Applying Topological Principles to Genomic Analysis

Topological data analysis helps to overcome difficulties that conventional tree-based phylogenetic analyses have with horizontal gene transfers

Kevin Emmett, Joseph M. Chan, Daniel S. Rosenbloom, and Raul Rabadan

In On the Origin of the Species in 1859, Charles Darwin proposed phylogenetic trees to depict evolutionary relationships on the basis of phenotypic attributes. Since then, biologists have shifted to using molecular traits to characterize evolutionary relationships between and among species. These approaches typically assume that traits are inherited vertically, and that evolution is strictly clonal. However, several other modes of genetic exchange add complexity to this process, including lateral gene transfers in bacteria, recombination and reassortment in viruses, viral integration in eukaryotes, and fusion of genomes of symbiotic species.

For several decades, evolutionary relationships among microorganisms were inferred mainly from sequences of genes encoding 16S ribosomal RNA molecules, a highly conserved genomic region for both bacterial and archaeal species. However, this region accounts for less than 1% of the complete genome in most species, meaning that this approach omits the vast majority of genetic information. Moreover, because horizontal inheritance is pervasive among microorganisms, the remaining 99% of genes may tell a very different evolutionary story. This problem becomes more acute for viruses, which lack 16S or other universally shared genes.

Alternative Approaches to Depicting Evolutionary Relationships Are Needed

These challenges underscore the need for alternative approaches with which to analyze evolutionary relationships, going beyond the traditional phylogenetic tree representation (see Microbe, August 2015, p. 319). The rapidly growing number of sequenced microbial genomes provides fertile ground for developing such alternative approaches to analyzing and depicting both vertical and horizontal evolutionary processes. While recent developments in phylogenetic networks provide ways to identify instances of reticulate evolution, the field does not yet have a widely accepted framework for visualizing and quantifying the frequency, scale, and significance of horizontal evolution.

Although phylogenetic trees can look complicated, in a sense they are very simple mathematical objects. They are composed of one-dimensional objects—branches—and contain no loops (Fig. 1).

To represent horizontal evolution, though, more complex objects are needed. Topology gives us a framework for characterizing the global properties of mathematical objects. According to some mathematicians, topology was born with Bernhard Riemann in the mid-1800s. Through conversations and letters to friends and colleagues, Riemann characterized the minimal number of cuts needed to transform a complex object into something simpler and without loops,

SUMMARY

➤ Horizontal gene transfers, particularly among microorganisms, complicate efforts to analyze and depict phylogenetic relationships.
➤ Topological data analysis provides a novel approach to characterizing both vertical and horizontal modes of evolution that are embedded in genomic datasets.
➤ Aligned genomic sequences can be viewed as points in a high-dimensional sequence space; applying persistent homology analysis to these spaces then enables us to read off phylogenetic information from resulting barcode diagrams.
➤ This analytic approach can be used to gain insights into the recent evolution of viral and bacterial pathogens.
like a tree. The number of cuts needed to make this change is called a topological invariant, meaning that if we deform the space continuously by stretching and compressing it, this number will not change. For example, an object with a loop can be continuously deformed into a circle (Fig. 1A). However if we cut that loop once, we obtain a tree that can be deformed into a point. The minimal number of cuts needed to transform an object into a tree is a topological invariant, called the first Betti number.

Here, we describe a topological approach to characterize both vertical and horizontal modes of evolution in genomic data simultaneously. The approach is based on recent developments in the field of computational topology and its application to large datasets, loosely known as topological data analysis.

Intuitively, our aim in applying topology to phylogenetic analyses is to represent and quantify loops in genomic space. We will approximate the space of genomes by an object that is neither a tree nor a network, but rather a higher-dimensional mathematical object known as a simplicial complex. Computational approaches can be used to read off properties of these complexes.

We first applied this approach to viral evolution—particularly that of the influenza virus, where genomic reassortment can lead to new versions of this virus that can cause pandemics.
Further work established the characteristic genomic scales at which recombination in bacteria leads to the emergence of antibiotic resistance in several strains of pathogens. While the application of topological data analysis to genomic data remains in its infancy, we hope our introduction to this elegant new framework spurs interest in its development and, more broadly, recognition that mathematical ideas once thought to be without application can help in addressing unsolved biological questions.

**Topology and Its Use in Data Analysis**

Topology is the branch of mathematics that characterizes spaces and how they can be deformed. If we take a tree and change the lengths of its branches or extend new branches from a point, the tree remains a tree. In the same way, if we take a cup, we can smoothly mold it into a donut without cutting or tearing it apart. We say that the cup and the donut have the same topology.

How can we assess if two separate spaces share the same topology? One trick is to assign to each space some algebraic object—such as a number, for instance—that does not change under deformation. That object is called a topological invariant. For instance, we can assign to an object the number of loops, which in the case of the cup will be one, and in the case of the tree will be zero. We can continuously deform only those spaces with the same invariants. Thus, for example, we cannot deform a tree into a donut without cutting or merging different sections. Algebraic topology provides tools to compute invariants of these spaces.

However, genomic sequence data—or other data from real-world analyses—does not come in the form of a perfect continuous space, as dealt with by classical topology. Instead, such data can be viewed as a high-dimensional point cloud forming a discrete representation of a space. Topological data analysis (TDA) refers to a framework that took shape during the last 15 years to compute topological properties from finite point clouds, building on developments in computational topology and statistics.

Our primary tool is persistent homology, a branch of TDA that computes topological invariants representing multiscale information about the connectivity and holes within a dataset. “Per-
Persistent homology is a method that captures structures in a range of spatial resolutions, often called topological features. The method uses a scale parameter to define a notion of distance in a high-dimensional space, connecting points that are closer than the scale. When three points connect with a line, a triangle is filled in, forming a simplicial complex. As the scale parameter increases, new connections are made, and topological features are identified and tracked in a barcode diagram. This allows for the extraction of information about the shape of the data, including the number of connected components, loops, and higher-dimensional features.

Using TDA to handle genomic sequence data involves aligning and comparing genomes to create a phylogenetic tree. However, horizontal gene transfer and other evolutionary events can complicate this view, creating a non-tree-like structure. Persistent homology can detect these non-tree-like features, providing a more nuanced understanding of evolutionary relationships.
will be captured in the barcode diagram as higher-dimensional homology.

Here is a simple example of how TDA can capture horizontal evolution from population data (Fig. 2). Consider this reticulate phylogeny (Fig. 2A): five genetic sequences sampled recently (yellow circles) originate from a single common ancestor due to clonal evolution (solid blue lines, tracing parent to offspring) and recombinant evolution (dotted red lines).

These five sequences can also be placed in the context of a larger sample, where the data are projected into two dimensions using PCA (Fig. 2B). PH is applied to this larger sample (Fig. 2C), resulting in a barcode pattern (Fig. 2D). The $H_1$ bar near the center (Fig. 2D) identifies a recombinant event involving the five highlighted sequences. The scale over which this bar persists represents the evolutionary distance between two parents of the recombinant offspring. PH additionally returns the set of sequences producing the loop, known as the homology generators. The number of $H_1$ bars is the first Betti number and can be used to measure the frequency of recombination events.

**Applying Topological Analysis to Influenza**

Our first application of TDA is to influenza A virus, a common human pathogen. It is well known that frequent reassortments punctuate the evolution of influenza, a segmented single-stranded RNA orthomyxovirus. Reassortant strains can lead to antigenic novelty when internal segments adapted to humans pair up with novel external segments. Such events led to the human influenza pandemics of 1957 and 1968.

To analyze recent trends in influenza evolution, we applied persistent homology to data for more than 3,000 avian influenza genomes from the NIH Influenza Sequence Database. Examining each genome segment separately, we recovered only zero-dimensional homology, consistent with the absence of intra-segmental recombination (Fig. 3A). In other words, segment-by-segment the evolution of influenza is tree-like and amenable to phylogenetic tree representation.

However, a similar analysis of the concatenated full genome reveals a complex topology, with a large number of loops in one and two
dimensions, reflecting pervasive reassortment (Fig. 3B). Statistical inference on the loops corresponding to reassortments identified segments that tend to co-segregate with each other during reassortment (Fig. 3C). In particular, polymerases are more likely to co-segregate, in contrast to envelope and capsid proteins that show independent reassortment patterns. Co-segregation of polymerases suggests that effective protein–protein interaction between the polymerase complex and the NP protein constrain reassortment. Finally, we identified two distinct scales of one-dimensional loops, corresponding to reassortment within a single subtype at smaller scales and reassortment across subtypes at larger scales (Fig. 3D).

**Applying Topological Analysis to Pathogenic Bacteria**

Horizontal gene transfers among bacteria are a well-known source of genetic diversity. As with viral reassortment, topology can be used to characterize bacterial lateral gene transfers. We constructed phyletic profiles for a range of pathogenic bacteria using FigFam annotations from the Pathosystems Resource Institute Center (PATRIC) database, which consists of more than 100,000 protein families curated from more than 950,000 unique proteins.

Protein family annotations groups proteins into sets of isofunctional homologs, that is, clusters of proteins with both similar sequence compositions and functions. A particular strain is represented as a binary vector, indicating the presence or absence of a given protein family. Correlations between strains can reveal genome-wide patterns of genetic exchange.

For each strain, we can compute a transformation into FigFam space and construct a strain-strain correlation matrix on which we can then compute persistent homology. For example, we can show relative recombination rates between different species, computing them by using persistent homology (Fig. 4A). From this analysis, we see that different species have a diverse topological structure in this space over a wide variety of recombination scales. For instance, the large scales of genomic exchange in *Haemophilus influenzae* suggest that it acquires novel genetic material from distantly related strains.

In Fig. 4B we show a topological network representing *Staphylococcus aureus*, a gram-positive bacterium commonly found in the nostrils and upper respiratory tract. The network was generated using Ayasdi Cure, a program for topological analysis and visualization of biological data. In the network, color corresponds to enrichment for *mecA*, an antibiotic resistance gene.

---

**FIGURE 4**

(A) Relative recombination rates in pathogenic bacteria. (B) Topological network representation of *S. aureus* genomes. Color corresponds to enrichment in *mecA*, an antibiotic resistance gene.
resistance gene. The network depicts two large clusters connected by a narrow bridge of strains. Strikingly, we observe that while mecA enrichment is not as strong in the left cluster, there is a distinct path of enrichment emanating along the connecting bridge between the two clusters and into the less-enriched cluster. This suggests the hypothesis that antibiotic resistance has spread from the right cluster into the left cluster via strains intermediate to the two, and will likely continue to be selected for in the left cluster.

Conclusions

Horizontal genetic exchange is pervasive. Here we describe a novel framework for characterizing its presence in genomic data using ideas from topology and show how it can be used to analyze both viral and bacterial datasets. The method allows us to describe in quantitative terms both the scale and frequency of horizontal gene transfers. As the quantity and complexity of genomic data increase, we hope new approaches for analyzing and depicting those data while drawing from rich mathematical frameworks will continue to be developed.

Kevin Emmett is graduate student at in the Departments of Physics and Systems Biology at Columbia University, Joseph M. Chan is a former M.D./Ph.D. student at Columbia University and is currently a resident at Weill Cornell Medical Center and fellow at Memorial Sloan Kettering, Daniel Scholes Rosenbloom is a Postdoctoral Research Scientist in the Department of Biomedical Informatics, Columbia University, New York, N.Y., and Raul Rabadan is an associate professor in the Departments of Systems Biology and Biomedical Informatics at Columbia University, New York, N.Y.

Acknowledgments

This work was performed thanks to NIH funding (U54 CA193313-01, 1R01GM117591, 1R21CA192854-01).

Suggested Reading