Class 10: PDB

Linh Tran (PID:A16435846)

About

For this class, we will explore the PDB website to analyze proteins.

Introduction to the RCSB Protein Data Bank (PDB)

PDB Statistics

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

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

```
Rows: 6 Columns: 8
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (1): Molecular Type
dbl (3): 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.
```

```r
CSV
```

```
# A tibble: 6 x 8
  `Molecular Type`   `X-ray`    EM   NMR `Multiple methods` Neutron Other  Total
  <chr>                <dbl> <dbl> <dbl>              <dbl>   <dbl> <dbl>  <dbl>
1 Protein (only)      163468 13582 12390                204      74    32 189750
2 Protein/Oligosacc~    9437  2287    34                  8       2     0  11768
3 Protein/NA            8482  4181   286                  7       0     0  12956
4 Nucleic acid (onl~    2800   132  1488                 14       3     1   4438
5 Other                  164     9    33                  0       0     0    206
6 Oligosaccharide (~      11     0     6                  1       0     4     22
```

Percentage solved by EM and X-ray

```
sum(CSV$`X-ray`, CSV$EM)/sum(CSV$Total)*100
```

```
[1] 93.34352
```

93.34%

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

Percentage of structures that are protein

```
sum(CSV[1:3,2:7])/sum(CSV$Total)*100
```

```
[1] 97.87077
```

> 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?

Based on the search, there are *4445 structures*

PDB Format

1HSG was downloaded

Visualizing the HIV-1 protease structure

Using Mol*

Download and moved to Class10 folder ## The important role of water >Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

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

> Q7: [Optional] As you have hopefully observed HIV protease is a homodimer (i.e. it is composed of two identical chains). With the aid of the graphic display can you identify secondary structure elements that are likely to only form in the dimer rather than the monomer?

Introduction to Bio3D in R

```r
library(bio3d)
```

Reading PDB file data into R

```r
pdb <- read.pdb("1hsg")
```

```
  Note: Accessing on-line PDB file
```

```r
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.

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

MK1 is a non-protein residue.

> Q9: How many protein chains are in this structure?

There are 2 protein chains.

We can lok at the attributes with `attributes()` and get access to a particular attribute with `pdb$attribute`

```r
attributes(pdb)
```

```
$names
[1] "atom"   "xyz"    "seqres" "helix"  "sheet"  "calpha" "remark" "call"

$class
[1] "pdb" "sse"
```

```r
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

Next, We read a new PDB structure of Adenylate Kinase

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

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

```r
adk
```

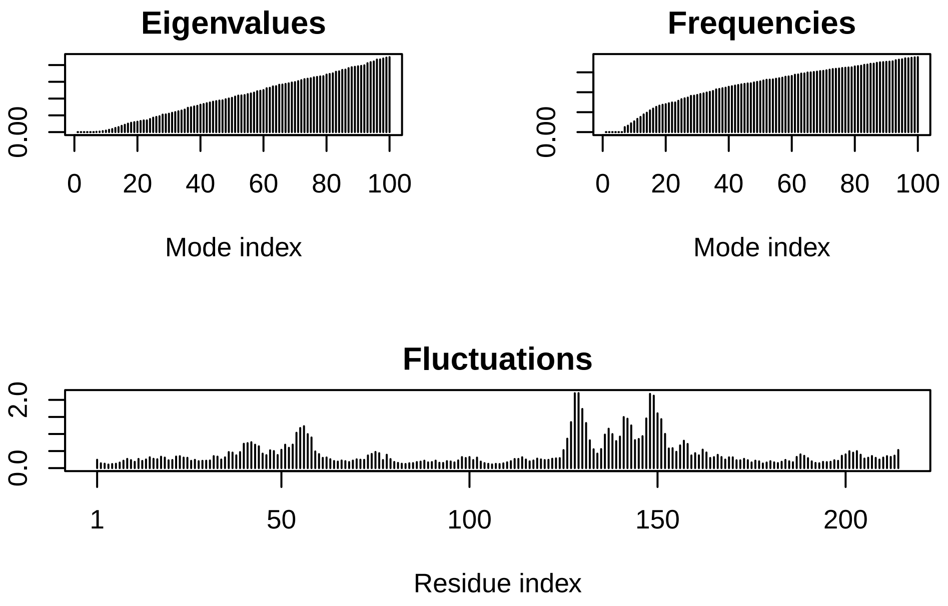
```

 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
```

We can then perform Normal mode analysis (NMA) and plot it with this PDB

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

```
 Building Hessian...        Done in 0.041 seconds.
 Diagonalizing Hessian...   Done in 0.487 seconds.
```

```r
plot(m)
```



We can also view a “movie” of these predicted motions by generating a molecular “trajectory” with `mktrj()`

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

File was opened on Mol* and played

Comparative structure analysis of Adenylate Kinase

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

The package found only on BioConductor and not CRAN is *msa*

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

The package not found on BioConductor or CRAN is *"Grantlab/bio3d-view"*

> 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

```
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?

There are 214 amino acids.

Align and superpose structures

Annotate collected PDB structures

We can annotate structures with `pdb.annotate()`

Principal component analysis

Normal mode analysis [optional]

> Q14. What do you note about this plot? Are the black and colored lines similar or different? Where do you think they differ most and why?

The the black and colored lines seem to have similar shape but the heights (fluctuations) are different. They differ most around residues 30-60 and 120-160.