class09

Download a CSV file from the PDB site (accessible from "Analyze" > "PDB Statistics" > "by Experimental Method and Molecular Type". Move this CSV file into your RStudio project and use it to answer the following questions:

```
pdb <- read.csv("pdbdata.csv")
pdb
```

```
           Molecular.Type   X.ray     EM    NMR Multiple.methods Neutron Other
1          Protein (only) 161,663 12,592 12,337              200      74    32
2 Protein/Oligosaccharide   9,348  2,167     34                8       2     0
3              Protein/NA   8,404  3,924    286                7       0     0
4     Nucleic acid (only)   2,758    125  1,477               14       3     1
5                   Other     164      9     33                0       0     0
6  Oligosaccharide (only)      11      0      6                1       0     4
    Total
1 186,898
2  11,559
3  12,621
4   4,378
5     206
6      22
```

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

(182348 + 18817) / 215684 = 93.27%

Q2: What proportion of structures in the PDB are protein?

(161663 + 9348 + 8404) / 215684 = 83.18%

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?

4410

Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

Each circle represents all 3 atoms in the water molecule.

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

308? It appears between branches and to be interacting with the ligand.

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.

Discussion Topic: Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site?

Maybe if the protein changed in conformation to make the space even larger and allow the ligand in?

```
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)
```

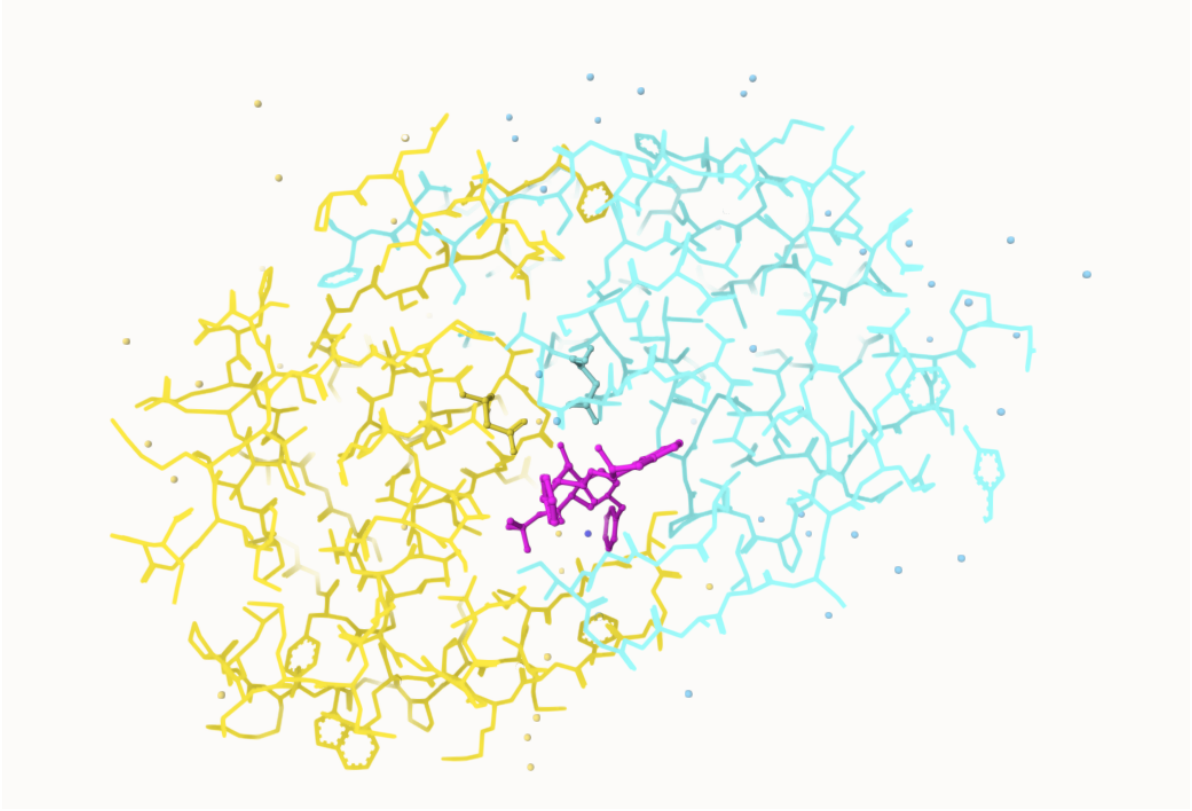


Figure 1: 1HSG in yellow and blue with ASP25 on Chain A and B and critical water in green highlight.

```
     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?

HOH or MK1

Q9: How many protein chains are in this structure?

2

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

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

```
adk
```

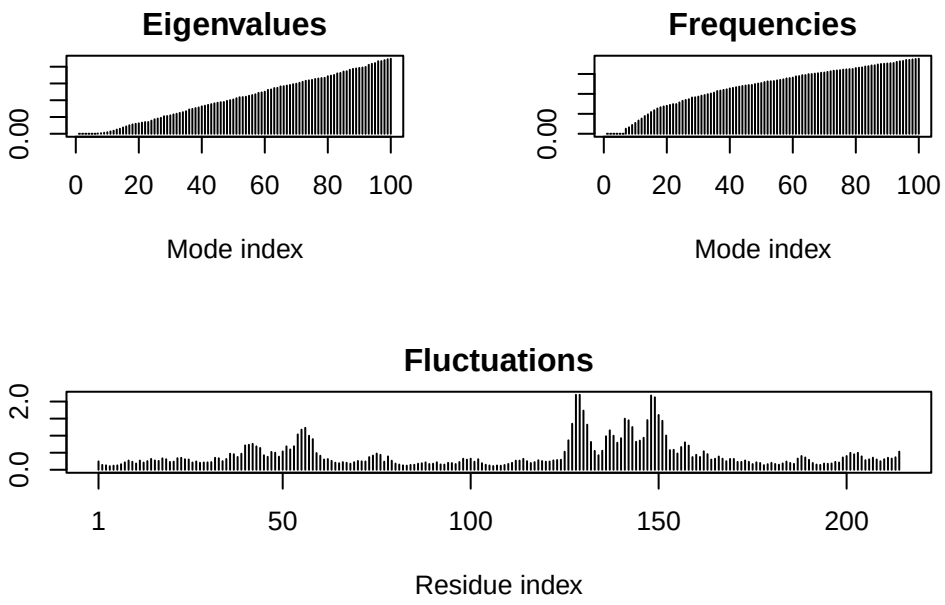
```
 Call:  read.pdb(file = "6s36")
```

```
   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:
      MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
      DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDKI
      VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
      YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG

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

```
# Perform flexiblity prediction
m <- nma(adk)
```

```
 Building Hessian...        Done in 0.021 seconds.
 Diagonalizing Hessian...   Done in 0.444 seconds.
```

```
plot(m)
```




```
# mktrj() section skipped for PDF
```

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?

bio3d-view

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

TRUE

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

```
Fetching... Please wait. Done.
```

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

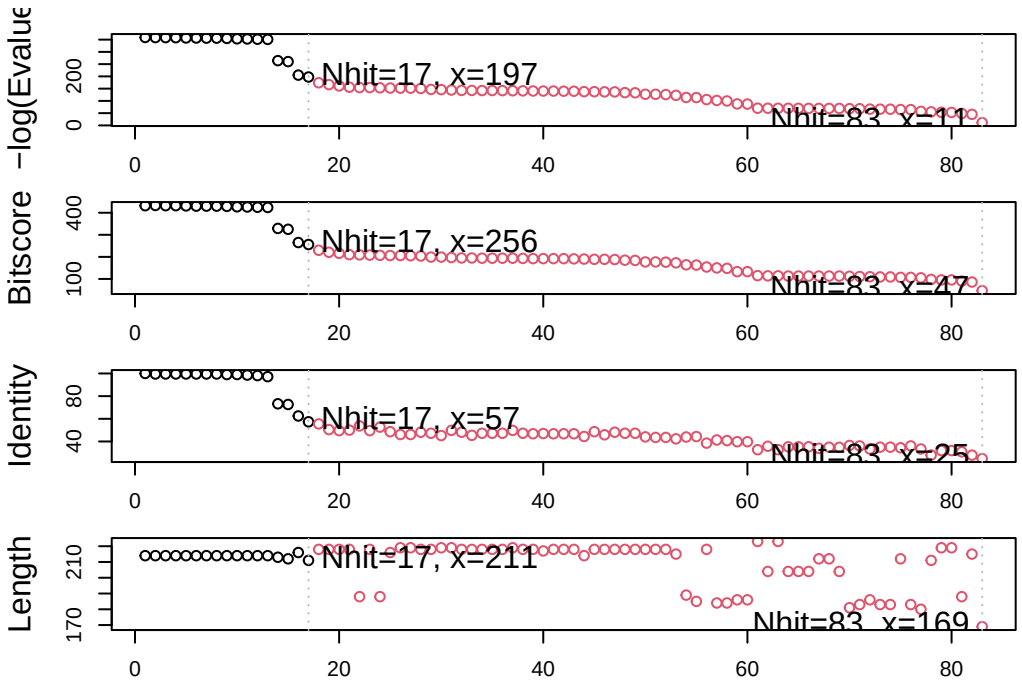
214

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

```
Searching ... please wait (updates every 5 seconds) RID = WY6H7XPT016
.....
Reporting 83 hits
```

```
# Plot a summary of search results
hits <- plot(b)
```

```
 * Possible cutoff values:    197 11
           Yielding Nhits:    17 83

 * Chosen cutoff value of:    197
           Yielding Nhits:    17
```



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

```
[1] "1AKE_A" "8BQF_A" "4X8M_A" "6S36_A" "6RZE_A" "4X8H_A"
```

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

```
  |
  |                                                                      |   0%
  |
  |====                                                                  |   6%
  |
  |========                                                              |  12%
  |
  |============                                                          |  18%
  |
  |================                                                      |  24%
  |
  |====================                                                  |  29%
  |
  |========================                                              |  35%
  |
  |============================                                          |  41%
  |
  |================================                                      |  47%
  |
  |====================================                                  |  53%
  |
  |========================================                              |  59%
  |
  |============================================                          |  65%
  |
  |================================================                      |  71%
  |
  |====================================================                  |  76%
  |
  |============================================================          |  82%
  |
```

```
  |=============================================================         |  88%
  |
  |=================================================================     |  94%
  |
  |======================================================================| 100%
```

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

```
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/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
....

Extracting sequences

pdb/seq: 1   name: pdbs/split_chain/1AKE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
```

```
pdb/seq: 2   name: pdbs/split_chain/8BQF_A.pdb
   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
pdb/seq: 5   name: pdbs/split_chain/6RZE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 6   name: pdbs/split_chain/4X8H_A.pdb
pdb/seq: 7   name: pdbs/split_chain/3HPR_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 8   name: pdbs/split_chain/1E4V_A.pdb
pdb/seq: 9   name: pdbs/split_chain/5EJE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 10   name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 11   name: pdbs/split_chain/3X2S_A.pdb
pdb/seq: 12   name: pdbs/split_chain/6HAP_A.pdb
pdb/seq: 13   name: pdbs/split_chain/6HAM_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 14   name: pdbs/split_chain/4K46_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 15   name: pdbs/split_chain/4NP6_A.pdb
pdb/seq: 16   name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 17   name: pdbs/split_chain/4PZL_A.pdb
```

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

```
# Draw schematic alignment
# plot(pdbs, labels=ids)
# omitting running this due to error "figure margins too large"
```

```
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] "Vibrio cholerae O1 biovar El Tor str. N16961"
[7] "Burkholderia pseudomallei 1710b"
```

```
[8] "Francisella tularensis subsp. tularensis SCHU S4"
```

```
anno
```

```
       structureId chainId macromoleculeType chainLength experimentalTechnique
1AKE_A        1AKE       A           Protein         214                 X-ray
8BQF_A        8BQF       A           Protein         234                 X-ray
4X8M_A        4X8M       A           Protein         214                 X-ray
6S36_A        6S36       A           Protein         214                 X-ray
6RZE_A        6RZE       A           Protein         214                 X-ray
4X8H_A        4X8H       A           Protein         214                 X-ray
3HPR_A        3HPR       A           Protein         214                 X-ray
1E4V_A        1E4V       A           Protein         214                 X-ray
5EJE_A        5EJE       A           Protein         214                 X-ray
1E4Y_A        1E4Y       A           Protein         214                 X-ray
3X2S_A        3X2S       A           Protein         214                 X-ray
6HAP_A        6HAP       A           Protein         214                 X-ray
6HAM_A        6HAM       A           Protein         214                 X-ray
4K46_A        4K46       A           Protein         214                 X-ray
4NP6_A        4NP6       A           Protein         217                 X-ray
3GMT_A        3GMT       A           Protein         230                 X-ray
4PZL_A        4PZL       A           Protein         242                 X-ray
       resolution       scopDomain                                    pfam
1AKE_A      2.000 Adenylate kinase Adenylate kinase, active site lid (ADK_lid)
8BQF_A      2.050             <NA>                      Adenylate kinase (ADK)
4X8M_A      2.600             <NA> Adenylate kinase, active site lid (ADK_lid)
6S36_A      1.600             <NA> Adenylate kinase, active site lid (ADK_lid)
6RZE_A      1.690             <NA> Adenylate kinase, active site lid (ADK_lid)
4X8H_A      2.500             <NA> Adenylate kinase, active site lid (ADK_lid)
3HPR_A      2.000             <NA>                      Adenylate kinase (ADK)
1E4V_A      1.850 Adenylate kinase                      Adenylate kinase (ADK)
5EJE_A      1.900             <NA> Adenylate kinase, active site lid (ADK_lid)
1E4Y_A      1.850 Adenylate kinase Adenylate kinase, active site lid (ADK_lid)
3X2S_A      2.800             <NA> Adenylate kinase, active site lid (ADK_lid)
6HAP_A      2.700             <NA>                      Adenylate kinase (ADK)
6HAM_A      2.550             <NA> Adenylate kinase, active site lid (ADK_lid)
4K46_A      2.010             <NA> Adenylate kinase, active site lid (ADK_lid)
4NP6_A      2.004             <NA> Adenylate kinase, active site lid (ADK_lid)
3GMT_A      2.100             <NA> Adenylate kinase, active site lid (ADK_lid)
4PZL_A      2.100             <NA> Adenylate kinase, active site lid (ADK_lid)
            ligandId
1AKE_A           AP5
```

8BQF_A	AP5
4X8M_A	<NA>
6S36_A	CL (3),NA,MG (2)
6RZE_A	NA (3),CL (2)
4X8H_A	<NA>
3HPR_A	AP5
1E4V_A	AP5
5EJE_A	AP5,CO
1E4Y_A	AP5
3X2S_A	JPY (2),AP5,MG
6HAP_A	AP5
6HAM_A	AP5
4K46_A	ADP,AMP,PO4
4NP6_A	<NA>
3GMT_A	SO4 (2)
4PZL_A	CA,FMT,GOL

	ligandName
1AKE_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
8BQF_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
4X8M_A	<NA>
6S36_A	CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)
6RZE_A	SODIUM ION (3),CHLORIDE ION (2)
4X8H_A	<NA>
3HPR_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
1E4V_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
5EJE_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE,COBALT (II) ION
1E4Y_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
3X2S_A	N-(pyren-1-ylmethyl)acetamide (2),BIS(ADENOSINE)-5'-PENTAPHOSPHATE,MAGNESIUM ION
6HAP_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
6HAM_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
4K46_A	ADENOSINE-5'-DIPHOSPHATE,ADENOSINE MONOPHOSPHATE,PHOSPHATE ION
4NP6_A	<NA>
3GMT_A	SULFATE ION (2)
4PZL_A	CALCIUM ION,FORMIC ACID,GLYCEROL

	source
1AKE_A	Escherichia coli
8BQF_A	Escherichia coli
4X8M_A	Escherichia coli
6S36_A	Escherichia coli
6RZE_A	Escherichia coli
4X8H_A	Escherichia coli
3HPR_A	Escherichia coli K-12
1E4V_A	Escherichia coli

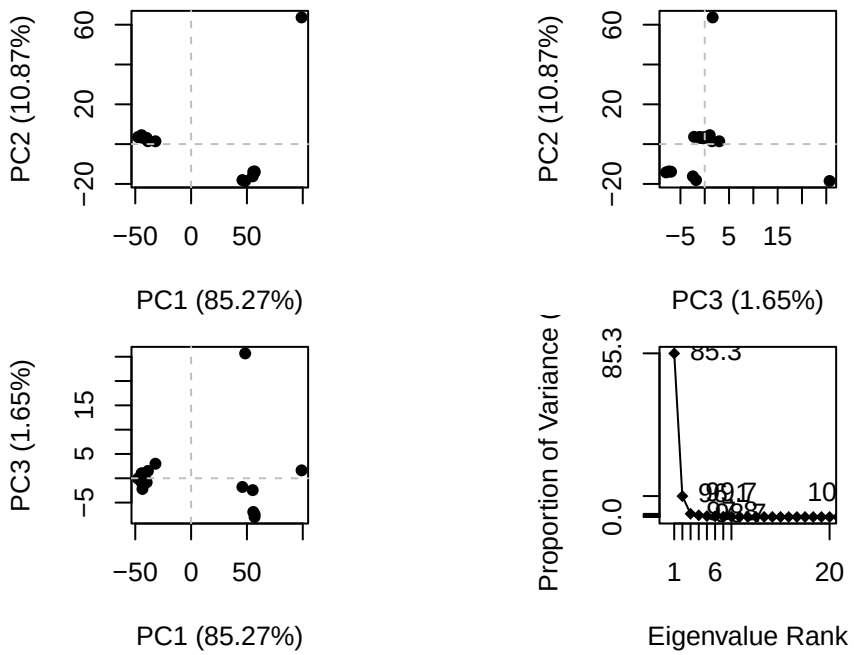
5EJE_A	Escherichia coli O139:H28 str. E24377A				
1E4Y_A	Escherichia coli				
3X2S_A	Escherichia coli str. K-12 substr. MDS42				
6HAP_A	Escherichia coli O139:H28 str. E24377A				
6HAM_A	Escherichia coli K-12				
4K46_A	Photobacterium profundum				
4NP6_A	Vibrio cholerae O1 biovar El Tor str. N16961				
3GMT_A	Burkholderia pseudomallei 1710b				
4PZL_A	Francisella tularensis subsp. tularensis SCHU S4				

1AKE_A	STRUCTURE OF THE COMPLEX BETWEEN ADENYLATE KINASE FROM ESCHERICHIA COLI AND THE INHIBI				
8BQF_A					
4X8M_A					
6S36_A					
6RZE_A					
4X8H_A					
3HPR_A					
1E4V_A					
5EJE_A					Cryst
1E4Y_A					
3X2S_A					
6HAP_A					
6HAM_A					
4K46_A					
4NP6_A					
3GMT_A					
4PZL_A					The cryst

	citation	rObserved	rFree
1AKE_A	Muller, C.W., et al. J Mol Biol (1992)	0.19600	NA
8BQF_A	Scheerer, D., et al. Proc Natl Acad Sci U S A (2023)	0.22073	0.25789
4X8M_A	Kovermann, M., et al. Nat Commun (2015)	0.24910	0.30890
6S36_A	Rogne, P., et al. Biochemistry (2019)	0.16320	0.23560
6RZE_A	Rogne, P., et al. Biochemistry (2019)	0.18650	0.23500
4X8H_A	Kovermann, M., et al. Nat Commun (2015)	0.19610	0.28950
3HPR_A	Schrank, T.P., et al. Proc Natl Acad Sci U S A (2009)	0.21000	0.24320
1E4V_A	Muller, C.W., et al. Proteins (1993)	0.19600	NA
5EJE_A	Kovermann, M., et al. Proc Natl Acad Sci U S A (2017)	0.18890	0.23580
1E4Y_A	Muller, C.W., et al. Proteins (1993)	0.17800	NA
3X2S_A	Fujii, A., et al. Bioconjug Chem (2015)	0.20700	0.25600
6HAP_A	Kantaev, R., et al. J Phys Chem B (2018)	0.22630	0.27760
6HAM_A	Kantaev, R., et al. J Phys Chem B (2018)	0.20511	0.24325
4K46_A	Cho, Y.-J., et al. To be published	0.17000	0.22290
4NP6_A	Kim, Y., et al. To be published	0.18800	0.22200

```
3GMT_A Buchko, G.W., et al. Biochem Biophys Res Commun (2010)   0.23800 0.29500
4PZL_A                        Tan, K., et al. To be published   0.19360 0.23680
         rWork spaceGroup
1AKE_A 0.19600  P 21 2 21
8BQF_A 0.21882  P 2 21 21
4X8M_A 0.24630    C 1 2 1
6S36_A 0.15940    C 1 2 1
6RZE_A 0.18190    C 1 2 1
4X8H_A 0.19140    C 1 2 1
3HPR_A 0.20620  P 21 21 2
1E4V_A 0.19600  P 21 2 21
5EJE_A 0.18630  P 21 2 21
1E4Y_A 0.17800   P 1 21 1
3X2S_A 0.20700 P 21 21 21
6HAP_A 0.22370    I 2 2 2
6HAM_A 0.20311       P 43
4K46_A 0.16730 P 21 21 21
4NP6_A 0.18600       P 43
3GMT_A 0.23500   P 1 21 1
4PZL_A 0.19130       P 32
```

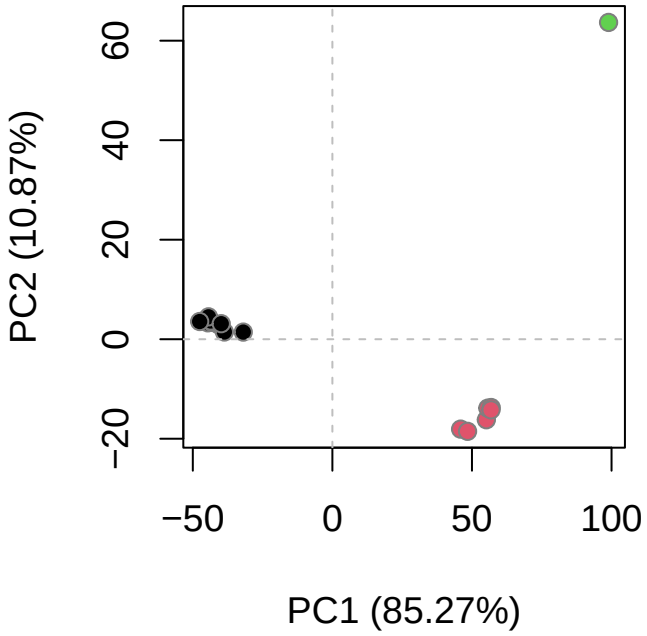
```
# Perform PCA
pc.xray <- pca(pdbs)
plot(pc.xray)
```




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

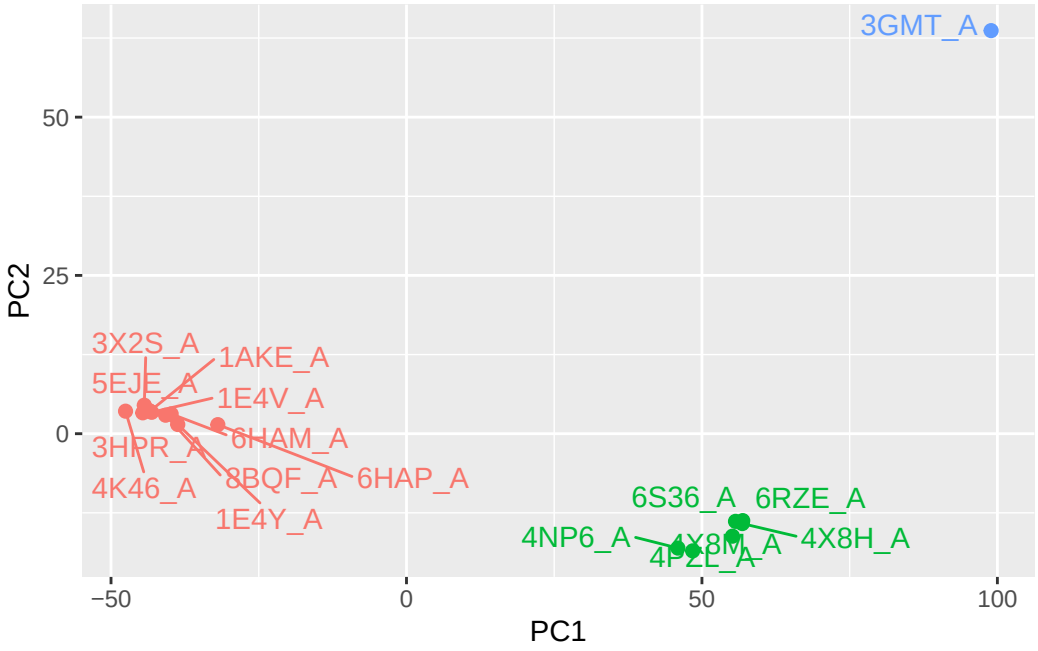
```
Warning in rmsd(pdbs): No indices provided, using the 199 non NA positions
```

```
# 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)
```




```
# mktrj omitted for pdf
```

```
#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
```



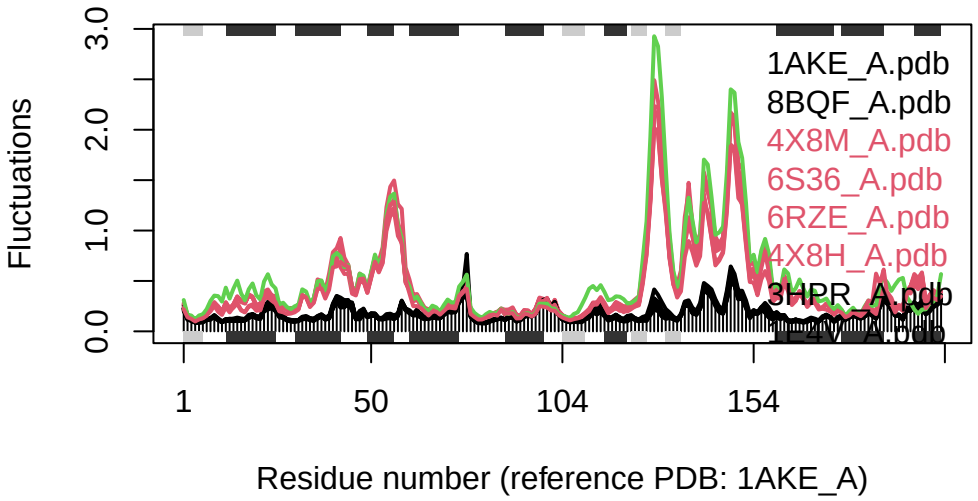

```
# NMA of all structures
modes <- nma(pdbs)
```

```
Warning in nma.pdbs(pdbs): 8BQF_A.pdb might have missing residue(s) in structure:
   Fluctuations at neighboring positions may be affected.


Details of Scheduled Calculation:
  ... 17 input structures
  ... storing 591 eigenvectors for each structure
  ... dimension of x$U.subspace: ( 597x591x17 )
  ... coordinate superposition prior to NM calculation
  ... aligned eigenvectors (gap containing positions removed)
  ... estimated memory usage of final 'eNMA' object: 45.9 Mb


  |
  |                                                                      |   0%
  |
  |====                                                                  |   6%
  |
```

```
  |========                                                              |  12%
  |
  |============                                                          |  18%
  |
  |================                                                      |  24%
  |
  |====================                                                  |  29%
  |
  |========================                                              |  35%
  |
  |============================                                          |  41%
  |
  |================================                                      |  47%
  |
  |====================================                                  |  53%
  |
  |========================================                              |  59%
  |
  |============================================                          |  65%
  |
  |================================================                      |  71%
  |
  |====================================================                  |  76%
  |
  |============================================================          |  82%
  |
  |================================================================      |  88%
  |
  |====================================================================  |  94%
  |
  |======================================================================| 100%
```

```
plot(modes, pdbs, col=grps.rd)
```

```
Extracting SSE from pdbs$sse attribute
```



Q14. What do you note about this plot? Are the black and colored lines similar or different? Where do you think they differ most and why?

The black lines are definitely different from the colored lines, especially around the 50 and 120-160 residues. This might because of the difference in conformations producing the 2 different (B&W vs. colored) plots. The residues mentioned indicate nucleotide-binding site regions that change the most in displacement during nucleotide binding.