Class 9: Structural Bioinformatics

Aparajita Pranjal

5/3/23

PDB Statistics

Importing and reading in csv file:

```
pdb_stats <- read.csv("data_export_summary.csv", row.names = 1)
head(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. **92.99%**

```
xray_total <- sum(as.numeric(gsub(",", "", pdb_stats$X.ray)))
em_total <- sum(as.numeric(gsub(",", "", pdb_stats$EM)))
n_total <- sum(as.numeric(gsub(",", "", pdb_stats$Total)))
xray_em_percent <- ((xray_total + em_total)/n_total)*100
xray_em_percent
```

```
[1] 92.99297
```

- **Q2:** What proportion of structures in the PDB are protein? **86.81%**

```
protein_total <- as.numeric(gsub(",", "", pdb_stats[1,7]))
protein_percentage <- (protein_total/n_total)*100
protein_percentage
```

```
[1] 86.81246
```

- **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? **204,352 structures**

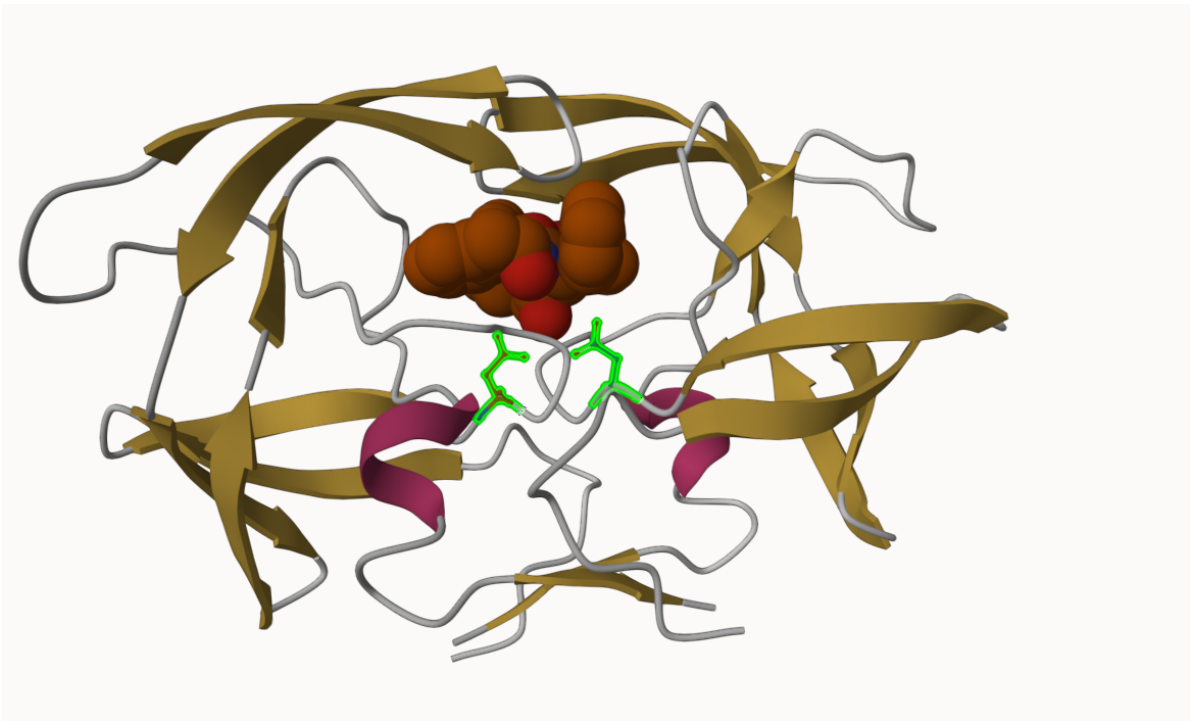
Mol* Exploration

Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure? **The hydrogen atoms are not resolved in X-ray crystallography hence only the oxygen molecule is used to represent water.**

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? **HOH 308**

Visualizing the HIV-1 protease structure:

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.



Intro 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
```

```
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>
```

- **Q7:** How many amino acid residues are there in this pdb object? **198 residues**
- **Q8:** Name one of the two non-protein residues? **HOH**
- **Q9:** How many protein chains are in this structure? **2 protein chains**

Predicting functional motions of a single structure by normal mode analysis

New protein: Adenylate Kinase

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

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

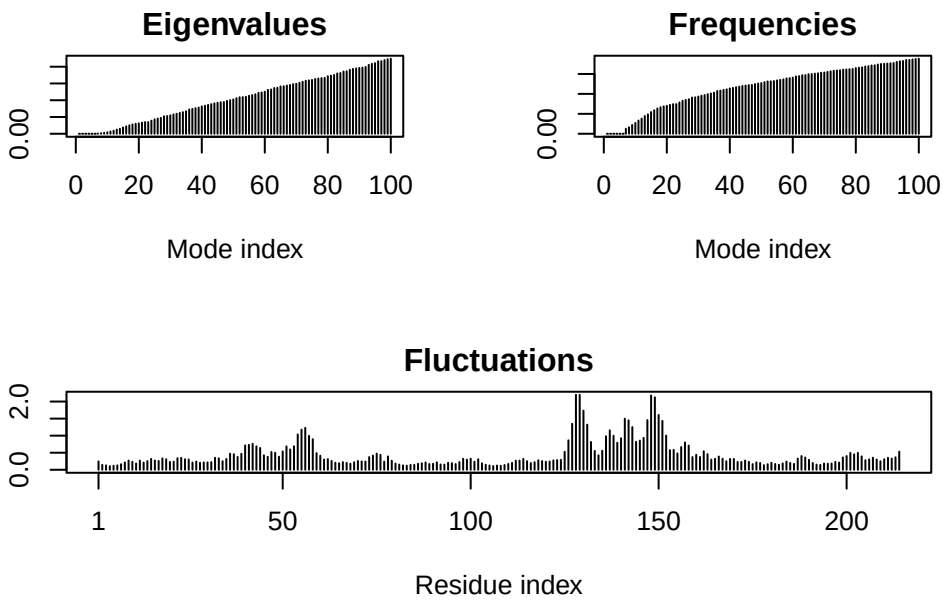
```
adk
```

```
 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 0.038 seconds.
 Diagonalizing Hessian...   Done in 0.537 seconds.
```

```
plot(m)
```



Creating movie of protein structure to view in Mol*:

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

Comparative structure analysis of Adenylate Kinase

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

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

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

- **Q10.** Which of the packages above is found only on BioConductor and not CRAN? **msa**
- **Q11.** Which of the above packages is not found on BioConductor or CRAN? **None of them**

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

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

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

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

```
aa
```

```
              1        .         .         .         .         .         60
pdb|1AKE|A    MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
              1        .         .         .         .         .         60

             61        .         .         .         .         .         120
pdb|1AKE|A    DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
             61        .         .         .         .         .         120

            121        .         .         .         .         .         180
pdb|1AKE|A    VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
            121        .         .         .         .         .         180

            181        .         .         .   214
pdb|1AKE|A    YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG
            181        .         .         .   214

Call:
  read.fasta(file = outfile)

Class:
  fasta

Alignment dimensions:
  1 sequence rows; 214 position columns (214 non-gap, 0 gap)

+ attr: id, ali, call
```

Q13. How many amino acids are in this sequence, i.e. how long is this sequence? **214 amino acids**