Bioinformatics Assignment in R

Raman Butta

2025-09-22

Contents

Part I: Gene Data Frame — Human Genes

I will create a data frame containing metadata for 10 human genes including: - Gene Name - NCBI Accession ID (RefSeq mRNA) - Chromosome Location - Sequence Length (in base pairs)

I will then: - Print the full table - Identify the longest and shortest genes - Calculate the average gene length

```
library(Biostrings)

# Create gene data frame for Homo sapiens
genes_df <- data.frame(
  Gene_Name = c(
    "BRCA1", "TP53", "CFTR", "MYC", "EGFR",
    "KRAS", "PTEN", "RB1", "APC", "VEGFA"
  ),
  Accession_Id = c(
    "NM_007294", "NM_000546", "NM_000492", "NM_002467", "NM_005228",
    "NM_033360", "NM_000314", "NM_000321", "NM_000038", "NM_001025366"
  ),
  Chromosome = c(
    "17", "17", "7", "8", "7",
    "12", "10", "13", "5", "6"
  ),
  Sequence_Length = c(
    7088, 2512, 6070, 3721, 9905,
    5430, 8515, 4768, 10704, 3660
  ),
  stringsAsFactors = FALSE
)

# Print entire data frame
cat("HUMAN GENE DATA FRAME:\n")
```

```
## HUMAN GENE DATA FRAME:
```

```
print(genes_df)
```

```
##    Gene_Name Accession_Id Chromosome Sequence_Length
## 1      BRCA1    NM_007294         17            7088
## 2       TP53    NM_000546         17            2512
## 3       CFTR    NM_000492          7            6070
## 4        MYC    NM_002467          8            3721
## 5       EGFR    NM_005228          7            9905
## 6       KRAS    NM_033360         12            5430
## 7       PTEN    NM_000314         10            8515
## 8        RB1    NM_000321         13            4768
## 9        APC    NM_000038          5           10704
## 10     VEGFA NM_001025366          6            3660
```

```
# Find longest and shortest sequence
longest_gene <- genes_df[which.max(genes_df$Sequence_Length), ]
shortest_gene <- genes_df[which.min(genes_df$Sequence_Length), ]

cat("\n Longest Gene:\n")
```

```
##
##  Longest Gene:
```

```
print(longest_gene$Gene_Name)
```

```
## [1] "APC"
```

```
cat("\n Shortest Gene:\n")
```

```
##
##  Shortest Gene:
```

```
print(shortest_gene$Gene_Name)
```

```
## [1] "TP53"
```

```
# Calculate average sequence length
avg_length <- mean(genes_df$Sequence_Length)
cat("\n Average Sequence Length:", round(avg_length, 2), "bp\n")
```

```
##
##  Average Sequence Length: 6237.3 bp
```

Part II: FASTA Analysis — GC Content & Translation

I will use `rentrez` to: - Fetch the **genomic RefSeq accession** `NG_005905.2` — which is the genomic region for BRCA1 on chromosome 17 - Parse the FASTA to extract the DNA sequence - Compute **GC content** - **Translate** the first ORF (for demonstration) into a protein sequence.

```r
library(Biostrings)
library(rentrez)

# Genomic RefSeq for BRCA1 region (includes introns, exons)
genomic_accession <- "NG_005905.2"  # Homo sapiens BRCA1 RefSeqGene, 193689 bp

cat("Downloading genomic DNA for:", genomic_accession, "\n")
```

```
## Downloading genomic DNA for: NG_005905.2
```

```r
# Fetch genomic FASTA
fasta_record <- entrez_fetch(
  db = "nucleotide",
  id = genomic_accession,
  rettype = "fasta",
  retmode = "text"
)

# Parse FASTA: split lines, remove header, collapse sequence
fasta_lines <- unlist(strsplit(fasta_record, "\n"))
sequence_lines <- fasta_lines[grepl("^[^>]", fasta_lines) & nchar(fasta_lines) > 0]
dna_sequence <- paste(sequence_lines, collapse = "")

# Create DNAString object
dna_seq <- DNAString(dna_sequence)

cat("Genomic DNA downloaded and parsed.\n")
```

```
## Genomic DNA downloaded and parsed.
```

```r
cat("Genomic sequence length:", length(dna_seq), "bp\n")
```

```
## Genomic sequence length: 193689 bp
```

```r
# Compute GC Content
gc_freq <- letterFrequency(dna_seq, letters = c("G", "C"), as.prob = TRUE)
gc_percent <- sum(gc_freq) * 100
cat("GC Content:", round(gc_percent, 2), "%\n")
```

```
## GC Content: 45.05 %
```

```r
# Translate - from position 1, Frame 0 (for demonstration only)
# In reality, CDS starts at position 181 in this RefSeqGene record (see GenBank)
protein_seq <- translate(dna_seq)

cat("\n First 120 nucleotides of genomic DNA:\n")
```

```
##
##  First 120 nucleotides of genomic DNA:
```

```r
cat(toString(subseq(dna_seq, 1, 120)), "\n")
```

```
## TGTGTGTATGAAGTTAACTTCAAAGCAAGCTTCCTGTGCTGAGGGGGTGGGAGGTAAGGGTGTGATGAGGCAGGGCTTCTCCTTTGGCAAAGCCTCTGTAC
```

```r
cat("\n Translated Protein from Frame 0 (first 20 aa - may not be biological):\n")
```

```
##
##  Translated Protein from Frame 0 (first 20 aa - may not be biological):
```

```r
cat(toString(subseq(protein_seq, 1, 20)), "\n")
```

```
## CVYEVNFKASFLC*GGGR*G
```

Part III: NCBI Queries with rentrez — Allobates kingsburyi

I will use `rentrez` to: - Search NCBI Nucleotide database for sequences from Allobates kingsburyi - Retrieve and save FASTA format sequences to a file - Repeat for Protein database

```r
library(rentrez)

species <- "Allobates kingsburyi[Organism]"

cat("Searching NCBI for:", species, "\n\n")
```

```
## Searching NCBI for: Allobates kingsburyi[Organism]
```

```r
# ----------------------------
# NUCLEOTIDE SEQUENCES
# ----------------------------
nucl_search <- entrez_search(db = "nucleotide", term = species)

if (length(nucl_search$ids) == 0) {
  cat("No nucleotide sequences found for", species, "\n")
} else {
  cat("Found", length(nucl_search$ids), "nucleotide records.\n")

  nucl_fasta <- entrez_fetch(
    db = "nucleotide",
    id = nucl_search$ids,
    rettype = "fasta",
    retmode = "text"
  )

  cat("\n Sample Nucleotide FASTA (first 200 chars):\n")
  cat(substr(nucl_fasta, 1, 200), "...\n")

  write(nucl_fasta, file = "Allobates_kingsburyi_nucleotide.fasta")
  cat("\n Saved to: 'Allobates_kingsburyi_nucleotide.fasta'\n")
}
```

```
## Found 17 nucleotide records.
##
##  Sample Nucleotide FASTA (first 200 chars):
## >MT524123.1 Allobates kingsburyi voucher QCAZA68477 large subunit ribosomal RNA gene, partial sequen
## CCTGATTAACCATAAGAGGTCAAGCCTGCCCAGTGACATTTGTTTAACGGCCGCGGTATCCTAACCGTGC
## GAAGGTAGCGT ...
##
##  Saved to: 'Allobates_kingsburyi_nucleotide.fasta'
```

```r
# ----------------------------
# PROTEIN SEQUENCES
# ----------------------------
prot_search <- entrez_search(db = "protein", term = species)

if (length(prot_search$ids) == 0) {
  cat("\n No protein sequences found for", species, "\n")
} else {
  cat("\n Found", length(prot_search$ids), "protein records.\n")

  prot_fasta <- entrez_fetch(
    db = "protein",
    id = prot_search$ids,
    rettype = "fasta",
    retmode = "text"
  )

  cat("\n Sample Protein FASTA (first 200 chars):\n")
  cat(substr(prot_fasta, 1, 200), "...\n")

  write(prot_fasta, file = "Allobates_kingsburyi_protein.fasta")
  cat("\n Saved to: 'Allobates_kingsburyi_protein.fasta'\n")
}
```

```
##
##  Found 11 protein records.
##
##  Sample Protein FASTA (first 200 chars):
## >ATG31804.1 nicotinic acetylcholine receptor beta-2, partial [Allobates kingsburyi]
## MTVLLLLLHLSLFGLVTRSMGTDTEERLVEFLLDPSQYNKLIRPATNGSEQVTVQLMVSLAQLISVHERE
## QIMTTNVWLTQEWXXXXXXXXXXXXXXXXXXXXXXXXWLPDVVL ...
##
##  Saved to: 'Allobates_kingsburyi_protein.fasta'
```