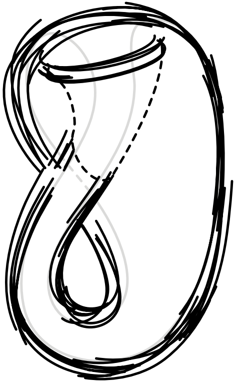
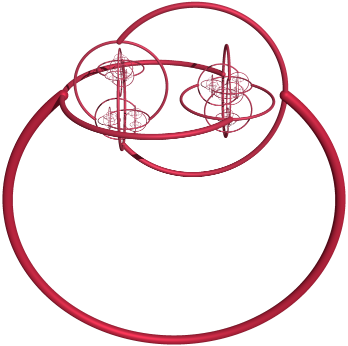
The Many Faces of Manifolds

Bastian Rieck

What this talk is *not* about

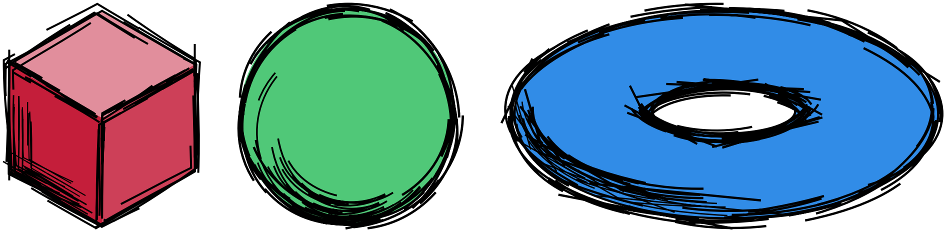
What this talk *is* about


Klein bottle


Alexander horned sphere

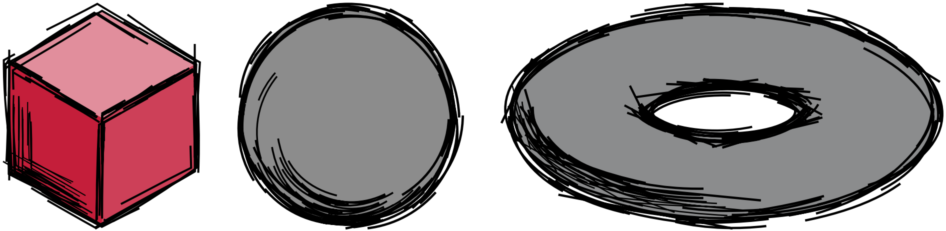
What is topology?

Studying the abstract shape of objects



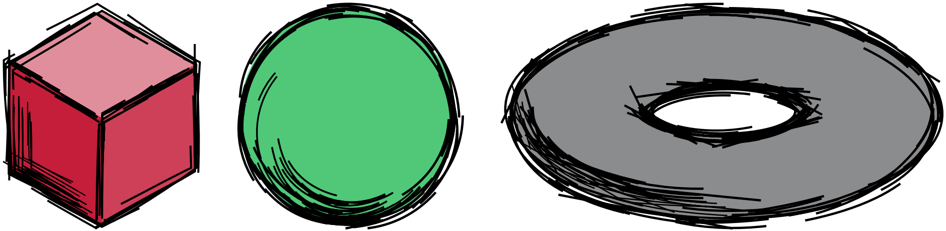
What is topology?

Studying the abstract shape of objects



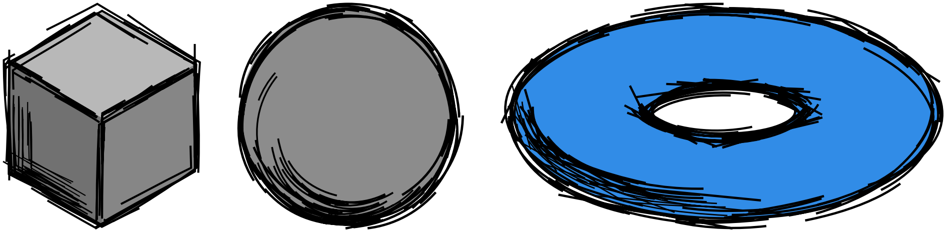
What is topology?

Studying the abstract shape of objects



What is topology?

Studying the abstract shape of objects



What is a manifold?

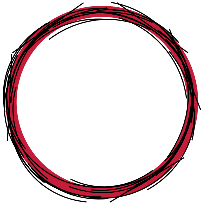
Informal definition

An object (or a space) that *locally* looks like some d-dimensional Euclidean space, i.e. we have d independent coordinates to describe our position.

What is a manifold?

Informal definition

An object (or a space) that *locally* looks like some d-dimensional Euclidean space, i.e. we have d independent coordinates to describe our position.

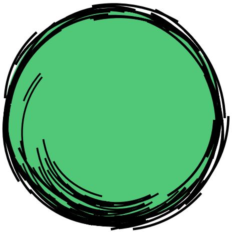


$$d = 1$$

What is a manifold?

Informal definition

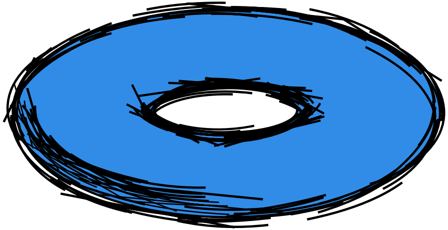
An object (or a space) that *locally* looks like some d-dimensional Euclidean space, i.e. we have d independent coordinates to describe our position.



$d = 2$

What is a manifold?
Informal definition

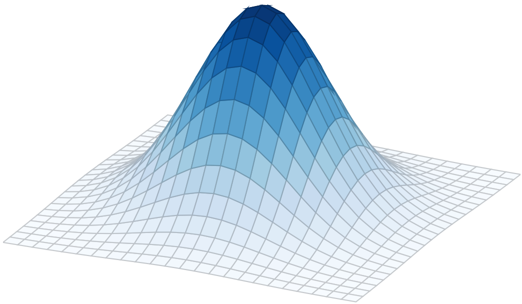
An object (or a space) that *locally* looks like some d-dimensional Euclidean space, i.e. we have d independent coordinates to describe our position.

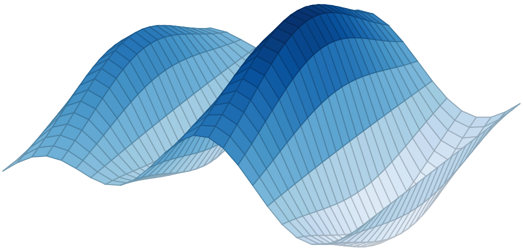


$$d = 2$$

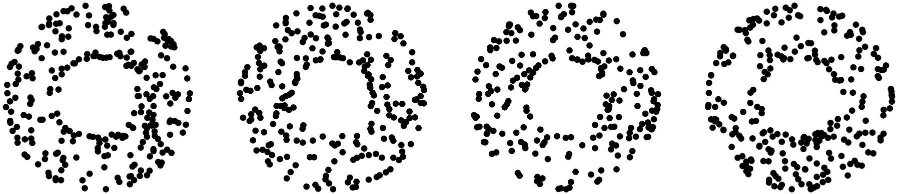
More manifolds

Let us *ignore* boundaries for now…





No manifolds, but *samples* of manifolds



Data analysis

Manifold hypothesis

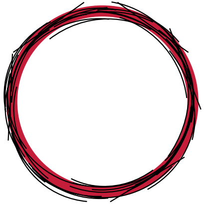
Many people hope that real-world data sets can often be adequately described by one (or more) manifolds.

Questions

1. How to *distinguish* between manifolds?
2. How to *classify* manifolds?

A simple taxonomy

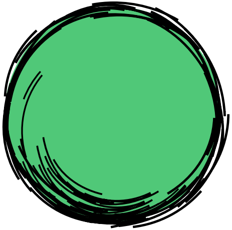
Betti numbers: counting d-dimensional holes



$\beta_0 = 1, \beta_1 = 1$

A simple taxonomy

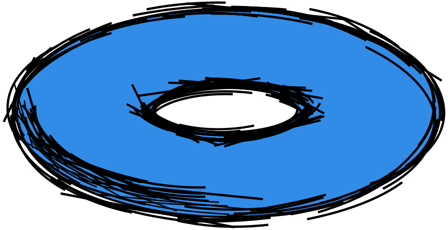
Betti numbers: counting d-dimensional holes



$$\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$$

A simple taxonomy

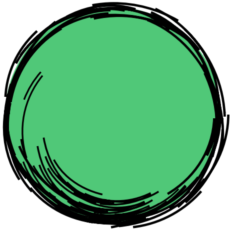
Betti numbers: counting d-dimensional holes



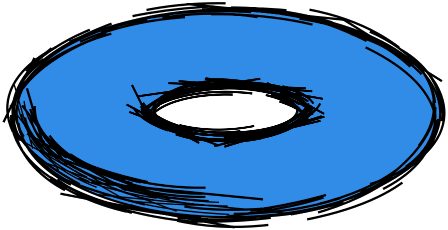
$$\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$$

A simple taxonomy

Betti numbers: counting d-dimensional holes



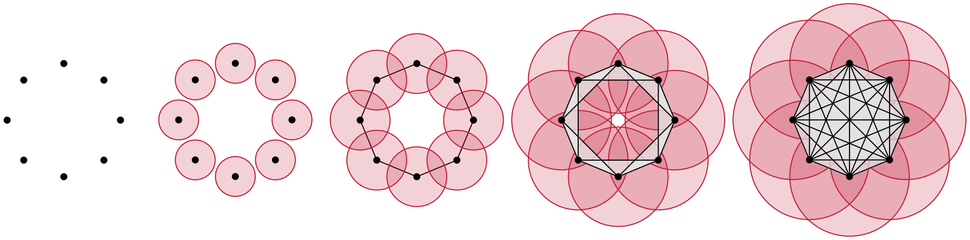
$\beta_0 = 1, \beta_1 = 0, \beta_2 = 1$



$\beta_0 = 1, \beta_1 = 2, \beta_2 = 1$

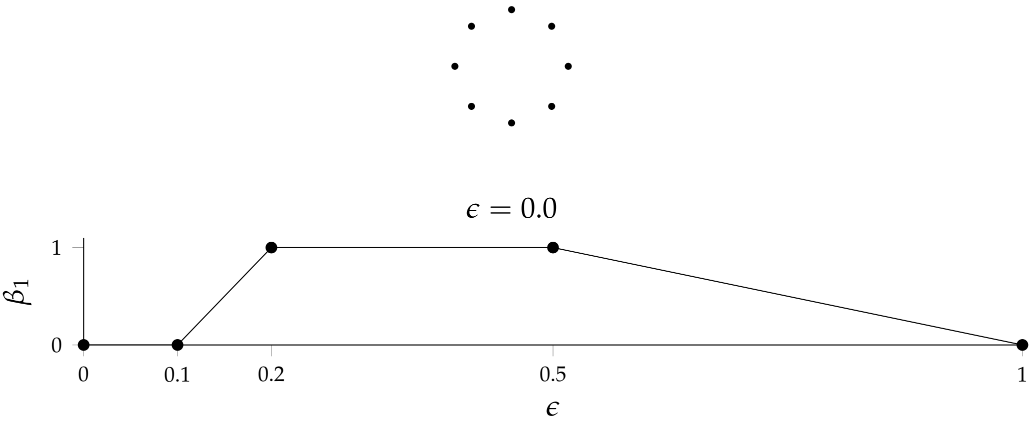
Back to reality

How to *approximate* the shape of data?

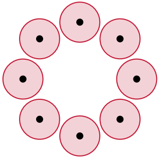


$\epsilon = 0.0$ $\epsilon = 0.1$ $\epsilon = 0.2$ $\epsilon = 0.5$ $\epsilon = 1.0$

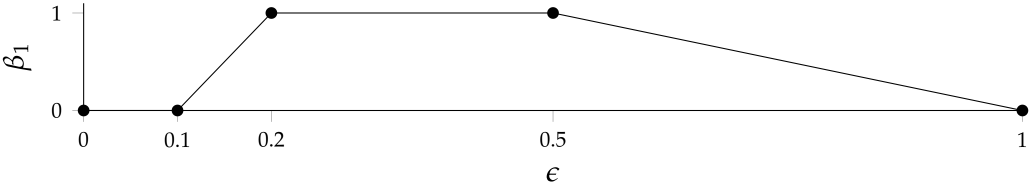
Instabilities



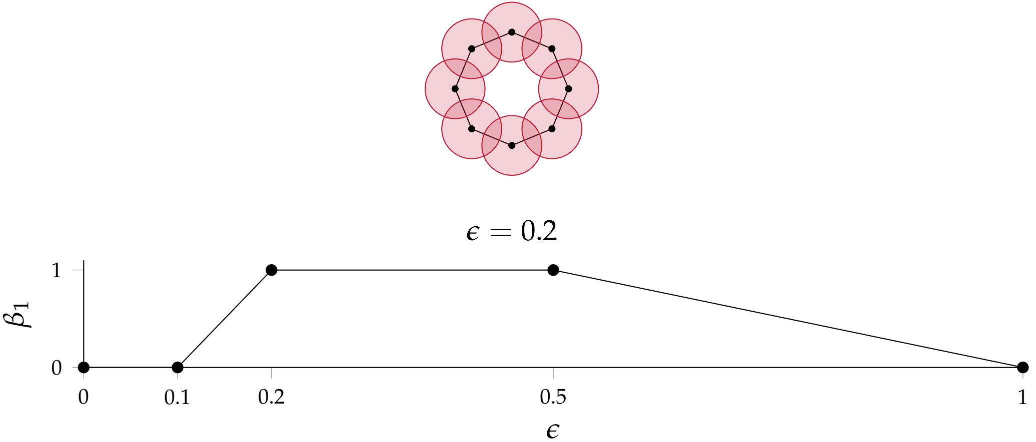
Instabilities



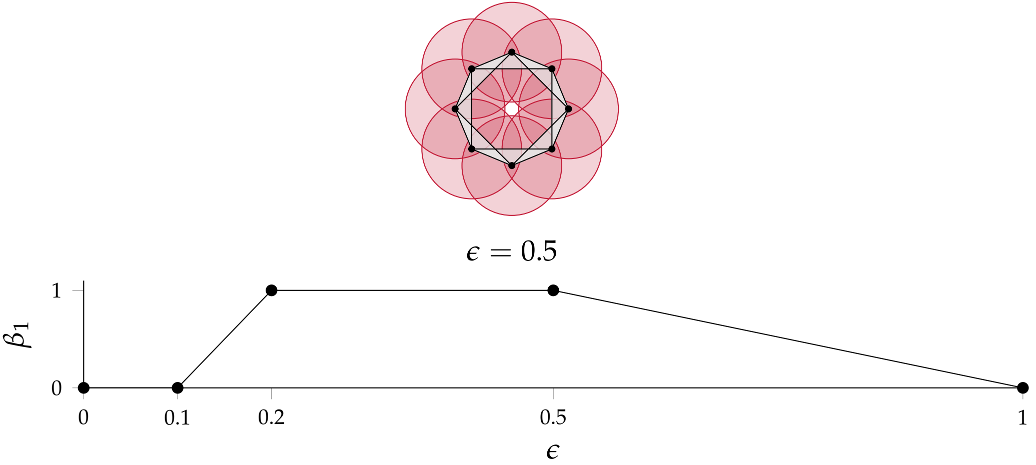
$\epsilon = 0.1$



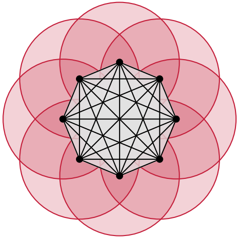
Instabilities



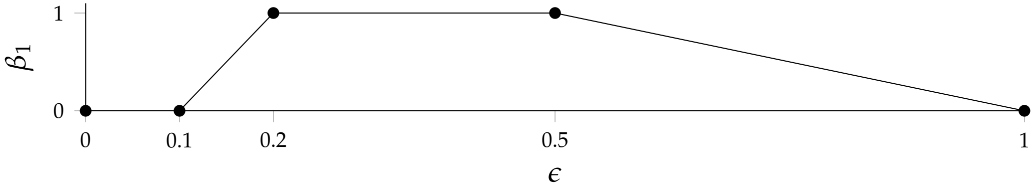
Instabilities



Instabilities



$\epsilon = 1.0$



Persistent homology

Homology

Homology refers to similarity that arises from a shared ancestry between a pair of structures or genes in different taxa.

Persistent homology

Homology

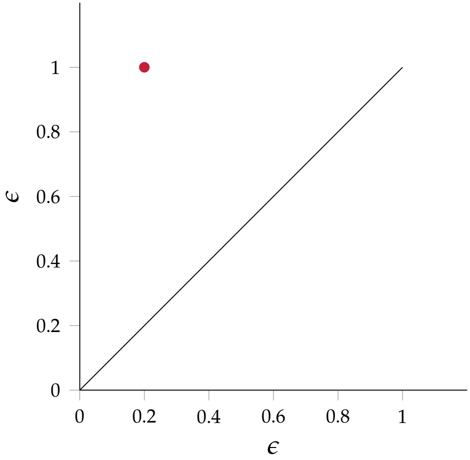
Homology refers to ~~*similarity that arises from a shared ancestry between a pair of structures or genes in different taxa*~~ a generic way of associating a sequence of algebraic objects, such as (Abelian) groups to other objects, such as topological spaces.

Persistence diagrams

Information about the Betti numbers is stored in a multi-scale topological descriptor, the *persistence diagram*. Points with a large distance from the diagonal are *persistent*.

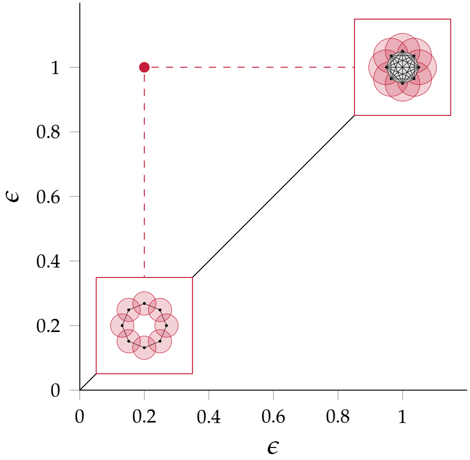
Persistent homology, continued

Example persistence diagram



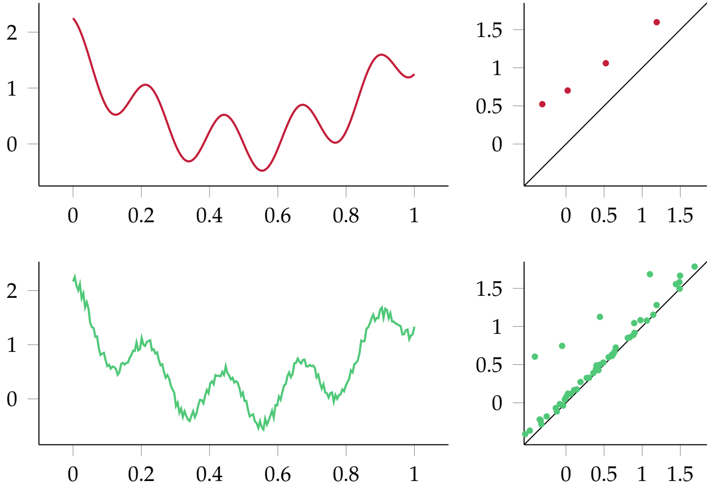
Persistent homology, continued

Example persistence diagram



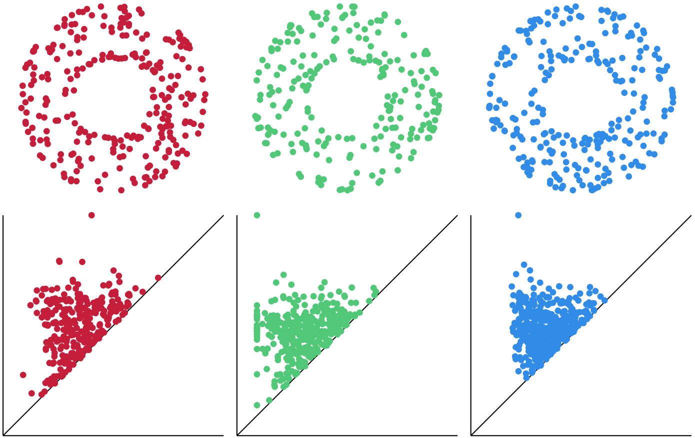
A real-world example

Analysing a function f over $\mathbb{R}$



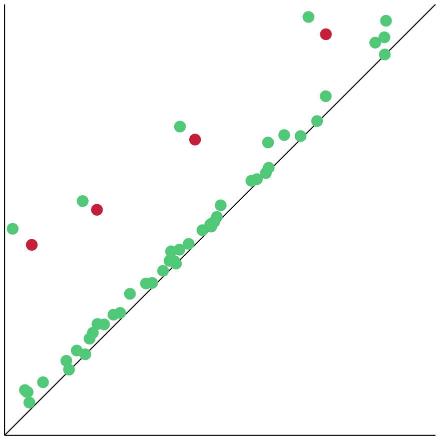
Another real-world example

Analysing the distance function of point clouds in $\mathbb{R}^3$



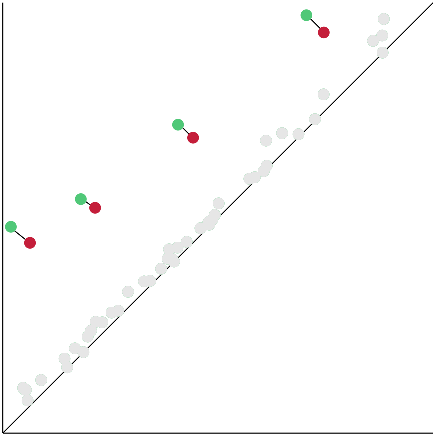
Calculating distances

Optimal transport



Calculating distances

Optimal transport



Applications

Topology of viral evolution

PNAS

Topology of viral evolution

Joseph Minhow Chan[a,b], Gunnar Carlsson[c], and Raul Rabadan[a,b,d,1]

[a]Center for Computational Biology and Bioinformatics and Departments of [b]Biomedical Informatics and [d]Systems Biology, Columbia University College of Physicians and Surgeons, New York, NY 10032; and [c]Department of Mathematics, Stanford University, Stanford, CA 94305

Edited* by Arnold J. Levine, Institute for Advanced Study, Princeton, NJ, and approved October 11, 2013 (received for review July 18, 2013)

The tree structure is currently the accepted paradigm to represent evolutionary relationships between organisms, species or other taxa. However, horizontal, or reticulate, genomic exchanges are pervasive in nature and confound characterization of phylogenetic trees. Drawing from algebraic topology, we present a unique evolutionary framework that comprehensively captures both clonal and reticulate evolution. We show that whereas clonal evolution can be summarized as a tree, reticulate evolution exhibits nontrivial topology of dimension greater than zero. Our method effectively characterizes clonal evolution, reassortment, and recombination in RNA viruses. Beyond detecting reticulate evolution, we succinctly recapitulate the history of complex genetic exchanges involving more than two parental strains, such as the triple reassortment of H7N9 avian influenza and the formation of circulating HIV-1 recombinants. In addition, we identify recurrent, large-scale patterns of reticulate evolution, including frequent PB2-PB1-PA-NP cosegregation during avian influenza reassortment. Finally, we bound the rate of reticulate events (i.e., 20 reassortments per year in avian influenza). Our method provides an evolutionary perspective that not only captures reticulate events precluding phylogeny, but also indicates the evolutionary scales where phylogenetic inference could be accurate.

persistent homology | gene flow | topological data analysis

In *On the Origin of the Species* in 1859, Darwin first proposed the phylogenetic tree as a structure to describe the evolution of phenotypic attributes. Since then, the advancement of modern sequencing has spurred development of a number of phylogenetic inference methods (1, 2). The tree structure effectively models vertical or clonal evolution, mediated by random mutations over multiple generations (Fig. 1*A*). Phylogenetic trees, however, cannot capture horizontal, or reticulate, events, which occur when distinct clades merge together to form a new hybrid lineage (Fig. 1*B* and *SI Appendix*, Fig. S1). In nature, horizontal evolution can occur through species hybridization in eukaryotes, lateral gene transfer in bacteria, recombination and reassortment in viruses, viral integration in eukaryotes, and fusion of genomes of symbiotic species (e.g., mitochondria). These horizontal genetic exchanges create incompatibilities that mischaracterize the species tree (3). Doolittle (4) argued that molecular phylogeneticists have failed to identify the "true tree of life" because the history of living organisms cannot be understood as a tree. One may wonder what other mathematical structures beyond phylogeny can capture the richness of evolutionary processes.

Current techniques that detect reticulate events can be divided into phylogenetic and nonphylogenetic methodologies. Phylogenetic methods detect incongruence in the tree structure of different segments (5–8). Nonphylogenetic methods probe for homoplasies (shared character traits independently arising in different lineages by convergent or parallel evolution) or similar inconsistencies in sequence alignment (9–13). Although many of these methods are designed for sensitive detection of viral recombination and bacterial lateral gene transfer, they do not provide comprehensive representation of the evolutionary process.

Perhaps phylogenetic networks exemplify the largest departure from trees, allowing multiple paths between any two leaves. These methods visualize incompatibilities of sequence patterns or tree topologies as reticulation cycles in a network (14–16). Only the subfield of evolutionary networks is amenable to reticulate detection. However, major stumbling blocks abound for such methods. Although phylogenetic network structure is not necessarily unique, all current implementations produce only one network that may represent a suboptimal solution; results may depend on factors as arbitrary as the ordering of samples in the data matrix (16, 17). Moreover, many methods have impractical running times for even small datasets owing to the nondeterministic polynomial-time hard (NP-hard) problem of determining whether a tree exists in an evolutionary network (18). To address these obstacles, ad hoc methods simplify the search space of network structures: k-level, galled, tree-child, and tree-sibling networks. Although some of these methods cease to be NP-hard (19), all prioritize computational tractability over biological modeling (20). For example, galled tree networks minimize the number of inferred recombinations by ensuring that reticulation cycles share no nodes (21). This heuristic is appropriate only for low recombination rates and is not universally applicable.

Here, we propose a comprehensive and fast method of extracting large-scale patterns from genomic data that captures both vertical and horizontal evolutionary events at the same time. The structure we propose is not a tree or a network, but a set of higher-dimensional objects with well-defined topological properties. Using the branch of algebraic topology called persistent homology (throughout this paper, we refer to mathematical homology, not the notion of genetic or structural similarity), we extract robust global features from these high-dimensional complexes. Unlike phylogenetic methods that produce a single, possibly suboptimal, tree or network, persistent homology considers all topologies and their relationships across the entire parameter space of genetic distance. Through analysis of viral and simulated genomic datasets, we show how persistent homology captures fundamental evolutionary aspects not directly inferred

Significance

Evolution is mediated not only by random mutations over a number of generations (vertical evolution), but also through the mixture of genomic material between individuals of different lineages (horizontal evolution). The standard evolutionary representation, the phylogenetic tree, faithfully represents the former but not the latter scenario. Although many elaborations have been developed to address this issue, there is still no agreed-upon method of incorporating both vertical and horizontal evolution. Here, we present an alternative strategy based on algebraic topology to study evolution. This method extends beyond the limits of a tree to capture directly even complex horizontal exchanges between multiple parental strains, as well as uncover broader reticulate patterns, including the segregation of segments during reassortment.

Author contributions: J.M.C., and R.R. designed research; J.M.C. performed research; G.C. and R.R. contributed analytic tools; J.M.C. and R.R. analyzed data; and J.M.C., G.C., and R.R. wrote the paper.

The authors declare no conflict of interest.

*This Direct Submission article had a prearranged editor.

Freely available online through the PNAS open access option.

[1]To whom correspondence should be addressed. E-mail: rabadan@dbmi.columbia.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1313480110/-/DCSupplemental.

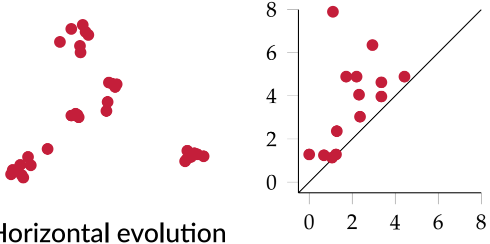
18566–18571 | PNAS | November 12, 2013 | vol. 110 | no. 46

www.pnas.org/cgi/doi/10.1073/pnas.1313480110

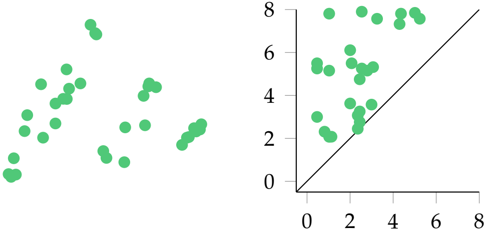
- Assess evolutionary behaviour of viruses
- Particularly interested in *reticulate* evolution
- Perform analysis based on genomic sequences
- Horizontal evolution results in non-trivial 1D topology

Comparing two sets of genomic sequences

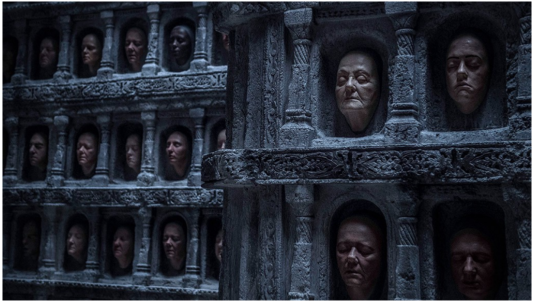
Vertical evolution



Horizontal evolution



Summary



1. Manifolds come in many forms, dimensions, and sizes.
2. Not everything that is high-dimensional *is* a manifold.
3. Assuming that samples come from some manifold $\mathcal{M}$ can be helpful.
4. Topological methods can be useful for describing and understanding data.