Topological Data Analysis

Geelon So (ags2191)

July 19, 2018

Manifolds

"*Manifolds* are spaces that locally look like Euclidean space."

Manifolds

"*Manifolds* are spaces that locally look like Euclidean space."

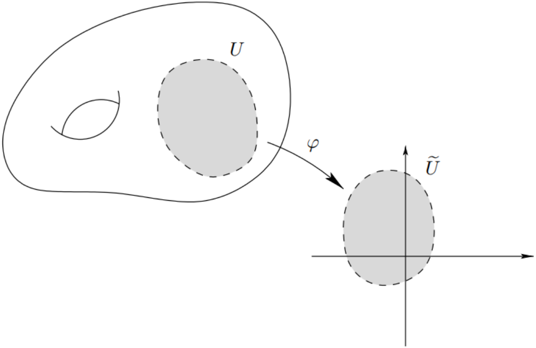


Figure 1: 'Local' region U identified with a piece of Euclidean space $\mathbb{R}^n$. [L2003]

Manifolds

But what does *locally* mean? Why care about *locality*?

Metric spaces

Metric spaces have a notion of distance:

$$d : X \times X \to \mathbb{R}.$$

Topological Space

A *topological space* is one on which similar points behave similarly.

Topological Space

A *topological space* is one on which similar points behave similarly.

- comes with a notion of *similarity* (a ‘topology’)
- *continuous functions* are maps that ‘respect’ similarity

Topology

Definition

A **topology** $\mathcal{T}$ *of a space* X *is a collection of subsets of* X *(called* **open sets***) such that:*

(i) $\varnothing, X \in \mathcal{T}$

(ii) $\mathcal{T}$ *is closed under finite intersection*

(iii) $\mathcal{T}$ *is closed under arbitrary union*

Learning Setup: Interlude

- X is the **space** our data comes from
- f a **computation/measurement** on X
    - we think of f as a *partial function* with some domain $A \subset X$
    - if $x \in A$, then $f(x)$ returns $\top$ in finite time
    - otherwise, $f(x)$ does not halt
- $f_1, f_2, \ldots$ a collection of *'primitive measurements'* on X
    - these are all the physical measurements we can make on X in finite time

Learning Setup: Interlude

Definition

A function f is **computable** *if it is either:*

- *a primitive measurement*
- *the conjunction of a finite number of computable functions*
- *the disjunction of a countable number of computable functions*

Topological Invariants

- cardinality
- number of connected components
- compactness
- metrizability
- separation
- homology group
- etc.

Topological Invariants

If we know how points in a space are related, we can say something about its **shape**.

- How are points connected?

The Classical Example

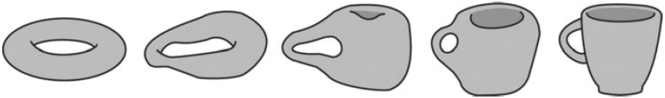


Figure 2: The donut is topologically equivalent to the coffee mug. [link]

Topological Data Analysis: the hope

Topology studies the ‘shape’ of objects.

- What is the shape of data?

Topological Data Analysis: the hope

Topology studies the ‘shape’ of objects.

- What is the shape of data?

Topological invariants are indifferent to ‘nice deformations’.

- Any robust statistics about our data?

Topological Data Analysis: the hope

Topology studies the ‘shape’ of objects.

- What is the shape of data?

Topological invariants are indifferent to ‘nice deformations’.

- Any robust statistics about our data?

The topology of a space determine which functions are possible.

- Does knowing the topology of data improve learnability?

Example: space determines the possible functions

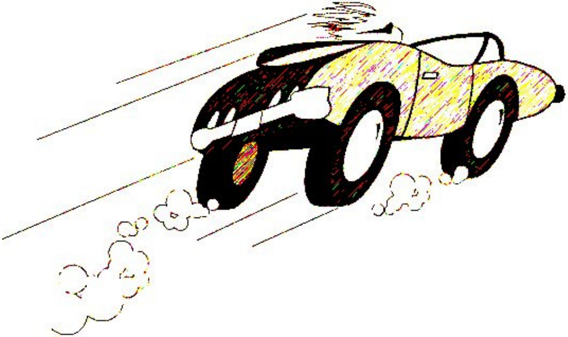


Figure 3: If you travel along a straight road from A to B, I could deduce how fast you were traveling at some point just by observing how long it took. (No wormholes). [link]

Intuition into Topology

1. Defining simplicial complexes as model spaces
2. Generating simplicial complexes from data
3. Understanding the shape of data (simplicial homology)
4. Summarizing data (persistence homology)

How are points connected?

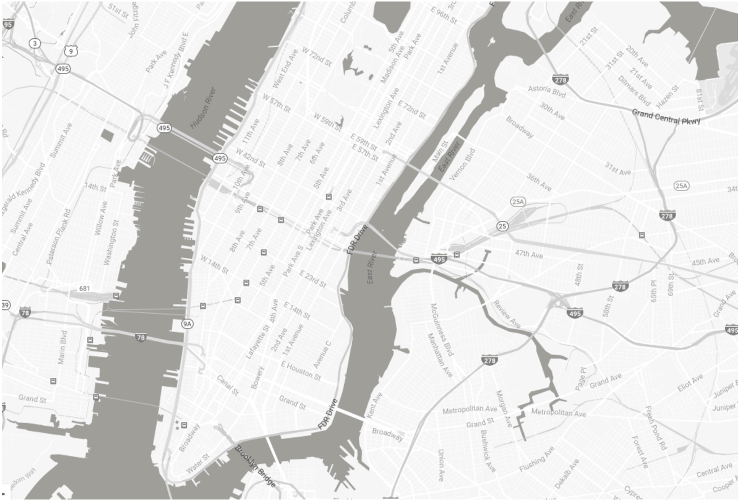


Figure 4: To model how NYC is connected for cars, use a graph. [link]

1-simplex is an edge

Definition

A **1-simplex** $[v_0, v_1]$ *is a graph* $G = (S_0, S_1)$ *where*

$$S_0 = \{\{v_0\}, \{v_1\}\} \qquad S_1 = \{\{v_0, v_1\}\},$$

where S_0 *are the vertices and* S_1 *are the edges.*

Model 1-dimensional connection with 1-simplices

A **1-dimensional simplicial complex** (a union of 1-simplexes) can represent the 1D ‘topology’ for a car.

Model 1-dimensional connection with 1-simplices

A **1-dimensional simplicial complex** (a union of 1-simplexes) can represent the 1D ‘topology’ for a car.

- i.e. a graph

Model n-dimensional connections with n-simplexes

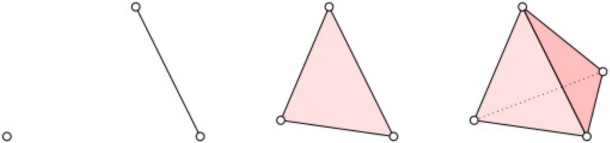


Figure 5: The 0-, 1-, 2-, and 3-simplex are ‘high-dimensional edges’. [link]

Approximation of Topological Spaces

Definition

*An n-**simplex** $[v_0, \ldots, v_n]$ is a hypergraph $K = (S_0, S_1, \ldots, S_n)$ where*

$$\begin{aligned} S_0 &= \big\{\{v_0\}, \ldots, \{v_n\}\big\}, \\ S_1 &= \big\{\{v_i, v_j\} : 0 \leq i < j \leq n\big\} \\ &\vdots \\ S_n &= \big\{\{v_0, \ldots, v_n\}\big\}. \end{aligned}$$

*An element of S_k is a k-**dimensional face** of K.*

Approximation of Topological Spaces

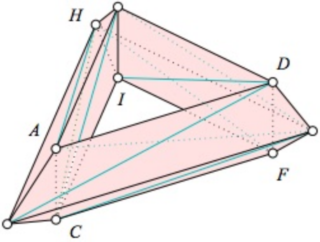


Figure 6: Approximation of the torus. [link]

Simplicial Complex

Definition

A **simplicial complex** K *is a set of simplexes such that:*[1]

- *every face of a simplex of* K *is also in* K
- *the intersection of any two simplexes* $\sigma_1, \sigma_2 \in K$ *is a face of both* σ_1 *and* σ_2

[1] Definition statement from Wikipedia.

Simplicial Complex

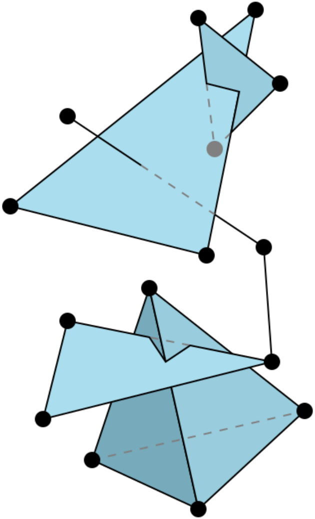


Figure 7: Definition of simplicial complex so that this is excluded. [link]

Notation

In the following, we will tend to use:

- X is an underlying space where data is generated from
  - often a metric space (X, d)
- $\mathbb{X} = \{x_1, \ldots, x_n\}$ is the point data

Graph from Data

Given data $\mathbb{X}$ and some measure of similarity/distance, how to generate graph?

- k-nearest neighbor
- ϵ-graph
- etc.

Simplicial Complex from Data: Vietoris-Rips complex

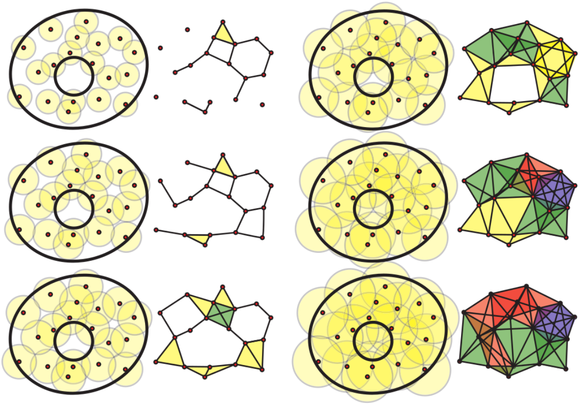


Figure 8: Vietoris-Rips complex for data sampled from an annulus. [G2008]

Vietoris-Rips complex

Definition

Let $\mathbb{X}$ be a set of points from a metric space, and $\epsilon \geq 0$. The **Vietoris-Rips complex** $\mathrm{Rips}_\epsilon(\mathbb{X})$ *is the set of simplices $[x_0, \ldots, x_k]$ such that $d(x_i, x_j) \leq \epsilon$ for all (i, j).*[2]

[2] Definition statement from [C2017].

Vietoris-Rips complex

Definition

Let $\mathbb{X}$ be a set of points from a metric space, and $\epsilon \geq 0$. The **Vietoris-Rips complex** $\mathrm{Rips}_\epsilon(\mathbb{X})$ *is the set of simplices* $[x_0, \ldots, x_k]$ *such that* $d(x_i, x_j) \leq \epsilon$ *for all* (i, j).[2]

- i.e. generate ϵ-graph from data, then 'fill in' any k-clique with the k-simplex.

[2] Definition statement from [C2017].

Simplicial Complex from Data: Čech complex

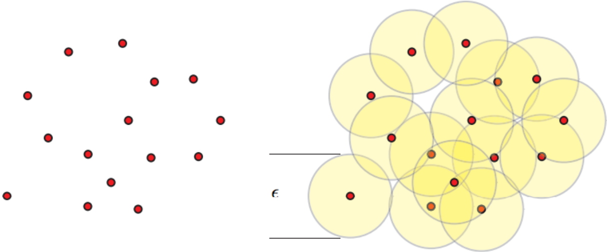


Figure 9: Generating a Čech complex for point data. [G2008]

Čech complex

Definition

Let $\mathbb{X}$ be as above. The **Čech complex** $\mathrm{Cech}_\epsilon(\mathbb{X})$ *is the set of simplices* $[x_0, \ldots, x_k]$ *such that the* $k+1$ *closed balls* $\overline{B(x_i, \epsilon)}$ *have a nonempty intersection.*[3]

[3] Definition statement from [C2017].

Čech complex

Definition

Let $\mathbb{X}$ *be as above. The* **Čech complex** $\mathrm{Cech}_\epsilon(\mathbb{X})$ *is the set of simplices* $[x_0, \ldots, x_k]$ *such that the* $k+1$ *closed balls* $\overline{B(x_i, \epsilon)}$ *have a nonempty intersection.*[3]

- i.e. draw ϵ-ball around points x_i, and k-way intersections become k-simplexes.

[3] Definition statement from [C2017].

Čech complex vs. Rips

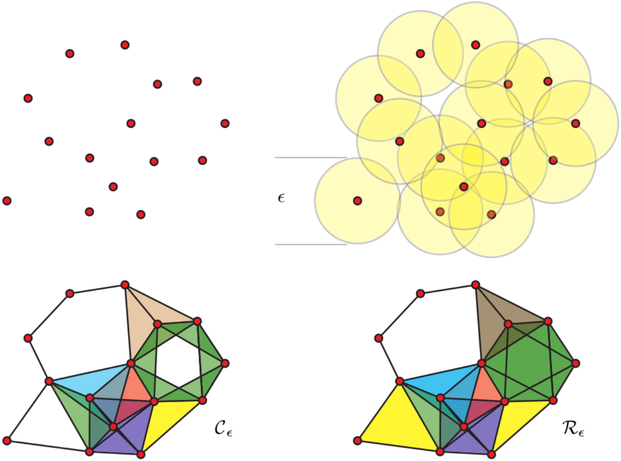


Figure 10: The Čech complex is not the same as the Vietoris-Rips complex. [G2008]

Rips vs. Čech complex

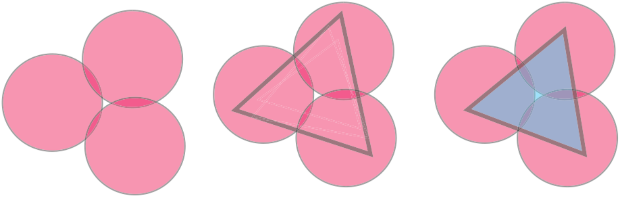


Figure 11: Zoom in to example where Čech (middle) and Rips (right) complex differ. [C20097]

Rips vs. Čech complex

$$\mathrm{Rips}_{\epsilon}(\mathbb{X}) \subset \mathrm{Cech}_{\epsilon}(\mathbb{X}) \subset \mathrm{Rips}_{2\epsilon}(\mathbb{X}).$$

$\mathbb{X}$ and X

What about the relationship between the complex generated on point data $\mathbb{X}$ and its underlying space X?

Topological Relationships

Definition

Let X and Y be topological spaces. They are **homeomorphic** *if there exists a continuous map $f : X \to Y$ with continuous inverse.*

Topological Relationships

Definition

Let X and Y be topological spaces. They are **homeomorphic** *if there exists a continuous map $f : X \to Y$ with continuous inverse.*

- f is just a ‘renaming’ of points
- so, one can choose to study either X or Y

Topological Relationships

Definition (Informal)

X and Y are **homotopy equivalent** *if they can be continuously deformed into each other.*

Topological Relationships

Definition (Informal)

X and Y are **homotopy equivalent** *if they can be continuously deformed into each other.*

- weaker than homeomorphism
  (i.e. X, Y homeomorphic $\implies$ homotopy equivalent)
- guarantees certain topological properties (shapes) are shared

Topological Property

Definition

A space X is **contractible** *if it is homotopy equivalent to a point within X.*

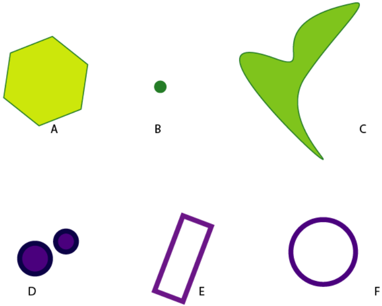


Figure 12: Which spaces are contractible? [link]

What properties are preserved by the complexes?

Under certain conditions, the Čech complex will be *homotopy equivalent* to the underlying space X where the data came from.

What properties are preserved by the complexes?

Under certain conditions, the Čech complex will be *homotopy equivalent* to the underlying space X where the data came from.

- i.e. we can learn something about the shape of X from $\mathbb{X}$

Good Cover

Definition

A **good cover** *of X is an open cover such that all finite intersections of open sets are either empty or contractible.*

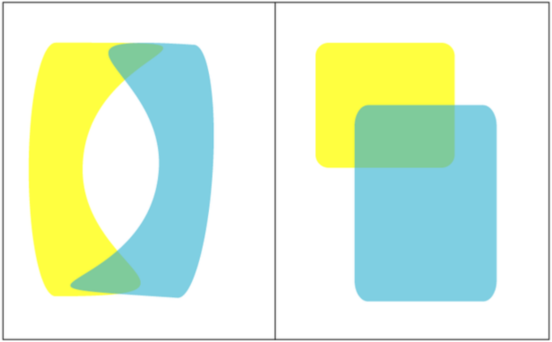


Figure 13: A bad (left) and good (right) cover. [link]

Nerve Theorem

Theorem (Nerve Theorem)

Let $\mathbb{X} := \{x_1, \ldots, x_n\} \subset X$. *If the open cover*

$$\{B(x_i, \epsilon) : x_i \in \mathbb{X}\}$$

is a good cover of X, *then* $\mathrm{Cech}_\epsilon(\mathbb{X})$ *and* X *are homotopy equivalent.*

Nerve Theorem (general)

Definition

Let $\mathcal{U}$ be a collection of open sets in X. Then, the **nerve** $C(\mathcal{U})$ *is the Čech complex generated from $\mathcal{U}$.*

- this contrasts to using the open cover $B(x_i, \epsilon)$

Nerve Theorem (general)

Theorem

Let $\mathcal{U}$ be a good cover of X. Then, X and the nerve $C(\mathcal{U})$ are homotopy equivalent.

Nerve Theorem (general)

Theorem

Let $\mathcal{U}$ be a good cover of X. Then, X and the nerve $C(\mathcal{U})$ are homotopy equivalent.

- this suggests an algorithm to study the topological properties of X—generate a good cover of X and compute the nerve

Mapper Algorithm Intuition

Let X be the space we want to understand.

Mapper Algorithm Intuition

Let X be the space we want to understand.

1. Let $f : X \to \mathbb{R}^d$ summarize the important aspects of X

Mapper Algorithm Intuition

Let X be the space we want to understand.

1. Let $f : X \to \mathbb{R}^d$ summarize the important aspects of X
2. Form open cover $\mathcal{V}$ of $\mathbb{R}^d$

Mapper Algorithm Intuition

Let X be the space we want to understand.

1. Let $f : X \to \mathbb{R}^d$ summarize the important aspects of X
2. Form open cover $\mathcal{V}$ of $\mathbb{R}^d$
3. Induces an open cover $f^{-1}(\mathcal{V})$ on X

Mapper Algorithm Intuition

Let X be the space we want to understand.

1. Let $f : X \to \mathbb{R}^d$ summarize the important aspects of X
2. Form open cover $\mathcal{V}$ of $\mathbb{R}^d$
3. Induces an open cover $f^{-1}(\mathcal{V})$ on X
4. Ensure $f^{-1}(\mathcal{V}$ is a good cover

Mapper Algorithm Intuition

Let X be the space we want to understand.

1. Let $f : X \to \mathbb{R}^d$ summarize the important aspects of X
2. Form open cover $\mathcal{V}$ of $\mathbb{R}^d$
3. Induces an open cover $f^{-1}(\mathcal{V})$ on X
4. Ensure $f^{-1}(\mathcal{V}$ is a good cover
5. Generate and study the nerve of open cover

Computing the Nerve

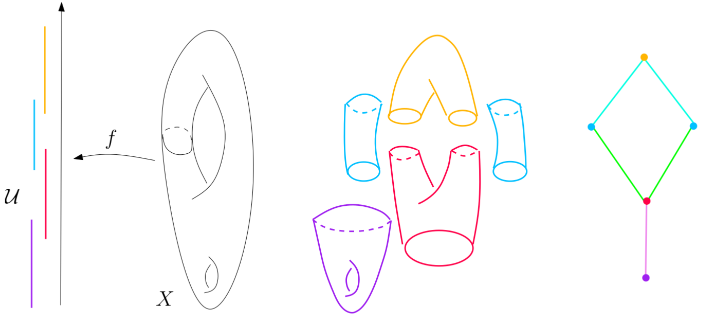


Figure 14: The refined pullback cover of the height function and its nerve.[4] [C2017]

[4] Caption statement from [C2017]

Pullback Cover

Definition

Let $f : X \to Y$ be continuous. If $\mathcal{V}$ is an open cover of Y, then:

$$f^{-1}(\mathcal{V}) := \{f^{-1}(V) : V \in \mathcal{V}\}$$

is the **pullback cover** *of X induced by $(f, \mathcal{V})$. The* **refined pullback cover** *is the collection of connected components of $f^{-1}(V)$ for $V \in \mathcal{V}$.*

Mapper Algorithm

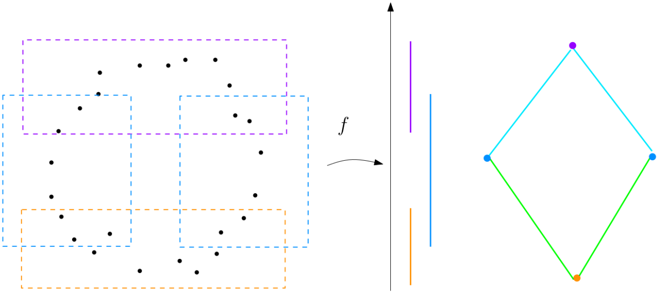


Figure 15: Example of mapper algorithm on point cloud. [C2017]

Mapper Algorithm

Algorithm 1 Mapper

```
input $\mathbb{X}$ a data set, a distance/similarity measure, $f : X \to \mathbb{R}^d$,
    and a cover $\mathcal{U}$ of $f(\mathbb{X})$
1: For each $U \in \mathcal{U}$, decompose $f^{-1}(U)$ into clusters $C_{U,1}, \dots, C_{U,k}$
2: Compute the nerve of $C_{U,i}$'s.
3: return simplicial complex, the nerve
```

- choice of f, the **filter** or **lens** function
- choice of **cover** $\mathcal{U}$ (resolution/gain)
- choice of **clustering** algorithm

Choice of Filter

The choice of f determines similarity with respect to what:

- PCA coordinates/nonlinear dimensionality reduction coordinates
- centrality function and eccentricity function:

$$\mathrm{central}(x) = \sum_{y \in \mathbb{X}} d(x,y) \qquad \mathrm{ecc}(x) = \max_{y \in \mathbb{X}} d(x,y)$$

- density estimates

Choice of Cover

Mapper may be very sensitive to choice of cover. A standard choice is evenly sized and spaced intervals

- resolution: the size of the intervals
- gain: the percent overlap of the intervals

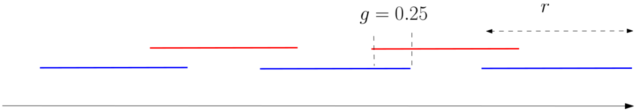


Figure 16: A cover of $\mathbb{R}$ with resolution r and gain 25%. [C2017]

Choice of Clustering

We need to cluster the preimages $f^{-1}(U)$, to generate a good cover.

- apply any clustering algorithm, or
- build neighborhood graph (k-NN or ϵ-graph)

Neural Codes

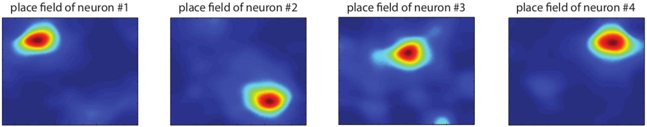


Figure 17: Place fields of for four place cells, recorded while a rat explored a 2-dimensional square box environment.[5] [Cu2017]

[5] Caption statement from [Cu2017]

Neural Codes

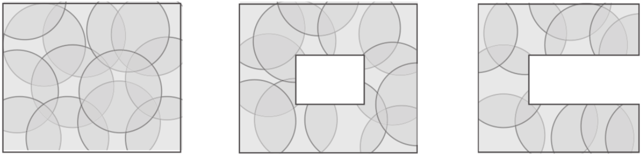
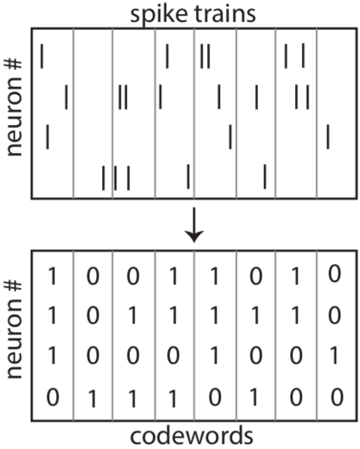


Figure 18: Three environments and place fields that cover the underlying space. [Cu2017]

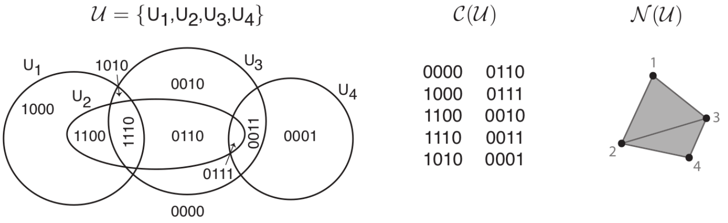
Neural Codes



1	0	0	1	1	0	1	0
1	0	1	1	1	1	1	0
1	0	0	0	1	0	0	1
0	1	1	1	0	1	0	0

Figure 19: Spike trains and neural codes. [Cu2017]

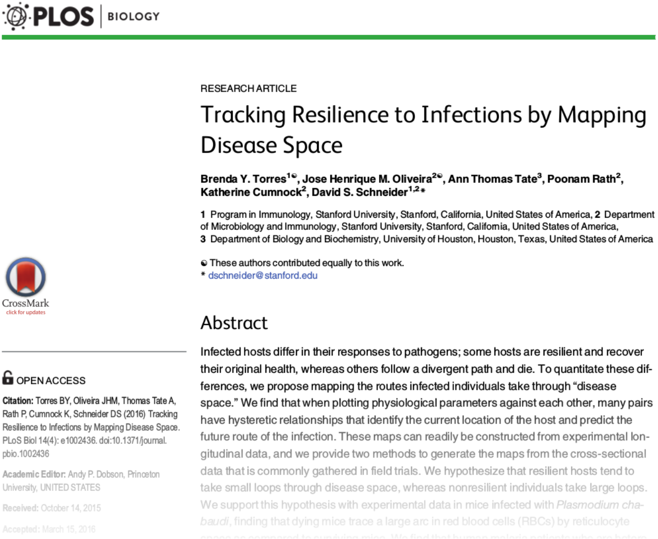
Neural Codes


$\mathcal{U} = \{U_1, U_2, U_3, U_4\}$
U_1
U_2
U_3
U_4
1010
0010
1000
1100
1110
0110
0011
0001
0111
0000
$\mathcal{C}(\mathcal{U})$
$\mathcal{N}(\mathcal{U})$
1
2
3
4

0000 0110
1000 0111
1100 0010
1110 0011
1010 0001

Figure 20: Codes and nerves of open covers. [Cu2017]

Application to Mapping Disease Space



PLOS | BIOLOGY

RESEARCH ARTICLE

Tracking Resilience to Infections by Mapping Disease Space

Brenda Y. Torres[1]◉, Jose Henrique M. Oliveira[2]◉, Ann Thomas Tate[3], Poonam Rath[2], Katherine Cumnock[2], David S. Schneider[1,2]*

1 Program in Immunology, Stanford University, Stanford, California, United States of America, **2** Department of Microbiology and Immunology, Stanford University, Stanford, California, United States of America, **3** Department of Biology and Biochemistry, University of Houston, Houston, Texas, United States of America

◉ These authors contributed equally to this work.
* dschneider@stanford.edu

CrossMark
click for updates

OPEN ACCESS

Citation: Torres BY, Oliveira JHM, Thomas Tate A, Rath P, Cumnock K, Schneider DS (2016) Tracking Resilience to Infections by Mapping Disease Space. PLoS Biol 14(4): e1002436. doi:10.1371/journal.pbio.1002436

Academic Editor: Andy P. Dobson, Princeton University, UNITED STATES

Received: October 14, 2015

Accepted: March 15, 2016

Abstract

Infected hosts differ in their responses to pathogens; some hosts are resilient and recover their original health, whereas others follow a divergent path and die. To quantitate these differences, we propose mapping the routes infected individuals take through "disease space." We find that when plotting physiological parameters against each other, many pairs have hysteretic relationships that identify the current location of the host and predict the future route of the infection. These maps can readily be constructed from experimental longitudinal data, and we provide two methods to generate the maps from the cross-sectional data that is commonly gathered in field trials. We hypothesize that resilient hosts tend to take small loops through disease space, whereas nonresilient individuals take large loops. We support this hypothesis with experimental data in mice infected with *Plasmodium chabaudi*, finding that dying mice trace a large arc in red blood cells (RBCs) by reticulocyte space as compared to surviving mice. We find that human malaria patients who are

Application to Mapping Disease Space

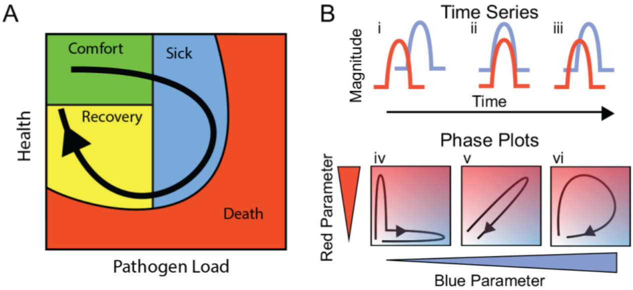


Figure 21: The path of an individual through a disease space. The choice of filter through partially out-of-phase heath statistics can generate a space with nontrivial fundamental group. [B2016]

Application to Mapping Disease Space

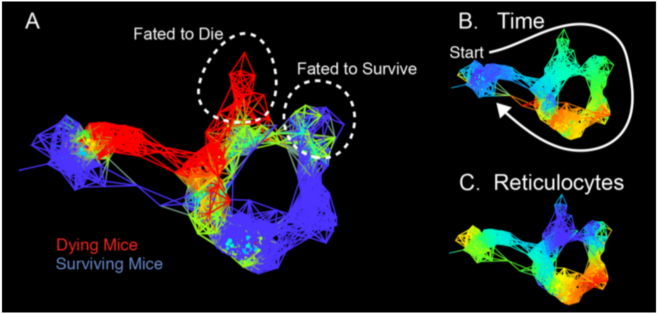


Figure 22: Disease space of malaria-infected mice. [B2016]

Geometric Reconstruction

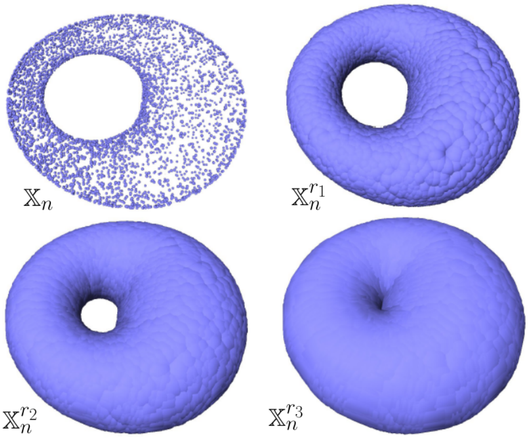


Figure 23: A point cloud sampled from a torus with varying offset values $r_1 < r_2 < r_3$. [C2017]

Geometric Reconstruction

If K is a compact set, let $d_K(x) := \inf_{y \in K} d(x, y)$. Then, the reconstruction $\mathbb{X}^r$ is:

$$\mathbb{X}^r = d_{\mathbb{X}}^{-1}([0, r]).$$

Geometric Reconstruction

Let $\phi = d_X$ and $\psi = d_{\mathbb{X}}$.

Theorem (Reconstruction Theorem)

Suppose that the α-reach of ϕ is at least R. If

$$\|\phi - \psi\|_\infty < \varepsilon(\alpha, R),$$

then there is some $r(\varepsilon, \alpha, R)$ such that $\psi^{-1}([0, r])$ is homotopy equivalent to X.

Geometric Reconstruction

Let $\phi = d_X$ and $\psi = d_{\mathbb{X}}$.

Theorem (Reconstruction Theorem)

Suppose that the α-reach of ϕ is at least R. If

$$\|\phi - \psi\|_\infty < \varepsilon(\alpha, R),$$

then there is some $r(\varepsilon, \alpha, R)$ such that $\psi^{-1}([0, r])$ is homotopy equivalent to X.

- the α-reach is just a regularity condition

Geometric Reconstruction

Let $\phi = d_X$ and $\psi = d_{\mathbb{X}}$.

Theorem (Reconstruction Theorem)

Suppose that the α-reach of ϕ is at least R. If

$$\|\phi - \psi\|_\infty < \varepsilon(\alpha, R),$$

then there is some $r(\varepsilon, \alpha, R)$ such that $\psi^{-1}([0, r])$ is homotopy equivalent to X.

- the α-reach is just a regularity condition
- it could be the case that given ϕ and ψ no reconstruction is possible

Geometric Reconstruction

The Reconstruction theorem tells us when the Čech complex is homotopy equivalent to the original space.

Geometric Reconstruction

The Reconstruction theorem tells us when the Čech complex is homotopy equivalent to the original space.

- But what can we deduce about X from $\mathrm{Cech}(X)$?

Homology Theory

Homology theory, **homotopy theory**, and more generally, **algebraic topology** studies the topological features of a space algebraically.

Example: Hairy Ball Theorem

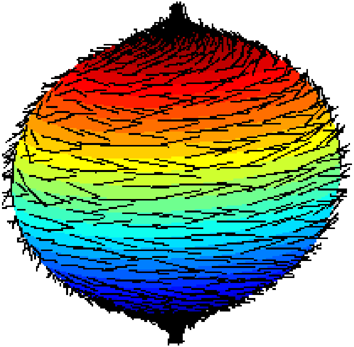


Figure 24: No nonzero smooth vector field exists on S^2. [link]

Example: Brouwer's Fixed Point Theorem

Theorem (Brouwer)

Let $X \subset \mathbb{R}^n$ *be a convex compact set. If* $f : X \to X$ *is continuous, then* f *has a fixed point.*

Example: Closed vs. Exact

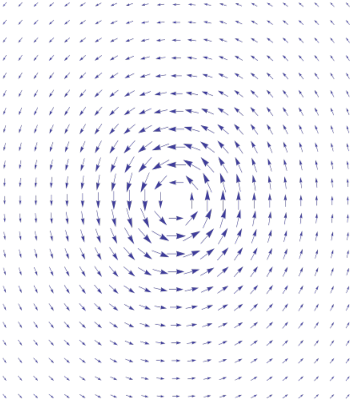


Figure 25: A vector field V on $\mathbb{R}^2$ where $\nabla \times V = 0$ everywhere except at 0. On $\mathbb{R}^2$, $\nabla \times V \equiv 0 \Leftrightarrow V = \nabla\phi$. What about $\mathbb{R}^2 \setminus \{0\}$? [link]

Example: Stokes' theorem

Theorem (Stokes)

Let Ω be an orientable manifold and ω be a differential form over its boundary $\partial\Omega$. Then:

$$\int_{\partial\Omega} \omega = \int_{\Omega} \mathrm{d}\omega.$$

k-chains

Definition

Let K be a simplicial complex and k a nonnegative number. Denote by $C_k(K)$ the **space of k-chains** *on K, the set of formal linear combinations of k-simplexes of K.*

[6] Definition statement from [C2017].

k-chains

Definition

Let K be a simplicial complex and k a nonnegative number. Denote by $C_k(K)$ the **space of k-chains** *on K, the set of formal linear combinations of k-simplexes of K.*

For example, if $\{\sigma_1, \dots, \sigma_p\} \in K$ are k-simplexes, a k-chain is:

$$c = \sum_{i=1}^{p} \alpha_i \sigma_i,$$

where the α_i's are scalars.[6]

[6] Definition statement from [C2017].

k-chains

The intuition is that $C_k(K)$ is the vector space built over all the k-dimensional subcomponents of K.

Boundary operator

$$\partial : C_k(K) \to C_{k-1}(K)$$

Boundary operator

$$\partial : C_k(K) \to C_{k-1}(K)$$

- maps a k-dimensional object into its $(k-1)$-dimensional boundary

Boundary operator

Definition

Let $\sigma = [v_0, \ldots, v_k]$ be a k-simplex. Then the $(k-1)$-chain:

$$\partial(\sigma) := \sum_{i=1}^{k} (-1)^{i+1} [v_0, \ldots, \hat{v}_i, \ldots, v_k]$$

is the boundary of σ. The **boundary operator** *is the linear extension to $C_k(K)$.*

Boundary operator

Example

An edge $e = [u, v]$ *is a 1-simplex, and its boundary is:*

$$\partial([u, v]) = v - u.$$

Boundary of Boundary

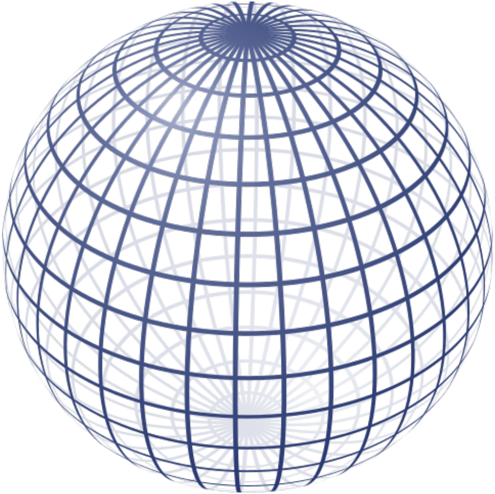


Figure 26: What is the boundary of a sphere? What is the sphere a boundary of? [link]

Boundary of Boundary

$$\partial_{k-1} \circ \partial_k \equiv 0.$$

Boundaries and Cycles

The following are (linear) subspaces of $C_k(K)$:

Definition

The image $\mathrm{im}(\partial_{k+1})$ *is the* **space of boundaries** $B_k(K)$ *of* K*.*

Definition

The kernel $\ker(\partial_k)$ *is the* **space of cycles** $Z_k(K)$ *of* K*.*

Boundaries and Cycles

Because the boundary of a boundary is zero,

$$B_k(K) \subset Z_k(K).$$

Boundaries and Cycles

Because the boundary of a boundary is zero,

$$B_k(K) \subset Z_k(K).$$

- the dimension of $Z_k(K)$ gives 'how many ways we can make subcomplexes of K that can be filled in?'
- the dimension of $B_k(K)$ gives 'how many of these subcomplexes are actually filled in?'

Betti Number

Definition (Informal)

The **Betti number** β_k *of* K *is:*

$$\beta_k := \dim Z_k(K) - \dim B_k(K).$$

Betti Number

Definition (Informal)

The **Betti number** β_k *of* K *is:*

$$\beta_k := \dim Z_k(K) - \dim B_k(K).$$

- i.e. how many k-dimensional holes are there?
    - $\beta_0 =$ number of connected components
    - $\beta_1 =$ number of ‘circular’ holes (punctures)
    - $\beta_2 =$ number of ‘voids’

Example

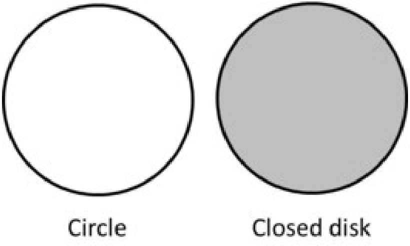


Figure 27: The circle itself is a cycle; however, because the whole space is 1-dimensional, it is not the image of a 2-dimensional subspace. It is not a boundary. So, $\beta_1(S^1) = 1$, as there is a '1D hole' in S^1. [link]

Example

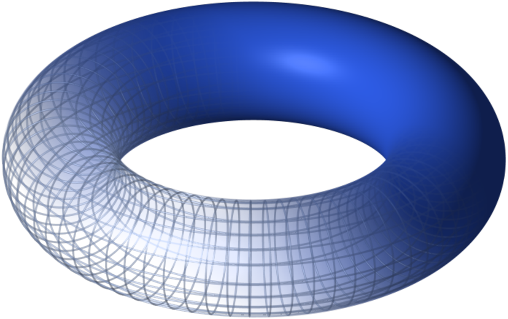


Figure 28: The torus has $\beta_0 = 1$, $\beta_1 = 2$, $\beta_2 = 1$.

Homology Group

Definition (Formal)

The kth **simplicial homology group** *of K is the quotient vector space*

$$H_k(K) = Z_k(K)/C_k(K).$$

The kth **Betti number** *β_k is $\beta_k(K) = \dim H_k(K)$.*

Homology Group

Definition (Formal)

The kth **simplicial homology group** *of K is the quotient vector space*

$$H_k(K) = Z_k(K)/C_k(K).$$

The kth **Betti number** *β_k is $\beta_k(K) = \dim H_k(K)$.*

- the homology group can be extended to general topological spaces.

Homotopy Invariance

Theorem

If $f : X \to Y$ is a homeomorphism or a homotopy equivalence, then

$$H_k(X) \cong H_k(Y)$$

are isomorphic.

Simplex and Space

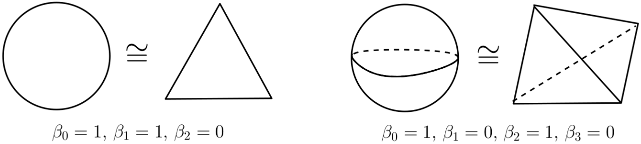

$\cong$
$\beta_0 = 1,\ \beta_1 = 1,\ \beta_2 = 0$
$\cong$
$\beta_0 = 1,\ \beta_1 = 0,\ \beta_2 = 1,\ \beta_3 = 0$

Figure 29: A space can be studied by considering the appropriate simplicial complex. [C2017]

High-Level View

- the homology group summarizes the shape of an object

High-Level View

- the homology group summarizes the shape of an object
- homotopy equivalent spaces have the same homology group ("they have the same shape")

High-Level View

- the homology group summarizes the shape of an object
- homotopy equivalent spaces have the same homology group ("they have the same shape")
- given certain conditions, the Čech complex of point data is homotopy equivalent to underlying space

Challenge

We don’t have access to underlying space, so cannot prove regularity conditions about it.

Challenge

We don't have access to underlying space, so cannot prove regularity conditions about it.

- So, study the Čech complex over a range of parameters.

Persistence Barcodes

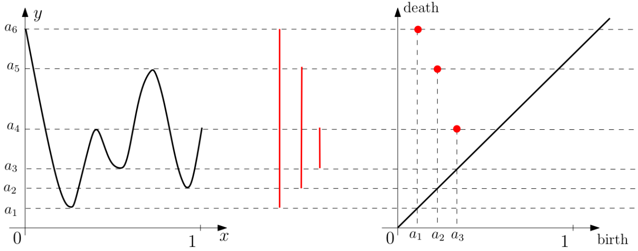


Figure 30: Persistence barcode for a function $f : [0,1] \to \mathbb{R}$. The diagram on the right plots β_0, the number of connected components. Imagine water filling up the graph on the left. [C2017]

Persistence Diagrams

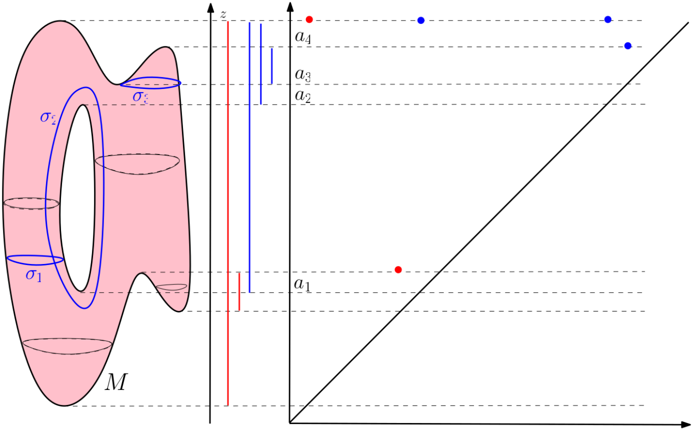


Figure 31: Persistence barcode, plotting β_1. [C2017]

Persistence Barcodes for Data

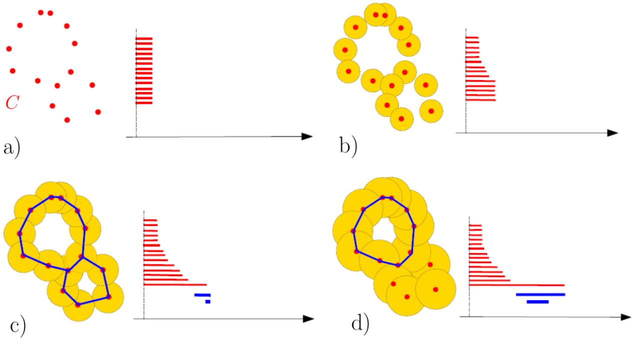
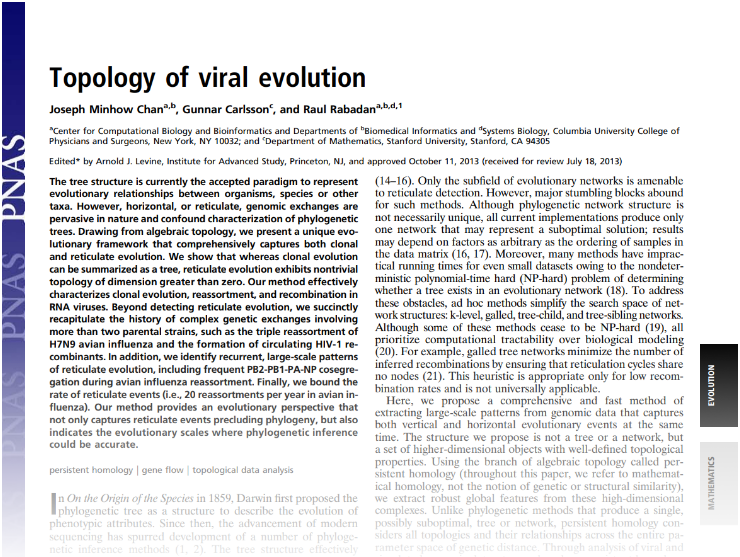


Figure 32: Persistence barcode, plotting β_0 and β_1. The size of the ball is called the *filtration value*. [C2017]

Application to Viral Evolution



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

(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

EVOLUTION

MATHEMATICS

Application to Viral Evolution

- data: viral genomes, with genetic distance as metric
- compute persistence homology; at a filtration value ε:
  - β_0 represents the number of strains/subclades
  - 1D topology provides information about horizontal evolution

Application to Viral Evolution

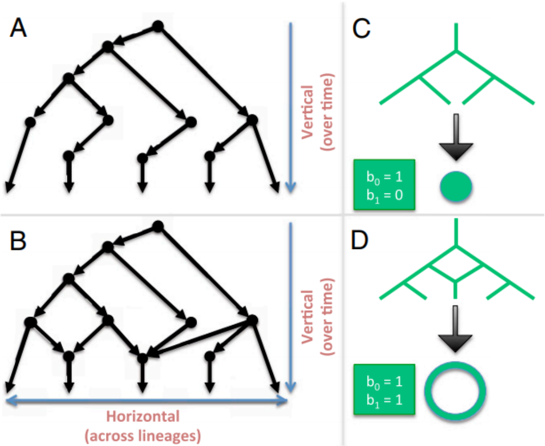


Figure 33: (A) tree shows vertical evolution (B) DAG with horizontal evolution (C) trees are contractible (D) DAGs are not.[7] [C2013]

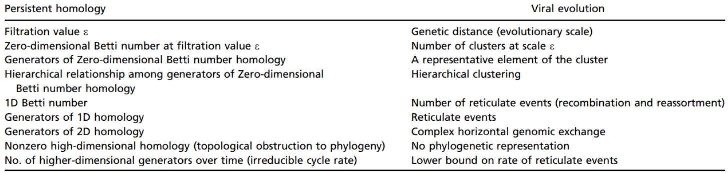
[7] Caption statement from [C2013].

Application to Viral Evolution

Hypothesis: higher dimensional homology groups capture even more complex/multiple horizontal genetic exchange.

Application to Viral Evolution

Hypothesis: higher dimensional homology groups capture even more complex/multiple horizontal genetic exchange.



Persistent homology	Viral evolution
Filtration value ε	Genetic distance (evolutionary scale)
Zero-dimensional Betti number at filtration value ε	Number of clusters at scale ε
Generators of Zero-dimensional Betti number homology	A representative element of the cluster
Hierarchical relationship among generators of Zero-dimensional Betti number homology	Hierarchical clustering
1D Betti number	Number of reticulate events (recombination and reassortment)
Generators of 1D homology	Reticulate events
Generators of 2D homology	Complex horizontal genomic exchange
Nonzero high-dimensional homology (topological obstruction to phylogeny)	No phylogenetic representation
No. of higher-dimensional generators over time (irreducible cycle rate)	Lower bound on rate of reticulate events

Application to Viral Evolution

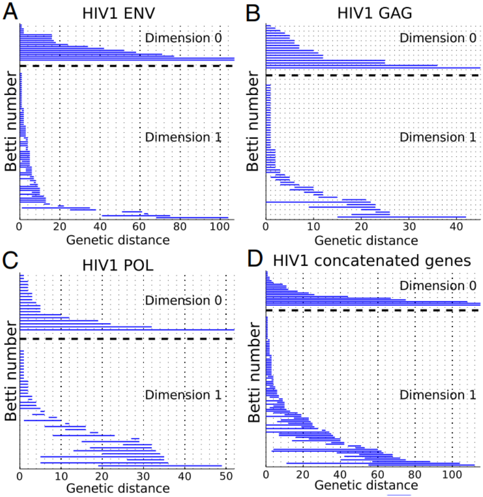


Figure 34: Persistence barcode based on genetic distances in three genes: ENV, GAG, POL. Plot (D) is based on their concatenation. [C2013]

Application to Viral Evolution

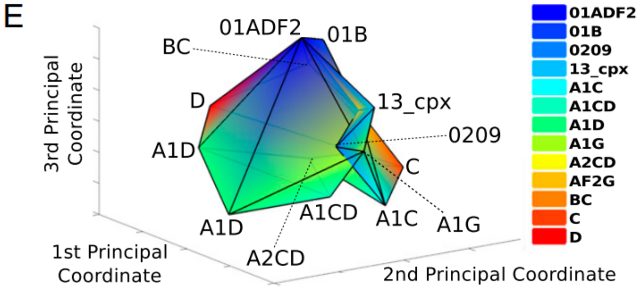


Figure 35: Representation of the recombination event with multiple parental strains. The vertices correspond to a HIV-1 subtype. [C2017]

References

[B2016] Torres, Brenda Y., et al. "Tracking resilience to infections by mapping disease space." *PLoS biology* 14.4 (2016): e1002436.

[C2009] Carlsson, Gunnar. "Topology and data." *Bulletin of the American Mathematical Society* 46.2 (2009): 255-308.

[C2013] Chan, Joseph Minhow, Gunnar Carlsson, and Raul Rabadan. "Topology of viral evolution." *Proceedings of the National Academy of Sciences* 110.46 (2013): 18566-18571.

[C2017] Chazal, Frédéric, and Bertrand Michel. "An introduction to Topological Data Analysis: fundamental and practical aspects for data scientists." *arXiv preprint arXiv:1710.04019* (2017).

[Cu2017] Curto, Carina. "What can topology tell us about the neural code?." *Bulletin of the American Mathematical Society* 54.1 (2017): 63-78.

[G2008] Ghrist, Robert. "Barcodes: the persistent topology of data." *Bulletin of the American Mathematical Society* 45.1 (2008): 61-75.

[L2003] Lee, John M. *Introduction to Smooth Manifolds*. Springer, New York, NY, 2003. 1-29.

[S2007] De Silva, Vin, and Robert Ghrist. "Coverage in sensor networks via persistent homology." *Algebraic & Geometric Topology* 7.1 (2007): 339-358.