Project09

Haroon Riyaz (PID A15377799)

2/19/2022

Data from CSV file

```
ExpMetMol <- read.csv("DataExportSummary.csv", row.names = 1)
ExpMetMol
```

```
##                           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.

```
PercXrayEM_Results <- 100*((ExpMetMol$X.ray + ExpMetMol$EM)/ExpMetMol$Total)
PercXrayEM_Names <- row.names(ExpMetMol)
PercXrayEM <- data.frame(PercXrayEM_Names, PercXrayEM_Results)
PercXrayEM
```

```
##          PercXrayEM_Names PercXrayEM_Results
## 1          Protein (only)           92.55189
## 2 Protein/Oligosaccharide           99.62902
## 3              Protein/NA           97.24734
## 4     Nucleic acid (only)           63.53763
## 5                   Other           83.15217
## 6  Oligosaccharide (only)           50.00000
```

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

```
ProteinPerc <- ExpMetMol$Total
ProteinPerc[1]/sum(ProteinPerc)
```

```
## [1] 0.8726292
```

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 *860* HIV-1 protease structures in the current PDB

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

The program may not display all 3 atoms as the orientation of water molecules is constantly changing. The program could also be displaying 1 atom per water molecule to reduce visual clutter.

> 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 near the binding site is: *HOH308:0*

> 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?

The beta-pleated sheets formed between the two chains (containing LEU 97 and ASN 98) likely formed as a result of the dimer as the residues on each strand stabilize the other strand to form a sheet.

3. Introduction to Bio3D in R

Bio3D is used for structural bioinformatics!

Load Bio3D package

```
library(bio3d)
```

Read and inspect 1HSG PDB file

```
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 amino acid residues

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

HOH

Q9: How many protein chains are in this structure?

2 protein chains

```
Access Attributes

'''r
attributes(pdb)
'''

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

'''r
#Access specific 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>
```

```
# 4. 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?:

bio3d-view

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

TRUE
```

A BLAST search will be performed to identify structures related to chain A of ADK.

```
library(bio3d)

# Query sequence obtained with 'get.seq()'

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

The BLAST search can now be performed

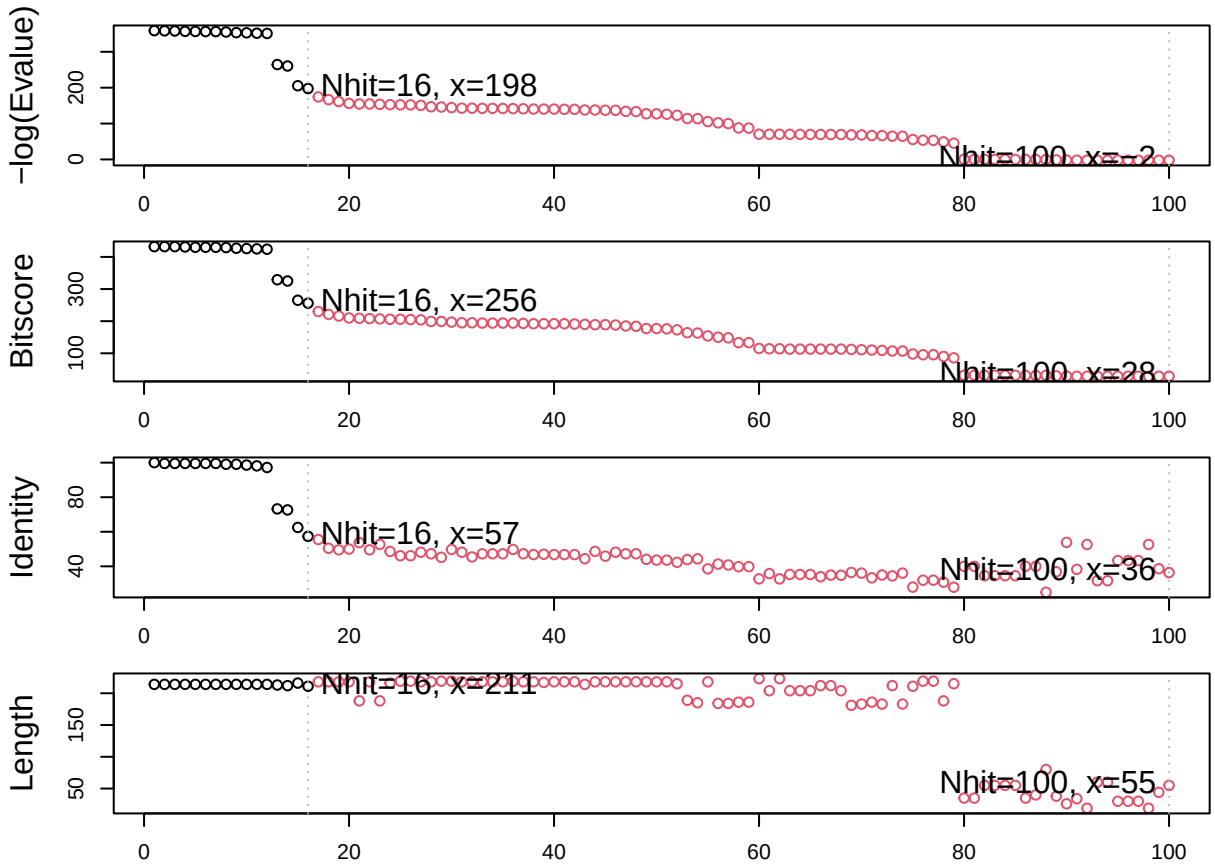
```
# Blast search
b <- blast.pdb(aa)
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 169HTGTP013
##  ..................
##  Reporting 100 hits
```

Visaulize top hits from BLAST

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

```
##   * Possible cutoff values:    197 -3
##              Yielding Nhits:    16 100
##
##   * Chosen cutoff value of:    197
##              Yielding Nhits:    16
```



Listing the top his only

```
# List the 'top hits'
head(hits$pdb.id)
```

```
## [1] "1AKE_A" "4X8M_A" "6S36_A" "6RZE_A" "4X8H_A" "3HPR_A"
```

Not all results were returned so vector will be used

```
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
```

PDB files will be downloaded now

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

```
## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 1AKE.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 6S36.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 6RZE.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 3HPR.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 1E4V.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 5EJE.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 1E4Y.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 3X2S.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 6HAP.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 6HAM.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 4K46.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 3GMT.pdb exists. Skipping download

## Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
## 4PZL.pdb exists. Skipping download

##   |                                                                              |
```

Align and Superpose Structures

FOLLOWING STEPS SKIPPED SINCE MUSCLE.EXE COULD NOT BE DOWNLOADED

```
# Align releated PDBs
#pdbs <- pdbaln(files, fit = TRUE)
```

```
# Vector for axis
#ids <- basename.pdb(pdbs$id)

# Draw Alignment
#plot(pdbs, labels=ids)
```

Principal Component Analysis

Describe variance in data (SKIPPED)

```
# PCA performed
#pc.xray <- pca(pdbs)
#plot(pc.xray)
```

> 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?

They're different from one another as the colored and black lines have a different number of fluctuations at different residue numbers; the colored lines have more fluctuations for nearly all residues (except ~ residue 75). The lines differ the most at around residue number 125 as the fluctuations differ the most, indicating a different flexibility.