Class 9: Structural Bioinformatics

Michelle Woo

1. Introduction to PDB

loading data in and reading it:

```r
# improving the dataframe
pdb_stats <- read.csv("Data Export Summary.csv", row.names = 1)
pdb_stats
```

```
                          X.ray     EM    NMR Multiple.methods Neutron Other
Protein (only)          154,766 10,155 12,187              191      72    32
Protein/Oligosaccharide   9,083  1,802     32                7       1     0
Protein/NA                8,110  3,176    283                6       0     0
Nucleic acid (only)       2,664     94  1,450               12       2     1
Other                       163      9     32                0       0     0
Oligosaccharide (only)       11      0      6                1       0     4
                          Total
Protein (only)          177,403
Protein/Oligosaccharide  10,925
Protein/NA               11,575
Nucleic acid (only)       4,223
Other                       204
Oligosaccharide (only)       22
```

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

```r
# X-Ray
pdb_stats$X.ray
```

```
[1] "154,766" "9,083"   "8,110"   "2,664"   "163"     "11"
```

```
as.numeric(gsub(',','', pdb_stats$X.ray))
```

```
[1] 154766   9083   8110   2664    163     11
```

```
XRay <- sum(as.numeric(gsub(',','', pdb_stats$X.ray)))

# EM
as.numeric(gsub(',','', pdb_stats$EM))
```

```
[1] 10155  1802  3176    94     9     0
```

```
EM <- sum(as.numeric(gsub(',','', pdb_stats$EM)))

# sum
n_total <- sum(as.numeric(gsub(',','', pdb_stats$Total)))
```

First, we can sum up the elements of the X-Ray column, then of the EM column.

When there are commas in the data set, R can't understand it or read it as numeric, making it not possible to add that column. Another command can be used (`gsub` - to remove the commas and replace it with nothing). Then we need to tell R that we want these characters to be numeric, essentially removing the quotations around our numbers (`as.numeric`)

Then we can divide that by the total in the dataset.

0.93

93%

```
(XRay) / n_total
```

```
[1] 0.8553721
```

```
(EM) / n_total
```

```
[1] 0.07455763
```

```
(XRay + EM) / n_total
```

```
[1] 0.9299297
```

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

```r
pdb_stats[1,]
```

```
                 X.ray     EM    NMR Multiple.methods Neutron Other   Total
Protein (only) 154,766 10,155 12,187              191      72    32 177,403
```

```r
# the proportion of total proteins in first row
total_protein <- as.numeric(gsub(',','',pdb_stats[1,7]))
```

```r
total_protein/n_total
```

```
[1] 0.8681246
```

There are 177403 total proteins in the PDB data set. The proportion of structures in the PDB that are proteins is around 86.81%

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?

Given the overwhelming amount of information, it was difficult to determine how many HIV-1 protease structures there were in the current PDB.

2. Visualizing the HIV-1 protease structure

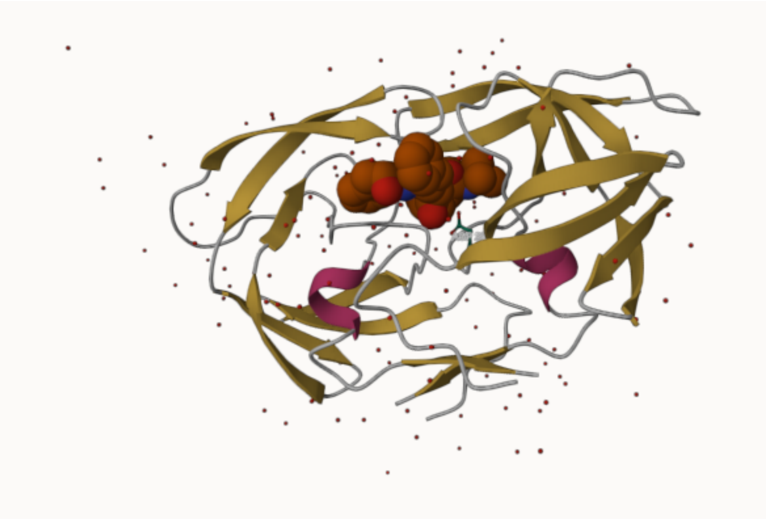
Q4. Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

We just see oxygen and not the two hydrogen molecules due to the xray resolution. From a MOL file, the structure may not all be displayed depending on the settings and limits of the software.

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?

The water molecule has a residue number of HOH-308 which interacts with the two Asp's at A-Asp 25 and B-Asp 25.

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.



Discussion Topic: Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site?

Indinavir is a protease inhibitor to treat HIV which binds directly to the active site. Entering the binding site is the first step for the inhibitor and the enzyme to interact, from there the inhibitor is able to bind to the active site by being a match.

3. Introduction to Bio3D in R

loading in Bio3D:

```r
library(bio3d)
```

Reading PDB file data into R:

```r
pdb <- read.pdb("1hsg")
```

```
  Note: Accessing on-line PDB file
```

```r
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
```

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

There are 198 amino acid residues from the pdb object. The label is falls under is "(residues/Calpha atoms#: 198)"

Q8. Name one of the two non-protein residues.

One of the non-protein residues is HOH 127. This is found at "Non-protein/nucleic resid values: [HOH (127), MK1 (1)]".

Q9. How many protein chains are in this structure?

There are 2 protein chains in this structure under the label “Chains#: 2 (values: A B)”

Finding the attributes:

```
attributes(pdb)
```

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

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

Accessing individual attributes:

```
# accessing atom attribute
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 read a new PDB structure of Adenylate Kinase and perform Normal mode analysis:

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

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

```
adk
```

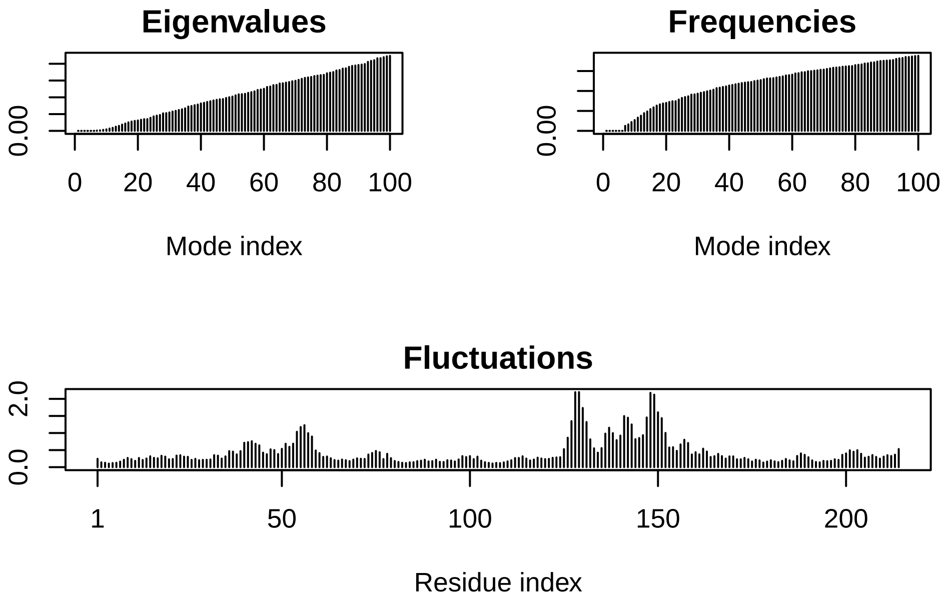
```
 Call:  read.pdb(file = "6s36")

   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
```

Using normal mode analysis (NMA) to predict protein flexibility and potential functional motions:

```
# Perform flexiblity prediction
m <- nma(adk)
```

```
 Building Hessian...        Done in 0.03 seconds.
 Diagonalizing Hessian...   Done in 0.26 seconds.
```

```
plot(m)
```



```
# viewing these predicted motions
mktrj(m, file="adk_m7.pdb")
```

The motion can be captured on Mol*