Class 09: Structural Bioinformatics 1

AUTHOR
Nicholas Yousefi

The RCSB Protein Data Bank (PDB)

Protein structures by X-ray crystallography dominate this database. We are skipping Q1-3 as the website was too slow for us (although I was able to do them later, when the website was working again).

```
des <- read.csv("Data Export Summary.csv", row.names=1)
head(des)
```

```
                         X.ray    NMR    EM Multiple.methods Neutron
Other
Protein (only)         150,342 12,053 8,534              188      72
32
Protein/Oligosaccharide  8,866     32 1,540                6       0
0
Protein/NA               7,911    278 2,681                6       0
0
Nucleic acid (only)      2,510  1,425    74               13       2
1
Other                      154     31     6                0       0
0
Oligosaccharide (only)      11      6     0                1       0
4
                         Total
Protein (only)         171,221
Protein/Oligosaccharide 10,444
Protein/NA              10,876
Nucleic acid (only)      4,025
Other                      191
Oligosaccharide (only)      22
```

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

$$\frac{169794 + 12835}{196779} \times 100 = 92.8\%$$

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

$$\frac{150342}{196779} \times 100 = 87.0\%$$

> 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?

There are 4703 HIV-1 structures in the PDB.

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

This experiment is not high enough resolution to show the hydrogen atoms. The hydrogen atoms are way too small to show up. We can only see the oxygen atom.

> 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

The water molecule is H_2O 308. Here, it is shown bound to the protein and ligand:

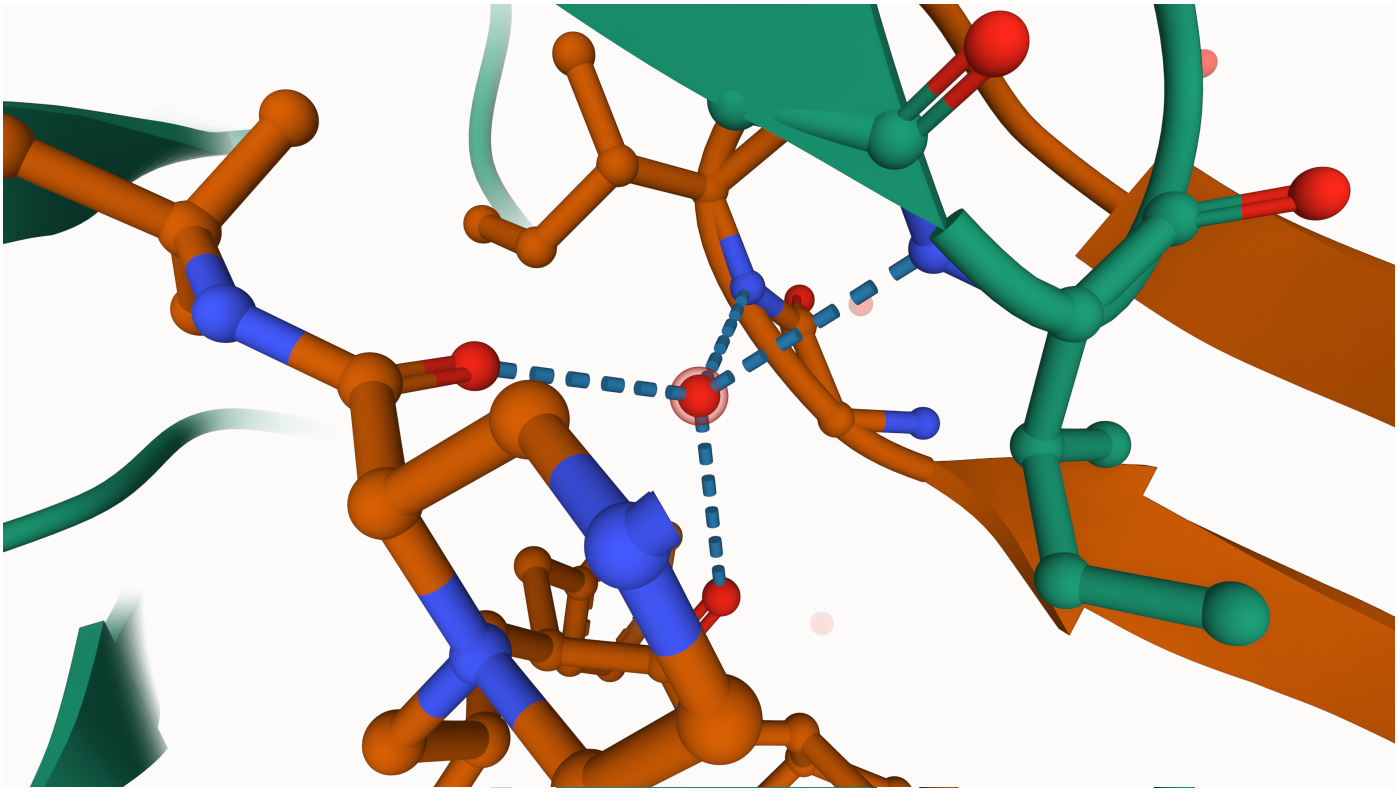


Figure 1: H_2O 308, in the binding site of 1HSG

Figure 1: H_2O 308, in the binding site of 1HSG

> 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 (we recommend "Ball & Stick" for these side-chains). Add this figure to your Quarto document.

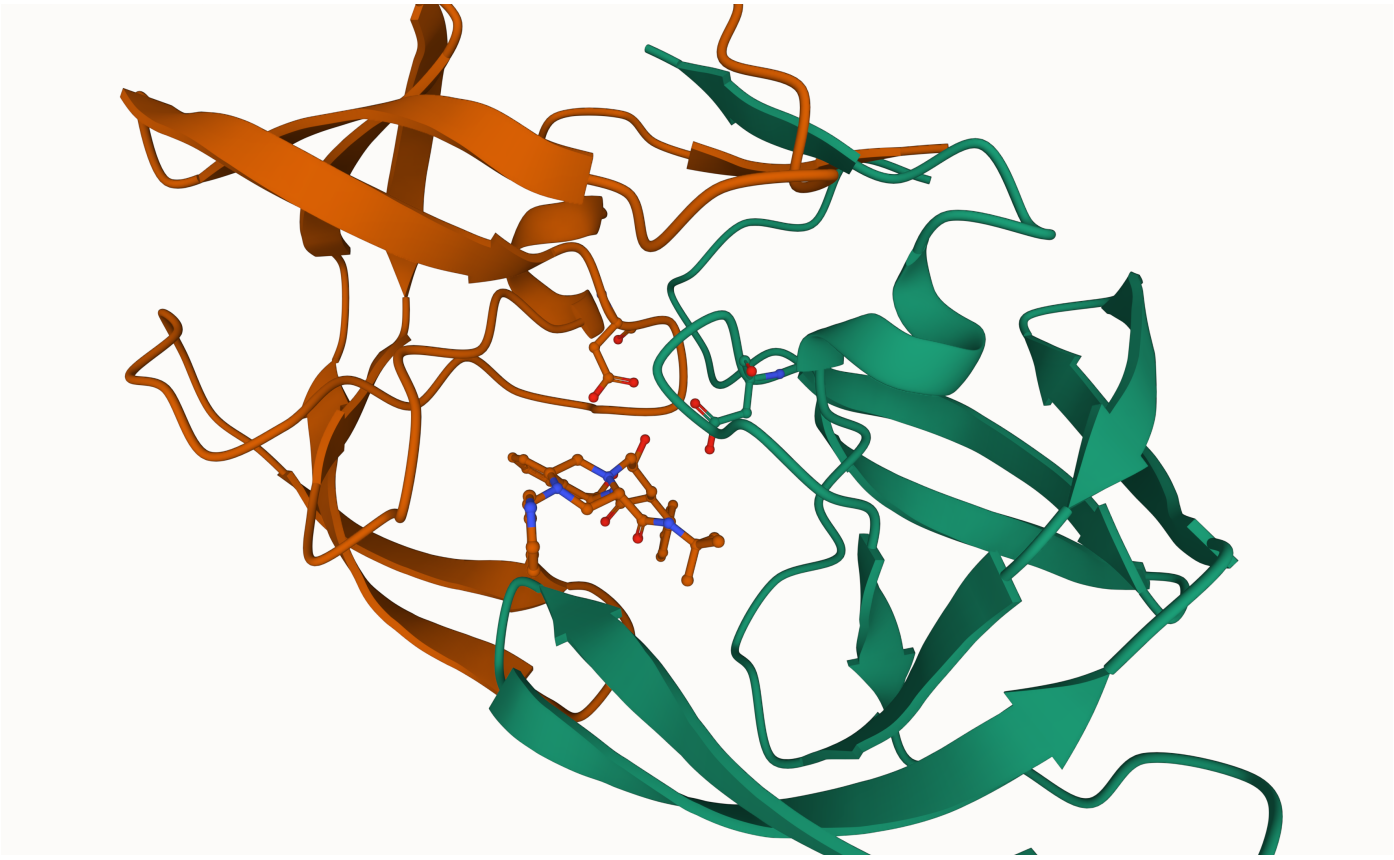


Figure 2: 1HSG protein with ASP 25 shown in each side chain.

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

Larger ligands like indinavir could have a group that hydrogen bonds where H_2O 308 normally binds when MK1 is in the molecule.

3. Introduction to Bio3D in R

Bio3D is an R package for structural bioinformatics. To use it, we need to call it up with the `library()` function (just like any package).

```
library(bio3d)
```

To read a PDB file we can use `read.pdb()`.

```
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 in this pdb object?

198 amino acid residues are in this PDB object.

> Q8: Name one of the two non-protein residues?

MK1

> Q9: How many protein chains are in this structure?

2 chains

Side note: you can figure out what attribute an object has using the `attributes()`

function:

```
attributes(pdb)
```

```
$names
[1] "atom"   "xyz"    "seqres" "helix"  "sheet"  "calpha" "remark"
"call"

$class
[1] "pdb" "sse"
```

The ATOM records of a PDB file are stored in `pdb$atom`

```
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>
```

Let's do a Normal mode analysis on a new PDB structure: Adenylate Kinase.

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

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

```
adk
```

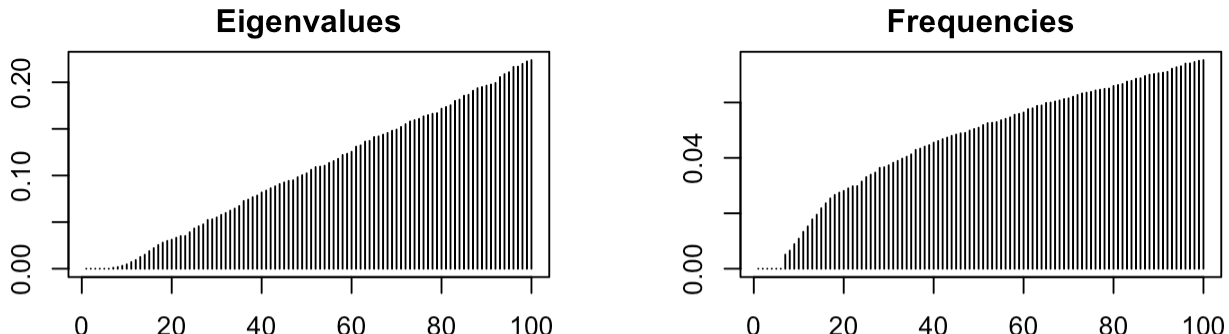
```
 Call:  read.pdb(file = "6s36")

   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
```

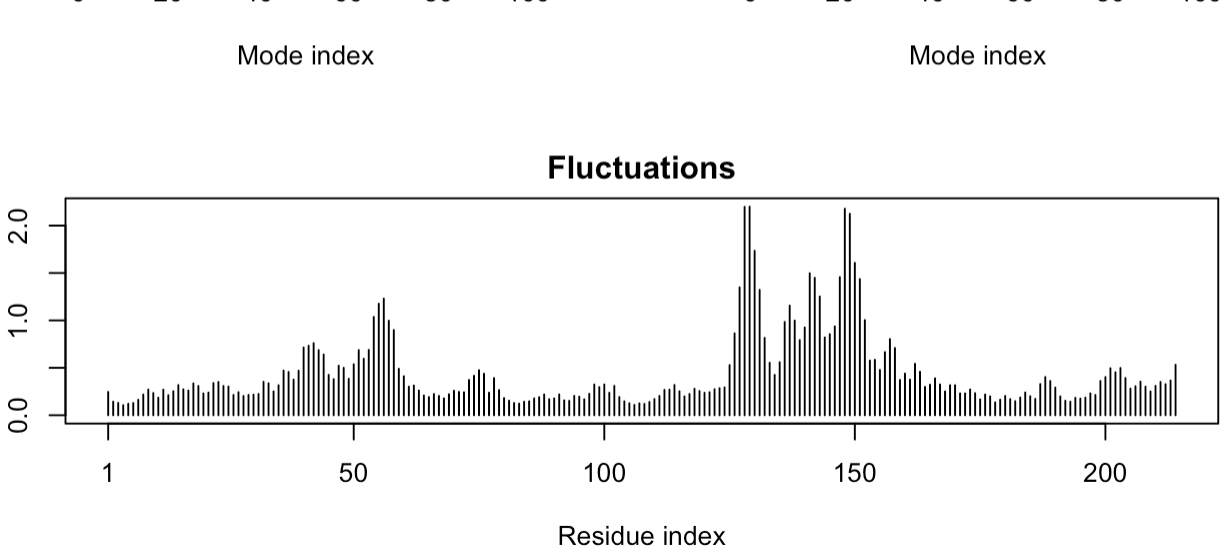
Normal Mode Analysis is used to predict protein flexibility and potential functinal motions, such as conformational changes.

```
# This is a flexibility prediction
m <- nma(adk)
```

```
 Building Hessian...        Done in 0.037 seconds.
 Diagonalizing Hessian...   Done in 0.407 seconds.
```

```
plot(m)
```





To view an animation showing the predicted protein motions, we can create a "trajectory" using `mktrj()`.

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

This file can be loaded into Mol* and the animation can be played.

Comparative Analysis of Adenylate kinase (ADK)

> 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
```

We will start our analysis with a single PDB id (code form the PDB database): 1AKE

First, we get its primary sequence:

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

214 amino acids

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

```
#hits <- plot(b)
# List out some 'top hits'
#head(hits$pdb.id)
```

Alternatively, if BLAST doesn't work, you can use these:

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

Download all these PDB files from the online database...

```
# Download related PDB files
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/
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/
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/
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%
  |
  |=====
|   8%
  |
  |===========
|  15%
  |
```

```
  |================
|  23%
  |
  |=====================
|  31%
  |
  |=========================
|  38%
  |
  |===============================
|  46%
  |
  |======================================
|  54%
  |
  |============================================
|  62%
  |
  |=================================================
|  69%
  |
  |=========================================================
|  77%
  |
  |=================================================================
|  85%
  |
  |=========================================================================
|  92%
  |

|=============================================================================|
100%
```

We downloaded a bunch of structures from the PDB database. Now, let's align them.

```
pdbs <- pdbaln(files, fit=T, 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
   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
```

```
pdbs
```

```
                                1       .       .       .
40
[Truncated_Name:1]1AKE_A.pdb                MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
```

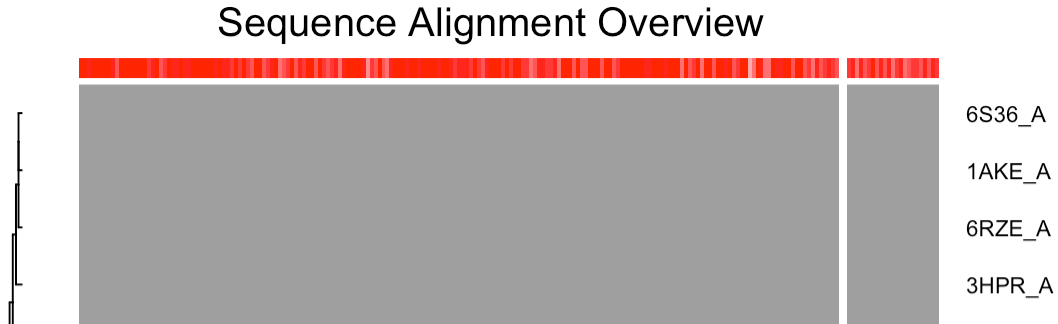
```
[Truncated_Name:1]1AKE_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:2]6S36_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:3]6RZE_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:4]3HPR_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:5]1E4V_A.pdb    ----------MRIILLGAPVAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:6]5EJE_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:7]1E4Y_A.pdb    ----------MRIILLGALVAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:8]3X2S_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:9]6HAP_A.pdb    ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:10]6HAM_A.pdb   ----------MRIILLGAPGAGKGTQAQFIMEKYGIPQIS
[Truncated_Name:11]4K46_A.pdb   ----------MRIILLGAPGAGKGTQAQFIMAKFGIPQIS
[Truncated_Name:12]3GMT_A.pdb   ----------MRLILLGAPGAGKGTQANFIKEKFGIPQIS
[Truncated_Name:13]4PZL_A.pdb   TENLYFQSNAMRIILLGAPGAGKGTQAKIIEQKYNIAHIS
                                          **^*****  *******  *  *^ *  **
                                1         .         .         .
40

                                41        .         .         .
80
[Truncated_Name:1]1AKE_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:2]6S36_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:3]6RZE_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:4]3HPR_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:5]1E4V_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:6]5EJE_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDACKLVTDELVIALVKE
[Truncated_Name:7]1E4Y_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVKE
[Truncated_Name:8]3X2S_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDCGKLVTDELVIALVKE
[Truncated_Name:9]6HAP_A.pdb    TGDMLRAAVKSGSELGKQAKDIMDAGKLVTDELVIALVRE
[Truncated_Name:10]6HAM_A.pdb   TGDMLRAAIKSGSELGKQAKDIMDAGKLVTDEIIIALVKE
[Truncated_Name:11]4K46_A.pdb   TGDMLRAAIKAGTELGKQAKSVIDAGQLVSDDIILGLVKE
[Truncated_Name:12]3GMT_A.pdb   TGDMLRAAVKAGTPLGVEAKTYMDEGKLVPDSLIIGLVKE
[Truncated_Name:13]4PZL_A.pdb   TGDMIRETIKSGSALGQELKKVLDAGELVSDEFIIKIVKD
                                ****^*  ^* *^ **   *  ^*   ** *  ^^ ^*^^
                                41        .         .         .
80

                                81        .         .         .
120
[Truncated_Name:1]1AKE_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:2]6S36_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:3]6RZE_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:4]3HPR_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:5]1E4V_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:6]5EJE_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:7]1E4Y_A.pdb    RIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
```

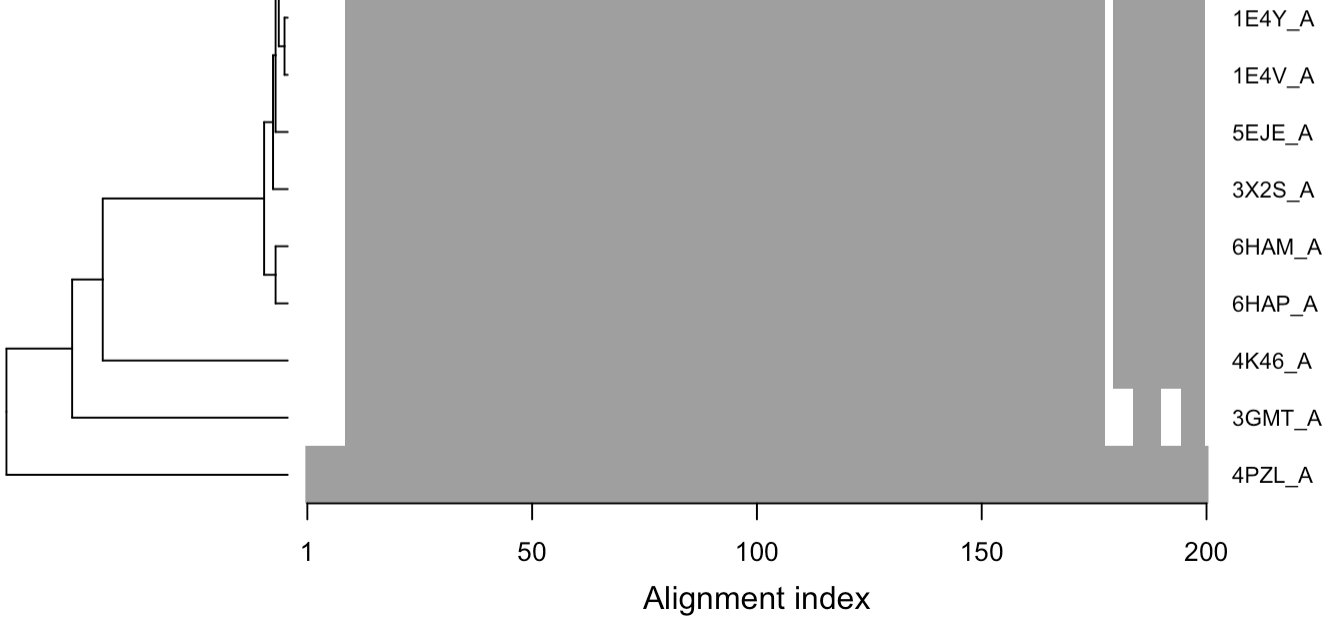
```
[Truncated_Name:8]3X2S_A.pdb    RIAQEDSRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:9]6HAP_A.pdb    RICQEDSRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:10]6HAM_A.pdb   RICQEDSRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFD
[Truncated_Name:11]4K46_A.pdb   RIAQDDCAKGFLLDGFPRTIPQADGLKEVGVVVDYVIEFD
[Truncated_Name:12]3GMT_A.pdb   RLKEADCANGYLFDGFPRTIAQADAMKEAGVAIDYVLEID
[Truncated_Name:13]4PZL_A.pdb   RISKNDCNNGFLLDGVPRTIPQAQELDKLGVNIDYIVEVD
                                *^   *   *^* ** **** **  ^   *^ ^**^^* *
                               81        .         .         .
120

                              121        .         .         .
160
[Truncated_Name:1]1AKE_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:2]6S36_A.pdb    VPDELIVDKIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:3]6RZE_A.pdb    VPDELIVDAIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:4]3HPR_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDGTG
[Truncated_Name:5]1E4V_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:6]5EJE_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:7]1E4Y_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:8]3X2S_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:9]6HAP_A.pdb    VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:10]6HAM_A.pdb   VPDELIVDRIVGRRVHAPSGRVYHVKFNPPKVEGKDDVTG
[Truncated_Name:11]4K46_A.pdb   VADSVIVERMAGRRAHLASGRTYHNVYNPPKVEGKDDVTG
[Truncated_Name:12]3GMT_A.pdb   VPFSEIIERMSGRRTHPASGRTYHVKFNPPKVEGKDDVTG
[Truncated_Name:13]4PZL_A.pdb   VADNLLIERITGRRIHPASGRTYHTKFNPPKVADKDDVTG
                                *    ^^^ ^ *** *  *** **  ^*****  *** **
                              121        .         .         .
160

                              161        .         .         .
200
[Truncated_Name:1]1AKE_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:2]6S36_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:3]6RZE_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:4]3HPR_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:5]1E4V_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:6]5EJE_A.pdb    EELTTRKDDQEECVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:7]1E4Y_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:8]3X2S_A.pdb    EELTTRKDDQEETVRKRLCEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:9]6HAP_A.pdb    EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:10]6HAM_A.pdb   EELTTRKDDQEETVRKRLVEYHQMTAPLIGYYSKEAEAGN
[Truncated_Name:11]4K46_A.pdb   EDLVIREDDKEETVLARLGVYHNQTAPLIAYYGKEAEAGN
[Truncated_Name:12]3GMT_A.pdb   EPLVQRDDDKEETVKKRLDVYEAQTKPLITYYGDWARRGA
[Truncated_Name:13]4PZL_A.pdb   EPLITRTDDNEDTVKQRLSVYHAQTAKLIDFYRNFSSTNT
                                * *  * ** *^ *  **  *   *  ** ^*
```

```
                                161       .       .       .
200

                                201       .       .     227
[Truncated_Name:1]1AKE_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:2]6S36_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:3]6RZE_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:4]3HPR_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:5]1E4V_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:6]5EJE_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:7]1E4Y_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:8]3X2S_A.pdb    T--KYAKVDGTKPVAEVRADLEKILG-
[Truncated_Name:9]6HAP_A.pdb    T--KYAKVDGTKPVCEVRADLEKILG-
[Truncated_Name:10]6HAM_A.pdb   T--KYAKVDGTKPVCEVRADLEKILG-
[Truncated_Name:11]4K46_A.pdb   T--QYLKFDGTKAVAEVSAELEKALA-
[Truncated_Name:12]3GMT_A.pdb   E-------NGLKAPA-----YRKISG-
[Truncated_Name:13]4PZL_A.pdb   KIPKYIKINGDQAVEKVSQDIFDQLNK
                                          *
                                201       .       .     227

Call:
  pdbaln(files = files, fit = T, exefile = "msa")

Class:
  pdbs, fasta

Alignment dimensions:
  13 sequence rows; 227 position columns (204 non-gap, 23 gap)

+ attr: xyz, resno, b, chain, id, ali, resid, sse, call
```

We will now plot the sequence alignment. Gray areas are aligned residues and white areas are areas that are not aligned. The red bar shows sequence conservation.

```
ids <- basename.pdb(pdbs$id) # create a vector of the PDB codes to use f
plot(pdbs, labels=ids) # plot sequence alignment
```





We can use `pdb.annotate()` to match each structure to its source species.

```
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
```

1L4T_A 1L4T 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, active site lid (ADK_lid)
6S36_A	1.60	<NA>	Adenylate kinase, active site lid (ADK_lid)
6RZE_A	1.69	<NA>	Adenylate kinase, active site lid (ADK_lid)
3HPR_A	2.00	<NA>	Adenylate kinase, active site lid (ADK_lid)
1E4V_A	1.85	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)
5EJE_A	1.90	<NA>	Adenylate kinase, active site lid (ADK_lid)
1E4Y_A	1.85	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)
3X2S_A	2.80	<NA>	Adenylate kinase, active site lid (ADK_lid)
6HAP_A	2.70	<NA>	Adenylate kinase, active site lid (ADK_lid)
6HAM_A	2.55	<NA>	Adenylate kinase, active site lid (ADK_lid)
4K46_A	2.01	<NA>	Adenylate kinase, active site lid (ADK_lid)
3GMT_A	2.10	<NA>	Adenylate kinase, active site lid (ADK_lid)
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

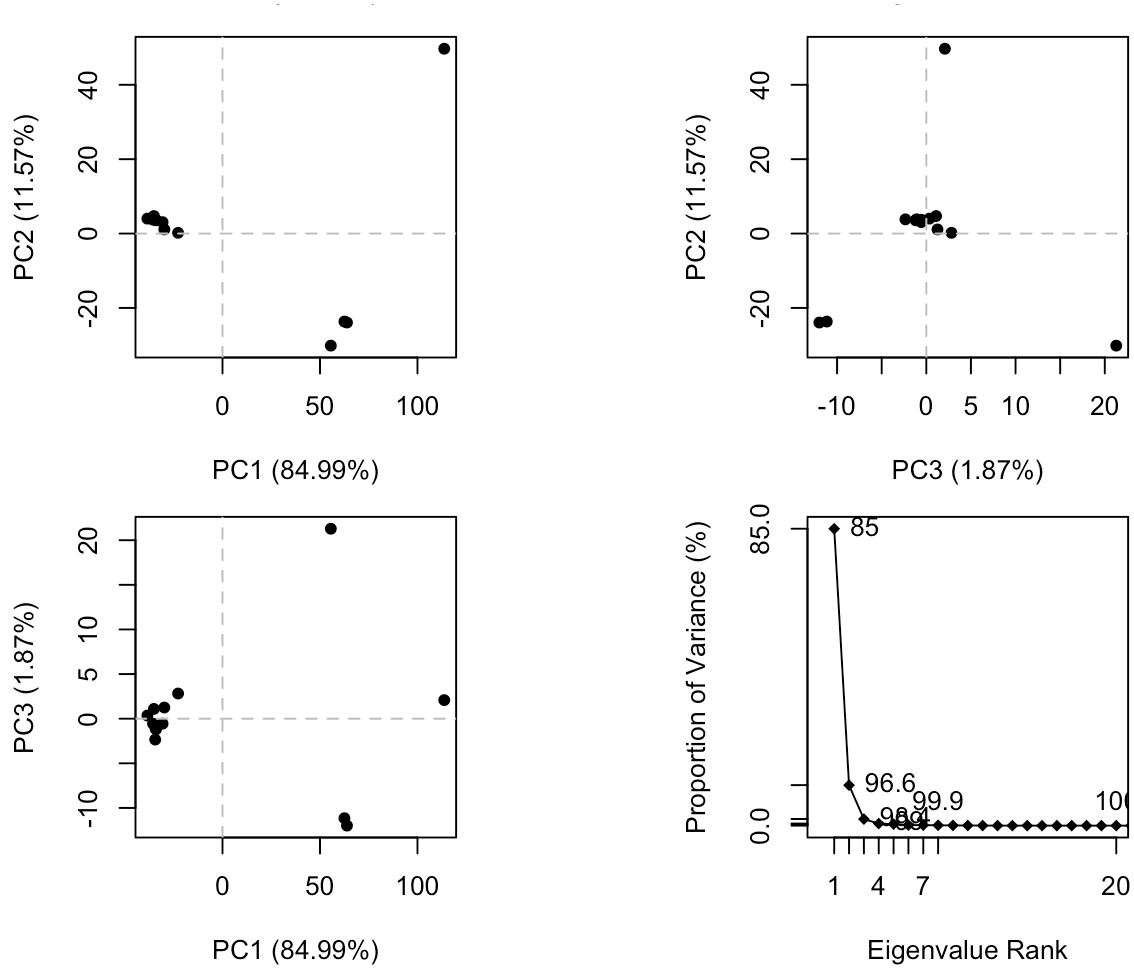
structureTitle
1AKE_A STRUCTURE OF THE COMPLEX BETWEEN ADENYLATE KINASE FROM ESCHERICHIA COLI AND THE INHIBITOR AP5A REFINED AT 1.9 ANGSTROMS RESOLUTION: A MODEL FOR A CATALYTIC TRANSITION STATE
6S36_A
Crystal structure of E. coli Adenylate kinase R119K mutant
6RZE_A
Crystal structure of E. coli Adenylate kinase R119A mutant
3HPR_A
Crystal structure of V148G adenylate kinase from E. coli, in complex with Ap5A
1E4V_A
Mutant G10V of adenylate kinase from E. coli, modified in the Gly-loop
5EJE_A
Crystal structure of E. coli Adenylate kinase G56C/T163C double mutant in complex with Ap5a
1E4Y_A
Mutant P9L of adenylate kinase from E. coli, modified in the Gly-loop
3X2S_A
Crystal structure of pyrene-conjugated adenylate kinase
6HAP_A
Adenylate kinase
6HAM_A
Adenylate kinase
4K46_A
Crystal Structure of Adenylate Kinase from Photobacterium profundum
3GMT_A
Crystal structure of adenylate kinase from burkholderia pseudomallei
4PZL_A
The crystal structure of adenylate kinase from Francisella tularensis subsp. tularensis SCHU S4

citation rObserved rFree
1AKE_A Muller, C.W., et al. J Mol Biol (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
```

Jump to PCA

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

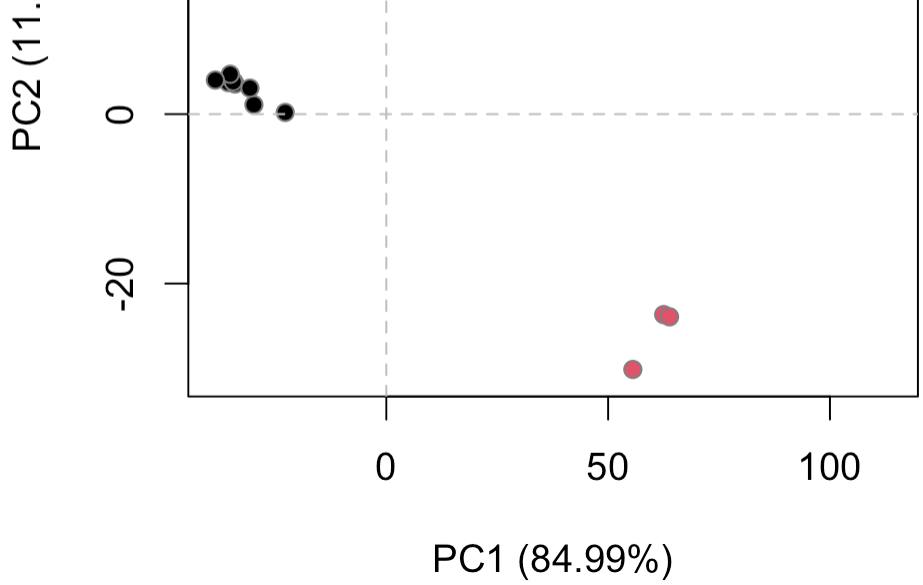


We can use the `rmsd()` function to calculate the pairwise RMSD (root mean square deviation) values. This lets us do clustering analysis based on the pairwise structural deviation.

```
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)
```



Each dot in this plot represents a single PDB structure.

Further Visualization

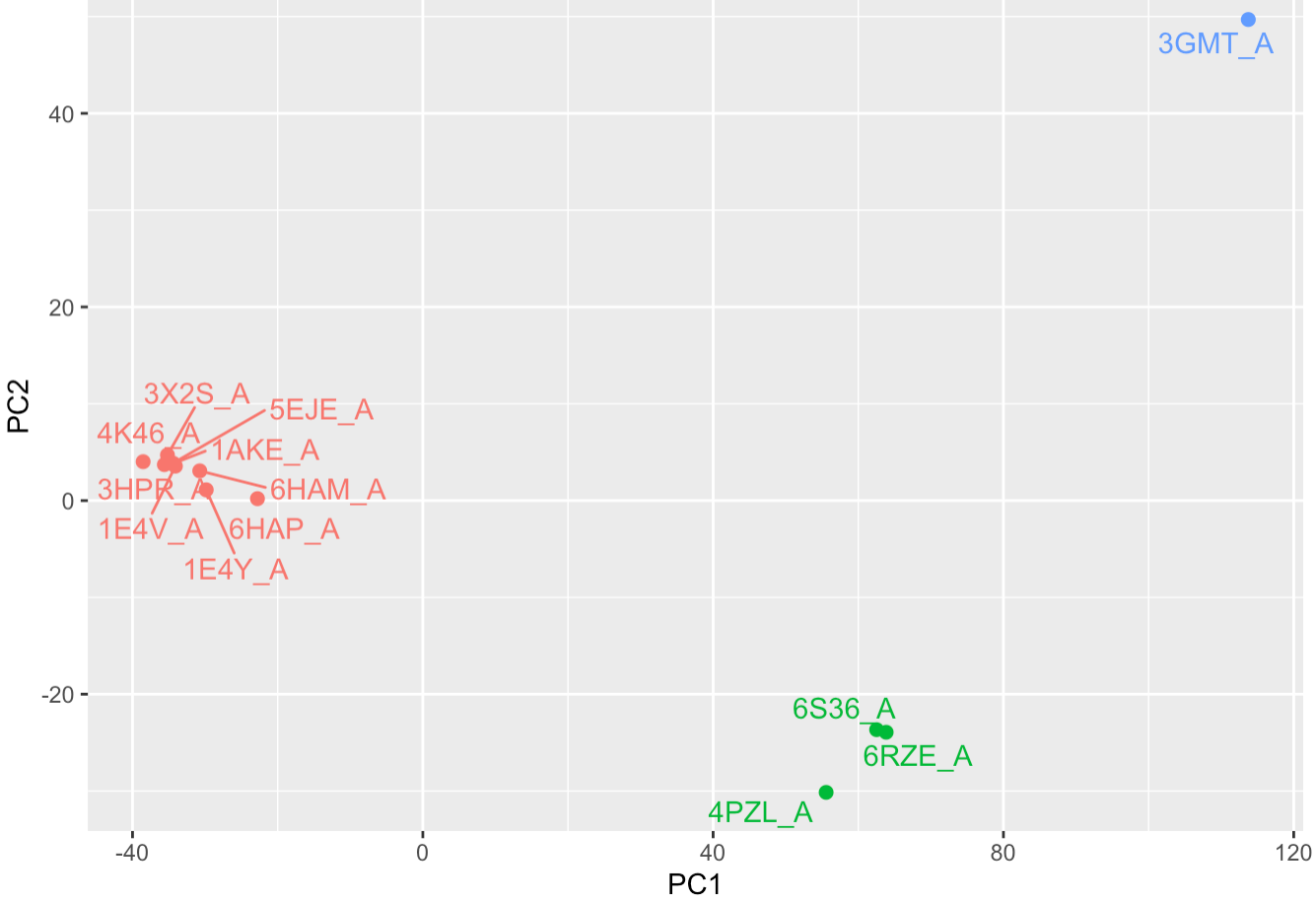
We can visualize along PC1 to see major structural variations in the protein:

```
# Visualize first principal component
pc1 <- mktrj(pc.xray, pc=1, file="pc_1.pdb")
```

We can open this file in Mole* to see the animation.

We can also plot our PCA results with ggplot:

```
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
```


3GMT_A
40
20
0
-20
PC2
3X2S_A
5EJE_A
4K46_A
1AKE_A
3HPR_A
6HAM_A
1E4V_A
6HAP_A
1E4Y_A
6S36_A
6RZE_A
4PZL_A
-40
0
40
80
120
PC1