Class09

Morgan Farrell

2/16/2022

Import the data set from the PBD. Download and then add to working directory file. Location of data https://www.rcsb.org/stats/summary

SEt the row names as equal to the first row

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

```
##                           X.ray   NMR   EM Multiple.methods Neutron Other  Total
## Protein (only)           144433 11881 6732               182      70    32 163330
## 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: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.

```
#First try- didn't sum column so not right
((tbl$X.ray + tbl$EM)/ tbl$Total)*100
```

```
## [1] 92.55189 99.62902 97.24734 63.53763 83.15217 50.00000
```

```
#Summing the entire columns
((sum(tbl$X.ray + tbl$EM))/sum(tbl$Total))*100
```

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

```
#Using colSums- best method
n.type<-colSums(tbl)
ans <- round(n.type/n.type["Total"]*100, digits = 3 )
ans
```

```
##           X.ray             NMR              EM Multiple.methods 
##          87.169           7.278           5.389           0.106 
##         Neutron           Other           Total 
##           0.038           0.020         100.000
```

The proportion or percent of Xray structures is 87.169%

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

```
#First way to get the value of protein total in total column
tbl$Total[1]
```

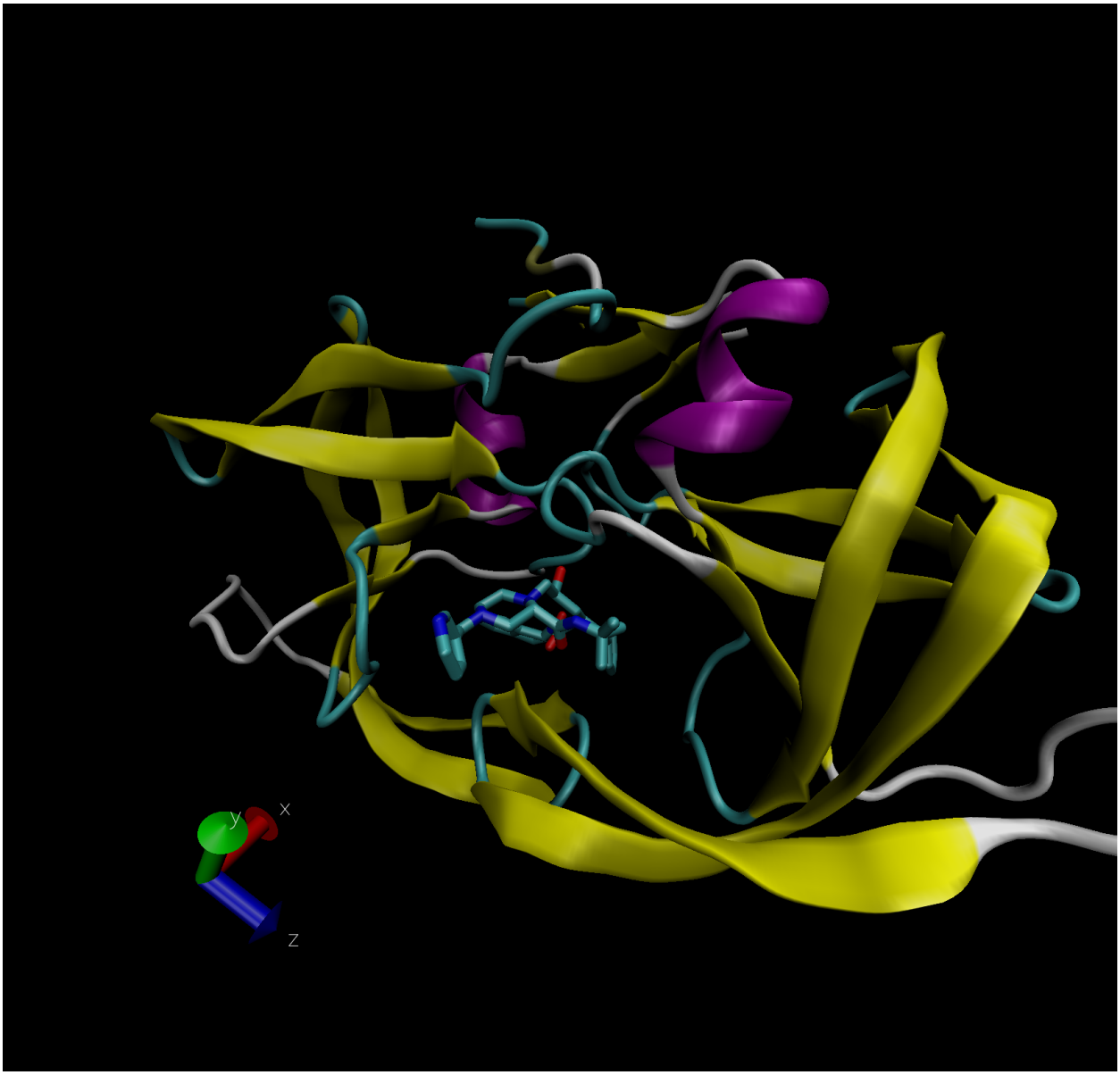
```
## [1] 163330
```

```
#Second method adding everything up
protein.total <- tbl["Protein (only)", "Total"]/n.type["Total"]*100
protein.total
```

```
##    Total
## 87.26292
```

The proportion of structure in the PDB are protein is 87.2629161%

Inserting image file



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?

The search results showed 8427 results

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

The resolution for the PMD is 2 A. Hydrogen is much smaller than this resolution therefore we cannot see it.

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)?

MK 1902

Bio3D 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?

Water H20

Q9: How many protein chains are in this structure?

2

```
#Converts the three letter protein code to one letter code
aa321("GLN")
```

```
## [1] "Q"
```

```
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>
```

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?:

bitbucket

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

TRUE

Comparative analysis of protein structures

Read a single ADK structure from the database

```
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

Let's find related sequences to 'aa' using blast.pdb

```
#blast <- blast.pdb(aa)
```

Plot the blast search to see the graphs of E-values and top hits. Save it to a vector to look at the results more.

```
#hits <- plot(blast)
#hits
```

Show the names of all the hits from the blast search

```
#hits$pdb.id
```

Alpha fold predicted protein for ROMO1