class9

Yunchong Zhao

10/16/2022

Section 1 skipped

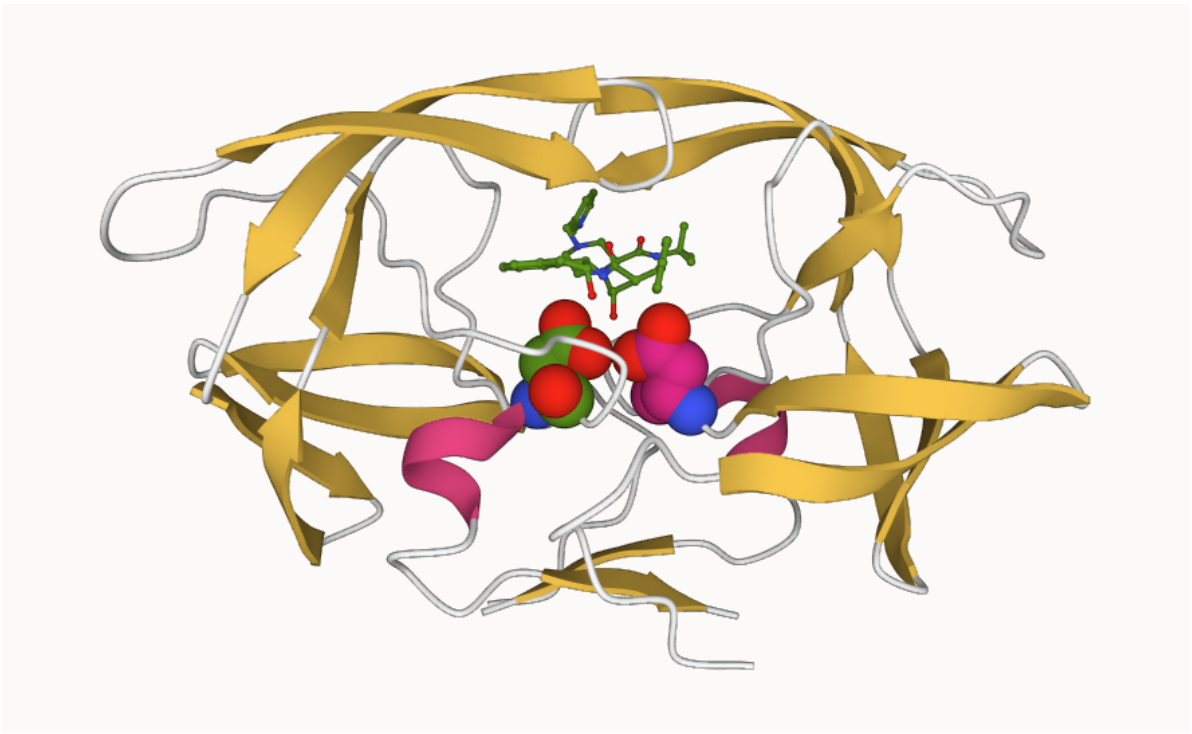
Q4: Water molecules normally have 3 atoms. Why do we see just one atom per water molecule in this structure?

To simplify the view of this proteins structure. The red molecules are representing oxygen and all hydrogens are omitted for a cleaner view.

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

A critical conserved water molecule would be 25D of both chain A and B. They seem to be at the connecting point of chain A and chain B of this protein structure. The hydrogen bonds from the water molecules must be important for the structure of the ligand binding.

#Viewing PDB structures with Molstart



#Reading and working with structures in R

The `bio3d` package for structural bioinfo has lot's of features for reading and working with biolomecular sequences and structures.

```
library(bio3d)
```

```
Warning: package 'bio3d' was built under R version 4.1.3
```

```
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
```

```
pdb$seqres
```

```
    A     A     A     A     A     A     A     A     A     A     A     A     A
"PRO" "GLN" "ILE" "THR" "LEU" "TRP" "GLN" "ARG" "PRO" "LEU" "VAL" "THR" "ILE"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"LYS" "ILE" "GLY" "GLY" "GLN" "LEU" "LYS" "GLU" "ALA" "LEU" "LEU" "ASP" "THR"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"GLY" "ALA" "ASP" "ASP" "THR" "VAL" "LEU" "GLU" "GLU" "MET" "SER" "LEU" "PRO"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"GLY" "ARG" "TRP" "LYS" "PRO" "LYS" "MET" "ILE" "GLY" "GLY" "ILE" "GLY" "GLY"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"PHE" "ILE" "LYS" "VAL" "ARG" "GLN" "TYR" "ASP" "GLN" "ILE" "LEU" "ILE" "GLU"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"ILE" "CYS" "GLY" "HIS" "LYS" "ALA" "ILE" "GLY" "THR" "VAL" "LEU" "VAL" "GLY"
    A     A     A     A     A     A     A     A     A     A     A     A     A
"PRO" "THR" "PRO" "VAL" "ASN" "ILE" "ILE" "GLY" "ARG" "ASN" "LEU" "LEU" "THR"
    A     A     A     A     A     A     A     A     B     B     B     B     B
"GLN" "ILE" "GLY" "CYS" "THR" "LEU" "ASN" "PHE" "PRO" "GLN" "ILE" "THR" "LEU"
    B     B     B     B     B     B     B     B     B     B     B     B     B
"TRP" "GLN" "ARG" "PRO" "LEU" "VAL" "THR" "ILE" "LYS" "ILE" "GLY" "GLY" "GLN"
    B     B     B     B     B     B     B     B     B     B     B     B     B
"LEU" "LYS" "GLU" "ALA" "LEU" "LEU" "ASP" "THR" "GLY" "ALA" "ASP" "ASP" "THR"
    B     B     B     B     B     B     B     B     B     B     B     B     B
```

```
"VAL" "LEU" "GLU" "GLU" "MET" "SER" "LEU" "PRO" "GLY" "ARG" "TRP" "LYS" "PRO"
   B     B     B     B     B     B     B     B     B     B     B     B     B
"LYS" "MET" "ILE" "GLY" "GLY" "ILE" "GLY" "GLY" "PHE" "ILE" "LYS" "VAL" "ARG"
   B     B     B     B     B     B     B     B     B     B     B     B     B
"GLN" "TYR" "ASP" "GLN" "ILE" "LEU" "ILE" "GLU" "ILE" "CYS" "GLY" "HIS" "LYS"
   B     B     B     B     B     B     B     B     B     B     B     B     B
"ALA" "ILE" "GLY" "THR" "VAL" "LEU" "VAL" "GLY" "PRO" "THR" "PRO" "VAL" "ASN"
   B     B     B     B     B     B     B     B     B     B     B     B     B
"ILE" "ILE" "GLY" "ARG" "ASN" "LEU" "LEU" "THR" "GLN" "ILE" "GLY" "CYS" "THR"
   B     B     B
"LEU" "ASN" "PHE"
```

```
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.

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

Water.

Q9: How many protein chains are in this structure?

Two.

creating a new PDB object

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

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

```
adk
```

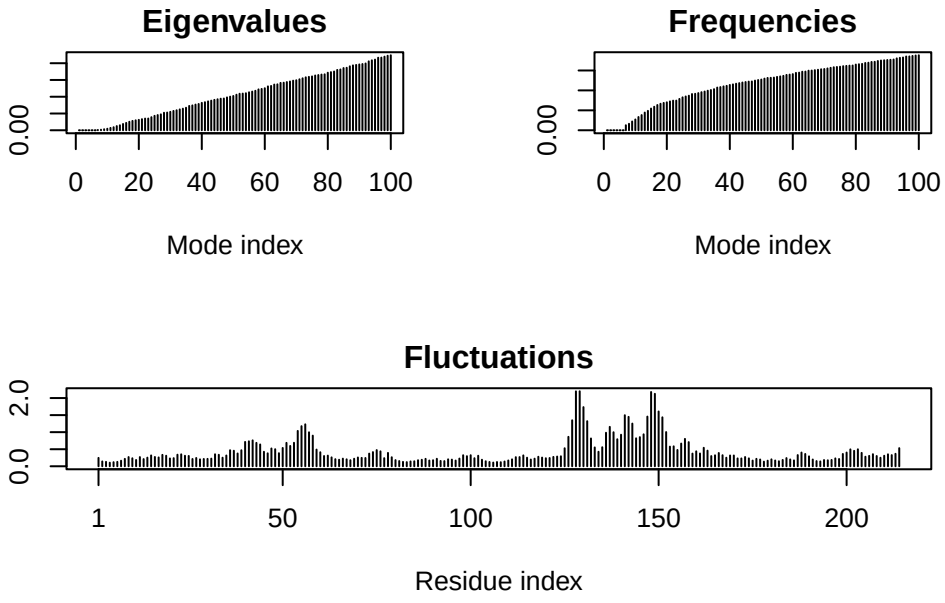
```
 Call:  read.pdb(file = "6s36")

   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:
      MRIILLGAPGAGKGTQAQFIMEKYGIPQISTGDMLRAAVKSGSELGKQAKDIMDAGKLVT
      DELVIALVKERIAQEDCRNGFLLDGFPRTIPQADAMKEAGINVDYVLEFDVPDELIVDKI
      VGRRVHAPSGRVYHVKFNPPKVEGKDDVTGEELTTRKDDQEETVRKRLVEYHQMTAPLIG
      YYSKEAEAGNTKYAKVDGTKPVAEVRADLEKILG

+ attr: atom, xyz, seqres, helix, sheet,
        calpha, remark, call
```

Normal mode analysis (NMA) is a bioinfo method for predicting functional motions. It will show uis the parts of the protein that are "flexible"

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

```
Building Hessian...        Done in 0.03 seconds.
Diagonalizing Hessian...   Done in 0.36 seconds.
```

```
plot(m)
```



Make a movie of this thing moving:

```
mktrj(m, file="adk.nma.pdb")
```

```
# Install packages in the R console NOT your Rmd/Quarto file

#install.packages("bio3d")
#install.packages("devtools")
#install.packages("BiocManager")

#BiocManager::install("msa")
#devtools::install_bitbucket("Grantlab/bio3d-view")
```

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.

Comparative analysis of all ADK structures.

First we get the sequence of ADK and use this to search the PDB database.

```
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?

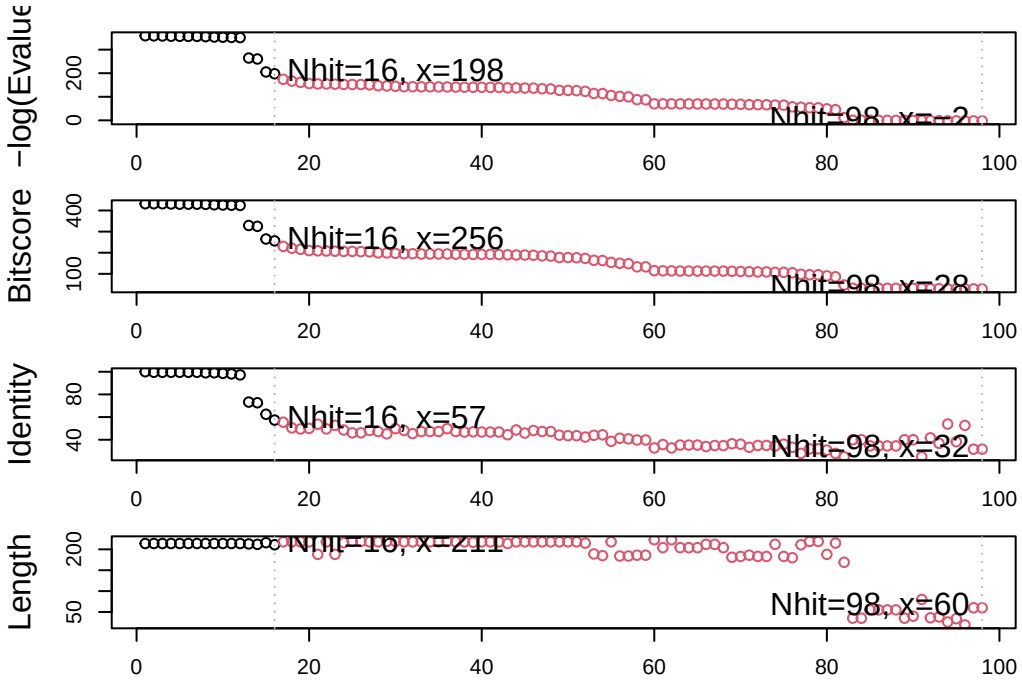
There are 214 a.a. in this sequence.

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

```
Searching ... please wait (updates every 5 seconds) RID = NKF6202401R
.
Reporting 98 hits
```

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

```
  * Possible cutoff values:    197 -3
            Yielding Nhits:    16 98

  * 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"
```

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

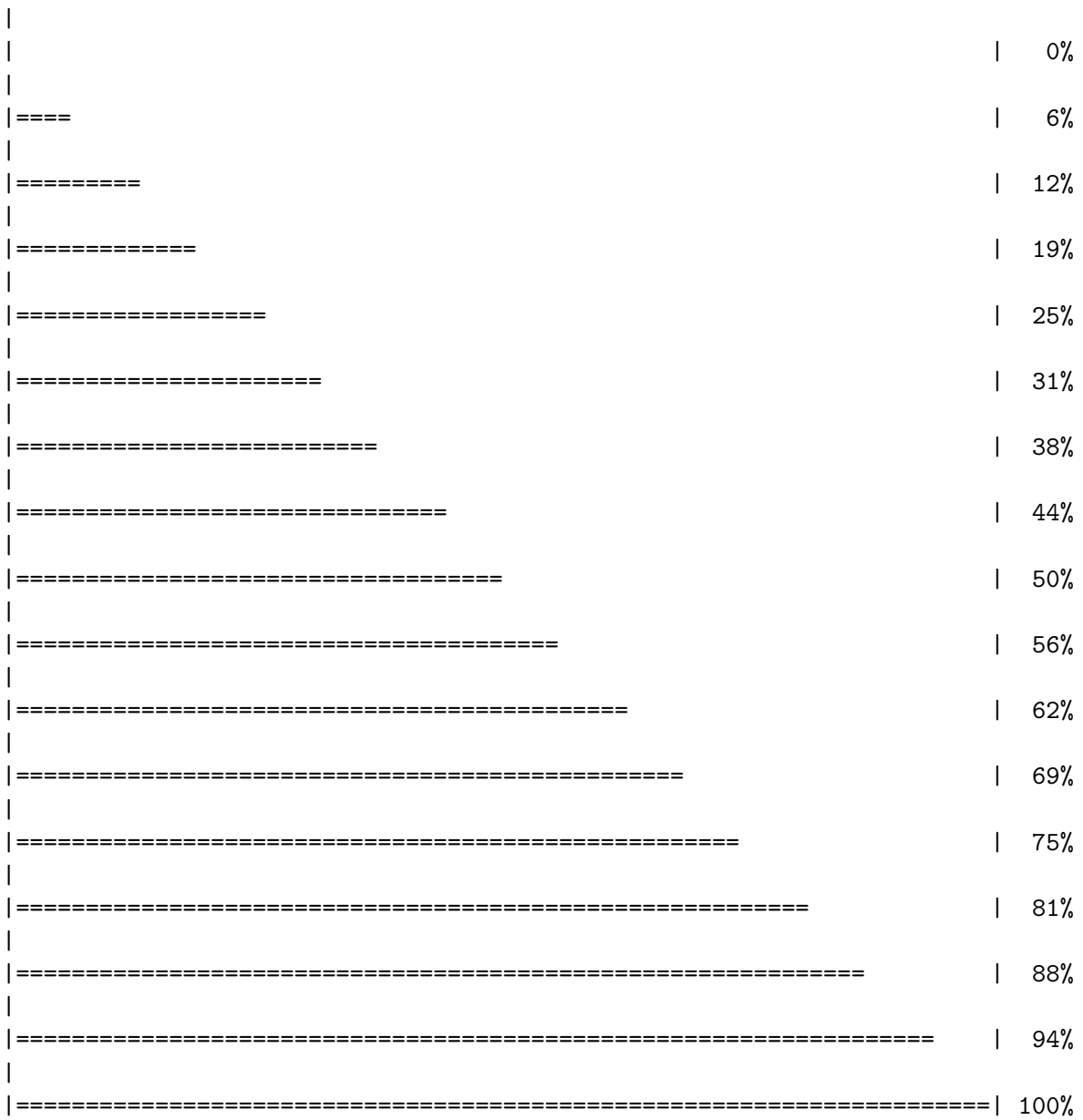
```
Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
1AKE.pdb exists. Skipping download
```

```
Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE): pdbs/
4X8M.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/
4X8H.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/
4NP6.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
```


|
| 0%
|
|==== | 6%
|
|========= | 12%
|
|========== | 19%
|
|========== | 25%
|
|========== | 31%
|
|========== | 38%
|
|========== | 44%
|
|========== | 50%
|
|========== | 56%
|
|========== | 62%
|
|========== | 69%
|
|========== | 75%
|
|========== | 81%
|
|========== | 88%
|
|========== | 94%
|
|==========| 100%

Align and superpose structures

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

```
Reading PDB files:
pdbs/split_chain/1AKE_A.pdb
pdbs/split_chain/4X8M_A.pdb
pdbs/split_chain/6S36_A.pdb
pdbs/split_chain/6RZE_A.pdb
pdbs/split_chain/4X8H_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/3X2S_A.pdb
pdbs/split_chain/6HAP_A.pdb
pdbs/split_chain/6HAM_A.pdb
pdbs/split_chain/4K46_A.pdb
pdbs/split_chain/4NP6_A.pdb
pdbs/split_chain/3GMT_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/4X8M_A.pdb
pdb/seq: 3   name: pdbs/split_chain/6S36_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 4   name: pdbs/split_chain/6RZE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 5   name: pdbs/split_chain/4X8H_A.pdb
```

```
pdb/seq: 6   name: pdbs/split_chain/3HPR_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 7   name: pdbs/split_chain/1E4V_A.pdb
pdb/seq: 8   name: pdbs/split_chain/5EJE_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 9   name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 10   name: pdbs/split_chain/3X2S_A.pdb
pdb/seq: 11   name: pdbs/split_chain/6HAP_A.pdb
pdb/seq: 12   name: pdbs/split_chain/6HAM_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 13   name: pdbs/split_chain/4K46_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 14   name: pdbs/split_chain/4NP6_A.pdb
pdb/seq: 15   name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 16   name: pdbs/split_chain/4PZL_A.pdb
```

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

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

Annotate collected PDB structures

```
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. MDS42"
[5] "Photobacterium profundum"
[6] "Vibrio cholerae O1 biovar El Tor str. N16961"
[7] "Burkholderia pseudomallei 1710b"
[8] "Francisella tularensis subsp. tularensis SCHU S4"
```

```
anno
```

	structureId	chainId	macromoleculeType	chainLength	experimentalTechnique
1AKE_A	1AKE	A	Protein	214	X-ray
4X8M_A	4X8M	A	Protein	214	X-ray
6S36_A	6S36	A	Protein	214	X-ray
6RZE_A	6RZE	A	Protein	214	X-ray
4X8H_A	4X8H	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
4NP6_A	4NP6	A	Protein	217	X-ray
3GMT_A	3GMT	A	Protein	230	X-ray
4PZL_A	4PZL	A	Protein	242	X-ray

	resolution	scopDomain	pfam
1AKE_A	2.000	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)
4X8M_A	2.600	<NA>	Adenylate kinase, active site lid (ADK_lid)
6S36_A	1.600	<NA>	Adenylate kinase, active site lid (ADK_lid)
6RZE_A	1.690	<NA>	Adenylate kinase, active site lid (ADK_lid)
4X8H_A	2.500	<NA>	Adenylate kinase, active site lid (ADK_lid)
3HPR_A	2.000	<NA>	Adenylate kinase, active site lid (ADK_lid)
1E4V_A	1.850	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)
5EJE_A	1.900	<NA>	Adenylate kinase, active site lid (ADK_lid)
1E4Y_A	1.850	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)
3X2S_A	2.800	<NA>	Adenylate kinase, active site lid (ADK_lid)
6HAP_A	2.700	<NA>	Adenylate kinase, active site lid (ADK_lid)
6HAM_A	2.550	<NA>	Adenylate kinase, active site lid (ADK_lid)
4K46_A	2.010	<NA>	Adenylate kinase, active site lid (ADK_lid)
4NP6_A	2.004	<NA>	Adenylate kinase, active site lid (ADK_lid)
3GMT_A	2.100	<NA>	Adenylate kinase, active site lid (ADK_lid)
4PZL_A	2.100	<NA>	Adenylate kinase, active site lid (ADK_lid)

	ligandId
1AKE_A	AP5
4X8M_A	<NA>
6S36_A	CL (3),NA,MG (2)
6RZE_A	NA (3),CL (2)
4X8H_A	<NA>
3HPR_A	AP5
1E4V_A	AP5
5EJE_A	AP5,CO

1E4Y_A	AP5
3X2S_A	JPY (2),AP5,MG
6HAP_A	AP5
6HAM_A	AP5
4K46_A	ADP,AMP,PO4
4NP6_A	<NA>
3GMT_A	SO4 (2)
4PZL_A	CA,FMT,GOL

	ligandName
1AKE_A	BIS(ADENOSINE)-5'-PENTAPHOSPHATE
4X8M_A	<NA>
6S36_A	CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)
6RZE_A	SODIUM ION (3),CHLORIDE ION (2)
4X8H_A	<NA>
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
4NP6_A	<NA>
3GMT_A	SULFATE ION (2)
4PZL_A	CALCIUM ION,FORMIC ACID,GLYCEROL

	source
1AKE_A	Escherichia coli
4X8M_A	Escherichia coli
6S36_A	Escherichia coli
6RZE_A	Escherichia coli
4X8H_A	Escherichia coli
3HPR_A	Escherichia coli K-12
1E4V_A	Escherichia coli
5EJE_A	Escherichia coli O139:H28 str. E24377A
1E4Y_A	Escherichia coli
3X2S_A	Escherichia coli str. K-12 substr. MDS42
6HAP_A	Escherichia coli O139:H28 str. E24377A
6HAM_A	Escherichia coli K-12
4K46_A	Photobacterium profundum
4NP6_A	Vibrio cholerae O1 biovar El Tor str. N16961
3GMT_A	Burkholderia pseudomallei 1710b
4PZL_A	Francisella tularensis subsp. tularensis SCHU S4

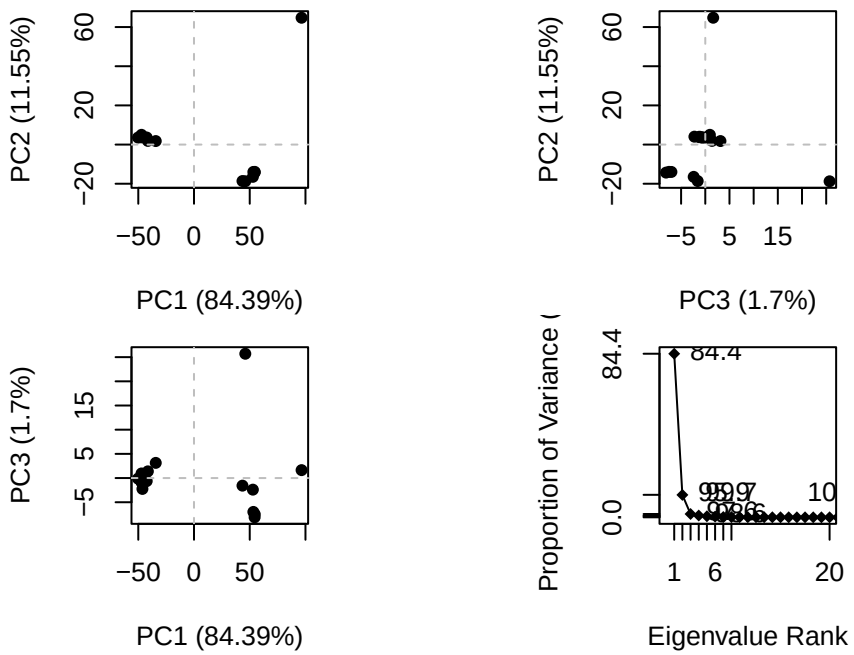
```
1AKE_A STRUCTURE OF THE COMPLEX BETWEEN ADENYLATE KINASE FROM ESCHERICHIA COLI AND THE INHIBI
4X8M_A
6S36_A
6RZE_A
4X8H_A
3HPR_A
1E4V_A
5EJE_A                                                                          Cryst
1E4Y_A
3X2S_A
6HAP_A
6HAM_A
4K46_A
4NP6_A
3GMT_A
4PZL_A                                                                      The cryst
```

	citation	rObserved	rFree	rWork	spaceGroup
1AKE_A	Muller, C.W., et al. J Mol Biol (1992)	0.19600	NA	0.19600	P 21 2 21
4X8M_A	Kovermann, M., et al. Nat Commun (2015)	0.24910	0.30890	0.24630	C 1 2 1
6S36_A	Rogne, P., et al. Biochemistry (2019)	0.16320	0.23560	0.15940	C 1 2 1
6RZE_A	Rogne, P., et al. Biochemistry (2019)	0.18650	0.23500	0.18190	C 1 2 1
4X8H_A	Kovermann, M., et al. Nat Commun (2015)	0.19610	0.28950	0.19140	C 1 2 1
3HPR_A	Schrank, T.P., et al. Proc Natl Acad Sci U S A (2009)	0.21000	0.24320	0.20620	P 21 21 2
1E4V_A	Muller, C.W., et al. Proteins (1993)	0.19600	NA	0.19600	P 21 2 21
5EJE_A	Kovermann, M., et al. Proc Natl Acad Sci U S A (2017)	0.18890	0.23580	0.18630	P 21 2 21
1E4Y_A	Muller, C.W., et al. Proteins (1993)	0.17800	NA	0.17800	P 1 21 1
3X2S_A	Fujii, A., et al. Bioconjug Chem (2015)	0.20700	0.25600	0.20400	P 21 21 2
6HAP_A	Kantaev, R., et al. J Phys Chem B (2018)	0.22630	0.27760	0.22350	P 21 21 2
6HAM_A	Kantaev, R., et al. J Phys Chem B (2018)	0.20511	0.24325	0.20230	P 21 21 2
4K46_A	Cho, Y.-J., et al. To be published	0.17000	0.22290	0.16720	P 21 21 2
4NP6_A	Kim, Y., et al. To be published	0.18800	0.22200	0.18520	P 21 21 2
3GMT_A	Buchko, G.W., et al. Biochem Biophys Res Commun (2010)	0.23800	0.29500	0.23520	P 21 21 2
4PZL_A	Tan, K., et al. To be published	0.19360	0.23680	0.19080	P 21 21 2

```
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
4NP6_A 0.18600       P 43
3GMT_A 0.23500   P 1 21 1
4PZL_A 0.19130       P 32
```

PCA to the rescue!!

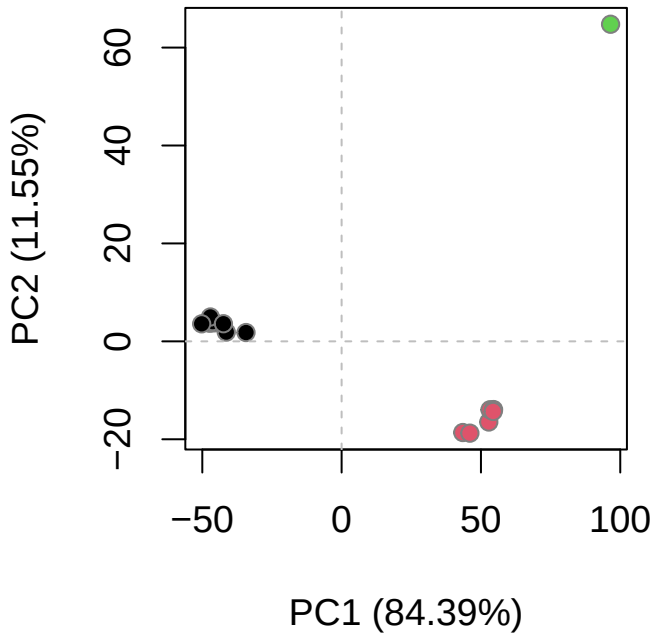
```
# Perform PCA
pc.xray <- pca(pdbs)
plot(pc.xray)
```




```
# 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", bg=grps.rd, pch=21, cex=1)
```



5. Optional further visualization

```
# Visualize first principal component
pc1 <- mktrj(pc.xray, pc=1, file="pc_1.pdb")
```

ggplotting

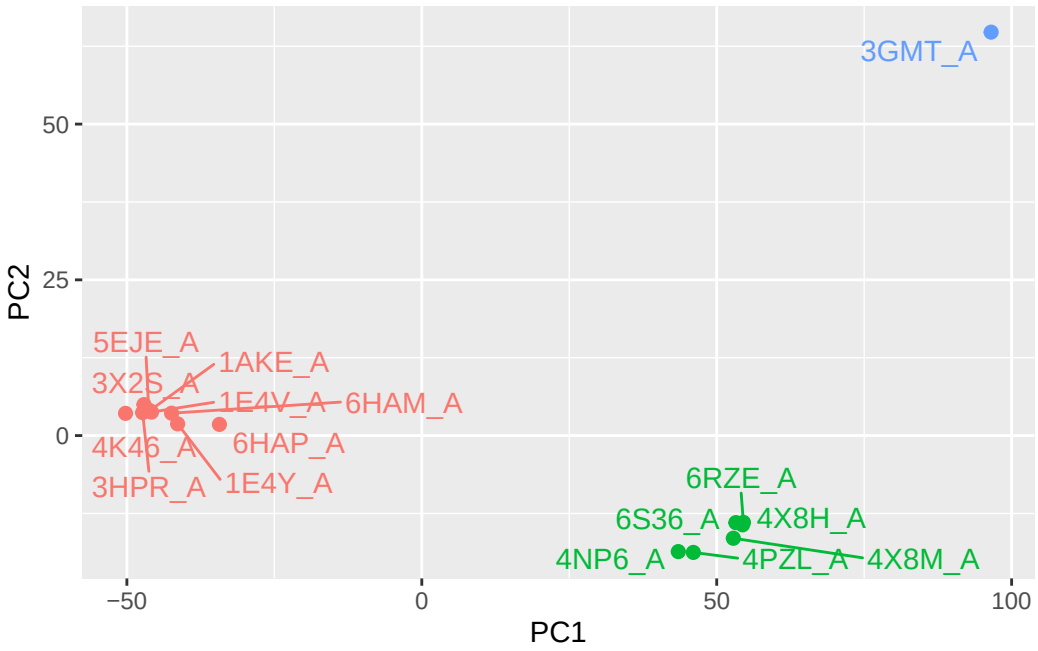
```
#Plotting results with ggplot2
library(ggplot2)
```

```
Warning: package 'ggplot2' was built under R version 4.1.3
```

```
library(ggrepel)
```

```
Warning: package 'ggrepel' was built under R version 4.1.3
```

```
df <- data.frame(PC1=pc.xray$z[,1],
                 PC2=pc.xray$z[,2],
                 col=as.factor(grps.rd),
                 ids=ids)

p <- ggplot(df) +
  aes(PC1, PC2, col=col, label=ids) +
  geom_point(size=2) +
  geom_text_repel(max.overlaps = 20) +
  theme(legend.position = "none")
p
```



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?

The black and colored lineso are diferent. The differences are at selected local regions and are probably important catalytic sites for ligand or drug binding activities, or that those regions

are at the outside of the protein and can be less conserved comparing to basic secondary structures that provides basic structural integrity of the protein.