Class 10

Kristiana Wong A16281367

##1. Introduction to the RCSB Protein Data Bank (PDB) The PDB archive is the major repository of information about the 3D structures of large biological molecules, including proteins and nucleic acids. Understanding the shape of these molecules helps to understand how they work. This knowledge can be used to help deduce a structure's role in human health and disease, and in drug development. The structures in the PDB range from tiny proteins and bits of DNA or RNA to complex molecular machines like the ribosome composed of many chains of protein and RNA.

First, let's download the CSV file

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

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

	Total
Protein (only)	186,898
Protein/Oligosaccharide	11,559
Protein/NA	12,621
Nucleic acid (only)	4,378
Other	206
Oligosaccharide (only)	22

My pdb data frame has commas in them, which may prove to be a problem

```
pdb$X.ray
```

```
[1] "161,663" "9,348"   "8,404"   "2,758"   "164"     "11"
```

```
pdb$EM
```

```
[1] "12,592" "2,167"  "3,924"  "125"    "9"      "0"
```

This will be a problem, so we need to change these from characters to numerics.

```
num <- function(sum) {
  sum(as.numeric(gsub(",", "", s)))
}

structures <- list(pdb$X.ray, pdb$EM, pdb$Total)
for (s in structures) {
  print(num(structures))
}
```

```
[1] 182348
[1] 18817
[1] 215684
```

Alternatively:

```
x.ray <- as.numeric(gsub(",", "", pdb$X.ray))
em <- as.numeric(gsub(",", "", pdb$EM))
total <- as.numeric(gsub(",", "", pdb$Total))
```

Now, we can use these numeric values in our calculations.

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

```
#From first code box
round((182348/215684 * 100), 2) #x.ray percentage
```

```
[1] 84.54
```

```
round((18817/215684 * 100), 2) #em percentage
```

```
[1] 8.72
```

```
#From second code box
round((sum(x.ray)/sum(total)) * 100, 2)
```

```
[1] 84.54
```

```
round((sum(em)/sum(total)) * 100, 2)
```

```
[1] 8.72
```

84.54% of structures in PDB are solved by X-ray. 8.72% of structures in PDB are solved by EM.

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

```
protein_only <- as.numeric(gsub(",", "", pdb[1,7]))

#Find the proportion
round((protein_only/sum(total)) * 100, 2)
```

```
[1] 86.65
```

86.65% of structures 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?

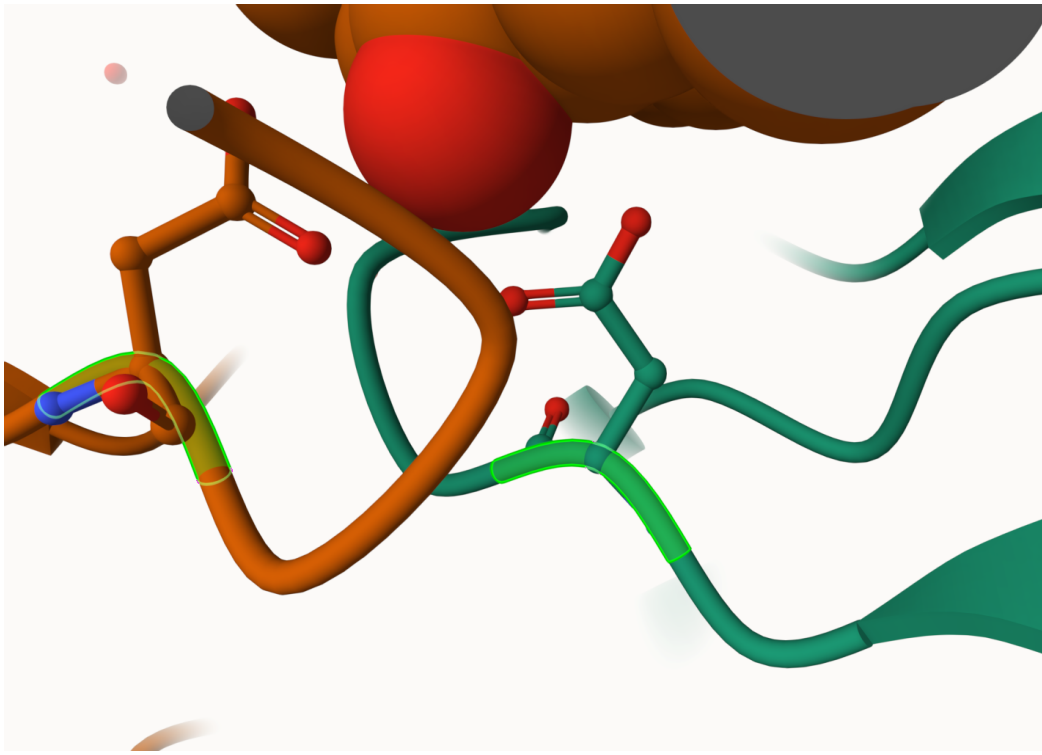
```
#4,410 searches found in the database
round(4410/sum(total) * 100, 2)
```

```
[1] 2.04
```

2.04% of HIV-1 protease structures in the current PDB.

##2. Visualizing Protein Structures We will learn the basics of Mol* (mol-star) homepage: https://molstar.org/viewer/

We will play with PDB code 1HSG



Show the Asp25 amino acids:
##Introduction to Bio3d in R

Predict the dynamics (flexibility) of an important protein:

```
library(bio3d)

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

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

```
hiv
```

```
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
```

```
head(hiv$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(hiv)

nma normal mode analysis to predict functional motions of kinase protein

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

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

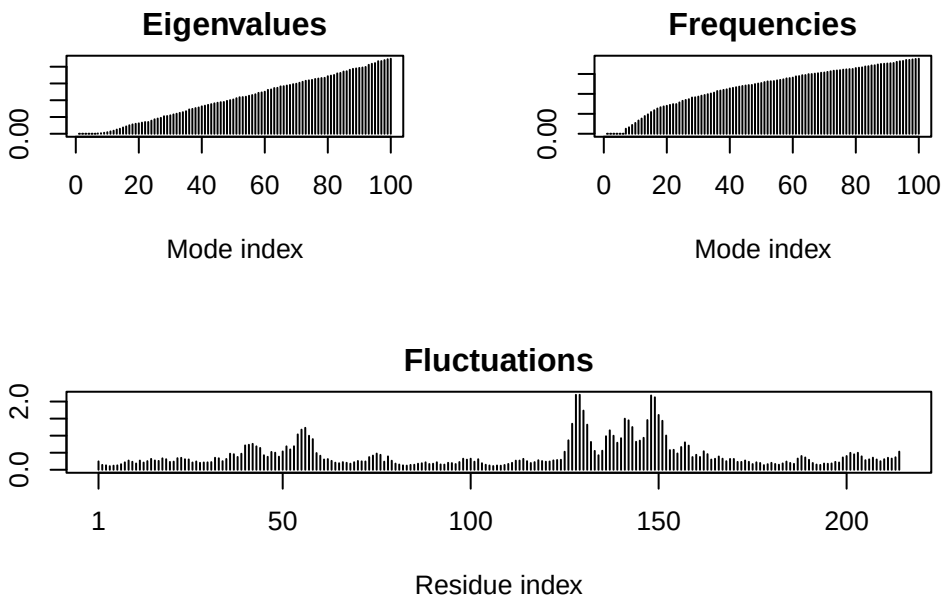
```
adk
```

```
 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
```

```
#bioinformatics calculation of motions of this protein
modes <-nma(adk)
```

```
 Building Hessian...        Done in 0.093 seconds.
 Diagonalizing Hessian...   Done in 1.15 seconds.
```

```
plot(modes)
```



Make a "movie" called a trajectory of the predicted motions:

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

Then I can open this file in Mol*…