VISUALIZING BACTERIOPHAGE EVOLUTION THROUGH SEQUENCE AND STRUCTURAL PHYLOGENY OF LYSIN A AND TERMINASE PROTEINS

An Analysis of Protein Structure across Phage Clusters

Abstract

Understanding how genes evolve and persist is a critical part of viral genomics. Bacteriophages can provide unique insight about viral evolution because of their abundance and largely unexplored history. Traditionally, phylogenetic trees have used DNA sequence comparison to visualize evolutionary paths between organisms. However, DNA sequence similarity does not reflect key alterations to protein structure and therefore how the protein performs its function. Phylogenetic trees based on predicted protein structure could provide an alternative lens through which to view evolutionary paths.

From each of the 10 largest clusters included in the Actinobacteriophage Database, three mycobacteriophage genomes were selected. Lysin A and terminase proteins are encoded by all of the mycobacteriophage genomes and were therefore selected for analysis. Protein structural predictions were generated from amino acid sequences using Phyre2 and compared with the PyMol Molecular Graphics System, Version 2.0. Structural alignment scores from PyMol were used to quantify the structural homology of lysin A and terminase across different clusters. Five phylogenetic trees were constructed: one was based on structural homology of lysin A, one was based on structural homology of terminase, two were based on amino acid sequence of these individual proteins, and one was based on overall genomic sequence alignment. Phylogenetic trees were compared to evaluate differences between amino acid sequence and structural homology. Visualizing the predicted relationships from amino acid sequences and structural analysis of phage proteins will provide a new perspective on the evolution of the virosphere.

Keywords

biotechnology, bioinformatics, evolution, bacteriophage

Student Authors

MAANSI ASTHANA is a biological engineering student with minors in Spanish, biotechnology, and biological sciences. She currently serves as an undergraduate teaching assistant for the Purdue Agricultural and Biological Engineering (ABE)

program's biotechnology laboratories and has conducted research under Dr. Elsje Pienaar in the Biomedical Engineering Department for the past year and a half. Asthana is currently working with researchers across the globe on a project to develop an agent-based model of SARS-CoV-2. After graduation, Maansi plans to pursue her MD degree, ultimately practicing medicine.

ALYSSA EASTON is a biological engineering student with minors in French and biotechnology and has been an undergraduate teaching assistant for the ABE's bioinformatics courses since the fall of 2020. She is a

Martin Agricultural Research Scholar in biochemistry and has worked with Dr. Hana Hall for two years investigating R-loops as a mechanism of age-related neurodegeneration in photoreceptor neurons. Easton plans to pursue an MD-PhD to continue studying epigenetics and neurodegeneration as a medical researcher.

JULIA MOLLENHAUER is a biological engineering student with minors in Spanish and biotechnology. She is the current president of Purdue's chapter of the American Society of Agricultural and Biological Engineers and is also an undergrad-

uate teaching assistant for biotechnology courses in the ABE department. Over the past two years, Mollenhauer has completed internships in R&D and engineering with Grande Cheese Company in Wisconsin. After graduation, she plans to apply to yearlong international research programs prior to enrolling in graduate school.

SEAN RENWICK is a biological engineering student with minors in German, biotechnology, and global engineering. He is currently doing research in Professor Torbert Rocheford's lab investigating bio-fortified corn with higher vitamin A

content; his main contributions involve creating programs with which Professor Rocheford and his team can objectively choose corn for breeding. Following graduation, Renwick hopes to find work in industry in Switzerland.

ANITA GOPALRATHNAM is a biological engineering student with minors in Spanish and biotechnology. She has served as an undergraduate teaching assistant for the biotechnology courses within the ABE program. In the spring of 2019,

Gopalrathnam participated in a study abroad program in Quito and Cuenca, Ecuador, where she was able to develop a prosthetic for a dog's femur and introduce several light boards to aid in neural stimulation of adolescents affected by cerebral palsy and other audio/visual deficits. She plans on pursuing her MAs and PhD in biomedical engineering and work in the fields of biotechnology and medical devices.

Mentors

GILLIAN SMITH is a graduate teaching assistant in agricultural and biological engineering and teaches two biotechnology laboratories. She is also a second-year MA student in agricultural and biological engineering studying bioengineering and

biotechnology. In addition to being a teaching assistant, Smith does research in bacteriophage proteomics and lipidomics using mass spectrometry as well as in STEM education.

KARI CLASE is a professor in the Department of Agricultural and Biological Engineering in the College of Agriculture and College of Engineering at Purdue University. She is also the director of the Biotechnology Innovation and

Regulatory Science (BIRS) Center, whose mission is to develop global programs to ensure sustainable access to medicines for Africa and developing nations and to advance discovery in manufacturing technology, quality of medicines, and rare disease research. Clase teaches multiple courses covering topics in biotechnology, bioinformatics, drug discovery, and development to engineers, scientists, and technologists. Her currently funded projects include collaborators from multiple disciplines and an impact that spans K–12 to graduate education.

INTRODUCTION

Bacteriophage—viruses that infect bacteria—are the most abundant biological entities on Earth and have been evolving for billions of years. Bacteriophages are a compelling research topic due to their diversity and vast potential applications. The bacteriophage population is incredibly diverse as a result of both vertical and horizontal genetic exchange. Vertical exchange is the transfer of genetic material to phage progeny, while horizontal exchange is the movement of genetic material between phage (Hatfull, 2008). Understanding the evolution of phages presents a unique challenge because there is no historical record prior to their discovery in 1915. The SEA-PHAGES program sponsored by Howard Hughes Medical Institute (HHMI) aims to build a bacteriophage record through an undergraduate research course held at universities worldwide. Students collect and purify bacteriophage samples from soil and send the samples to HHMI. The SEA-PHAGES initiative has led to the development of a large database of bacteriophage genomes worldwide, paving the way for a better understanding of our biosphere.

Understanding the evolutionary relationships between bacteriophages helps create a clearer portrait of the virosphere, simplifying the phage-selection process for antibiotic resistance and drug delivery applications. Bacteriophages have been proposed as a viable alternative to antibiotic treatment, especially in the case of antibiotic resistance (Bragg et al., 2014). Due to the imminent threat of antibiotic resistance, there is an urgent need to understand how bacteriophage may be used to target pathogenic bacteria by breaking down the bacterial cell wall (Bragg et al., 2014). Bacteriophages have also been proposed as vessels for drug delivery. New techniques in synthetic biology could allow the genetic engineering of bacteriophages to precisely deliver pharmaceuticals (Garg, 2019). Precision drug delivery could improve upon current treatment methods. For example, phage-mediated cancer therapy could reduce the need for nonspecific treatments such as chemotherapy (Garg, 2019).

BACKGROUND

A bacteriophage cluster is a group of bacteriophages with high overall genomic similarity, meaning they share

similar genetic material. Bacteriophages in the same cluster tend to encode similar proteins, although there may be slight variation in their genomic sequences that cause differences in the final protein structures. While there is typically conservation of genes for phages contained in the same cluster, many genes are present in all bacteriophage clusters, as they are essential to phage function and reproduction. In this study, two proteins that are present in all mycobacteriophages were chosen for further investigation, lysin A and terminase. Lysin A is an endolysin protein that hydrolyzes the bacterial membrane, resulting in cell lysis and the release of more phages (Pohane, Joshi, & Jain, 2014). Terminase is a protein that recognizes phage DNA and initiates DNA packaging during the formation of new phages (Shen et al., 2012). Due to their conservation across all phages, comparative sequence and structural analysis of these proteins can provide insight into the evolution and diversification of bacteriophages.

Phylogeny, which is the prediction of evolutionary history, may be applied to bacteriophages. Phylogeny can be visually presented with phylogenetic trees in which organisms that are closely related evolutionarily are closer and more distantly related organisms branch off farther away. Different methods of building these trees exist. The neighbor-joining method is a simpler method that takes a matrix of divergence values between pairs of organisms and organizes them based on each respective divergence (Saitou & Nei, 1987). This method, while fast, has a tendency to make flawed trees. Maximum likelihood is a more complex method that accounts for the likelihood of certain mutations to occur when assessing phylogeny, which yields more accurate phylogenetic trees (Guindon et al., 2010).

The values used to build these trees are usually determined through the comparison of either amino acid or nucleic acid sequences. However, structural comparison between proteins may be an alternative method for predicting evolutionary relationships. Programs exist to predict the structure of proteins based on their amino acid sequence. Hidden Markov models use a probabilistic method to create complex protein models using intuition and statistics. The algorithm decides on the next most probable outcome based on its current state, optimizing the overall output through step-by-step analysis and ultimately working to output the most likely

protein structure (Eddy, 2004). Protein comparison and analysis can be done by superimposing two predicted protein structures and using an alignment tool to quantify structural homology.

Our investigation of protein conservation was inspired by studies conducted by Graham Hatfull and Roger Hendrix (Hatfull & Hendrix, 2011). In their fundamental essay, Hatfull and Hendrix discuss a hallmark feature of bacteriophage genomes: mosaicism. They found that sections of bacteriophage genomes have different evolutionary histories due to high levels of horizontal exchange. Different genes' mobility varies, however, because genes encoding for proteins that interact "travel together." We can identify mosaicism by comparing the particular genes' evolution with the entire genome. In addition, generating phylogenetic trees using two methods and comparing findings could indicate the efficacy of protein structure comparison in determining phage evolutionary trends. Our goal with this research was to determine if evolution of bacteriophages could be analyzed through their protein structures instead of their amino acid and nucleic acid sequences.

METHODS

Cluster and Phage Selection

Ten mycobacteriophage clusters were chosen with a high number of nondraft genomes in the Actinobacteriophage database (Russell & Hatfull, 2016). If the overarching cluster had three subclusters of at least 20 nondraft members, the genomes were chosen from three different subclusters. All samples selected were isolated in 2016 or later.

Amino Acid and Nucleic Acid Sequence Prediction

As a control, a whole-genome dot plot analysis was executed using the program Gepard (Krumisiek, Arnold, & Rattei, 2007). The resulting dot matrix reveals DNA sequence homology in the form of dots; comparing two identical sequences results in a diagonal black line.

An amino acid sequence-based phylogenetic tree was created for both terminase and lysin A proteins. These

sequence homology trees were generated by Phylogeny.fr (Dereeper et al., 2008), which runs MUSCLE multiple sequence alignment, followed by PhyML for tree building (Guindon et al., 2010).

Structural Prediction

Protein structure was predicted using Phyre2, an open-access structural prediction software that utilizes hidden Markov models (Kelley, Mezulis, Yates, Wass, & Sternberg, 2015). These structural predictions were aligned on Pymol using the superalign function, which benchmarks structural homology. The superalign scores were collected for each pair of proteins. Higher superalign scores indicated higher structural similarity; however, the neighbor-joining algorithm in the PHYLIP computer program assumes that smaller scores indicate higher similarity. Therefore, the inverse of each superalign score was organized into a matrix and input in PHYLIP's neighbor-joining algorithm to generate a phylogenetic tree (Felsenstein, 2005).

Comparative Analysis

The trees created using nucleic acid sequence, protein amino acid sequence, and protein structural alignment scores were compared to determine the viability of protein structural alignment as a means of predicting evolutionary relationships between phages.

RESULTS

Based on cluster and phage genome selection criteria, the phages shown in Table 1 were selected for analysis. For each phage, the whole genome, the lysin A amino acid sequence, and the terminase amino acid sequence were collected.

Overall Nucleic Acid Sequence Similarity

A whole-genome dot plot was developed to visualize the overall evolutionary relationship between the phages. The dot plot is a two-dimensional matrix with sequence comparison along both the horizontal and vertical axes

TABLE 1. Cluster and phage selections from the Actinobacteriophage Database.

Cluster	Phage Name	Year Found
B1	Chaelin	2019
B2	BananaFish	2018
B3	Rita1961	2018
C1	Ronan	2019
	Quasimodo	2019
	Janiyra	2019
E	Tomaszewski	2018
	Gator	2018
	Flypotenuse	2018
F1	Sebastian	2016
	JoeyJr	2017
	Donkeykong	2018
G1	Jonghyun	2018
	Rabbs	2017
	Olga	2017
J1	Hanaconda	2016
	Dallas	2018
	Ejimix	2018
K1	Beezoo	2016
K4	Patt	2017
K6	Yuna	2018
L1	Acquire49	2016
L2	Gabriela	2018
L3	Kingsolomon	2016
N	Fulbright	2017
	Chewbacca	2017
	Smurph	2018
P1	Camster	2019
	Necropolis	2018
	Megiddo	2018

(Smith et al., 2013). The dot plot is shaded based on regions of homology, with darker regions indicating more similarity. Therefore, identical sequences are shown as diagonal black lines. The dot plot in Figure 1 shows all 30 phage genomic sequences and provides a baseline genomic similarity that was later used for comparative analysis.

The conservation of phage genomes within a cluster was analyzed along the diagonal of the dot plot. Clusters with

FIGURE 1. Dot plot comparison of entire genomic sequence from 30 mycobacteriophage across 10 major clusters. Notably, clusters F, L, and J appear the least similar to other clusters, while clusters K, B, G, N, and P are more similar. Plot generated using Gepard (Krumsiek et al., 2007).

darker shading along the diagonal suggest that the genomes of phages selected from that cluster are highly conserved, such as in clusters K and G. The conservation of phage genomes between clusters was identified at the intersection of two clusters along the horizontal and vertical axes. The intersection of clusters P and K had the darkest shading, suggesting high conservation between the phage genomes in those clusters. Clusters P, G, B, and K all showed higher conservation than other clusters compared in this dot plot.

Amino Acid Sequence Prediction

Phylogenetic trees were then developed using the amino acid sequence of the proteins of interest using the program PhyML (Guindon et al., 2010). The trees for terminase and lysin A are shown in Figures 2A and 2B, respectively.

Cluster relationships were highly conserved for both lysin A and terminase when relationships were

FIGURE 2. The above phylogenetic trees were built using PhyML based on the amino acid sequence of either lysin A or terminase. (A) Phylogenetic tree predicted by PhyML (Guindon & Gascuel, 2003) using amino acid sequences of lysin A genes. Clusters K (red) and L (navy) diverge from one another, while cluster B (purple) has even more variation. These variations from overall genomic similarity (which designate clusters) may suggest recent horizontal gene transfer. (B) Phylogenetic tree predicted by PhyML (Guindon & Gascuel, 2003) using amino acid sequences of terminase genes. Cluster relationships are widely conserved.

predicted with amino acid sequence. Phages belonging to a cluster were in close proximity on the tree, reflecting the basis of clusters: overall genetic similarity. For lysin A proteins, the majority of the phage clusters stayed together in the amino acid phylogenetic tree (see Figure 2A). However, clusters K (red) and L (navy) diverged from each other, while cluster B (purple) showed an even greater divergence from the cluster conservation observed for the other lysin A clusters tested and the terminase proteins.

When the two trees were compared to each other, the terminase sequence was found to be more conserved within clusters than the lysin A proteins. In other words, groups of similar phages showed little variation in the amino acid sequence of terminase and slightly more variation in that of lysin A. This difference could be attributed to the variation of catalytic domains in the lysin A protein. Lysin A proteins are made up of a combination of catalytic domains, and these domains can vary between phages, giving lysin A more opportunity for variation: most endolysins contain two or more catalytic domains and one cell-binding domain (Fischetti, 2008). The cell-binding domain is conserved because all mycobacteriophages target mycobacterium

specifically, but there may be different combinations of catalytic domains between mycobacteriophages. Terminase, on the other hand, performs a very specific function and has less room for variation (Shen et al., 2012).

Structural Predictions

Structural predictions for each phage's terminase and lysin A protein from Phyre2 were visualized in Pymol. Protein structure visuals were superimposed, as shown in Figure 3, to determine the structural similarity of lysin A or terminase proteins produced by each phage. A superalign function on Pymol generated quantitative scores to measure structural alignment where high alignment scores indicate higher structural similarity, which can be confirmed visually.

The inverse of alignment scores for each phage combination were input in PHYLIP's neighbor-joining algorithm to generate phylogenetic trees from structural alignment. The resulting trees are shown in Figure 4. The neighbor-joining method identifies pairs of operational taxonomic units that minimize the branch length at

FIGURE 3. The image above shows two examples of proteins being superimposed in PyMol. (A) Structural predictions of lysin A proteins from mycobacteriophage Acquire49 and Rita1961 aligned in the PyMol Molecular Graphics System. These models had a structural alignment score of 338.382, indicating high structural similarity. (B) Structural predictions of lysin A proteins from mycobacteriophage Acquire49 and Yuna aligned in the PyMOL Molecular Graphics System. These models had a structural alignment score of 78.41, indicating low structural similarity.

FIGURE 4. The trees above are built based on the phage's superalign scores. (A) Phylogenetic tree built with the neighbor function in PHYLIP based on superalign scores between lysin A proteins phage. (B) Phylogenetic tree built with the neighbor function in PHYLIP based on superalign scores between terminase proteins of each phage.

stages of clustering, quickly identifying the branch lengths (Saitou & Nei, 1987).

As observed from Figures 4A and 4B, cluster relationships were generally conserved between both lysin A and terminase proteins. The most highly conserved clusters for lysin A were clusters G, N, and P, while the remaining clusters showed some evolutionary divergence. However, the terminase proteins showed strong conservation between the clusters. Most strong cluster conservation is observed in clusters B, G, L, and P. The remaining clusters demonstrated general conservation, with two of the three phages closely related.

Comparative Analysis and Discussion

Generally, the phylogenetic trees based on structural alignment conserved cluster relationships as well, as indicated by clusters G and P, which were conserved in both Figure 4A and Figure 4B. Many clusters, such as L and B, only had minimal protein structural variation in

either lysin A or terminase but not both. There was less structural variation of terminase within clusters than of lysin A (see Figure 4B), which was to be expected because this trend was also seen in the trees based on amino acid sequence (see Figure 2). The fact that clusters were loosely maintained and that similar trends were seen between amino acid sequence phylogeny and structural phylogeny reflects a key motif in biology: structure determines function.

Some of the variation in cluster conservation in Figure 4 can be attributed to error introduced by the structural prediction algorithm used, Phyre2. In general, the protein models from Phyre2 yielded a wide range of coverage, many as low as 30%. This resulted in incomplete models, an error compounded as structural analysis was conducted.

Error was also likely introduced to the structure-based trees by the neighbor-joining method, which is a very basic algorithm. By contrast, the amino acid-based trees were predicted by PhyML, which uses a maximum

likelihood algorithm (Guindon & Gascuel, 2003). The maximum likelihood algorithm is a more modern improvement upon simpler methods such as neighbor-joining. The use of different methods may have resulted in inconsistencies in phylogenetic tree prediction.

There are some notable differences between the structural alignment trees and amino acid trees. For example, Figure 4 demonstrates that the phage Dallas is structurally distant from the rest of its cluster and is closely related to Tomaszewski. This relationship is conserved in the lysin A and terminase phylogenetic trees, suggesting that there may be an evolutionary relationship between the two phages that was not obvious from only amino acid sequence. This is especially relevant to phage genomes because evolution occurs through not inherited mutations but also horizontal transfer, or mosaicism. In terms of evolution, one change in an amino acid has the potential to significantly change the structure of a protein, thus affecting the protein's function. A small change in the DNA sequence could alter the properties of the amino acid, causing a drastic change in structure. For this reason, structural comparison could provide valuable insight into evolutionary history. The structural similarities between both lysin A and terminase structures in Dallas and Tomaszewski, despite a lack of sequence similarity, may suggest evolutionary convergence to these structures.

Furthermore, cluster organization is based on overall genetic similarity but is constantly changing; some clusters have almost 100% sequence alignment, while others have almost none (Hatfull et al., 2010). Hatfull and his team developed clusters with the goal of organization, not to demonstrate phylogeny, so while there are evolutionary trends, as seen in the trees based on amino acid sequence (see Figure 2), this is not always the case. Therefore, while using clusters to track the consistency of structure-based and sequence-based phylogenetic trees is useful, the clusters themselves should not be used to assess the evolution of the phage. It is also important to note that comparing overall genetic similarity is not sufficient, because bacteriophage can transfer their DNA to other phages. Thus, structural analysis could be useful in understanding evolutionary changes not reflected in the analysis of DNA and amino acid sequences.

CONCLUSION

Cluster conservation was observed across all phylogenetic trees generated (see Figures 1, 2, and 4). Notably, stronger conservation was shown in the terminase protein compared to lysin A, which may be attributed to lysin A's diversity in catalytic domains (Fischetti, 2008). There may also be less room for variation in terminase because its function is crucial to phage replication. Terminase's higher cluster conservation was observed in both the amino acid-based phylogenetic trees and the structural alignment trees.

While all trees demonstrated a degree of cluster conservation, predicted evolutionary relationships between clusters varied for the structural alignment and amino acid sequence-based trees. Therefore, the trees generated from structural alignment could help validate cluster relationships and help determine phages' evolutionary history. Further investigation of structural phylogenetic trees may determine their effectiveness as tools in predicting evolutionary relationships between bacteriophages.

Further analysis should be conducted using the structural prediction protein server I-TASSER, a program that has proven to provide robust and accurate protein predictions. I-TASSER, ranked best in automated 3D structure prediction by the Protein Structure Prediction Center (Zhang, n.d.), uses a three-step algorithm that optimizes probability and uses verified template protein sequences as a basis for the prediction. The algorithm then assembles full-length protein models using *ab initio* modeling and maximizing the low free-energy states, which are identified by the database SPICKER. Structural outputs from I-TASSER may be compared in PyMol using the same superalign function. Then, trees generated using the neighbor-joining method could be compared with the amino acid sequence trees. They could also be compared to the structural trees generated from Phyre2 to assess the significance of the changes in output.

ACKNOWLEDGMENTS

First, we would like to thank Dr. Kari Clase for her support, encouragement, and guidance over the past two years. We thank graduate students Gillian Smith and

Emily Kersteins for their mentorship and Shruthi Garimella for her insight. Finally, we thank the SEA-PHAGES program at HHMI for bringing this research experience to the classroom.

REFERENCES

- Bragg, R.t, van der Westhuizen, W., Lee, J-Y., Coetsee, E., & Boucher, C. (2014). Bacteriophages as potential treatment option for antibiotic resistant bacteria. *Advances in Experimental Medicine and Biology*, 807, 97–110. https://doi.org/10.1007/978-81-322-1777-0_7
- Dereeper, A., Guignon, V., Blanc, G., Audic, S., Buffet, S., Chevenet, F., . . . Gascuel, O. (2008). Phylogeny.fr: Robust phylogenetic analysis for the non-specialist. *Nucleic Acids Research*, 36 (Web Server issue). <https://doi.org/10.1093/nar/gkn180>
- Eddy, S. (2004). What is a hidden Markov model? *Nature Biotechnology* 22, 1315–1316. <https://doi.org/10.1038/nbt1004-1315>
- Felsenstein, J. (2005). PHYLIP (Phylogeny Inference Package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.
- Fischetti, V. A. (2008). Bacteriophage lysins as effective antibacterials. *Current Opinion in Microbiology*, 11(5), 393–400. <https://doi.org/10.1016/j.mib.2008.09.012>
- Garg, P. (2019). Filamentous bacteriophage: A prospective platform for targeting drugs in phage-mediated cancer therapy. *Journal of Cancer Research and Therapeutics*, 15 (8), 1–10. https://doi.org/10.4103/jcrt.JCRT_218_18
- Guindon, S., Dufayard, J.-F., Lefort, V., Anisimova, M., Hordijk, W., & Gascuel, O. (2010). New algorithms and methods to estimate maximum-likelihood phylogenies: Assessing the performance of PhyML 3.0. *Systematic Biology*, 59(3), 307–321. <https://doi.org/10.1093/sysbio/syq010>
- Guindon, S., & Gascuel, O. (2003). A Simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696–704. <https://doi.org/10.1080/10635150390235520>
- Hatfull, G. F. (2008). Bacteriophage genomics. *Current Opinion in Microbiology*, 11(5), 447–453. <https://doi.org/10.1016/j.mib.2008.09.004>
- Hatfull, G. F., & Hendrix, R. W. (2011). Bacteriophages and their genomes. *Current Opinion in Virology* 1(4), 298–303. <https://doi.org/10.1016/j.coviro.2011.06.009>
- Hatfull, G. F., Jacobs-Sera, D., Lawrence, J. G., Pope, W. H., Russell, D. A., Ko, C.-C., . . . Hendrix, R. W. (2010). Comparative genomic analysis of 60 mycobacteriophage genomes: Genome clustering, gene acquisition, and gene size. *Journal of Molecular Biology*, 397(1), 119–143. <https://doi.org/10.1016/j.jmb.2010.01.011>
- Kelley, L. A., Mezulis, S., Yates, C. M., Wass, M. N., & Sternberg, M. J. (2015). The Phyre2 web portal for protein modeling, prediction and analysis. *Nature Protocols*, 10(6), 845–858. <https://doi.org/10.1038/nprot.2015.053>
- Krumsiek, J., Arnold, R., & Rattei, T. (2007). Gepard: A rapid and sensitive tool for creating dotplots on genome scale. *Bioinformatics*, 23(8), 1026–1028. <https://doi.org/10.1093/bioinformatics/btm039>
- Pohane, A. A., Joshi, H., & Jain, V. (2014). Molecular dissection of phage endolysin. *Journal of Biological Chemistry*, 289(17), 12085–12095. <https://doi.org/10.1074/jbc.m113.529594>
- Russell, D. A., & Hatfull, G. F. (2016). PhagesDB: The actinobacteriophage database. *Bioinformatics*, 33(5), 784–786. <https://doi.org/10.1093/bioinformatics/btw711>
- Saitou, N., & Nei, M. (1987). The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution*, 4(4), 406–425. <https://doi.org/10.1093/oxfordjournals.molbev.a040454>
- Shen, X., Li, M., Zeng, Y., Hu, X., Tan, Y., Rao, X., Jin, X., Li, S., Zhu, J., Zhang, K., & Hu, F. (2012). Functional identification of the DNA packaging terminase from *Pseudomonas aeruginosa* phage PaP3. *Archives of Virology*, 157 (11), 2133–2141. <https://doi.org/10.1007/s00705-012-1409-5>
- Smith, K. C., Castro-Nallar, E., Fisher, J. N. (2013). Phage cluster relationships identified through single gene analysis. *BMC Genomics*, 14(14). <https://doi.org/10.1186/1471-2164-14-410>
- Zhang, Y. (n.d.). Zhang Lab. <https://zhanglab.dcm.med.umich.edu/>