Employing Genetic Similarities to Predict Characteristics of Microbes

Natalie Zhou, zhounatalie00@gmail.com | Diamond Bar High School, Class of 2024
USC Viterbi Department of Electrical & Computer Engineering, SHINE 2022

Introduction

- The conventional method to compare two genomic sequences is to traverse each length and compare the nucleotides
- We can use a different method that involves **transition probability matrices (TPMs)**
- The TPMs of genomic sequences can be used to build a network
- We can analyze the network to determine the relationships between the various genomic sequences

Objective & Impact of Research

In Professor Paul Bogdan's Cyber-Physical Systems Lab, we are analyzing data sets of microbes throughout history and applying network science and machine learning concepts to determine the relationships between the microbes. We generate transition probability matrices (TPMs) for genomic sequences of microbes and then run the TPMs through a state machine [1]. Then, a network (in which the microbes are nodes and edge weights are indicated by TPM similarity) can be created. We can use the network to determine similar characteristics between microbes. The genomic sequences of new, unknown microbes can also be analyzed and compared with existing, known microbes to quickly determine the characteristics of a new microbe.

What are the social impacts of this research?

Being able 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 [2]
- decreased economic activity
- increased unemployment rates

Skills Learned + Data Analysis

1. Network science and graph theory: components and characteristics of a network (Fig. 1)
2. Python programming and Networkx graph library
3. Generating transition probability matrices (TPMs) (Fig. 2)
4. Generating and using the following graphs to analyze the network:
   - adjacency matrices (Fig. 3)
   - histograms measuring centrality over time
   - a graph measuring CDF (Fig. 4)

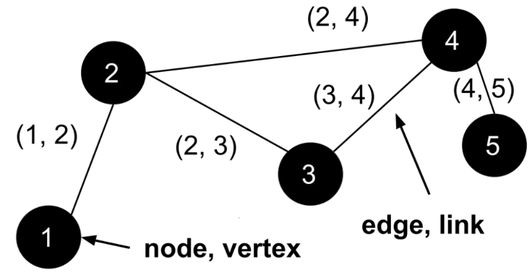

Fig. 1. The components of a network. N is the number of nodes, or the size of the network. Each node is labeled with a number. L is the number of links, and each link is labeled using the two nodes that it connects. PC: Natalie Zhou

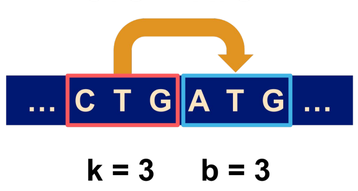

Fig. 2. *k* and *b* segments of a genomic sequence. *k* and *b* represent different segment lengths on a genomic sequence. We choose set values for *k* and *b* and determine the probabilities of the *b* segment having certain nucleotides in certain sequences if it is after a certain *k* segment. We then generate a TPM using these probabilities. PC: Natalie Zhou

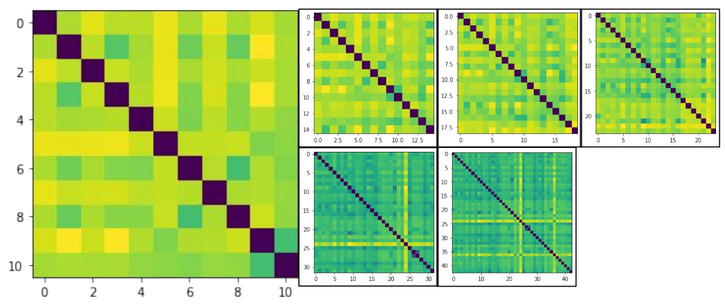

Fig. 3. Adjacency matrices to analyze the similarity of the network. The leftmost adjacency matrix is for the network in the year 2000. Moving clockwise, the rest of the adjacency matrices are for 2004, 2008, 2012, 2016, and 2020, respectively. Each number on the axes represents one microbe. The color 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

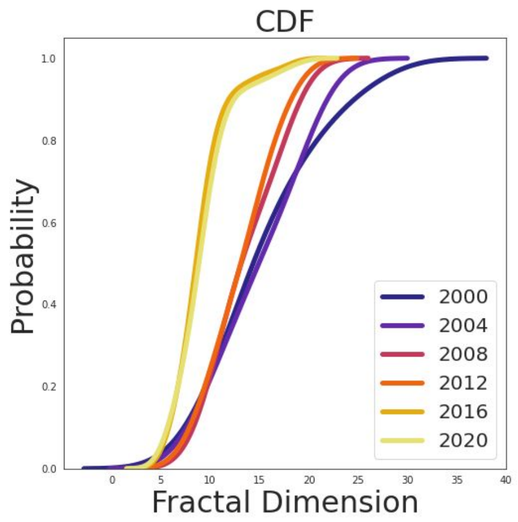

Fig. 4. Fractal dimension over time. Fractal dimension indicates the centrality of the network. The graph compares how the fractal dimension changes in the years 2000, 2004, 2008, 2012, 2016, and 2020. The fractal dimension is greater in more recent years, showing that the microbial network has decreased in centrality. PC: Natalie Zhou

Conclusion

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, unknown microbe we can predict its characteristics by observing how it fits into the network.

Acknowledgements

Thank you Professor Bogdan for allowing me to have this amazing opportunity to work in your lab! I also want to thank Dr. Mills and the rest of the SHINE team for providing support and guidance throughout my SHINE journey. In addition, a massive thank you to Xiongye Xiao for mentoring me and helping me develop skills for SHINE. Lastly, thank you to my lab partner Julia Gong for always making my day!

References

[1]. Jain, Siddharth, Xiongye Xiao, Paul Bogdan, and Jehoshua Bruck. "Generator based approach to analyze mutations in genomic datasets." Scientific reports 11, no. 1 (2021): 1-12.
[2]. World Health Organization. "Coronavirus disease 2019 (COVID-19): situation report, 73." (2020)

SCAN ME

Scan to see the methodology and additional data analysis!