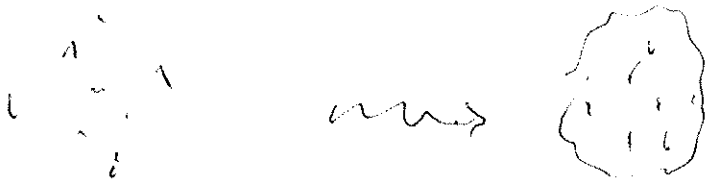
TOPOLOGICAL DATA ANALYSIS (TDA)

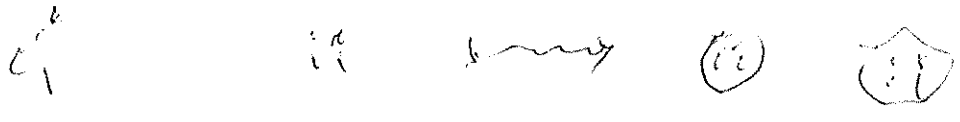
Dogma:

Data has shape,

{data} → {simplicial complex} → {algebraic invariants}

Intuitively





What is data?

Data = finite subset of $\mathbb{R}^n$

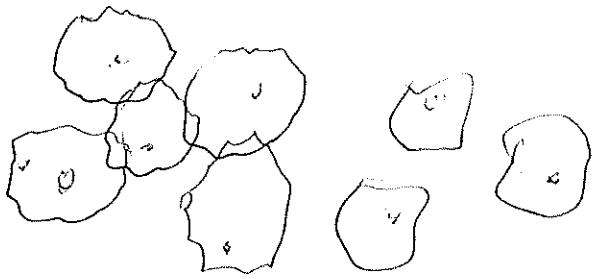
Ex. Blood tests of 100 people.

One test = 10 numbers

Get 100 points of $\mathbb{R}^{10}$.

This is data $\subset \mathbb{R}^{10}$. A set $X \subset \mathbb{R}^{10}$

Pick $\varepsilon > 0$.



$$\bigcup_{x \in X} B_\varepsilon(x) \subset \mathbb{R}^{10}.$$

object embedded in $\mathbb{R}^n$ (e.g., a compact smooth manifold). In this case, it is clear that we need to somehow "fill in the gaps" between the samples. If we have a rough sense of the average distance between points that are supposed to be connected, there is an evident construction: just take the union of balls around the points.

Definition 2.1.1 (Union of balls). Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon \geq 0$. The *union of balls* is the union

$$\bigcup_{x \in X} B_\epsilon(x) \subset \mathbb{R}^n.$$

However, from a practical perspective, the union of balls is not ideal; it is not evidently algorithmically tractable, and it requires that (X, ∂_X) arise as a subspace of $\mathbb{R}^n$. To fix the first problem, we would like to produce an abstract simplicial complex that encodes the information of the union of balls. We can adapt this construction to the discrete setting by regarding the ϵ-balls around a finite set X as a cover. That is, the idea is to associate a k-simplex to a set of k points whose ϵ-neighborhoods intersect.

Definition 2.1.2 (Čech complex). Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. The *Čech complex* $C_\epsilon(X, \partial_X)$ is the abstract simplicial complex with

1. vertices the points of X, and
2. a k-simplex $[v_0, v_1, \ldots, v_k]$ when a set of points $\{v_0, v_1, \ldots, v_k\} \subset X$ satisfies

$$\bigcap_i B_\epsilon(v_i) \neq \emptyset.$$

In fact, the Čech complex (Figure 2.2) is a special case of a standard construction from algebraic topology that associates a simplicial complex to a *cover* of a space. Recall from Definition 1.3.15 that an open cover $\{U_i\}$ of a space X is a collection of open sets such that $\cup_i U_i = X$. Given a cover $\{U_i\}$ of X, we define the *nerve* of the cover as follows.

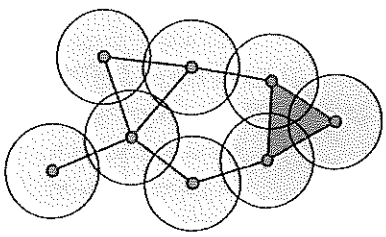


Figure 2.2 The Čech complex is a combinatorial approximation to the union of balls.

arbitrary finite metric space (X, ∂_X). The Vietoris-Rips complex is the maximal simplicial complex determined by the vertices and 1-simplices specified by the graph G.

Definition 2.1.6 (Vietoris-Rips complex). Let (X, ∂_X) be a finite metric space and fix $\epsilon > 0$. The *Vietoris-Rips complex* $\mathrm{VR}_\epsilon(X, \partial_X)$ is the abstract simplicial complex with

1. vertices the points of X, and
2. a k-simplex $[v_0, v_1, \ldots, v_k]$ when

$$\partial_X(v_i, v_j) \leq 2\epsilon \qquad \text{for all} \qquad 0 \leq i, j \leq k.$$

For a point cloud in $\mathbb{R}^n$, the Vietoris-Rips complex and the Čech complex can be different; for instance, notice that there is a difference between the Čech complex in Figure 2.2 and the Vietoris-Rips complex in Figure 2.3, which are generated by the same underlying metric space. The next example highlights the kind of phenomenon that leads to such differences.

Example 2.1.7. Consider the finite metric space $X = \{(0,0), (1,0), (\frac{1}{2}, \frac{\sqrt{3}}{2})\} \subset \mathbb{R}^2$. These points are the vertices of an equilateral triangle with side length 1. Choose an ϵ in the open interval $(\frac{1}{2}, \frac{\sqrt{3}}{3})$, i.e., $\frac{1}{2} < \epsilon < \frac{\sqrt{3}}{3}$. (For concreteness, $\frac{\sqrt{3}}{3} \approx 0.577$.)

1. The Vietoris-Rips complex $\mathrm{VR}_\epsilon(X, \partial_X)$ has three vertices (one for each point of X), three 1-simplices (connecting the points), and therefore has a single 2-simplex filling in the triangle.
2. In contrast, the Čech complex $C_\epsilon(X, \partial_X)$ has three vertices (one for each point of X) and three 1-simplices (connecting the points), but does not have the 2-simplex spanned by all the points since there is no point in the intersection of the balls of radius ϵ.

(See Figure 2.4 for a corresponding picture.)

The use of the Čech complex is justified by the Nerve Lemma (Theorem 2.1.4); there is no analogous result for the Vietoris-Rips complex. However, despite the

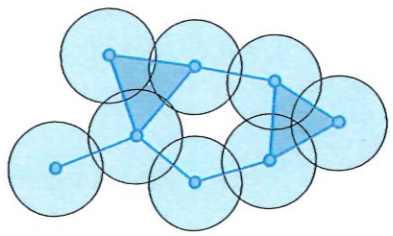


Figure 2.3 The Vietoris-Rips complex is completely determined by its 1-skeleton.

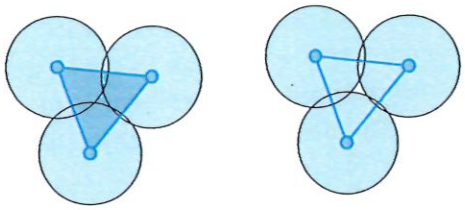


Figure 2.4 The Vietoris-Rips complex (on the left) is completely determined by its 1-skeleton, whereas the Čech complex (on the right) can potentially omit higher simplices.

fact that they are sometimes different, there is a close relationship between the Vietoris-Rips and Čech complexes.

Lemma 2.1.8. *Let $X \subset \mathbb{R}^n$ be a finite subspace and fix $\epsilon > 0$. There are natural simplicial inclusions*

$$\check{C}_{\epsilon/2}(X, \partial_X) \subseteq \mathrm{VR}_{\epsilon/2}(X, \partial_X) \subseteq \check{C}_{\epsilon}(X, \partial_X) \subseteq \mathrm{VR}_{\epsilon}(X, \partial_X) \subseteq \check{C}_{2\epsilon}(X, \partial_X) \subseteq \mathrm{VR}_{2\epsilon}(X, \partial_X)$$

An essential property of the constructions of the Čech complex and the Vietoris-Rips complex is that they are functorial. To be precise, these constructions are functorial in both X and ϵ. (In the following discussion, we focus on the Vietoris-Rips complex; the properties of the Čech complex are analogous.) For $\epsilon < \epsilon'$ and any metric space (X, ∂_X), there is an induced simplicial map

$$\mathrm{VR}_{\epsilon}(X, \partial_X) \to \mathrm{VR}_{\epsilon'}(X, \partial_X),$$

since increasing the scale parameter adds more simplices.

Next, recall that a map $f\colon X \to Y$ between metric spaces (X, ∂_X) and (Y, ∂_Y) is Lipschitz continuous with constant k if $\partial_Y(f(x_1), f(x_2)) \leq k\partial_X(x_1, x_2)$. Given a Lipschitz map $f\colon X \to Y$ with Lipschitz constant k, there is an induced simplicial map

$$f\colon \mathrm{VR}_{\epsilon}(X, \partial_X) \to \mathrm{VR}_{k\epsilon}(Y, \partial_Y)$$

for any ϵ. Summarizing, we have the following theorem.

Theorem 2.1.9. *The construction $\mathrm{VR}_{\epsilon}(-)$ specifies a functor from the category of finite metric spaces and Lipschitz maps with constant 1 to Simp. The construction $\mathrm{VR}_{(-)}(X, \partial_X)$ specifies a functor from $\mathbb{R}$ to Simp.*

This means that when we vary the scale ϵ, there is a map between the associated complexes for a given data set (X, ∂_X). And if we change a data set (X, ∂_X) to produce a new data set (Y, ∂_Y) related via a Lipschitz map, there is a map connecting the associated complexes. For example, if we add some data points, so that

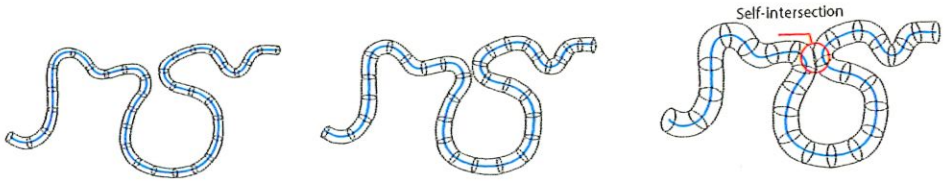


Figure 2.7 As the tubular neighborhood of a curve expands, eventually it self-intersects at the narrowest "pinch."

be thickened out to a *tubular neighborhood* of radius r; this is what one gets by extending out along the normal at any point. (See Figure 2.6 for some examples.)

The condition number is the minimum radius at which a tubular neighborhood of a manifold self-intersects; clearly, this can happen either because the manifold itself has small features (e.g., small holes) or because the embedding twists the manifold around on itself. (See Figure 2.7 for an example.)

The following theorem, due to Niyogi, Smale, and Weinberger [384], now provides a concrete result guaranteeing correct estimation of the homology.

Theorem 2.2.1. *Let M be a compact submanifold of $\mathbb{R}^n$ with condition number τ and let $\{x_1, \ldots, x_k\}$ be a set of points drawn from M according to the volume measure. Fix $0 < \epsilon < \frac{\tau}{2}$. Then if*

$$k > \beta_1 \left(\log(\beta_2) + \log\left(\frac{1}{\delta}\right) \right),$$

there is a homotopy equivalence

$$\bigcup_{z \in \{x_1, \ldots, x_k\}} B_\epsilon(z) \simeq M$$

between the union of balls and M (and in particular the homology groups coincide) with probability $> 1 - \delta$.

Here

$$\beta_1 = \frac{\text{vol}(M)}{\cos^n(\theta_1)\text{vol}(B^n_{\frac{\epsilon}{4}})}$$

and

$$\beta_2 = \frac{\text{vol}(M)}{\cos^n(\theta_2)\text{vol}(B^n_{\frac{\epsilon}{8}})},$$

where $\theta_1 = \arcsin\left(\frac{\epsilon}{8\tau}\right)$, $\theta_2 = \arcsin\left(\frac{\epsilon}{16\tau}\right)$, and $\text{vol}(B^n_r)$ denotes the volume of the n-dimensional ball of radius r.

finite metric space (X, ∂_X), $H_0(\mathrm{VR}_\epsilon(X, \partial_X))$ computes the single-linkage clustering at scale ϵ of (X, ∂_X).

When considering the persistent homology, observe that each cluster at time $p+i$ can be thought of as resulting from the merger of clusters at i. This is clearly closely related to the information encoded in the hierarchical clustering dendrogram associated to single-linkage clustering. (See Figure 2.10 for comparison of the barcode and dendrogram for a synthetic data set.)

In Figure 2.11, we see an idealized situation involving sampling from an object in $\mathbb{R}^2$. In practice, however, the barcodes are often not so easy to interpret. Even for geometrically simple situations, complications can arise. In Figure 2.12, we illustrate how the barcode can change due to perturbation of the data by considering a sequence of nested circles.

In Figure 2.13, persistent homology of genomic sequence data generated by coalescent simulation is shown. As explained in Section 5.7, this is a way of modeling evolutionary phenomena. Typically, one fits phylogenetic trees to the finite metric space of sequences; here, we compute the persistent homology instead. Computing the first persistent homology group detects when "non-tree-like" events are occurring, i.e., when there is genetic recombination. Another example of this kind of application of persistent homology in studying recombination rates in the evolution of bacteria is discussed in Section 5.6.3; see Figure 2.14. In both of these applications, increased recombination can be detected by a large number of bars in the PH_1 barcode.

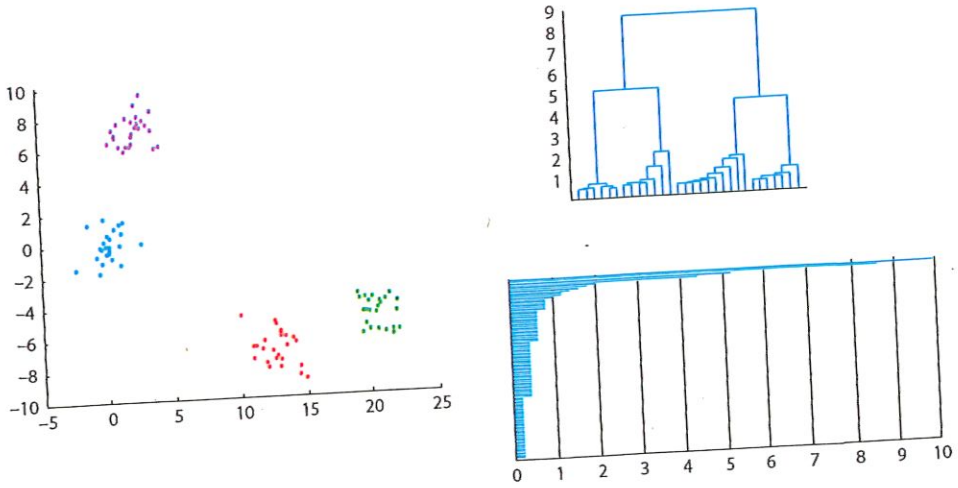


Figure 2.10 For the data set on the left, both the dendrogram and the zeroth persistent homology barcode capture how clusters merge as ϵ increases.

Lemma 2.3.8. *Let $X\colon \mathbb{R} \to \mathrm{Simp}$ be a tame filtered simplicial complex. The filtered vector space produced as $H_i(X(-);\mathbb{F})$ has the property that*

1. *each vector space $H_i(X(\epsilon);\mathbb{F})$ is of finite rank and*
2. *there exists N such that $H_i(X(\epsilon_1);\mathbb{F}) \to H_i(X(\epsilon_2);\mathbb{F})$ is an isomorphism for $\epsilon_2 > \epsilon_1 > N$.*

We say such a filtered vector space is of finite type.

Remark 2.3.9. A filtered vector space of finite type can be regarded as indexed on $\mathbb{Z}$, where the integral indices correspond to values in $\mathbb{R}$ where the homology changes.

The key classification result of Zomorodian and Carlsson [551] is then the following.

Theorem 2.3.10. *Let $\mathbb{F}$ be a field. There is a bijection between the set of finite barcodes and the set of isomorphism classes of filtered $\mathbb{F}$-vector spaces of finite type.*

The basic idea of this classification is quite simple; we define *interval modules*, which are filtered systems I_{ab} of $\mathbb{F}$-vector spaces $\{V_i\}$ where for $i \in [a,b]$, $V_i = \mathbb{F}$, and all the maps $\mathbb{F}' \to \mathbb{F}$ are the identity (and the others are necessarily zero). Then any filtered system of $\mathbb{F}$-vector spaces is a direct sum of interval modules; the interval modules correspond to the bars in the barcode representing the lifetime of particular elements in homology.

Theorem 2.3.10 tells us that all of the information in the filtered system of vector spaces can be encoded as barcodes. It is often useful to think of a barcode as a collection of points in $\mathbb{R}^2$, specified by the endpoints of the intervals. Such a set is referred to as a *persistence diagram*, and often it is regarded as containing the entire diagonal (consisting of size zero bars).

In conclusion, we have the "persistent homology pipeline"

$$\left\{\begin{matrix}\text{finite}\\ \text{metric}\\ \text{spaces}\end{matrix}\right\} \to \left\{\begin{matrix}\text{filtered}\\ \text{simplicial}\\ \text{complexes}\end{matrix}\right\} \to \left\{\begin{matrix}\text{barcodes / persistence}\\ \text{diagrams}\end{matrix}\right\}.$$

We now turn to some examples of the use of barcodes to describe shape. When $k = 0$, the persistent homology is describing a standard hierarchical clustering construction.

Example 2.3.11. Recall from Theorem 1.10.10 that for a simplicial complex X, $H_0(X)$ is computing the free abelian group on the components. In the case of $\mathrm{VR}_\epsilon(X, \partial_X)$ for a

4. Compute the collection of vector spaces

$$\{H_2(|\mathrm{VR}_{\epsilon_1}(X, \partial_X)|), H_2(|\mathrm{VR}_{\epsilon_2}(X, \partial_X)|), \ldots, H_2(|\mathrm{VR}_{\epsilon_m}(X, \partial_X)|)\}.$$

5. Compare these abelian groups; for example, make a graph of the ranks of the free parts. If these are all non-zero and all the same, it suggests that there are stable topological features of M at the feature scales in the interval $[\epsilon_{\min}, \epsilon_{\max}]$. If there is a subinterval $[a; b] \subset [\epsilon_{\min}, \epsilon_{\max}]$ on which the ranks are the same, we might conclude that there are stable topological features at those ranges of scales. (Of course, there is no guarantee that we are not seeing different features at the different scales; this procedure does not really help us match topological features across scales.)

For an example of how this might work, consider the situation depicted in Figure 2.8. When ϵ is smaller than the distance between points, the Vietoris-Rips

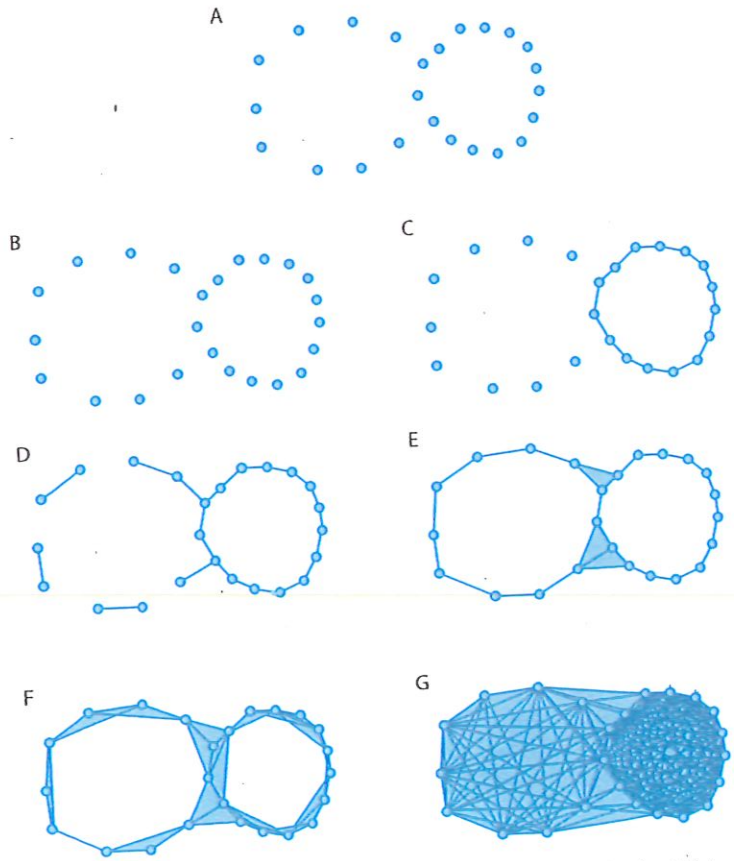


Figure 2.8 As ϵ increases, more and more simplices appear in the Vietoris-Rips complex.

Roughly speaking, an element $\gamma \in H_k(\mathrm{VR}_{\epsilon_i}(X, \partial_X))$ represents a k-dimensional hole in the geometric realization of the Vietoris-Rips complex at ϵ_i. If γ does not exist for $\epsilon' < \epsilon_i$, we think of this feature as being "born" at ϵ_i. When $\theta_{ij}(\gamma) = 0$, it means that the hole has been filled in by a collection of simplices with boundary γ. This suggests that it makes sense to try to figure out the "lifespan" of a particular element in homology, i.e., when it first appears and when it vanishes. More precisely, for a filtered simplicial complex $X_\bullet$, an element $\gamma \in H_k(X_i; \mathbb{F})$ is

1. *born* at i if it is not in the image of $H_k(X_{i-q}; \mathbb{F}) \to H_k(X_i; \mathbb{F})$ for any $q > 0$, and
2. *dies* at $\ell > i$ if it becomes zero in $H_k(X_\ell; \mathbb{F})$ or its image in $H_k(X_\ell; \mathbb{F})$ coincides with the image of another class that was born earlier.

Thus, we can think of the information contained in the filtered system of vector spaces as a series of elements with intervals representing their lifetime. Precisely, the persistent homology of a finite metric space can be described via a "barcode," a collection of intervals. Each interval represents the lifespan of a homological feature. (See Figure 2.9 for a simple representative example.)

Definition 2.3.6. A barcode is a multiset of non-empty intervals of the form either $[x, y) \subset \mathbb{R}$ or $[x, \infty)$. (A multiset is a generalization of a set where repeated elements are allowed, e.g., {1,1,2}.)

To be precise about the connection between persistent homology and barcodes, we require some finiteness hypotheses that always hold in practice, since we only have finitely many data points. We fix a field $\mathbb{F}$ for the remainder of this section.

Definition 2.3.7. A filtered simplicial complex is *tame* if the homology groups $H_i(-; \mathbb{F})$ are always of finite rank and change at only a finite number of indices.

By Lemma 2.3.1, the filtered complexes produced by applying the Vietoris-Rips complex construction to a finite metric space are always tame.

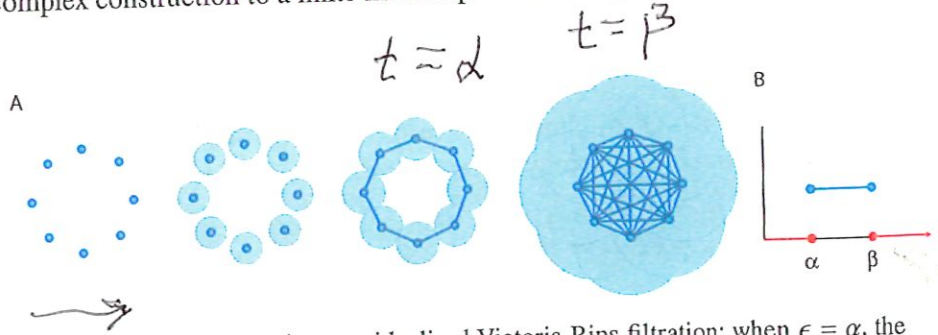


Figure 2.9 In (A), we have an idealized Vietoris-Rips filtration: when $\epsilon = \alpha$, the circle appears, and when $\epsilon = \beta$, the circle is filled in. In (B), the barcode has a single bar (representing a $\mathbb{Z}$ in homology) that appears at α and vanishes at β; this is the homology of the circle, for as long as it lasts.

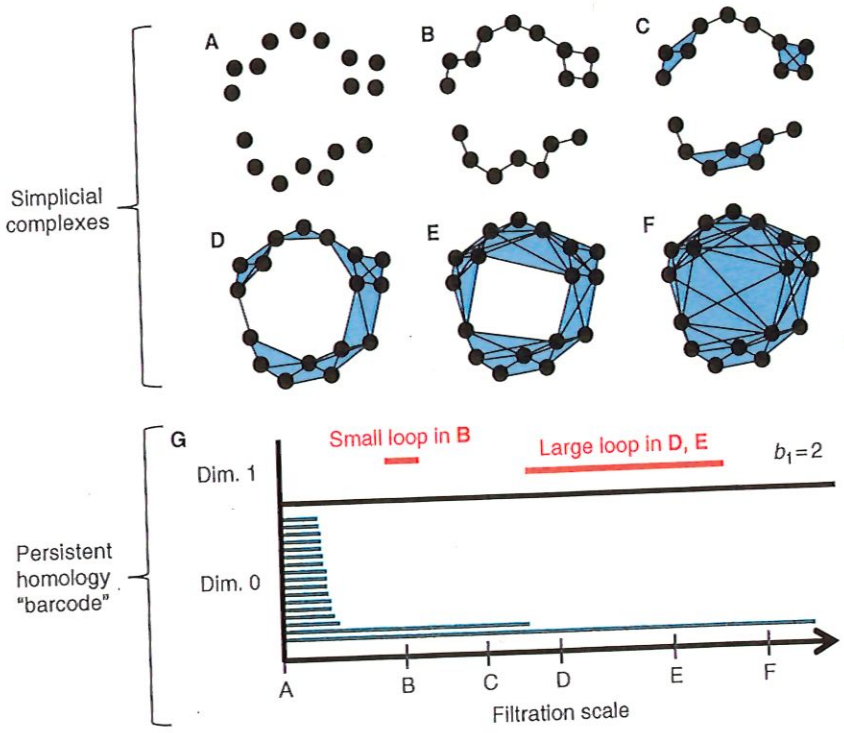


Figure 2.11 The points in panel A form a circle, with a horizontal gap separating upper and lower points. Panels A-F show the Vietoris-Rips filtration on these points as ϵ increases. Panel G shows the barcode. PH_0 (dimension 0) shows clustering of the data at different scales; each horizontal bar in the barcode is a cluster. In panel A (filtration scale 0), no points are connected; each is its own cluster (represented as 17 horizontal bars). As the scale increases, points in the simplicial complex connect, represented in the barcode as termination of a bar. There are two distinct clusters through panels B and C and one cluster in panels D, E, and F. PH_1 (dimension 1) shows loops in the data at different scales. Each bar in this part of the barcode identifies a different loop. There are two loops in this data: a short-lived loop in the top-right of the simplicial complex at scale B, and a long-lived loop appearing in panel D and persisting through panel E – this loop is represented as the long bar in the dimension 1 barcode. Robust features of the data set are captured in the barcode: the data clusters into two groups (two dimension 0 bars through scale C), and forms a loop (one long dimension 1 bar). The persistent first Betti number (b_1) is the total number of dimension 1 bars; here it is equal to 2.

In Section 8.3, we discuss an application of persistent homology to study the physical structure of DNA. Modeling DNA as a sequence of repeated units that have prescribed interaction points, persistent homology can be used to extract information about loops in the strands from a similarity matrix encoding the contact of sites with other sites. (See Figure 2.15.)

Claim. If two finite metric spaces are "close" to each other then their bar-codes are close to each other.

In technical terms: if $D \geq 0$ is the Gromov-Hausdorff distance between two finite metric spaces then the bottleneck distance between their bar-codes $\leq D$.

Figure 2.16 The Hausdorff distance is determined by the point in A with the largest distance to the closest point in B (and vice versa).

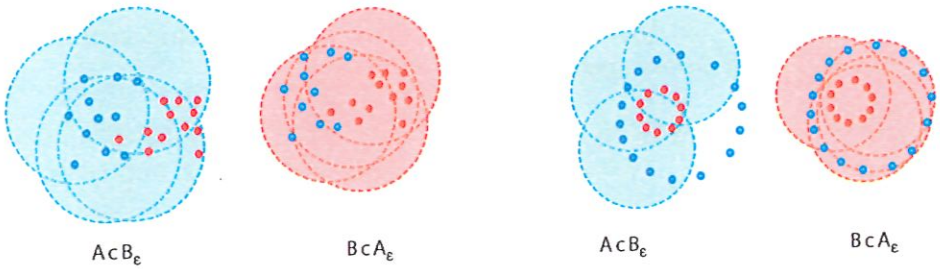


Figure 2.17 The Hausdorff distance can be computed by considering the smallest ϵ fattening of each set that contains the other.

$$d_H(A,B) = \inf_{\epsilon>0}\{B \subseteq A_\epsilon, A \subseteq B_\epsilon\},$$

where A_ϵ and B_ϵ denotes the sets of all points within distance ϵ of A and B, respectively (see Figures 2.16, 2.17).

Example 2.4.2.

1. Let $A \subset X$ and suppose that B is generated from A by perturbing each point $a \in A$ by at most ϵ; i.e., the points of B are in bijection with those of A and (denoting the bijection by θ) we have $\partial_X(a, \theta(a)) \leq \epsilon$. For instance, consider $A = \{[0,0,0],[1,2,3],[-1,0,5]\} \subset \mathbb{R}^3$ and $B = \{[\epsilon,0,0],[1,2+\epsilon,3],[-1,0,5-\epsilon]\}$. Then $d_H(A,B) \leq \epsilon$.
2. The Hausdorff distance is heavily influenced by the single most extreme point; given $A \subset X$, let $A' = A \cup \{x\}$. Then $d_H(A,A') = \min_{a\in A} \partial_X(x,a)$.

Lemma 2.4.3. *The Hausdorff distance imposes a metric on the set of non-empty subsets of a metric space (X, ∂_X).*

However, we cannot in general assume that the metric spaces we consider are given as subsets of a common ambient metric space. A key insight of Gromov is to circumvent this issue by considering the infimum of the Hausdorff distance over all isometric embeddings of the two metric spaces into a larger ambient metric space. Here an isometric embedding

$$\phi\colon (X, \partial_X) \to (Y, \partial_Y)$$

is an injective map $X \to Y$ such that

$$\partial_X(x_1, x_2) = \partial_Y(\phi(x_1), \phi(x_2)).$$

That is, an isometric embedding identifies X with a submetric space of Y.

Definition 2.4.4. Let (X_1, ∂_{X_1}) and (X_2, ∂_{X_2}) be compact metric spaces. Th *Gromov-Hausdorff distance* between X_1 and X_2 is defined to be

$$d_{GH}((X_1, \partial_{X_1}), (X_2, \partial_{X_2})) = \inf_{\substack{\theta_1\colon X_1 \to Z \\ \theta_2\colon X_2 \to Z}} d_H(X_1, X_2).$$

Here θ_1 and θ_2 are isometric embeddings of (X_1, ∂_{X_1}) and (X_2, ∂_{X_2}) in (Z, ∂_Z) respe tively (see Figure 2.18 for an example); the infimum is taken over all such $(Z, \partial$ and embeddings θ_1 and θ_2.

We will say that two metric spaces are *isometric* if there exists an isomorphis $f\colon X \to Y$ that preserves all distances. This clearly defines an equivalence relatio on the set of metric spaces.

Theorem 2.4.5. *The Gromov-Hausdorff distance is a metric on the set of isometr classes of compact metric spaces.*

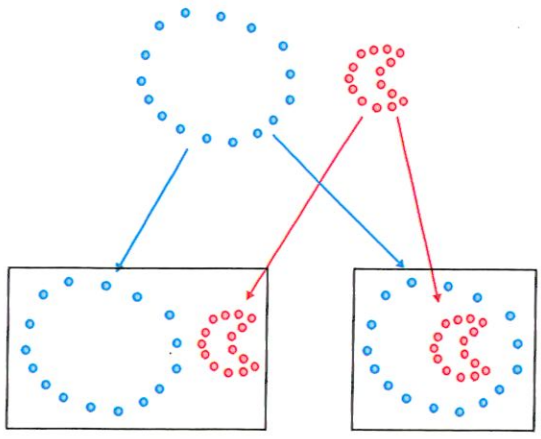


Figure 2.18 The Gromov-Hausdorff distance is computed by minimizing over all embeddings; here, the embedding on the right has a much smaller Hausdorff distance between the two image sets than the embedding on the left.

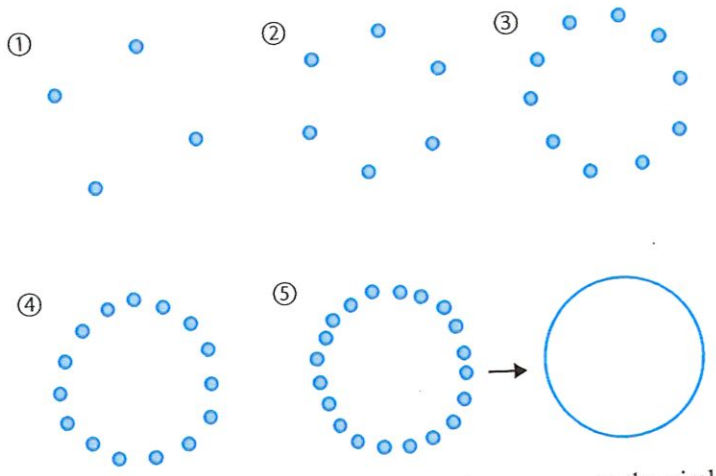


Figure 2.19 Samples of points that lie on a circle converge to the circle in the Gromov-Hausdorff distance as the sampling density increases.

We now turn to the description of various metrics on the set of barcodes. We begin with the *bottleneck distance*. Given two intervals $[a_1, b_1)$ and $[a_2, b_2)$, define

$$d_\infty([a_1, b_1), [a_2, b_2)) = \max(|a_1 - a_2|, |b_1 - b_2|).$$

We extend d_∞ to include $\emptyset$ by defining

$$d_\infty([a, b), \emptyset) = \frac{|b - a|}{2}.$$

Now given two barcodes B_1 and B_2, we define a matching between B_1 and B_2 as follows. Without loss of generality, assume that $|B_1| \leq |B_2|$. Then a matching is specified by a bijection $\phi\colon A_1 \to A_2$, where A_1 is a multi-subset of B_1 and A_2 is a multi-subset of B_2. We formally add $\emptyset$ to B_1 and B_2, and we regard the elements of $B_1 \setminus A_1$ and $B_2 \setminus A_2$ as matched with $\emptyset$.

Definition 2.4.8. Let B_1 and B_2 be barcodes. The *bottleneck distance* is defined to be

$$d_B(B_1, B_2) = \inf_\phi \sup_{Z \in B_1} d_\infty(Z, \phi(Z)),$$

where ϕ varies over all matchings between B_1 and B_2 and the supremum is taken over bars in B_1.

Roughly speaking, the bottleneck distance measures the worst discrepancy in the best matching between the two barcodes. Note that two barcodes which are a distance ϵ apart in the bottleneck distance could differ in an essentially arbitrary number of short bars of length less than $\frac{\epsilon}{2}$. Put another way, two barcodes are close

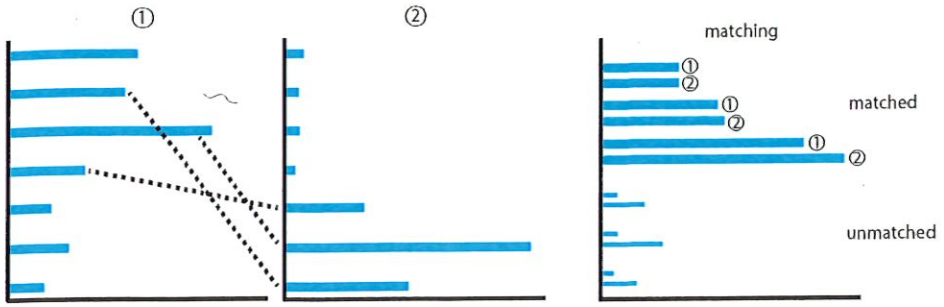


Figure 2.20 The bottleneck distance on barcodes is computed by matching long bars. Figure from experiment performed by Elena Kandror, Abbas Rizvi, and Tom Maniatis at Columbia University, with permission.

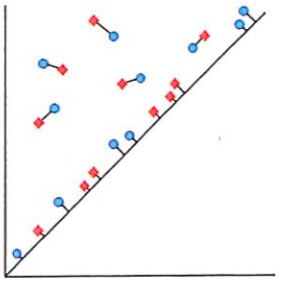


Figure 2.21 The bottleneck distance when expressed in terms of persistence diagrams is computed by matching nearby points and assigning points close to the diagonal to the nearest diagonal point.

in the bottleneck distance if after ignoring "short" bars, the endpoints of matching "long" bars are close (see Figures 2.20 and 2.21 for examples.)

There are other sensible metrics on the space of barcodes, most notably including mass transportation (Wasserstein) metrics. Since it will be convenient for later use, we will also introduce the *Wasserstein metric* here.

Definition 2.4.9. Let B_1 and B_2 be barcodes. For $p > 0$, the *p-Wasserstein distance* is defined to be

$$d_{W_p}(B_1, B_2) = \left(\inf_{\phi} \sum_{Z \in B_1} d_\infty(Z, \phi(Z))^p \right)^{\frac{1}{p}}.$$

We can now state the stability theorem for persistent homology, arguably the most important theorem in the subject [117]. (See Figure 2.22 for an illustration.)

Theorem 2.4.10. *Let (X, ∂_X) and (Y, ∂_Y) be finite metric spaces. Then for all $k \geq 0$,*

$$d_B(\mathrm{PH}_k(\mathrm{VR}(X, \partial_X)), \mathrm{PH}_k(\mathrm{VR}(Y, \partial_Y))) \leq d_{GH}((X, \partial_X), (Y, \partial_Y)).$$

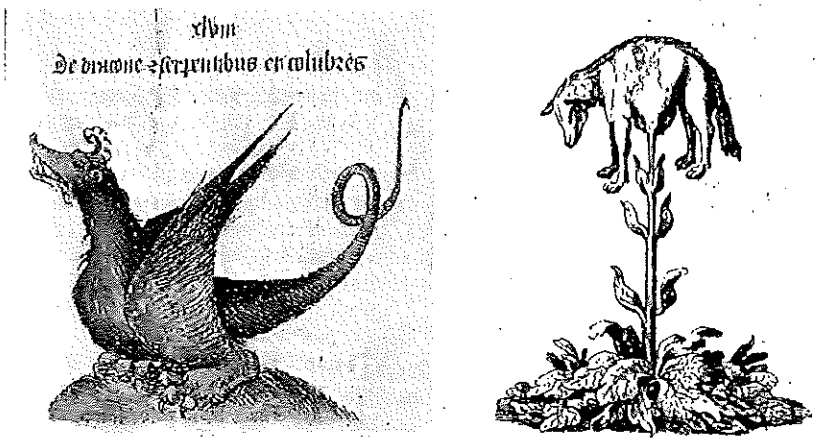


Figure 5.2 The "animalia paradoxa" were animals that challenged Linnaean taxonomy because they possessed similarities with organisms belonging to at least two different higher taxa. The dragon, for instance, had a body similar to that of a reptile but also wings like birds (illustration from the *Liber Floridus*, or *Book of Flowers*, circa 1100AD, public domain). The Borometz, or Scythian Lamb, was a plant that grew lambs. Source: Lee, H. 1887. The Vegetable Lamb of Tartary: a Curious Fable of the Cotton Plant, to Which Is Added a Sketch of the History of Cotton and the Cotton Trade. S. Low, Marston, Searle & Rivington, London.

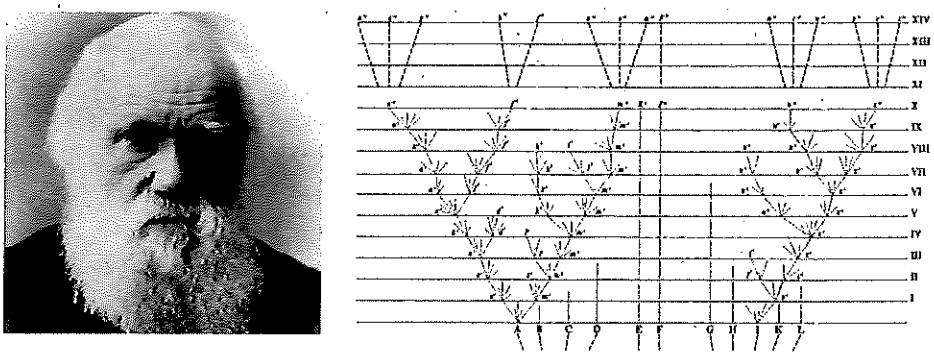


Figure 5.3 This tree appeared in Darwin's *On the Origin of Species* as a means of capturing the divergence of species. In this figure, time advances moving up the tree. The roots of the tree represent the original species that diversified according to a branching process through progeny variation and selection. The top branches constitute the modern species, and the branches that do not persist to the top represent extinct species. Source: Left: Library of Congress, Prints & Photographs Division, reproduction number, LC-DIG-ggbain-03485. Right: Darwin, C. R. 1859. *On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life*. London: John Murray.

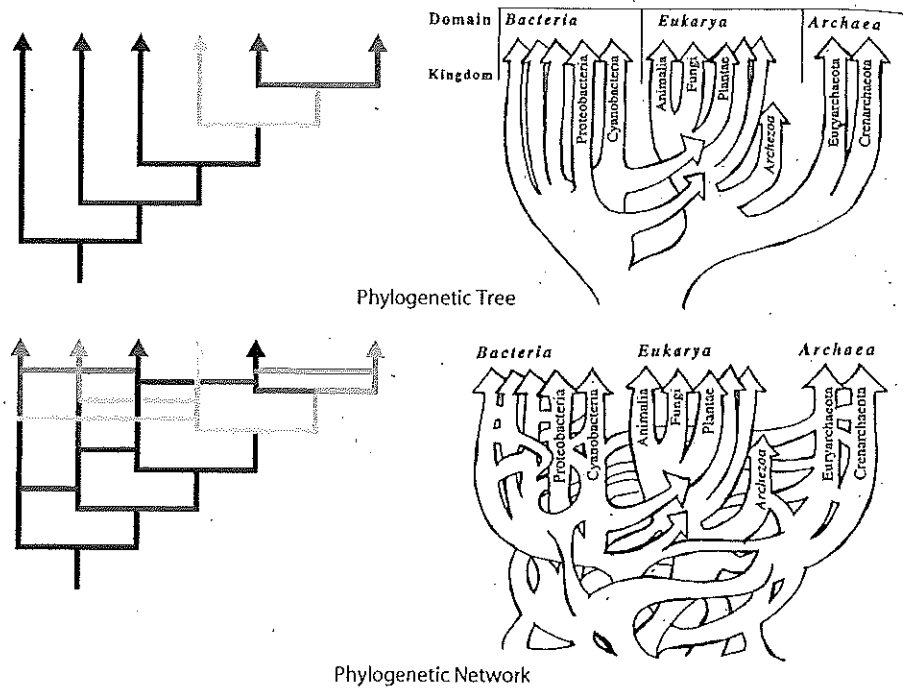


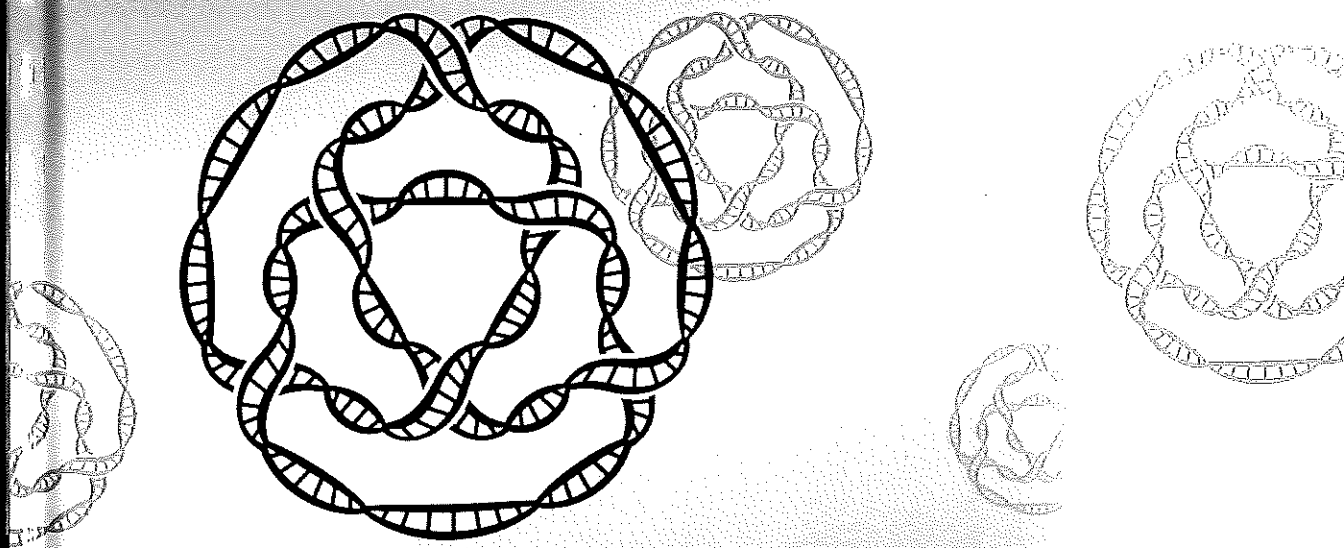
Figure 5.4 Idealized, simplistic phylogenetic trees contrast with more realistic, complex reticulate networks. On the top right is the Doolittle representation of the Tree of Life, made before the advent of sequencing technologies. It was thought that most evolution occurred through branching processes, with the notable exceptions of mitochondria and chloroplasts – believed to be symbiotic bacteria that fused part of their genome to their host's. This picture is changing as the significance of horizontal exchange of genomic information is becoming more evident. Source: [147]. From Doolittle, W. F., Phylogenetic Classification and the Universal Tree, *Science*, 1999, 284 (5423): 2124–2128. © 1999 Reprinted with permission from AAAS.

found in viruses, for instance. As we will see later in detail, viral influenza undergoes horizontal evolution through reassortment and HIV undergoes horizontal evolution through recombination. Phylogenetic trees, however, are not able to capture these horizontal evolutionary events. Representing these events graphically requires a new structure called a *reticulate network*, in which branches are allowed to both join and split. Places in the network where branches merge are known as cycles and correspond to individual reticulate events (Figure 5.5, right). The resulting network is the result of merging many different trees with different topologies.

To detect reticulate events by phylogenetic means, one must first construct a tree for each gene in the genome and then cross-reference each pair of trees for conflicts in lineal history. The simple example of the Network of Life, depicted

TOPOLOGICAL DATA ANALYSIS FOR GENOMICS AND EVOLUTION

TOPOLOGY IN BIOLOGY



Raúl Rabadán and Andrew J. Blumberg