Employing Genetic Similarities to Predict Characteristics of Microbes

Natalie Zhou, Xiongye Xiao, Paul Bogdan, Ph.D.

Dept of Electrical and Computer Engineering, Viterbi School of Engineering, University of Southern California, Los Angeles, CA , USA

Bridge UnderGrad Science (BUGS) ***Summer Research Program***

USC University of Southern California

Introduction

"The world is ill-prepared to respond to a severe influenza pandemic or to any similarly global, sustained, and threatening public health emergency," concluded an investigation into the World Health Organization's response to the 2009 H1N1 pandemic. Since then, our global community has been challenged with several outbreaks of pathogens, including MERS, Ebola, the Zika virus, and COVID-19, which have amounted to hundreds of millions of cases, millions of deaths, losses in the global Gross Domestic Product, and mass unemployment. In order to prepare for future pandemics, we must consider how we can use existing microbial information to identify similarities between microbes and predict characteristics of future microbes.

We can construct a microbial network on which we can conduct analysis by collecting the genomic sequences of 43 microbes from the National Center for Biotechnology Information (NCBI). Rather than the conventional method of comparing sequences by examining individual nucleotides, we propose the usage of the algorithmic-based genomic (AbG) distance. Thus, we construct a microbial network in which the nodes are microbes and the link between two microbes have weights equal to the genetic distance encoded in the microbes' state machines.

Methods

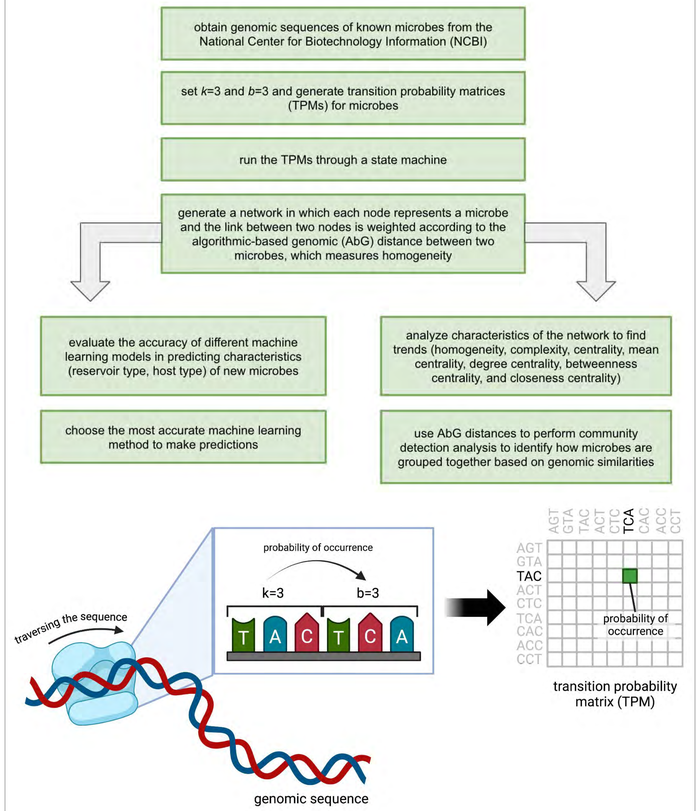

Fig. 1. Traversing the genomic sequence. *k* and *b* represent different segment lengths on a genomic sequence. We choose set values for *k* and *b* and determine the probabilities of the *b* segment having certain nucleotides in certain sequences if it is after a certain *k* segment. We then generate a transition probability matrix (TPM) using these probabilities. PC: Natalie Zhou

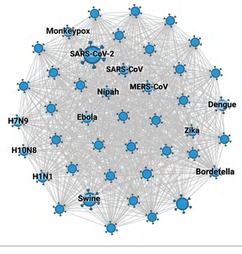

Fig. 2. Generating the network. Running the TPMs through a state machine generates a network in which each node represents the genomic sequence of a microbe, and all nodes have connecting links with a weight proportional to the algorithmic-based genomic (AbG) distance between the microbes. PC: Natalie Zhou

Network Analysis

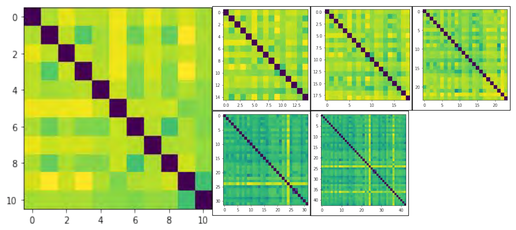

Fig. 3. Adjacency matrices to analyze the similarity of the network. The leftmost adjacency matrix is for the network in the year 2000. Moving clockwise, the rest of the adjacency matrices are for 2004, 2008, 2012, 2016, and 2020, respectively. Each number on the axes represents one microbe. The color at the intersection of the locations of two microbes shows the degree of similarity between the two. Darker colors correspond to greater similarity, which is why there is a dark, diagonal line running through each matrix (this is where variants intersect themselves). In recent years, the network has gained a greater degree of similarity. PC: Natalie Zhou

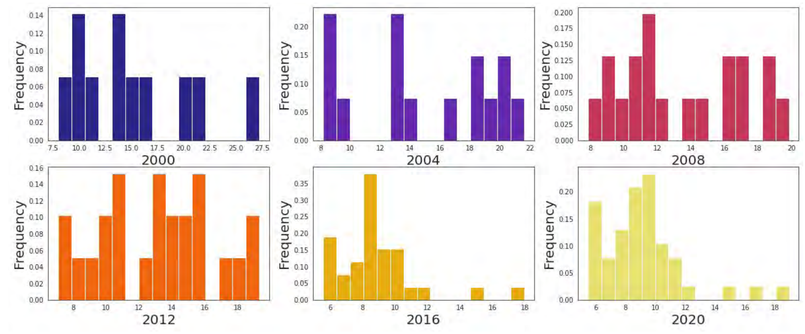

Fig. 4. Centrality of the network over time. The histograms show the frequencies at which the centrality values occur over time. The highest centrality occurred by 2000, and the centrality shows a trend of decreasing from 2000 to 2020. PC: Natalie Zhou

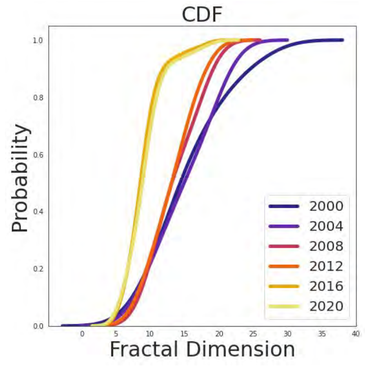

Fig. 5. Fractal dimension over time. Fractal dimension indicates the centrality of the network. The graph compares how the fractal dimension changes in the years 2000, 2004, 2008, 2012, 2016, and 2020. The fractal dimension is greater in more recent years, showing that the microbial network has decreased in centrality. PC: Natalie Zhou

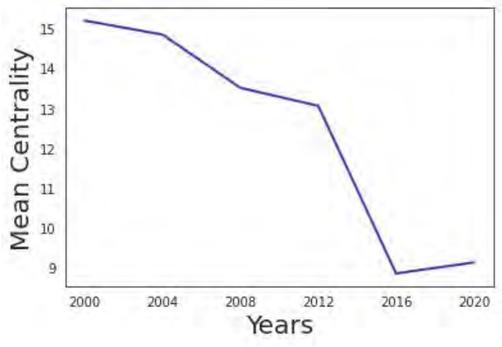

Fig. 6. Mean centrality of the network over time. We determine the mean centrality of the network for each year (2000, 2004, 2008, 2012, 2016, and 2020). The graph shows that the mean centrality has decreased over time, which is reasonable because other data analysis indicates that overall network centrality has also decreased over time. PC: Natalie Zhou

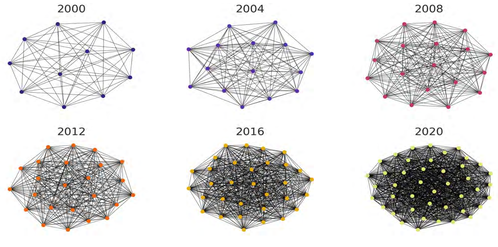

Fig. 7. Visual representations of the network over time. We start with a network consisting of 11 nodes (microbes) in 2000, and as more microbes are added to the network, more links are formed in between nodes. The links in this graphic are unweighted, but in the actual network each link between two nodes is weighted according to the algorithmic-based genomic (AbG) distance between the two microbes. The AbG distance indicates the degree of homogeneity, showing how two microbes are related—thus, we can see that homogeneity of the overall network increases with time. PC: Natalie Zhou

Community Detection

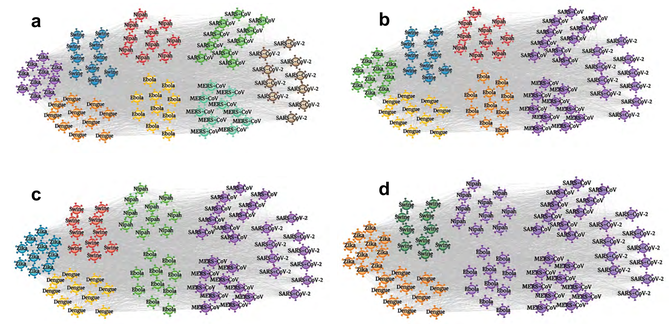

Fig. 8. Community detection analysis. We can compare various AbG distances between microbes to the mean AbG distance in order to group similar genomic sequences. The results of community detection analysis with 8, 6, 5, 3 communities are shown in (a, b, c, d) respectively. Thus, we can understand that the microbial network has a hierarchical structure and communities may share characteristics such as reservoir type due to similarities in the genomic sequence. PC: Natalie Zhou

Conclusions, Implications, and Future Direction

We can find the TPMs of various genomic sequences and put them through a state machine to generate a network. We analyze the network to determine the relationships between microbes, and when we have a new, emerging microbe we can predict its characteristics by observing where it fits into the network. We tested various machine learning models to predict characteristics of emerging microbes and were able to reach an accuracy of nearly 0.99 with the OvO (One-vs-One) and OvR (One-vs-Rest) models. The ability to quickly determine the characteristics of a new, unknown pathogen will allow us to:

- manufacture and stockpile personal protective equipment (PPE)
- create diagnosis and treatment plans earlier
- build appropriate infrastructure and supplies

This can help reduce the lasting effects of a pandemic, which include:

- high number of cases and deaths [1]
- decreased economic activity
- increased unemployment rates

In the future, we plan to do a similar project with SARS-CoV-2 variants instead of microbes so that we generate a network of SARS-CoV-2 variants and use machine learning to predict characteristics such as host type and reservoir type of an emerging variant based on its genomic similarities to other, existing variants.

Tools Used

python™ NetworkX Network Analysis in Python bio RENDER

Acknowledgements

Thank you Professor Bogdan for allowing me this amazing opportunity to work in your lab, and my mentor, Xiongye Xiao, for advising me throughout my research process.

References

[1]. World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 73." (2020)

CONTACT US

zhounatalie00@gmail.com | bridge.usc.edu/bugs