Lab09: Structural Bioinformatics pt. 1

Adam Bisharat

The main database for structural data is called the PDB (protein Data Bank). Let's see what it contains:

```
PDB <- read.csv("ProteinData.csv" , row.names = 1)

PDB
```

```
                          X.ray     EM    NMR Multiple.methods Neutron Other
Protein (only)          167,192 15,572 12,529              208      77    32
Protein/Oligosaccharide   9,639  2,635     34                8       2     0
Protein/NA                8,730  4,697    286                7       0     0
Nucleic acid (only)       2,869    137  1,507               14       3     1
Other                       170     10     33                0       0     0
Oligosaccharide (only)       11      0      6                1       0     4
                          Total
Protein (only)          195,610
Protein/Oligosaccharide  12,318
Protein/NA               13,720
Nucleic acid (only)       4,531
Other                       213
Oligosaccharide (only)       22
```

```
PDB$Total
```

```
[1] "195,610" "12,318"  "13,720"  "4,531"   "213"     "22"
```

```
as.numeric(sub(",", "", PDB$Total))
```

```
[1] 195610  12318  13720   4531    213     22
```

I could turn this into a function to fix the whole table or any future table I read like this:

```
x <- PDB$Total
as.numeric(sub(",", "", x))
```

```
[1] 195610  12318  13720   4531    213     22
```

```
comma2numeric <- function(x) {
  as.numeric(sub(",", "", x))
}
```

```
comma2numeric(PDB$X.ray)
```

```
[1] 167192   9639   8730   2869    170     11
```

```
apply(PDB, 2, comma2numeric)
```

```
      X.ray    EM   NMR Multiple.methods Neutron Other  Total
[1,] 167192 15572 12529              208      77    32 195610
[2,]   9639  2635    34                8       2     0  12318
[3,]   8730  4697   286                7       0     0  13720
[4,]   2869   137  1507               14       3     1   4531
[5,]    170    10    33                0       0     0    213
[6,]     11     0     6                1       0     4     22
```

##Or try a differnt read/import funciton:

```
library(readr)
PDBN <- read_csv("ProteinData.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.
```

```
PDBN$Total
```

```
[1] 195610  12318  13720   4531    213     22
```

```
sum(PDBN$Total)
```

```
[1] 226414
```

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

```
Combined.Percent <- ((sum(PDBN$'X-ray') + sum(PDBN$'EM') ) / sum(PDBN$Total) ) * 100

XRay.Percent <- (sum(PDBN$'X-ray') / sum(PDBN$Total) ) * 100

EM.Percent <- sum(PDBN$'EM') / sum(PDBN$Total)  * 100


Combined.Percent
```

```
[1] 93.4845
```

```
XRay.Percent
```

```
[1] 83.30359
```

```
EM.Percent
```

```
[1] 10.18091
```

```
PDBN
```

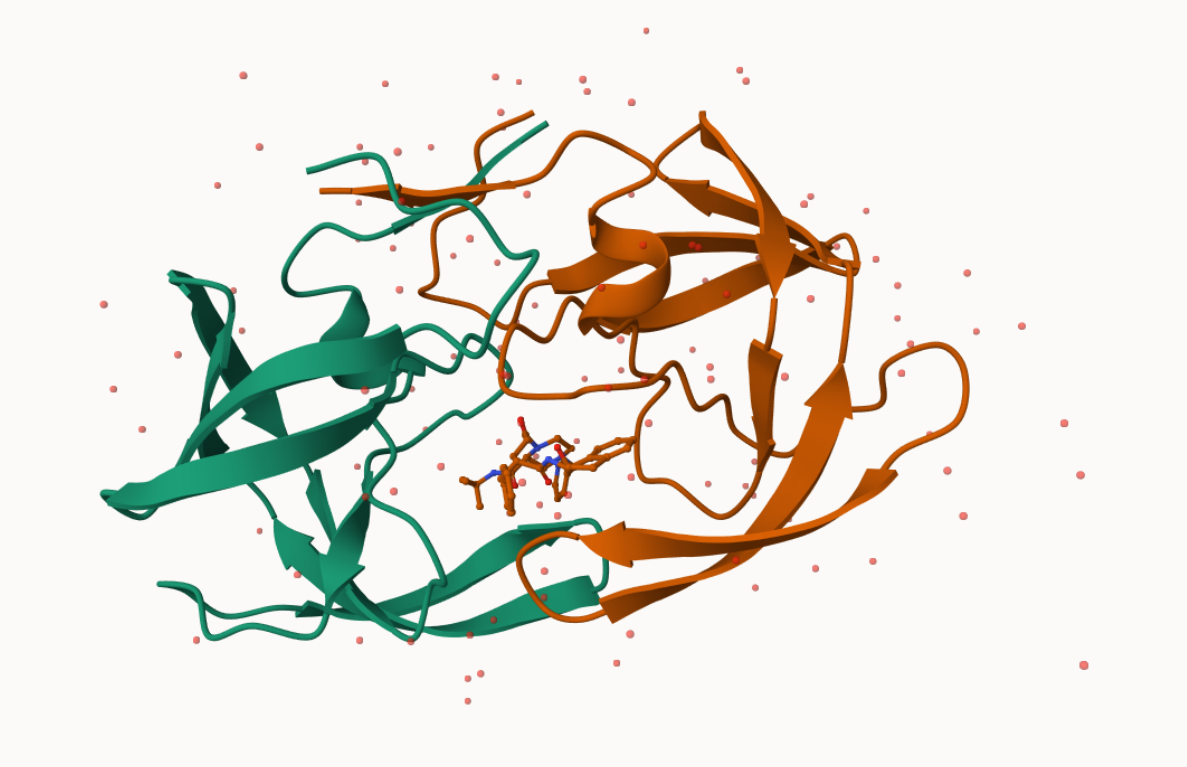
```
# A tibble: 6 x 8
  `Molecular Type`   `X-ray`    EM   NMR `Multiple methods` Neutron Other  Total
  <chr>                <dbl> <dbl> <dbl>              <dbl>   <dbl> <dbl>  <dbl>
1 Protein (only)      167192 15572 12529                208      77    32 195610
2 Protein/Oligosacc~    9639  2635    34                  8       2     0  12318
3 Protein/NA            8730  4697   286                  7       0     0  13720
4 Nucleic acid (onl~    2869   137  1507                 14       3     1   4531
5 Other                  170    10    33                  0       0     0    213
6 Oligosaccharide (~      11     0     6                  1       0     4     22
```

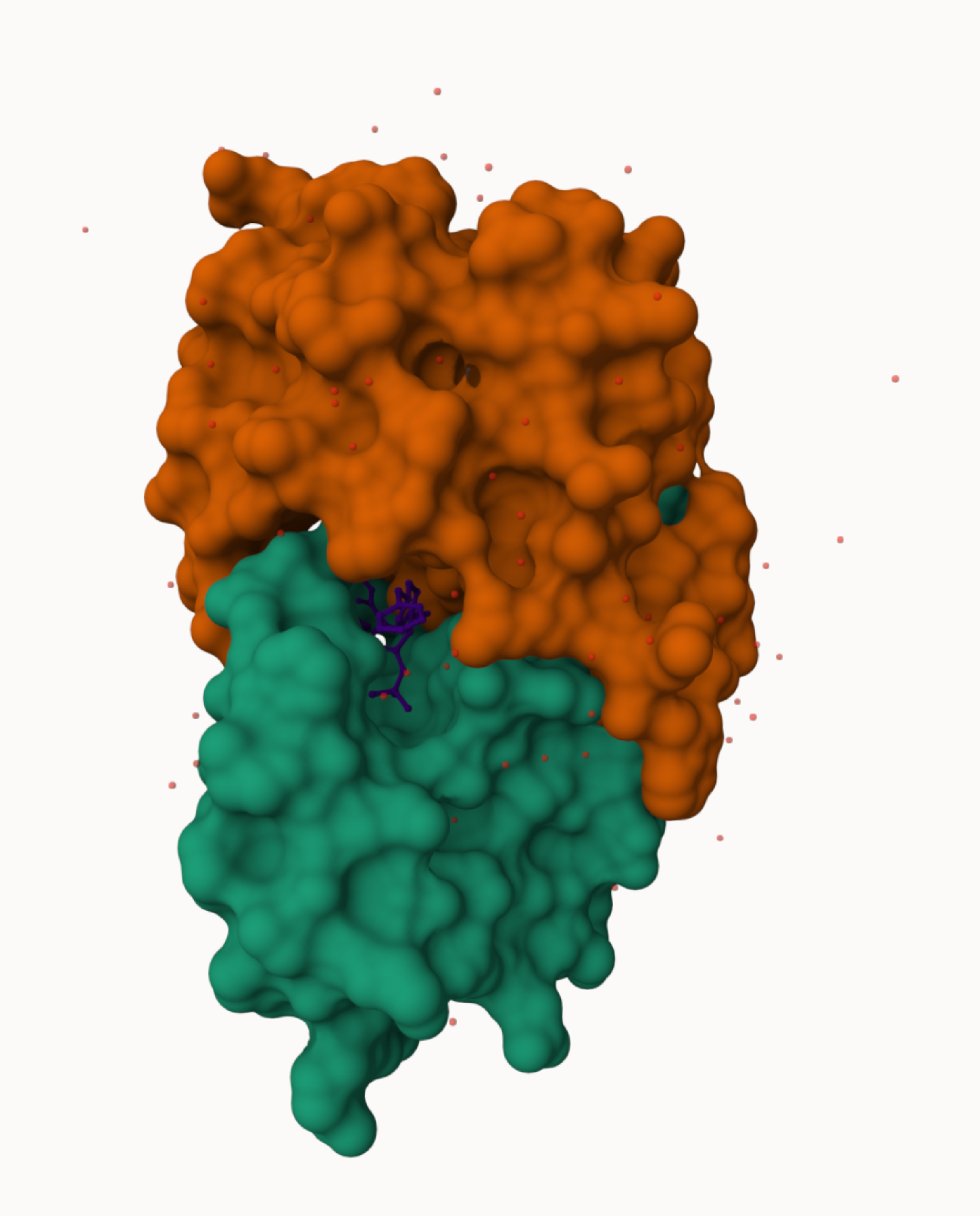
Q2: What proportion of structures in the PDB are protein?

```r
PDBN$Total[1] / sum(PDBN$'Total')
```

```
[1] 0.8639483
```

Molstar Viewer: 1HSG





> 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?

226114

Bio3D Package very useful

```
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>
```

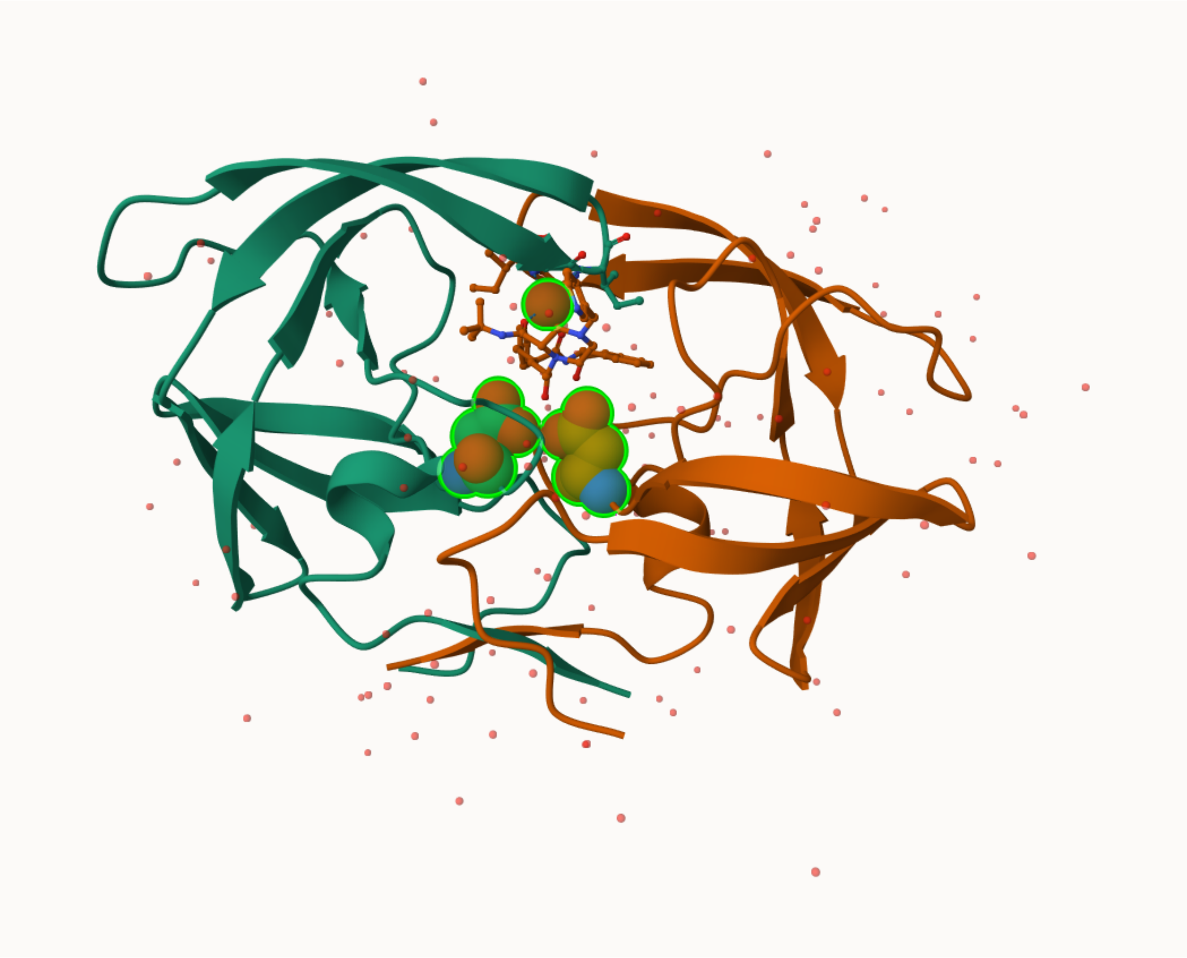
```
3  <NA>     C   <NA>
4  <NA>     O   <NA>
5  <NA>     C   <NA>
6  <NA>     C   <NA>
```

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

Hydrogens have little to no electron density.

> 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

308



> 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?

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

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

```
[1] 198
```

```r
length (pdbseq(pdb))
```

```
[1] 198
```

198 amino acids

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

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

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

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

2 chains

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

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

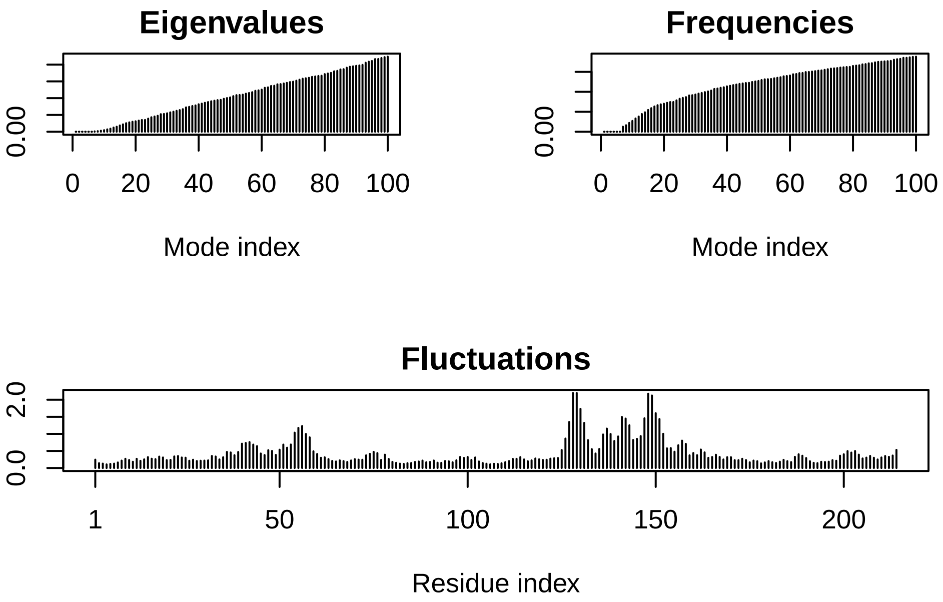
```r
adk
```

```
 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
```

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

```
 Building Hessian...        Done in 0.021 seconds.
 Diagonalizing Hessian...   Done in 0.472 seconds.
```

```
plot(m)
```



Writes PDB file to make animation of predicted motions.

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

I can open this in Mol* to play the trajectory…

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

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

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

> Q13. How many amino acids are in this sequence, i.e. how long is this sequence?

> Q14. What do you note about this plot? Are the black and colored lines similar or different? Where do you think they differ most and why?