Data Requirmetns

Predictive Analysis of Antibiotic Resistance in Bacterial Strains via DNA Sequence Motifs

Author: Michał Raczkowski
Student number: 4465024
Fontys ICT - AI-core-AI4-RB02

September 22, 2023

Contents

1 Genomic Data

Sequence Data:

- **Type:** DNA sequences of bacterial strains.
- **Format:** FASTA. Each entry should have a unique identifier and the corresponding DNA sequence.
- **Details:** Complete genomes or specific genes associated with antibiotic resistance (e.g., *bla* genes for *E. coli*).

Metadata:

- **Type:** Supplementary information for each bacterial strain.
- **Details:** Bacterial strain identifier, source, collection date, location, clinical outcomes, resistance phenotypes, etc.

Antibiotic Resistance Phenotype Data

Resistance Profile:

- **Type:** Results from antibiotic susceptibility tests.
- **Format:** Tabular data linking strain identifiers to susceptibility results.
- **Details:** Includes MIC values, interpretation (Resistant, Intermediate, Susceptible), antibiotic type, etc.

2 Auxiliary Data

Reference Data:

- **Type:** Known antibiotic resistance genes and variants.
- **Source:** Databases like ResFinder, CARD, etc.
- **Format:** FASTA or similar with sequences of known resistance genes.
- **Details:** Aid in identifying genes or motifs in the genome data associated with resistance.

Control Sequences:

- **Type:** DNA sequences of strains known to be susceptible.
- **Purpose:** For building a balanced dataset.

3 Data Quality Requirements

- **Accuracy:** Accurate sequences without errors.
- **Completeness:** Full genomes or comprehensive gene sequences.
- **Consistency:** Matching metadata and phenotype data.
- **Resolution:** High-resolution genomic data, preferably from WGS.

4 Data Storage and Management

- **Database System:** A relational database system like MySQL or PostgreSQL.
- **Backup:** Regular data backups.

5 Possible Data Sources

NCBI GenBank: A comprehensive public database of nucleotide sequences and supporting bibliographic and biological annotation.

ResFinder: A database hosted by the Center for Genomic Epidemiology, focusing specifically on antibiotic resistance genes in bacteria.

CARD (Comprehensive Antibiotic Resistance Database): A rigorously curated collection of known resistance genes.

MicrobesOnline: A platform offering integrated tools for visualizing and analyzing microbial genomes and their associated functional information.

ENA (European Nucleotide Archive): A globally comprehensive data resource for nucleotide sequence, spanning raw data, alignments, and assembled/annotated sequences.

PATRIC (Pathosystems Resource Integration Center): A bioinformatics resource center that provides comprehensive bacterial infectious disease data.