Class 9: Structural Bioinformatics pt. 1

Taylor F. (A59010460)

2/16/2022

Taking a Quick Look at PDB

The PDB is the main repository of biomolecular structure data.

Here we grab the current composition statistics from the web page: https://www.rcsb.org/stats/summary

```
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
```

Question 1: What percentage of structures in the PDB are solved by X-Ray and Electron Microscopy.

```
#Check the sums of all the columns in the data set
#colSums(tbl)

#Sum the relevant columns and divide that number by the sum of the "total" column, multiplying the answ
n.type <- colSums(tbl)
n.type / n.type["Total"] * 100
```

```
##            X.ray              NMR               EM Multiple.methods
##      87.16888390       7.27787573       5.38868408       0.10632046
##          Neutron            Other            Total
##       0.03846770       0.01976813     100.00000000
```

```
#If we were to use the above method the generate the answer for the question, we would want store n.typ

#The less elegant way I came up with
XR <- sum(tbl[,1]) / sum(tbl[,7]) * 100
EM <- sum(tbl[,3]) / sum(tbl[,7]) * 100
XR
```

```
## [1] 87.16888
```

```
EM
```

```
## [1] 5.388684
```

```
#How do we get an output with only 3 decimal places?
XRr <- round(XR, digits = 3)
XRr
```

```
## [1] 87.169
```

```
EMr <- round(EM, digits = 3)
EMr
```

```
## [1] 5.389
```

The proportion of of X-ray structures is 87.169% of the total structures

The proportion of of EM structures is 5.389% of the total structures

Question 2: What proportion of structures in the PDB are protein?

```
#Take the total number of protein entries (located in row 1, column 7) and divide it by the sum of the
Prot <- round(tbl[1,7] / sum(tbl[,7]) * 100, digits = 3)
Prot
```

```
## [1] 87.263
```

```
#Barry's more elegant solution
#tbl$Total[1]
#This allows you to not have to know the column number, just the name, and you can still specify the ro
#This also protects you from issues if the database changes at all, still searching for the 'Total' col
```

The proportion of entries that are protein structures is 87.263%

Question 3: 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?

Using the HIV query, and the protease subquery, I found 1225 structures

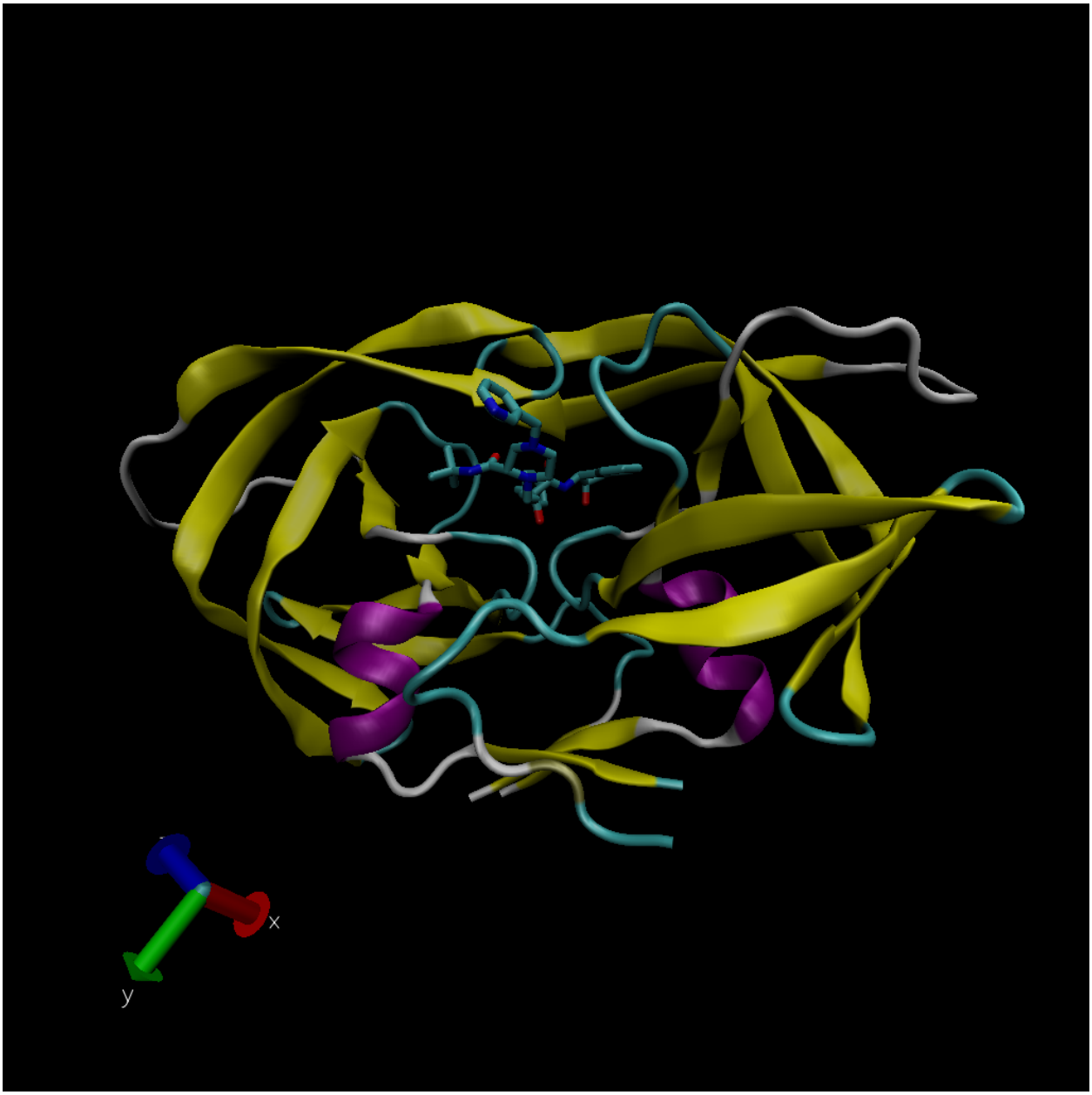
Question 4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

We can see the oxygen, but not the hydrogens, because the resolution of the structure is 2 angstroms, which is larger than the size of hydrogen but smaller than the size of oxygen.

Question 5: 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 water is at residue number 308

Inserting an Image File



Bio3D for Structural Bioinformatics

```
library(bio3d)

#We need to access the online PDB file for 1HSG
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
```

```
#Looking at the atom column of the pdb dataset
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>
```

Q7: How many amino acid residues are there in this pdb object?

198 amino acids

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

Water (also the small molecule)

Q9: How many protein chains are in this structure?

There are two chains

Comparative Analysis of Protein Structures

Read a single ADK structure from the database

```
# Install packages in the R console not your Rmd

#install.packages("bio3d")
#install.packages("ggplot2")
#install.packages("ggrepel")
#install.packages("devtools")
#install.packages("BiocManager")
#BiocManager::install("msa")
#devtools::install_bitbucket("Grantlab/bio3d-view")

#Be sure to load in all of the packages necessary for this
library("bio3d")
library("ggplot2")
library("ggrepel")
library("devtools")
```

```
## Loading required package: usethis
```

```
library("BiocManager")
```

```
##
## Attaching package: 'BiocManager'

## The following object is masked from 'package:devtools':
##
##     install
```

```
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
```

Question 10: Which of the packages above is found only on BioConductor and not CRAN?

The 'msa' package

Question 11: Which of the above packages is not found on BioConductor or CRAN?:

The 'bio3d-view' package

Question 12: True or False? Functions from the devtools package can be used to install packages from GitHub and BitBucket?

True

Question 13: How many amino acids are in this sequence, i.e. how long is this sequence?

This sequence contains 214 amino acid residues

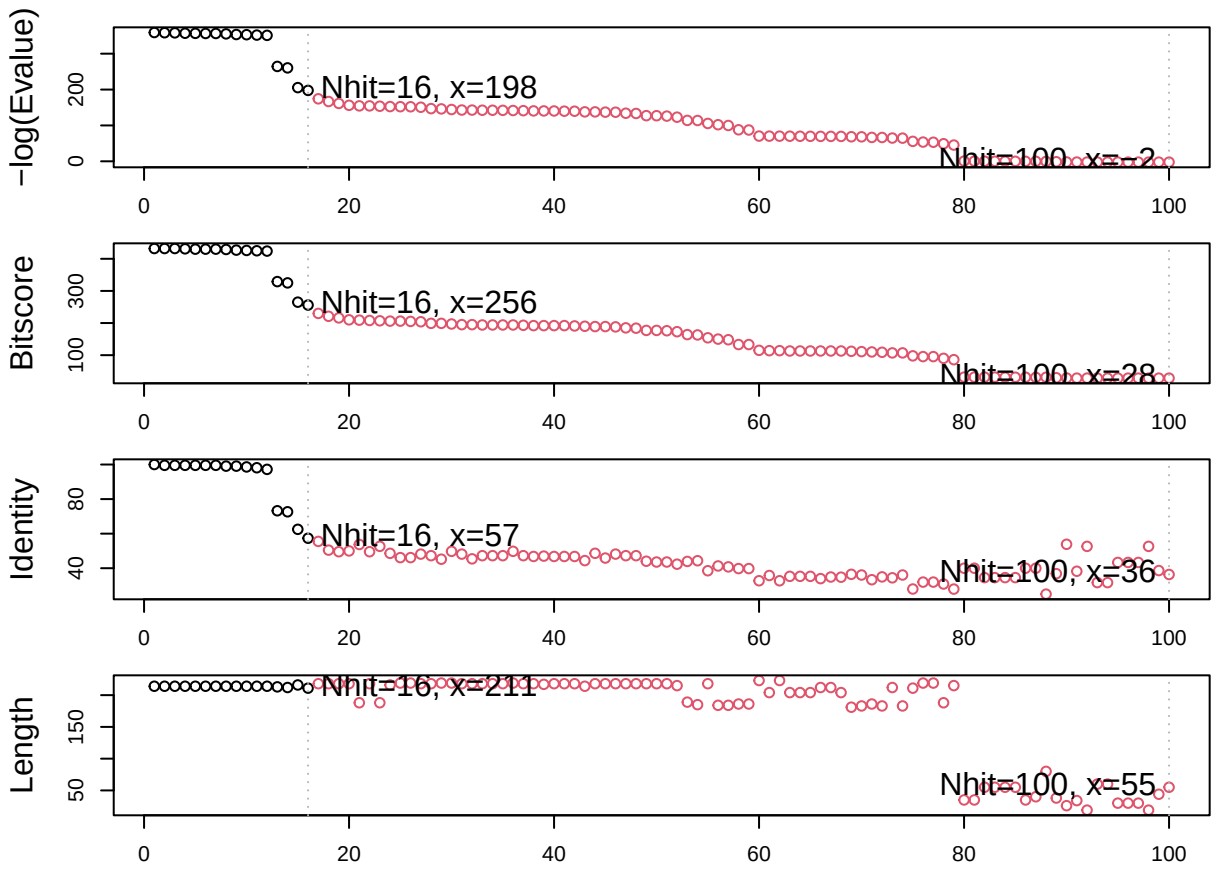
Now let's blast for related sequences

```
#Blast search for matches to 'aa' sequence
blast <- blast.pdb(aa)
```

```
##  Searching ... please wait (updates every 5 seconds) RID = 0V46EPWK013
##  ..............................
##  Reporting 100 hits
```

```
#Plotting results of the blast
hits <- plot(blast)
```

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

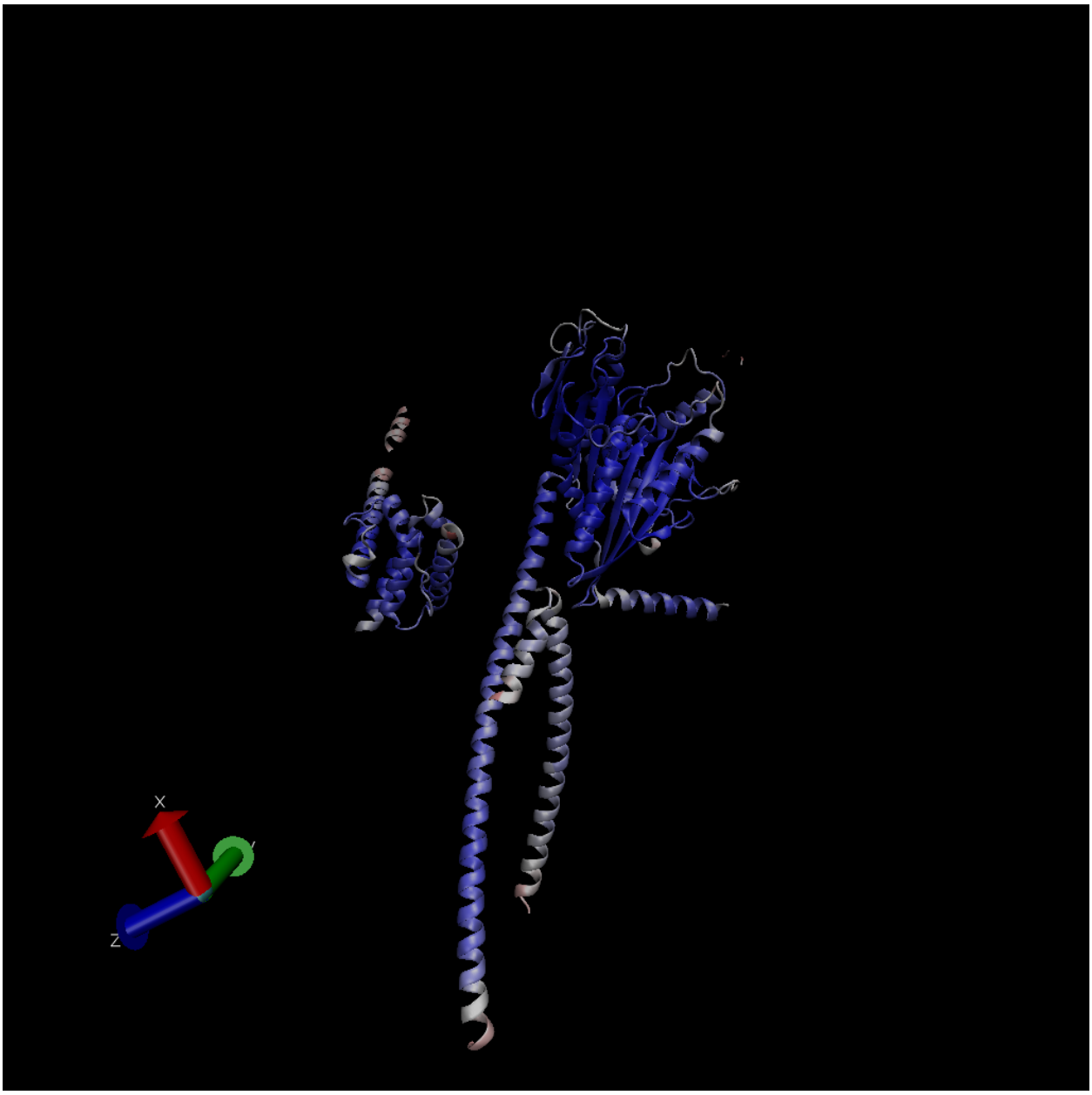



```
#We should look at some of the top hits
hits$pdb.id
```

```
##  [1] "1AKE_A" "4X8M_A" "6S36_A" "6RZE_A" "4X8H_A" "3HPR_A" "1E4V_A" "5EJE_A"
##  [9] "1E4Y_A" "3X2S_A" "6HAP_A" "6HAM_A" "4K46_A" "4NP6_A" "3GMT_A" "4PZL_A"
```

Using AlphaFold

Here is an image of a new structure based on sequence


X
Z