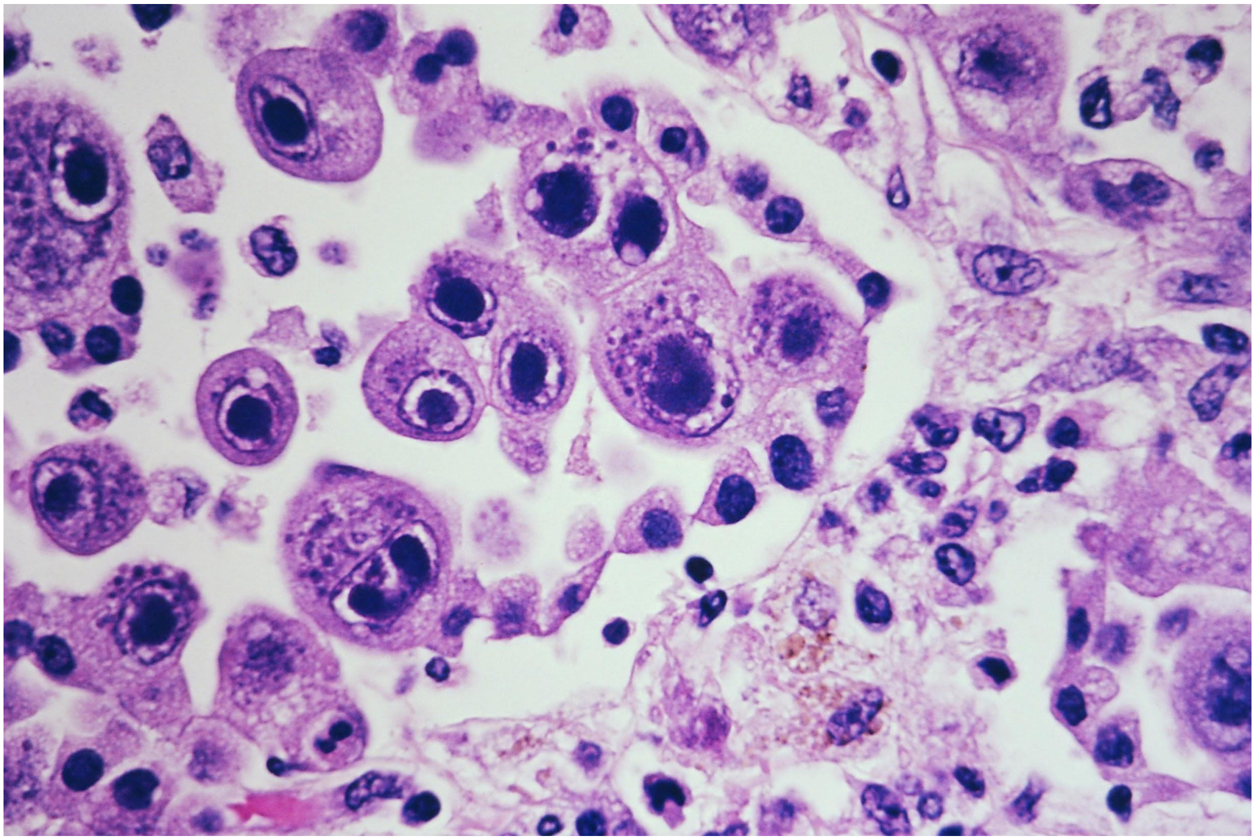
CS 676: Data Science Algorithms

A Phylogenetic Analysis of Mammalian Hosts with Cytomegalovirus Infection

By: Kelsey Woods & Aayushi Verma



Introduction

Cytomegalovirus, also commonly known as CMV, affects 1 in 200 babies every year, in 20 percent of cases causing major birth defects. However, most of the population does not know about it, including the most at risk population: pregnant mothers. There is no known model organism for CMV, meaning there is no animal that appropriately models how CMV would infect and affect a human. There are, however, similar animals that act as a precursor to human research for ethical and IRB purposes. Kelsey's lab at UMASS Chan

Medical School is one of the leading labs in transmission research. They are currently partnered with Moderna Therapeutics to look into daycare transmission rates and causes.

Methodology

Three samples were extracted from the NCBI Virus database: all of which are from projects Kelsey has analyzed previously. The samples were from three distinct species: the rhesus monkey, macaque monkey, and human. The host was not extracted from the samples and they had already been cleaned previously by Kelsey before their addition to NCBI. Further analysis was performed in Google Colaboratory.

The files were uploaded to Google Colab using a corpus format. They were uploaded in text format to create the corpus and then were subsetted into ‘id’ and ‘sequence’. Packages were uploaded and installed including sgt which is from a lab at Cincinatti Children’s that Kelsey worked with.

The sgt package allowed for a PCA and k means analysis to be performed between the three protein sequences after a nucleotide comparison. K was set to three to indicate the three sequences in use.

The SGT algorithm, which stands for Spectrum-based Graph Kernels, is an algorithm that computes similarity scores based on the similarity of graph spectra, which is the disctribution of eigenvalues of the graphs' Laplacian matrices. The SGT algorithm works like this:

Compute the Laplacian matrices of two graphs.

Compute the eigendecomposition of the Laplacian matrices to obtain corresponding eigenvalues and eigenvectors.

Compute the normalized eigenvector dot product between the two graphs' corresponding eigenvalues to obtain the similarity score.

Repeat steps 1-3 for all pairs of graphs in the dataset to obtain a graph similarity matrix.

The SGT algorithm can be applied to a variety of types of graphs. It is used very widely in the bioinformatics field, and some applications of the algorithm include for molecular graphs, protein structures, social networks, and more.

There is a Python package, SGT bioinformatics, that implements the SGT algorithm.

Results

None of the sequences appear to be highly relational in terms of phylogeny. This is to be expected because they are not known to be phylogenetically similar and none are models for the other.

Discussion

Further analysis should be performed using more species and samples. With the current grouping there are not enough samples to truly allow for the sgt algorithm to work effectively. Although the results are as expected, more accurate and precise results would come with a greater number of samples.

Conclusion

Although further and greater sampling would help immensely, it should be noted that the hypothesis was supported. The analytics team did expect to see highly dissimilar sequences because they are not highly related other than being from the same genus of mammalia. Further analysis would likely only support the findings and strengthen the hypothesis.

Citations

Sgt. PyPI. (n.d.). Retrieved April 23, 2023, from https://pypi.org/project/sgt/