Exploration of a large spectrum of machine-learning techniques to predict phage-bacterium interactions

Diogo Leite[1], Grégory Resch[4], Yok-Ai Que[3], Xavier Brochet[1], Miguel Barreto[1], and Carlos Peña[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]*Department of Fundamental Microbiology, University of Lausanne, Lausanne, Switzerland.*

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

Abstract

Antibiotic resistance threatens the efficacy of currently-used medical treatments and call for novel, innovative approaches to manage multi-drug resistant infections. Phage therapy use viruses (phages) to specifically infect and kill bacteria during their life cycle. Currently, there is no method to predict phage-bacterium interactions, and these pairs must be empirically tested in laboratory, a costly process in terms of time and money. To overcome such situation, we are currently exploring several computational approaches intended at predicting if a given phage-bacterium pair may interact reducing, thus, the number of required in vivo experiments.

Data aquisition and management

1—Data

Public databases

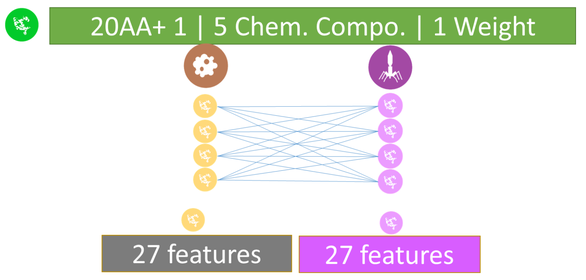
In vivo experiments

Our data:

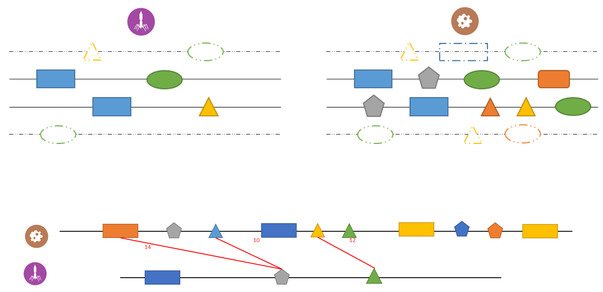
- 2’028 bacteria
- 3’810 phages
- 2’301 positive interactions
- 295 positive interactions (in vivo)
- 132 negative interactions (in vivo)

2—Feature engineering

Chemical composition based



Protein-protein interactions based

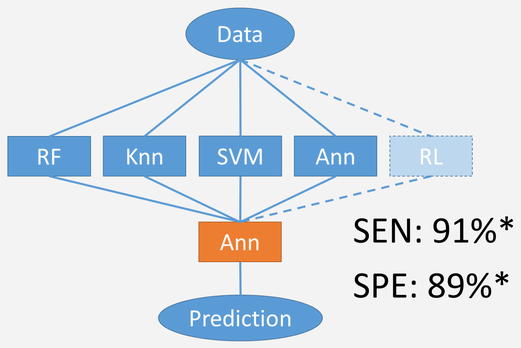


Machine-Learning approches

A—Ensemble-learning

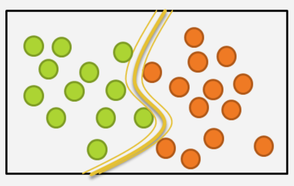
Creation of a stacking approach composed by an odd number of supervised machine-learning models plus one meta-learner model that receive the results of the other models to make its prediction.

*Previous work



SEN: 91%*

SPE: 89%*

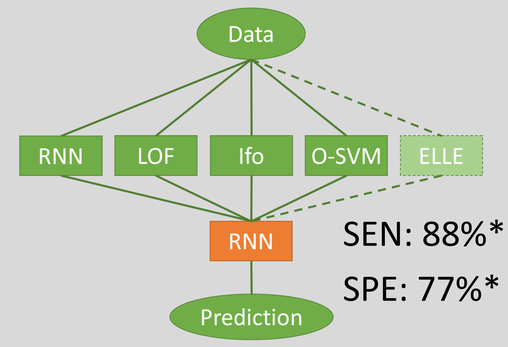


B—One-class learning

Utilization of models which can learn only with one class. We develop two workflows:

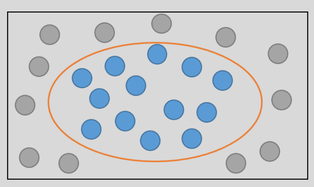
- Predict the interactions
- Validate our negative set

*Previous work



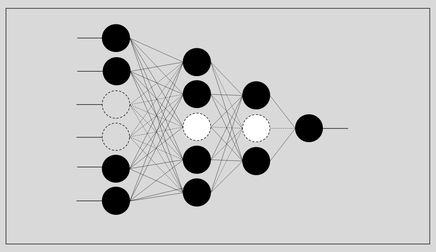
SEN: 88%*

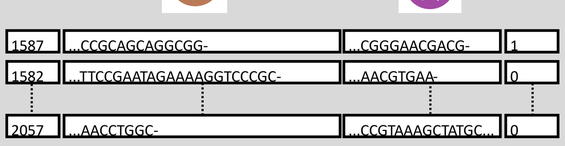
SPE: 77%*



C—Deep-Learning: Recurrent neural Networks (RNN)

Application of deep-learning directly on the sequence using RNN.

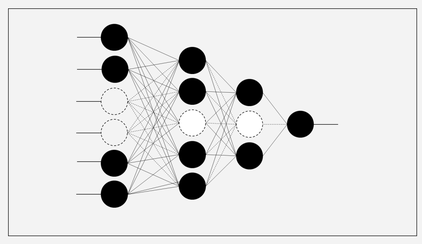


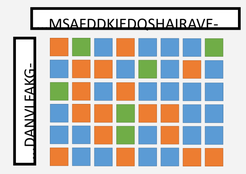


On going

D—Deep-learning: Convolutional Neural Network (CNN)

Transformation of the features extracted into image that can be analyzed by a CNN.





On going

Conclusions and future work

Conclusions

- These approaches use different phage-bacterium representation to train machine-learning models e.g.: from extracted features, complete genome, and informative images. This allow us to analyze and detect which are the most performant techniques to predict phage-bacterium interactions;
- We have obtained 87% of sensitivity and 56% of specificity for the one-class learning approach which indicates that is a good a path to follow (see poster N°42).

Future work

- Perform hyperparameter search for all approaches with different datasets configuration;
- Increase our database with new organisms and interactions to allow us predict at the strain level;
- Test the models with data extract from *in vivo* experiments ;
- Analyze and determine new ways to transform the genome information into informative images.

Found by: