Performance of various protein representations for predicting phage-host interaction

Franz Stewart V. Dizon
Bioinformatics Lab
Advanced Research Institute for Informatics, Computing and Networking
De La Salle University
Manila, Philippines
franz_dizon@dlsu.edu.ph

Jennifer C. Ureta
Bioinformatics Lab
Advanced Research Institute for Informatics, Computing and Networking
De La Salle University
Manila, Philippines
jennifer.ureta@dlsu.edu.ph

Anish M.S. Shrestha
Bioinformatics Lab
Advanced Research Institute for Informatics, Computing and Networking
De La Salle University
Manila, Philippines
anish.shrestha@dlsu.edu.ph

ABSTRACT

With the rise of antimicrobial resistance that decreases the effectiveness of antibiotics in treating bacterial infections, phage therapy is being studied as an alternative to antibiotics. Phage therapy is the use of phages to treat bacterial infections by letting the phages infect and lyse the bacterial pathogen at the site of infection. Phages are known to be able to infect a narrow range of hosts only, but laboratory experiments to verify an interaction between a phage and a bacterium are both costly and time-consuming. To mitigate this, several studies have explored the use of machine learning classifiers to predict whether a phage-host pair interacts or not. In this study, we formulated the prediction problem as a binary classification problem with the host and phage proteomes as input, and explored different kinds of protein representations, including protein embeddings that are generated by protein language models, that can serve as an input to machine learning classifiers. In our experiments, under a phylogeny-based train-test data split, protein embeddings did not necessarily improve classifier performance compared to using the conventional k-mer profile representation.

KEYWORDS

Phage, anti-microbial resistance, phage-host pair interaction prediction, protein language models, machine learning classifier

1 INTRODUCTION

Antimicrobial resistance (AMR) threatens the use of antibiotics by decreasing its effectiveness in treating bacterial infections [19, 24], which results in longer duration of illness, higher rates of mortality, increased costs of treatment, and inability to perform procedures that rely on effective antibiotics to prevent infection [19]. In 2019 alone, there were an estimated 1.27 million deaths globally directly attributable to drug resistance [27]. Resistance arises due to the excessive use of antibiotics, which exerts selective pressure that allows microorganisms that have developed resistance, to have a competitive advantage to survive and proliferate [14, 19].

With the emergence of antibiotic-resistant strains of life threatening microbes [4], phage therapy is being studied as an alternative to antibiotics [11, 15]. Phages are viruses that are capable of infecting and replicating within bacterial cells [13]. Phage therapy is the use of phages to treat bacterial infection, which is done by letting the phages infect and lyse the bacterial pathogen at the site of infection [13, 23]. A specific strain of phage is known to be able to infect a narrow range of hosts only, which can be an advantage compared to a wider host range of antibiotics, that not only kills the target pathogenic bacteria but also other bacteria, some of which might be beneficial [13, 33].

Experimental methods used to identify phage-specific hosts require costly and time-consuming lab experiments to verify whether there is an interaction between the phage and the host [21]. To mitigate the cost and time consumed by lab experiments, computational approaches have been developed.

Recently, the use of machine learning classifiers for predicting phage-host interaction have been explored. One group of studies uses genomic information from only the phages. Young et al. [37] formulated the problem as taking the genome representation (DNA k-mer, amino acid k-mer, physio-chemical k-mer, protein domains) of the phage as input, and predicting a possible host. They trained a Support Vector Machine classifier separately for each host. Boeckaerts et al. [5] used only the phages' Receptor-Binding Proteins (RBP) as input. An RBP was represented by a vector composed of handcrafted DNA and protein features such as nucleotide frequencies, GC-content, codon frequencies, and others. The RBP vector representation served as the input to their multi-class classifier to identify the phage's possible hosts. Mark et al. [12] extended Boeckaerts et al.'s study by using Protein Language Models (PLM) to acquire vector representations, also called as protein embeddings, which serve as the input to their multi-class classifier.

There are also studies that formulate the phage-host interaction prediction problem as a binary classification problem that takes as input features derived from both the phage and the host, and predict whether the phage-host pair will interact or not. PredPHI [21] is a neural network classifier consisting of convolutional and fully connected layers. The input to PredPHI is a matrix that consists of derived protein features across the whole proteome of both the phage and the host, such as frequency of amino acids, abundance of each chemical component, and molecular weights. PHIAF [22], similar to PredPHI, incorporates an attention layer into its neural network, adds DNA-derived features for its input, and uses a Generative Adversarial Network (GAN) for data augmentation. PhageHostLearn [7] uses RBPs of phages and K-locus proteins as the input, and predicts whether the phage-host pair will interact or not. Embeddings for the protein sequence, obtained using the ESM-2 protein language model, served as an input to an XGBoost classifier.

In this study, we defined the phage-host prediction problem as taking the proteome (entire set of proteins) of both the phage and the host as the input, and deciding whether the phage-host pair interacts or not. We tested embeddings produced by different PLMs, namely ProtVec[2], Seq2Vec[18], ProtBert[10], and Prot5[10] to represent the protein sequences of both the host and the phage. We also tested non-PLM-based protein representations prior studies have used, such as k-mer profile and statistical profile[21], to determine if the use of protein embeddings does improve the prediction of phage-host interaction. In a departure from previous studies, we separated our training and test set according to the host's phylum classification instead of just randomly splitting the dataset, so that the host sequences seen on the test set are as unrelated/independent as possible from the one found on the training set. We found that, under this strict train-test data split, protein embeddings did not necessarily improve classifier performance compared to using the conventional k-mer profile representation.

2 MATERIALS AND METHODS

Shown in Figure 1 is the overview of the methodology used in this study.

2.1 Problem definition

We defined the phage-host prediction problem as taking the proteome (entire set of proteins) of both the phage and the host as input, and deciding whether the phage-host pair interacts or not. We tested various protein representations to determine the best protein representation to use as an input to classifiers when predicting phage-host interaction. We used several machine learning classifiers to predict whether the phage-host pair interacts or not.

2.2 Data collection

We collected 5,257 phage-host interaction records from VirusHostDB [25] consisting of 780 distinct hosts and 4,854 distinct phages. We also acquired 4,830 proteomes of phages from VirusHostDB [25] and 308 proteomes of bacteria from NCBI [31]. Since there are phage-host interaction records where either the phage or the host's proteome is not accessible online, we removed those phage-host interaction records in our dataset, which resulted in our final dataset containing 3,390 phage-host interaction records. We accessed VirusHostDB [25] last January 2023 and accessed NCBI [31] datasets last February 2023.

We also collected protein sequences from the Swiss-Prot database [9], containing 570,157 protein sequences, which we accessed last June 2023. We used it for training protein language models not available online such as ProtVec [2] and Seq2Vec [18].

2.3 Train-test split and negative pairs generation

Our test set consists of phage-host interaction pairs where the host's phylum classifications are *Actinobacteria*, *Cyanobacteria*, *Spirochaetes*, *Tenericutes*, *Bacteroidetes*, *Chlamydiae*, *Fusobacteria*, and *Deinococcus-Thermus*, while the remaining phage-host interaction pairs not found on the test set, where its host's phylum classification are *Firmicutes* and *Proteobacteria*, were included in the training set. The split according to the host's phylum classification was done so that hosts found on the training set are as unrelated as possible to the hosts found on the test set. After splitting the training and test set, negative pairs of phage-host were generated independently for both sets, where we randomly paired phage-host found in its respective set, and the pair had no previously recorded interactions. Lastly, the training set was split into 4 folds using sklearn's GroupKFold [29] and was grouped according to the host's order classification. The GroupKFold split is done to ensure that the same group is not seen in both training and validation sets. The overview of dataset composition is visualized in Figure 2. Looking at the train-test split at the species level, the test set contains 55 distinct species, while the training folds contain 38, 58, 51, and 53 distinct species respectively from fold 1 to fold 4.

2.4 Protein representations

In this study, we mainly explored the use of protein embeddings generated by protein language models to the phage-host interaction problem. The use of protein embeddings is motivated by the advancements of large language models for natural languages, which provide representations for words in the form of vectors, often called word embeddings [3]. Similar ideas have been recently implemented for biological sequences. [17]

Protein language models that provide protein embeddings have been recently applied to different bioinformatics problems [17]. These embeddings when applied to classification tasks, have been shown to provide better results compared to using hand-crafted features [34, 36].

2.4.1 k-mer profile. The k-mer profile of a protein sequence is a vector that contains the frequencies of all the possible substrings of length k found in the sequence. Since a proteome consists of multiple protein sequences, we calculated the summation of the k-mer profiles over all the protein sequences, and did this separately for the phages and the hosts. The value of k used in this study is set to 3.

2.4.2 Statistical profile. The PredPHI classifier built by Li et al. [21] proposed a proteome representation which extracted 27 features, where 21 of those are the frequency of amino acids (20 amino acids, 1 indicating unknown amino acid), 5 of those are the abundance of each chemical component in the sequences (carbon, hydrogen, oxygen, nitrogen, and sulfur), and the last 1 indicates the sum of molecular weights of all amino acids in the sequence. These 27 protein features across the entire proteome were combined using statistical measures such as mean, standard deviation, maximum, minimum, median, and variance. The final encoded features are in the form of a matrix with a shape of 6 x 27 x 2, where *6* denotes the six statistical measures (mean, std, max, min, median, var), *27* denotes the 27 protein features, and *2* represents the features for the phage and its host. For convenience, we termed this representation as *statistical profile* in this study. Lastly, we flattened the statistical profile into a vector to serve as an input to machine learning classifiers.

2.4.3 ProtVec. ProtVec [2] is a type of protein language model based on Word2Vec [26], which outputs a 100-dimensional vector embedding given a 3-mer of amino acids. To generate a ProtVec

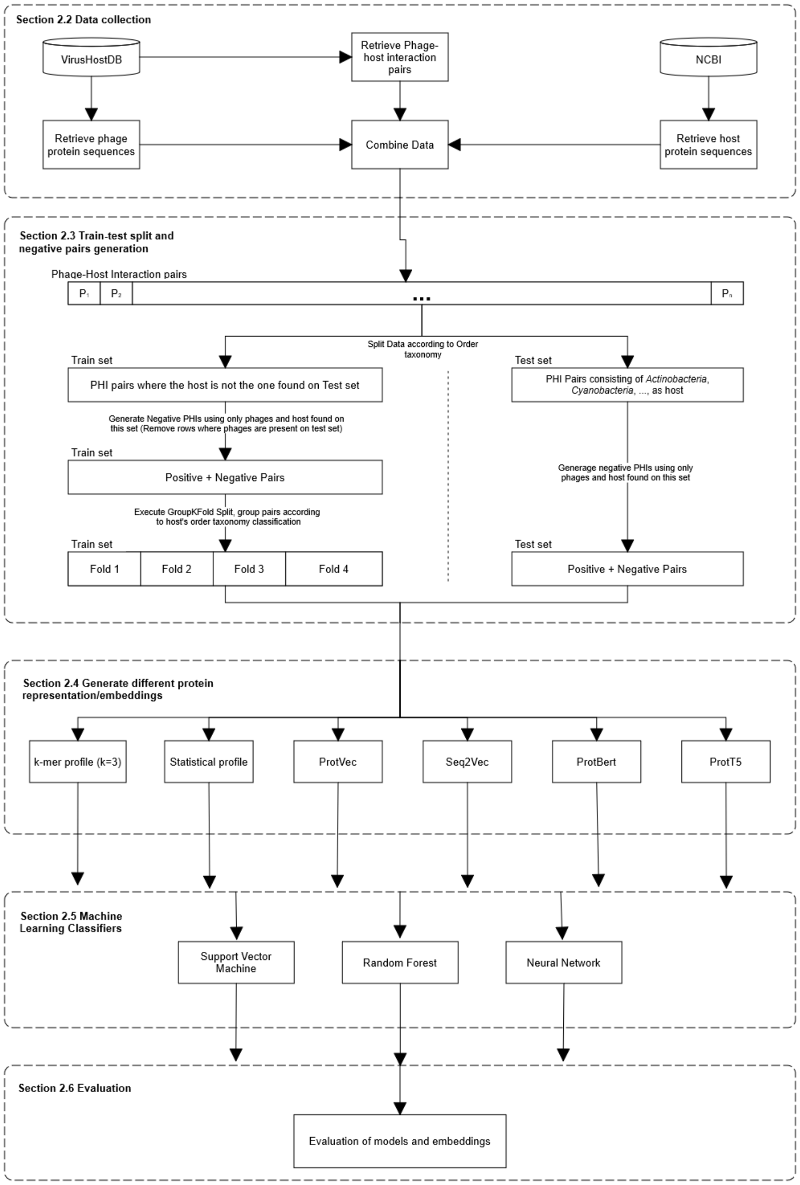

Section 2.2 Data collection
VirusHostDB
Retrieve Phage-host interaction pairs
NCBI
Retrieve phage protein sequences
Combine Data
Retrieve host protein sequences
Section 2.3 Train-test split and negative pairs generation
Phage-Host Interaction pairs
P1
P2
...
Pn
Split Data according to Order taxonomy
Train set
PHI pairs where the host is not the one found on Test set
Test set
PHI Pairs consisting of Actinobacteria, Cyanobacteria, ..., as host
Generate Negative PHIs using only phages and host found on this set (Remove rows where phages are present on test set)
Train set
Positive + Negative Pairs
Generage negative PHIs using only phages and host found on this set
Execute GroupKFold Split, group pairs according to host's order taxonomy classification
Train set
Fold 1
Fold 2
Fold 3
Fold 4
Test set
Positive + Negative Pairs
Section 2.4 Generate different protein representation/embeddings
k-mer profile (k=3)
Statistical profile
ProtVec
Seq2Vec
ProtBert
ProtT5
Section 2.5 Machine Learning Classifiers
Support Vector Machine
Random Forest
Neural Network
Section 2.6 Evaluation
Evaluation of models and embeddings

Figure 1: Flowchart summarizing the methodology of this study, which consists of 5 core components shown as dashed boxes. The first component is about collecting phage-host interaction records and protein sequences. The second component is about the train-test split method. The third component is about the generation of different protein representations. The fourth component is about the use of different machine learning classifier. And the last component is about the evaluation of protein representations and classifiers. Each component is discussed on Materials and methods section.

representation for a protein sequence, we generated ProtVec embeddings for each 3-mers that make up the protein sequence, summed up all the 3-mer embeddings element-wise, and then divided it by the number of 3-mer embeddings generated for that protein sequence. Since phages and hosts proteome consists of multiple proteins, to acquire the proteome representation, we summed up element-wise all the ProtVec representation of each protein sequence found in the proteome, then divided it by the number of proteins found in the proteome.

The ProtVec model is not available online, so we built it by training a Word2Vec model available on the gensim python library [30]. We trained the model using the Swiss-Prot database [9], and used the hyper-parameters specified in ProtVec's paper (*vector_size*: 100, *window*: 25, *sg*: 1).

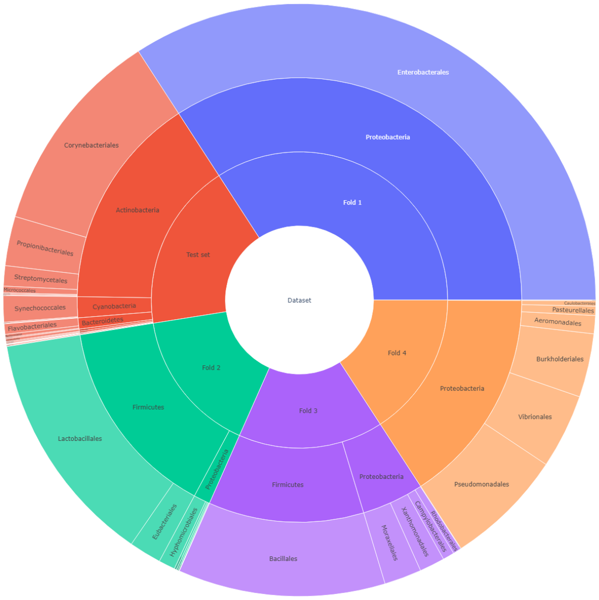


Figure 2: Sunburst chart visualizing the data for training and validation composition. The innermost pie chart visualizes the folds and the test set. The middle pie chart visualizes the hosts composition of each folds and test set at order level. The outermost pie chart visualizes the hosts composition of each folds and test set at species level.

2.4.4 Seq2Vec. Seq2Vec [18] is a type of protein embedding model based on Doc2Vec [20], where the main difference over ProtVec is the ability to generate a single embedding for the entire protein sequences regardless of its length, rather than having an embedding for each 3-mers. To acquire the proteome representation, we summed element-wise up all the embeddings of each protein sequence found in the proteome and divided it by the number of proteins found in the proteome. The Seq2Vec model is unavailable online, so we built it by training a Doc2Vec model available on the gensim python library [30]. We trained the model using the Swiss-Prot database [9], and used the hyper-parameters specified in Seq2Vec's paper (*vector_size*: 250, *window*: 5).

2.4.5 ProtTrans. ProtTrans [10] developed several protein embedding models, two of which are used in this study namely ProtBert and Prot5-XL. Prot5-XL was considered as one of their top-performing models when tested on different tasks such as protein subcellular localization prediction and secondary structure prediction, and it is used in this study to generate embeddings for each phage. However, host protein sequences are very long and we lack computing resources to generate Prot5-XL embeddings for each host, so we used ProtBert instead to generate embeddings for each host. To acquire the final proteome representation, we summed up element-wise all the embeddings of each protein sequence found in the proteome and divided it by the number of proteins found in the proteome.

Pre-trained ProtTrans models are available on HuggingFace's transformer library [35].

2.5 Machine learning classifiers

We used three classifiers: random forest, support vector machine, and a neural network. They take in the protein sequence representation of the phage-host pair as an input, and predict whether the phage-host pair interacts or not.

2.5.1 Random Forest. We trained a random forest classifier from sklearn [29], and used grid search to tune hyperparameters such as *n_estimators* to identify the best number of trees (the range was [25, 50, 100, 175, 250]), *criterion* to identify the best function to measure the quality of split (the range was [*gini*, *entropy*, *log_loss*]), and *max_features* to identify the best number of features to consider when looking for the best split (the range was [*sqrt*, *log2*, None]).

2.5.2 Support Vector Machine (SVM). We trained an SVM classifier using sklearn [29], and used grid search to tune hyperparameters such as *kernel* to identify the best kernel type to use (the range was [*linear*, *rbf*, *sigmoid*, *poly*]), *C* to identify the strength of the regularization (the range was [0.1, 1, 10, 50, 100]), *gamma* to identify the best value for the kernel coefficient (the range was [10, 5, 1, 0.1, 0.001], not applicable when kernel=*linear*), and *degree* to identify the best value for the degree of the polynomial kernel function (the range was [1,3,5], only applicable when kernel=*poly*).

2.5.3 Neural Network. Our neural network model accepts the combined vector representation of both the phage and the host, where the first half of the vector contains the representation of the host proteome, while the other half contains the representation of the phage proteome. The model consists of 3 fully connected hidden layers. The first layer contains 128 nodes, followed by a layer with 32 nodes, and followed by a layer with 8 nodes. The activation function used for the hidden layers is the Rectified Linear Unit (ReLU). The last layer in the neural network is a softmax layer containing 2 nodes, where the argmax function is used between the 2 nodes to predict whether an interaction between the phage-host pair is positive or not. The model was built using Pytorch [28]. We used the Adam function as our optimizer, with an initial learning rate of 0.75, which is multiplied by 0.75 every 75 epochs, to smoothen out the training loss as the number of epochs increases. We used a batch size of 256.

2.6 Evaluation

We evaluated our models on our test set. The performance metrics used to evaluate our models are accuracy, precision, recall, specificity, and F1 score.

Our data and codes are available at: https://github.com/bioinfodlsu/phi-prediction

3 RESULTS AND DISCUSSION

3.1 Evaluation result

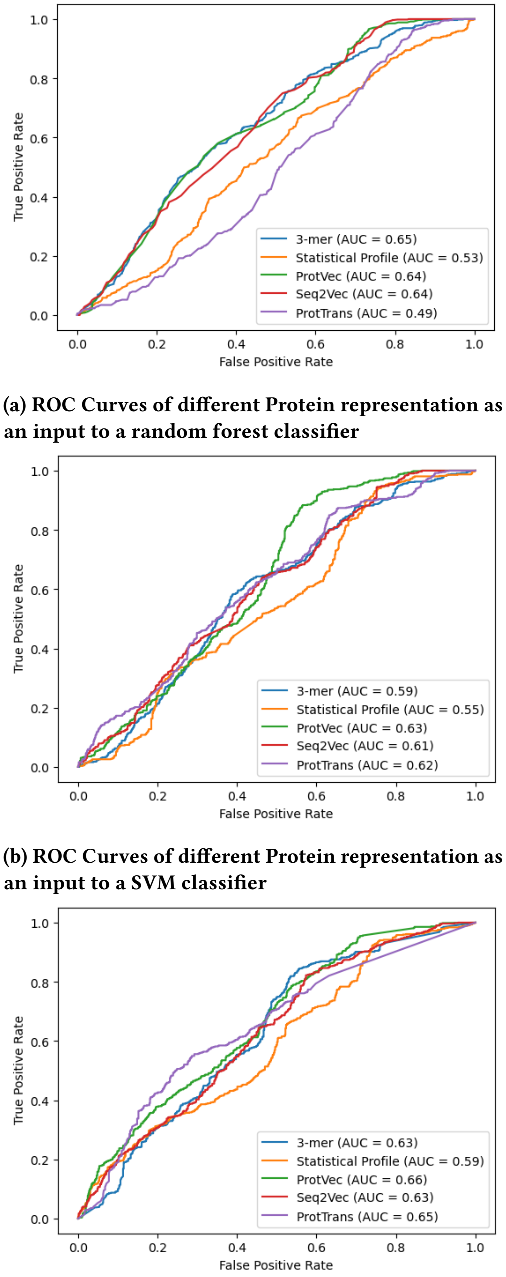
The result of evaluating different protein representations as input to different machine learning classifiers is shown in Table 1. The ROC curve, together with its AUC value is shown in Figure 3.

As can be observed from Table 1 and Figure 3, the different protein representations provide similar results across different classifiers (except for the statistical profile), with each representation outperforming the others in one kind of performance metric, but

Table 1: Results of evaluating different protein representations as input to random forest, SVM, and neural network classifiers, for predicting phage-host interaction.

Protein representation	Metric	Random Forest	SVM	Neural Network
3-mer	Accuracy	0.596931	0.598546	0.653473
	Precision	0.585960	0.592988	0.627346
	Recall	0.660743	0.628433	0.756058
	Specificity	0.533118	0.568659	0.550889
	F1 Score	0.621109	0.610196	0.685714
Statistics profile	Accuracy	0.492730	0.522617	0.423263
	Precision	0.484642	0.525180	0.413793
	Recall	0.229402	0.471729	0.368336
	Specificity	0.756058	0.573506	0.478191
	F1 Score	0.311404	0.497021	0.389744
ProtVec	Accuracy	0.590468	0.650242	0.627625
	Precision	0.564815	0.606651	0.586623
	Recall	0.788368	0.854604	0.864297
	Specificity	0.392569	0.445880	0.390953
	F1 Score	0.658125	0.709591	0.698890
Seq2Vec	Accuracy	0.596931	0.570275	0.608239
	Precision	0.567114	0.573854	0.573465
	Recall	0.819063	0.546042	0.844911
	Specificity	0.374798	0.594507	0.371567
	F1 Score	0.670192	0.559603	0.683214
ProtTrans (ProtBFD + Prot5)	Accuracy	0.489499	0.533118	0.627625
	Precision	0.341463	0.572438	0.589569
	Recall	0.022617	0.261712	0.840065
	Specificity	0.956381	0.804523	0.415186
	F1 Score	0.042424	0.359202	0.692871

being outperformed in another. For example, as shown in Table 1, 3-mer as input to the neural network slightly outperforms protein embeddings as input to the neural network in terms of accuracy, but as shown in Figure 3c, some protein embeddings such as ProtVec and ProtTrans slightly outperformed 3-mer in terms of AUC. Protein embeddings acquired from protein language models are shown to provide good results compared to the conventional k-mer profile representation when compared to other tasks found in bioinformatics such as predicting molecular function [34, 36]. However, protein embeddings do not necessarily improve classifier performance compared to a more conventional 3-mer profile in our phage-host interaction problem setting, possibly because protein embeddings on other tasks where it performed well are used individually as an input, whereas in this study, protein embeddings undergo averaging or mean-pooling due to the hosts and phages having multiple proteins. This averaging of embeddings may lead to the loss of information. Given our current results with the accuracy peaking only at about 65% and similarly low values of precision and recall, there is still room for improvement.



(a) ROC Curves of different Protein representation as an input to a random forest classifier

(b) ROC Curves of different Protein representation as an input to a SVM classifier

(c) ROC Curves of different Protein representation as an input to a Neural Network classifier

Figure 3: ROC Curves of different protein representation on different classifiers.

The statistical profile proposed by Li et al. [21] for their PredPHI classifier does not provide good results when used as an input to our classifiers as shown in Table 1 and Figure 3. A possible reason for the statistical profile not providing good results is that calculating the statistical values of protein-derived features over a large number of protein sequences (such as the mean of amino acid 'A' frequency overall protein sequences of a phage), might not provide predictive signals for the machine learning models.

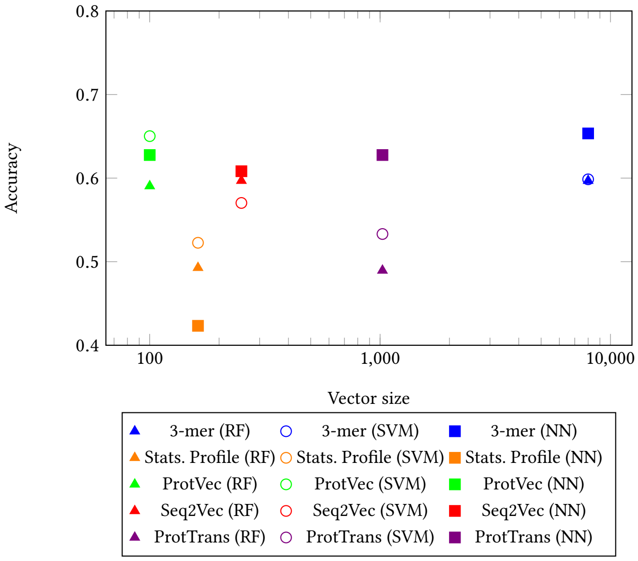


Figure 4: Visualization of the effect of different protein representation vector sizes in terms of the accuracy of classifiers for predicting phage-host interaction. RF indicates Random Forest, SVM indicates Support Vector Machine, NN indicates Neural Network

3.2 Host-specific evaluation

Shown in Table 2 is the performance of the SVM classifier using ProtVec protein embeddings as the input, for each type of host found in the test set. ProtVec embeddings and SVM were selected in this discussion since they achieved the highest accuracy when pairing a protein embedding with a machine learning model as shown in Table 1. For cases where a certain host only has a small number of recorded phage-host pairs, Young et al. [37] disregarded the hosts where the host does not have a minimum of 28 phages infecting it. However, in our current problem formulation, our classifier can still predict phages even for a host with a small number of recorded phages that infect it such as *Mycobacterium tuberculosis* with only 14 samples.

3.3 Effects of protein representation vector sizes

As can be observed from Figure 4, the vector size of protein representation does not correlate with the accuracy of the classifiers. For instance, ProtVec embedding is one of the top-performing representations even if it has the smallest vector size, outperforming the ProtTrans embedding, the largest vector size embeddings tested in this study, and even matching the performance of the 3-mer profile, which is 80 times larger in vector size. This suggests that the vector size does not directly affect the performance of the classifiers.

3.4 Future directions

Due to the large amount of proteins found on both the phage's proteome and the host's proteome, performing averaging or mean-pooling of protein representation for each proteome might lead to loss of information. It would be interesting to filter the proteins that are related to the adsorption process, such as receptor-binding proteins of the phages or surface proteins of the hosts.

For filtering the phage's receptor-binding proteins, there has been previous work that filters the proteins based on the gene annotations [5, 12]. For cases where a gene does not have annotations, there are tools available for annotation such as Prokka [32]. There's also a recent study by Boeckaerts et al. [6], which aims to predict whether a protein is a receptor-binding protein or not, given the protein sequence. The authors proposed two approaches: the first one uses Hidden Markov Models that represent protein domains strictly related to phage RBPs, and the other one generates protein embeddings and uses them as input to a machine learning classifier.

For filtering the host's surface proteins, to the best of our knowledge, there are currently no studies that can directly identify whether a protein can be found on the bacterial surface or not. However, there are studies that can identify the subcellular localization of a given protein [1, 8, 16].

4 CONCLUSIONS

In this study, we defined the phage-host prediction problem as taking the proteome of both the phage and the host as the input, and deciding whether the phage-host pair interacts or not. We tested different protein representations and different classifiers. Protein embeddings acquired from protein language models are shown to provide good results compared to a more conventional feature such as a k-mer profile for various tasks in bioinformatics [34, 36]. However, protein embeddings did not necessarily improve classifier performance in our phage-host interaction problem setting. For future directions, it would be interesting to filter the proteins that are related to the adsorption process, such as receptor binding proteins of the phages or surface proteins of the hosts.

REFERENCES

[1] José Juan Almagro Armenteros, Casper Kaae Sønderby, Søren Kaae Sønderby, Henrik Nielsen, and Ole Winther. 2017. DeepLoc: prediction of protein subcellular localization using deep learning. *Bioinformatics* 33, 21 (07 2017), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431 arXiv:https://academic.oup.com/bioinformatics/article-pdf/33/21/3387/50315453/bioinformatics_33_21_3387.pdf

[2] Ehsaneddin Asgari and Mohammad R. Mofrad. 2015. Continuous distributed representation of biological sequences for deep proteomics and Genomics. *PLOS ONE* 10, 11 (2015). https://doi.org/10.1371/journal.pone.0141287

[3] Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2014. Representation Learning: A Review and New Perspectives. arXiv:1206.5538 [cs.LG]

[4] Juliano Bertozzi Silva, Zachary Storms, and Dominic Sauvageau. 2016. Host receptors for bacteriophage adsorption. *FEMS Microbiology Letters* 363, 4 (01 2016). https://doi.org/10.1093/femsle/fnw002 arXiv:https://academic.oup.com/femsle/article-pdf/363/4/fnw002/23927805/fnw002.pdf fnw002.

[5] Dimitri Boeckaerts, Michiel Stock, Bjorn Criel, Hans Gerstmans, Bernard De Baets, and Yves Briers. 2021. Predicting bacteriophage hosts based on sequences of annotated receptor-binding proteins. *Scientific Reports* 11, 1 (2021). https://doi.org/10.1038/s41598-021-81063-4

[6] Dimitri Boeckaerts, Michiel Stock, Bernard De Baets, and Yves Briers. 2022. Identification of Phage Receptor-Binding Protein Sequences with Hidden Markov Models and an Extreme Gradient Boosting Classifier. *Viruses* 14, 6 (June 2022), 1329. https://doi.org/10.3390/v14061329

[7] Yves Briers, Dimitri Boeckaerts, Michiel Stock, Celia Ferriol-González, Jesús Oteo-Iglesias, Rafael Sanjuan, Pilar Domingo-Calap, and Bernard De Baets. 2023. *Actionable prediction of Klebsiella phage-host specificity at the subspecies level* (Jul 2023). https://doi.org/10.21203/rs.3.rs-3101607/v1

[8] Sebastian Briesemeister, Torsten Blum, Scott Brady, Yin Lam, Oliver Kohlbacher, and Hagit Shatkay. 2009. SherLoc2: A High-Accuracy Hybrid Method for Predicting Subcellular Localization of Proteins. *Journal of Proteome Research* 8, 11 (2009), 5363–5366. https://doi.org/10.1021/pr900665y

Table 2: Evaluation of SVM classifier using ProtVec protein embeddings as the input for some hosts found on our test set.

Host Name	# of samples	Accuracy	Precision	Recall	Specificity	F1 Score
Leptospira noguchii serovar	2	1.0000	1.0000	1.0000	1.0000	1.0000
Microcystis aeruginosa	2	0.5000	0.5000	1.0000	0.0000	0.6667
Synechococcus sp.	44	0.5227	0.4857	0.8500	0.2500	0.6182
Prochlorococcus	14	0.7857	0.7000	1.0000	0.5714	0.8235
Prochlorococcus marinus	8	1.0000	1.0000	1.0000	1.0000	1.0000
Bifidobacterium asteroides	4	0.7500	1.0000	0.5000	1.0000	0.6667
Corynebacterium glutamicum	6	0.6667	0.5000	1.0000	0.5000	0.6667
Propionibacterium freudenreichii	18	0.8333	0.8667	0.9286	0.5000	0.8966
Cutibacterium acnes	162	0.6667	0.6000	0.8961	0.4588	0.7187
Mycobacterium avium	2	0.5000	0.0000	0.0000	0.5000	0.0000
Mycolicibacterium phlei	6	0.6667	0.6667	0.6667	0.6667	0.6667
Mycolicibacterium smegmatis	200	0.6250	0.5870	0.8182	0.4356	0.6835
Mycobacterium tuberculosis	14	0.9286	0.9167	1.0000	0.6667	0.9565
Rhodococcus rhodochrous	2	0.5000	0.0000	0.0000	0.5000	0.0000
Rhodococcus erythropolis	24	0.6250	0.5714	1.0000	0.2500	0.7273
Streptomyces griseus	52	0.5577	0.5250	0.8400	0.2963	0.6462
Streptomyces hygroscopicus	2	0.5000	0.5000	1.0000	0.0000	0.6667
Streptomyces scabiei	8	0.7500	1.0000	0.7143	1.0000	0.8333
Microbacterium paraoxydans	16	0.7500	0.6667	0.8571	0.6667	0.7500
Gordonia terrae	352	0.6676	0.6230	0.8588	0.4743	0.7221
Mycoplasma	4	0.5000	0.3333	1.0000	0.3333	0.5000
Rhodococcus opacus	2	0.5000	0.5000	1.0000	0.0000	0.6667
Rhodococcus hoagii	16	0.8125	0.9091	0.8333	0.7500	0.8696
Parabacteroides merdae	2	0.5000	0.5000	1.0000	0.0000	0.6667
Chlamydia psittaci	6	0.8333	0.6667	1.0000	0.7500	0.8000
Corynebacteriales	6	0.6667	0.6667	1.0000	0.0000	0.8000
Fusobacterium nucleatum	4	0.7500	1.0000	0.5000	1.0000	0.6667
Gordonia amicalis	2	0.5000	0.0000	0.0000	1.0000	0.0000
Flavobacterium psychrophilum	30	0.6333	0.6087	0.8750	0.3571	0.7179
Flavobacterium columnare	4	1.0000	1.0000	1.0000	1.0000	1.0000

arXiv:https://doi.org/10.1021/pr900665y PMID: 19764776.

[9] The UniProt Consortium. 2022. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Research* 51, D1 (11 2022), D523–D531. https://doi.org/10.1093/nar/gkac1052 arXiv:https://academic.oup.com/nar/article-pdf/51/D1/D523/48441158/gkac1052.pdf

[10] Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. 2022. ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 44, 10 (2022), 7112–7127. https://doi.org/10.1109/TPAMI.2021.3095381

[11] Zhabiz Golkar, Omar Bagasra, and Donald Gene Pace. 2014. Bacteriophage therapy: a potential solution for the antibiotic resistance crisis. *The Journal of Infection in Developing Countries* 8, 02 (2014), 129–136. https://doi.org/10.3855/jidc.3573

[12] Mark Edward Gonzales, Jennifer C. Ureta, and Anish M. Shrestha. 2023. Protein embeddings improve phage-host interaction prediction. *PLOS ONE* 18, 7 (Jul 2023). https://doi.org/10.1371/journal.pone.0289030

[13] Fernando L. Gordillo Altamirano and Jeremy J. Barr. 2019. Phage Therapy in the Postantibiotic Era. *Clinical Microbiology Reviews* 32, 2 (2019). https://doi.org/10.1128/cmr.00066-18

[14] Alison H Holmes, Luke S P Moore, Arnfinn Sundsfjord, Martin Steinbakk, Sadie Regmi, Abhilasha Karkey, Philippe J Guerin, and Laura J V Piddock. 2015. Understanding the mechanisms and drivers of antimicrobial resistance. *The Lancet* 387, 10014 (2015), 176–187. https://doi.org/10.1016/s0140-6736(15)00473-0

[15] John N. Housby and Nicholas H. Mann. 2009. Phage therapy. *Drug Discovery Today* 14, 11 (2009), 536–540. https://doi.org/10.1016/j.drudis.2009.03.006

[16] Annette Höglund, Pierre Dönnes, Torsten Blum, Hans-Werner Adolph, and Oliver Kohlbacher. 2006. MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition. *Bioinformatics* 22, 10 (01 2006), 1158–1165. https://doi.org/10.1093/bioinformatics/btl002 arXiv:https://academic.oup.com/bioinformatics/article-pdf/22/10/1158/48838310/bioinformatics_22_10_1158.pdf

[17] Hitoshi Iuchi, Taro Matsutani, Keisuke Yamada, Natsuki Iwano, Shunsuke Sumi, Shion Hosoda, Shitao Zhao, Tsukasa Fukunaga, and Michiaki Hamada. 2021. Representation learning applications in biological sequence analysis. *Computational and Structural Biotechnology Journal* 19 (2021), 3198–3208. https://doi.org/10.1016/j.csbj.2021.05.039

[18] Dhananjay Kimothi, Akshay Soni, Pravesh Biyani, and James M. Hogan. 2016. Distributed Representations for Biological Sequence Analysis. *CoRR* abs/1608.05949 (2016). arXiv:1608.05949 http://arxiv.org/abs/1608.05949

[19] Ramanan Laxminarayan, Adriano Duse, Chand Wattal, Anita K Zaidi, Heiman F Wertheim, Nithima Sumpradit, Erika Vlieghe, Gabriel Levy Hara, Ian M Gould, Herman Goossens, and et al. 2013. Antibiotic resistance—the need for global solutions. *The Lancet Infectious Diseases* 13, 12 (Nov 2013), 1057–1098. https://doi.org/10.1016/s1473-3099(13)70318-9

[20] Quoc V. Le and Tomas Mikolov. 2014. Distributed Representations of Sentences and Documents. https://doi.org/10.48550/ARXIV.1405.4053

[21] Menglu Li, Yanan Wang, Fuyi Li, Yun Zhao, Mengya Liu, Sijia Zhang, Yannan Bin, A. Ian Smith, Geoffrey I. Webb, Jian Li, Jiangning Song, and Junfeng Xia. 2021. A Deep Learning-Based Method for Identification of Bacteriophage-Host Interaction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 18, 5 (2021), 1801–1810. https://doi.org/10.1109/TCBB.2020.3017386

[22] Menglu Li and Wen Zhang. 2021. PHIAF: prediction of phage-host interactions with GAN-based data augmentation and sequence-based feature fusion. *Briefings in Bioinformatics* 23, 1 (09 2021), bbab348.

https://doi.org/10.1093/bib/bbab348 arXiv:https://academic.oup.com/bib/article-pdf/23/1/bbab348/42229736/bbab348.pdf

[23] Derek M Lin, Britt Koskella, and Henry C Lin. 2017. Phage therapy: An alternative to antibiotics in the age of multi-drug resistance. *World Journal of Gastrointestinal Pharmacology and Therapeutics* 8, 3 (2017), 162–173. https://doi.org/10.4292/wjgpt.v8.i3.162

[24] Hilary D. Marston, Dennis M. Dixon, Jane M. Knisely, Tara N. Palmore, and Anthony S. Fauci. 2016. Antimicrobial Resistance. *JAMA* 316, 11 (09 2016), 1193–1204. https://doi.org/10.1001/jama.2016.11764 arXiv:https://jamanetwork.com/journals/jama/articlepdf/2553454/jsc160016.pdf

[25] Tomoko Mihara, Yosuke Nishimura, Yugo Shimizu, Hiroki Nishiyama, Genki Yoshikawa, Hideya Uehara, Pascal Hingamp, Susumu Goto, and Hiroyuki Ogata. 2016. Linking Virus Genomes with Host Taxonomy. *Viruses* 8, 3 (2016). https://doi.org/10.3390/v8030066

[26] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. https://doi.org/10.48550/ARXIV.1301.3781

[27] Christopher JL Murray, Kevin Shunji Ikuta, Fablina Sharara, Lucien Swetschinski, Gisela Robles Aguilar, Authia Gray, Chieh Han, Catherine Bisignano, Puja Rao, and Eve et al. Wool. 2022. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The Lancet* 399, 10325 (2022), 629–655. https://doi.org/10.1016/s0140-6736(21)02724-0

[28] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. arXiv:1912.01703 [cs.LG]

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[30] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[31] Eric W Sayers, Evan E Bolton, J Rodney Brister, Kathi Canese, Jessica Chan, Donaldnbsp;C Comeau, Ryan Connor, Kathryn Funk, Chris Kelly, Sunghwan Kim, and et al. 2021. Database Resources of the National Center for Biotechnology Information. *Nucleic Acids Research* 50, D1 (2021). https://doi.org/10.1093/nar/gkab1112

[32] Torsten Seemann. 2014. Prokka: rapid prokaryotic genome annotation. *Bioinformatics* 30, 14 (03 2014), 2068–2069. https://doi.org/10.1093/bioinformatics/btu153 arXiv:https://academic.oup.com/bioinformatics/article-pdf/30/14/2068/48924770/bioinformatics_30_14_2068.pdf

[33] Saptarshi Sinha, Rajdeep K. Grewal, and Soumen Roy. 2018. Chapter Three - Modeling Bacteria–Phage Interactions and Its Implications for Phage Therapy. Advances in Applied Microbiology, Vol. 103. Academic Press, 103–141. https://doi.org/10.1016/bs.aambs.2018.01.005

[34] Amelia Villegas-Morcillo, Stavros Makrodimitris, Roeland C H J van Ham, Angel M Gomez, Victoria Sanchez, and Marcel J T Reinders. 2020. Unsupervised protein embeddings outperform hand-crafted sequence and structure features at predicting molecular function. *Bioinformatics* 37, 2 (08 2020), 162–170. https://doi.org/10.1093/bioinformatics/btaa701 arXiv:https://academic.oup.com/bioinformatics/article-pdf/37/2/162/50322153/btaa701_supplementary_data.pdf

[35] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 [cs.CL]

[36] Kevin K Yang, Zachary Wu, Claire N Bedbrook, and Frances H Arnold. 2018. Learned protein embeddings for machine learning. *Bioinformatics* 34, 15 (03 2018), 2642–2648. https://doi.org/10.1093/bioinformatics/bty178 arXiv:https://academic.oup.com/bioinformatics/article-pdf/34/15/2642/48935464/bioinformatics_34_15_2642.pdf

[37] Francesca Young, Simon Rogers, and David L. Robertson. 2020. Predicting host taxonomic information from viral genomes: A comparison of feature representations. *PLOS Computational Biology* 16, 5 (05 2020), 1–24. https://doi.org/10.1371/journal.pcbi.1007894