Class09 R

Katelyn Brown PID 15891811

2/15/2022

The PDB database

The PDB is the main repository for 3D structure data of biomolecules.

Here we explore its composition.

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

```
#Downloaded CSV file
x <- read.csv("Data Export Summary.csv", row.names = 1)
x
```

```
##                           X.ray   NMR   EM Multiple.methods Neutron Other  Total
## Protein (only)           144301 11877 6676              182      70    32 163138
## Protein/Oligosaccharide    8528    31 1116                5       0     0   9680
## Protein/NA                 7617   274 2153                3       0     0  10047
## Nucleic acid (only)        2393  1398   61                8       2     1   3863
## Other                       150    31    3                0       0     0    184
## Oligosaccharide (only)       11     6    0                1       0     4     22
```

```
# Find percentage of structures solved by X-Ray and EM
column.sums <- colSums(x)
column.sums
```

```
##            X.ray              NMR               EM Multiple.methods
##           163000            13617            10009              199
##          Neutron            Other            Total
##               72               37           186934
```

```
round(column.sums/column.sums["Total"]*100, 3)
```

```
##            X.ray              NMR               EM Multiple.methods
##           87.197            7.284            5.354            0.106
##          Neutron            Other            Total
##            0.039            0.020          100.000
```

87.197%% of the structures in PDB are solved by X-Ray, and 5.354% by Electron Microscopy.

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

```
column.sums["Total"]
```

```
##  Total
## 186934
```

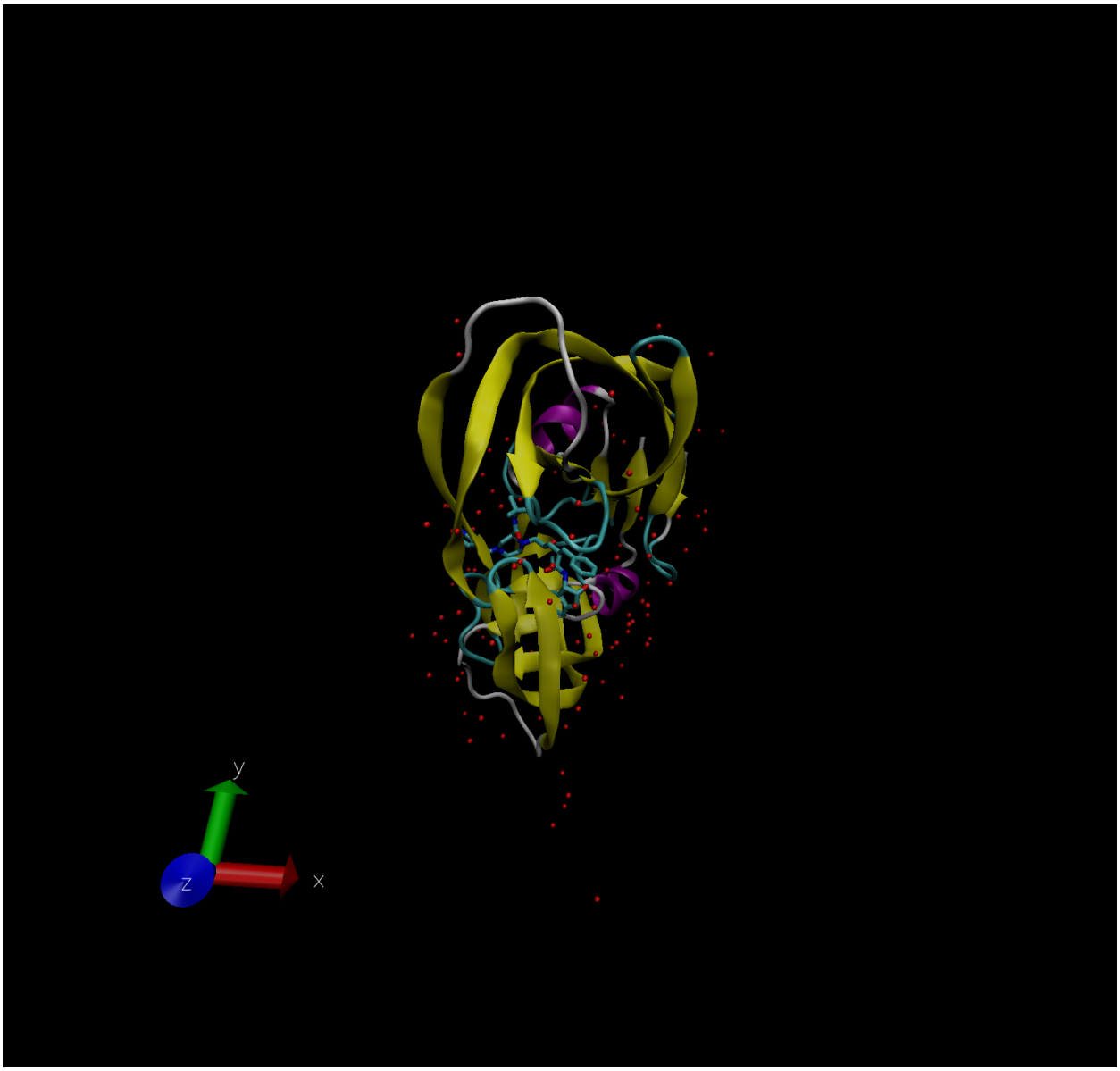
```
proteins <- x["Protein (only)","Total"]
round((proteins/column.sums["Total"])*100, 3)
```

```
## Total
## 87.27
```

87.27% 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?

Here is a VMD generated image of HIV-protease, PDB code: 1HSG



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

We see just one atom per water molecule because this atom is oxygen, and the VMD program does not have a high enough resolution to view hydrogen, which is very small.

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)?

This water molecule is residue number 308.

Bio3D package for structural bioinformatics

Load the Bio3D package.

```
# Installed Bio3D package
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
```

```
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>
```

Extract the sequence for ADK:

```
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
```

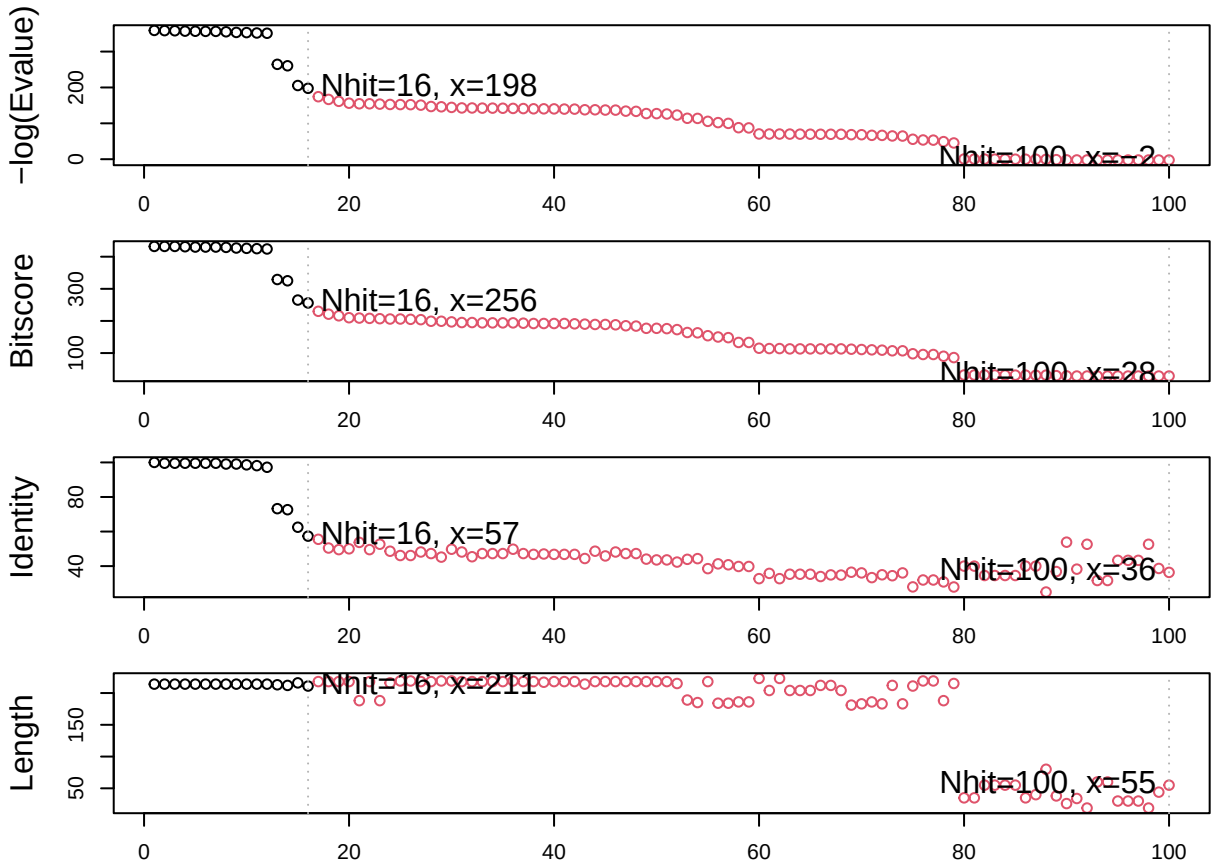
```
##              1        .         .         .         .         .         60
## pdb|1AKE|A   MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
##              1        .         .         .         .         .         60
##
##             61        .         .         .         .         .         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
```

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

```
##  Searching ... please wait (updates every 5 seconds) RID = 0WWYESD8016
##  ..............................
##  Reporting 100 hits
```

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

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




```
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"
```

Normal mode analysis (NMA)

```
pdb <- read.pdb("1ake")
```

```
##   Note: Accessing on-line PDB file
##    PDB has ALT records, taking A only, rm.alt=TRUE
```

```
pdb
```

```
##
##  Call:  read.pdb(file = "1ake")
##
##    Total Models#: 1
##      Total Atoms#: 3804,  XYZs#: 11412  Chains#: 2  (values: A B)
##
##      Protein Atoms#: 3312  (residues/Calpha atoms#: 428)
##      Nucleic acid Atoms#: 0  (residues/phosphate atoms#: 0)
##
##      Non-protein/nucleic Atoms#: 492  (residues: 380)
##      Non-protein/nucleic resid values: [ AP5 (2), HOH (378) ]
##
##    Protein sequence:
##       MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
##       DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
##       VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
##       YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILGMRIILLGAPGA...<cut>...KILG
##
## + attr: atom, xyz, seqres, helix, sheet,
##         calpha, remark, call
```

Trim to chain A only.

```
chain <- trim.pdb(pdb, chain = "A")
chain
```

```
##
##  Call:  trim.pdb(pdb = pdb, chain = "A")
##
##    Total Models#: 1
##      Total Atoms#: 1954,  XYZs#: 5862  Chains#: 1  (values: A)
##
##      Protein Atoms#: 1656  (residues/Calpha atoms#: 214)
##      Nucleic acid Atoms#: 0  (residues/phosphate atoms#: 0)
##
##      Non-protein/nucleic Atoms#: 298  (residues: 242)
##      Non-protein/nucleic resid values: [ AP5 (1), HOH (241) ]
##
##    Protein sequence:
##       MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
##       DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
##       VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
##       YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG
##
## + attr: atom, helix, sheet, seqres, xyz,
##         calpha, call
```

Run a bioinformatics method to predict the flexibility and "functional motions" of this protein chain.

```
modes <- nma(chain)
```

```
##  Building Hessian...     Done in 0.071 seconds.
##  Diagonalizing Hessian...    Done in 0.423 seconds.
```

```
mktrj.nma(modes, mode=7, file="mode_7.pdb")
```

Align and superimpose structures.

AlphaFold rendering of Find a Gene Protein