Class 9: Structural Bioinformatic, Part 1

Saba Heydari Seradj (A17002175)

Introduction to the RCSB Protein Data Bank (PDB)

```
pdbdb <- read.csv("Data Export Summary.csv")
```

```
pdbdb
```

```
           Molecular.Type   X.ray     EM    NMR Multiple.methods Neutron Other
1          Protein (only) 167,317 15,698 12,534              208      77    32
2 Protein/Oligosaccharide   9,645  2,639     34                8       2     0
3              Protein/NA   8,735  4,718    286                7       0     0
4     Nucleic acid (only)   2,869    138  1,507               14       3     1
5                   Other     170     10     33                0       0     0
6  Oligosaccharide (only)      11      0      6                1       0     4
    Total
1 195,866
2  12,328
3  13,746
4   4,532
5     213
6      22
```

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

```
# Install tidyverse
# install.packages('tidyverse')
```

```
library(readr)
pdbdb <- read_csv('Data Export Summary.csv')
```

```
Rows: 6 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Molecular Type
dbl (3): Multiple methods, Neutron, Other
num (4): X-ray, EM, NMR, Total

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

```
# characters changed in this table now
pdbdb
```

```
# A tibble: 6 x 8
  `Molecular Type`   `X-ray`    EM   NMR `Multiple methods` Neutron Other  Total
  <chr>                <dbl> <dbl> <dbl>              <dbl>   <dbl> <dbl>  <dbl>
1 Protein (only)      167317 15698 12534                208      77    32 195866
2 Protein/Oligosacc~    9645  2639    34                  8       2     0  12328
3 Protein/NA            8735  4718   286                  7       0     0  13746
4 Nucleic acid (onl~    2869   138  1507                 14       3     1   4532
5 Other                  170    10    33                  0       0     0    213
6 Oligosaccharide (~      11     0     6                  1       0     4     22
```

```
sum(pdbdb$`X-ray`) / sum(pdbdb$Total) * 100
```

```
[1] 83.25592
```

```
sum(pdbdb$EM) / sum(pdbdb$Total) * 100
```

```
[1] 10.2348
```

83.26% of structures in PDB are solved by X-Ray and 10.23% solved by electron microscopy.

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

```
pdbdb$Total[1]/sum(pdbdb$Total) * 100
```

```
[1] 86.3961
```

86.39% of structures in the PDB are protein (only 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?

There are 4563 structures in the current PDB database.

Visualizing the HIV-1 protease structure

We will use Mol and PDB code 1HSG.

```
knitr::include_graphics("1HSG.png")
```

Figure 1: Image of 1HSG from molstar

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

> Mol viewer is simplifying water molecules so they don't cluter the image and we can focus on the important structural information.

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?

```
knitr::include_graphics("water.png")
```

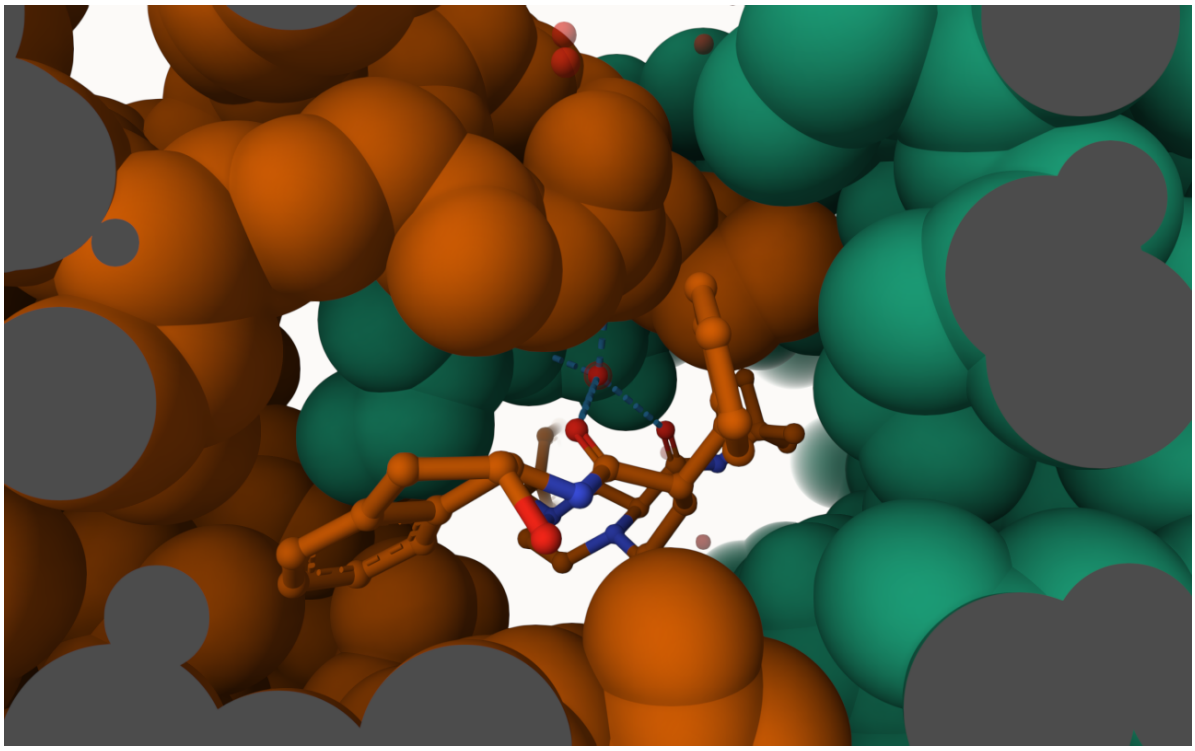


Figure 2: Water 308 zoomed in

Water molecule #308 is the conserved water molecule that is inside the binding site making important hydrogen bonds.

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.

```
knitr::include_graphics("fig3.png")
```

Figure 3: HIV-protease chains with ASP 25 highlighted

Introduction to Bio3D in R

```
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>
```

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

```
sum(pdb$calpha)
```

```
[1] 198
```

There are 198 amino acid residues

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

MK1

Q9: How many protein chains are in this structure?

```
unique(pdb$atom$chain)
```

```
[1] "A" "B"
```

Two unique chains, A and B.

Predicting functional motions of a single strucutre

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

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

```
adk
```

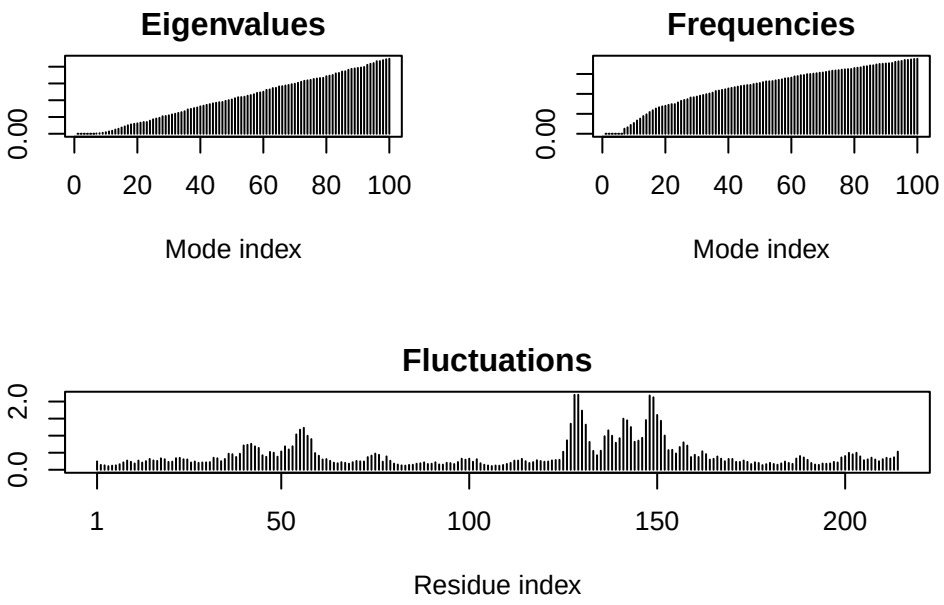
```
 Call:  read.pdb(file = "6s36")

   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
```

Performing flexibility prediction:

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

```
 Building Hessian...        Done in 0.021 seconds.
 Diagonalizing Hessian...   Done in 0.454 seconds.
```

```
plot(m)
```




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