Published in *Proceedings of 2020 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 16-19 December 2020, Seoul, South Korea, which should be cited to refer to this work.
DOI: 10.1109/BIBM49941.2020.9313487

Towards BacterioPhage Genetic Edition:

Deep Learning Prediction of Phage-Bacterium Interactions

Shabnam Ataee
HEIG-VD, HES-SO, SIB
Yverdon-les-bains, Switzerland
shabnam.ataee@heig-vd.ch

Óscar Rodríguez
HEIG-VD, HES-SO
Yverdon-les-bains, Switzerland
oscar.rodriguezzalona@heig-vd.ch

Xavier Brochet
HEIG-VD, HES-SO, SIB
Yverdon-les-bains, Switzerland
xavier.brochet@heig-vd.ch

Carlos Andrés Pena
HEIG-VD, HES-SO, SIB
Yverdon-les-bains, Switzerland
carlos.pena@heig-vd.ch

***Abstract*—In this paper, a novel approach is proposed for genetically engineering bacteriophages. It is formed of two main modules: *a predictor* and *a genome sequence generator*. *Convolutional Neural Networks* are used to build the predictor while the generator is constructed based on *Deep Generative Models*. This paper concentrates in the architecture and the results for the predictor module. The evaluation results suggest that the proposed model has the potential to be further used to guide genetic edition of phages so as to improve phage therapy against bacterial infections.**

Keywords—Phage Therapy, Deep Learning, Convolutional Neural Networks (CNN), Deep Generative Models (DGM)

I. Introduction

Antimicrobial resistance can lead to difficulties or even the impossibility to treat some infections. This serious situation among other things has motivated a renewed interest in Phage Therapy (PT). In PT, bacteriophages (viruses) are used to attack infection-causing bacteria as an alternative or complementary approach to antibiotics for treating bacterial infections.

Basically, as PT uses natural phages against bacterial infections, it may exhibit limitations such as narrow host range or inability of a single phage to treat infections caused by several bacteria [1-2]. The available phages are not always sufficient to find a treatment, especially in the absence of adequate lytic phages. One forward-thinking modernization of PT is to use genetically engineered phages which could provide substantial advantages in terms of host range, immune system recognition, and environmental stability [3]. To achieve this goal, we propose *PERPHECT (Deep Generative Models for Phage Genetic Edition)*, a novel approach for genetically engineering bacteriophages so as to improve their therapeutical value. This approach relies on two machine-learning-driven phases: A) predicting interactions between bacteria and phages and B) generating novel phage genome sequences. This paper concentrates on phase A, presenting a novel model to predict interactions between bacteria and phages.

II. Model Architecture

The proposed PERPHECT architecture is formed of two fundamental components: The *Phage-Bacterium Interaction Predictor* and the *Phage Genome Sequence Generator* as illustrated by Figure 1.

A. Phage-Bacterium Interaction Predictor

The close evolutionary relationship between phages and bacteria hosts entails that their genetic information can be used to predict their interaction. This fundamental component is used to predict the potential interaction between a bacterium and a bacteriophage based solely on their genome sequences. To address this classification problem, a Deep Learning (DL) model composed of a stack of 1-D Convolutional Neural Networks (1-D CNN) is used to build a predictor, as shown in Figure 2.

The proposed predictor architecture has a non-linear network topology. The two inputs (bacteria genome sequences and phages genome sequences) are processed separately by two parallel convolutional branches whose outputs are then merged together and passed through two subsequential dense layers. A *dropout layer* is also used to *reduce overfitting* and to *improve the generalization* of the proposed deep neural network. This multi-input and non-sequential architecture is implemented using *Keras functional API* [4]. The *sigmoid* is used as the last-layer activation and *binary cross-entropy* is used as the loss function when optimizing this binary

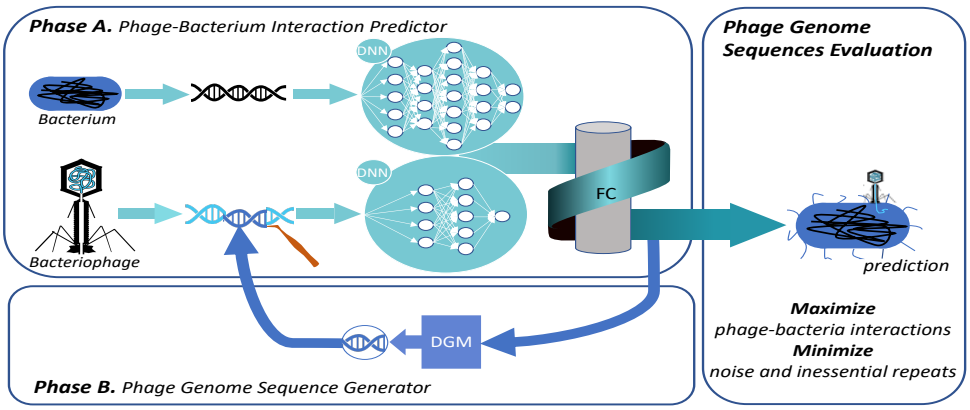

Figure 1. PERPHECT Model Architecture

classifier. In the next section, we evaluate the performance of the proposed predictor.

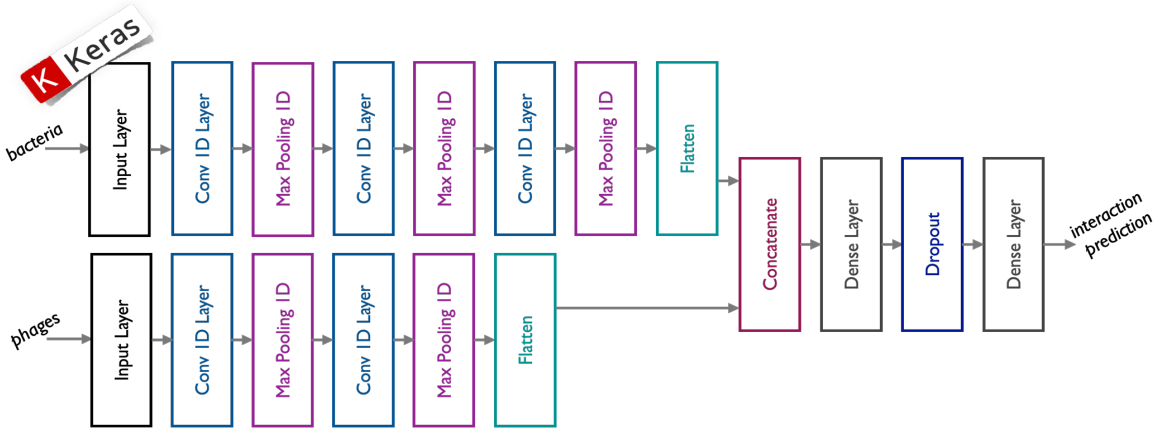

Figure 2. Predictor Model Architecture

B. *Phage Genome Sequence Generator*

This component generates phage genome sequences that maximize the interaction with a target bacterium while minimizing potentially perturbing length-associated effects such as noise and repeats. It is built based on *Deep Generative Models,* able to generate new samples (phage genome sequences in this case) from the learning domain [4].

III. Data Sets & Data Preparation

Two different data sets are used to train, validate and evaluate the predictor model: *A public data set* and *a private data set*. The public data set was created from public databases such as *PhageDB* [5] and *GenBank* [6] and it is described in [7]. After data preparation, i.e., removing missing values and dealing with duplicated values, the public data set is composed of 94 bacteria and 3121 phages for a total of 4202 reported phage-bacterium interactions, while the private data set includes 133 bacteria, 87 phages, and 3518 interactions.

Then, based on results from Exploratory Data Analysis in terms of distribution of sequence lengths for both bacteria and phages available in our data sets, fixed sequence lengths are defined and set to 7M bases for bacteria and 200K bases for phages. Some few longer genome sequences are cut from the end while shorter sequences are padded with zeros at the end of genome sequence (*zero padding*). The fixed-length sequences are then transformed to binary representation based on nucleic acid notations, as shown in Table I.

TABLE I. Binary Representation of Nucleotides

symbol	Binary Representation			
A	1	0	0	0
C	0	1	0	0
G	0	0	1	0
T	0	0	0	1
W	1	0	0	1
S	0	1	1	0
M	1	1	0	0
K	0	0	1	1
R	1	0	1	0
Y	0	1	0	1
B	0	1	1	1
D	1	0	1	1
N	1	1	1	1

Next, these two data sets are mixed and divided into a *train set* (70%) used for building models, a *validation set* (15%) used to tune algorithm's hyper-parameters, and a *test set* (15%) unused during training and used for final evaluation as described in the following section.

IV. Evaluation Results

This section presents evaluation results for the predictor model only. To evaluate the performance of the proposed classifier *accuracy*, *recall*, *precision*, *specificity* and *f1-score* are used as evaluation metrics.

Figure 3 shows the confusion matrix of the classification results obtained by the predictor on the test set. It can be

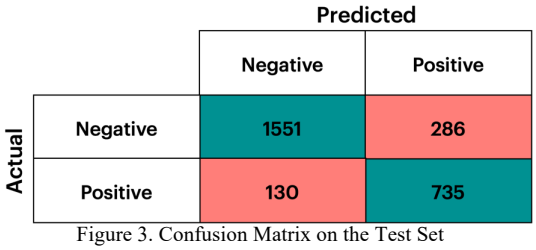

Figure 3. Confusion Matrix on the Test Set

noticed that most cases are correctly classified and among the misclassifications, it exhibits a higher amount (and proportion) of false positives than false negatives. In summary, these figures correspond to 85% accuracy, 85% recall, 72% precision, 84% specificity and 78% f1-score on the test set.

Table II presents the classification performance obtained by our predictor on both validation and test sets. It attains almost the same results in terms of the different evaluation metrics in both validation and test sets, suggesting a very good generalization. In general, these performances confirm the potential of the deep learning predictor to be further used for guiding genetic edition of phages so as to improve their therapeutical power against bacterial infections.

TABLE II. Evaluation Results on Validation Set vs. Test Set

Metric	Validation Results	Test Results
Accuracy	86%	85%
Recall	86%	85%
Precision	74%	72%
Specificity	85%	84%
F1-score	79%	78%

REFERENCES

[1] S. Matsuzaki, J. Uchiyama, I. Takemura-Uchiyama, and M. Daibata, "Perspective: The age of the phage", Nature, 2014, 509:S9–S9.

[2] A.S. Nilsson, "Phage therapy-constraints and possibilities", Upsala Journal of Medical Sciences, 2014, 119:192–198.

[3] D.P. Pires, S. Cleto, S. Sillankorva, J. Azeredo, and T. K. Lu, "Genetically Engineered Phages: a Review of Advances over the Last Decade", Microbiology & Molecular Biology Reviews, 2014.

[4] F . Chollet, "Deep Learning with Python", published in 2018.

[5] D.A. Russell, and G. F. Hatfull, "PhagesDB: the actinobacteriophage database", *Bioinformatics*, 2017, Volume 33, Issue 5, Pages 784–786.

[6] E.W. Sayers, M. Cavanaugh, K. Clark, J. Ostell, K.D. Pruitt, and I. Karsch-Mizrachi, "GenBank. Nucleic Acids Research", 2020, Volume 48, Pages 84–86.

[7] D. M. Carvalho Leite, X. Brochet, G. Resch, Y. Que, A. Neves and C. Peña-Reyes, "Computational prediction of inter-species relationships through omics data analysis and machine learning", BMC Bioinformatics, 2018, 19(Suppl 14): 420