Class 09: Structural Bioinformatics

Max Gruber

5/3/23

1: Introduction to the RCSB Protein Data Bank (PDB)

To read the file, we are going to use the command `read.csv`.

```
pdb_stats <- read.csv('Data Export Summary.csv', row.names = 1)
```

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

I need to sum all the elements of the X.ray column.

```
pdb_stats$X.ray
```

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

We are going to use `gsub` to remove the commas.

```
xray_without_commas <- gsub(',',"",pdb_stats$X.ray)
as.numeric(xray_without_commas)
```

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

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

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

```
total_without_commas <- gsub(',',"",pdb_stats$Total)
as.numeric(total_without_commas)
```

```
[1] 177403  10925  11575   4223    204     22
```

I use the `sum` command to get the sum.

```
n_xray <- sum(as.numeric(xray_without_commas))
n_xray
```

```
[1] 174797
```

```
n_em <- sum(as.numeric(em_without_commas))
n_em
```

```
[1] 15236
```

```
n_total <- sum(as.numeric(total_without_commas))
n_total
```

```
[1] 204352
```

Now I can find the percentage.

```
p_xray <- (n_xray)/ n_total
p_xray
```

```
[1] 0.8553721
```

```
p_em <- (n_em)/ n_total
p_em
```

```
[1] 0.07455763
```

```
((n_xray + n_em) / n_total)*100
```

```
[1] 92.99297
```

92.99%

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

```
protein_without_commas <- gsub(',',"",pdb_stats[1,7])
as.numeric(protein_without_commas)
```

```
[1] 177403
```

```
n_protein <- sum(as.numeric(protein_without_commas))

n_protein/n_total
```

```
[1] 0.8681246
```

.8681, or 86.81%, are protein.

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?

1,292 was one possible answer. Overall, it was difficult to search/find an exact value for the specific criteria because different values came up depending on how and where search criteria was entered.

2. Visualizing the HIV-1 protease structure

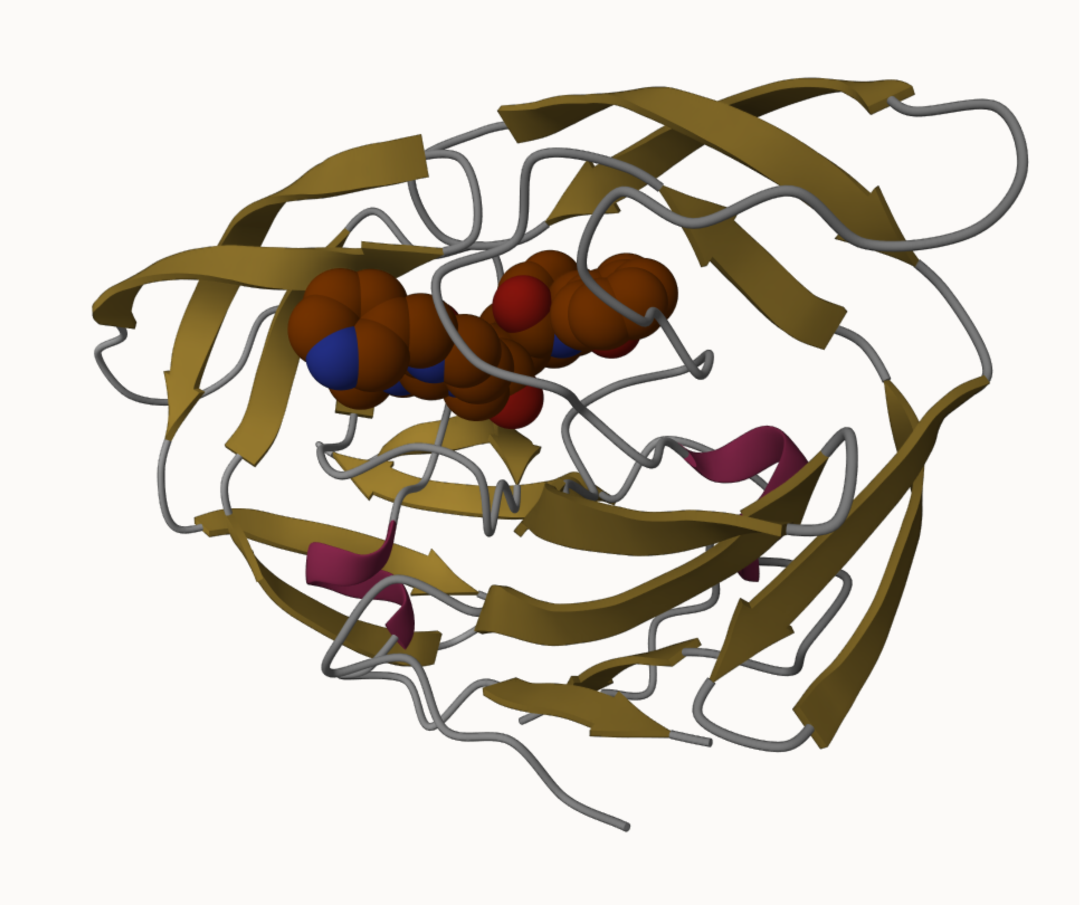
Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

We see one because in x-ray crystallography it's hard to detect hydrogen (as its density is so small), and so in the structures only oxygen is seen (one atom) because it's simply easier to see.

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?

Yes, you can identify the water molecule, and it has a residue number of 308 (HOH 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.



And a second image with the catalytic residues ASP 25 in each chain and the critical water (highlighted in green).

3. Introduction to Bio3D in R

```
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?

There are 198 amino acid residues in this pdb object.

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

One of the two non-protein residues is HOH, or water.

Q9: How many protein chains are in this structure?

There are 2 protein chains in this structure.

```
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 by NMA

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

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

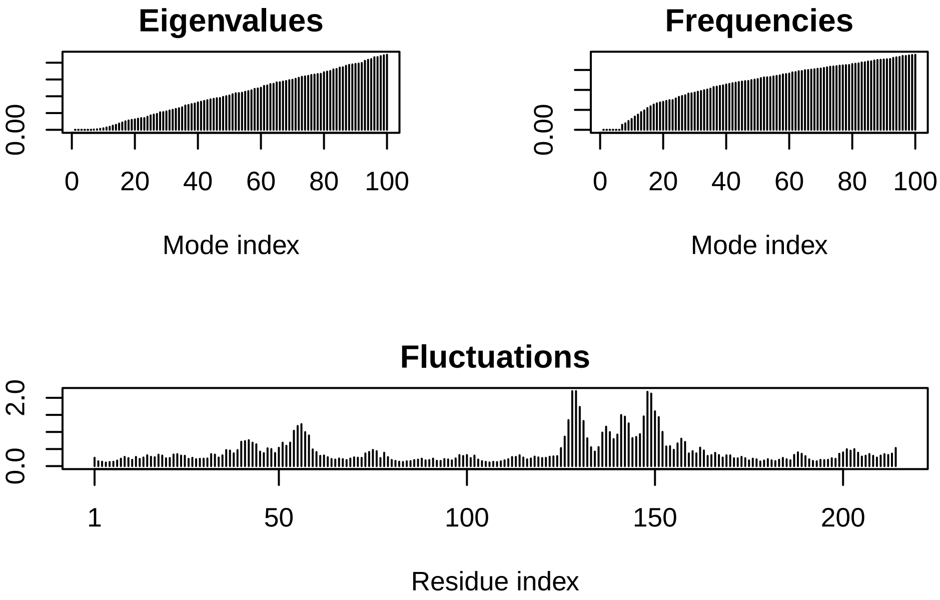
```
adk
```

```
 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 0.012 seconds.
 Diagonalizing Hessian...   Done in 0.26 seconds.
```

```
plot(m)
```



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

4. Comparative structure analysis of Adenylate Kinase

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

The msa package.

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

The bio3d view package.

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

True

Search and retrieve ADK structures

```
#install.packages("bio3d")
#install.packages("devtools")
#install.packages("BiocManager")

#BiocManager::install("msa")
#devtools::install_bitbucket("Grantlab/bio3d-view")

library(bio3d)
aa <- get.seq("1ake_A")
```

```
Warning in get.seq("1ake_A"): Removing existing file: seqs.fasta
```

```
Fetching... Please wait. Done.
```

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

214