Sophia Bałdysz

A new identification method of phage cell wall-associated lytic proteins

Bacteriophage lytic proteins (lysins) are the key factors in the viral infection process. Their role is twofold - they enable the injection of the phage genomic material into the host cell through local degradation of the peptidoglycan and cause the breakdown of the wall at the end of the replication cycle, resulting in progeny release. Despite common biological function, these proteins belong to many different families, have a wide variety of structures and display several distinct types of enzymatic activity. Hence, it is difficult to identify new candidate lytic enzymes. However, these proteins are used in many areas of the industry and medicine. In the age of a growing number of infections caused by antibiotic-resistant bacteria, proteins with bactericidal properties may be promising alternative therapeuticals and preparations containing lysins are at various stages of clinical trials. Additionally, several products containing bactericidal enzymes are already used in cosmetics. Thus, the demand for new lysins for medical and industrial use is high but finding them is not an easy task. Traditionally, the search for new lytic proteins required the isolation of phages, the sequencing of their genomes and subsequent laboratory screening for the right enzyme. This work was tedious, costly and often did not bring satisfactory results. Hence, researchers have developed bioinformatic tools to help identify novel lysins. Unfortunately, many of these programs are currently unavailable and those that are have not been updated since their release. Additionally, models that are at the heart of these tools were trained and tested on small, biased, redundant or misannotated datasets. This limits the reliability of their predictions, and hence, their applicability in the lysin screening process.

Therefore, the aim of this project was to develop a novel identification method/tool for phage cell wall-associated lytic proteins, based on the use of machine learning techniques, which emerged in the last decade and are rapidly developed. The construction of such a tool required an in-depth analysis of the physicochemical properties of lytic enzymes, the conserved domains associated with bacteriophage lysins and the protein families that group them.

Several machine learning models, as well as sequence representations were tested and optimized. The best method was based on a relatively simple artificial neural network and a combination of two amino acid composition representation methods, with an explicit encoding of the physicochemical properties of the whole protein. The benchmark of the trained model and currently available tools showed that the proposed solution outperforms existing ones in terms of several key metrics, including F1 score and recall.

The data collected in this study enabled setting forth a comprehensive set of lysin-associated domains and families. The results of the conducted experiments indicate that the selected domains and families enable more effective identification of lytic proteins compared to the previously developed motif collections. The application of the compiled domain set facilitated the identification of lytic proteins from environmental and patient-derived

metagenomic data. Therefore this set may be used as a supplement to the developed machine learning classifier.

After *in silico* evaluation, both the predictions of the classifier and hits of the selected models were validated *in vitro*. To do this, selected putative lysins were produced and their enzymatic activity was assessed by zymography. These experiments led to the successful production and activity confirmation of a virion-associated lysin that digests the peptidoglycan of *Proteus mirabilis*. To the best of knowledge, this is the first virion-associated lytic protein identified up to date from a bacteriophage infecting *Proteus mirabilis,* an opportunistic pathogen. Additionally, the selected model set enabled the identification from patient-derived samples of a novel lytic enzyme that targets *Rothia*, as well as several lysins from animal-associated metagenomic samples that target *Enterococcus.*

To summarize, the result of this work is a machine learning algorithm that enables the rapid search for new bacteriophage lytic proteins. These proteins may find application in medicine and the industry and may contribute to the development of genomics bioinformatics in the future.