Lab9_inclass

Qianqian Tao

```
db <- read.csv("Data Export Summary_2.csv")
#To remove comma in the original dataset use global substitution (gsub)
modified_data <- gsub(",", "", db$X.ray)
#To change the value to numeric, use as.numeric()
numeric_data <- as.numeric(modified_data)
#you can also turn this into a function
make_sum_numeric <- function(x){
    sum(as.numeric(gsub(",", "",x)))
}
```

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

```
total <-make_sum_numeric(db$Total)
X.Ray_sum <- make_sum_numeric(db$X.ray)
X.Ray_protion <- (X.Ray_sum)/total*100
X.Ray_protion <- round(X.Ray_protion,3)
EM_sum <- make_sum_numeric(db$EM)
EM_protion <- (EM_sum)/total
EM_protion <- round(EM_protion, 3)
```

85.537 of structures are solved by X-Ray and 0.075 of structures are solved by EM.

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

```
head(db)
```

```
           Molecular.Type   X.ray     EM    NMR Multiple.methods Neutron Other
1          Protein (only) 154,766 10,155 12,187              191      72    32
2 Protein/Oligosaccharide   9,083  1,802     32                7       1     0
3              Protein/NA   8,110  3,176    283                6       0     0
```

```
4     Nucleic acid (only)   2,664     94  1,450                12       2     1
5                   Other     163      9     32                 0       0     0
6  Oligosaccharide (only)      11      0      6                 1       0     4
    Total
1 177,403
2  10,925
3  11,575
4   4,223
5     204
6      22
```

```
#how to access to the last column
column_number <- ncol(db)
db[1,column_number]
```

```
[1] "177,403"
```

```
protein_structure <- make_sum_numeric(db[1,8])
protein_protion <- protein_structure/total
protein_protion <- round(protein_protion,2)
```

0.87 of structures are proteins

> 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?

Skipped

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

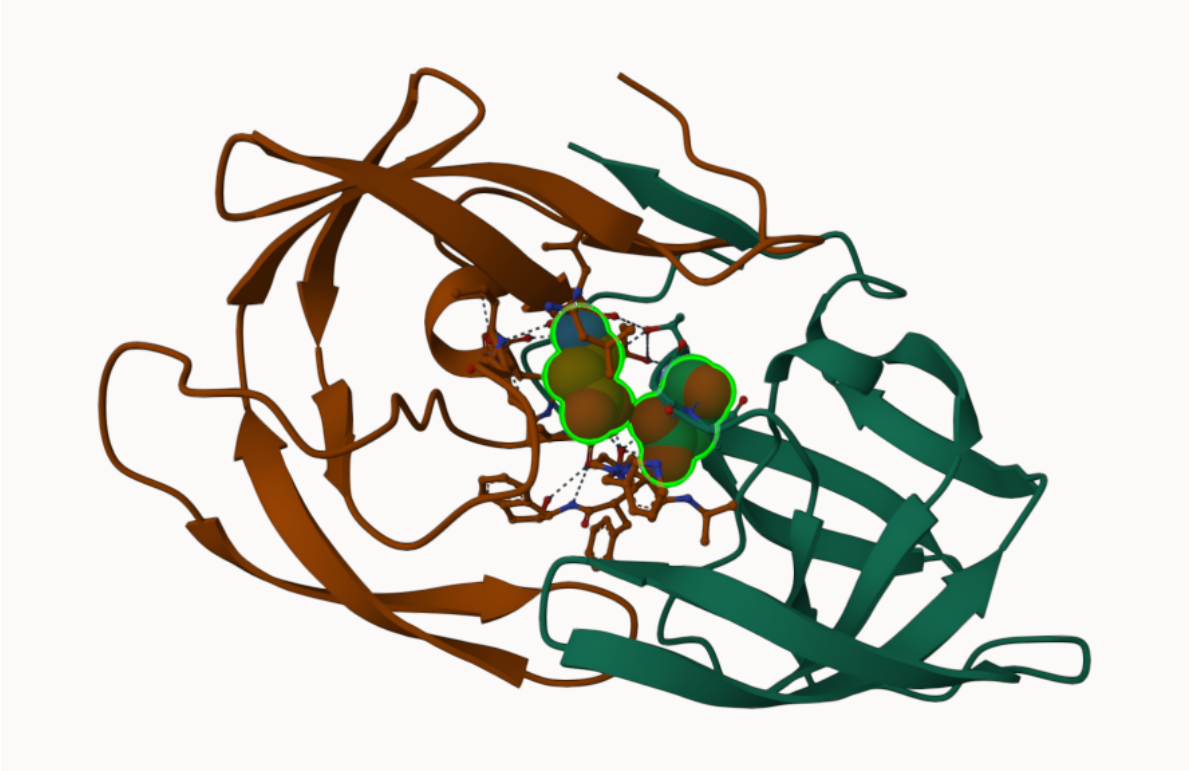
Hydrogen is the smallest atom. It is hard to be captured. Oxygen is much larger than hydrogen and can be visualized.

The structure is too low a resolution to see a H atom. You need a sub 1 Angstrom resolution to see the Hydrogen 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?

HOH308

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.



#Working with Structures in R We can use the `bio3d` package to read and perform bioinformatics calculations on PDB structures

```
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
```

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

198

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

H2O and MK1

Q9: How many protein chains are in this structure?

2 (A and B)

```
#A list object, combine vectors... all things together
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>
```

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

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

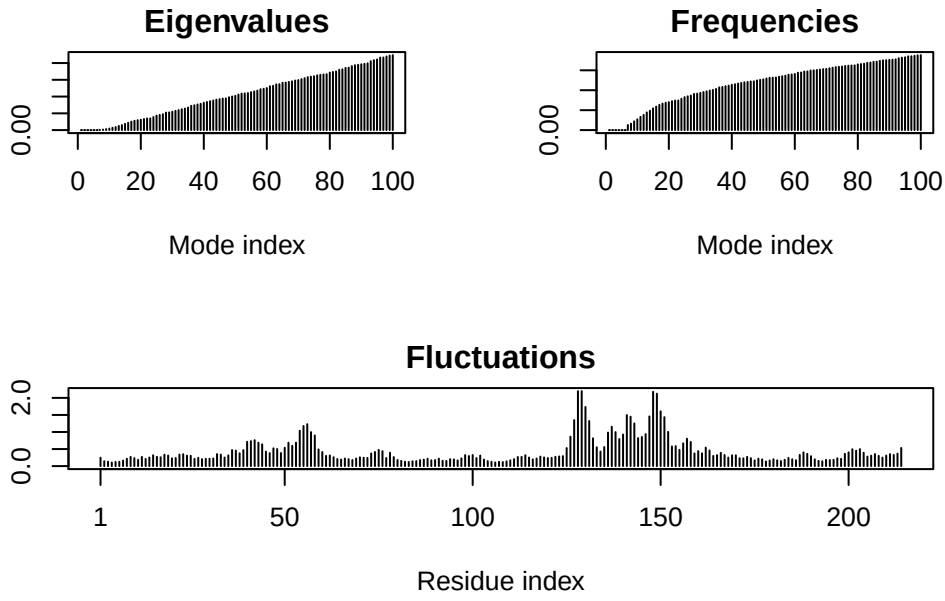
```
#has one chain A
```

```
# Perform flexiblity prediction,Normal mode analysis (NMA) is a structural bioinformatics

m <- nma(adk)
```

```
Building Hessian...        Done in 0.031 seconds.
Diagonalizing Hessian...   Done in 0.37 seconds.
```

```
plot(m)
```




```
#bottom graphs interested to explore, peaks are regions that are more likely to move, more
```

Write out a "movie" (a.k.a. trajectory) of the motion for viewing in MOLstar

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