# Plasmid Identification Through Graph Neural Networks

Broňa Brejová1, Veronika Tordová1, Kristián Andraščík1, Cedric Chauve2, Tomáš Vinař1

1 Faculty of Mathematics, Physics and Informatics, Comenius University in Bratislava, Slovakia

2 Dept. of Mathematics, Faculty of Science, Simon Fraser University, Burnaby, BC, Canada

Identification of plasmids from sequencing data is an important and challenging problem related to antimicrobial resistance spread. The problem is typically addressed by machine learning methods combining features derived from input contigs. We provide a new architecture for identifying plasmid contigs in fragmented genome assemblies built from short-read data. We employ graph neural networks (GNNs) and the assembly graph to propagate the information from nearby nodes, which leads to more accurate classification, especially for short contigs that are difficult to classify based on sequence features or database searches alone. Our software tool, plASgraph2, either outperforms or performs on par with a wide range of state-of-the-art methods.

Some methods also use additional features based on homology to sequences typical for known plasmids or chromosomes. We propose a method for creating such features using log-odds scores based on ideas similar to those traditionally used in sequence alignment scoring, such as BLOSUM scoring matrices. The framework is flexible as it can handle both close homolog tags derived from a pangenome of training sequences as well as protein domains capturing distant homology. Inclusion of these features into the plASgraph2 graph neural network further significantly improves the accuracy of the predictions.

Availability: Our software is available at <https://github.com/cchauve/plasgraph2> and the training and testing data sets are available at <https://github.com/fmfi-compbio/plasgraph2-datasets>.

Acknowledgements: This work was partially supported by a grant from the European Union Horizon 2020 research and innovation program No. [872539] (PANGAIA); and grants from the Slovak Research and Development Agency [APVV-22-0144] and the Scientific Grant Agency VEGA [1/0140/25, 1/0538/22]. This research was enabled in part by computational infrastructure support provided by Digital Research Alliance of Canada (https://alliancecan.ca).