

class09

Vina Nguyen A16151678

2023-05-08

```
library(dplyr)
```

```
##  
## Attaching package: 'dplyr'  
  
## The following objects are masked from 'package:stats':  
##  
##   filter, lag  
  
## The following objects are masked from 'package:base':  
##  
##   intersect, setdiff, setequal, union
```

1: Introduction to the RCSB Protein Data Bank (PDB)

PDB Statistics

```
EMMT.df <- read.csv("Data Export Summary.csv", row.names = 1)
```

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

92.99%

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

0.8681 - Protein(only) 0.9782 - Protein included

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?

4,926 Structures

2. Visualizing the HIV-1 Protease Structure

Using Mol*

Mol* homepage at: <https://molstar.org/viewer/> To load a structure from the PDB we can enter the PDB code and click “Apply” in the “Download Structure” menu.

Getting to Know HIV-Pr

Let's temporarily toggle OFF/ON the display of water molecules and change the display representation of the Ligand to Spacefill. Three dots for "Ligand" > Add Representation > Spacefill

Let's also change the protein "Polymer" > "Set Coloring" > "Residue Property" > "Secondary Structure".

Saving an Image

Save high-resolution image to computer: find "iris-like" screenshot icon on right side of display region and select resolution and click download.

```
# insert an image  
knitr::include_graphics("1HSG(1).png.png")
```



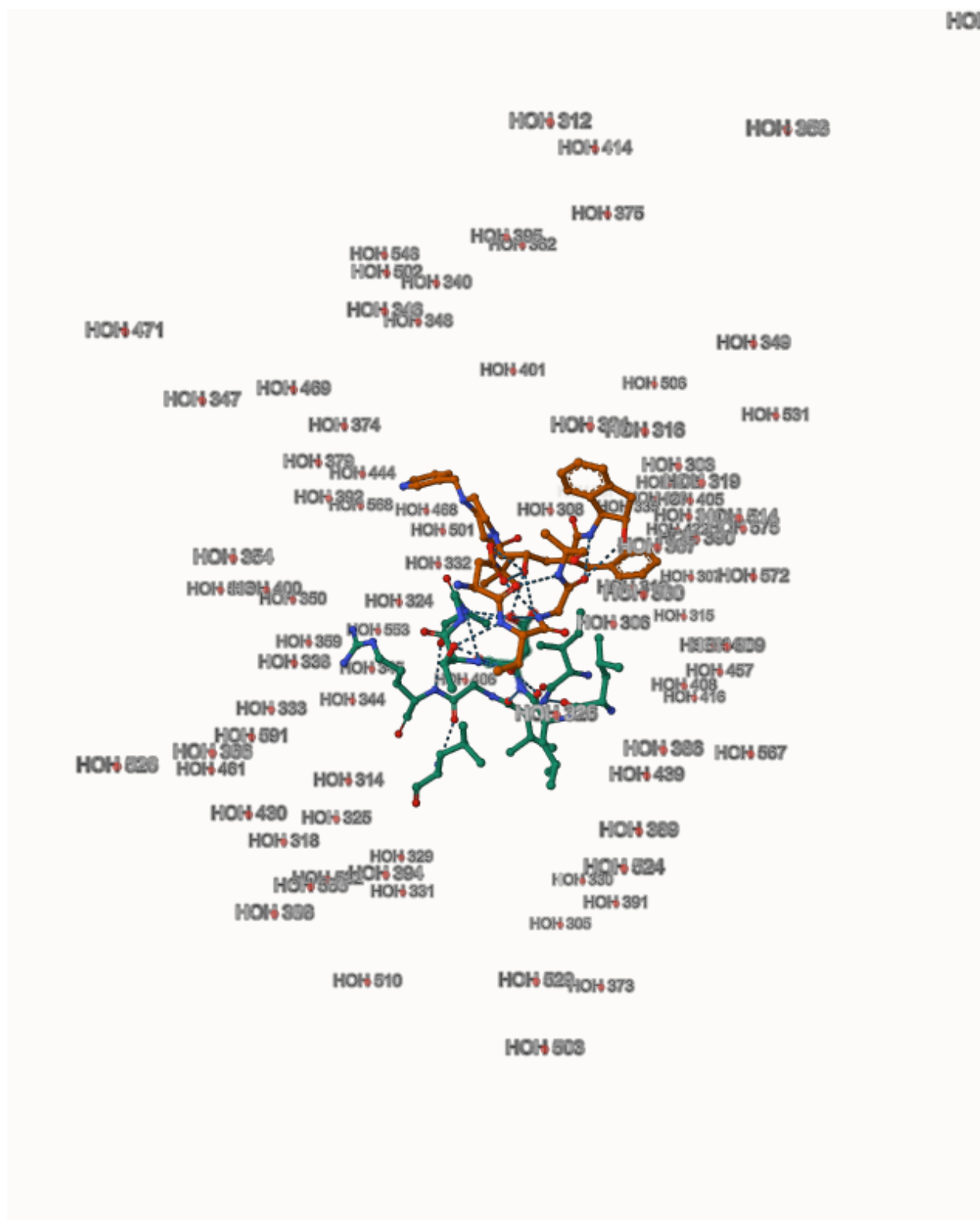
The Important Role of Water

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

When label representation is on, we can see that the one atom per water molecule is named HOH. The red spheres is O.

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

```
knitr::include_graphics("1HSGWater.png")
```

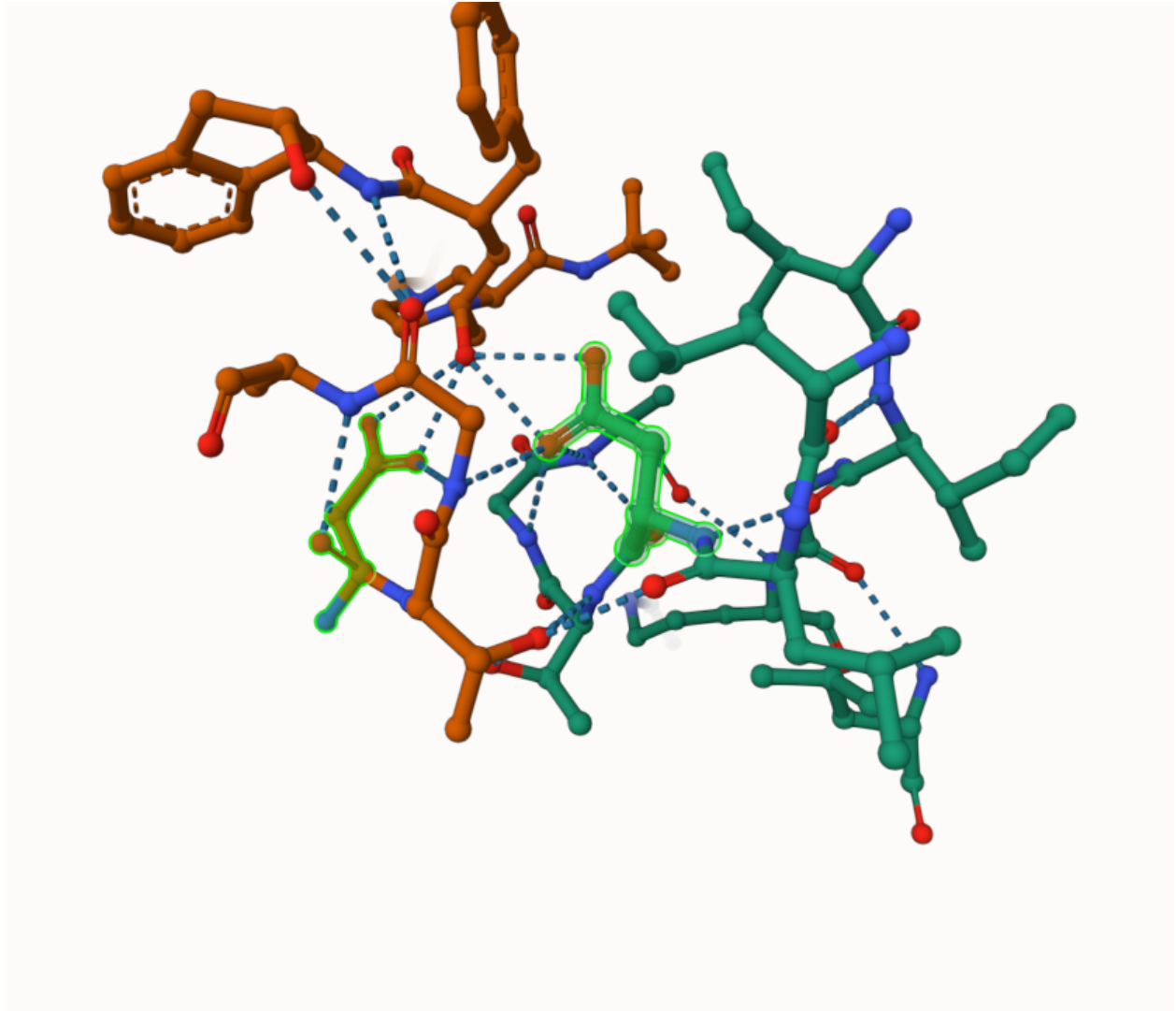


HOH 308

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 and the critical water (we recommend “Ball & Stick” for these side-chains). Add this figure to

your Quarto document.

```
knitr::include_graphics("1HSG.png")
```



The image above shows two Asp 25 positions (B and A) in the 3D structure of 1HSG.

3.Introduction to Bio3D in R

```
# load Bio3D package  
library(bio3d)
```

Reading PDB File Data into R

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

```
## Note: Accessing on-line PDB file
```

```
# quick summary of contents of pdb object  
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

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

HOH (127), MK1 (1)

Q9: How many protein chains are in this structure?

2

Note that the attributes (+ attr:) of this object are listed on the last couple of lines. To find the attributes of any such object you can use:

```
attributes(pdb)
```

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

To access these individual attributes we use the dollar-attribute name convention that is common with R list objects. For example, to access the atom attribute or component use `pdb$atom`:

```
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>
```

Predicting functional motions of a single structure

Let's read a new PDB structure of Adenylate Kinase and perform Normal mode analysis.

```
adk <- read.pdb("6s36")
```

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

```
adk
```

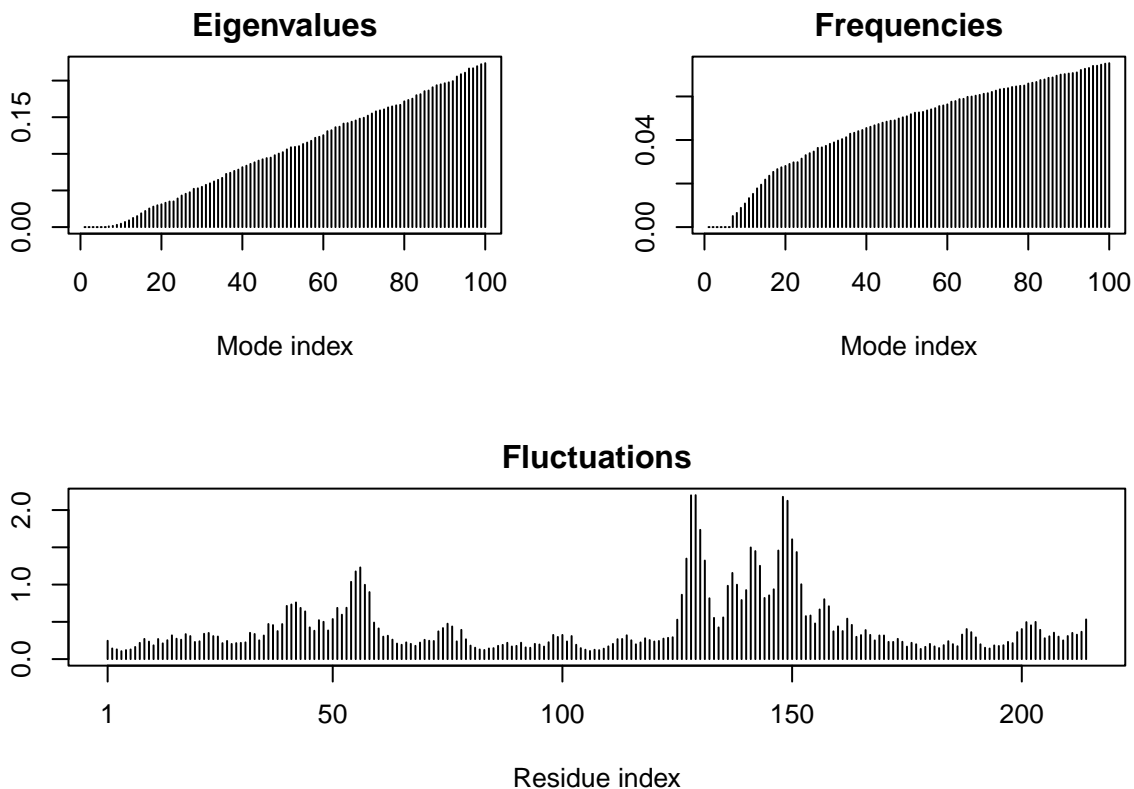
```
##
## Call: read.pdb(file = "6s36")
##
## Total Models#: 1
## Total Atoms#: 1898, XYZs#: 5694 Chains#: 1 (values: A)
##
## Protein Atoms#: 1654 (residues/Calpha atoms#: 214)
## Nucleic acid Atoms#: 0 (residues/phosphate atoms#: 0)
##
## Non-protein/nucleic Atoms#: 244 (residues: 244)
## Non-protein/nucleic resid values: [ CL (3), HOH (238), MG (2), NA (1) ]
##
## Protein sequence:
## MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMRLRAAVKSGSELGKQAKDIMDAGKLV
## DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDKI
## VGRRVHAPSGRVRVHVKNPPKVEGKDDVTGEELTRKDDQEETVRKRLVEYHQMTAPLIG
## YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG
##
## + attr: atom, xyz, seqres, helix, sheet,
## calpha, remark, call
```



```
# Perform flexibility prediction
m <- nma(adk)
```

```
## Building Hessian... Done in 0.03 seconds.
## Diagonalizing Hessian... Done in 0.2 seconds.
```

```
plot(m)
```



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?:

devtools::install_bitbucket("Grantlab/bio3d-view")

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

TRUE

Search and retrieve ADK structures

```
library(bio3d)
aa <- get.seq("lake_A")
```

```
## Warning in get.seq("lake_A"): Removing existing file: seqs.fasta
```

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

```
aa
```

```
##           1           .           .           .           .           .           60
## pdb|1AKE|A  MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLV
##           1           .           .           .           .           .           60
##
##           61           .           .           .           .           .           120
## pdb|1AKE|A  DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDRI
##           61           .           .           .           .           .           120
##
##           121          .           .           .           .           .           180
## pdb|1AKE|A  VGRRVHAPSGRVYHVKNPPKVEGKDDVTGEELTRKDDQEETVRKRLVEYHQMTAPLIG
##           121          .           .           .           .           .           180
##
##           181          .           .           .           214
## pdb|1AKE|A  YYSKEAEAGNTKYAKVDGTPVAEVRADLEKILG
##           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?

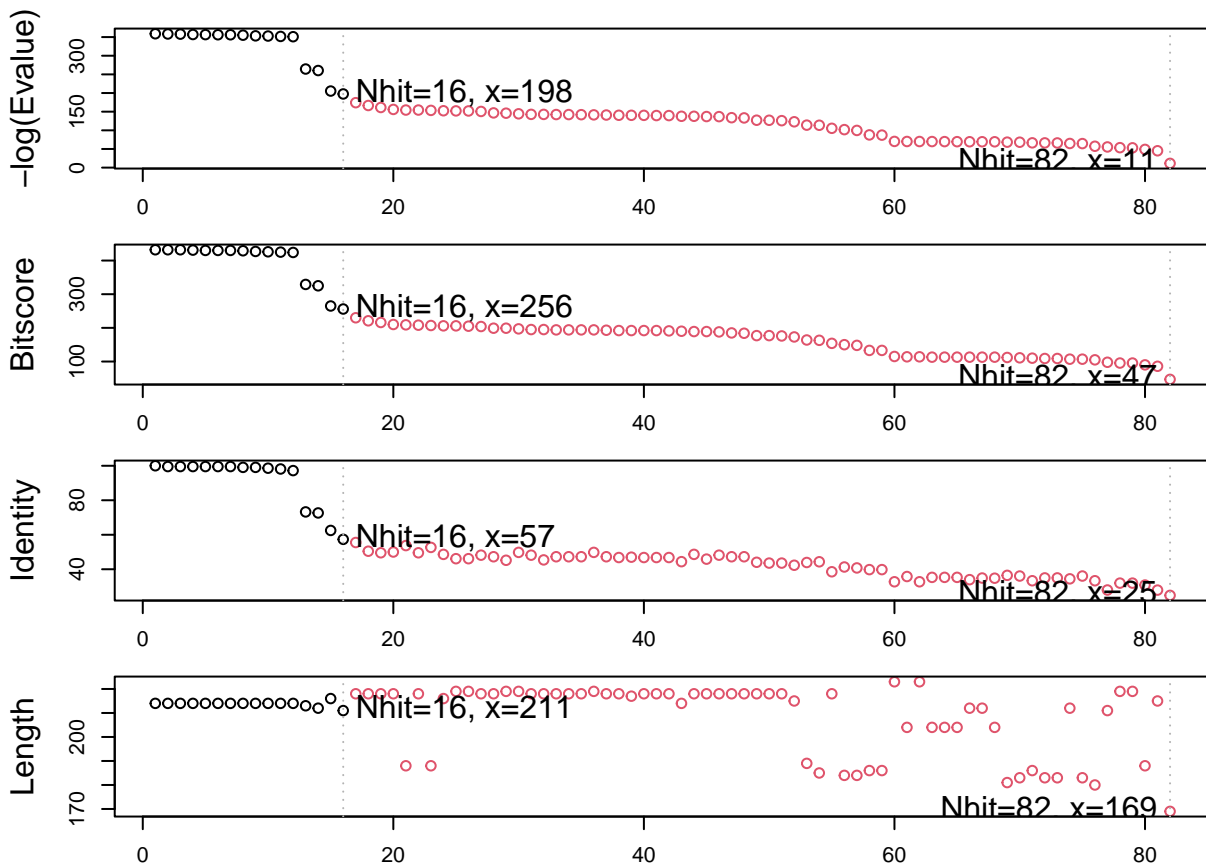
214

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

```
## Searching ... please wait (updates every 5 seconds) RID = 5JHB69E901N
## .
## Reporting 82 hits
```

```
# Plot a summary of search results
hits <- plot(b)
```

```
## * Possible cutoff values: 197 11
##      Yielding Nhits: 16 82
##
## * Chosen cutoff value of: 197
##      Yielding Nhits: 16
```



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

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

```
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_A')
```

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

```
## Warning in get.pdb(hits$ pdb.id, path = "pds", split = TRUE, gzip = TRUE):
## pds/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

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

## 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
```

```

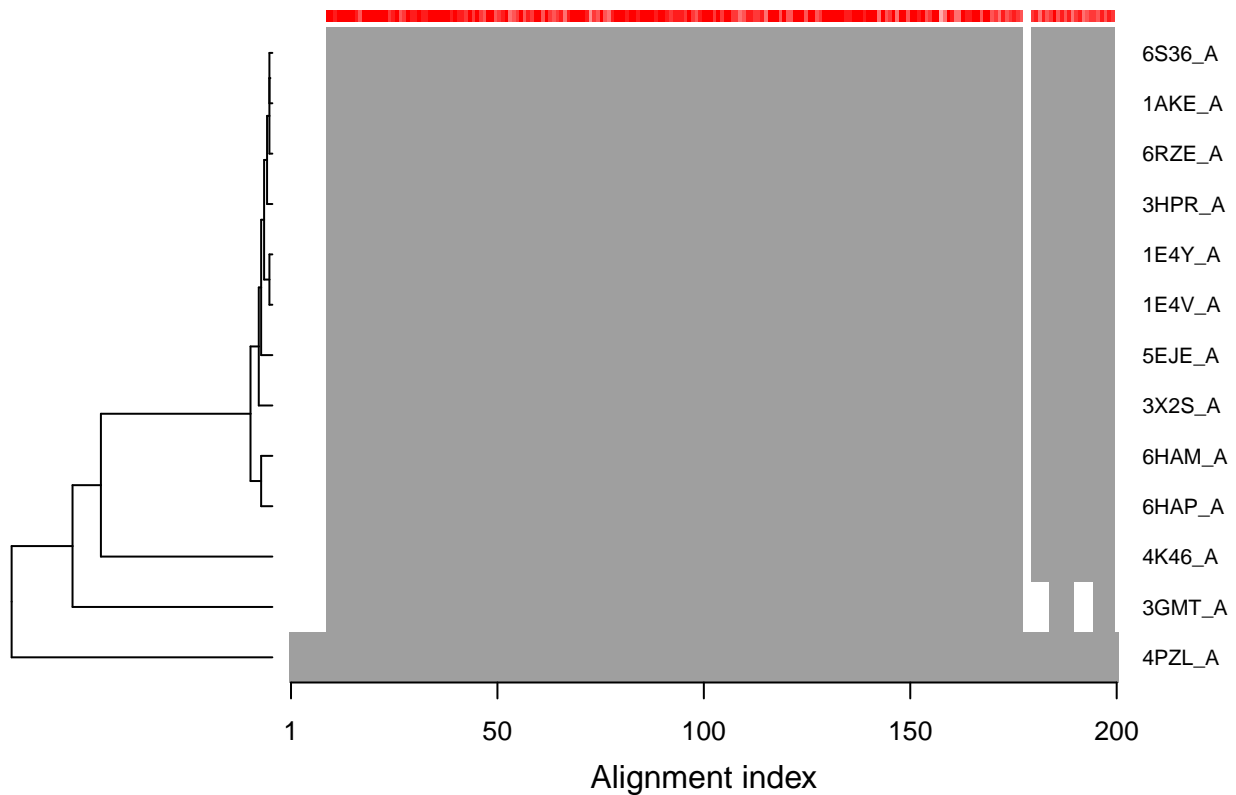
## pdbc/split_chain/5EJE_A.pdb
## pdbc/split_chain/1E4Y_A.pdb
## pdbc/split_chain/3X2S_A.pdb
## pdbc/split_chain/6HAP_A.pdb
## pdbc/split_chain/6HAM_A.pdb
## pdbc/split_chain/4K46_A.pdb
## pdbc/split_chain/3GMT_A.pdb
## pdbc/split_chain/4PZL_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## .   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: pdbc/split_chain/1AKE_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 2   name: pdbc/split_chain/6S36_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 3   name: pdbc/split_chain/6RZE_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 4   name: pdbc/split_chain/3HPR_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 5   name: pdbc/split_chain/1E4V_A.pdb
## pdb/seq: 6   name: pdbc/split_chain/5EJE_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 7   name: pdbc/split_chain/1E4Y_A.pdb
## pdb/seq: 8   name: pdbc/split_chain/3X2S_A.pdb
## pdb/seq: 9   name: pdbc/split_chain/6HAP_A.pdb
## pdb/seq: 10  name: pdbc/split_chain/6HAM_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 11  name: pdbc/split_chain/4K46_A.pdb
##   PDB has ALT records, taking A only, rm.alt=TRUE
## pdb/seq: 12  name: pdbc/split_chain/3GMT_A.pdb
## pdb/seq: 13  name: pdbc/split_chain/4PZL_A.pdb

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

# Draw schematic alignment
plot(pdbc, labels=ids)

```

Sequence Alignment Overview



Annotate collected PDB structures

The function `pdb.annotate()` provides a convenient way of annotating the PDB files we have collected.

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

```
## [1] "Escherichia coli"
## [2] "Escherichia coli K-12"
## [3] "Escherichia coli 0139:H28 str. E24377A"
## [4] "Escherichia coli str. K-12 substr. MDS42"
## [5] "Photobacterium profundum"
## [6] "Burkholderia pseudomallei 1710b"
## [7] "Francisella tularensis subsp. tularensis SCHU S4"
```

view all available annotation data:

```
anno
```

```
##      structureId chainId macromoleculeType chainLength experimentalTechnique
## 1AKE_A         1AKE      A             Protein           214             X-ray
## 6S36_A         6S36      A             Protein           214             X-ray
## 6RZE_A         6RZE      A             Protein           214             X-ray
## 3HPR_A         3HPR      A             Protein           214             X-ray
```

##	1E4V_A	1E4V	A	Protein	214	X-ray
##	5EJE_A	5EJE	A	Protein	214	X-ray
##	1E4Y_A	1E4Y	A	Protein	214	X-ray
##	3X2S_A	3X2S	A	Protein	214	X-ray
##	6HAP_A	6HAP	A	Protein	214	X-ray
##	6HAM_A	6HAM	A	Protein	214	X-ray
##	4K46_A	4K46	A	Protein	214	X-ray
##	3GMT_A	3GMT	A	Protein	230	X-ray
##	4PZL_A	4PZL	A	Protein	242	X-ray

##	resolution	scopDomain	pfam	ligandId
##	1AKE_A	2.00 Adenylate kinase	Adenylate kinase (ADK)	AP5
##	6S36_A	1.60 <NA>	Adenylate kinase (ADK)	CL (3),NA,MG (2)
##	6RZE_A	1.69 <NA>	Adenylate kinase (ADK)	NA (3),CL (2)
##	3HPR_A	2.00 <NA>	Adenylate kinase (ADK)	AP5
##	1E4V_A	1.85 Adenylate kinase	Adenylate kinase (ADK)	AP5
##	5EJE_A	1.90 <NA>	Adenylate kinase (ADK)	AP5,CO
##	1E4Y_A	1.85 Adenylate kinase	Adenylate kinase (ADK)	AP5
##	3X2S_A	2.80 <NA>	Adenylate kinase (ADK)	JPY (2),AP5,MG
##	6HAP_A	2.70 <NA>	Adenylate kinase (ADK)	AP5
##	6HAM_A	2.55 <NA>	Adenylate kinase (ADK)	AP5
##	4K46_A	2.01 <NA>	Adenylate kinase (ADK)	ADP,AMP,P04
##	3GMT_A	2.10 <NA>	Adenylate kinase (ADK)	S04 (2)
##	4PZL_A	2.10 <NA>	Adenylate kinase (ADK)	CA,FMT,GOL

##	ligandName	
##	1AKE_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	6S36_A	CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)
##	6RZE_A	SODIUM ION (3),CHLORIDE ION (2)
##	3HPR_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	1E4V_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	5EJE_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE,COBALT (II) ION
##	1E4Y_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	3X2S_A	N-(pyren-1-ylmethyl)acetamide (2),BIS(ADENOSINE)-5'-PENTAPHOSPHATE,MAGNESIUM ION
##	6HAP_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	6HAM_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
##	4K46_A	ADENOSINE-5'-DIPHOSPHATE,ADENOSINE MONOPHOSPHATE,PHOSPHATE ION
##	3GMT_A	SULFATE ION (2)
##	4PZL_A	CALCIUM ION,FORMIC ACID,GLYCEROL

##	source	
##	1AKE_A	Escherichia coli
##	6S36_A	Escherichia coli
##	6RZE_A	Escherichia coli
##	3HPR_A	Escherichia coli K-12
##	1E4V_A	Escherichia coli
##	5EJE_A	Escherichia coli 0139:H28 str. E24377A
##	1E4Y_A	Escherichia coli
##	3X2S_A	Escherichia coli str. K-12 substr. MDS42
##	6HAP_A	Escherichia coli 0139:H28 str. E24377A
##	6HAM_A	Escherichia coli K-12
##	4K46_A	Photobacterium profundum
##	3GMT_A	Burkholderia pseudomallei 1710b
##	4PZL_A	Francisella tularensis subsp. tularensis SCHU S4

1AKE_A STRUCTURE OF THE COMPLEX BETWEEN ADENYLATE KINASE FROM ESCHERICHIA COLI AND THE INHIBITOR AP5.
6S36_A

```
## 6RZE_A
## 3HPR_A
## 1E4V_A
## 5EJE_A
## 1E4Y_A
## 3X2S_A
## 6HAP_A
## 6HAM_A
## 4K46_A
## 3GMT_A
## 4PZL_A
```

Crystal stru

The crystal stru

```
##
## citation rObserved rFree
## 1AKE_A Muller, C.W., et al. J Mol Biol (1992) 0.19600 NA
## 6S36_A Rogne, P., et al. Biochemistry (2019) 0.16320 0.23560
## 6RZE_A Rogne, P., et al. Biochemistry (2019) 0.18650 0.23500
## 3HPR_A Schrank, T.P., et al. Proc Natl Acad Sci U S A (2009) 0.21000 0.24320
## 1E4V_A Muller, C.W., et al. Proteins (1993) 0.19600 NA
## 5EJE_A Kovermann, M., et al. Proc Natl Acad Sci U S A (2017) 0.18890 0.23580
## 1E4Y_A Muller, C.W., et al. Proteins (1993) 0.17800 NA
## 3X2S_A Fujii, A., et al. Bioconjug Chem (2015) 0.20700 0.25600
## 6HAP_A Kantaev, R., et al. J Phys Chem B (2018) 0.22630 0.27760
## 6HAM_A Kantaev, R., et al. J Phys Chem B (2018) 0.20511 0.24325
## 4K46_A Cho, Y.-J., et al. To be published 0.17000 0.22290
## 3GMT_A Buchko, G.W., et al. Biochem Biophys Res Commun (2010) 0.23800 0.29500
## 4PZL_A Tan, K., et al. To be published 0.19360 0.23680
##
## rWork spaceGroup
## 1AKE_A 0.19600 P 21 2 21
## 6S36_A 0.15940 C 1 2 1
## 6RZE_A 0.18190 C 1 2 1
## 3HPR_A 0.20620 P 21 21 2
## 1E4V_A 0.19600 P 21 2 21
## 5EJE_A 0.18630 P 21 2 21
## 1E4Y_A 0.17800 P 1 21 1
## 3X2S_A 0.20700 P 21 21 21
## 6HAP_A 0.22370 I 2 2 2
## 6HAM_A 0.20311 P 43
## 4K46_A 0.16730 P 21 21 21
## 3GMT_A 0.23500 P 1 21 1
## 4PZL_A 0.19130 P 32
```

```
sessionInfo()
```

```
## R version 4.2.3 (2023-03-15 ucrt)
## Platform: x86_64-w64-mingw32/x64 (64-bit)
## Running under: Windows 10 x64 (build 19044)
##
## Matrix products: default
##
## locale:
## [1] LC_COLLATE=English_United States.utf8
## [2] LC_CTYPE=English_United States.utf8
## [3] LC_MONETARY=English_United States.utf8
## [4] LC_NUMERIC=C
## [5] LC_TIME=English_United States.utf8
```



```
##
## attached base packages:
## [1] stats      graphics  grDevices utils      datasets  methods   base
##
## other attached packages:
## [1] bio3d_2.4-4 dplyr_1.1.1
##
## loaded via a namespace (and not attached):
## [1] Rcpp_1.0.10      GenomeInfoDb_1.34.9 pillar_1.9.0
## [4] compiler_4.2.3   highr_0.10        XVector_0.38.0
## [7] bitops_1.0-7     tools_4.2.3       zlibbioc_1.44.0
## [10] digest_0.6.31    jsonlite_1.8.4    evaluate_0.20
## [13] lifecycle_1.0.3  tibble_3.2.1      pkgconfig_2.0.3
## [16] rlang_1.1.0      cli_3.6.1         rstudioapi_0.14
## [19] curl_5.0.0       yaml_2.3.7        parallel_4.2.3
## [22] xfun_0.38        fastmap_1.1.1     GenomeInfoDbData_1.2.9
## [25] httr_1.4.6       knitr_1.42        generics_0.1.3
## [28] Biostrings_2.66.0 vctrs_0.6.1       S4Vectors_0.36.2
## [31] IRanges_2.32.0   stats4_4.2.3      grid_4.2.3
## [34] tidyselect_1.2.0 glue_1.6.2        R6_2.5.1
## [37] fansi_1.0.4      rmarkdown_2.21    msa_1.30.1
## [40] magrittr_2.0.3   htmltools_0.5.5   BiocGenerics_0.44.0
## [43] utf8_1.2.3       RCurl_1.98-1.12   crayon_1.5.2
```