Machine-Learning, Natural Language Processing, and Genomics approaches for type III effectors identification

By Naama Wagner

Under the supervision of Prof. Tal Pupko

A wide range of Gram-negative bacteria use the type III secretion system (T3SS) to inject dozens of type III effectors (T3Es) into their eukaryotic host. These effectors alter molecular processes within the host for the benefit of the bacteria, leading to disease.

T3Es are highly diverse in function and sequence and their repertoire varies between different species and strains. This fact makes the task of unravelling the full list of T3E coding genes within a given bacterial genome complex.

To tackle this problem, I applied machine-learning classification algorithm using a set of informative features extracted for all the protein coding genes in the genome, derived from the genomic sequences. In this approach, the classification algorithms are trained on a set of known T3Es and non-T3Es from within the genome and are then applied to predict for each gene its likelihood to encode a T3E. I applied it to the mouse pathogen *Citrobacter rodentium* and following my prediction two novel effectors were validated experimentally. These results were included in a paper published in *Science*.

I then generalized and implemented this pipeline within a web server I developed, called *Effectidor*. Effectidor is a freely available user-friendly web server, for prediction of T3Es within bacterial genome input. It is the most accurate web server for this purpose existing today. A paper describing it was published in *Bioinformatics*.

To improve Effectidor's predictions and enhance our biological understanding of the T3Es recognition for injection, I decided to focus on the type III secretion signal. The secretion signal is known to be encoded within the N-terminal region of the protein, but it is non-conserved. We used a Natural Language Processing model that was developed and pre-trained by Facebook on all the available protein multiple sequence alignments to model the secretion signal of T3Es. A classifier trained on a non-redundant set of effectors and non-effectors from various species modeled by this model was successfully used to predict the existence of a type III secretion signal in new sequences. It was included in Effectidor and improved its already leading accuracy. A paper describing this work was published in *Frontiers in Plant Science.*

Lastly, I focused on two plant pathogens, *Xanthomonas fragariae* and *Xanthomonas hortorum pv. pelargonii*. I assembled the genome of the latter after full genome sequencing and used Effectidor to identify possible novel effectors within it and within *X. fragariae* genome. Highly scoring samples with no significant sequence similarity to previously validated T3Es were carried on for experimental validation, which resulted with four and two novel T3Es in *X. hortorum* and *X. fragariae*, respectively. A paper describing this work was published in *Frontiers in Plant Science*.