Class 10: Structural Bioinformatics (Pt1)

James Garza (PID:A16300772)

```r
library(bio3d)
```

The PDB Database

Here we examine the size and compostion of the main datbase of biomolecular structures

Download a CSV file from the PDB site (accessible from "Analyze" > "PDB Statistics" > "by Experimental Method and Molecular Type". Move this CSV file into your RStudio project and use it to answer the following questions:

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

```
                          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
```

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

My pdbstats data frame has numbers with commas in them. This may cause us problems. Let's see:

```
pdbstats$X.ray
```

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

We found a function called `gsub()` now we can figure out how it works.

```
as.numeric(gsub(",","",pdbstats$X.ray))
```

```
[1] 161663   9348   8404   2758    164     11
```

I can turn this snippet into a function that I can use for every column in the table.

```
commasum <- function(x) {
    sum(as.numeric(gsub(",","",x)))
}

commasum(pdbstats$X.ray)
```

```
[1] 182348
```

```
totals <- apply(pdbstats, 2, commasum)
totals
```

```
           X.ray               EM              NMR Multiple.methods 
          182348            18817            14173              230 
         Neutron            Other            Total 
              79               37           215684 
```

```
round(totals/totals["Total"] * 100, 2)
```

```
           X.ray               EM              NMR Multiple.methods 
           84.54             8.72             6.57             0.11 
         Neutron            Other            Total 
            0.04             0.02           100.00 
```

8.72% of structures for EM and 84.5% of structures via X-ray

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

```
totals_pro <- commasum(pdbstats[1,7])
totals_pro
```

```
[1] 186898
```

```
round(totals_pro/totals["Total"] * 100, 2)
```

```
Total
86.65
```

86.65% of the stuctures are protein.

> 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?

```
# Ignore
```

> Q What percentage of Uniprot are we looking at relative to all the structures?

```
215684/249751891 * 100
```

```
[1] 0.08635931
```

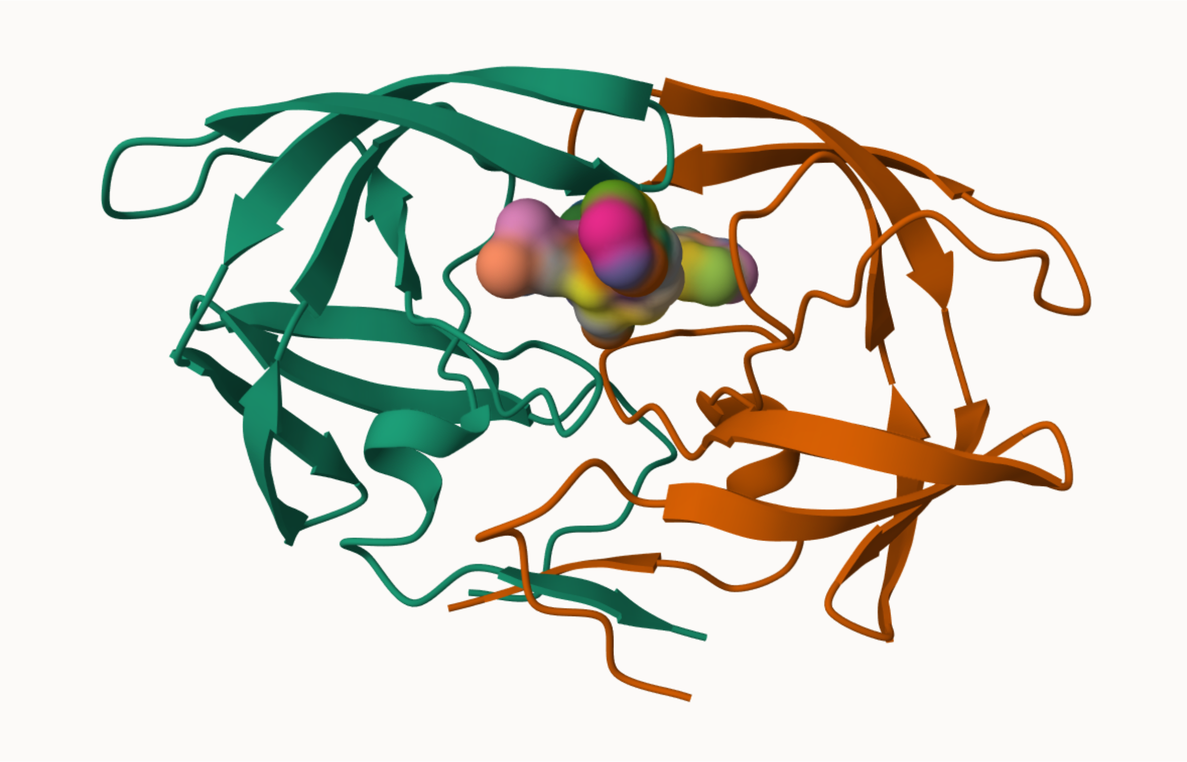
2. Visualizing Protein Structure

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

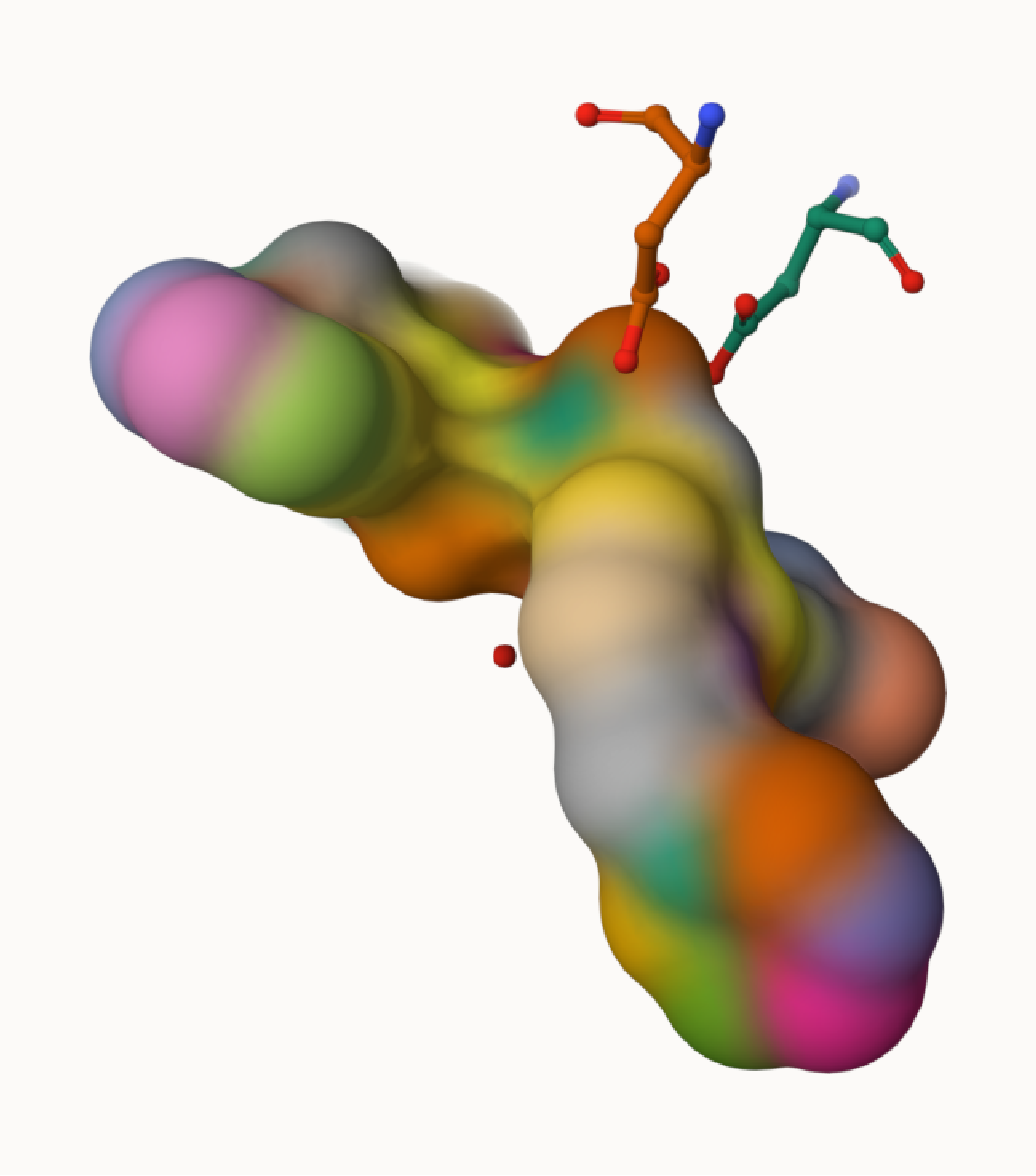
We will play with PDB code 1HSG

> Q6. Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site?

Images are below. As for getting larger molecules to enter binding sites the flaps of the enzyme are able to flap up and down to be able to allow larger substrates into the active site.



Show the ASP 25 amino acids:



> 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 because the scale is to about 2 angstroms while hydrogens are smaller than that so it does not appear because the scale of the H atoms is too small.

> 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 it is H308 and it is in the image above in the form of the singular red dot. It is conserved because it aids in the hydrogen bonding of the inhibitor in the active site.

> Q7: [Optional] As you have hopefully observed HIV protease is a homodimer (i.e. it is composed of two identical chains). With the aid of the graphic display can you identify secondary structure elements that are likely to only form in the dimer rather than the monomer?

There is only one beta sheet that forms in between the dimers. All of the other secondary structure elements form within each dimer itself.

Back to R and working with PDB structures

Predict the dynamics (felxibility) of an important protein:

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

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

```r
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)
```

```
  1   2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20 
"P" "Q" "I" "T" "L" "W" "Q" "R" "P" "L" "V" "T" "I" "K" "I" "G" "G" "Q" "L" "K" 
 21  22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40 
"E" "A" "L" "L" "D" "T" "G" "A" "D" "D" "T" "V" "L" "E" "E" "M" "S" "L" "P" "G" 
 41  42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60 
"R" "W" "K" "P" "K" "M" "I" "G" "G" "I" "G" "G" "F" "I" "K" "V" "R" "Q" "Y" "D" 
 61  62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80 
"Q" "I" "L" "I" "E" "I" "C" "G" "H" "K" "A" "I" "G" "T" "V" "L" "V" "G" "P" "T" 
 81  82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99   1 
"P" "V" "N" "I" "I" "G" "R" "N" "L" "L" "T" "Q" "I" "G" "C" "T" "L" "N" "F" "P" 
  2   3   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21 
"Q" "I" "T" "L" "W" "Q" "R" "P" "L" "V" "T" "I" "K" "I" "G" "G" "Q" "L" "K" "E" 
 22  23  24  25  26  27  28  29  30  31  32  33  34  35  36  37  38  39  40  41 
"A" "L" "L" "D" "T" "G" "A" "D" "D" "T" "V" "L" "E" "E" "M" "S" "L" "P" "G" "R" 
 42  43  44  45  46  47  48  49  50  51  52  53  54  55  56  57  58  59  60  61 
"W" "K" "P" "K" "M" "I" "G" "G" "I" "G" "G" "F" "I" "K" "V" "R" "Q" "Y" "D" "Q" 
 62  63  64  65  66  67  68  69  70  71  72  73  74  75  76  77  78  79  80  81 
```

```
"I" "L" "I" "E" "I" "C" "G" "H" "K" "A" "I" "G" "T" "V" "L" "V" "G" "P" "T" "P"
 82  83  84  85  86  87  88  89  90  91  92  93  94  95  96  97  98  99
"V" "N" "I" "I" "G" "R" "N" "L" "L" "T" "Q" "I" "G" "C" "T" "L" "N" "F"
```

Here we will do a Normal Mode Analysis (NMA) to predict funtional motions of a kinase protein.

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

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

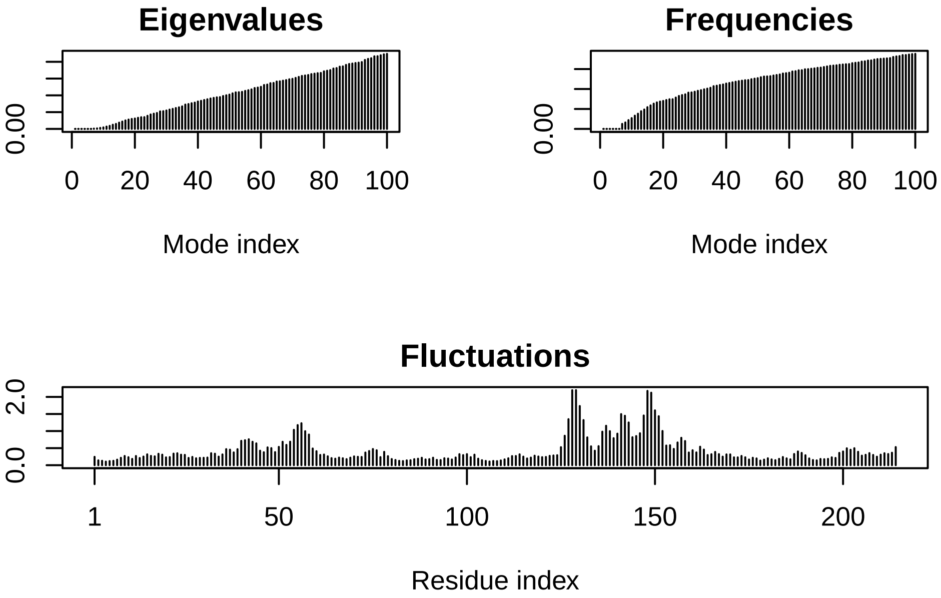
```
adk
```

```

 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
```

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

```
 Building Hessian...        Done in 0.083 seconds.
 Diagonalizing Hessian...   Done in 0.246 seconds.
```

```r
plot(modes)
```



Make a “movie” alled a trajectory of the predicted motions:

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

Then I can open this in Mol*…

> Q7: How many amino acid residues are there in this pdb object?

There are 198 amino acid residues.

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

One of the non-protein residues is water.

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

There are 2 protein chains in the structure.