09_Lab_Report

Angela Abraham

2025-10-28

Introduction to the RCSB Protein Data Bank (PDB)

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

```
head(pdb_stats)
```

```
##            Molecular.Type   X.ray     EM    NMR Integrative Multiple.methods
## 1          Protein (only) 176,378 20,438 12,709         342               221
## 2 Protein/Oligosaccharide  10,284  3,396     34           8                11
## 3              Protein/NA   9,007  5,931    287          24                 7
## 4     Nucleic acid (only)   3,077    200  1,554           2                15
## 5                   Other     174     13     33           3                 0
## 6  Oligosaccharide (only)      11      0      6           0                 1
##   Neutron Other   Total
## 1      83    32 210,203
## 2       1     0  13,734
## 3       0     0  15,256
## 4       3     1   4,852
## 5       0     0     223
## 6       0     4      22
```

```
str(pdb_stats)
```

```
## 'data.frame':    6 obs. of  9 variables:
##  $ Molecular.Type  : chr  "Protein (only)" "Protein/Oligosaccharide" "Protein/NA" "Nucleic acid (only
##  $ X.ray           : chr  "176,378" "10,284" "9,007" "3,077" ...
##  $ EM              : chr  "20,438" "3,396" "5,931" "200" ...
##  $ NMR             : chr  "12,709" "34" "287" "1,554" ...
##  $ Integrative     : int  342 8 24 2 3 0
##  $ Multiple.methods: int  221 11 7 15 0 1
##  $ Neutron         : int  83 1 0 3 0 0
##  $ Other           : int  32 0 0 1 0 4
##  $ Total           : chr  "210,203" "13,734" "15,256" "4,852" ...
```

```
clean_numeric <- function(x) {
  as.numeric(gsub(",", "", x))
}
pdb_stats_clean <- pdb_stats
```

```
pdb_stats_clean[, 2:9] <- lapply(pdb_stats_clean[, 2:9], clean_numeric)

head(pdb_stats_clean)
```

```
##            Molecular.Type  X.ray    EM   NMR Integrative Multiple.methods
## 1          Protein (only) 176378 20438 12709         342               221
## 2 Protein/Oligosaccharide  10284  3396    34           8                11
## 3              Protein/NA   9007  5931   287          24                 7
## 4     Nucleic acid (only)   3077   200  1554           2                15
## 5                   Other    174    13    33           3                 0
## 6  Oligosaccharide (only)     11     0     6           0                 1
##   Neutron Other  Total
## 1      83    32 210203
## 2       1     0  13734
## 3       0     0  15256
## 4       3     1   4852
## 5       0     0    223
## 6       0     4     22
```

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

```
total_structures <- sum(pdb_stats_clean$Total)
xray_structures <- sum(pdb_stats_clean$X.ray)

em_structures <- sum(pdb_stats_clean$EM)

xray_percent <- (xray_structures / total_structures) * 100
em_percent <- (em_structures / total_structures) * 100
combined_percent <- xray_percent + em_percent
print(xray_percent)
```

```
## [1] 81.43231
```

```
print(em_percent)
```

```
## [1] 12.27148
```

#X-Ray percent is 81.43% and Electron Microscopy is 12.27%.

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

```
protein_only<-pdb_stats_clean[pdb_stats_clean$Molecular.Type=="Protein (only)", "Total"]
protein_proportion<-(protein_only/total_structures) * 100
print(protein_proportion)
```

```
## [1] 86.0465
```

#86.04% of structures in the PDB are proteins.

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?

#There is approximately 1,150 HIV-1 protease structures in the current PDB.

```
pdb_text<-readLines("1hsg.pdb")
head(pdb_text, 50)
```

```
##  [1] "HEADER    HYDROLASE (ACID PROTEINASE)             31-MAR-95   1HSG              "
##  [2] "TITLE     CRYSTAL STRUCTURE AT 1.9 ANGSTROMS RESOLUTION OF HUMAN                "
##  [3] "TITLE    2 IMMUNODEFICIENCY VIRUS (HIV) II PROTEASE COMPLEXED WITH L-735,524, AN"
##  [4] "TITLE    3 ORALLY BIOAVAILABLE INHIBITOR OF THE HIV PROTEASES                   "
##  [5] "COMPND    MOL_ID: 1;                                                           "
##  [6] "COMPND   2 MOLECULE: HIV-1 PROTEASE;                                           "
##  [7] "COMPND   3 CHAIN: A, B;                                                        "
##  [8] "COMPND   4 EC: 3.4.23.-;                                                       "
##  [9] "COMPND   5 ENGINEERED: YES;                                                    "
## [10] "COMPND   6 OTHER_DETAILS: NY5 ISOLATE                                          "
## [11] "SOURCE    MOL_ID: 1;                                                          "
## [12] "SOURCE   2 ORGANISM_SCIENTIFIC: HUMAN IMMUNODEFICIENCY VIRUS 1;               "
## [13] "SOURCE   3 ORGANISM_TAXID: 11676;                                             "
## [14] "SOURCE   4 GENE: HIV-1 PROTEASE FROM THE NY5 ISOLATE;                         "
## [15] "SOURCE   5 EXPRESSION_SYSTEM: ESCHERICHIA COLI;                               "
## [16] "SOURCE   6 EXPRESSION_SYSTEM_TAXID: 562                                       "
## [17] "KEYWDS    HYDROLASE (ACID PROTEINASE)                                         "
## [18] "EXPDTA    X-RAY DIFFRACTION                                                   "
## [19] "AUTHOR    Z.CHEN                                                              "
## [20] "REVDAT   4   07-FEB-24 1HSG    1       REMARK                                 "
## [21] "REVDAT   3   24-FEB-09 1HSG    1       VERSN                                  "
## [22] "REVDAT   2   01-APR-03 1HSG    1       JRNL                                   "
## [23] "REVDAT   1   03-APR-96 1HSG    0                                              "
## [24] "JRNL        AUTH   Z.CHEN,Y.LI,E.CHEN,D.L.HALL,P.L.DARKE,C.CULBERSON,          "
## [25] "JRNL        AUTH 2 J.A.SHAFER,L.C.KUO                                         "
## [26] "JRNL        TITL   CRYSTAL STRUCTURE AT 1.9-A RESOLUTION OF HUMAN             "
## [27] "JRNL        TITL 2 IMMUNODEFICIENCY VIRUS (HIV) II PROTEASE COMPLEXED WITH    "
## [28] "JRNL        TITL 3 L-735,524, AN ORALLY BIOAVAILABLE INHIBITOR OF THE HIV     "
## [29] "JRNL        TITL 4 PROTEASES.                                                 "
## [30] "JRNL        REF    J.BIOL.CHEM.                 V. 269 26344 1994            "
## [31] "JRNL        REFN                   ISSN 0021-9258                            "
## [32] "JRNL        PMID   7929352                                                   "
## [33] "REMARK   2                                                                    "
## [34] "REMARK   2 RESOLUTION.    2.00 ANGSTROMS.                                     "
## [35] "REMARK   3                                                                    "
## [36] "REMARK   3 REFINEMENT.                                                        "
## [37] "REMARK   3   PROGRAM     : X-PLOR                                             "
## [38] "REMARK   3   AUTHORS     : BRUNGER                                            "
## [39] "REMARK   3                                                                    "
## [40] "REMARK   3  DATA USED IN REFINEMENT.                                          "
## [41] "REMARK   3   RESOLUTION RANGE HIGH (ANGSTROMS) : 2.00                         "
## [42] "REMARK   3   RESOLUTION RANGE LOW  (ANGSTROMS) : NULL                         "
## [43] "REMARK   3   DATA CUTOFF            (SIGMA(F)) : NULL                         "
## [44] "REMARK   3   DATA CUTOFF HIGH         (ABS(F)) : NULL                         "
## [45] "REMARK   3   DATA CUTOFF LOW          (ABS(F)) : NULL                         "
```

```
## [46] "REMARK   3   COMPLETENESS (WORKING+TEST)   (%) : NULL                          "
## [47] "REMARK   3   NUMBER OF REFLECTIONS             : NULL                          "
## [48] "REMARK   3                                                                     "
## [49] "REMARK   3  FIT TO DATA USED IN REFINEMENT.                                    "
## [50] "REMARK   3   CROSS-VALIDATION METHOD          : NULL                           "
```

```
atom_lines<-grep("^ATOM", pdb_text, value=TRUE)
head(atom_lines)
```

```
## [1] "ATOM      1  N   PRO A   1      29.361  39.686   5.862  1.00 38.10           N  "
## [2] "ATOM      2  CA  PRO A   1      30.307  38.663   5.319  1.00 40.62           C  "
## [3] "ATOM      3  C   PRO A   1      29.760  38.071   4.022  1.00 42.64           C  "
## [4] "ATOM      4  O   PRO A   1      28.600  38.302   3.676  1.00 43.40           O  "
## [5] "ATOM      5  CB  PRO A   1      30.508  37.541   6.342  1.00 37.87           C  "
## [6] "ATOM      6  CG  PRO A   1      29.296  37.591   7.162  1.00 38.40           C  "
```

Visualizing the HIV-1 protease structure

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

#We only see one atom per water molecule in this structure because this is the water molecule that is interacting with the other molecules; this is showing an interaction and not the whole picture.

> 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.

#Yes, W301; residue number is 301.

> 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).

```
knitr::include_graphics("1HSG_1.png")
```

Introduction to Bio3D in R

```
library(bio3d)
```

```
## Warning: package 'bio3d' was built under R version 4.5.2
```

```
list.files()
```

```
##  [1] "09_Lab_Report.pdf"
##  [2] "09_Lab_Report.Rmd"
##  [3] "09_Lab_Report_files"
##  [4] "1HSG.pdb"
##  [5] "1HSG_1.png"
##  [6] "adk_m7.pdb"
##  [7] "alignment_plot.png"
##  [8] "aln.fa"
##  [9] "Data Export Summary.csv"
## [10] "nma_plot.png"
## [11] "pc_1.pdb"
## [12] "pdbs"
## [13] "pdbs_aligned.rds"
## [14] "seqs.fasta"
## [15] "Structural Bioinformatics (Part 1).Rproj"
```

```
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. How many amino acid residues are there is this pdb object?

#198 amino acid residues

Q8. Name one of the two non-protein residues?

#HOH

Q9. How many protein chains are in this structure?

#2

```
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>
```

```
library(bio3d)
pdb<-read.pdb("1HSG")
```

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

```
## Warning in get.pdb(file, path = tempdir(), verbose = FALSE):
## C:\Users\Angela\AppData\Local\Temp\Rtmpyuq4C9/1HSG.pdb exists. Skipping
## download
```

```
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
```

```
sele <- atom.select(pdb, resno=25)
print(sele)
```

```
##
##  Call:  atom.select.pdb(pdb = pdb, resno = 25)
##
##    Atom Indices#: 16  ($atom)
##    XYZ  Indices#: 48  ($xyz)
##
## + attr: atom, xyz, call
```

```
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
```

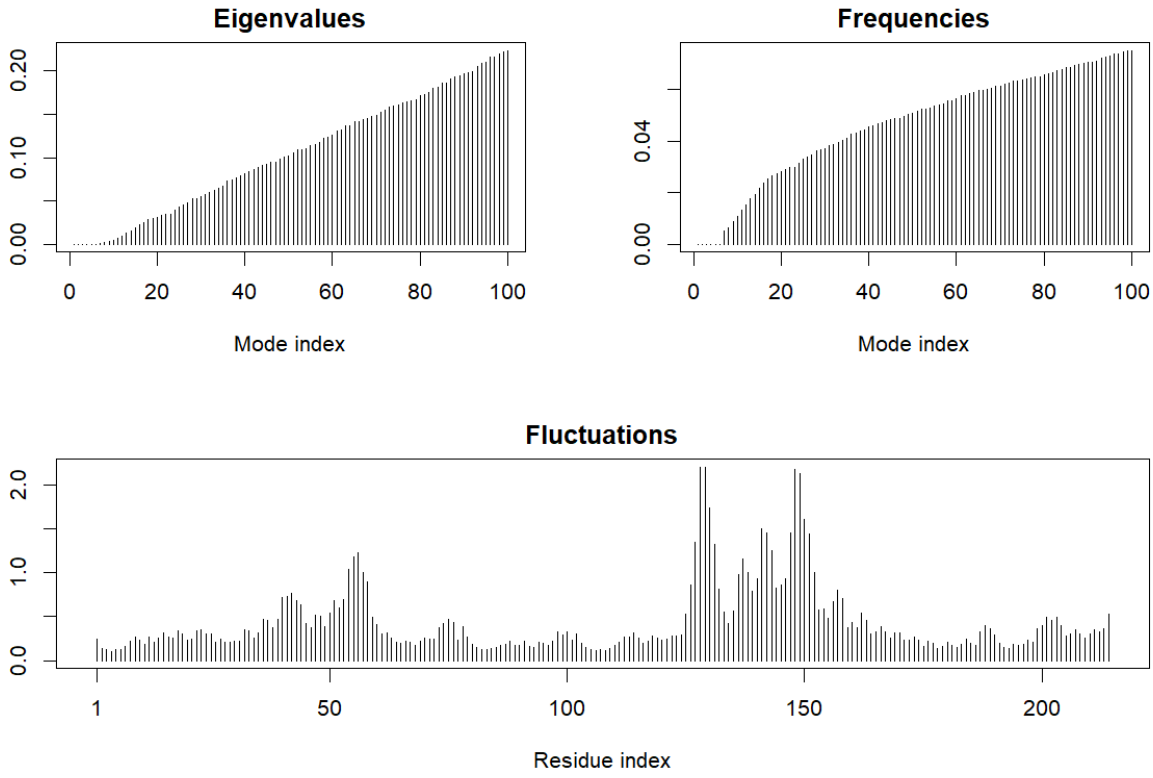
```
m<-nma(adk)
```

```
##  Building Hessian...     Done in 0.02 seconds.
##  Diagonalizing Hessian...    Done in 0.25 seconds.
```

```
png("nma_plot.png", width=1200,height=800,res=150)
plot(m)
dev.off()
```

```
## pdf
##   2
```

```
knitr::include_graphics("nma_plot.png")
```




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

Comparitive structure analysis of Adenylate Kinase

```
library(bio3d)
```

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?

#bio3dview

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

#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   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
```

```
length(aa$ali)
```

```
## [1] 214
```

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

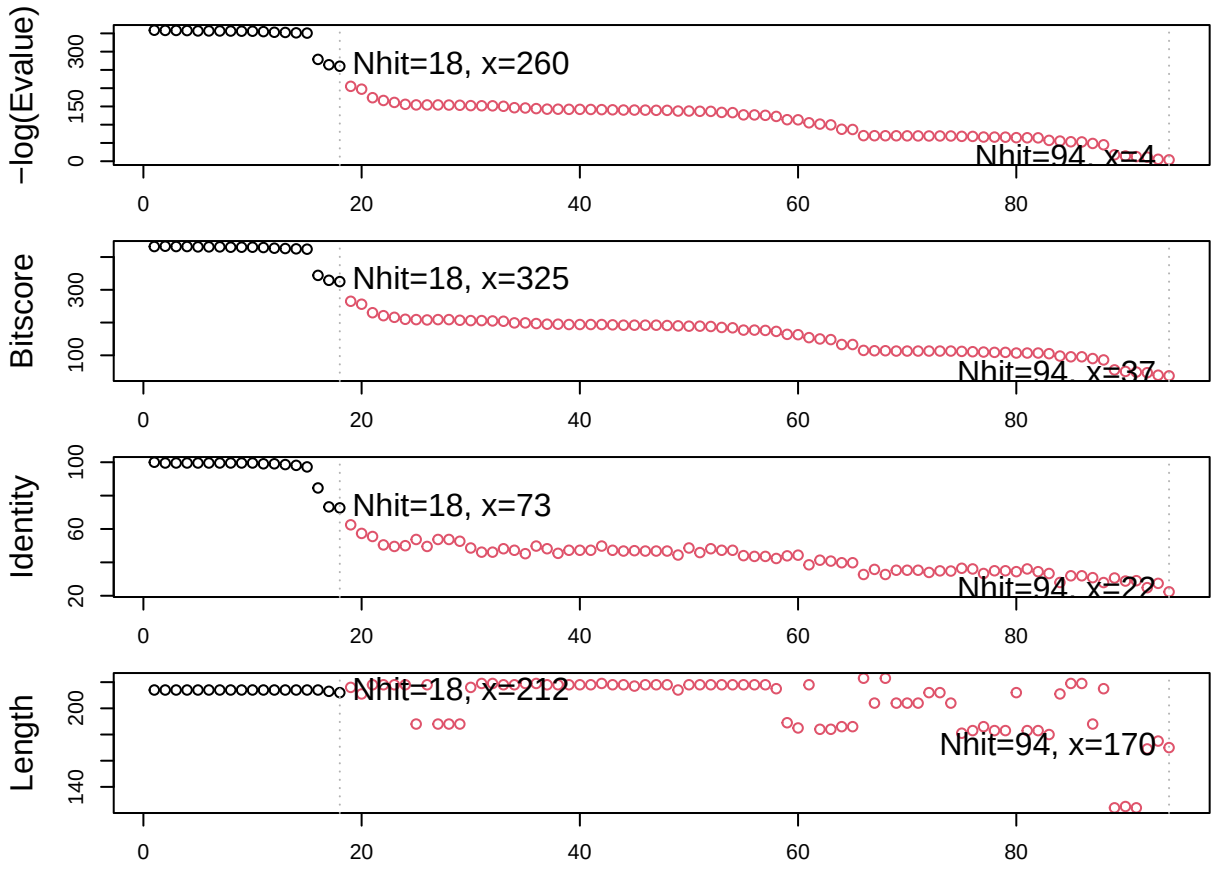
#214 amino acids in this sequence

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

```
##  Searching ... please wait (updates every 5 seconds) RID = GFVPKYS3014
##  .
##  Reporting 94 hits
```

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

```
##   * Possible cutoff values:    260 3
##              Yielding Nhits:    18 94
##
##   * Chosen cutoff value of:    260
##              Yielding Nhits:    18
```




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

```
##  [1] "1AKE_A" "6S36_A" "6RZE_A" "3HPR_A" "1E4V_A" "5EJE_A" "1E4Y_A" "3X2S_A"
##  [9] "6HAP_A" "6HAM_A" "4K46_A" "3GMT_A" "4PZL_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

##   |                                                                      |
```

```
pdbs<-pdbaln(files, web.args=list(email="ala021@ucsd.edu"))
```

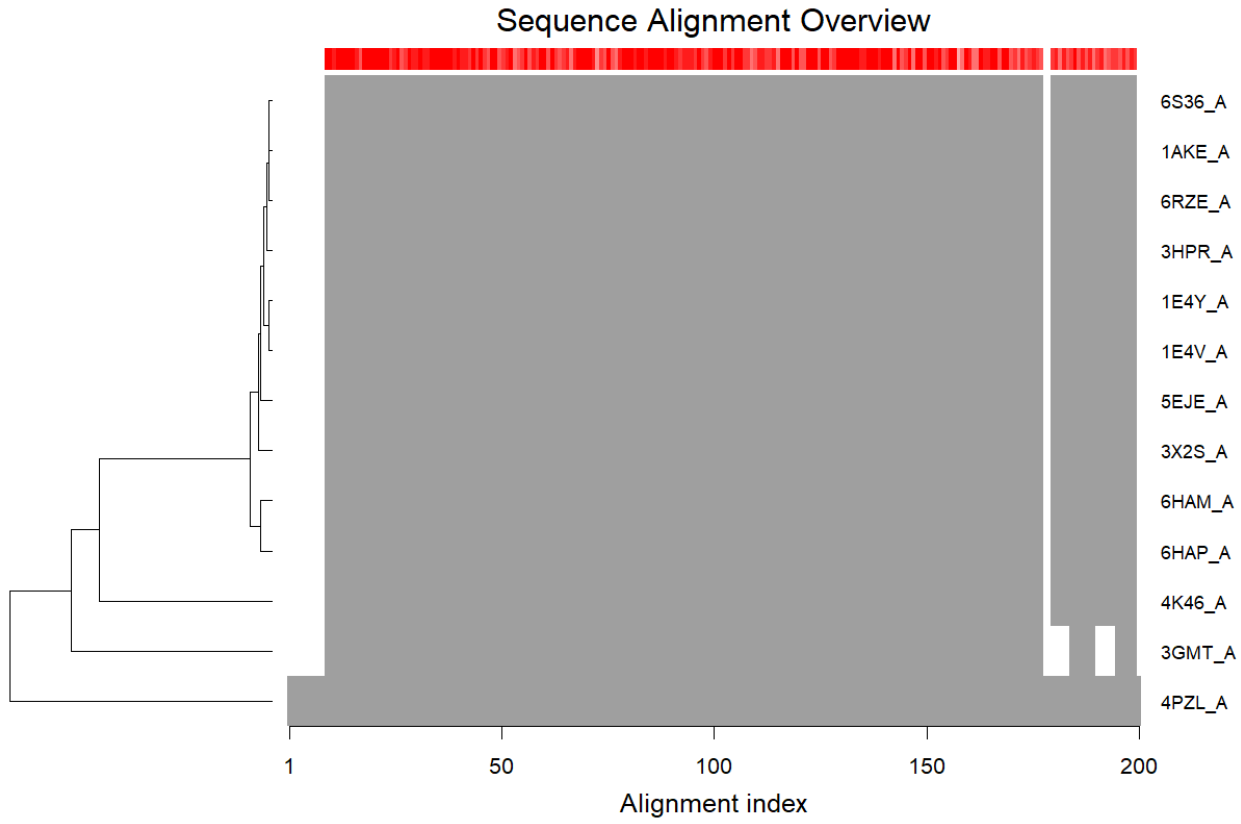
```
## Reading PDB files:
## pdbs/split_chain/1AKE_A.pdb
## pdbs/split_chain/6S36_A.pdb
## pdbs/split_chain/6RZE_A.pdb
## pdbs/split_chain/3HPR_A.pdb
## pdbs/split_chain/1E4V_A.pdb
## pdbs/split_chain/5EJE_A.pdb
## pdbs/split_chain/1E4Y_A.pdb
## pdbs/split_chain/3X2S_A.pdb
## pdbs/split_chain/6HAP_A.pdb
## 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
##
##
## Will try to align sequences online...
##
## Job successfully submited (job ID: muscle-R20251102-235308-0293-33436624-p1m)
## Waiting for job to finish...Done.
##
## 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/6S36_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 3   name: pdbs/split_chain/6RZE_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 4   name: pdbs/split_chain/3HPR_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 5   name: pdbs/split_chain/1E4V_A.pdb
## pdb/seq: 6   name: pdbs/split_chain/5EJE_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 7   name: pdbs/split_chain/1E4Y_A.pdb
## pdb/seq: 8   name: pdbs/split_chain/3X2S_A.pdb
## pdb/seq: 9   name: pdbs/split_chain/6HAP_A.pdb
## pdb/seq: 10   name: pdbs/split_chain/6HAM_A.pdb
##    PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 11   name: pdbs/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
```

```
saveRDS(pdbs, "pdbs_aligned.rds")
ids<-basename.pdb(pdbs$id)
```

```
png("alignment_plot.png", width=1200, height=800, res=150)
plot(pdbs, labels=ids)
dev.off()
```

```
## pdf
##   2
```

```
knitr::include_graphics("alignment_plot.png")
```




```
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"
```

```
anno
```

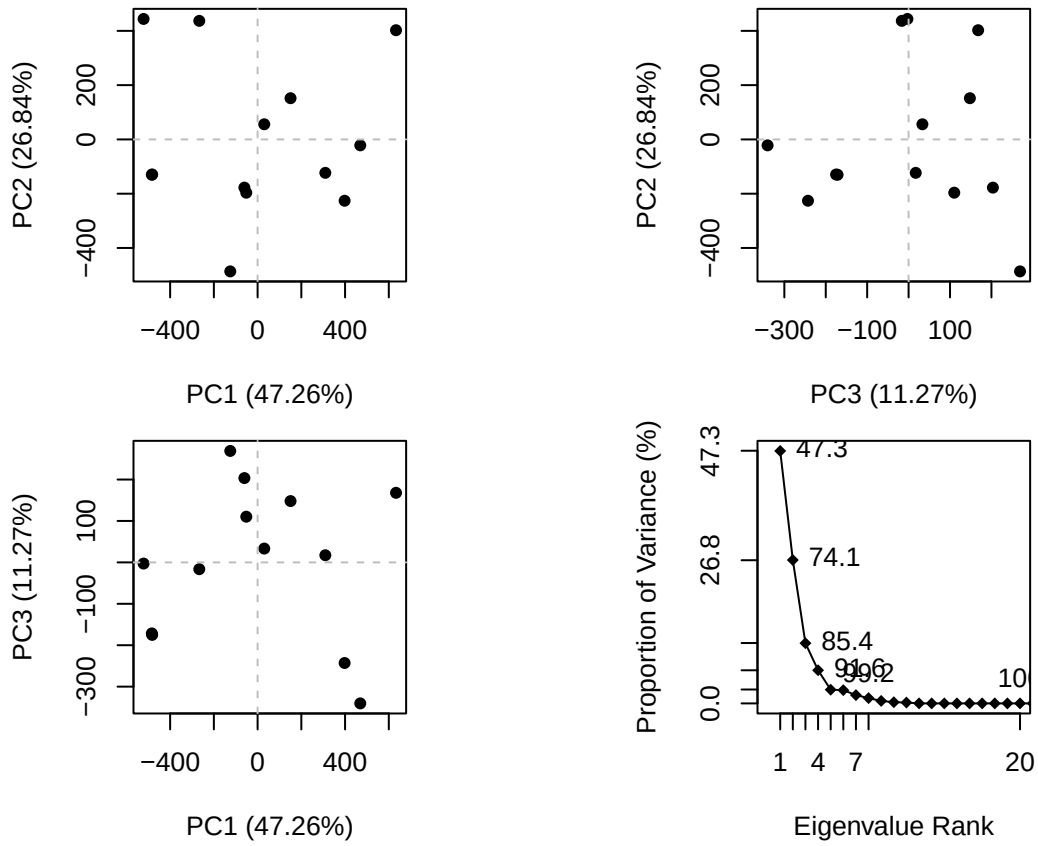
```
##        structureId chainId macromoleculeType chainLength experimentalTechnique
## 1AKE_A        1AKE       A           Protein         214                 X-ray
```

```
## 6S36_A        6S36       A           Protein         214                 X-ray
## 6RZE_A        6RZE       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
## 3GMT_A        3GMT       A           Protein         230                 X-ray
## 4PZL_A        4PZL       A           Protein         242                 X-ray
##        resolution       scopDomain                                    pfam
## 1AKE_A       2.00 Adenylate kinase                    Adenylate kinase (ADK)
## 6S36_A       1.60             <NA> Adenylate kinase, active site lid (ADK_lid)
## 6RZE_A       1.69             <NA>                    Adenylate kinase (ADK)
## 3HPR_A       2.00             <NA>                    Adenylate kinase (ADK)
## 1E4V_A       1.85 Adenylate kinase                    Adenylate kinase (ADK)
## 5EJE_A       1.90             <NA>                    Adenylate kinase (ADK)
## 1E4Y_A       1.85 Adenylate kinase                    Adenylate kinase (ADK)
## 3X2S_A       2.80             <NA>                    Adenylate kinase (ADK)
## 6HAP_A       2.70             <NA> Adenylate kinase, active site lid (ADK_lid)
## 6HAM_A       2.55             <NA>                    Adenylate kinase (ADK)
## 4K46_A       2.01             <NA>                    Adenylate kinase (ADK)
## 3GMT_A       2.10             <NA>                    Adenylate kinase (ADK)
## 4PZL_A       2.10             <NA> Adenylate kinase, active site lid (ADK_lid)
##                ligandId
## 1AKE_A              AP5
## 6S36_A CL (3),NA,MG (2)
## 6RZE_A    NA (3),CL (2)
## 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
## 3GMT_A          SO4 (2)
## 4PZL_A       CA,FMT,GOL
##                                                                        ligandName
## 1AKE_A                                                 BIS(ADENOSINE)-5'-PENTAPHOSPHATE
## 6S36_A                                  CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)
## 6RZE_A                                                  SODIUM ION (3),CHLORIDE ION (2)
## 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
## 3GMT_A                                                                  SULFATE ION (2)
## 4PZL_A                                                 CALCIUM ION,FORMIC ACID,GLYCEROL
```

```
##                                                 source
## 1AKE_A                                Escherichia coli
## 6S36_A                                Escherichia coli
## 6RZE_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
## 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 INHIBITOR AP5A
## 6S36_A
## 6RZE_A
## 3HPR_A
## 1E4V_A
## 5EJE_A                                                                                  Crystal struc
## 1E4Y_A
## 3X2S_A
## 6HAP_A
## 6HAM_A
## 4K46_A
## 3GMT_A
## 4PZL_A                                                                                The crystal struc
##                                                    citation rObserved   rFree
## 1AKE_A              Muller, C.W., et al. J Mol Biology (1992)   0.19600      NA
## 6S36_A                  Rogne, P., et al. Biochemistry (2019)   0.16320 0.23560
## 6RZE_A                  Rogne, P., et al. Biochemistry (2019)   0.18650 0.23500
## 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
## 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
## 6S36_A 0.15940    C 1 2 1
## 6RZE_A 0.18190    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
```

```
## 3GMT_A 0.23500   P 1 21 1
## 4PZL_A 0.19130       P 32
```

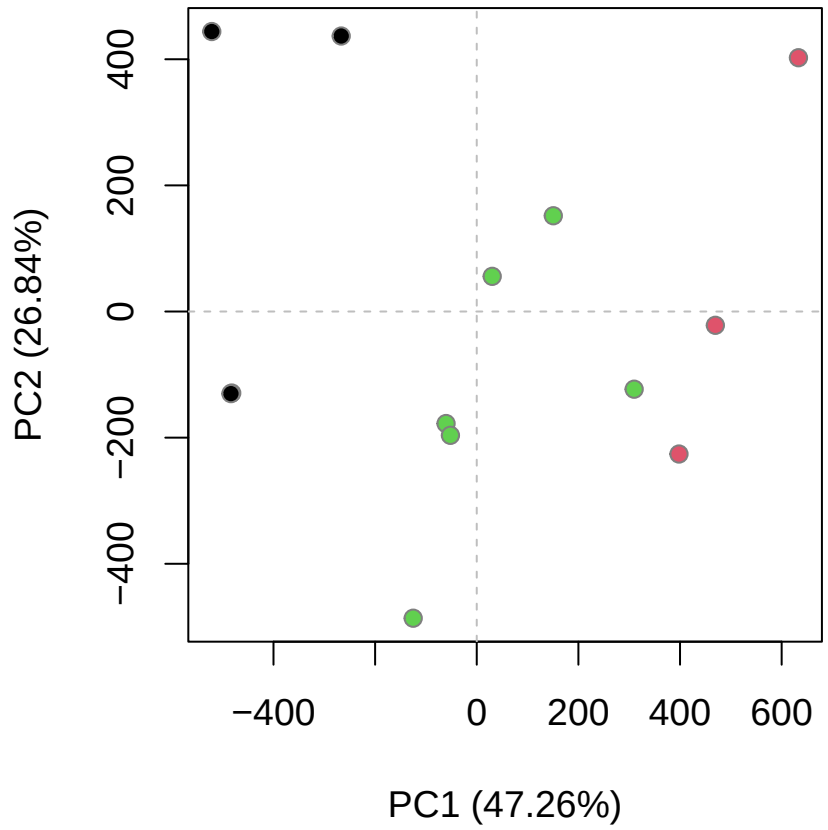
```
pc.xray<-pca(pdbs)
plot(pc.xray)
```




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

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

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

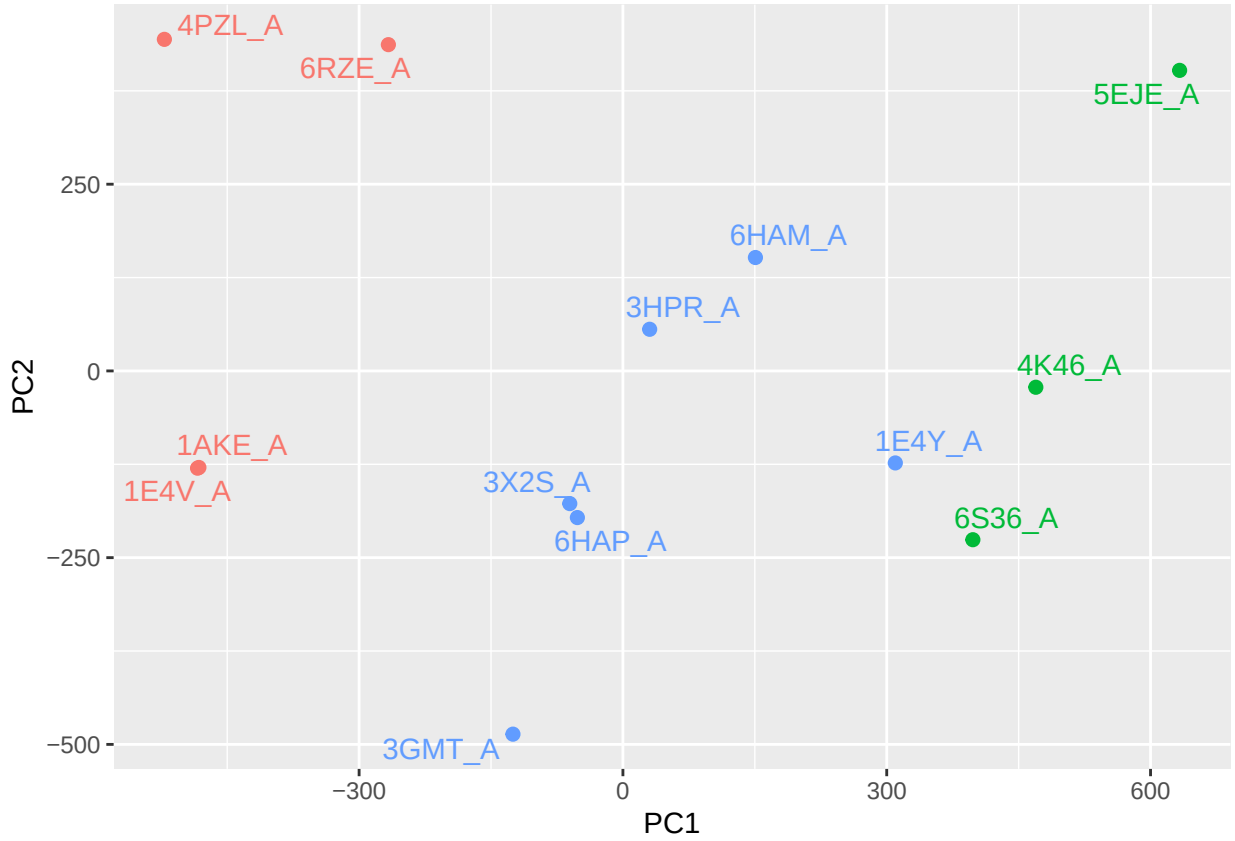


Optional further visualization

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

```
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
```



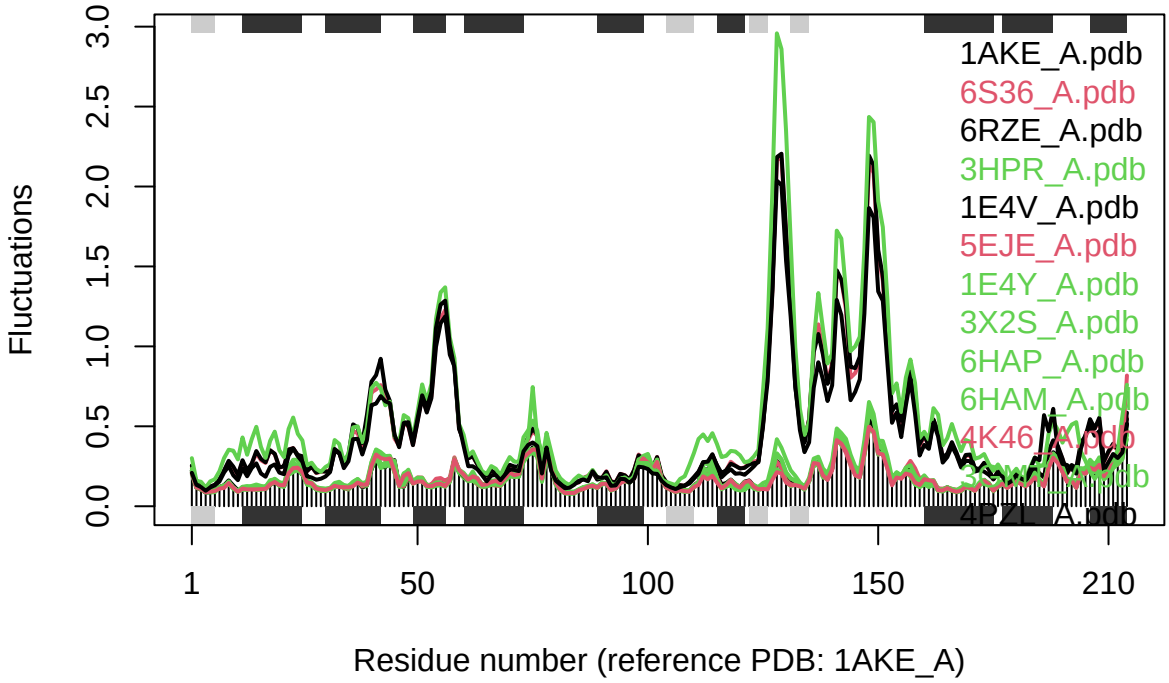
Normal mode analysis [optional]

```
modes <-nma(pdbs)
```

```
##
## Details of Scheduled Calculation:
##   ... 13 input structures
##   ... storing 606 eigenvectors for each structure
##   ... dimension of x$U.subspace: ( 612x606x13 )
##   ... coordinate superposition prior to NM calculation
##   ... aligned eigenvectors (gap containing positions removed)
##   ... estimated memory usage of final 'eNMA' object: 36.9 Mb
##
##   |                                                                      |
```

```
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 similiar or different? Where do you think they differ most and why?

#The black and colored lines are similiar for the most part except in certain areas where the green line increased in residues 10-30 and 130-160, but the black line was higher in residues 190-210. I think this happens due to changes in the nucleotide binding (that is what the black and colored lines represent).