Bioinformatics in Bacteriophages: Leveraging the Levenshtein Distance for Enhanced Bacteriophage Genome Analysis

By

Nathini Sion

Thesis Submitted to Flinders University for The Degree of

Master of Biotechnology

College of Medicine and Public Health

October 2024

Supervised by

Professor Robert Edwards

Susanna Grigson

College of Science and Engineering

TABLE OF CONTENTS

ABSTRACT

Viruses that infect and eliminate bacteria are known as bacteriophages, or 'phage'. Through various ecosystems, such as the human microbiome, the animal gut, marine, and soil, they are remarkably abundant and diverse. Phages have mosaic genome are composed of modules with unique evolutionary histories. The phages mosaicism contributes to the incredibly diversity observed among phages and poses challenges for their classification. In recent years, bacteriophage taxonomy has developed from the morphology-based to the genome-based classification principle. This reflects the genome classification provides more comprehensive and accurate basis for understanding phage relationships with evolution.

This study utilises the Levenshtein distance as an important tool for assessing the similarity among phage genomes. The distance is computed by evaluating the least number of string modifications, which encompass insertions, deletions, and substitutions, required to transform one string another. This has the potential to effectively recognise between phage genomes based on their evolutionary relationships with functional diversity.

The International Committee on Taxonomy of Viruses (ICTV) has recently implemented a genome-based taxonomy to enhance the classification of viruses, including phages. This transition highlights the importances of bioinformatics. To increase our understanding of viruses and to enable researchers to analyse genomes using computational algorithms, to facilitate the identification of phage functions and the comparison of phages genomes to understand their relationship together with potential applications.

The number of phage genomes in the National Center for Biotechnology Information (NCBI) database has increased significantly as a result of advancements in molecular techniques since the late 20th century. Phage taxonomy has been substantially enhanced by bioinformatics algorithms. Nevertheless, the current methodologies continue to have their limitations. In particular, the approaches for comparing phage genomes may not fully convey the complexity of genome arrangement and synteny. This serves to highlight the importance of conducting further research on algorithms that would accurately and comprehensively represent the entire spectrum of phage diversity.

The objective of this thesis is to evaluate the efficacy of genome similarity analysis by utilising the Levenshtein distance and generating phylogenetic trees. This method functions as both an alternative method for phage classification and an investigation of the extent to which the gene

arrangement within genomes is consistent with the current taxonomic classification. Moreover, its efficacy is evaluated in comparison to the current classification principle. The analysis has the potential to offer valuable insights into phage classification, which could be instrumental in the comprehension of phage biology, the prediction of phage-host interactions, and the development of precise classification systems for the effective use of phages in therapy and other applications.

The phage genomes datasets were compiled from the NCBI Genbank database. Genome similarity was then computed using the Levenshtein and Mash distances. The phylogenies were constructed from the Levenshtein distance and visualised using the Interactive Tree Of Life (iTOL) online tool. Numerous phage characteristics, including genome length, bacterial host, and viral taxonomy, were employed to analyse these trees. To evaluate the correlation between the two-distance metrics, tanglegrams were generated. The potential of this method to investigate the relationship between phage gene arrangement and phage taxonomy is illustrated by the results of this study. Research in the future should expand the phage dataset and investigate in several algorithm methods.

Keywords: Bioinformatics, Bacteriophages, Levenshtein Distance, Phage Taxonomy, Phage Classification

DECLARATION

I certify that this thesis:

1. does not incorporate without acknowledgment any material previously submitted for a degree of diploma in any university; and
2. the research within will not be submitted for any other future degree or diploma without the permission of Flinders University; and
3. to the best of my knowledge and belief, does not contain any material previously published or written by another person except where due reference is made in the text.

Signed:

Nathini Sion

Date October 2024

ACKNOWLEDGEMENTS

I am deeply grateful to Professor Robert Edwards and Susie Grigson, my supervisors, for their invaluable guidance and insightful feedback during the course of this thesis. Their proficiency in bacteriophages and bioinformatics has been genuinely indispensable. This thesis would not have been feasible without their contributions.

Additionally, I am also thankful of the FAME Lab members, particularly Bhavya Nalagampalli Papudeshi, Vijini Mallawaarachchi, and everyone for their offering feedback in this thesis presentation. Moreover, the Flinders University Go Beyond Global scholarship has allowed me to concentrate on my academic pursuits, and I am pleased for the financial support it has provided.

I would like to acknowledge the use of Gemini Advanced [gemini.google.com] to assist with this undertaking. It was helpful to me to review the thesis outline, explain key concepts, and perform a thorough proofreading of the grammar for this project. I carefully evaluated and analysed the suggestions provided by Gemini and revised the writing to use my own words and expressions.

I would be remiss if I did not mentor the unconditional support and confidence that my family and friends have shown in me. This includes my spouse, my sister, my parents, and even my cats.

CHAPTER 1 – INTRODUCTION

Bacteriophages, also known as phages, are viruses uniquely capable of infecting and replicating within bacteria hosts. With approximately 10^{31} phages on our planet, phages are the most abundant biological entities in various biome, from the human gut to soil and the marine, to fossil stool specimens. As a result, they have a significant impact on the diversity of microbial ecosystems and have facilitated the development of complex evolution (Dion et al., 2020; Rohwer & Edwards, 2002). As illustrated in Figure 1, phage genomes frequently exhibit mosaic architecture, which is characterised by variable regions, unknown functions genes, and open reading frames, as a result of their uniquity and diverse in environments (Belcaid et al., 2010; Kang et al., 2017). This has challenged phage taxonomy, which organised phage by structure, bacterial hosts, and genomic types (Dion et al., 2020).

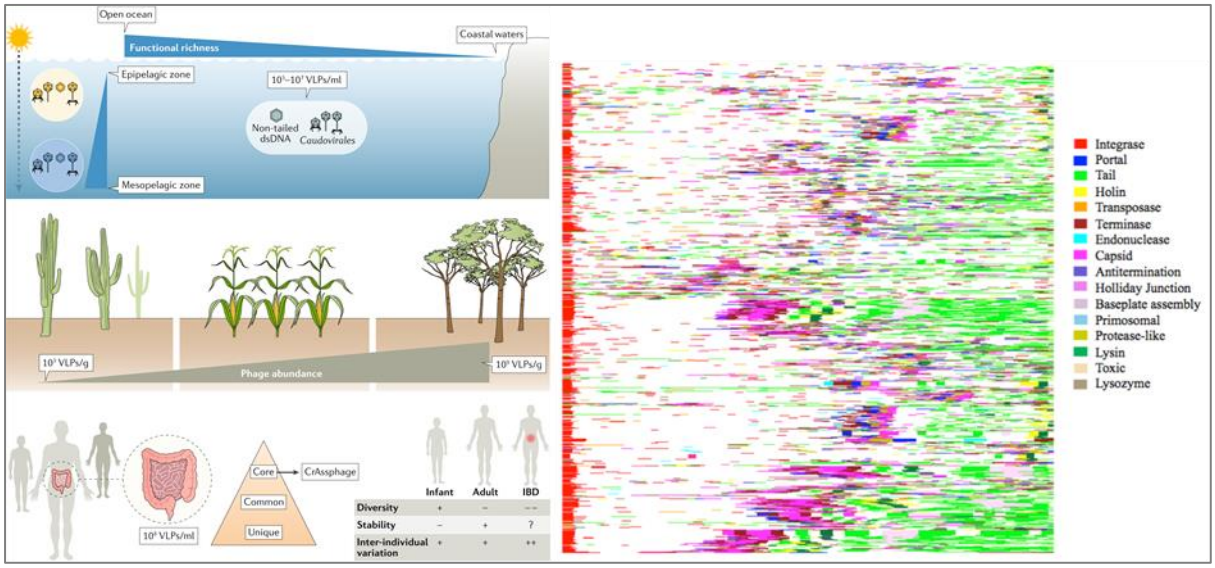


Figure 1. *(Right) The diagram shows the phages abundant in various environments including the marine, soil, and human microbiome, adapted from Dion et al. (2020). (Left) Protein functions diversity in phage genome, which are annotated by colour and a line represents a phage genomic pattern, derived from Kang et al. (2017)*

This comprehension of the gene arrangements in phage genomes can offer valuable insights into their evolutionary history. In Figure 2A, the gene arrangements and orders show how species share a common ancestor, how genes move from one species to another, or how they have changed to survive in different features (Moura de Sousa et al., 2021). Analysis gene arrangements represent the phages evolution and their capabilities. This is important for the

application of phages in medicine and biotechnology, particularly in gene transfer, phage treatment, and the drug development for against antibiotic-resistant bacteria (Susanna R Grigson et al., 2023; Mallawaarachchi et al., 2023; Strathdee et al., 2023).

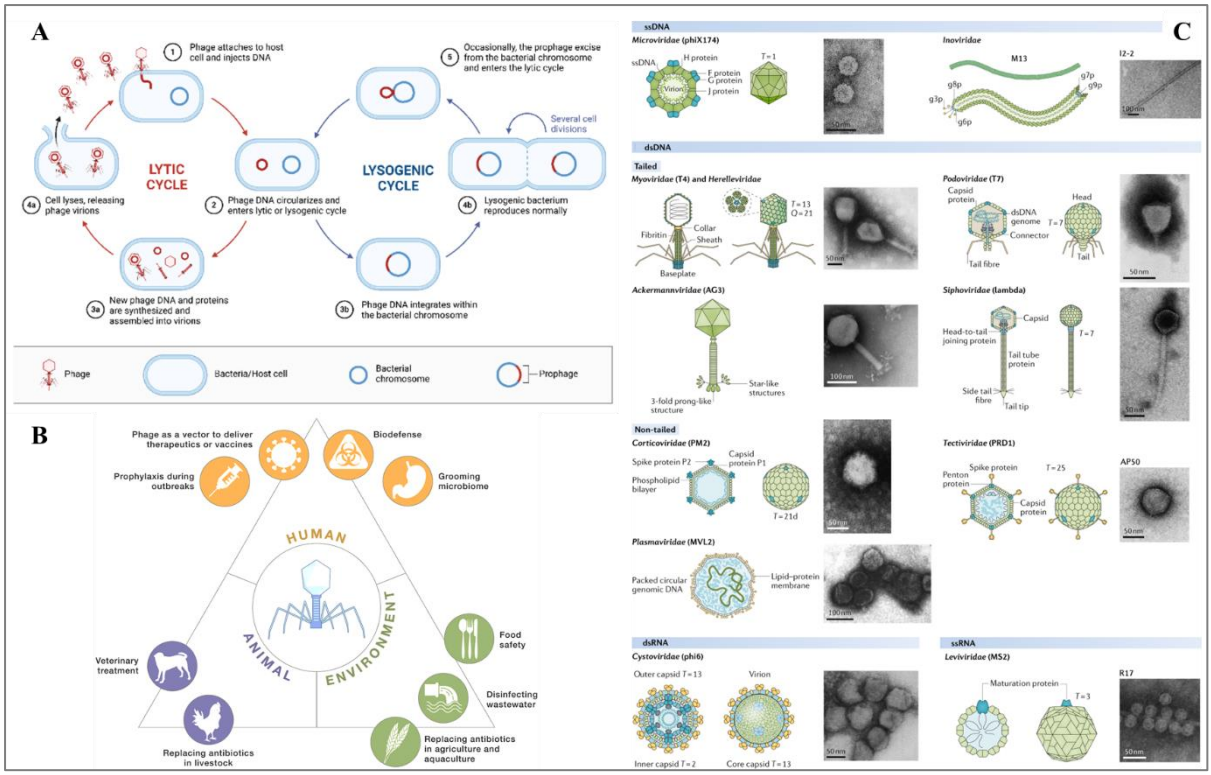


Figure 2. ***(A)*** *Diagram of the bacteriophage mechanism in the lytic and lysogenic cycle, created with BioRender (2021).* ***(B)*** *Illustration of phage therapy applications in environmental management, human and animal health (Strathdee et al., 2023).* ***(C)*** *Schema of phage classification based on morphology and genome type, including single- and double-stranded DNA and RNA viruses, adapted from Dion et al. (2020).*

1.1 Aims

In this project, the Levenshtein distance method is used to compare phages within each gene function group and phylogenetic trees to demonstrate phage genome relationships. Employing bioinformatics techniques to demonstrate that the Levenshtein distance could provide novel insights into phage diversity and evolution (Dorlass & Amgarten, 2024).

There are main goals for the project. The primary objective is to assess the efficacy of genome similarity analysis through the algorithm implication and the phylogenetic trees. Develop a preliminary prototype to apply this method to phage genomes and evaluate genome similarity to accurately reflect evolutionary relationships.

Secondly, the methodology was developed and used to an expanded phage genome dataset, which is larger as well as more varied than the first phage dataset, to validate the algorithm and assess its generalisability.

Lastly, this study aims to determine the distance outcome of the Levenshtein distance in comparison to established classification techniques such as Mash distance.

1.2 Bacteriophages

Bacteriophages are commonly called phages. These are viruses with diverse structures and highly mosaic genomes. They can infect, destroy, and replicate within bacteria cells (Alsayed & Permana, 2024; Kang et al., 2017). Phages exhibit high host specificity, targeting bacterial species or stains. As seen in Figure 2A that show primary life cycles, including lytic and lysogenic cycles, which provide lytic and temperate phages. Lytic phages, also known as virulent phages, are crucial in medical applications to control bacteria populations. They can synthesise their genetic materials and proteins before being released from the host cells to infect other bacteria (Alsayed & Permana, 2024; E. White & V. Orlova, 2020; Susanna R. Grigson et al., 2023)

Phages are crucial in shaping microbial ecosystems and driving their evolution. These bacterial viruses infect and lyse their bacteria hosts, helping to maintain microbial balance and prevent the pathogens from overgrowing. Moreover, phages contribute their own evolution and bacterial evolution through horizontal gene transfer. They encourage the spread of beneficial traits like against antimicrobial resistance (Casey et al., 2021).

Phages have unique qualities that are employed in medicine, agriculture, and biotechnology. In medicine and pharmacology, phages are explored as alternative antibiotics against antibiotic-resistant pathogens. Phage therapy offers a precision approach to infection control through specific targeting bacteria and avoids destroying beneficial microbiota (Alsayed & Permana, 2024). In agriculture, phage cocktails are utilised as biocontrol agents to combat plant diseases caused by bacteria. Also, phages serve as valuable tools in biotechnology for genetic engineering and molecular studies, including, CRISPR-Cas, diagnostic probes, genes and protein transfer (Jo et al., 2023; Strathdee et al., 2023), as outlined in Figure 2B.

The study of comparative analysis in phage genomes would have the potential to gain insight into phage taxonomy classification and allow researchers to modify phages, which would have significant implications for phage therapy, evolution, and bacterial resistance drug novelty.

1.3 Phage Taxonomy: The Classification of Bacteriophages

Phage Taxonomy provides a framework for organising and understanding the vast diversity of bacteriophages. Traditionally, the classification has been based on various characteristics, for example, morphology, nucleic acid type, replication, host, and diseases (E. White & V. Orlova, 2020). The International Committee on Taxonomy of Viruses (ICTV), the authoritative body responsible for classifying and naming viruses, developed a taxonomy order system based on culturing phages and visualising them under electron microscopy (Rohwer & Edwards, 2002; Simmonds et al., 2023). In 2022, the Bacterial Viruses Subcommittee (BVS) of the ICTV implemented significant changes, adopting a classification system based on genome analysis. This reflects the growing importance of genomics in understanding viral insight and has the goal of developing a universal virus taxonomy (Turner et al., 2023), as presented in Figure 3.

Comparative analysis focuses on identifying similarities and differences between phages at the genetic level. That allows us to characterise phages and interpret meaningful information (Rossi et al., 2024). Due to the phage diversity with advances in genomic technology, numerous new strains are found and added to the GenBank database as the vast number illustrated in Figure 4 (Adriaenssens & Brister, 2017). To gain insight at the genomic level, comparative analysis through bioinformatics will provide more understanding of complex phages genomes (Dorlass & Amgarten, 2024).

Figure removed due to copyright restriction.

Figure 3. *Phage classification between the years 1991 to 2019 and 2019 represents the increased the taxonomy structure since the ICTV employed genomic-based classification (Gorbalenya et al., 2020; International Committee on Taxonomy of Viruses Executive et al., 2020).*

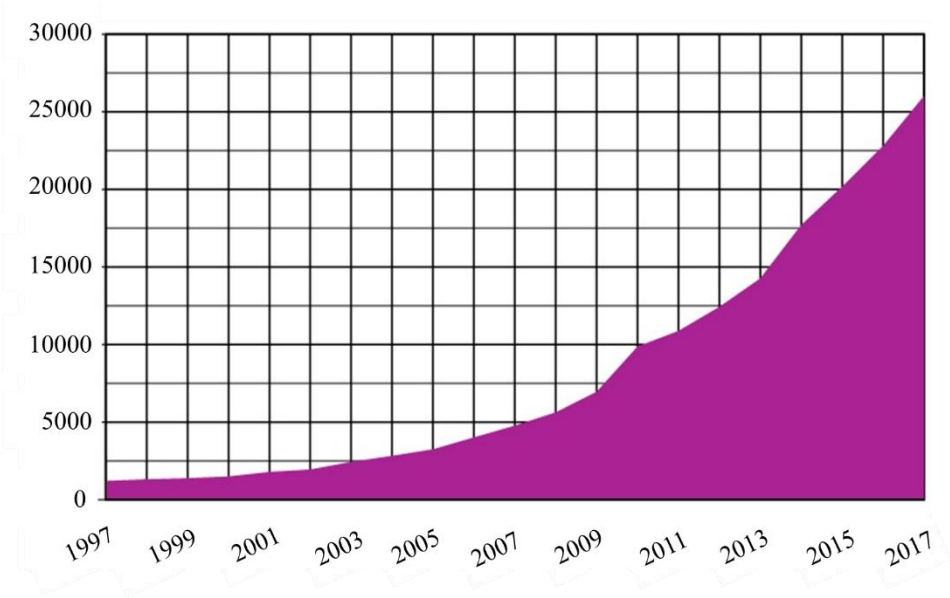


Figure 4. The increasing of phage databases since 1997, which present the phage nucleotide sequences published each year in the International Nucleotide Sequence Database Consortium (INSDC) that searching in the GenBank NCBI database, adapted from Adriaenssens and Brister (2017).

1.4 Bioinformatics: Computational Methods for Biological

Hogeweg (2011) provided a definition of bioinformatics as "the study of informatic processes in biotic systems" in the early 1970s (Hogeweg & Hesper, 1978), then after the 1980s, this term represents the computational approaches for genome comparative analysis and interpreting biological data. Currently, bioinformatics is related to various science disciplines and associated with many support systems in the computational area, as shown in Figure 5 (Pathak et al., 2022). Especially, the genomic database has increased in number as well as the phage genome data (Adriaenssens & Brister, 2017; Mount, 2001).

Bioinformatics also has emerged as a critical tool in the study of bacteriophages (Cresawn et al., 2011). This research utilises computational algorithms, including the Levenshtein distances, for genome function and similarity analysis. Moreover, we are expected to gain insight into their diversity and function through phylogenetic analysis (Mount, 2001). Quantifying the genomic relationship between phages could uncover the evolutionary relationships and offer an alternative to refine phage taxonomy based on genomic data (Edwards et al., 2016). Furthermore, that could enable the identification of specific gene arrangements associated with phage functions at the genomic level.

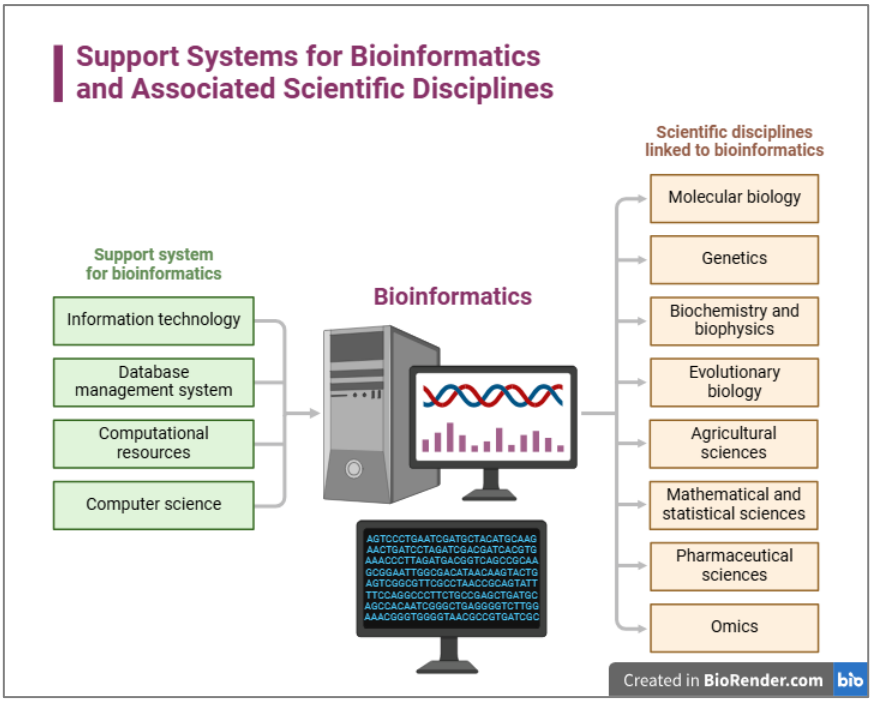


Figure 5. *Schema of support system for bioinformatics associated with scientific disciplines, created in BioRender.com.*

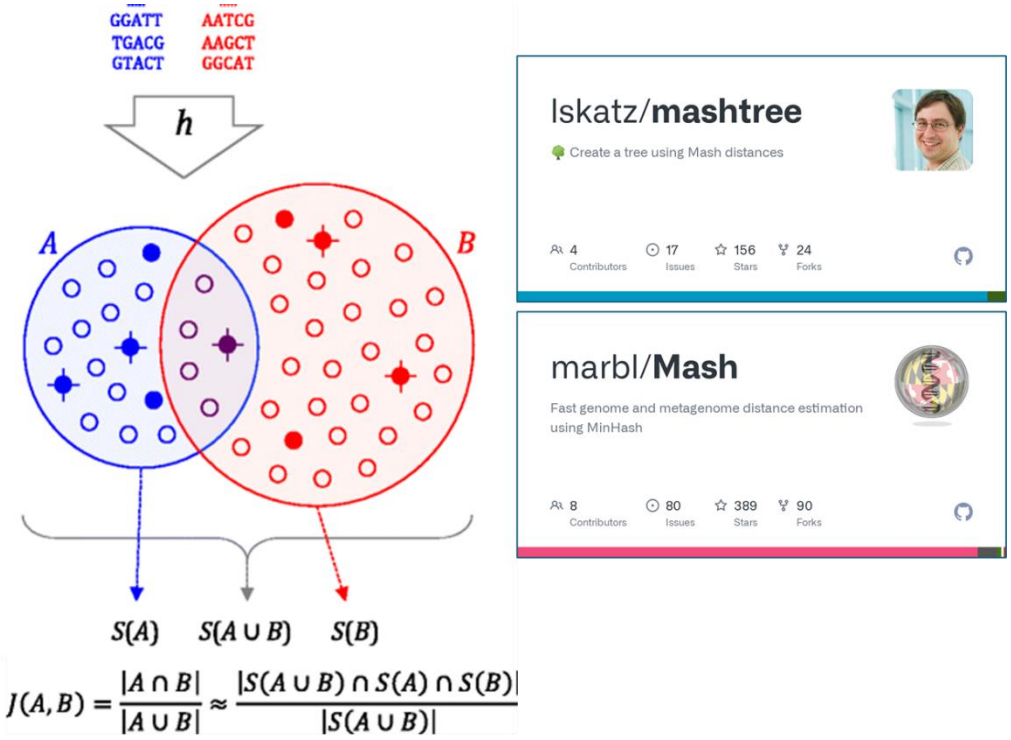


Figure 6. *Diagram of the Mash algorithm using the MinHash technique that reduces two large sequences to small sequence sets using k-mers and estimates the distance by the Jaccard equation, J(A,B). Mashtree then uses the neighbour-joining algorithm to generate a dendrogram, adapted from Ondov et al. (2016) and Katz et al. (2019).*

1.5 Mash Distance: A Sequences Comparison Tool

Mash distance is a bioinformatics tool that uses the MinHash module to calculate similarity based on the k-mers and the Jaccard index (Ondov et al., 2016). That had been employed in the Mashtree algorithm package for rapidly and efficiently generating large phylogenetic trees (Katz et al., 2019). Figure 6 illustrates the Mash distance concept for analysing two genome sequences and highlights the GitHub packages used in this research.

1.6 Levenshtein Distance: From Linguistics to Bioinformatics

The Levenstein distance, developed in 1995, is an algorithm for measuring the similarity between two words (Levenshtein, 1966). This calculates the smallest number of edits consisting of insertions, deletions, and substitutions comparing the transformation of one string into another in the field of computer science (Bachmann et al., 2021). Initially, The Levenstein distance was implemented in linguistics. Now, this method has proven valuable in a broad range of applications, for instance, generating destination matrices in transportation (Behara et al., 2020), developing dictionary lookup methods (Haldar & Mukhopadhyay, 2011), searching biological database (Berger et al., 2021), and applying machine learning for health records (Hossain et al., 2023). The algorithm has also found applications in bioinformatics for clustering sequences data (Logan et al., 2022) and analysing multiplex DNA sequences (Buschmann & Bystrykh, 2013).

This research employs The Levenshtein distance to compute phage genome similarity computation from gene functions, as outlined in Figure 7, that the grouped gene function strings were calculated similarity by following mathematic equations in two-dimension metric. The python-Levenshtein package serve as the main tool for the vast data analysis (Bachmann et al., 2021).

This approach offers novel alternative for phage classification based on genomics data. Which could significantly increase understanding of phage diversity, phage-host interactions, genomic region properties, and evolution. These would contribute to accelerating classification, provide valuable insight and lead to development in therapeutic applications.

Figure(s) removed due to copyright restriction.

Figure 7. *Demonstration of use of the Levenshtein algorithm to phage genome similarity computations concept. The two annotated function codes of the phage genes shown at the top are converted into letter sets, which are then analysed using the Levenshtein distance GitHub package (Bachmann et al., 2021) to determine the similarity between the genomes.*

CHAPTER 2 – MATERIALS AND METHODS

This study applies the Levenshtein distance and phylogenetic tree to determine the potential uses for phage genome classification. The methodology is outlined in Figure 8 below.

2.1 Phage Genomes Dataset.

The phage genomes dataset was compiled from the NCBI GenBank database with the INfrastructure for a PHAge REference Database (INPHARED) published by the Millard lab (https://millardlab.org/phage-genomes-may-2024/). Gene functions annotation was performed using pharokka v1.4.0 (Bouras et al., 2023). This compilation was facilitated by the Flinders HPC DeepThought (Flinders University, 2021). Subsequently, the dataset was processed into nested dictionaries in pickle format. This methodology was developed with the support of Susie Grigson, a PhD candidate in the FAME lab, College of Science and Engineering, Flinders University.

2.2 Calculation of Levenshtein Distance.

The Levenshtein distance was calculated using Python implementation sourced from GitHub (Bachmann et al., 2021). As presented in Figure 7, this distance metric quantifies the minimum number of edit operations, consisting of insertions, deletions, and substitutions (Logan et al., 2022), required to transform one genome into another. This package is employed to compute the distance between phage genome functions in the dataset using optimised algorithms in Python v3.10.12 within Google Colab notebooks. The results provided a measure of similarity between genome functions in the dataset, which was then converted to Newick format using average linkage hierarchical clustering before generating the phylogenetic trees. The method code is shown in Appendix A.

2.3 Calculation of Mash Distance.

Multi-Phage genomes as a FASTA file was separated into individual genome sequences. MashTree was downloaded and installed from GitHub (Katz et al., 2019; Ondov et al., 2016) and implemented in Bash v4.4.20 on Jupyter notebook within the Flinders DeepThought HPC (Flinders University, 2021) following the command in Appendix B. The Mash distance results were also encoded into Newick format.

2.4 Phylogenetic Tree Generation.

The Interactive Tree Of Life (https://itol.embl.de), an online tool, was used to produce a tree from the Newick format distance matrices (Letunic & Bork, 2024). The phage genome tree was analysed based on genome length, bacterial host, viral family, viral sub-family, viral genus, lower viral taxa (Cook et al., 2021), and by using different tree display types e.g. circular rooted, circular unrooted, and dendrogram. The iTOL setting provided in Appendix C.

2.5 Tanglegrams Generation.

Tanglegrams were also generated from the Newick format, which were the results of the distance computation methods. This was performed in the Colab environment using the R programming language v4.4.2 with the dendextend package (Galili, 2015; Grigson et al., 2022), according to appendix D. This approach allowed for the manipulation of complex data inputs and facilitated the comparison of trees through visualisation (Galili, 2015).

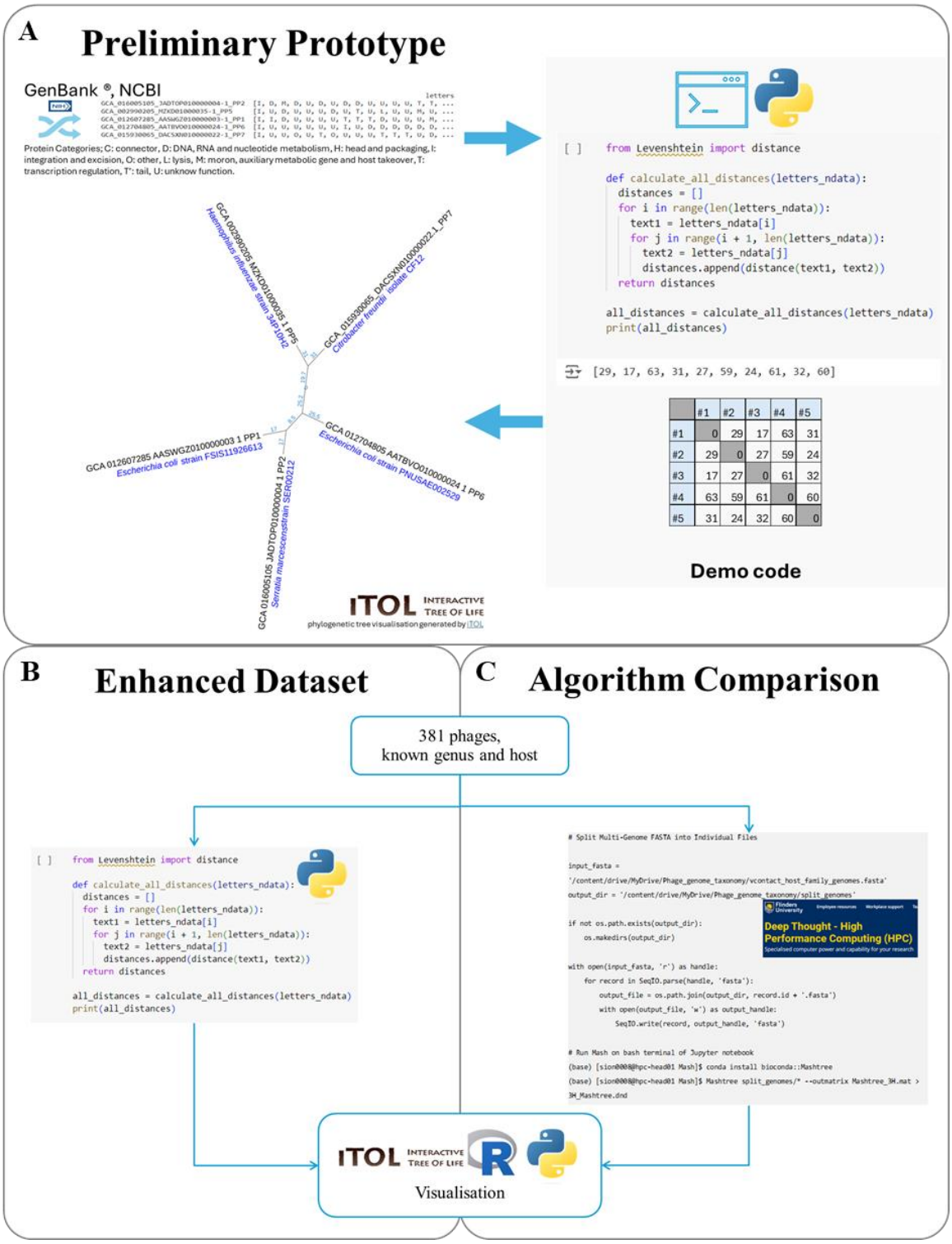


Figure 8. *Overview of the approach utilised to achieve the research objectives.* ***(A)*** *The preliminary process and the phylogenetic tree of demo dataset. This stage includes dataset import and preparation, demonstration code with the Levenshtein distance metric, and the phylogenetic tree generated from the result matrix.* ***(B)*** *Analysis of the enhanced dataset for genome similarity using the Levenshtein distance.* ***(C)*** *Comparison of the previous results with those obtained using Mash distance, employing visualisation approaches.*

CHAPTER 3 – RESULTS

3.1 Levenshtein Distance for Phage Genome Similarity Analysis

3.1.1 Preliminary Prototype

The initial goal of this research was to demonstrate proof of the concept. This was achieved by analysing five randomly selected phage genomes, regardless of phage families and their host bacteria. Python code, executed in the Colab, was implemented to calculate the Levenshtein distance between the phages and reformat the output to Newick format for phylogenetic tree construction. The code was designed to be modular and scalable, facilitating future expansion to larger datasets. The phylogenetic tree in Figure 7A, generated using iTOL, reflects genome similarity and has the potential to provide insights into phage taxonomy classification.

3.1.2 Enhanced Dataset

The Levenshtein distance code from the preliminary method was applied to an enhanced dataset consisting of 381 phage genomes. The resulting phylogenetic trees were analysed and visualised using iTOL, exploring various features such as genome length, bacterial host, and viral taxonomy. Different modes were employed, including rectangular, circular rooted and unrooted. Figure 9A and 10 demonstrated that the Levenshtein distance effectively captured relationships associated with genome length. In the circular tree, the position progresses from the highest to lowest genome length, then transitions to respect the viral family. These results indicate that the similarity measured by the Levenshtein distance reflects both genome length and the viral family, which are related to gene function.

3.2 Comparison of The Levenshtein and The Mash Distances

The Levenshtein distances calculated from the increased dataset were compared with those obtained using the Mash distance on the same dataset. Figure 9 presents a comparison of the circular phylogenetic trees created using both distance algorithms, illustrating their relationship across different annotated dataset groups with the same varies of phage classification. Additionally, the comparison is visualised in the Tanglegram presented in Figure 11. The tanglegram reveals overview relationships between the phylogenetic trees rendered using the Levenshtein distance (Left) and the Mash distance (Right), with an entanglement value of 0.576.

Figure removed due to copyright restriction.

Figure 9. *Comparison of circular phylogenies of the 381 phages dataset illustrating the relationship between the phages, based on bacterial host, viral family, sub-family, genera, lowest taxonomic level, and genome length.* ***(A)*** *Phylogenetic trees obtained from the Levenshtein distance and* ***(B)*** *The Mash distance, generated by iTOL (Letunic & Bork, 2024).*

Figure removed due to copyright restriction.

Figure 10. *Phylogenetic trees of the 381-phage dataset using the Levenshtein distance, displayed in circular (Left) and dendrogram (Right). The colors are annotated to the viral family of the phages, and the light blue bar presented to the phage genome length. The trees generated by iTOL (Letunic & Bork, 2024).*

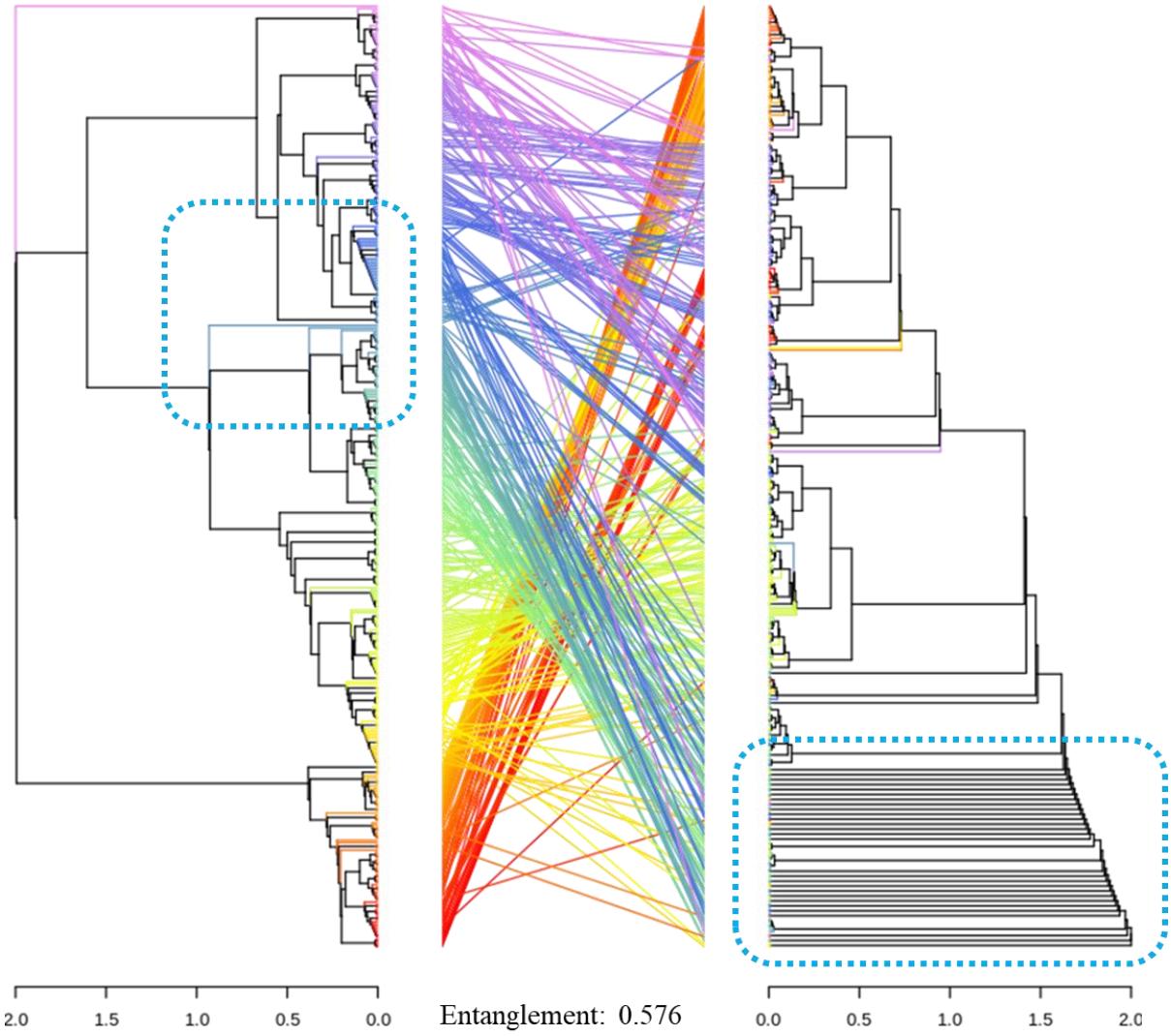


Figure 11. *Tanglegram comparing the Levenshtein distance (Left) and the Mash distance (Right) phylogenies. The coloured clusters enable comparison of distance results for the same 381 phages dataset, with the entanglement value of 0.576. The blue dot clusters represent polytomy areas in the Mash distance tress, and the significance of the difference patterns. (Generated using R studio in the Colab)*

CHAPTER 4 – DISCUSSION

4.1 Levenshtein Distance for Phage Genome Similarity Analysis

The preliminary method provided insight into phage similarity using the Levenshtein algorithm and displayed its potential for phage classification field. Therefore, we focused on analysing the expanded phage dataset comprising 381 entities. Figure 9A illustrates the capacity of the Levenshtein distance to capture some taxonomic relationships. When examining viral family relationships between phages, Figure 10 shows that the phylogenetic trees clearly support the validity of the existing viral families *Straboviridae* and *Autographiviridae*, represented as yellow and light blue groups, respectively. Both viral families were recently abolished in the ICTV taxonomy update (Turner et al., 2023).

While the clustering of other phages is distinctly separated, most of them are grouped according to genome length, from the longest to shortest. This observation supports the notion that the Levenshtein distance is sensitive to data length, consistent with previous findings (Berger et al., 2021; Buschmann & Bystrykh, 2013; Logan et al., 2022). This sensitivity could be attributed to longer genomes naturally having a higher potential for accumulating insertions, deletions, and substitutions, which directly influence the Levenshtein distance calculation (Bachmann et al., 2021).

4.2 Comparison of The Levenshtein and The Mash Distances

The Tanglegram in Figure 11 provides a visual comparison of the phylogenies generated using the Levenshtein and the Mash distances. The entanglement value of 0.576 indicates a moderate level of incongruence between the two dendrograms, that represents there have some harmony clusters in the overall, but also observable differences across the dendrograms (Galili, 2015).

Notably, the clustering highlighted by the blue dotted squares in Figure 11 signifies polytomies in the Mash distance dendrogram. Polytomy indicates unsolved relationships, where the Mash distance analysis had not sufficient information to determine the branching order within those groups (Krone et al., 2021). This suggests that the Levenshtein distance might offer a higher resolution in certain cases, potentially due to its sensitivity to the length and positions of edits within the genomes.

These discrepancies emphasise the limitations of relying solely on overall phage similarity, as captured by the Mash distance. The Levenshtein distance provided a more nuanced view of

phage relationships by considering the specific functions and positions of editing between genomes, which could be crucial for understanding the evolutionary and phages diversity.

4.3 Limitations and Future Directions

This study had limitations regarding the time available to expand the dataset and to conduct a comprehensive comparison with other related algorithms.

Future research should address those limitations to gain more complete understanding of the strengths and weaknesses of the Levenshtein distance in phage genome analysis. In addition, studying more diverse phage databases would enhance the resolution and generalisation of the algorithm. Furthermore, comparison with other genome analysis approaches could provide their relative performance and identify the most suitable methods for the phage universal classification.

4.4 Conclusion

This thesis examined the effectiveness of the Levenshtein distance for enhanced bacteriophage genome analysis. Applying this approach to the phage genome dataset demonstrated the potential of the Levenshtein distance to compute genome similarity. These findings highlight the adeptness of this method for exploring the relationship between gene functions arrangement and taxonomy. While this study focused on a specific set of phages, future research could expand this analysis to a wider range of phage databases and explore the integration of other genomic features into the phage classification framework. Finally, this research exhibits the value of the Levenshtein distance as a potent tool for enhancing phage genome analysis and advancing the understanding of these significant biological entities.

BIBLIOGRAPHY

Ackermann, H. W. (2011). Bacteriophage taxonomy. *Microbiology Australia*, *32*(2), 90-94. https://doi.org/https://doi.org/10.1071/MA11090

Adriaenssens, E., & Brister, J. R. (2017). How to Name and Classify Your Phage: An Informal Guide. *Viruses*, *9*(4). https://doi.org/10.3390/v9040070

Aljabali, A. A. A., Aljbaly, M. B. M., Obeid, M. A., Shahcheraghi, S. H., & Tambuwala, M. M. (2024). The next generation of drug delivery: Harnessing the power of bacteriophages. In *Methods in Molecular Biology* (pp. 279-315). Springer US. https://doi.org/10.1007/978-1-0716-3549-0_18

Alsayed, A. R., & Permana, A. D. (2024). Bacteriophages therapy: Exploring their promising role in microbiome modulation and combatting antibiotic resistance. *OBM Genet.*, *08*(02), 1-8. https://doi.org/10.21926/obm.genet.2402237

Arahal, D. R. (2014). Chapter 6 - Whole-Genome Analyses: Average Nucleotide Identity. In M. Goodfellow, I. Sutcliffe, & J. Chun (Eds.), *Methods in Microbiology* (Vol. 41, pp. 103-122). Academic Press. https://doi.org/10.1016/bs.mim.2014.07.002

Bachmann, M., Górny, M., Larsson, A., Rosin, G., Tavant, A., atomflunder, & DatGuy1. (2021). *Levenshtein*. In (Version 0.25.1) rapidfuzz. https://github.com/rapidfuzz/Levenshtein

Behara, K. N. S., Bhaskar, A., & Chung, E. (2020). A novel approach for the structural comparison of origin-destination matrices: Levenshtein distance. *Transportation Research Part C: Emerging Technologies*, *111*, 513-530. https://doi.org/https://doi.org/10.1016/j.trc.2020.01.005

Belcaid, M., Bergeron, A., & Poisson, G. (2010). Mosaic graphs and comparative genomics in phage communities. *J Comput Biol*, *17*(9), 1315-1326. https://doi.org/10.1089/cmb.2010.0108

Berger, B., Waterman, M. S., & Yu, Y. W. (2021). Levenshtein Distance, Sequence Comparison and Biological Database Search. *IEEE Trans. Inf. Theory*, *67*(6), 3287-3294. https://doi.org/10.1109/tit.2020.2996543

BioRender. (2021). *Lytic and Lysogenic Cycle*. https://app.biorender.com/biorender-templates/figures/all/t-603d11bcefa06000ad8b8f02-lytic-and-lysogenic-cycle

Bouras, G., Nepal, R., Houtak, G., Psaltis, A. J., Wormald, P.-J., & Vreugde, S. (2023). Pharokka: a fast scalable bacteriophage annotation tool. *Bioinformatics*, *39*(1). https://doi.org/10.1093/bioinformatics/btac776

Buschmann, T., & Bystrykh, L. V. (2013). Levenshtein error-correcting barcodes for multiplexed DNA sequencing. *BMC Bioinformatics*, *14*(1), 272. https://doi.org/10.1186/1471-2105-14-272

Casey, A., Coffey, A., & McAuliffe, O. (2021). Genetics and Genomics of Bacteriophages. In D. R. Harper, S. T. Abedon, B. H. Burrowes, & M. L. McConville (Eds.), *Bacteriophages: Biology, Technology, Therapy* (pp. 193-218). Springer International Publishing. https://doi.org/10.1007/978-3-319-41986-2_5

Chibani, C. M., Farr, A., Klama, S., Dietrich, S., & Liesegang, H. (2019). Classifying the Unclassified: A Phage Classification Method. *Viruses*, *11*(2), 195. https://www.mdpi.com/1999-4915/11/2/195

Clark, M. A., Douglas, M., & Choi, J. (2018). *Biology 2e* (2nd ed.). OpenStax. https://openstax.org/books/biology-2e/pages/21-1-viral-evolution-morphology-and-classification

Cook, R., Brown, N., Redgwell, T., Rihtman, B., Barnes, M., Clokie, M., Stekel, D. J., Hobman, J., Jones, M. A., & Millard, A. (2021). INfrastructure for a PHAge REference database:

Identification of large-scale biases in the current collection of cultured phage genomes. *Phage (New Rochelle)*, *2*(4), 214-223. https://doi.org/10.1089/phage.2021.0007

Cresawn, S. G., Bogel, M., Day, N., Jacobs-Sera, D., Hendrix, R. W., & Hatfull, G. F. (2011). Phamerator: a bioinformatic tool for comparative bacteriophage genomics. *BMC Bioinformatics*, *12*, 395. https://doi.org/10.1186/1471-2105-12-395

Dion, M. B., Oechslin, F., & Moineau, S. (2020). Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.*, *18*(3), 125-138. https://doi.org/10.1038/s41579-019-0311-5

Dorlass, E. G., & Amgarten, D. E. (2024). Bioinformatic Approaches for Comparative Analysis of Viruses. *Methods Mol. Biol.*, *2802*, 395-425. https://doi.org/10.1007/978-1-0716-3838-5_13

E. White, H., & V. Orlova, E. (2020). Bacteriophages: Their Structural Organisation and Function. In *Bacteriophages - Perspectives and Future*. IntechOpen. https://doi.org/10.5772/intechopen.85484

Edwards, R. A., McNair, K., Faust, K., Raes, J., & Dutilh, B. E. (2016). Computational approaches to predict bacteriophage-host relationships. *FEMS Microbiol. Rev.*, *40*(2), 258-272. https://doi.org/10.1093/femsre/fuv048

El-Kassas, W. S., Salama, C. R., Rafea, A. A., & Mohamed, H. K. (2021). Automatic text summarization: A comprehensive survey. *Expert Systems with Applications*, *165*, 113679. https://doi.org/https://doi.org/10.1016/j.eswa.2020.113679

Ezekannagha, C., Welzel, M., Heider, D., & Hattab, G. (2023). DNAsmart: Multiple attribute ranking tool for DNA data storage systems. *Computational and Structural Biotechnology Journal*, *21*, 1448-1460. https://doi.org/https://doi.org/10.1016/j.csbj.2023.02.016

Flinders University. (2021). *Deep Thought (HPC)*. In https://doi.org/10.25957/FLINDERS.HPC.DEEPTHOUGHT

Galili, T. (2015). dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718-3720. https://doi.org/10.1093/bioinformatics/btv428

Giri, N. (2021). Bacteriophage Structure, Classification, Assembly and Phage Therapy. *Biosciences Biotechnology Research Asia*, *18*, 239-250. https://doi.org/10.13005/bbra/2911

Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., Sanfaçon, H., . . . International Committee on Taxonomy of Viruses Executive, C. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nature Microbiology*, *5*(5), 668-674. https://doi.org/10.1038/s41564-020-0709-x

Grigson, S. R., Giles, S. K., Edwards, R. A., & Papudeshi, B. (2023). Knowing and naming: Phage annotation and nomenclature for phage therapy. *Clin. Infect. Dis.*, *77*(Suppl 5), S352-S359. https://doi.org/10.1093/cid/ciad539

Grigson, S. R., Giles, S. K., Edwards, R. A., & Papudeshi, B. (2023). Knowing and Naming: Phage Annotation and Nomenclature for Phage Therapy. *Clinical Infectious Diseases*, *77*(Supplement_5), S352-S359. https://doi.org/10.1093/cid/ciad539

Grigson, S. R., McKerral, J. C., Mitchell, J. G., & Edwards, R. A. (2022). Organizing the bacterial annotation space with amino acid sequence embeddings. *BMC Bioinformatics*, *23*(1), 385. https://doi.org/10.1186/s12859-022-04930-5

Guerin, E., Shkoporov, A., Stockdale, S. R., Clooney, A. G., Ryan, F. J., Sutton, T. D. S., Draper, L. A., Gonzalez-Tortuero, E., Ross, R. P., & Hill, C. (2018). Biology and Taxonomy of

crAss-like Bacteriophages, the Most Abundant Virus in the Human Gut. *Cell Host & Microbe*, *24*(5), 653-664.e656. https://doi.org/10.1016/j.chom.2018.10.002

Gupta, A., Mirarab, S., & Turakhia, Y. (2024). Accurate, scalable, and fully automated inference of species trees from raw genome assemblies using ROADIES. *bioRxiv*. https://doi.org/10.1101/2024.05.27.596098

Haldar, R., & Mukhopadhyay, D. (2011). Levenshtein Distance Technique in Dictionary Lookup Methods: An Improved Approach. *arXiv [cs.IT]*. http://arxiv.org/abs/1101.1232

Higgs, P. G., & Attwood, T. K. (2005). *Bioinformatics and molecular evolution* (1st ed.). Blackwell Pub.

Hogeweg, P. (2011). The Roots of Bioinformatics in Theoretical Biology. *PLOS Computational Biology*, *7*(3), e1002021. https://doi.org/10.1371/journal.pcbi.1002021

Hogeweg, P., & Hesper, B. (1978). Interactive instruction on population interactions. *Computers in Biology and Medicine*, *8*(4), 319-327. https://doi.org/https://doi.org/10.1016/0010-4825(78)90032-X

Hossain, E., Rana, R., Higgins, N., Soar, J., Barua, P. D., Pisani, A. R., & Turner, K. (2023). Natural Language Processing in Electronic Health Records in relation to healthcare decision-making: A systematic review. *Computers in Biology and Medicine*, *155*, 106649. https://doi.org/https://doi.org/10.1016/j.compbiomed.2023.106649

Huson, D. H., & Bryant, D. (2006). Application of phylogenetic networks in evolutionary studies. *Mol. Biol. Evol.*, *23*(2), 254-267. https://doi.org/10.1093/molbev/msj030

International Committee on Taxonomy of Viruses Executive, C., Gorbalenya, A. E., Krupovic, M., Mushegian, A., Kropinski, A. M., Siddell, S. G., Varsani, A., Adams, M. J., Davison, A. J., Dutilh, B. E., Harrach, B., Harrison, R. L., Junglen, S., King, A. M. Q., Knowles, N. J., Lefkowitz, E. J., Nibert, M. L., Rubino, L., Sabanadzovic, S., . . . Kuhn, J. H. (2020). The new scope of virus taxonomy: partitioning the virosphere into 15 hierarchical ranks. *Nat. Microbiol.*, *5*(5), 668-674. https://doi.org/10.1038/s41564-020-0709-x

Jiang, W. (2022). *Support Systems for Bioinformatics and Associated Scientific Disciplines*. BioRender. https://app.biorender.com/biorender-templates/figures/all/t-62978c2e31178c6c18323e60-support-systems-for-bioinformatics-and-associated-scientific

Jo, S. J., Kwon, J., Kim, S. G., & Lee, S.-J. (2023). The Biotechnological Application of Bacteriophages: What to Do and Where to Go in the Middle of the Post-Antibiotic Era. *Microorganisms*, *11*(9). https://doi.org/10.3390/microorganisms11092311

Kang, H. S., McNair, K., Cuevas, D. A., Bailey, B. A., Segall, A. M., & Edwards, R. A. (2017). Prophage genomics reveals patterns in phage genome organization and replication. *bioRxiv*. https://doi.org/10.1101/114819

Katz, L. S., Griswold, T., Morrison, S. S., Caravas, J. A., Zhang, S., den Bakker, H. C., Deng, X., & Carleton, H. A. (2019). Mashtree: a rapid comparison of whole genome sequence files. *J. Open Source Softw.*, *4*(44), 1762. https://doi.org/10.21105/joss.01762

Korf, I. H. E., Meier-Kolthoff, J. P., Adriaenssens, E. M., Kropinski, A. M., Nimtz, M., Rohde, M., van Raaij, M. J., & Wittmann, J. (2019). Still Something to Discover: Novel Insights into Escherichia coli Phage Diversity and Taxonomy. *Viruses*, *11*(5), 454. https://doi.org/https://doi.org/10.3390/v11050454

Krone, I., Thanukos, A., Collins, A., & Frankel, J. (2021). Phylogenetic pitchforks - Understanding Evolution. *University of California Museum of Paleontology*. https://evolution.berkeley.edu/phylogenetic-systematics/reading-trees-a-quick-review/phylogenetic-pitchforks/

Lefkowitz, E. J., Dempsey, D. M., Hendrickson, R. C., Orton, R. J., Siddell, S. G., & Smith, D. B. (2018). Virus taxonomy: the database of the International Committee on Taxonomy of Viruses (ICTV). *Nucleic Acids Res.*, *46*(D1), D708-D717. https://doi.org/10.1093/nar/gkx932

Letunic, I., & Bork, P. (2024). Interactive Tree of Life (iTOL) v6: recent updates to the phylogenetic tree display and annotation tool. *Nucleic Acids Res.*, *52*(W1), W78-W82. https://doi.org/10.1093/nar/gkae268

Levenshtein, V. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Proceedings of the Soviet physics doklady*.

Liu, H., Kheirvari, M., & Tumban, E. (2023). Potential Applications of Thermophilic Bacteriophages in One Health. *International Journal of Molecular Sciences*, *24*(9), 8222. https://www.mdpi.com/1422-0067/24/9/8222

Logan, R., Fleischmann, Z., Annis, S., Wehe, A. W., Tilly, J. L., Woods, D. C., & Khrapko, K. (2022). 3GOLD: optimized Levenshtein distance for clustering third-generation sequencing data. *BMC Bioinformatics*, *23*(1), 95. https://doi.org/10.1186/s12859-022-04637-7

Lomeli-Ortega, C. O., & Balcázar, J. L. (2024). Why tRNA acquisition could be relevant to bacteriophages? *Microb. Biotechnol.*, *17*(4), e14464. https://doi.org/10.1111/1751-7915.14464

Mallawaarachchi, V., Roach, M. J., Decewicz, P., Papudeshi, B., Giles, S. K., Grigson, S. R., Bouras, G., Hesse, R. D., Inglis, L. K., Hutton, A. L. K., Dinsdale, E. A., & Edwards, R. A. (2023). Phables: from fragmented assemblies to high-quality bacteriophage genomes. *Bioinformatics*, *39*(10). https://doi.org/10.1093/bioinformatics/btad586

McKerral, J. C., Papudeshi, B., Inglis, L. K., Roach, M. J., Decewicz, P., McNair, K., Luque, A., Dinsdale, E. A., & Edwards, R. A. (2023). The Promise and Pitfalls of Prophages. *bioRxiv*. https://doi.org/10.1101/2023.04.20.537752

Mount, D. W. (2001). *Bioinformatics : sequence and genome analysis*. Cold Spring Harbor Laboratory Press.

Moura de Sousa, J. A., Pfeifer, E., Touchon, M., & Rocha, E. P. C. (2021). Causes and Consequences of Bacteriophage Diversification via Genetic Exchanges across Lifestyles and Bacterial Taxa. *Molecular Biology and Evolution*, *38*(6), 2497-2512. https://doi.org/10.1093/molbev/msab044

Ona, S. (2023). *Bioinformatics (AI vs Traditional Techniques)*. BioRender. https://app.biorender.com/biorender-templates/figures/all/t-6527127ab71fc88dc7533f70-bioinformatics-ai-vs-traditional-techniques

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol.*, *17*(1). https://doi.org/10.1186/s13059-016-0997-x

Paradis, E., Claude, J., & Strimmer, K. (2004). APE: Analyses of Phylogenetics and Evolution in R Language. *Bioinformatics (Oxford, England)*, *20*, 289-290. https://doi.org/10.1093/bioinformatics/btg412

Parks, D. H., Chuvochina, M., Chaumeil, P.-A., Rinke, C., Mussig, A. J., & Hugenholtz, P. (2020). A complete domain-to-species taxonomy for Bacteria and Archaea. *Nat. Biotechnol.*, *38*(9), 1079-1086. https://doi.org/10.1038/s41587-020-0501-8

Pathak, R. K., Singh, D. B., & Singh, R. (2022). Chapter 1 - Introduction to basics of bioinformatics. In D. B. Singh & R. K. Pathak (Eds.), *Bioinformatics* (pp. 1-15). Academic Press. https://doi.org/https://doi.org/10.1016/B978-0-323-89775-4.00006-7

Rangel-Pineros, G., Millard, A., Michniewski, S., Scanlan, D., Siren, K., Reyes, A., Petersen, B., Clokie, M., & Sicheritz-Ponten, T. (2021). From Trees to Clouds: PhageClouds for Fast Comparison of ~640,000 Phage Genomic Sequences and Host-Centric

Visualization Using Genomic Network Graphs. *PHAGE*, *2*, 194-203. https://doi.org/10.1089/phage.2021.0008

Revell, L. J. (2012). phytools: an R package for phylogenetic comparative biology (and other things). *Methods in Ecology and Evolution*, *3*(2), 217-223. https://doi.org/10.1111/j.2041-210X.2011.00169.x

Röhling, S., Linne, A., Schellhorn, J., Hosseini, M., Dencker, T., & Morgenstern, B. (2020). The number of k-mer matches between two DNA sequences as a function of k and applications to estimate phylogenetic distances. *PLoS One*, *15*(2), e0228070. https://doi.org/10.1371/journal.pone.0228070

Rohwer, F., & Edwards, R. (2002). The Phage Proteomic Tree: a genome-based taxonomy for phage. *J. Bacteriol.*, *184*(16), 4529-4535. https://doi.org/10.1128/JB.184.16.4529-4535.2002

Rossi, F. P. N., Flores, V. S., Uceda-Campos, G., Amgarten, D. E., Setubal, J. C., & da Silva, A. M. (2024). Comparative Analyses of Bacteriophage Genomes. *Methods Mol. Biol.*, *2802*, 427-453. https://doi.org/10.1007/978-1-0716-3838-5_14

Schackart, K. E., 3rd, Graham, J. B., Ponsero, A. J., & Hurwitz, B. L. (2023). Evaluation of computational phage detection tools for metagenomic datasets. *Front. Microbiol.*, *14*, 1078760. https://doi.org/10.3389/fmicb.2023.1078760

Sharma, S., Chatterjee, S., Datta, S., Prasad, R., Dubey, D., Prasad, R. K., & Vairale, M. G. (2017). Bacteriophages and its applications: an overview. *Folia Microbiol. (Praha)*, *62*(1), 17-55. https://doi.org/10.1007/s12223-016-0471-x

Shaukat, M. A., Nguyen, T. T., Hsu, E. B., Yang, S., & Bhatti, A. (2023). Comparative study of encoded and alignment-based methods for virus taxonomy classification. *Sci. Rep.*, *13*(1). https://doi.org/10.1038/s41598-023-45461-0

Siddell, S. (2018). *Why Virus Taxonomy Is Important*. https://microbiologysociety.org/publication/past-issues/imaging/article/why-virus-taxonomy-is-important.html

Simmonds, P., Adriaenssens, E. M., Zerbini, F. M., Abrescia, N. G. A., Aiewsakun, P., Alfenas-Zerbini, P., Bao, Y., Barylski, J., Drosten, C., Duffy, S., Duprex, W. P., Dutilh, B. E., Elena, S. F., García, M. L., Junglen, S., Katzourakis, A., Koonin, E. V., Krupovic, M., Kuhn, J. H., . . . Vasilakis, N. (2023). Four principles to establish a universal virus taxonomy. *PLoS Biol.*, *21*(2), e3001922. https://doi.org/10.1371/journal.pbio.3001922

Strathdee, S. A., Hatfull, G. F., Mutalik, V. K., & Schooley, R. T. (2023). Phage therapy: From biological mechanisms to future directions. *Cell*, *186*(1), 17-31. https://doi.org/10.1016/j.cell.2022.11.017

Turner, D., Adriaenssens, E. M., Lehman, S. M., Moraru, C., & Kropinski, A. M. (2024). Bacteriophage taxonomy: A continually evolving discipline. In *Methods in Molecular Biology* (pp. 27-45). Springer US. https://doi.org/10.1007/978-1-0716-3523-0_3

Turner, D., Kropinski, A. M., & Adriaenssens, E. M. (2021). A Roadmap for Genome-Based Phage Taxonomy. *Viruses*, *13*(3), 506. https://www.mdpi.com/1999-4915/13/3/506

Turner, D., Shkoporov, A. N., Lood, C., Millard, A. D., Dutilh, B. E., Alfenas-Zerbini, P., van Zyl, L. J., Aziz, R. K., Oksanen, H. M., Poranen, M. M., Kropinski, A. M., Barylski, J., Brister, J. R., Chanisvili, N., Edwards, R. A., Enault, F., Gillis, A., Knezevic, P., Krupovic, M., . . . Adriaenssens, E. M. (2023). Abolishment of morphology-based taxa and change to binomial species names: 2022 taxonomy update of the ICTV bacterial viruses subcommittee. *Arch. Virol.*, *168*(2), 74. https://doi.org/10.1007/s00705-022-05694-2

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson,

E., . . . SciPy, C. (2020). SciPy 1.0: fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261-272. https://doi.org/10.1038/s41592-019-0686-2

Vita, F., Taiti, C., Pompeiano, A., Gu, Z., Lo Presti, E., Whitney, L., Monti, M., Di Miceli, G., Giambalvo, D., Ruisi, P., Guglielminetti, L., & Mancuso, S. (2016). Aromatic and proteomic analyses corroborate the distinction between Mediterranean landraces and modern varieties of durum wheat. *Sci. Rep.*, *6*, 34619. https://doi.org/10.1038/srep34619

Wang, Z., & Noda, M. (2018). Alarm Data Analysis for Safe Plant Operations: Case Study of Ethylene Plant. In M. R. Eden, M. G. Ierapetritou, & G. P. Towler (Eds.), *Computer Aided Chemical Engineering* (Vol. 44, pp. 2311-2316). Elsevier. https://doi.org/https://doi.org/10.1016/B978-0-444-64241-7.50380-3

Wiley, E. O., & Lieberman, B. S. (2011). *Phylogenetics: Theory and Practice of Phylogenetic Systematics*. John Wiley & Sons.

Yang, Z., & Rannala, B. (2012). Molecular phylogenetics: principles and practice. *Nat. Rev. Genet.*, *13*(5), 303-314. https://doi.org/10.1038/nrg3186

Young, B., Faris, T., & Armogida, L. (2021). Levenshtein distance as a measure of accuracy and precision in forensic PCR-MPS methods. *Forensic Science International: Genetics*, *55*, 102594. https://doi.org/https://doi.org/10.1016/j.fsigen.2021.102594

Zhu, Q., Mai, U., Pfeiffer, W., Janssen, S., Asnicar, F., Sanders, J. G., Belda-Ferre, P., Al-Ghalith, G. A., Kopylova, E., McDonald, D., Kosciolek, T., Yin, J. B., Huang, S., Salam, N., Jiao, J.-Y., Wu, Z., Xu, Z. Z., Cantrell, K., Yang, Y., . . . Knight, R. (2019). Phylogenomics of 10,575 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat. Commun.*, *10*(1), 1-14. https://doi.org/10.1038/s41467-019-13443-4

Zhu, X., Tang, L., Wang, Z., Xie, F., Zhang, W., & Li, Y. (2024). A comparative analysis of phage classification methods in light of the recent ICTV taxonomic revisions. *Virology*, *594*(110016), 110016. https://doi.org/10.1016/j.virol.2024.110016

APPENDICES

Appendix A – Levenshtein Distance Calculation

```
#Python work on Colab environment

import pickle
with open('/content/drive/MyDrive/Colab Notebooks/prophage_category_subset_16042024.pkl',
'rb') as f:
  phage_data = pickle.load(f)
data_type = type(phage_data)

#install Levenshtein python package
!pip install python-Levenshtein

# Print the new dictionary which replace '.' to '-'
new_phage_data = {}
for key, value in phage_data.items():
    new_key = key.replace(".", "-")
    new_phage_data[new_key] = value
print(new_phage_data)

#convert list to dataframe for analysis
import pandas as pd
df_n = pd.DataFrame.from_dict(new_phage_data, orient='index')
#Letters are represented to Categories, the different types of the gene in the phage.
letters_ndata = df_n['letters']

#Analyse Levenshtein distance from the dataframe
from Levenshtein import distance
def calculate_all_distances(letters_ndata):
  distances = []
  for i in range(len(letters_ndata)):
    text1 = letters_ndata[i]
    for j in range(i + 1, len(letters_ndata)):
      text2 = letters_ndata[j]
      distances.append(distance(text1, text2))
  return distances

all_distances = calculate_all_distances(letters_ndata)
print(all_distances)
```

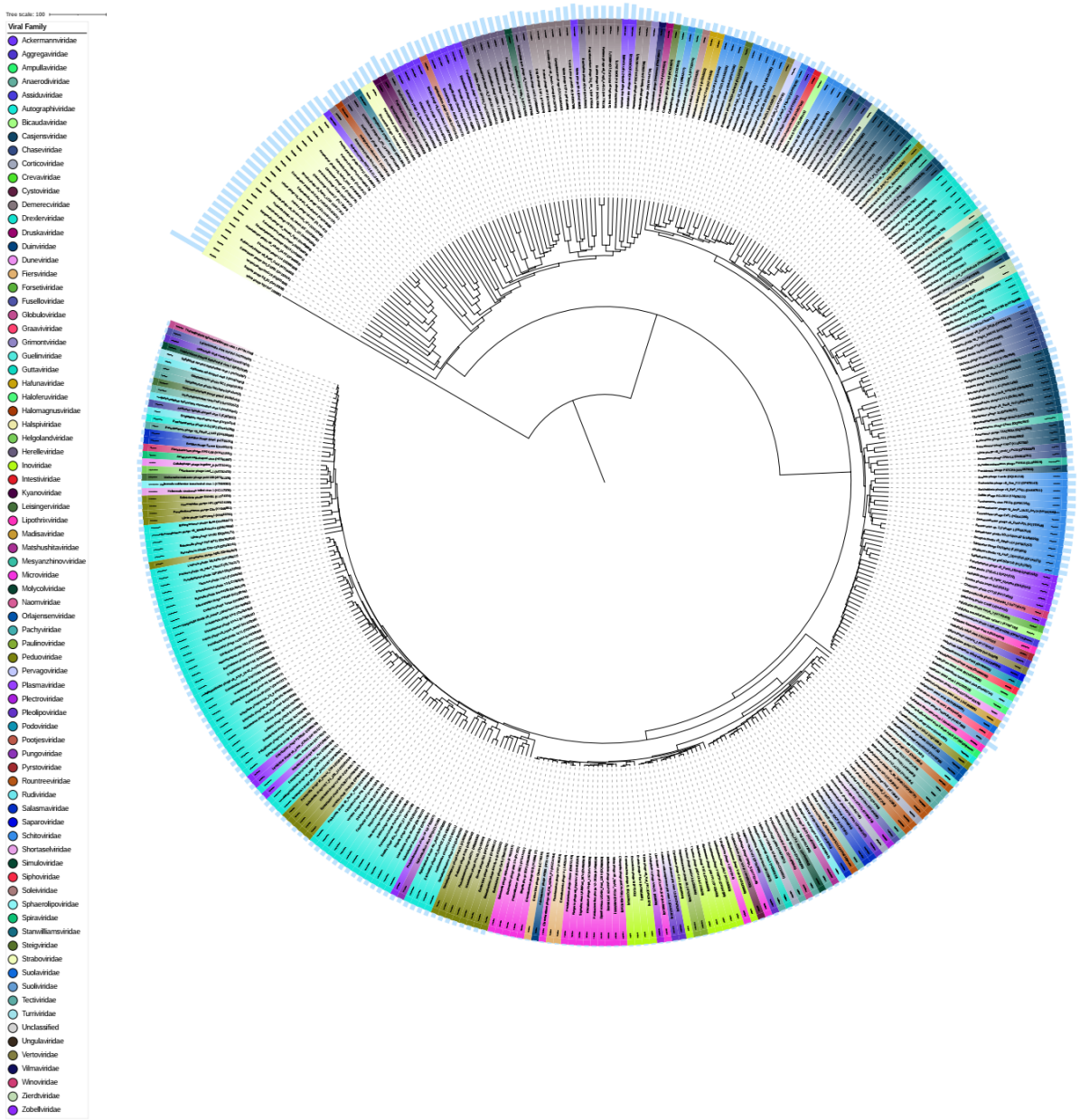
```
# Conversion Levenshtein distance to Newick in Python

import numpy as np
from scipy.cluster.hierarchy import linkage, to_tree

# take keys and categories to comparation
keys_and_categories = []
for key, value in phages_data.items():
    categories = value.get('categories', [])  # Handle cases where 'categories' might be
missing
    keys_and_categories.append((key, categories))
keys = [key for key, _ in keys_and_categories]
cate_ndata = [categories for _, categories in keys_and_categories]

# Perform hierarchical clustering by average linkage
Z = linkage(all_distances, method='average')

# Convert the linkage matrix to a tree object
tree = to_tree(Z, False)

# Define a function to get the Newick string
def NewickForm(node, newick, treedist, label_names):
    if node.is_leaf():
        return "%s:%.2f%s" % (leaf_names[node.id], treedist - node.dist, newick)
    else:
        if len(newick) > 0:
            newick = "):%.2f%s" % (treedist - node.dist, newick)
        else:
            newick = ");"
        newick = NewickForm(node.get_left(), newick, node.dist, label_names)
        newick = NewickForm(node.get_right(), ",%s" % (newick), node.dist, label_names)
        newick = "(%s" % (newick)
        return newick

# Get the Newick string with keys
newick_str = NewickForm(tree, "", tree.dist, keys)
print(newick_str)

#Get the Newick format for iTOL visualization by copy the print or save to *.txt or *.nwk
file directly
```

This same concept was also applied to the enhanced dataset of 381 phages, with the analysis demonstrated through the demo code.

Appendix B – Mash Distance Calculation

```
# Data preparation @Python runtime on Colab environment
!pip install biopython
from Bio import SeqIO

# Split Multi-Genome FASTA into Individual Files

input_fasta =
'/content/drive/MyDrive/Phage_genome_taxonomy/vcontact_host_family_genomes.fasta'
output_dir = '/content/drive/MyDrive/Phage_genome_taxonomy/split_genomes'

if not os.path.exists(output_dir):
    os.makedirs(output_dir)

with open(input_fasta, 'r') as handle:
    for record in SeqIO.parse(handle, 'fasta'):
        output_file = os.path.join(output_dir, record.id + '.fasta')
        with open(output_file, 'w') as output_handle:
            SeqIO.write(record, output_handle, 'fasta')

# Run Mash on bash terminal of Jupyter notebook
(base) [sion0008@hpc-head01 Mash]$ conda install bioconda::Mashtree
(base) [sion0008@hpc-head01 Mash]$ Mashtree split_genomes/* --outmatrix Mashtree_3H.mat >
3H_Mashtree.dnd
```

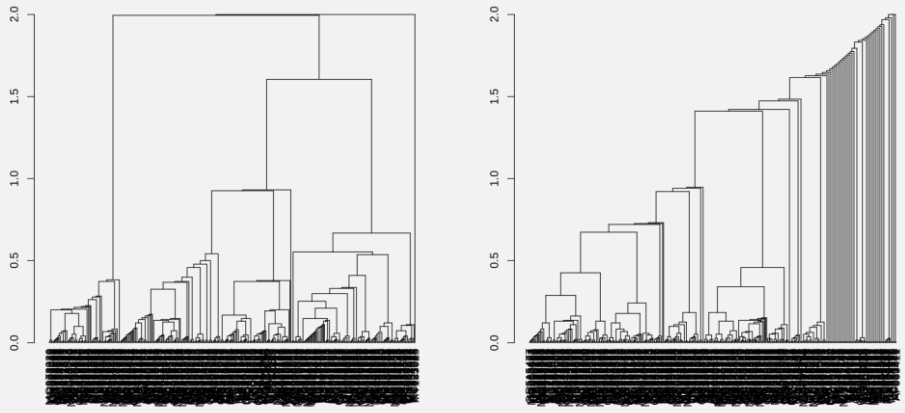
Appendix C – Phylogenetic Tree Generation



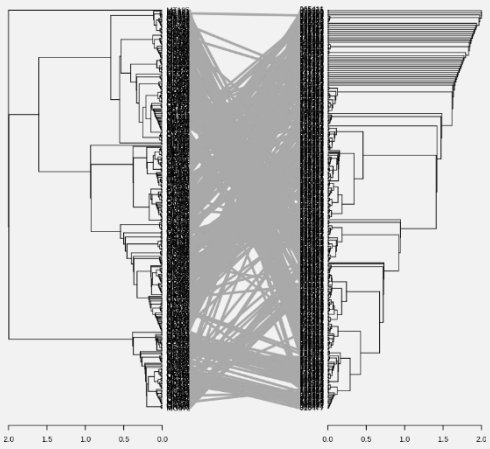
The phylogenetic trees of 381 phages, shown in Figure 9 and 10, were generated in iTOL. The project can be accessed at https://itol.embl.de/tree/1299685104484117237828 53 for the Levenshtein and https://itol.embl.de/tree/1299685104718917237833 11 for the Mash distance.

Appendix D – Tanglegram Generation

```
# R language work on Colab environment

install.packages('dendextend') #stable CRAN version
install.packages('ape',      repos      =      c('https://emmanuelparadis.r-universe.dev',
'https://cloud.r-project.org'))
install.packages("phytools")
install.packages("factoextra")
install.packages('hash')

library("dendextend") # load the package to make tanglegrams
library(ape)
library(phytools)
library(factoextra) #plot single dendrograms
library(cluster) #does the clustering
library(dplyr)
library(RColorBrewer)
library(hash) #R quivalent of dictionary

#Trees from external files
TreeLeven <-ape::read.tree("/content/3HLevens.nwk")
TreeMash <-ape::read.tree("/content/3HMashtree.nwk")

#Convert both trees to dendrograms
TreeLeven_cladogram <- compute.brlen(TreeLeven, method = "ED")  #Equal splits (Yule) model
TreeLeven_cladogram_rooted  <-  midpoint.root(TreeLeven_cladogram)  #Root  the  tree  at  its
midpoint
dend_leven  <-  as.dendrogram(TreeLeven_cladogram_rooted)  #Convert  the  rooted  tree  to  a
dendrogram
plot(dend_leven) # Convert to dendrograms

TreeMash_cladogram <- compute.brlen(TreeMash, method = "ED")
TreeMash_cladogram_rooted <- midpoint.root(TreeMash_cladogram)
dend_Mash <- as.dendrogram(TreeMash_cladogram_rooted)
plot(dend_Mash)
```



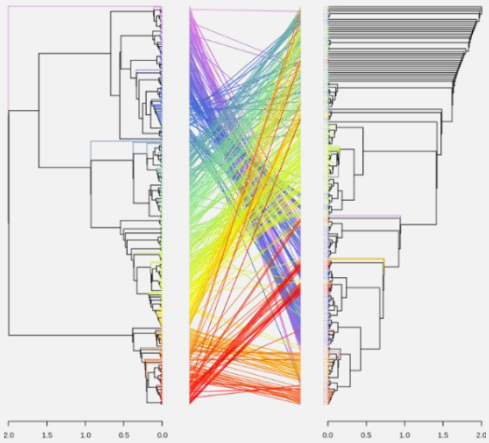
```
#Create a list of dendrograms
dend_list <- dendlist(dend_leven, dend_Mash)

#Create the tanglegram
tanglegram(dend_list, common_subtrees_color_lines = FALSE, highlight_distinct_edges =
FALSE, sort = FALSE, highlight_branches_lwd=FALSE, faster = TRUE)
```



```
#colours to pick from
print('Generating colours...')
#Plain tanglegram
n <- 381 #number of phages in dateset
colfunc<-colorRampPalette(c("red","yellow","lightgreen","royalblue", 'violet'))
col_n=(colfunc(n))

# Custom these dendrograms, and place them in a list
levensh_tree <- dend_leven %>%
    set("labels_col", value = '#FFFFFF', k=n) %>%
    set("branches_lty", 1) %>%
    set("branches_k_color", value = col_n, k = n)
```

```
#get the labels and colours
levensh_labels <- labels(levensh_tree)
levensh_colours <- get_leaves_branches_col(levensh_tree)

#make a vector of white colours to colour the labels in the dendrogram
white_col <- replicate(length(levensh_labels), '#FFFFFF')

#make dict
levensh_dict <- hash(levensh_labels, levensh_colours)

#get the Mash labels
Mash_labels <- labels(dend_Mash)

#generate the Mash colours
Mash_colours <- integer(length(Mash_labels))

for (i in 1:length(Mash_labels)){
  Mash_colours[i] <- levensh_dict[[Mash_labels[[i]]]]
}

print('Plotting Tanglegram...')
dl <- dendlist(
  levensh_tree,
  dend_Mash %>%
    set("branches_k_color", value = Mash_colours, k = length(Mash_colours)) %>%
      set("labels_col", value = '#FFFFFF', k = n)
)

tanglegram(dl,
           common_subtrees_color_lines = FALSE,
           highlight_distinct_edges  = FALSE,
           sort = FALSE,
           color_lines = get_leaves_branches_col(levensh_tree),
           highlight_branches_lwd=FALSE,
           faster = TRUE,
           lwd = 1,
           lab.cex = 0.1
)
#Left - Levensthine AND right – Mash
print('entanglement: ')
print(entanglement(dl))

[1] "Plotting tanglegram..."
[1] "entanglement: "
```

```
[1] 0.5554369
```



```
# Flip the right dendrogram (from 'dl' is a list of two dendrograms)
dl[[2]] <- rev(dl[[2]])
tanglegram(dl,
           common_subtrees_color_lines = FALSE,
           highlight_distinct_edges  = FALSE,
           sort = FALSE,
           color_lines = get_leaves_branches_col(levensh_tree),
           highlight_branches_lwd=FALSE,
           faster = TRUE,
           lwd = 1,
           lab.cex = 0.1
)
print('entanglement: ')
print(entanglement(dl))
[1] "entanglement: "
[1] 0.5755887
```