BACPHLIP: Predicting bacteriophage lifestyle from conserved protein domains

Adam J. Hockenberry[1,*], Claus O. Wilke[1],

1 Department of Integrative Biology, The University of Texas at Austin, Austin, TX 78712, USA
* Corresponding author

Abstract
Motivation: Bacteriophages are broadly classified into two distinct lifestyles: temperate (lysogenic) and virulent (lytic). Temperate phages are capable of a latent phase of infection within a host cell, whereas virulent phages directly replicate and lyse host cells upon infection. Accurate lifestyle identification is critical for determining the role of individual phage species within ecosystems and their effect on host evolution.
Results: Here, we present BACPHLIP, a BACterioPHage LIfestyle Predictor. BACPHLIP detects the presence of a set of conserved protein domains within an input genome and uses this data to predict lifestyle via a Random Forest classifier. The classifier was trained on 634 phage genomes. On an independent test set of 423 phages, BACPHLIP has an accuracy of 98%, greatly exceeding that of the best existing available tool (79%).
Availability: BACPHLIP is freely available on GitHub (`https://github.com/adamhockenberry/bacphlip`) and the code used to build and test the classifier is provided in a separate repository (`https://github.com/adamhockenberry/bacphlip-model-dev`).

1 Introduction

Bacteriophages play important ecological roles (1–4) and influence both the physiology and evolution of host species (5–7). Additionally, phages possess a number of unique traits (8, 9) that are relevant to ongoing biotechnological and medical applications (10–12).

The availability of phage genome sequences has expanded greatly in recent years—owing largely to the development of meta-genomics (13–18). A particularly important phage phenotype is whether the phage has a temperate (lysogenic) or virulent (lytic) lifestyle (19–22). Phage lifestyle ultimately needs to be determined experimentally, but it is infeasible to perform the required experiments exhaustively on the large number of newly discovered phage genomes.

McNair *et al.* [23] developed a machine learning classifier (PHACTS) to predict phage lifestyle from the protein sequences encoded within the genome. However, PHACTS was trained and tested on a set of phages that were available in 2012—a total of 227. More recently, Mavrich *et al.* [24] compiled a dataset of >1,000 phage genomes with lifestyle annotations and describe a distinct classification method (which we abbreviate hereafter as "Mavrich") based on the presence of specific conserved protein domains. Their approach

reported a high prediction accuracy of ∼96%, but training and testing sets were not explicitly defined. Moreover, software development was not the focus of that study, and their classification tool is not available for general use.

Here, we combined the distinct approaches used in previous studies (23, 24) to create an up-to-date, user-friendly software package for predicting phage lifestyles based solely on genome sequence data. Our BACterioPHage LIfestyle Predictor (BACPHLIP) is a python library with an optional command-line interface that relies on the HMMER3 software suite (25) to identify the presence/absence of a set of conserved protein domains. BACPHLIP is fully open-source and achieves high prediction accuracy (98.3%) on an independent test set of 423 phages—greatly exceeding the performance of existing available methods.

2 Development and implementation

We developed BACPHLIP by first searching the Conserved Domain Database (26; accessed: 03/2020) for protein domains with descriptions containing any one of several search terms that are hypothesized to be enriched in temperate phages given their unique integration and excision capabilities (*e.g.* "integrase", "recombinase", "excisionase", *etc.*). To determine the relative importance of each of the 371 initial protein domains, we relied on the 1,057 phages with annotated lifestyles collected by Mavrich *et al.* [24]. For each genome sequence in the labeled dataset, we created a list of all possible 6-frame translation products $\geq$ 40 amino acids. Next, we used HMMER3 to search for the presence of the aforementioned protein domains within each genome. The output of this step was a vector for each phage describing the presence (1) or absence (0) of each domain.

We next built a machine learning pipeline to prioritize among the starting 371 protein domains in our database and identify higher-order associations. We randomly split the phage dataset into training and testing sets (60:40 split, 634 and 423 phages). Using only the training data, we then removed any protein domain from subsequent analysis that was present in two or fewer genomes, as well as any for which the number of hits in virulent phage genomes equaled or exceeded that of temperate phages. We thus established a condensed dataset of 206 putatively useful protein domains for downstream phage classification.

Finally, we fit a Random Forest classifier (implemented in python using `scikit-learn`) to our labeled training data using cross-validation to tune hyper-parameters and mitigate against overfitting. The best performing model from this search, when re-fit to the entire training set, achieved 99.8% predictive accuracy (633/634 correct predictions) on the training data.

3 Validation

When applied to an independent test set of 423 phages, BACPHLIP achieved a 98.3% classification accuracy (415/423 correct predictions, Figure 1). This accuracy greatly exceeded that of PHACTS (79% accuracy on the same set of phages) and improved upon the predictions reported by Mavrich *et al.* [24] (95.5% accuracy)—this despite the fact that some fraction

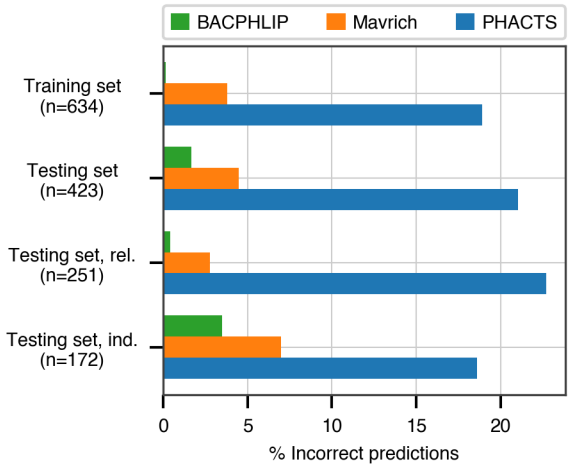

Figure 1. Classification accuracy of each compared method across all datasets analyzed. The labels "rel." (related) and "ind."(independent) refer to subsets of the testing set with (and without) related genomes contained in the training set.

of our test set may have been used in training these previous approaches. For assessing PHACTS, we note that we report the lifestyle as the category with the highest probability *regardless* of the confidence placed on that prediction, reasoning that "no prediction" should be viewed as an error.

The high accuracy of BACPHLIP could result from correctly identifying general principles governing phage proteomes, identifying phylogenetic similarities between query and training set species, or some combination of these two. To ascertain whether hidden phylogenetic structure in our dataset was inflating test set accuracies, we conservatively clustered all labeled phages (training and testing sets) using fastANI (27) and assessed BACPHLIP accuracy separately on test data for which there were (and were not) genomes in the training set with >80% sequence identity across >80% of the genome. In the most strict test set, where the training set contained only distant phylogenetic representatives, BACPHLIP achieved 96.5% accuracy (compared with 82% and 93% for PHACTS and Mavrich, respectively; Figure 1).

4 Discussion

Across all tested datasets, BACPHLIP substantially outperforms existing available software for classifying phage lifestyles. Our method relies on having sufficiently large training data (phages with known lifestyles) and we encourage users of BACPHLIP to recognize this important contribution made by Mavrich *et al.* [24]. However, we also wish to emphasize that the existing dataset is biased towards a small number of hosts (95% of labeled phages infect species within the orders *Actinobacteria*, *Gammaproteobacteria*, and *Bacilli*) and urge caution when predicting the lifestyle of phages infecting species that are not well-represented within the training data. We anticipate that the accuracy of BACPHLIP will increase in future releases as: i) more phylogenetically diverse phages become available for training

the classifier, and ii) discovery and annotation of conserved protein domains improves to encapsulate new domains and more phylogenetic diversity amongst the 206 domains that BACPHLIP currently relies on.

Funding

This work was supported by National Institutes of Health grants F32 GM130113 to A.J.H. and R01 GM088344 to C.O.W.

References

[1] Brum, J. R. *et al.* Patterns and ecological drivers of ocean viral communities. *Science* **348** (2015). URL `https://science.sciencemag.org/content/348/6237/1261498`.

[2] Nishimura, Y. *et al.* Environmental Viral Genomes Shed New Light on Virus-Host Interactions in the Ocean. *mSphere* **2**, e00359–16 (2017). URL `https://mSphere.asm.org/lookup/doi/10.1128/mSphere.00359-16`.

[3] Emerson, J. B. *et al.* Host-linked soil viral ecology along a permafrost thaw gradient. *Nature Microbiology* **3**, 870–880 (2018). URL `http://www.nature.com/articles/s41564-018-0190-y`.

[4] Daly, R. A. *et al.* Viruses control dominant bacteria colonizing the terrestrial deep biosphere after hydraulic fracturing. *Nature Microbiology* **4**, 352–361 (2019). URL `http://www.nature.com/articles/s41564-018-0312-6`.

[5] Touchon, M., Bernheim, A. & Rocha, E. P. Genetic and life-history traits associated with the distribution of prophages in bacteria. *The ISME Journal* **10**, 2744–2754 (2016). URL `http://www.nature.com/articles/ismej201647`.

[6] Bondy-Denomy, J. *et al.* Prophages mediate defense against phage infection through diverse mechanisms. *The ISME Journal* **10**, 2854–2866 (2016). URL `http://www.nature.com/articles/ismej201679`.

[7] Carey, J. N. *et al.* Phage integration alters the respiratory strategy of its host. *eLife* **8**, e49081 (2019). URL `https://doi.org/10.7554/eLife.49081`.

[8] Bobay, L.-M. & Ochman, H. Biological species in the viral world. *Proceedings of the National Academy of Sciences* **115**, 6040–6045 (2018). URL `http://www.pnas.org/lookup/doi/10.1073/pnas.1717593115`.

[9] Ofir, G. & Sorek, R. Contemporary Phage Biology: From Classic Models to New Insights. *Cell* **172**, 1260–1270 (2018). URL `http://www.sciencedirect.com/science/article/pii/S0092867417313053`.

[10] Rodriguez-Gonzalez, R. A., Leung, C. Y., Chan, B. K., Turner, P. E. & Weitz, J. S. Quantitative Models of Phage-Antibiotic Combination Therapy. *mSystems* **5**, e00756–19 (2020). URL `http://msystems.asm.org/lookup/doi/10.1128/mSystems.00756-19`.

[11] Dedrick, R. M. *et al.* Engineered bacteriophages for treatment of a patient with a disseminated drug-resistant Mycobacterium abscessus. *Nature Medicine* **25**, 730–733 (2019). URL http://www.nature.com/articles/s41591-019-0437-z.

[12] Roach, D. R. *et al.* Synergy between the Host Immune System and Bacteriophage Is Essential for Successful Phage Therapy against an Acute Respiratory Pathogen. *Cell Host & Microbe* **22**, 38–47 (2017). URL http://www.sciencedirect.com/science/article/pii/S1931312817302585.

[13] Deng, L. *et al.* Viral tagging reveals discrete populations in Synechococcus viral genome sequence space. *Nature* **513**, 242–245 (2014). URL http://www.nature.com/articles/nature13459.

[14] Pope, W. H. *et al.* Whole genome comparison of a large collection of mycobacteriophages reveals a continuum of phage genetic diversity. *eLife* **4**, e06416 (2015). URL https://doi.org/10.7554/eLife.06416.

[15] Roux, S., Hallam, S. J., Woyke, T. & Sullivan, M. B. Viral dark matter and virus–host interactions resolved from publicly available microbial genomes. *eLife* **4**, e08490 (2015). URL https://doi.org/10.7554/eLife.08490.

[16] Simmonds, P. *et al.* Virus taxonomy in the age of metagenomics. *Nature Reviews Microbiology* **15**, 161–168 (2017). URL http://www.nature.com/articles/nrmicro.2016.177.

[17] Mizuno, C. M. *et al.* Numerous cultivated and uncultivated viruses encode ribosomal proteins. *Nature Communications* **10**, 752 (2019). URL http://www.nature.com/articles/s41467-019-08672-6.

[18] Tisza, M. J. *et al.* Discovery of several thousand highly diverse circular DNA viruses. *eLife* **9**, e51971 (2020). URL https://doi.org/10.7554/eLife.51971.

[19] Bobay, L.-M., Rocha, E. P. & Touchon, M. The Adaptation of Temperate Bacteriophages to Their Host Genomes. *Molecular Biology and Evolution* **30**, 737–751 (2013). URL https://academic.oup.com/mbe/article-lookup/doi/10.1093/molbev/mss279.

[20] Harrison, E. & Brockhurst, M. A. Ecological and Evolutionary Benefits of Temperate Phage: What Does or Doesn't Kill You Makes You Stronger. *BioEssays* **39**, 1700112 (2017). URL https://onlinelibrary.wiley.com/doi/abs/10.1002/bies.201700112.

[21] Li, G., Cortez, M. H., Dushoff, J. & Weitz, J. S. When to be Temperate: On the Fitness Benefits of Lysis vs. Lysogeny. *bioRxiv* (2019). URL http://biorxiv.org/lookup/doi/10.1101/709758.

[22] Moura de Sousa, J. A., Pfeifer, E., Touchon, M. & Rocha, E. P. C. Genome diversification via genetic exchanges between temperate and virulent bacteriophages. *bioRxiv* (2020). URL http://biorxiv.org/lookup/doi/10.1101/2020.04.14.041137.

[23] McNair, K., Bailey, B. A. & Edwards, R. A. PHACTS, a computational approach to classifying the lifestyle of phages. *Bioinformatics* **28**, 614–618 (2012). URL https://academic.oup.com/bioinformatics/article-lookup/doi/10.1093/bioinformatics/bts014.

[24] Mavrich, T. N. & Hatfull, G. F. Bacteriophage evolution differs by host, lifestyle and genome. *Nature Microbiology* **2**, 17112 (2017). URL http://www.nature.com/articles/nmicrobiol2017112.

[25] Eddy, S. R. Accelerated Profile HMM Searches. *PLOS Computational Biology* **7**, e1002195 (2011). URL https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002195.

[26] Lu, S. *et al.* CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Research* **48**, D265–D268 (2020). URL https://academic.oup.com/nar/article/48/D1/D265/5645006.

[27] Jain, C., Rodriguez-R, L. M., Phillippy, A. M., Konstantinidis, K. T. & Aluru, S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**, 5114 (2018). URL http://www.nature.com/articles/s41467-018-07641-9.