

Class 10 Structural Bioinformatics (pt1)

Erin McTavish PID: A17300519

PDB statistics

The Protein Data Bank (PDB) is the main repository of biomolecular structures. Let's see what it contains:

```
stats <- read.csv("Data Export Summary.csv")
stats
```

	Molecular.Type	X.ray	EM	NMR	Integrative	Multiple.methods
1	Protein (only)	178,795	21,825	12,773	343	226
2	Protein/Oligosaccharide	10,363	3,564	34	8	11
3	Protein/NA	9,106	6,335	287	24	7
4	Nucleic acid (only)	3,132	221	1,566	3	15
5	Other	175	25	33	4	0
6	Oligosaccharide (only)	11	0	6	0	1

	Neutron	Other	Total
1	84	32	214,078
2	1	0	13,981
3	0	0	15,759
4	3	1	4,941
5	0	0	237
6	0	4	22

```
stats$X.ray
```

```
[1] "178,795" "10,363" "9,106" "3,132" "175" "11"
```

```
sum(stats$Neutron)
```

```
[1] 88
```

The comma in these numbers leads to the numbers here being read as character.

```
c(100, 10, "barry")
```

```
[1] "100"  "10"   "barry"
```

```
library(readr)
stats <- read_csv("Data Export Summary.csv")
```

```
Rows: 6 Columns: 9
```

```
-- Column specification -----
```

```
Delimiter: ","
```

```
chr (1): Molecular Type
```

```
dbl (4): Integrative, 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.

```
stats
```

```
# A tibble: 6 x 9
```

	<code>`Molecular Type`</code>	<code>`X-ray`</code>	EM	NMR	Integrative	<code>`Multiple methods`</code>	Neutron
	<code><chr></code>	<code><dbl></code>	<code><dbl></code>	<code><dbl></code>	<code><dbl></code>	<code><dbl></code>	<code><dbl></code>
1	Protein (only)	178795	21825	12773	343	226	84
2	Protein/Oligosacch~	10363	3564	34	8	11	1
3	Protein/NA	9106	6335	287	24	7	0
4	Nucleic acid (only)	3132	221	1566	3	15	3
5	Other	175	25	33	4	0	0
6	Oligosaccharide (o~	11	0	6	0	1	0

```
# i 2 more variables: Other <dbl>, Total <dbl>
```

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

```
n.xray <- sum(stats$`X-ray`)
#n.em <-
n.total <- sum(stats$Total)
n.xray/n.total
```

```
[1] 0.8095077
```

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

```
library(dplyr)
```

Attaching package: 'dplyr'

The following objects are masked from 'package:stats':

```
filter, lag
```

The following objects are masked from 'package:base':

```
intersect, setdiff, setequal, union
```

```
stats2 <- stats %>%  
  mutate(Total = rowSums(across(where(is.numeric)), na.rm = TRUE))
```

```
n.protein_only <- sum(stats2$Total[stats2$`Molecular Type` == "Protein (only)"])  
n.total <- sum(stats2$Total)  
  
n.protein_only / n.total
```

```
[1] 0.8596889
```

Q3. SKIP..

Visualizing the HIV-1 protease structure

We can use the Molstar viewer online: <https://molstar.org/viewer/>

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

PDB is only representing the Oxygen atoms in the water molecule and not hydrogen. This is a simplified representation.

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

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.

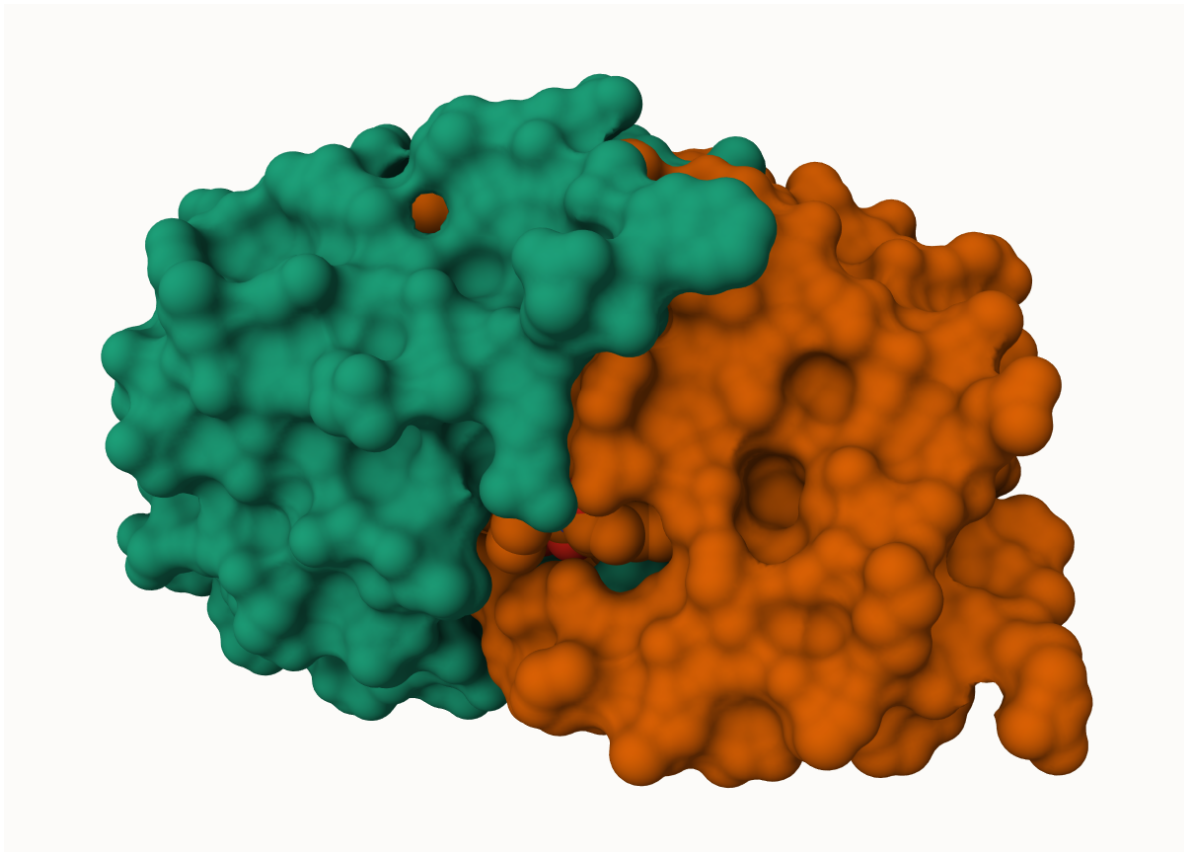


Figure 1: My first image of HIV-Pr with surface display showing ligand binding

A new clean image showing the catalytic ASP25 amino acids in both chains of the HIV-PR dimer along with the inhibitor and all important active site water.



Bio3D package for structural bioinformatics

```
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?

HOH

Q9: How many protein chains are in this structure?

2

```
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>										

```
##library(bio3dview)
```

```
##view.pdb(pdb)
```

```
# Select the important ASP 25 residue  
#sele <- atom.select(pdb, resno=25)  
  
# and highlight them in spacefill representation  
#view.pdb(pdb, cols=c("navy","teal"),  
          # highlight = sele,  
          # highlight.style = "spacefill")
```

Predicting functional motions of a single structure

Read an ADK structure from the PDB database:

```
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:

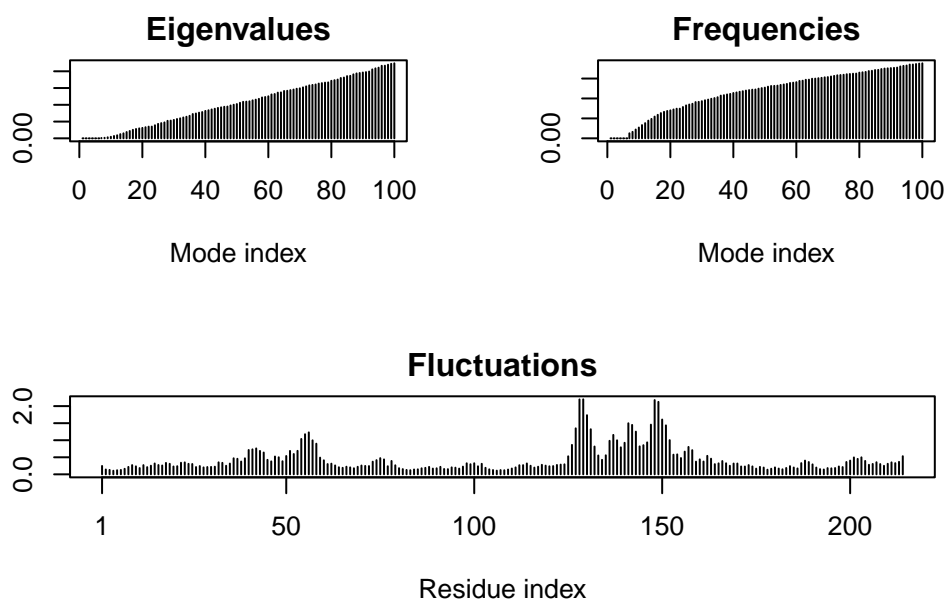
```
MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLV  
DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPELIVDKI  
VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG  
YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG
```

```
+ attr: atom, xyz, seqres, helix, sheet,  
      calpha, remark, call
```

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

```
Building Hessian...      Done in 0.011 seconds.  
Diagonalizing Hessian... Done in 0.051 seconds.
```

```
plot(m)
```



write out results as a wee trajectory/movie of predicted motions:

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