BioTechnology: An Indian Journal

Review|Vol18Iss01

Bacteriophage Host Ranges Prediction by Computation

Nikolas Brown*

University of Ljubljana, Slovania

***Corresponding author**: Nikolas Brown, University of Ljubljana, Slovania. E-Mail: brown.nikolas69671@gmail.com

Received date: January 03, 2022; **Accepted date:** January 06, 2022; **Published date:** January 29, 2022

Abstract

The development of bacteriophage agents for a variety of uses in agriculture, biotechnology, and medicine has been motivated by increased antibiotic resistance. The host range of a bacteriophage, or the bacterial genera, species, and strains that a bacteriophage can infect, is an important consideration in the selection of agents for these applications. Although experimental host range studies are still the gold standard, they are naturally confined to a small number of viruses and bacteria that can be cultivated. We present a review of recently developed bioinformatic methods that provide a potential and high-throughput alternative by computationally predicting the possible host ranges of bacteriophages, even those that are difficult to culture in laboratory settings.

***Keywords:** Bacteriophages; Bacterial hosts; Host ranges; Bioinformatics tools*

Introduction

On Earth, there are roughly 10^{31} viruses, which is more than the number of stars in the visible universe. Bacteriophages, or viruses that infect and feast on bacteria, make up the great majority of this varied virosphere. Frederick William Twort and Félix d'Herelle independently discovered these abundant biological entities in the early 1900s, and they have since been routinely used for a variety of purposes, including diagnostics, drug design and discovery, vaccine development, agriculture, food preservation and safety, and wastewater treatment. To use bacteriophages' bactericidal effects in these applications, bacterial host ranges (i.e., collections of bacterial species and strains that sustain the bacteriophage's life cycle) must be established. The study of bacteriophage–host connections is possible because to a variety of experimental approaches (such as spot, plaque, and liquid assays, viral tagging, microfluidic PCR, phageFISH, and single-cell genomics). They are, however, frequently time and labour consuming, costly, and scientifically challenging (e.g., due to inconclusive or absent signs of infection). Due to both the bacterial cultures utilised in the experiments—a limited number of microbial hosts and viruses are amenable to cultivation—and the conditions under which they are done in the

laboratory, these approaches are inherently limited in scope. Recent developments in sequencing technologies have made it possible to locate and identify bacteriophages and their hosts from environmental (rather than cultured) samples, opening up a new path for studying natural viral diversity. Many bioinformatic algorithms have been developed in tandem with these technological breakthroughs to computationally forecast possible virus host ranges on a wide scale, based on genetic characteristics shared by bacteriophages and their bacterial hosts during their co-evolution over time. Although predictive in nature, these approaches can help identify the most promising candidates for further testing to confirm the bacteriophage's capacity to recognise and adsorb to the host, as well as define infection cycles, bacteriophage–host interactions, and lysis efficacy. In this study, we address the similarities and variations in the design of many computational host prediction methods, as well as essential issues to keep in mind while choose between them.

Computational Methods for Predicting Bacteriophage Host Ranges

Bioinformatic approaches to computationally predict putative bacteriophage host ranges can be broadly classified into three categories: (i) alignment-based methods based on sequence homology and sequence similarity, (ii) alignment-free methods based on sequence composition and genomic features, and (iii) machine-learning-based methods.

Alignment-Based Methods

The host specificity of bacteriophages is influenced by a variety of circumstances. As lysogenic prophages, temperate bacteriophages can integrate their own genomes into those of their bacterial hosts. This process frequently changes the host's phenotypic, which can result in higher fitness (for example, by providing antibiotic resistance, boosting virulence, creating toxins, or preventing additional (super) infections; see Touchon and colleagues' overview). At the same time, many bacterial hosts use a range of restriction-modification (RM) and clustered regularly interspaced short palindromic repeats (CRISPRs)/Cas (CRISPR-associated protein) techniques to defend themselves against aggressive bacteriophages and other invaders. In the latter situation, following infection (adaptation), a stretch of nucleotides from the invasive genetic material is incorporated into a CRISPR spacer array, which is then used as a guide to construct site-specific cleavages, ultimately leading to the demise of the invading bacteriophage (immunity). In both instances, the invading bacteriophage alters or causes changes to the host genome. Sequence homology (i.e., the common evolutionary lineage between sequences) and sequence similarity are used in alignment-based approaches to predict host ranges from sequence homology and similarity. Many alignment-based methods are simple to use, such as the most well-known example, the Basic Local Alignment Search Tool (BLAST), which compares a user-provided viral sequence with those of potential bacterial hosts publically available in well-maintained (reference) databases. As a result, the comprehensiveness and completeness of the employed datasets limit the inference of virus–host connections using alignment-based approaches.

Databases of Bacteriophages and their Hosts

When possible, experimental evidence obtained through bacteriophage isolation and cultivation is used to determine bacteriophage host ranges. Experimental validation, on the other hand, is often time and labour intensive. Between the initial prediction and clear experimental confirmation that crAssphage–a highly prevalent

bacteriophage in the human gut microbiome–can infect bacteria of the species Bacteroides–nearly half a decade passed. As a result, data on bacteriophage–host relationships is scarce, with information in the well-known National Center for Biotechnology Information (NCBI) RefSeq and GenBank databases frequently restricted to the genus and/or species level or limited to a few of samples. The Viral Host Range database (VHRdb), a web-based platform that incorporates host range data as an analysis tool and search engine, was recently launched with the goal of collecting more data by allowing researchers to directly share their experimental findings with the scientific community (at the time of writing, 16,715 interactions between 760 viruses and 1923 hosts have been recorded). Bacteriophage–host databases, such as VHRdb, are likely to play an important role in the development of future machine learning approaches, given the necessity for validated training datasets.

Key considerations

Prediction accuracy

In addition to their underlying algorithms, bacteriophage–host prediction systems differ in their prediction accuracy, or the percentage of bacteriophages for whom the taxonomy of their predicted and known hosts agree. Prediction accuracy can be given at many taxonomic levels, including family, genus, and species, as well as phylum and domain levels. When choosing the best instrument for any investigation, it's crucial to think about which taxonomic levels were measured. Variables in prediction accuracy between tools can also be attributed to methodological differences (such as the type of data used in the benchmarking process).

As a result, comparisons should ideally be made using a standardised benchmarking dataset. Zielezinski and colleagues compared a number of alignment-based, alignment-free, and ML-based host-range prediction techniques using such consistent benchmarking data, revealing that strategies based on sequence homology have a greater predictive accuracy than those based on sequence composition similarity. The non-uniform number of microbial species found in a metagenomic sample is a barrier for researchers dealing with environmental samples. Metagenomic samples often result in diverse read coverage profiles across various genomes since sequencing technologies are optimised for moderate- to high-coverage individual samples. Contigs (a gapless length of nucleotide sequence created by overlapping sequencing reads) derived from metagenomic samples are usually short as a result of these variations, resulting in fragmented and/or incomplete genome assemblies. Short viral contigs (less than 10 kb) have a considerable decline in prediction accuracy, which is a non-negligible factor in most tools' prediction accuracy. WIsH, a tool that matches VirHostMatcher's full-length genome prediction accuracy with only 3 kb of nucleotide sequence, has established itself as an alignment-free option for samples including small viral contigs.

Usability

Operating system limitations can play a big role in deciding which bacteriophage–host prediction method to use. The bulk of prediction programmes rely on the command line interface (CLI) embedded into UNIX-based operating systems to permit both automation and reproducibility (such as Linux and macOS). As a result, users of other operating systems (such as Windows and Chrome OS) will need to either buy a dedicated workstation or install the required operating system on an existing machine, such as via dual boot or virtualization. Users of Windows can also utilise the Windows Subsystem for Linux (WSL) to run native Linux programmes on their computers. Prediction tools on the web (such as HostPhinder and PHP) are a good option. Web-based tools, in

addition to being user-friendly and straightforward, eliminate the hassle of installation and potential dependence concerns by requiring only a suitable browser. However, one of the biggest disadvantages of web-based tools is that they have a limit on the amount of data that can be entered. The web-based version of PHP, for example, is restricted to 100 viruses, whereas the standalone version can analyse datasets that are orders of magnitude larger. Multi-threading is an additional benefit of many phage–host prediction CLI programmes (including Phirbo, WIsH, and VirHostMatcher-Net), which speeds up the studies.

Conclusion

Bacteriophages are now frequently employed for a variety of biotechnological and therapeutic objectives, including individualised phage therapy to treat multi-drug resistant illnesses, thanks to their bactericidal powers. Although large-scale bacteriophage banks (such as the Phage Directory) provide a wide range of bacteriophages to the scientific community, knowing the host range that a bacteriophage can infect is necessary to successfully guide the use of bacteriophages in various disciplines. The gold standard for experimentally characterising host ranges is still phage isolation and cultivation. They are, however, time-consuming and consequently unsuitable for large-scale analysis. Recently developed computer prediction methods offer a promising alternative, allowing researchers to narrow down the vast number of prospective hosts to a small number that can be tested in a laboratory setting feasible (and more cost-effectively). Because different tools use different strategies to predict bacteriophage–host relationships, each with its own set of benefits and drawbacks, using multiple, complementary prediction tools can aid in the selection of the most promising candidates, especially for bacteriophages with broad host ranges. If time and computational resources allow, a three-way combination of alignment-based, alignment-free, and machine learning approaches could be used to select those that have been predicted by all three strategies for experimental validation and characterization of infection cycles and bacteriophage–host interactions. Although there is still a lot more to learn about bacteriophages and their hosts, developments in genomic databases, machine learning, and high-performance computing have started to pave the way for even more complex and accurate computational methods in the near future.

REFERENCES

1. Hussain B. Modernization in plant breeding approaches for improving biotic stress resistance in crop plants. Turk J Agric For. 2015 Jul 23;39(4):515-30. [**Google Scholar**] [**CrossRef**]
2. Lin Z, Cogan NO, Pembleton LW, Spangenberg GC, Forster JW, Hayes BJ, Daetwyler HD. Genetic gain and inbreeding from genomic selection in a simulated commercial breeding program for perennial ryegrass. Plant Genome. 2016 Mar;9(1):plantgenome2015-06. [**Google Scholar**] [**CrossRef**]
3. Moose SP, Mumm RH. Molecular plant breeding as the foundation for 21st century crop improvement. Plant phys. 2008 Jul;147(3):969-77. [**Google Scholar**] [**CrossRef**]
4. Bohra A, Chand Jha U, Godwin ID, Kumar Varshney R. Genomic interventions for sustainable agriculture. Plant Biotech J. 2020 Dec;18(12):2388-405. [**Google Scholar**] [**CrossRef**]
5. Sinha P, Singh VK, Bohra A, Kumar A, Reif JC, Varshney RK. Genomics and breeding innovations for enhancing genetic gain for climate resilience and nutrition traits. Theor Appl Genet. 2021 Jun;134(6):1829-43. [**Google Scholar**] [**CrossRef**]
6. Varshney RK, Bohra A, Yu J, Graner A, Zhang Q, Sorrells ME. Designing future crops: genomics-assisted breeding comes of age. Trends Plant Sci. 2021 Jun 1;26(6):631-49. [**Google Scholar**] [**CrossRef**]