ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Application of persistent homology to viral evolution

Bachelor Thesis

Luisa S. Meinecke

July 5, 2023

Advisor: Dr. Sara Kalisnik Hintz
Department of Mathematics, ETH Zürich

Abstract

In this thesis I introduce persistent homology, a variant of homology adapted to the setting of finite metric spaces. Intuitively speaking, persistent homology counts the occurrence of holes and higher dimensional cavities in a data set. I also discuss an application of persistent homology to viral evolution. For a long time, the tree structure was the best way to represent processes in evolution of phenotypic attributes. however, horizontal events such as recombination and reassortment in viruses cannot be captured in a tree as they generate holes. Persistent homology can ”count” these holes. I mainly follow the main paper written by Joseph Minhow Chan, Gunnar Carlsson and Raul Rabadan called *Topology of viral evolution* [5].

Acknowledgements

I would like to express my gratitude to Dr. Sara Kalisnik Hintz for supervising my thesis. I thank her especially for answering all my questions, giving very detailed comments on my first drafts, suggesting me this very interesting paper to work on as well as helping me get a deep understanding of the topic. I am also thankful for the support of my family and friends during the writing period of my thesis.[1]

[1]Moreover, I am grateful that I could use the thesis template provided by CADMO.

Contents

Chapter 1

Introduction

The goal of this thesis is to show how persistent homology can be applied to viral evolution. Persistent homology is a tool that measures the shape of a point cloud or a finite metric space. In the simplest setting, the idea is to take the point cloud and thicken the points by a radius r and then track the appearance of holes and higher dimensional cavities as we vary r. Look at the topology of Figure 1.1. In the fist step we see four connected components, in the second step all points are connected with its two neighbours but there is still a hole in the middle, because the distance between the diagonal points is smaller than the radius. In the third step there is no longer a hole in the middle. The geometric structure is captured by the so-called simplicial complexes. A 0-simplex is a point, a 1-simplex is a line, a 2-simplex is a triangle, a 3-simplex is a tetrahedron and so on.

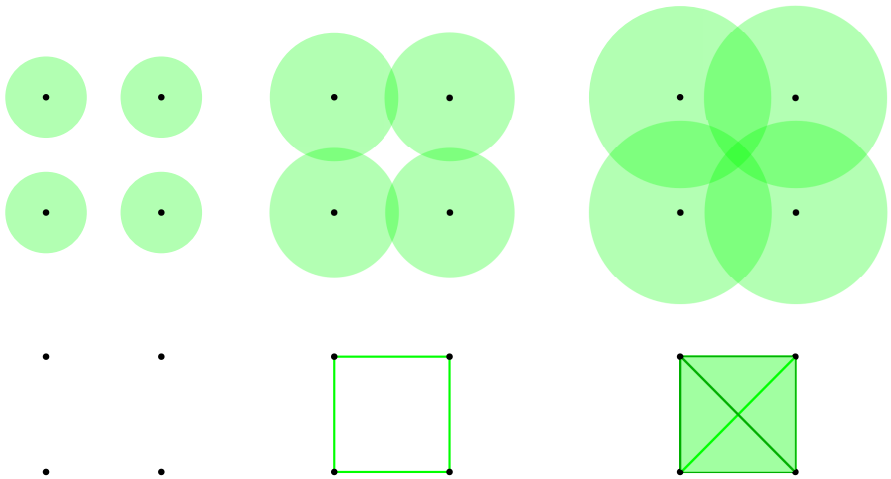


Figure 1.1: Point cloud data thickened by various r. The simplicial complexes modelling the spaces in the topology are depicted.

For a long time, evolution was modelled by phylogenetic trees. A phylogenetic tree is a tree that tries to reconstruct evolutionary relationships between species. Figure 1.2 shows an example of a phylogenetic tree and indicates the phenotypic (concerning observable traits) splitting reason between the species. We will see that phylogenetic trees are still the best way to model

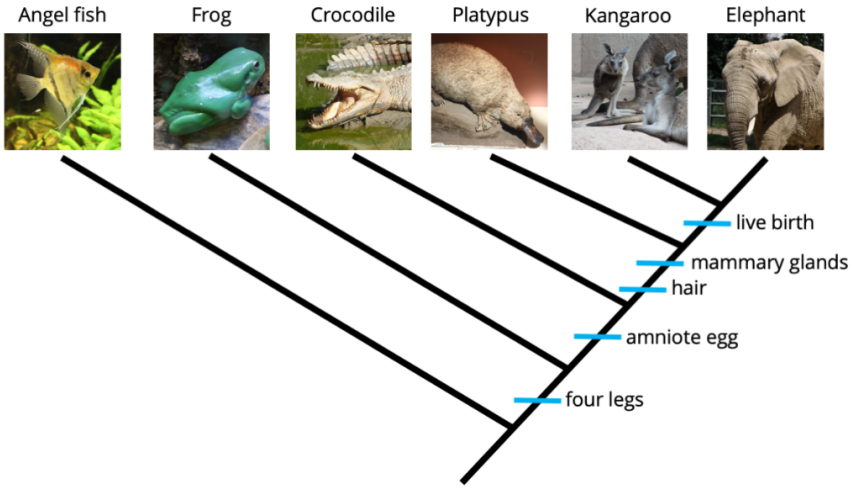


Figure 1.2: Phylogenetic tree of angel fish, frog, crocodile, platypus, kangaroo and elephant [1].

evolutionary processes that contain only vertical events, but they are not convenient to model horizontal events. We now apply this idea of thickening point clouds to genomic data with the goal of giving a criterion for how non-tree-like the evolution of viruses is. The distance between the data points is the genetic distance, which we will measure with the Hamming distance. The Hamming distance considers two genetic sequences of the same length and counts the number of positions at which their corresponding characters differ. So the Hamming distance between ”cloud” and ”proud” is 2. Detected holes at some genetic distance correspond to horizontal evolutionary processes.

All processes in viral evolution that have no reassortment or recombination in it can be modelled as a phylogenetic tree. One of the main results about trees discussed in this thesis is that for a tree-like metric space there are no holes and higher dimensional cavities in the Vietoris-Rips complex for any given radius. This theorem is stated and proved in Section 3.2 and topological obstruction to phylogeny in Section 4.3.

Using tree-like metric spaces and persistent homology we can now study the evolution of viruses and can decide whether it has a underlying phylogenetic tree structure or a more complex structure with horizontal evolutionary events.

Chapter 2

Persistent homology

In this chapter we define persistent homology. We first look at abstract simplicial complexes and homology and then discuss persistence vector spaces and the main theorem in this chapter, the Decomposition Theorem. We refer the reader to [10] for basic homology theory and for notation and contents of the following sections we follow [4].

2.1 Abstract simplicial complexes

Definition 2.1 *A set of points $X = \{x_0, x_1, \ldots, x_n\}$ as a subset of Euclidean space $\mathbb{R}^m$ is called **affinely independent** if it is not contained in an affine hyperplane of $\mathbb{R}^m$ that has smaller dimension than n for $m > n$.*

Definition 2.2 *A **n-simplex** σ in a Euclidean space $\mathbb{R}^m$ is defined as the convex hull of the set $X = \{x_0, x_1, \ldots, x_n\}$ of n affinely independent points. The x_i s are called **vertices** and the simplices spanned by $Y \subseteq X$, where $|Y| = i$ are called **i-faces**. 1-simplices are called edges.*

Definition 2.3 *A **geometric simplicial complex** is a finite collection $\mathcal{X}$ of simplices in a Euclidean space, such that:*

- *For a simplex σ of $\mathcal{X}$ all the faces of σ are also in $\mathcal{X}$.*
- *For two simplices σ and τ of $\mathcal{X}$ $\sigma \cap \tau$ is a simplex of $\mathcal{X}$ and a face of σ and τ.*

Example 2.4 In Figure 2.1 we see a geometric simplicial complex. Figures b) and c) are not geometric simplicial complexes. In b) there are intersections of edges, that are not vertices (marked green) so the second condition for a geometric simplicial complex is violated. In c) we have a green 2-simplex but not all of its faces are in the drawn structure so the first condition is violated.

Definition 2.5 *An **abstract simplicial complex** X is a pair $(V(X), \Sigma(X))$, where $V(X)$ is the is the vertex set of X and $\Sigma(X)$ is a set of subsets of the collection of non-*

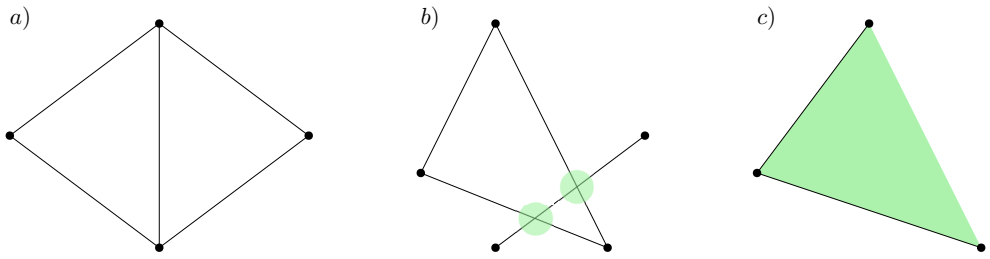


Figure 2.1: a) A geometric simplicial complex, b) and c) are no geometric simplicial complexes.

empty subsets of $V(X)$, satisfying the condition that for $\sigma \in \Sigma(X)$ and $0 \neq \tau \subseteq \sigma$ it holds that $\tau \in \Sigma(X)$. $\Sigma(X)$ is called the ***set of simplices****.*

A special case of an abstract simplicial complex is a graph.

Definition 2.6 *A* ***graph*** *is a pair (V, E), where V denotes the set of vertices and E the set of edges between these vertices.*

Definition 2.7 *An* ***additive graph*** *is a graph Γ where each edge (v, v') gets a positive length assigned, which is denoted by $\lambda_\Gamma(v, v')$.*

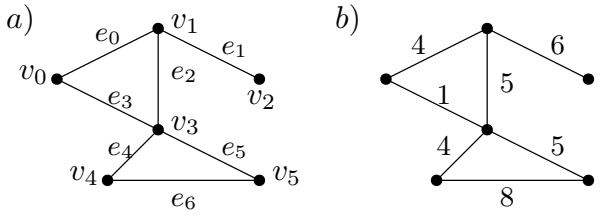


Figure 2.2: a) A graph with vertices $V = \{v_0, \ldots, v_5\}$ and edges $E = \{e_0 \ldots, e_6\}$, b) an additive graph.

Note that we can derive an abstract simplicial complex from a given geometric simplicial complex $\mathcal{X}$ by choosing the vertex set $V(\mathcal{X})$ as the set of all vertices of simplices of $\mathcal{X}$ and by letting a subset of $V(\mathcal{X})$ be in $\Sigma(\mathcal{X})$ if and only if the set is the set of vertices of some simplex of $\mathcal{X}$.

Definition 2.8 *For abstract simplicial complexes X and Y, a* ***map of abstract simplicial complexes*** *$f\colon X \to Y$ is a function of sets $f_V\colon V(X) \to V(Y)$, such that for all simplices $\sigma \in \Sigma(X)$, the subset $f_V(\sigma) \in \Sigma(Y)$.*

Example 2.9 Let $X = (V(X), \Sigma(X))$ with

$$V(X) = \{x_1, x_2, x_3\},$$
$$\Sigma(X) = \{\{x_1\}, \{x_2\}, \{x_3\}, \{x_1, x_2\}, \{x_2, x_3\}, \{x_1, x_3\}\}$$

and $Y = (V(Y), \Sigma(Y))$ with

$$V(Y) = \{y_1, y_2, y_3\},$$
$$\Sigma(Y) = \{\{y_1\}, \{y_2\}, \{y_3\}, \{y_1, y_2\}, \{y_2, y_3\}, \{y_1, y_3\}, \{y_1, y_2, y_3\}\}.$$

We define $f\colon V(X) \to V(Y)$ with

$$f(\{x_1\}) = \{y_2\}, f(\{x_2\}) = \{y_2\} \text{ and } f(\{x_3\}) = \{y_1\}.$$

It follows that

$$f(\{x_1, x_2\}) = \{y_2\}, f(\{x_2, x_3\}) = \{y_1, y_2\} \text{ and } f(\{x_1, x_3\}) = \{y_1, y_2\}.$$

All these images are simplices in Y. Illustration of the associated geometric simplicial complexes of this example is shown in 2.3.

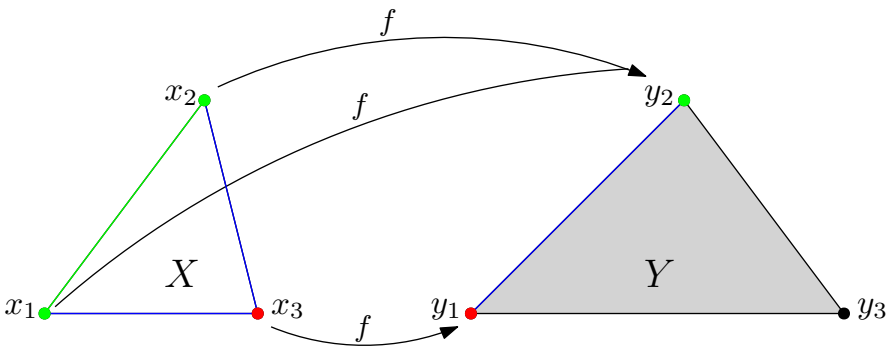


Figure 2.3: A map between simplices is determined by the images of its vertices.

Definition 2.10 *A geometric simplicial complex X in $\mathbb{R}^m$ is called a **geometric realization** of an abstract simplicial complex X' if and only if there is an embedding $e\colon V(X') \to \mathbb{R}^m$ that takes every i-simplex $\{v_0, \ldots, v_i\}$ in X' to a i-simplex in X that is the convex hull of $e(v_0), \ldots, e(v_i)$. We denote this by $|X|$.*

Note that the geometric realization construction is functorial in the sense that every map of abstract simplicial complexes $f\colon X \to Y$ induces a map on the corresponding geometric realizations $|f|\colon |X| \to |Y|$, such that $|id_X| = id_{|X|}$ and $|f \circ g| = |f| \circ |g|$.

Remark 2.11 One can determine a geometric simplicial complex up to homeomorphism from an abstract simplicial complex, in the sense that a geometric simplicial complex is homeomorphic to the geometric realisation of its associated abstract simplicial complex.

Example 2.12 In Figure 2.4 we see the geometric realizations of two abstract simplicial complexes X_1 and X_2, where

$$V(X_1) = \{1,2,3\},$$
$$\Sigma(X_1) = \{\{1\},\{2\},\{3\},\{1,2\},\{2,3\},\{1,3\}\},$$
$$V(X_2) = \{1,2,3,4\},$$
$$\Sigma(X_2) = \{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{2,3\},\{1,3\},\{1,2,3\},\{3,4\}\}.$$

X_3 with

$$V(X_3) = \{1,2,3\} \text{ and}$$
$$\Sigma(X_3) = \{\{1\},\{2\},\{3\},\{1,2\},\{2,3\},\{1,2,3\}\}$$

is not an abstract simplicial complex, because $\{1,3\} \subseteq \{1,2,3\}$ but $\{1,3\} \notin \Sigma(X_3)$.

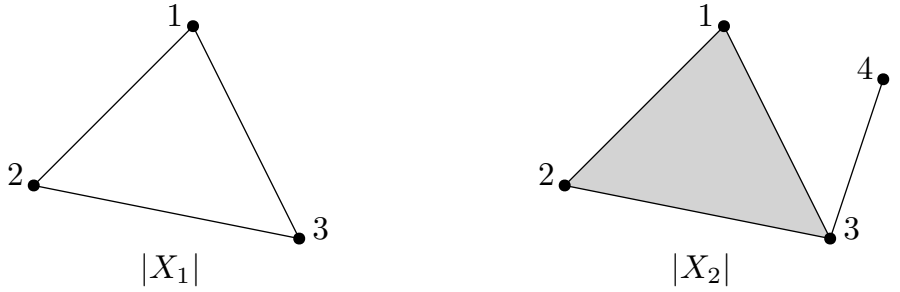


Figure 2.4: Geometric realizations $|X_1|$ and $|X_2|$ of abstract simplicial complexes in Example 2.12

Definition 2.13 *A **filtered simplicial complex** (or a **filtration**) of a set X is a family $(S_a | a \in \mathbb{R})$ of subcomplexes of some fixed simplicial complex $\bar{S}$ with vertex set X such that $S_a \subseteq S_b$ for any $a \leq b$.*

2.2 Homology

Homology is an approach to study shapes of topological spaces and to characterize them through occurring patterns. The idea is to count holes or higher dimensional cavities in a space. To study homology we only consider coefficients over $\mathbb{Z}_2$.

Definition 2.14 *For a field k and a finite set X, the **free k-vector space on** X $V_k(X)$ is defined as the k-span of the set of elements of X.*

Definition 2.15 *The **set of i-dimensional simplices of** X is the set of subsets of the vertex set $V(X)$ that have cardinality $i+1$. We will denote this set by $\Sigma_i(X)$.*

Example 2.16 In Figure 2.5 we can see an example of the associated geometric simplicial complex of an abstract simplicial complex X with

$$V(X) = \{1,2,3,4\},$$
$$\Sigma(X) = \{\{1\},\{2\},\{3\},\{4\},\{1,2\},\{2,3\},\{3,4\},\{1,4\},\{2,4\},\{2,3,4\}\}$$

and its sets of i-dimensional simplices for $i = 0,1,2$.

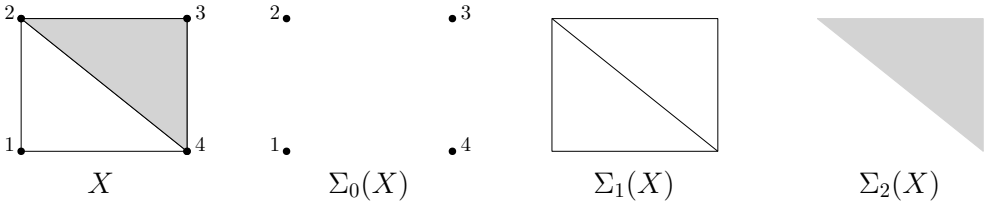


Figure 2.5: The geometric simplicial complex X and the sets of its i-simplices.

For the following definitions we follow [8] and [10].

Definition 2.17 *Let X be an abstract simplicial complex and i a dimension. A **i-chain** is a sum of i-simplices in X. We write $c = \sum a_j\sigma_j$, where the a_j s are coefficients and the σ_j s are i-simplices.*

Definition 2.18 *The **group of i-chains** $C_i(X)$ is the i-chains together with the addition operation defined through $c + c' = \sum(a_j + a'_j)\sigma_j$.*

Definition 2.19 *For a simplex σ spanned by $\{v_0, v_1, \ldots, v_i\}$ the **boundary** is defined by*

$$\partial_i(\sigma) = \sum_{j=0}^{i}(-1)^j\sigma|_{[v_0,\ldots,\hat{v_j},\ldots,v_i]}.$$

We see that the boundary takes a i-chain to a $i-1$-chain and can therefore write $\partial_i \colon C_i \to C_{i-1}$. Notice that $\partial_i(c + c') = \partial_i(c) + \partial_i(c')$ which is the defining property of a homomorphism. We call ∂_i therefore **boundary map**.

Definition 2.20 *The **chain complex** is the sequence of chain groups connected by boundary maps*

$$\cdots \xrightarrow{\partial_{i+2}} C_{i+1} \xrightarrow{\partial_{i+1}} C_i \xrightarrow{\partial_i} C_{i-1} \xrightarrow{\partial_{i-1}} \cdots \tag{2.1}$$

Definition 2.21 *A **i-cycle** is a i-chain c such that $\partial c = 0$. The group of i-cycles is denoted by Z_i.*

Notice that $Z_i = Ker(\partial_i)$.

Definition 2.22 *A **i-boundary** is a i-chain c such that $c = \partial d$ for $d \in C_{i+1}$. The group of i-boundaries is denoted by B_i.*

Notice that $B_i = Im(\partial_{i+1})$.

Definition 2.23 *The **i-th homology group** is $H_i = Z_i / B_i$.*

Definition 2.24 *The **i-th Betti number** β_i is the dimension of the i-th homology group, $\beta_i = dim(H_i)$.*

Definition 2.25 *A topological space X is called **acyclic** if it is connected and $H_i(X) = 0$ for all $i \geq 1$.*

Example 2.26 A circle is not acyclic, because $H_1(S^1) = \mathbb{Z} \neq 0$. But a disk, for example, is acyclic.

Definition 2.27 *The **boundary matrix** ∂_i is defined to be the matrix, that has a one-to-one correspondence between the columns and $\Sigma_i(X)$ and one between the rows and $\Sigma_{i-1}(X)$. So if you look at the entry $e_{\tau',\tau}$ in the column corresponding to some i-simplex τ and the row corresponding to some $(i-1)$-simplex τ', we have that*

$$e_{\tau',\tau} = \begin{cases} 1, & \tau' \subseteq \tau \text{ (as set of vertices)} \\ 0, & otherwise. \end{cases} \tag{2.2}$$

Example 2.28 We take the same abstract simplicial complex as in Example 2.16. The corresponding boundary matrices ∂_1 and ∂_2 are the following:

$$\partial_1 = \begin{array}{c} \\ \{1\} \\ \{2\} \\ \{3\} \\ \{4\} \end{array} \begin{array}{c} \begin{array}{ccccc} \{1,2\} & \{2,3\} & \{3,4\} & \{1,4\} & \{2,4\} \end{array} \\ \begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \end{array}, \quad \partial_2 = \begin{array}{c} \\ \{1,2\} \\ \{2,3\} \\ \{3,4\} \\ \{1,4\} \\ \{2,4\} \end{array} \begin{array}{c} \begin{array}{c} \{2,3,4\} \end{array} \\ \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} \end{array}$$

An important result for boundary matrices is the following proposition.

Proposition 2.29 *The matrix product $\partial_{i-1} \cdot \partial_i$ is equal to the zero matrix over $\mathbb{Z}_2$.*

Example 2.30 Again we take the same complex as in Example 2.16 and compute the matrix product $\partial_1 \cdot \partial_2$ with the matrices from Example 2.28:

$$\begin{pmatrix} 1 & 0 & 0 & 1 & 0 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix} \cdot \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 2 \\ 2 \\ 2 \end{pmatrix}, \text{ which is equal to } \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix} \text{ over } \mathbb{Z}_2.$$

Definition 2.31 *For a finite metric space $M = (M, d)$, the **Vietoris-Rips complexes** $V(M, r)$ of M are defined as follows. M is the vertex set of $V(M, r)$ and a k-tuple $\{m_0, m_1, \ldots, m_k\}$ spans a k-simplex of $V(M, r)$ if and only if $d(m_i, m_j) \leq r$ for all $0 \leq i, j \leq r$.*

Example 2.32 In Figure 2.6 we see three Vietoris-Rips complexes on the same finite metric space with three points and varying radii r_i for $i \in \{1, 2, 3\}$.

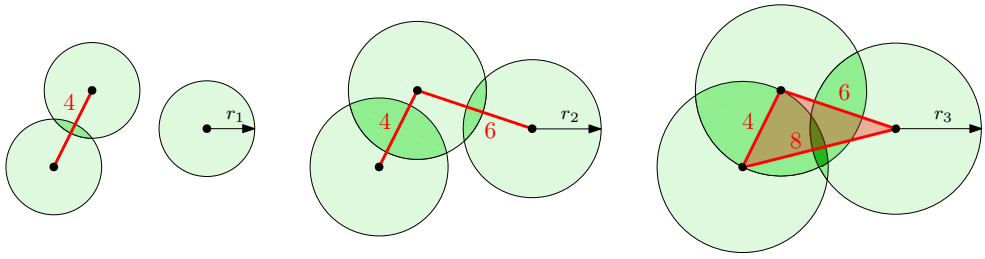


Figure 2.6: Example of three Vietoris-Rips complexes with $2 \leq r_1 < 3 \leq r_2 < 4 \leq r_3$.

2.3 Persistent vector spaces

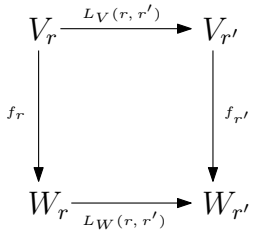
Definition 2.33 *Let k be a field. A **persistence vector space** over k is a family of k-vector spaces $\{V_r\}_{r \in \mathbb{R}}$ together with linear transformations $L_V(r, r') \colon V_r \to V_{r'}$ for $r \leq r'$, such that $L_V(r', r'') \cdot L_V(r, r') = L_V(r, r'')$ for all $r \leq r' \leq r''$.*

Definition 2.34 *A **sub-persistent vector space** of $\{V_r\}_r$ is $\{U_r\}_r$ with all $U_r \subseteq V_r$ for $r \in [0, \infty)$, such that $L_V(r, r')(U_r) \subseteq U_{r'}$ for all $r \leq r'$.*

Definition 2.35 *A **linear transformation f** between persistence vector spaces $\{V_r\}$ and $\{W_r\}$ is a family of linear transformations $f_r \colon V_r \to W_r$, such that for all $r \leq r'$ it holds that*

$$f_{r'} \circ L_V(r, r') = L_W(r, r') \circ f_r. \tag{2.3}$$

This means in particular that the following diagram commutes:



Definition 2.36 *The **quotient space** of a persistence vector space $\{V_r\}$ and one of its sub-persistent vector spaces $\{U_r\}$, where $\{U_r\} \subseteq \{V_r\}$, is defined as the pair $\{V_r/U_r\}$ together with the linear transformation $L_{V/U}(r, r') :: V_r/U_r \to V_{r'}/U_{r'}$, given by $[v] \mapsto [L_V(r, r')(v)]$ for all $v \in V_r$.*

Definition 2.37 *Let X be a set and $\rho\colon X \to [0, \infty)$ a function. Then (X, ρ) is called an $\mathbb{R}_+$**-filtered set**.*

Definition 2.38 *The **free persistence vector space on the pair** (X, ρ), is the vector space $\{V_k(X, \rho)_r\}$ with $V_k(X, \rho)_r \subseteq V_k(X)$, which is equal to the k-linear span of the set $X[r] = \{x \in X \mid \rho(x) \leq r\} \subseteq X$. A persistence vector space $\{V_r\}$ is called **free** if there exists an isomorphism, such that $\{V_r\} \cong \{V_k(X, \rho)_r\}$ for some pair (X, ρ).*

Lemma 2.39 *A linear combination of the form $\sum_x a_x x \in V_k(X)$ lies in $V_k(X, \rho)_r$ if and only if $a_x = 0$ for all x, such that $\rho(x) > r$. [4]*

Example 2.40 We will observe the filtration F in Figure 2.7 of simplicial complexes. We can see that all the vertices appear at time 0. At time 1 the edge $\{a, b\}$ appears and so on. So our persistence vector space of 0-chains is

$$(C_0(F))_r = \langle a, b, c \rangle \text{ for } r \in [0, \infty). \tag{2.4}$$

Similarly for the 1-chains we get

$$(C_1(F))_r = \begin{cases} 0 & \text{for } r \in [0, 1) \\ \langle \{a, b\} \rangle & \text{for } r \in [1, 2) \\ \langle \{a, c\} \rangle & \text{for } r \in [2, 3) \\ \langle \{b, c\} \rangle & \text{for } r \in [3, \infty) \end{cases} \tag{2.5}$$

and for the 2-chains

$$(C_2(F))_r = \begin{cases} 0 & \text{for } r \in [0, 4) \\ \langle \{a, b, c\} \rangle & \text{for } r \in [4, \infty) \end{cases} \tag{2.6}$$

We can define persistence linear maps from $\{(C_i(F))_\rho\}$ to $\{(C_{i-1}(F))_\rho\}$ a family

$$(\partial_i)_\rho\colon (C_i(F))_\rho \to (C_{i-1}(F))_\rho \tag{2.7}$$

of boundary maps, where $\rho \in [0, \infty)$.

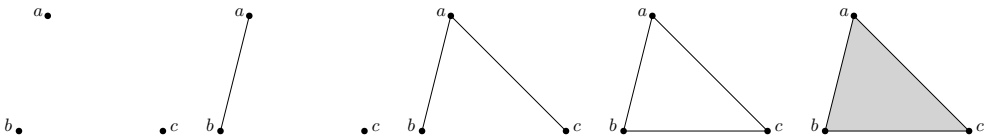


Figure 2.7: A Filtration of the chain complex on the right.

Definition 2.41 *A persistence vector space as in Definition 2.38 is **finitely generated**, if X can be taken to be finite.*

Definition 2.42 *A persistence vector space $\{V_r\}$ over k is called **finitely presented**, if there exists an isomorphism, such that $\{V_r\} \cong \{W_r\}/Im(f)$ for a linear transformation f between the two finitely generated free persistence vector spaces $\{V_r\}$ and $\{W_r\}$. Note that $\{W_r\}/Im(f)$ is also a persistence vector space.*

Definition 2.43 *For two finite sets X, Y and a field k, the (X,Y)-**matrix** is given by an array $[a_{xy}]$ where the a_{xy}'s are elements in k. For two $\mathbb{R}_+$-filtered sets (X,ρ) and (Y,σ), a (X,Y)-matrix $A(f)$ is called (ρ,σ)-**adapted** if*

$$a_{xy} = 0 \text{ for } \rho(x) > \sigma(y). \tag{2.8}$$

For a finite set X we can see that for a finitely generated free persistence vector space $\{V_r\} = \{V_k(X,\rho)_r\}$ and big enough r we have that $V_k(X,\rho) = V_k(X)$. Therefore we can conclude that any persistent linear map between finitely generated free persistence vector spaces $f\colon \{V_k(Y,\sigma)_r\} \to \{V_k(X,\rho)_r\}$ gives a linear map $f_\infty\colon V_k(Y) \to V_k(X)$ between finite dimensional vector spaces over k, where we have a label for every generator. By using the bases $\{\varphi_x\}$ of $V_k(X)$ and $\{\varphi_y\}$ of $V_k(Y)$ we can determine an (X,Y)-matrix $A(f) = [a_{xy}]$ with entries in k.

Lemma 2.44 *Suppose we have two $\mathbb{R}_+$-filtered sets (X,ρ) and (Y,σ). Then all matrices $A = [a_{xy}]$ satisfying 2.8 determine a persistence vector space in the following way:*

$$\theta\colon A \to V_k(X,\rho)/Im(f_A), \tag{2.9}$$

where the linear transformation $f_A\colon \{V_k(Y,\sigma)_r\} \to \{V_k(X,\rho)_r\}$ between persistence vector spaces is uniquely determined by the (X,Y)-matrix A.

Remark 2.45 For A as above, $\theta(A)$ is finitely presented.

Example 2.46 We now want to write out the boundary matrices from Example 2.40.

$$(\partial_1)_\infty = \begin{array}{c} \\ (a,0) \\ (b,0) \\ (c,0) \end{array} \begin{array}{c} \begin{array}{ccc} (ab,1) & (ac,2) & (bc,3) \end{array} \\ \begin{pmatrix} 1 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \end{pmatrix} \end{array}$$

$$(\partial_2)_\infty = \begin{array}{c} \\ (ab,1) \\ (ac,2) \\ (bc,3) \end{array} \begin{array}{c} \begin{array}{c} ((a,b,c),4) \end{array} \\ \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} \end{array}$$

At the left of $(\partial_1)_\infty$ we see that $\rho(x)$ is 0 everywhere so the condition $\rho(x) > \sigma(y)$ is never given. Therefore the matrix is (ρ,σ)-adapted. Similar for $(\partial_2)_\infty$ we have that $1,2,3 < 4$ and this matrix is also (ρ,σ)-adapted.

Definition 2.47 *An **interval persistent vector space** $P(a,b)$ for $a \in \mathbb{R}_+$, $b \in \mathbb{R}_+ \cup \{\infty\}$ and $a < b$ is defined by*

$$P(a,b)_r = \begin{cases} k, & r \in [a,b) \\ 0, & r \notin [a,b) \end{cases} \tag{2.10}$$

and where $L(r,r') = id_k$ for $r,r' \in [a,b)$.

Remark 2.48 This definition is obvious to interpret in the case where $b = \infty$. Note that $P(a,b)$ is finitely presented. In the case where $b < \infty$ we have the following: Let (X,ρ) and (Y,σ) be two $\mathbb{R}_+$-filtered sets and $X = \{x\}$, $Y = \{y\}$. Furthermore we have $\rho(x) = a$ and $\sigma(y) = b$. Then it follows that the (1×1)-matrix (1) is (ρ,σ)-adapted and $P(a,b) \cong \theta((1))$. When $b = \infty$ we get the isomorphism $P(a,b) \cong V_k(X,\rho)$ and we can write $P(a,b) = \theta(0)$, where 0 is the zero linear transformation from the persistence vector space 0.

Proposition 2.49 *The matrix $A(f)$ satisfies 2.8 and (X,Y)-matrix A satisfying these conditions uniquely determines a linear transformation between the two persistence vector spaces $\{V_k(Y,\sigma)_r\}$ and $\{V_k(X,\rho)_r\}$, which we will denote by f_A. Moreover the linear correspondence $f \to A(f)$ and the matrix correspondence $A \to f_A$ are inverses to one another.*

2.4 The Decomposition Theorem

Theorem 2.50 (Decomposition Theorem) *Every finitely presented persistence vector space $\{V_r\}$ over k is isomorphic to a finite direct sum of the form*

$$\{V_r\} \cong \bigoplus_{i \in I} P(a_i, b_i), \tag{2.11}$$

where I is a finite set, $a_i \in [0, \infty)$, $b_i \in [0, \infty]$ and $a_i < b_i$ for all i. Moreover, this decomposition is unique in the sense that the collection of pairs $\{(a_i, b_i)\}_i$ is unique up to the ordering of the factors.

We will not proof this theorem here, but we will state all the propositions that are needed for the proof. A complete proof is written in [4]. We already used a result from the following proposition in Lemma 2.44. Now we want to state and proof it to derive another proposition that is important for the Decomposition Theorem.

Proposition 2.51 *For an $\mathbb{R}_+$-filtered set (X, ρ) the isomorphisms*

$$V_k(X, \rho) \to V_k(X, \rho)$$

are identified with the group of invertible (ρ, ρ)-adapted (X, X)-matrices under the correspondences stated in Proposition 2.44.

Proposition 2.52 *Take $\mathbb{R}_+$-filtered sets (X, ρ), (Y, σ) and a (ρ, σ)-adapted (X, Y)-matrix A. Let now B be a (ρ, ρ)-adapted (X, X)-matrix and C be a (σ, σ)-adapted (Y, Y)-matrix. Then BAC is (ρ, σ)-adapted and $\theta(A) \cong \theta(BAC)$, for θ as in 2.9.*

This proposition allows us to define adapted row- and column operations and we want to give an intuition to from where they come. In an (X, Y)-matrix we write $r(X)$ for the row corresponding to $x \in X$ and $c(y)$ for the column corresponding to $y \in Y$. Let now A be a (ρ, σ)-adapted (X, Y)-matrix for two $\mathbb{R}_+$-filtered sets (X, ρ) and (Y, σ). We can now define *adapted row and column operations* as follows:

- Additions of a multiple of $r(X)$ to $r(x')$ whenever $\rho(x) \geq \rho(x')$;
- Additions of a multiple of $c(y)$ to $c(y')$ whenever $\sigma(y) \leq \sigma(y')$;
- Multiplications of $0 \neq a \in k$ to a row or a column, where k is the field.

These operations are important for the proof of the first section of Theorem 2.50. With all these propositions it is possible to prove the Decomposition Theorem. In particular, the proof gives us an algorithm to compute persistent homology using the adapted row and column operations. We give an example below. Before we start with the example, we want to look at the allowed operations on the set of two matrices in this setting. We need a pair of

matrices because for computing homology we always need two boundary matrices. We name the left matrix in the set of two matrices A and the right one B. We will use the matrix on the left to find $Ker(A)$ and the one on the right to find $Im(B)$. To get the kernel and the image in consistant bases, we perform operations on both matrices simultaneously.

- Any adapted row operation on the left matrix A;
- Any adapted column operation on the right matrix B;
- A combination of an adapted column operation on A and an adapted row operation on B in one of the following pairs of operations:
  - Multiplication of the i-th column with $0 \neq a \in k$ and multiplication of the i-th row of B with a^{-1}.
  - Transposition of the i-th column with the j-th column in A and transposition of the i-th row with the j-th row in B.
  - Addition of a times the i-th column to the j-th column in A and substraction of a times the j-th row from the i-th row in B.

Each isomorphism class of finitely presented persistence vector spaces has a direct correspondence with a subset of $\{(a,b)|a \in [0,\infty), b \in [0,\infty], a < b\}$. We have two ways of interpreting this visually:

- As a collection of intervals $\{[x,y]|x \geq 0, y > x\}$ in the upper right quadrant of $\mathbb{R}^2$; $\mathbb{R}_+ = \{(x,y)|x \geq 0, y \geq 0\}$, called a *barcode*,
- As a collection of points in $\{(x,y)|x \geq 0, y > x\}$ as a subset of the plane, called a *persistence diagram*.

So we can now replace the concept of Betti numbers as the dimension of the homology groups in singular homology by the barcodes for persistent homology.

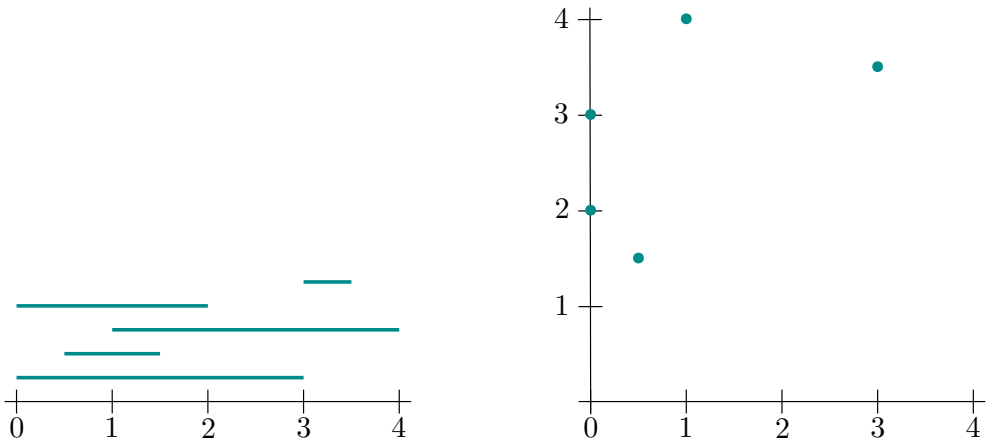
Example 2.53 Recall from Example 2.40 that

$$(C_0(F))_r = \langle a,b,c\rangle \text{ for } r \in [0,\infty). \tag{2.12}$$

$$(C_1(F))_r = \begin{cases} 0 & \text{for } r \in [0,1) \\ \langle\{a,b\}\rangle & \text{for } r \in [1,2) \\ \langle\{a,c\}\rangle & \text{for } r \in [2,3) \\ \langle\{b,c\}\rangle & \text{for } r \in [3,\infty) \end{cases} \tag{2.13}$$

$$(C_2(F))_r = \begin{cases} 0 & \text{for } r \in [0,4) \\ \langle\{a,b,c\}\rangle & \text{for } r \in [4,\infty) \end{cases} \tag{2.14}$$

We now want to compute the 1-dimensional persistent homology group. So we write down the pair of adapted matrices and perform adapted row and



a) barcode *b*) persistence diagram

Figure 2.8: We can represent a persistence vector space as barcodes where each bar represents the lifetime a hole or cavity or as persistence diagram where the x-coordinates are the birthtime of a cavity and the y-coordinates is the deathtime.

column operations on it. Remember that we are in $\mathbb{Z}_2$.

$$\left(\begin{array}{cc} & \begin{array}{ccc}(ab,1) & (ac,2) & (bc,3)\end{array} \\ \begin{array}{c}(0)\\(0)\\(0)\end{array} & \begin{pmatrix}1 & 1 & 0\\ 1 & 0 & 1\\ 0 & 1 & 1\end{pmatrix} \end{array} \quad \begin{array}{c}(ab,1)\\(ac,2)\\(bc,3)\end{array} \begin{pmatrix}1\\1\\1\end{pmatrix} \right) \rightarrow$$

Adding the first row to the second row in the left matrix yields:

$$\left(\begin{array}{cc} & \begin{array}{ccc}(ab,1) & (ac,2) & (bc,3)\end{array} \\ \begin{array}{c}(0)\\(0)\\(0)\end{array} & \begin{pmatrix}1 & 1 & 0\\ 0 & 1 & 1\\ 0 & 1 & 1\end{pmatrix} \end{array} \quad \begin{array}{c}(ab,1)\\(ac,2)\\(bc,3)\end{array} \begin{pmatrix}1\\1\\1\end{pmatrix} \right) \rightarrow$$

Now we add the first column to the second column in the left matrix and we

add the second row to the first row in the right matrix.

$$\begin{pmatrix} & (ab,1) & (ab+ac,2) & (bc,3) & & (abc,4) \\ (0) & 1 & 0 & 0 & (ab,1) & 0 \\ (0) & 0 & 1 & 1 & , (ac,2) & 1 \\ (0) & 0 & 1 & 1 & (bc,3) & 1 \end{pmatrix} \rightarrow$$

We now add the second row to the third row in the left matrix.

$$\begin{pmatrix} & (ab,1) & (ab+ac,2) & (bc,3) & & (abc,4) \\ (0) & 1 & 0 & 0 & (ab,1) & 0 \\ (0) & 0 & 1 & 1 & , (ac,2) & 1 \\ (0) & 0 & 0 & 0 & (bc,3) & 1 \end{pmatrix} \rightarrow$$

Finally we add the second column to the third column in the left matrix and we add the third row to the second row in the right matrix.

$$\begin{pmatrix} & (ab,1) & (ab+ac,2) & (ab+ac+bc,3) & & (abc,4) \\ (0) & 1 & 0 & 0 & (ab,1) & 0 \\ (0) & 0 & 1 & 0 & , (ac,2) & 0 \\ (0) & 0 & 0 & 0 & (bc,3) & 1 \end{pmatrix}$$

We can see from the left matrix that $Ker(\partial_1)$ is isomorphic to (X, ρ), where $X = \{\partial(abc)\}$ and $\rho(\partial(abc)) = 3$. We now look at the persistence linear map $(\partial_2)_r \colon (C_2(F))_r \rightarrow (Ker(\partial_1))_r$, $r \in [0, \infty)$. We get the following adapted matrix:

$$\begin{pmatrix} & (abc,4) \\ (\partial(abc),3) & 1 \end{pmatrix}$$

So the persistent homology group in dimension 1 is isomorphic to $P(3,4)$. In the resulting adapted matrix, the appearence time of $\partial(abc)$ is 3. this is the first enty in $P(a,b)$. The second one is the appearence time of abc which is 4.

We want to study the degree to which the barcode changes when we have small changes in the data. To formalize the understanding of small changes, we will now define three metrics we will need to state the theorem below. For these definitions and the proof of the following theorem we will follow [6].

Definition 2.54 *For two compact subsets X, Y of a metric space (Z, d_Z) the **Hausdorff distance** is defined as follows:*

$$d_H^Z(X,Y) = \max\{\max_{x\in X}\min_{y\in Y} d_Z(x,y), \max_{y\in Y}\min_{x\in X} d_Z(x,y)\}. \tag{2.15}$$

The Gromov-Hausdorff distance can be defined in different ways but we will need the following to state Theorem 2.57.

Definition 2.55 *The **Gromov-Hausdorff distance** between two compact metric spaces (X, d_X) and (Y, d_Y) is defined as:*

$$d_{GH}((X,d_X),(Y,d_Y)) = \inf_{Z,\gamma_X,\gamma_Y}\{d_H^Z(\gamma_X(X),\gamma_Y(Y))\}, \tag{2.16}$$

where γ_X and γ_Y are isometric embeddings from X, Y into the space (Z, d_Z).

Definition 2.56 *The **Bottleneck distance** between two diagrams $A, B \in (\overline{\mathbb{R}}^2, \ell^\infty)$ is defined as:*

$$d_B^\infty(A,B) = \min_{\gamma}\max_{a\in A} \parallel a - \gamma(a) \parallel_\infty, \tag{2.17}$$

where γ goes over all bijections from A to B.

Let $\mathcal{M}^{fin}$ be the family of finite metric spaces and $\mathcal{B}$ the family of persistence barcodes. For $k \geq 0$ an integer, we have an assignment between finite metric spaces and persistence barcodes $H_k \colon \mathcal{M}^{fin} \to \mathcal{B}$. The following theorem, known as the stability theorem, guarantees that the Bottleneck distance between the homology groups of two spaces is smaller or equal to the Gromov-Hausdorff distance between the underlying spaces.

Theorem 2.57 (Stability Theorem) *For two compact metric spaces X, Y, we have*

$$d_B(H_k(X), H_k(Y)) \leq d_{GH}(X,Y), \tag{2.18}$$

which means that each of the assignments is non-increasing.

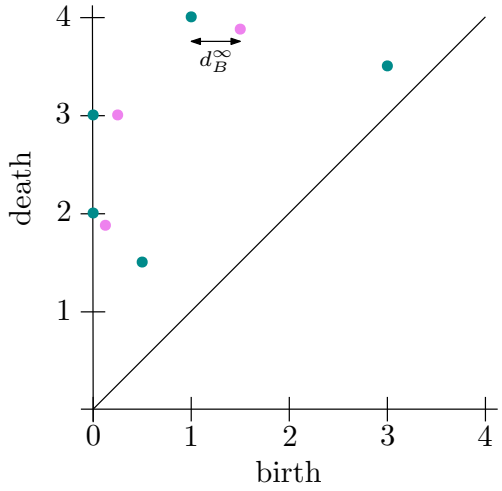


Figure 2.9: The bottleneck distance between two persistence diagrams [2].

Chapter 3

Tree-like spaces

In this chapter we will look at trees, tree-like finite metric spaces and their homology to derive an important result discussed in 3.2. It states that for all radii $r \geq 0$, the complex $V(M, r)$ is a disjoint union of acyclic complexes. We need these structures because evolution is often modelled with trees. In general, we are following the Appendix of our main paper [5]. Some other references are [3] for the theorem over the isometric embedding of a finite metric space into the metric space of vertices of an additive tree and [9] for the four-point condition.

3.1 Definitions

Definition 3.1 *A **simple cycle** in a graph is a cycle where every edge and every vertex appears only once.*

Definition 3.2 *A **tree** T is a finite connected one-dimensional simplicial complex which has no simple cycles. An **additive tree** (T, λ_T) is a tree T equipped with a (real valued) weight function λ_T on the edges, which maps to the positive real numbers.*

Example 3.3 In Figure 3.1a) we see a simple cycle. b) and c) do not contain any cycles. In b) we see a tree and we get the additive tree in c) from b) by placing values on the edges.

Definition 3.4 *The set of **leaves** of a graph Γ is the set of vertices which occur in exactly one edge.*

Definition 3.5 *An **edge path** in a tree T is a sequence with k vertices $\{v_0, \ldots, v_k\}$ such that for every two v_i, v_{i+1} it holds that (v_i, v_{i+1}) is an edge in T. For an additive tree, the **length of an edge path** is the sum over all weights of the edges that are in the edge path. The **edge path distance** between two vertices in an additive tree is the length of the shortest edge path that connects the two vertices.*

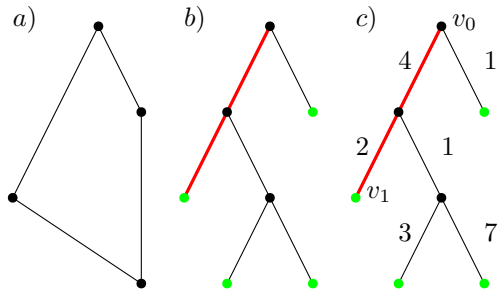


Figure 3.1: a) A simple cycle, b) a tree, c) an additive tree, where the green vertices are leaves and the red edges denote an edge path.

Example 3.6 The red edges in Figure 3.1b) and 3.1c) are edge paths with 3 vertices. The length of the edge path in 3.1c) is $4+2=6$, which is also the edge path distance between v_0 and v_1.

Note that not all finite metric spaces come from additive trees, but we name those finite metric spaces that do come from additive trees **tree-like**. We get a metric space from an additive tree in the following way.

Definition 3.7 *Let the **distance between two vertices v and w** be the minimum* $\min\{\sum_i \lambda(v_i, v_{i+1})\}$ *over the set of all edge paths* $\{(v_0, v_1), (v_1, v_2), \ldots, (v_{n-1}, v_n)\}$ *where $v_0 = v$ and $v_n = w$. Since different edge paths have different lengths the n can vary over the set of all edge paths. The metric obtaind from this distance is the **associated edge path metric** and will be denoted by d_λ. The corresponding metric space will be denoted by $\mathfrak{M}(\Gamma) = (V(T), d_\lambda)$ for some additive tree $T = (T, \lambda_T)$.*

We want to use persistent homology to develop a measure to see how close any finite metric space is to a tree-like space.

Lemma 3.8 *If Γ is a tree, then d_Γ satisfies the four point condition*

$$d_\Gamma(x,y) + d_\Gamma(z,t) \leq \max\{d_\Gamma(x,z) + d_\Gamma(y,t), d_\Gamma(x,t) + d_\Gamma(y,z)\}, \tag{3.1}$$

where x, y, z and t are vertices in Γ.

To prove this Lemma we need Proposition 0.21 and Corollary 0.20 from [9]:

Proposition 3.9 (Proposition 0.21) *For G a tree and x, y, z and t are vertices in G, let $v_0, v_1, \ldots, v_g$ be path from x to y. Let i, j be elements in $\{0, 1, \ldots, g\}$ such that $d(z, v_i)$ and $d(t, v_j)$ are minimal.*

$$\text{If } i \leq j \text{ then } d(x,t) + d(y,z) \geq d(x,y) + d(z,t). \tag{3.2}$$

$$\text{If } i \geq j \text{ then } d(x,z) + d(y,t) \geq d(x,y) + d(z,t). \tag{3.3}$$

Corollary 3.10 (Corollary 0.20) *For G a tree and x, y, z and t are vertices in G, let $v_0, v_1, \ldots, v_g$ be path from x to y. Let i be the element in $\{0, 1, \ldots, g\}$ such that $d(z, v_i)$ is minimal and let $h = d(z, v_i)$. Then we have:*

- $d(x,z) = i + h$,
- $d(y,z) = g - i + h$,
- *If* $y \in \{0,1,\ldots,g\}$ *is such that* $i \leq j$, *then* $d(z,v_j) = j - i + h$.

In the proof of Lemma 3.8 we will write d instead of d_Γ for simplicity.

Proof (Lemma 3.8) We suppose that the converse is true, i.e.

$$d(x,y) + d(z,t) > \max\{d(x,z) + d(y,t), d(x,t) + d(y,z)\}. \tag{3.4}$$

It follows directly that

$$d(x,y) + d(z,t) > \max\{d(x,z) + d(y,t), d(x,t) + d(y,z)\} \geq d(x,z) + d(y,t) \tag{3.5}$$

and

$$d(x,y) + d(z,t) > \max\{d(x,z) + d(y,t), d(x,t) + d(y,z)\} \geq d(x,t) + d(y,z). \tag{3.6}$$

Since Γ is a tree, it is connected and there exists at least one path from x to y. We fix one of these paths and denote it as $v_0, v_1, \ldots, v_g$, where $v_0 = x$ and $v_g = y$. We now take the indices $i, j \in \{0,1,\ldots,g\}$, such that the distances $d(z,v_i)$ and $d(t,v_j)$ are minimized. We have to check two cases:
$\underline{i \leq j}$: With Proposition 3.9 and Corollary 3.10 from [9] we know that

$$d(x,t) + d(y,z) \geq d(x,y) + d(z,t), \tag{3.7}$$

which is a contradiction to the inequality in 3.6.
$\underline{i \geq j}$: Again with the Proposition and Corollary mentioned above, we know that

$$d(x,z) + d(y,t) \geq d(x,y) + d(z,t), \tag{3.8}$$

which is a contradiction to the inequality in 3.5.
Since we get a contradiction in both cases, our first assumption 3.4 was false and we are done. □

Example 3.11 In Figure 3.2 we see an example of a tree where we have an equality in the four point condition 3.1 and one example, where the four point condition 3.1 is strict. All the edges have length 1.

Theorem 3.12 *Any finite metric space that satisfies the four point condition can be embedded isometrically in the metric space of vertices on an additive tree. [3]*

Definition 3.13 *A finite metric space is called* ***tree-like*** *if it satisfies the four point condition 3.1.*

Remark 3.14 An equivalent definition of a **tree-like** finite metric space is that it can be embedded isometrically in $\mathfrak{M}(T)$ for some additive tree $T = (T, \lambda_T)$.

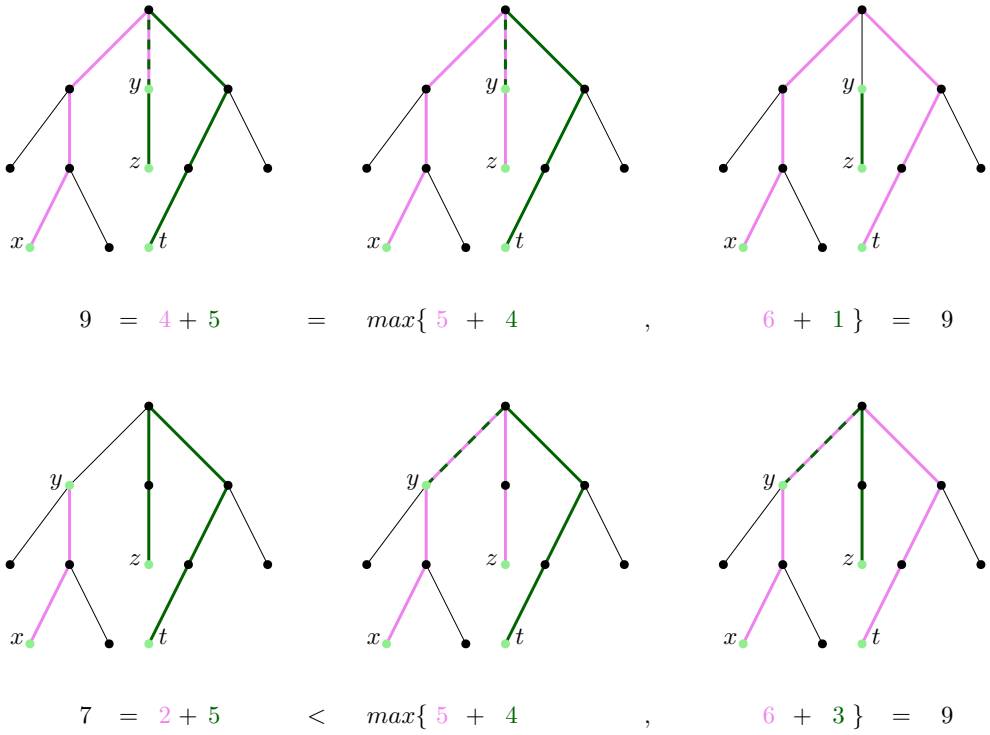


Figure 3.2: Examples for the four point condition.

Remark 3.15 For $r \leq r'$ there exists a natural inclusion $\iota\colon V(M,r) \hookrightarrow V(M,r')$ where ι is the identity on the vertex set.

Definition 3.16 *An* ***r-path*** *from x to x' in a finite metric space X is a sequence of elements $x_0, x_1, \ldots, x_s \in X$ such that $x_0 = x$, $x_s = x'$ and $d(x_i, x_{i+1}) \leq r$ for $i = 0, 1, \ldots, s-1$. A finite metric space is called* ***r-connected*** *if every pair of points in X can be connected by an r-path.*

Note that an r-path is the correspondence to an edge path from x to x' in $V(X,r)$.

Definition 3.17 *A vertex is called* ***linear****, if it is contained in exactly two edges. A vertex is called a* ***junction****, if it is not linear.*

In the case where there is at least one junction in a tree, there exists a unique nearest junction $j(v)$ for every leaf v. In the case without any junctions in the tree, the graph is a line by definition of a junction.

Definition 3.18 *For a tree T and a pair (v,e) of a vertex v in the tree and an edge e containing v, we define the* ***branch of*** *T* ***through*** *v* ***and*** *e to be the subtree containing all vertices v', such that the unique shortest edge path from v' to v contains e. We will denote this by $Br(e,v)$.*

Example 3.19 We take a tree as in Figure 3.3. The red path is a 4-path, but the tree is not 4-connected, because the distance between the green leave v and its neighbour is $7 > 4$. But since 7 is the biggest distance between two vertices, the tree is 7-connected. All the blue vertices are linear and all the green ones are junctions.

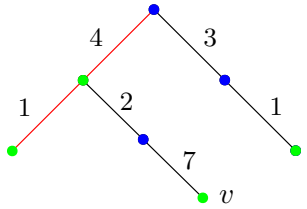


Figure 3.3: An additive tree with red 4-path, blue linear vertices and green junctions.

3.2 Homology of tree-like finite metric spaces

The goal of this section is to show that $H_i(V(M,r)) = 0$ for all $i \geq 1$ for a tree-like metric space M. This means that for all $r \geq 0$, the complex $V(M,r)$ is a disjoint union of acyclic complexes.

Lemma 3.20 *For an additive tree T the tree-like finite metric space M can be isometrically embedded in $\mathfrak{M}(T)$, such that all the leaves of T are included in the image of M, contained in the vertex set of T.*

Proof Let T be an additive tree, and $i\colon M \to V(T)$ be the isometric embedding in the vertex set of T, which exists by Theorem 3.12. Let x be any leaf which is not contained in the image of i, i.e. $x \in V(T)\backslash i(M)$. Then M is included in $T\backslash\{x\}$ and any minimal edge path from m_0 to m_1, for $m_0, m_1 \in M$, will not include x. This is shown by the fact that if (y,x) is the unique edge containing x, then any occurrence of x in an edge path in T is of the form $\ldots(y,x)(x,y)\ldots$ and can be deleted. So we see that M embeds isometrically in $\mathfrak{M}(T\backslash\{x\},\lambda)$. Now we can continue eliminating leaves and the edges attached to them from the tree until we get to the setting where all leaves are contained in the image of M. □

The following proposition describes the relationship between the edge path metrics on additive trees and the edge path metrics on additive subtrees.

Proposition 3.21 *Let T be an additive tree and suppose that $T' \subseteq T$ is a subadditive tree which we get by removing a set of leaves $\{x_1,\ldots,x_n\}$. Then we have the equality*

$$d_{T'} = d_T\big|_{V(T')\times V(T')}. \tag{3.9}$$

So $\mathfrak{M}(T')$ is a metric subspace of $\mathfrak{M}(T)$ in the sense that the distance function on $\mathfrak{M}(T')$ is just the restriction of the distance function on $\mathfrak{M}(T)$.

Proof We first prove the case where one leaf is removed and then complete the proof by induction. Let x be the one removed leaf, such that the attached edge is also removed and v_0, v_1 be any vertices in T'. Let y be the unique vertex of T such that (x, y) is an edge in T. All the edge paths between v_0 and v_1 containing x are of the form $\dots(y, x)(x, y)\dots$ and can be shortened by removing $(y, x)(x, y)$. So the edge path of minimal length from v_0 to v_1 does not contain x. Now suppose that we already proved the case for n removed leaves. Let T' be the tree with $n+1$ leaves $\{x_1, \dots, x_{n+1}\}$ removed. Then we know that for all edge paths from v_0 to v_1, not containing x_{n+1} the proposition holds. Consider an edge path containing x_{n+1}. Then we can prove the case exactly as in the case $n = 1$. □

We need two more propositions to prove our goal. The fist one allows us to perform an induction on the number of points in a tree-like space. The second one is an important homological result.

Proposition 3.22 *Let M be a finite metric space and $m_0, m_1 \in M$ two points such that $d(m_0, m_1) \geq d(m, m')$ for any $m, m' \in M$. Let $M_0 = M \backslash \{m_0\}$ and $M_1 = M \backslash \{m_1\}$, regarded as metric subspaces. Then for all $r < d(m_0, m_1)$, we have that*

$$V(M, r) = V(M_0, r) \cup V(M_1, r) \tag{3.10}$$

and $V(M_0, r) \cap V(M_1, r) = V(M_0 \cap M_1, r)$ since $M_0 \cap M_1 = M \backslash \{m_0, m_1\}$.

Proof By definition of the Vietoris-Rips complexes it holds for $r < d(m_0, m_1)$ that m_0 and m_1 do not span a 2-simplex of $V(M, r)$. So no simplex of $V(M, r)$ contains both m_0 and m_1 and $V(M, r)$ can be covered by $V(M_0, r)$ and $V(M_1, r)$. □

Example 3.23 (Proposition 3.22) We look at the finite metric space with three points, where the two points m_0 and m_1 have maximal distance from each other. The proposition only makes a statement for $r < d(m_0, m_1)$, so Figure 3.4 only shows these cases.

Proposition 3.24 *Let X be a simplicial complex with two subcomplexes U and V such that $X = U \cup V$. If all three U, V and $U \cap V$ are acyclic, then X is also acyclic.*

Proof To prove this we write out the reduced Mayer-Vietoris sequence [10, p. 149] for this scenario:

$$\dots \to H_i(U \cap V) \to H_i(U) \oplus H_i(V) \to H_i(X) \to H_{i-1}(U \cap V) \to \dots \tag{3.11}$$

Since U, V and $U \cap V$ are acyclic their homology groups vanish for $i \geq 1$. We get

$$\dots \to 0 \to 0 \to H_i(X) \to 0 \to 0 \to \dots \tag{3.12}$$

and therefore $H_i(X)$ also vanishes for $i \geq 1$. □

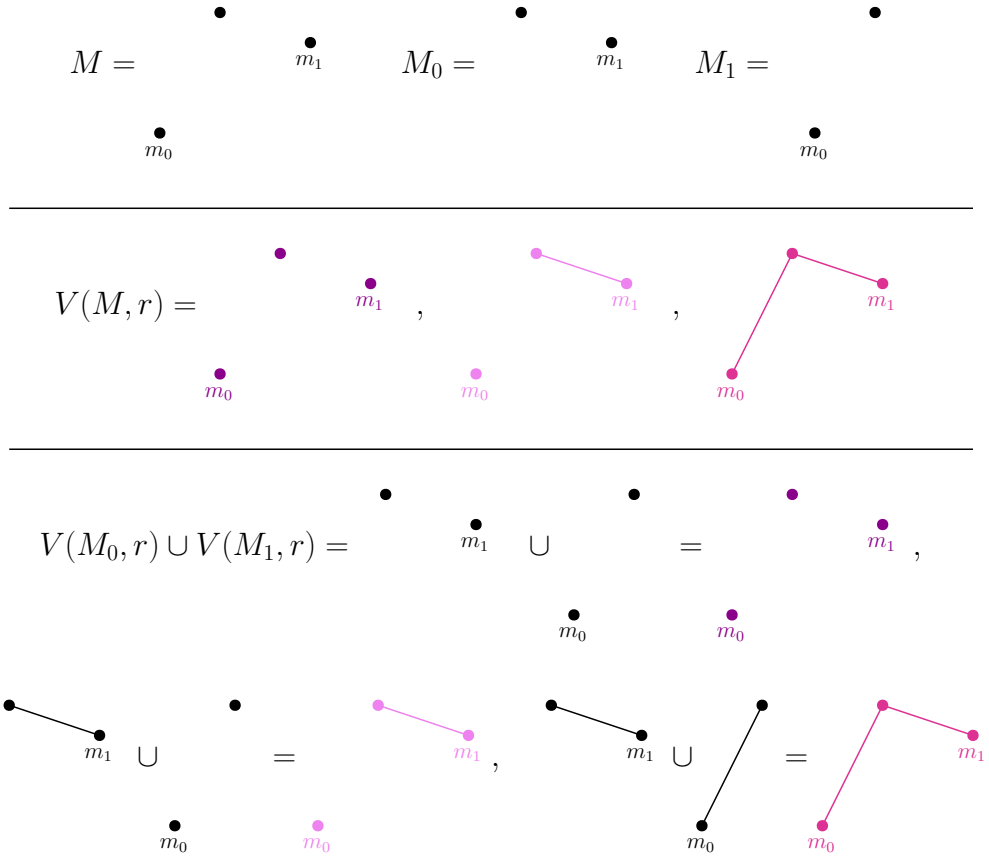


Figure 3.4: Example for Proposition 3.22.

Example 3.25 We want to illustrate the Proposition 3.24 with the example of a simplicial complex X with three vertices and two edges, that looks like in Figure 3.5. Clearly, A, B and $A \cap B$ are acyclic, because they are all homotopy

Figure 3.5: A simplicial complex X with subcomplexes A, B and $A \cap B$.

equivalent to the one-point space and so is X.

Now we get to our main theorem from this section.

Theorem 3.26 *Let M be a tree-like finite metric space and let $r \geq 0$. Then the complex $V(M,r)$ is a disjoint union of acyclic complexes. In particular, $H_i(V(M,r)) = 0$ for $i \geq 1$.*

Before we start with the proof of this theorem, we look at an example.

Example 3.27 In Figure 3.6a) we see a tree and our M will be the tree-like finite metric space coming from this additive tree. In 3.6b) we see the Vietoris-Rips complex $V(M,r)$ for $r = 1.5$ of M. In 3.6c) we chose two subcomplexes A and B. They are acyclic and we see that $V(M,1.5))$ is the disjoint union of these two acyclic complexes. In particular, we have that $H_i(V(M,1.5)) = 0$ for $i \geq 1$.

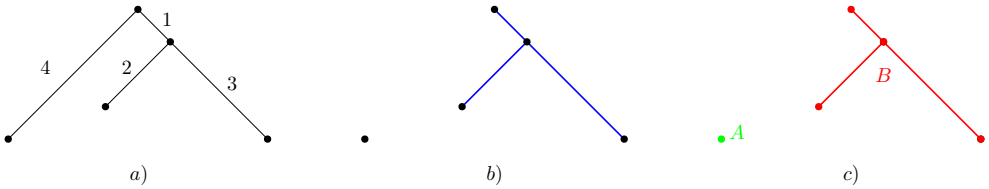


Figure 3.6: a) An additive tree T, b) Vietoris-Rips complex of the metric space obtained from T, c) disjoint subsets A and B of b).

Proof (of Theorem 3.26) We will show this theorem by induction on the cardinality of M. Let $\simeq_r$ be the equivalence relation that is defined by $m \sim_r m'$ if and only if $d(m,m') \leq r$. So in each equivalence class M_α all the elements have distance smaller equal r to each other. For equivalence classes M_α we get that our metric space $M = \bigsqcup_\alpha M_\alpha$ under $\simeq_r$, and it is clear that

$$M = \bigsqcup_\alpha V(M_\alpha, d_M\big|_{M_\alpha \times M_\alpha}, r). \tag{3.13}$$

This follows from the definition of Vietoris-Rips complexes in the following way. Look at the $V(M_\alpha, d_M\big|_{M_\alpha \times M_\alpha}, r)$ s. We know from the definition that each k-tuple of elements in M_α spans a k-simplex, because we defined $m \sim_r m'$ if and only if $d(m,m') \leq r$. The union of all these gives us back our space M. Since M was tree-like, each of the metric spaces $(M_\alpha, d_M\big|_{M_\alpha \times M_\alpha})$ is also tree-like. For the induction step we suppose that the theorem holds for all metric spaces with cardinality $< n$. We can assume that $V(M,r)$ is connected, since otherwise it would be a disjoint union of acyclic complexes by the induction hypothesis. According to Lemma 3.20 we can suppose that M is isometrically embedded as a metric subspace of $\mathfrak{M}(T)$, where all the leaves are included in the image of M.
Now we select the two points in M which have maximal distance from each other and call them m_0 and m_1. Suppose that m_0 and m_1 are not leaves. Now take any leaf that has minimal distance to m_0 and name it l_0. Do the same for m_1 and name it l_1. These leaves can not be contained in the maximal path between m_0 and m_1. So we get can construct a path from l_0 to l_1 containing the maximal path between m_0 and m_1. the resulting path is clearly longer than the original which is a contradiction to our assumption. We can therefore assume that m_0 and m_1 are leaves. Looking at the results

from Proposition 3.22 we see that it suffices to prove that the three spaces $M\backslash\{m_0\}$, $M\backslash\{m_1\}$ and $M\backslash\{m_0, m_1\}$ have acyclic Vietoris-Rips complexes. By induction on the cardinality and using Proposition 3.21 it suffices to show that these metric spaces are connected.
We have seen that for a tree-like space with cardinality n we may assume in the induction over n that M is r-connected. Now it suffices to show that $M\backslash\{m_0\}$, $M\backslash\{m_1\}$ and $M\backslash\{m_0, m_1\}$ are r-connected.
To prove the theorem for the case where there are no junctions in the tree, we look at m_0 and m_1. These are clearly the two leaves of the line-graph, we will denote as L. With our assumption from above that L is r-connected, we know that $d(m_0, m_1) \leq r$. Name the vertex that is connected to m_0, m_0' and the one connected to m_1, m_1'. Then it is obvious that the distance between m_0' and m_1' is also less or equal than r. Thus $L\backslash\{m_0\}$, $L\backslash\{m_1\}$ and $L\backslash\{m_0, m_1\}$ are r-connected, which shows the theorem for the line-case.
Now we can assume that there exists a junction in our graph. Now consider the set of nodes $\{v_0, \ldots, v_n\}$ such that $\{j(m_0), v_i\}$ is an edge in T for every i, and construct all branches $\mathfrak{B}_i = Br(j(m_0), \{j(m_0), v_i\})$. The $\mathfrak{B}_i$'s cover all of T and we may assume that $m_i \in \mathfrak{B}_i$ for $i = 0, 1$. Notice that $\mathfrak{B}_0$ is a linear tree that starts in $j(m_0)$ and ends in m_0.
We now assume for a contradiction that $d(v, j(m_0)) > d(m_0, j(m_0))$ for a vertex $v \in \mathfrak{B}_i$ and $i \geq 2$. This implies $d(m_1, v) = d(m_1, j(m_0)) + d(j(m_0), v) > d(m_1, j(m_0)) + d(j(m_0), m_0) = d(m_1, m_0)$. But this contradicts the maximality of $d(m_0, m_1)$ and we can conclude that

$$d(v, j(m_0)) \leq d(m_0, j(m_0)). \tag{3.14}$$

For $v \in \mathfrak{B}_i$ ($i \geq 2$) and $w \in \mathfrak{B}_i$ ($i \geq 1$) it follows that

$$d(w, m_0) \geq d(w, v). \tag{3.15}$$

We now prove the connectivity of $V(M\backslash\{m_0\}, r)$. Suppose that we have $m, m' \in M\backslash\{m_0\}$. Because M is r-connected there exists an r-path $v_0, v_1, \ldots, v_k$ from m to m' in M. Recall from the definition of an r-path that $v_0 = m$, $v_k = m'$ and $d(v_l, v_{l+1}) \leq r$ for $l = 0, 1, \ldots, k-1$. In the case where $v_l \neq m_0$ for all l, the r-path lies completely in $M\backslash\{m_0\}$. So we can assume that $v_l = m_0$ for some v_l. We get $d(v_{l-1}, m_0) \leq r$ and $d(m_0, v_{l+1}) \leq r$. We have to check the following two cases:
$M \cap (\mathfrak{B}_0\backslash\{m_0\})$ is non-empty: Let $\bar{m}$ be the point in $M \cap (\mathfrak{B}_0\backslash\{m_0\})$ which is the nearest to m_0. Now we can replace the segment $v_{l-1}m_0v_{l+1}$ in the r-path by the segment $v_{l-1}\bar{m}v_{l+1}$. This gives us an r-path from m to m' without m_0 in it.
$M \cap (\mathfrak{B}_0\backslash\{m_0\})$ is empty: For $i \geq 2$ we select a leaf m^* in $\mathfrak{B}_i$. Since v_{l-1} and v_{l+1} are elements of $\mathfrak{B}_i$ for some $i \geq 1$ we can conclude from 3.15 that we can replace $v_{l-1}m_0v_{l+1}$ with $v_{l-1}m^*v_{l+1}$ to get an r-path from m to m' without m_0 in it.

We have now replaced our r-path with one that has a smaller count of occurrences of m_0 and proceeding this way, we get an r-path lying completely in $M\backslash\{m_0\}$, which shows that $V(M\backslash\{m_0\}, r)$ is connected. The same result for $M\backslash\{m_1\}$ follows identically. Since m_0 is never replaced by m_1 and m_1 never by m_0, the result follows also for $M\backslash\{m_0, m_1\}$, because we can remove the occurrences of m_0 and m_1 independently. □

Chapter 4

Application to viral evolution

Equipped with all the theory about tree-like spaces and persistent homology we can now discuss their application to detect horizontal events in viral evolution as in *Topology of viral evolution* [5]. Another book we will look at for this section is called *Topological Data Analysis for Genomics and Evolution* [12]. For the biological definitions we refer to [14, Chapter 6], [7, Chapter 1 and 2] and [13].

4.1 Structures to describe viral evolution

Charles Darwin first introduced the idea of modelling evolution of phenotypic attributes with phylogenetic trees in his book *On the Origin of the Species* in 1859. **Phylogenetics** is the study of evolutionary relationships by inferring or estimating the evolutionary past among biological entities, that can be for example organisms, species or genes. **Phylogeny**, on the other hand, is the evolution of a genetically related group of organisms via the study of protein or gene evolution by involving the comparison of homologous sequences. Phylogeny is often modelled by a tree. A **phylogenetic tree** is a kind of molecular archaeology that tries to reconstruct possible evolutionary relationships by extrapolating backward from a small dataset from surviving organisms. So it is a tree that explains how different species evolved over time, how they are related with each other and how they came to their present form. Species or individuals that share specific derived characters are grouped more closely together than those who do not. The groups are called **clades**; each clade consists of an ancestor and all of its descendants.
While phylogenetic trees are great to capture clonal or vertical evolution, they cannot capture any reticulate or horizontal evolutionary events. These **reticulate** events occur whenever different clades merge together and produce a new lineage. Figure 4.2 shows the difference of a phylogenetic tree with only clonal evolution (Figure 4.2a)) and a reticulate network that captures

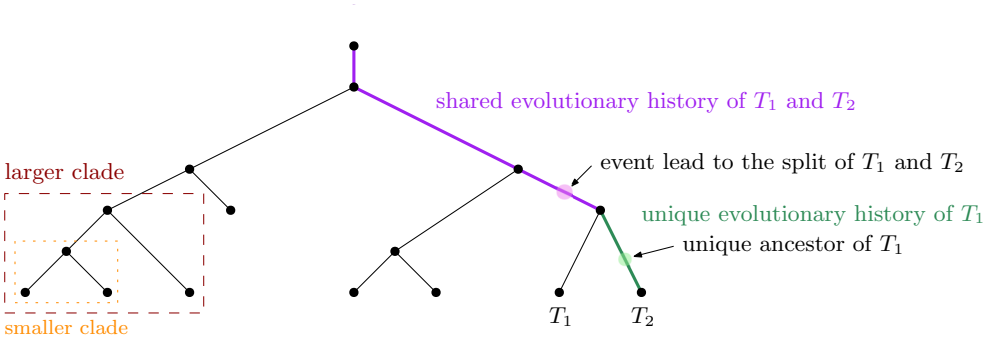


Figure 4.1: Example of a phylogenetic tree [7].

horizontal events (Figure 4.2b)).

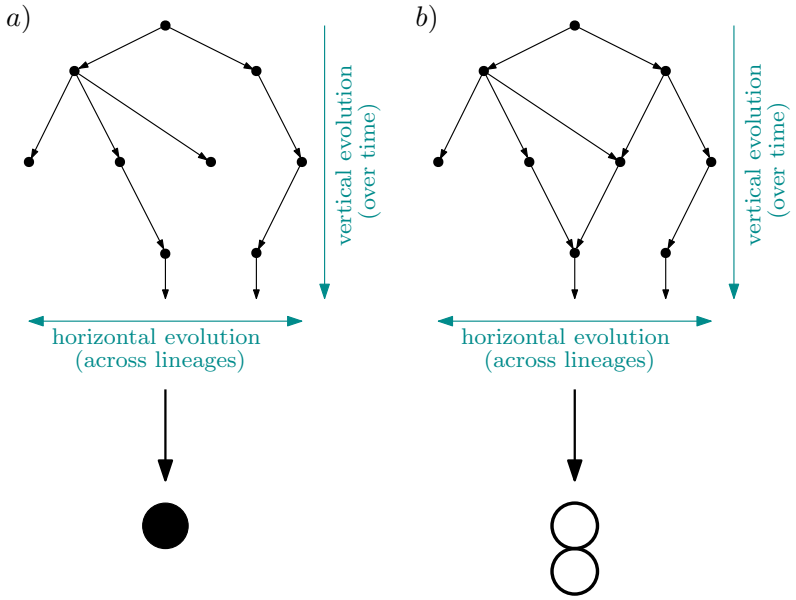


Figure 4.2: The clonal tree in a) gets compressed to a point, whereas the network in b) has holes and therefore can be compressed to $S^1 \vee S^1$.

For viruses these reticulate merges are better known as recombination and reassortment. **Genetic recombination** is the process where a DNA segment moves from one DNA molecule to another DNA molecule. **Reassortment** is the process where different viruses exchange genomic material in the form of gene segments. Nowadays there are different methods to detect reticulate events in evolution. The phylogenetic methods search for discrepancies in the tree structure of phylogenetic trees. Nonphylogenetic methods look for shared character traits that occur independently in different lineages, called homoplasies. Although these methods detect the reticulate events in viral

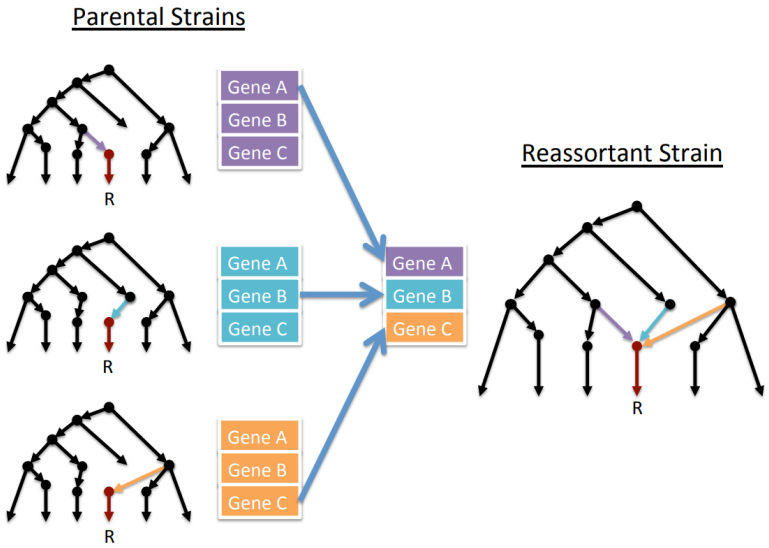


Figure 4.3: Reticulate structure representing the reassortment of three parental strains. The reassortment stain results from merging the three phylogenetic trees at the left [5].

evolution, they do not provide a simple way to represent the evolutionary processes. Phylogenetic networks may be the best way to illustrate reticulate events in evolution. But we have the problem that all implementations we have to this point only give one phylogenetic network structure even though the structure is not unique.
So we need a structure to study the patterns of evolutionary processes that captures vertical events as well as horizontal events. We already introduced the main ideas of persistent homology and this is exactly the structure we will use. The big advantage of persistent homology over trees or networks is that it captures all possible topologies and their relations over the space of genetic distance. Not only can persistent homology detect horizontal evolution between two individuals but also more complex exchanges of genomic data and statistical patterns of cosegregation. **Cosegregation** is the name for the circumstance that two or more genes are more likely to be passed to the next generation as a set. **Phenotypic** means concerning the phenotype, which are all observable traits of an individual. Exemplary for a phenotype are eye color, height and weight. When discussing evolution we have to look at the component of DNA and RNA. DNA and RNA are constructed out of four **nucleotides** each. The four nucleotides for DNA have the bases **adenin** (A), **cytosine** (C), **guanine** (G) and **thymine** (T). And for RNA the thymine is replaced by **uracil** (C).

4.2 Persistent homology in evolution

Definition 4.1 *The **hamming distance** is a distance measure between two genetic sequences of the same length and is defined by the number of positions in the two nucleotide strings at which their corresponding characters are different.*

Example 4.2 The hamming distance for letters between the words "cake" and "take" is 1 and the hamming distance between "fly" and "try" is 2.

In Figure 4.2 we see an example of a tree deforming to a point and a reticulate network deforming to the wedge product of two circles. As the figure suggests, the presence of holes in the compressed structure corresponds to the reticulate events. So to study the reticulate events we want to compute the number of holes and higher dimensional cavities in a evolutionary network. In the language of persistent homology this means that we want to find irreducible cycles, which are cycles in dimension k, that are not a boundary of a $(k+1)$-dimensional simplex. Our correspondence is now the following; the homology group H_k contains all the holes in dimension k and the Betti number b_k counts the holes. As mentioned before the persistent analogue of Betti numbers are barcodes. To link biological expressions to persistent homology terms we provide a table of correspondences as in [5], see 4.1.

Persistent homology	Viral evolution
Filtration value ρ	Genetic distance
0-dimensional Betti number at filtration value ρ	Number of clusters at scale ρ
Generators of 0-dimensional Betti number homology	A representative element of the cluster
Hierarchical relationship among generators of 0-dimensional Betti number homology	Hierarchical clustering
1-dimensional Betti number	Number of reticulate events
Generators of 1-dimensional homology	Reticulate events
Generators of 2-dimensional homology	Complex reticulate genomic exchanges
Non-zero higher dimensional homology	No phylogenetic representation
Number of higher dimensional generators over time (irreducible cycle rate)	Lower bound on rate of reticulate events

Table 4.1: Dictionary between persistent homology and evolutionary concepts.

We saw the definition of a barcode earlier and want to give the genetic intuition behind it. For different values of the genetic distance ρ we get different simplicial complexes and therefore also different irreducible cycles. So if we have an irreducible cycle C that is present over a filtration, we denote the "birth" of C a_C and its "death" b_C. we have therefore an interval

$[a_C, b_C]$ over which C "lives". When we then compute the homology groups of dimension k at all scales ρ, we get a barcode plot as in 2.8a).
So we take a set of genomes and calculate the genomic distance between each pair of points and the corresponding distance matrix. Then we can construct a phylogenetic tree out of the matrix, if it exists. To construct the tree we use the neighbor-joining method. It starts with a star-like tree. Then, a pair of sequences is chosen at random, gets removed from the star, and attached to a second internal node which is connected by a branch to the center of the star-like pattern. The branch lengths are calculated. These two sequences are then returned to their original positions and another pair is selected to repeat the same procedure. The goal is to check all possible pairs to find out the combination of neighbors that minimizes the total length of the phylogenetic tree [7], page 213. We will give an example of this process in Figure 4.4. We can compute the homology groups for different ρ's in different dimensions. We can now assume that 0-dimensional homology gives us information over vertical evolution in the sense that we can for example compute the number of distinct subclades in our dataset. 1-dimensional homology provides information over reticulate events, because these structures contain loops, see 4.2b). The distance metric we use is the Hamming distance, which we will define as in [15] adapted to our genomic environment.

Example 4.3 We want to look at an example where four genetic sequences yield a distance matrix and finally a phylogenetic tree, see Figure 4.4.

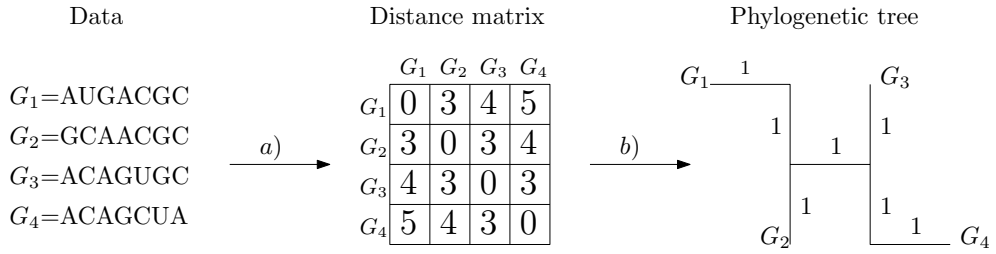


Figure 4.4: In a) we apply the hamming distance to get a distance matrix. In b) we construct a phylogenetic tree out of the distance matrix.

We mentioned above that we can construct a phylogenetic tree from the distance matrix, if one exists. In the case where there is no such tree we can use topology to represent the structure behind evolutionary processes. In Example 4.5 we will see how persistent homology can detect recombination and reassortment events in evolution.

4.3 Topological obstruction to phylogeny

We want to formalize the role of phylogeny in our framework of persistent homology. First notice, that a tree can not contain holes and that it is therefore not possible to place a phylogenetic structure on a simplicial complex coming from genomic data with distance parameter ρ that contains higher dimensional irreducible cycles. That is what we proved in Theorem 3.26. Given some genomic data and the fact that there is a Betti number $b_i \neq 0$ for some $i > 0$ we know that there is a deviation from a tree metric. To measure this deviation, we define the *topological obstruction to phylogeny* or shorter TOP. To define this we take a distribution B_K on the lengths of the persistent bars in a barcode for k-dimensional cycles and $k > 0$.

Definition 4.4 ***TOP*** *is the L^∞-norm, or maximum of the lengths of the bars in the barcode.*

For a filtration with non-zero TOP there exists no additive tree that can represent the underlying data. In a set of data one always has some kind of statistical noise, sequencing errors or incomplete samplings. But TOP is stable and bounded by the Gromov-Hausdorff distance to the additive tree according to Theorem 2.57. Since higher dimensional homology vanishes in additive trees, small deviations generate only small bars in the barcode. We want to give an example to illustrate how irreducible cycles correspond with reticulate events.

Example 4.5 In Figure 4.5 with graphics and data from [5] we see how persistent homology can detect reticulate events in evolution. In a) we have the reticulate network, where the yellow dots are the genetic sequence we look at as subspace of a bigger sample, whose pairwise genetic distances are illustrated in b). All these yellow samples have a common ancestor through clonal evolution (turquoise lines) or reticulate evolution (dotted red lines). We apply persistent homology to the whole sample and we see three different filtrations in c). Computing the barcode yields us d) and we see that the red bar around 0.15 corresponds to the red cycle of our five yellow data points. It is a recombination event involving our selected sequences. The length of the bar corresponding to this reticulate event is $\rho = [0.13, 0.16]$ Hamming distance. So we can conclude that this is the genetic distance between the parents of the recombinant.

We now want to consider the L^0-norm instead of the L^∞-norm. It is simply the count of the number of the higher dimensional bars and it is proportional to the horizontal evolution rate r. To approximate r more closely we define the follwing.

Definition 4.6 *The **irreducible cycle rate** (ICR) is the total number of 1-dimensional bars for all values ρ per a specific time frame.*

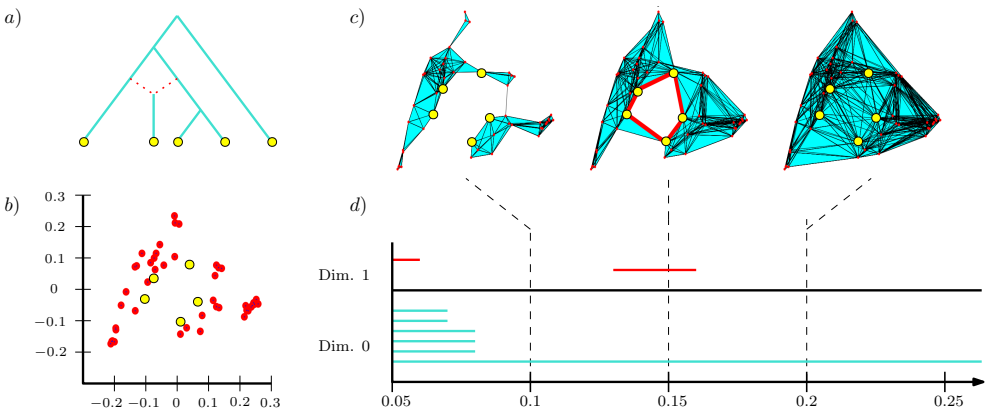


Figure 4.5: a) A reticulate network, b) dataset modelled by their genetic distance, c) three different filtrations of the dataset, d) barcode plot of the dataset [12].

Simulations have shown that ICR is proportional to the recombination/ reassortment rate and moreover yields a lower bound for it.
Since, as we mentioned earlier, persistent homology can give us information on the obstruction to tree-like metrics caused through recombination, reassortment, homoplasies or other horizontal genomic exchanges, we can study the cycles that generate higher dimensional homology to conclude what horizontal event precludes the tree-like structure. To specify the type of these horizontal evolutionary events one can look at the size of the barcodes in non-zero homology. So when looking at the generators of different bars, one can distinguish between two underlying reassortment processes. The mixing of closely related viruses generates smaller bars while the mixing of more distant viruses, not from the same subtype, generates longer barcodes.

4.4 Testing the detection of simulated reticulate events

Before we can look at some examples of genomic data of viruses, we have to check that persistent homology is an appropriate method to study complex processes in evolution. Therefore it was tested on simulated reticulate events in [5] by modelling a population undergoing clonal evolution for 30 generations with random reassortment events at time step 15. Then persistent homology was applied to this model. In the simulation were four different scenarios: clonal evolution, population admixture, reassortment and homologous recombination. The constant populations were observed over several generations under a Wright-Fisher model defined as in [11]. We use this specific model because the Wright-Fisher model ignores mutations and recombinations to study distributions of alleles in a population. We first need the definition of an allele.

Definition 4.7 *An **allele** is a specific molecular variant at a locus. A **locus** is the position in the genome where an allele resides.*

Definition 4.8 *The **Wright-Fisher model** describes a constant population where the generations are discrete and non-overlapping. For each generation $g = n$ an allele is picked from a parent allele from generation $g = n - 1$. This process is random. For a population of size N, the probability that an allele that is present in i individuals at generation $g = n$ is present in j individuals at generation $g = n + 1$ is*

$$P_{ij} = \binom{N}{j} \left(\frac{i}{N}\right)^j \left(1 - \frac{i}{N}\right)^{N-j}. \tag{4.1}$$

We now introduce some parameters: the substitution rate μ, the recombination/ reassortment rate r and the number of reassorting segments S. For the test with simulated reticulate events, evolution was considered over a range for the values μ and r. Simulations have shown three interesting results:

- If $r \neq 0$ we have nontrivial homology,
- 1-dimensional ICR increases proportionally to r,
- Complex recombination/ reassortment events can cause 2-dimensional topology.

4.5 Viral evolution in different viruses

4.5.1 Influenza A

We now want to have a closer look at the influenza virus. Influenza is a segmented single-stranded RNA in the family of orthomyxoviruses. It is a virus that often reassorts and can be found in many different hosts. These hosts are mostly birds but also humans, seals and pigs. In [5] they applied persistent homology to influenza datasets from humans, birds and pigs. Application of persistent homology on a single viral segment with no reassortment history shows that higher dimensional homology groups vanish as we would suppose. In this scenario the 0-dimensional homology generates trees. In Figure 4.6 we see the reconstruction of a phylogeny from persistent homology of avian influenza HA.

Remark 4.9 (Figure 4.6) For a given Hamming distance ρ each bar incorporates a connected simplex of sequences. If a bar ends at a given ρ it merges with another one which is shown nicely in b). If two simplices of the same subtype merge at ρ the corresponding bar is grey. For two simplices of different subtype but same major clade merging together, the bars have only one color. If two simplices of different major clade merge together, we have a color gradient in the bars.

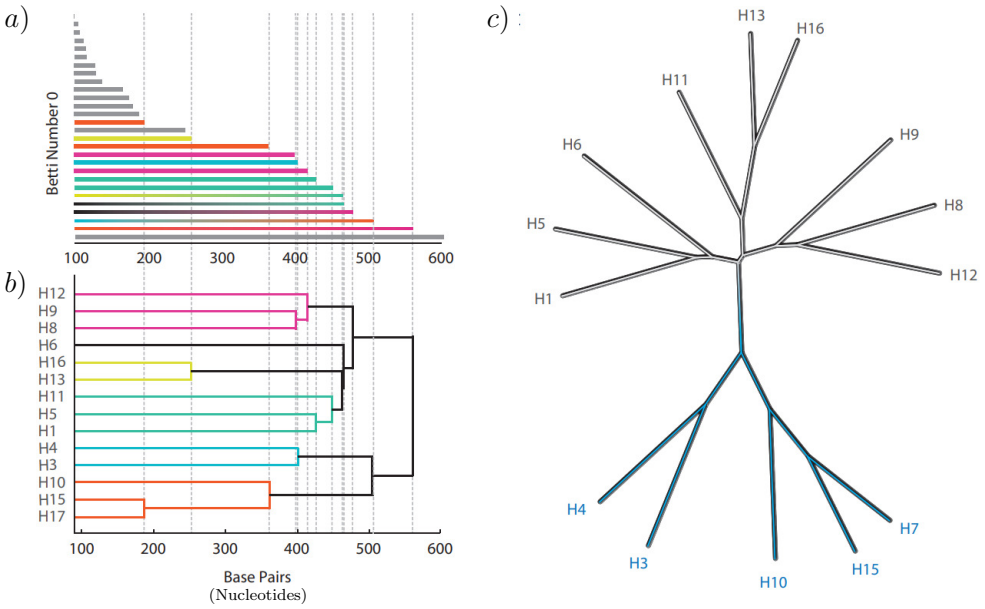


Figure 4.6: a) Barcode plot of all avian HA subtypes in dimension 0, where the only significant homology occurs. b) Phylogeny of avian HA reconstructed from the barcode plot in a) showing clustering. c) Neighbour-joining tree of avian HA showing the relation between HA subtypes (from [5] and [12]).

If we look at several concatenated genes at once, higher dimensional homology classes appear, which indicate reticulate events in the evolution. An example for this circumstance is given below. Influenza viruses are classified through the antigentic properties of the proteins in the viral envelope hemagglutinin (HA) and neuraminidase (NA). The properties range from H1 to H16 for HA and from N1 to N9 for NA.

Example 4.10 In Figure 4.7a) we see the persistence analysis of HA in avian influenza. In b) it is the same but now for NA. In c) we see the analysis for several concatenated segments and that there are non-zero higher dimensional homology groups, that indicate reassortment.

For avian influenza the work of [5] calculated a rate of 22.16 reassortments per year, whereas this rate is smaller than one for swine and human influenza. Therefore the example of avian influenza is the most interesting to look at. It was shown earlier that avian influenza has a high reassortment rate, but it was not possible to study the patterns of the associated gene segments. With persistent homology it is now possible to check whether there are gene segments that cosegrate more than expected. In [5] it was found that the following four gene segments of avian influenza cosegrate: polymerase basic 2 (PB2), polymerase basic 1, polymerase acidic (PA) and nucleoprotein (NP), visualized in Figure 4.8.

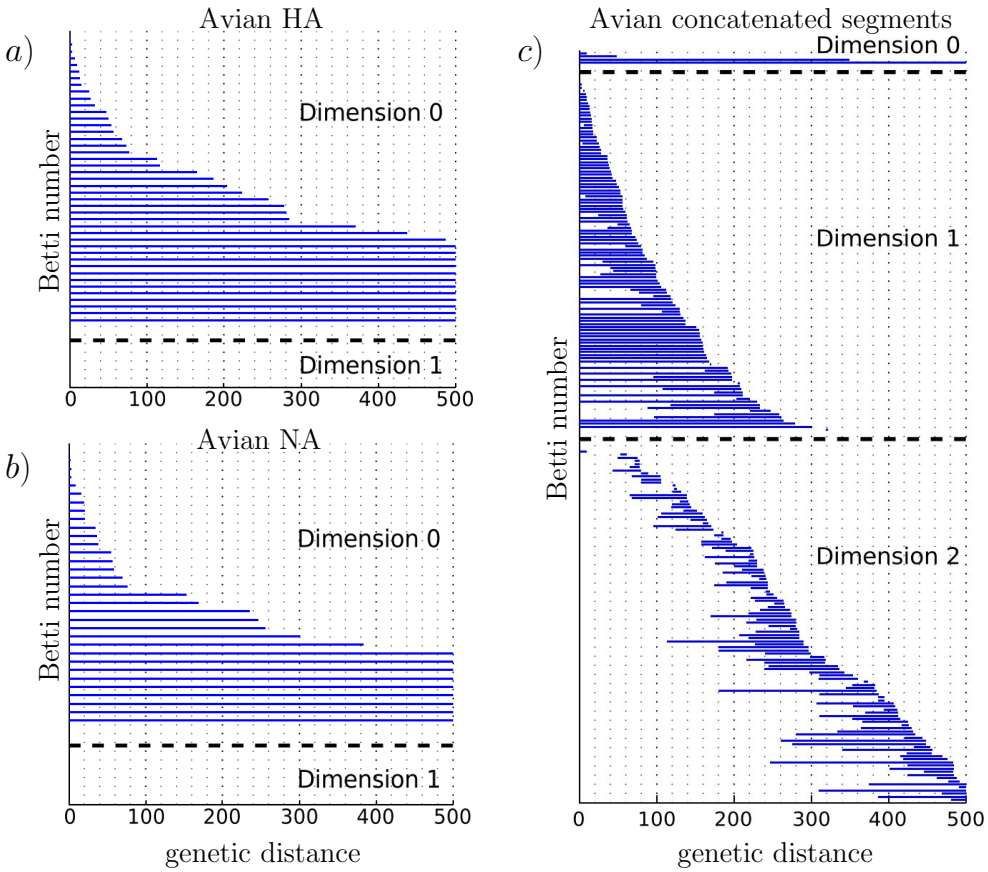


Figure 4.7: Persistence analysis of HA (a)), NA (b)) and several concatenated segments (c)) in avian influenza [5].

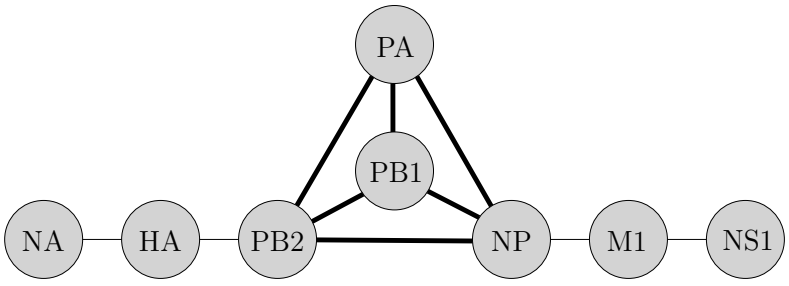


Figure 4.8: Cosegregation of avian influenza segments. Thicker lines indicate a higher probability for two segments to reassort together [5].

4.5.2 HIV

The Human Immunodeficiency Virus, or HIV is famous for its high diversity that comes not only through a high mutation rate but also through frequent recombination leading for example to antiretroviral resistance. It is also famous because more than one million people die every year as a consequence to infection with HIV. HIV is a *retrovirus*. This type of virus has a positive single-stranded RNA. When infecting a cell, a retrovirus releases its RNA into the cytoplasm of the cell, where it generates a DNA strand out of the RNA and implements this strand into the cell's DNA. In this way the virus gets access to the cell's genomic machinery and can therefore replicate the viral genome. The process of converting RNA to DNA is called *reversed transcription*. The three largest genes in the genome of the HIV-1 retrovirus are:

- *gag*: Codes for the proteins generating the shell of a viral particle.
- *pol*: Carries all the information about the enzymes that are needed for replication and reverse transcription, integrating the viral DNA into the genome of the host and cleaving viral polyproteins to activate them.
- *env*: Codes for the glycoproteins that bind to the T-cell's receptors and it enables the virus to get into the host cell.

Because the genome of HIV is not segmented there can not be a reassortment and all the horizontal evolution is caused by recombination. Figure 4.9 is an illustration of a recombination of two HI viruses. In [5] persistent

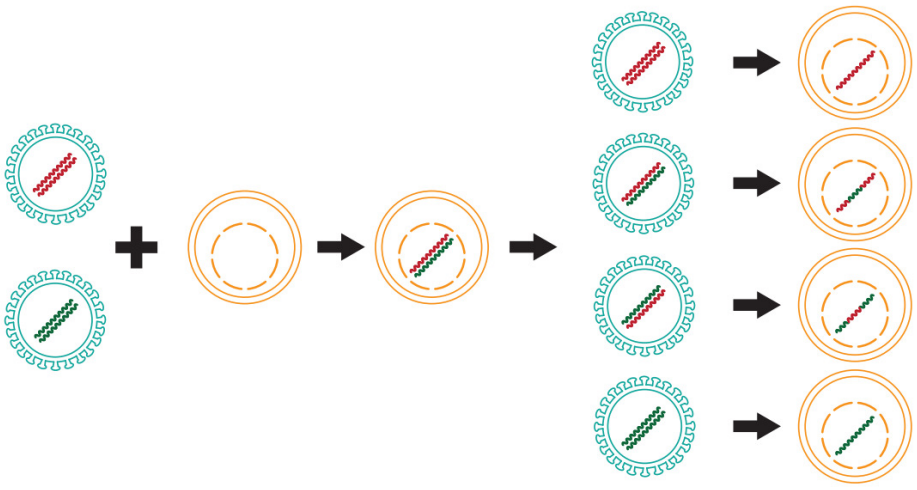


Figure 4.9: Different recombination of two HI viruses [12].

homology was applied to the concatenated alignments of the three largest genes mentioned above. It produced 1-dimensional homology, but so it did

for the individual genes, see Figure 4.10. This indicates a horizontal event, so recombination in the evolution of these genes as well as between the individual genes. Persistent homology applied to the concatenation also produced 2-dimensional homology, hence we have a complex recombination.

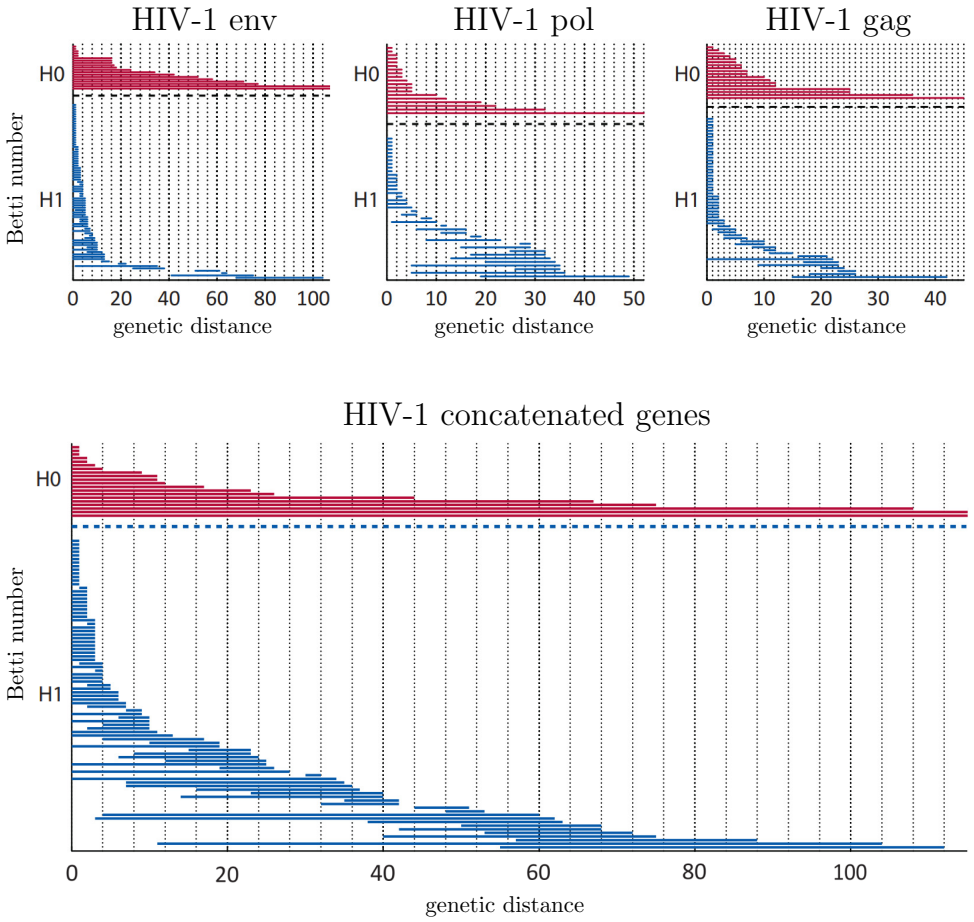


Figure 4.10: Persistent homology suggests recombination within as well as between the genes [12].

4.5.3 Flaviviruses

Flaviviruses are positive single-stranded RNA viruses. It has been debated if they are able to perform homologous recombination through RNA polymerase template switching. Some flaviviruses are hepatitis C, dengue virus and West Nile virus. In all of these there have been found sporadic recombinants, but for example for the dengue virus it has been shown that the detected recombinants are a product of sequencing error. So if recombination occurs in flaviviruses, it is rare. The persistent homology approach that was

used to study reassortment in influenza A and recombination in HIV, was now applied to recombination in flaviviruses. In particular it was applied to hepatitis subtype 1, dengue subtypes 1-4 and West Nile virus. In [5] they found high dimensional topology for hepatitis C virus , but little to no high dimensional topology in dengue and no high dimensional topology in West Nile virus. See Figure 4.11 for a comparison between the TOP (size of longest bar) and ICR (number of bars per time) of different viruses.

Remark 4.11 In viruses with negative-sense RNA recombination is even more rare. An example of such a virus is the Newcastle virus that has a low ICR but a non-zero TOP, also shown in Figure 4.11.

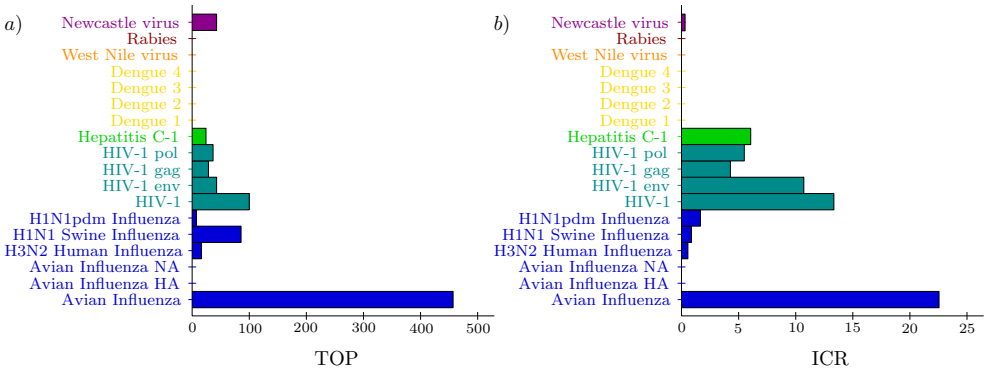


Figure 4.11: a) TOP is calculated using the maximum barcode length in non-zero dimensions. b) The ICR is the number of higher-dimensional barcodes normalized by the time span of sequence collection (data from [5]).

Bibliography

[1] Digital encyclopedia of ancient life. https://www.digitalatlasofancientlife.org/learn/systematics/phylogenetics/character-mapping/, accessed: 29.06.2023, Paleontological Research Institution, Ithaca, New York.

[2] J. Boissonnat, F. Chazal, and M. Yvinec. *Geometric and Topological Inference*. Cambridge University Press, 2018.

[3] P. Buneman. A note on the metric properties of trees. In *Journal of Combinatorial Theory, Series B*, pages 48–50. 1974.

[4] G. Carlsson. *Topological pattern recognition for point cloud data*. Cambridge University Press, 2014.

[5] J. M. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. In *Proceedings of the National Academy of Sciences*, volume 110.46, pages 18566–18571. National Academy of Sciences, 2013.

[6] F. Chazal, D. Cohen-Steiner, L. Guibas, F. Mémoli, and S. Oudot. *Gromov-Hausdorff Stable Signatures for Shapes using Persistence*. Eurographics Symposium on Geometry Processing, 2009.

[7] S. Choudhuri. *Bioinformatics for Beginners*. Academic Press, 2014.

[8] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.

[9] D. Grinberg. Math 5707 spring 2017 (darij grinberg): midterm 2 solution sketches, 2017. http://www.cip.ifi.lmu.de/~grinberg/t/17s/mt2s.pdf, accessed: 20.06.2023.

[10] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002.

[11] P. W. Messer. Neutral Models of Genetic Drift and Mutation. In R. M. Kliman, editor, *Encyclopedia of Evolutionary Biology*, pages 119–123. Academic Press, 2016.

[12] R. Rabadan and A. J. Blumberg. *Topological Data Analysis for Genomics and Evolution*. Cambridge University Press, 2019.

[13] J. Steel and A. C. Lowen. Influenza A Virus Reassortment. In R. W. Compans and M. B. A. Oldstone, editors, *Influenza Pathogenesis and Control*, volume I, pages 377–401. Springer Cham, 2014.

[14] J. Waikagul and U. Thaenkham. *Approaches to Research on the Systematics of Fish-Borne Trematodes*. Academic Press, 2014.

[15] X. Yang. *Nature-Inspired Optimization Algorithms*. Academic Press, 2021.

ETH

Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Declaration of originality

The signed declaration of originality is a component of every semester paper, Bachelor's thesis, Master's thesis and any other degree paper undertaken during the course of studies, including the respective electronic versions.

Lecturers may also require a declaration of originality for other written papers compiled for their courses.

I hereby confirm that I am the sole author of the written work here enclosed and that I have compiled it in my own words. Parts excepted are corrections of form and content by the supervisor.

Title of work (in block letters):

Application of persistent homology to viral evolution

Authored by (in block letters):

For papers written by groups the names of all authors are required.

Name(s):	**First name(s):**
Meinecke	Luisa Sophie

With my signature I confirm that

- I have committed none of the forms of plagiarism described in the 'Citation etiquette' information sheet.
- I have documented all methods, data and processes truthfully.
- I have not manipulated any data.
- I have mentioned all persons who were significant facilitators of the work.

I am aware that the work may be screened electronically for plagiarism.

Place, date	**Signature(s)**
Brugg, 05.07.2023	L Meinecke

For papers written by groups the names of all authors are required. Their signatures collectively guarantee the entire content of the written paper.