class09

Steven Gan

2022-10-26

Table of contents

1: Introduction to the RCSB Protein Data Bank (PDB)

PDB statistics

```
library(dplyr)
```

```
Attaching package: 'dplyr'
```

```
The following objects are masked from 'package:stats':

    filter, lag
```

```
The following objects are masked from 'package:base':

    intersect, setdiff, setequal, union
```

```
setClass("num.with.commas")
setAs("character", "num.with.commas",
        function(from) as.numeric(gsub(",", "", from)))

pdb.stat <- read.csv(file = "Data Export Summary.csv", colClasses = "num.with.commas")
```

```
Warning in asMethod(object): NAs introduced by coercion
```

```
pdb.stat[1] <- read.csv(file = "Data Export Summary.csv")[1]
row.names(pdb.stat) <- pdb.stat[,1]
tot.pdb <- colSums(pdb.stat[-1], na.rm = T)
pdb.stat <- rbind(pdb.stat[-1], tot.pdb)
row.names(pdb.stat)[length(pdb.stat)] <- "Total"
```

Q1:

```
xemR = round((pdb.stat["Total", "X.ray"] + pdb.stat["Total", "EM"]) /
               pdb.stat["Total", "Total"] * 100, 2)
print(paste0("Percentage of structures are solved by X-Ray and EM: ",
             xemR, "%"))
```

```
[1] "Percentage of structures are solved by X-Ray and EM: 92.81%"
```

Q2:

```
proR = round((pdb.stat["Protein (only)", "Total"] +
                pdb.stat["Protein/Oligosaccharide", "Total"]) /
               pdb.stat["Total", "Total"] * 100, 2)
print(paste0("Proportion of structures in the PDB are protein: ", proR, "%"))
```

```
[1] "Proportion of structures in the PDB are protein: 92.3%"
```

Q3:

1264 HIV-1 protease structures are in the current PDB

2. Visualizing the HIV-1 protease structure

Using Mol*; Getting to knowHIV-Pr; Saving an image

Devling deeper

Q4:

Since crystallographic experiments do not resolve hydrogen molecules, thus only oxygen molecules are displayed.

Q5:

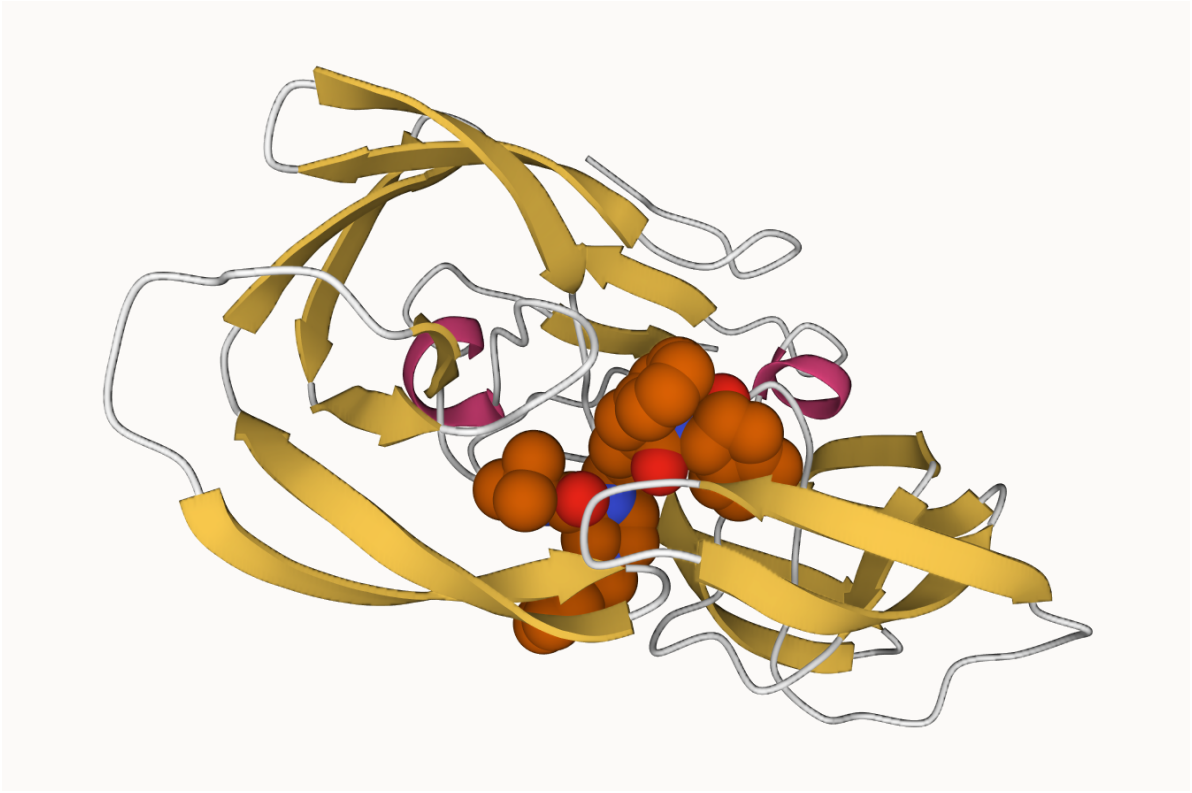
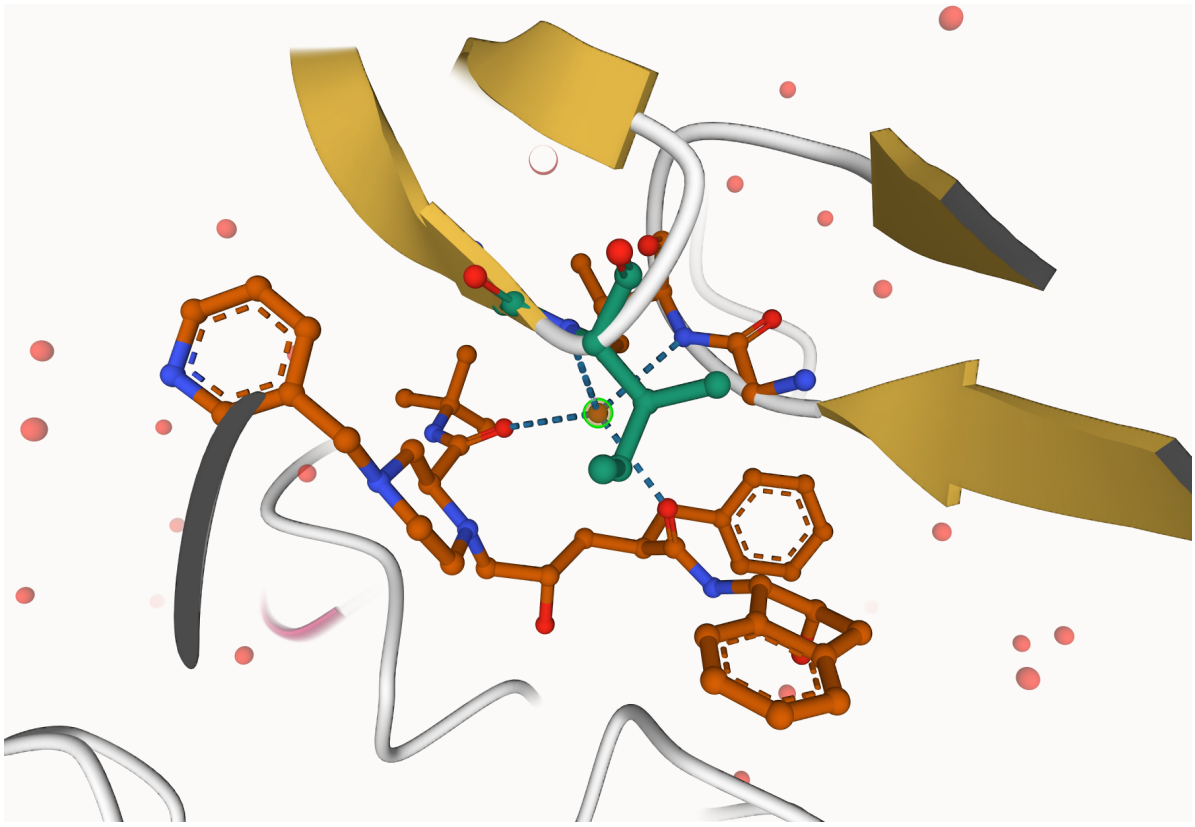


Figure 1: 1HSG Image



Identified in the graph above, residue 308.

Q6:

3. Introduction to Bio3D in R

```
library(bio3d)
```

Reading PDB file data into R

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

```
Note: Accessing on-line PDB file
```

```
pdb
```

Figure 2: 1HSG Water D25 Image

```
 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:

198 amino acid residues are there in this pdb object.

Q8:

HOH is one of the two non-protein residues.

Q9:

2 protein chains are in this structure.

```
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>
```

Predicting functional motions of a single structure

```
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
```

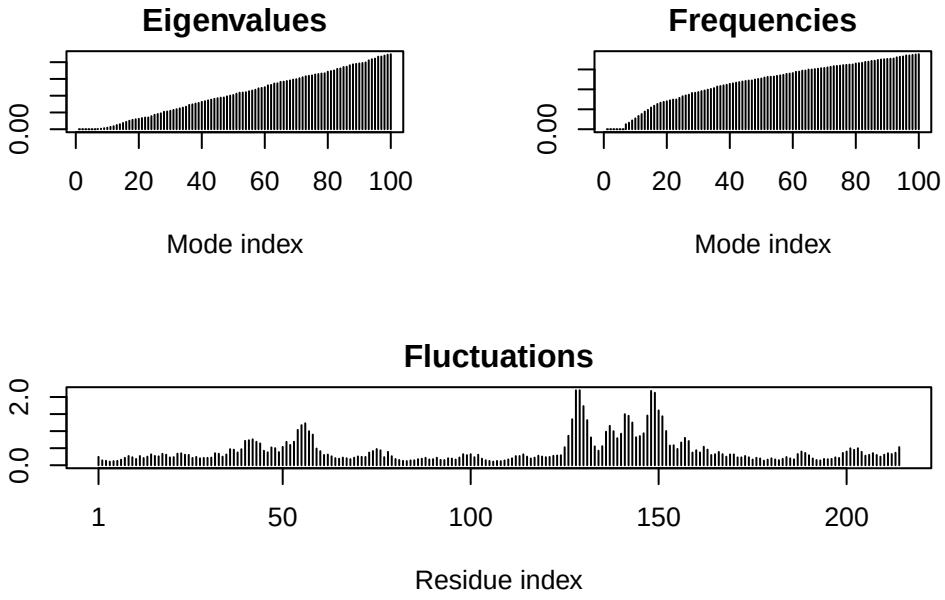
```
      YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG

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

```
adk.nma <- nma(adk)
```

```
 Building Hessian...        Done in 0.109 seconds.
 Diagonalizing Hessian...   Done in 0.341 seconds.
```

```
plot(adk.nma)
```




```
mktrj(adk.nma, file="adk_m7.pdb")
```

Figure 3: ADK Movement Image

4. Comparative structure analysis of Adenylate Kinase

Setup

```
# Install packages in the R console NOT your Rmd/Quarto file

# install.packages("bio3d")
# install.packages("devtools")
# install.packages("BiocManager")

# BiocManager::install("msa")
# devtools::install_bitbucket("Grantlab/bio3d-view")
```

Q10:

msa package is found only on BioConductor and not CRAN.

Q11:

bio3d-view is not found on BioConductor or CRAN

Q12:

TRUE: Functions from the devtools package can be used to install packages from GitHub and BitBucket

Search and retrieve ADK structures

```
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   DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
            61        .         .         .         .         .         120

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

           181        .         .         .   214
pdb|1AKE|A   YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG
           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:

214 amino acids in this sequence

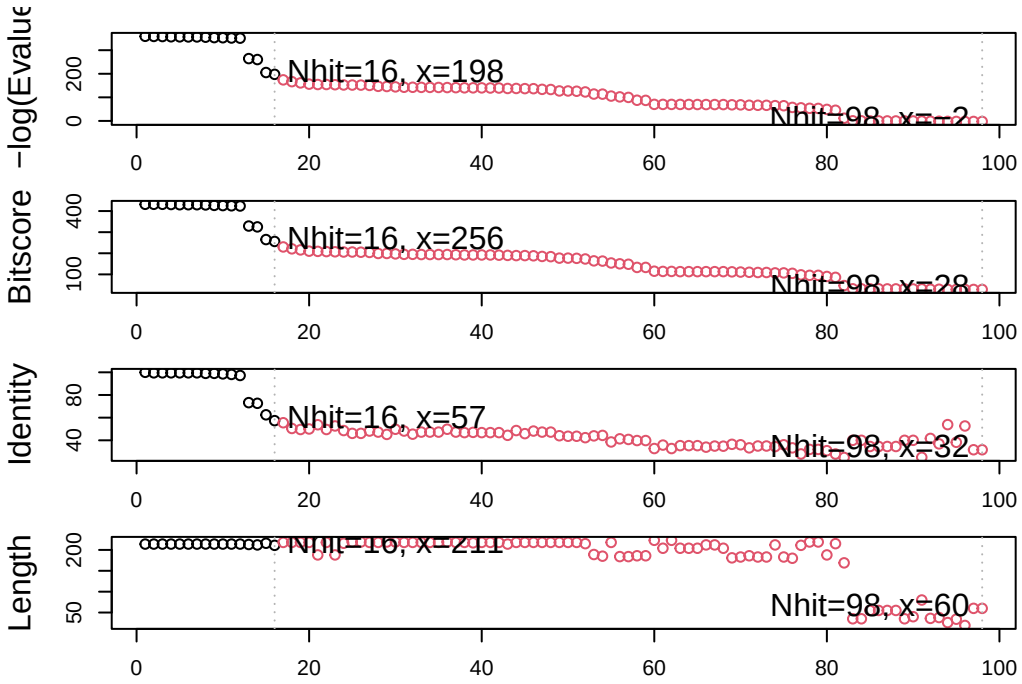
```
b <- blast.pdb(aa)
```

```
Searching ... please wait (updates every 5 seconds) RID = NKHVKUCD016
...................
Reporting 98 hits
```

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

```
 * Possible cutoff values:    197 -3
           Yielding Nhits:    16 98
```

```
 * Chosen cutoff value of:    197
          Yielding Nhits:    16
```




```
head(hits$pdb.id)
```

```
[1] "1AKE_A" "4X8M_A" "6S36_A" "6RZE_A" "4X8H_A" "3HPR_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.gz exists. Skipping download
```

```
Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
4X8M.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/
4X8H.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/
4NP6.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%
  |
  |====                                                                  |   6%
  |
  |=========                                                             |  12%
  |
  |=============                                                         |  19%
  |
  |==================                                                    |  25%
  |
  |======================                                                |  31%
  |
  |==========================                                            |  38%
  |
  |==============================                                        |  44%
  |
  |==================================                                    |  50%
  |
  |======================================                                |  56%
  |
  |==========================================                            |  62%
  |
  |==============================================                        |  69%
  |
  |==================================================                    |  75%
  |
  |======================================================                |  81%
  |
  |==========================================================            |  88%
  |
  |==============================================================        |  94%
  |
  |======================================================================| 100%
```

Align and superpose structures

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

```
Reading PDB files:
pdbs/split_chain/1AKE_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
....

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/4X8M_A.pdb
pdb/seq: 3   name: pdbs/split_chain/6S36_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 4   name: pdbs/split_chain/6RZE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 5   name: pdbs/split_chain/4X8H_A.pdb
pdb/seq: 6   name: pdbs/split_chain/3HPR_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 7   name: pdbs/split_chain/1E4V_A.pdb
pdb/seq: 8   name: pdbs/split_chain/5EJE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 9   name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 10   name: pdbs/split_chain/3X2S_A.pdb
pdb/seq: 11   name: pdbs/split_chain/6HAP_A.pdb
pdb/seq: 12   name: pdbs/split_chain/6HAM_A.pdb
```

```
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 13   name: pdbs/split_chain/4K46_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 14   name: pdbs/split_chain/4NP6_A.pdb
pdb/seq: 15   name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 16   name: pdbs/split_chain/4PZL_A.pdb
```

```
ids <- basename.pdb(pdbs$id)
#plot(pdbs, labels = ids)
```

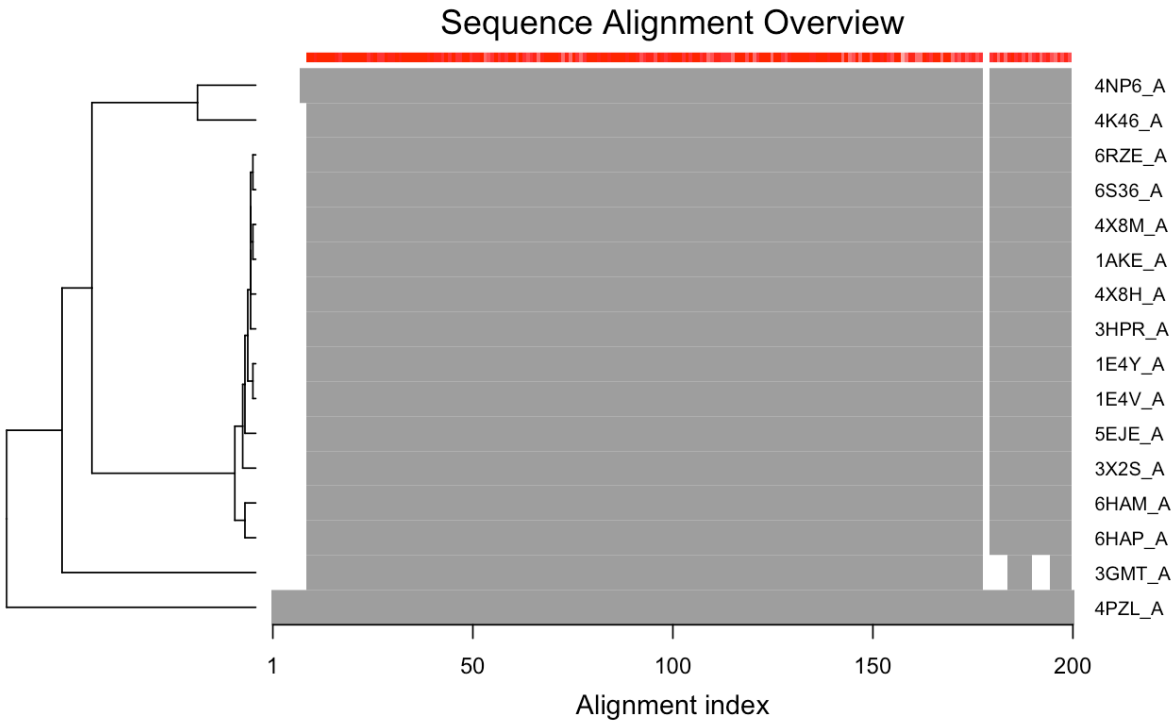


Figure 4: Seq Align Overview

Optional: Viewing our superposed structures

```
library(bio3d.view)
library(rgl)

view.pdbs(pdbs)
```

Annotate collected PDB structures

```
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
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)
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, active site lid (ADK_lid)
```

1E4V_A	1.850	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)	
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, active site lid (ADK_lid)	
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	ligandName
1AKE_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
4X8M_A	<NA>	<NA>
6S36_A	CL (3),NA,MG (2)	CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)
6RZE_A	NA (3),CL (2)	SODIUM ION (3),CHLORIDE ION (2)
4X8H_A	<NA>	<NA>
3HPR_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
1E4V_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
5EJE_A	AP5,CO	BIS(ADENOSINE)-5'-PENTAPHOSPHATE,COBALT (II) ION
1E4Y_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
3X2S_A	JPY (2),AP5,MG	N-(pyren-1-ylmethyl)acetamide (2),BIS(ADENOSINE)-5'-PENTAPHOSPHATE,MAGNESIUM ION
6HAP_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
6HAM_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
4K46_A	ADP,AMP,PO4	ADENOSINE-5'-DIPHOSPHATE,ADENOSINE MONOPHOSPHATE,PHOSPHATE ION
4NP6_A	<NA>	<NA>
3GMT_A	SO4 (2)	SULFATE ION (2)
4PZL_A	CA,FMT,GOL	

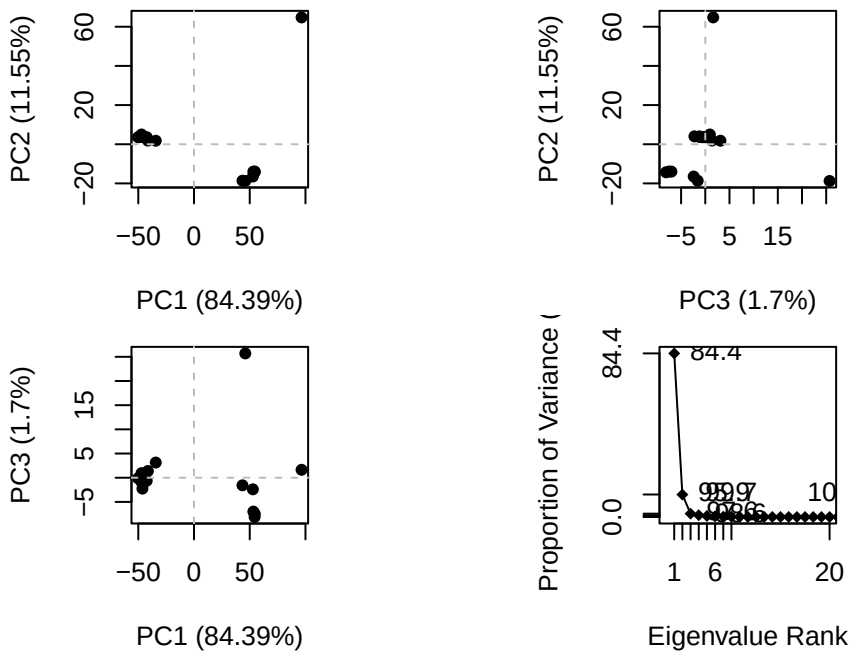
```
4PZL_A                                                CALCIUM ION,FORMIC ACID,GLYCEROL
                                                 source
1AKE_A                                  Escherichia coli
4X8M_A                                  Escherichia coli
6S36_A                                  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
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
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
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
```

Principal component analysis

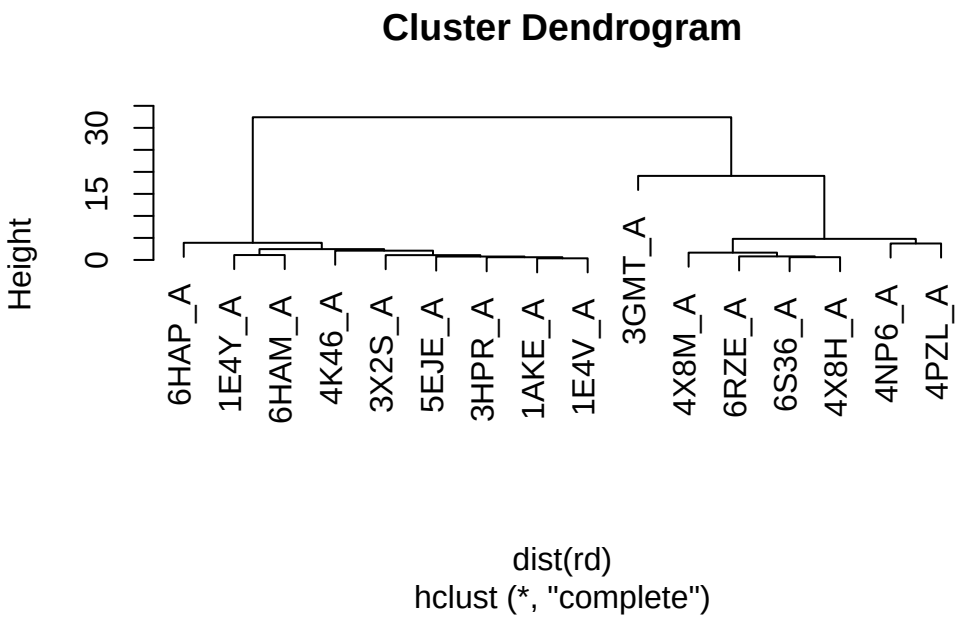
```
pc.xray <- pca(pdbs)
plot(pc.xray)
```



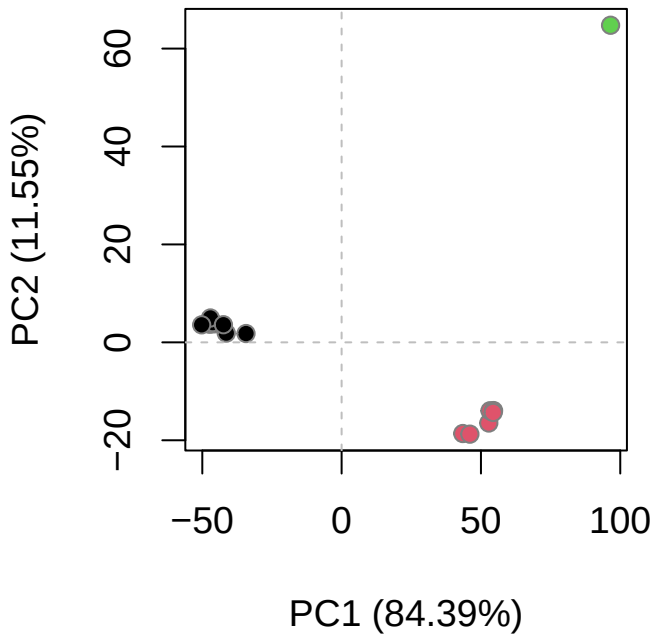

```
rd <- rmsd(pdbs)
```

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

```
hc.rd <- hclust(dist(rd))
plot(hc.rd)
```




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



5. Optional further visualization

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

```
view.xyz(pc1)
```

```
Potential all C-alpha atom structure(s) detected: Using calpha.connectivity()
```

```
view.xyz(pc1, col = vec2color(rmsf(pc1)))
```

```
Potential all C-alpha atom structure(s) detected: Using calpha.connectivity()
```

```
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
```

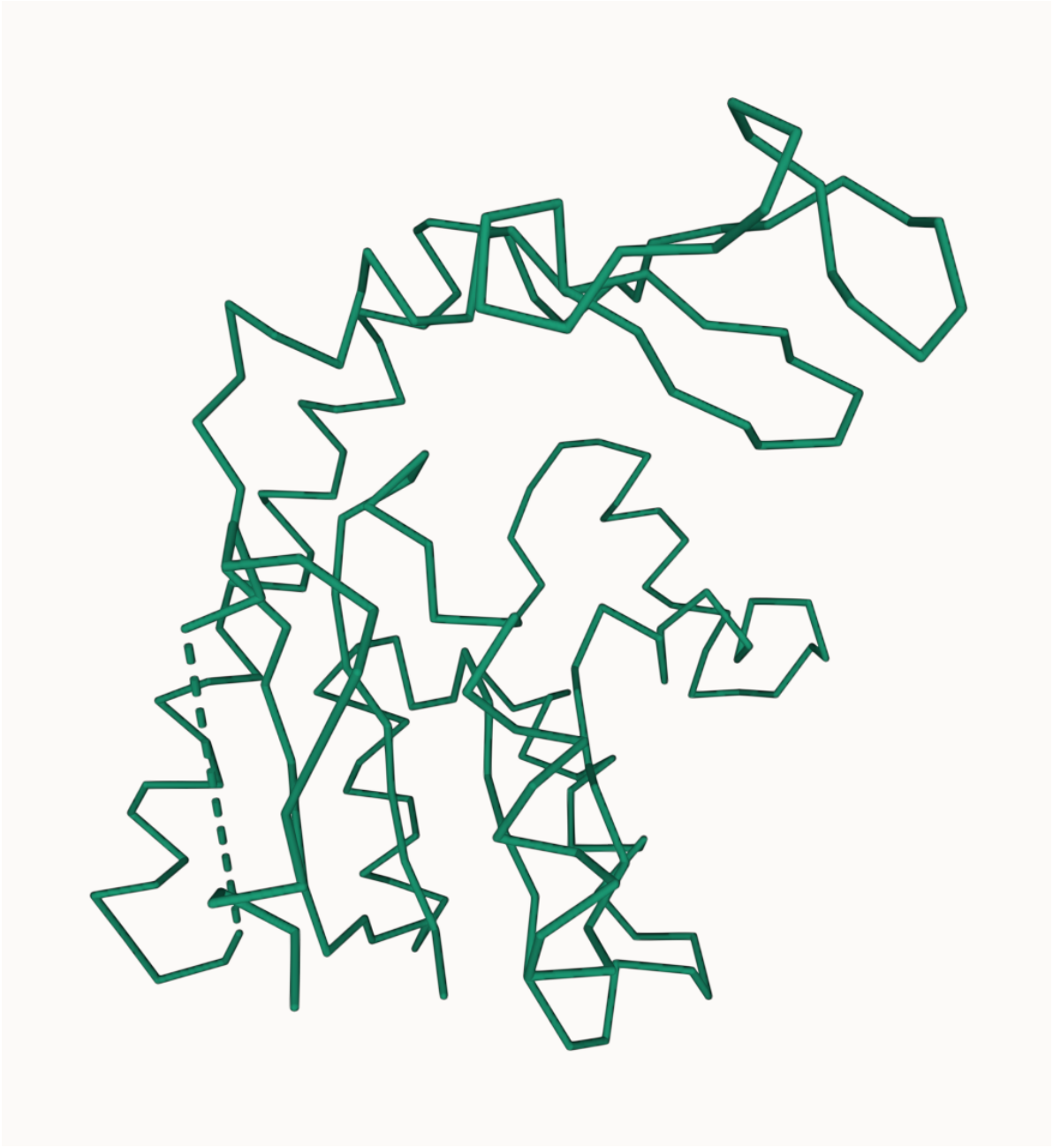
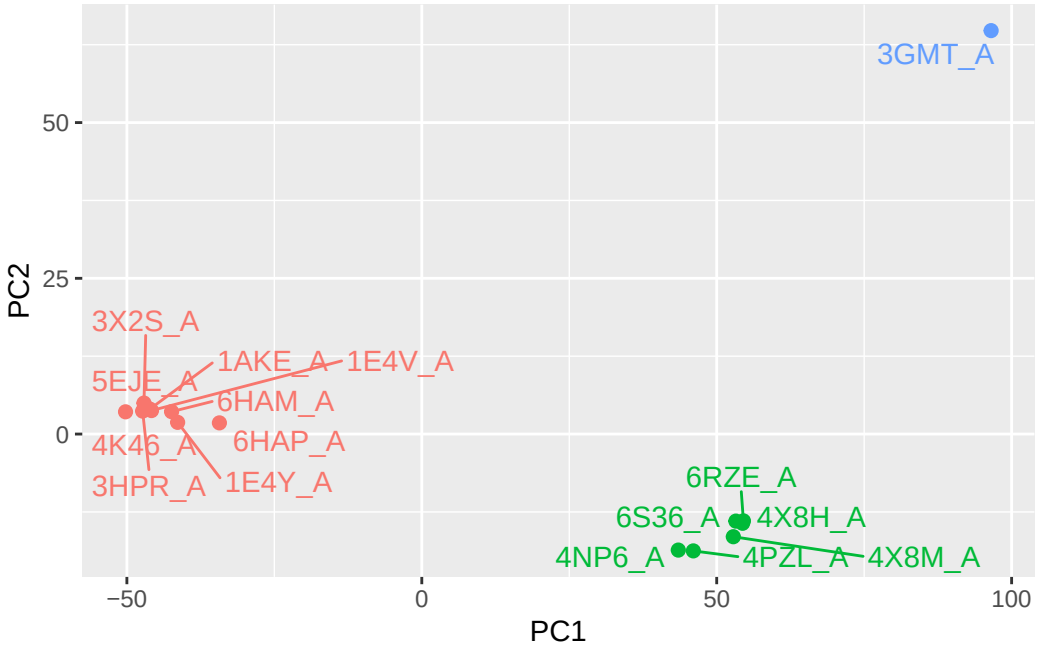


Figure 5: PCA Image



6. Normal mode analysis [optional]

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

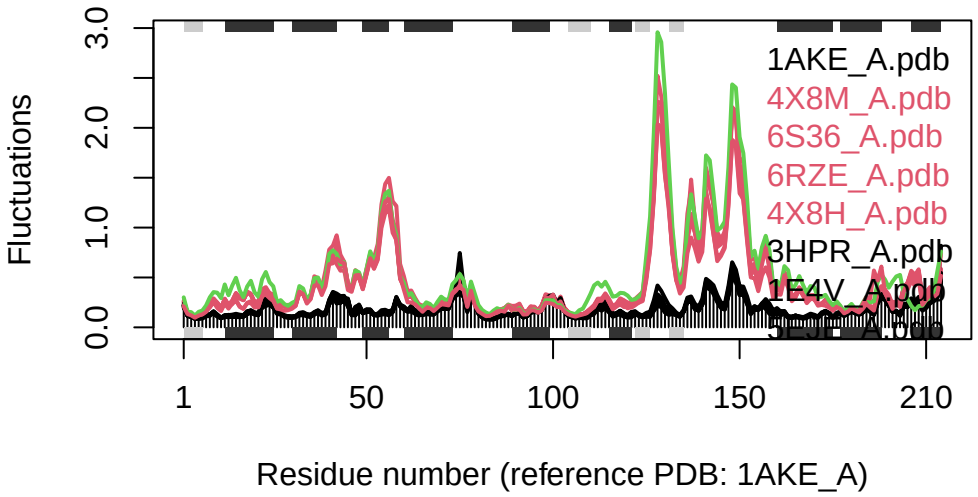
```
Details of Scheduled Calculation:
  ... 16 input structures
  ... storing 606 eigenvectors for each structure
  ... dimension of x$U.subspace: ( 612x606x16 )
  ... coordinate superposition prior to NM calculation
  ... aligned eigenvectors (gap containing positions removed)
  ... estimated memory usage of final 'eNMA' object: 45.4 Mb


  |
  |                                                                      |   0%
  |
  |====                                                                  |   6%
  |
  |=========                                                             |  12%
```

```
  |
  |=============                                                         |  19%
  |
  |==================                                                    |  25%
  |
  |======================                                                |  31%
  |
  |==========================                                            |  38%
  |
  |================================                                      |  44%
  |
  |====================================                                  |  50%
  |
  |========================================                              |  56%
  |
  |============================================                          |  62%
  |
  |================================================                      |  69%
  |
  |====================================================                  |  75%
  |
  |============================================================          |  81%
  |
  |================================================================      |  88%
  |
  |====================================================================  |  94%
  |
  |======================================================================| 100%
```

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

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



Q14:

Red and Green groups exhibit similar flexible states accessibility, and are significantly distinguished with Black group. This corresponds with the PCA results, as Red and Green groups are closer and Black group is much far away from them. High flexibility regions tend to appear at regions without fixed secondary structures.