Assessment 4 : SLE777

Nishita Sood

2025-10-01

Assignment 4: R Project

PART-2: Examining biological sequence diversity

Downloading and loading sequences for both organisms

This code first loads the R.utils library, which is used to decompress files. It then defines the URL for compressed FASTA files containing Escherichia coli and Salmonella coding sequences, downloads the files, decompresses them using gunzip, and finally lists the files in the current directory to confirm the successful download and extraction.This step ensures we have complete coding-sequence (CDS) data for both organisms.

```
suppressPackageStartupMessages({
  library("R.utils") # general utilities like zip and unzip
  library("seqinr") # is a package designed to process and analyse sequence data

# Downloading E. coli coding sequences
URL="http://ftp.ensemblgenomes.org/pub/bacteria/release-53/fasta/bacteria_0_collection/escherichia_coli
download.file(URL,destfile="ecoli_cds.fa.gz")
gunzip("ecoli_cds.fa.gz", overwrite=TRUE)
list.files()

# Generating the summary of first few sequences of data
cds_ecoli <- seqinr::read.fasta("ecoli_cds.fa")
str(head(cds_ecoli))

# Download Salmonella coding sequences
URL="https://ftp.ensemblgenomes.ebi.ac.uk/pub/bacteria/release-62/fasta/bacteria_50_collection/salmonel
download.file(URL,destfile="salmonella_cds.fa.gz")
gunzip("salmonella_cds.fa.gz", overwrite=TRUE)
list.files()

# Generating the summary of first few sequences of data
cds_salmonella <- seqinr::read.fasta("salmonella_cds.fa")
str(head(cds_salmonella))
})
```

```
## List of 6
##  $ AAC73112: 'SeqFastadna' chr [1:66] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73112"
##   ..- attr(*, "Annot")= chr ">AAC73112 cds chromosome:ASM584v2:Chromosome:190:255:1 gene:b0001 gene_
##  $ AAC73113: 'SeqFastadna' chr [1:2463] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "AAC73113"
##   ..- attr(*, "Annot")= chr ">AAC73113 cds chromosome:ASM584v2:Chromosome:337:2799:1 gene:b0002 gene
##  $ AAC73114: 'SeqFastadna' chr [1:933] "a" "t" "g" "g" ...
```

```
##   ..- attr(*, "name")= chr "AAC73114"
##   ..- attr(*, "Annot")= chr ">AAC73114 cds chromosome:ASM584v2:Chromosome:2801:3733:1 gene:b0003 gene
##  $ AAC73115: 'SeqFastadna' chr [1:1287] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73115"
##   ..- attr(*, "Annot")= chr ">AAC73115 cds chromosome:ASM584v2:Chromosome:3734:5020:1 gene:b0004 gene
##  $ AAC73116: 'SeqFastadna' chr [1:297] "g" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "AAC73116"
##   ..- attr(*, "Annot")= chr ">AAC73116 cds chromosome:ASM584v2:Chromosome:5234:5530:1 gene:b0005 gene
##  $ AAC73117: 'SeqFastadna' chr [1:777] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "AAC73117"
##   ..- attr(*, "Annot")= chr ">AAC73117 cds chromosome:ASM584v2:Chromosome:5683:6459:-1 gene:b0006 ge
## List of 6
##  $ ENSB:rzMRPrOXj2f6n3A: 'SeqFastadna' chr [1:417] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "ENSB:rzMRPrOXj2f6n3A"
##   ..- attr(*, "Annot")= chr ">ENSB:rzMRPrOXj2f6n3A cds primary_assembly:ASM551873v1:contig00038:4502
##  $ ENSB:x3MzlaR1iM6Op6v: 'SeqFastadna' chr [1:402] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:x3MzlaR1iM6Op6v"
##   ..- attr(*, "Annot")= chr ">ENSB:x3MzlaR1iM6Op6v cds primary_assembly:ASM551873v1:contig00003:3217
##  $ ENSB:btfdQZt4_vWjqOV: 'SeqFastadna' chr [1:2613] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:btfdQZt4_vWjqOV"
##   ..- attr(*, "Annot")= chr ">ENSB:btfdQZt4_vWjqOV cds primary_assembly:ASM551873v1:contig00009:9004
##  $ ENSB:IZ64ldiL3-oOAYf: 'SeqFastadna' chr [1:1314] "a" "t" "g" "t" ...
##   ..- attr(*, "name")= chr "ENSB:IZ64ldiL3-oOAYf"
##   ..- attr(*, "Annot")= chr ">ENSB:IZ64ldiL3-oOAYf cds primary_assembly:ASM551873v1:contig00025:3363
##  $ ENSB:9CMZHlFso1POuRQ: 'SeqFastadna' chr [1:165] "a" "t" "g" "c" ...
##   ..- attr(*, "name")= chr "ENSB:9CMZHlFso1POuRQ"
##   ..- attr(*, "Annot")= chr ">ENSB:9CMZHlFso1POuRQ cds primary_assembly:ASM551873v1:contig00002:3745
##  $ ENSB:WDGo7WHsQXy-ZVK: 'SeqFastadna' chr [1:918] "a" "t" "g" "a" ...
##   ..- attr(*, "name")= chr "ENSB:WDGo7WHsQXy-ZVK"
##   ..- attr(*, "Annot")= chr ">ENSB:WDGo7WHsQXy-ZVK cds primary_assembly:ASM551873v1:contig00004:2703
```

Both E. coli and Salmonella CDS files loaded successfully. These lists contain thousands of coding sequences that will be used for comparative genome analysis.

STEP 1 Counting the Number of coding sequences

Here at very first step,we count the total number of CDS enteries in each organism and display them in a comparison table.

```
# Counting the number of CDS in each organism
n_ecoli <- length(cds_ecoli)
n_salmonella <- length(cds_salmonella)

# Creating a comparison table
cds_count_table <- data.frame(
  Organism = c("Escherichia coli (K-12 MG1655)",
               "Salmonella enterica subsp. enterica serovar Weltevreden"),
  Coding_Sequences = c(n_ecoli, n_salmonella)
)
cds_count_table
```

```
##                                              Organism Coding_Sequences
## 1                          Escherichia coli (K-12 MG1655)            4239
## 2 Salmonella enterica subsp. enterica serovar Weltevreden            4585
```

The table shows that E. coli has a slightly different number of coding sequences compared to Salmonella,

reflecting species-specific genome size and gene content.

STEP 2: Total Coding DNA Length

We calculate the total number of base pairs (bp) contributed by all coding sequences.This helps quantify overall genome coding capacity.

```
# Extracting gene lengths for both organisms
len_ecoli <- as.numeric(summary(cds_ecoli)[,1])
len_salmonella <- as.numeric(summary(cds_salmonella)[,1])

# Calculating total coding DNA (sum of all CDS lengths)
total_ecoli <- sum(len_ecoli)
total_salmonella <- sum(len_salmonella)

# Creating comparison table
total_coding_table <- data.frame(
  Organism = c("Escherichia coli (K-12 MG1655)",
               "Salmonella enterica subsp. enterica serovar Weltevreden"),
  Total_Coding_bp = c(total_ecoli, total_salmonella)
)
# Displaying the table
total_coding_table
```

```
##                                                Organism Total_Coding_bp
## 1                          Escherichia coli (K-12 MG1655)         3978528
## 2 Salmonella enterica subsp. enterica serovar Weltevreden         4294851
```

Salmonella and E. coli differ in total coding base pairs, which relates to genome complexity and the number of functional genes.

STEP 3: CDS Length Distribution (Mean, Median & Boxplot)

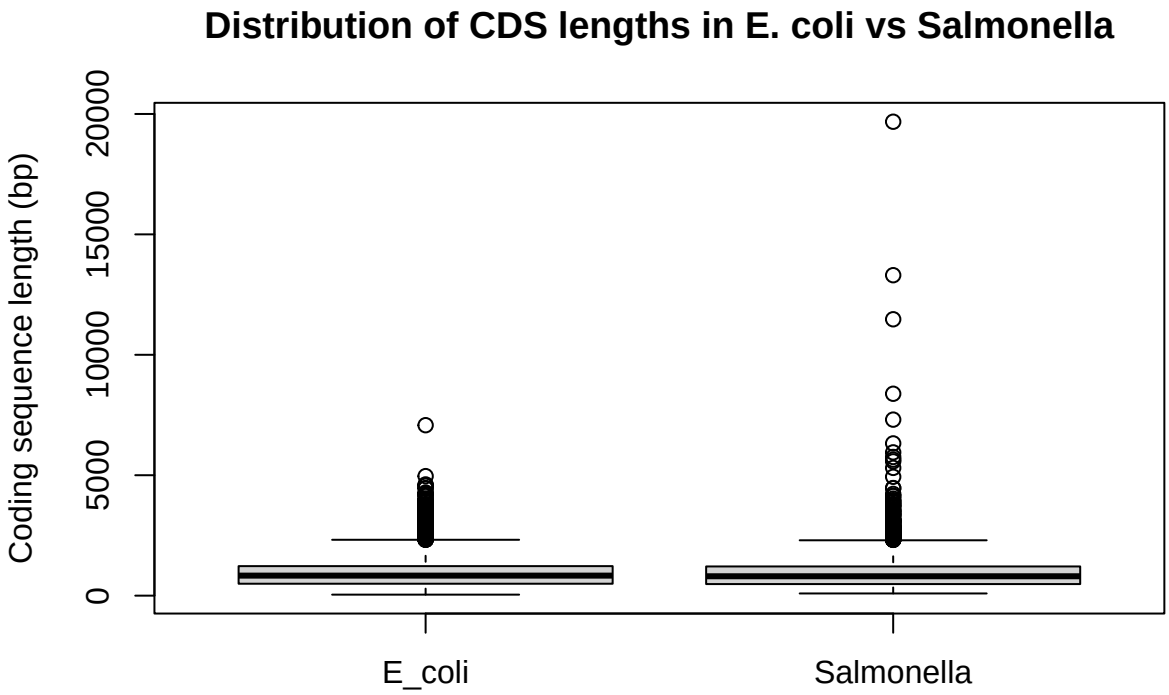
We compute descriptive statistics (mean & median) and visualise the CDS length distributions using a boxplot to show variability.

```
# Mean and median CDS lengths
mean_ecoli <- mean(len_ecoli)
median_ecoli <- median(len_ecoli)

mean_salmonella <- mean(len_salmonella)
median_salmonella <- median(len_salmonella)

# Comparison Table
length_summary <- data.frame(
  Organism = c("Escherichia coli (K-12 MG1655)",
               "Salmonella enterica subsp. enterica serovar Weltevreden"),
  Mean_Length_bp = c(mean_ecoli, mean_salmonella),
  Median_Length_bp = c(median_ecoli, median_salmonella)
)
# Displaying the table
length_summary
```

```
##                                                Organism Mean_Length_bp
## 1                          Escherichia coli (K-12 MG1655)       938.5534
## 2 Salmonella enterica subsp. enterica serovar Weltevreden       936.7178
##   Median_Length_bp
```

```
## 1              831
## 2              804
```

```r
# Boxplot comparison of CDS lengths
boxplot(list(E_coli = len_ecoli, Salmonella = len_salmonella),
        ylab = "Coding sequence length (bp)",
        main = "Distribution of CDS lengths in E. coli vs Salmonella")
```



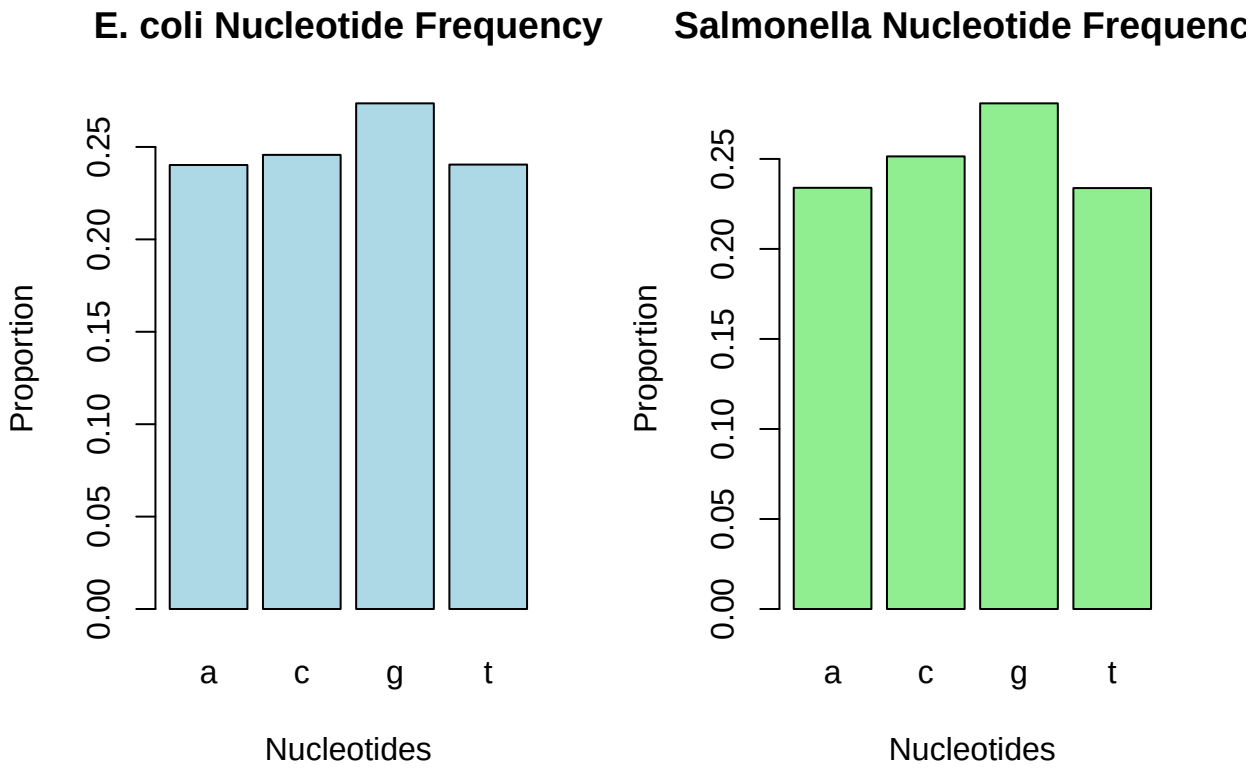
The mean and median CDS lengths are similar between both species, though E. coli shows slightly longer genes overall. The boxplot indicates comparable variability in gene lengths.

STEP 4: Nucleotide and Amino Acid frequency

Next, We compare the base composition (A, T, G, C) and amino-acid composition of the two organisms. Differences in these frequencies can reveal genomic and proteomic biases.

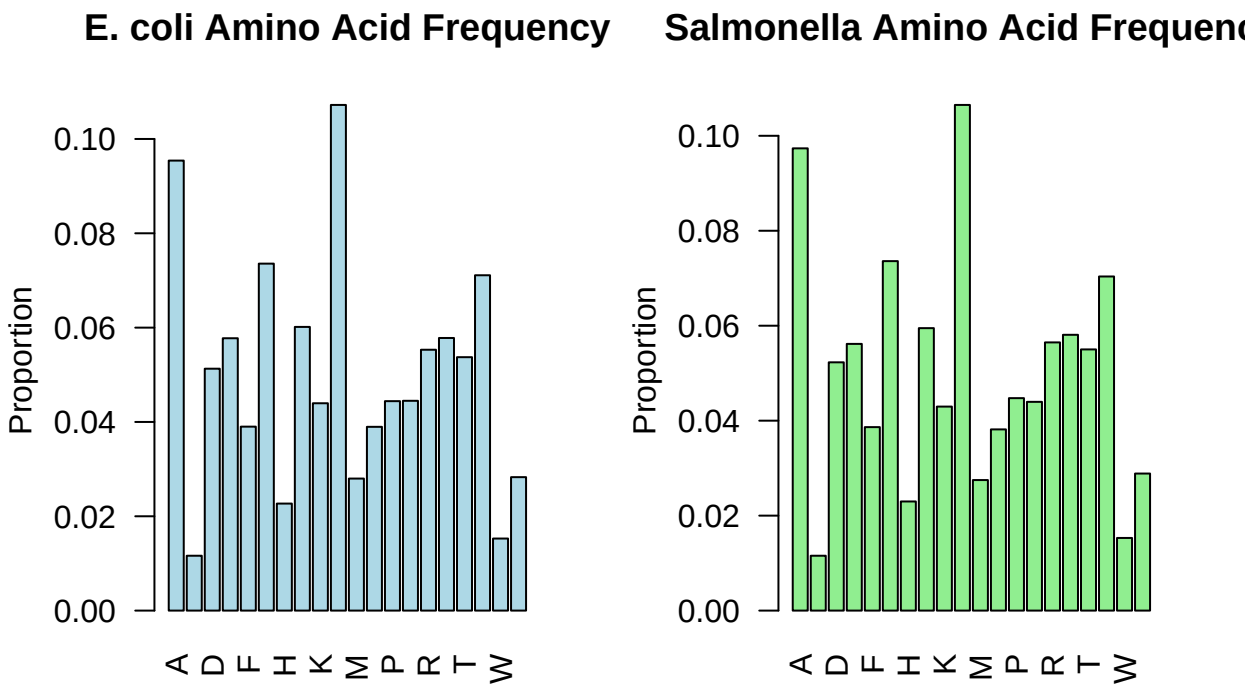
```r
# Combining all the CDS for each organism into single DNA sequences
dna_ecoli <- unlist(cds_ecoli)
dna_salmonella <- unlist(cds_salmonella)

# Counting and normalising nucleotide frequencies (C,G,A,T)
freq_ecoli <- count(dna_ecoli, 1)
freq_salmonella <- count(dna_salmonella, 1)
prop_ecoli <- freq_ecoli / sum(freq_ecoli)
prop_salmonella <- freq_salmonella / sum(freq_salmonella)

# Barplots of nucleotide proportions
par(mfrow = c(1, 2))
```

```r
barplot(prop_ecoli,
        main = "E. coli Nucleotide Frequency",
        xlab = "Nucleotides", ylab = "Proportion", col = "lightblue")
barplot(prop_salmonella,
        main = "Salmonella Nucleotide Frequency",
        xlab = "Nucleotides", ylab = "Proportion", col = "lightgreen")
```




```r
par(mfrow = c(1, 1))   # reset layout

# Translating CDS to proteins
prot_ecoli <- lapply(cds_ecoli, translate)
prot_salmonella <- lapply(cds_salmonella, translate)

# Defining amino acid alphabet (exclude stop codon '*')
aa <- unique(unlist(prot_ecoli))
aa <- aa[aa != "*"]

# Counting amino-acid frequencies (excluding stop codon '*')
aa_freq_ecoli <- count(unlist(prot_ecoli), 1, alphabet = aa)
aa_freq_salmonella <- count(unlist(prot_salmonella), 1, alphabet = aa)
aa_prop_ecoli <- aa_freq_ecoli / sum(aa_freq_ecoli)
aa_prop_salmonella <- aa_freq_salmonella / sum(aa_freq_salmonella)

# Barplots for amino acid proprotions
par(mfrow = c(1, 2))
barplot(aa_prop_ecoli,
```

```
        main = "E. coli Amino Acid Frequency",
        las = 2, ylab = "Proportion", col = "lightblue")
barplot(aa_prop_salmonella,
        main = "Salmonella Amino Acid Frequency",
        las = 2, ylab = "Proportion", col = "lightgreen")
```




```
par(mfrow = c(1, 1))
```

Both organisms show similar base compositions dominated by A and T, typical for bacteria. Amino acid frequencies show conservation in major residues like Leu, Ala, and Gly, but minor variations suggest species-specific coding preferences.

STEP 5: Codon Usage Bias

We examine Relative Synonymous Codon Usage (RSCU) to assess codon-bias patterns.The mean & SD of RSCU values are compared between the two organisms.

```
# Codon usage for E. coli
uco_ecoli <- uco(unlist(cds_ecoli), index="rscu", as.data.frame=TRUE)

# Codon usage for Salmonella
uco_salmonella <- uco(unlist(cds_salmonella), index="rscu", as.data.frame=TRUE)

# Displaying first few rows to check structure
head(uco_ecoli)
```

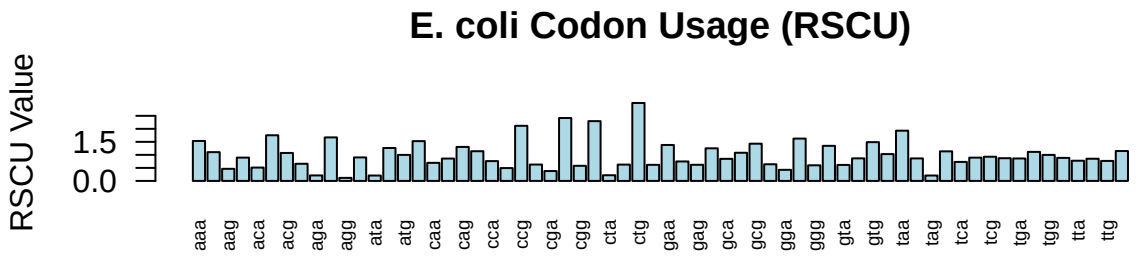
```
##      AA codon   eff        freq      RSCU
```

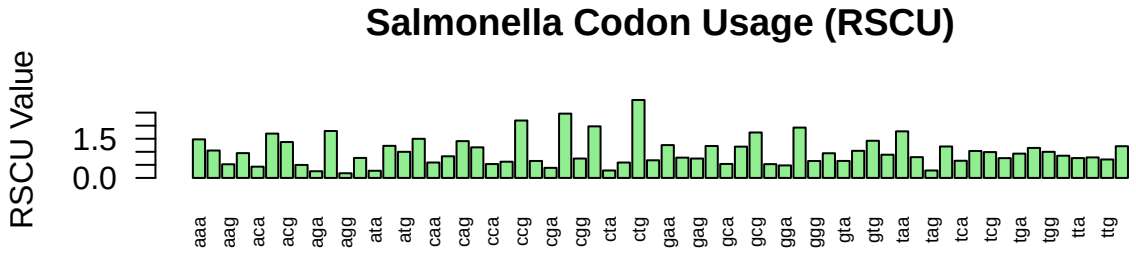
```
## aaa Lys   aaa 44592 0.033624496 1.5346652
## aac Asn   aac 28454 0.021455674 1.1049453
## aag Lys   aag 13521 0.010195479 0.4653348
## aat Asn   aat 23049 0.017380046 0.8950547
## aca Thr   aca  9116 0.006873899 0.5133967
## acc Thr   acc 31139 0.023480292 1.7536924
```

```
head(uco_salmonella)
```

```
##      AA codon   eff        freq      RSCU
## aaa Lys   aaa 45272 0.031622983 1.4767264
## aac Asn   aac 28589 0.019969727 1.0498889
## aag Lys   aag 16042 0.011205511 0.5232736
## aat Asn   aat 25872 0.018071873 0.9501111
## aca Thr   aca  8453 0.005904512 0.4307590
## acc Thr   acc 33302 0.023261808 1.6970469
```

```
# Plot comparison of RSCU values
par(mfrow=c(2,1))
barplot(uco_ecoli$RSCU, names.arg=uco_ecoli$codon, las=2, cex.names=0.6,
        main="E. coli Codon Usage (RSCU)", col="lightblue", ylab="RSCU Value")
barplot(uco_salmonella$RSCU, names.arg=uco_salmonella$codon, las=2, cex.names=0.6,
        main="Salmonella Codon Usage (RSCU)", col="lightgreen", ylab="RSCU Value")
```






```
par(mfrow=c(1,1))

# Calculating basic statistics for codon bias
mean_rscu_ecoli <- mean(uco_ecoli$RSCU, na.rm=TRUE)
```

```r
sd_rscu_ecoli <- sd(uco_ecoli$RSCU, na.rm=TRUE)

mean_rscu_salmonella <- mean(uco_salmonella$RSCU, na.rm=TRUE)
sd_rscu_salmonella <- sd(uco_salmonella$RSCU, na.rm=TRUE)

# Creating summary table for codon bias comparison
codon_bias_summary <- data.frame(
  Organism = c("E. coli (K-12 MG1655)",
               "Salmonella enterica subsp. enterica serovar Weltevreden"),
  Mean_RSCU = c(mean_rscu_ecoli, mean_rscu_salmonella),
  SD_RSCU = c(sd_rscu_ecoli, sd_rscu_salmonella)
)
codon_bias_summary
```

```
##                                                Organism Mean_RSCU   SD_RSCU
## 1                                E. coli (K-12 MG1655)         1 0.5549891
## 2 Salmonella enterica subsp. enterica serovar Weltevreden         1 0.5561407
```

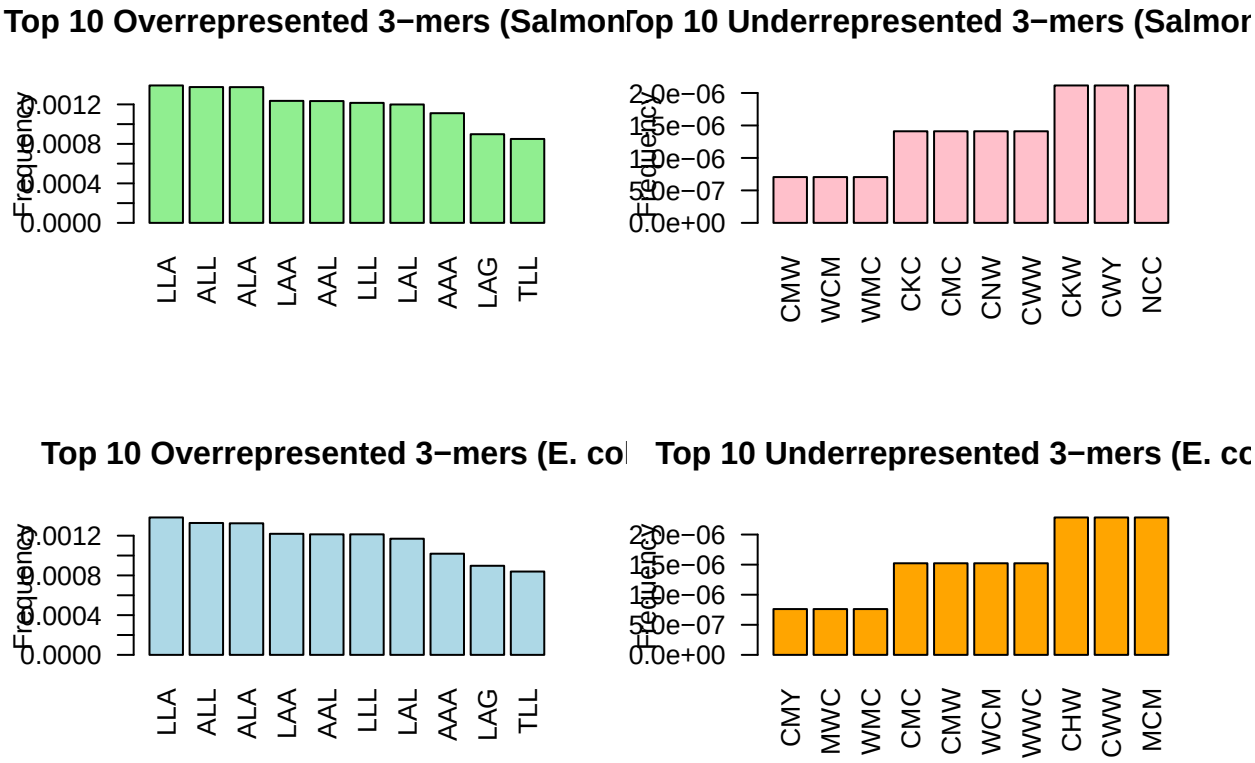
The RSCU plots reveal that certain codons are used more frequently than others, indicating codon bias. The average RSCU and SD values suggest E. coli has slightly stronger codon preference, which can affect translation efficiency.

STEP 6: Protein Sequence and K-mer profiling

Finally, we identify the most over- and under-represented amino-acid motifs (k-mers) of length 3 – 5 in the proteins of both organisms.

```r
# Combining all protein sequences into one vector
prot_all_salmonella <- unlist(prot_salmonella)
prot_all_ecoli <- unlist(prot_ecoli)

# Counting k-mers (3 to 5 amino acids long)
# k = 3
k3_salmonella <- count(prot_all_salmonella, wordsize=3, alphabet=aa, freq=TRUE)
# k = 4
k4_salmonella <- count(prot_all_salmonella, wordsize=4, alphabet=aa, freq=TRUE)
# k = 5
k5_salmonella <- count(prot_all_salmonella, wordsize=5, alphabet=aa, freq=TRUE)

# Identifying top and bottom 10 k-mers for Salmonella
top10_k3_salmonella <- head(sort(k3_salmonella, decreasing=TRUE), 10)
bottom10_k3_salmonella <- head(sort(k3_salmonella, decreasing=FALSE), 10)

# Comparison with e coli
k3_ecoli <- count(prot_all_ecoli, wordsize=3, alphabet=aa, freq=TRUE)
top10_k3_ecoli <- head(sort(k3_ecoli, decreasing=TRUE), 10)
bottom10_k3_ecoli <- head(sort(k3_ecoli, decreasing=FALSE), 10)

# Creating the barplots for visualisation
par(mfrow=c(2,2))
barplot(top10_k3_salmonella, las=2, col="lightgreen",
        main="Top 10 Overrepresented 3-mers (Salmonella)",
        ylab="Frequency")
barplot(bottom10_k3_salmonella, las=2, col="pink",
        main="Top 10 Underrepresented 3-mers (Salmonella)",
```

```
        ylab="Frequency")
barplot(top10_k3_ecoli, las=2, col="lightblue",
        main="Top 10 Overrepresented 3-mers (E. coli)",
        ylab="Frequency")
barplot(bottom10_k3_ecoli, las=2, col="orange",
        main="Top 10 Underrepresented 3-mers (E. coli)",
        ylab="Frequency")
```




```
par(mfrow=c(1,1))
```

The k-mer plots show which amino acid motifs are most and least frequent. Differences between E. coli and Salmonella may relate to variations in protein structure, adaptation, or selective pressure.

In summary, the comparative sequence analysis showed that E. coli and Salmonella share similar overall genomic patterns but differ slightly in coding sequence lengths, nucleotide composition, codon usage bias, and k-mer profiles, reflecting subtle species-specific adaptations in their genomes.