Class10 - Structural Bioinformatics pt 1

AUTHOR
Gabriella Tanoto (A18024184)

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

The main repository of biomolecular structure data is called **PDB** (Protein Data Bank). It's the second oldest database, after GenBank.

PDB Statistics

The CSV file of the PDB data distributions. It was accessed by: "Analyze" > "PDB Statistics" > "by Experimental Method and Molecular Type"

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

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

```
xray <- gsub(",", "", PDBdata$X.ray) #substitute comma for n  h
xray <- as.integer(xray) # converting to nurmeric.
sum(xray)
```

```
[1] 193952
```

Now, we turn this snippet into a function, so we can reuse it for something else.

```
comma.sum <- function(x) {
  y <- gsub(",", "", x)
  y <- as.integer(y)
  return(sum(y))
}

#Try it!
comma.sum(PDBdata$X.ray)
```

```
[1] 193952
```

Now we answer the question!

```
xray.sum <- comma.sum(PDBdata$X.ray)
```

```
EM.sum <- comma.sum(PDBdata$EM)
total.sum <- comma.sum(PDBdata$Total)

#Division:
(xray.sum+EM.sum)/total.sum*100 #93.7% For both EM and X Ray
```

```
[1] 93.6787
```

```
xray.sum/total.sum*100 #82.4%
```

```
[1] 82.37223
```

```
EM.sum/total.sum*100 #11.3%
```

```
[1] 11.30648
```

- X-ray solved **82.4%** of PBD structures
- EM solved **11.3%** of PDB structures

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

```
comma.sum(PDBdata["Protein (only)","Total"])/comma.sum(PDBdata$
```

```
[1] 86.2107
```

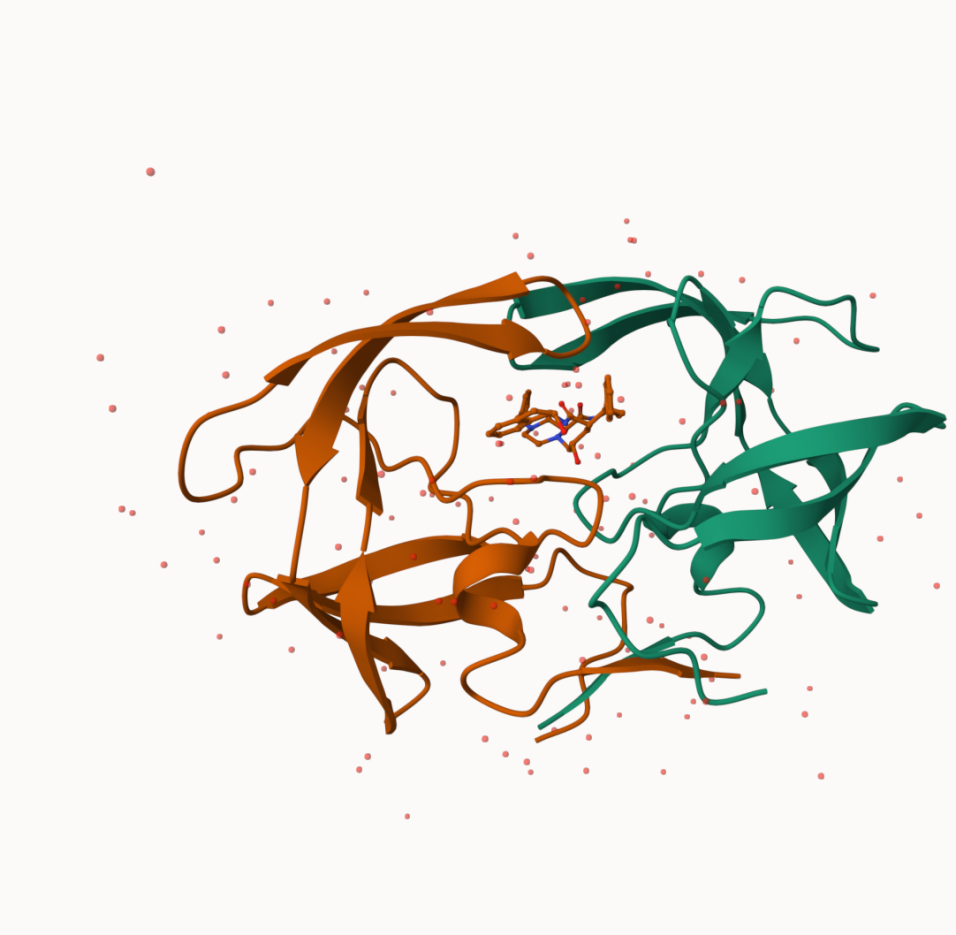
86.2% of the structures in PDB is just the `protein (only)`

> 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?

(SKIPPED)

2. Visualizing the HIV-1 protease structure

Explore the HIV-1 protease structure with PDB code: `1HSG` Molstar homepage here.



This is a protease protein, that cuts off aspartic amino acid in HIV replication. The small molecule inhibitor is a Merck Drug (MK1). The hydrogen is not shown in this structure, because it's too small compared to the resolution (of 2Å). We can change the molecular views

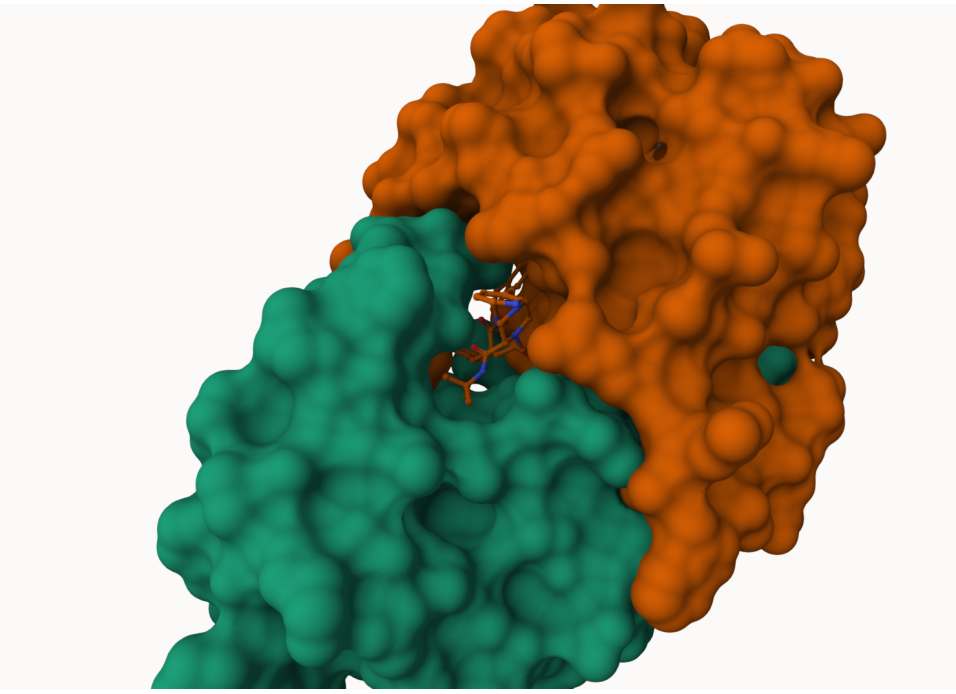


Fig 2. A molecular surface view showing binding cavity

Fig 3. The catalitically important Asp25 amino acids and drug-interacting HOH 308 water molecule

So far, we've used the MolStar website to view protein structures, but we haven't done much bioinformatics.

3. Using the Bio3D in R

First, install & load the bio3D package. It is focused on structural bioinformatics analysis and allows us to read and analyze PB (and related) data.

```
library(bio3d)
```

Read in some data from PDB

```
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
```

```
#lookng at the attributes in the pdb data:
attributes(pdb)
```

```
$names
[1] "atom"   "xyz"    "seqres" "helix"  "sheet"  "calpha"
"remark" "call"

$class
[1] "pdb" "sse"
```

```
#open one of the attributes.
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(pdbseq(pdb))
```

```
  1   2   3   4   5   6
"P" "Q" "I" "T" "L" "W"
```

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

There are **198 amino acid** (198 alpha carbons, each is an amino acid)

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

Water (HOH) is one of them.

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

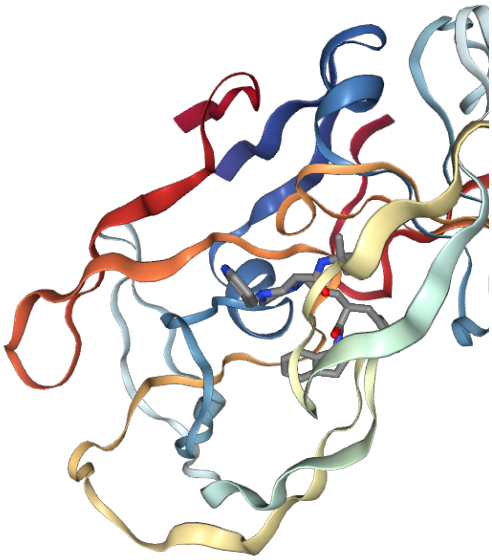
There are **2** protein chains, A and B.

Molecular Visualization in R (interactive!)

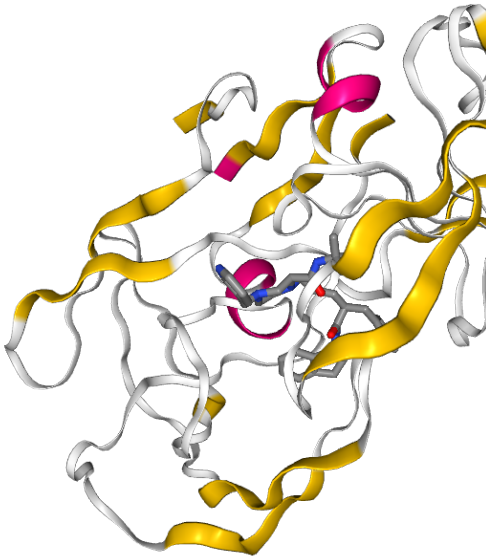
We can make a quick 3D vizualisation with the `bio3dview` package. Use the following to install.

{r} install.packages("pak") pak::pak("bioboot/bio3dview") install.packages("NGLVieweR")

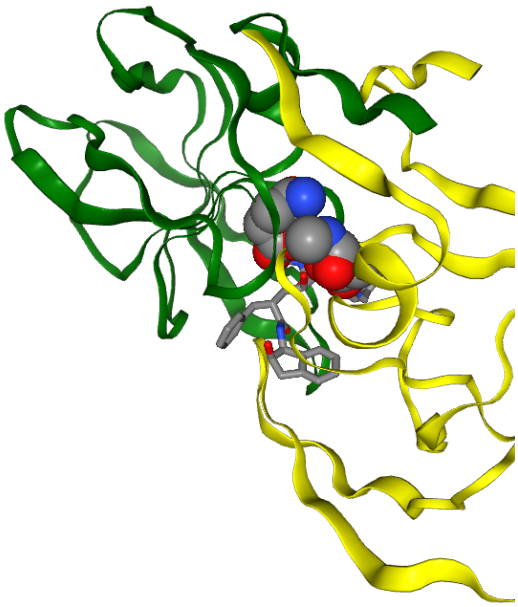
```
library(bio3dview)
library(NGLVieweR)
view.pdb(pdb) |>
  setSpin()
```



```
view.pdb(pdb, backgroundColor = "black", colorScheme = "sse") |>
```



```
sel <- atom.select(pdb, resno=25)
view.pdb(pdb, cols = c("green", "yellow"),
         highlight= sel, highlight.style= "spacefill") |> setRo
```



Predicting Functional motions of a single structure

We can finish off today with a bioinformatics prediction of functional motions od a protein.

We will run a Normal Moda Analysis (NMA) – it inserts an energy into the prediction

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

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

```
adk
```

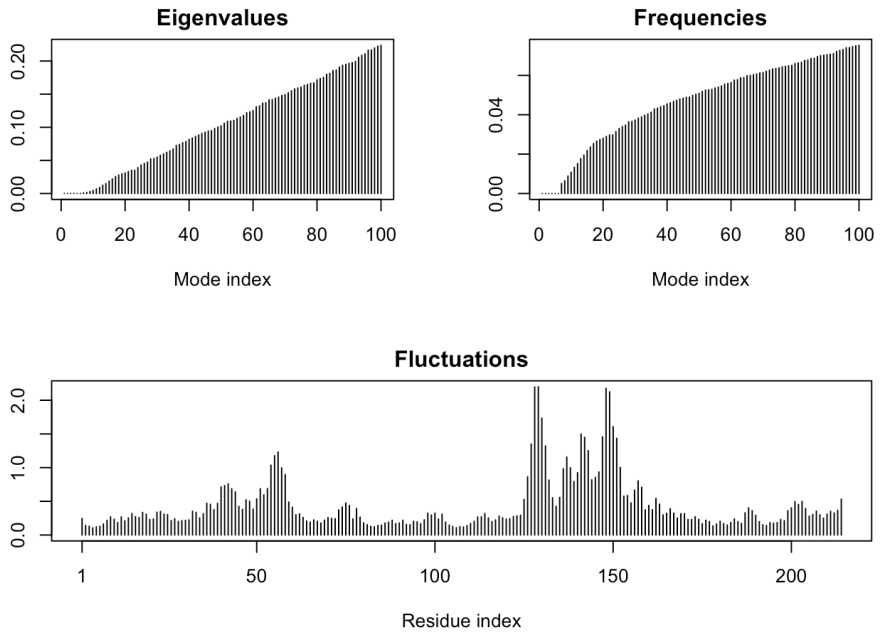
```
Call:  read.pdb(file = "6s36")

  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
```

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

```
 Building Hessian...        Done in 0.044 seconds.
 Diagonalizing Hessian...   Done in 0.516 seconds.
```

```
plot(m)
```



We can write our a trajectory of the predicted dynamics and view in molstar.

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

This makes it into a playable animation in Molstar if we import it from files.