Class 10: Structural Bioinformatics

Nathaniel Lightle (A16669288)

1. Introduction to RCSB Protein Data Bank (PDB)

Downloading CSV file for data distribution

```
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

Making a function to remove the commas from the numbers

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

Removing the commas from the dataset

```
rm.comma(stats$EM)
```

```
[1] 11759  2054  3667   113     9     0
```

We can use `apply()` to fix the whole table

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

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

```
       X.ray               EM              NMR Multiple.methods 
 84.83231383        8.32730146        6.67953467        0.10691797 
     Neutron             Other             Total 
  0.03642780        0.01750427      100.00000000 
```

84.83% are solved by X-ray and 8.33% are solved by EM.

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

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

```
      Protein (only) Protein/Oligosaccharide              Protein/NA
               86.67                    5.37                    5.80
Nucleic acid (only)                   Other Oligosaccharide (only)
                2.05                    0.10                    0.01
```

86.67% are protein only.

> Q3: Skip

Here is a lovely figure of HIP-OR with the catalytic ASP residues, the MK1 compound and the all important water 308

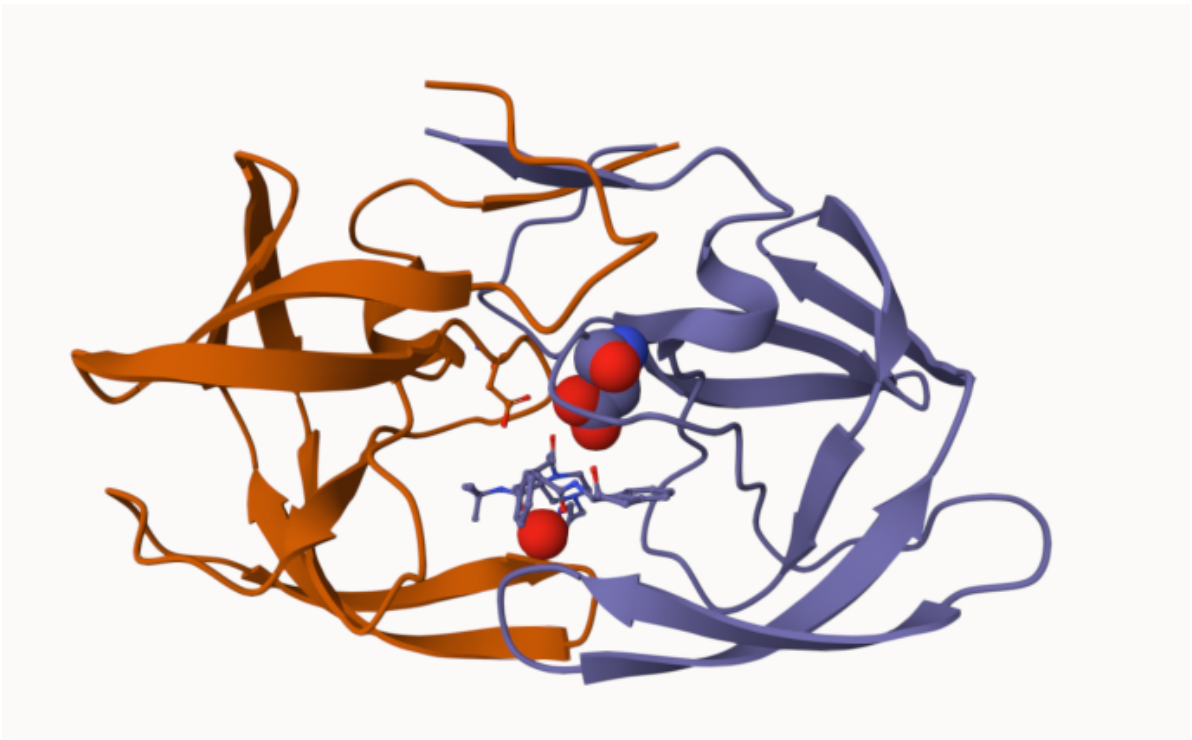
> Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

Hydrogens are smaller than the resolution. So, they don't show up.

> 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?

It's water molecule 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



The bio3d package for structural bioinformatics

```
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
```

```
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>
```

Predicting functional motions of a single structure

Let's finish today with a bioinformatics calculation to predict the functional motions of a PDB structure.

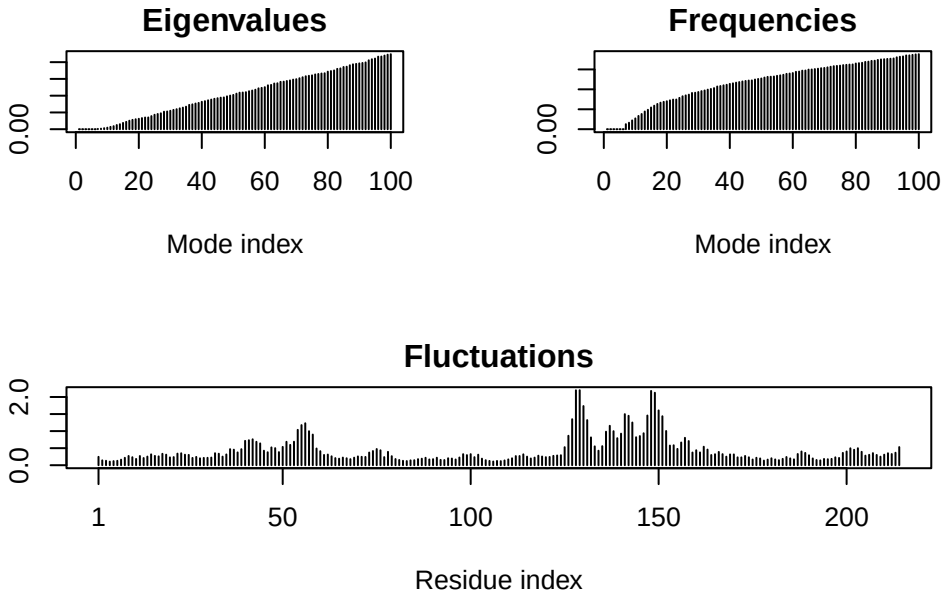
```
adk <- read.pdb("6s36")
```

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

```
m <- nma(adk)
```

```
Building Hessian...        Done in 0.058 seconds.
Diagonalizing Hessian...   Done in 1.046 seconds.
```

```
plot(m)
```




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

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 two non-protein residues is HOH.

Q9. How many protein chains are in this structure?

There are 2 protein chains in this structure.