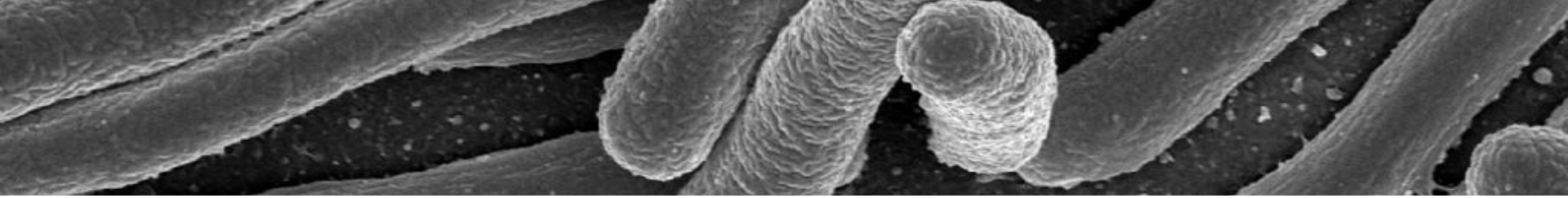


Plasmid Identification Through Graph Neural Networks

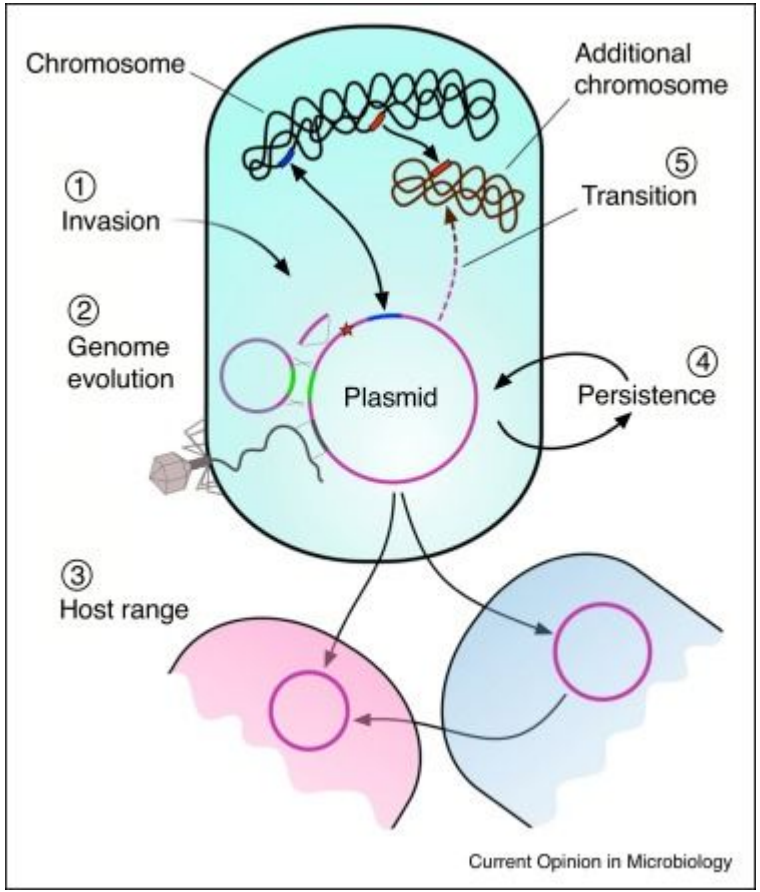
Broňa Brejová[1], Veronika Tordová[1], Kristián Andraščík[1], Cedric Chauve[2], Tomáš Vinař[1]

[1] Comenius University in Bratislava, Slovakia
[2] Simon Fraser University, Burnaby, BC, Canada

What are plasmids and why are we interested?

- small circular DNA molecules, typically in bacteria
- replicate independently of chromosomal DNA
- carry genes that confer selective advantage - e.g. antibiotic resistance
- facilitate horizontal transfer between individuals

This typically means **observable differences between chromosome and plasmid sequences.**

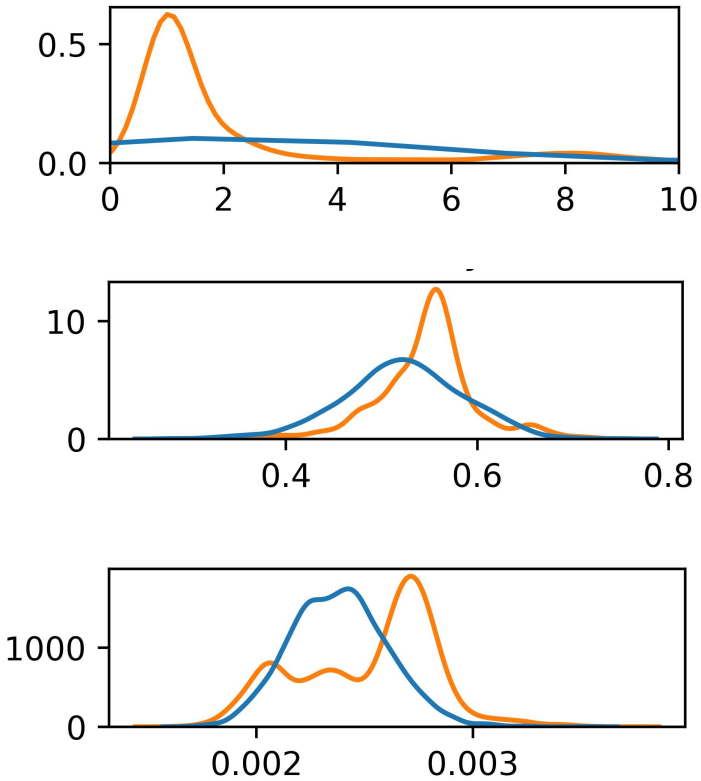


Sequence-based properties for *Klebsiella oxytoca*

normalized read coverage

GC content

relative *k*-mer content (dot product of contig k-mer content and whole-sample k-mer content vectors)



orange: chromosomes
blue: plasmids

Problem

- In a bacterial genome assembled from a bacterial isolate **identify which contigs correspond to chromosomes and which contigs correspond to plasmids**

Is this a difficult task?

- Depends largely on the **quality of the assembly**

Example: hybrid assembly *C. freundii* SAMN15148288

>1 length=5061881 depth=1.00x circular=true
>2 length=230132 depth=0.58x circular=true
>3 length=103762 depth=2.48x circular=true
>4 length=35077 depth=8.77x circular=true
>5 length=6790 depth=30.30x circular=true
>6 length=3370 depth=22.65x circular=true
>7 length=2001 depth=27.23x circular=true

Example: short-read assembly *C. freundii* SAMN15148288

>1 length=1221865 depth=1.00x
>2 length=527709 depth=1.04x
>3 length=460930 depth=1.00x
>4 length=456861 depth=0.98x
>5 length=366649 depth=1.01x

….

>170 length=108 depth=1.13x
>171 length=106 depth=1.29x
>172 length=106 depth=1.24x
>173 length=102 depth=0.70x

Problem

- In a bacterial genome assembled from a bacterial isolate
  identify which contigs correspond to chromosomes
  and which contigs correspond to plasmids

Is this a difficult task?

- Depends largely on the **quality of the assembly**
- **Long contigs** easily classified, **short contigs** almost impossible

Our goal:

- Contig classification for **short-read assemblies** with **many short contigs**

Standard approaches

Classification of individual contigs:

- **based on sequence features**
  k-mer composition, GC content (specific for each species)
  coverage by sequencing reads
  [mlplasmids, PlasClass, …]

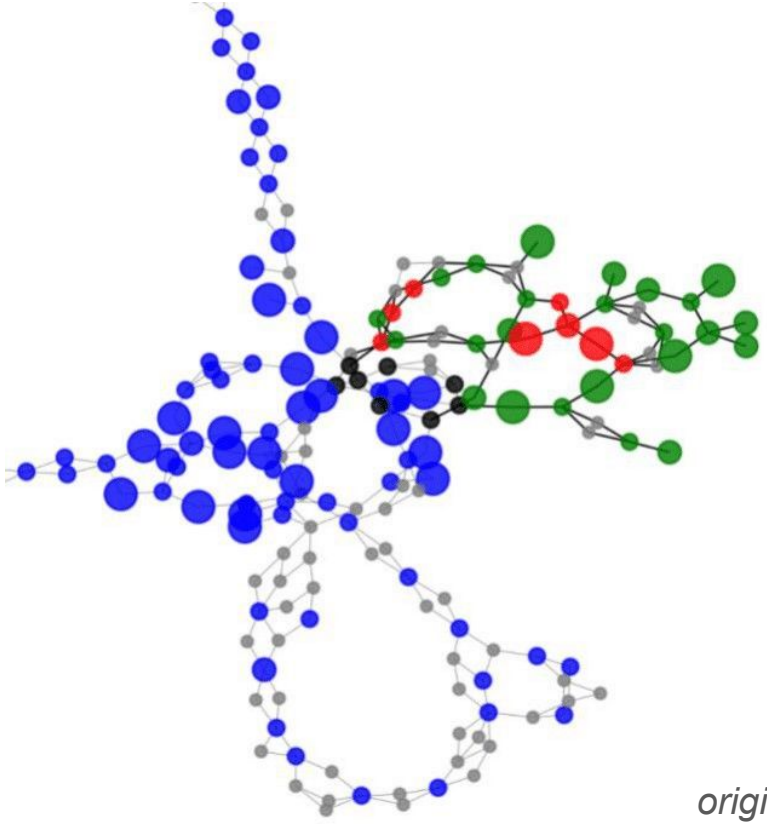
- **based on similarity to known chromosomes / plasmids**
  [PlasForest, Platon, Deeplasmid, RFPlasmid, …]

… but context matters

"… Analysing the pattern of ABR gene occurrence in the genomes of 2635 *Enterobacteriaceae* isolates, we find that **33% of the 416 ABR genes are shared between chromosomes and plasmids**. Phylogenetic reconstruction of ABR genes occurring on both plasmids and chromosomes supports their **evolution by lateral gene transfer**. …"

Wang et al. 2022, Philos Trans R Soc Lond B Biol Sci.

Assembly graph



label
chromosome
ambiguous
plasmid 1
plasmid 2

lengths
<= 100 bp
<= 1 kbp
<= 10 kbp
> 10 kbp

original idea Pu and Shamir, 2022, Bioinformatics

Graph Neural Networks

Graph convolution layers (GCLs):

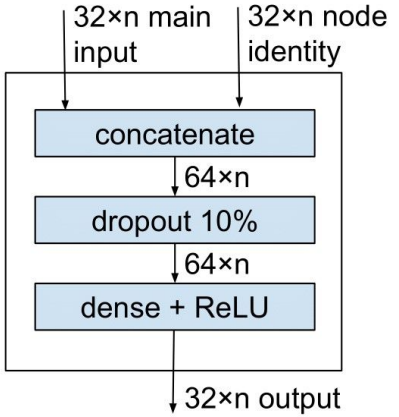
$Z = \sigma(\tilde{D}^{-1/2} \tilde{A} \tilde{D}^{-1/2} X\Theta + b)$

- combine feature values of each node with its neighbors
  (scaled by a function of node degree to prevent numerical explosion)
- combine values of features within node using linear combination
  (trainable weights Θ and b)
- pass through a non-linear activation function σ

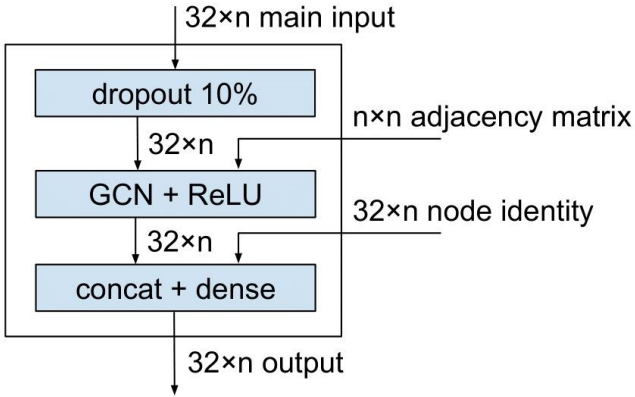
Using *d* GCL layers, information is integrated from nodes at a distance <=*d*.

Kipf and Welling, 2016

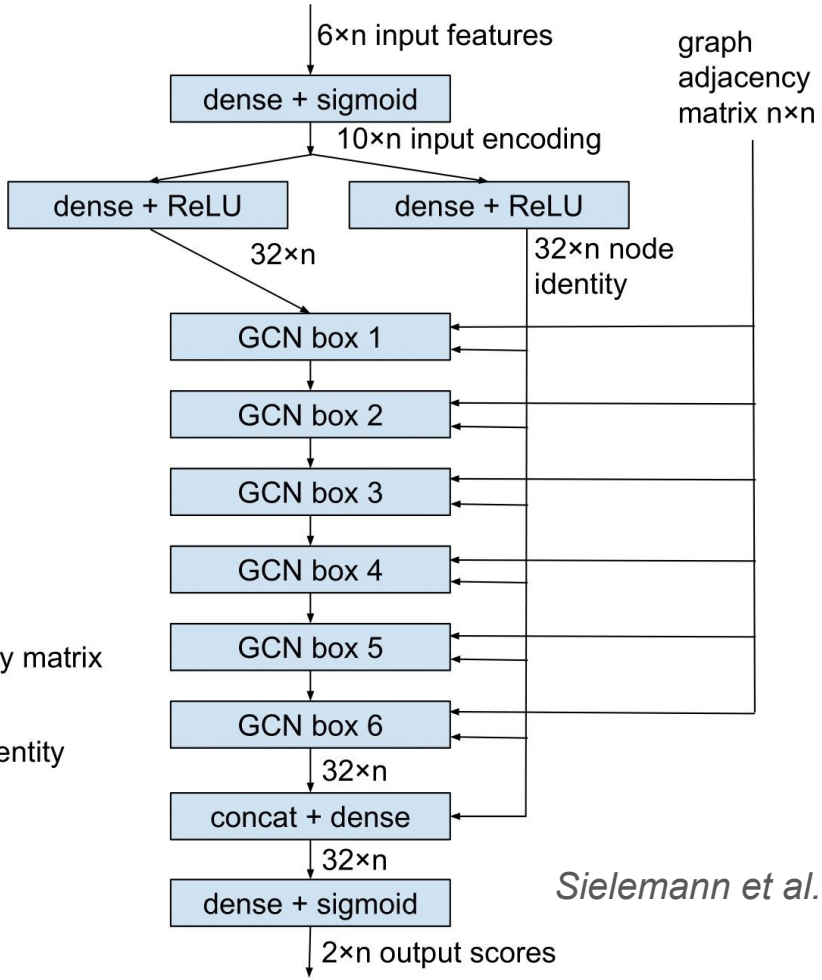
Concat + dense box:



GCN box:



Overall architecture:



Sielemann et al., 2023, Front. Microbiol.

Relative features - example: relative k-mer content

- vector p = 5-mer content of whole assembly (dominated by chromosomes)
- vector q = 5-mer content of a given contig
- relative k-mer content: ⟨p, q⟩
  - chromosome contigs will have a 5-mer composition **similar** to the overall assembly
    => large values
  - plasmid contigs will have 5-mer composition **different** from the overall assembly
    => small values
- prevents network learning specific k-mers included in specific genomes from the training set

Sielemann et al., 2023, Front. Microbiol.

plASgraph2 - experimental evaluation

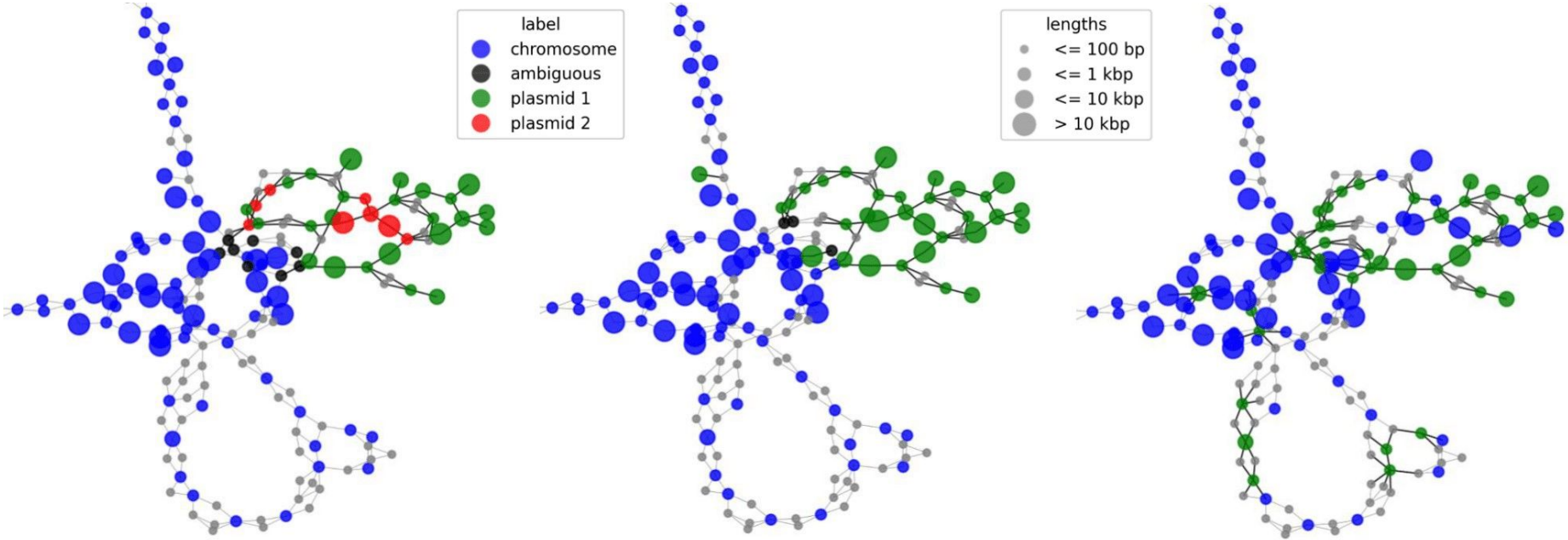
- trained on a dataset of 140 short-read assemblies from the ESKAPEE isolates
- tested on a dataset of 224 short-read assemblies from the ESKAPEE isolates
- annotated using hybrid assemblies of the same isolates
- additional testing on closely-related and more distant species that were not included in the training set

ESKAPEE = *Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter* spp., *Escherichia coli*

Sielemann et al., 2023, Front. Microbiol.

Method	SS	DB	AUROC	Precision	Recall	F1	Accuracy
A: Plasmid classification, contigs >100 bp, n =38,110							
plASgraph2	–	–	**0.991**	0.906	0.908	**0.808**	**0.935**
mlplasmids	X	–	0.896	0.273	**0.957**	0.480	0.641
PlasClass	–	–	0.892	0.381	0.939	0.617	0.794
PlasForest	–	X	n/a	0.486	0.939	0.711	0.852
Platon	–	X	n/a	**1**	0.5	0.667	0.924
Deeplasmid	–	X	n/a	n/a	n/a	n/a	n/a
RFPlasmid	X	X	0.973	0.854	0.789	0.667	0.885

Sielemann et al., 2023, Front. Microbiol.

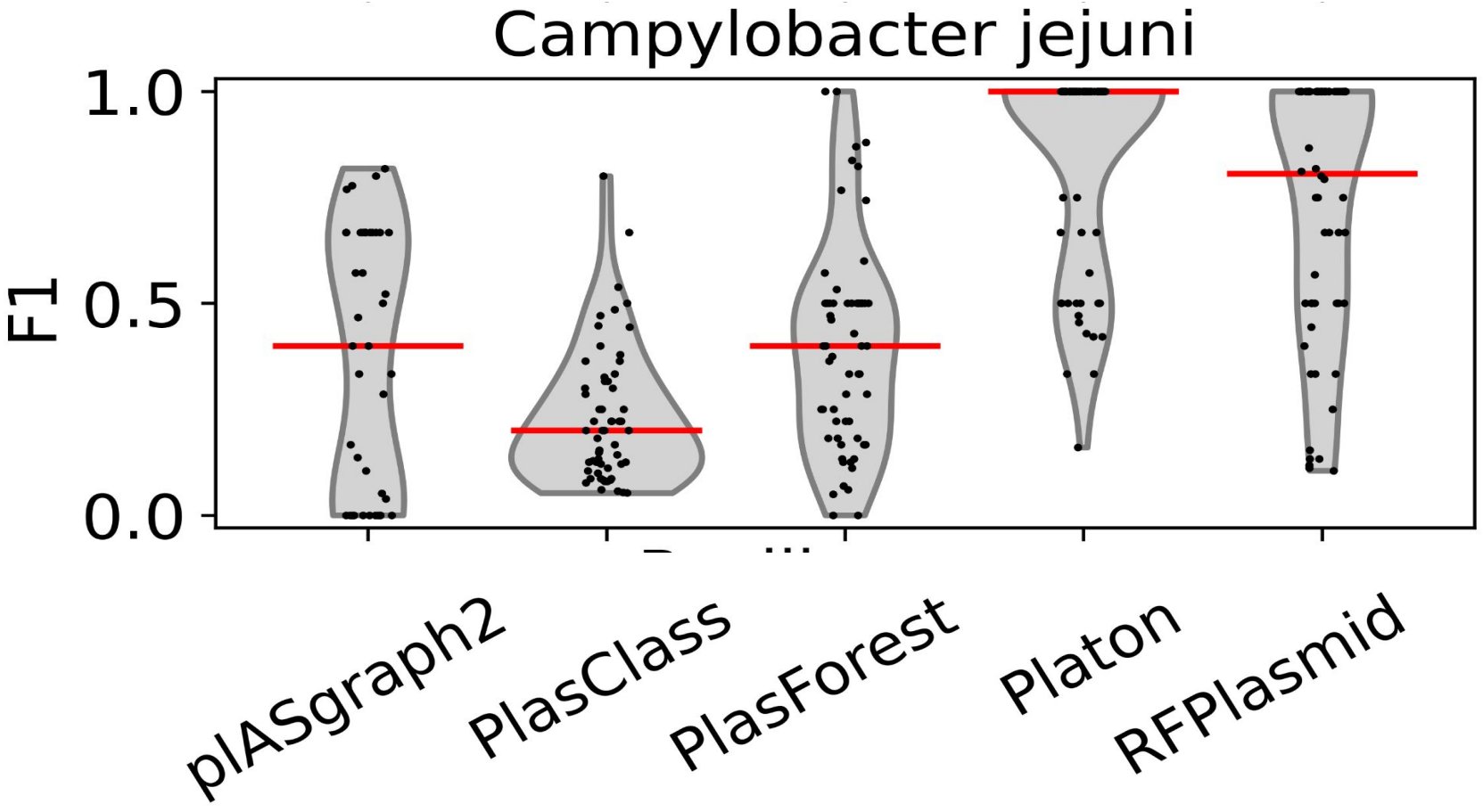


ground-truth

plASgraph prediction

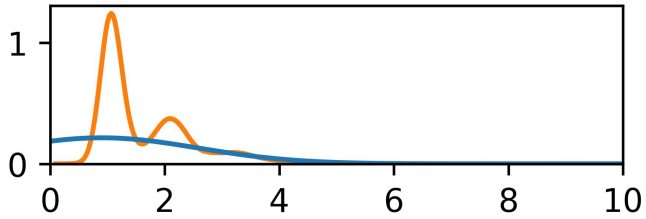
PlasForest prediction

Sielemann et al., 2023, Front. Microbiol.

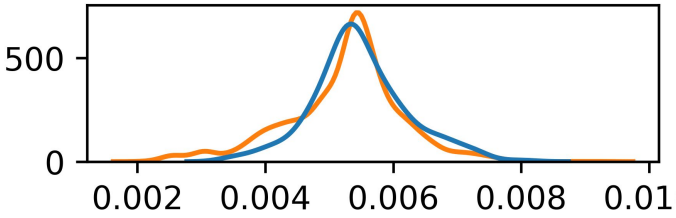

Campylobacter jejuni
F1
1.0
0.5
0.0
plASgraph2
PlasClass
PlasForest
Platon
RFPlasmid

Sequence-based properties for *Campylobacter jejuni*

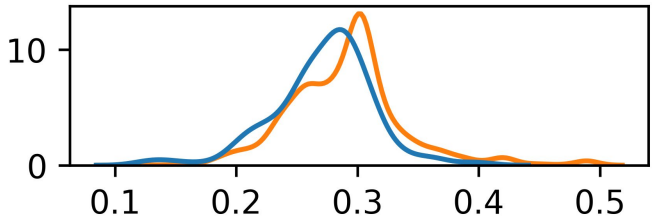
normalized read coverage



relative GC content



relative *k*-mer content (dot product of contig k-mer content and whole-sample k-mer content vectors)

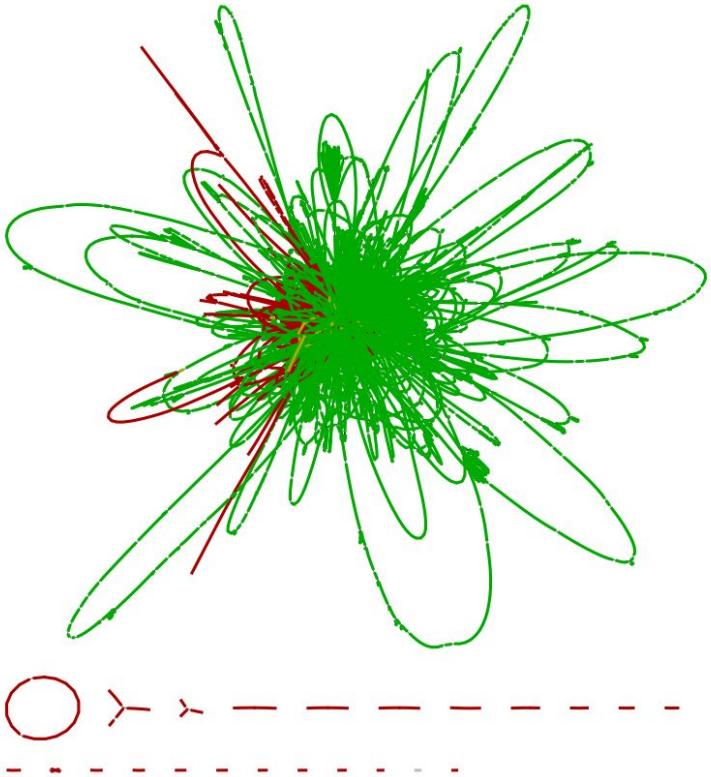


orange: chromosomes
blue: plasmids

How to introduce homology?

- introduce **tags** that are easily located in sequences and may indicate **plasmids** or **chromosomes**
  - Pfam protein families
  - simple homology-based tags
- logodds score based on **number of occurrences in training** (additive, similar to BLOSUM scores)
  - positive = more likely plasmid
  - zero = neutral
  - negative = more likely chromosome

Pan-genome approach with Geese



Method	SS	DB	AUROC	Precision	Recall	F1	Accuracy
plASgraph2	–	–	0.991	0.906	0.908	0.808	0.935
plASgraph2 + Pfam	–	X	0.996	0.980	1.000	0.926	0.970
plASgraph2 + Geese	–	X	1.000	1.000	1.000	0.970	0.988
mlplasmids	X	–	0.896	0.273	0.957	0.480	0.641
PlasClass	–	–	0.892	0.381	0.939	0.617	0.794
PlasForest	–	X	n/a	0.486	0.939	0.711	0.852
Platon	–	X	n/a	1	0.5	0.667	0.924
RFPlasmid	X	X	0.973	0.854	0.789	0.667	0.885

Brejová et al., 2025, unpublished

Acknowledgements

Computational Biology Group at Comenius University in Bratislava

- Broňa Brejová

Dept. of Mathematics, Simon Fraser University, Canada



- Cedric Chauve
- Aniket Mane
- Mahsa Faizrahnemoon

Bielefeld University


- Janik Sielemann
- Katharina Sielemann

PANGAIA

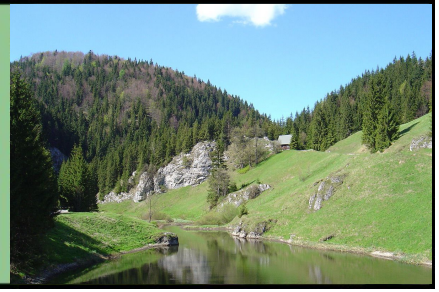
VEGA 1/0538/22 (TV)
VEGA 1/0140/25 (BB)
SRDA APVV-22-0144

Digital Research Alliance of Canada

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 872539

ITAT

26.9.-30.9.2025
Hotel Telgárt,
Nízke Tatry



WBCB 2025: Workshop on Bioinformatics and Computational Biology

The workshop is organized to help foster contacts within the bioinfo and compbio communities, especially those in Central Europe, within the settings of a larger IT meeting. It will provide a forum for the exchange of ideas on the newest developments in the disciplines, as well as an opportunity to introduce the history and interests of individual labs/groups.

Key dates:
full paper submission: June 27, 2025
abstract submission: August 7, 2025

https://wbcb.biocenter.sk/

Invited speaker:
Fereydoun Hormozdiari, University of California at Davis, USA

Program chairs:
Monika Čechová, Luca Denti