Lab 9 Structural Bioinformatics

AUTHOR
Tiffany Chin 15700705

Structural Bioinformatics Pt. 1

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

Read CSV file from PDB site and load. This dataset has some column titles as characters, which will be an issue when trying to perform math functions. How do we fix this?

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

```
                          X.ray     EM    NMR Multiple.methods Neutron Other
Protein (only)          167,317 15,698 12,534              208      77    32
Protein/Oligosaccharide   9,645  2,639     34                8       2     0
Protein/NA                8,735  4,718    286                7       0     0
Nucleic acid (only)       2,869    138  1,507               14       3     1
Other                       170     10     33                0       0     0
Oligosaccharide (only)       11      0      6                1       0     4
                          Total
Protein (only)          195,866
Protein/Oligosaccharide  12,328
Protein/NA               13,746
Nucleic acid (only)       4,532
Other                       213
Oligosaccharide (only)       22
```

```r
x <- pdbstats$X.ray
x
```

```
[1] "167,317" "9,645"   "8,735"   "2,869"   "170"     "11"
```

```r
#comma will be an issue when converting to as.numeric
#can substitute comma for nothing using gsub
#sub only replaces for the first iterance while gsub will work multiple times
x <- gsub(",", "", x)
as.numeric(x)
```

```
[1] 167317   9645   8735   2869    170     11
```

Can now set the above as a function we can call back. We can use the apply function to apply this function for all columns we’re interested in.

```r
convert_comma_numbers <- function(x) {
  #remove commas
  x <- gsub(",", "", x)
  #convert to numeric
  x <- as.numeric(x)
  return(x)
}
```

First, apply our function to the dataset. We have removed the column title for the first column, so it should not return NA, since the rest of the columns are all numeric values.

```
pdb <- apply(pdbstats, c(1, 2), convert_comma_numbers)
pdb
```

```
                          X.ray    EM   NMR Multiple.methods Neutron Other
Protein (only)           167317 15698 12534              208      77    32
Protein/Oligosaccharide    9645  2639    34                8       2     0
Protein/NA                 8735  4718   286                7       0     0
Nucleic acid (only)        2869   138  1507               14       3     1
Other                       170    10    33                0       0     0
Oligosaccharide (only)       11     0     6                1       0     4
                          Total
Protein (only)           195866
Protein/Oligosaccharide   12328
Protein/NA                13746
Nucleic acid (only)        4532
Other                       213
Oligosaccharide (only)       22
```

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

```
#get sums
xray_total <- sum(pdb[,1])
xray_total
```

```
[1] 188747
```

```
em_total <- sum(pdb[,2])
em_total
```

```
[1] 23203
```

```
total <- sum(pdb[,7])
total
```

```
[1] 226707
```

```
#get percentage
#for X-Ray
xray_total / total * 100
```

```
[1] 83.25592
```

```
#for EM
em_total / total * 100
```

```
[1] 10.2348
```

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

```
protein <- sum(pdb[1:3,])
protein
```

```
[1] 443880
```

```
all_structures <- sum(pdb[,])
all_structures
```

```
[1] 453414
```

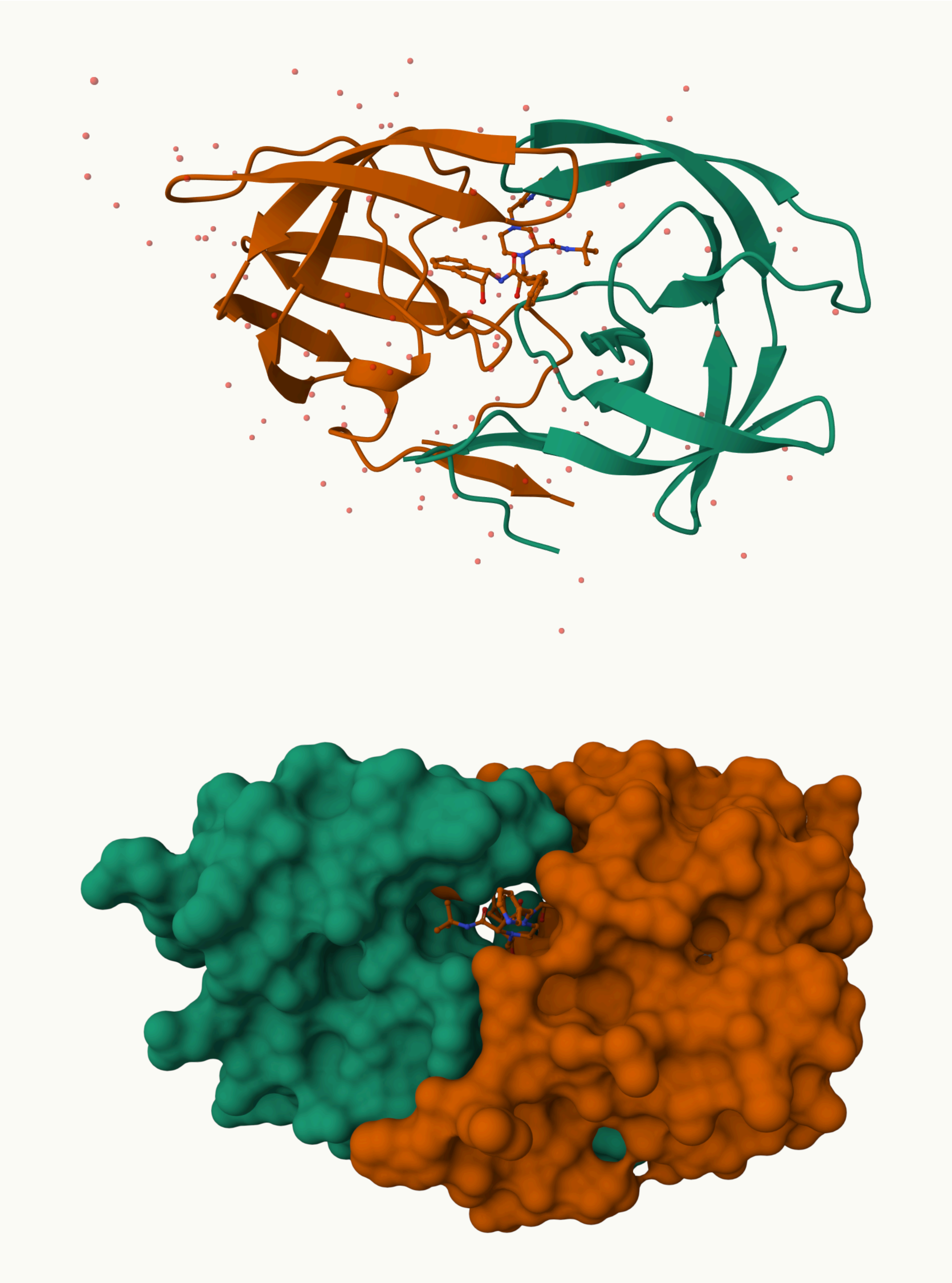
```
protein/all_structures * 100
```

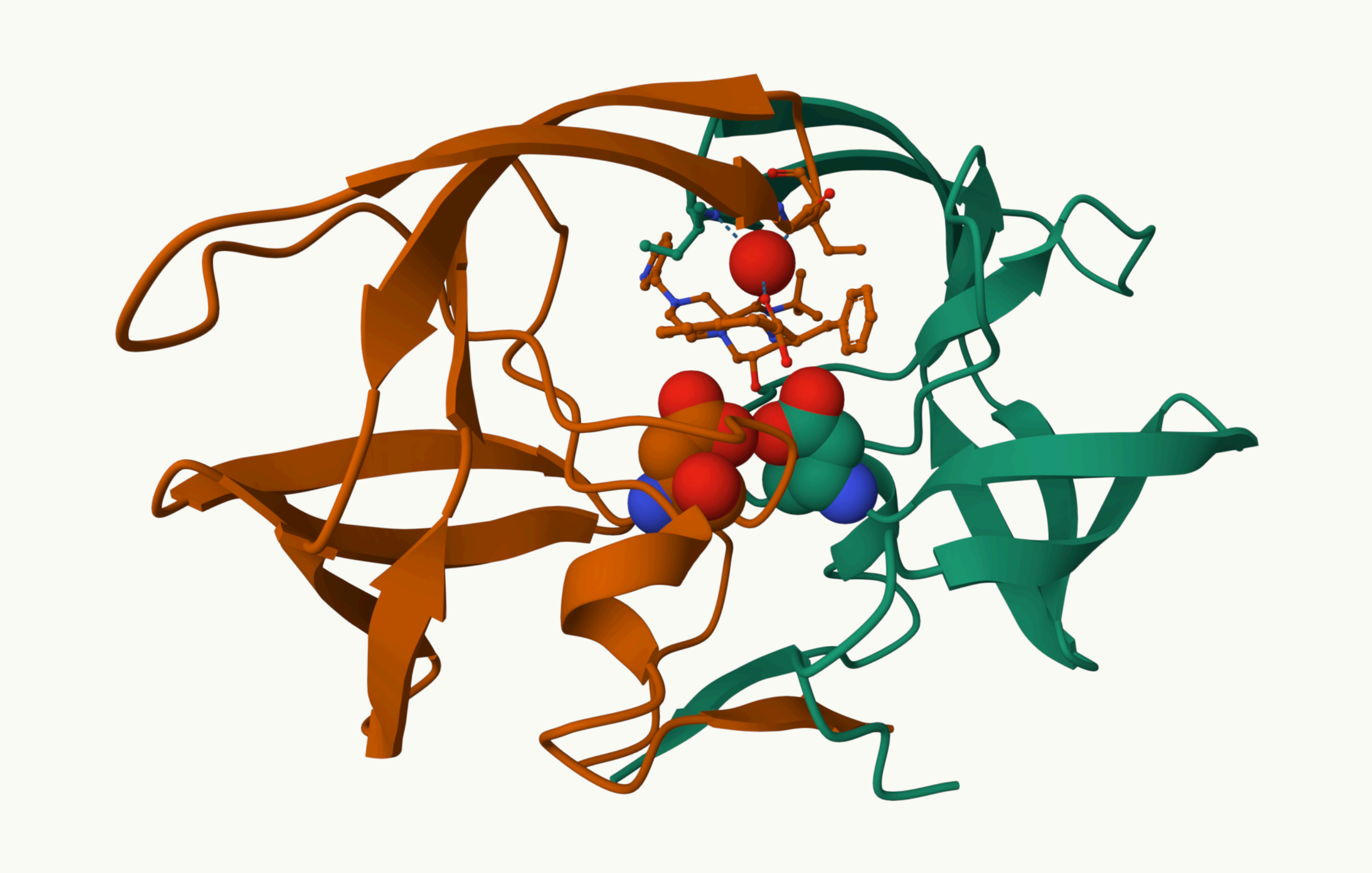
```
[1] 97.89729
```

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 4,563 structures!

Using Mol*





water

Bio3D package for structural Bioinformatics

```r
library(bio3d)

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

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

```r
pdb
```

```

 Call:  read.pdb(file = "1hsg")

   Total Models#: 1
     Total Atoms#: 1686,  XYZs#: 5058  Chains#: 2  (values: A B)

     Protein Atoms#: 1514  (residues/Calpha atoms#: 198)
     Nucleic acid Atoms#: 0  (residues/phosphate atoms#: 0)

     Non-protein/nucleic Atoms#: 172  (residues: 128)
     Non-protein/nucleic resid values: [ HOH (127), MK1 (1) ]

   Protein sequence:
      PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYD
      QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPQITLWQRPLVTIKIGGQLKE
      ALLDTGADDTVLEEMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTP
      VNIIGRNLLTQIGCTLNF

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

```
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>
```

```
pdbseq(pdb)[25]
```

```
 25 
"D" 
```

Q. How many amino acids are in this structure?

```
length(pdbseq(pdb))
```

```
[1] 198
```

Functional dynamics prediction

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

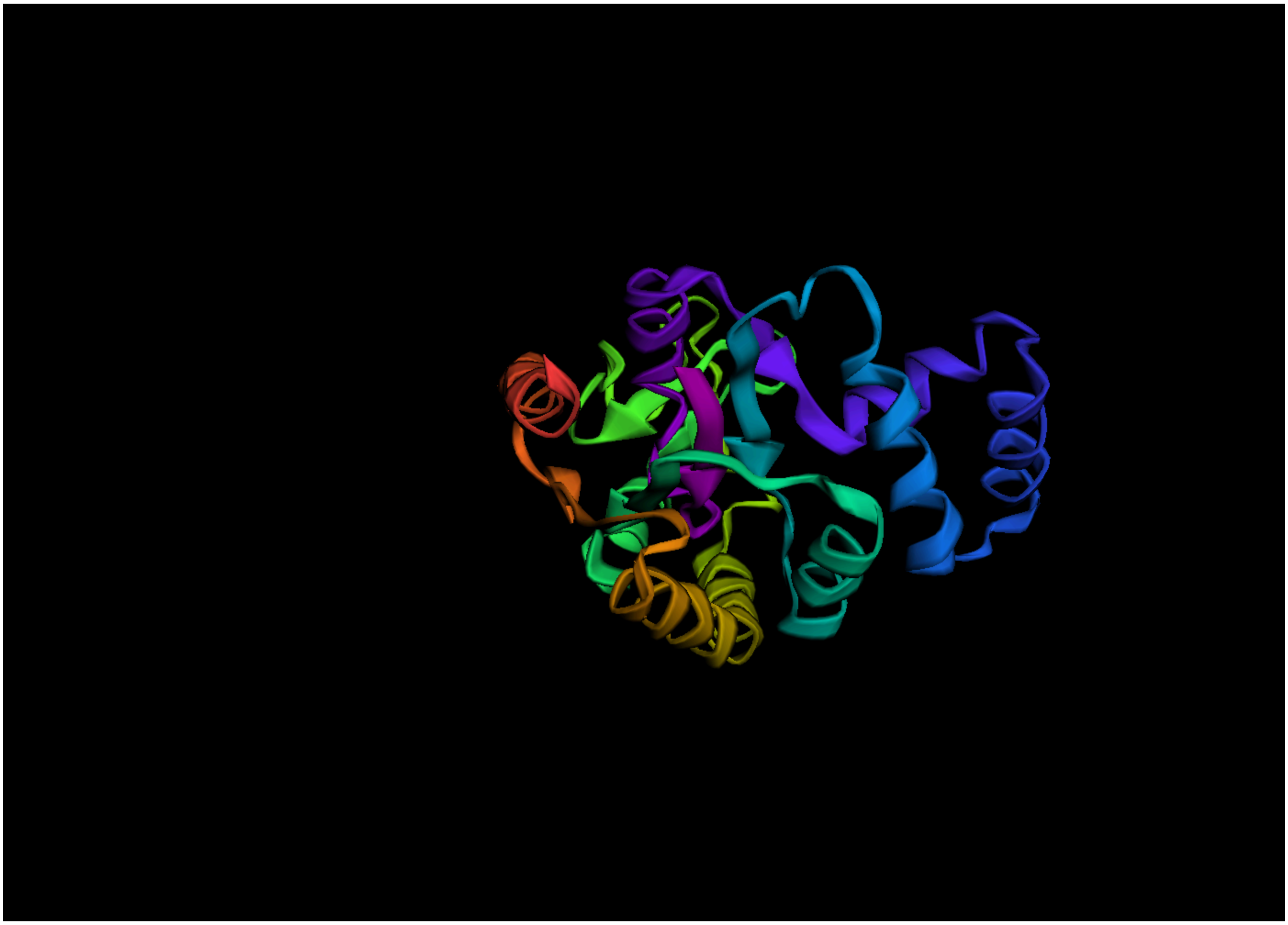
```
 Note: Accessing on-line PDB file
  PDB has ALT records, taking A only, rm.alt=TRUE
```

```r
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
```

```r
source("https://tinyurl.com/viewpdb")
library(r3dmol)
library(shiny)

view.pdb(pdb, backgroundColor = "pink")
```

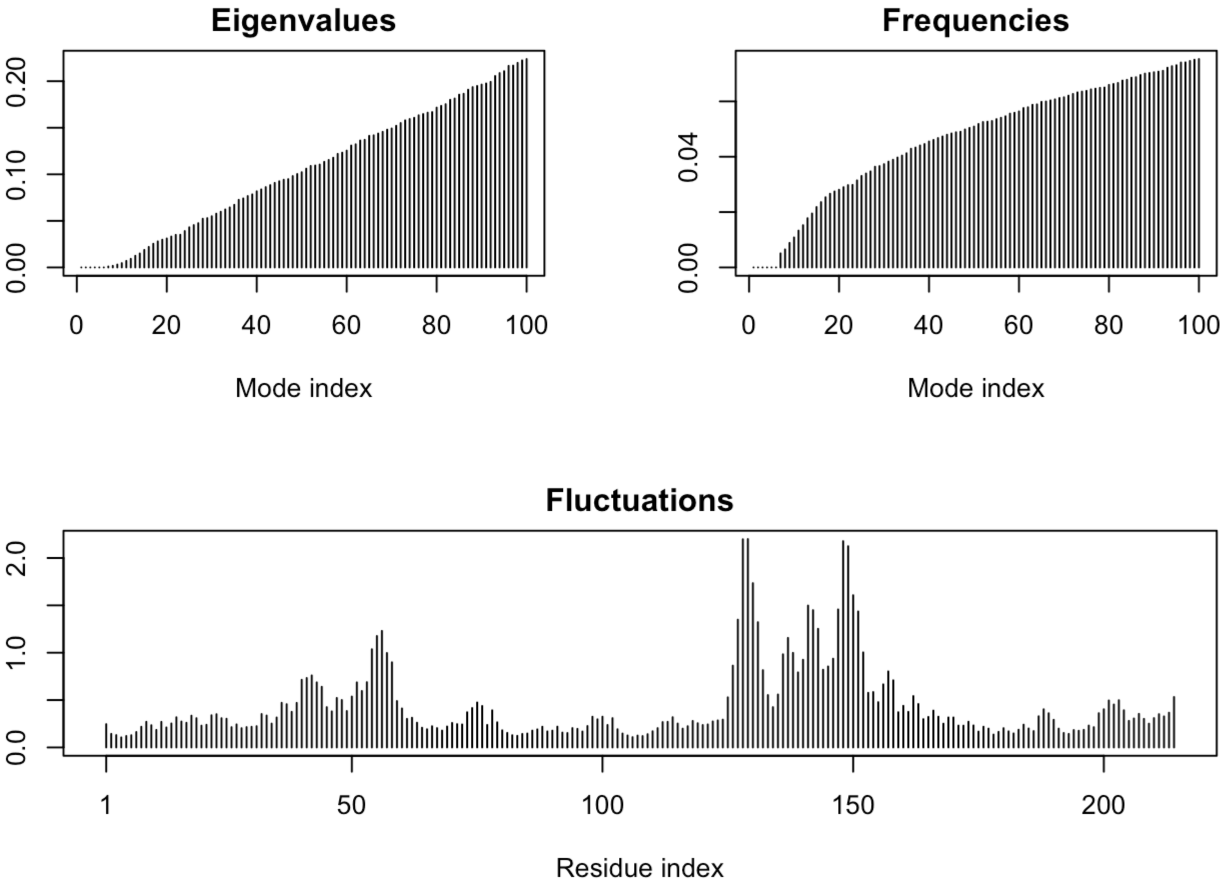
```r
view.pdb(adk)
```



```r
modes <- nma(adk)
```

```
 Building Hessian...        Done in 0.012 seconds.
 Diagonalizing Hessian...   Done in 0.263 seconds.
```

```r
plot(modes)
```



```
mktrj(modes, pdb=adk, file="adk.pdb")
```