Class 09: Structural Bioinformatics Pt. 1

Chantal Rabay

2/25/2022

```
r = getOption("repos")
r["CRAN"] = "http://cran.us.r-project.org"
options(repos = r)
```

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

```
(163351+10139)/187423 * 100
```

```
## [1] 92.56601
```

92.57% are solved by X-Ray and Electron Microscopy

[Q2] What proportion of structures in the PDB are protein?

```
163543/187423 * 100
```

```
## [1] 87.25877
```

87.26% 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?

There are 187423 HIV-1 protease structures in the current PDB.

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

In the visual created by VMD we assigned spheres to the HOH residues. Therefore, while water does in fact have 3 atoms, the visual is marking every spot in which the HOH residue is present. So one atom being marked by the green sphere in VMD is three atoms making up the water molecule.

[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)?

OH308:0 around residue MK1

Reading PDB file data into R

```
install.packages("bio3d", dependencies=TRUE)
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
library(bio3d)
pdb <- read.pdb("1hsg.pdb")
```

Note:Accessing on-line PDB file

```
pdb
```

```
##
##  Call:  read.pdb(file = "1hsg.pdb")
##
##    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 amino acid residues

[Q8] Name one of the two non-protein residues?

HOH and MK1

[Q9] How many proteins are in this structure?

There are two proteins in this structure.

Comparative Structure Analysis of Adenylate Kinase

```
# Install packages in the R console not your Rmd

install.packages("bio3d")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
install.packages("ggplot2")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
install.packages("ggrepel")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
install.packages("devtools")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
install.packages("BiocManager")
```

```
##
## The downloaded binary packages are in
##  /var/folders/9r/2f14lwsd285gxhxcsztcvzp40000gn/T//RtmpAAvft7/downloaded_packages
```

```
BiocManager::install("msa")
```

```
## 'getOption("repos")' replaces Bioconductor standard repositories, see
## '?repositories' for details
##
## replacement repositories:
##     CRAN: http://cran.us.r-project.org
```

```
## Bioconductor version 3.14 (BiocManager 1.30.16), R 4.1.2 (2021-11-01)

## Warning: package(s) not installed when version(s) same as current; use `force = TRUE` to
##   re-install: 'msa'

## Old packages: 'class', 'cli', 'colorspace', 'crayon', 'evaluate', 'foreign',
##   'glue', 'jsonlite', 'MASS', 'Matrix', 'mgcv', 'nlme', 'nnet', 'rpart',
##   'spatial', 'tinytex', 'yaml'
```

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

```
## Skipping install of 'bio3d.view' from a bitbucket remote, the SHA1 (dd153987) has not changed since l
##   Use `force = TRUE` to force installation
```

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

msa is only found on BioConductor

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

bio3d-view is not found on Bioconductor or CRAN

[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("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
```

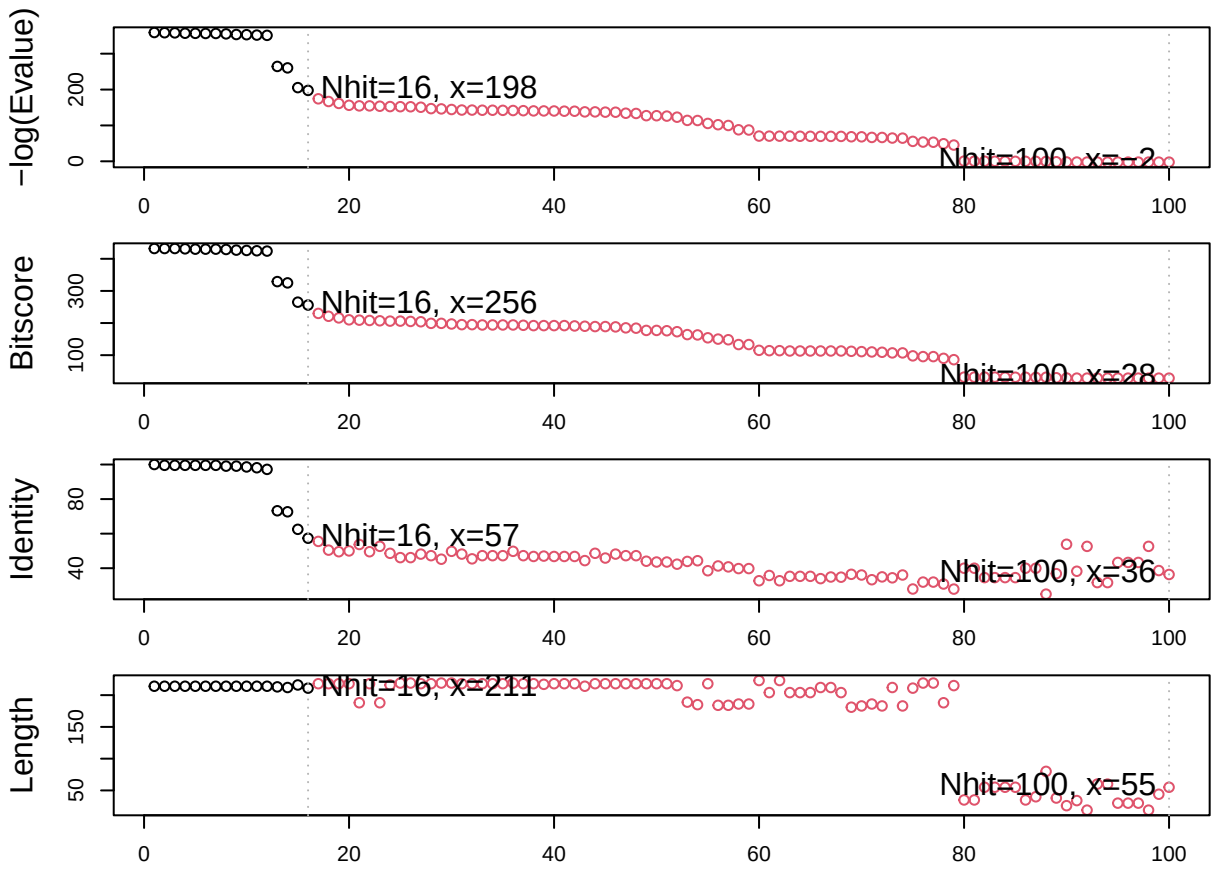
#[Q13] How many amino acids are in this sequence, i.e. how long is this sequence?

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

```
##  Searching ... please wait (updates every 5 seconds) RID = 1JKKTAGK013
##  .
##  Reporting 100 hits
```

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

```
##   * Possible cutoff values:    197 -3
##              Yielding Nhits:    16 100
##
##   * 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.gz exists. Skipping download
```

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

##   |                                                                              |
```

Align and Superpose Structures

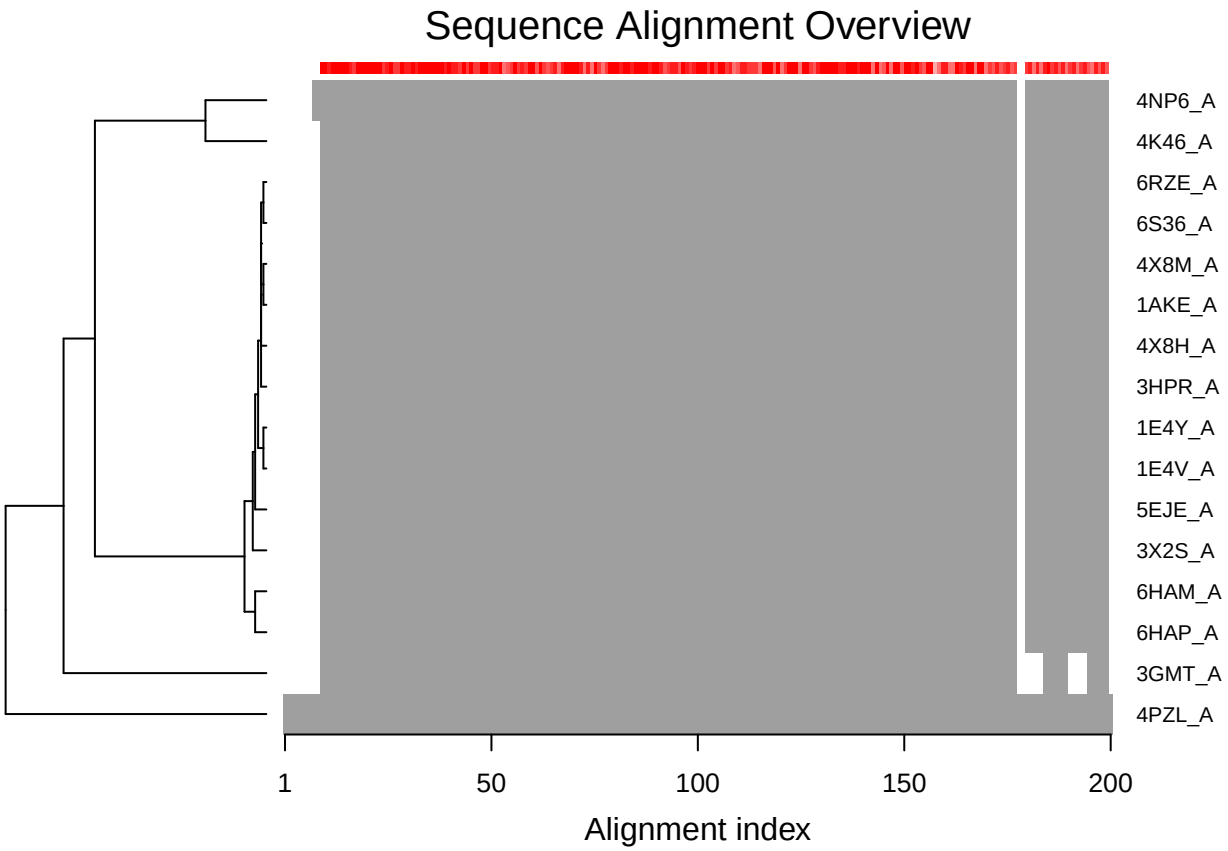
```
# Align releated PDBs
pdbs <- pdbaln(files, fit = TRUE)#, exefile="msa")
```

```
## 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
```

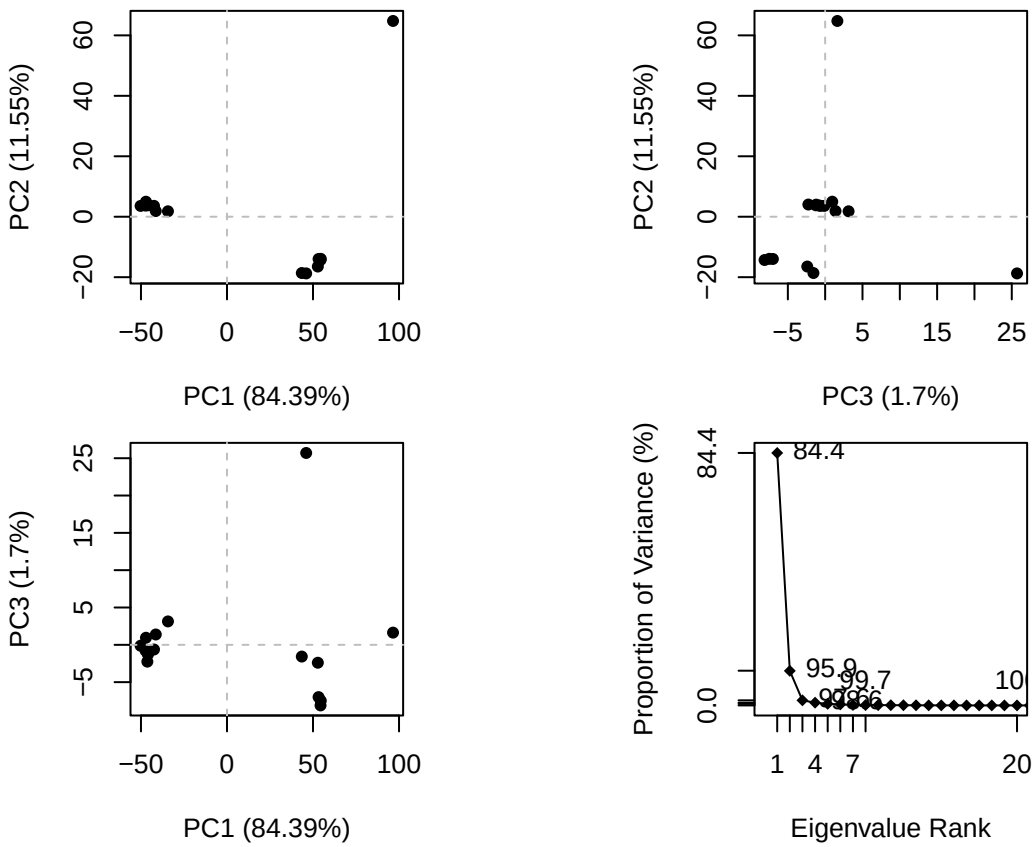
```
# Vector containing PDB codes for figure axis
ids <- basename.pdb(pdbs$id)

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



Principal Component Analysis

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

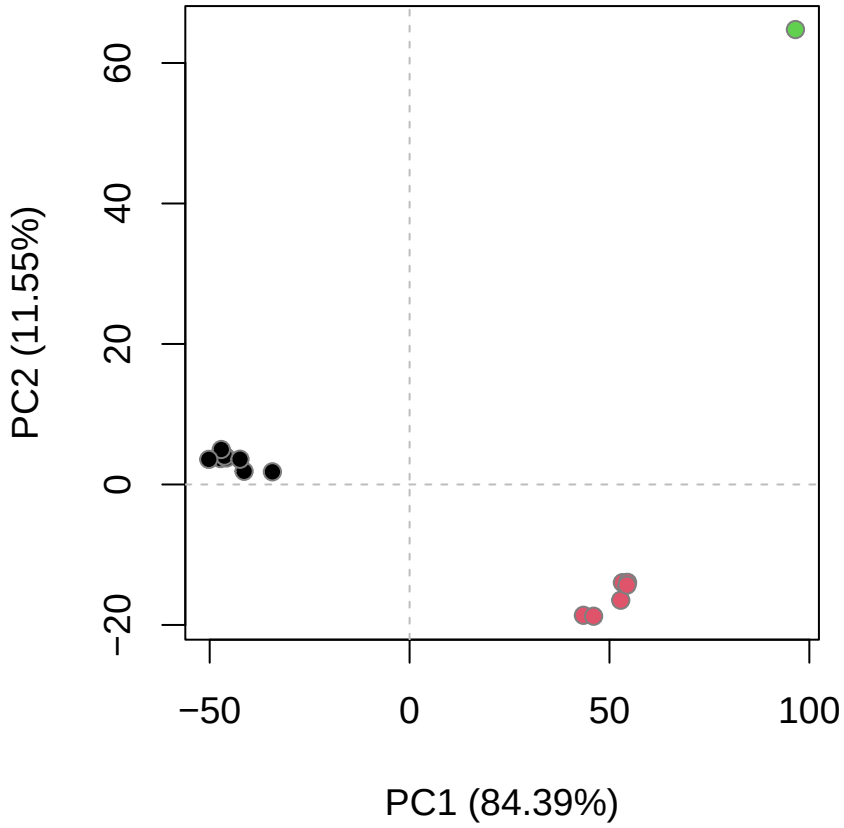



```
# Calculate RMSD
rd <- rmsd(pdbs)
```

```
## 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)
```


PC2 (11.55%)
60
40
20
0
−20
−50
0
50
100
PC1 (84.39%)