Class 11: Protein Structure Prediction with AlphaFold

Noel Lim (PID: A17652474)

Q10. Which of the packages above is found only on BioConductor and not CRAN?

MSA Package

Q11. Which of the above packages is not found on BioConductor or CRAN?:

Q12. T/F? Functions from the devtools package can be used to install packages from GitHub and BitBucket?

True

```
library("bio3d")
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
```

Q13. How many amino acids are in this sequence, i.e. how long is this sequence?

Search PDB database for related sequences

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

```
# hits <- plot(b)
```

```
# head(hits$pdb.id)
```

```
hits <- NULL
hits$pdb.id <- c('1AKE_A','6S36_A','6RZE_A','3HPR_A','1E4V_A','5EJE_A','1E4Y_A','3X2S_A','6HA
```

Download related structures

```
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/3X2S.pdb.gz exists. Skipping download
```

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

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

Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE):
pdbs/4K46.pdb.gz exists. Skipping download

Warning in get.pdb(hits$pdb.id, path = "pdbs", split = TRUE, gzip = TRUE):
pdbs/3GMT.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%
  |
  |=====                                                                 |   8%
  |
  |===========                                                           |  15%
  |
  |================                                                      |  23%
  |
  |======================                                                |  31%
  |
  |============================                                          |  38%
  |
  |================================                                      |  46%
  |
  |====================================                                  |  54%
  |
  |========================================                              |  62%
  |
  |============================================                          |  69%
  |
  |====================================================                  |  77%
  |
  |========================================================              |  85%
  |
  |================================================================      |  92%
  |
  |======================================================================| 100%
```

```
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
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/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/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 has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 7   name: pdbs/split_chain/1E4Y_A.pdb
pdb/seq: 8   name: pdbs/split_chain/3X2S_A.pdb
pdb/seq: 9   name: pdbs/split_chain/6HAP_A.pdb
pdb/seq: 10   name: pdbs/split_chain/6HAM_A.pdb
```

```
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 11   name: pdbs/split_chain/4K46_A.pdb
   PDB has ALT records, taking A only, rm.alt=TRUE
pdb/seq: 12   name: pdbs/split_chain/3GMT_A.pdb
pdb/seq: 13   name: pdbs/split_chain/4PZL_A.pdb
```

```
ids <- basename.pdb(pdbs$id)
```

```
#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. MDS42"
[5] "Photobacterium profundum"
[6] "Burkholderia pseudomallei 1710b"
[7] "Francisella tularensis subsp. tularensis SCHU S4"
```

```
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
1AKE_A       2.00 Adenylate kinase Adenylate kinase, active site lid (ADK_lid)
6S36_A       1.60             <NA>                      Adenylate kinase (ADK)
6RZE_A       1.69             <NA>                      Adenylate kinase (ADK)
```

3HPR_A	2.00	<NA>	Adenylate kinase (ADK)	
1E4V_A	1.85	Adenylate kinase	Adenylate kinase (ADK)	
5EJE_A	1.90	<NA>	Adenylate kinase (ADK)	
1E4Y_A	1.85	Adenylate kinase	Adenylate kinase, active site lid (ADK_lid)	
3X2S_A	2.80	<NA>	Adenylate kinase (ADK)	
6HAP_A	2.70	<NA>	Adenylate kinase, active site lid (ADK_lid)	
6HAM_A	2.55	<NA>	Adenylate kinase (ADK)	
4K46_A	2.01	<NA>	Adenylate kinase, active site lid (ADK_lid)	
3GMT_A	2.10	<NA>	Adenylate kinase (ADK)	
4PZL_A	2.10	<NA>	Adenylate kinase (ADK)	

	ligandId	ligandName	source
1AKE_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	Escherichia coli
6S36_A	CL (3),NA,MG (2)	CHLORIDE ION (3),SODIUM ION,MAGNESIUM ION (2)	Escherichia coli
6RZE_A	CL (2),NA (3)	CHLORIDE ION (2),SODIUM ION (3)	Escherichia coli
3HPR_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	Escherichia coli K-12
1E4V_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	
5EJE_A	AP5,CO	BIS(ADENOSINE)-5'-PENTAPHOSPHATE,COBALT (II) ION	
1E4Y_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	
3X2S_A	JPY (2),AP5,MG	N-(pyren-1-ylmethyl)acetamide (2),BIS(ADENOSINE)-5'-PENTAPHOSPHATE,MAGNESIUM ION	
6HAP_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	
6HAM_A	AP5	BIS(ADENOSINE)-5'-PENTAPHOSPHATE	
4K46_A	ADP,AMP,PO4	ADENOSINE-5'-DIPHOSPHATE,ADENOSINE MONOPHOSPHATE,PHOSPHATE ION	
3GMT_A	SO4 (2)	SULFATE ION (2)	
4PZL_A	CA,FMT,GOL	CALCIUM ION,FORMIC ACID,GLYCEROL	

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
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
6S36_A	
6RZE_A	
3HPR_A	
1E4V_A	
5EJE_A	Cryst
1E4Y_A	
3X2S_A	
6HAP_A	
6HAM_A	
4K46_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
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
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		
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		

```
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
```

Performing PCA

```
pc.xray <- pca(pdbs)
#plot(pc.xray, pc.axes = c(1,2))
```

```
rd <- rmsd(pdbs)
```

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

```
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)
```

Visualize first principal component

```
pc1 <- mktrj(pc.xray, pc=1, file="pc_1.pdb")
```

Plotting results with ggplot2

```
library(ggplot2)
library(ggrepel)

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
```