Detection of tail fibre proteins via machine learning methods

Jozef Bača[1], Andrej Baláž[1,2]

[1] *Department of Applied Informatics, Faculty of Mathematics, Physics and Informatics, Comenius University, Bratislava, Slovakia*
[2] *Geneton Ltd., Bratislava, Slovakia*

Bacteriophages are viral organisms infecting and killing bacteria. They contain tail fibre proteins, which can be utilised in the diagnostics and consequent treatment of bacterial infections. The identification of these proteins in the laboratory is expensive and time-consuming. In this work, we present a bioinformatics tool, capable of detecting tail fibre proteins from a set of proteins with unknown functions. To create this tool, we utilised the information from public databases, identified common features of tail fibre proteins, and used those features to train multiple machine learning models. The models were evaluated and the best performing model was deployed to the Anaconda cloud for the ease of installation and use. The comparison of the model's performance with current state-of-art models on the test set showed significant improvement in F1 score for the tail fibre detection.