Computational prediction of inter-species relationships through omics data analysis and machine learning

Diogo Leite[1,2], Grégory Resch[3], Yok-Ai Que[4], Carlos Peña[1], Aitana Neves[1] and Xavier Brochet[1]

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

[2]*School of Engineering of Polytechnic Institute of Porto (ISEP), Porto, Portugal*

[3]*Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland.*

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

Abstract

A main challenge for phage therapy consists in finding the matching phage for a given bacteria. This is currently achieved empirically in the laboratory, which is time-consuming and expensive. The new approach developed here consists in using machine-learning to predict if a phage can infect a bacteria or not (i.e. if they interact). Using publicly multi-omics available data, we obtained more than 90% of accuracy, f-measure, speficity ;and sensitivity. Future work will address improvement of the methods and its validation with clinically relevant data that will be generated.

Overview

Phage-therapy is a promising alternative to antibiotics and may contribute to reduce the burden of antibiotics-resistance, which arises mainly because of their over-use in hospitals, agriculture ;and animals (Figure 1). If no solutions are found, the WHO predicts that, by 2050, a simple flu could have strong complications by 2050 (Figure 2). In phage therapy the treatment consists in using viruses (phages) to kill the pathogenic bacteria (Figure 3).

Figure 1. Source of consumption

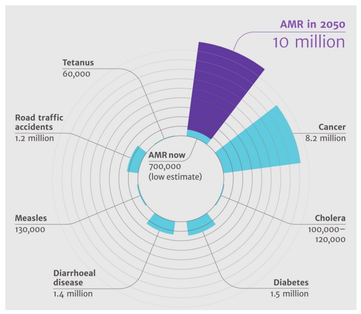

Figure 2. Data from World Health Organization [1]

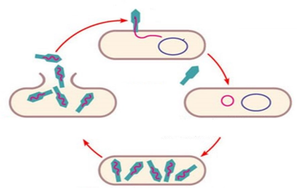

Figure 3. Cycle of infection of a Lytic phage

The project has three steps: (1) collecting data and create the datasets; (2) training and evaluating a machine-learning model and (3), in future, elaboraition of a network phage-bacteria (Figure 4).

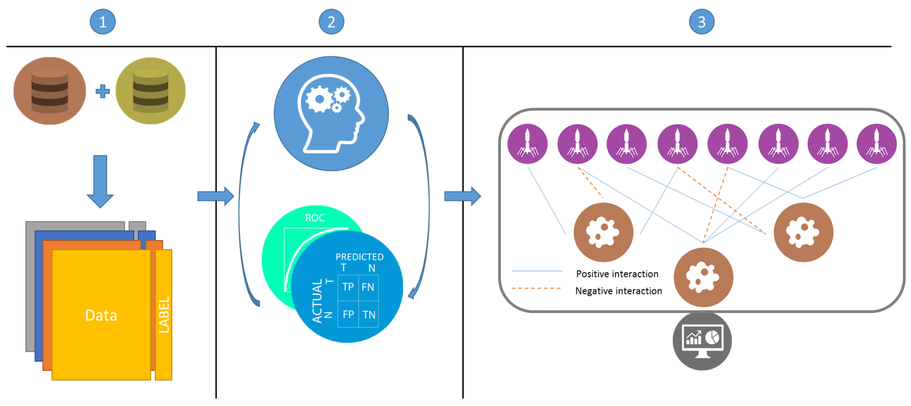

Figure 4. Overview of the project

Data acquisition

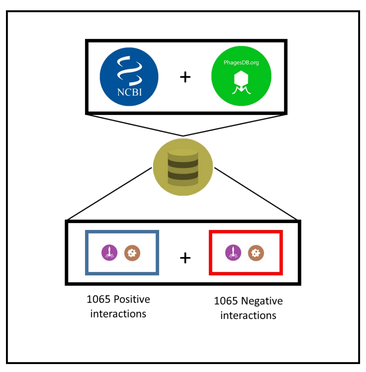

Figure 5. Overview of the data acquisition

To acquire data we (Figure 5):

- Collect all publicly available phage genomes from NCBI [2] and Phagesdb.org [3];
- If the reported host bacteria is sequenced (NCBI [2]), retrieve its genome as well;
- Predict genes in bacterial and phage genomes using GeneMarkS [4], if necessary;
- Constitute a positive dataset of 1065 entries. A negative dataset of equal size was also built artificially by matching phages with bacteria from a different species than their known host.

Feature extraction

Predict domains using HMMER-PFAM [5] for all proteins of all phages and bacteria in the database (Figure 6)

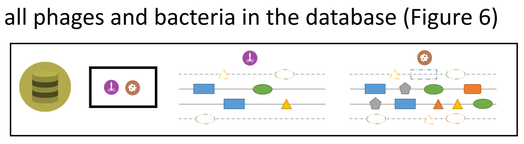

Figure 6. Domain acquisition

Identify interacting domains (DOMINE [6]) (Figures 7 & 8)

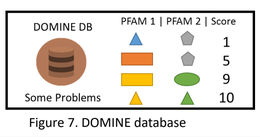

Figure 7. DOMINE database

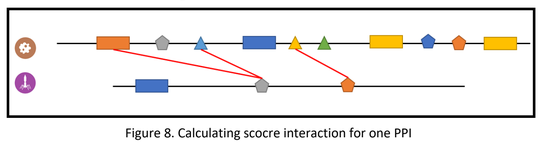

Figure 8. Calculating scocre interaction for one PPI

Build data set from PPI scores: use histogram of scores to define a binning and get the same number of features for each pair of phage-bacteria (Figure 9)

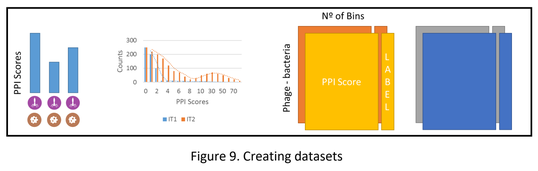

Figure 9. Creating datasets

Try the combination of all proteins between each bacteria and phage (Figure 10)

- Extract % of amino acid, % chemical components (C,H,O,N,S) and weight -> 27 features;
- Concatenate both vectors to get a PPI -> 27 + 27 = 54 features;
- For a given pair phage-bacteria, obtain a matrix of size -> Nº PPI × 54;
- Use PCA reduction to limit the number of features per phage-bacteria to ≈ 100 features.

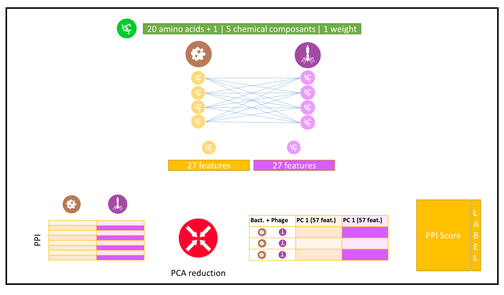

Figure 10. Feature based in the protein composition

Machine-Learning model

Our developed machine-learning model use:

- 19 datasets (Figure 11)
  - 10% of data use to create a test set
  - 90% of data use to create a training to perform 10-fold cross-validation
- Cross-validation with 10 folds
- Ensemble-learning approach with k-Nearest Neighbors (kNN), Random Forest (RF), Support Vector Machines (SVM) ; and Artifical Neural Networks (ANN)

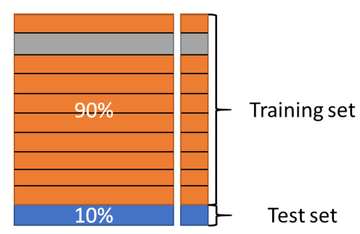

Figure 11. elaboration of the test and training set

In order to identify the best dataset and configurations of algorithms, we performed the following (Figure 12)

1. Find the datasets with the highest score
2. Refine the parameters
3. Build the final model based on a voting system (Figure 13) and assess its performance on the test set with unseen data

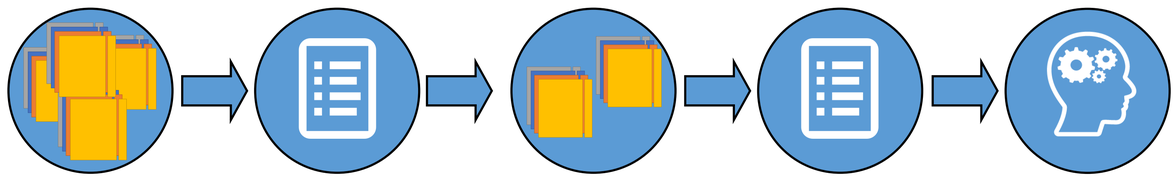

Figure 12. 19 datasets– first test phase– select better dataset—second test phase—refine the best configurations—develop the model

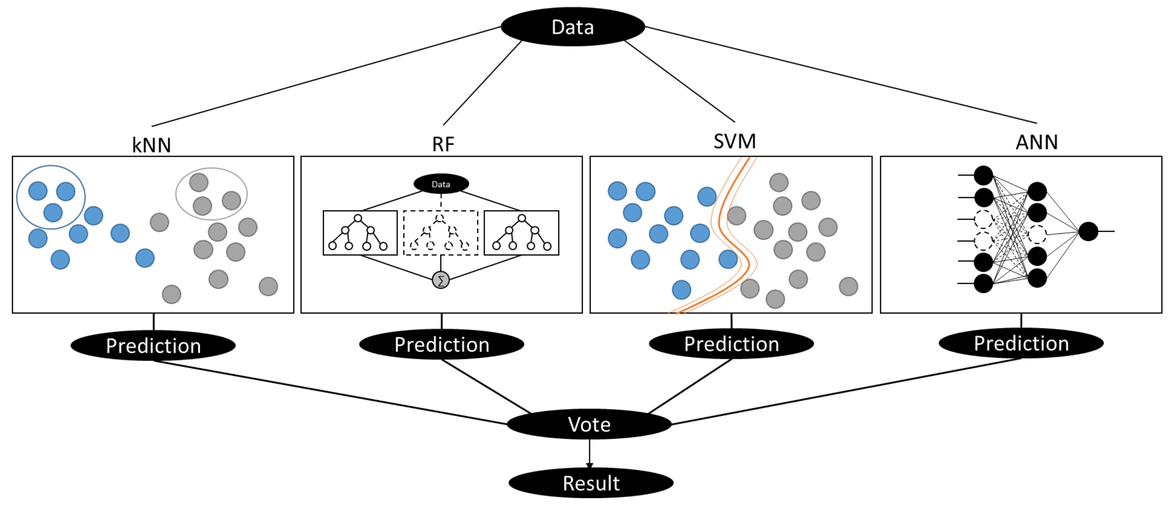

Figure 13. Ensemble learning machine learning developed

Main results

We present below the results for each bacteria in the validation (Figure 14) and test (Figure 15) sets. The table represents the prediction results obtained during the validation and test phases (Table. 1)

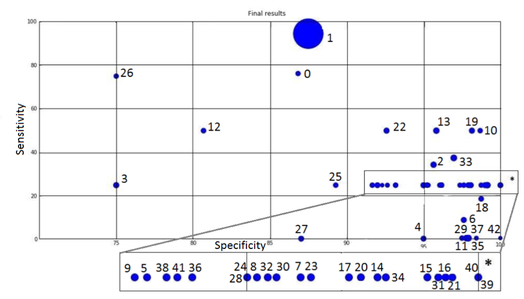

Figure 14. Results for Cross-validation by bacteria in validation sets

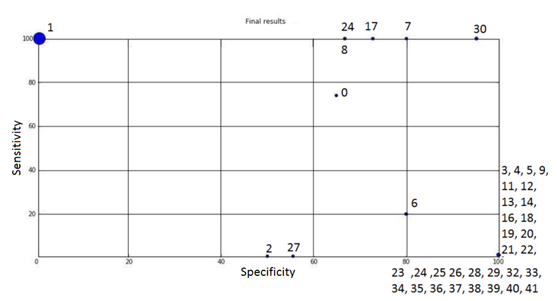

Figure 15. Results for Cross-validation by bacteria in test sets

	Accuracy	F1	Precision	Sensitivity	Specificity
Validation	91%	91%	94%	88%	95%
Test	88%	87%	85%	90%	86%

Table 1. Results in general model for validation and test sets

Conclusions and future work

These promising results demonstrate that our computational approach based on machine- learning is able to predict phage-bacteria interactions.

Future work:

- Improve the voting system of the ensemble-learning algorithm to better account for the performance of each individual algorithm;
- Preprocess more carefully each dataset:
  - Remove the redundant/correlated variables;
  - Remove those variables that provide little information gain;
- Use more data, and in particular, test the model on more clinically relevant data consisting of different strains of the same bacterial species.
- Investigate other types of features to yet increase model performance.
- Improve domain interaction scores, e.g. by further developing existing databases like DOMINE.

References:

[1] Antimicrobial Resistance : Tackling a crisis for the health and wealth of nations; [2] http://www.ncbi.nlm.nih.gov/; [3] http://phagesdb.org/; [4]http://www.ncbi.nlm.nih.gov/genomes/MICROBES/genemark.cgi; [5] http://www.ebi.ac.uk/Tools/hmmer/search/hmmscan ; [6] http://domine.utdallas.edu/cgi-bin/Domine; [Features extraction based on article by E. Coelho] Computational prediction of the human-microbial oral interactome