Class 09: Structual Bioinformatics

Patricia Chen A16138722

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

PDB statistics

The PDB is the main database for structural information on biomolecules.

Download a CSV file from the PDB site (accessible from "Analyze" > "PDB Statistics" > "by Experimental Method and Molecular Type". Move this CSV file into your RStudio project and use it to answer the following questions:

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

	X.ray	EM	NMR	Multiple.methods	Neutron	Other
Protein (only)	152,809	9,421	12,117	191	72	32
Protein/Oligosaccharide	9,008	1,654	32	7	1	0
Protein/NA	8,061	2,944	281	6	0	0
Nucleic acid (only)	2,602	77	1,433	12	2	1
Other	163	9	31	0	0	0
Oligosaccharide (only)	11	0	6	1	0	4

	Total
Protein (only)	174,642
Protein/Oligosaccharide	10,702
Protein/NA	11,292
Nucleic acid (only)	4,127
Other	203
Oligosaccharide (only)	22

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

```
  gsub(",","",db$Total)
```

```
[1] "174642" "10702"  "11292"  "4127"   "203"    "22"
```

```
  gsub("","",db$EM)
```

```
[1] "9,421" "1,654" "2,944" "77"    "9"     "0"
```

```
  gsub(",","",db$X.ray)
```

```
[1] "152809" "9008"   "8061"   "2602"   "163"    "11"
```

```
  sum(as.numeric(gsub(",","",db$X.ray)))
```

```
[1] 172654
```

I'm doing the same thing over and over to write a function.

```
  # I will work with x as input

    sum_comma <-function(x){

      # Substitute the comma and convert to numeric

    sum(as.numeric(gsub(",","",x)))

  }
```

For Xray:

```
  sum_comma(db$X.ray)/sum_comma(db$Total)
```

```
[1] 0.8590264
```

> Answer:0.859 or 85.9% percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.

For EM:

```
round(sum_comma(db$EM)/sum_comma(db$Total),2) #should yield 0.07
```

```
[1] 0.07
```

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

The proportion of protein structures in the PDB is yielded as 0.87 or 87% after rounding.

```
round(sum_comma(db$Total[1])/sum_comma(db$Total))
```

```
[1] 1
```

Q3: SKIPPED

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?

Answer: We only see one atom molecule in the structure because the hydrogen atom with atomic mass of 1 u is too small to be displayed on the structure. So the two hydrogen atoms in the water molecule is hidden and only the one oxygen atom is displayed, which is what we observed.

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?

Answer: Yes, I was able to identify the critical "conserved" water molecule. This molecule is at the position Asp (D25), with the residue number of this water molecule is HOH308.

Now you should be able to produce an image similar or even superior to Figure 2 and save it to an image file.

Q6: Generate and save a figure clearly showing the two distinct chains of HIV-pro tease 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.

Answer: Insert captured image from molstar for question 6.

Discussion Topic: Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site?

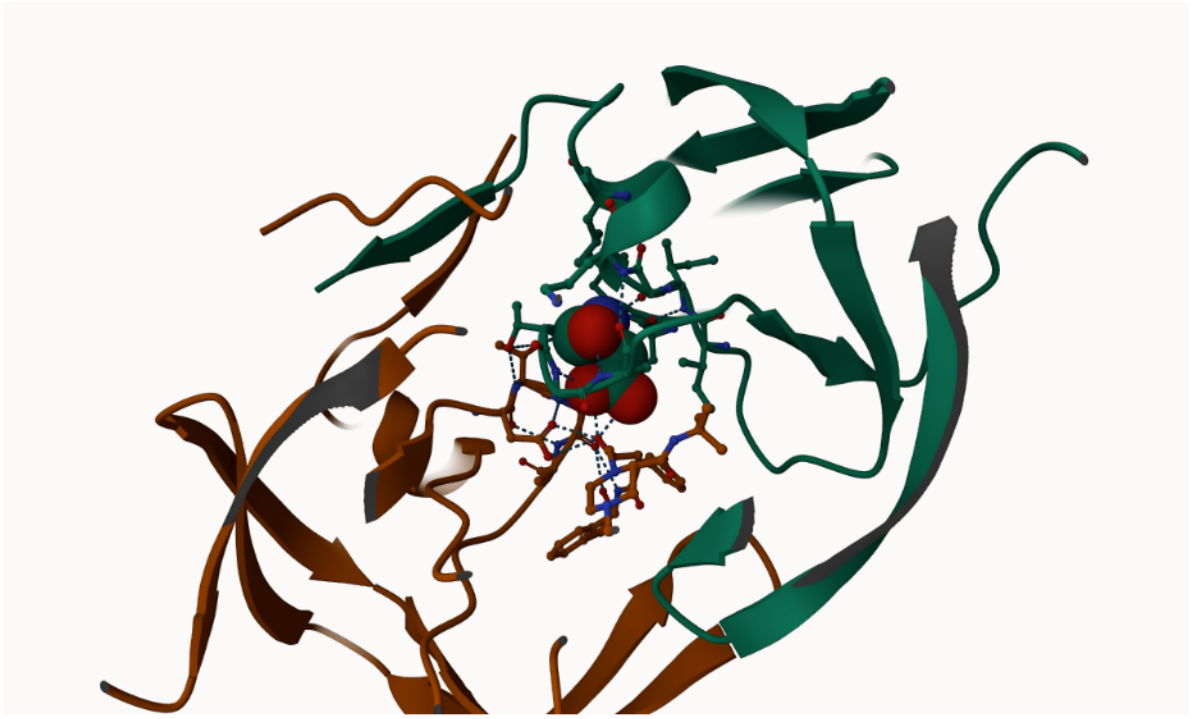


Figure 1: HIV-PR structure from MERK with a bound drug

3. Introduction to Bio3D in R

Working with Structures 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? Answer: There are 198 amino acid residues in the pdb object.

Q8: Name one of the two non-protein residues? Answer: The name of the two non-protein residues is called HOH.

Q9: How many protein chains are in this structure? Answer: There are two protein chains in this structure.

```
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>
```

2.Predicting functional motions of a single structure

Read on ADK structure

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

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

```
adk
```

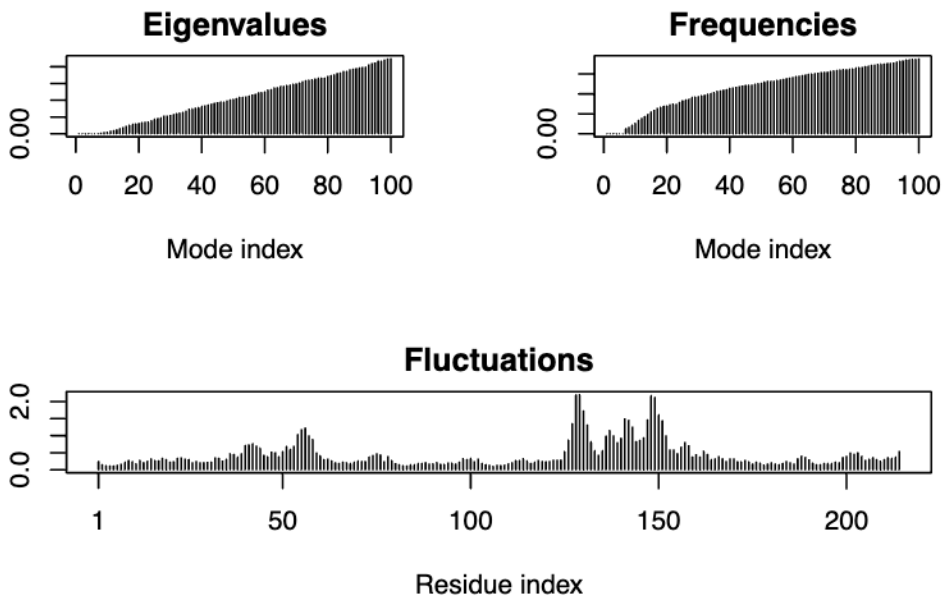
```
 Call:  read.pdb(file = "6s36")

   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
```

Perform a prediction of flexibility with a technique called NMA (normal mode)

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

```
 Building Hessian...        Done in 0.021 seconds.
 Diagonalizing Hessian...   Done in 0.444 seconds.
```

```
plot(m)
```



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

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

```
##  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
```