Bacterial and Viral Co-occurrence Across Samples in Lake Michigan to Predict Potential Infection

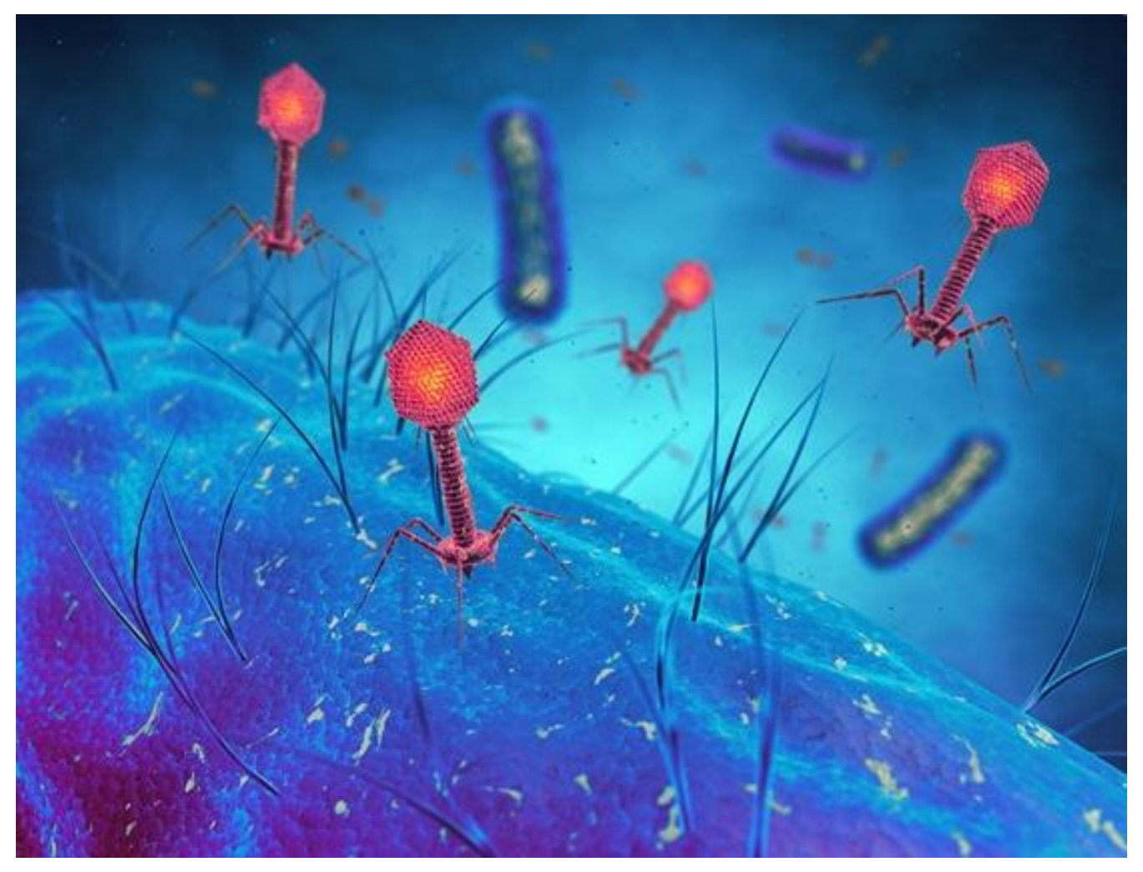
Tien T. VoNguyen[3], Ashley J. Mulford[1], Elyse C. Geoffroy[1], Mark V. Albert[3], Catherine Putonti[2,3]

(1) Bioinformatics Interdisciplinary Program, (2) Department of Biology, (3) Department of Computer Science

Loyola University Chicago

Background

- Viruses are unable to reproduce without a host
- Bacteriophages, a type of virus, use bacterial cells as their hosts
- Cell machinery is hijacked for replication of genetic material
- Host cell undergoes lysis and new viruses emerge



Bacteriophages invading and infecting a bacterial cell

Bacteriophages breaking out of host cell

Relevance to the Medical Field

- Growing antibiotic resistance of bacteria necessitates the development of alternative treatments for infection
- Bacteriophages infect bacterial cells that are harmful to humans, and can act as this alternative treatment
- Genome networks show the infection relationship between phages and bacteria, which is vital in developing these new treatments

Abstract

There is much unknown about the functions of viruses in the field of bioinformatics. Despite being present on all surfaces, only a small fraction of viruses are classified, whereas most bacterial functions are known. To further understand the function of viruses, one must also look at their interactions with certain bacteria. Viruses cannot live on their own, as they require a host cell to reproduce. In the case of bacteriophages, a specific type of virus, the host is a bacterial cell. The best way to increase the understanding of the infection relationship between specific phages and their host cells is through the analysis of co-occurrence across lake water samples. By seeing which bacteria and viruses are present together in multiple samples, the correlation between the two can be found, as viruses will only be able to reproduce through their corresponding bacterial host cells. The goal for this project is to create genome networks from Lake Michigan water samples to illustrate the potential bacterial and viral infection relationships. By placing various thresholds on the data and then running it through code, edge lists were created. These edge lists include the bacteria name, virus name, edge weight, and node weight. From there, the edge lists were imported into Cytoscape, a bioinformatics program that outputs genome networks, showing the co-occurrence in varying degrees between viruses and bacteria. While the networks cannot determine with certainty whether the virus actually infects the bacteria, it provides the visual representation of the correlation that exists across samples, which can be useful to bioinformaticians as they continue their study on viruses.

Data Collection

1. Thirty-one water samples collected from Lake Michigan over the course of twelve weeks
2. Genetic material isolated, sequenced, and inputted into BLAST algorithm for identification
3. Known species compiled into data tables with corresponding quantity value

Corresponding Viral Sample						V2S1	V1S5	V2S2	V2S3
Date						140513	140513	140513	140513
Kingdom	Phylum	Class	Order	Family	Genus				
Bacteria	Actinobacteria	Actinobacteria	Actinomycetales	unclassified	unclassified	16423	24034	13561	7675
Bacteria	Bacteroidetes	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Flavobacterium	26406	21836	32021	12547
Bacteria	Proteobacteria	Betaproteobacteria	Burkholderiales	Comamonadaceae	unclassified	8169	13392	13219	2306
Bacteria	Bacteroidetes	Sphingobacteria	Sphingobacteriales	Cytophagaceae	unclassified	8770	8606	17086	3725
Bacteria	unclassified	unclassified	unclassified	unclassified	unclassified	4523	3482	1285	2615

Bacterial data showing five species in table from four different lake water samples

Accession Number	Spp	Number of Proteins in Genome	Number of Hits to Genome	Percent Proteins Hit
NC_011165	Pseudomonas phage LBL3	88	248	95.45
NC_019503	Escherichia phage ime09	268	106	10.07
NC_024625	Peridroma alphabaculovirus	139	32	3.6
NC_002321	Staphylococcus phage PVL	62	20	11.29
NC_022987	Xylella phage Prado	52	24	9.62

Viral data showing the first five phages in table from a single lake water sample

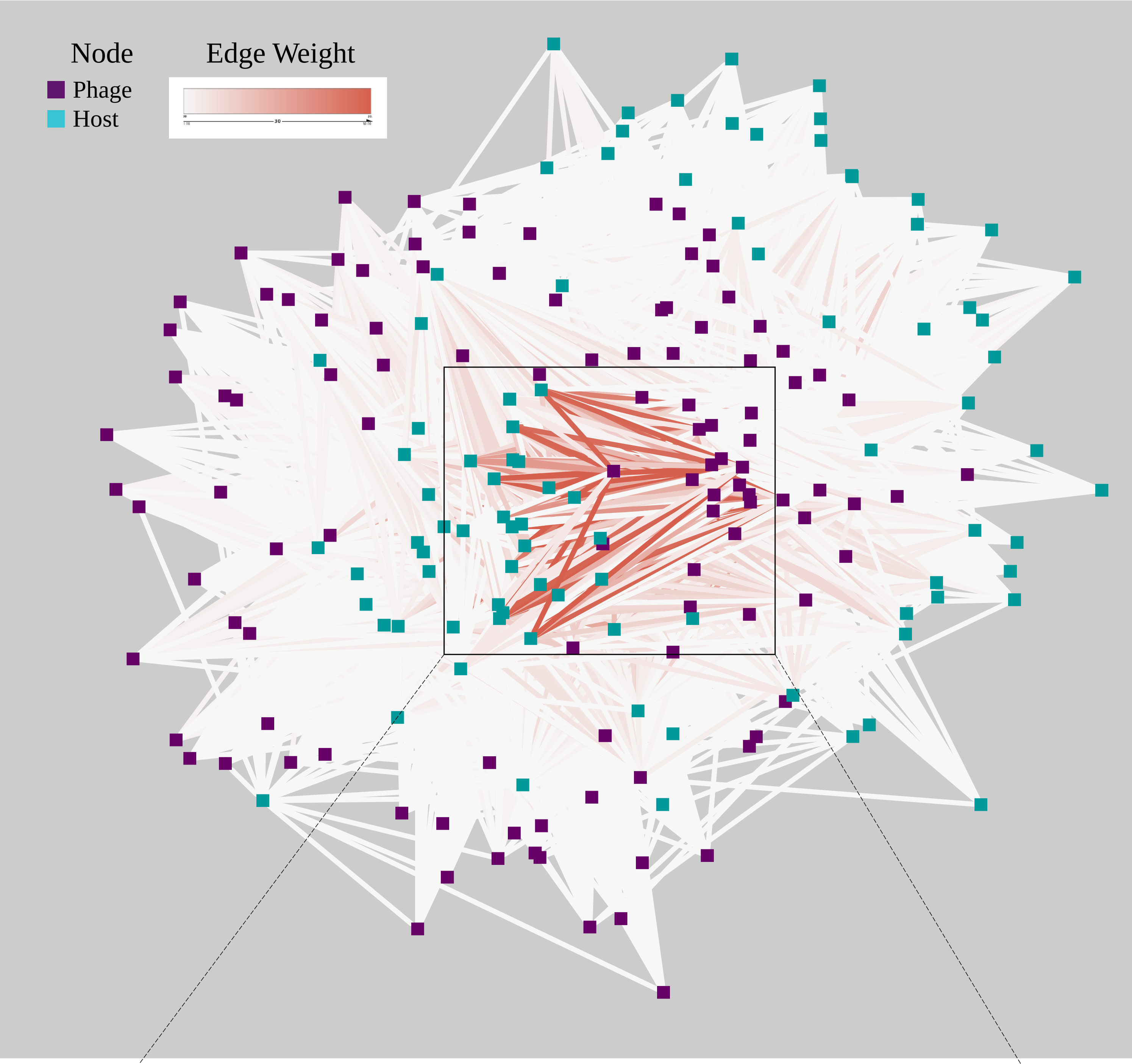
Data Processing

1. Begin with raw data, in the form of tables, containing species name and quantity
2. Python code in Jupyter Notebook reads in the data files and uses packages ‘numpy’ and ‘pandas’ to determine co-occurrence for different thresholds, set by the user
3. Edge lists containing two nodes for the species and a numerical value for the edge weight are produced
4. Insert edge lists into Cytoscape to produce the genome networks, depicting the relationships between the viruses and bacteria across samples

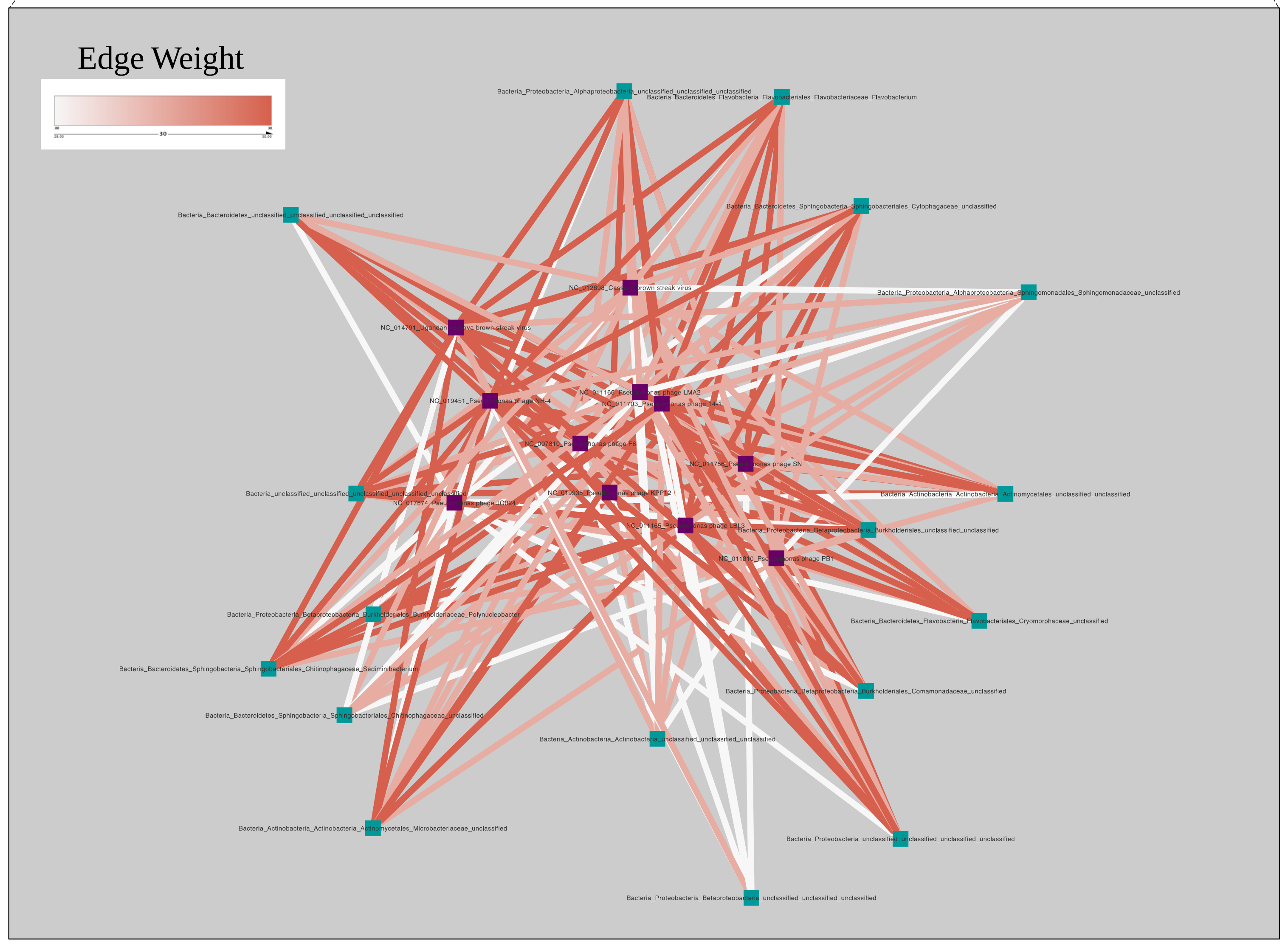
Edge Weight	Virus	Bacteria
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Actinobacteria_Actinobacteria_Actinomycetales_unclassified_unclassified
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Bacteroidetes_Flavobacteria_Flavobacteriales_Flavobacteriaceae_Flavobacterium
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Comamonadaceae_unclassified
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Bacteroidetes_Sphingobacteria_Sphingobacteriales_Cytophagaceae_unclassified
30	NC_011165_Pseudomonas phage LBL3	Bacteria_unclassified_unclassified_unclassified_unclassified_unclassified
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Bacteroidetes_Sphingobacteria_Sphingobacteriales_Chitinophagaceae_Sediminibacterium
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_Burkholderiaceae_Polynucleobacter
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Proteobacteria_Betaproteobacteria_Burkholderiales_unclassified_unclassified
30	NC_011165_Pseudomonas phage LBL3	Bacteria_Proteobacteria_unclassified_unclassified_unclassified_unclassified

Edge list with a 50% viral genome threshold and a bacterial cell count of at least 100. List includes quantity of corresponding samples, virus name, and bacterium name

Results



Genome network at a threshold of 50% viral genome found with a bacterial cell count of 100, showing all correlations across samples that could be indicative of infection relationship



Genome network at a threshold of 50% viral genome found with a bacterial cell count of 100, filtered to show strongest correlations only, indicating potential infection relationship

Acknowledgments

This project was made possible by:

- Loyola University Chicago Summer Research Program
- Loyola University Chicago F.Y.R.E. Program (2018)
- Loyola University Chicago Computer Science and Bioinformatics Departments