Class 9: Structural Bioinformatics (Pt 1)

Meg Robinson

2/18/2022

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

Download a CSV file from the PDB site

```
data <- read.csv("dataexportsummary.csv",row.names = 1)
head(data)
```

```
##                            X.ray   NMR   EM Multiple.methods Neutron Other
Total
## Protein (only)          144433 11881 6732              182      70    32 1
63330
## Protein/Oligosaccharide   8543    31 1125                5       0     0
9704
## Protein/NA                7621   274 2165                3       0     0
10063
## Nucleic acid (only)       2396  1399   61                8       2     1
3867
## Other                      150    31    3                0       0     0
184
## Oligosaccharide (only)      11     6    0                1       0     4
22
```

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

```
xray <- round((sum(data$X.ray)/sum(data$Total)*100),2)

em <- round((sum(data$EM)/sum(data$Total)*100),2)

xray
```

```
## [1] 87.17
```

```
em
```

```
## [1] 5.39
```

ANSWER: 87.17% of structures in PDB are solved by X-ray, and 5.39% of structures in PDB are solved by EM

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

```
protein = round(data$Total[1]/sum(data$Total)*100,2)
protein
```

```
## [1] 87.26
```

ANSWER: 87.26% of structures in the PDB 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?

ANSWER: After typing 'HIV' and adding a 'protease' term, I see there are 1225 structures with this match. Note for HIV alone, there are 4486 structures.

Now download the "PDB File" for the HIV-1 protease structure with the PDB identifier 1HSG (1hsg.pdb)

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 are only seeing a single atom which is the oxygen atom in the water molecule. This is because oxygen is large compared to hydrogen, so we would need a bigger resolution in order to view hydrogen as well.

Q5: There is a conserved water molecule in the binding site. Can you identify this water molecule? What residue number does this water molecule have (see note below)?

The conserved water molecule residue number is 308.

3. Introduction to Bio3D in R

```
# install.packages("bio3d")
library(bio3d)
```

```
## Warning: package 'bio3d' was built under R version 4.1.2
```

```
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 this pdb object

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

ANSWER: HOH

Q9: How many protein chains are in this structure?

ANSWER: 2

4. Comparative structure analysis of Adenylate Kinase

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

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

ANSWER: bio3d-view

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

ANSWER: True

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

ANSWER: 214