UNIVERSIDAD
NACIONAL
DE COLOMBIA

Persistent homology analysis of RNA secondary structures

Astrid Arena Olave Herrera

Supervisors: Gustavo Nevardo Rubiano Ortegón

In Partial Fulfillment of the requirements for the

Degree of Bachelor of Mathematics

Department of Mathematics
Science Faculty
Universidad Nacional de Colombia
Bogotá D.C, Colombia
2018

Dedication

To my mother, my angel on earth.
Thanks for all your efforts
and always believing in me

Acknowledgements

Prima facea, I am grateful to God for all the blessings all alone my life.

I thank my thesis supervisor Gustavo Nevardo Rubiano of the department of mathematics and my co-supervisor Clara Isabel Bermudez Santana of the department of biology , both, at Universidad Nacional de Colombia. They always support me with their knowledge in mathematics, biology and steered me in the right direction whenever I needed it.

I take this opportunity to express gratitude to the group of ”RNomica Teórica y Computacional” for letting me work in their magnificent computational laboratory and gave me the opportunity to interact with people on the biological fields, your suggestions enrich my work.

I also want to thank my friends of the bachelor , my football friends and my friends in God for sharing with me moments of sadness but over all, moments of joy through this five years of University; you guys make my life interesting everyday.

Last but not least, I would like to thank my mother, the most beautiful person that I know. Thanks for being by my side everyday, raised me with so much love and helped me to become who I am today, and who I will be tomorrow.

Abstract

Persistent homology provides a tool to infer qualitative and quantitative information about the structure of a data set, in particular the 0-homology classes can be interpreted as a clustering of the data. A subset of the space of secondary structures of the micro RNA's is chosen to applied this topological analysis. This set defines a metric space regarding the hamming, levenshtein, base pair and tree edit distances. Then, a nested family of Vietoris-Rips complexes from the data set is built in order to convert the data in a topological object and persistent homology is applied. In the thesis we found this method of clustering not to be the best one since it may cluster elements far from each other, making it susceptible to forming a large dominant cluster. But, it performs better when the base pair distance is used, thus, this distance constitutes a better choice of metric. Even though it is not as good as other methods of clustering, it is sensible to small changes of the data points, it identifies distant points in the set and it reveals a prevalent shape of the data set.

Keywords: persistent homology, RNA secondary structure, clustering technique.

Resumen

La homología persistente constituye una herramienta para inferir información cualitativa y cuantitativa de la estructura de un conjunto de datos, en particular las 0-clases de homología pueden ser interpretadas como una forma de aglomeración de los datos. Este análisis topológico es aplicado a un subespacio de las estructuras secundarias de los micro ARN's. Este conjunto define un espacio métrico para las distancias de hamming, levenshtein, bases pareadas y de edición de árboles. Luego de los datos se construye una familia anidada de complejos de Vietoris-Rips convirtiéndolo en un objeto topológico y así aplicar la homología persistente. En la tesis se encuentra que el método de aglomeración no es el mejor ya que puede agrupar elementos distantes entre si, haciendo que sea susceptible a la formación de un agrupamiento grande y dominante, pero trabaja de mejor manera cuando la distancia de bases pareadas es usada, así que esta distancia constituye una mejor opción de métrica. A pesar que el método no es tan bueno como otros métodos de agrupamiento, es sensible a pequeños cambios en los datos, identifica puntos distantes en el conjunto y revela una forma general del conjunto de datos.

Palabras clave: homología persistente, estructura secundaria de ARN, técnica de agrupamiento.

Contents

Background

Topological data analysis (TDA) refers to a collection of methods and tools that enable researches find and study topological invariants structure in data [1]. The most known tool is persistent homology [8] which measures topological features of shapes and functions. These tools have been applied to reveal useful information from the data not always discover by other analysis techniques.

On the other hand, ribonucleic acid (RNA) is an ubiquitous molecule playing an important role in various biological process. The secondary structure of RNA is the set of base pairs that occur when the RNA folds in on itself in a complex three-dimensional shape due to chemical interactions. The RNA secondary structure comparison is essential for classification of RNA by similarities, identification of similar functionalities among RNA and identification of mutations related to mis-functionality.

The first work research in which persistent homology is applied on the space of RNA secondary structures is from Mamuye and Rucco and collaborators in 2015 [23] clustering the space of suboptimal structures and analyzing the structure similarity among optimal structures of family species.

Objectives

General

Apply the analysis of persistent homology as a method to compare and cluster RNA secondary structures.

Specifics

- Choose a suitable set of secondary structures in order to define a consistent and meaningful metric space regarding the selected metrics.
- Interpret the results from the analysis of persistent homology of the secondary structure space as relevant information to compare and cluster the underlying set of RNA foldings.
- Determine the advantages and disadvantages of the topological analysis in relation to the common techniques to compare RNA secondary structures.

State of art

1.1 Topology

1.1.1 Simplicial Complexes

The simplicial complexes are the primary structure to represent topological spaces.

Definition 1.1.1. An *abstract simplicial complex* A is a finite collection of sets such that if $\alpha \in A$ and $\beta \subset \alpha$ implies $\beta \in A$.

We call the sets in A *simplices* and they are finite. The *dimension* of a simplex α is dim $\alpha = |\alpha| - 1$ where $|\alpha|$ is the cardinal of the set α. We define the dimension of the complex as the maximum dimension of any of its simplices. The simplex α it is called a *k-simplex* if its dimension is k. Also the simplex α is *maximal* if there is not simplex β such that $\alpha \subset \beta$.

A *subsimplex* of α is a non-empty subset $\beta \subseteq \alpha$, which is proper if $\beta \neq \alpha$. Sometimes is noted as $\beta \leq \alpha$. If besides, dim $\beta + 1 =$ dim α, then β is called a *face* of α. The *vertex set* is the union of all simplices, $V(A) = \bigcup_{\alpha \in A} \alpha$, that is, the set of all elements that lie in at least one simplex $\alpha \in A$. Finally a subcomplex B of A is an abstract simplicial complex such that $B \subseteq A$.

It is important to see that every abstract simplicial complex A of dimension d has a geometric realization in $\mathbb{R}^{2d+1}$:

1. If k is the number of vertices of A, they are injected in a set of points $v_0, ..., v_k$ in $\mathbb{R}^{2d+1}$ such that they are *affinely independent*, meaning the k vectors $v_i - v_0$, for $1 \leq i \leq k$, are linearly independent. This is possible since if α and β are simplices in A with $n =$ dim α and $m =$ dim β, the union of the two has size card $(\alpha \cup \beta) =$ card $\alpha +$ card $\beta -$ card $(\alpha \cap \beta) \leq n + m + 2 \leq 2d + 2$ and the fact that any $2d + 2$ or fewer of the points are affinely independent.

2. A point $x = \sum_{i=0}^{k} \lambda_i v_i$ with each $\lambda_i \in \mathbb{R}$, is an *affine combination* of the v_i if $\sum_{i=0}^{k} \lambda_i = 1$ and is a *convex combination* if all λ_i are non negative.

3. A *k-simplex* is the set of all the convex combinations of $k + 1$ affinely independent points $\sigma = \{v_0, ..., v_k\}$ as in 1. .

The geometric realization of the k-simplex for $k = 0, 1, 2, 3$ gives the familiar geometric figures: *vertex* for 0-simplex, *edge* for 1-simplex, *triangle* for 2-simplex, and *tetrahedron* for 3-simplex, as shown in Figure 1.1

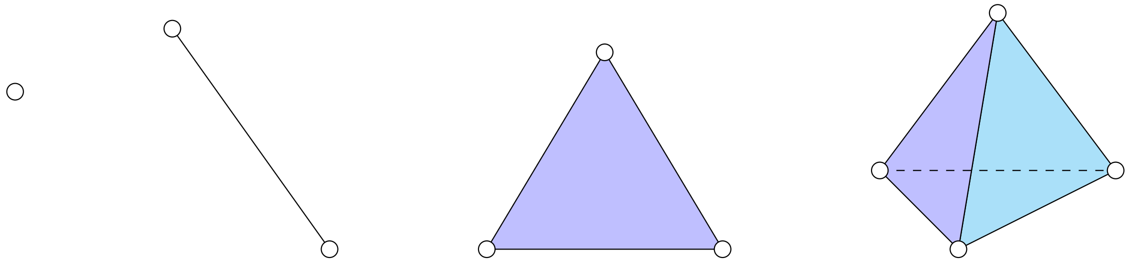


Figure 1.1: From left to right each figure represents: a vertex, an edge, a triangle, and a tetrahedron. Note that an edge has two vertices, a triangle has three edges, and a tetrahedron has four triangles as faces.

A *subsimplex* of σ is the set of all the convex combinations of a non empty subset of the v_i and its proper if is not the entire set. The *boundary* of σ, denoted as $\partial\sigma$ is the union of all faces, and the interior denoted as $\mathring{\sigma}$ is everything else, $\mathring{\sigma} = \sigma - \partial\sigma$. A point $x \in \sigma$ is interior if all its λ_i coefficients are positive. Since x has unique coefficients λ_i, it belongs to the interior of exactly one face.

Note that for a simplex K in a simplicial complex if $\sigma, \sigma' \in K$ then $\sigma \cap \sigma'$ is either empty or a subsimplex of σ and σ' and if $\tau \leq \sigma$ then $\tau \in K$. See for example the representations shown in Figure 1.2. Additionally the underlying space, denoted as $|K|$, is the union of its simplices together with the topology inherited from the ambient Euclidean space in which the simplices live.

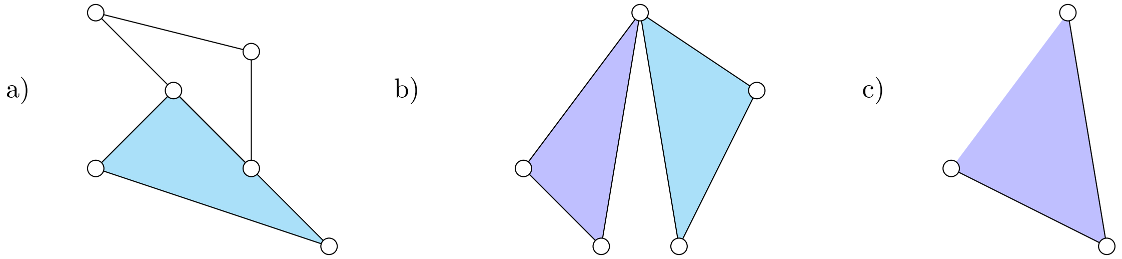


Figure 1.2: a) It is not a simplicial complex since the intersection of the triangle and the edge is not a subsimplex, b) Is a complex since the intersection of the triangles is indeed subsimplex of both and c) Is not a simplex because the triangle has a missing edge.

We say two simplices σ and τ are k-connected if there is a sequence of simplices $\sigma, \sigma_1, ..., \sigma_n, \tau$ such that any two consecutive ones share a k-subsimplex, implying that they have at least $k+1$ vertices in common. Such a chain is called a $k-$chain. The complex K is k-connected if any two simplices in K of dimensionality greater than k are k-connected.

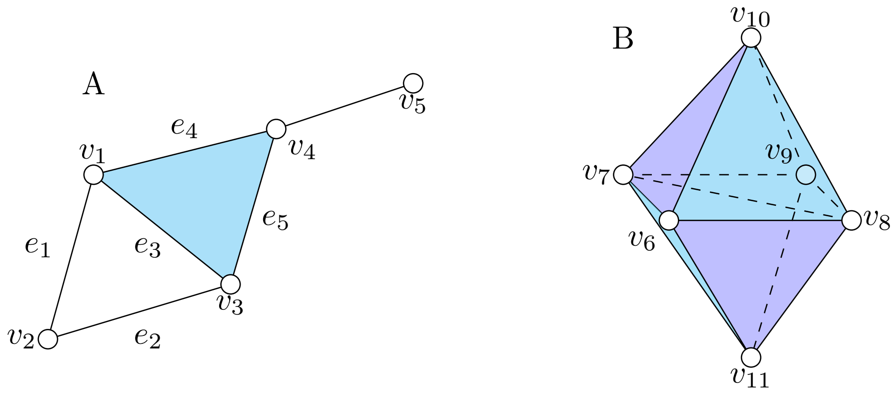


Figure 1.3: A is the realization of the simplicial complex $\{\{v_1\},\{v_2\},\{v_3\},\{v_4\},\{v_5\},\{v_1,v_2\},\{v_1,v_3\},\{v_2,v_3\},\{v_1,v_4\},\{v_3,v_4\},\{v_4,v_5\},\{v_1,v_3,v_4\}\}$. The dimension of A is 2 and A is 0-connected. B is a simplicial complex with dimension 3 and is 2-connected, also is 1-connected.

Simplicial maps

Now, in the most natural way, we define the continuous maps between simplicial complexes.

Let K be a simplicial complex with vertices $u_0, ..., u_n$. It is known for $x \in |K|$, $x = \sum_{i=0}^{n} \lambda_i u_i$ with $\sum_{i=0}^{n} \lambda_i = 1$ and $\lambda_i \geq 0$ for all i.

Let ϕ be a function with $\phi : V(K) \to V(L)$ such that every simplex $\{u_{i_1}, ..., u_{i_k}\}$ in K map to a simplex $\{v_{i_1}, ..., v_{i_k}\}$ in L. Then ϕ can be extended to a continuous function f

$$\begin{aligned} f : |K| &\to |L| \\ x &\mapsto \sum_{i=0}^{n} \lambda_i \phi(u_i) \end{aligned}$$

Just written as $f : K \to L$.

If the vertex map $\phi : V(K) \to V(L)$ is bijective and $\phi^{-1} : V(L) \to V(K)$ is also a vertex map, then the induced simplicial map f is a *homeomorphism* or an *isomorphism* between K and L.

1.1.2 Convex Set Systems

It is possible that simplicial complexes arise as intersection patterns of collections of sets. We present two fundamental theorems for convex sets, remind that a set C is convex if for any x and y in C and all t in the interval $[0, 1]$, the point $(1-t)x+ty$ also belongs to C.

Theorem 1.1.2. HELLY'S THEOREM. Let $\mathcal{F}$ be a finite collection of closed and convex sets in $\mathbb{R}^d$. Every $d+1$ sets have a non-empty common intersection if and only if they all have a non-empty common intersection. (For its proof see [7])

Definition 1.1.3. Let $\mathcal{F}$ be a finite collection of sets. The *nerve* of $\mathcal{F}$ consist of all non-empty subcollections whose sets have a non-empty common intersection. We noted as Nrv$\mathcal{F}$

$$\mathrm{Nrv}\mathcal{F} = \{X \subset \mathcal{F} | \bigcap X \neq \emptyset\}$$

The nerve of $\mathcal{F}$ defines an abstract simplicial complex since satisfies the Definition 1.1.1.

Theorem 1.1.4. NERVE THEOREM. Let $\mathcal{F}$ be a finite collection of closed, convex sets in an Euclidean space. Then the nerve of $\mathcal{F}$ and the union of the sets in $\mathcal{F}$ have the same topological properties.

Čech complexes

Now let us consider the special case in which the convex sets are closed geometric balls, all of the same radius r.

Definition 1.1.5. Let S be a finite set of points in $\mathbb{R}^d$ and let $B_r(x)$ be the closed ball with center x and radius r. The Čech complex of S and r is the nerve of this collection of balls, that is:

$$\check{\mathrm{C}}\mathrm{ech}(S, r) = \{\sigma \subseteq S | \bigcap_{x \in \sigma} B_r(x) \neq \emptyset\}$$

Clearly, a set of balls has a non-empty intersection if and only if their centers lie inside a common ball of the same radius. An easy consequence of Helly's Theorem is therefore that every $d+1$ points in S are contained in a common ball of radius r iff all points in S are.

The Čech complex was introduced by Eduard Čech[1]. We show an example of a Čech complex in the Figure 1.4.

Vietoris-Rips complexes

This complex was first introduced by Leopold Vietoris [2] in 1927. Instead of checking all subcollections of S as the Čech complex, *the Vietoris-Rips complex* just check the pairs and add 2- and higher-dimensional simplices whenever all their edges are in the complex. This simplification lead us to the following definition:

$$\mathrm{VR}(S, r) = \{\sigma \subseteq S \mid \mathrm{diam}\sigma \leq 2r\}$$

[1] Eduard Čech was a Czech mathematician born in Stračov in 1893 and die in Prague in 1960.

[2] Leopold Vietoris was an Austrian mathematician. He was born in Radkersburg in 1891 and died in Innsbruck in 2002.
Eliyahu Rips is an Israeli mathematician born in 1964 in Latvia

Where diamσ is the maximum distance between the points on σ.

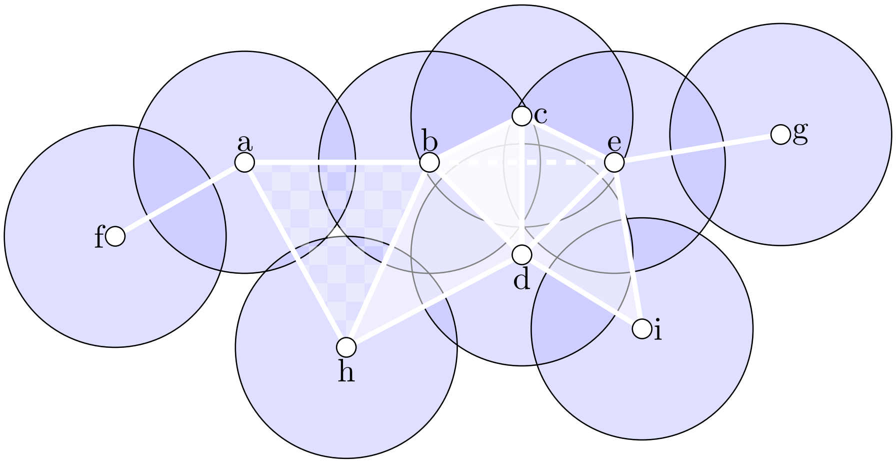


Figure 1.4: Vietoris-Rips complex of nine points with pairwise intersections among the disks indicated by straight white edges connecting their centers. As maximal simplices it has $\{a, f\}$, $\{a, b, h\}$, $\{g, e\}$, $\{b, d, h\}$, $\{d, e, i\}$ and $\{b, c, d, e\}$. Notice that $\check{\mathrm{C}}\mathrm{ech}(S, r)$ has the same maximal simplices except for $\{a, b, h\}$, instead, it has $\{a, h\}$ and $\{b, h\}$ as maximal components.

Note that the previous definitions can be generalized to a finite metric space if we set that for σ subset of S, $\sigma \in \check{\mathrm{C}}\mathrm{ech}(S, r)$ if $d(i, j) \leq 2r$ for all $i, j \in \sigma$ and $\sigma \in \mathrm{VR}(S, r)$ if diam$\sigma \leq 2r$.

Clearly, $\check{\mathrm{C}}\mathrm{ech}(S, r) \subseteq \mathrm{VR}(S, r)$, because the latter contains every simplex warranted by the given edges, as shown in Figure 1.4. Also it is true that $\mathrm{VR}(S, r) \subseteq \check{\mathrm{C}}\mathrm{ech}(S, \sqrt{2}r)$. Then

$$\mathrm{VR}(S, r) \subseteq \check{\mathrm{C}}\mathrm{ech}(S, \sqrt{2}r) \subseteq \mathrm{VR}(S, \sqrt{2}r) \tag{1.1.1}$$

1.1.3 Homology

“Homology is a mathematical formalism for talking in a quantitative and unambiguous manner about how a space is connected. Compared to most other, competing formalisms, homology has faster algorithms but captures less of the topological information, however is not necessarily a drawback”. [7]

Chain Complexes

Definition 1.1.6. Let K be a simplicial complex and p a dimension. c is called a *p-chain* of K if $c = \sum_i a_i \sigma_i$ where σ_i are p-simplices in k and a_i the coefficients.

The coefficients can be integers, rational or even real numbers, but is enough for us to consider them in $\mathbb{Z}_2$. For now on, unless stated otherwise, the coefficients are always in $\mathbb{Z}_2$.

We sum the p-chains componentwise. Specifically, if $c = \sum_i a_i\sigma_i$ and $c' = \sum_i b_i\sigma_i$ then $c + c' = \sum_i(a_i + b_i)\sigma_i$. The set of p-chains together with the addition operation form the abelian *group of p-chains* denoted as $C_p(K)$ or just C_p. At the same time the p-chains form a vectorial space over $\mathbb{Z}_2$. For p less than zero and greater than the dimension of K the group is trivial, consisting only with the zero.

Let $\sigma = [v_0, ..., v_p]$ be the simplex spanned by the listed vertices and let $\tau = [v_0, ..., \hat{v}_i, ..., v_p]$ be the simplex generated but all except for v_i. Clearly $\tau \leq \sigma$.

Definition 1.1.7. The *boundary* of a p-simplex, noted by $\partial_p\sigma$ is the sum of its faces:

$$\partial_p\sigma = \sum_{i=0}^{p}[v_0, ..., \hat{v}_i, ..., v_p]$$

Hence ∂_p defines a function from C_p to C_{p-1}. Moreover, is a linear map since $\partial_p(\lambda c) = \lambda\partial_p(c)$ for $\lambda = 0, 1$ and $\partial_p(c + c') = \partial_p c + \partial_p c'$. Therefore ∂_p is refer as the *boundary map.*

Definition 1.1.8. The *chain complex* is the sequence of chain groups connected by boundary maps

$$... \xrightarrow{\partial_{p+2}} C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \xrightarrow{\partial_{p-1}} ...$$

The chain complex is noted as (C, ∂) where C and ∂ are the collections of C_i and ∂_i respectively.

Cycles and boundaries

The linear transformation ∂_p establishes two well known subgroups, the kernel $K(\partial_p)$ and the image $\mathrm{Im}(\partial_p)$. If $c \in K(\partial_p)$ is going to be called a *p-cycle.* Also the group of p-cycles is noted as Z_p. On the other hand a *p-boundary* is an element of $\mathrm{Im}(\partial_{p+1})$ and the group of p-boundaries is noted as B_p.

Example 1.1.9. Let $\tau = [a, b, c]$ be the triangle abc as shown in Figure 1.5 and now $\partial_2(\tau)$ is calculated.

$$\partial_2(\tau) = [b, c] + [a, c] + [a, b]$$

And now let us calculate $\partial_1(\partial_2(\tau))$

$$\begin{aligned}\partial_1(\partial_2(\tau)) &= \partial_1([b, c]) + \partial_1([a, c]) + \partial_1([a, b])\\ &= c + b + c + a + b + a\\ &= 2c + 2b + 2a\\ &= 0\end{aligned}$$

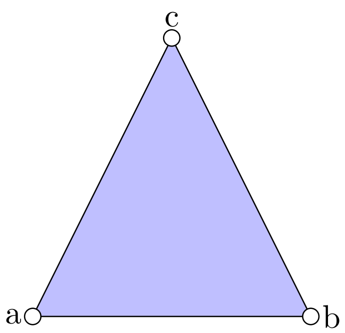


Figure 1.5: 2-simplex $[a, b, c]$

Remembering that coefficients are in $\mathbb{Z}_2$.

Is not an accident that $\partial_1(\partial_2(\tau)) = 0$, it is true for all simplices and the result is shown in the next lemma.

Lemma 1.1.10. FUNDAMENTAL LEMMA OF HOMOLOGY. $\partial_{p-1}(\partial_p(c)) = 0$ for every non negative integer p and every p-chain c.

Proof. It is enough if it is true for every p-simplex τ. Let $\tau = [v_0, ..., v_p]$

$$\begin{aligned}\partial_{p-1}(\partial_p(\tau)) &= \partial_{p-1}(\sum_{i=0}^{p}[v_0, ..., \hat{v}_i, ..., v_p]) \\ &= \sum_{i=0}^{p} \partial_{p-1}([v_0, ..., \hat{v}_i, ..., v_p]) \\ &= \sum_{i=0}^{p} \sum_{j=0, j\neq i}^{p} [v_0, ..., \hat{v}_j, ..., \hat{v}_i, ..., v_p] \\ &= \sum_{i=1, j<i}^{p} 2[v_0, ..., \hat{v}_j, ..., \hat{v}_i, ..., v_p] \\ &= 0\end{aligned}$$

□

It follows that every p-boundary is also a p-cycle or, equivalently, that $B_p \subseteq Z_p$. Figure 1.6 illustrates the subgroup relations among the three types of groups and their connection across dimensions established by the boundary homomorphisms.

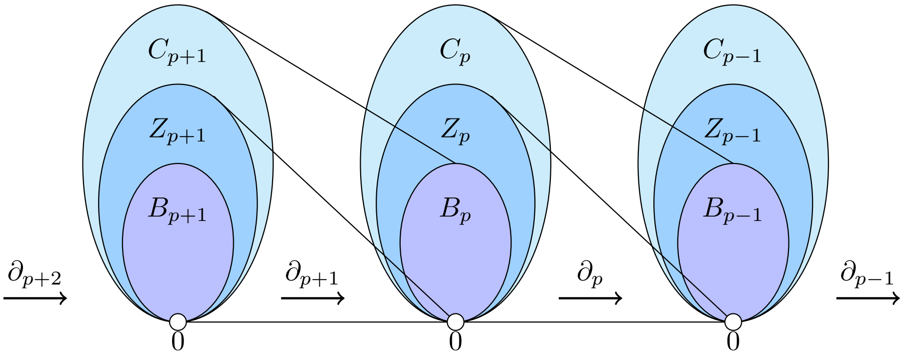


Figure 1.6: The chain complex consisting of a linear sequence of chain, cycle and boundary groups connected by homomorphisms.

Homology groups

Definition 1.1.11. The *p-th homology group* H_p is the p-th group modulo the p-th boundary group, $H_p = Z_p/B_p$. The *p-th Betti*[3] *number* β_p is the rank of H_p.

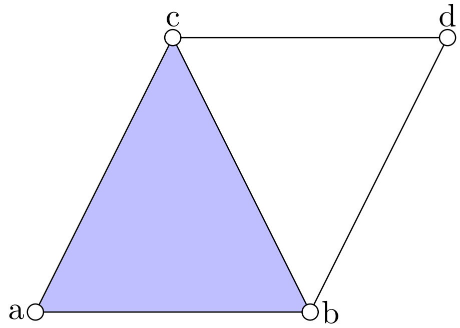
[3] Enrico Betti Glaoui was an Italian mathematician born in Pistoia in 1823 and die in Soiana in 1892

We call each coset of H_p a *homology class*. Any two cycles c and c' in the same homology class are called *homologous*, which is denoted as $c \sim c'$. H_p is indeed a group, and because Z_p is abelian, so is H_p. For example in Figure 1.3 the 1-cycles $e_1 + e_2 + e_3$ and $e_1+e_2+e_4+e_5$ are homologous since its sum its equal to $e_3+e_4+e_5$ which is a 2-boundary.

The cardinality of a group is called its *order*. Since the coefficients are modulo 2, a group with n generators has order 2^n, furthermore is isomorphic to $\mathbb{Z}_2^n$. The *dimension* is referred to the rank of the vector space, $n = \text{rank } \mathbb{Z}_2^n$. The number of cycles in each homology class is the order of H_p; hence the number of classes in the homology group is $\text{ord}H_p = \text{ord}Z_p/\text{ord}B_p$, equivalently $\beta_p = \text{rank}Z_p - \text{rank}B_p$.

Example 1.1.12. Consider $K = \{[a, b, c], [b, d], [c, d]\}$ as shown in the figure below.

First, since $C_i = 0$ for $i > 2$, then $H_i = 0$ for $i > 2$.



Now the chain to be considered is:

$$0 \xrightarrow{\partial_3} C_2 \xrightarrow{\partial_2} C_1 \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0$$

Clearly $B_2 = 0$ and $Z_0 = C_0 \cong \mathbb{Z}_2^4$.

Second, using the example 1.5 is found that

$$\partial_2(C_2) = \{u([b, c] + [a, c] + [a, b]) \mid u = 0, 1\}$$

Thus $Z_2 = 0$ and $B_1 \cong \mathbb{Z}_2$.

Third, let $c = u_1[a, b] + u_2[b, c] + u_3[a, c] + u_4[b, d] + u_5[c, d]$ be an arbitrary chain from C_1, then

$$\begin{aligned}\partial_1(c) &= u_1(a + b) + u_2(b + c) + u_3(a + c) + u_4(b + d) + u_5(c + d)\\ &= (u_1 + u_3)a + (u_1 + u_2 + u_4)b + (u_2 + u_3 + u_5)c + (u_4 + u_5)d\end{aligned}$$

If $c \in Z_1$, therefore $u_1 + u_3 = 0$ or equivalently $u_3 = u_1$. In the same way, $u_5 = u_4$ and $u_2 = u_1 + u_4$ implying $c = u_1([a, b] + [a, c] + [b, c]) + u_4([b, d] + [c, d] + [b, c])$ taking u_1, u_4 in $\mathbb{Z}_2$. Therefore $Z_1 \cong \mathbb{Z}_2^2$. In addition, the chains in B_0 are of the form $u_1a + u_2b + u_3c + (u_1 + u_2 + u_3)d$, in this way $B_0 \cong \mathbb{Z}_2^3$.

Consequently, $H_2 \cong 0/0 \cong 0$, $H_1 \cong \mathbb{Z}_2^2/\mathbb{Z}_2 \cong \mathbb{Z}_2$ and $H_0 \cong \mathbb{Z}_2^4/\mathbb{Z}_2^3 \cong \mathbb{Z}_2$. □

Intuitively, the cycles that are boundaries of higher-dimensional subcomplex from the set of all p-cycles are removed, so that the ones that remain carry information about the p-dimensional holes of the complex. In simplest terms β_p counts the number of p-dimensional holes in a simplicial complex. There is an intuitive depiction of the first three Betti numbers, for β_0, a theorem is given in [11] by John B. Fraleigh :

Theorem 1.1.13. Let K be a simplicial complex, then $\beta_0(K)$ is equal to the number of 0-connected components of K.

Proof. First, a *path* between two vertices u and w is a sequence of edges $e_i = \{v_{i1}, v_{i2}\}$ with $i = 1, 2, ...n$ such that $e_{11} = u$, $e_{n2} = w$ and $e_{i2} = e_{(i+1)1}$. This path is noted as $u, v_{21}, ...v_{n1}, w$ or simply $u, v_2, ...v_n, w$. Clearly, K is 0-connected if any two vertices can be joined by a path.

Now, $C_0(K)$ is composed by chains of the form $\sum_i a_i v_i$ where v_i are the vertices of K. Fix a vertex u, then for any vertex w in the same connected component there is a path

$$u, v_2, ...v_n, w$$

Then

$$w = u + (u + v_2) + (v_2 + v_3) + ... + (v_{n-1} + v_n) + (v_n + w)$$

showing that $w \in u + B_0(K)$ since $\partial_1(\{v_i, v_j\}) = v_i + v_j$.

It is easy to see that if w is not in the same connected component as u, then $w \notin u + B_0(K)$. Otherwise there are n edges such that

$$w = u + (v_{11} + v_{12}) + ...(v_{n1} + v_{n2}) \tag{1.1.2}$$

With out losing generality, $v_{11} = u$ in order to eliminate u, if $v_{12} = w$ then u and w are in the same connected component, thus $v_{12} \neq w$. By induction $v_{i2} = v_{(i+1)1}$, if $v_{(i+1)2} = w$ then there is a path between u and w implying they are in the same component, but then, there is not $v_{ij} = w$, contradicting 1.1.2.

Thus each coset of $H_0(K)$ represents exactly one and only one connected component proving the assertion.

□

Due to an Alexander[4] duality property [15] for H_1, the non-bounding 1-cycle represents a collection of non-contractible closed curves in K, or a set of tunnels formed by K. So, β_1 represents the dimension of the basis for the tunnels. The non-bounding 2-cycle represents the set of non-contractible closed surfaces in K, or a set of voids. The dimension of the basis for voids is represented by β_2.

Induced homomorphisms

Let K and L be simplicial complexes and f a continuous function from K to L, then f naturally gives rise to a homomorphism f_p of $C_p(K)$ into $C_p(L)$ which has the important property that commutes with ∂_p, meaning that

$$\partial_p f_p = f_{p-1} \partial_p$$

And implies that f defines a homomorphism $\hat{f}$ between $H_p(K)$ and $H_p(L)$ for each dimension p, which is consequence of the next theorem.

[4] James Waddell Alexander II is a mathematician born in Sea Bright in 1888 an die in Princeton in 1971

Theorem 1.1.14. Let (C, ∂) and (D, ∂') be chain complexes and suppose there is a collection f of homomorphisms $f_p : C_p \to D_p$ such that $\partial'_p f_p = f_{p-1}\partial_p$

$$\begin{tikzcd}
\cdots \arrow[r, "\partial_{p+2}"] & C_{p+1} \arrow[r, "\partial_{p+1}"] \arrow[d, "fp+1"] & C_p \arrow[r, "\partial_p"] \arrow[d, "fp"] & C_{p-1} \arrow[r, "\partial_{p-1}"] \arrow[d, "fp-1"] & \cdots \\
\cdots \arrow[r, "\partial'_{p+2}"] & D_{p+1} \arrow[r, "\partial'_{p+1}"] & D_p \arrow[r, "\partial'_p"] & D_{p-1} \arrow[r, "\partial'_{p-1}"] & \cdots
\end{tikzcd}$$

Then f_k induces a natural homomorphism $\hat{f}_k : H_k(C) \to H_k(D)$.

Proof. Let $z \in Z_k(C)$. Now

$$\partial'_p(f_p(z)) = f_{p-1}(\partial_p(z)) = \partial_p(0) = 0$$

so $f_p(z) \in Z_p(D)$. Let's define $\hat{f}$

$$\begin{aligned} \hat{f}_k &: H_k(C) \to H_k(D) \\ z + B_k(C) &\mapsto f_k(z) + B_k(D) \end{aligned} \tag{1.1.3}$$

Take $z_1 \in z + B_k(C)$, then $z_1 - z \in B_k(C)$, so there exists $c \in C_{p+1}$ such that $\mathbb{Z}_1 - z = \partial_{p+1}(c)$. But then

$$f_p(z_1) - f_p(z) = f_p(z_1 - z) = f_p(\partial_{p+1}(c)) = \partial'_{p+1}(f_{p+1}(c))$$

Hence $f_k(z_1) \in f_k(z) + B_k(C)$ and in this way $\hat{f}_k$ 1.1.3 is well defined since is independent from the choice of the representative.

As f_k is a homomorphism between $Z_k(C)$ and $Z_k(D)$ then $\hat{f}_k$ is a homomorphism of $H_k(C)$ into $H_k(D)$. □

1.1.4 Persistent Homology

"The concept of persistence emerged independently in the work of Frosini, Ferri, and collaborators in Bologna, Italy, in the doctoral work of Robins at Boulder, Colorado, and within the biogeometry project of Edelsbrunner at Duke, North Carolina." [8]

Filtrations

Definition 1.1.15. A *filtration* of a complex K is a nested sequence of subcomplexes,

$$\emptyset = K_0 \subset K_1 \subset K_2 \subset ... \subset K_m = K$$

A complex K with a filtration is called a *filtered complex*.

Before continuing, notice for a finite set of points S in $\mathbb{R}^d$ and a sequence of real numbers $r_0, ..., r_n$ with $r_0 = 0$ and $r_n = diam(S)$, $\mathrm{VR}(S, r_i)$ defines a filtration:

$$\emptyset \subset \mathrm{VR}(S, r_0) \subset \mathrm{VR}(S, r_1) \subset ... \subset \mathrm{VR}(S, r_n)$$

Similarly there is a filtration for $\check{\mathrm{C}}\mathrm{ech}(S, r_i)$. For example in Figure 1.7 is shown a sequence of Vietoris Rips complexes for a set $S \subset \mathbb{R}^2$ sampled from an annulus.

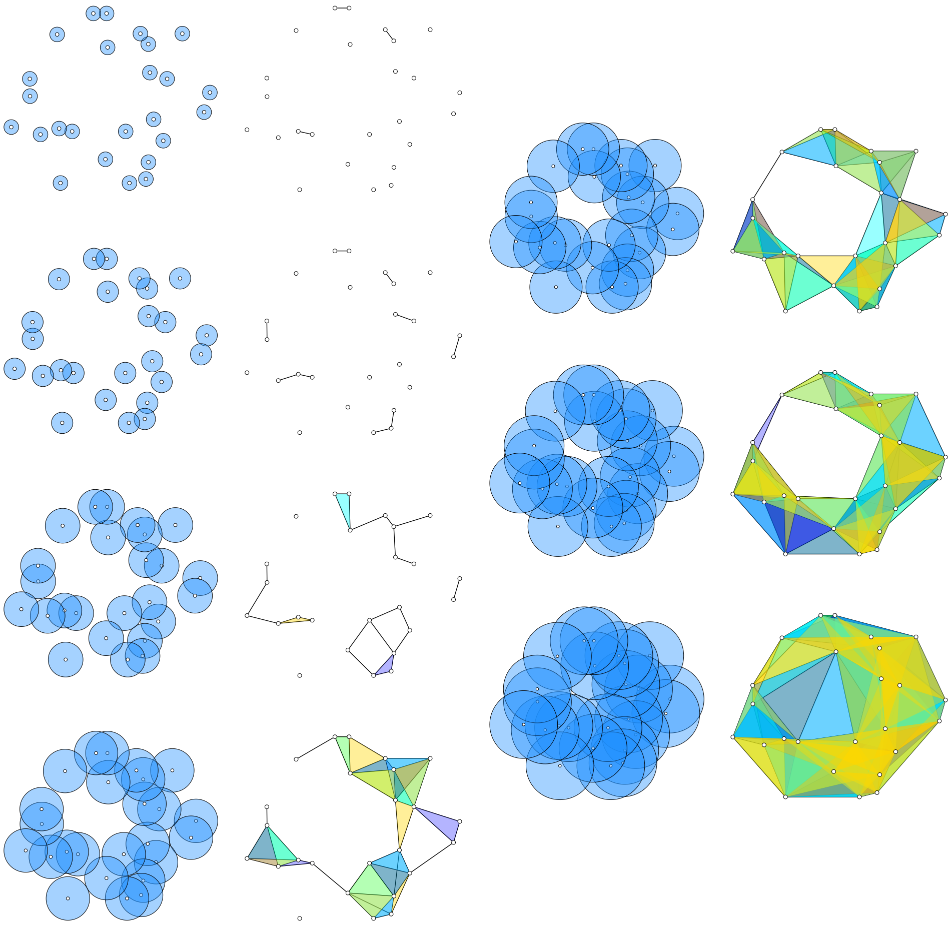


Figure 1.7: A sequence of Vietoris-Rips complexes for a point cloud data set S sampled from an annulus.

Let C_p^i, Z_p^i and B_p^i represent the p-th chain group, the p-th cycle group and the p-boundary group, respectively, of the i-th complex K_i in the filtration sequence.

For every $i \leq j$ there is an inclusion map $f^{i,j}$ from K_i to K_j. Using the result from the Theorem 1.1.14 $f^{i,j}$ gives rise to a homomorphism $f_p^{i,j} : H_p(K_i) \rightarrow H_p(K_j)$ for each dimension p. The filtration thus corresponds to a sequence of homology groups connected by homomorphisms,

$$0 = H_p(K_0) \xrightarrow{f_p^{0,1}} H_p(K_1) \rightarrow ... \rightarrow H_p(K_{n-1}) \xrightarrow{f_p^{n-1,n}} H_p(K_n) = H_p(K)$$

again one for each dimension p. May in K_i there are homology classes that are not in K_{i-1} or some classes in K_{i-1} become trivial or merge with others in K_i. The classes that are born at or before a given threshold and die after another threshold are collected in groups.

Definition 1.1.16. The *p-th persistent homology groups* are the images of the homomorphisms induced by inclusion, $H_p^{i,j} = \text{Im } f_p^{i,j}$, for $0 \leq i \leq j \leq n$. The corresponding *p-th persistent Betti numbers* are the ranks of these groups, $\beta_p^{i,j} = \text{rank} H_p^{i,j}$.

Lemma 1.1.17. Let $f_p^{i,j}$ be defined as before. Then Im $f_p^{i,j} \cong Z_p^i/(B_p^j \cap Z_p^i)$

Proof. Let's define

$$\begin{aligned} F : Z_p^i \rightarrow & \text{ Im}(f_p^{i,j}) \\ z \mapsto & f_p^{i,j}([z]) = z + B_p^j \end{aligned}$$

F is clearly surjective and $ker(F) = \{z \in Z_p^i | z + B_p^j = B_p^j\} = \{z \in Z_p^i | z \in B_p^j\} = Z_p^i \cap B_p^j$. Thus , by the fundamental homomorphism theorem Im $f_p^{i,j} \cong Z_p^i/(B_p^j \cap Z_p^i)$ □

Hence $\beta_p^{i,j}$ counts homological classes in the complex K_j which were created during filtration in the complex K_i or earlier. Let $[z]$ be a class in $H_p(K_i)$, it borns at K_i if $[z] \notin H_p^{i-1,i}$. Furthermore, if $[z]$ is born at K_i , then it dies entering K_j if it merges with an older class from K_{j-1}, that is $f_p^{i,j-1}([z]) \notin H_p^{i-1,j-1}$ but $f_p^{i,j}([z]) \in H_p^{i-1,j}$.

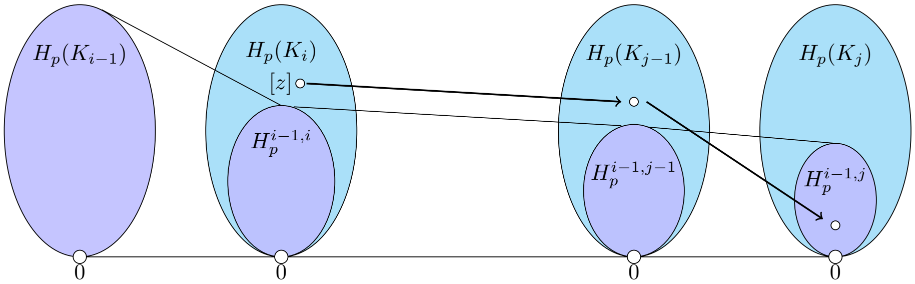


Figure 1.8: Representation of a class $[z]$ that is born at K_i and dies in K_j

To get a more intuitive illustration of persistence concept, the following ideas are shown in [18]: let z be a non-bounding p-cycle created in K_i as a consequence of the appearance of the simplex σ in the complex. The simplex σ is labeled as a *creator simplex*, or

$\sigma+$, (*positive simplex*). Consider the appearance of another simplex τ in K_j with $j \geq i$ which turns a cycle z' in $[z]$ into a boundary, so that $z' \in B_p$. This causes the decrease of the rank of the homology group since the class $[z]$ is joined with the older class of cycles. The simplex τ is labeled as an *annihilator simplex*, $\tau-$, (*negative simplex*) since it annihilates $[z]$.

If $[z]$ is born at K_i and dies at K_j it is said that $[z]$ is born at time (step) i and dies at time (step) j. The *index persistence* of $[z]$ is defined as $j - i$. If $[z]$ never dies its index persistence is set to infinity. If $j - i$ is large (long enough), $[z]$ can be considered as pertinent information about homology groups and Betti numbers. Meanwhile for short $j - i$ are possibly topological noise.

Persistence diagrams

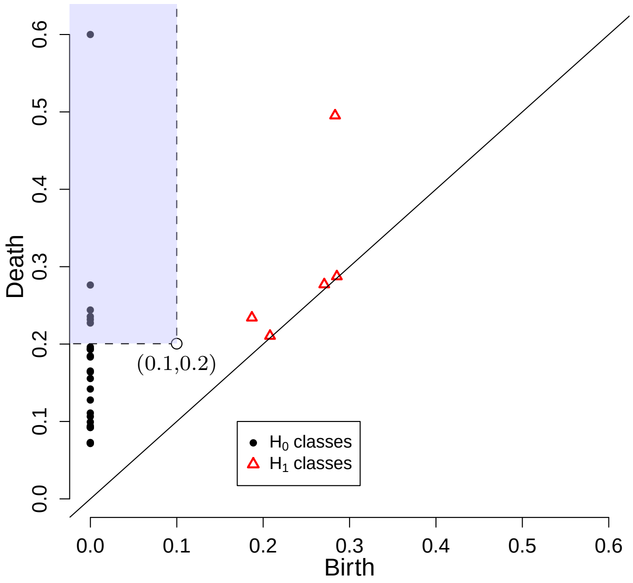


Figure 1.9: Persistence diagram of the set of points S. The 5 black points in the shaded region correspond to $\beta_0^{0.10,0.15}$, meanwhile $\beta_1^{0.10,0.15} = 0$, since there are not triangles in the shaded region.

Let $\mu_p^{i,j}$ be the number of independent p-dimensional classes that are born at K_i and die entering K_j, then

$$\mu_p^{i,j} = (\beta_p^{i,j-1} - \beta_p^{i,j}) - (\beta_p^{i-1,j-1} - \beta_p^{i-1,j}) \tag{1.1.4}$$

for all $i < j$ and all p. Indeed, the first difference on the right side counts the classes that are born at or before K_i and die entering K_j, while the second difference counts the classes that are born at or before K_{i-1} and die entering K_j.

The *p-persistence diagram* of the filtration, denoted as $Dgm_p(f)$ is made drawing each point (i, j) with multiplicity $\mu_p^{i,j}$. It represents a class whose index persistence is the vertical distance to the diagonal. Since the multiplicities are defined only for $i < j$, all points lie above the diagonal. For instance Figure 1.9 is a 0-persistence diagram and 1-persistence diagram combined of the sampling of points in an annulus from Figure 1.7. It is easy to read off the persistent Betti numbers. Specifically, $\beta_p^{k,l}$ is the number of points in the upper left quadrant with corner point (k, l). A class that is born at K_i and dies entering K_j is counted iff $i \leq k$ and $j > l$. The quadrant is therefore closed along its vertical right side and open along its horizontal lower side.

Lemma 1.1.18. FUNDAMENTAL LEMMA OF PERSISTENT HOMOLOGY Let $\emptyset = K_0 \subset K_1 \subset K_2 \subset ... \subset K_m = K$ be a filtration. For every pair of indices $0 \leq k \leq l \leq n$ and every dimension p, the p-th persistent Betti number is $\beta_p^{k,l} = \sum_{i \geq k} \sum_{j > l} \mu_p^{i,j}$

This is an important property since makes the diagram to encode all information about persistent homology groups.

Barcodes

There is other graphical way to represent persistent homology. Since persistent homology detects the birth and death of each topological feature as the complex evolves in time (step), this inspires a visual snapshot of H_p in the form of a *barcode*.

A barcode represents H_p as a collection of horizontal line segments in a plane whose horizontal axis corresponds to the parameter (time) and whose vertical axis represents an arbitrary ordering of p-cycles that are homology generators. If l is the segment representing the cycle σ, l is in the height of σ and l starts at i and finish at j if σ is born at time i and dies at time j. The Figure 1.10 gives an example of the barcode representation of the 0-homology and 1-homology of the sampled points in Figure 1.7.

Clearly, the barcodes do not provide information on the delicate structure of the homology, however, $\beta_p^{i,j}$ is equal to the number of intervals in the barcode of H_p spanning the parameter interval $[i, j]$ and has the ability to filter out topological noise and capture significant features. Indeed, in Figure 1.10 the point cloud likely represents one connected object with one significant 1-hole, as expected, since S was sampled from an annulus.

1.1.5 Computation

Let S be a finite set of points from a metric space. From a computational point of view, the Rips complex of S for any r is less expensive that the corresponding Čech complex, even though the Vietoris Rips complex has in general more simplices. The reason of that is the Rips complex is completely determined by the skeleton of its edges and can be stored as a graph and reconstituted instead of storing the entire complex.

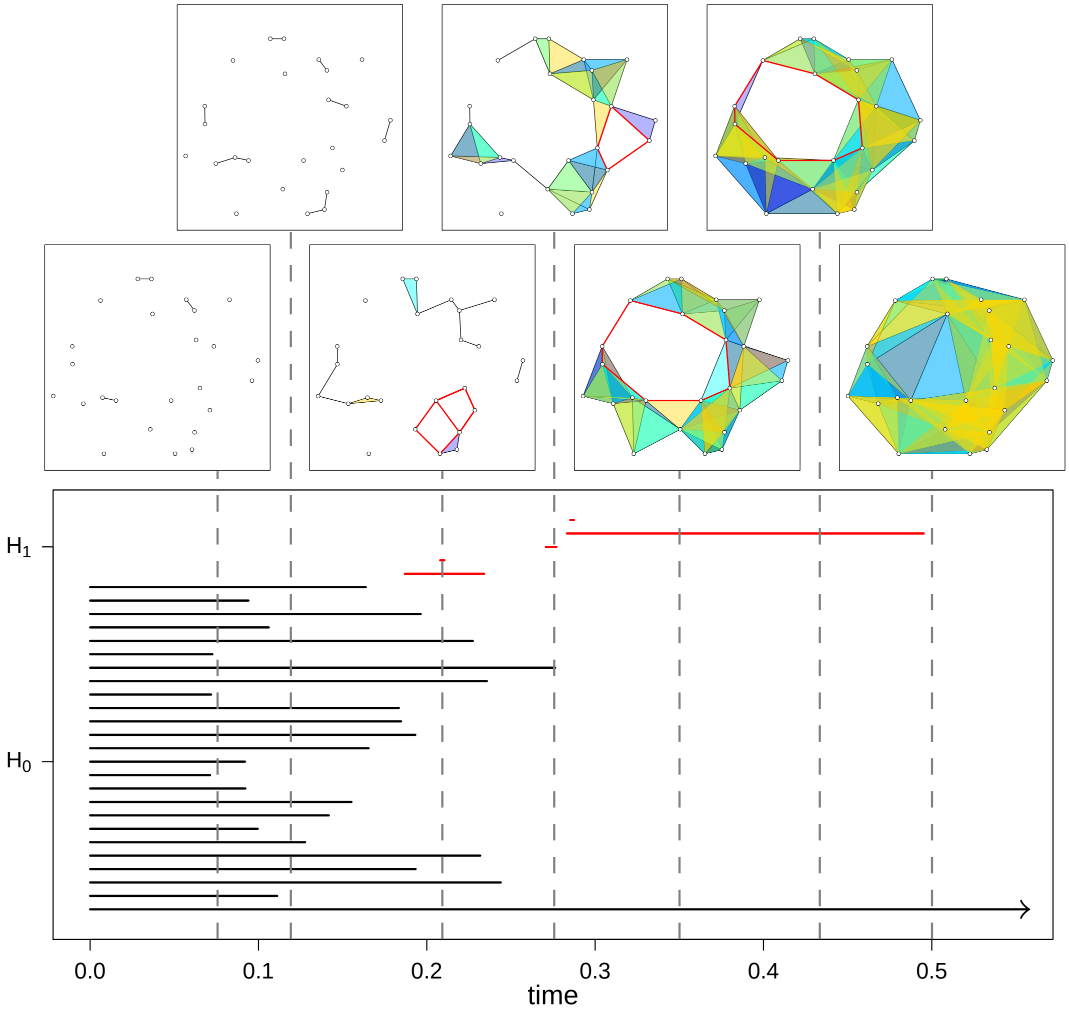


Figure 1.10: Barcode diagram of the set of points S

Nevertheless, this virtue is not without cost. While the Čech complex behaves exactly like its underlying set (see Nerve theorem, 1.1.4), the Rips complex does not necessarily. However, using the equation 1.1.1 it is clear that any topological feature that persists from $\mathrm{VR}(S,r)$ to $\mathrm{VR}(S,\sqrt{2}r)$ is in fact a topological feature of $\check{\mathrm{C}}\mathrm{ech}(S\sqrt{2}r)$, thus a filtration of Vietoris-Rips complexes reveals the topological information given by the Čech complexes.[12]

Matrix reduction

On the other hand, we can compute persistence efficiently in just one matrix reduction.

First, let $\{K_i\}$ be a filtration of K such that $K_i - K_{i-1} = \sigma_i$ where σ_i is a simplex. Then there is a natural total order over the simplices of K: $\sigma_0, ..., \sigma_m$.

Second, define the matrix $\partial \in \mathcal{M}_m[\mathbb{Z}_2]$ as:

$$\partial[i,j] = \begin{cases} 1 & \text{if } \sigma_i \leq \sigma_j \text{ and } \dim(\sigma_i) = \dim(\sigma_j) - 1 \\ 0 & \text{otherwise} \end{cases}$$

In simple terms, $\partial[i,j] = 1$ if σ_i is a face of σ_j.

Subsequently, the matrix ∂ is reduced to a matrix R using the algorithm shown below. Let $low(i)$ be the index of the row with the lowest 1 of a non-zero column R_i. Then

```
$R = \partial$
for $j = 1$ to $m$ do
    while there exists $j_0 < j$ with $low(j_0) = low(j)$ do
        add $R_{j_0}$ to $R_j$
    done
done
```

As summing columns is a matrix operation is equivalent to multiply ∂ by an upper triangular matrix V such that $V[i,i] = 1$, $V[i,j] = 1$ if R_i was added to R_j and the other entries of V are 0. Thus R_j stores the boundary of the chain in V_j. Keep in mind that R is still a 0-1 matrix because the coefficients are in $\mathbb{Z}_2$. R is called *reduced* if whenever R_i and R_j are non-zero columns, $low(i) \neq low(j)$.

Note that the matrix associated to the linear map ∂_p is a submatrix of ∂ taking the columns that corresponds to p-simplices and the rows that correspond to $p-1$-simplices. Then, the number of zero columns that correspond to p-simplices is the rank of Z_p and the number of non-zero columns of p-simplices gives the rank of B_{p-1}.

Example 1.1.19. Consider again the complex $K = \{[a,b,c],[b,d],[c,d]\}$ from the Example 1.1.12.

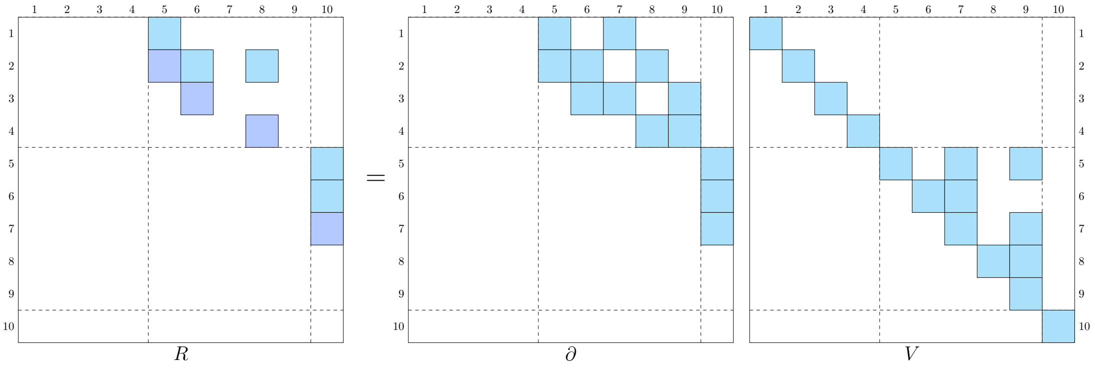


Figure 1.11: Reducing the matrix ∂ of K. The ligth blue squared mark the ones and the gray squares mark the lowest 1's in R

Define an order of the simplices:

$$\emptyset, [a], [b], [c], [d], [a,b], [b,c], [a,c], [b,d], [c,d], [a,b,c]$$

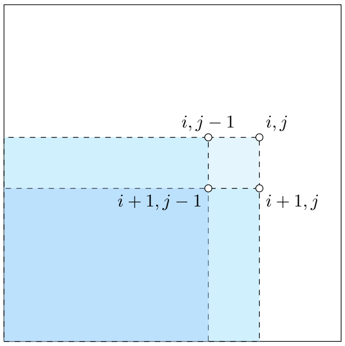
and make ∂ and R
Remember that

$$\beta_p = \text{rank} Z_p - \text{rank} B_p$$

In this way, R gives the following information: rank $Z_0 = 4$, rank $B_0 = 3$, rank $Z_1 = 2$, rank $B_1 = 1$, rank $Z_2 = 0$ and rank $B_2 = 0$. Thus, by definition $\beta_0 = 4 - 3 = 1$, $\beta_1 = 2 - 1 = 1$, $\beta_2 = 0 - 0 = 0$ and $\beta_i = 0$ for $i > 2$ agreeing with what was known before.

□

That is not all, the matrix R also gives information about persistent homology. But first we present a lemma in order to achieve this information.



Lemma 1.1.20. Let R and R' be reduced matrices from ∂. Then $low_R(j) = low_{R'}(j)$ for all $j = 1, 2, ..., m$

Proof. Consider the lower left submatrix R_i^j of R whose corner element is $R[i,j]$. In other words, R_i^j is obtained from R by removing the first $i-1$ rows and the last $n-j$ columns. Since left-to-right column operations preserve the rank of every such submatrix, the rank of R_i^j is the same as that of the corresponding submatrix of ∂. Define

$$r_R(i,j) = \text{rank} R_i^j - \text{rank} R_{i+1}^j + \text{rank} R_{i+1}^{j-1} - \text{rank} R_i^{j-1} \tag{1.1.5}$$

Clearly $r_R(i,j) = r_{R'}(i,j) = r_\partial(i,j)$. Also the rank of R_i^j is equal to its number of non-zero columns.

Now, if $R[i,j]$ is a lowest 1, then R_i^j has one more non-zero column than R_{i+1}^j and $\text{rank} R_{i+1}^{j-1} = \text{rank} R_i^{j-1}$, implying $r_R(i,j) = 1$. If $R[i,j]$ is not a lowest 1, then are two subcases.

- If none of the columns from 1 to $j-1$ has its lowest 1 in row i, then $\text{rank} R_{i+1}^j = \text{rank} R_i^j$ and so do R_{i+1}^{j-1} and R_i^{j-1}
- If one of the columns from 1 to $j-1$ has its lowest 1 in row i, then R_i^j has one more non-zero column than R_{i+1}^j and R_i^{j-1} has one more non-zero column than R_{i+1}^{j-1}.

In eiher case $r_R(i,j) = 0$.

Since the ranks of the lower left submatrices of R are the same as those of R', we have a characterization of the lowest 1s that does not depend on the reduction process.

□

Corollary 1.1.21. For $r_\partial(i,j)$ defined in 1.1.5, $i = low(j)$ iff $r_\partial(i,j) = 1$.

Now, knowing that the lowest 1's are not an artifact of the particular strategy used for reduction it is presented its meaning. Let σ_j be a p-simplex and consider the column j in R.

- If R_j is zero then the rank of Z_p increase by one implying β_p increase by one too. Thus σ_j is positive since its addition creates gives birth to a new p-homology class.
- If R_j is non-zero, then the rank of B_{p-1} increase by one and makes that β_{p-1} decrease by one. Thus σ_j is negative because its addition gives death to a $(p-1)$-homology class.

Furthermore, it is claimed the class that dies by the addition of σ_j is born with the addition of σ_i where $i = low(j)$. In first place, if R_j is non-zero, by construction, stores the boundary of σ_j. Second, as σ_j is negative implies that the dying cycle d becomes a boundary, indeed the boundary of σ_j. Therefore,

$$\sigma_k \text{ is part of the representative cycle } d \text{ if and only if } \partial[k,j] = 1 \tag{1.1.6}$$

Thus the class did not exist before the addition of σ_i. Suppose the class is born with the addition of $\sigma_{i'}$ with $i' > i$. Using equation 1.1.6 , $\sigma_{i'}$ is not part of d, otherwise $\partial[i',j] = 1$ but contradicts $i = low(j)$. Thus, there is a cycle $d' \neq d$ such that $\sigma_{i'}$ is part of it and d and d' are homologous. This is possible if and only if there is a p-simplex σ'_j at K_i such that $\sigma_{i'}$ is part of the boundary of σ'_j, but σ'_j can not be added before $\sigma_{i'}$ since $\{K_i\}$ is a filtration, again is a contradiction. In this manner, the assertion is true.

We use the Figure 1.11 to show how this idea works. Adding 1 gives birth to a new 0-cycle, since row 1 does not contain a lowest one implies the homology class never dies. The first lowest one is in row 2 and column 5. In simple terms, the vertex 2 gives birth to the 0-cycle that the edge 5 kills. Similarly, the vertex 3 gives birth to the 0-cycle that the edge 6 kills and the vertex 4 gives birth to the 0-cycle that the edge 8 kills. Adding the edge 7 does not kill anything, it rather gives birth to a 1-cycle corresponding to the sum of 5, 6 and 7 as shown in V_7. In the same way, the edge 9 gives birth to a 1-cycle corresponding to the sum of 5, 7, 8 and 9 . Meanwhile the surface 10 kills the 1-cycle created by the addition of 7 , the last 1-cycle never dies, thus, it persists over time.

Bibliography

This section is mostly based on the book of Edelsbrunner and Harer: Computational Topology [7]. Also were used “A first introduction to abstract algebra” from Jhon B. Fraleigh [11] and “Barcodes: The persistent topology of data” by Rober Ghrist [12]

Remark

We made all the figures using the language TikZ [33] and the programming language R [9][27][10]

1.2 Application of persistent homology

The primary application of persistent homology is in data analysis, an activity that reaches into every discipline in science and engineering. Our data are the RNA secondary structures and our purpose in this work is revealing similarity information hidden in the RNA secondary structure space trough the application of persistent homology.

This application is inspired in the conference paper “Persistent Homology on RNA Secondary Structure Space” of 2015 [23] were the space of 5s rRNA foldings is clustered using the information of the 0-homological classes that represents the connected components of the data. Nevertheless, our work uses a larger data set and more distances to compare the structures. The preliminaries of the data set are presented next.

1.2.1 RNA secondary structures

Ribonucleic acid (RNA) is a molecule ubiquitous in the cell and important in various biological process as coding, decoding, regulation and expression of the genes. A RNA molecule consists of a chain of ribonucleotides linked together by covalent chemical bonds. All nucleotide contains one of the four bases: adenine (A), cytosine (C), guanine (G) or uracil (U). This linear string is called its *primary structure*. The number of ribonucleotides is called the *length* of the molecule.

Each nucleotide of the backbone can form a *base pair* following the symmetric Watson-Crick rules A-U, G-C and Wobble rule U-G. These interactions forced the molecule to fold in on itself and form a complex, three-dimensional shape called the *RNA tertiary structure* .

In order to simplify the study, the biologist just focus their attention on the base pairs involved. This collection of base pairs is referred to as its *secondary structure*. Predicting secondary structure first and then proceeding on to tertiary structure has been a fruitful, if not infallible approach.

Before continuing, a formally definition of the secondary structure is given.

Let $\mathcal{S}_n$ be the space of all possible secondary structures of a RNA sequence of length n.

Definition 1.2.1. Given $S \in \mathcal{S}_n$, the bases are numbered form 1 (called the 5’ terminus) to n (the 3’ terminus). A *secondary structure* S is a graph whose vertices $V(S)$ are the nucleotides of the RNA and the edges $E(S)$ are base pairs, such that if $\{i, j\}$ and $\{p, q\}$ are in $E(S)$ then

i) $j - i > 3$

ii) $i = p$ if and only if $j = q$

iii) $p \leq j$ implies that $i < p < q < j$ or $p < q < i < j$

An edge $\{i,j\}$ of S is noted as $i.j$ when $i < j$. Those vertices not contained in a base pair are called *unpaired.* Condition ii) implies that each vertex (i.e., nucleotide) is allowed to belong to at most one base pair and condition (iii) excludes the formation of what are called pseudoknots. The first condition is going to be explained few paragraphs later.

Due to the last conditions, the secondary structure can be decomposed into well defined substructures such that each base is contained in just on of them.

Suppose $i.j \in S$ are paired in S and $i < r < j$ and there is not a pair $s.t$ such that $i < s < r < t < j$, then i is *accesible* from $i.j$. If $p.q \in S$ such that p and q are accessible from $i.j$, then the pair $p.q$ is accessible from $i.j$. The $k-1$ pairs and u unpaired terms accessible from $i.j$ constitute the k-cycle (also k-loop) closed by $i.j$. Those sequence terms contained in no k-cycle are called external. [40]

The k-loops defined by the pair $i.j$ are classified in this form:

$k = 1$: Forms a *hairpin* loop. Note that condition i) from definition 1.2.1 implies that $u \geq 3$.

$k = 2$:

- If $u = 0$ is called a *stack.*
- If $u > 0$ and either $i + 1$ or $j - 1$ is paired but no both is called a *bulge.*
- If $u > 0$ and neither $i + 1$ or $j - 1$ is paired is called an *interior loop.*

$k > 2$: This loops are called a *multiple loop, multi-branched loop* or *multiloop*

Each of these substructures is called a motif. For illustration see the Figure 1.12.

1.2.2 Representation of secondary structures

There are many ways to represent a secondary structure S, however, two representations are of interest for this work.

Bracket Representation This compact representation was introduced in 1994 by Hofacker and colaborators [17]. For each element $i \in S$, if i is unpaired is replaced by a dot "." in the i-th position, if is not, the pair $i.j$ is replaced by "(" and ")" i-th and j-th positions, respectively. For instance there is the Figure 1.12 b).

Tree Representation: The tree representation was first described by Hofacker and collaborators [17] in 1994. This representation describes the secondary structure using a root labeled ordered tree. A tree T is called a *labeled tree* if each node is assigned a symbol from a fixed finite alphabet and is called a *ordered tree* if a left-to-right order among siblings in T is given. For illustration see the Figure 1.12 c). The root does not correspond to a part of the RNA secondary structure and for each unpaired nucleotide i corresponds a vertex whose label is $l(i)$ and each pair $i.j$ corresponds a vertex whose label is $l(i,j)$. The three is constructed using the following rules:

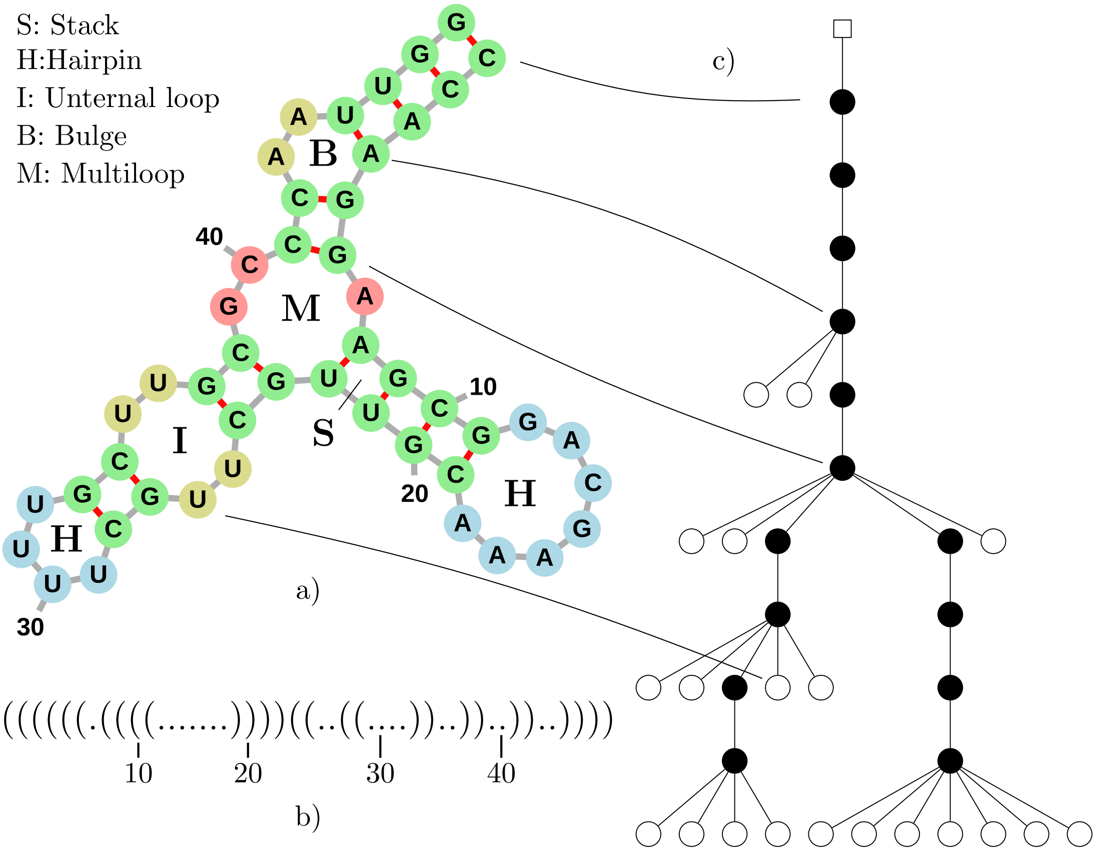


Figure 1.12: RNA secondary structure a) Structure with characteristic motifs , b) Tree representation, c) Bracket representation

i) $l(i)$ and $l(i,j)$ are children of the root if they are external.

ii) $l(i)$ is child of $l(p,q)$ if i is accessible from $p.q$.

iii) $l(i,j)$ is child of $l(p,q)$ if $i.j$ is accessible from $p.q$

In this way internal nodes correspond to base pairs, and leaves correspond to unpaired vertices. Notice that is really important the condition ii) from the definition 1.2.1 that avoid the formation of pseudoknots.

Also the siblings are ordered following the next rules:

i) $l(i) < l(j)$ if $i < j$

ii) $l(i) < l(j,k)$ if $i < j$

iii) $l(j,k) < l(i)$ if $k < i$

1.2.3 Distances between secondary structures

The computer scientists have been developed different methods in order to find the optimal secondary structure for a given RNA sequences [5], [24],[37], [39]. In general these methods are based on loop dependent energy rules where each motif has an associated energy determined experimentally and theoretically and the energy of the structure is the sum all over its motifs.

Besides, it is not hard to see that the cardinality of $\mathcal{S}_n$ grows extremely rapidly with n, in fact if $T(n)$ is the number of structures that can be formed with n nucleotides [40]

$$T(n) \sim \left(\frac{15 + 7\sqrt{5}}{8\pi} n^{\frac{-3}{2}} \left(\frac{3 + \sqrt{5}}{2} \right)^n \right)^{\frac{1}{2}}$$

Thus, even if the optimal folding is computed it does not mean that only one structure exists, it just represents the best one that fits certain parameters and may not adequately describes a real situation for two major reasons. First, the energy parameters on which the folding algorithm relies are inevitably imprecise and second the folding not only rely on the base pairs it also depends on external factors of the RNA or potential tertiary interactions that are not considered. Hence, it is necessary to considered alternative folding that are called *suboptimal structures*; structures that does not necessarily fits all the requirements; and are fundamental to compare these heterogeneous structures.

On the other hand, “RNA secondary structure comparison is essential for (i) identification of highly conserved structures during evolution (which cannot always be detected in the primary sequence, since it is often unpreserved) which suggest a significant common function for the studied RNA molecules, for instance the families form the Rfam [26], for (ii) RNA classification of various species (phylogeny) , describe in *Evolutionary genomics and systems biology* [32], (iii) identification of a consensus structure and consequently of a common role for molecules developed in [34] ” [4] and (iv) to identify non conservative mutations that produces malfunction of the RNA in different biological processes, for example the polymorphic miRNA targeting in the coronary artery disease [2].

For those and more reasons , distances have been defined on the space of secondary structures.

Base pair distance

One of the simplest metrics that one can define on $\mathcal{S}_n$ was described by Zuker in 1989 [38]. The distance between a pair S_1, S_2 in $\mathcal{S}_n$ equal to the cardinality of the symmetric difference of $E(S_1)$ and $E(S_2)$. However, it does not capture much of the secondary structure information and in general is better to be used to compare sequences of same length.

Using the bracket notation are defined two metrics:

Hamming distance

The Hamming distance was introduced by Richard Hamming[5] in 1950 [14] to measure the minimum number of substitutions in order to change one string to another. Formally

Definition 1.2.2. Let $\sum$ be an alphabet. Given $u, v \in \sum$ both of length n, the *hamming distance* between u and v is the number of place where u and v differ.

Note that Hamming is restricted to strings with same length.

Example 1.2.3.

```
.......(((((...)))))
....(((((...)))))(((
```

The Hamming distance between these strings is 12

Levenshtein distance

The Levenshtein distance was introduced in 1966 by Vladimir Levenshtein[6] [21]. Like the Hamming distance it measure the "cost" to transform one string to another but using this set of edit operations:

Insertion of a single symbol. If $a = wv$, inserting the symbol x produces wxv.

Deletion of a single symbol. If $a = wxv$, deleting the symbol x produces wv.

Substitution of a single symbol. If $a = wxv$, and $x \neq y$ changes x for y produces wyv.

And each of the operations has cost one.

Example 1.2.4.

```
.......(((((...)))))
....(((((...)))))
....(((((...)))))(((
```

1. Delete the first three "."
2. Insert the last three "("

The Levenshtein distance of the same sequences as Example 1.2.3 is 6: 3 deletions and 3 insertions.

Generally if the strings are of the same size, the Hamming distance is an upper bound on the Levenshtein distance.

And using the tree representation.

[5] Richard Wesley Hamming was a mathematician born in 1915 in Chicago U.S and die in 1998 in Monterrey U.S

[6] Vladimir Iosifovich Levenshtein was a Russian mathematician. He was born in 1935 in Moscow and die in 2017 in the same city.

Tree edit distance

Tree edit operations are defined in order to convert a labeled tree into another as shown in the next figure (Figure 1.13)

relabel Change the label of a vertex v in T.

delete Delete a non-root node v in T with parent v', making the children of v' become the children of v'. The children are inserted in the place of v as a subsequence in the left-to-right order of the children of v'.

insert The complement of delete. Insert a node v as a child of v' in T making v the parent of a consecutive subsequence of the children of v'.

Now, let T_1 and T_2 be labeled ordered trees and let $D_T(T_1, T_2)$ be the *tree edit distance.* Define an *edit script* $F = (f_0, ..., f_n)$ as a sequence of edit operations turning T_1 into T_2. Now each edit operation has an assigned cost and the cost of F, is defined as the sum of the costs of each operation f_i, meaning $cost(F) = \sum_{i=0}^{n} f_i$. At the end

$$D_T(T_1, T_2) = \min \{cost(F) \mid F \text{ is a edit script from } T_1 \text{ to } T_2\}$$

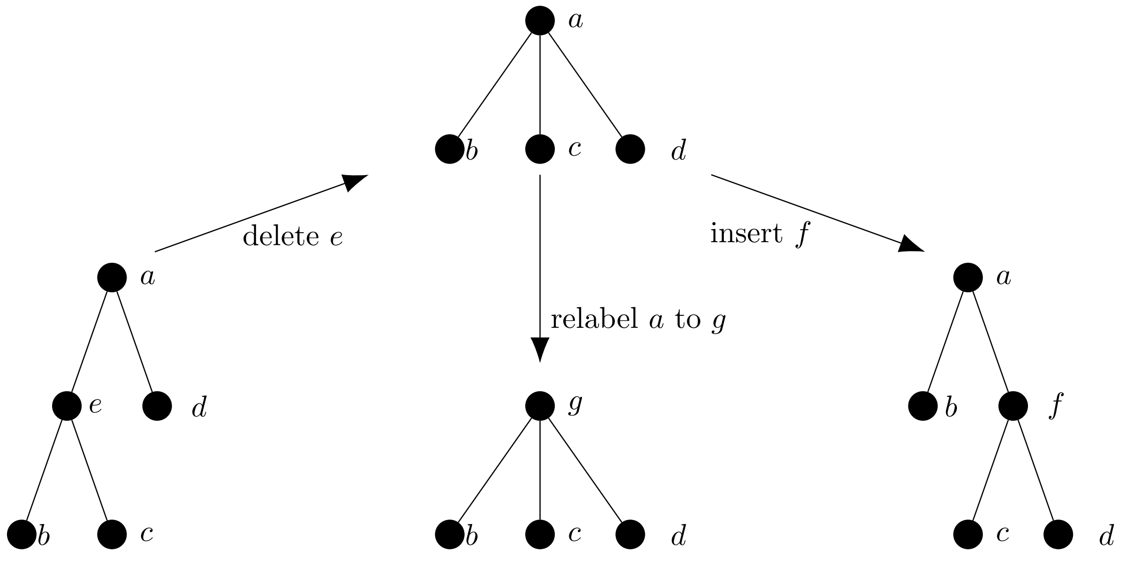


Figure 1.13: An example of tree edit operations

Bibliography

This section is mostly based on the articles: RNA secondary structures and their prediction [40] and Metrics on RNA Secondary Structures [25]

Methods and Implementation

2.1 Data and clustering

We work with the set of microRNA (miRNA) recovered from the miRbase[13]. These sequences are about 28645 between 39 and 2354 base length, then the data is filtered to select miRNA of length 110, giving a total of 775 sequences.

Afterwards, the function *RNAfold* from the *ViennaRNA Package* [22] is used to compute the minimum free secondary structure for each of the sequences. Let S be the set of all these structures, then, based on the folding structures, four distances are calculated: Tree edit distance and base pair (BP) distance using the function *RNAdistance* again from the *ViennaRNA Package* and the Hamming and the Levenshtein distance calculated from its bracket representation. Each of the distances creates a matrix N where $N[i, j] = d(i, j)$, i is the $i - th$ structure and $d(i, j)$ is the distance between i and j. Clearly $d(i, j)$=0 if and only if i and j are the same structure.

2.1.1 Clustering methods

We call "Own method" the work develop here to cluster the set S. Also, in the interest of compare the Own method, other methods of clustering are presented.

Own method

We use for the topological analysis the Statistical Tools for Topological Data Analysis (TDA) package implemented in R [10]. The function RipsDiag from TDA computes for the matrix distance N the corresponding Rips filtration and the corresponding persistence homology of S. Note that the output obtained relies completely on the chosen distance.

Using that for a simplicial complex K, each coset of $H_0(K)$ represents exactly one connected component of K (Theorem 1.1.13), each class can be seen as a cluster. If $K = \mathrm{VR}(S, r)$, the clusters in K are the clusters formed at the moment r. We highlight that the clusters are not explicit from the output of ripsDiag, but in particular gives the following information for each 0-class $[z]$:

- The birth moment and the death moment of $[z]$.

- The annihilator edge of $[z]$ that is added precisely in the death moment. The edge joins the representative of $[z]$ with the representative of the class $[z']$ that merges with $[z]$.

Observe that at $r = 0$, i and j belongs to the same homology class if and only if $d(i, j) = 0$. In the output of ripsDiag those structures were taken as an only point. Thus every 0-class is born at the moment 0. Using the above information, a script in R is implemented to find the clusters for a given r, following the algorithm defined next:

```
Data: - $S$ the set of secondary structures
- $death$ a list storing the moments of death
Result: $C$ list of 0 classes
foreach $i$ in $S$ do
    $C[i] = \{i\}$
    foreach $j \leq i$ in $S$ do
        if $d(i, j) = 0$ then
            $C[j] = C[j] \cup C[i]$
            $C[i] = \{\}$
            break
        end
    end
end
foreach $r' \leq r$ in $death$ do
    foreach annihilator edge $(e_1, e_2)$ in moment $r'$ do
        $C[e_2] = C[e_1] \cup C[e_2]$ $C[e_1] = \{\}$
    end
end
```

Finally, as for each r there is a different clustering we want the best one. The clusterings that have too many clusters or few clusters are discarded. The remaining clusterings are evaluated using the criteria shown in subsection(2.2.1). For this method the best clustering must have a non negative Average Silhouette Width and the best Dunn index uses the diameter of a cluster the complete distance. This last choice of diameter is going to be well explained in the chapter of Discussion.

Unweighted Pair Group Method with Arithmetic Mean (UPGMA)

Attributed to Sokal[7] and Michener[8] [31] is an agglomerative hierarchical clustering method.

[7] Robert Reuven Sokal was an Austrian biostatistician and entomologist born in Vienna in 1926 and die in Stony Brook U.S

[8] Charles Duncan Michener was an US entomologist born in Pasadena in 1918 and die in Lawrence in 2015

In this method, the distance between any two clusters C and D with sizes $|C|$ and $|D|$ is defined as the average distance between the elements in C and D:

$$d(C,D) = \frac{1}{|C| \cdot |D|} \sum_{i \in C} \sum_{j \in D} d(i,j)$$

And works as follows:

1. Each cluster is made with structures whose distances between each other is 0.
2. Calculate $\mathcal{D} = \{d(C,D) \mid C, D \text{ clusters } \}$.
3. Take $C' = C_0 \cup D_0$ as the new cluster if $d(C_0, D_0) = \min \mathcal{D}$ and go back to step 2.
4. Finish if there is just one cluster.

As the "Own method" each step creates a different form of clustering S. Thus, the best clustering is chosen in the same way as before but the Dunn index uses the average distance as the form to calculate the diameter of a cluster.

To cluster S using the UPGMA method is used the function hclust implemented in R.[27]

Partitioning Around Medoids (PAM)

In 1987 Kaufman and Rousseeuw[9] [20] introduced this centroid-based clustering.
PAM uses a greedy search , might be not optimal but it is faster and works as follows:

1. Select k of the structures as the medoids. It may be chosen as the most central elements.
2. Associate each structure to the closest medoid. Calculate the **cost** of this clustering as the sum of the distances between each structure and its medoid.
3. For each medoid m and for each non medoid n, make n the new medoid and recompute the cost , if the cost increases undo the swap.
4. Continue until there are not more possible reductions.

It is clear that choice of k is fundamental to get the best clustering. In particular, k can be computed for a range of numbers and choose the clustering with the best silhouette average width; defined in the next section (2.2.1). The function pamk implemented in the package fpc in R [16] is used to applied PAM method on S.

2.2 Validation

As for each ratio of the filtration there are different clusters it is important to choose the best clustering. Assume S have been clustered into k clusters with $k > 1$, avoiding trivial clusterings. We use two methods for internal cluster validation: the *Average Silhouette Width* (ASW) and the *Dunn Index* (DI) in order to choose the best k's in each case.

[9] Peter J. Rousseeuw is a Belgian statistician born in Wilrijk, Belgium in 1956

2.2.1 Internal cluster validation

When a clustering result is evaluated based on the data that was clustered itself, is called internal evaluation. Are presented two criteria:

Average silhouette width

This method for interpretation and validation of cluster analysis was proposed by Peter Rousseeuw[10] in 1986 [29]. Let i be a structure of S and let C be the cluster such that $i \in C$. Define:

$$a(i) = \frac{1}{|C| - 1} \sum_{\substack{j \in C \\ i \neq j}} d(i, j)$$

Simply, $a(i)$ is the average distance between s_i and the other structures in C. Thus, $a(i)$ can be interpreted as a measure of how well i is assigned to C; the smaller the value, the better its assignment.

Then, for any cluster D distinct of C define:

$$d(i, D) = \frac{1}{|D|} \sum_{j \in D} d(i, j)$$

And take

$$b(i) = \min\{d(i, D) \mid D \text{ cluster}\}$$

The cluster D that attaches the minimum is said to be the *neighbor cluster* of i because it is the next best cluster for i.

With these elements the *silhouette index* for i is defined as

$$s(i) = \frac{b(i) - a(i)}{max\{a(i), b(i)\}} = \begin{cases} 1 - a(i)/b(i), & \text{if } a(i) < b(i) \\ 0, & \text{if } a(i) = b(i) \\ b(i)/a(i) - 1, & \text{if } a(i) > b(i) \end{cases}$$

If $s(i)$ is close to 1, then $a(i) << b(i)$. A small $a(i)$ means it is well matched and a large $b(i)$ implies i is badly matched to its neighbor cluster. Thus, an $s(i)$ close to 1 means that the structure is appropriately clustered. In the same way, $s(i)$ is close to -1 if its neighbor cluster represents a better matching. A $s(i)$ near zero means that i is on the border of two natural clusters.

Finally, the *average silhouette width* is define as the average of $s(i)$ for all structures of S. The ASW might be used to select an appropriate number of clusters trying to maximize its value.

[10] Peter Rousseeuw is a statician born in 1956 in Wilrijk, Belgium

Dunn Index

The Dunn index was defined by J. C. Dunn in 1973 [6] . Let C_i and C_j be clusters of S and define the inter-cluster distance between C_i and C_j, in particular:

- $d(C_i, C_j) = \min\{d(s,t) \mid s \in C_i, t \in C_j\}$, is the distance between the closest two points. It is called the single distance.
- $d(C_i, C_j) = \dfrac{1}{|C_i||C_j|} \displaystyle\sum_{s \in C_i, t \in C_j} d(s,t)$, is the average distance between the elements of C_i and C_j. It is called the average distance.

In the same way, the diameter of C_i is defined, noted as δ_i.

- $\delta_i = \max\{d(s,t) \mid s,t \in C_i\}$, is the distance between the farthest two points.
- $\delta_i = \dbinom{|C_i|}{2}^{-1} \displaystyle\sum_{s,t \in C_i} d(s,t)$, is the average distance between the elements of C_i.

With the above notation, if there are k clusters, then the Dunn Index (DI) is defined as:

$$\mathrm{DI}(S) = \frac{\min_{1 \leq i \leq j \leq k} \; d(C_i, C_j)}{\max_{1 \leq i \leq k} \; \delta_i}$$

2.2.2 External cluster validation

In external evaluation, clustering results are evaluated based on other clustering that is assumed to be the gold standard. To do that let $\mathcal{C}$ and $\mathcal{D}$ two forms of clustering S and partition the set of pairs (i,j) of S:

- A=$\{(i,j) \mid$ i and j are in the same cluster in $\mathcal{C}$ and in the same cluster in $\mathcal{D}\}$.
- B=$\{(i,j) \mid$ i and j are in different clusters in $\mathcal{C}$ and in different clusters in $\mathcal{D}\}$.
- C=$\{(i,j) \mid$ i and j are in the same cluster in $\mathcal{C}$ and in different clusters in $\mathcal{D}\}$.
- D=$\{(i,j)$ — i and j are in different clusters in $\mathcal{C}$ and in the same cluster in $\mathcal{D}\}$.

And take

- $\mathrm{TP}(\mathcal{C}, \mathcal{D}) = |A|$, (true positives)
- $\mathrm{TN}(\mathcal{C}, \mathcal{D}) = |B|$, (true negatives)
- $\mathrm{FP}(\mathcal{C}, \mathcal{D}) = |C|$, (false positives)
- $\mathrm{FN}(\mathcal{C}, \mathcal{D}) = |A|$, (false negatives)

If the context is clear $\mathrm{TP}(\mathcal{C}, \mathcal{D})$ is simply TP, analogous with TN, FP and FN.

Intuitively, $\mathrm{TP} + \mathrm{TN}$ can be considered as the number of agreements between $\mathcal{C}$ and $\mathcal{D}$ meanwhile $\mathrm{FP} + \mathrm{FN}$ as the number of disagreements, note that the definition of TP and TN is symmetric respect to $\mathcal{C}$ and $\mathcal{D}$ but not for FP and FN; this is particular useful when $\mathcal{D}$ is considered a gold standard.

Using the above classification of the pairs of S we define two criteria:

Jaccard Index

Paul Jaccard[11] developed the Jaccard index of similarity in 1901 [19]. The Jaccard index J of $\mathcal{C}$ and $\mathcal{D}$ is defined as:

$$\mathrm{J}(\mathcal{C},\mathcal{D}) = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}+\mathrm{FN}}$$

Rand Index

The Rand Index was introduced by William Rand in 1971 [28] as a criteria of evaluation of clustering methods. The Rand index of $\mathcal{C}$ and $\mathcal{D}$ is defined as:

$$\mathrm{Rand}(\mathcal{C},\mathcal{D}) = \frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{TN}+\mathrm{FP}+\mathrm{FN}}$$

2.2.3 Random sets

To find out how sensitive is the algorithm to calculate the properties of persistent homology, we want to know how it behaves with a random set of structures. The worst behavior with the random set , the better the method is.

A set M is created from S mutating each sequence in S. The mutation is made choosing 5 random positions between the positions 21 to 40 and replacing the bases in the positions with random bases. In this way, each structure i in S has it corresponding mutated structure i in M, thus, $|M| = |S|$. Afterwards, the process of clustering we made for the original sequences is applied again on M in order to cluster this new set and proceed to compare with the clusters in S for each distance.

M doesn't seem as a random set since M clearly depends on S, but an "aggressive" random set of structure might not have biological sense since they are supposed to be RNA molecules.

False positive rate

To compare M and S is going to be used a variation of the false positive rate.

Take some clustering for M and S. Each pair of elements (i, j) is classified in one of this groups:

TP: The pair is in the same cluster in S and different clusters in M.

TN: The pair is in different clusters in S and same cluster in M.

FP: The pair is in the same cluster in S and same clusters in M.

FN: The pair is in different clusters in S and different clusters in M.

[11] Paul Jaccard was a professor of botany and plant physiology. He was born in Sainte-Croix in 1868 and die in Zurich in 1944

Intuitively a pair (i, j) is consider false if it coincides with the original set, meaning the method find the "right" clustering despite the randomization and it is true if it differs. In addition, is considered positive if i and j are in the same cluster in S and negative if not.

The false positive rate (FPR) is defined by:

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TP} + \mathrm{TN}}$$

It is taken in this way because in FP, TP, TN, each counted pair is either in S or M in the same cluster, meaning the method took a decision of clustering . With a better look in the partition made in subsection 2.2.2, FPR coincides with the Jaccard index, swapping the false ones with the positive ones.

Results

Previously we mentioned the initial data set is the set of 775 miRNA's of length 110 . After we computed the set S of foldings and we calculated each distance we found that for 223 sequences there is at least one distinct whose distance form the first one is 0, meaning they have the same secondary structure. The larger set of sequences with the same structure is 16. The set S has 606 structures, confirmed by the number of connected components in the first step of the Rips filtration for each distance.

3.1 Hamming Distance

3.1.1 Clustering

The non zero distances between the structures in S varies between 2 and 94. The figure 3.14 shows the barcode corresponding to 0-homology of S.

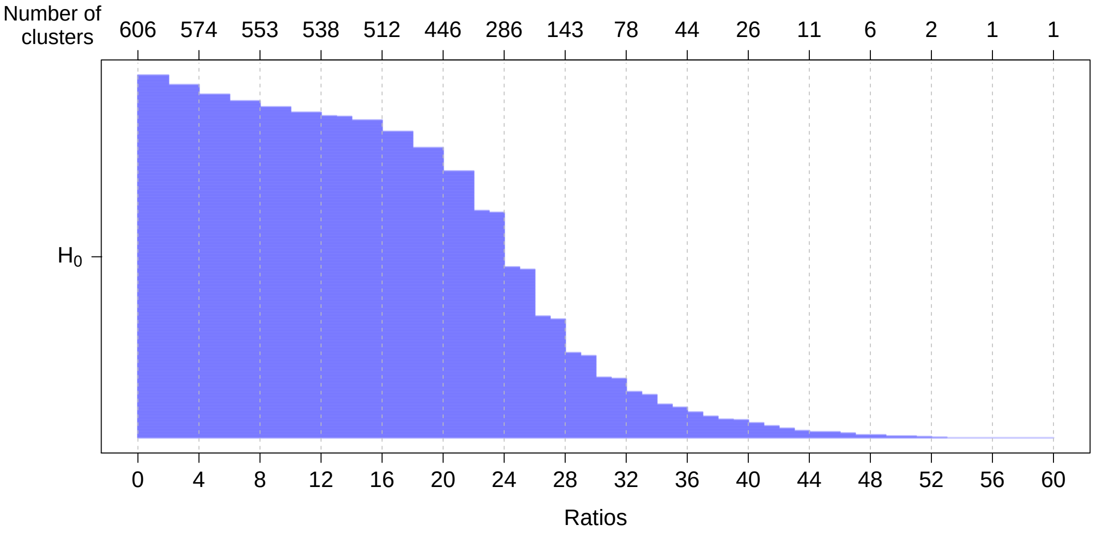


Figure 3.14: Barcode diagram of the metric space S using the hamming distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

We call a ratio **representative** if is 0 or is the death point for some 0-class, subsequently the clustering for each representative ratio is computed. The total of representative ratios otained is 41 and the first ratio when it starts to be an only cluster is 53.

Some of the clusterings are illustrated in the figure 3.15. Each color represents a different cluster and the difference between where it ends and where it starts is the percentage of elements that are in the cluster.

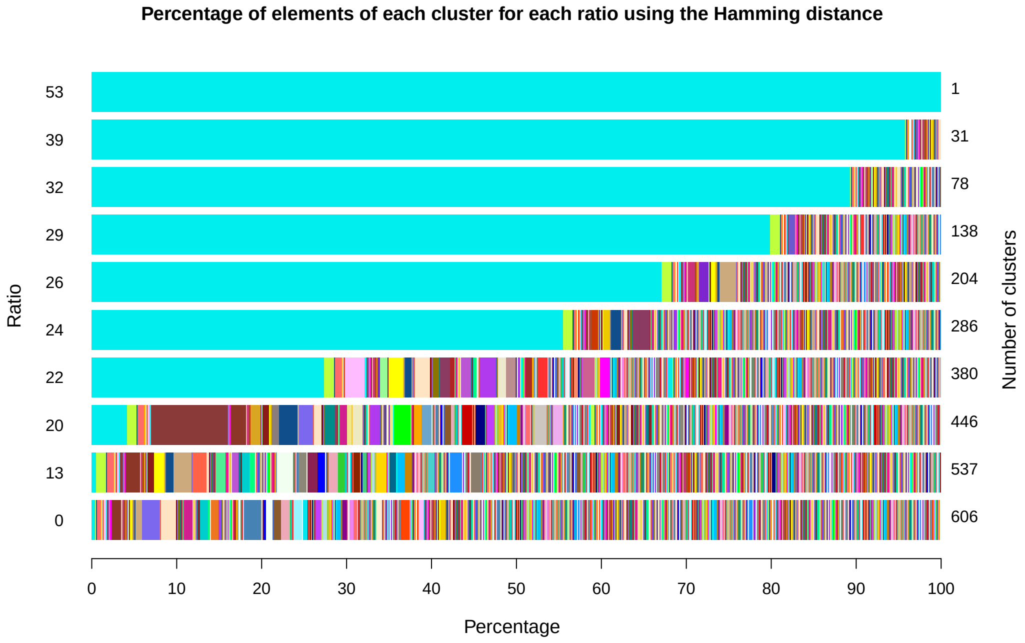


Figure 3.15: Percentage of elements for each cluster for some representative ratios. The number of clusters for each ratio are in the right side of the graphic.

3.1.2 Validation

Following the indications in Methods we choose the best clustering and its information is described in the table 3.1:

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
22	380	27.35%	0.1346	0.2191	0.2116

Table 3.1: Information of best clustering. a means average, c means complete, s means single, and a/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter.

It is interesting to point here that this cluster is in the 9th place respect to the Dunn index but is the first with non-negative ASW.

On the other side, we computed the reference clusterings using the methods UPGMA and PAM. Some of the clusterings of the method UPGMA are illustrated in Figure 3.16. The comparison between methods is shown graphical (Figure 3.17)and numerical (Table 3.2).

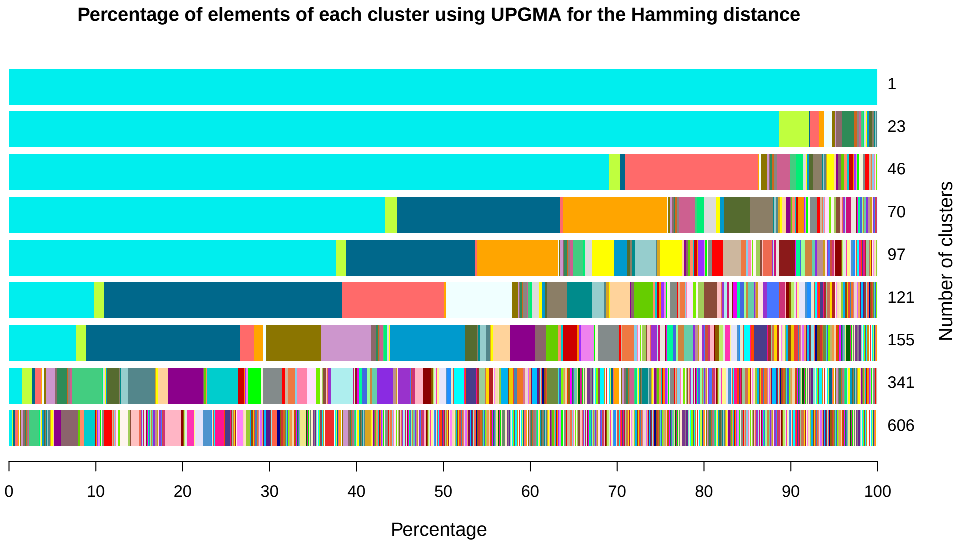


Figure 3.16: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

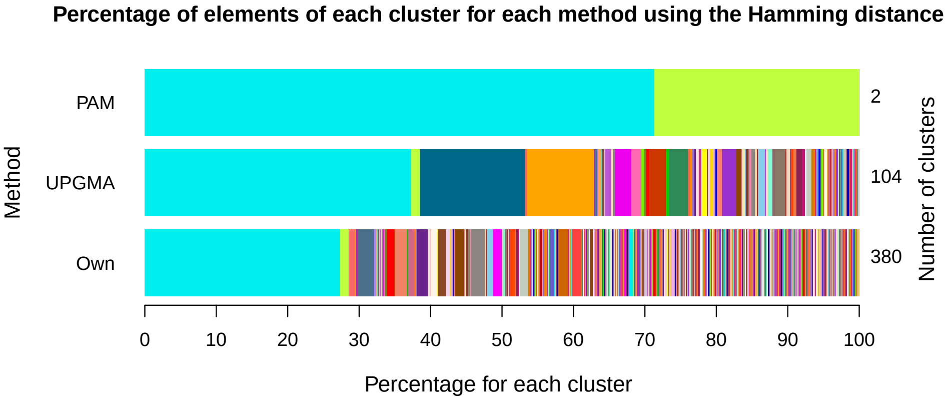


Figure 3.17: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

	TP	TN	FN	FP	Rand Index	Jaccard Index
UPGMA	13563	238609	38122	9631	0.841	0.221
PAM	15801	115373	161358	7393	0.437	0.086

Table 3.2: Comparison between best clustering from Own method against best clusterings from UPGMA and PAM

On the other hand, the set of mutants M is computed. The distances between the structures in M varies between 2 and 110. The figure 3.18 shows the barcode corresponding to 0-homology of M.

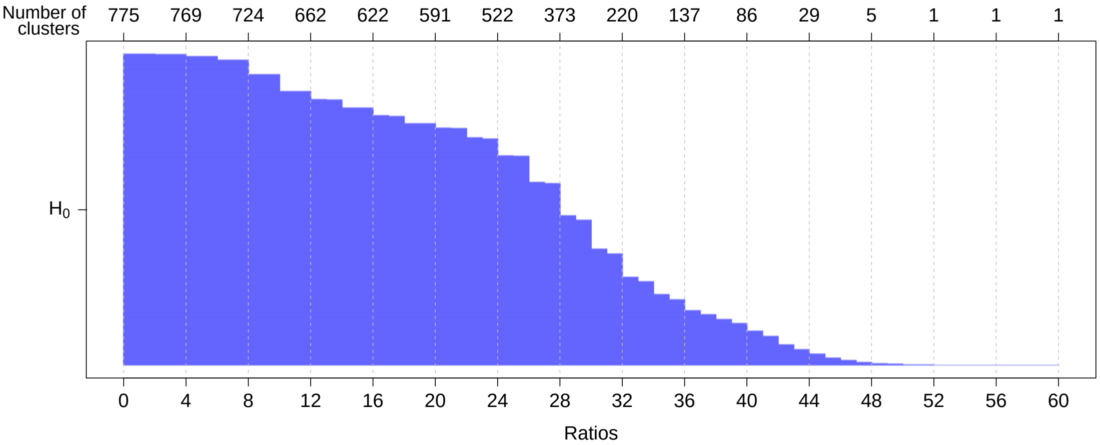


Figure 3.18: Barcode diagram of the metric space M using the hamming distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

The total of representative ratios is 44 and the first ratio when there is an only cluster is 52.

As the best clustering has a total of 380 clusters, from the mutated sequences we chose the clusterings with almost similar number of clusters and for each is calculated the false positive rate in order to pick the one with the worst FPR (Table 3.3).

Different mutant clusterings

FPR	0.061	0.061	0.232	0.235	0.205	0.197	0.153	0.146	0.123 0.114

Table 3.3: FPR between the best clustering of S and some clusterings with almost similar number of clusters of M

Based on Table 3.3 we chose the fourth clustering of that list; we call this clustering the analogous clustering. The information of both is shown in the Table 3.4 :

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
27	453	27.87 %	−0.0087	0.2687	0.2687

Table 3.4: Information of analogous clusterings.

And the best clustering is compared with the analogous clustering. The information is described in Table 3.5.

FP	FN	TN	TP	FPR
8895	262076	14299	14655	0.235

Table 3.5: Comparison between best clustering and analogous clustering.

3.2 Levenshtein Distance

3.2.1 Clustering

The non zero distances between the structures in S varies between 2 and 82. The figure 3.19 shows the barcode corresponding to 0-homology of S.

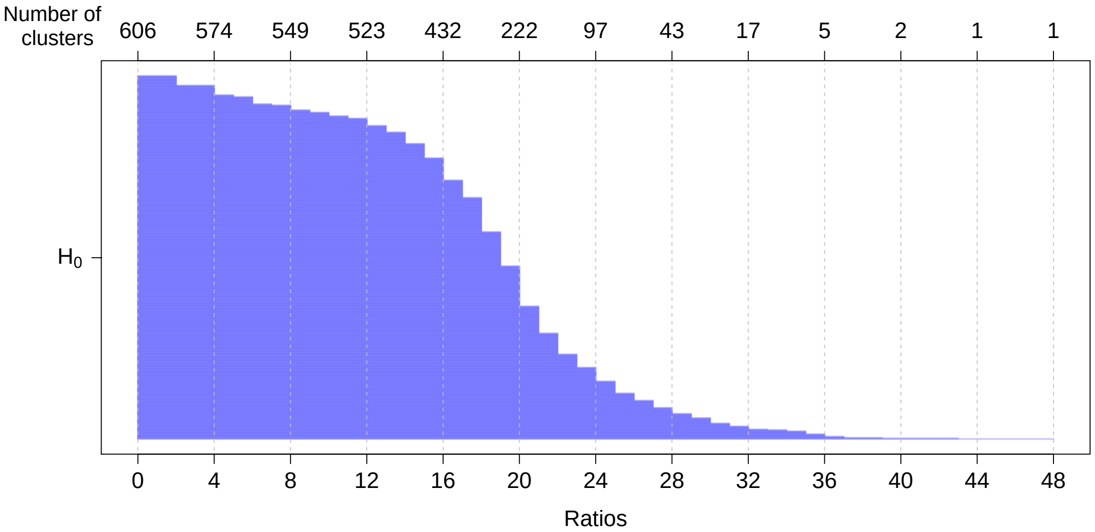


Figure 3.19: Barcode diagram of the metric space S using the levenshtein distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

Subsequently the clustering for each representative ratio was computed. The total of representative ratios is 39 and the first ratio when it starts to have an only cluster is 43.

Some of the clusterings are illustrated in the figure 3.20.

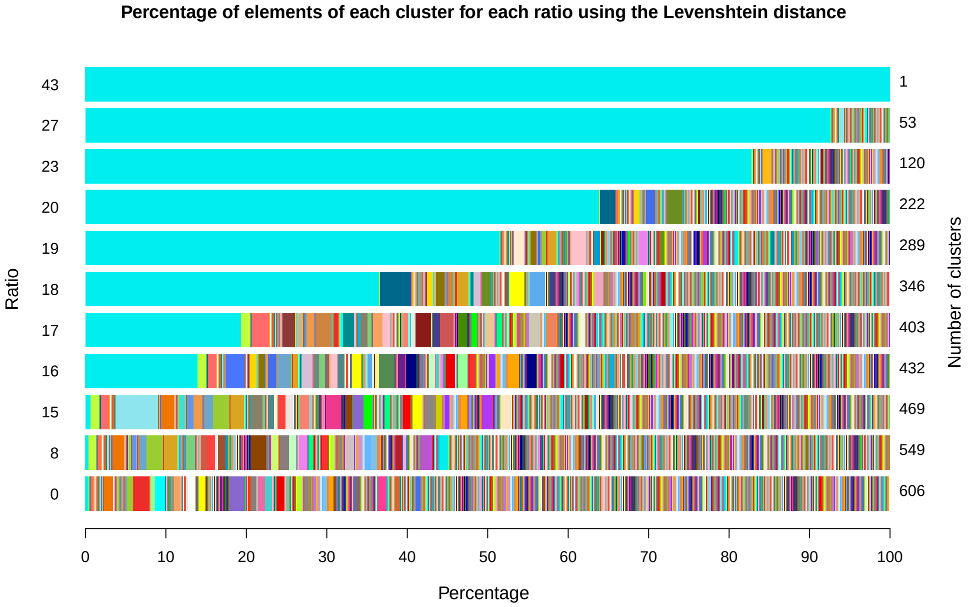


Figure 3.20: Percentage of elements for each cluster for some representative ratios. The number of clusters for each ratio are in the right side of the graphic.

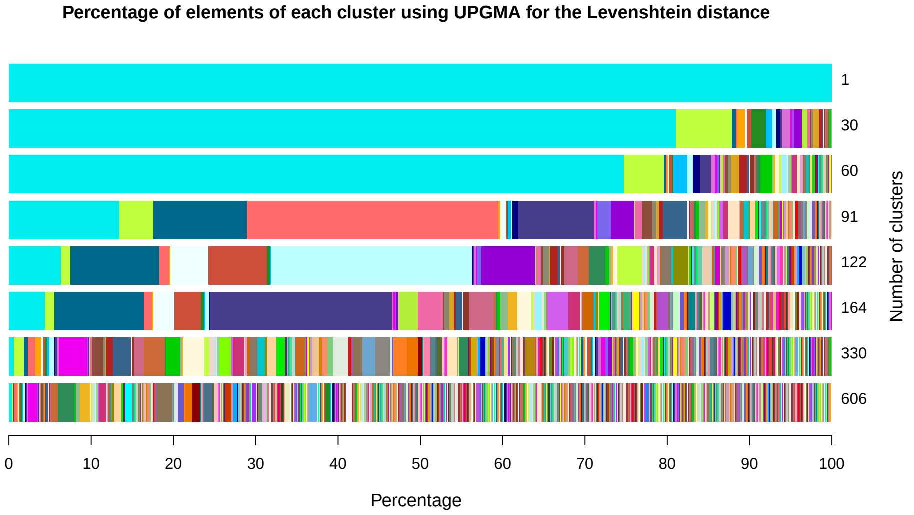


Figure 3.21: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

3.2.2 Validation

We chose the best clustering for this distance and its information is described in table 3.6:

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
17	403	19.35%	0.1603	0.2339	0.2298

Table 3.6: Information of best clustering. a means average, c means complete, s means single, and a/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter

On the other side we computed the reference clusterings using the methods UPGMA and PAM. As before some of the clusterings of UPGMA are shown in Figure 3.21 . Their comparison is shown graphical (Figure 3.22)and numerical (Table 3.7) below.

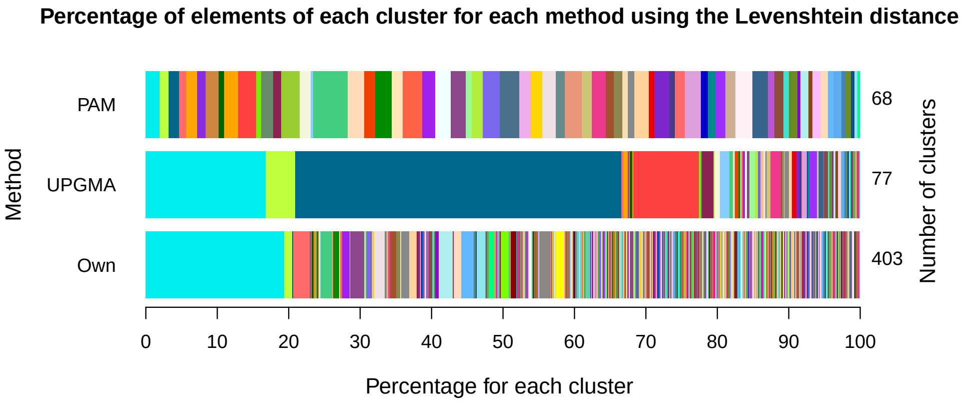


Figure 3.22: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

	TP	TN	FN	FP	Rand Index	Jaccard Index
UPGMA	8631	222177	65568	3549	0.77	0.111
PAM	1668	284242	3503	10512	0.953	0.106

Table 3.7: Comparison between best clustering from Own method against best clusterings from UPGMA and PAM. The order of the comparisons are (UPGMA,Own) and (PAM,Own)

On the other hand, the set of mutants M was computed. The distances between the structures in M varies between 2 and 79. The figure 3.23 shows the barcode corresponding to 0-homology of M.

The total of representative ratios is 39 and the first ratio when it is an only cluster is 39.

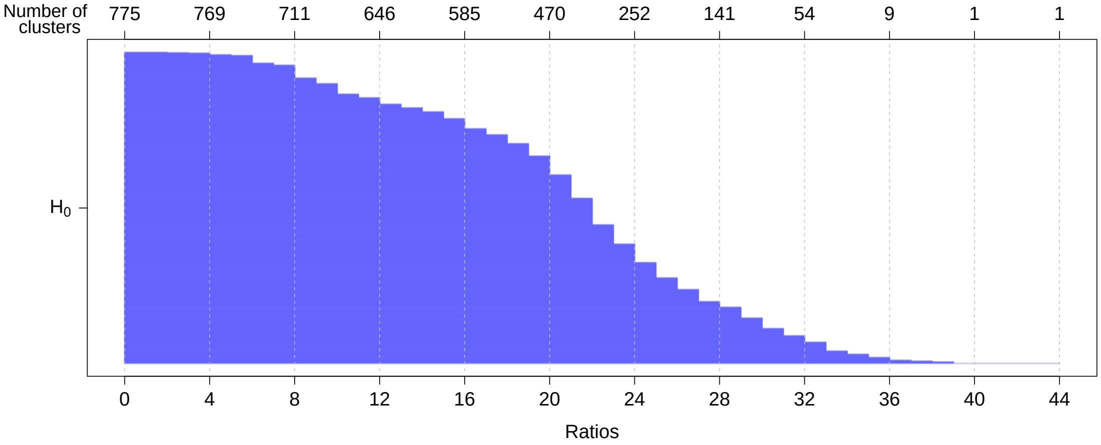


Figure 3.23: Barcode diagram of the metric space M using the levenshtein distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

Since the best clustering has a total of 403 clusters from the mutated sequences we chose the clusterings with almost similar number of clusters and for each is calculated the false positive rate in order to pick the one with the worst FPR (Table 3.8).

Different mutant clusterings

FPR	0.056	0.062	0.11	0.171	0.169	0.124	0.1	0.083	0.069	0.062

Table 3.8: FPR between the best clustering of S and some clusterings with almost similar number of clusters of M

Based on Table 3.8 the fourth clustering of that list was chosen; it is called the analogous clustering. Its information is shown in the Table 3.9.

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
19	470	30.32 %	−0.0484	0.2083	0.1919

Table 3.9: Information of analogous and best mutant clustering. s means single, c means complete, and s/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter

Finally, the best clustering was compared with the analogous mutant clustering as shown in Table 3.10.

FP	FN	TN	TP	FPR
2642	284494	9538	3251	0.169

Table 3.10: Comparison between best clustering and analogous clustering.

3.3 Base pair Distance

3.3.1 Clustering

The non zero distances between the structures in S varies between 1 and 97. The figure 3.24 shows the barcode corresponding to 0-homology of S.

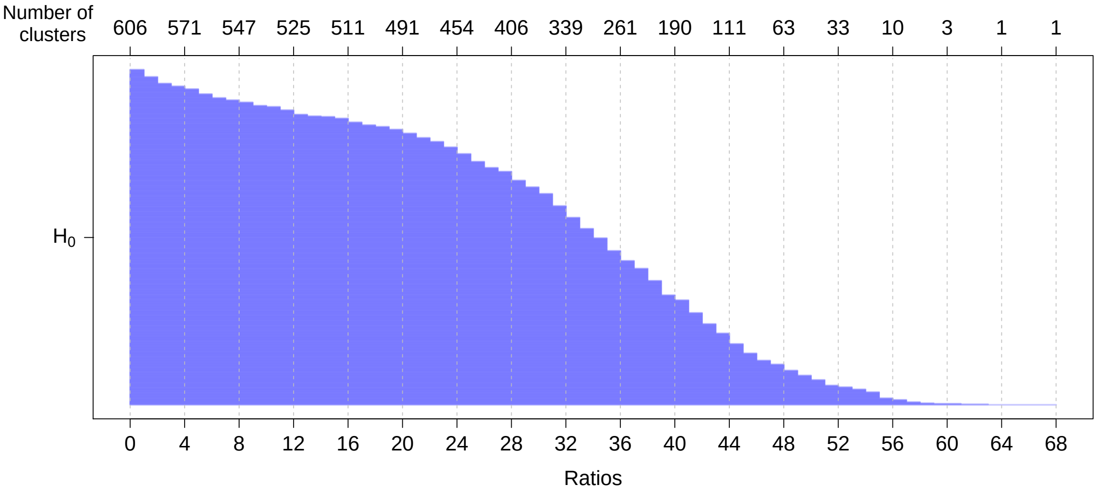


Figure 3.24: Barcode diagram of the metric space S using the base pair distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

Afterwards we computed the clustering for each representative ratio. The total of representative ratios is 62 and from the ratio 63 the clusterings have an only cluster.

Some of the clusterings are illustrated in the figure 3.25.

3.3.2 Validation

Next, the best clustering was chosen. The information of the clustering is found below (Table 3.11):

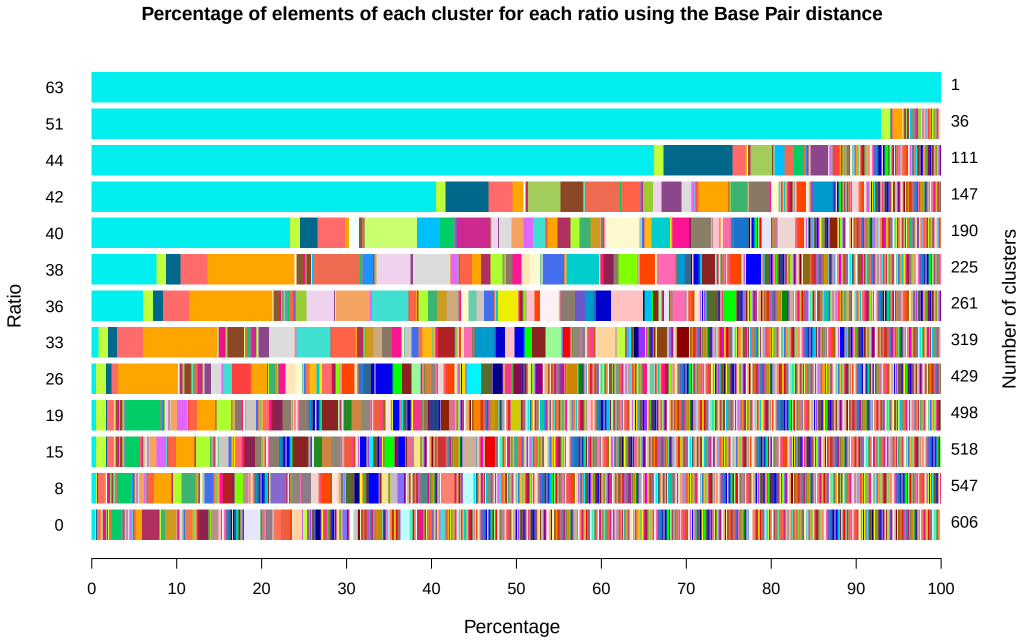


Figure 3.25: Percentage of elements for each cluster for some representative ratios. The number of clusters for each ratio are in the right side of the graphic.

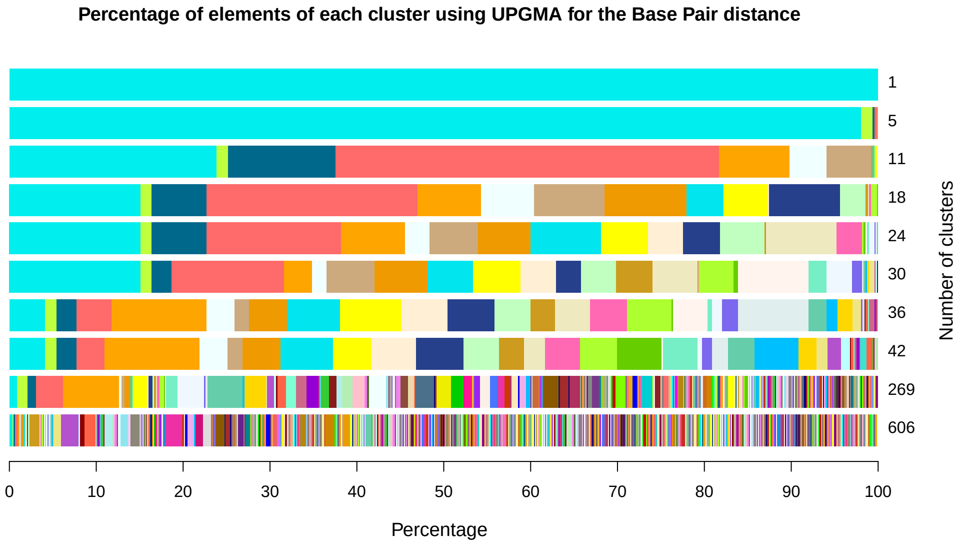


Figure 3.26: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
38	225	10.19%	0.2002	0.1612	0.1612

Table 3.11: Information of best clustering. a means average, c means complete, s means single, and s/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter.

On the other hand we computed the reference clusterings using the methods UPGMA and PAM. Some of the clusterings of the method UPGMA are illustrated in Figure 3.26. The comparison of the methods is shown in a graphical (Figure 3.27) and numerical (Table 3.12) way.

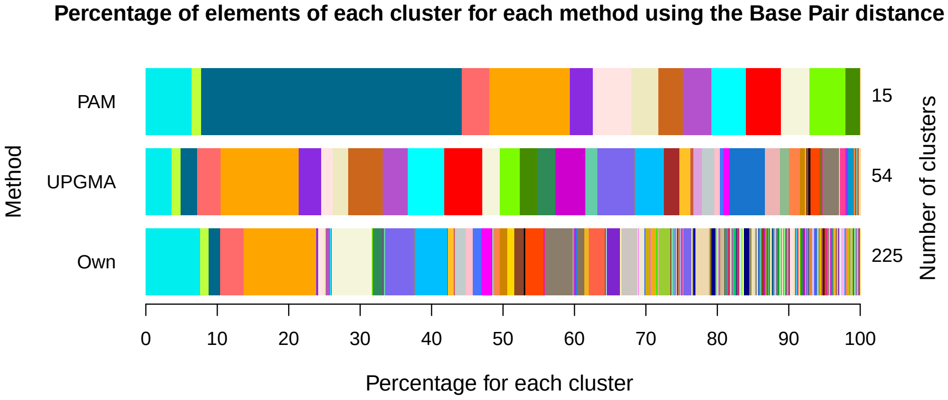


Figure 3.27: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

	TP	TN	FN	FP	Rand Index	Jaccard Index
UPGMA	7135	286551	4592	1647	0.979	0.533
PAM	6878	247594	43549	1904	0.848	0.131

Table 3.12: Comparison between best clustering from Own method against best clusterings from UPGMA and PAM

On the other side, we computed the set of mutants M. The distances between the structures in M varies between 1 and 93. The figure 3.28 shows the barcode corresponding to 0-homology of M.

The total of representative ratios was 60 and the first ratio when there is an only cluster is 65.

After this, from the mutated sequences were chosen the clusterings with a number of

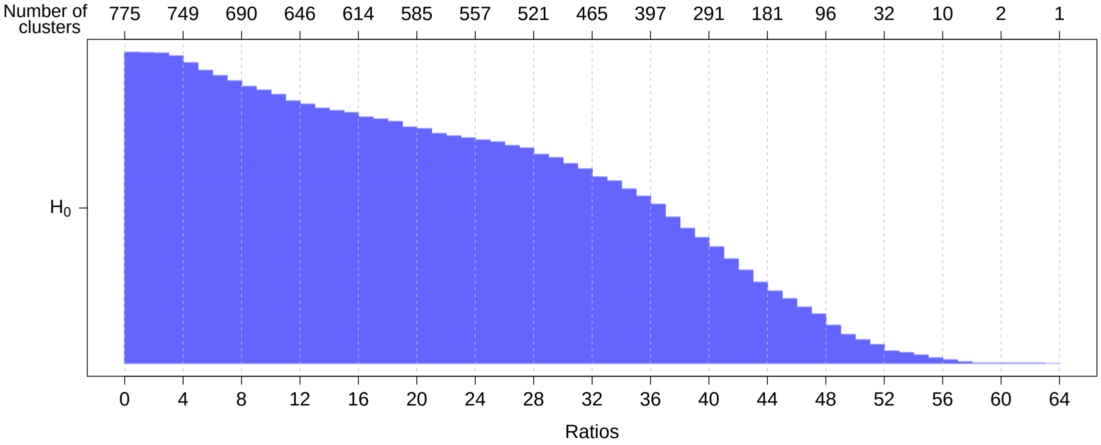


Figure 3.28: Barcode diagram of the metric space M with the base pair distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters.

clusters close to 225 and for each is calculated the false positive rate in order to pick the one with the worst FPR (Table 3.13).

Different mutant clusterings

FPR	0.295	0.308	0.324	0.277	0.181	0.159	0.124	0.069	0.049	0.044

Table 3.13: FPR between the best clustering of S and some clusterings with almost similar number of clusters of M

Using the information of the Table 3.13 we chose the third clustering of the list above. We call this clustering the analogous clustering. The information of the clustering is shown in the Table 3.14:

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
38	337	13.41 %	0.1038	0.118	0.118

Table 3.14: Information of analogous clustering.

And the best clustering was compared with the analogous clustering in and numerical way (Table 3.15).

FP	FN	TN	TP	FPR
2642	284494	9538	3251	0.324

Table 3.15: Comparison between best clustering and analogous clustering.

3.4 Tree edit distance

3.4.1 Clustering

The non zero tree edit distances between the structures in S varies between 2 and 148. The figure 3.29 shows the barcode corresponding to 0-homology of S.

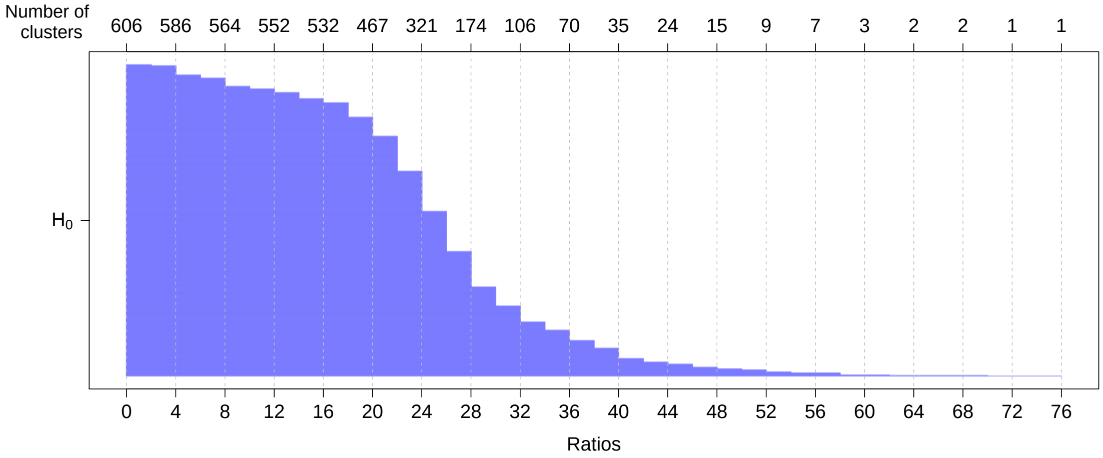


Figure 3.29: Barcode diagram of the metric space S using the tree edit distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters

Then, the clustering for each representative ratio is computed. The total of representative ratios is 31 and from 70 the clustering has an only cluster.

Some of the clusterings are illustrated in the figure 3.30.

3.4.2 Validation

Afterwards, as with the other distances, we chose the best clustering. The information of the clustering is found below (Table 3.16):

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
20	467	6.06%	0.2737	0.2372	0.2372

Table 3.16: Information of best clustering. a means average, c means complete, s means single, and s/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter.

On the other hand the best clusterings were computed using the methods UPGMA and PAM. The method UPGMA is illustrated in Figure 3.31 Its comparison is shown in a graphical (Figure 3.32) and numerical (Table 3.17) way.

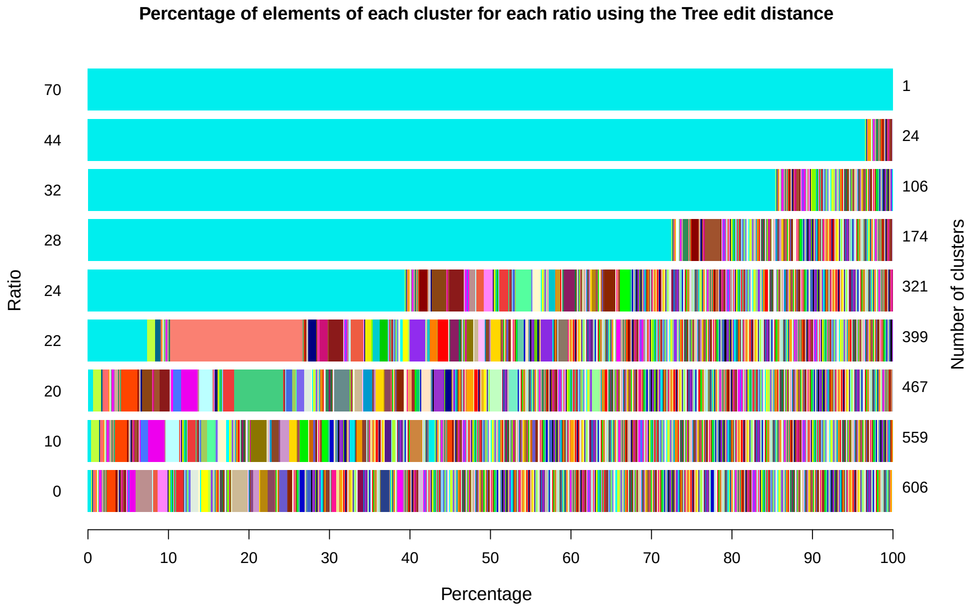


Figure 3.30: Percentage of elements for each cluster for some representative ratios. The number of clusters for each ratio are in the right side of the graphic.

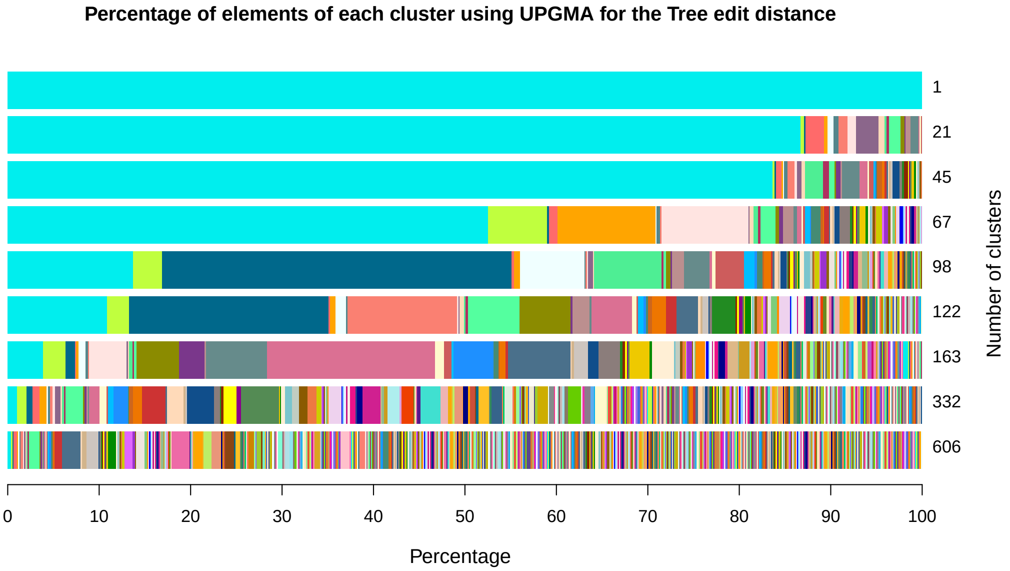


Figure 3.31: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

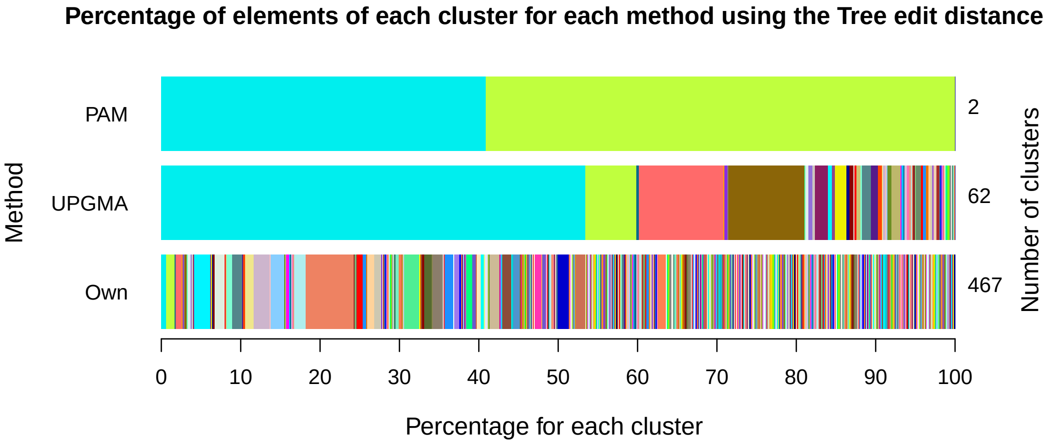


Figure 3.32: Percentage of elements for each cluster for Own, UPGMA and PAM method.The number of clusters for each method are in the right side of the graphic.

	TP	TN	FN	FP	Rand Index	Jaccard Index
UPGMA	2170	206773	90959	23	0.697	0.023
PAM	2014	145007	152725	179	0.49	0.013

Table 3.17: Comparison between best clustering from Own method against best clusterings from UPGMA and PAM

On the other side, the set of mutants M was computed. The distances between the structures in M varies between 4 and 158. The figure 3.33 shows the 0-classes of M using the barcode representation.

The total of representative ratios is 31 and the ratio when starts to have an only cluster is 66.

Afterwards,from the mutated sequences we chose the clusterings with a number of clusters close to 467 and for each is calculated the false positive rate in order to pick the one with the worst FPR (Table 3.18).

Different mutant clusterings									
0.14	0.187	0.2	0.24	0.248	0.143	0.07	0.031	0.022	0.016

Table 3.18: FPR between the best clustering of S and some clusterings with almost similar number of clusters of M

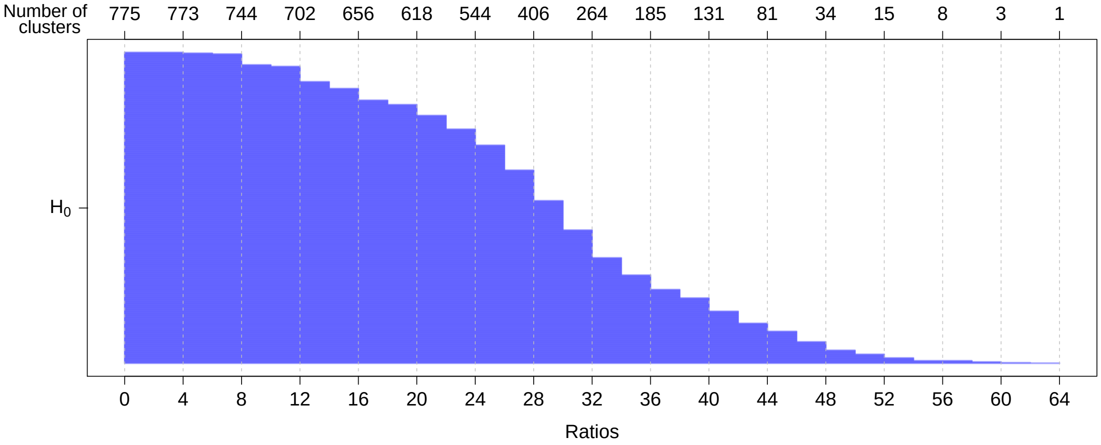


Figure 3.33: Barcode diagram of the metric space M using the tree edit distance. On bottom are the ratios corresponding to the Rips filtration and on top their corresponding β_0 that corresponds to the number of clusters.

Using the information of the Table 3.18 the fifth clustering of that list was chosen. This clustering is called the analogous clustering. The information of the clustering is shown below in the Table 3.19

ratio	# of clusters	Per. of greater cluster	ASW	DI a/c	DI s/c
22	584	3.87 %	0.0897	0.2524	0.2222

Table 3.19: Information of analogous clustering. a means average, c means complete, and a/c means the Dunn Index is calculated using average distance as inter-cluster distance and complete distance as diameter

Finally, the best clustering was compared with the analogous clustering. See the Table 3.20).

FP	FN	TN	TP	FPR
624	297412	1569	320	0.248

Table 3.20: Comparison between best clustering and analogous clustering.

Discussion

Is observed that even though the original set of data has 775 sequences, for all distances, the maximum number of clusters is 606. This happens since calculating secondary structures is not an injective function, it is not true that each folding comes from a unique sequence, even more there is an exponential number of them [40] .

From now are discussed the results obtained for the **hamming , levenshtein and tree edit distances** since their results are similar.

Observing the number of clusters as a function of the ratio (Figures 3.14, 3.19 and 3.29) , clearly its slope is non positive as it is a decreasing function. In the extremes the slope can be greater than -4 but decreases in a vertiginous way , reaching an average slope of -30 in the middle part of the total range. At the same time, the diminution of the number of clusters is accompanied by the formation of a large cluster that grows bigger as r increases as we can see in Figures 3.15, 3.20 and 3.30

Consider again the fundamentals . It is clear that an edge (i, j) belongs to $\text{VR}(S, r)$ if $d(i, j) \leq r$, then the clusters(classes) $[z]$ and $[z']$ merges at step r if $i \in [z]$ and $j \in [z]$.

For instance, consider the cluster $C = \{i_1, i_2, ...i_m\} \subset S$ such that i_k is a structure of length $m >> 1$. Using the hamming distance might be the case that $d(i_j, i_{j+1}) = 1$ with the difference in the j-position of the bracket representation for all $j = 1, 2, ..., n-1$,then, $d(i_1, i_m) = m - 1$. Thus , C is a connected component of $\text{VR}(S, 1)$, but $d(i_1, i_m) >> 1$.

We call this behavior the chaining problem: a chain of points can be extended for long distances without regard to the overall shape of the emerging cluster, also, that is why the complete distance is chosen to measure the diameter of a cluster in the Dunn Index, it this way, the index penalties the cluster if it spreads out too much.

Going back to the clusterings, it can be deduced:

- The three distances reach an only cluster in the half of its maximum distance. Thus even the most apart sequences are connected with a path of structures with distance pair by pair less than the half of the maximum distance. At the same time, almost all structures are connected with only the third part of the maximum distance.

- Besides the big cluster, there is no presence of clusters that surpass the 5% of the elements. This behavior indicated that the other clusters remains almost with the same size as r increases. Then it is deduced there are structures that persist isolated, meaning they are really far from each other.

- Due to chaining problem, the clusters are not well defined, that is why the ASW and Dunn index are low.

Now, we consider the results from the comparison with the UPGMA method and PAM method (Figures 3.17, 3.22 and 3.32). In the UPGMA method, still is present a big cluster , but there are other clusters of medium size. Also the numbers of clusters differ: meanwhile the number of cluster in the own method is over 300, PAM may choose 2 as the best number of clusters, UPGMA is in the middle of both.

The validation analysis described in Tables 3.2, 3.7 and 3.17 indicates that for all the distances are a lot of true negatives (TN), this is because the large number of clusters in the "Own method" makes difficult that two elements belong in the same set. Also, this is the reason of the huge difference between the Rand Index and the Jaccard index, which is really low. In general, the clusterings are not similar. In particular:

- In the hamming distance the number of false negatives is large indicating the alternative methods cluster elements that "Own method" does not, but in the PAM method is almost trivial since there are only two clusters . At the same time, FN is not all the set, thus, the elements in the big cluster of the "Own method" are for sure in the same cluster in PAM .

- In the levenshtein distance it is relevant the number of false positives respect to the PAM method, this is due to the elements in the big cluster in the "Own method" that are not in different clusters in PAM, which have a large number of clusters.

In contrast, **base pair distance** has a different behavior, as it presents the following characteristics:

- The Figure 3.24 shows that the number of clusters do not decrease abruptly when the ratio increases, its average slope is not less than -20. At the same time, there is a large cluster that grows bigger as r increases.

- A one cluster is reached in the two third of its maximum distance, contrasting the half part of before. At the same time, almost all structures are connected in the half of the maximum distance.

- The main difference observed here with the other distances is that at each step there are clusters merging that do not involve the larger cluster, thus, there exists also clusters of middle size (Figure 3.25). Hence, the clusters are better defined and suggests that other clusters do not remain the same, as r increases, then with this distance there are a less number of isolated structures.

- Even though the Dunn index and the ASW shown lower values in comparison with other distances (Compare Table 3.11 with Tables 3.1,3.6 and 3.16), the BP distance gives a better clustering of elements. This is because these criteria are internal, thus, there are not a good way to compare clusterings of different distances since its condition of maximum is local.

Now, in the UPGMA method (Figure 3.26), there is not a presence of a big cluster contrasting with the other distances. Also the numbers of clusters differ between the three methods giving a larger number for the Own method, followed by UPGMA and with less clusters the PAM method.

In the validation analysis described in Table 3.12, there is still a presence of a lot of true negatives, that is the reason why they present a good Rand Index. Opposite to before, the Jaccard index for the UPGMA is greater than 0.5 because the number of false negatives is not as large as before, because it presents a greater number of true pairs. The Jaccard index for the PAM method is still low, due to the difference of the number of clusters of both methods.

Regarding the mutant set for all distances, is noted that the number of structures coincides with the number of sequences, thus, after the mutation the sequences that used to have the same structure has lost this similarity. In general, the clustering of the mutant set behaves similar as its non mutated peer.

On the other hand, even though it has a similar behavior does not reach a similar best clustering. This is shown in the false positive rates that are specifically 0.169, 0.235, 0.248, 0.324 .

It can be concluded that the “Own method” is sensible to the changes between structures, even the slightly ones. Despite the clusterings are not the best, is a good way to distinguish the differences between sets that in principle seems similar component by component.

Conclusions

1. The Own method is not as good as other methods of clustering, since the chaining problem makes it susceptible to form one big cluster rather than many clusters of medium size.

2. The own method can be used to recognize general properties of the connected components. Can identify isolated data points and data points close in a chaining sense: sometimes might be useful to know if two data points are precisely joined by a chain of data points with small distance pair by pair.

3. As the application of the Own method in different metric spaces gives different results can be conjectured that the change depends on the metric space. Then, it suggests that base pair distance defines a promising metric space that allows a better clustering.

4. The Own method is sensible to changes on the data points , even the little ones. Thus , rather than similarities , it is a good way to distinguish differences between close data sets , for example , noisy data .

Future Work

There are a lot of questions that arise from the results and are beyond the scope defined for the thesis.

Some of the interrogations will remain unanswered until there is a biological analysis of the obtained clusterings, for example:

- The sequences which have exactly the same secondary structure question if they are the same RNA but in MiRBase are annotated with different names or if they are different, question whether belong to the same specie or if they have the same function.

- The RNA's that represents isolated point question if they are the same for all the distances and the reason why they are far from the rest of the sequences.

- If the chaining condition may be an useful information for this set of miRNA.

And other questions depends on a further work , like:

- Since only the 0-homology from the Rips filtration was used in this work, question if using greater dimensions for the topological data analysis gives more and meaningful information about the data set.

- Since a particular set of RNA's were chosen and particular distances, is worth to question if with other distances or other sets the results are improved.

Bibliography

[1] Alaa, H., and Mohamed, S., *On the topological data analysis extensions and comparisons*, Journal of the Egyptian Mathematical Society **25** (2017), no. 4, 406 – 413.

[2] Bastami, M., Nariman-Saleh-Fam, Z., Saadatian, Z., Nariman-Saleh-Fam, L., Omrani, M. D., Ghaderian, S. M. H., and Masotti, A., *The miRNA targetome of coronary artery disease is perturbed by functional polymorphisms identified and prioritized by in-depth bioinformatics analyses exploiting genome-wide association studies*, Gene **594** (2016), no. 1, 74–81.

[3] Bille, P., *A survey on tree edit distance and related problems*, Theor. Comput. Sci. **337** (2005), no. 1-3, 217–239.

[4] Blin, G., Fertin, G., Rusu, I., and Sinoquet, C., *Extending the hardness of rna secondary structure comparison*, Combinatorics, Algorithms, Probabilistic and Experimental Methodologies (Berlin, Heidelberg) (Chen, B., Paterson, M., and Zhang, G., eds.), Springer Berlin Heidelberg, 2007, pp. 140–151.

[5] Ding, Y., Chan, C., and Lawrence, C., *Rna secondary structure prediction by centroids in a boltzmann weighted ensemble.*, RNA (New York, N.Y.) **11** (2005), 1157–1166.

[6] Dunn†, J. C., *Well-separated clusters and optimal fuzzy partitions*, Journal of Cybernetics **4** (1974), no. 1, 95–104.

[7] Edelsbrunner, H., and Harer, J., *Computational topology: an introduction*, American Mathematical Society, 2010.

[8] Edelsbrunner, H., and Harer, J., *Persistent homology – a survey*, 2008.

[9] et mult. al., A. S., *Desctools: Tools for descriptive statistics*, 2018, R package version 0.99.24.

[10] Fasy, B. T., Kim, J., Lecci, F., Maria, C., included GUDHI is authored by Clement Maria, V. R. T., by Dmitriy Morozov, D., by Ulrich Bauer, P., Kerber, M., and Reininghaus., J., *Tda: Statistical tools for topological data analysis*, 2017, R package version 1.6.

[11] Fraleigh, J. B., and Katz, V. J., *A first course in abstract algebra*, 7th ed ed., Addison-Wesley, Boston, 2003.

[12] Ghrist, R., *Barcodes: The persistent topology of data*, Bulletin of the American Mathematical Society **45** (2007), no. 01, 61–76.

[13] Griffiths-Jones, S., Saini, H. K., van Dongen, S., and Enright, A. J., *mirbase: tools for microrna genomics*, Nucleic Acids Research **36** (2008), 154–158.

[14] Hamming, R. W., *Error detecting and error correcting codes*, The Bell System Technical Journal **29** (1950), no. 2, 147–160.

[15] Hatcher, A., *Algebraic topology*, Cambridge University Press, Cambridge, 2002.

[16] Hennig, C., *fpc: Flexible procedures for clustering*, 2018, R package version 2.1-11.

[17] Hofacker, I. L., Fontana, W., Stadler, P. F., Bonhoeffer, L. S., Tacker, M., and Schuster, P., *Fast folding and comparison of rna secondary structures*, Monatshefte für Chemie / Chemical Monthly **125** (1994), no. 2, 167–188.

[18] Horak, D., Maletic, S., and Rajkovic, M., *Persistent homology of complex networks*, Journal of Statistical Mechanics: Theory and Experiment (2009).

[19] Jaccard, P., *Étude comparative de la distribution florale dans une portion des Alpes et des Jura*, Bulletin del la Société Vaudoise des Sciences Naturelles **37** (1901), 547–579.

[20] Kaufman, L., and Rousseeuw, P. J., *Clustering by means of medoids*, 1987, p. 405–416.

[21] Levenshtein, V., *Binary Codes Capable of Correcting Deletions, Insertions and Reversals*, Soviet Physics Doklady **10** (1966), 707.

[22] Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., and Hofacker, I. L., *Viennarna package 2.0*, Algorithms for Molecular Biology **6** (2011), no. 1, 26.

[23] Mamuye, A., and Rucco, M., *Persistent homology on rna secondary structure space*, Proceedings of the 9th EAI International Conference on Bio-inspired Information and Communications Technologies (Formerly BIONETICS) (ICST, Brussels, Belgium, Belgium), 189–192, ICST (Institute for Computer Sciences, Social-Informatics and Telecommunications Engineering), ICST, Brussels, Belgium, Belgium, 2016, pp. 189–192.

[24] McCaskill, J. S., *The equilibrium partition function and base pair binding probabilities for rna secondary structure*, Biopolymers **29** (1990), no. 6-7, 1105–19.

[25] Moulton, V., Zuker, M., Steel, M., Pointon, R., and Penny, D., *Metrics on RNA secondary structures*, Journal of Computational Biology **7** (2000), no. 1-2, 277–292.

[26] Nawrocki, E. P., Burge, S. W., Bateman, A., Daub, J., Eberhardt, R. Y., Eddy, S. R., Floden, E. W., Gardner, P. P., Jones, T. A., Tate, J., and Finn, R. D., *Rfam 12.0: updates to the rna families database*, Nucleic Acids Research **43** (2015), no. D1, D130–D137.

[27] R Core Team, *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, 2017.

[28] Rand, W. M., *Objective criteria for the evaluation of clustering methods*, Journal of the American Statistical Association **66** (1971), no. 336, 846–850.

[29] Rousseeuw, P. J., *Silhouettes: A graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics **20** (1987), 53–65.

[30] Shapiro, B. A., *An algorithm for comparing multiple RNA secondary structures*, Bioinformatics **4** (1988), no. 3, 387–393.

[31] Sokal, R. R., and Michener, C. D., *A statistical method for evaluating systematic relationships*, University of Kansas Science Bulletin **38** (1958), 1409–1438.

[32] Sun, F.-J., Harish, A., and Caetano-Anolles, G., *Phylogenetic utility of rna structure: Evolution's arrow and emergence of early biochemistry and diversified life*, Evolutionary genomics and systems biology, Wiley-Blackwell, 2010.

[33] Tantau, T., *Tikz and pgf manual for version 2.10.*

[34] Vandivier, L. E., Anderson, S. J., Foley, S. W., and Gregory, B. D., *The conservation and function of rna secondary structure in plants*, Annual Review of Plant Biology **67** (2016), no. 1, 463–488, PMID: 26865341.

[35] Zomorodian, A., and Carlsson, G., *Computing persistent homology*, Discrete & Computational Geometry **33** (2005), no. 2, 249–274.

[36] Zomorodian, A. J., *Topology for computing*, Cambridge University Press, 2005.

[37] Zuker, M., Mathews, D. H., and Turner, D. H., *Algorithms and thermodynamics for rna secondary structure prediction: A practical guide*, RNA Biochemistry and Biotechnology (Dordrecht), 11–43, Springer Netherlands, Dordrecht, 1999, pp. 11–43.

[38] Zuker, M., *The use of dynamic programming algorithms in rna secondary structure prediction., chapter 7*, Mathematical methods for DNA sequences, 159–184, Waterman M.S., Ed. CRC Press, Inc., 1989, pp. 159–184.

[39] Zuker, M., *Computational methods for rna secondary structure*, 03 2006.

[40] Zuker, M., and Sankoff, D., *Rna secondary structures and their prediction*, Bulletin of Mathematical Biology **46** (1984), no. 4, 591–621.