Analyze Genetic Distance Among Microbes by Microbial Network Community Detection

Zonglin Zhang, 24zhangzonglin@gmail.com | Portola High School, Class of 2024
USC Viterbi Department of Electrical & Computer Engineering, SHINE 2023

Introduction

- Building the **Transition Probability Matrix(TPM)** of the gene sequence for comparing the genetic similarity of sequences with different lengths
- Using the network community detection function to classify network nodes into cohesive communities on the basis of their connection patterns
- We use the **Ollivier-Ricci Curvature(ORC)** community detection method to obtain k(k=2 to 8) communities in order to determine the hierarchical community structure of the microbial network
- Using Gephi to visualize the network with hierarchical community structure

Objective & Impact of Research

The Cyber-Physical Systems Lab, led by Professor Paul Bogdan, employs network science and Community Detection principles to determine the relationships between microorganisms by analyzing historical microbial datasets. We generate transition probability matrices (TPMs) [1] for microbial genome sequences and then apply a state machine to these TPMs. Using constructed microbial networks and Ollivier-Ricci curvature community detection, communities were classified. The ORC Community Detection eliminates the edge with lowest ORC value in the network, as the edge with lowest ORC value is the bridge of two communities. Using ORC Community Detection, we can classify unknown viruses as community rapidly.

This research will have the following impacts:

- It will enhance the ability to quickly identify the characteristics of previously unidentified pathogens.
- It will aid in formulating procedures for early diagnosis and treatment of newly discovered diseases.
- It will contribute to the development of necessary infrastructure and assets needed to combat these diseases effectively.
- It will improve our capacity to infer the host of newly emerging microbes, which is crucial in understanding their transmission and implementing control measures.

Methods & Skills learned

Skills learned :

1. Building Transition Probability Matrix(TPM) (Fig. 1)
2. Ollivier-Ricci Curvature and Networkx Community Detection Algorithm (Fig. 2)
3. Gephi - The application that enables the network community to visualize (Fig. 4)

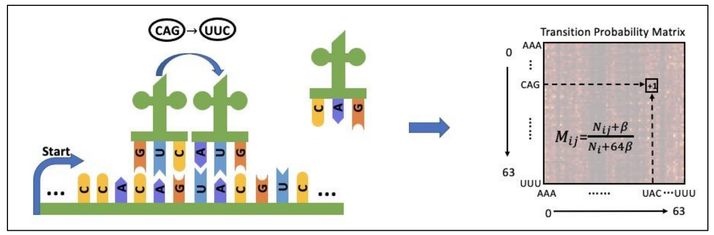

Fig. 1. Transition probability matrix of the gene sequence. We consider each 3-mers as the state in a Markov chain process, and traverse the whole sequences to get the transition probability matrix. PC: Zonglin Zhang

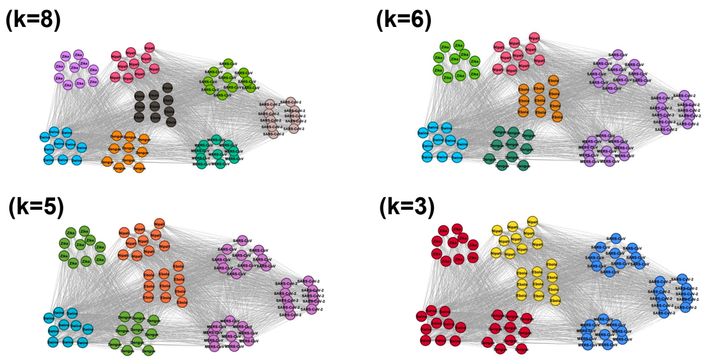

Fig. 4. The visual hierarchical microbial community detection. When k=8, the virus samples in each microbe are correctly identified show as graph above. When k=6, SARS-CoV, SARS-CoV-2, and MERS-CoV samples combine to create one community, while other microbial samples establish other communities. This shows that SARS-CoV, MERS-CoV, and SARS-CoV-2 have the smallest genetic distances at this time.when k=5, Nipah and Ebola merge into a community. When k=3. Dengue and Zika virus communities converge. Coincidentally, these two viruses are of the same reservoir type (insect). PC: Zonglin Zhang

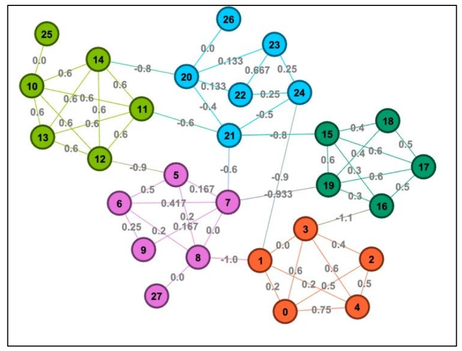

Fig. 2. The concept of the Ollivier-Ricci curvature(ORC). Two fundamental structural properties of complex networks are illustrated by the ORC: 1.Deleting the edge with the lowest ORC value. 2.The occurrence of triangles increases the cluster coefficient.[2] PC: Bogdan Paul

Suppose we have two gene sequences with corresponding TPMs A and B, the AbG distance between these two sequences should be:

$$\|A-B\|_F=\sum_{i=1}^{N}\sum_{i=1}^{N}\left|(a-b)_{ij}\right|^2$$

3(a)

	source	target	weight
0	0	1	0.042893
1	0	2	0.045524
2	0	3	0.045970
3	0	4	0.031861
4	0	5	0.036039
...	...	...	...
3154	76	78	0.021430
3155	76	79	0.039296
3156	77	78	0.018268
3157	77	79	0.037509
3158	78	79	0.032239

3(b)

	id	label	cluster_8	cluster_7	cluster_6	cluster_5	cluster_4
0	0	SARS-CoV	1	1	1	1	1
1	1	SARS-CoV	1	1	1	1	1
2	2	SARS-CoV	1	1	1	1	1
3	3	SARS-CoV	1	1	1	1	1
4	4	SARS-CoV	1	1	1	1	1
..	..	...	...	...	...	...	...
75	75	SARS-CoV-2	7	1	1	1	1
76	76	SARS-CoV-2	7	1	1	1	1
77	77	SARS-CoV-2	7	1	1	1	1
78	78	SARS-CoV-2	7	1	1	1	1
79	79	SARS-CoV-2	7	1	1	1	1

Fig. 3 The dataset for Gephi. Import the cluster result of ORC community detection in the csv document. **Fig. 3(a)** is the edge list including: source, target, weighted. **Fig. 3(b)** is the node table including: id, label, cluster 2 to 8(the result of the ORC community detection). PC: Zonglin Zhang

Conclusion

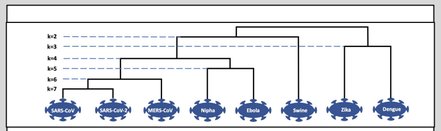

Fig.5. PC: Zonglin Zhang

In this figure, when k = 3, the Dengue and Zika populations coalesce into one. These two microbes share the same reservoir. Notable is the fact that SARS-CoV, SARS-CoV-2, MERS-CoV, Nipah, and Ebola viruses form a single community and share the same reservoir (bats), whereas Swine flu is an independent organism belonging to avian influenza. Therefore, we are able to infer the reservoir of the new emerging microbe.

Acknowledgements

I'm deeply appreciative of the opportunity Professor Bogdan has given me to work in the Cyber Physical lab. I'd also like to extend my heartfelt thanks to Monica Lopez and the entire SHINE team for their invaluable assistance so far. Furthermore, I am immensely grateful to Xiongye Xiao, who has played a crucial role as a mentor and has generously helped me prepare for high level research..

References

[1]. Jain, S., Xiao, X., Bogdan, P. et al. Generator based approach to analyze mutations in genomic datasets. Sci Rep 11, 21084 (2021). https://doi.org/10.1038/s41598-021-00609-8
[2]. Sia, J., Jonckheere, E. & Bogdan, P. Ollivier-Ricci Curvature-Based Method to Community Detection in Complex Networks. Sci Rep 9, 9800 (2019). https://doi.org/10.1038/s41598-019-46079-x