Machine-learning models able to predict phage-bacteria interactions

Diogo Leite[1], Grégory Resch[2], Yok-Ai Que[3], Xavier Brochet[1], Carlos Peña[1]

[1]*School of Business and Engineering Vaud (HEIG-VD), University of Applied Sciences Western Switzerland (HES-SO), Swiss Institute of Bioinformatics (SIB), Switzerland*

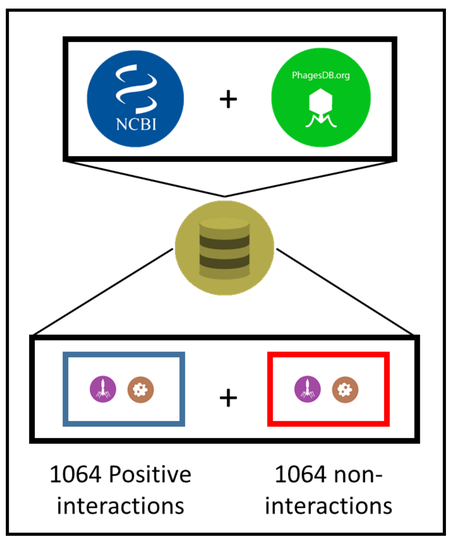
[2]*Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland*

[3]*Department of Intensive Care Medicine, Bern University Hospital (Inselspital), Bern, Switzerland*

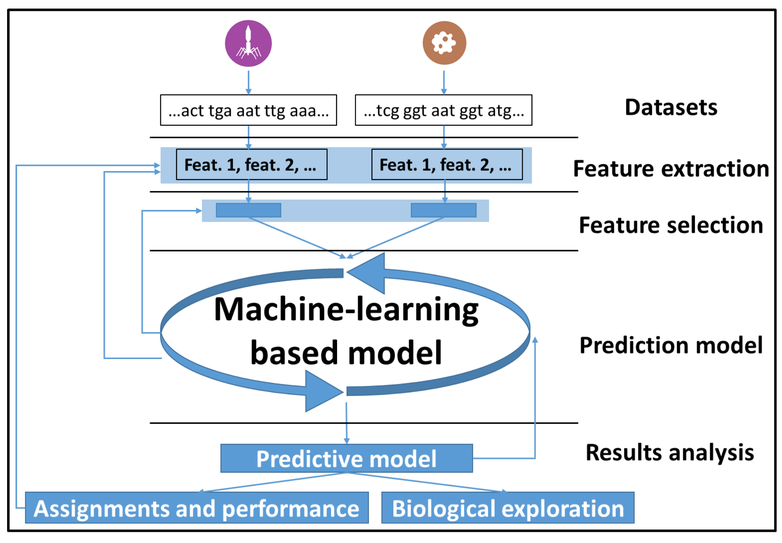
Abstract

Phage-therapy, a promising alternative to antibiotic-resistance, uses phages to infect and kill pathogenic bacteria. It requires finding perfectly matching phage-bacterium pairs, a time and money-consuming task, currently achieved empirically in laboratory. Our project aims at improving this task by predicting, *in-silico,* if a given phage-bacterium pair would interact. Predictions are performed on the base of public genomic data combined with machine-learning algorithms. With such an approach we have obtained around 90% of predictive power. In order to improve these results, we will extend our methodology and we will validate it with newly-generated clinically-relevant data.

Overview: prediction of phage-bacteria interactions



Data acquisition



Project overview

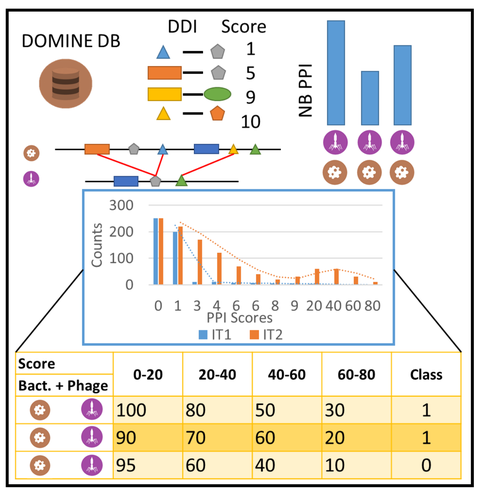
For our project we:

- Acquired data from public databases as NCBI [1] and phagesdb.org [2]. From GeneMarkS [3] we predict genes in bacterial and phages genomes.
- Constituted a positive dataset with 1064 phage-bacteria interactions pairs.
- Developed a model able to predict phage-bacteria interactions

Feature engineering: obtention of informative features

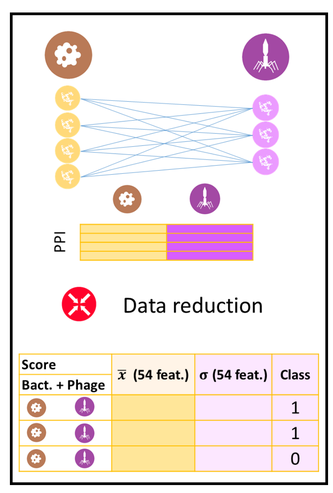
On our data we:

- Extracted two kinds of features based on:
  - Protein interactions (PFAM Domain-domain interactions) - 18 datasets
  - Genomic sequences (% of amino acids, chemical components and weight) - 1 dataset
- Corrected the over-representation of the Mycobacterium smegmatis mc2 155 bacterium, reducing it from 86% to 14%



Score Bact. + Phage	0-20	20-40	40-60	60-80	Class
	100	80	50	30	1
	90	70	60	20	1
	95	60	40	10	0

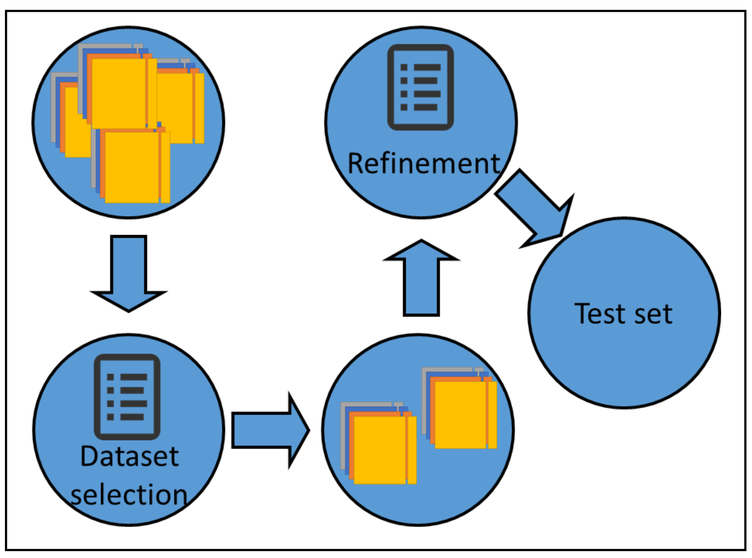
Domain-domain scoring



Score Bact. + Phage	$\bar{x}$ (54 feat.)	σ (54 feat.)	Class
			1
			1
			0

Sequence quantification

Machine-Learning based modeling



Process of research used

In our machine-learning search we used:

- 19 datasets (12'918 samples)
- Predictive model building with 10-fold cross-validation
- Four approaches: K-Nearest Neighbors (kNN), Random Forests (RF), Support Vector Machines (SVM), and Artificial Neural Networks (ANN)

In order to identify the best datasets and configurations of algorithms, we performed the following:

1. Selected the datasets exhibiting the highest scores across the four approaches
2. Further modelling to refine the algorithms' hyperparameters

Main results and conclusions

The best results on the test, obtained by ANN with 9 neurons and 50 epochs are presented below:

Datasets	Accuracy	F-Score	Sensitivity	Specificity
NB50	89.78%	90.13%	89.56%	90.12%
NBN50	89.79%	90.13%	89.56%	90.12%
S1e^{-6}	85.79%	86.24%	85.43%	86.35%

Future work:

- Explore other types of features to further increase model performance.
- Improve the predicitvity of the dataset by prunning redundant/ correlated variables that may perturb modeling
- Search for new relevant interactions allowing to predict interactions for different strains of a given bacterial species
- Improve domain interactions scores

References:

[1] http://www.ncbi.nlm.nih.gov/; [2] http://phagesdb.org/; [3] http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi [Features extraction based on article by E. Coelho] Computational prediction of the human-microbial oral interactome

Found by: FNSNF