Supporting Text for "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

In the primary text of the manuscript describing our bacteriophage lifestyle predictor (BACPHLIP), we focused on presenting software use cases, limitations, and motivation. Here, we more thoroughly detail the development, structure, and accuracy of BACPHLIP and describe extended results and methods to further validate and emphasize the details of our approach for interested readers and users of the software.

Supplementary results and discussion

Evaluating model performance

The purpose of the BACPHLIP software is to predict the lifestyle (temperate or virulent) of a given bacteriophage from genome sequence input. The most important question a potential user would likely wish to know the answer to is: how well this software performs at this stated task? When applied to an independent test set of 423 phages (240 temperate and 183 virulent, withheld for the entirety of model training and development), BACPHLIP achieved a 98.3% classification accuracy (415/423 correct predictions). On the same set of phages, this accuracy exceeded that of the previously existing PHACTS software [1], as well as the results reported by Mavrich *et al.* [2] (79% and 95.5% accuracy, respectively).

The impact of phylogenetic relatedness on classification accuracy. To ascertain the effect of phylogenetic structure in our dataset (which could inflate accuracy metrics since

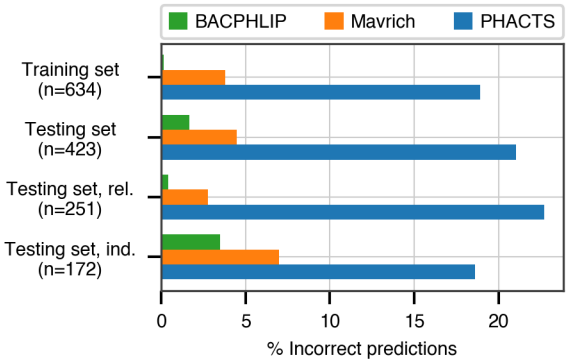

Figure S1. Classification accuracy of each compared method across all datasets analyzed. The labels "rel." (related) and "ind."(independent) refer to subsets of the testing set with (and without) related genomes contained in the training set.

testing set data may not be fully independent from training set data), we clustered all labeled phages using FastANI [3] and evaluated accuracy separately on testing set data for which there were no genomes in the training set with >80% sequence identity across >80% of the genome. Although the training set contained only distant phylogenetic relatives (at best) to these 172 genomes, BACPHLIP still achieved 96.5% classification accuracy (82% and 93% for PHACTS and Mavrich *et al.*, respectively).

Classification accuracy across all data subsets. We report classification accuracy as our primary indicator of model performance, defined as the number of correct predictions divided by the total number of predictions made. Here, in Figure S1, we show for the three tested methods (BACPHLIP, PHACTS [1], and Mavrich [2]), the % of incorrect predictions (defined simply as 1−accuracy) for the various datasets that we considered. These datasets include: i) the full training dataset, ii) the full testing dataset, iii) isolated members of the testing dataset for which there *were* identifiable relatives in the training set (see clustering procedure above), and iv) isolated members of the testing set for which there *were not* any close relatives in the training set (at the 80% sequence identity level, across at least 80% of the genome).

Alternative measures of model accuracy. While these results highlight the superior performance of BACPHLIP compared to existing approaches, both accuracy and error rate can be misleading in classification tasks. For instance, our testing set of 423 phages contains 240 temperate phages and 183 virulent phages. A model that simply predicted a temperate lifestyle for every phage, would have an accuracy of 57% (seemingly better than random chance) despite this hypothetical model not being much of a model at all. If we imagine a far more extreme (and purely hypothetical scenario) where our testing set contained 400 temperate phages and 23 virulent, then this same strategy of always guessing temperate would achieve an accuracy of 95%. In general, accuracy is slightly counter-intuitive in binary classification tasks because even in the best case scenario (equal class numbers), the expected accuracy from random guessing is 0.5 (rather than what our intuition might expect to be 0).

	BACPHLIP	Mavrich	PHACTS
Accuracy	0.983	0.955	0.79
Balanced accuracy	0.97	0.917	0.528
MCC	0.967	0.911	0.586
F1-score	0.985	0.939	0.837

Table S1. Performance measures for the testing set data ($n = 423$). For all measures, higher values represent better model predictions. Balanced accuracy considers the uneven balance of classes ('temperate' and 'virulent') and was calculated with 'adjusted=True' such that random guessing would have a score of 0 and a perfect model would have a score of 1. 'MCC' stands for Matthew's correlation coefficient.

Given that our classes are relatively well balanced, accuracy is an appropriate and easy to interpret measurement, which is why we use it as our primary performance indicator. However, a number of other metrics have been developed to deal with the nuance of binary classification problems. In Table S1 we show our accuracy results alongside three other

measures of model performance (balanced accuracy, Matthew's Correlation Coefficient, and F1-score) for all three phage classification methods on the testing set data. While the numbers capture subtly different effects, the end result is always the same: i) BACPHLIP consistently outperforms all other methods and ii) on an absolute scale, its performance is indicative of a very accurate and robust model.

Exploration of classification errors. The metrics discussed above summarize classifier performance in a single number, but the reality is that different types of classification errors can occur and the nature of these errors may provide insight that can aid in future model improvements. BACPHLIP only makes 7 incorrect predictions on the testing set data, but it is useful to know whether these incorrect predictions are always in one direction (i.e. classifying a true temperate phage as virulent or classifying a true virulent phage as temperate). The gold standard for depicting binary classification results is a confusion matrix, where the matrix rows indicate the true phage lifestyles and the matrix columns are predictions. Table S2 shows these results for BACPHLIP, highlighting that 6 of the 7 errors occur when BACPHLIP classifies true temperate phages as virulent. These results are quite similar for Mavrich (Table S3), which makes more than twice as many errors but does so largely in the same general pattern as BACPHLIP. We note that this similarity is perhaps not surprising given that both of these methods rely on targeted identification of protein domains associated with lysogeny.

The incorrect classification of several true temperate phages as virulent (6 instances for BACPHLIP) suggests that either these particular temperate phages have incomplete genomes (and are thus missing critical lysogenic machinery for bioinformatic methods to identify) or more likely that the lysogenic machinery that they have is novel, uncharacterized, or highly diverged and thus not picked up by the (limited) set of protein domains that we used to build BACPHLIP. A deeper understanding of sequence level diversity within integrase, recombinase, excisionase, *etc.* families may eventually allow us to improve our method and correct these errors.

	Predicted class	
	Virulent	Temperate
Virulent	182	1
Temperate	6	234

Table S2. Testing set confusion matrix for BACPHLIP. Actual classes are depicted in rows, predicted classes in columns.

	Predicted class	
	Virulent	Temperate
Virulent	180	3
Temperate	16	224

Table S3. Testing set confusion matrix for Mavrich. Actual classes are depicted in rows, predicted classes in columns.

By contrast, the confusion matrix for PHACTS is heavily biased in the opposite manner (Table S4). While PHACTS makes many more errors compared to the other methods,

its errors are heavily biased towards classifying true virulent phages as temperate. The precise reason for these errors is unknown but the information could be valuable for future improvements to the PHACTS methodology.

	Predicted class	
	Virulent	Temperate
Virulent	105	78
Temperate	11	229

Table S4. Testing set confusion matrix for PHACTS. Actual classes are depicted in rows, predicted classes in columns.

Model errors and genome size. As one final test of potential factors that impact model accuracy, we investigated whether genome size had any impact on BACPHLIP's testing set accuracy. The results were insignificant (Wilcoxon rank-sum test, $p = 0.45$) when comparing the genome sizes of incorrectly and correctly classified phages. However, given the small number of incorrect test set predictions (7 out of 423) we note that the magnitude of the difference in genome sizes between incorrectly and correctly predicted phage lifestyles would have to be very large to detect in a statistically significant manner. At present, we do not have any reason to suspect that genome-size will play an important role in the ability to classify a given phage, provided that the genome is complete.

Relative feature importance in the BACPHLIP model

Although we have demonstrated the strong performance of BACPHLIP relative to existing methods, the presented analyses thus far have not yet shown why or how BACPHLIP makes its predictions–which can be a common and difficult problem in understanding machine learning models such as random forest classifiers (the basis of BACPHLIP).

Selection of protein domains. We began our analysis by searching descriptions within the conserved domain database [4] using several (case insensitive) search terms focused on identifying lysogeny-associated domains. These search terms were: 'integrase', 'excisionase', 'recombinase', 'transposase', 'lysogen', and 'temperate'. We additionally included a case sensitive search for 'parA|ParA|parB|ParB' due to its short length and potential overlap with many common words. Collectively, this search strategy identified 371 protein domains that formed our starting set. We stress that the "description" field of the selected domains must contain one or more of the above words at some point within it, but that the selected domains may or may not actually be related to these search terms. This strategy allowed us to select a potentially broad set of starting domains but it is by definition likely to include some erroneously included domains that we fully expect to not be useful for the task of delineating temperate from virulent phages. Our strategy is nevertheless sufficiently constrained so as to hopefully avoid over-fitting to phylogenetic signals or noise.

After removal of domains that were present in two or fewer *training* set genomes or which were actually more prevalent in the virulent phage genomes (again, only considering the training set), we established a more condensed dataset of 206 putatively useful lysogeny-associated protein domains. At this stage, we still did not know if any/all of these 206

domains would be useful for delineating temperate and virulent phages, which is why we next used this data as input into a random forest classifier that should be able to disregard un-important features (domains) and detect higher-level patterns in the data.

Evaluating the importance of individual protein domains. After fitting our model to the training data, we found (as with many machine learning tasks) that the overall importance of the 206 features (individual protein domains) was highly variable. Fig. S2 shows the distribution of feature importance values in the final BACPHLIP model. Six features have a final feature weight of 0 indicating that they were discarded by the model entirely. The "Top 50" protein domains (ranked in order of feature importance) accounted for 85% of the model weights and the "Top 20" protein domains accounted for 59%. Table S5 depicts the relevant number of domains in each category whose description contains the various search terms that we used at the outset. It is clear from this table that the most important search terms were 'integrase' and 'recombinase'. We again reiterate that this does not mean that these domains are from either integrases or recombinases but rather that the *description* of the domain from the Conserved Domain Database contains these respective words.

Each term in our search strategy apparently contributed something meaningful as the "Top 50" most important domains contained at least one domain hit to each term. Of course, as we note in the table caption, many domain descriptions contain multiple search term hits such that it is unclear from this whether each term made a *unique* contribution. We separately looked at these "Top 50" hits, while filtering out any description that contained hits to multiple search terms. Only 'excisionase' had zero hits given these constraints, which is not surprising as this was the term with the fewest number of domain hits to start. With the exception of the 'excisonase' term, this tells us that any conserved domain search strategy that omitted one of these search terms would have invariably produced a worse model. Of course, it is likely that many terms that we did not consider here may result in further improvements to future versions of BACPHLIP.

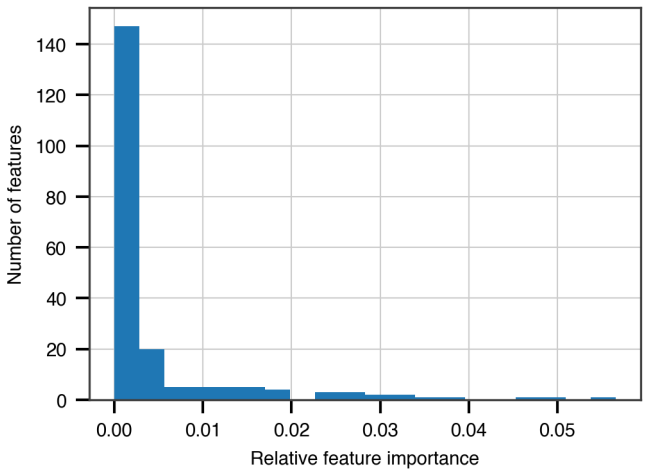

Figure S2. The distribution of feature importance values used by BACPHLIP. The sum of all feature importance values is 1.

Search term	Starting ($n = 371$)	Model ($n = 206$)	Top 50	Top 20
integrase	101	72	24	13
excisionase	5	4	2	0
recombinase	72	52	28	17
transposase	143	70	14	3
lysogen	23	10	2	1
temperate	11	10	3	0
parA|ParA|parB|ParB	65	29	7	0

Table S5. For each search term used to identify putatively important conserved protein domains, we show the number of domain descriptions that contain this term for various categories including: i) the starting set, ii) the input to the random forest model, iii) the top 50 and iv) the top 20 features after model fitting. Note that column entries will not sum up to the n depicted at the top of each column as many descriptions contain multiple search terms.

Leveraging prediction confidence to improve accuracy

Rather than outputting a simple prediction (temperate or virulent), BACPHLIP outputs a probability of belonging to either class. This is similar to the PHACTS model [1], which is also based on a random forest classifier. In our evaluation, we have selected the lifestyle for an individual phage by considering simply which lifestyle is predicted with >50% probability. However, we note to interested users of BACPHLIP that varying degrees of confidence are (and should be placed by users) on the predicted lifestyle of individual phages that have 51% vs 99% probability, for instance. Of the 423 testing set genomes, BACPHLIP predicted a lifestyle with ≥95% probability for 333 of these genomes. Of those 333, there was only one classification error (compared to 6 errors in the 90 genomes that were predicted with <95% probability). **Thus, in order to enrich accuracies and depending on the application, users may wish to restrict their analyses to genomes with ≥95% class probability and treat those genomes with lifestyle prediction between 50 and 95% as uncertain**.

Supplementary methods

We developed BACPHLIP by searching the Conserved Domain Database (4; accessed: 03/2020) for protein domains that are hypothesized to be enriched in temperate phages (*i.e.* mechanistically involved in lysogeny). We did not include a broad set of protein domains in our search strategy in order to ensure interpretability of our model and to limit the possibility of over-fitting. To determine which of the 371 initial protein domain hits preferentially associate with temperate phages, we leveraged 1,057 phages with annotated lifestyles collected by Mavrich *et al.* [2]. For each genome sequence, we created a list of all possible 6-frame translation products ≥ 40 amino acids. Next, we used HMMER3 to search for the presence of the aforementioned protein domains, resulting in a vector for each phage describing the presence (1) or absence (0) of each domain.

At this stage, we randomly split the phage dataset into training and testing sets (60:40 split, 634 and 423 phages). Using only the training data, we removed any protein domain that was present in two or fewer genomes or which was more prevalent in the virulent phage genomes. We thus established a condensed dataset of 206 putatively useful protein domains for downstream phage classification. Finally, we fit a Random Forest classifier to our labeled training data using cross-validation to tune hyper-parameters (20 separate randomly selected validation sets drawn from within the training set. 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.

Training, validation, and testing set construction

For readers who are unfamiliar with certain terminology and common approaches used machine learning applications, we wish to briefly expand upon potentially confusing points regarding the terminology of training, validation, and testing sets.

Both the training and testing of the BACPHLIP model relied on the dataset collected by Mavrich *et al.* [2], which was itself a composite dataset that the researchers assembled. To build and evaluate a classification model, we took this initial dataset and randomly split it into two completely separate groups (using a 60:40 split, as noted in the main text) that we refer to as training and testing sets. At this stage, the testing set was fully set aside for the remainder of any discussion of model training/fitting.

The training set, however, gets further split up during the process of model training into (what is still, rather unfortunately called) training and validation sets. This splitting of the training set is often necessary in machine learning applications because the random forest model has numerous hyper-parameters whose tuning can alter model performance. For instance, we assessed numerous combinations that varied whether to use bootstrapping, how to consider uneven class weights, the minimum number of samples per leaf, the maximum depth of the trees, and the number of estimators.

To learn which combination of these various parameters to use, we followed the commonly used validation set approach. For each hyper-parameter combination, we randomly split the overall training data up into separate and smaller training and validation sets. We then fit a model to the smaller training dataset, and assessed the accuracy of that fitted model on the validation set in question. Next, we repeated this process 20 different times for each hyper-parameter combination. At the end of this hyper-parameter search, we were thus left with 20 different measurements of validation set accuracy (where each validation set was independently drawn from the larger training set) for each possible hyper-parameter combination. We selected the 'best' model by choosing the hyper-parameter combination that yielded the highest minimum accuracy across the 20 independent validation set tests. The parameters of that model were then re-fit to the *entire* training set of data to become BACPHLIP.

At this stage, the random forest model has still not seen *any* data from the testing set, which makes it a fully independent dataset to test how well BACPHLIP performs. Indeed, at this stage the held out 423 phages could could have come from the same initial starting

dataset or somewhere else entirely and this distinction should make no difference our ability to assess the model. We therefore report (and emphasize) most of our results in terms of this independent testing set of phages.

While this testing set was randomly split solely for the purposes of developing/evaluating BACPHLIP, we test all of the different methods (including Mavrich and PHACTS) on the same sets of genomes in order to make for an equal comparison. However, we note that it's entirely possible/likely that some (or many) of the genomes in our independent testing set were actually used to develop and fit the models of the other methods. Which is to say, genomes in what we define as our testing set may have been a part of the training set for these other methods. Our approach in evaluating the accuracy of other methods and comparing them to BACPHLIP is thus highly conservative since, ideally, any phage genome used to train either previous approach should not be considered. Nevertheless, even this conservative approach which is likely to slightly inflate the accuracies of the other methods, shows that BACPHLIP has superior predictive power across all splits of the dataset that we analyzed.

Model fitting and parameter details

As previously noted, we trained the random forest classifier while tuning various hyper-parameters. Here, we explicitly note the range of parameters that we tested. Within the `scikit-learn` Random Forest framework, we evaluated 'bootstrap' (True, False), 'class_weight' (balanced, balanced_subsample), 'min_samples_leaf' (1, 2), 'n_estimators' (range(10, 105, 5)), and 'max_depth' (range(10, 42, 2)). We used 'GridSearchCV' to evaluate all possible combinations of these parameters (using 'f1' as the scoring function) across 20 different randomly selected validation sets (discussed at length above). All of the code necessary to perform/replicate our model fitting is freely available at: `https://github.com/adamhockenberry/bacphlip-model-dev`.

The final BACPHLIP model selected: 'bootstrap = False', 'class_weight = balanced_subsample', 'min_samples_leaf = 1', 'n_estimators = 80', and 'max_depth = 40'. All other parameters followed `scikit-learn` defaults as of version 0.23.1. The final model was fit, as noted above, to the entire training set using this set of parameters.

We note briefly that the 'class_weight' parameter in particular was used to assign variable weights given that the number of temperate and virulent phages in our training set were uneven. The 'balanced_subsample' parameter will correct for this un-evenness and ensure the accuracy of a model despite the (slightly) imbalanced classes.

Assessing the PHACTS model.

For assessing PHACTS, we report the lifestyle as the category with the highest probability *regardless* of the confidence, reasoning that "no prediction" should be viewed as an error.

References

[1] 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.

[2] 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.

[3] 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.

[4] 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.