Project 4 Membereship detection with profile HMMs

Team C - Minghang Li, Xiaocheng Yang, Xinyi Chen

March 15, 2024

Problem 8-10 are all based on the MSA below:

Table 1: Multiple sequence alignment for Problem 8-10

	M	M	I	I	I	M
bat	A	G	-	-	-	C
rat	A	-	A	G	-	C
cat	A	G	-	-	-	C
gnat	-	G	A	A	A	C
goat	A	G	-	-	A	C
	1	2	·	·	·	3

Problem 8: Estimating match emission probabilities

The count $E_i(a)$ and insert emission probability $e_i(a)$, $a \in \mathcal{A} = \{A, C, G, T\}$ are:

Table 2: Estimated match emission probabilities of the profile HMM

	1		**2**		**3**	
	E_1	e_1	E_2	e_2	E_3	e_3
A	5	5/8	1	1/8	1	5/8
C	1	1/8	1	1/8	6	1/8
G	1	1/8	5	5/8	1	1/8
T	1	1/8	1	1/8	1	1/8

Problem 9: Estimating insert emission probabilities

Since the contiguous insert states have the same position in the model, the count $E_i(a)$ and insert emission probability $e_i(a)$, $a \in \mathcal{A} = \{A, C, G, T\}$ are:

Table 3: Estimated insert emission probabilities of the profile HMM

	I	
	E_1	e_1
A	6	6/10
C	1	1/10
G	2	2/10
T	1	1/10

Problem 10: Estimating transition probabilities

The path of each sequence (bat, rat, cat, gnat, goat) from Begin to end is summarized below,

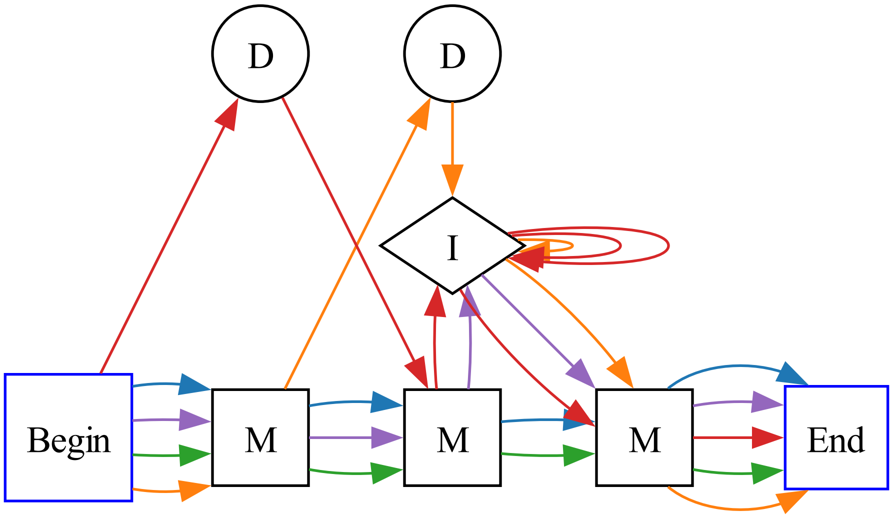


Figure 1: HMM profile

Nodes that are not visited in any path have uniform probability of transitting to its next states (as it only has pseudocounts). The transition probabilities are summarized in the following table:

Table 4: Estimated transition probabilities of the profile HMM

	0		**1**		**2**		**3**	
	E_0	e_0	E_1	e_1	E_2	e_2	E_3	e_3
$M \to M$	5	5/8	4	4/7	3	3/7	6	6/8
$M \to I$	1	1/8	1	1/7	3	3/7	1	1/8
$M \to D$	2	2/8	2	2/7	1	1/7	1	1/8
$I \to M$	1	1/3	1	1/3	4	4/9	1	1/3
$I \to I$	1	1/3	1	1/3	4	4/9	1	1/3
$I \to D$	1	1/3	1	1/3	1	1/9	1	1/3
$D \to M$	1	1/3	2	2/4	1	1/4	1	1/3
$D \to I$	1	1/3	1	1/4	2	2/4	1	1/3
$D \to D$	1	1/3	1	1/4	1	1/4	1	1/3

Problem 11: Protein family membership classification

Import functions and read alignments

```
# import functions
source("code/profileHMM.R", local = knitr::knit_global())
```

```
# read alignments
GTPase <- parseAlignment("./data/GTP_binding_proteins.txt")
ATPase <- parseAlignment("./data/ATPases.txt")
```

Learn HMM from two protein families

```
HMM_GTPase <- learnHMM(GTPase)
HMM_ATPase <- learnHMM(ATPase)
```

Identify position(s) with the highest match and insert emission frequencies over all symbols.

In ATPase

The position(s) with the highest **match** emission frequency over all symbols is/are:

```
mE_max_ATPase <- which(HMM_ATPase$mE == max(HMM_ATPase$mE, na.rm = TRUE),
                       arr.ind = TRUE)
mE_max_ATPase[2] - 1
```

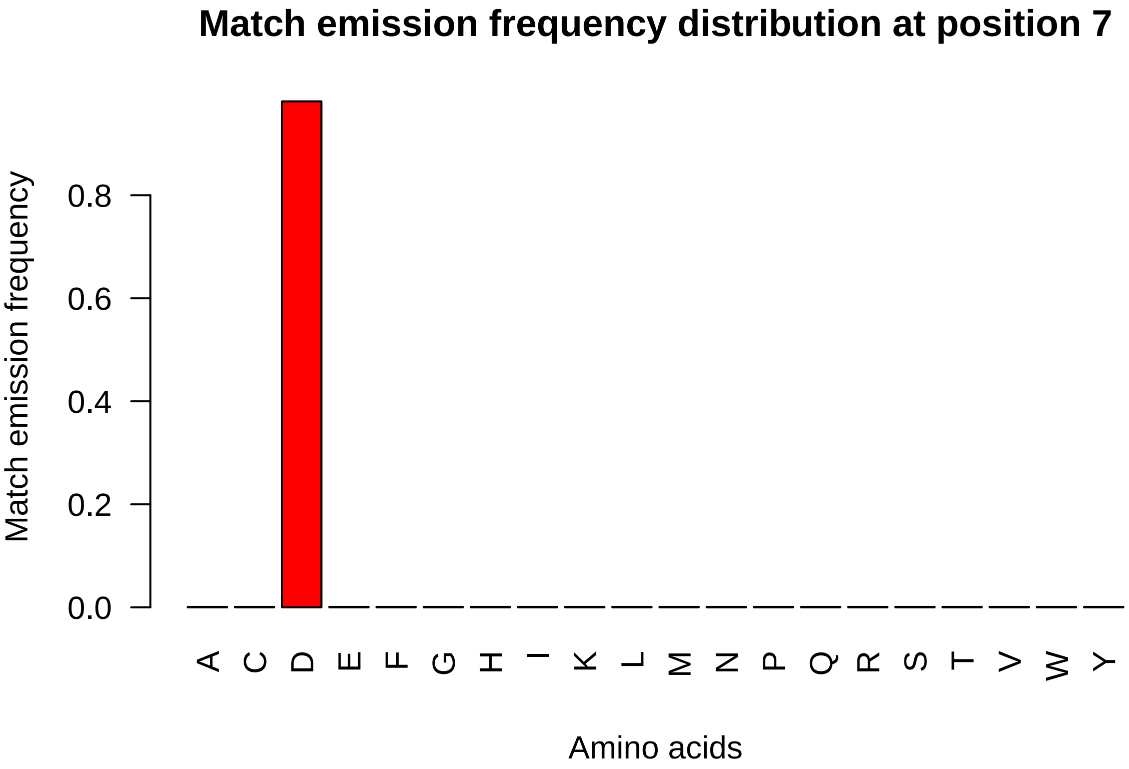
```
## [1] 7
```

The position(s) with the highest **insert** emission frequency over all symbols is/are:

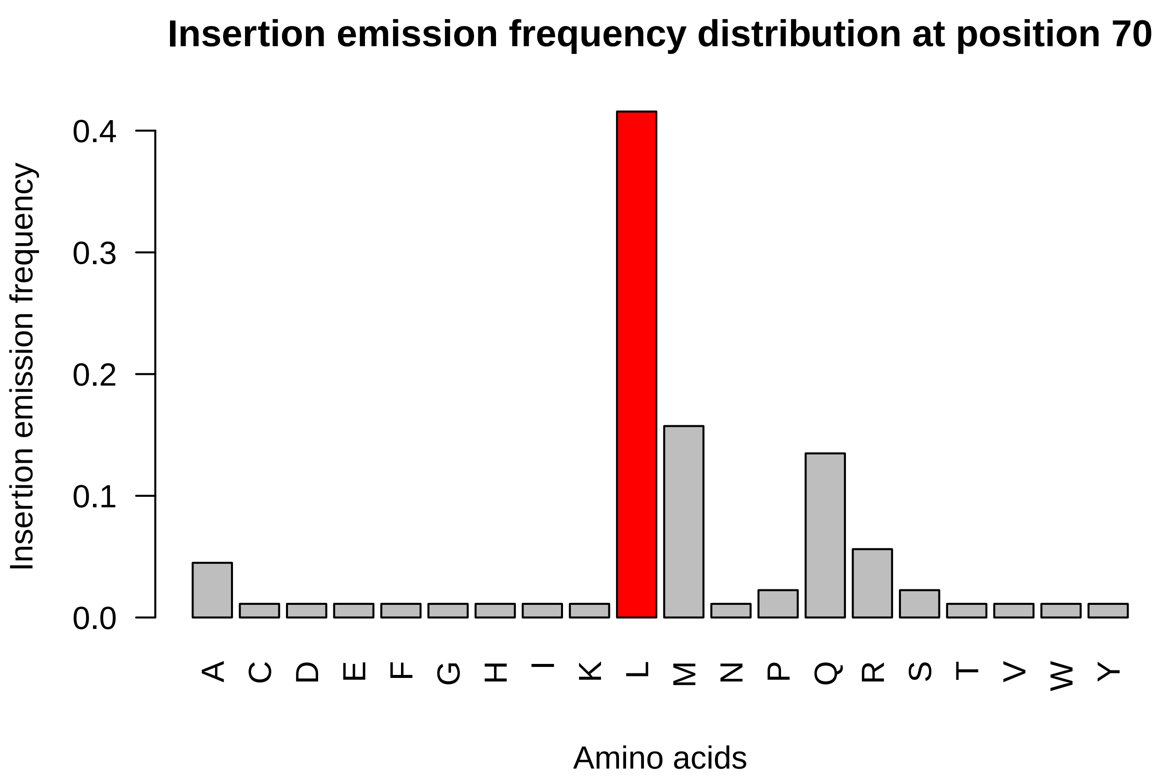
```
iE_max_ATPase <- which(HMM_ATPase$iE == max(HMM_ATPase$iE, na.rm = TRUE),
                       arr.ind = TRUE)
iE_max_ATPase[2] - 1
```

```
## [1] 70
```

```r
# generate barplot color sequence c() with length of alphabet
# all colors are gray except for position at mE_max_ATPase[1] is red
color_mE_max_ATPase <- c(rep("gray", length(HMM_ATPase$alphabet)))
color_mE_max_ATPase[mE_max_ATPase[1]] <- "red"
barplot(HMM_ATPase$mE[, mE_max_ATPase[2]],
        col = color_mE_max_ATPase,
        names = HMM_ATPase$alphabet,
        las = 2,
        main = paste("Match emission frequency distribution at position",
                     mE_max_ATPase[2] - 1),
        xlab = "Amino acids",
        ylab = "Match emission frequency")
```



```r
# generate barplot color sequence c() with length of alphabet
# all colors are gray except for position at mE_max_ATPase[1] is red
color_iE_max_ATPase <- c(rep("gray", length(HMM_ATPase$alphabet)))
color_iE_max_ATPase[iE_max_ATPase[1]] <- "red"
barplot(HMM_ATPase$iE[, iE_max_ATPase[2]],
        col = color_iE_max_ATPase,
        names = HMM_ATPase$alphabet,
        las = 2,
        main = paste("Insertion emission frequency distribution at position",
                     iE_max_ATPase[2] - 1),
        xlab = "Amino acids",
        ylab = "Insertion emission frequency")
```



In GTPase (GTP binding protein)

The position(s) with the highest **match** emission frequency over all symbols is/are:

```
mE_max_GTPase <- which(HMM_GTPase$mE == max(HMM_GTPase$mE, na.rm = TRUE),
                       arr.ind = TRUE)
mE_max_GTPase[2] - 1
```

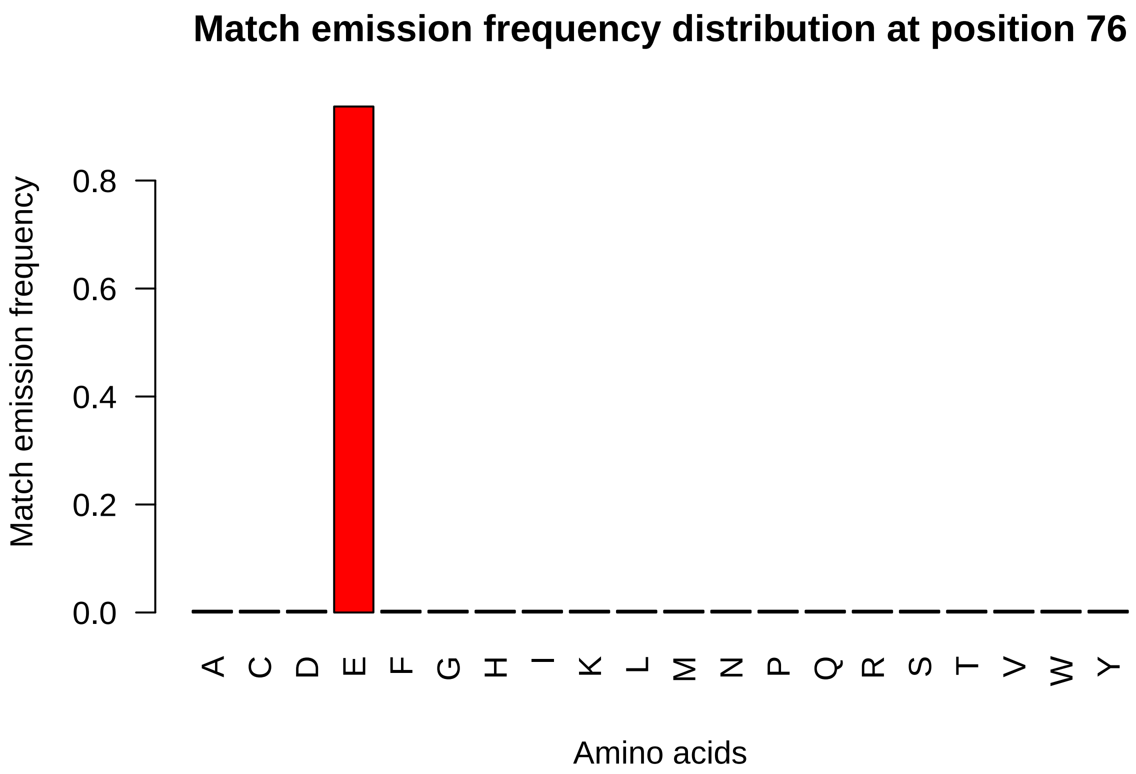
```
## [1] 76
```

The position(s) with the highest **insert** emission frequency over all symbols is/are:

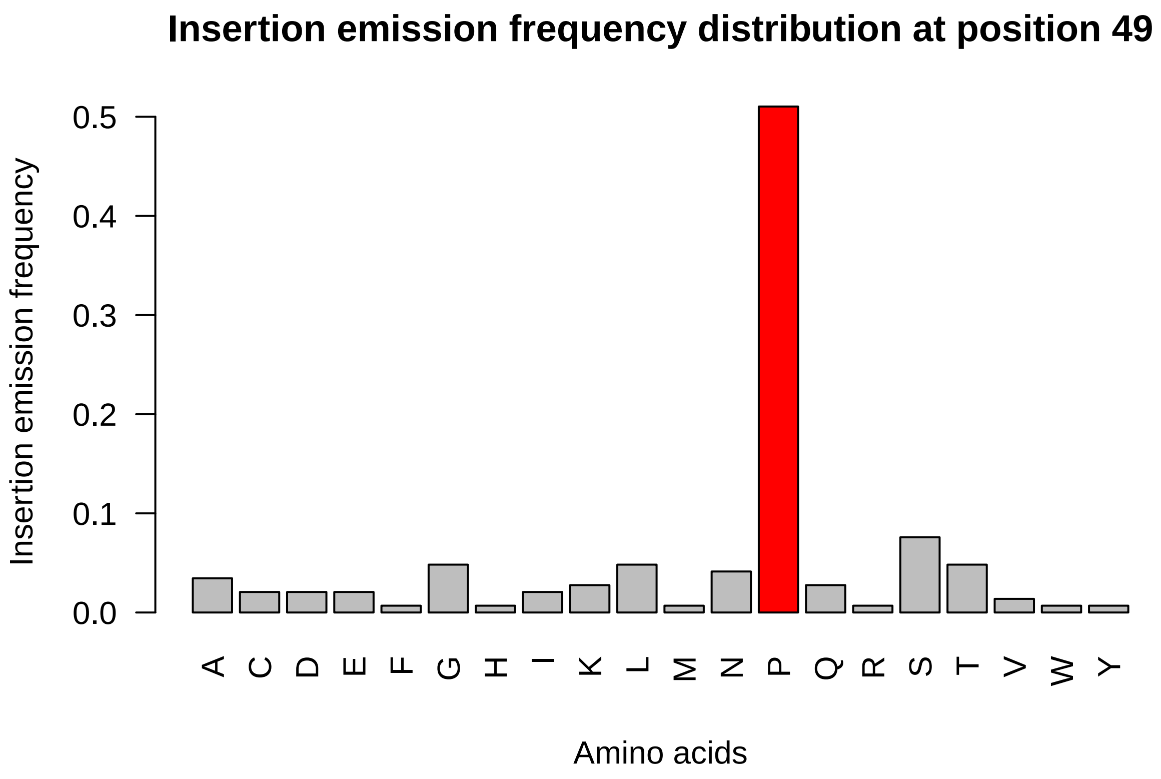
```
iE_max_GTPase <- which(HMM_GTPase$iE == max(HMM_GTPase$iE, na.rm = TRUE),
                       arr.ind = TRUE)
iE_max_GTPase[2] - 1
```

```
## [1] 49
```

```
# generate barplot color sequence c() with length of alphabet
# all colors are gray except for position at mE_max_ATPase[1] is red
color_mE_max_GTPase <- c(rep("gray", length(HMM_GTPase$alphabet)))
color_mE_max_GTPase[mE_max_GTPase[1]] <- "red"
barplot(HMM_GTPase$mE[, mE_max_GTPase[2]],
        col = color_mE_max_GTPase,
        names = HMM_GTPase$alphabet,
        las = 2,
        main = paste("Match emission frequency distribution at position",
                     mE_max_GTPase[2] - 1),
        xlab = "Amino acids",
        ylab = "Match emission frequency")
```



```r
# generate barplot color sequence c() with length of alphabet
# all colors are gray except for position at mE_max_ATPase[1] is red
color_iE_max_GTPase <- c(rep("gray", length(HMM_GTPase$alphabet)))
color_iE_max_GTPase[iE_max_GTPase[1]] <- "red"
barplot(HMM_GTPase$iE[, iE_max_GTPase[2]],
        col = color_iE_max_GTPase,
        names = HMM_GTPase$alphabet,
        las = 2,
        main = paste("Insertion emission frequency distribution at position",
                     iE_max_GTPase[2] - 1),
        xlab = "Amino acids",
        ylab = "Insertion emission frequency")
```

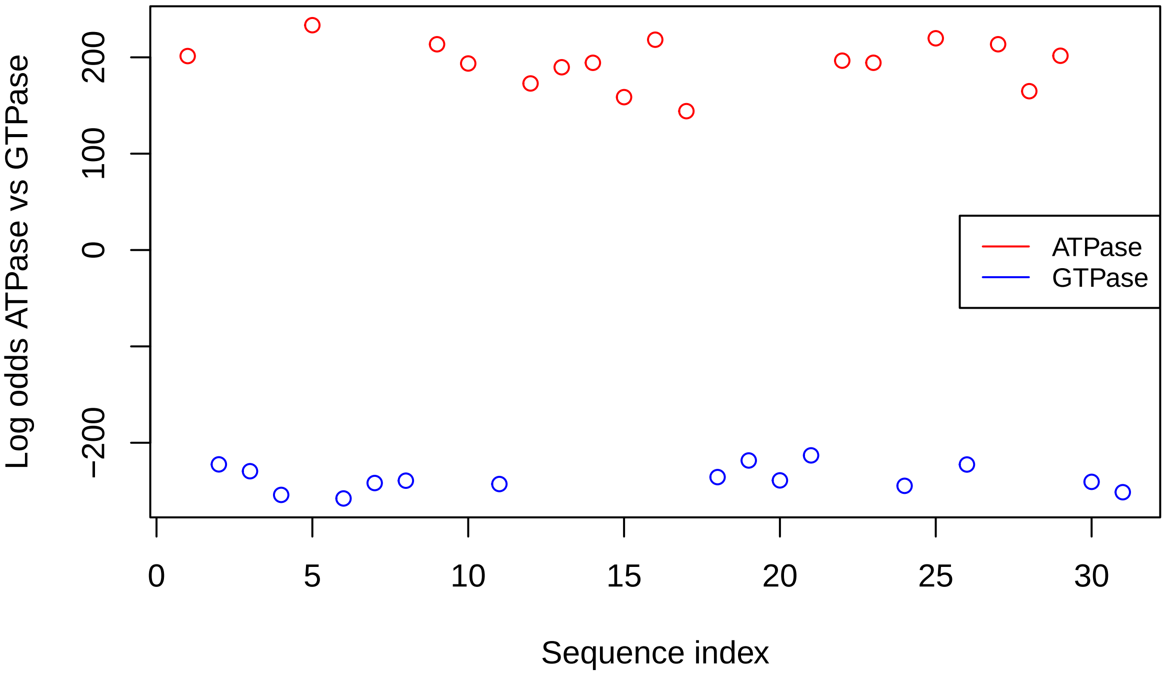


Classify unclassified proteins

```
unclassified_proteins <- parseProteins(proteinsFile = "./data/Unclassified_proteins.txt")
```

```
num_cores <- detectCores()
registerDoParallel(num_cores)
log_odds_ATPase <- foreach(seq=unclassified_proteins, .combine = c) %dopar% {
  forward(HMM = HMM_ATPase, seq = seq)
}
log_odds_GTPase <- foreach(seq=unclassified_proteins, .combine = c) %dopar% {
  forward(HMM = HMM_GTPase, seq = seq)
}
stopImplicitCluster()
```

```
q <- log_odds_ATPase - log_odds_GTPase
names(q) <- NULL
plot(q, col=ifelse(q >= 0, "red", "blue"),
     xlab = "Sequence index",
     ylab = "Log odds ATPase vs GTPase")
legend("right", legend=c("ATPase", "GTPase"), col=c("red", "blue"), lty=1, cex=0.8)
```



We can see a clear separation between points that are classified as ATPases and those classified as GTPases. The log ratio of ATPase vs GTPase is always way larger than 0 or way smaller than 0, indicating high likelihood.

The index of proteins that are classified as ATPases are:

```
# remove list name
which(q >= 0)
```

```
##  [1]  1  5  9 10 12 13 14 15 16 17 22 23 25 27 28 29
```

The index of proteins that are classified as GTPases are:

```
which(q < 0)
```

```
##  [1]  2  3  4  6  7  8 11 18 19 20 21 24 26 30 31
```