class09pdf

Katie Chau

2/21/2022

Introduction to the RCSB Protein Data Bank

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

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

```
##            Molecular.Type  X.ray   NMR   EM Multiple.methods Neutron Other
## 1          Protein (only) 144433 11881 6732              182      70    32
## 2 Protein/Oligosaccharide   8543    31 1125                5       0     0
## 3              Protein/NA   7621   274 2165                3       0     0
## 4     Nucleic acid (only)   2396  1399   61                8       2     1
## 5                   Other    150    31    3                0       0     0
## 6  Oligosaccharide (only)     11     6    0                1       0     4
##    Total
## 1 163330
## 2   9704
## 3  10063
## 4   3867
## 5    184
## 6     22
```

```
sumXEM <- sum(data$X.ray)+sum(data$EM)
sumXEM
```

```
## [1] 173240
```

```
sumAll <- sum(data$Total)
sumAll
```

```
## [1] 187170
```

```
percentage <- sumXEM/sumAll
percentage*100
```

```
## [1] 92.55757
```

92.5% of the structures in the PDB are solved by X-ray and Electron Microscopy.

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

```
protein <- data[1,8]
  protein
```

```
## [1] 163330
```

```
  proportion <- protein/sumAll
  proportion*100
```

```
## [1] 87.26292
```

87.3% of the 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?

There are 4486 structures when using a text search of just "HIV" in the PDB website search box. When using the HIV-1 protease sequence and doing a sequence search on the PDB website advanced search, there are 860 structures.

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 see only one atom per water molecule so that there's more noise reduction with the amount of atoms in the structure. We don't need to see all the atoms of this molecule if we already know what the molecular structure of that molecule is we can represent it as one sphere.

> 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)?

HOH 308

> Q6: 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 and the sequence viewer extension can you identify secondary structure elements that are likely to only form in the dimer rather than the monomer?

B-pleated sheet secondary structure of the backbone.

##Introduction to Bio3D in R

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

198

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

HOH

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

2

```r
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>
```

Comparative structure analysis of Adenylate Kinase

> 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?:

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?

214 amino acids

```r
#blast search
#b <- blast.pdb(aa)
```

```r
#hits <- plot(b)

hits <- NULL
hits$pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A','3X2S_A','6HAP_A','6HAM_

#head(hits$pdb.id)
#plot.blast(b)
```

```r
#files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

#Align and superpose structures

```r
#pdbs <- pdbaln(files, fit = TRUE)#, exefile="msa")
```

Could not get muscle installed. I installed the mac arm and intel but didn't know where to go from there.

```r
#vector of PDB codes
#ids <- basename.pdb(pdbs$id)

#schematic alignment
#plot(pdbs, labels=ids)
```