PROJECT 4

Marie-Claire Indlewitsch, Anna Toidze, Sophia Houhamdi, Ieva Pudziuvelyte

Problem 8: $e_i(a) = \dfrac{t_i(a) + 1}{\sum_{a' \in A} (E_i(a') + 1)}$

$e_1(A) = \dfrac{4+1}{(4+1)+(0+1)+(0+1)+(0+1)} = 5/8$

$e_2(A) = \dfrac{0+1}{(0+1)+(0+1)+(0+1)+(0+1)} = 1/8 = e_2(C) = e_2(T)$

$e_1(C) = \dfrac{0+1}{(4+1)+3\cdot(0+1)} = 1/8 \quad e_1(G) = 1/8 \quad e_1(T) = 1/8 \quad e_2(G) = \dfrac{4+1}{8} = 5/8$

$e_3(A) = \dfrac{0+1}{(0+1)+(5+1)+(0+1)+(0+1)} = 1/9 = e_3(G) = e_3(T)$

$e_3(C) = \dfrac{5+1}{9} = 6/9$

e_{match}

	1	2	3
A	5/8	1/8	1/9
C	1/8	1/8	6/9
G	1/8	5/8	1/9
T	1/8	1/8	1/9

Problem 9 $e_i(a) = \dfrac{t_i(a) + 1}{\sum_{a' \in A} (E_i(a') + 1)}$

$e_1(A) = \dfrac{0+1}{4\cdot(0+1)} = 1/4 = e_1(C) = e_1(G) = e_1(T) = e_3(A) = e_3(C) = e_3(G) = e_3(T)$

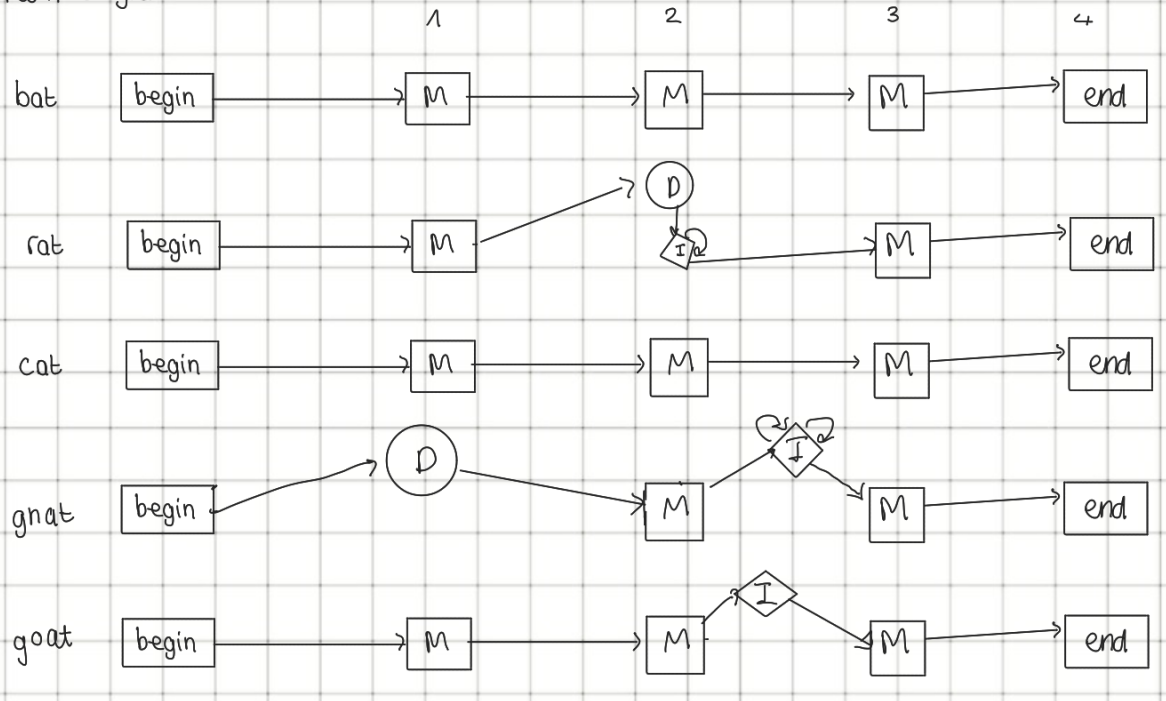
$e_2(A) = \dfrac{5+1}{(5+1)+(1+1)+(0+1)+(0+1)} = 6/10 \quad e_2(G) = \dfrac{1+1}{10} = \dfrac{2}{10}$

$e_2(C) = e_2(T) = 1/10$

e_{insert}

	1	2	3
A	1/4	6/10	1/4
C	1/4	1/10	1/4
G	1/4	2/10	1/4
T	1/4	1/10	1/4

Problem 3 Path diagrams:



$$t_i(k \to \ell) = \frac{T_i(k \to \ell) + 1}{\sum_{\ell' \in S} (T_i(k \to \ell') + 1)}$$

0 → 1 :

$t_0(M \to M) = \frac{4+1}{(4+1)+(1+1)+(0+1)} = 5/8$ $\quad t_0(I \to M) = \frac{0+1}{3.(0+1)} = 1/3 = t_0(I \to I) = t_0(I \to D)$

$t_0(M \to I) = \frac{0+1}{8} = 1/8$ $\quad t_0(D \to M) = t_0(D \to I) = t_0(D \to D) = 1/3$

$t_0(M \to D) = \frac{1+1}{8} = \frac{2}{8}$

$$T_0 = \begin{array}{c} \\ M \\ I \\ D \end{array} \begin{array}{c} \begin{array}{ccc} M & I & D \end{array} \\ \begin{pmatrix} 5/8 & 1/8 & 2/8 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

1 → 2

$t_1(M \to M) = \frac{3+1}{(3+1)+(1+1)+(0+1)} = 4/7$ $\quad t_1(I \to M) = \frac{0+1}{3.(0+1)} = 1/3 = t_1(I \to I) = t_1(I \to D)$

$t_1(M \to I) = \frac{0+1}{7} = 1/7$ $\quad t_1(D \to M) = \frac{1+1}{1+1+0+1+0+1} = 1/2$

$t_1(M \to D) = \frac{1+1}{7} = \frac{2}{7}$ $\quad t_1(D \to D) = t_1(D \to I) = 1/4$

$$T_1 = \begin{array}{c} \\ M \\ I \\ D \end{array} \begin{array}{c} \begin{array}{ccc} M & I & D \end{array} \\ \begin{pmatrix} 4/7 & 1/7 & 2/7 \\ 1/3 & 1/3 & 1/3 \\ 1/2 & 1/4 & 1/4 \end{pmatrix} \end{array}$$

2 → 3

$t_2(M \to M) = \frac{2+1}{(2+1)+(2+1)+(0+1)} = 3/7$ $\quad t_2(I \to M) = \frac{3+1}{3+1+3+1+0+1} = 4/9$

$t_2(M \to I) = \frac{2+1}{7} = 3/7$ $\quad t_2(I \to I) = \frac{3+1}{9} = 4/9$

$t_2(M \to D) = \frac{0+1}{7} = 1/7$ $\quad t_2(I \to D) = \frac{0+1}{9} = 1/9$

$t_2(D \to M) = \frac{0+1}{0+1+1+1+0+1} = 1/4$ $\quad t_2(D \to I) = \frac{1+1}{4} = 1/2$

$t_2(D \to D) = \frac{0+1}{4} = 1/4$

$$T_2 = \begin{array}{c} \\ M \\ I \\ D \end{array} \begin{array}{c} \begin{array}{ccc} M & I & D \end{array} \\ \begin{pmatrix} 3/7 & 3/7 & 1/7 \\ 4/9 & 4/9 & 1/9 \\ 1/4 & 1/2 & 1/4 \end{pmatrix} \end{array}$$

3→4

$t_3(M \to M) = \frac{5+1}{(5+1)+(0+1)+(0+1)} = 6/8$

$t_3(M \to I) = \frac{0+1}{8} = 1/8$

$t_3(M \to D) = \frac{0+1}{8} = \frac{1}{8}$

$t_3(I \to M) = \frac{0+1}{3 \cdot (0+1)} = 1/3 = t_3(I \to I) = t_3(I \to D)$

$t_3(D \to M) = t_3(D \to I) = t_3(D \to D) = 1/3$

$$T_3 = \begin{array}{c} \\ M \\ I \\ D \end{array} \begin{array}{c} \begin{array}{ccc} M & I & D \end{array} \\ \begin{pmatrix} 6/8 & 1/8 & 1/8 \\ 1/3 & 1/3 & 1/3 \\ 1/3 & 1/3 & 1/3 \end{pmatrix} \end{array}$$

	0	1	2	3
M→M	5/8	4/7	3/7	6/8
M→I	2/8	1/7	3/7	1/8
M→D	1/8	2/7	1/7	1/8
I→M	1/3	1/3	4/9	1/3
I→I	1/3	1/3	4/9	1/3
I→D	1/3	1/3	1/9	1/3
D→M	1/3	1/2	1/4	1/3
D→I	1/3	1/4	1/2	1/3
D→D	1/3	1/4	1/4	1/3

Project 4

Problem 11

You are given multiple alignments of protein sequences for two protein families: GTP binding proteins, and a family of ATPases. The task is to determine to which family certain unclassified proteins belong.

1. Run source("profileHMM.R") to import functions which you will use below

```
library(ggplot2)
source("profileHMM.R")
```

2. Read the two alignments 'GTP binding proteins.txt' and 'ATPases.txt' into memory using the function parseAlignment().

```
GTP_alignment <- parseAlignment("C:/Users/marie/Desktop/CBB/SMCB/Project_4_student/data/GTP_binding_pro
ATP_alignment <- parseAlignment("C:/Users/marie/Desktop/CBB/SMCB/Project_4_student/data/ATPases.txt")
```

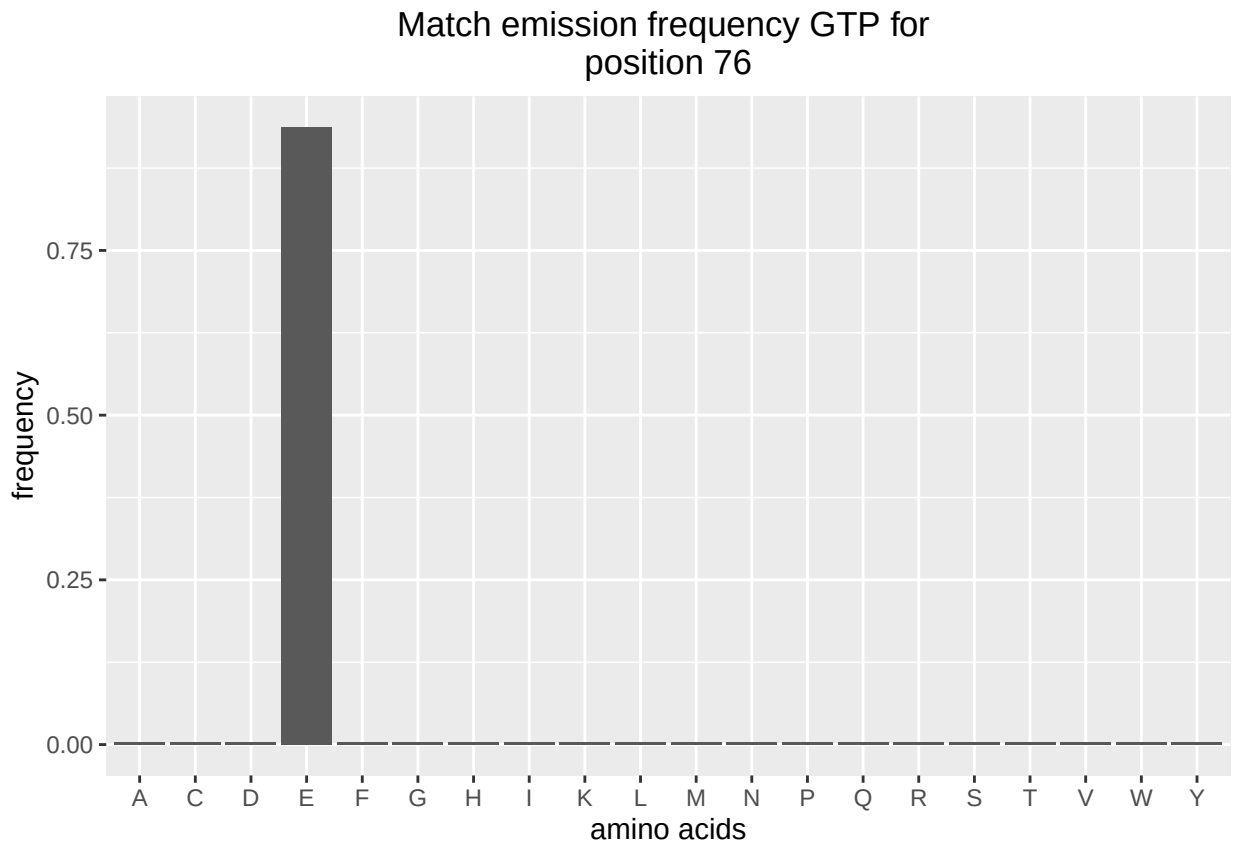
3. Use the function learnHMM() to parametrise two profile HMMs: one for each protein family (multiple alignment).

```
GTP_HMM <- learnHMM(GTP_alignment)
ATP_HMM <- learnHMM(ATP_alignment)
```

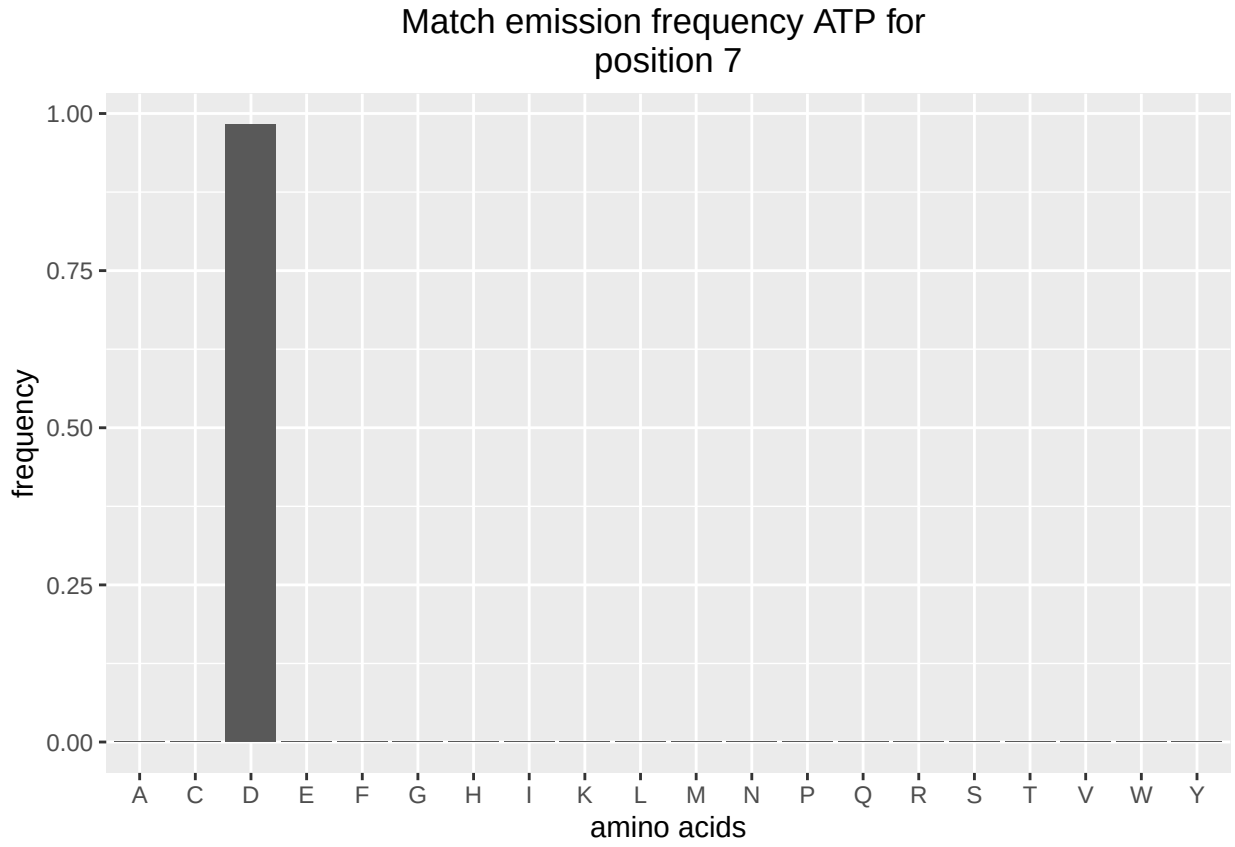
4.Identify the position(s) with the highest match and with the highest insert emission frequencies over all symbols. Plot the respective match and insert emission frequencies for the identified positions.

```
max_emission_GTP <- apply(GTP_HMM$mE, 2, max)
max_pos_emi_GTP <- which.max(max_emission_GTP)

max_emission_ATP <- apply(ATP_HMM$mE, 2, max)
max_pos_emi_ATP <- which.max(max_emission_ATP)

max_insertion_GTP <- apply(GTP_HMM$iE, 2, max)
max_pos_ins_GTP <- which.max(max_insertion_GTP)

max_insertion_ATP <- apply(ATP_HMM$iE, 2, max)
max_pos_ins_ATP <- which.max(max_insertion_ATP)

df_emi_GTP <-data.frame(freq = GTP_HMM$mE[,max_pos_emi_GTP], rownames = rownames(GTP_HMM$mE))
df_emi_ATP <-data.frame(freq = ATP_HMM$mE[,max_pos_emi_ATP], rownames = rownames(ATP_HMM$mE))
df_ins_GTP <-data.frame(freq = GTP_HMM$iE[,max_pos_ins_GTP], rownames = rownames(GTP_HMM$mE))
df_ins_ATP <-data.frame(freq = ATP_HMM$iE[,max_pos_ins_ATP], rownames = rownames(ATP_HMM$mE))
```

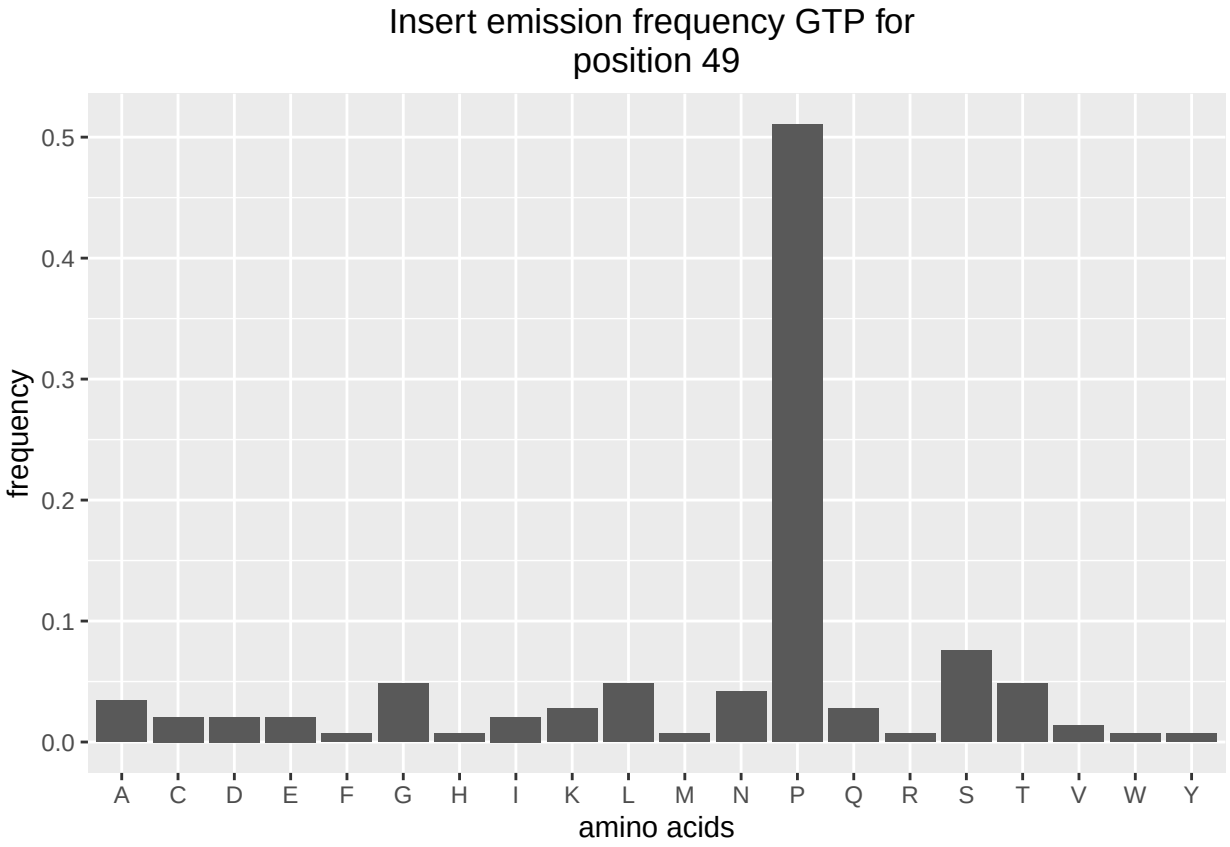
```
ggplot(df_emi_GTP, aes(x = rownames, y = freq))+
  geom_bar(stat = "identity")+
  labs(x="amino acids", y = "frequency", title = paste("Match emission frequency GTP for \nposition",ma
  theme(plot.title = element_text(hjust = 0.5))
```



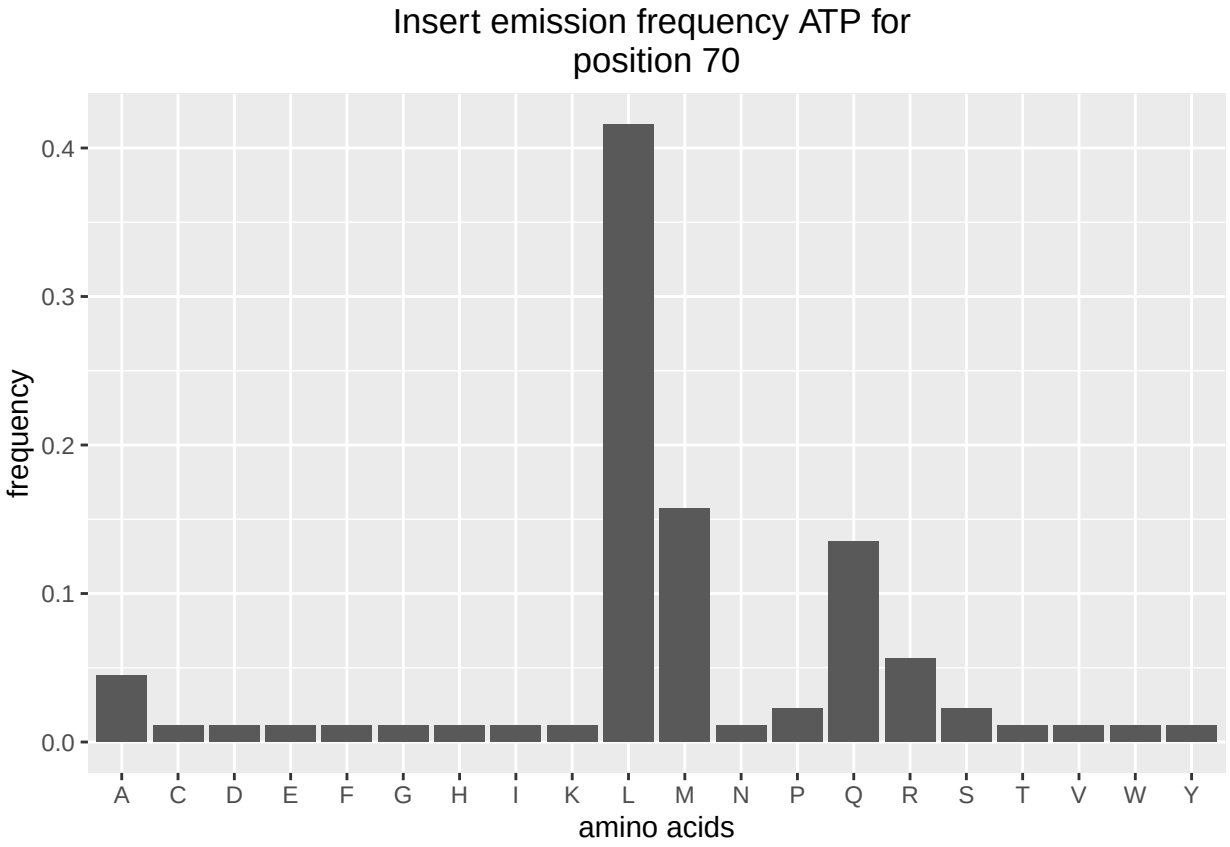

```
ggplot(df_emi_ATP, aes(x = rownames, y = freq))+
  geom_bar(stat = "identity")+
  labs(x="amino acids", y = "frequency", title = paste("Match emission frequency ATP for \nposition",ma
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(df_ins_GTP, aes(x = rownames, y = freq))+
  geom_bar(stat = "identity")+
  labs(x="amino acids", y = "frequency", title = paste("Insert emission frequency GTP for \nposition",ma
  theme(plot.title = element_text(hjust = 0.5))
```




```
ggplot(df_ins_ATP, aes(x = rownames, y = freq))+
  geom_bar(stat = "identity")+
  labs(x="amino acids", y = "frequency", title = paste("Insert emission frequency ATP for \nposition",ma
  theme(plot.title = element_text(hjust = 0.5))
```



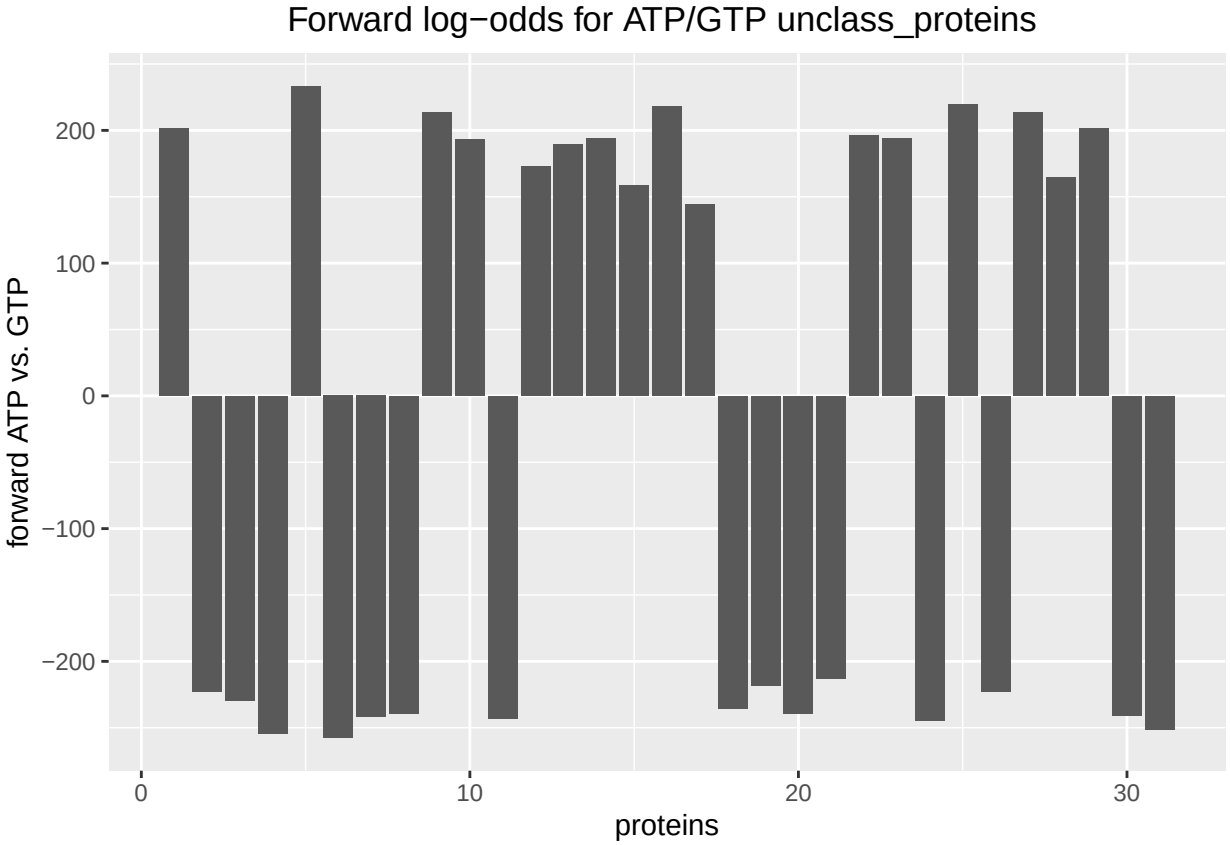
5. The file Unclassified proteins.txt contains 31 protein sequences from unknown families. Load the protein sequences into a list using the parseProteins() function.

```
unclass_proteins <- parseProteins("C:/Users/marie/Desktop/CBB/SMCB/Project_4_student/data/Unclassified_
```

6. The function forward() takes as input a profile HMM M and a sequence x. It returns the log odds ratio of the probability of observing the sequence x given the model M versus the probability of observing the sequence x given the random model R. For each unclassified protein x(i) in the list, apply the forward algorithm for both models M1 and M2 to obtain the log odds ratio. Plot the values q(x_i) and include this in your report. Which proteins in the list belong to which family? Can you clearly decide for each protein?

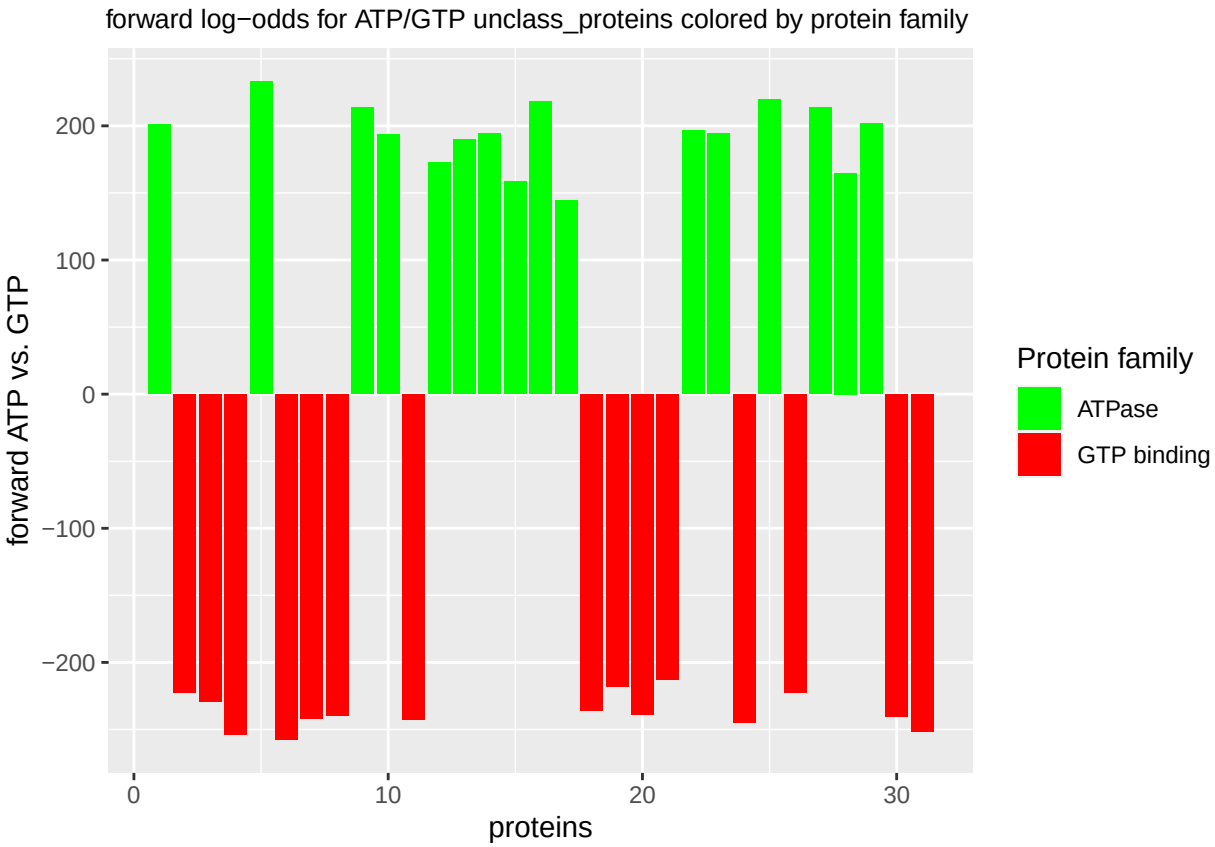
```
#apply forward to GTP and ATP

GTP_forward <- sapply(unclass_proteins, forward, HMM = GTP_HMM)
ATP_forward <- sapply(unclass_proteins, forward, HMM = ATP_HMM)

#get the log odds ratio by taking the difference (calculate q values)
log_odds <- ATP_forward - GTP_forward
names(log_odds) <- 1:length(unclass_proteins)

#plot q values
df_log_odds <-data.frame(log_odds = log_odds, num_proteins = 1:length(log_odds))
ggplot(df_log_odds, aes(x = num_proteins, y = log_odds))+
  geom_bar(stat = "identity")+
```

```
  labs(x="proteins", y = "forward ATP vs. GTP", title = ("Forward log-odds for ATP/GTP unclass_proteins
  theme(plot.title = element_text(hjust = 0.5))
```



One can clearly decide for each protein to which family it belongs to based on the log odds ratio ATP/GTP sign. If its positive then the proteins belongs to the ATPase protein family, if its negative to the GTP binding protein family. This can be visualized:

```
df_log_odds$color <- ifelse(df_log_odds$log_odds > 0, "ATPase", "GTP binding")
ggplot(df_log_odds, aes(x = num_proteins, y = log_odds, fill = color)) +
  geom_bar(stat = "identity") +
  labs(x = "proteins", y = "forward ATP vs. GTP",
       title = "forward log-odds for ATP/GTP unclass_proteins colored by protein family ") +
  theme(plot.title = element_text(hjust = 0.5, size = 10)) +
    scale_fill_manual(values = c("ATPase" = "green", "GTP binding" = "red"),
                      name = "Protein family",
                      labels = c("ATPase", "GTP binding"))
```



So proteins number 1,5,9,10,12,13,14,15,16,17,22,23,25,27,28,29 belong to the ATPase family and proteins 2,3,4,6,7,8,11,18,19,20,21,24,26,30,31 to the GTP binding family.