Network Inference and Dimensionality Reduction in Biome-Specific Virus Data

Gorka Buenvarón-Campo[1], Mar Cuevas-Blanco[1], Carlos M. Duarte[2], and Victor M. Eguíluz[3,4]
[1]Institute for Cross-Disciplinary Physics and Complex Systems, E-07122 Palma, Mallorca, Spain
[2]Red Sea Research Center, King Abdullah University of Science and Technology, Kingdom of Saudi Arabia
[3]Basque Centre for Climate Change (BC3), Scientific Campus of the University of the Basque Country, 48940 Leioa, Spain
[4]IKERBASQUE, Basque Foundation for Science, 48009 Bilbao, Spain

Viruses play vital roles in diverse ecosystems, influencing dynamics and genetic diversity through their interactions with host organisms. Investigating the structure and distribution of viruses in across biomes is crucial for understanding their ecological significance and the potential impacts on ecosystem functioning. In this study, we analyze virus data provided by the KAUST Metagenomic Analysis Platform (KMAP) [1]. We aim to unravel taxonomic relationships and occurrences of viruses, shedding light on their role in different environments.

The data consists of files associated with different PFAM domains containing measurements of virus taxa at various biomes. By aggregating virus families and combining these files, we constructed a comprehensive dataset encompassing a wide range of virus taxa and their occurrences across biomes. The participation of viruses across PFAM domains and biomes exhibits significant heterogeneity, ranging from generalist viruses that occur in multiple domains and biomes, to highly specific or even exclusive viruses that are confined to particular PFAM domains and reported biomes. For the analysis of this complex dataset, we employed two distinct approaches that share similar conceptual foundations, providing complementary insights into the taxonomic relationships and distributions of viruses in different ecosystems. The first approach involved **inferring networks**, and the second utilized clustering algorithms, specifically **UMAP** (Uniform Manifold Approximation and Projection) [2], for data visualization and exploration.

We employed a two-step approach for network inference. Firstly, we combined the individual data frames associated with every PFAM domain. Pairwise similarity Eq. (1) between viruses were then calculated using a metric derived from the **Jensen-Shannon divergence** [3]. Significance was assessed through randomization, and a network was constructed, with virus families represented as nodes and edge weights determined by similarity measures. Additionally, a multilayer network analysis was conducted, treating each PFAM dataset as a separate layer. Using the Infomap algorithm [4], we perform community detection on both networks to unveil modular structures across diverse biomes.

As an alternative approach to network inference, we employed **UMAP** for data visualization and exploration. Using the combined data frame described earlier, we applied UMAP with the Jensen-Shannon distance as the metric. By iteratively adjusting the parameters, we aimed to strike a balance between capturing global and local structure, ultimately obtaining a visualization, shown in Fig. 1, that revealed clustering patterns and provided insights into the overall structure of the virus data.

$$s_{PQ} = 1 - \frac{1}{\sqrt{2}} \left[\sum P \log \frac{2P}{P+Q} + \sum Q \log \frac{2Q}{P+Q} \right]^{\frac{1}{2}}. \tag{1}$$

Here we present two figures (Fig. 1) showcasing the outcomes of UMAP analysis. These visualizations depict the embedding of virus families in a two-dimensional space, allowing us to observe the formation of clusters and the relative distances between viruses. Each virus is represented by its name. Fig. 1A shows the color-coding of viruses based on the communities detected in the multilayer network, while Fig. 1B represents the color-coding based on the communities detected in the combined network. The spatial distribution of the virus families is better explained by the communities found at the joint network, while they do not match completely.

This study investigates the taxonomic relationships and occurrences of viruses in diverse environments. The results demonstrate the potential of network inference in capturing the underlying structure of virus communities. Additionally, various avenues, including embedding projection, statistical analysis, and bipartite network construction, are being pursued to further elucidate the complexity and dynamics of viral ecosystems.

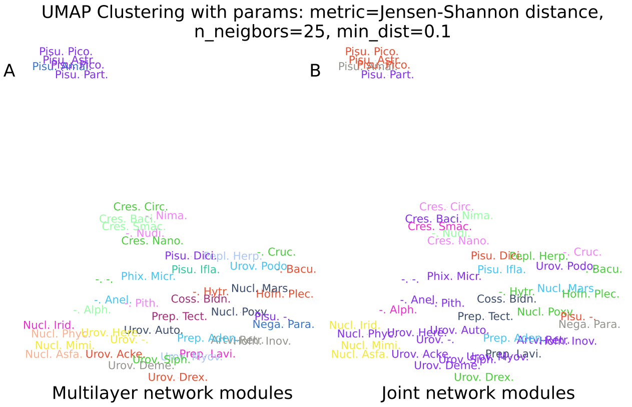

Fig. 1. UMAP results after embedding the virus space in 2 dimensions. The color of names refers to the communities detected in the multilayer (A) and joint (B) networks

[1] I. Alam, A. A. Kamau, D. K. Ngugi, et al., *KAUST Metagenomic Analysis Platform (KMAP), enabling access to massive analytics of re-annotated metagenomic data*, *Scientific Reports* **11**, 11511 (2021), https://doi.org/10.1038/s41598-021-90799-y.

[2] L. McInnes, J. Healy, and J. Melville, *UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction*, arXiv preprint arXiv:1802.03426 (2018).

[3] J. Lin, *Divergence measures based on the Shannon entropy*, Information Theory, IEEE Transactions on **37**(1), 145-151 (1991).

[4] D. Edler, A. Holmgren, and M. Rosvall, *The MapEquation software package*, https://mapequation.org, 2023.