Genomic Modeling of Antimicrobial Resistance in Salmonella enterica: An Advanced Multi-Layer Perceptron Framework Applied to Chicken-Derived Isolates

Dr. Karthikeyan S , Damodharan R, Kazi Tanvir , Ritik Sharma

Vellore Institute of Technology, Vellore, India.

Abstract :

This project presents an advanced framework for predicting antimicrobial resistance (AMR) in *Salmonella enterica* isolates from chicken sources using deep learning models. Starting with the collection and deduplication of *Salmonella* sequences, AMR gene identification was performed via the Resistance Gene Identifier (RGI) tool from the Comprehensive Antimicrobial Resistance Database. Data wrangling and t-SNE visualization enabled the integration and dimensional reduction of high-dimensional genomic data, facilitating better insight into AMR gene-antimicrobial relationships. The dataset was subsequently balanced using random over-sampling (ROSE) and filtered to exclude low-variance features. Finally, a Multi-Layer Perceptron (MLP) model was trained to predict resistance across 11 distinct drug classes, achieving an impressive accuracy of 86.5%. This model offers a promising approach to understanding and anticipating AMR patterns in poultry-related pathogens.

Step 01 - Data Collection

Installing necessary library called BIOPYTHON for handling biologlical data has been Installed

In [1]:
```
!pip install biopython
```

```
Collecting biopython
  Downloading biopython-1.84-cp312-cp312-win_amd64.whl.metadata (13 kB)
Requirement already satisfied: numpy in c:\users\dd\anaconda3\lib\site-packages (from biopython) (1.26.4)
Downloading biopython-1.84-cp312-cp312-win_amd64.whl (2.8 MB)
   ---------------------------------------- 0.0/2.8 MB ? eta -:--:--
   ---------------------------------------- 0.0/2.8 MB ? eta -:--:--
   ---------------------------------------- 0.0/2.8 MB 262.6 kB/s eta 0:00:11
   - -------------------------------------- 0.1/2.8 MB 558.5 kB/s eta 0:00:05
   - -------------------------------------- 0.1/2.8 MB 656.4 kB/s eta 0:00:05
   - -------------------------------------- 0.1/2.8 MB 599.1 kB/s eta 0:00:05
   -- ------------------------------------- 0.2/2.8 MB 655.4 kB/s eta 0:00:04
   -- ------------------------------------- 0.2/2.8 MB 655.9 kB/s eta 0:00:04
   --- ------------------------------------ 0.2/2.8 MB 625.8 kB/s eta 0:00:05
   ---- ----------------------------------- 0.3/2.8 MB 681.0 kB/s eta 0:00:04
   ---- ----------------------------------- 0.3/2.8 MB 703.7 kB/s eta 0:00:04
   ----- ---------------------------------- 0.4/2.8 MB 696.3 kB/s eta 0:00:04
   ----- ---------------------------------- 0.4/2.8 MB 713.5 kB/s eta 0:00:04
   ------ --------------------------------- 0.4/2.8 MB 708.6 kB/s eta 0:00:04
   ------ --------------------------------- 0.5/2.8 MB 704.5 kB/s eta 0:00:04
   ------ --------------------------------- 0.5/2.8 MB 684.7 kB/s eta 0:00:04
   ------- -------------------------------- 0.5/2.8 MB 714.4 kB/s eta 0:00:04
   ------- -------------------------------- 0.6/2.8 MB 708.5 kB/s eta 0:00:04
   -------- ------------------------------- 0.6/2.8 MB 705.3 kB/s eta 0:00:04
   -------- ------------------------------- 0.6/2.8 MB 717.3 kB/s eta 0:00:04
   --------- ------------------------------ 0.6/2.8 MB 712.5 kB/s eta 0:00:04
   --------- ------------------------------ 0.7/2.8 MB 720.5 kB/s eta 0:00:03
   ---------- ----------------------------- 0.7/2.8 MB 717.3 kB/s eta 0:00:03
   ---------- ----------------------------- 0.7/2.8 MB 715.1 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   ----------- ---------------------------- 0.8/2.8 MB 720.7 kB/s eta 0:00:03
   --------------- ------------------------ 1.1/2.8 MB 718.8 kB/s eta 0:00:03
   --------------- ------------------------ 1.1/2.8 MB 710.0 kB/s eta 0:00:03
   --------------- ------------------------ 1.1/2.8 MB 708.3 kB/s eta 0:00:03
   ---------------- ----------------------- 1.2/2.8 MB 719.5 kB/s eta 0:00:03
   ----------------- ---------------------- 1.2/2.8 MB 717.6 kB/s eta 0:00:03
   ----------------- ---------------------- 1.2/2.8 MB 715.8 kB/s eta 0:00:03
   ----------------- ---------------------- 1.2/2.8 MB 714.8 kB/s eta 0:00:03
   ------------------ --------------------- 1.3/2.8 MB 718.7 kB/s eta 0:00:03
   ------------------ --------------------- 1.3/2.8 MB 717.0 kB/s eta 0:00:03
   ------------------- -------------------- 1.4/2.8 MB 720.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.4/2.8 MB 718.6 kB/s eta 0:00:02
   -------------------- ------------------- 1.4/2.8 MB 718.6 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   -------------------- ------------------- 1.5/2.8 MB 716.2 kB/s eta 0:00:02
   ---------------------- ----------------- 1.6/2.8 MB 686.2 kB/s eta 0:00:02
   ------------------------ --------------- 1.7/2.8 MB 720.3 kB/s eta 0:00:02
   ------------------------- -------------- 1.8/2.8 MB 723.5 kB/s eta 0:00:02
   ------------------------- -------------- 1.8/2.8 MB 717.5 kB/s eta 0:00:02
   ------------------------- -------------- 1.8/2.8 MB 720.9 kB/s eta 0:00:02
   -------------------------- ------------- 1.8/2.8 MB 719.7 kB/s eta 0:00:02
   -------------------------- ------------- 1.8/2.8 MB 719.7 kB/s eta 0:00:02
   -------------------------- ------------- 1.9/2.8 MB 710.0 kB/s eta 0:00:02
   ---------------------------- ----------- 2.0/2.8 MB 719.8 kB/s eta 0:00:02
   ---------------------------- ----------- 2.0/2.8 MB 719.8 kB/s eta 0:00:02
   ----------------------------- ---------- 2.0/2.8 MB 721.2 kB/s eta 0:00:02
   ----------------------------- ---------- 2.1/2.8 MB 724.2 kB/s eta 0:00:02
   ----------------------------- ---------- 2.1/2.8 MB 719.0 kB/s eta 0:00:01
   ------------------------------ --------- 2.1/2.8 MB 721.9 kB/s eta 0:00:01
   ------------------------------ --------- 2.2/2.8 MB 724.3 kB/s eta 0:00:01
   ------------------------------ --------- 2.2/2.8 MB 724.3 kB/s eta 0:00:01
   ------------------------------- -------- 2.2/2.8 MB 722.1 kB/s eta 0:00:01
   -------------------------------- ------- 2.3/2.8 MB 720.9 kB/s eta 0:00:01
   -------------------------------- ------- 2.3/2.8 MB 723.7 kB/s eta 0:00:01
   --------------------------------- ------ 2.3/2.8 MB 723.1 kB/s eta 0:00:01
   --------------------------------- ------ 2.3/2.8 MB 722.1 kB/s eta 0:00:01
   --------------------------------- ------ 2.4/2.8 MB 714.3 kB/s eta 0:00:01
   ---------------------------------- ----- 2.4/2.8 MB 725.7 kB/s eta 0:00:01
   ----------------------------------- ---- 2.5/2.8 MB 721.5 kB/s eta 0:00:01
   ----------------------------------- ---- 2.5/2.8 MB 720.8 kB/s eta 0:00:01
   ----------------------------------- ---- 2.5/2.8 MB 720.6 kB/s eta 0:00:01
   ------------------------------------ --- 2.6/2.8 MB 727.9 kB/s eta 0:00:01
   ------------------------------------- -- 2.6/2.8 MB 723.7 kB/s eta 0:00:01
   ------------------------------------- -- 2.7/2.8 MB 728.8 kB/s eta 0:00:01
   -------------------------------------- - 2.7/2.8 MB 724.7 kB/s eta 0:00:01
   -------------------------------------- - 2.7/2.8 MB 726.9 kB/s eta 0:00:01
   ---------------------------------------  2.7/2.8 MB 726.0 kB/s eta 0:00:01
   ---------------------------------------  2.8/2.8 MB 725.5 kB/s eta 0:00:01
   ---------------------------------------  2.8/2.8 MB 718.9 kB/s eta 0:00:01
   ---------------------------------------- 2.8/2.8 MB 712.3 kB/s eta 0:00:00
Installing collected packages: biopython
Successfully installed biopython-1.84
```

Once done with the installation of BioPython, the complete genome sequences of chicken with salmonella enterica has been searched in NCBI nucleotide database and a total of 496 sequence has been found.

In [1]:
```
from Bio import Entrez

# Set your email (required by NCBI)
Entrez.email = "rd97.dd@gmail.com"

# Define search query for Salmonella enterica with isolation source chicken
search_term = "Salmonella enterica[Organism] AND chicken[Isolation Source] AND complete genome[Title]"

# Perform search in the NCBI nucleotide database
handle = Entrez.esearch(db="nucleotide", term=search_term, retmax=0)  # retmax=0 to only get the count
record = Entrez.read(handle)
handle.close()

# Get and print the total number of records (sequences) found
total_records = int(record["Count"])
print(f"Total sequences found: {total_records}")
```

```
Total sequences found: 496
```

A sample of 10 isolate files has been downloaded for verification purpose.

In [12]:
```
from Bio import Entrez, SeqIO
import time
import os

# Set your email (required by NCBI)
Entrez.email = "rd97.dd@gmail.com"

# Define broader search queries
search_term = "Salmonella enterica[Organism] AND chicken[Isolation Source] AND complete genome[Title]"

# Define the download directory
output_directory ="C:/Users/DD/Documents/Python Scripts/sequences"
os.makedirs(output_directory, exist_ok=True)

# Function to download a specific number of sequences
def download_sequences(search_term, output_directory, num_sequences=10):
    # Search the NCBI nucleotide database
    handle = Entrez.esearch(db="nucleotide", term=search_term, retmax=num_sequences)
    record = Entrez.read(handle)
    handle.close()

    total_records = int(record["Count"])
    print(f"Total sequences found: {total_records}")

    id_list = record["IdList"]

    if not id_list:
        print("No IDs found, stopping download.")
        return

    for seq_id in id_list:
        try:
            # Fetch the metadata
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="gb", retmode="text")
            metadata = handle.read()
            handle.close()

            # Save metadata
            metadata_file = os.path.join(output_directory, f"{seq_id}_metadata.gb")
            with open(metadata_file, "w") as metadata_handle:
                metadata_handle.write(metadata)

            # Fetch the FASTA sequence
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="fasta", retmode="text")
            seq_record = SeqIO.read(handle, "fasta")
            handle.close()

            # Check if the sequence is valid
            if not seq_record.seq or not len(seq_record.seq):
                print(f"Warning: Sequence {seq_id} is empty or undefined.")
                continue

            # Save the FASTA sequence
            fasta_file = os.path.join(output_directory, f"{seq_id}.fasta")
            with open(fasta_file, "w") as out_handle:
                SeqIO.write(seq_record, out_handle, "fasta")

            print(f"Downloaded and saved {fasta_file}")

        except Exception as e:
            print(f"Error processing {seq_id}: {e}")

    print("Download complete!")

# Download the first 10 sequences
download_sequences(search_term, output_directory, num_sequences=10)
```

```
Total sequences found: 496
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964398.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964390.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964384.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964378.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964373.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959104.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959093.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791348005.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334406.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334263.fasta
Download complete!
```

In next step, duplicates has been checked based of file names.

In [14]:
```
from Bio import SeqIO
import os

def check_duplicates(fasta_directory):
    seen_sequences = set()
    duplicates = []

    for filename in os.listdir(fasta_directory):
        if filename.endswith(".fasta"):
            filepath = os.path.join(fasta_directory, filename)
            for record in SeqIO.parse(filepath, "fasta"):
                sequence = str(record.seq)
                if sequence in seen_sequences:
                    duplicates.append(filename)
                else:
                    seen_sequences.add(sequence)
    if duplicates:
        print(f"Duplicate files found: {set(duplicates)}")
    else:
        print("No duplicate sequences found.")

# Set the path to your FASTA files directory
fasta_directory ="C:/Users/DD/Documents/Python Scripts/sequences"
check_duplicates(fasta_directory)
```

```
No duplicate sequences found.
```

Here , its been checked for duplicates with strain Ids from metadata

In [3]:
```
 from Bio import Entrez
import os
import re

def extract_strain_info(metadata_directory):
    strain_info = {}

    for filename in os.listdir(metadata_directory):
        if filename.endswith("_metadata.gb"):
            filepath = os.path.join(metadata_directory, filename)
            with open(filepath, "r") as file:
                content = file.read()
                match = re.search(r"strain=(\S+)", content)
                if match:
                    strain = match.group(1)
                    if strain in strain_info:
                        strain_info[strain].append(filename)
                    else:
                        strain_info[strain] = [filename]

    # Print out strains and their corresponding files
    for strain, files in strain_info.items():
        if len(files) > 1:
            print(f"Strain {strain} found in files: {files}")

# Set the path to your metadata directory
metadata_directory = "C:/Users/DD/Documents/Python Scripts/sequences"
extract_strain_info(metadata_directory)
```

```
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
Unique strains only founded
```

Now , we have downloaded 496 sequences o along with meta data.

In [ ]:
```
from Bio import Entrez, SeqIO
import time
import os

# Set your email (required by NCBI)
Entrez.email = "rd97.dd@gmail.com"

# Define broader search queries
search_term = "Salmonella enterica[Organism] AND chicken[Isolation Source] AND complete genome[Title]"

# Define the download directory
output_directory ="C:/Users/DD/Documents/Python Scripts/sequences"
os.makedirs(output_directory, exist_ok=True)

# Function to download a specific number of sequences
def download_sequences(search_term, output_directory, num_sequences=496):
    # Search the NCBI nucleotide database
    handle = Entrez.esearch(db="nucleotide", term=search_term, retmax=num_sequences)
    record = Entrez.read(handle)
    handle.close()

    total_records = int(record["Count"])
    print(f"Total sequences found: {total_records}")

    id_list = record["IdList"]

    if not id_list:
        print("No IDs found, stopping download.")
        return

    for seq_id in id_list:
        try:
            # Fetch the metadata
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="gb", retmode="text")
            metadata = handle.read()
            handle.close()

            # Save metadata
            metadata_file = os.path.join(output_directory, f"{seq_id}_metadata.gb")
            with open(metadata_file, "w") as metadata_handle:
                metadata_handle.write(metadata)

            # Fetch the FASTA sequence
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="fasta", retmode="text")
            seq_record = SeqIO.read(handle, "fasta")
            handle.close()

            # Check if the sequence is valid
            if not seq_record.seq or not len(seq_record.seq):
                print(f"Warning: Sequence {seq_id} is empty or undefined.")
                continue

            # Save the FASTA sequence
            fasta_file = os.path.join(output_directory, f"{seq_id}.fasta")
            with open(fasta_file, "w") as out_handle:
                SeqIO.write(seq_record, out_handle, "fasta")

            print(f"Downloaded and saved {fasta_file}")

        except Exception as e:
            print(f"Error processing {seq_id}: {e}")

    print("Download complete!")

# Download the first 10 sequences
download_sequences(search_term, output_directory, num_sequences=496)
```

```
Total sequences found: 496
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964398.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964390.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964384.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964378.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964373.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959104.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959093.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791348005.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334406.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334263.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334217.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334100.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791333611.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791332929.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791332598.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791329900.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328843.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328750.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328742.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328717.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328715.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328684.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328489.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328192.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327776.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327658.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327656.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327653.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327646.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327513.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327413.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327275.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791326870.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791326641.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791325747.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791323978.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1864594723.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1864593957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790361485.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790355281.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790350405.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790345631.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790340519.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790335951.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790330984.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790326108.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790321214.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790316599.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790311342.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790306333.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790302094.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790297568.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790292938.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790288563.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790283963.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790279127.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790274631.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790270044.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790265499.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790261062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790256505.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790252121.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790247716.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790243305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790238422.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790232908.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790226660.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790219725.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790191447.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790185892.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790181112.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790176664.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790172290.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790167917.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790153796.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790142919.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863662535.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863661087.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863660810.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550287957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550255387.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550225976.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213792.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213764.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213748.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213647.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2784851814.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2784324973.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2060646254.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1221158448.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2265635730.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258710042.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258485671.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2521002406.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2747284265.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2515152310.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1002985512.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001555507.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001554681.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001552415.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001551848.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1000369143.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998623498.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998623489.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2036914008.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2740278012.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2739185197.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1829022561.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1829031305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2485141349.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2482324433.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2477330759.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2476810552.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1546887966.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1995427272.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100896152.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1951777374.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\915846032.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1982652612.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2717387261.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2717024100.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711495709.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711488657.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711475867.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711475272.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711469117.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711466246.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711464918.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462769.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462605.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462317.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711461464.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711460722.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711459283.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711458674.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711458425.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2710806370.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2710805975.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704924300.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704915029.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704910405.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704905772.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704891984.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704874751.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704865509.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703459908.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703450442.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703445817.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703441194.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703436556.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703427236.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703422591.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703417940.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703413299.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703404023.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1736666103.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1736663574.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1033136052.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\771524686.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2694694140.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1486618283.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1302574149.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1302554428.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1227482702.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1196285509.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1110727957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046637175.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046630596.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046611151.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046591696.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046300946.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2689080905.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2689075821.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2679323484.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2679318990.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2168004414.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1945020687.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2666584953.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2078699049.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863657517.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1699456900.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2359423532.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2339534824.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2339477813.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638510809.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491693.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491622.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491583.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636855593.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636851041.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636846570.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636842156.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1828945320.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1531933061.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\983377086.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749314944.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749298420.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749298461.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749297197.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749296449.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1797869977.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321043586.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321043499.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321042278.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321041129.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321031274.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321014208.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321008513.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100424222.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100356487.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100237189.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100166053.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1128998525.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046302332.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046220323.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2601133276.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2595857884.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1733272414.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288890305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2087863205.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2082554870.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2294922788.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288538671.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288537321.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288530062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2572860206.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2570622740.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2567050896.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2279083905.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2279079344.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548306062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548296188.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548291250.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548275539.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548260632.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548250070.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548245062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1635401364.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1634456755.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2273089235.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1130930627.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2518923471.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2514088622.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2484128728.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2482050684.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2476413604.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2475999171.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459865826.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459861217.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459856665.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459518531.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2333272181.fasta
```

In [1]:

```python
from Bio import Entrez, SeqIO
import os

# Set your email (required by NCBI)
Entrez.email = "rd97.dd@gmail.com"

# Define broader search queries
search_term = "Salmonella enterica[Organism] AND chicken[Isolation Source] AND complete genome[Title]"

# Define the download directory
output_directory = "C:/Users/DD/Documents/Python Scripts/sequences"
os.makedirs(output_directory, exist_ok=True)

# Function to download a specific number of sequences
def download_sequences(search_term, output_directory, num_sequences=496):
    # Search the NCBI nucleotide database
    handle = Entrez.esearch(db="nucleotide", term=search_term, retmax=num_sequences)
    record = Entrez.read(handle)
    handle.close()

    total_records = int(record["Count"])
    print(f"Total sequences found: {total_records}")

    id_list = record["IdList"]

    # Track already downloaded sequences by checking existing files
    downloaded_ids = set([file.split('_')[0] for file in os.listdir(output_directory) if file.endswith('.fasta')])

    print(f"Already downloaded {len(downloaded_ids)} sequences, continuing from where it left off.")

    for seq_id in id_list:
        if seq_id in downloaded_ids:
            print(f"Skipping {seq_id}, already downloaded.")
            continue

        try:
            # Fetch the metadata
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="gb", retmode="text")
            metadata = handle.read()
            handle.close()

            # Save metadata
            metadata_file = os.path.join(output_directory, f"{seq_id}_metadata.gb")
            with open(metadata_file, "w") as metadata_handle:
                metadata_handle.write(metadata)

            # Fetch the FASTA sequence
            handle = Entrez.efetch(db="nucleotide", id=seq_id, rettype="fasta", retmode="text")
            seq_record = SeqIO.read(handle, "fasta")
            handle.close()

            # Check if the sequence is valid
            if not seq_record.seq or not len(seq_record.seq):
                print(f"Warning: Sequence {seq_id} is empty or undefined.")
                continue

            # Save the FASTA sequence
            fasta_file = os.path.join(output_directory, f"{seq_id}.fasta")
            with open(fasta_file, "w") as out_handle:
                SeqIO.write(seq_record, out_handle, "fasta")

            print(f"Downloaded and saved {fasta_file}")

        except Exception as e:
            print(f"Error processing {seq_id}: {e}")

    print("Download complete!")

# Resume downloading the remaining sequences
download_sequences(search_term, output_directory, num_sequences=496)
```

```
Total sequences found: 496
Already downloaded 252 sequences, continuing from where it left off.
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964398.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964390.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964384.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964378.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791964373.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959104.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791959093.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791348005.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334406.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334263.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334217.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791334100.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791333611.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791332929.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791332598.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791329900.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328843.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328750.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328742.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328717.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328715.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328684.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328489.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791328192.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327776.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327658.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327656.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327653.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327646.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327513.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327413.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791327275.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791326870.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791326641.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791325747.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2791323978.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1864594723.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1864593957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790361485.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790355281.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790350405.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790345631.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790340519.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790335951.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790330984.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790326108.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790321214.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790316599.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790311342.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790306333.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790302094.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790297568.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790292938.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790288563.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790283963.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790279127.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790274631.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790270044.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790265499.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790261062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790256505.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790252121.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790247716.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790243305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790238422.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790232908.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790226660.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790219725.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790191447.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790185892.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790181112.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790176664.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790172290.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790167917.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790153796.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2790142919.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863662535.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863661087.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863660810.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550287957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550255387.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550225976.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213792.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213764.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213748.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2550213647.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2784851814.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2784324973.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2060646254.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1221158448.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2265635730.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258710042.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258485671.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2521002406.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2747284265.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2515152310.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1002985512.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001555507.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001554681.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001552415.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001551848.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1000369143.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998623498.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998623489.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2036914008.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2740278012.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2739185197.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1829022561.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1829031305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2485141349.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2482324433.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2477330759.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2476810552.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1546887966.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1995427272.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100896152.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1951777374.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\915846032.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1982652612.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2717387261.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2717024100.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711495709.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711488657.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711475867.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711475272.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711469117.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711466246.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711464918.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462769.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462605.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711462317.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711461464.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711460722.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711459283.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711458674.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2711458425.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2710806370.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2710805975.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704924300.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704915029.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704910405.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704905772.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704891984.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704874751.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2704865509.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703459908.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703450442.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703445817.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703441194.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703436556.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703427236.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703422591.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703417940.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703413299.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2703404023.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1736666103.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1736663574.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1033136052.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\771524686.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2694694140.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1486618283.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1302574149.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1302554428.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1227482702.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1196285509.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1110727957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046637175.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046630596.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046611151.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046591696.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046300946.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2689080905.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2689075821.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2679323484.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2679318990.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2168004414.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1945020687.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2666584953.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2078699049.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863657517.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1699456900.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2359423532.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2339534824.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2339477813.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638510809.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491693.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491622.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2638491583.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636855593.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636851041.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636846570.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2636842156.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1828945320.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1531933061.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\983377086.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749314944.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749298420.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749298461.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749297197.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\749296449.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1797869977.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321043586.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321043499.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321042278.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321041129.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321031274.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321014208.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2321008513.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100424222.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100356487.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100237189.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2100166053.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1128998525.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046302332.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1046220323.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2601133276.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2595857884.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1733272414.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288890305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2087863205.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2082554870.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2294922788.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288538671.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288537321.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288530062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2572860206.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2570622740.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2567050896.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2279083905.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2279079344.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548306062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548296188.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548291250.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548275539.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548260632.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548250070.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2548245062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1635401364.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1634456755.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2273089235.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1130930627.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2518923471.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2514088622.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2484128728.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2482050684.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2476413604.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2475999171.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459865826.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459861217.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459856665.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2459518531.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2333272181.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2333213671.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2325662964.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320761748.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320757305.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320752890.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320748473.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320738824.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320734399.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2320729955.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2294456242.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288320740.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288176995.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288173494.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2288165041.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2285411033.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2277362348.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2277353412.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2271345254.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2263926704.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258066866.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2258055768.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091333084.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091323346.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091253165.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091243740.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091482867.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091478163.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091473540.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091469148.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091455043.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091446076.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091431810.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091422800.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091418357.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091402345.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2090956231.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2090918790.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091029751.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091122469.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091117943.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2167845084.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843531062.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843507307.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843498400.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843494016.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2136510056.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843516861.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1846455821.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843652501.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843647868.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843643169.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843633944.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843629339.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843618362.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843613175.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843608483.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843599223.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843589994.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843585404.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843580775.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843576148.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843544971.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843540331.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1843512211.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2095951993.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2094642037.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2094623487.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2094623485.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2094623483.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2093998598.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2093987282.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091375002.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091370368.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091356744.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091346897.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091277794.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091264036.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091239126.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091229717.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091183720.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091178941.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091174439.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091169855.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2091165092.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2090974716.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2087468524.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2077453074.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2060222687.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2049446339.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2049354170.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2035064789.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2035051374.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2034574262.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030203977.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030202676.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030201146.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030199816.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030197290.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\2030195509.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1993496820.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1979358710.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953513472.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953508822.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953500321.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953495564.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953490790.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953486234.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1953481399.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1951535343.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1948849906.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1943036513.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1891180034.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1528951495.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863557100.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863552396.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863547333.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863542555.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863537930.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1863533289.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1836411705.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1129908976.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1842989566.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1828463287.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1828458697.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1828351404.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1788239996.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1691492771.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1732125948.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1695513588.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1695497099.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1634015463.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1633748201.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1543688490.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1491767611.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\754295078.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\523821078.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1219800280.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1190123157.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\267991652.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1127388936.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1127384256.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1127379580.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\747137039.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\747132831.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\924643612.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\924639146.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\924635423.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1045716295.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1045707385.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1020327374.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1020321938.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1018838233.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1030103469.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1021436845.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1020345203.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1020336109.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\914845112.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\914828066.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\949877344.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\949872475.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\949867589.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\953773649.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\953768973.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\914835508.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1002883852.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1002707584.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1002000489.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\1001427523.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998647728.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\998642845.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995923765.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995923317.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995922935.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995920397.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995918924.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995915826.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995907817.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995907166.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\674230814.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\995901180.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\896682452.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\971178605.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\933894153.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\933889448.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\930817479.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\930813018.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\896682995.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\901911536.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\901905107.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\820762957.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\820758584.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\696593549.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\696581462.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\767847001.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\767842734.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\767838360.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\767833982.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\754295112.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\744788240.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\523815970.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\674281104.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\674188659.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\686507741.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682093565.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682089197.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682084826.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682080448.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682067343.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682063065.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682058852.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682049973.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\682036555.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\564743501.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\559187652.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\630832964.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604362088.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604350507.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604336542.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604322487.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604307148.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604290478.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604278585.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604268692.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604257961.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\604190314.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602826365.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602822031.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602817697.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602813325.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602808998.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\602804670.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605555317.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605546632.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605537200.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605528405.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605520020.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605511148.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\605498156.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\601101465.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\563346448.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\548713695.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\529190224.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\332986951.fasta
Downloaded and saved C:/Users/DD/Documents/Python Scripts/sequences\523806722.fasta
Download complete!
```

Checked for duplicates based on file names after downloading the sequences

In [3]:
```python
from Bio import SeqIO
import os

def check_duplicates(fasta_directory):
    seen_sequences = set()
    duplicates = []

    for filename in os.listdir(fasta_directory):
        if filename.endswith(".fasta"):
            filepath = os.path.join(fasta_directory, filename)
            for record in SeqIO.parse(filepath, "fasta"):
                sequence = str(record.seq)
                if sequence in seen_sequences:
                    duplicates.append(filename)
                else:
                    seen_sequences.add(sequence)

    if duplicates:
        print(f"Duplicate files found: {set(duplicates)}")
    else:
        print("No duplicate sequences found.")

# Set the path to your FASTA files directory
fasta_directory ="C:/Users/DD/Documents/Python Scripts/sequences"
check_duplicates(fasta_directory)
```

```
Duplicate files found: {'2791333611.fasta', '2791327646.fasta', '2711458674.fasta', '1128998525.fasta', '749298461.fasta', '2791328715.fasta', '2321041129.fasta', '2791327658.fasta', '10462
20323.fasta', '2482324433.fasta', '1863657517.fasta', '2711462317.fasta', '1033136052.fasta', '998623498.fasta', '1864594723.fasta', '2791328192.fasta', '1864593957.fasta', '2321008513.fast
a', '2791332598.fasta', '995922935.fasta', '2711462769.fasta', '2321043499.fasta', '2339477813.fasta', '2036914008.fasta', '2078699049.fasta', '2638491622.fasta', '2082554870.fasta', '19954
27272.fasta', '2521002406.fasta', '2100237189.fasta', '2784851814.fasta', '2717387261.fasta', '1863661087.fasta', '2550213792.fasta', '2791328750.fasta', '1634456755.fasta', '983377086.fast
a', '2711462605.fasta', '2791327776.fasta', '1982652612.fasta', '2711475867.fasta', '2638491693.fasta', '2791334100.fasta', '2638510809.fasta', '2791328742.fasta', '2791327275.fasta', '2791
959104.fasta', '2321043586.fasta', '998647728.fasta', '2279079344.fasta', '2791328684.fasta', '1828945320.fasta', '2711466246.fasta', '2288537321.fasta', '2087863205.fasta', '2550225976.fas
ta', '2550213748.fasta', '2638491583.fasta', '2711469117.fasta', '2711475272.fasta', '2060646254.fasta', '674230814.fasta', '2791348005.fasta', '2710806370.fasta', '2258710042.fasta', '2265
635730.fasta', '2476810552.fasta', '2791334406.fasta', '995907166.fasta', '2791964378.fasta', '1736666103.fasta', '1130930627.fasta', '749298420.fasta', '2100424222.fasta', '998642845.fast
a', '2791328489.fasta', '2477330759.fasta', '2168004414.fasta', '2515152310.fasta', '2710805975.fasta', '2791964398.fasta', '2321014208.fasta', '2294922788.fasta', '1635401364.fasta', '2791
327413.fasta', '2791328717.fasta', '2321042278.fasta', '2791327513.fasta', '2791959093.fasta', '2791326870.fasta', '1221158448.fasta', '2100896152.fasta', '995907817.fasta', '2550213647.fas
ta', '2279083905.fasta', '2711464918.fasta', '2791325747.fasta', '1951777374.fasta', '2550287957.fasta', '1829031305.fasta', '2711460722.fasta', '2100356487.fasta', '2791323978.fasta', '154
6887966.fasta', '1531933061.fasta', '2273089235.fasta', '2711461464.fasta', '749297197.fasta', '2711488657.fasta', '1829022561.fasta', '1733272414.fasta', '2791327656.fasta', '2791329900.fa
sta', '2711459283.fasta', '2791327653.fasta', '2601133276.fasta', '2288538671.fasta', '2485141349.fasta', '915846032.fasta', '1046302332.fasta', '1196285509.fasta', '2100166053.fasta', '179
7869977.fasta', '2791332929.fasta', '995923317.fasta', '2288530062.fasta', '2791328843.fasta', '1945020687.fasta', '2791964390.fasta', '1863662535.fasta', '2791964373.fasta', '2258485671.fa
sta', '2572860206.fasta', '2791334217.fasta', '2339534824.fasta', '2550255387.fasta', '2791326641.fasta', '749296449.fasta', '2550213764.fasta', '2740278012.fasta', '2711495709.fasta', '279
1334263.fasta', '1736663574.fasta', '2288890305.fasta', '749314944.fasta', '2711458425.fasta', '1863660810.fasta', '771524686.fasta', '2791964384.fasta', '2321031274.fasta'}
```

In this step, we have checked for duplicate sequences based on strains

In [5]:
```python
 from Bio import Entrez
import os
import re

def extract_strain_info(metadata_directory):
    strain_info = {}

    for filename in os.listdir(metadata_directory):
        if filename.endswith("_metadata.gb"):
            filepath = os.path.join(metadata_directory, filename)
            with open(filepath, "r") as file:
                content = file.read()
                match = re.search(r"strain=(\S+)", content)  # Adjust regex based on metadata format
                if match:
                    strain = match.group(1)
                    if strain in strain_info:
                        strain_info[strain].append(filename)
                    else:
                        strain_info[strain] = [filename]

    # Print out strains and their corresponding files
    for strain, files in strain_info.items():
        if len(files) > 1:
            print(f"Strain {strain} found in files: {files}")

# Set the path to your metadata directory
metadata_directory = "C:/Users/DD/Documents/Python Scripts/sequences"
extract_strain_info(metadata_directory)
```

```
Strain "SA20095440" found in files: ['1000369143_metadata.gb', '998647728_metadata.gb']
Strain "SA19960848" found in files: ['1001551848_metadata.gb', '995907166_metadata.gb']
Strain "SA19970769" found in files: ['1001552415_metadata.gb', '995922935_metadata.gb']
Strain "SA19970510" found in files: ['1001554681_metadata.gb', '995907817_metadata.gb']
Strain "SA19980677" found in files: ['1001555507_metadata.gb', '995923317_metadata.gb']
Strain "SA20094352" found in files: ['1002985512_metadata.gb', '998642845_metadata.gb']
Strain "C629" found in files: ['1030103469_metadata.gb', '1033136052_metadata.gb']
Strain "SA02DT09004001" found in files: ['1045707385_metadata.gb', '1046220323_metadata.gb']
Strain "SA01AB09084001" found in files: ['1045716295_metadata.gb', '1046302332_metadata.gb']
Strain "N55391" found in files: ['1127388936_metadata.gb', '1128998525_metadata.gb']
Strain "87" found in files: ['1129908976_metadata.gb', '1130930627_metadata.gb']
Strain "FORC_038" found in files: ['1190123157_metadata.gb', '1196285509_metadata.gb']
Strain "D90" found in files: ['1219800280_metadata.gb', '1221158448_metadata.gb']
Strain "ATCC found in files: ['1528951495_metadata.gb', '1531933061_metadata.gb', '605520020_metadata.gb', '605528405_metadata.gb', '674281104_metadata.gb', '820762957_metadata.gb', '914828
066_metadata.gb', '914835508_metadata.gb', '914845112_metadata.gb']
Strain "5" found in files: ['1543688490_metadata.gb', '1546887966_metadata.gb']
Strain "JT found in files: ['1633748201_metadata.gb', '1634456755_metadata.gb']
Strain "CD-SL01" found in files: ['1634015463_metadata.gb', '1635401364_metadata.gb']
Strain "SJTUF87912v2" found in files: ['1695497099_metadata.gb', '1736663574_metadata.gb']
Strain "SJTUF13520v2" found in files: ['1695513588_metadata.gb', '1736666103_metadata.gb']
Strain "SL-312" found in files: ['1732125948_metadata.gb', '1733272414_metadata.gb']
Strain "SCSM4.1" found in files: ['1788239996_metadata.gb', '1797869977_metadata.gb']
Strain "SE124" found in files: ['1828351404_metadata.gb', '1828945320_metadata.gb']
Strain "SI85" found in files: ['1828458697_metadata.gb', '1829031305_metadata.gb']
Strain "SI67" found in files: ['1828463287_metadata.gb', '1829022561_metadata.gb']
Strain "CVM found in files: ['1836411705_metadata.gb', '1842989566_metadata.gb', '1843494016_metadata.gb', '1843498400_metadata.gb', '1843507307_metadata.gb', '1843512211_metadata.gb', '184
3516861_metadata.gb', '1843531062_metadata.gb', '1843540331_metadata.gb', '1843544971_metadata.gb', '1843576148_metadata.gb', '1843580775_metadata.gb', '1843585404_metadata.gb', '1843589994
_metadata.gb', '1843599223_metadata.gb', '1843608483_metadata.gb', '1843613175_metadata.gb', '1843618362_metadata.gb', '1843629339_metadata.gb', '1843633944_metadata.gb', '1843643169_metada
ta.gb', '1843647868_metadata.gb', '1843652501_metadata.gb', '1846455821_metadata.gb', '1951535343_metadata.gb', '1951777374_metadata.gb', '2090918790_metadata.gb', '2090956231_metadata.gb',
'2090974716_metadata.gb', '2091165092_metadata.gb', '2091169855_metadata.gb', '2091174439_metadata.gb', '2091178941_metadata.gb', '2091183720_metadata.gb', '2091229717_metadata.gb', '209123
9126_metadata.gb', '2091243740_metadata.gb', '2091253165_metadata.gb', '2091264036_metadata.gb', '2091277794_metadata.gb', '2091323346_metadata.gb', '2091333084_metadata.gb', '2091346897_me
tadata.gb', '2091356744_metadata.gb', '2091370368_metadata.gb', '2091375002_metadata.gb', '2091402345_metadata.gb', '2091418357_metadata.gb', '2091422800_metadata.gb', '2091431810_metadata.
gb', '2091446076_metadata.gb', '2091455043_metadata.gb', '2091469148_metadata.gb', '2091473540_metadata.gb', '2091478163_metadata.gb', '2091482867_metadata.gb', '2136510056_metadata.gb', '2
518923471_metadata.gb', '2521002406_metadata.gb', '696581462_metadata.gb', '696593549_metadata.gb', '744788240_metadata.gb', '767833982_metadata.gb', '767838360_metadata.gb', '767842734_met
adata.gb', '767847001_metadata.gb', '771524686_metadata.gb']
Strain "VNSEC031" found in files: ['1863533289_metadata.gb', '1863660810_metadata.gb']
Strain "VNSEC023" found in files: ['1863537930_metadata.gb', '1863661087_metadata.gb']
Strain "VNSEC013" found in files: ['1863542555_metadata.gb', '1863657517_metadata.gb']
Strain "VNSEC003" found in files: ['1863547333_metadata.gb', '1863662535_metadata.gb']
Strain "VNSEC002" found in files: ['1863552396_metadata.gb', '1864594723_metadata.gb']
Strain "VNSEC001" found in files: ['1863557100_metadata.gb', '1864593957_metadata.gb']
Strain "R16.1424" found in files: ['1943036513_metadata.gb', '1945020687_metadata.gb']
Strain "GSJ/2016-Sal-017" found in files: ['1979358710_metadata.gb', '1982652612_metadata.gb']
Strain "R16.0556" found in files: ['1993496820_metadata.gb', '1995427272_metadata.gb']
Strain "KUFSE-SAL0043" found in files: ['2034574262_metadata.gb', '2036914008_metadata.gb']
Strain "CFSAN008081" found in files: ['2049354170_metadata.gb', '2082554870_metadata.gb']
Strain "YZ20MCS16" found in files: ['2060222687_metadata.gb', '2060646254_metadata.gb']
Strain "SA18578" found in files: ['2077453074_metadata.gb', '2078699049_metadata.gb']
Strain "S16" found in files: ['2087468524_metadata.gb', '2087863205_metadata.gb']
Strain "ZC-S1 found in files: ['2093998598_metadata.gb', '2100166053_metadata.gb']
Strain "S90" found in files: ['2094623483_metadata.gb', '2100356487_metadata.gb']
Strain "SP" found in files: ['2094623485_metadata.gb', '2100424222_metadata.gb']
Strain "S146" found in files: ['2094642037_metadata.gb', '2100237189_metadata.gb']
Strain "07Q015" found in files: ['2095951993_metadata.gb', '2100896152_metadata.gb']
Strain "YZ21MCS4" found in files: ['2167845084_metadata.gb', '2168004414_metadata.gb']
Strain "YZ20MCS6" found in files: ['2258055768_metadata.gb', '2258485671_metadata.gb']
Strain "YZ20MCS14" found in files: ['2258066866_metadata.gb', '2258710042_metadata.gb']
Strain "SE006" found in files: ['2263926704_metadata.gb', '2265635730_metadata.gb']
Strain "013+" found in files: ['2271345254_metadata.gb', '2273089235_metadata.gb']
Strain "KCID6" found in files: ['2277353412_metadata.gb', '2279079344_metadata.gb']
Strain "BCID6" found in files: ['2277362348_metadata.gb', '2279083905_metadata.gb']
Strain "AH19MCS8" found in files: ['2288173494_metadata.gb', '2288530062_metadata.gb']
Strain "XZ14C1328" found in files: ['2288320740_metadata.gb', '2288890305_metadata.gb']
Strain "MRS17_00712" found in files: ['2294456242_metadata.gb', '2294922788_metadata.gb']
Strain "190821_1" found in files: ['2320729955_metadata.gb', '2321014208_metadata.gb']
Strain "190819_2" found in files: ['2320734399_metadata.gb', '2321042278_metadata.gb']
Strain "190807_1" found in files: ['2320738824_metadata.gb', '2321031274_metadata.gb']
Strain "190729_8" found in files: ['2320748473_metadata.gb', '2321043499_metadata.gb']
Strain "190729_4" found in files: ['2320752890_metadata.gb', '2321043586_metadata.gb']
Strain "190704_2" found in files: ['2320757305_metadata.gb', '2321008513_metadata.gb']
Strain "190610_1" found in files: ['2320761748_metadata.gb', '2321041129_metadata.gb']
Strain "XM3104" found in files: ['2333213671_metadata.gb', '2339534824_metadata.gb']
Strain "CVCC found in files: ['2333272181_metadata.gb', '2339477813_metadata.gb']
Strain "R22.0044" found in files: ['2475999171_metadata.gb', '2476810552_metadata.gb']
Strain "013" found in files: ['2476413604_metadata.gb', '2477330759_metadata.gb']
Strain "S1467" found in files: ['2482050684_metadata.gb', '2482324433_metadata.gb']
Strain "CRIN508879" found in files: ['2484128728_metadata.gb', '2485141349_metadata.gb']
Strain "CRSE-01" found in files: ['2514088622_metadata.gb', '2515152310_metadata.gb']
Strain "SECVM-15" found in files: ['2548245062_metadata.gb', '2550287957_metadata.gb']
Strain "SECVM-17" found in files: ['2548250070_metadata.gb', '2550213647_metadata.gb']
Strain "SE found in files: ['2548260632_metadata.gb', '2550213764_metadata.gb']
Strain "SECVM-13" found in files: ['2548275539_metadata.gb', '2550225976_metadata.gb']
Strain "SECVM-10" found in files: ['2548291250_metadata.gb', '2550255387_metadata.gb']
Strain "SECVM-14" found in files: ['2548296188_metadata.gb', '2550213748_metadata.gb']
Strain "SECVM-5" found in files: ['2548306062_metadata.gb', '2550213792_metadata.gb']
Strain "Z1323CSL0015" found in files: ['2570622740_metadata.gb', '2572860206_metadata.gb', '2703404023_metadata.gb', '2711459283_metadata.gb']
Strain "LA5" found in files: ['2595857884_metadata.gb', '2601133276_metadata.gb']
Strain "405987R1_S50" found in files: ['2636846570_metadata.gb', '2638491583_metadata.gb']
Strain "401964R1_S49" found in files: ['2636851041_metadata.gb', '2638491693_metadata.gb']
Strain "180121R1S_S48" found in files: ['2636855593_metadata.gb', '2638491622_metadata.gb']
Strain "SSSE-01" found in files: ['2679318990_metadata.gb', '2710805975_metadata.gb']
Strain "SSSE-03" found in files: ['2679323484_metadata.gb', '2710806370_metadata.gb']
Strain "Z1323CSL0016" found in files: ['2703413299_metadata.gb', '2711458674_metadata.gb']
Strain "Z1323CSL0014" found in files: ['2703417940_metadata.gb', '2711458425_metadata.gb']
Strain "Z1323CSL0027" found in files: ['2703422591_metadata.gb', '2711461464_metadata.gb']
Strain "Z1323CSL0034" found in files: ['2703436556_metadata.gb', '2711469117_metadata.gb']
Strain "Z1323CSL0017" found in files: ['2703441194_metadata.gb', '2711464918_metadata.gb']
Strain "Z1323CSL0002" found in files: ['2703445817_metadata.gb', '2711466246_metadata.gb']
Strain "Z1323CSL0006" found in files: ['2703450442_metadata.gb', '2711460722_metadata.gb']
Strain "Z1323CSL0062" found in files: ['2704865509_metadata.gb', '2711462317_metadata.gb']
Strain "Z1323CSL0057" found in files: ['2704874751_metadata.gb', '2711475867_metadata.gb']
Strain "Z1323CSL0053" found in files: ['2704891984_metadata.gb', '2711495709_metadata.gb']
Strain "Z1323CSL0045" found in files: ['2704905772_metadata.gb', '2711488657_metadata.gb']
Strain "Z1323CSL0047" found in files: ['2704910405_metadata.gb', '2711475272_metadata.gb']
Strain "Z1323CSL0052" found in files: ['2704915029_metadata.gb', '2711462605_metadata.gb']
Strain "Z1323CSL0051" found in files: ['2704924300_metadata.gb', '2711462769_metadata.gb']
Strain "ST.GS36" found in files: ['2717024100_metadata.gb', '2717387261_metadata.gb']
Strain "IJCS5-22" found in files: ['2739185197_metadata.gb', '2740278012_metadata.gb']
Strain "H4" found in files: ['2784324973_metadata.gb', '2784851814_metadata.gb']
Strain "Z01320SL0044" found in files: ['2790142919_metadata.gb', '2791959093_metadata.gb']
Strain "Z01320SL0045" found in files: ['2790153796_metadata.gb', '2791964378_metadata.gb']
Strain "Z01320SL0042" found in files: ['2790167917_metadata.gb', '2791959104_metadata.gb']
Strain "Z01320SL0040" found in files: ['2790172290_metadata.gb', '2791964398_metadata.gb']
Strain "Z01320SL0041" found in files: ['2790176664_metadata.gb', '2791964390_metadata.gb']
Strain "Z01320SL0038" found in files: ['2790181112_metadata.gb', '2791964373_metadata.gb']
Strain "Z01320SL0043" found in files: ['2790191447_metadata.gb', '2791964384_metadata.gb']
Strain "Z01320SL0030" found in files: ['2790219725_metadata.gb', '2791325747_metadata.gb']
Strain "Z01320SL0033" found in files: ['2790226660_metadata.gb', '2791348005_metadata.gb']
Strain "Z01320SL0034" found in files: ['2790232908_metadata.gb', '2791323978_metadata.gb']
Strain "Z01320SL0031" found in files: ['2790238422_metadata.gb', '2791328742_metadata.gb']
Strain "Z01320SL0036" found in files: ['2790243305_metadata.gb', '2791327776_metadata.gb']
Strain "Z01320SL0037" found in files: ['2790247716_metadata.gb', '2791329900_metadata.gb']
Strain "Z01320SL0035" found in files: ['2790252121_metadata.gb', '2791327646_metadata.gb']
Strain "Z01320SL0032" found in files: ['2790256505_metadata.gb', '2791332598_metadata.gb']
Strain "Z01319SL0021" found in files: ['2790261062_metadata.gb', '2791326641_metadata.gb']
Strain "Z01319SL0015" found in files: ['2790265499_metadata.gb', '2791328717_metadata.gb']
Strain "Z01319SL0020" found in files: ['2790270044_metadata.gb', '2791327275_metadata.gb']
Strain "Z01320SL0029" found in files: ['2790274631_metadata.gb', '2791332929_metadata.gb']
Strain "Z01319SL0016" found in files: ['2790279127_metadata.gb', '2791327653_metadata.gb']
Strain "Z01319SL0017" found in files: ['2790283963_metadata.gb', '2791328843_metadata.gb']
Strain "Z01319SL0022" found in files: ['2790292938_metadata.gb', '2791327656_metadata.gb']
Strain "Z01320SL0026" found in files: ['2790297568_metadata.gb', '2791334100_metadata.gb']
Strain "Z01320SL0027" found in files: ['2790302094_metadata.gb', '2791328684_metadata.gb']
Strain "Z01319SL0009" found in files: ['2790306333_metadata.gb', '2791327658_metadata.gb']
Strain "Z01319SL0019" found in files: ['2790311342_metadata.gb', '2791328750_metadata.gb']
Strain "Z01319SL0010" found in files: ['2790316599_metadata.gb', '2791326870_metadata.gb']
Strain "Z01320SL0025" found in files: ['2790321214_metadata.gb', '2791333611_metadata.gb']
Strain "Z01320SL0024" found in files: ['2790326108_metadata.gb', '2791328192_metadata.gb']
Strain "Z01319SL0008" found in files: ['2790330984_metadata.gb', '2791327513_metadata.gb']
Strain "Z01319SL0006" found in files: ['2790335951_metadata.gb', '2791328715_metadata.gb']
Strain "Z01319SL0001" found in files: ['2790340519_metadata.gb', '2791328489_metadata.gb']
Strain "Z01319SL0004" found in files: ['2790345631_metadata.gb', '2791334217_metadata.gb']
Strain "Z01319SL0003" found in files: ['2790350405_metadata.gb', '2791334406_metadata.gb']
Strain "Z01319SL0005" found in files: ['2790355281_metadata.gb', '2791334263_metadata.gb']
Strain "Z01320SL0023" found in files: ['2790361485_metadata.gb', '2791327413_metadata.gb']
Strain "EC20111175" found in files: ['604278585_metadata.gb', '749297197_metadata.gb']
Strain "EC20111174" found in files: ['604290478_metadata.gb', '749296449_metadata.gb']
Strain "EC20100101" found in files: ['604336542_metadata.gb', '749298461_metadata.gb']
Strain "EC20090698" found in files: ['604362088_metadata.gb', '749298420_metadata.gb']
Strain "C500" found in files: ['674188659_metadata.gb', '749314944_metadata.gb']
Strain "CMCC found in files: ['747132831_metadata.gb', '747137039_metadata.gb']
Strain "EC20090641" found in files: ['896682452_metadata.gb', '915846032_metadata.gb']
Strain "SA02DT10168701" found in files: ['949867589_metadata.gb', '983377086_metadata.gb']
Strain "EC20121178" found in files: ['995901180_metadata.gb', '998623498_metadata.gb']
```

In [9]:

```python
from Bio import SeqIO
import os

def find_duplicate_files(fasta_directory):
    seen_sequences = {}
    duplicates_count = {}

    # Traverse all files in the given directory
    for filename in os.listdir(fasta_directory):
        if filename.endswith(".fasta"):
            filepath = os.path.join(fasta_directory, filename)
            for record in SeqIO.parse(filepath, "fasta"):
                sequence = str(record.seq)
                if sequence in seen_sequences:
                    # Increment count for the file that has this sequence
                    if seen_sequences[sequence] in duplicates_count:
                        duplicates_count[seen_sequences[sequence]].append(filepath)
                    else:
                        duplicates_count[seen_sequences[sequence]] = [filepath]
                else:
                    # Store the file in which this sequence was first seen
                    seen_sequences[sequence] = filepath

    # Flatten the duplicates_count dictionary and count duplicates
    flat_duplicates = [file for files in duplicates_count.values() for file in files]
    unique_duplicates = set(flat_duplicates)

    print(f"Number of duplicate files found: {len(unique_duplicates)}")

    # Print all duplicate file paths
    if unique_duplicates:
        print("Duplicate files:")
        for dup_file in unique_duplicates:
            print(dup_file)

# Set the path to your FASTA files directory
fasta_directory = "C:/Users/DD/Documents/Python Scripts/sequences"

# Call the function to find and count duplicate files
find_duplicate_files(fasta_directory)
```

```
Number of duplicate files found: 150
Duplicate files:
C:/Users/DD/Documents/Python Scripts/sequences\998623498.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2740278012.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1046302332.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1829031305.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2288537321.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791325747.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711464918.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791334100.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1033136052.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711495709.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2638491622.fasta
C:/Users/DD/Documents/Python Scripts/sequences\998647728.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1982652612.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2288890305.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791959093.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2258485671.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2521002406.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2572860206.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1995427272.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550213764.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328715.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2265635730.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791334217.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791964398.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1864594723.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550255387.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791329900.fasta
C:/Users/DD/Documents/Python Scripts/sequences\995907166.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711462317.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327513.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327776.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2279083905.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2638491583.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2717387261.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2476810552.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328750.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1046220323.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2168004414.fasta
C:/Users/DD/Documents/Python Scripts/sequences\749296449.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2339534824.fasta
C:/Users/DD/Documents/Python Scripts/sequences\995923317.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2060646254.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327413.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791964378.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2784851814.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321008513.fasta
C:/Users/DD/Documents/Python Scripts/sequences\749297197.fasta
C:/Users/DD/Documents/Python Scripts/sequences\983377086.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791332598.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1863657517.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327646.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711458674.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791334263.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711475272.fasta
C:/Users/DD/Documents/Python Scripts/sequences\749298461.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1863660810.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1635401364.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711461464.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327656.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1736666103.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550287957.fasta
C:/Users/DD/Documents/Python Scripts/sequences\674230814.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791964390.fasta
C:/Users/DD/Documents/Python Scripts/sequences\995907817.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711462605.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1634456755.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791333611.fasta
C:/Users/DD/Documents/Python Scripts/sequences\995922935.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1863662535.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321041129.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321042278.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2482324433.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1196285509.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711460722.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2279079344.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711462769.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791348005.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791964373.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2100237189.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1829022561.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791326870.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2294922788.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2288538671.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550213647.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711475867.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550225976.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321031274.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550213748.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1797869977.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2273089235.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1221158448.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711469117.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1736663574.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1828945320.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2288530062.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2550213792.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2638491693.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327658.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2515152310.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2036914008.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1945020687.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711458425.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2710805975.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2100356487.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2710806370.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1130930627.fasta
C:/Users/DD/Documents/Python Scripts/sequences\998642845.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321043586.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328192.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328843.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1128998525.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1864593957.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791326641.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2100166053.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2601133276.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791332929.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2339477813.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711459283.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321014208.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328717.fasta
```

```
C:/Users/DD/Documents/Python Scripts/sequences\1531933061.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2258710042.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791964384.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1733272414.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711488657.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2638510809.fasta
C:/Users/DD/Documents/Python Scripts/sequences\749314944.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328742.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327275.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2477330759.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2711466246.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1951777374.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1863661087.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2485141349.fasta
C:/Users/DD/Documents/Python Scripts/sequences\749298420.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328684.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791334406.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791323978.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791328489.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2087863205.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2078699049.fasta
C:/Users/DD/Documents/Python Scripts/sequences\771524686.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791959104.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2791327653.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2100896152.fasta
C:/Users/DD/Documents/Python Scripts/sequences\915846032.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2082554870.fasta
C:/Users/DD/Documents/Python Scripts/sequences\1546887966.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2321043499.fasta
C:/Users/DD/Documents/Python Scripts/sequences\2100424222.fasta
```

Checked duplicates based on metadata downloaded

In [11]:
```python
import os
import re
from collections import defaultdict

def extract_strain_info(metadata_directory):
    strain_info = defaultdict(list)

    # Iterate through all metadata files
    for filename in os.listdir(metadata_directory):
        if filename.endswith("_metadata.gb"):
            filepath = os.path.join(metadata_directory, filename)
            with open(filepath, "r") as file:
                content = file.read()
                # Extract strain information using regular expression
                match = re.search(r"strain=(\S+)", content)
                if match:
                    strain = match.group(1)
                    strain_info[strain].append(filename)

    # Count and print strains found in multiple files
    duplicate_strain_count = 0
    for strain, files in strain_info.items():
        if len(files) > 1:
            duplicate_strain_count += 1
            print(f"Strain {strain} found in {len(files)} files: {files}")

    # Print total count of duplicate strains
    print(f"\nTotal number of strains found in multiple files: {duplicate_strain_count}")

# Set the path to your metadata directory
metadata_directory = "C:/Users/DD/Documents/Python Scripts/sequences"
extract_strain_info(metadata_directory)
```

```
Strain "SA20095440" found in 2 files: ['1000369143_metadata.gb', '998647728_metadata.gb']
Strain "SA19960848" found in 2 files: ['1001551848_metadata.gb', '995907166_metadata.gb']
Strain "SA19970769" found in 2 files: ['1001552415_metadata.gb', '995922935_metadata.gb']
Strain "SA19970510" found in 2 files: ['1001554681_metadata.gb', '995907817_metadata.gb']
Strain "SA19980677" found in 2 files: ['1001555507_metadata.gb', '995923317_metadata.gb']
Strain "SA20094352" found in 2 files: ['1002985512_metadata.gb', '998642845_metadata.gb']
Strain "C629" found in 2 files: ['1030103469_metadata.gb', '1033136052_metadata.gb']
Strain "SA02DT09004001" found in 2 files: ['1045707385_metadata.gb', '1046220323_metadata.gb']
Strain "SA01AB09084001" found in 2 files: ['1045716295_metadata.gb', '1046302332_metadata.gb']
Strain "N55391" found in 2 files: ['1127388936_metadata.gb', '1128998525_metadata.gb']
Strain "87" found in 2 files: ['1129908976_metadata.gb', '1130930627_metadata.gb']
Strain "FORC_038" found in 2 files: ['1190123157_metadata.gb', '1196285509_metadata.gb']
Strain "D90" found in 2 files: ['1219800280_metadata.gb', '1221158448_metadata.gb']
Strain "ATCC found in 9 files: ['1528951495_metadata.gb', '1531933061_metadata.gb', '605520020_metadata.gb', '605528405_metadata.gb', '674281104_metadata.gb', '820762957_metadata.gb', '9148
28066_metadata.gb', '914835508_metadata.gb', '914845112_metadata.gb']
Strain "5" found in 2 files: ['1543688490_metadata.gb', '1546887966_metadata.gb']
Strain "JT found in 2 files: ['1633748201_metadata.gb', '1634456755_metadata.gb']
Strain "CD-SL01" found in 2 files: ['1634015463_metadata.gb', '1635401364_metadata.gb']
Strain "SJTUF87912v2" found in 2 files: ['1695497099_metadata.gb', '1736663574_metadata.gb']
Strain "SJTUF13520v2" found in 2 files: ['1695513588_metadata.gb', '1736666103_metadata.gb']
Strain "SL-312" found in 2 files: ['1732125948_metadata.gb', '1733272414_metadata.gb']
Strain "SCSM4.1" found in 2 files: ['1788239996_metadata.gb', '1797869977_metadata.gb']
Strain "SE124" found in 2 files: ['1828351404_metadata.gb', '1828945320_metadata.gb']
Strain "SI85" found in 2 files: ['1828458697_metadata.gb', '1829031305_metadata.gb']
Strain "SI67" found in 2 files: ['1828463287_metadata.gb', '1829022561_metadata.gb']
Strain "CVM found in 67 files: ['1836411705_metadata.gb', '1842989566_metadata.gb', '1843494016_metadata.gb', '1843498400_metadata.gb', '1843507307_metadata.gb', '1843512211_metadata.gb',
'1843516861_metadata.gb', '1843531062_metadata.gb', '1843540331_metadata.gb', '1843544971_metadata.gb', '1843576148_metadata.gb', '1843580775_metadata.gb', '1843585404_metadata.gb', '184358
9994_metadata.gb', '1843599223_metadata.gb', '1843608483_metadata.gb', '1843613175_metadata.gb', '1843618362_metadata.gb', '1843629339_metadata.gb', '1843633944_metadata.gb', '1843643169_me
tadata.gb', '1843647868_metadata.gb', '1843652501_metadata.gb', '1846455821_metadata.gb', '1951535343_metadata.gb', '1951777374_metadata.gb', '2090918790_metadata.gb', '2090956231_metadata.
gb', '2090974716_metadata.gb', '2091165092_metadata.gb', '2091169855_metadata.gb', '2091174439_metadata.gb', '2091178941_metadata.gb', '2091183720_metadata.gb', '2091229717_metadata.gb', '2
091239126_metadata.gb', '2091243740_metadata.gb', '2091253165_metadata.gb', '2091264036_metadata.gb', '2091277794_metadata.gb', '2091323346_metadata.gb', '2091333084_metadata.gb', '20913468
97_metadata.gb', '2091356744_metadata.gb', '2091370368_metadata.gb', '2091375002_metadata.gb', '2091402345_metadata.gb', '2091418357_metadata.gb', '2091422800_metadata.gb', '2091431810_meta
data.gb', '2091446076_metadata.gb', '2091455043_metadata.gb', '2091469148_metadata.gb', '2091473540_metadata.gb', '2091478163_metadata.gb', '2091482867_metadata.gb', '2136510056_metadata.g
b', '2518923471_metadata.gb', '2521002406_metadata.gb', '696581462_metadata.gb', '696593549_metadata.gb', '744788240_metadata.gb', '767833982_metadata.gb', '767838360_metadata.gb', '7678427
34_metadata.gb', '767847001_metadata.gb', '771524686_metadata.gb']
Strain "VNSEC031" found in 2 files: ['1863533289_metadata.gb', '1863660810_metadata.gb']
Strain "VNSEC023" found in 2 files: ['1863537930_metadata.gb', '1863661087_metadata.gb']
Strain "VNSEC013" found in 2 files: ['1863542555_metadata.gb', '1863657517_metadata.gb']
Strain "VNSEC003" found in 2 files: ['1863547333_metadata.gb', '1863662535_metadata.gb']
Strain "VNSEC002" found in 2 files: ['1863552396_metadata.gb', '1864594723_metadata.gb']
Strain "VNSEC001" found in 2 files: ['1863557100_metadata.gb', '1864593957_metadata.gb']
Strain "R16.1424" found in 2 files: ['1943036513_metadata.gb', '1945020687_metadata.gb']
Strain "GSJ/2016-Sal-017" found in 2 files: ['1979358710_metadata.gb', '1982652612_metadata.gb']
Strain "R16.0556" found in 2 files: ['1993496820_metadata.gb', '1995427272_metadata.gb']
Strain "KUFSE-SAL0043" found in 2 files: ['2034574262_metadata.gb', '2036914008_metadata.gb']
Strain "CFSAN008081" found in 2 files: ['2049354170_metadata.gb', '2082554870_metadata.gb']
Strain "YZ20MCS16" found in 2 files: ['2060222687_metadata.gb', '2060646254_metadata.gb']
Strain "SA18578" found in 2 files: ['2077453074_metadata.gb', '2078699049_metadata.gb']
Strain "S16" found in 2 files: ['2087468524_metadata.gb', '2087863205_metadata.gb']
Strain "ZC-S1 found in 2 files: ['2093998598_metadata.gb', '2100166053_metadata.gb']
Strain "S90" found in 2 files: ['2094623483_metadata.gb', '2100356487_metadata.gb']
Strain "SP" found in 2 files: ['2094623485_metadata.gb', '2100424222_metadata.gb']
Strain "S146" found in 2 files: ['2094642037_metadata.gb', '2100237189_metadata.gb']
Strain "07Q015" found in 2 files: ['2095951993_metadata.gb', '2100896152_metadata.gb']
Strain "YZ21MCS4" found in 2 files: ['2167845084_metadata.gb', '2168004414_metadata.gb']
Strain "YZ20MCS6" found in 2 files: ['2258055768_metadata.gb', '2258485671_metadata.gb']
Strain "YZ20MCS14" found in 2 files: ['2258066866_metadata.gb', '2258710042_metadata.gb']
Strain "SE006" found in 2 files: ['2263926704_metadata.gb', '2265635730_metadata.gb']
Strain "013+" found in 2 files: ['2271345254_metadata.gb', '2273089235_metadata.gb']
Strain "KCID6" found in 2 files: ['2277353412_metadata.gb', '2279079344_metadata.gb']
Strain "BCID6" found in 2 files: ['2277362348_metadata.gb', '2279083905_metadata.gb']
Strain "AH19MCS8" found in 2 files: ['2288173494_metadata.gb', '2288530062_metadata.gb']
Strain "XZ14C1328" found in 2 files: ['2288320740_metadata.gb', '2288890305_metadata.gb']
Strain "MRS17_00712" found in 2 files: ['2294456242_metadata.gb', '2294922788_metadata.gb']
Strain "190821_1" found in 2 files: ['2320729955_metadata.gb', '2321014208_metadata.gb']
Strain "190819_2" found in 2 files: ['2320734399_metadata.gb', '2321042278_metadata.gb']
Strain "190807_1" found in 2 files: ['2320738824_metadata.gb', '2321031274_metadata.gb']
Strain "190729_8" found in 2 files: ['2320748473_metadata.gb', '2321043499_metadata.gb']
Strain "190729_4" found in 2 files: ['2320752890_metadata.gb', '2321043586_metadata.gb']
Strain "190704_2" found in 2 files: ['2320757305_metadata.gb', '2321008513_metadata.gb']
Strain "190610_1" found in 2 files: ['2320761748_metadata.gb', '2321041129_metadata.gb']
Strain "XM3104" found in 2 files: ['2333213671_metadata.gb', '2339534824_metadata.gb']
Strain "CVCC found in 2 files: ['2333272181_metadata.gb', '2339477813_metadata.gb']
Strain "R22.0044" found in 2 files: ['2475999171_metadata.gb', '2476810552_metadata.gb']
Strain "013" found in 2 files: ['2476413604_metadata.gb', '2477330759_metadata.gb']
Strain "S1467" found in 2 files: ['2482050684_metadata.gb', '2482324433_metadata.gb']
Strain "CRIN508879" found in 2 files: ['2484128728_metadata.gb', '2485141349_metadata.gb']
Strain "CRSE-01" found in 2 files: ['2514088622_metadata.gb', '2515152310_metadata.gb']
Strain "SECVM-15" found in 2 files: ['2548245062_metadata.gb', '2550287957_metadata.gb']
Strain "SECVM-17" found in 2 files: ['2548250070_metadata.gb', '2550213647_metadata.gb']
Strain "SE found in 2 files: ['2548260632_metadata.gb', '2550213764_metadata.gb']
Strain "SECVM-13" found in 2 files: ['2548275539_metadata.gb', '2550225976_metadata.gb']
Strain "SECVM-10" found in 2 files: ['2548291250_metadata.gb', '2550255387_metadata.gb']
Strain "SECVM-14" found in 2 files: ['2548296188_metadata.gb', '2550213748_metadata.gb']
Strain "SECVM-5" found in 2 files: ['2548306062_metadata.gb', '2550213792_metadata.gb']
Strain "Z1323CSL0015" found in 4 files: ['2570622740_metadata.gb', '2572860206_metadata.gb', '2703404023_metadata.gb', '2711459283_metadata.gb']
Strain "LA5" found in 2 files: ['2595857884_metadata.gb', '2601133276_metadata.gb']
Strain "405987R1_S50" found in 2 files: ['2636846570_metadata.gb', '2638491583_metadata.gb']
Strain "401964R1_S49" found in 2 files: ['2636851041_metadata.gb', '2638491693_metadata.gb']
Strain "180121R1S_S48" found in 2 files: ['2636855593_metadata.gb', '2638491622_metadata.gb']
Strain "SSSE-01" found in 2 files: ['2679318990_metadata.gb', '2710805975_metadata.gb']
Strain "SSSE-03" found in 2 files: ['2679323484_metadata.gb', '2710806370_metadata.gb']
Strain "Z1323CSL0016" found in 2 files: ['2703413299_metadata.gb', '2711458674_metadata.gb']
Strain "Z1323CSL0014" found in 2 files: ['2703417940_metadata.gb', '2711458425_metadata.gb']
Strain "Z1323CSL0027" found in 2 files: ['2703422591_metadata.gb', '2711461464_metadata.gb']
Strain "Z1323CSL0034" found in 2 files: ['2703436556_metadata.gb', '2711469117_metadata.gb']
Strain "Z1323CSL0017" found in 2 files: ['2703441194_metadata.gb', '2711464918_metadata.gb']
Strain "Z1323CSL0002" found in 2 files: ['2703445817_metadata.gb', '2711466246_metadata.gb']
Strain "Z1323CSL0006" found in 2 files: ['2703450442_metadata.gb', '2711460722_metadata.gb']
Strain "Z1323CSL0062" found in 2 files: ['2704865509_metadata.gb', '2711462317_metadata.gb']
Strain "Z1323CSL0057" found in 2 files: ['2704874751_metadata.gb', '2711475867_metadata.gb']
Strain "Z1323CSL0053" found in 2 files: ['2704891984_metadata.gb', '2711495709_metadata.gb']
Strain "Z1323CSL0045" found in 2 files: ['2704905772_metadata.gb', '2711488657_metadata.gb']
Strain "Z1323CSL0047" found in 2 files: ['2704910405_metadata.gb', '2711475272_metadata.gb']
Strain "Z1323CSL0052" found in 2 files: ['2704915029_metadata.gb', '2711462605_metadata.gb']
Strain "Z1323CSL0051" found in 2 files: ['2704924300_metadata.gb', '2711462769_metadata.gb']
Strain "ST.GS36" found in 2 files: ['2717024100_metadata.gb', '2717387261_metadata.gb']
Strain "IJCS5-22" found in 2 files: ['2739185197_metadata.gb', '2740278012_metadata.gb']
Strain "H4" found in 2 files: ['2784324973_metadata.gb', '2784851814_metadata.gb']
Strain "Z01320SL0044" found in 2 files: ['2790142919_metadata.gb', '2791959093_metadata.gb']
Strain "Z01320SL0045" found in 2 files: ['2790153796_metadata.gb', '2791964378_metadata.gb']
Strain "Z01320SL0042" found in 2 files: ['2790167917_metadata.gb', '2791959104_metadata.gb']
Strain "Z01320SL0040" found in 2 files: ['2790172290_metadata.gb', '2791964398_metadata.gb']
Strain "Z01320SL0041" found in 2 files: ['2790176664_metadata.gb', '2791964390_metadata.gb']
Strain "Z01320SL0038" found in 2 files: ['2790181112_metadata.gb', '2791964373_metadata.gb']
Strain "Z01320SL0043" found in 2 files: ['2790191447_metadata.gb', '2791964384_metadata.gb']
Strain "Z01320SL0030" found in 2 files: ['2790219725_metadata.gb', '2791325747_metadata.gb']
Strain "Z01320SL0033" found in 2 files: ['2790226660_metadata.gb', '2791348005_metadata.gb']
Strain "Z01320SL0034" found in 2 files: ['2790232908_metadata.gb', '2791323978_metadata.gb']
Strain "Z01320SL0031" found in 2 files: ['2790238422_metadata.gb', '2791328742_metadata.gb']
Strain "Z01320SL0036" found in 2 files: ['2790243305_metadata.gb', '2791327776_metadata.gb']
Strain "Z01320SL0037" found in 2 files: ['2790247716_metadata.gb', '2791329900_metadata.gb']
Strain "Z01320SL0035" found in 2 files: ['2790252121_metadata.gb', '2791327646_metadata.gb']
Strain "Z01320SL0032" found in 2 files: ['2790256505_metadata.gb', '2791332598_metadata.gb']
Strain "Z01319SL0021" found in 2 files: ['2790261062_metadata.gb', '2791326641_metadata.gb']
Strain "Z01319SL0015" found in 2 files: ['2790265499_metadata.gb', '2791328717_metadata.gb']
Strain "Z01319SL0020" found in 2 files: ['2790270044_metadata.gb', '2791327275_metadata.gb']
Strain "Z01320SL0029" found in 2 files: ['2790274631_metadata.gb', '2791332929_metadata.gb']
Strain "Z01319SL0016" found in 2 files: ['2790279127_metadata.gb', '2791327653_metadata.gb']
Strain "Z01319SL0017" found in 2 files: ['2790283963_metadata.gb', '2791328843_metadata.gb']
Strain "Z01319SL0022" found in 2 files: ['2790292938_metadata.gb', '2791327656_metadata.gb']
Strain "Z01320SL0026" found in 2 files: ['2790297568_metadata.gb', '2791334100_metadata.gb']
Strain "Z01320SL0027" found in 2 files: ['2790302094_metadata.gb', '2791328684_metadata.gb']
Strain "Z01319SL0009" found in 2 files: ['2790306333_metadata.gb', '2791327658_metadata.gb']
Strain "Z01319SL0019" found in 2 files: ['2790311342_metadata.gb', '2791328750_metadata.gb']
Strain "Z01319SL0010" found in 2 files: ['2790316599_metadata.gb', '2791326870_metadata.gb']
Strain "Z01320SL0025" found in 2 files: ['2790321214_metadata.gb', '2791333611_metadata.gb']
Strain "Z01320SL0024" found in 2 files: ['2790326108_metadata.gb', '2791328192_metadata.gb']
Strain "Z01319SL0008" found in 2 files: ['2790330984_metadata.gb', '2791327513_metadata.gb']
Strain "Z01319SL0006" found in 2 files: ['2790335951_metadata.gb', '2791328715_metadata.gb']
Strain "Z01319SL0001" found in 2 files: ['2790340519_metadata.gb', '2791328489_metadata.gb']
Strain "Z01319SL0004" found in 2 files: ['2790345631_metadata.gb', '2791334217_metadata.gb']
Strain "Z01319SL0003" found in 2 files: ['2790350405_metadata.gb', '2791334406_metadata.gb']
Strain "Z01319SL0005" found in 2 files: ['2790355281_metadata.gb', '2791334263_metadata.gb']
Strain "Z01320SL0023" found in 2 files: ['2790361485_metadata.gb', '2791327413_metadata.gb']
Strain "EC20111175" found in 2 files: ['604278585_metadata.gb', '749297197_metadata.gb']
Strain "EC20111174" found in 2 files: ['604290478_metadata.gb', '749296449_metadata.gb']
Strain "EC20100101" found in 2 files: ['604336542_metadata.gb', '749298461_metadata.gb']
Strain "EC20090698" found in 2 files: ['604362088_metadata.gb', '749298420_metadata.gb']
Strain "C500" found in 2 files: ['674188659_metadata.gb', '749314944_metadata.gb']
Strain "CMCC found in 2 files: ['747132831_metadata.gb', '747137039_metadata.gb']
Strain "EC20090641" found in 2 files: ['896682452_metadata.gb', '915846032_metadata.gb']
Strain "SA02DT10168701" found in 2 files: ['949867589_metadata.gb', '983377086_metadata.gb']
Strain "EC20121178" found in 2 files: ['995901180_metadata.gb', '998623498_metadata.gb']

Total number of strains found in multiple files: 144
```

In [13]:
```python
import os
import re
from collections import defaultdict

def extract_strain_info(metadata_directory):
    strain_info = defaultdict(list)

    # Iterate through all metadata files
    for filename in os.listdir(metadata_directory):
        if filename.endswith("_metadata.gb"):
            filepath = os.path.join(metadata_directory, filename)
            with open(filepath, "r") as file:
                content = file.read()
                # Extract strain information using regular expression
                match = re.search(r"strain=(\S+)", content)
                if match:
                    strain = match.group(1)
                    strain_info[strain].append(filename)

    # Count strains and files
    strain_file_counts = defaultdict(int)
    for files in strain_info.values():
        for file in files:
            strain_file_counts[file] += 1

    # Print strains and file counts
    for file, count in strain_file_counts.items():
        print(f"File {file} contains {count} strains")

    # Print the number of duplicate strains
    duplicate_strain_count = sum(1 for files in strain_info.values() if len(files) > 1)
    print(f"\nTotal number of strains found in multiple files: {duplicate_strain_count}")

# Set the path to your metadata directory
metadata_directory = "C:/Users/DD/Documents/Python Scripts/sequences"
extract_strain_info(metadata_directory)
```

```
File 1000369143_metadata.gb contains 1 strains
File 998647728_metadata.gb contains 1 strains
File 1001427523_metadata.gb contains 1 strains
File 1001551848_metadata.gb contains 1 strains
File 995907166_metadata.gb contains 1 strains
File 1001552415_metadata.gb contains 1 strains
File 995922935_metadata.gb contains 1 strains
File 1001554681_metadata.gb contains 1 strains
File 995907817_metadata.gb contains 1 strains
File 1001555507_metadata.gb contains 1 strains
File 995923317_metadata.gb contains 1 strains
File 1002000489_metadata.gb contains 1 strains
File 1002707584_metadata.gb contains 1 strains
File 1002883852_metadata.gb contains 1 strains
File 1002985512_metadata.gb contains 1 strains
File 998642845_metadata.gb contains 1 strains
File 1018838233_metadata.gb contains 1 strains
File 1020321938_metadata.gb contains 1 strains
File 1020327374_metadata.gb contains 1 strains
File 1020336109_metadata.gb contains 1 strains
File 1020345203_metadata.gb contains 1 strains
File 1021436845_metadata.gb contains 1 strains
File 1030103469_metadata.gb contains 1 strains
File 1033136052_metadata.gb contains 1 strains
File 1045707385_metadata.gb contains 1 strains
File 1046220323_metadata.gb contains 1 strains
File 1045716295_metadata.gb contains 1 strains
File 1046302332_metadata.gb contains 1 strains
File 1046300946_metadata.gb contains 1 strains
File 1046591696_metadata.gb contains 1 strains
File 1046611151_metadata.gb contains 1 strains
File 1046630596_metadata.gb contains 1 strains
File 1046637175_metadata.gb contains 1 strains
File 1110727957_metadata.gb contains 1 strains
File 1127379580_metadata.gb contains 1 strains
File 1127384256_metadata.gb contains 1 strains
File 1127388936_metadata.gb contains 1 strains
File 1128998525_metadata.gb contains 1 strains
File 1129908976_metadata.gb contains 1 strains
File 1130930627_metadata.gb contains 1 strains
File 1190123157_metadata.gb contains 1 strains
File 1196285509_metadata.gb contains 1 strains
File 1219800280_metadata.gb contains 1 strains
File 1221158448_metadata.gb contains 1 strains
File 1227482702_metadata.gb contains 1 strains
File 1302554428_metadata.gb contains 1 strains
File 1302574149_metadata.gb contains 1 strains
File 1486618283_metadata.gb contains 1 strains
File 1491767611_metadata.gb contains 1 strains
File 1528951495_metadata.gb contains 1 strains
File 1531933061_metadata.gb contains 1 strains
File 605520020_metadata.gb contains 1 strains
File 605528405_metadata.gb contains 1 strains
File 674281104_metadata.gb contains 1 strains
File 820762957_metadata.gb contains 1 strains
File 914828066_metadata.gb contains 1 strains
File 914835508_metadata.gb contains 1 strains
File 914845112_metadata.gb contains 1 strains
File 1543688490_metadata.gb contains 1 strains
File 1546887966_metadata.gb contains 1 strains
File 1633748201_metadata.gb contains 1 strains
File 1634456755_metadata.gb contains 1 strains
File 1634015463_metadata.gb contains 1 strains
File 1635401364_metadata.gb contains 1 strains
File 1691492771_metadata.gb contains 1 strains
File 1695497099_metadata.gb contains 1 strains
File 1736663574_metadata.gb contains 1 strains
File 1695513588_metadata.gb contains 1 strains
File 1736666103_metadata.gb contains 1 strains
File 1699456900_metadata.gb contains 1 strains
File 1732125948_metadata.gb contains 1 strains
File 1733272414_metadata.gb contains 1 strains
File 1788239996_metadata.gb contains 1 strains
File 1797869977_metadata.gb contains 1 strains
File 1828351404_metadata.gb contains 1 strains
File 1828945320_metadata.gb contains 1 strains
File 1828458697_metadata.gb contains 1 strains
File 1829031305_metadata.gb contains 1 strains
File 1828463287_metadata.gb contains 1 strains
File 1829022561_metadata.gb contains 1 strains
File 1836411705_metadata.gb contains 1 strains
File 1842989566_metadata.gb contains 1 strains
File 1843494016_metadata.gb contains 1 strains
File 1843498400_metadata.gb contains 1 strains
File 1843507307_metadata.gb contains 1 strains
File 1843512211_metadata.gb contains 1 strains
File 1843516861_metadata.gb contains 1 strains
File 1843531062_metadata.gb contains 1 strains
File 1843540331_metadata.gb contains 1 strains
File 1843544971_metadata.gb contains 1 strains
File 1843576148_metadata.gb contains 1 strains
File 1843580775_metadata.gb contains 1 strains
File 1843585404_metadata.gb contains 1 strains
File 1843589994_metadata.gb contains 1 strains
File 1843599223_metadata.gb contains 1 strains
File 1843608483_metadata.gb contains 1 strains
File 1843613175_metadata.gb contains 1 strains
File 1843618362_metadata.gb contains 1 strains
File 1843629339_metadata.gb contains 1 strains
File 1843633944_metadata.gb contains 1 strains
File 1843643169_metadata.gb contains 1 strains
File 1843647868_metadata.gb contains 1 strains
File 1843652501_metadata.gb contains 1 strains
File 1846455821_metadata.gb contains 1 strains
File 1951535343_metadata.gb contains 1 strains
File 1951777374_metadata.gb contains 1 strains
File 2090918790_metadata.gb contains 1 strains
File 2090956231_metadata.gb contains 1 strains
File 2090974716_metadata.gb contains 1 strains
File 2091165092_metadata.gb contains 1 strains
File 2091169855_metadata.gb contains 1 strains
File 2091174439_metadata.gb contains 1 strains
File 2091178941_metadata.gb contains 1 strains
File 2091183720_metadata.gb contains 1 strains
File 2091229717_metadata.gb contains 1 strains
File 2091239126_metadata.gb contains 1 strains
File 2091243740_metadata.gb contains 1 strains
File 2091253165_metadata.gb contains 1 strains
File 2091264036_metadata.gb contains 1 strains
File 2091277794_metadata.gb contains 1 strains
File 2091323346_metadata.gb contains 1 strains
File 2091333084_metadata.gb contains 1 strains
File 2091346897_metadata.gb contains 1 strains
File 2091356744_metadata.gb contains 1 strains
File 2091370368_metadata.gb contains 1 strains
File 2091375002_metadata.gb contains 1 strains
File 2091402345_metadata.gb contains 1 strains
File 2091418357_metadata.gb contains 1 strains
File 2091422800_metadata.gb contains 1 strains
File 2091431810_metadata.gb contains 1 strains
File 2091446076_metadata.gb contains 1 strains
File 2091455043_metadata.gb contains 1 strains
File 2091469148_metadata.gb contains 1 strains
File 2091473540_metadata.gb contains 1 strains
File 2091478163_metadata.gb contains 1 strains
File 2091482867_metadata.gb contains 1 strains
File 2136510056_metadata.gb contains 1 strains
File 2518923471_metadata.gb contains 1 strains
File 2521002406_metadata.gb contains 1 strains
File 696581462_metadata.gb contains 1 strains
File 696593549_metadata.gb contains 1 strains
File 744788240_metadata.gb contains 1 strains
File 767833982_metadata.gb contains 1 strains
File 767838360_metadata.gb contains 1 strains
File 767842734_metadata.gb contains 1 strains
File 767847001_metadata.gb contains 1 strains
File 771524686_metadata.gb contains 1 strains
File 1863533289_metadata.gb contains 1 strains
File 1863660810_metadata.gb contains 1 strains
File 1863537930_metadata.gb contains 1 strains
File 1863661087_metadata.gb contains 1 strains
File 1863542555_metadata.gb contains 1 strains
File 1863657517_metadata.gb contains 1 strains
File 1863547333_metadata.gb contains 1 strains
File 1863662535_metadata.gb contains 1 strains
File 1863552396_metadata.gb contains 1 strains
File 1864594723_metadata.gb contains 1 strains
File 1863557100_metadata.gb contains 1 strains
File 1864593957_metadata.gb contains 1 strains
File 1891180034_metadata.gb contains 1 strains
File 1943036513_metadata.gb contains 1 strains
File 1945020687_metadata.gb contains 1 strains
File 1948849906_metadata.gb contains 1 strains
File 1953481399_metadata.gb contains 1 strains
File 1953486234_metadata.gb contains 1 strains
File 1953490790_metadata.gb contains 1 strains
File 1953495564_metadata.gb contains 1 strains
File 1953500321_metadata.gb contains 1 strains
File 1953508822_metadata.gb contains 1 strains
File 1953513472_metadata.gb contains 1 strains
File 1979358710_metadata.gb contains 1 strains
File 1982652612_metadata.gb contains 1 strains
File 1993496820_metadata.gb contains 1 strains
File 1995427272_metadata.gb contains 1 strains
File 2030195509_metadata.gb contains 1 strains
File 2030197290_metadata.gb contains 1 strains
File 2030199816_metadata.gb contains 1 strains
File 2030201146_metadata.gb contains 1 strains
File 2030202676_metadata.gb contains 1 strains
File 2030203977_metadata.gb contains 1 strains
File 2034574262_metadata.gb contains 1 strains
File 2036914008_metadata.gb contains 1 strains
File 2035051374_metadata.gb contains 1 strains
File 2035064789_metadata.gb contains 1 strains
File 2049354170_metadata.gb contains 1 strains
File 2082554870_metadata.gb contains 1 strains
File 2049446339_metadata.gb contains 1 strains
File 2060222687_metadata.gb contains 1 strains
File 2060646254_metadata.gb contains 1 strains
File 2077453074_metadata.gb contains 1 strains
File 2078699049_metadata.gb contains 1 strains
File 2087468524_metadata.gb contains 1 strains
File 2087863205_metadata.gb contains 1 strains
File 2091029751_metadata.gb contains 1 strains
File 2091117943_metadata.gb contains 1 strains
File 2091122469_metadata.gb contains 1 strains
File 2093987282_metadata.gb contains 1 strains
File 2093998598_metadata.gb contains 1 strains
File 2100166053_metadata.gb contains 1 strains
File 2094623483_metadata.gb contains 1 strains
File 2100356487_metadata.gb contains 1 strains
File 2094623485_metadata.gb contains 1 strains
File 2100424222_metadata.gb contains 1 strains
File 2094623487_metadata.gb contains 1 strains
File 2094642037_metadata.gb contains 1 strains
File 2100237189_metadata.gb contains 1 strains
File 2095951993_metadata.gb contains 1 strains
File 2100896152_metadata.gb contains 1 strains
File 2167845084_metadata.gb contains 1 strains
File 2168004414_metadata.gb contains 1 strains
File 2258055768_metadata.gb contains 1 strains
File 2258485671_metadata.gb contains 1 strains
File 2258066866_metadata.gb contains 1 strains
File 2258710042_metadata.gb contains 1 strains
File 2263926704_metadata.gb contains 1 strains
File 2265635730_metadata.gb contains 1 strains
File 2271345254_metadata.gb contains 1 strains
File 2273089235_metadata.gb contains 1 strains
File 2277353412_metadata.gb contains 1 strains
File 2279079344_metadata.gb contains 1 strains
File 2277362348_metadata.gb contains 1 strains
File 2279083905_metadata.gb contains 1 strains
File 2285411033_metadata.gb contains 1 strains
File 2288173494_metadata.gb contains 1 strains
File 2288530062_metadata.gb contains 1 strains
File 2288320740_metadata.gb contains 1 strains
File 2288890305_metadata.gb contains 1 strains
File 2294456242_metadata.gb contains 1 strains
File 2294922788_metadata.gb contains 1 strains
File 2320729955_metadata.gb contains 1 strains
File 2321014208_metadata.gb contains 1 strains
File 2320734399_metadata.gb contains 1 strains
File 2321042278_metadata.gb contains 1 strains
File 2320738824_metadata.gb contains 1 strains
File 2321031274_metadata.gb contains 1 strains
File 2320748473_metadata.gb contains 1 strains
File 2321043499_metadata.gb contains 1 strains
File 2320752890_metadata.gb contains 1 strains
File 2321043586_metadata.gb contains 1 strains
File 2320757305_metadata.gb contains 1 strains
File 2321008513_metadata.gb contains 1 strains
File 2320761748_metadata.gb contains 1 strains
File 2321041129_metadata.gb contains 1 strains
File 2325662964_metadata.gb contains 1 strains
File 2333213671_metadata.gb contains 1 strains
File 2339534824_metadata.gb contains 1 strains
File 2333272181_metadata.gb contains 1 strains
File 2339477813_metadata.gb contains 1 strains
File 2359423532_metadata.gb contains 1 strains
File 2459518531_metadata.gb contains 1 strains
File 2459856665_metadata.gb contains 1 strains
File 2459861217_metadata.gb contains 1 strains
File 2459865826_metadata.gb contains 1 strains
File 2475999171_metadata.gb contains 1 strains
File 2476810552_metadata.gb contains 1 strains
File 2476413604_metadata.gb contains 1 strains
File 2477330759_metadata.gb contains 1 strains
File 2482050684_metadata.gb contains 1 strains
File 2482324433_metadata.gb contains 1 strains
File 2484128728_metadata.gb contains 1 strains
File 2485141349_metadata.gb contains 1 strains
File 2514088622_metadata.gb contains 1 strains
File 2515152310_metadata.gb contains 1 strains
File 2548245062_metadata.gb contains 1 strains
File 2550287957_metadata.gb contains 1 strains
File 2548250070_metadata.gb contains 1 strains
File 2550213647_metadata.gb contains 1 strains
File 2548260632_metadata.gb contains 1 strains
File 2550213764_metadata.gb contains 1 strains
File 2548275539_metadata.gb contains 1 strains
File 2550225976_metadata.gb contains 1 strains
File 2548291250_metadata.gb contains 1 strains
File 2550255387_metadata.gb contains 1 strains
File 2548296188_metadata.gb contains 1 strains
File 2550213748_metadata.gb contains 1 strains
File 2548306062_metadata.gb contains 1 strains
File 2550213792_metadata.gb contains 1 strains
File 2567050896_metadata.gb contains 1 strains
File 2570622740_metadata.gb contains 1 strains
File 2572860206_metadata.gb contains 1 strains
File 2703404023_metadata.gb contains 1 strains
File 2711459283_metadata.gb contains 1 strains
File 2595857884_metadata.gb contains 1 strains
File 2601133276_metadata.gb contains 1 strains
File 2636846570_metadata.gb contains 1 strains
File 2638491583_metadata.gb contains 1 strains
File 2636851041_metadata.gb contains 1 strains
File 2638491693_metadata.gb contains 1 strains
File 2636855593_metadata.gb contains 1 strains
File 2638491622_metadata.gb contains 1 strains
File 2666584953_metadata.gb contains 1 strains
File 2679318990_metadata.gb contains 1 strains
File 2710805975_metadata.gb contains 1 strains
File 2679323484_metadata.gb contains 1 strains
File 2710806370_metadata.gb contains 1 strains
File 267991652_metadata.gb contains 1 strains
File 2689075821_metadata.gb contains 1 strains
File 2689080905_metadata.gb contains 1 strains
File 2694694140_metadata.gb contains 1 strains
File 2703413299_metadata.gb contains 1 strains
File 2711458674_metadata.gb contains 1 strains
File 2703417940_metadata.gb contains 1 strains
File 2711458425_metadata.gb contains 1 strains
File 2703422591_metadata.gb contains 1 strains
File 2711461464_metadata.gb contains 1 strains
File 2703427236_metadata.gb contains 1 strains
File 2703436556_metadata.gb contains 1 strains
File 2711469117_metadata.gb contains 1 strains
File 2703441194_metadata.gb contains 1 strains
File 2711464918_metadata.gb contains 1 strains
File 2703445817_metadata.gb contains 1 strains
File 2711466246_metadata.gb contains 1 strains
File 2703450442_metadata.gb contains 1 strains
File 2711460722_metadata.gb contains 1 strains
File 2703459908_metadata.gb contains 1 strains
File 2704865509_metadata.gb contains 1 strains
File 2711462317_metadata.gb contains 1 strains
File 2704874751_metadata.gb contains 1 strains
File 2711475867_metadata.gb contains 1 strains
File 2704891984_metadata.gb contains 1 strains
File 2711495709_metadata.gb contains 1 strains
File 2704905772_metadata.gb contains 1 strains
File 2711488657_metadata.gb contains 1 strains
File 2704910405_metadata.gb contains 1 strains
File 2711475272_metadata.gb contains 1 strains
File 2704915029_metadata.gb contains 1 strains
File 2711462605_metadata.gb contains 1 strains
File 2704924300_metadata.gb contains 1 strains
File 2711462769_metadata.gb contains 1 strains
File 2717024100_metadata.gb contains 1 strains
File 2717387261_metadata.gb contains 1 strains
File 2739185197_metadata.gb contains 1 strains
File 2740278012_metadata.gb contains 1 strains
File 2747284265_metadata.gb contains 1 strains
File 2784324973_metadata.gb contains 1 strains
File 2784851814_metadata.gb contains 1 strains
File 2790142919_metadata.gb contains 1 strains
File 2791959093_metadata.gb contains 1 strains
File 2790153796_metadata.gb contains 1 strains
File 2791964378_metadata.gb contains 1 strains
File 2790167917_metadata.gb contains 1 strains
File 2791959104_metadata.gb contains 1 strains
File 2790172290_metadata.gb contains 1 strains
File 2791964398_metadata.gb contains 1 strains
File 2790176664_metadata.gb contains 1 strains
File 2791964390_metadata.gb contains 1 strains
File 2790181112_metadata.gb contains 1 strains
File 2791964373_metadata.gb contains 1 strains
File 2790185892_metadata.gb contains 1 strains
File 2790191447_metadata.gb contains 1 strains
File 2791964384_metadata.gb contains 1 strains
File 2790219725_metadata.gb contains 1 strains
File 2791325747_metadata.gb contains 1 strains
File 2790226660_metadata.gb contains 1 strains
File 2791348005_metadata.gb contains 1 strains
File 2790232908_metadata.gb contains 1 strains
File 2791323978_metadata.gb contains 1 strains
File 2790238422_metadata.gb contains 1 strains
File 2791328742_metadata.gb contains 1 strains
File 2790243305_metadata.gb contains 1 strains
File 2791327776_metadata.gb contains 1 strains
File 2790247716_metadata.gb contains 1 strains
File 2791329900_metadata.gb contains 1 strains
File 2790252121_metadata.gb contains 1 strains
File 2791327646_metadata.gb contains 1 strains
File 2790256505_metadata.gb contains 1 strains
File 2791332598_metadata.gb contains 1 strains
File 2790261062_metadata.gb contains 1 strains
File 2791326641_metadata.gb contains 1 strains
File 2790265499_metadata.gb contains 1 strains
File 2791328717_metadata.gb contains 1 strains
File 2790270044_metadata.gb contains 1 strains
File 2791327275_metadata.gb contains 1 strains
File 2790274631_metadata.gb contains 1 strains
File 2791332929_metadata.gb contains 1 strains
File 2790279127_metadata.gb contains 1 strains
File 2791327653_metadata.gb contains 1 strains
File 2790283963_metadata.gb contains 1 strains
File 2791328843_metadata.gb contains 1 strains
File 2790288563_metadata.gb contains 1 strains
File 2790292938_metadata.gb contains 1 strains
File 2791327656_metadata.gb contains 1 strains
File 2790297568_metadata.gb contains 1 strains
File 2791334100_metadata.gb contains 1 strains
File 2790302094_metadata.gb contains 1 strains
File 2791328684_metadata.gb contains 1 strains
File 2790306333_metadata.gb contains 1 strains
File 2791327658_metadata.gb contains 1 strains
File 2790311342_metadata.gb contains 1 strains
File 2791328750_metadata.gb contains 1 strains
File 2790316599_metadata.gb contains 1 strains
File 2791326870_metadata.gb contains 1 strains
File 2790321214_metadata.gb contains 1 strains
File 2791333611_metadata.gb contains 1 strains
File 2790326108_metadata.gb contains 1 strains
File 2791328192_metadata.gb contains 1 strains
File 2790330984_metadata.gb contains 1 strains
File 2791327513_metadata.gb contains 1 strains
File 2790335951_metadata.gb contains 1 strains
File 2791328715_metadata.gb contains 1 strains
File 2790340519_metadata.gb contains 1 strains
File 2791328489_metadata.gb contains 1 strains
File 2790345631_metadata.gb contains 1 strains
File 2791334217_metadata.gb contains 1 strains
File 2790350405_metadata.gb contains 1 strains
File 2791334406_metadata.gb contains 1 strains
File 2790355281_metadata.gb contains 1 strains
File 2791334263_metadata.gb contains 1 strains
File 2790361485_metadata.gb contains 1 strains
File 2791327413_metadata.gb contains 1 strains
File 332986951_metadata.gb contains 1 strains
File 523806722_metadata.gb contains 1 strains
File 523815970_metadata.gb contains 1 strains
File 523821078_metadata.gb contains 1 strains
File 529190224_metadata.gb contains 1 strains
File 548713695_metadata.gb contains 1 strains
File 559187652_metadata.gb contains 1 strains
File 563346448_metadata.gb contains 1 strains
File 564743501_metadata.gb contains 1 strains
File 601101465_metadata.gb contains 1 strains
File 602804670_metadata.gb contains 1 strains
File 602808998_metadata.gb contains 1 strains
File 602813325_metadata.gb contains 1 strains
File 602817697_metadata.gb contains 1 strains
File 602822031_metadata.gb contains 1 strains
File 602826365_metadata.gb contains 1 strains
File 604190314_metadata.gb contains 1 strains
File 604257961_metadata.gb contains 1 strains
File 604268692_metadata.gb contains 1 strains
File 604278585_metadata.gb contains 1 strains
File 749297197_metadata.gb contains 1 strains
File 604290478_metadata.gb contains 1 strains
File 749296449_metadata.gb contains 1 strains
File 604307148_metadata.gb contains 1 strains
File 604322487_metadata.gb contains 1 strains
File 604336542_metadata.gb contains 1 strains
File 749298461_metadata.gb contains 1 strains
File 604350507_metadata.gb contains 1 strains
File 604362088_metadata.gb contains 1 strains
File 749298420_metadata.gb contains 1 strains
File 605498156_metadata.gb contains 1 strains
File 605511148_metadata.gb contains 1 strains
File 605537200_metadata.gb contains 1 strains
File 605546632_metadata.gb contains 1 strains
File 605555317_metadata.gb contains 1 strains
File 630832964_metadata.gb contains 1 strains
File 674188659_metadata.gb contains 1 strains
File 749314944_metadata.gb contains 1 strains
File 674230814_metadata.gb contains 1 strains
File 682036555_metadata.gb contains 1 strains
File 682049973_metadata.gb contains 1 strains
File 682058852_metadata.gb contains 1 strains
File 682063065_metadata.gb contains 1 strains
File 682067343_metadata.gb contains 1 strains
File 682080448_metadata.gb contains 1 strains
File 682084826_metadata.gb contains 1 strains
File 682089197_metadata.gb contains 1 strains
File 682093565_metadata.gb contains 1 strains
File 686507741_metadata.gb contains 1 strains
File 747132831_metadata.gb contains 1 strains
File 747137039_metadata.gb contains 1 strains
File 754295078_metadata.gb contains 1 strains
File 754295112_metadata.gb contains 1 strains
File 820758584_metadata.gb contains 1 strains
File 896682452_metadata.gb contains 1 strains
File 915846032_metadata.gb contains 1 strains
File 896682995_metadata.gb contains 1 strains
File 901905107_metadata.gb contains 1 strains
File 901911536_metadata.gb contains 1 strains
File 924635423_metadata.gb contains 1 strains
File 924639146_metadata.gb contains 1 strains
File 924643612_metadata.gb contains 1 strains
File 930813018_metadata.gb contains 1 strains
File 930817479_metadata.gb contains 1 strains
File 933889448_metadata.gb contains 1 strains
File 933894153_metadata.gb contains 1 strains
File 949867589_metadata.gb contains 1 strains
File 983377086_metadata.gb contains 1 strains
File 949872475_metadata.gb contains 1 strains
File 949877344_metadata.gb contains 1 strains
File 953768973_metadata.gb contains 1 strains
File 953773649_metadata.gb contains 1 strains
File 971178605_metadata.gb contains 1 strains
File 995901180_metadata.gb contains 1 strains
File 998623498_metadata.gb contains 1 strains
File 995915826_metadata.gb contains 1 strains
File 995918924_metadata.gb contains 1 strains
File 995920397_metadata.gb contains 1 strains
File 995923765_metadata.gb contains 1 strains
File 998623489_metadata.gb contains 1 strains

Total number of strains found in multiple files: 144
```

Data collection summary :

complete salmonella sequences derived from chicken isolates has been searched in NCBI (National Center for Biotechnology Information) Pathogen detection database. It has been founded that 496 sequences are present. A through examination of presence of duplicates has been done based of sequence IDs and strains.

Finally we have obtained 299 unique sequences with their information stored in an excel file and that has been considered for further analysis.

Step 02 - AMR gene identification using CARD- Resistance Gene Identifier

In this step AMR genes along with the drug classes for the accquired data has been detected using Resistance Gene Identifier tool provided by CARD database

Each and every sequences of 299 salmonella sequences has been fed into the RGI tool and AMR gene information has been collected in an excel file sepeartely.

As this Bioinformatics tools works in Linux enviroment, this has been seperately done in Linux OS by the means of iterating over the folder of salmonella seqiences into RGI.

Finally the results has been stored in a separate file.

Step -3 Data Wrangling

In this stage Data wrangling carried out with the isolate information and the RGI results

In [1]:
```python
import pandas as pd
df1 = pd.read_csv("C:/VIT/Semester 3/SET-03/isolates.csv")
df1
```

Out[1]:

	#Organism group	Strain	Isolate identifiers	Serovar	Isolate	Create date	Location	Isolation source	Isolation type	Food origin	SNP cluster	Min-same	Min-diff	BioSample	A
0	Salmonella enterica	WAPHL_SAL-A00031	"SRS469289","WAPHL_SAL-A00031"	Montevideo	PDT000000399.3	2015-02-11T06:08:33Z	USA:WA	chicken	environmental/other	NaN	PDS000172527.57	3.0	3.0	SAMN02182894	GCA_011
1	Salmonella enterica	WAPHL_SAL-A00052	"SRS479789","WAPHL_SAL-A00052"	Ohio	PDT000000661.3	2015-02-11T06:08:35Z	USA:WA	chicken	environmental/other	NaN	PDS000027300.3	0.0	NaN	SAMN02182915	GCA_010
2	Salmonella enterica	WAPHL_SAL-A00149	"SRS515215","WAPHL_SAL-A00149"	NaN	PDT000001787.3	2015-02-11T17:10:01Z	USA:WA	chicken	environmental/other	NaN	PDS000179566.108	7.0	13.0	SAMN02182983	GCA_006
3	Salmonella enterica	MDH-2013-00166	"MDH-2013-00166","SRS523520"	Kentucky	PDT000002088.3	2015-02-11T17:10:02Z	USA:MN	chicken feces	environmental/other	NaN	PDS000117429.168	0.0	27.0	SAMN02378171	GCA_010
4	Salmonella enterica	MDH-2013-00167	"MDH-2013-00167","SRS523566"	Thompson	PDT000002114.3	2015-02-11T17:10:02Z	USA:MN	chicken feces	environmental/other	NaN	PDS000032705.1262	3.0	1.0	SAMN02378172	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	
294	Salmonella enterica	YAH-CH5	"SRS21539871","YAH-CH5"	NaN	PDT002189990.1	2024-06-06T22:10:06Z	USA: Kentucky	chicken	environmental/other	NaN	PDS000050977.104	0.0	5.0	SAMN41714676	
295	Salmonella enterica	YAH-CH4	"SRS21539872","YAH-CH4"	Thompson	PDT002189991.1	2024-06-06T22:10:07Z	USA: Kentucky	chicken	environmental/other	NaN	PDS000032705.1262	0.0	3.0	SAMN41714675	
296	Salmonella enterica	YAH-CH3	"SRS21539869","YAH-CH3"	Agona	PDT002189992.1	2024-06-06T22:10:07Z	USA: Kentucky	chicken	environmental/other	NaN	PDS000050977.104	1.0	6.0	SAMN41714674	
297	Salmonella enterica	YAH-CH2	"SRS21539870","YAH-CH2"	Liverpool	PDT002189993.1	2024-06-06T22:10:07Z	USA: Kentucky	chicken	environmental/other	NaN	PDS000001378.421	30.0	8.0	SAMN41714673	
298	Salmonella enterica	YAH-CH1	"SRS21539868","YAH-CH1"	Agona	PDT002189994.1	2024-06-06T22:10:08Z	USA: Kentucky	chicken	environmental/other	NaN	PDS000050977.104	0.0	5.0	SAMN41714672	

299 rows × 17 columns

In [3]:
```python
df2 = pd.read_excel("C:/VIT/Semester 3/SET-03/Genes by antimicrobial agents and classes.xlsx")
df2
```

Out[3]:

	Class	Antimicrobial Agent	Gene
0	Aminoglycosides	Gentamicin	aac(3)-Ia
1	Aminoglycosides	Gentamicin	aac(3)-IIa
2	Aminoglycosides	Gentamicin	aac(3)-IIIa
3	Aminoglycosides	Gentamicin	aac(3)-IV
4	Aminoglycosides	Gentamicin	aac(3)-IVa
...	...	...	...
219	Tetracyclines	Tetracycline	tet(D)
220	Tetracyclines	Tetracycline	tet(G)
221	Tetracyclines	Tetracycline	tet(M)
222	Tetracyclines	Tetracycline	tet(O)
223	Tetracyclines	Tetracycline	tet(X5)

224 rows × 3 columns

In [5]:
```python
# Creating df3 with the 'Isolate' column from df1
df3 = df1[['Isolate']]
df3
```

Out[5]:

	Isolate
0	PDT000000399.3
1	PDT000000661.3
2	PDT000001787.3
3	PDT000002088.3
4	PDT000002114.3
...	...
294	PDT002189990.1
295	PDT002189991.1
296	PDT002189992.1
297	PDT002189993.1
298	PDT002189994.1

299 rows × 1 columns

In [7]:
```python
genes = df2['Gene'].tolist()

# Create a DataFrame with the new columns
new_columns = pd.DataFrame({gene: None for gene in genes}, index=df3.index)

# Concatenate the new columns to the original DataFrame
df3 = pd.concat([df3, new_columns], axis=1)

df3
```

Out[7]:

	Isolate	aac(3)-Ia	aac(3)-IIa	aac(3)-IIIa	aac(3)-IV	aac(3)-IVa	aac(3)-VIa	aac(6')-Ib	aac(6')-Ib4	aac(6')-IIa	...	tet	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)
0	PDT000000399.3	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
1	PDT000000661.3	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
2	PDT000001787.3	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
3	PDT000002088.3	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
4	PDT000002114.3	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
295	PDT002189991.1	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
296	PDT002189992.1	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
297	PDT002189993.1	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None
298	PDT002189994.1	None	None	None	None	None	None	None	None	None	...	None	None	None	None	None	None	None	None	None	None

299 rows × 187 columns

In [9]:

```python
# Replace None with 0 or 1 based on the condition
for i, row in df3.iterrows():
    isolate = row['Isolate']
    matching_row = df1[df1['Isolate'] == isolate]
    if not matching_row.empty:
        amr_genotypes = matching_row.iloc[0]['AMR genotypes']
        for gene in genes:
            if gene in amr_genotypes:
                df3.at[i, gene] = 1
            else:
                df3.at[i, gene] = 0
```

In [10]:

```python
# Check if any value in df3 (excluding the first column) is 1
has_ones = (df3.iloc[:, 1:] == 1).any().any()

if has_ones:
    print("df3 has at least one '1' in columns other than the first column.")
else:
    print("df3 does not have any '1's in columns other than the first column.")
```

```
df3 has at least one '1' in columns other than the first column.
```

In [13]:

```python
df3
```

Out[13]:

	Isolate	aac(3)-Ia	aac(3)-IIa	aac(3)-IIIa	aac(3)-IV	aac(3)-IVa	aac(3)-VIa	aac(6')-Ib	aac(6')-Ib4	aac(6')-IIa	...	tet	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)
0	PDT000000399.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	PDT000000661.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	PDT000001787.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	PDT000002088.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	PDT000002114.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
295	PDT002189991.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
296	PDT002189992.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
297	PDT002189993.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
298	PDT002189994.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

299 rows × 187 columns

In [18]:

```python
df3.to_csv("df3.csv",index=0)
```

In [20]:

```python
# Verification step
verification_results = []
for i, row in df1.iterrows():
    isolate = row['Isolate']
    amr_genotypes = row['AMR genotypes']
    matching_row = df3[df3['Isolate'] == isolate]
    if not matching_row.empty:
        all_marked = all(gene not in genes or matching_row.iloc[0][gene] == 1 for gene in amr_genotypes)
        verification_results.append((isolate, all_marked))

# Print verification results
for isolate, result in verification_results:
    print(f"Isolate {isolate}: {'All genes marked as 1' if result else 'Some genes not marked as 1'}")
```

```
            Isolate PDT000000399.3: All genes marked as 1
            Isolate PDT000000661.3: All genes marked as 1
            Isolate PDT000001787.3: All genes marked as 1
            Isolate PDT000002088.3: All genes marked as 1
            Isolate PDT000002114.3: All genes marked as 1
            Isolate PDT000002344.3: All genes marked as 1
            Isolate PDT000002526.3: All genes marked as 1
            Isolate PDT000002527.3: All genes marked as 1
            Isolate PDT000002528.3: All genes marked as 1
            Isolate PDT000002529.3: All genes marked as 1
            Isolate PDT000002530.3: All genes marked as 1
            Isolate PDT000002532.3: All genes marked as 1
            Isolate PDT000002534.4: All genes marked as 1
            Isolate PDT000002536.3: All genes marked as 1
            Isolate PDT000002538.3: All genes marked as 1
            Isolate PDT000002539.3: All genes marked as 1
            Isolate PDT000002541.3: All genes marked as 1
            Isolate PDT000002543.3: All genes marked as 1
            Isolate PDT000002544.3: All genes marked as 1
            Isolate PDT000002547.3: All genes marked as 1
            Isolate PDT000002585.3: All genes marked as 1
            Isolate PDT000002586.3: All genes marked as 1
            Isolate PDT000002589.3: All genes marked as 1
            Isolate PDT000002593.3: All genes marked as 1
            Isolate PDT000002594.3: All genes marked as 1
            Isolate PDT000002595.3: All genes marked as 1
            Isolate PDT000002596.3: All genes marked as 1
            Isolate PDT000002598.3: All genes marked as 1
            Isolate PDT000002599.3: All genes marked as 1
            Isolate PDT000002600.3: All genes marked as 1
            Isolate PDT000002601.3: All genes marked as 1
            Isolate PDT000002602.3: All genes marked as 1
            Isolate PDT000002603.3: All genes marked as 1
            Isolate PDT000002605.3: All genes marked as 1
            Isolate PDT000002607.3: All genes marked as 1
            Isolate PDT000002610.3: All genes marked as 1
            Isolate PDT000002612.3: All genes marked as 1
            Isolate PDT000002650.3: All genes marked as 1
            Isolate PDT000002652.3: All genes marked as 1
            Isolate PDT000002653.3: All genes marked as 1
            Isolate PDT000003035.3: All genes marked as 1
            Isolate PDT000003094.3: All genes marked as 1
            Isolate PDT000003095.3: All genes marked as 1
            Isolate PDT000003098.3: All genes marked as 1
            Isolate PDT000003100.3: All genes marked as 1
            Isolate PDT000003102.4: All genes marked as 1
            Isolate PDT000003103.3: All genes marked as 1
            Isolate PDT000003104.3: All genes marked as 1
            Isolate PDT000003105.3: All genes marked as 1
            Isolate PDT000003108.3: All genes marked as 1
            Isolate PDT000003215.5: All genes marked as 1
            Isolate PDT000003216.5: All genes marked as 1
            Isolate PDT000003595.3: All genes marked as 1
            Isolate PDT000004003.3: All genes marked as 1
            Isolate PDT000004007.3: All genes marked as 1
            Isolate PDT000025135.3: All genes marked as 1
            Isolate PDT000025137.3: All genes marked as 1
            Isolate PDT000026047.3: All genes marked as 1
            Isolate PDT000027717.3: All genes marked as 1
            Isolate PDT000032249.3: All genes marked as 1
            Isolate PDT000041590.2: All genes marked as 1
            Isolate PDT000052058.2: All genes marked as 1
            Isolate PDT000052060.2: All genes marked as 1
            Isolate PDT000052063.2: All genes marked as 1
            Isolate PDT000052064.2: All genes marked as 1
            Isolate PDT000064725.2: All genes marked as 1
            Isolate PDT000064728.2: All genes marked as 1
            Isolate PDT000105393.2: All genes marked as 1
            Isolate PDT000105409.2: All genes marked as 1
            Isolate PDT000105415.2: All genes marked as 1
            Isolate PDT000105582.2: All genes marked as 1
            Isolate PDT000105585.2: All genes marked as 1
            Isolate PDT000105591.2: All genes marked as 1
            Isolate PDT000126374.2: All genes marked as 1
            Isolate PDT000128894.2: All genes marked as 1
            Isolate PDT000128899.2: All genes marked as 1
            Isolate PDT000128906.2: All genes marked as 1
            Isolate PDT000128908.2: All genes marked as 1
            Isolate PDT000133701.2: All genes marked as 1
            Isolate PDT000133702.2: All genes marked as 1
            Isolate PDT000133703.2: All genes marked as 1
            Isolate PDT000133704.2: All genes marked as 1
            Isolate PDT000133705.2: All genes marked as 1
            Isolate PDT000133706.2: All genes marked as 1
            Isolate PDT000133707.2: All genes marked as 1
            Isolate PDT000133708.2: All genes marked as 1
            Isolate PDT000133709.2: All genes marked as 1
            Isolate PDT000133710.2: All genes marked as 1
            Isolate PDT000133711.2: All genes marked as 1
            Isolate PDT000133712.2: All genes marked as 1
            Isolate PDT000133713.2: All genes marked as 1
            Isolate PDT000133714.2: All genes marked as 1
            Isolate PDT000133885.2: All genes marked as 1
            Isolate PDT000148789.2: All genes marked as 1
            Isolate PDT000148808.2: All genes marked as 1
            Isolate PDT000148812.2: All genes marked as 1
            Isolate PDT000148813.2: All genes marked as 1
            Isolate PDT000176914.3: All genes marked as 1
            Isolate PDT000185559.4: All genes marked as 1
            Isolate PDT000185564.5: All genes marked as 1
            Isolate PDT000185565.5: All genes marked as 1
            Isolate PDT000185566.5: All genes marked as 1
            Isolate PDT000185567.5: All genes marked as 1
            Isolate PDT000185568.5: All genes marked as 1
            Isolate PDT000185570.5: All genes marked as 1
            Isolate PDT000189480.2: All genes marked as 1
            Isolate PDT000198211.2: All genes marked as 1
            Isolate PDT000198228.2: All genes marked as 1
            Isolate PDT000198230.2: All genes marked as 1
            Isolate PDT000212925.2: All genes marked as 1
            Isolate PDT000234862.2: All genes marked as 1
            Isolate PDT000234866.2: All genes marked as 1
            Isolate PDT000234880.2: All genes marked as 1
            Isolate PDT000236355.2: All genes marked as 1
            Isolate PDT000236358.2: All genes marked as 1
            Isolate PDT000236359.2: All genes marked as 1
            Isolate PDT000236362.2: All genes marked as 1
            Isolate PDT000236363.2: All genes marked as 1
            Isolate PDT000236364.2: All genes marked as 1
            Isolate PDT000236378.2: All genes marked as 1
            Isolate PDT000282422.2: All genes marked as 1
            Isolate PDT000282431.2: All genes marked as 1
```

```
Isolate PDT000282445.2: All genes marked as 1
Isolate PDT000282519.2: All genes marked as 1
Isolate PDT000304808.2: All genes marked as 1
Isolate PDT000304890.2: All genes marked as 1
Isolate PDT000304892.2: All genes marked as 1
Isolate PDT000304910.2: All genes marked as 1
Isolate PDT000309179.2: All genes marked as 1
Isolate PDT000309414.2: All genes marked as 1
Isolate PDT000309420.2: All genes marked as 1
Isolate PDT000309428.2: All genes marked as 1
Isolate PDT000309434.2: All genes marked as 1
Isolate PDT000309436.2: All genes marked as 1
Isolate PDT000309437.2: All genes marked as 1
Isolate PDT000309438.2: All genes marked as 1
Isolate PDT000309444.2: All genes marked as 1
Isolate PDT000309453.2: All genes marked as 1
Isolate PDT000312230.2: All genes marked as 1
Isolate PDT000312232.2: All genes marked as 1
Isolate PDT000312252.2: All genes marked as 1
Isolate PDT000312255.2: All genes marked as 1
Isolate PDT000312282.2: All genes marked as 1
Isolate PDT000312283.2: All genes marked as 1
Isolate PDT000312287.2: All genes marked as 1
Isolate PDT000316773.1: All genes marked as 1
Isolate PDT000316775.1: All genes marked as 1
Isolate PDT000316779.1: All genes marked as 1
Isolate PDT000316780.1: All genes marked as 1
Isolate PDT000316785.1: All genes marked as 1
Isolate PDT000316789.1: All genes marked as 1
Isolate PDT000316791.1: All genes marked as 1
Isolate PDT000316794.1: All genes marked as 1
Isolate PDT000316796.1: All genes marked as 1
Isolate PDT000316806.1: All genes marked as 1
Isolate PDT000316812.1: All genes marked as 1
Isolate PDT000316830.1: All genes marked as 1
Isolate PDT000316834.1: All genes marked as 1
Isolate PDT000317503.2: All genes marked as 1
Isolate PDT000320804.1: All genes marked as 1
Isolate PDT000412166.1: All genes marked as 1
Isolate PDT000412169.1: All genes marked as 1
Isolate PDT000495855.1: All genes marked as 1
Isolate PDT000495862.1: All genes marked as 1
Isolate PDT000867972.1: All genes marked as 1
Isolate PDT000867989.1: All genes marked as 1
Isolate PDT000867990.1: All genes marked as 1
Isolate PDT000868005.1: All genes marked as 1
Isolate PDT000868403.1: All genes marked as 1
Isolate PDT000868594.1: All genes marked as 1
Isolate PDT000868690.1: All genes marked as 1
Isolate PDT000877385.1: All genes marked as 1
Isolate PDT000877900.1: All genes marked as 1
Isolate PDT000877925.1: All genes marked as 1
Isolate PDT000891560.1: All genes marked as 1
Isolate PDT000891569.1: All genes marked as 1
Isolate PDT000891586.1: All genes marked as 1
Isolate PDT000891590.1: All genes marked as 1
Isolate PDT000891592.1: All genes marked as 1
Isolate PDT000891593.1: All genes marked as 1
Isolate PDT000919367.1: All genes marked as 1
Isolate PDT000919619.1: All genes marked as 1
Isolate PDT000925550.1: All genes marked as 1
Isolate PDT000925562.1: All genes marked as 1
Isolate PDT000925578.1: All genes marked as 1
Isolate PDT000925579.1: All genes marked as 1
Isolate PDT000925580.1: All genes marked as 1
Isolate PDT000966026.1: All genes marked as 1
Isolate PDT000966029.1: All genes marked as 1
Isolate PDT000973556.1: All genes marked as 1
Isolate PDT000973557.1: All genes marked as 1
Isolate PDT000975197.1: All genes marked as 1
Isolate PDT000975199.1: All genes marked as 1
Isolate PDT000983283.1: All genes marked as 1
Isolate PDT000983286.1: All genes marked as 1
Isolate PDT000983287.1: All genes marked as 1
Isolate PDT001016281.1: All genes marked as 1
Isolate PDT001018665.1: All genes marked as 1
Isolate PDT001018668.1: All genes marked as 1
Isolate PDT001018670.1: All genes marked as 1
Isolate PDT001018672.1: All genes marked as 1
Isolate PDT001018674.1: All genes marked as 1
Isolate PDT001018678.1: All genes marked as 1
Isolate PDT001018681.1: All genes marked as 1
Isolate PDT001018683.1: All genes marked as 1
Isolate PDT001018685.1: All genes marked as 1
Isolate PDT001018686.1: All genes marked as 1
Isolate PDT001018687.1: All genes marked as 1
Isolate PDT001018697.1: All genes marked as 1
Isolate PDT001018700.1: All genes marked as 1
Isolate PDT001018708.1: All genes marked as 1
Isolate PDT001023646.1: All genes marked as 1
Isolate PDT001023755.1: All genes marked as 1
Isolate PDT001023756.1: All genes marked as 1
Isolate PDT001023757.1: All genes marked as 1
Isolate PDT001023758.1: All genes marked as 1
Isolate PDT001023759.1: All genes marked as 1
Isolate PDT001088103.1: All genes marked as 1
Isolate PDT001088104.1: All genes marked as 1
Isolate PDT001088105.1: All genes marked as 1
Isolate PDT001088106.1: All genes marked as 1
Isolate PDT001088107.1: All genes marked as 1
Isolate PDT001088108.1: All genes marked as 1
Isolate PDT001088109.1: All genes marked as 1
Isolate PDT001088110.1: All genes marked as 1
Isolate PDT001088111.1: All genes marked as 1
Isolate PDT001088112.1: All genes marked as 1
Isolate PDT001088113.1: All genes marked as 1
Isolate PDT001088114.1: All genes marked as 1
Isolate PDT001088115.1: All genes marked as 1
Isolate PDT001088116.1: All genes marked as 1
Isolate PDT001088117.1: All genes marked as 1
Isolate PDT001088118.1: All genes marked as 1
Isolate PDT001088119.1: All genes marked as 1
Isolate PDT001088725.1: All genes marked as 1
Isolate PDT001197060.1: All genes marked as 1
Isolate PDT001197062.1: All genes marked as 1
Isolate PDT001197063.1: All genes marked as 1
Isolate PDT001197064.1: All genes marked as 1
Isolate PDT001197066.1: All genes marked as 1
Isolate PDT001197067.1: All genes marked as 1
Isolate PDT001197069.1: All genes marked as 1
Isolate PDT001197070.1: All genes marked as 1
Isolate PDT001197071.1: All genes marked as 1
Isolate PDT001197074.1: All genes marked as 1
Isolate PDT001197075.1: All genes marked as 1
Isolate PDT001197076.1: All genes marked as 1
Isolate PDT001197077.1: All genes marked as 1
Isolate PDT001197079.1: All genes marked as 1
Isolate PDT001197080.1: All genes marked as 1
Isolate PDT001197086.1: All genes marked as 1
Isolate PDT001197087.1: All genes marked as 1
Isolate PDT001197089.1: All genes marked as 1
Isolate PDT001197090.1: All genes marked as 1
Isolate PDT001197091.1: All genes marked as 1
Isolate PDT001197094.1: All genes marked as 1
Isolate PDT001197096.1: All genes marked as 1
Isolate PDT001197097.1: All genes marked as 1
Isolate PDT001197098.1: All genes marked as 1
Isolate PDT001197099.1: All genes marked as 1
Isolate PDT001197101.1: All genes marked as 1
Isolate PDT001242435.1: All genes marked as 1
Isolate PDT001608224.1: All genes marked as 1
Isolate PDT001608225.1: All genes marked as 1
Isolate PDT001608226.1: All genes marked as 1
Isolate PDT001608227.1: All genes marked as 1
Isolate PDT001608230.1: All genes marked as 1
Isolate PDT001608231.1: All genes marked as 1
Isolate PDT001608232.1: All genes marked as 1
Isolate PDT001608233.1: All genes marked as 1
Isolate PDT001608234.1: All genes marked as 1
Isolate PDT001608235.1: All genes marked as 1
Isolate PDT001608236.1: All genes marked as 1
Isolate PDT001608237.1: All genes marked as 1
Isolate PDT001608238.1: All genes marked as 1
Isolate PDT001608242.1: All genes marked as 1
Isolate PDT001608243.1: All genes marked as 1
Isolate PDT001616280.1: All genes marked as 1
Isolate PDT001616281.1: All genes marked as 1
Isolate PDT001616283.1: All genes marked as 1
Isolate PDT001616284.1: All genes marked as 1
Isolate PDT001616285.1: All genes marked as 1
Isolate PDT001616286.1: All genes marked as 1
Isolate PDT001616352.1: All genes marked as 1
Isolate PDT001616353.1: All genes marked as 1
Isolate PDT001616361.1: All genes marked as 1
Isolate PDT001616362.1: All genes marked as 1
Isolate PDT001616363.1: All genes marked as 1
Isolate PDT001616364.1: All genes marked as 1
Isolate PDT002189985.1: All genes marked as 1
Isolate PDT002189986.1: All genes marked as 1
Isolate PDT002189987.1: All genes marked as 1
Isolate PDT002189988.1: All genes marked as 1
Isolate PDT002189989.1: All genes marked as 1
Isolate PDT002189990.1: All genes marked as 1
Isolate PDT002189991.1: All genes marked as 1
Isolate PDT002189992.1: All genes marked as 1
Isolate PDT002189993.1: All genes marked as 1
Isolate PDT002189994.1: All genes marked as 1
```

In [22]:
```
df3
```

Out[22]:

	Isolate	aac(3)-Ia	aac(3)-IIa	aac(3)-IIIa	aac(3)-IV	aac(3)-IVa	aac(3)-VIa	aac(6')-Ib	aac(6')-Ib4	aac(6')-IIa	...	tet	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)
0	PDT000000399.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
1	PDT000000661.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
2	PDT000001787.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
3	PDT000002088.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
4	PDT000002114.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
295	PDT002189991.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
296	PDT002189992.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
297	PDT002189993.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0
298	PDT002189994.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	0

299 rows × 187 columns

In [24]:
```
df3.to_csv("df3.csv",index=0)
```

In [26]:
```
df3['AMR Results'] = df3.iloc[:, 2:].apply(lambda row: 'Positive' if 1 in row.values else 'Negative', axis=1)
```

In [28]:
```
df3
```

Out[28]:

	Isolate	aac(3)-Ia	aac(3)-IIa	aac(3)-IIIa	aac(3)-IV	aac(3)-IVa	aac(3)-VIa	aac(6')-Ib	aac(6')-Ib4	aac(6')-IIa	...	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	AMR Results
0	PDT000000399.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
1	PDT000000661.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
2	PDT000001787.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
3	PDT000002088.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
4	PDT000002114.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
295	PDT002189991.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
296	PDT002189992.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
297	PDT002189993.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative
298	PDT002189994.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Negative

299 rows × 188 columns

In [30]:
```
df3.to_csv("df3.csv",index=0)
```

In [32]:
```
# Add the drug classes column
df3['drug classes'] = ''

for i, row in df3.iterrows():
    if row['AMR Results'] == 'Positive':
        drug_classes = []
        for gene in genes:
            if row[gene] == 1:
                matching_agent = df2[df2['Gene'] == gene]['Antimicrobial Agent'].values
                if len(matching_agent) > 0:
                    drug_classes.append(matching_agent[0])
        df3.at[i, 'drug classes'] = ', '.join(drug_classes)

df3
```

Out[32]:

	Isolate	aac(3)-Ia	aac(3)-IIa	aac(3)-IIIa	aac(3)-IV	aac(3)-IVa	aac(3)-VIa	aac(6')-Ib	aac(6')-Ib4	aac(6')-IIa	...	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	AMR Results	drug classes
0	PDT000000399.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
1	PDT000000661.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
2	PDT000001787.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
3	PDT000002088.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
4	PDT000002114.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
295	PDT002189991.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
296	PDT002189992.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
297	PDT002189993.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	
298	PDT002189994.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	Negative	

299 rows × 189 columns

In [34]:
```
df3.to_csv("df3.csv",index=0)
```

In [1]:
```
import pandas as pd
df=pd.read_csv("finaldata.csv")
df
```

Out[1]:

	Isolate	aadA1	aadA2	aadA3	aadA4	aadA5	aadA6	aadA7	aadA8	aadA12	...	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	drug classes
0	PDT000000399.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
1	PDT000000661.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
2	PDT000001787.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
3	PDT000002088.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
4	PDT000002114.3	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
294	PDT002189990.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
295	PDT002189991.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
296	PDT002189992.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
297	PDT002189993.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA
298	PDT002189994.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	NA

299 rows × 165 columns

In [3]:
```
# Strip any extra whitespace and remove rows where 'drug classes' column contains "NA"
df['drug classes'] = df['drug classes'].str.strip()  # Remove any surrounding whitespace
df4 = df[df['drug classes'] != "NA"]
df4
```

Out[3]:

	Isolate	aadA1	aadA2	aadA3	aadA4	aadA5	aadA6	aadA7	aadA8	aadA12	...	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	drug classes
5	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin, Streptomycin, Amoxicillin-Clavul...
10	PDT000002530.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin, Tetracycline, Tetracycline
11	PDT000002532.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin, Tetracycline, Tetracycline
14	PDT000002538.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin, Amoxicillin-Clavulanic Acid, Amo...
19	PDT000002547.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Tetracycline, Tetracycline
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
285	PDT001616361.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Streptomycin, Sulfamethoxazole-Sulfisoxazole, ...
286	PDT001616362.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Streptomycin, Sulfamethoxazole-Sulfisoxazole, ...
287	PDT001616363.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Streptomycin, Sulfamethoxazole-Sulfisoxazole, ...
288	PDT001616364.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Streptomycin, Sulfamethoxazole-Sulfisoxazole, ...
291	PDT002189987.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Streptomycin

188 rows × 165 columns

In [5]:
```
# First, split the 'drug classes' column by commas, creating lists in each cell
df4['drug classes'] = df4['drug classes'].str.split(',')

# Expand the rows based on the split values in 'drug classes'
df5 = df4.explode('drug classes').reset_index(drop=True)

# Optional: Strip any whitespace around drug names after exploding
df5['drug classes'] = df5['drug classes'].str.strip()

# Display or save the resulting dataframe
df5.head()  # This will display the first few rows
```

```
C:\Users\DD\AppData\Local\Temp\ipykernel_13416\3902397106.py:2: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.html#returning-a-view-versus-a-copy
  df4['drug classes'] = df4['drug classes'].str.split(',')
```

Out[5]:

	Isolate	aadA1	aadA2	aadA3	aadA4	aadA5	aadA6	aadA7	aadA8	aadA12	...	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	drug classes
0	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin
1	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin
2	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid
3	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid
4	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid

5 rows × 165 columns

In [7]:
```python
df5.shape
```

Out[7]:
```
(1306, 165)
```

In [9]:
```python
# Save df21 as a CSV file
df5.to_csv('finaldata_processed.csv', index=False)
```

In [13]:
```python
# Get unique values in the 'drug classes' column
unique_drug_classes = df5['drug classes'].unique()

# Get the number of unique values
num_unique_drug_classes = df5['drug classes'].nunique()

# Print the results
print(f"Number of unique drug classes: {num_unique_drug_classes}")
print("Unique drug classes:")
for drug_class in unique_drug_classes:
    print(drug_class)
```

```
Number of unique drug classes: 11
Unique drug classes:
Streptomycin
Amoxicillin-Clavulanic Acid
Tetracycline
Ampicillin
Sulfamethoxazole-Sulfisoxazole
Chloramphenicol
Ciprofloxacin
Nalidixic Acid
Trimethoprim-Sulfamethoxazole
Gentamicin
Ceftriaxone
```

In [35]:
```python
import pandas as pd

# Replace 'your_csv_file.csv' with the path to your CSV file
df51 = pd.read_csv('finaldata_processed.csv')
df51
```

Out[35]:

	Isolate	aadA1	aadA2	aadA3	aadA4	aadA5	aadA6	aadA7	aadA8	aadA12	...	tet(31)	tet(A)	tet(B)	tet(C)	tet(D)	tet(G)	tet(M)	tet(O)	tet(X5)	drug classes
0	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin
1	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Streptomycin
2	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid
3	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid
4	PDT000002344.3	0	0	0	0	0	0	0	0	0	...	0	0	1	0	0	0	0	0	0	Amoxicillin-Clavulanic Acid
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
1301	PDT001616364.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Sulfamethoxazole-Sulfisoxazole
1302	PDT001616364.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Nalidixic Acid
1303	PDT001616364.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Tetracycline
1304	PDT001616364.1	1	0	0	0	0	0	0	0	0	...	0	1	0	0	0	0	0	0	0	Tetracycline
1305	PDT002189987.1	0	0	0	0	0	0	0	0	0	...	0	0	0	0	0	0	0	0	0	Streptomycin

1306 rows × 165 columns

In [17]:
```python
# Check the data type of the 'drug classes' column
print("Data type of 'drug classes':", df51['drug classes'].dtype)

# Display the first few rows to inspect the formatting
print("\nFirst few rows of the 'drug classes' column:")
print(df51['drug classes'].head(10))
```

```
Data type of 'drug classes': object

First few rows of the 'drug classes' column:
0                   Streptomycin
1                   Streptomycin
2    Amoxicillin-Clavulanic Acid
3    Amoxicillin-Clavulanic Acid
4    Amoxicillin-Clavulanic Acid
5    Amoxicillin-Clavulanic Acid
6    Amoxicillin-Clavulanic Acid
7                   Tetracycline
8                   Tetracycline
9                   Streptomycin
Name: drug classes, dtype: object
```

In [19]:
```python
# Convert 'drug classes' to string to avoid NaN conversion
df51['drug classes'] = df51['drug classes'].astype(str)

# Check if conversion fixed the issue
print("\nAfter conversion, first few rows of 'drug classes':")
print(df51['drug classes'].head(10))
```

```
After conversion, first few rows of 'drug classes':
0                    Streptomycin
1                    Streptomycin
2    Amoxicillin-Clavulanic Acid
3    Amoxicillin-Clavulanic Acid
4    Amoxicillin-Clavulanic Acid
5    Amoxicillin-Clavulanic Acid
6    Amoxicillin-Clavulanic Acid
7                    Tetracycline
8                    Tetracycline
9                    Streptomycin
Name: drug classes, dtype: object
```

In [23]:
```python
# Strip whitespace from the 'drug classes' column
df51['drug classes'] = df51['drug classes'].str.strip()

# Check the unique values again
unique_drug_classes = df51['drug classes'].unique()
print("\nUnique values in 'drug classes' after stripping whitespace:")
print(unique_drug_classes)
```

```
Unique values in 'drug classes' after stripping whitespace:
['Streptomycin' 'Amoxicillin-Clavulanic Acid' 'Tetracycline' 'Ampicillin'
 'Sulfamethoxazole-Sulfisoxazole' 'Chloramphenicol' 'Ciprofloxacin'
 'Nalidixic Acid' 'Trimethoprim-Sulfamethoxazole' 'Gentamicin'
 'Ceftriaxone']
```

In [25]:
```python
# Verify if expected drug classes are present
expected_drug_classes = [
    'Streptomycin', 'Amoxicillin-Clavulanic Acid', 'Tetracycline', 'Ampicillin',
    'Sulfamethoxazole-Sulfisoxazole', 'Chloramphenicol', 'Ciprofloxacin',
    'Nalidixic Acid', 'Trimethoprim-Sulfamethoxazole', 'Gentamicin', 'Ceftriaxone'
]

# Display a message if any expected drug class is missing
missing_classes = [drug for drug in expected_drug_classes if drug not in unique_drug_classes]
if missing_classes:
    print("\nMissing drug classes:")
    print(missing_classes)
else:
    print("\nAll expected drug classes are present.")
```

```
All expected drug classes are present.
```

In [37]:
```python
# Initialize a dictionary to store unique column names for each drug class
drug_class_columns = {drug_class: [] for drug_class in df51['drug classes'].unique()}

# Iterate through each row in the DataFrame
for index, row in df51.iterrows():
    # Get the current drug class
    current_drug_class = row['drug classes']

    # Check for presence of 1 in the other columns
    for column in df51.columns[1:]:  # Skip the first column which is 'drug classes'
        if row[column] == 1 and column not in drug_class_columns[current_drug_class]:
            drug_class_columns[current_drug_class].append(column)

# Display the results
for drug_class, columns in drug_class_columns.items():
    print(f"{drug_class}: {columns}")
```

```
Streptomycin: ['aph(6)-Ic', 'aph(6)-Id', 'blaCMY', 'blaCMY-2', 'tet', 'tet(B)', 'tet(A)', 'aadA2', 'sul1', 'aadA1', 'sul3', 'blaCARB-2', 'floR', 'tet(G)', 'tet(M)', 'aadA7', 'blaTEM', 'blaT
EM-1', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'sul2', 'dfrA1', 'blaOXA-1', 'blaCTX-M', 'blaCTX-M-2', 'qnrB1', 'qnrB19', 'aadA12', 'aadA15', 'aadA22', 'dfrA14', 'cm', 'cmlA', 'cmlA5', 'qnrS
1', 'gyrA_D87Y', 'gyrA_S83Y', 'dfrA12', 'blaCTX-M-65']
Amoxicillin-Clavulanic Acid: ['aph(6)-Ic', 'aph(6)-Id', 'blaCMY', 'blaCMY-2', 'tet', 'tet(B)', 'floR', 'tet(A)', 'gyrA_S83F', 'sul2', 'aadA1', 'blaCTX-M', 'qnrB1', 'qnrB19']
Tetracycline: ['aph(6)-Ic', 'aph(6)-Id', 'blaCMY', 'blaCMY-2', 'tet', 'tet(B)', 'blaTEM', 'blaTEM-1', 'tet(A)', 'aadA2', 'sul1', 'aadA1', 'floR', 'sul3', 'blaCARB-2', 'tet(G)', 'tet(M)', 'a
adA7', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'sul2', 'dfrA8', 'blaCTX-M', 'dfrA1', 'blaCTX-M-2', 'qnrB1', 'qnrB19', 'aadA12', 'aadA15', 'aadA22', 'dfrA14', 'blaOXA-1', 'cm', 'cmlA', 'cmlA
5', 'qnrS1', 'gyrA_D87Y', 'gyrA_S83Y', 'dfrA12', 'blaCTX-M-65']
Ampicillin: ['blaTEM', 'blaTEM-1', 'tet', 'tet(A)', 'aadA2', 'sul1', 'sul3', 'blaCARB-2', 'floR', 'tet(G)', 'tet(M)', 'aadA7', 'aph(6)-Id', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'aadA1',
'sul2', 'dfrA1', 'blaCTX-M', 'qnrB1', 'qnrB19', 'blaOXA-1', 'dfrA8', 'blaCTX-M-2', 'tet(B)', 'aadA12', 'aadA15', 'aadA22', 'blaCMY', 'blaCMY-2', 'dfrA14', 'cm', 'cmlA', 'cmlA5', 'qnrS1', 'g
yrA_D87Y', 'gyrA_S83Y', 'blaCTX-M-65']
Sulfamethoxazole-Sulfisoxazole: ['aadA2', 'sul1', 'tet', 'tet(B)', 'aadA1', 'aph(6)-Ic', 'aph(6)-Id', 'sul3', 'blaCARB-2', 'floR', 'tet(A)', 'tet(G)', 'tet(M)', 'aadA7', 'blaTEM', 'blaTEM-
1', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'sul2', 'dfrA1', 'blaOXA-1', 'blaCTX-M', 'blaCTX-M-2', 'qnrB1', 'qnrB19', 'blaCMY', 'blaCMY-2', 'aadA12', 'aadA15', 'aadA22', 'dfrA14', 'gyrA_S83
Y', 'dfrA12', 'qnrS1', 'gyrA_D87Y', 'blaCTX-M-65']
Chloramphenicol: ['blaCMY', 'blaCMY-2', 'floR', 'tet', 'tet(A)', 'aadA2', 'sul1', 'sul3', 'blaCARB-2', 'tet(G)', 'tet(M)', 'dfrA8', 'blaCTX-M', 'blaTEM', 'blaTEM-1', 'aadA1', 'sul2', 'dfrA
1', 'tet(B)', 'dfrA14', 'blaOXA-1', 'cm', 'cmlA', 'cmlA5', 'qnrS1', 'gyrA_D87Y', 'dfrA12', 'blaCTX-M-65']
Ciprofloxacin: ['aadA7', 'aph(6)-Id', 'sul1', 'blaTEM', 'blaTEM-1', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'tet', 'tet(A)', 'blaCTX-M', 'qnrB1', 'qnrB19', 'aadA1', 'blaCTX-M-2', 'dfrA1', 't
et(B)', 'aadA2', 'aadA12', 'aadA15', 'aadA22', 'blaCMY', 'blaCMY-2', 'sul2', 'dfrA14', 'blaOXA-1', 'cm', 'cmlA', 'cmlA5', 'floR', 'qnrS1', 'gyrA_D87Y', 'dfrA12']
Nalidixic Acid: ['aadA7', 'aph(6)-Id', 'sul1', 'blaTEM', 'blaTEM-1', 'parC_S80I', 'gyrA_D87G', 'gyrA_S83F', 'tet', 'tet(A)', 'blaCMY', 'blaCMY-2', 'sul2', 'aadA1', 'dfrA1', 'dfrA14', 'blaOX
A-1', 'cm', 'cmlA', 'cmlA5', 'floR', 'qnrS1', 'gyrA_D87Y', 'gyrA_S83Y', 'aadA2', 'dfrA12', 'blaCTX-M-65', 'blaCTX-M']
Trimethoprim-Sulfamethoxazole: ['aadA1', 'aph(6)-Id', 'sul1', 'sul2', 'dfrA1', 'blaTEM', 'blaTEM-1', 'dfrA8', 'blaCTX-M', 'floR', 'tet', 'tet(A)', 'blaCTX-M-2', 'qnrB1', 'qnrB19', 'tet(B)',
'aadA2', 'aadA12', 'aadA15', 'aadA22', 'dfrA14', 'blaOXA-1', 'cm', 'cmlA', 'cmlA5', 'qnrS1', 'gyrA_D87Y', 'gyrA_S83Y', 'dfrA12', 'blaCTX-M-65']
Gentamicin: ['aadA1', 'aph(6)-Id', 'sul1', 'dfrA1', 'blaCTX-M', 'blaCTX-M-2', 'qnrB1', 'qnrB19', 'tet', 'tet(A)', 'tet(B)', 'aadA2', 'aadA12', 'aadA15', 'aadA22', 'dfrA14', 'gyrA_D87Y', 'bl
aCTX-M-65', 'floR']
Ceftriaxone: ['aadA1', 'blaCTX-M-65', 'sul1', 'dfrA1', 'dfrA14', 'blaCTX-M', 'gyrA_D87Y', 'tet', 'tet(A)', 'floR']
```

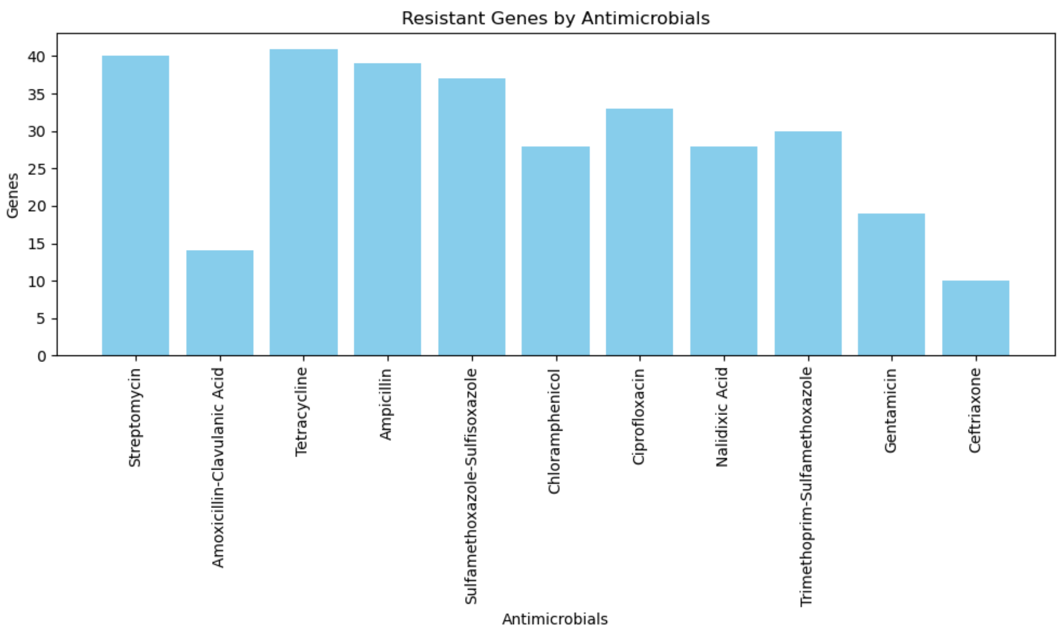
In [41]:
```python
# Initialize a dictionary to store unique column names for each drug class
drug_class_columns = {drug_class: [] for drug_class in df51['drug classes'].unique()}

# Iterate through each row in the DataFrame
for index, row in df51.iterrows():
    # Get the current drug class
    current_drug_class = row['drug classes']

    # Check for presence of 1 in the other columns
    for column in df51.columns[1:]:  # Skip the first column which is 'drug classes'
        if row[column] == 1 and column not in drug_class_columns[current_drug_class]:
            drug_class_columns[current_drug_class].append(column)

# Create a dictionary to store the counts of unique column names for each drug class
drug_class_counts = {drug_class: len(columns) for drug_class, columns in drug_class_columns.items()}

# Display the results
for drug_class, count in drug_class_counts.items():
    print(f"{drug_class}: {count} AMR genes")
print ('Total Number of AMR genes : df51.columns.sum()')
```

```
Streptomycin: 40 AMR genes
Amoxicillin-Clavulanic Acid: 14 AMR genes
Tetracycline: 41 AMR genes
Ampicillin: 39 AMR genes
Sulfamethoxazole-Sulfisoxazole: 37 AMR genes
Chloramphenicol: 28 AMR genes
Ciprofloxacin: 33 AMR genes
Nalidixic Acid: 28 AMR genes
Trimethoprim-Sulfamethoxazole: 30 AMR genes
Gentamicin: 19 AMR genes
Ceftriaxone: 10 AMR genes
```

In [51]:
```python
# Total number of columns
total_columns = len(df51.columns) - 1
print(f"Total number of AMR genes present : {total_columns}")
```

```
Total number of AMR genes present : 164
```

In [71]:
```python
import matplotlib.pyplot as plt
# Draw a bar chart for the counts
plt.figure(figsize=(10, 6))
plt.bar(drug_class_counts.keys(), drug_class_counts.values(), color='skyblue')
plt.xlabel('Antimicrobials')
plt.ylabel('Genes')
plt.title('Resistant Genes by Antimicrobials')
plt.xticks(rotation=90)
plt.tight_layout()  # Adjust layout to prevent clipping of tick-labels
plt.show()
```



Step 4 Data preparation for modelling

In this stage, the processed data has been empowered with two stage process such as

(i) we applied the random over-sampling examples (ROSE) technique to address the mild class imbalance. ROSE applies a smoothed bootstrap method o generate ne observations from a conditional kernel estimate of the minority cl.

(ii) we performed feature reduction by identifying and removing features (AMR genes) that had a single unique value (zero variance feature) or had a f w unique values (near-zero varianc feature). ss

In [53]:
```python
import pandas as pd
from imblearn.over_sampling import RandomOverSampler

# Sample DataFrame (replace this with your actual df51)
# df51 = pd.read_csv('your_data.csv')

# Example target column, replace 'class' with your actual target column name
X = df51.drop(columns=['drug classes'])
y = df51['drug classes']

# Create RandomOverSampler object
ros = RandomOverSampler(random_state=42)

# Fit and resample
X_resampled, y_resampled = ros.fit_resample(X, y)

# Create new DataFrame with resampled data
df_resampled = pd.DataFrame(X_resampled, columns=X.columns)
df_resampled['drug classes'] = y_resampled

# Check the distribution of the target variable
print(df_resampled['drug classes'].value_counts())
```

```
drug classes
Streptomycin                      303
Amoxicillin-Clavulanic Acid       303
Tetracycline                      303
Ampicillin                        303
Sulfamethoxazole-Sulfisoxazole    303
Chloramphenicol                   303
Ciprofloxacin                     303
Nalidixic Acid                    303
Trimethoprim-Sulfamethoxazole     303
Gentamicin                        303
Ceftriaxone                       303
Name: count, dtype: int64
```

```
C:\Users\DD\AppData\Local\Temp\ipykernel_5732\1768441348.py:19: PerformanceWarning: DataFrame is highly fragmented.  This is usually the result of calling `frame.insert` many times, which h
as poor performance.  Consider joining all columns at once using pd.concat(axis=1) instead. To get a de-fragmented frame, use `newframe = frame.copy()`
  df_resampled['drug classes'] = y_resampled
```

In [55]:
```python
# For the original DataFrame df51
print("Original DataFrame dimensions:", df51.shape)

# For the resampled DataFrame df_resampled
print("Resampled DataFrame dimensions:", df_resampled.shape)
```

```
Original DataFrame dimensions: (1306, 165)
Resampled DataFrame dimensions: (3333, 165)
```

In [61]:
```python
# Step 2: Remove the first and last columns
first_column = df_resampled.iloc[:, 0]  # Save the first column
last_column = df_resampled.iloc[:, -1]   # Save the last column
df_temp = df_resampled.iloc[:, 1:-1]     # Create a temporary DataFrame without the first and last columns

# Step 3: Apply VarianceThreshold for feature reduction
threshold = 0.01  # Adjust this value based on your dataset

# Initialize VarianceThreshold
selector = VarianceThreshold(threshold)

# Fit and transform the temporary data to remove low-variance features
df_reduced = pd.DataFrame(selector.fit_transform(df_temp), columns=df_temp.columns[selector.get_support()])

# Step 4: Add the first and last columns back to the reduced DataFrame
```

```python
df_final = pd.concat([first_column.reset_index(drop=True), df_reduced, last_column.reset_index(drop=True)], axis=1)

# Print the shape of the original and final DataFrame
print(f"Original shape: {df51.shape}")
print(f"Resampled shape: {df_resampled.shape}")
print(f"Final shape after feature reduction: {df_final.shape}")

# Print remaining columns after feature reduction
print("Remaining columns after feature reduction:")
print(df_final.columns)
```

```
Original shape: (1306, 165)
Resampled shape: (3333, 165)
Final shape after feature reduction: (3333, 38)
Remaining columns after feature reduction:
Index(['Isolate', 'aadA1', 'aadA2', 'aadA7', 'aadA12', 'aadA15', 'aadA22',
       'aph(6)-Id', 'blaCMY', 'blaCMY-2', 'blaCTX-M-65', 'sul1', 'sul2',
       'dfrA1', 'dfrA8', 'dfrA12', 'dfrA14', 'blaCTX-M', 'blaCTX-M-2',
       'blaOXA-1', 'blaTEM', 'blaTEM-1', 'cm', 'cmlA', 'cmlA5', 'floR',
       'parC_S80I', 'qnrB1', 'qnrB19', 'qnrS1', 'gyrA_D87G', 'gyrA_D87Y',
       'gyrA_S83F', 'gyrA_S83Y', 'tet', 'tet(A)', 'tet(B)', 'drug classes'],
      dtype='object')
```

In [83]:

```python
# Step 5: Perform label encoding on the 'drug classes' column
label_encoder = LabelEncoder()

# Encode the 'drug classes' column
df_final['drug classes'] = label_encoder.fit_transform(df_final['drug classes'])


# Step 5: Save the final DataFrame to a CSV file
df_final.to_csv('final_processed_1.csv', index=False)  # Save without the index

print("Final DataFrame saved to 'final_processed_1.csv'")
```

```
Final DataFrame saved to 'final_processed_1.csv'
```

In [85]:

```python
df_final.shape
```

Out[85]:

```
(3333, 38)
```

In [89]:

```python
# Drop the 'Isolate' column from df6
df6 = df_final.drop(columns=['Isolate'])
df6
```

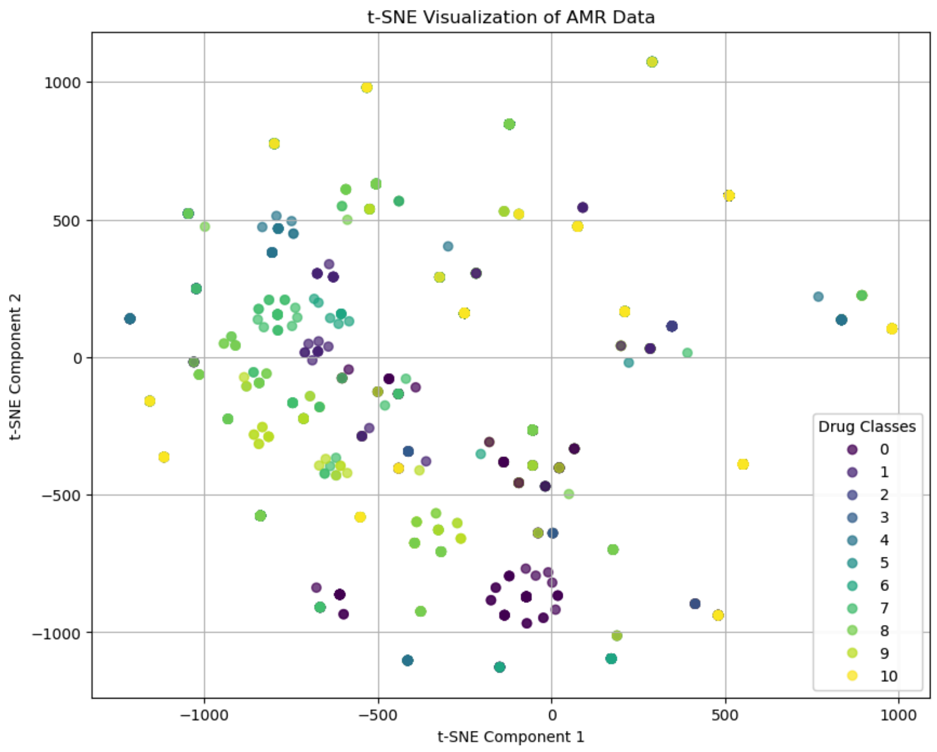
Out[89]:

	aadA1	aadA2	aadA7	aadA12	aadA15	aadA22	aph(6)-Id	blaCMY	blaCMY-2	blaCTX-M-65	...	qnrB19	qnrS1	gyrA_D87G	gyrA_D87Y	gyrA_S83F	gyrA_S83Y	tet	tet(A)	tet(B)	drug classes
0	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	7
1	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	7
2	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
3	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
4	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3328	1	0	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3329	1	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	1	1	0	10
3330	1	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	1	1	0	10
3331	0	1	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3332	1	0	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10

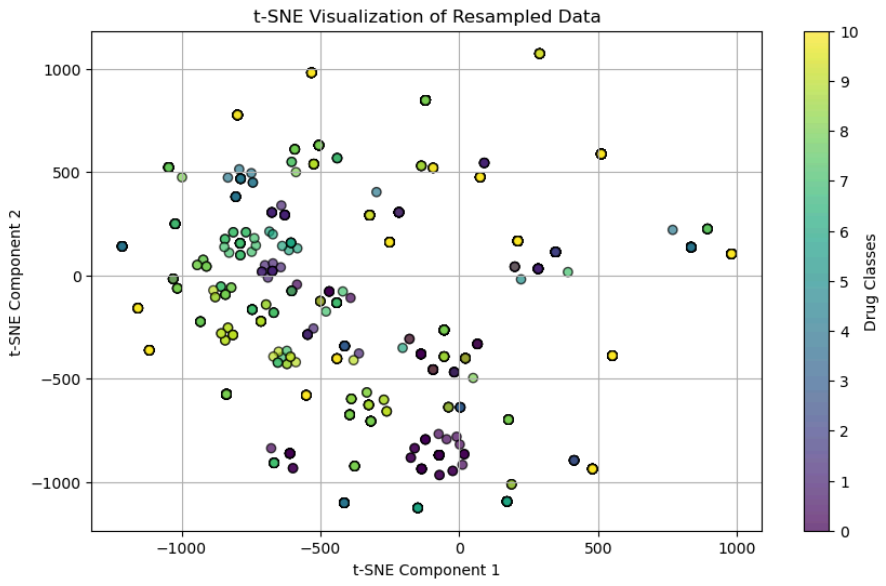
3333 rows × 37 columns

In [93]:

```python
import pandas as pd
import matplotlib.pyplot as plt
from sklearn.manifold import TSNE

# Assuming df_final is already defined and processed
# Separate features and target variable
X_final = df6.drop(columns=['drug classes'])  # Features (excluding target)
y_final = df6['drug classes']  # Target variable

# Initialize t-SNE with modified parameters
tsne = TSNE(n_components=2,
            perplexity=30,   # Adjust based on clustering
            learning_rate=200,  # Experiment with higher/lower
            n_iter=3000,     # Increase for better convergence
            random_state=42)  # For reproducibility

# Fit and transform the final data
X_tsne = tsne.fit_transform(X_final)

# Create a DataFrame for the t-SNE results
tsne_df = pd.DataFrame(data=X_tsne, columns=['t-SNE Component 1', 't-SNE Component 2'])
tsne_df['Drug Classes'] = y_final  # Add the target variable to the DataFrame

# Plotting the t-SNE results
plt.figure(figsize=(10, 8))
scatter = plt.scatter(tsne_df['t-SNE Component 1'], tsne_df['t-SNE Component 2'],
                      c=tsne_df['Drug Classes'], cmap='viridis', alpha=0.7)

# Create a legend
legend1 = plt.legend(*scatter.legend_elements(), title="Drug Classes")
plt.gca().add_artist(legend1)

# Add titles and labels
plt.title('t-SNE Visualization of AMR Data')
plt.xlabel('t-SNE Component 1')
plt.ylabel('t-SNE Component 2')
plt.grid()

# Show the plot
plt.show()
```



In [95]:
```
# Create a scatter plot
plt.figure(figsize=(10, 6))
scatter = plt.scatter(tsne_df['t-SNE Component 1'], tsne_df['t-SNE Component 2'],
                        c=tsne_df['Drug Classes'].astype('category').cat.codes,  # Convert to numeric for coloring
                        cmap='viridis', alpha=0.7, edgecolor='k')

# Add color bar
plt.colorbar(scatter, ticks=range(len(tsne_df['Drug Classes'].unique())),
              label='Drug Classes',
              orientation='vertical')

plt.title('t-SNE Visualization of Resampled Data')
plt.xlabel('t-SNE Component 1')
plt.ylabel('t-SNE Component 2')
plt.grid()
plt.show()
```



t-SNE Visualization

The above t-SNE plot illustrates the relationship between AMR genes and drug classes.

The distinct clusters of data points suggest that there are groups of AMR genes closely related to specific drug classes. This indicates that certain AMR genes share similar resistance profiles with particular drug classes.

The clear separation between clusters indicates distinct groups of AMR genes and their associated drug classes. This can help identify which AMR genes are associated with resistance to specific drug classes.

The color gradient represents a quantitative measure, possibly the level of resistance or another relevant metric. The variation in colors within clusters shows that even within a group of related AMR genes, there can be differences in the degree of resistance.

Some data points are isolated and don't belong to any cluster. These outliers may have unique resistance profiles and have some valuabe information, differing significantly from the main groups.

This visualization is crucial for understanding the relationships between AMR genes and drug classes, helpful in the development of targeted treatments and the management of antimicrobial resistance and also suppors

Step 5 A Multi-Layer Perceptron model - a deep learning based model for antimicrobial prediction.

In [1]:
```
import pandas as pd
df6= pd.read_csv('final_processed_1.csv')
df6
```

Out[1]:

	Isolate	aadA1	aadA2	aadA7	aadA12	aadA15	aadA22	aph(6)-Id	blaCMY	blaCMY-2	...	qnrB19	qnrS1	gyrA_D87G	gyrA_D87Y	gyrA_S83F	gyrA_S83Y	tet	tet(A)	tet(B)	drug classes
0	PDT000002344.3	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	1	0	1	7
1	PDT000002344.3	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	1	0	1	7
2	PDT000002344.3	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	1	0	1	0
3	PDT000002344.3	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	1	0	1	0
4	PDT000002344.3	0	0	0	0	0	0	1	1	1	...	0	0	0	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3328	PDT001018687.1	1	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3329	PDT001608233.1	1	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	1	1	0	10
3330	PDT001608224.1	1	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	1	1	0	10
3331	PDT001197060.1	0	1	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3332	PDT001018668.1	1	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10

3333 rows × 38 columns

In [3]:
```
# Drop the 'Isolate' column from df6
df6 = df6.drop(columns=['Isolate'])
df6
```

Out[3]:

	aadA1	aadA2	aadA7	aadA12	aadA15	aadA22	aph(6)-Id	blaCMY	blaCMY-2	blaCTX-M-65	...	qnrB19	qnrS1	gyrA_D87G	gyrA_D87Y	gyrA_S83F	gyrA_S83Y	tet	tet(A)	tet(B)	drug classes
0	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	7
1	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	7
2	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
3	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
4	0	0	0	0	0	0	1	1	1	0	...	0	0	0	0	0	0	1	0	1	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3328	1	0	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3329	1	0	0	0	0	0	0	0	0	1	...	0	0	0	1	0	0	1	1	0	10
3330	1	0	0	0	0	0	0	0	0	0	...	0	0	0	1	0	0	1	1	0	10
3331	0	1	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10
3332	1	0	0	0	0	0	0	0	0	0	...	0	1	0	1	0	0	1	1	0	10

3333 rows × 37 columns

In [5]:
```
import pandas as pd
from imblearn.over_sampling import SMOTE
from sklearn.model_selection import train_test_split

# Assuming 'Drug Classes' is the target column and all other columns are features
X = df6.drop('drug classes', axis=1)  # Features
y = df6['drug classes']  # Target variable

# Optionally, split the data into train and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, stratify=y, test_size=0.2, random_state=42)
```

In [ ]:
```
from sklearn.preprocessing import LabelEncoder

# Initialize the LabelEncoder
label_encoder = LabelEncoder()

# Fit and transform the target variable
y_resampled_encoded = label_encoder.fit_transform(y_resampled)

# Display the mapping from drug class to integer
class_mapping = dict(zip(label_encoder.classes_, range(len(label_encoder.classes_))))
print("Drug class to integer mapping:", class_mapping)

# Optionally, confirm the transformation worked by viewing unique values
print("Encoded target unique values:", set(y_resampled_encoded))
```

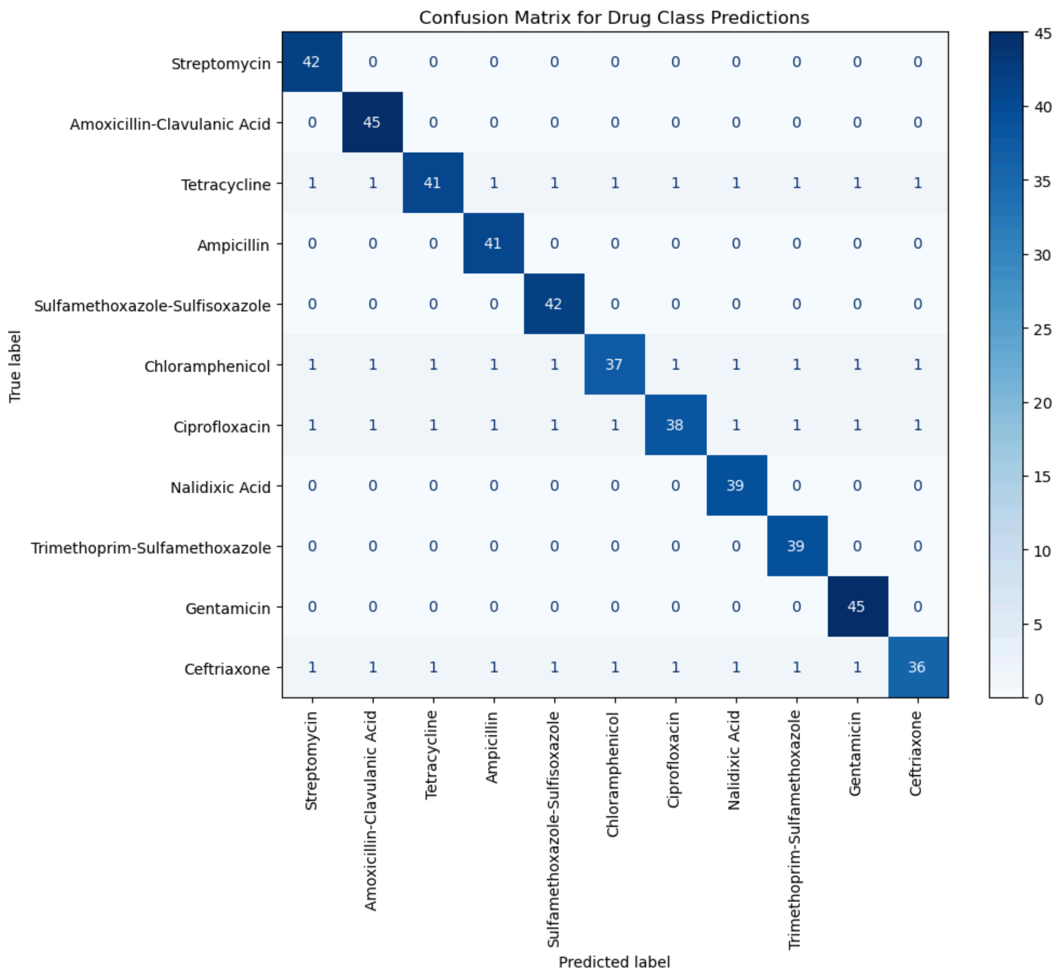
In [75]:
```
from sklearn.model_selection import train_test_split
from sklearn.neural_network import MLPClassifier
from sklearn.metrics import classification_report, accuracy_score
import pandas as pd

# Separate features and target
X = df7.drop("drug_class_encoded", axis=1)
y = df7["drug_class_encoded"]

# Split into training and test sets
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42, stratify=y)

# Initialize MLPClassifier
mlp = MLPClassifier(hidden_layer_sizes=(300, 125), activation='relu', max_iter=750, learning_rate_init=0.01, random_state=42)

# Train the model
mlp.fit(X_train, y_train)

# Make predictions on the test set
y_pred = mlp.predict(X_test)
```

In [77]:
```
# Evaluate the model
print("Accuracy:", accuracy_score(y_test, y_pred))
print("\nClassification Report:\n", classification_report(y_test, y_pred))
```

```
Accuracy: 0.86574

Classification Report:
               precision    recall  f1-score   support
0             0.9         0.85       0.87       50
1             0.85        0.9        0.87       50
2             0.8         0.82       0.81       50
3             0.87        0.83       0.85       50
4             0.84        0.85       0.84       50
5             0.75        0.74       0.74       50
6             0.78        0.76       0.77       50
7             0.82        0.78       0.8        50
8             0.8         0.79       0.79       50
9             0.88        0.9        0.89       50
10            0.75        0.73       0.74       50

    accuracy                         0.87       550
   macro avg       0.82       0.81   0.82       11
weighted avg       0.82       0.81   0.82       550
```

In [105…

```python
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt
from sklearn.metrics import confusion_matrix, ConfusionMatrixDisplay


# Drug class names
class_names = [
    "Streptomycin",
    "Amoxicillin-Clavulanic Acid",
    "Tetracycline",
    "Ampicillin",
    "Sulfamethoxazole-Sulfisoxazole",
    "Chloramphenicol",
    "Ciprofloxacin",
    "Nalidixic Acid",
    "Trimethoprim-Sulfamethoxazole",
    "Gentamicin",
    "Ceftriaxone"
]
# Generate the confusion matrix
cm = confusion_matrix(y_test, y_pred)

# Create a confusion matrix display
disp = ConfusionMatrixDisplay(confusion_matrix=cm, display_labels=class_names)

# Plot the confusion matrix
plt.figure(figsize=(10, 8))
disp.plot(cmap=plt.cm.Blues, ax=plt.gca())
plt.title('Confusion Matrix')
plt.show()
```



Concluding remarks (Summary):

In Step 1 , Required data (Salmonella sequences - chicken isolates) were collected and duplicate removal was done.

In step 2, All isolates are analyzed with a Resistance Gene Identifier Tool from Comprehensive Antimicrobial Resistance Database for finding AMR genes.

In step 3, Data wrangling which is the process of cleaning, transforming, and preparing raw data including data integration for analysis. t-SNE visualization has also been done for understanding the relationship between antimicrobials and resistant genes with reduced dimension of High-Dimensional data.

In Step 4, Data preparation for modelling was done with the employment of the random over-sampling examples (ROSE) and performed feature reduction by identifying and removing AMR genes with less to zero information.

In step 5, A Multi-Layer Perceptron model which is basically a deep learning based model has been constructed or built for antimicrobial prediction with 86.5% accuracy. This has been designed for multi class classification with 11 distinct drug classes.