Class 9 structural bioinformatics part 1

Serena Quezada (PID: A18556865)

The main database for structural biology is called the PDB. Let's have a look at what it contains:

1. Introduction to the RCSB Protein

```
read.csv("Data Export Summary.csv")
```

```
           Molecular.Type  X.ray     EM    NMR Integrative Multiple.methods
1          Protein (only) 176,204 20,299 12,708         342              218
2 Protein/Oligosaccharide  10,279  3,385     34           8               11
3              Protein/NA   9,007  5,897    287          24                7
4     Nucleic acid (only)   3,066    200  1,553           2               15
5                   Other     173     13     33           3                0
6  Oligosaccharide (only)      11      0      6           0                1
  Neutron Other   Total
1      83    32 209,886
2       1     0  13,718
3       0     0  15,222
4       3     1   4,840
5       0     0     222
6       0     4      22
```

```
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)       176204 20299 12708         342                218      83
2 Protein/Oligosacch~   10279  3385    34           8                 11       1
3 Protein/NA             9007  5897   287          24                  7       0
4 Nucleic acid (only)    3066   200  1553           2                 15       3
5 Other                   173    13    33           3                  0       0
6 Oligosaccharide (o~       11     0     6           0                  1       0
# i 2 more variables: Other <dbl>, Total <dbl>
```

```
n.total <- sum(stats$Total, na.rm = TRUE)
```

```
stats$Total
```

```
[1] 209886  13718  15222   4840    222     22
```

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

```
n.xray <- sum(stats$`X-ray`)
n.xray / n.total * 100
```

```
[1] 81.48087
```

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

```
round(stats$Total[1]/n.total * 100,2)
```

```
[1] 86.05
```

> 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 about 4,865 HIV-protease structures

2. Visualizing the HIV-1 protease structure

Package for structural bioinformatics

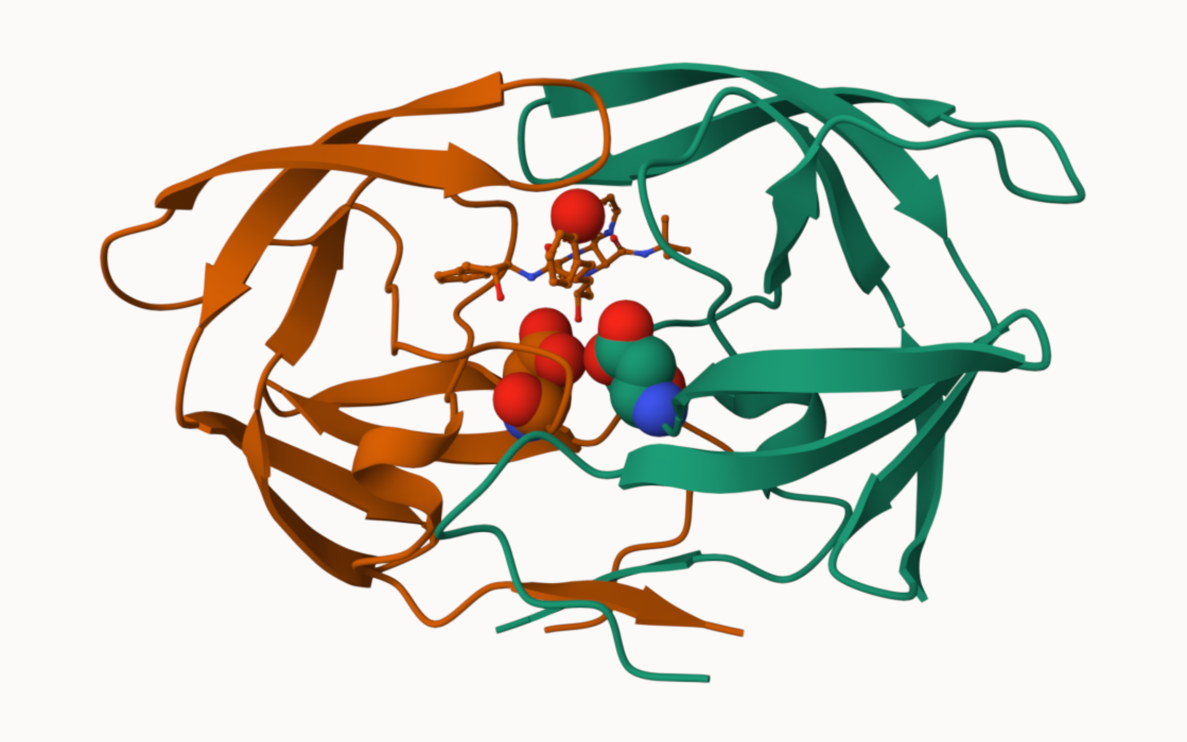
```
library(bio3d)

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

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

```
hiv
```

```


 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
```

Let's first use the Mol* viewer to explore this structure



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

We only see one atom per water molecule in this structure because its a simplified symbol representing the entire molecule, not a single real atom.

> 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 residue number this water molecule has is HOH 306

3. Introduction to Bio3D in R

> Q6. And a view of the ligand with catlytic ASP 25 amino-acids (spacefill) and the all important active site water molecule (spacefill): Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site?

Reading PDB file data in R

```
library(bio3d)

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

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

```
Warning in get.pdb(file, path = tempdir(), verbose = FALSE):
/var/folders/zk/6hldzf5n74scx33n9zx154500000gn/T//Rtmp64DTAV/1hsg.pdb exists.
Skipping download
```

```
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

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

HOH, which is a water molecule

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

There are 2 protein chains, labeled chain A and B

Quick PDB visualization in R

```r
attributes(pdb)
```

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

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

```r
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>
```

I can interactively view these PDB objects in R with the new **bio3dview** package. This is not yet on CRAN

To install this I can setup **pak** package and use it to install **bio3dview** from GitHub. In my console I first run

install.packages("pak") pak::pak("bioboot/bio3dview")

library(bio3dview) library(NGLVieweR)

view.pdb(hiv)

Due to the interactive photo it is not visible on pdf but below is the code that would be used to run the interactive photo

In order to change some settings

sel <- atom.select(hiv, resno= 25)

view.pdb(hiv, highlight = sel, colorScheme = "chain", col = c("blue", "orange"), backgroundColor ="pink")

Prediction protein flexibility

We can run bioinformatics calculation to predict protein dynamics - i.e. functional motions.

We will use the `nmac()` function:

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

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

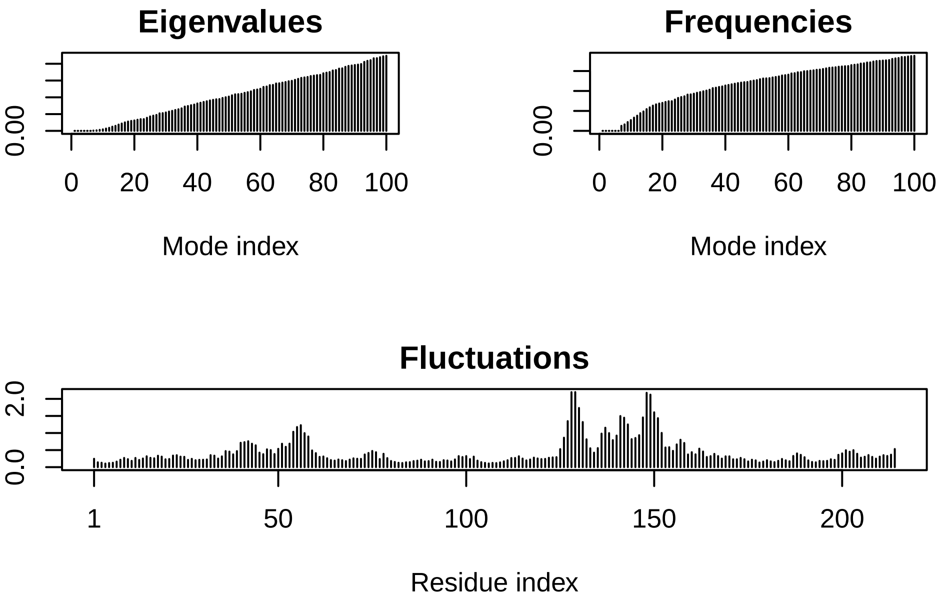
```r
adk
```

```

 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
```

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

```

 Building Hessian...        Done in 0.019 seconds.
 Diagonalizing Hessian...   Done in 0.719 seconds.
```

```r
plot(m)
```



Generate a “trajectory” of predicted motion

mktrj(m, file = “ADK_nma.pdb”)

view.nma(m)

4. Comparative structure analysis of Adenylate Kinase

Install packages in the R console NOT your Rmd/Quarto file

install.packages(“bio3d”) install.packages(“NGLVieweR”)

install.packages(“pak”) pak::pak(“bioboot/bio3dview”)

install.packages(“BiocManager”) BiocManager::install(“msa”)

Q10. Which of the packages above is found only on BioConductor and not CRAN?

msa if found only on BioConductor and not CRAN

Q11. Which of the above packages is not found on BioConductor or CRAN?

Bio3dview is not found on BioConductor or CRAN

Q12. True or False? Functions from the pak package can be used to install packages from GitHub and BitBucket?

True, the pak package is a modern R package manager that can install packages from multiple sources - including CRAN, BioConductor, GitHub and BitBucket