Predicting Promoters in Phage Genomes Using Machine Learning Models

Marta Sampaio, Miguel Rocha, Hugo Oliveira, and Oscar Dias(✉)

Centre of Biological Engineering, University of Minho, Braga, Portugal
{msampaio, odias}@ceb.uminho.pt, mrocha@di.uminho.pt,
hugooliveira@deb.uminho.pt

Abstract. The renewed interest in phages as antibacterial agents has led to the exponentially growing number of sequenced phage genomes. Therefore, the development of novel bioinformatics methods to automate and facilitate phage genome annotation is of utmost importance. The most difficult step of phage genome annotation is the identification of promoters. As the existing methods for predicting promoters are not well suited for phages, we used machine learning models for locating promoters in phage genomes. Several models were created, using different algorithms and datasets, which consisted of known phage promoter and non-promoter sequences. All models showed good performance, but the ANN model provided better results for the smaller dataset (92% of accuracy, 89% of precision and 87% of recall) and the SVM model returned better results for the larger dataset (93% of accuracy, 91% of precision and 80% of recall). Both models were applied to the genome of *Pseudomonas* phage phiPsa17 and were able to identify both types of promoters, host and phage, found in phage genomes.

Keywords: Machine learning · Genome analysis · Phages · Promoters

1 Introduction

Bacteriophages, or phages, are viruses that exclusively kill bacteria [1]. In the last decades, phages have been extensively studied and their genomic information has increased exponentially, mainly due to their therapeutic potential against bacterial infections, at a time when the rise of antibiotic resistance in pathogenic bacteria represents a serious health problem [2]. Thus, such abundance of data demands the development of bioinformatics methods to facilitate genome annotation. The main obstacle in genome annotation is the identification of promoters, which are specific DNA regions responsible for transcription initiation. Identification of promoters is difficult, because these are composed of short, non-conserved elements. However, it is crucial for understanding and characterising phage genetic regulatory networks, which may allow to design better phages with applications in biotechnology and medicine [3].

Promoters are poorly described for phage genomes. Indeed, only the phiSITE database provides a list of identified promoters for 29 phage genomes [4]. Some phages

F. Fdez-Riverola et al. (Eds.): PACBB 2019, AISC 1005, pp. 105–112, 2020.
https://doi.org/10.1007/978-3-030-23873-5_13

are able of encoding their own RNA polymerase (RNAP). Hence, besides host promoters, which are recognized by the host's RNAP, the genome of these phages contains promoters that are recognized by their intrinsic RNAP [5].

A few general-purpose promoter prediction tools for bacterial genomes have been developed, using diverse computational algorithms. The most recent tools are based on machine learning models, such as CNNpromoter_b, using deep learning networks [6], BPROM, using linear discriminant analysis (LDA) [7], and bTSS finder, using artificial neural networks (ANN) [8]. However, these tools still return numerous false positives [9]. Such tools were developed using bacterial promoters and only search for the typical bacterial motifs of the −35 and −10 elements (TTGACA and TATAAT, respectively), thus not being suitable for phages genomes. Therefore, these are not able to find promoters recognized by phage own RNAP nor host promoters with different motifs. Other tools, such as the widely-used PromoterHunter available on the phiSITE website, have additional disadvantages like requiring the weight matrices of the two promoter elements as input and limiting the size of the input genome sequence [4]. For phages, only PHIRE software was developed for predicting regulatory elements [10]. However, it only searches for conserved sequences with 20 base pairs or more and requires installation.

Therefore, in this work, machine learning models were trained using phage sequences for identifying both types of promoters found in phage genomes and different motifs of each promoter type.

2 Methods

2.1 Data

The positive data used to train the models was retrieved from the phiSITE database and available publications, consisting of 800 promoter sequences from 53 phage genomes. Since there are no sequences identified as non-promoters, sequence fragments of 65 base pairs were randomly selected from the 53 genomes to form the negative sets, provided that the selected fragment did not include a known promoter. There is no consensus length for promoters, so the fragment size was chosen according to the size of the biggest collected promoter.

As the number of promoters in the whole genome is several orders of magnitude lower than the non-promoters, having more negative than positive cases in the dataset should be more adequate for finding promoters in phage genomes. Therefore, two datasets were created, both comprising the 800 positive cases, though with different negative sets: Dataset1 including 1600 negative sequences and Dataset2 including 2400 negatives.

2.2 Features

Twelve different motifs were previously found in the collected phage promoters, using motif finder tools like MEME [11]. Eight of these motifs represent the elements of host promoters and four are recognized by the intrinsic RNAP (Table 1). Data features included the sizes and scores of these motifs, which were calculated using Position Specific Scoring Matrices (PSSM) with pseudocounts of 1, as well as information about phage lifecycle, family and host. The free energy value and the frequency of adenines and thymines were also calculated for each sequence, as these express the stability of the DNA molecule, which is expected to be lower in the promoter region. The free energy value was calculated by summing the unified nearest-neighbor (NN) energy values of each dinucleotide that were defined by SantaLucia et al. [12]. The datasets were standardized and the recursive feature elimination (RFE) method was used to select the most relevant features of the datasets. RFE was applied using Random Forests as estimator and removing one feature at each iteration. After applying this method, some features representing the host and motif sizes were eliminated from the datasets. The final datasets are available at: https://github.com/martaS95/PhagePromoter/Data.

Table 1. Description of the motifs identified in the collected data. In the consensus sequence, Y = C or T; M = A or C; R = A or G; W = A or T; N = A, C, G or T.

Type	Element	Size (bp)	Phages	Consensus sequence
Host	−10	6	Almost all (51)	TATAAT
Host	−35	6	Almost all (50)	TTGACA
Host	−10	8	T4 e CBB	TATAAATA
Host	−35	7	T4	GTTTACA
Host	−35	7	CBB	TGAAACG
Host	−35	9	T4	AWTGCTTTA
Host	−35	14	Lambda-like	TTGCN$_6$TTGC
Host	−35	14	Mu-like	CCATAACCCCGGAC
Phage	None	23	T7-like	TAATAAGACTCACTAAAGGGAGA
Phage	None	21	phi-C31	CCGGGTTGCCGACTCCCTTMC
Phage	None	27	phiKMV-like	CGACCCTGCCCTACTCCGGGCTYAAAT
Phage	None	32	KP34-like	AGCCTATAGCRTCCTACGGGGYGCTATGTGAA

2.3 Models

Machine learning models were built to classify sequence fragments as promoters or non-promoters, using four different models: artificial neural networks (ANN), support vector machines (SVM), random forests (RF) and k-nearest neighbors (KNN). For each algorithm, two models were trained with each dataset (Dataset1 or Dataset2), creating eight models. These models were optimized using the Grid Search method. Table 2 describes the tested values of model hyperparameters, which were selected empirically, and the best values obtained for the hyperparameters.

The models were further evaluated using cross-validation with 5-folds and the selected metrics were accuracy, precision and recall. Confusion matrices and Matthews correlation coefficients (MCCs) were also calculated for all models. These steps were performed using the Python library Scikit-learn [13].

Table 2. Hyperparameter values tested for each model, using Grid Search. The best values are highlighted by color: in red are the best values obtained using Dataset1 and in blue are the ones obtained using Dataset2. The values in green were the same for the two datasets.

Model	Parameter	Values tested for Grid Search
ANN	solver for weight optimization	lbfgs, sgd, **adam**
	activation function	identity, logistic, tanh, **relu**
	alpha	0.0001, **0.001**, **0.01**
	hidden layer size	**(15,)**,(25,),(50,),(75,),**(100,)**
SVM	C (regularization)	1,**2.26**,10,**15**,20
	gamma	**auto**,0.001,0.01, **0.05**, 0.1
	kernel	linear, **rbf**, poly, sigmoid
RF	number of trees in the forest	**20**, 40, 60, 80,**100**
	number of features	**auto**,2,3,6,10
	minimum number of samples to split an internal node	**2**, 3, **6**, 10
	minimum number of samples to be at a leaf node	**2**, 3, 6, 10
	maximum depth of the tree	2, 3, **None**
	bootstrap	**True**, **False**
	criterion	**gini**, **entropy**
KNN	Number of neighbors	3,5,7,**9**

3 Results and Discussion

The results of model evaluation are presented in Table 3. Globally, all models showed good results, as these presented high accuracy and precision and acceptable recall. The models trained with Dataset1 present higher recall while the models trained with Dataset2 have higher accuracy. The SVM and KNN models also show higher precision with Dataset2. For Dataset1, the model with best performance was the ANN model with 92% of accuracy, 89% of precision and 87% of recall. Although the RF model presented the highest precision (91%), it also had the lowest recall (83%). For Dataset2, both SVM and RF models presented the highest precision (91%) and the ANN model the highest recall (83%). The KNN model performed slightly worse for both datasets.

Since the differences between these metrics are not significant, confusion matrices were also generated to evaluate the performances of the models and MCCs were calculated from them. These results are represented in Tables 4, 5, 6 and 7.

As expected, the ANN model has the lowest number of false negatives (FN) and the highest number of false positives (FP) for both datasets. For Dataset1, the RF model has the lowest number of FP but the highest number of FN. For Dataset2, both RF and SVM models have the lowest number of FP, but the SVM has less FN than the RF model. Correlating these values using MCC, it is possible to see that for Dataset1, both SVM and ANN have the highest MCC value and SVM model has also the highest MCC value for Dataset2. Nevertheless, the small differences between the calculated MCCs indicate that all models present similar performances.

Table 3. Mean values of accuracy, precision and recall for each model after a 5-fold CV

	Dataset1			Dataset2		
Models	Accuracy	Precision	Recall	Accuracy	Precision	Recall
ANN	0.92	0.89	0.87	0.93	0.87	0.83
SVM	0.92	0.89	0.86	0.93	0.91	0.80
RF	0.92	0.91	0.83	0.93	0.91	0.79
KNN	0.91	0.89	0.84	0.92	0.90	0.78

Table 4. Confusion matrices of the ANN models for both datasets

Dataset1 (1600 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	694	106	800
Negative	87	1513	1600
Total	781	1619	2400
MCC	0.82		

Dataset2 (2400 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	661	139	800
Negative	94	2306	2400
Total	755	2445	3200
MCC	0.81		

Table 5. Confusion matrices of the SVM models for both datasets

Dataset1 (1600 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	685	115	800
Negative	75	1525	1600
Total	760	1640	2400
MCC	0.82		

Dataset2 (2400 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	641	159	800
Negative	68	2332	2400
Total	709	2491	3200
MCC	0.80		

Table 6. Confusion matrices of the RF models for both datasets

Dataset1 (1600 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	666	134	800
Negative	69	1531	1600
Total	735	1665	2400
MCC	0.81		

Dataset2 (2400 negatives)

Predicted Real	**Positive**	**Negative**	**Total**
Positive	628	172	800
Negative	68	2332	2400
Total	696	2504	3200
MCC	0.79		

Table 7. Confusion matrices of the KNN models for both datasets

Dataset1 (1600 negatives)

Predicted Real	Positive	Negative	Total
Positive	672	128	800
Negative	81	1519	1600
Total	753	1647	2400
MCC	0.80		

Dataset2 (2400 negatives)

Predicted Real	Positive	Negative	Total
Positive	627	173	800
Negative	69	2331	2400
Total	696	2504	3200
MCC	0.79		

Data from confusion matrices confirms that the number of FN is higher than the number of FP in all models, which might be explained by the fact that the negative sequences selected to train the model were putative and not proven to be negative. Thus, the negative cases set may encompass promoter sequences that have not yet been identified which can prompt the models to predict a promoter as negative. Another possible explanation for these results is that some known promoters have motif sequences very distinct from the consensus, thus their scores regarding these features will be low, inducing the model to classify them as negatives.

Although all models presented similar performance, only two were selected to be applied to the case study: the ANN model trained with Dataset1 and the SVM model trained with Dataset2.

3.1 Case Study: *Pseudomonas* Phage PhiPsa17

The two models were applied to the genome of *Pseudomonas* phage phiPsa17, to test their predictive capacity. This lytic phage belongs to the *Podoviridae* family and was extracted from *Pseudomonas syringae*. It encodes its own RNAP which means its early genes are transcribed by the host RNAP whereas middle and late gene are transcribed by phage intrinsic RNAP [14]. Thus, two types of promoters are expected to be found in its genome: host promoters, with the −10 and/or −35 elements, and phage promoters with sequence similar to those of T7-like virus. There are no promoters of this phage in phiSITE. Analysing the study of Frampton et al. [14], 1 host promoter was identified in the early region of the genome using BPROM tool [7] and 11 phage promoters were identified using MEME [11], considering only the 100 base pairs upstream of the predicted genes. As all predicted genes are in the direct strand, the models were only applied to the direct strand of this genome and searched the whole genome sequence for promoters. The results predicted by both models are summarized in the Venn diagrams of Fig. 1.

The SVM model predicted 16 promoters, 3 host and 13 phage promoters, while the ANN predicted 25 promoters, 8 host and 17 phage promoters. 14 promoters were identified by both models (2 host and 12 phage promoters) and 12 correspond to the promoters previously predicted by Frampton et al. [14]. The other promoters predicted by the models have lower scores and less common motifs for the −35 and −10 elements. The promoters predicted by both models are close to the predicted genes, except for the 2 host promoters. As expected, the ANN model predicted more promoters than

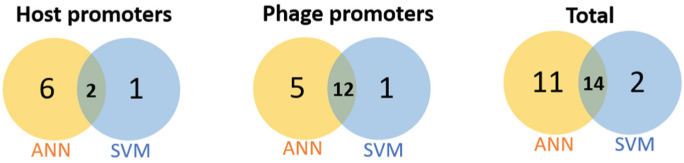

Fig. 1. Venn Diagrams representing the number of host and phage promoters predicted by SVM and ANN models for phiPsa17 phage genome

the SVM model. Thus, although a careful analysis of the results is required, these models presented good performance as both were able to identify all promoters predicted by Frampton et al. [14].

Comparing these results with the results of other tools, CNNpromoter_e predicted 31 host promoters for this genome, which is a higher number than expected for this genome since this phage only uses the host RNAP for transcribing early genes. None of these promoters were identified by the models. PromoterHunter tool only predicted three promoters which correspond to one host promoter predicted by both models and two predicted by the ANN model. Regarding phage promoters, PHIRE program identified 10 of the 11 phage promoters previously identified by MEME. Therefore, the developed models are better than these tools because they can recognize both promoter types, host and phage, and with different motifs of each, so there is no need to use different tools for identifying different promoters.

4 Conclusions

In this work, we propose methods to identify promoters in phage genomes, during genome annotation. Several machine learning models were trained with phage data, using two different datasets. All models showed good performance, but the ANN model provided better results for the smaller dataset whereas the SVM model returned better results for the larger dataset. The ANN model is expected to predict more promoters than the SVM model, so it can result in more false positives when applied to new data. On the other hand, the SVM predicts less promoters, so it may result in more false negatives.

The number of false negatives is higher than the number of false positives for all models, which might be explained by the high variety of phage promoter motifs and by the fact that the set of negative examples may encompass unidentified promoter sequences. In addition, the proportion between positive and negative cases in the datasets is much lower than the real proportion of promoters and non-promoters in a genome. The models identified phage promoters previously predicted by other tools and manually curated, but the number of phage genomes with identified promoters that were not used to train the models is very low. Therefore, having more phage promoter and non-promoter sequences experimentally identified is crucial to validate the models and improve promoter identification.

Nevertheless, as these models are the first to use phage data and to identify different motifs for both promoter types, they are undoubtedly useful for facilitating and speeding up the task of predicting promoters in phage genomes.

Acknowledgments. This study was supported by the Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2019 unit and the Project POCI-01-0145-FEDER-029628. This work was also supported by BioTecNorte operation (NORTE-01-0145-FEDER-000004) funded by the European Regional Development Fund under the scope of Norte2020 - Programa Operacional Regional do Norte.

References

1. Salmond, G.P.C., Fineran, P.C.: A century of the phage: past, present and future. Nat. Rev. Microbiol. **13**(12), 777–786 (2015)
2. Haq, I.U., Chaudhry, W.N., Akhtar, M.N., Andleeb, S., Qadri, I.: Bacteriophages and their implications on future biotechnology: a review. Virol. J. **9**(1), 9 (2012)
3. Guzina, J., Djordjevic, M.: Bioinformatics as a first-line approach for understanding bacteriophage transcription. Bacteriophage **5**(3), e1062588 (2015)
4. Klucar, L., Stano, M., Hajduk, M.: phiSITE: database of gene regulation in bacteriophages. Nucleic Acids Res. **38**(Database issue), D366–D370 (2010)
5. Yang, H., Ma, Y., Wang, Y., Yang, H., Shen, W., Chen, X.: Transcription regulation mechanisms of bacteriophages: recent advances and future prospects. Bioengineered **5**(5), 300–304 (2014)
6. Umarov, R.K., Solovyev, V.V.: Recognition of prokaryotic and eukaryotic promoters using convolutional deep learning neural networks. PLoS ONE **12**(2), e0171410 (2017). https://doi.org/10.1371/journal.pone.0171410
7. Solovyev, V., Salamov, A.: Automatic annotation of microbial genomes and metagenomic sequences, January 2016
8. Shahmuradov, I.A., Mohamad Razali, R., Bougouffa, S., Radovanovic, A., Bajic, V.B.: bTSSfinder: a novel tool for the prediction of promoters in cyanobacteria and Escherichia coli. Bioinformatics **33**(3), 334–340 (2017)
9. Silva, S., Echeverrigaray, S.: Bacterial promoter features description and their application on e. coli in silico prediction and recognition approaches. In: Bioinformatics inTech, November 2012
10. Lavigne, R., Sun, W., Volckaert, G.: PHIRE, a deterministic approach to reveal regulatory elements in bacteriophage genomes. Bioinformatics **20**(5), 629–635 (2004)
11. Bailey, T.L., Williams, N., Misleh, C., Li, W.W.: MEME: discovering and analyzing DNA and protein sequence motifs. Nucleic Acids Res. **34**, W369–W373 (2006)
12. SantaLucia, J.: A unified view of polymer, dumbbell, and oligonucleotide DNA nearest-neighbor thermodynamics. Proc. Natl. Acad. Sci. **95**(4), 1460–1465 (1998)
13. scikit-learn: machine learning in Python — scikit-learn 0.21.2 documentation. https://scikit-learn.org/stable/index.html
14. Frampton, R.A., Acedo, E.L., Young, V.L., Chen, D., Tong, B., Taylor, C., Easingwood, R. A., Pitman, A.R., Kleffmann, T., Bostina, M., Fineran, P.C.: Genome, proteome and structure of a T7-Like bacteriophage of the kiwifruit canker phytopathogen pseudomonas syringae pv. actinidiae. Viruses **7**(7), 3361–3379 (2015)