Class 10: Structural Bioinformatics 1

Katherine Quach (A18541014)

Table of contents

PDB Statistics

The Protein Data Bank (PDB) is the main repository of bio-molecular structures. Let's see what it contains:

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

```
           Molecular.Type   X.ray     EM    NMR Integrative Multiple.methods
1          Protein (only) 178,795 21,825 12,773         343              226
2 Protein/Oligosaccharide  10,363  3,564     34           8               11
3              Protein/NA   9,106  6,335    287          24                7
4     Nucleic acid (only)   3,132    221  1,566           3               15
5                   Other     175     25     33           4                0
6  Oligosaccharide (only)      11      0      6           0                1
  Neutron Other   Total
1      84    32 214,078
2       1     0  13,981
3       0     0  15,759
4       3     1   4,941
5       0     0     237
6       0     4      22
```

```
sum(stats$Neutron)
```

```
[1] 88
```

The commna in these numbers leads to the numbers here being read as characters

```
c(100, "Barry")
```

```
[1] "100"   "Barry"
```

```
library(readr)
stats <- read_csv("Data Export Summary.csv")
```

```
Rows: 6 Columns: 9
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Molecular Type
dbl (4): Integrative, 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.
```

```
stats
```

```
# A tibble: 6 x 9
  `Molecular Type`    `X-ray`    EM   NMR Integrative `Multiple methods` Neutron
  <chr>                 <dbl> <dbl> <dbl>       <dbl>               <dbl>   <dbl>
1 Protein (only)       178795 21825 12773         343                 226      84
2 Protein/Oligosacch~   10363  3564    34           8                  11       1
3 Protein/NA             9106  6335   287          24                   7       0
4 Nucleic acid (only)    3132   221  1566           3                  15       3
5 Other                   175    25    33           4                   0       0
6 Oligosaccharide (o~      11     0     6           0                   1       0
# i 2 more variables: Other <dbl>, Total <dbl>
```

```
n.xray <- sum(stats$`X-ray`)
#n.em <-
n.total <- sum(stats$Total)

n.xray/n.total
```

```
[1] 0.8095077
```

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

```
n.xray <- sum(stats$`X-ray`)
n.em <- sum(stats$EM)
n.total <- sum(stats$Total)

(n.xray/n.total) * 100
```

```
[1] 80.95077
```

```
(n.em/n.total) * 100
```

```
[1] 12.83843
```

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

```
n.protein <- sum(stats$Total[grep("Protein", stats$`Molecular Type`)])
(n.protein / n.total) * 100
```

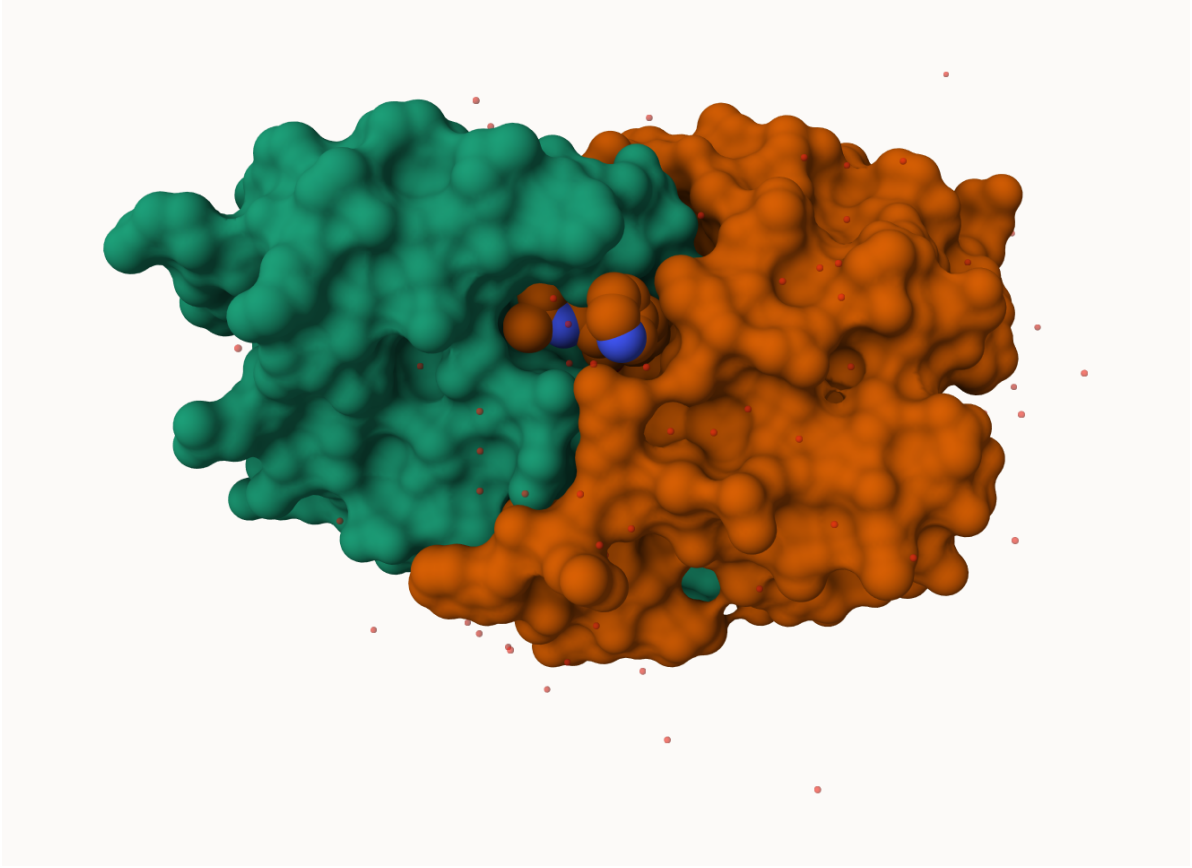
```
[1] 97.9118
```

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?

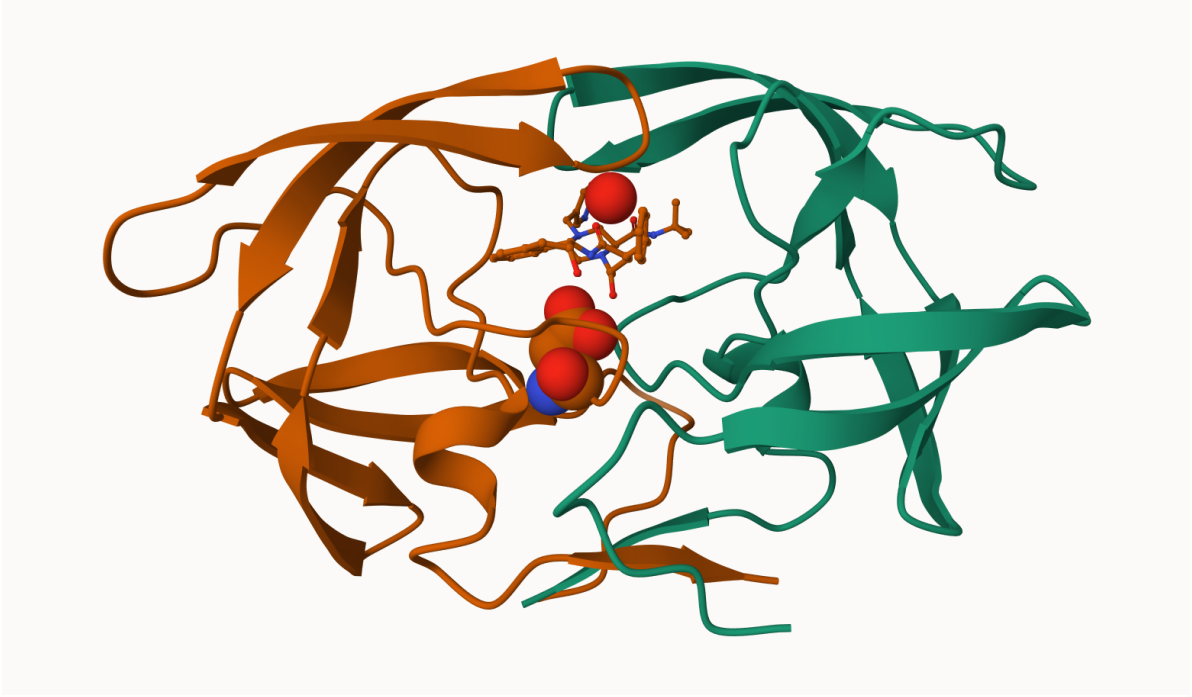
Skip...

Visualizing the HIV-1 protease structure

We can use Molstar viewer online: https://molstar.org/viewer/



A clean image showing the catalytic ASP25 amino acids in both chains of the HIV-PR dimer, along with the inhibitor and all important active site water.



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
```

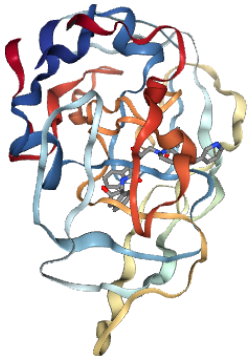
```
attributes(pdb)
```

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

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

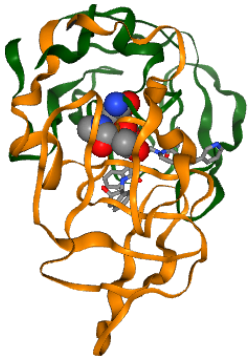
```
head(pdb$atom)
```

```
  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>
```

```
library(bio3dview)
library(knitr)
library(webshot2)

view.pdb(pdb)
```



```
library(knitr)
library(webshot2)

# Select the important ASP 25 residue
sele <- atom.select(pdb, resno=25)

# Highlight them in spacefill representation
view.pdb(pdb, cols=c("darkgreen","darkorange"),
       highlight = sele,
        highlight.style = "spacefill")
```



Predicting functional motions of a single structure

Read an ADK structure from the PDB database:

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

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

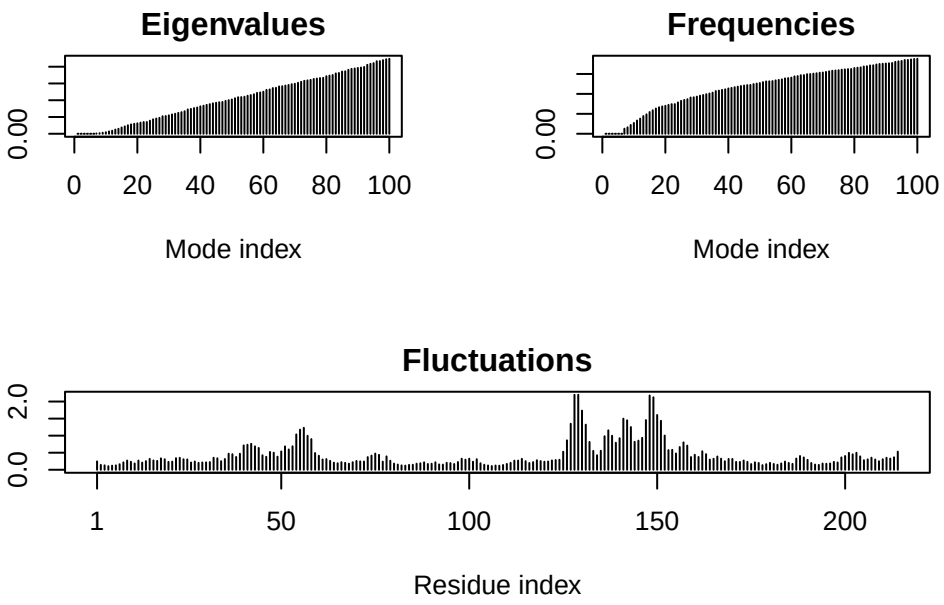
```
adk
```

```
 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 0.045 seconds.
 Diagonalizing Hessian...   Done in 0.672 seconds.
```

```
plot(m)
```



Write out our results as a trajectory/movie of predicted motions:

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