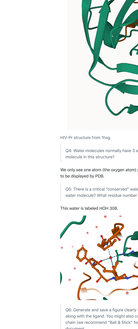
Class 9 Lab

AUTHOR
Catherine Diep

The RCSB Protein Data Bank (PDB)

Protein structures by X-ray crystallography dominate this database. We are skipping Q1-3 as the website was too slow for us.

2. Visualizing the HIV-1 protease structure



HIV-Pr structure from 1hsg

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

We only see one atom (the oxygen atom) per molecule because the hydrogen atoms are too small to be displayed by PDB.

Q5: There is a critical "conserved" water molecule in the binding site. Can you identify this water molecule? What residue number does this water molecule have?

This water is labeled HOH 308.

Q6: Generate and save a figure clearly showing the two distinct chains of HIV-protease along with the ligand. You might also consider showing the catalytic residues ASP 25 in each chain (we recommend "Ball & Stick" for these side-chains). Add this figure to your Quarto document.

1hsg with marked Asp

Discussion Topic: Can you think of a way in which indinavir, or even larger ligands and substrates, could enter the binding site? Larger ligands could enter the binding site when the HIV-protease's two chains move apart to expose the site.

3. Introduction to Bio3D in R.

Bio3D is an R package for structural bioinformatics. To use it we need to call it up with the 'library()' function.

```
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  Chain#: 2  Residues: # 0)

    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 sequences:
     PQITLWQRPLVTIKIGGQLKEALLDTGADDTVLEDMSLPGRWKPKMIGGIGGFIKVRQYD
     QILIEICGHKAIGTVLVGPTPVNIIGRNLLTQIGCTLNFPQITLWQRPLVTIKIGGQLKE
     ALLDTGADDTVLEDMSLPGRWKPKMIGGIGGFIKVRQYDQILIEICGHKAIGTVLVGPTP
     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? HOH

Q9: How many protein chains are in this structure? 2

The ATOM records of a PDB file are stored in 'pdb$atom':

```
head(pdb$atom)
```

```
  type elesy 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 37.543 6.342 1 37.87
6 ATOM    6   CG  <NA>   PRO     A    1   <NA> 29.296 37.191 7.162 1 38.40
  segid 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>
```

Comparative analysis of Adenylate kinase (ADK)

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 and BitBucket? TRUE

We will start our analysis with a single PDB id (code from the PDB database): 1AKE

First we get its primary sequence:

```
aa <- get.seq("1ake_A")
```

```
Warning in get.seq("1ake_A"): Removing existing file: seq.fasta
```

```
Fetching... Please wait. Done.
```

```
aa
```

```
               1                                                  49
pdb|1AKE|A     MRILILGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
               1          .         .         .         .         .  60

              50                                                  99
pdb|1AKE|A     DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDR
              61          .         .         .         .         . 120

             100                                                 149
pdb|1AKE|A     IVGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPL
             121          .         .         .         .         . 180

             150                                                 214
pdb|1AKE|A     IDFYRKEAKAGNTKYAKVDGTKPVAEVRADLEKILG
             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

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

#hits <- plot(b)
#List out some 'top hits'
#head(hits$pdb.id)
```

Use these ADK structures for analysis:

```
hits <- NULL
hits$pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A',
```

Download all these PDB files from the database...

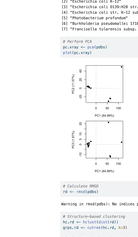
```
files <- get.pdb(hits$pdb.id, path="pdbs", split=TRUE, gzip=TRUE)
```

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

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

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

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

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

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

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

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

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

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

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

```
  |                                                                      |   0%
  |=                                                                     |   1%
  |==                                                                    |   3%
  |===                                                                   |   4%
  |====                                                                  |   6%
  |=====                                                                 |   7%
  |======                                                                |   9%
  |=======                                                               |  10%
  |========                                                              |  12%
  |=========                                                             |  13%
  |==========                                                            |  15%
  |===========                                                           |  16%
  |============                                                          |  18%
  |=============                                                         |  19%
  |==============                                                        |  21%
  |===============                                                       |  22%
  |================                                                      |  24%
  |=================                                                     |  25%
  |==================                                                    |  27%
  |===================                                                   |  28%
  |====================                                                  |  30%
  |=====================                                                 |  31%
  |======================                                                |  33%
  |=======================                                               |  34%
  |========================                                              |  36%
  |=========================                                             |  37%
  |==========================                                            |  39%
  |===========================                                           |  40%
  |============================                                          |  42%
  |=============================                                         |  43%
  |==============================                                        |  45%
  |===============================                                       |  46%
  |================================                                      |  48%
  |=================================                                     |  49%
  |==================================                                    |  51%
  |===================================                                   |  52%
  |====================================                                  |  54%
  |=====================================                                 |  55%
  |======================================                                |  57%
  |=======================================                               |  58%
  |========================================                              |  60%
  |=========================================                             |  61%
  |==========================================                            |  63%
  |===========================================                           |  64%
  |============================================                          |  66%
  |=============================================                         |  67%
  |==============================================                        |  69%
  |===============================================                       |  70%
  |================================================                      |  72%
  |=================================================                     |  73%
  |==================================================                    |  75%
  |===================================================                   |  76%
  |====================================================                  |  78%
  |=====================================================                 |  79%
  |======================================================                |  81%
  |=======================================================               |  82%
  |========================================================              |  84%
  |=========================================================             |  85%
  |==========================================================            |  87%
  |===========================================================           |  88%
  |============================================================          |  90%
  |=============================================================         |  91%
  |==============================================================        |  93%
  |===============================================================       |  94%
  |================================================================      |  96%
  |=================================================================     |  97%
  |==================================================================    |  99%
  |======================================================================| 100%
```

```
# Align related PDBs
pdbs <- pdbaln(files, fit = TRUE, exefile="muscle")
```

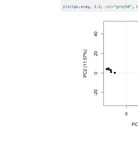
```
Reading PDB files:
pdbs/split_chain/1AKE_A.pdb
pdbs/split_chain/6S36_A.pdb
pdbs/split_chain/6RZE_A.pdb
pdbs/split_chain/3HPR_A.pdb
pdbs/split_chain/1E4V_A.pdb
pdbs/split_chain/5EJE_A.pdb
pdbs/split_chain/1E4Y_A.pdb
pdbs/split_chain/1DCH_A.pdb
pdbs/split_chain/4AKE_A.pdb
pdbs/split_chain/1ANK_A.pdb
pdbs/split_chain/4PZL_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
.  PDB has ALT records, taking A only, rm.alt=TRUE
.  PDB has ALT records, taking A only, rm.alt=TRUE
.  PDB has ALT records, taking A only, rm.alt=TRUE
..  PDB has ALT records, taking A only, rm.alt=TRUE
.... PDB has ALT records, taking A only, rm.alt=TRUE
.  PDB has ALT records, taking A only, rm.alt=TRUE
...

Extracting sequences

pdb/seq: 1   name: pdbs/split_chain/1AKE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 2   name: pdbs/split_chain/6S36_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 3   name: pdbs/split_chain/6RZE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 4   name: pdbs/split_chain/3HPR_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 5   name: pdbs/split_chain/1E4V_A.pdb
pdb/seq: 6   name: pdbs/split_chain/5EJE_A.pdb
pdb/seq: 7   name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 8   name: pdbs/split_chain/1DCH_A.pdb
pdb/seq: 9   name: pdbs/split_chain/4AKE_A.pdb
pdb/seq: 10   name: pdbs/split_chain/1ANK_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 11   name: pdbs/split_chain/4PZL_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE

# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

# Draw schematic alignment
plot(pdbs, labels=ids)
```




```
anno <- pdb.annotate(ids)
unique(anno$source)
```

```
[1] "Escherichia coli"
[2] "Escherichia coli K-12"
[3] "Escherichia coli O139:H28 str. E24377A"
[4] "Escherichia coli str. K-12 substr. MG1a"
[5] "Photobacterium profundum"
[6] "Escherichia coli UMN026 ST38b"
[7] "Francisella tularensis subsp. tularensis SCHU S4"
```

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




```
# Calculate RMSD
rd <- rmsd(pdbs)
```

```
Warning in rmsd(pdbs): No indices provided, using the 204 non NA positions
```

```
# Structure-based clustering
hc.rd <- hclust(dist(rd))
grps.rd <- cutree(hc.rd, k=3)

plot(pc.xray, 1:2, col="grey50", hc=hc.rd, grps=grps.rd, pch=20, cex=2)
```