OVERVIEW

Data analysis is a critical skill in this era of large biological data sets. For example, an abundance of genomic sequence data already exists *but it is meaningless until it is analyzed.* For example, the SARS-CoV2 genome had to be compared to other known coronaviruses in order to determine that it is most closely related to the SARS virus from 2003.[1] Annotation of the SARS-CoV2 genome revealed the presence of the spike protein, which directly interacts with the host cell, and bioinformatic comparison of this spike protein to other coronavirus spike proteins revealed a potential receptor on the host cell [2] that was later verified by wet lab experiments.[3] Further analysis of SARS-CoV2 genome sequences from around the world using phylogenetic tools allowed tracking of virus variants and identification of newly emerged variants as the virus spread across the globe.[1] Clearly, data analysis is powerful and provides meaning for sequence data.

Phylogenetic trees and networks are created by analysis of either nucleic acid or protein molecular sequence data, and they help support predictions of evolutionary relationships between organisms. Phylogeny is a complex field with several different methods used to create trees and networks, such as neighbor-joining, parsimony, maximum likelihood and Bayesian methods. Different tree building programs will use different methods and algorithms. One way to validate their results is to analyze the data using different methods and comparison of the results. In addition, bootstrapping is also used to validate phylogenetic trees and networks.

A key point to remember is that phylogenetic trees and networks are not right or wrong- just mostly right, and the goal is to get the best representation of your data that you can get.

These exercises will introduce you to bioinformatic tools used to analyze sequences using bacteriophage phylogeny as a focus. Bacteriophages are viruses that specifically infect and kill bacterial hosts, helping to maintain bacterial levels in nature and with potential uses to treat bacterial infections. Clustering of phages based upon genome sequence similarities is used to categorize phages, and phylogenetic analysis reveals evolutionary relationships of these clusters. The SEA-PHAGES program has exponentially increased the level of phage genomic data available for analysis. [4]

References:

1. Xiaolu Tang, Changcheng Wu, Xiang Li, Yuhe Song, Xinmin Yao, Xinkai Wu, Yuange Duan, Hong Zhang, Yirong Wang, Zhaohui Qian, Jie Cui, Jian Lu, On the origin and continuing evolution of SARS-CoV-2, *National Science Review*, Volume 7, Issue 6, June 2020, Pages 1012–1023, https://doi.org/10.1093/nsr/nwaa036
2. Wan Y, Shang J, Graham R, Baric RS, Li F. Receptor Recognition by the Novel Coronavirus from Wuhan: an Analysis Based on Decade-Long Structural Studies of SARS Coronavirus. *J Virol*. 2020;94(7):e00127-20. Published 2020 Mar 17. doi:10.1128/JVI.00127-20
3. Ou X, Liu Y, Lei X, et al. Characterization of spike glycoprotein of SARS-CoV-2 on virus entry and its immune cross-reactivity with SARS-CoV. *Nat Commun*. 2020;11(1):1620. Published 2020 Mar 27. doi:10.1038/s41467-020-15562-9
4. Russell DA, Hatfull GF. PhagesDB: the actinobacteriophage database. *Bioinformatics*. 2017;33(5):784-786. doi:10.1093/bioinformatics/btw711