BIOINFORMATICS ORIGINAL PAPER Vol. 28 no. 5 2012, pages 614–618
doi:10.1093/bioinformatics/bts014

Genome analysis Advance Access publication January 11, 2012

PHACTS, a computational approach to classifying the lifestyle of phages

Katelyn McNair[1,*], Barbara A. Bailey[2] and Robert A. Edwards[1,3,4,*]

[1]Computational Science Research Center, [2]Department of Mathematics and Statistics, [3]Department of Computer Science, San Diego State University, San Diego, CA 92182 and [4]Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL 60439, USA

Associate Editor: Martin Bishop

ABSTRACT

Motivation: Bacteriophages have two distinct lifestyles: virulent and temperate. The virulent lifestyle has many implications for phage therapy, genomics and microbiology. Determining which lifestyle a newly sequenced phage falls into is currently determined using standard culturing techniques. Such laboratory work is not only costly and time consuming, but also cannot be used on phage genomes constructed from environmental sequencing. Therefore, a computational method that utilizes the sequence data of phage genomes is needed.

Results: Phage Classification Tool Set (PHACTS) utilizes a novel similarity algorithm and a supervised Random Forest classifier to make a prediction whether the lifestyle of a phage, described by its proteome, is virulent or temperate. The similarity algorithm creates a training set from phages with known lifestyles and along with the lifestyle annotation, trains a Random Forest to classify the lifestyle of a phage. PHACTS predictions are shown to have a 99% precision rate.

Availability and implementation: PHACTS was implemented in the PERL programming language and utilizes the FASTA program (Pearson and Lipman, 1988) and the R programming language library 'Random Forest' (Liaw and Weiner, 2010). The PHACTS software is open source and is available as downloadable stand-alone version or can be accessed online as a user-friendly web interface. The source code, help files and online version are available at http://www.phantome.org/PHACTS/.

Contact: katelyn@rohan.sdsu.edu; redwards@sciences.sdsu.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on September 12, 2011; revised on January 4, 2012; accepted on January 5, 2012

1 INTRODUCTION

Viruses that infect bacteria are called bacteriophages or phages. It is estimated that there are 10^{30} bacterial cells in biosphere (Whitman *et al.*, 1998). Given that typical ratios of bacteria to phage are on the order of 1:10 (Wommack and Colwell, 2000), it is estimated that there exist 10^{31} phage particles on the planet. Viruses thus are the most abundant biological entities on the planet. Phages are ubiquitous and can be found in any environment where their bacterial hosts are present. Phages are found in high numbers in terrestrial environments such as soil, and in aquatic environments such as lakes and seawater (Srinivasiah *et al.*, 2008). Recent estimates suggest that there exist globally ~100 million phage species (Rohwer, 2003); however, only a small fraction of phages have so far been characterized.

When a phage infects a bacterial cell, the phage enters into one of two distinct lifestyles: virulent or temperate. During a virulent lifestyle a phage infects a bacteria; its genome is replicated many times; and the newly created copies are released into the surrounding environment through lysis, extrusion or budding. In contrast during a temperate lifestyle, a phage infects a bacteria and either integrates its DNA into the bacterial genome or re-circularizes its DNA into a stable plasmid. The temperate phage will live in this semi-stable lifestyle as a prophage as the host bacteria continues to grow and divide. The prophage will be carried through future bacterial cell divisions until appropriate environmental conditions cause the temperate phage to enter into a virulent lifestyle and release itself from the host bacterium. This switch into a virulent lifestyle is referred to as induction and is generally caused by host cell damage (Witkin, 1976) or environmental stressors (Clarke, 1998; Clark *et al.*, 1986). Not only does the characterization of phage lifestyles contribute to the understanding of phage population dynamics, genomics and microbiology; but also the virulent lifestyle has applications toward phage therapy and biocontrol (Housby and Mann, 2009).

Previously, the lifestyle of a phage was identified through culturing and isolation in the lab. This is not only time consuming but also costly. With the advent of shotgun sequencing, large numbers of phage are being sequenced at an increasing rate. As the ability to sequence new phages faster than culturing can identify the lifestyle, there is a need to computationally annotate genomic data and also to make predictions about the lifestyle. In addition, because many of these newly sequenced genomes are derived from entire environmental community sequencing methods, it may not be possible to isolate the phages for culturing.

Computationally classifying phages based on their genomes is difficult due to the highly mosaic organization of their genomes (Hendrix *et al.*, 1999). Unlike bacteria, which have 16S rRNA and various other conserved genes that can be used for taxonomy and phylogeny, phages have no universally present gene that can be used for analysis (Rohwer and Edwards, 2002). The first attempt at using genomic data to classify phage by comparing structural proteins does not work well across all clades of phages (Proux *et al.*, 2002). An alternative methodology was created by Rohwer and

*To whom correspondence should be addressed.

Edwards, and was used to create the Phage Proteomic Tree (Rohwer and Edwards, 2002). To deal with the mosaicism of phages, Lima-Mendez *et al.* (2008) implemented a framework for a reticulate classification based on gene content, by building a weighted graph where nodes represent phages and edges represent shared gene/protein similarities. Recently, this reticulate classification was extended to shared evolutionarily conserved modules consisting of groups of proteins that have a similar phylogenetic profile (Lima-Mendez *et al.*, 2011). Certain modules were found to be associated with either temperate or virulent phages, and it was suggested that a refining of the methodology might be used for an automated classification of phage lifestyle. An alternative method uses the tetranucleotide frequency differences between a phage and host to classify the lifestyle of the phage (Deschavanne *et al.*, 2010); however, this method is severely limited by the necessity to have a phages' host fully sequenced.

In this work, a Phage Classification Tool Set (PHACTS) was developed to classify whether a phage's preferred lifestyle is virulent or temperate. PHACTS utilizes a novel similarity algorithm and a supervised Random Forest classifier to make a prediction whether the lifestyle of a phage is virulent or temperate. The similarity algorithm creates a training set from phages with known lifestyles that, along with the lifestyle annotation, is used to train a Random Forest to classify the lifestyle of a phage. To test the accuracy of PHACTS, each phage with an annotated lifestyle was removed from the database one at a time and treated as a single phage with an unknown lifestyle. The lifestyle of the phage was predicted using PHACTS and the predicted lifestyle was then compared with the actual lifestyle.

2 METHODS

2.1 Implementation

2.1.1 Lifestyle database At the time of this work, the PHANTOME database of phages with complete genomes contained 654 phages (www.phantome.org). The lifestyles for 227 of these phages were manually curated by hand from various literature sources. In this subset of 227 phages with a known lifestyle, there were 148 temperate phages and 79 virulent phages, and thus temperate phages predominated the database 2:1. These phages with a known lifestyle were used to create a local database for use during PHACTS classifications.

2.1.2 Query proteins A set of query protein sequences $Q=\{P_1, P_2, \ldots, P_M\}$, is created by randomly selecting M proteins, where M is equal to the user-specified number of proteins to use for creating the training set. From each class, M/C proteins are selected at random that belong to phages of that class, where C is equal to the number of classes in the training set. For our experiments, it was empirically found that $M=600$ gave the best results. When M was decreased the accuracy went down, and when M was increased the runtimes went up without a corresponding increase in accuracy.

2.1.3 Training sets To create the training set for the Random Forest classifier, a set of N similarity vectors is assembled, where N is equal to the number of phages to use as training cases. From each class, N/C phage genomes are selected at random, without replacement. From these N phages the list $L=\{G_1, G_2, \ldots, G_N\}$ is created. The class with the fewest number of representative samples limits how many training cases can be used. For our purposes, it was empirically found that $N=100$ gave the best results. Having 50 phages per class was adequate to provide accurate results as well as allowing for a diverse random sampling. For each of these N genomes, a similarity vector X is assembled. The proteins of a phage are aligned against every protein in Q using the FASTA program. The percent identity score for each protein in that phage's proteome to the protein P is calculated as a percent identity corresponding to the highest scoring pair S. This percent identity score S is inserted into the similarity vector X, as shown below.

$$X_1=[S_{1,1}, S_{1,2}, \ldots, S_{1,M}]$$

$$X_2=[S_{2,1}, S_{2,2}, \ldots, S_{2,M}]$$

$$\ldots$$

$$X_N=[S_{N,1}, S_{N,2}, \ldots, S_{N,M}]$$

The manually curated lifestyles of the phages are retrieved from the locally stored database and are used as the classification factors.

2.1.4 Testing set The proteins of the input phage proteome are aligned against each protein in Q using the FASTA35 program. The percent identity score for each protein in the input phage's proteome to the protein P is calculated. The percent identity corresponding to the highest scoring pair is inserted into a vector X_{N+1}. A single similarity vector is assembled for the input phage's proteome as shown below.

$$X_{N+1}=[S_{N+1,1}, S_{N+1,2}, \ldots, S_{N+1,M}]$$

This vector becomes the testing set, and the Random Forest ensemble classifier is used to predict the lifestyle.

2.1.5 Random Forest To classify the testing set, PHACTS utilizes the Random Forest algorithm. In the Random Forest classifier, a set of decision trees is created. For each tree, bootstrapping is performed by selecting N cases with replacement from the training set of N cases. Each tree is grown by randomly selecting m number of variables at each node, where m is equal to the square root of the total number of variables. The best split at that node is calculated from these m variables, and the tree is grown to the largest extent possible. Each tree predicts a lifestyle and the final prediction is a majority-voting rule for the trees in the Random Forest. Random Forest also returns information on the voting as a percentage that corresponds to the number of trees that predicted a particular lifestyle divided by the total number of trees. Since the Random Forest algorithm does not overfit the data, large numbers of trees can be created. For our predictions, 1001 trees were created to provide enough coverage of the variable training set. In a Random Forest classification, a value in the form of a probability is output for each lifestyle. This value corresponds to the fraction of trees in the Random Forest that predict that particular lifestyle, thus the values vary from 0 to 1. The lifestyle with the higher probability is considered to be the predicted lifestyle for that phage.

2.1.6 Replicate iterations The resulting prediction from a single Random Forest calculation is based on N known phages, which are randomly selected as training cases, and M proteins, which are randomly chosen to create the Similarity Vectors. Because of this random selection of training data, an unknown phage might be predicted as a different lifestyle in each subsequent Random Forest classification. To better account for this variability in predictions, 10 replicates are performed with different training phages and a different set of Query Proteins. Ten replicates are chosen to balance runtime and accuracy. Predictions based on five replicates were less accurate, whereas predictions based on 20 replicates caused runtimes to greatly increase without a concomitant increase in accuracy. The 10 replicate predictions are averaged, and the lifestyle with the higher average is considered the predicted lifestyle of the phage. For some phages, the replicate predictions of which lifestyle they prefer might vary, with some of the replicate predictions voting for one lifestyle and some replicate predictions voting for the other lifestyle. The distribution of these predictions was calculated to be a normal distribution. The final probability score is considered 'confident' if a consensus of the 10 replicate predictions is for one particular lifestyle. To determine whether a prediction was confident, the mean and the SD of the 10 replicate predictions is calculated. The prediction is deemed 'confident'

if the averaged probability score of the predicted lifestyle is 2 SD away from the averaged probability score of the other lifestyle.

2.1.7 Initialization Not all proteins are useful in identifying the class of a phage. To increase the accuracy of predictions, an importance cutoff value was incorporated to include only proteins that are important toward predicting a phage's class into the creation of the set of Query Proteins Q. A similarity vector is created for each temperate and virulent phage. This set of similarity vectors is used by the Random Forest algorithm to calculate the Gini importance values (also known as the Gini-coefficient) for all the proteins in the database that belong to the phage with an annotated lifestyle. The Gini importance value is a measure of how important a protein is toward classifying a phage's lifestyle (Gini, 1912). A Gini value of zero corresponds to perfect equality (unimportant) and a value of one corresponds to perfect inequality (important). This step is only performed when any new phages, and thus new proteins, are added to the database. This importance value for a protein is used during runtime so that only the most important proteins are selected to create the similarity vectors. To empirically determine which proteins to include into PHACTS calculations, the importance cutoff value was set to various percentages at and above the mean, and the 227 phages were classified using these various importance value cutoff values. It was found that an importance cutoff value of twice the mean of the importance values, gave the best results for our dataset, and by excluding less important proteins both the speed and the accuracy increased. To speed up runtime, the percent identity scores of every protein to every other protein are calculated at initialization by the FASTA program, and results are stored in a data structure for optimized retrieval.

2.2 Partial genomes

Datasets were created that consisted of partial proteomes of various sizes. The first dataset contained 1000 partial proteomes that consisted of a single protein. Six more datasets were created by increasing the size of the partial proteomes in increments of five proteins until the final partial proteome dataset consisted of 1000 partial proteomes of 30 proteins. Testing partial proteomes >30 proteins causes a bias, since phages with small genomes become excluded. Each proteome was created by randomly choosing with replacement a phage with a known lifestyle and then randomly selecting a set of contiguous proteins in that phage. The partial proteomes were then used by PHACTS to predict the lifestyle of the phage. Accuracy scores were calculated by dividing the number of confident correct predictions by the total number of confident predictions.

3 RESULTS

3.1 Accuracy of the lifestyle predictions of PHACTS

To test the efficacy of PHACTS toward classifying a phage's lifestyle, each phage with an annotated lifestyle was sequentially removed from the known database, along with any phages that share >90% of their proteins with >90% percent identity, and PHACTS was used to predict its lifestyle. The predicted lifestyle was compared with the actual annotated lifestyle. Out of the 227 phages with a known lifestyle, PHACTS was able to confidently calculate the lifestyle of 199 phages (Fig. 1). The other 28 phages gave variable results, sometime replicates being classified as virulent and other times as temperate. Out of the 199 predictions that were confident, 197 of those predictions were correct, giving PHACTS a precision rate of 99% and sensitivity of 88%, for predicting the lifestyle of a phage. The results for each phage prediction, along with the SD, are listed in Supplementary Table S1.

The two phages that were consistently classified incorrectly were the Mycobacteriophage D29 (28369.1) and the Lactococcal bacteriophage ul36 (114416.1). To find out the reason for the incorrect predictions of D29 and ul36, the genomes of these virulent double-stranded DNA phages were analyzed. Both phages contain an integrase gene, and both of these integrases are indeed functional (Peña *et al.*, 1998; Labrie and Moineau, 2002). The fact that a virulent phage contains a functional integrase is counter to the current idea that only temperate phages contain integrase. In the case of the Mycobacteriophage D29, a truncated repressor gene that is necessary for temperate proliferation is the cause of the strictly virulent lifestyle (Peña *et al.*, 1998), whereas horizontal gene transfer seems to be responsible for the presence of the integrase in the Lactococcal bacteriophage ul36 (Labrie and Moineau, 2002). The reason that the lifestyle of 28 phages could not be predicted confidently was not as straightforward, but most likely, arises by a query phage having low similarity to phages with known lifestyles in the database.

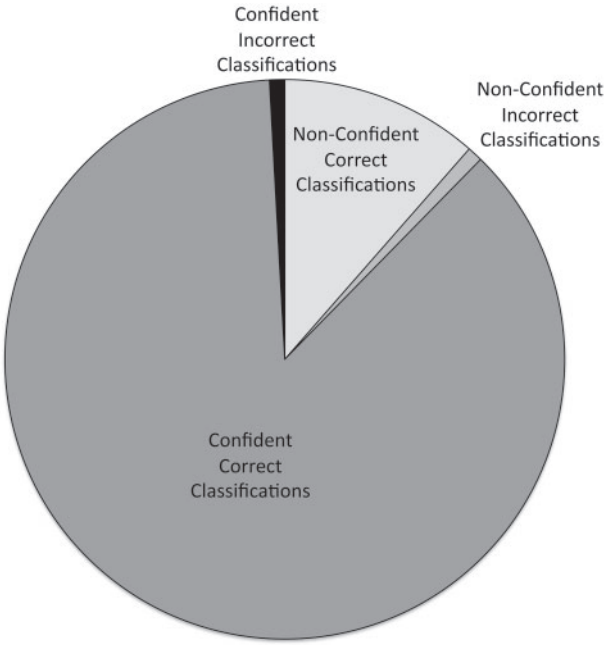

Fig. 1. Accuracy of PHACTS predictions when classifying the lifestyle of the 227 phages with known lifestyles. A confident classification is where the averaged replicate predictions are >2 SD apart.

To determine how the function of a protein correlated to the importance that a protein had on a prediction, the functional role was found for every protein in the Query Protein selection pool from the PHANTOME website (www.phantome.org). Proteins were grouped according to lifestyle, and for each functional role a percent importance value was calculated by summing the Gini importance scores for proteins in that functional role and dividing by the total number of proteins in all functional roles (Fig. 2). Even though a large percentage of the proteins have unknown function, it is clearly visible that Integration/Excision/Lysogeny, Regulation of Expression and Toxins genes are predominantly important toward classiying temperate phages, whereas Nucleotide Metabolism, Phage Lysis and Structural Proteins are predominantly important toward classifing virulent phages. The fact that Structural Proteins are one of the most important functional roles for clasifying

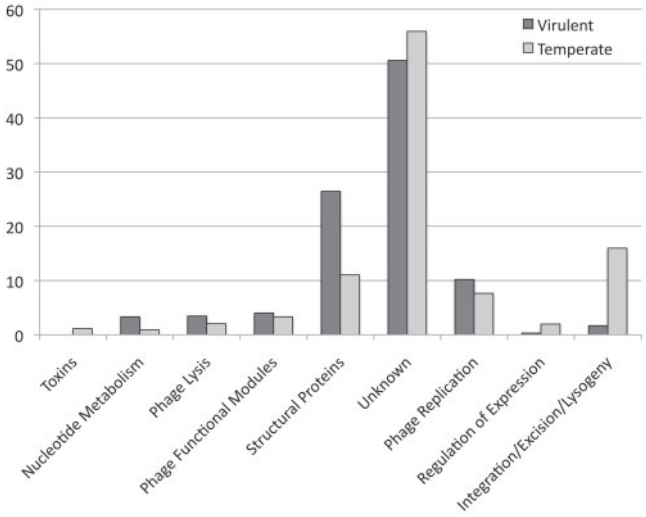

Fig. 2. The correlation between the protein function and the importance toward lifestyle predictions. Phage functional modules are proteins that have functions that are unique to phages, such as capsid assembly or phage DNA packaging.

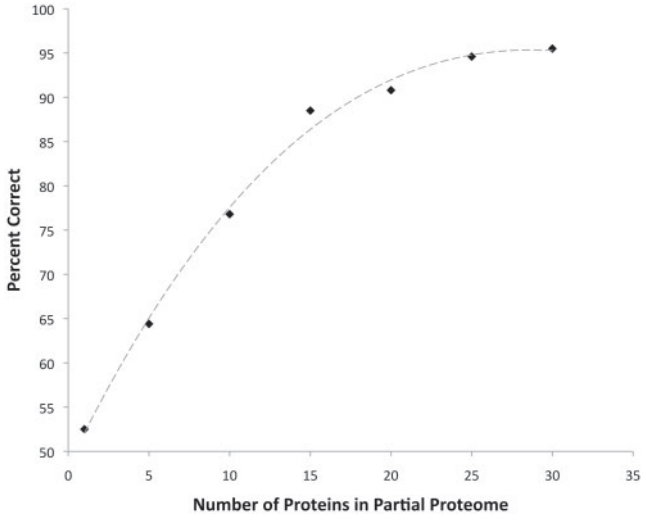

Fig. 3. The effect of incomplete phage proteomes on the accuracy of PHACTS lifestyle predictions.

both temperate and virulent phages shows that by utilizing sequence similaity, PHACTS is able to distinguish between temperate phage proteins and virulent phage proteins even if they share similar functions. These important proteins were compared with the evolutionarily conserved modules found by Lima-Mendez *et al.* (2011) to be associated with a specific lifestyle, and the same correlation between module 1 and virulent phages, and module 17 and temperate phages was observed (Supplementary Fig. S1).

3.2 Classification of partial genomes

PHACTS has been shown to be highly accurate for classifying the lifestyle of complete phage genomes. However, often times only partial genomes are sequenced. To determine how accurate PHACTS predictions are when incomplete proteomes are used, lifestyle predictions were made for phages using only partial proteomes. It was found that with only 20 proteins, PHACTS can identify the lifestyle of a phage with ∼90% precision rate (Fig. 3). The median number of proteins per phage genome in the database was 57 proteins, which suggests that at least a third of a phage's proteome is needed to accurately predict the lifestyle of a phage.

3.3 Classification of unknown phages

The lifestyle of each phage in the database that did not have an annotated lifestyle was predicted by PHACTS using the same methodology as above, but without excluding any phages from the training set (Supplementary Table S1). Out of the 417 phages, PHACTS was able to confidently predict the lifestyles of 217 phages, giving this dataset a specificity of <51%. This drop in specificity suggests that these phages without an annotated lifestyle are more diverse than the subset of phages with a known lifestyle. Also of note was the fact the ratio of phages predicted temperate to phages predicted virulent in this dataset was ∼1:1, which is different from the ratio of 2:1 observed in the set of phages with annotated lifestyles.

4 CONCLUSIONS AND FUTURE WORK

PHACTS provides a mechanism to determine the lifestyle of a phage without having to perform costly and time-consuming experimental lab techniques. PHACTS predictions were shown to have a 99% precision rate, and PHACTS can also determine the lifestyle of a phage using only genomic data, which previously could not be done.

One of the limitations of PHACTS currently is that for a small percentage of phages, a confident lifestyle prediction cannot be made. This is primarily caused by the variability and that arises from the random sampling during classifications. If an unknown phage does not have any similarity to phages with known lifestyles in the database, predictions will be less certain. It is expected that as more phages with known lifestyles are added to the database, the precision rate and sensitivity of predictions will increase.

The web version is simple and easy to use, and the stand-alone version allows for user customization and alternate training sets. The application of PHACTS on different classification schemes (Gram-stain of host and phage Family) has been shown to be moderately successful (data not shown). In the future, refinements to the methodology may lead to high precision rates when classifying the Gram stain of host and phylogenetic Family of phages, as well as other novel classification schemes.

ACKNOWLEDGEMENTS

We thank Drs Ramy Aziz, Elizabeth Dinsdale and Jeff Elhai for useful discussions and comments.

Funding: Advances in Bioinformatics from the National Science Foundation (grant DBI 0850356).

Conflict of Interest: none declared.

REFERENCES

Clarke,K.J. (1998) Virus particle production in lysogenic bacteria exposed to protozoan grazing. *FEMS Microbiol. Lett.*, **166**, 177–180.

Clark,D.W. *et al.* (1986) Effects of growth medium on phage production and induction in Escherichia coli K-12 lambda lysogens. *J. Biotechnol.*, **3**, 271–280.

Deschavanne,P. *et al.* (2010) The use of genomic signature distance between bacteriophages and their hosts displays evolutionary relationships and phage growth cycle determination. *Virol. J.*, **7**, 163–163.

Gini,C.W. (1912) Variability and Mutability, contribution to the study of statistical distributions and relations. *Studi Economico-Giuricici della R. Universita de Cagliari*.

Hendrix,R.W. *et al.* (1999) Evolutionary relationships among diverse bacteriophages and prophages: All the world's a phage. *Proc. Natl Acad. Sci. USA*, **96**, 2192–2197.

Housby,J.N. and Mann,N.H. (2009) Phage therapy. *Drug Discov. Today*, **14**, 536–540.

Labrie,S. and Moineau,S. (2002) Complete Genomic Sequence of Bacteriophage ul36: Demonstration of Phage Heterogeneity within the P335 Quasi-Species of Lactococcal Phages. *Virology*, **296**, 308–320.

Liaw,A. and Wiener,M. (2002). Classification and Regression by random Forest. *R News*, **2**, 18–22.

Lima-Mendez,G. *et al.* (2008) Reticulate Representation of Evolutionary and Functional Relationships between Phage Genomes. *Mol. Biol. Evol.*, **25**, 762–777.

Lima-Mendez,G. *et al.* (2011) A modular view of the bacteriophage genomic space: identification of host and lifestyle marker modules. *Res. Microbiol.*, **162**, 737–746.

Peña,C.E.A. *et al.* (1998) Mycobacteriophage D29 integrase-mediated recombination: specificity of mycobacteriophage integration. *Gene*, **225**, 143–151.

Proux,C. *et al.* (2002) The dilemma of phage taxonomy illustrated by comparative genomics of Sfi21-like Siphoviridae in lactic acid bacteria. *J. Bacteriol.*, **184**, 6026–6036.

Pearson,W.R. and Lipman,D.J. (1988) Improved tools for biological sequence comparison. In *Proceedings of the National Academy of Sciences*, **85**, pp. 2444–2448.

Rohwer,F. (2003) Global Phage Diversity. *Cell*, **113**, 141.

Rohwer,F. and Edwards,R. (2002) The Phage Proteomic Tree: a Genome-Based Taxonomy for Phage. *J. Bacteriol.*, **184**, 4529–4535.

Srinivasiah,S. *et al.* (2008) Phages across the biosphere: contrasts of viruses in soil and aquatic environments. *Res. Microbiol.*, **159**, 349–357.

Whitman,W.B. *et al.* (1998) Prokaryotes: The unseen majority. *Proc. Natl Acad. Sci.*, **95**, 6578–6583.

Witkin,E.M. (1976) Ultraviolet mutagenesis and inducible DNA repair in Escherichia coli. *Bacteriol. Rev.*, **40**, 869–907.

Wommack,K.E. and Colwell,R.R. (2000) Virioplankton: Viruses in Aquatic Ecosystems. *Microbiol. Mol. Biol. Rev.*, **64**, 69–114.