HW Class 6 (R Functions)

Audrey Nguyen

```
# (A. Can you improve this analysis code?
df <- data.frame(a=1:10, b=seq(200,400,length=10),c=11:20,d=NA)
df$a <- (df$a - min(df$a)) / (max(df$a) - min(df$a))
df$b <- (df$b - min(df$b)) / (max(df$b) - min(df$b))
df$c <- (df$c - min(df$c)) / (max(df$c) - min(df$c))
df$d <- (df$d - min(df$d)) / (max(df$d) - min(df$d))
df
```

```
          a         b         c  d
1  0.0000000 0.0000000 0.0000000 NA
2  0.1111111 0.1111111 0.1111111 NA
3  0.2222222 0.2222222 0.2222222 NA
4  0.3333333 0.3333333 0.3333333 NA
5  0.4444444 0.4444444 0.4444444 NA
6  0.5555556 0.5555556 0.5555556 NA
7  0.6666667 0.6666667 0.6666667 NA
8  0.7777778 0.7777778 0.7777778 NA
9  0.8888889 0.8888889 0.8888889 NA
10 1.0000000 1.0000000 1.0000000 NA
```

```
analyze <- function(x) {
  (x - min(x)) / (max(x) - min(x))
}
```

```
results <- apply(df, 2, analyze)
results
```

```
             a         b         c  d
 [1,] 0.0000000 0.0000000 0.0000000 NA
```

```
 [2,] 0.1111111 0.1111111 0.1111111 NA
 [3,] 0.2222222 0.2222222 0.2222222 NA
 [4,] 0.3333333 0.3333333 0.3333333 NA
 [5,] 0.4444444 0.4444444 0.4444444 NA
 [6,] 0.5555556 0.5555556 0.5555556 NA
 [7,] 0.6666667 0.6666667 0.6666667 NA
 [8,] 0.7777778 0.7777778 0.7777778 NA
 [9,] 0.8888889 0.8888889 0.8888889 NA
[10,] 1.0000000 1.0000000 1.0000000 NA
```

```
# Can you improve this analysis code?
library(bio3d)
s1 <- read.pdb("4AKE") # kinase with drug
```

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

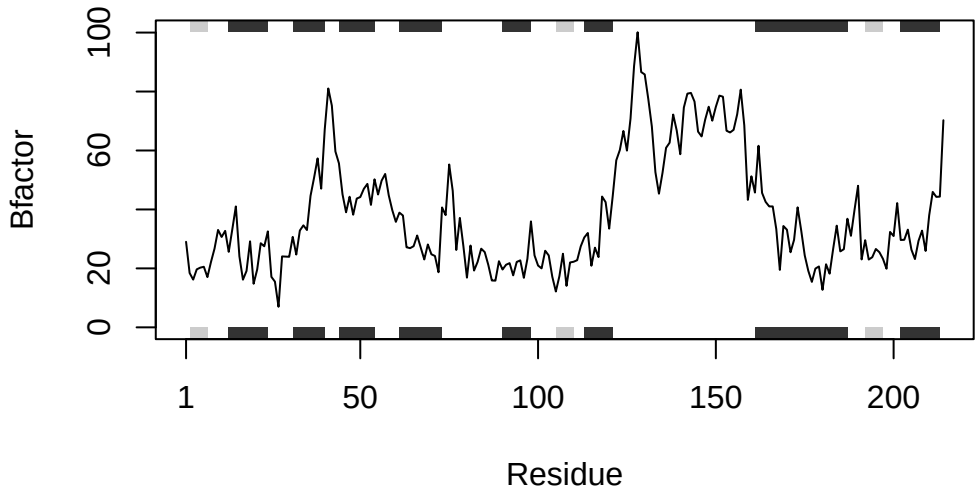
```
s2 <- read.pdb("1AKE") # kinase no drug
```

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

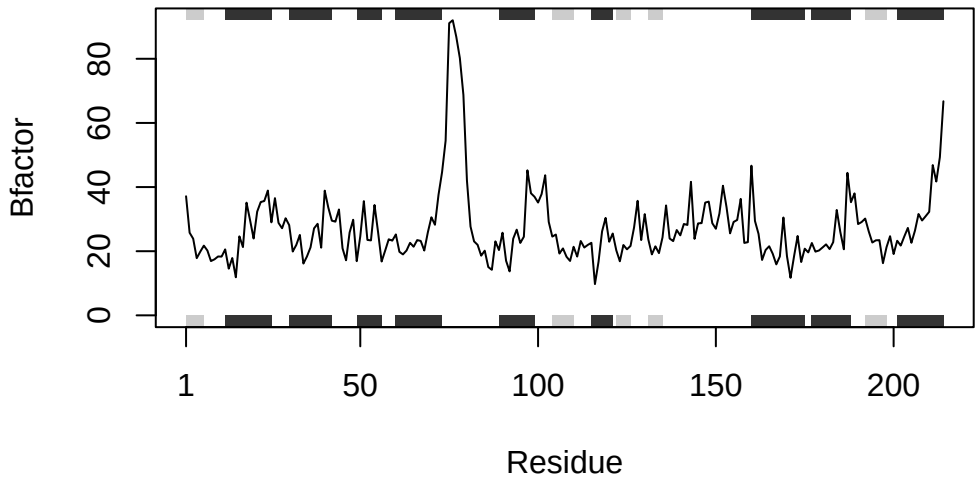
```
s3 <- read.pdb("1E4Y") # kinase with drug
```

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

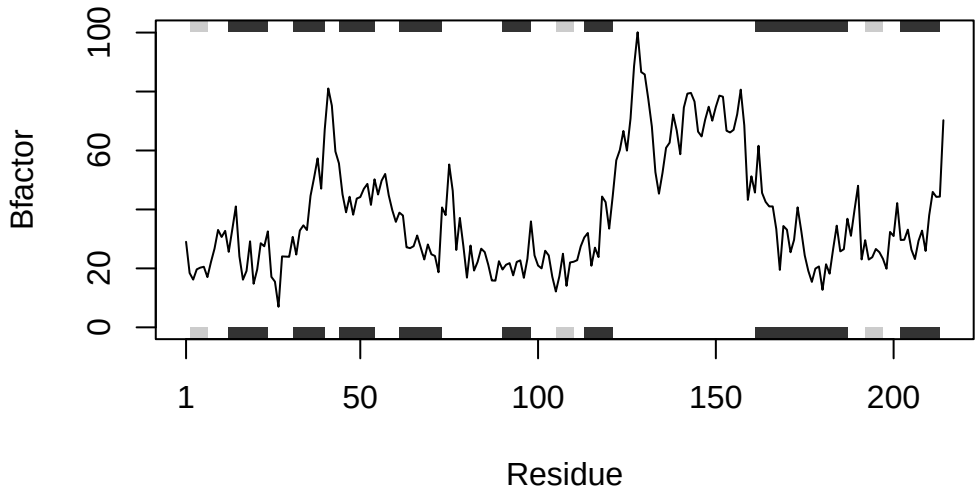
```
s1.chainA <- trim.pdb(s1, chain="A", elety="CA")
s2.chainA <- trim.pdb(s2, chain="A", elety="CA")
s3.chainA <- trim.pdb(s1, chain="A", elety="CA")

s1.b <- s1.chainA$atom$b
s2.b <- s2.chainA$atom$b
s3.b <- s3.chainA$atom$b

plotb3(s1.b, sse=s1.chainA, typ="l", ylab="Bfactor")
```




```
plotb3(s2.b, sse=s2.chainA, typ="l", ylab="Bfactor")
```




```
plotb3(s3.b, sse=s3.chainA, typ="l", ylab="Bfactor")
```



Q1. What type of object is returned from the `read.pdb()` function? The `read.pdb()` function returns a Protein Data Bank (PDB) coordinate file.

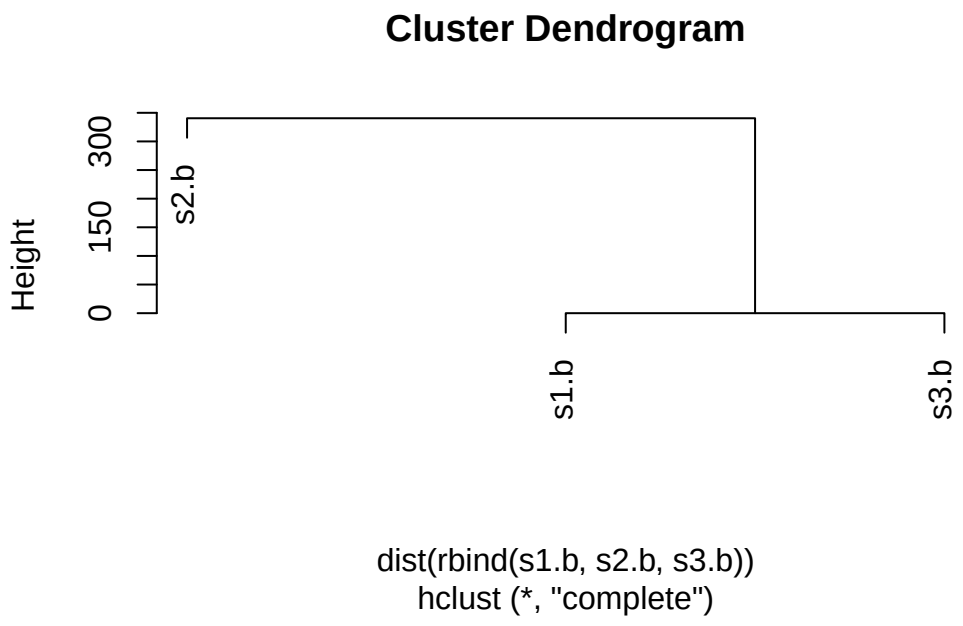
Q2. What does the `trim.pdb()` function do? The `trim.pdb()` function produces a smaller PDB object, containing a subset of atoms.

Q3. What input parameter would turn off the marginal black and grey rectangles in the plots and what do they represent in this case? **sse**, and they represent secondary structure objects returned from read.pdb in this case

Q4. What would be a better plot to compare across the different proteins? Maybe a **cluster dendrogram**, because it shows the hierarchical relationship between objects.

Q5. Which proteins are more similar to each other in their B-factor trends? How could you quantify this? *HINT: try the rbind(), dist(), and hclust() functions together with a resulting dendrogram plot. Look up the documentation to see what each of these functions do.*

```
hc <- hclust(dist(rbind(s1.b, s2.b, s3.b)))
plot(hc)
```



Q6. How would generalize the original code above to work with any set of input protein structures?

```
prot_drug_plot <- function(file, chain, elmnt, fctr) {

  # allows our data to be different colors in the graph
  plot_colors <- c("cyan", "orange", "magenta")


  # to iterate through every value of the file vector
  for (i in 1:length(file)) {
  s1 <- read.pdb(file[i])

  s1.chain <- trim.pdb(s1, chain = chain, elety = elmnt)

  atom_df <- s1.chain$atom

  # the "$" syntax cannot take a variable, so s1.fctr takes in all the atom information an
  s1.fctr <- atom_df[, fctr]

  # creates the first plot
  if (i == 1) {
    plotb3(s1.fctr, sse = s1.chain, typ = "l", ylab = paste(toupper(fctr), "factor", sep =
```

```
    # adds additional plots to first plot
  } else {
    lines(s1.fctr, col = plot_colors[i])
  }
  }

  # creates a legend for the graph
  legend("topright", title = "PDB File Name", file, fill = plot_colors, horiz=TRUE, cex =
}
```

```
files <- c("4AKE", "1AKE", "1E4Y")
chains <- "A"
elements <- "CA"
factors <- "b"

prot_drug_plot(files, chains, elements, factors)
```

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

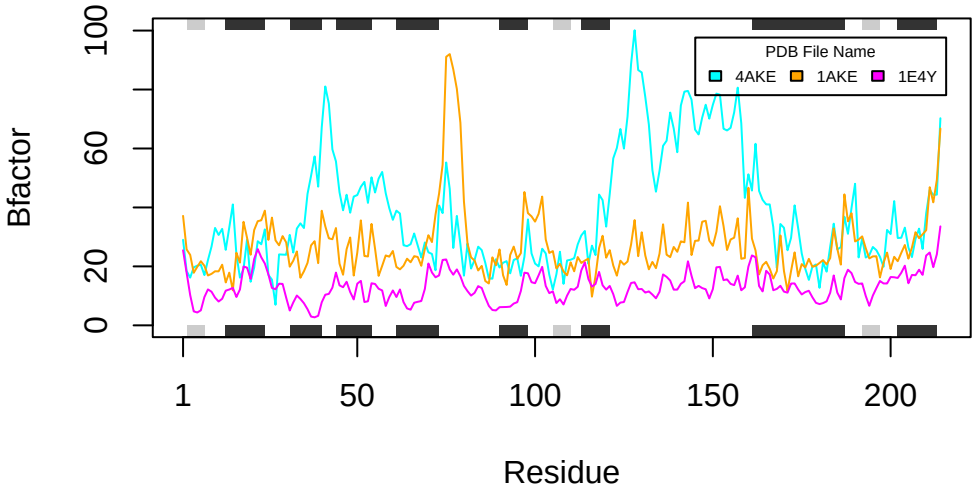
```
Warning in get.pdb(file, path = tempdir(), verbose = FALSE): /var/folders/br/
szcwkw792plb8lhwpfclxz_w0000gn/T//RtmpdztMVc/4AKE.pdb exists. Skipping download
```

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

```
Warning in get.pdb(file, path = tempdir(), verbose = FALSE): /var/folders/br/
szcwkw792plb8lhwpfclxz_w0000gn/T//RtmpdztMVc/1AKE.pdb exists. Skipping download
```

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

```
Warning in get.pdb(file, path = tempdir(), verbose = FALSE): /var/folders/br/
szcwkw792plb8lhwpfclxz_w0000gn/T//RtmpdztMVc/1E4Y.pdb exists. Skipping download
```


Bfactor
0
20
60
100
1
50
100
150
200
Residue
PDB File Name
4AKE
1AKE
1E4Y