Lab 09: Structural Bioinformatics

Sam Altshuler (PID: A59010373)

2/18/2022

Introduction to RCSB PDB

Download a CSV file from the PDB site. Move this CSV file into your RStudio project and use it to answer the following questions:

```
dat <- read.csv("Data_Export_Summary.csv", row.names = 1)
dat
```

```
##                           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
```

```
total <- apply(dat, 2, sum)
use_total <- as.data.frame(t(total))
dat <- rbind(dat, "Total" = total)
```

- Q1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.
  - 92.56% of the structures in the PDB are from X-Ray and EM.

```
# Number solved by X-Ray and EM
x_em <- use_total$X.ray + use_total$EM
#Percent solved by X-Ray and EM
per_x_em <- round(x_em/use_total$Total*100, 2)
per_x_em
```

```
## [1] 92.56
```

- Q2: What proportion of structures in the PDB are protein?
  - 87.26 percent of the structures in the PBD are purely protein. 97.82% of the structures have some form of protein, where it's protein/oligosaccharide, protein alone, or protein/NA.

```
per_pro_only <- round(dat$Total[1]/use_total$Total *100, 2)
per_pro_only
```

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

```
pro <- grep("Protein", rownames(dat) )
per_pro_total <- round((sum(dat$Total[pro]))/use_total$Total *100, 2)
per_pro_total
```

```
## [1] 97.82
```

- 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?
  - 4343 protein structures structures are in the current PDB. When typing in "HIV" into the search bar, 4486 structures show up. Of these roughly 4500 structures, 4343 are identified as proteins (or at least made from amino acids).
- Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?
  - The current structure doesn't show any of the hydrogens. So all water is only seen as the oxygen atom. This is because the chosen resolution is too large to portray hydrogen atoms
- 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)?
  - H2O 308 appears to be the conserved water found between the ligand and the protein.
- 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?
  - I believe the beta sheet's at the c-terminal end of the monomers likely only forms in the dimer as this is the clearest area where direct interactions between the two monomers seems to occur.

Intro to Bio3D

```
library(bio3d)

#Reading PDB file into R
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?
  - There are 198 residues in this pdb object.
- Q8: Name one of the two non-protein residues?
  - One of the two non-protein residues is MK1
- Q9: How many protein chains are in this structure?
  - There are 2 protein chains in this structure.

Look at the attributes to see what specific data can be pulled from this object

```
attributes(pdb)
```

```
## $names
## [1] "atom"   "xyz"    "seqres" "helix"  "sheet"  "calpha" "remark" "call"
##
## $class
## [1] "pdb" "sse"
```

See what pops up when you look at the atom attribute.

```
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>
```