning:
pdbs/6HAM.pdb.gz  exists. Skipping downloadWarning: pdbs/4K46.pdb.gz  exists. Skipping downloadWarning: pdbs/3GMT.pdb.gz
exists. Skipping downloadWarning: pdbs/4PZL.pdb.gz  exists. Skipping download
```

Align & Repurpose Structures

Align

```{r}
# Align releated 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
...

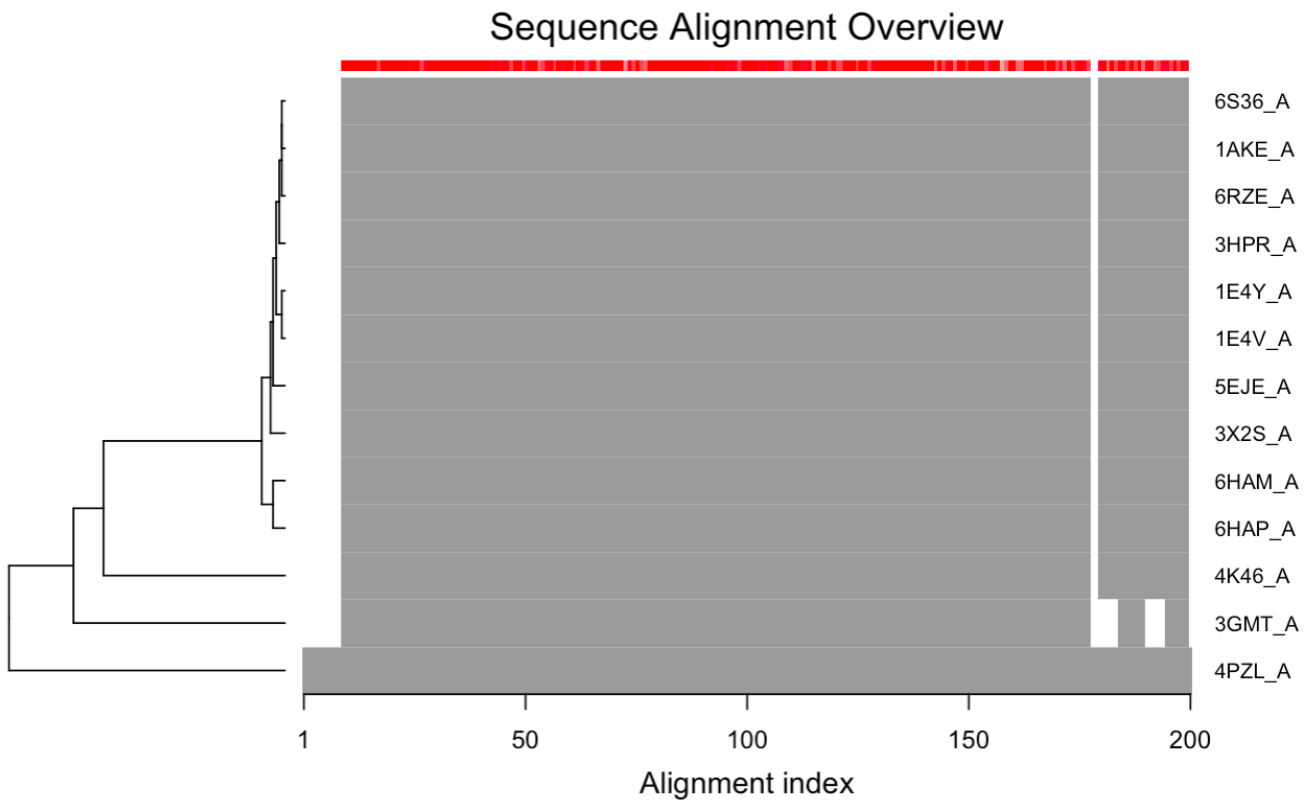
Extracting sequences

pdb/seq: 1   name: pdbs/split_chain/1AKE_A.pdb
```
```

Visualize:

```{r}
# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

# Draw schematic alignment
plot(pdbs, labels=ids)
```



Annotate PDB Structures

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

R Console

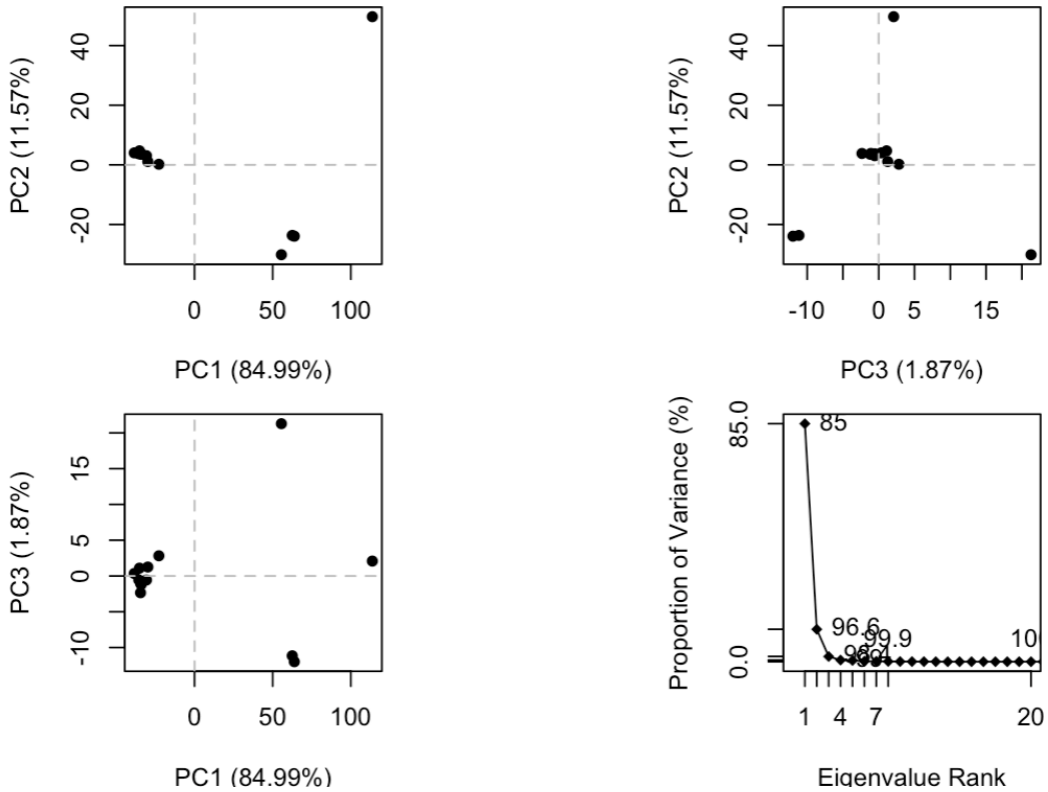
data.frame
13 x 17

	structureId <chr>	chainId <chr>	macromoleculeType <chr>	chainLength <int>	experimentalTechnique <chr>	resolution <dbl>
1AKE_A	1AKE	A	Protein	214	X-ray	2.00
6S36_A	6S36	A	Protein	214	X-ray	1.60
6RZE_A	6RZE	A	Protein	214	X-ray	1.69
3HPR_A	3HPR	A	Protein	214	X-ray	2.00
1E4V_A	1E4V	A	Protein	214	X-ray	1.85
5EJE_A	5EJE	A	Protein	214	X-ray	1.90
1E4Y_A	1E4Y	A	Protein	214	X-ray	1.85
3X2S_A	3X2S	A	Protein	214	X-ray	2.80
6HAP_A	6HAP	A	Protein	214	X-ray	2.70
6HAM_A	6HAM	A	Protein	214	X-ray	2.55

1–10 of 13 rows | 1–7 of 17 columns

Previous 1 2 Next

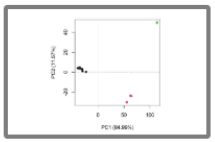
```
180
181 Principal Component Analysis
182 ```{r}
183 pc.xray <- pca(pdbs)
184 plot(pc.xray)
185 ```
```



```{r}
# Calculate RMSD
rd <- rmsd(pdbs)

# 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)
```



R Console