Class10: Structural Bioinformatics pt1

Hsiang-Ying Lu (PID: A15608316)

The main repository of structural data is the PDB. Let's examine what it contains.

I download conposition stats from: < https://www.rcsb.org/stats/summary >

At the time of writting there are 183,201 protein structures. In UniProt, there are 251,600,768 protein sequences.

```
round(183201/251600768*100, 2)
```

```
[1] 0.07
```

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

	X.ray	EM	NMR	Multiple.methods	Neutron	Other
Protein (only)	158,844	11,759	12,296	197	73	32
Protein/Oligosaccharide	9,260	2,054	34	8	1	0
Protein/NA	8,307	3,667	284	7	0	0
Nucleic acid (only)	2,730	113	1,467	13	3	1
Other	164	9	32	0	0	0
Oligosaccharide (only)	11	0	6	1	0	4

	Total
Protein (only)	183,201
Protein/Oligosaccharide	11,357
Protein/NA	12,265
Nucleic acid (only)	4,327
Other	205
Oligosaccharide (only)	22

```
string <- c("10", "100", 1, "1,000")
as.numeric(string) +1
```

```
Warning: NAs introduced by coercion
```

```
[1]  11 101   2  NA
```

Q. Write a function to fix this non numerix table... We can use the `gsub()` function.

```
x <-string
as.numeric(gsub(",", "", x))
```

```
[1]   10  100    1 1000
```

```
rm.comma <-  function(x){
  as.numeric(gsub(",", "", x))
}

pdbstats <- apply(stats, 2, rm.comma)
```

We will add the row names from the original wee table...

```
rownames(pdbstats) <- row.names(stats)
pdbstats
```

```
                        X.ray    EM   NMR Multiple.methods Neutron Other
Protein (only)         158844 11759 12296               197      73    32
Protein/Oligosaccharide  9260  2054    34                 8       1     0
Protein/NA               8307  3667   284                 7       0     0
Nucleic acid (only)      2730   113  1467                13       3     1
Other                     164     9    32                 0       0     0
Oligosaccharide (only)     11     0     6                 1       0     4
                        Total
Protein (only)         183201
Protein/Oligosaccharide  11357
Protein/NA               12265
Nucleic acid (only)       4327
Other                      205
Oligosaccharide (only)      22
```

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

```
apply(pdbstats, 2, sum)
```

```
       X.ray               EM              NMR Multiple.methods
      179316            17602            14119              226
     Neutron            Other            Total
          77               37           211377
```

```
totals <- apply(pdbstats, 2, sum)
round(totals/totals["Total"]*100, 2)
```

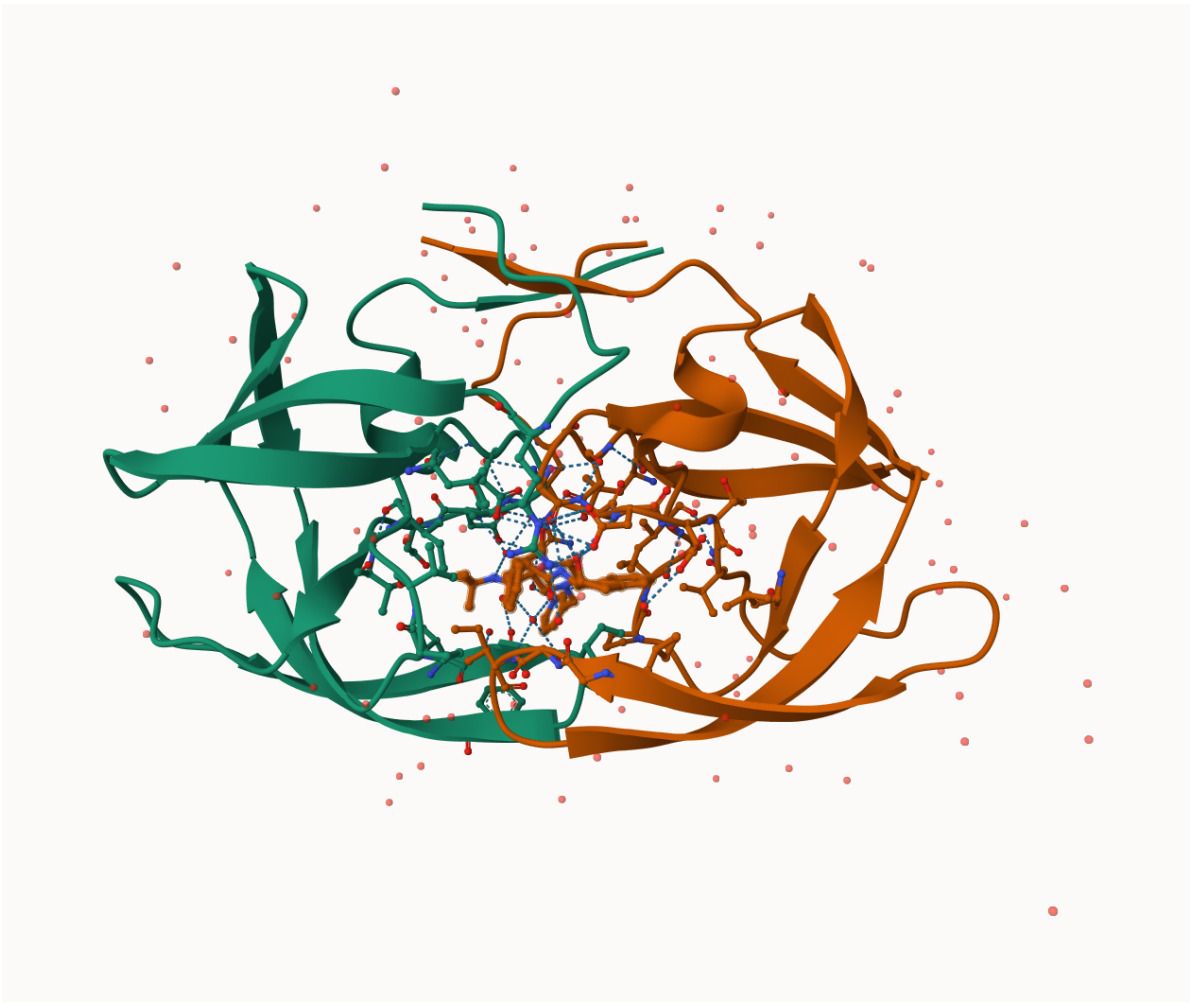
```
       X.ray               EM              NMR Multiple.methods
       84.83             8.33             6.68             0.11
     Neutron            Other            Total
        0.04             0.02           100.00
```

X-Ray: 84.83% Electron Microscopy: 8.33%

Q2-3: Let's skip these...

Using Nol* to examine HIV-Pr

Here is a rubbish pic of HIV-Pr that is not very useful yet.



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

We are only seeing the oxygen atom because water molecules are too small (0.5A). 1HSG Resolution: 2.00 A.

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's at 308

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). Add this figure to your Quarto document.

And a nicer pic colored by secondary structure with catalytic active site ASP25 shown in each chain along with MK1 drug and all important water...

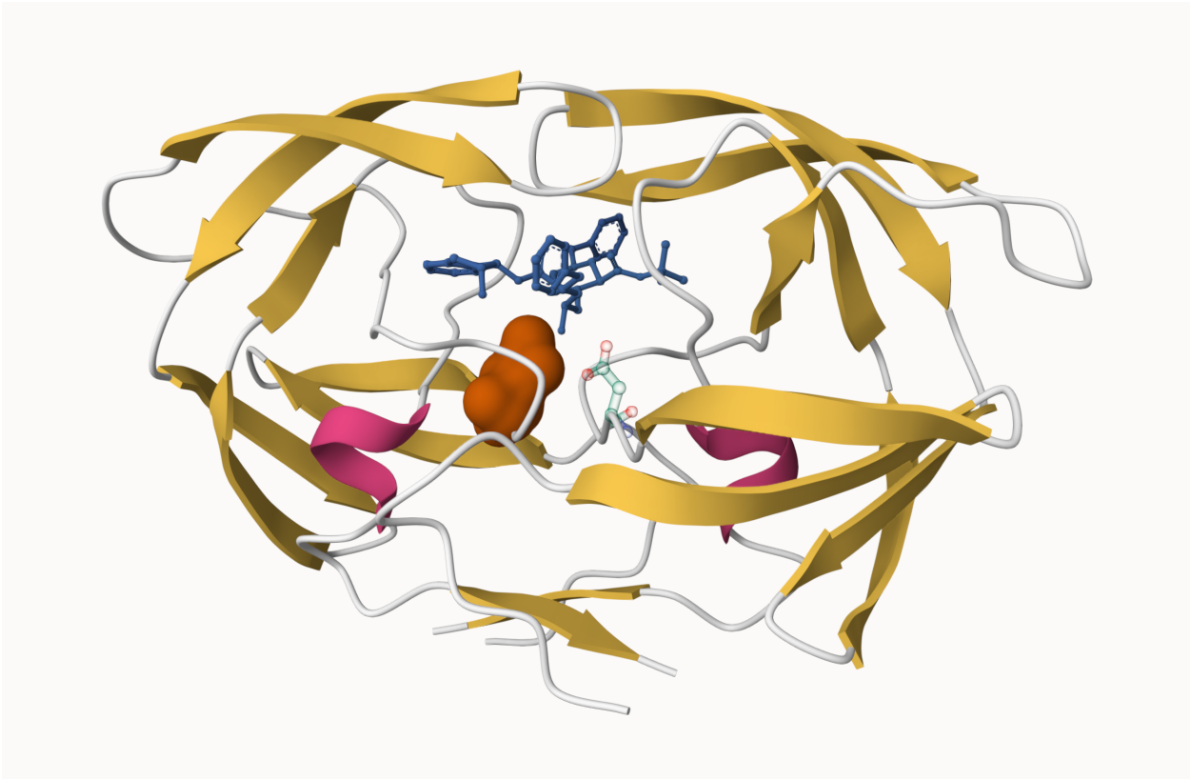


Figure 1: A lovely image

Using the bio3d package

```
library(bio3d)
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
```

```
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>
```

```
head(pdb$atom$resid)
```

```
[1] "PRO" "PRO" "PRO" "PRO" "PRO" "PRO"
```

```
aa321(pdb$atom$resid [pdb$calpha])
```

```
  [1] "P" "Q" "I" "T" "L" "W" "Q" "R" "P" "L" "V" "T" "I" "K" "I" "G" "G" "Q"
 [19] "L" "K" "E" "A" "L" "L" "D" "T" "G" "A" "D" "D" "T" "V" "L" "E" "E" "M"
 [37] "S" "L" "P" "G" "R" "W" "K" "P" "K" "M" "I" "G" "G" "I" "G" "G" "F" "I"
 [55] "K" "V" "R" "Q" "Y" "D" "Q" "I" "L" "I" "E" "I" "C" "G" "H" "K" "A" "I"
 [73] "G" "T" "V" "L" "V" "G" "P" "T" "P" "V" "N" "I" "I" "G" "R" "N" "L" "L"
 [91] "T" "Q" "I" "G" "C" "T" "L" "N" "F" "P" "Q" "I" "T" "L" "W" "Q" "R" "P"
[109] "L" "V" "T" "I" "K" "I" "G" "G" "Q" "L" "K" "E" "A" "L" "L" "D" "T" "G"
[127] "A" "D" "D" "T" "V" "L" "E" "E" "M" "S" "L" "P" "G" "R" "W" "K" "P" "K"
[145] "M" "I" "G" "G" "I" "G" "G" "F" "I" "K" "V" "R" "Q" "Y" "D" "Q" "I" "L"
[163] "I" "E" "I" "C" "G" "H" "K" "A" "I" "G" "T" "V" "L" "V" "G" "P" "T" "P"
[181] "V" "N" "I" "I" "G" "R" "N" "L" "L" "T" "Q" "I" "G" "C" "T" "L" "N" "F"
```

Predicting functional motions of a single structure

Run a Normal Mode Analysis (NMA) - a bioinformatics method to predict fucntional motions.

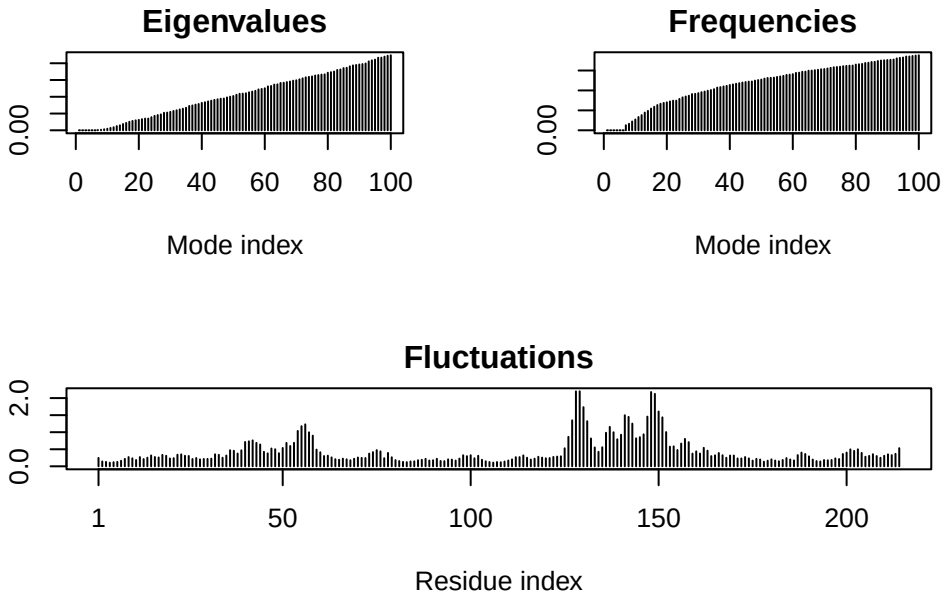
```
adk <- read.pdb("6s36")
```

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

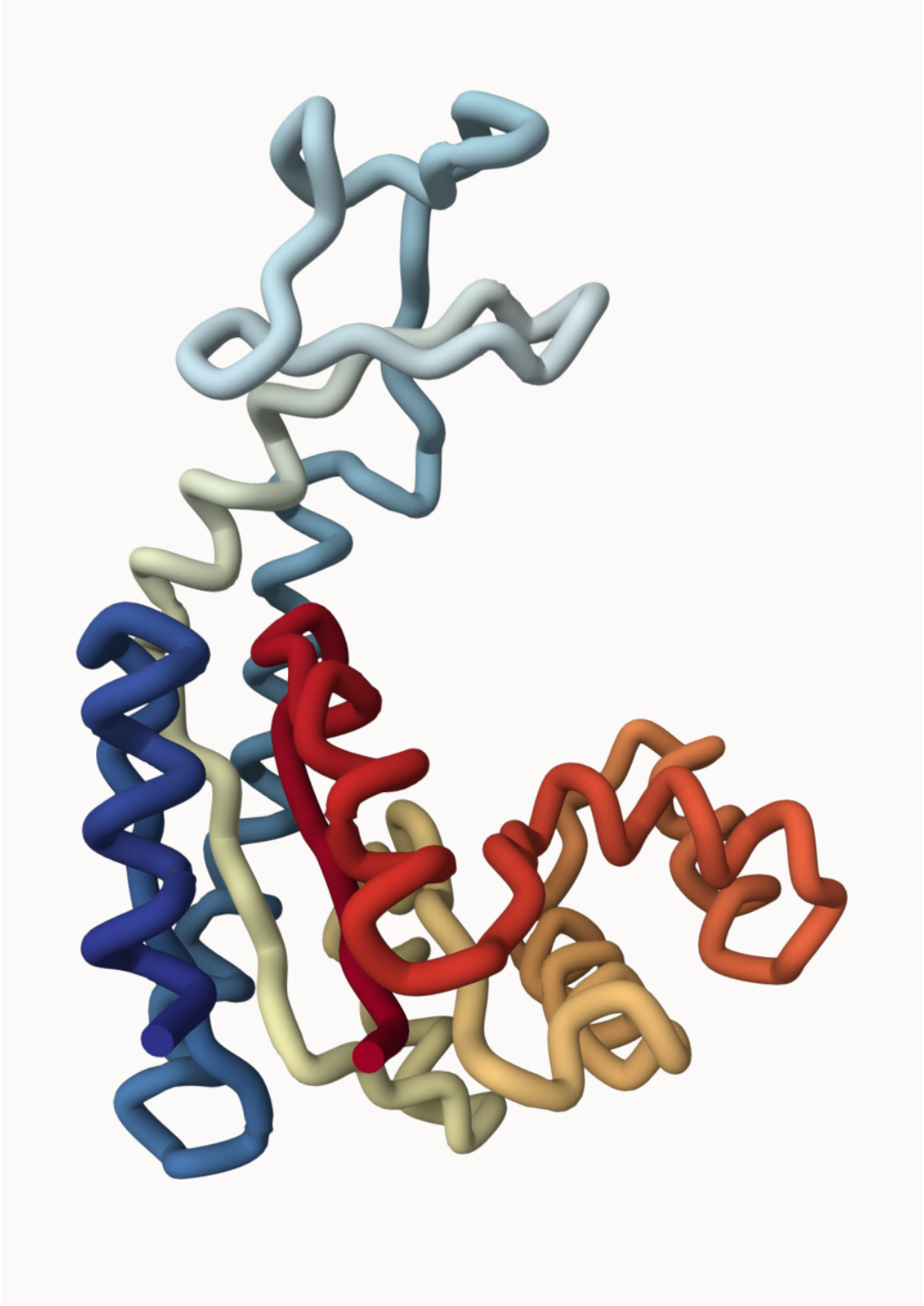
```
modes <- nma(adk)
```

```
Building Hessian...        Done in 0.014 seconds.
Diagonalizing Hessian...   Done in 0.255 seconds.
```

```
plot(modes)
```




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



Q7: How many amino acid residues are there in this pdb object? 198

Q8: Name one of the two non-protein residues? MK1

Q9: How many protein chains are in this structure? 2 chains.