Computational analysis of bacterial plasmids

MatFyz CONNECTIONS poster

Abstract

- Plasmids are small, circular, extrachromosomal DNA molecules commonly found in bacteria. They can be transferred between different bacterial cells through horizontal gene transfer and often carry genes conferring antimicrobial resistance (AMR), making them a critical focus in the study of antibiotic resistance.
- The goal of our work is to develop new bioinformatics methods for plasmid detection and comparison, using techniques from machine learning and comparative genomics. We explore computational approaches for classifying plasmid sequences based on high-throughput sequencing data. Using k-mer profiles, various sequence-derived features, and homology-based log-odds scores, we train machine learning models to distinguish plasmid reads from chromosomal ones.
- Our dataset, consisting of multiple E. coli isolates, presents significant challenges due to class imbalance – plasmid reads are markedly underrepresented. Experimental results show that data partitioning strategies and isolate-specific differences have a strong effect on classification performance.

Plasmids & our task

- Circular DNA molecules commonly found in bacteria
- Independent replication
- Often carry antibiotic resistance (AMR) genes

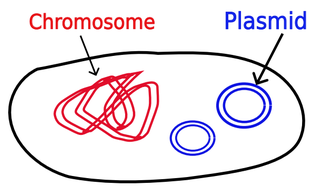


Figure 1. Bacterial cell: chromosomal and plasmid DNA

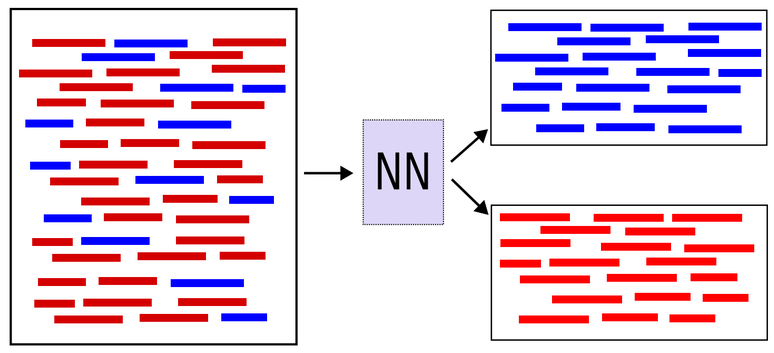
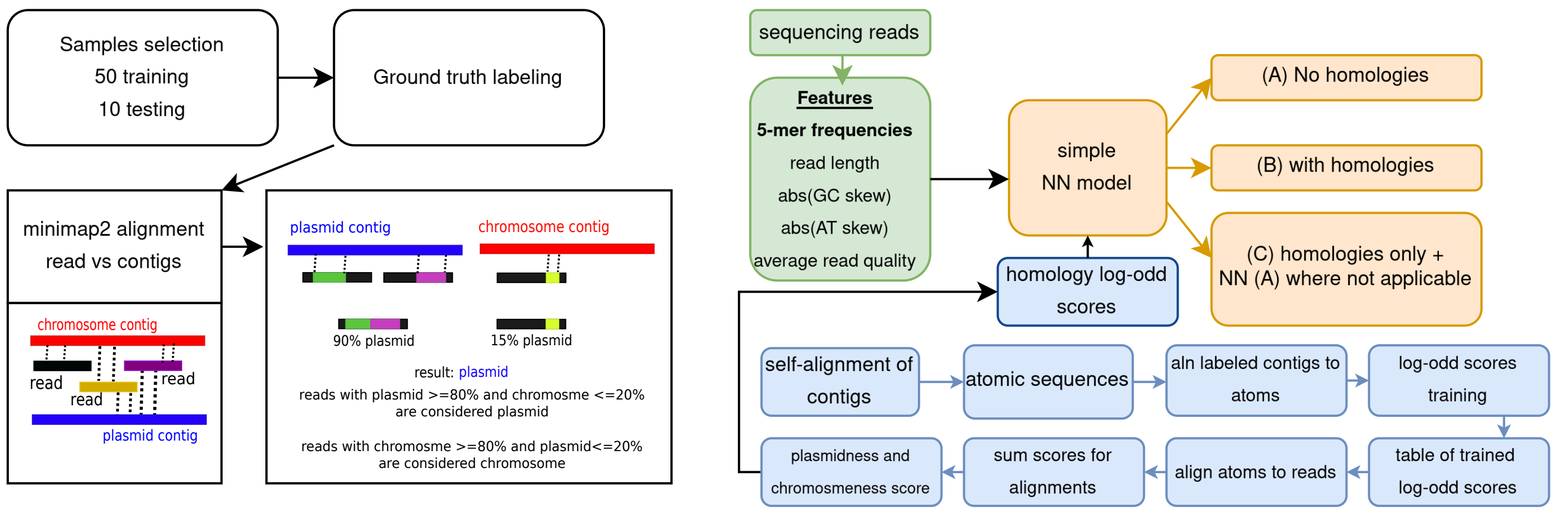


Figure 2. Our task: distinguishing plasmid and chromosomal reads

Graphical abstract



Results

Model	Accuracy	Precision	Recall	F1 score
(A) NN without homolgies	0.9437	0.8703	0.8622	0.8663
(B) NN with homology scores	0.9665	0.9238	0.9173	0.9205
(C) Homology-based rule + (A)	0.9738	0.9538	0.9209	0.9371
mlplasmids	0.7271	0.3963	0.5551	0.4625
plasclass	0.8881	0.6721	0.9195	0.7766
plasforest	0.8594	0.9203	0.3668	0.5245
platon	0.8108	0.999	0.1055	0.1908
rfplasmid	0.7967	0.9938	0.0392	0.0754

Table 1. Results of trained models and comparison with other tools. Plasmids considered as positive class. (A) *NN without homologies* is the simplest model, with only k-mer profile and non-homology features. (B) *NN with homology scores* uses all the features that were used in case (A) and homology scores for plasmids and for chromosomes. (C) *Homology-based rule* is a simple decision rule based on homologies where available, model (A) used where homologies are not available.

- We provide three models (A, B, C) for classification of plasmid sequencing reads.
- Results show that our models outperform other tested tools for plasmid read classification task, however, those tools were designed for classification of longer contigs, not sequencing reads.
- A simple homology-based decision rule (model C) outperformed the NN models.

References

[1] Sergio Arredondo-Alonso, Malbert R C Rogers, Johanna C Braat, Tess D Verschuuren, Janetta Top, Jukka Corander, Rob J L Willems, and Anita C Schürch.
Mlplasmids: A user-friendly tool to predict plasmid- and chromosome-derived sequences for single species.
Microb. Genom., 4(11), November 2018.

[2] David Pellow, Itzik Mizrahi, and Ron Shamir.
PlasClass improves plasmid sequence classification.
PLoS Comput. Biol., 16(4):e1007781, April 2020.

[3] Léa Pradier, Tazzio Tissot, Anna-Sophie Fiston-Lavier, and Stéphanie Bedhomme.
PlasForest: a homology-based random forest classifier for plasmid detection in genomic datasets.
BMC Bioinformatics, 22(1):349, June 2021.

[4] Oliver Schwengers, Patrick Barth, Linda Falgenhauer, Torsten Hain, Trinad Chakraborty, and Alexander Goesmann.
Platon: identification and characterization of bacterial plasmid contigs in short-read draft assemblies exploiting protein sequence-based replicon distribution scores.
Microb. Genom., 6(10), October 2020.

[5] Linda van der Graaf-van Bloois, Jaap A Wagenaar, and Aldert L Zomer.
RFPlasmid: predicting plasmid sequences from short-read assembly data using machine learning.
Microb. Genom., 7(11), November 2021.

Funding VEGA 1/0538/22 and 1/0140/25, H2020 872539 (PANGAIA)

FAKULTA MATEMATIKY, FYZIKY A INFORMATIKY
Univerzita Komenského v Bratislave

MATFYZ CONNECTIONS

Jana Černíková, Tomáš Vinař
Department of Applied Informatics FMFI UK
Broňa Brejová
Department of Computer Science FMFI UK