USC UNIVERSIDADE DE SANTIAGO DE COMPOSTELA
FACULTADE DE MATEMÁTICAS

Traballo Fin de Grao

Topología de la evolución de los virus

Lurdes María Quintela Tizón

2024/2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

GRADO DE MATEMÁTICAS

Traballo Fin de Grado

Topología de la evolución de los virus

Lurdes María Quintela Tizón

Febrero, 2025

UNIVERSIDADE DE SANTIAGO DE COMPOSTELA

A Jorge e a Malena
por apoiarme ata o final.

Trabajo propuesto

Área de Conocimiento: Xeometría e Topoloxía
Título: Topología de la evolución de los virus
Breve descripción del contenido
En los últimos 15-20 años las técnicas de análisis de datos basadas o inspiradas en ideas topológicas se están demostrando muy eficaces a la vez que complementan las técnicas clásicas. En este TFG veremos cómo se están aplicando dichas técnicas en el campo de la teoría evolutiva, centrándonos en la evolución de los virus.
Recomendaciones
Rabadan, R., & Blumberg, A. J. (2019). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press. Chan, J. M., Carlsson, G., & Rabadan, R. (2013). Topology of viral evolution. *Proceedings Of The National Academy Of Sciences, 110*(46), 18566-18571. `https://doi.org/10.1073/pnas.1313480110`.
Otras observaciones

Índice

Resumen

En las últimas décadas se han desarrollado herramientas topológicas para el análisis de datos en distintas áreas. En este trabajo se explicarán la homología simplicial y la homología persistente, y su aplicación en la biología como método para estudiar y predecir la evolución de los virus, no muy conocida ni controlada.

Particularmente, nos centraremos en el virus de la gripe (*Influenza A*) y en el Virus de la Inmunodeficiencia Humana (VIH), tanto por su prevalencia y mortalidad provocada en humanos, como por la disposición de sus datos e idoneidad con los métodos topológicos expuestos para su estudio.

Abstract

In the last decades, several topological tools for data analysis in different areas have been developed. The present work aims to explain simplicial homology and persistent homology, and their application in biology as a method to study and predict the viral evolution, not very well known nor controlled.

Specifically, we will focus on the flu virus (*Influenza A*) and the Human Immunodeficiency Virus (HIV), both for their prevalence and mortality rate in humans, as well as the disposition of its data and the suitability of the explained topological methods for its study.

Introducción

El análisis topológico de datos es un campo que ha sido desarrollado gradualmente a lo largo de las últimas décadas. A pesar de existir varias referencias a principios de los 2000, el análisis topológico de datos se hace popular en 2009 con un artículo de Gunnar Carlsson llamado *Topology and Data* ("Topología y datos"), en el cual expone cómo la topología es una herramienta adecuada en el estudio de datos por la abstracción que tiene sobre parámetros contextuales (como la métrica dada, o las coordenadas en un espacio), para estudiar las propiedades intrínsecas de los objetos y la relación entre los mismos. Estas "relaciones" se pueden interpretar como aplicaciones entre objetos, dando lugar a la idea de que no sólo es relevante estudiar los objetos en si mismos, si no también las aplicaciones entre ellos[1]. Esta es la base de la teoría de categorías, que cobra relevancia al adoptar esta perspectiva, y por lo cual se ha introducido también en el presente trabajo.

Al margen del avance teórico, es muy relevante destacar el avance computacional que ha permitido poner cosas en práctica que antes eran inviables. En la parte de análisis topológico, existe una librería e Python llamada *Mapper* que pone en práctica el algoritmo del mismo nombre, desarrollado en 2007 por Singh, Mémoli, y Carlsson; o el muy importante cálculo de grupos de homología y cohomología computacionalmente. La aplicación de este tipo de algoritmos a gran escala ha ido, por supuesto, acompañado de la tecnología.

Una de las aplicaciones más sorprendentes de la topología algebraica en los datos ha sido en la evolución de los virus, particularmente el virus de la Influenza (la gripe), y del virus de inmunodeficiencia humana (VIH). Mientras que en el estudio tradicional se representaba la evolución como un árbol binario, este no es adecuado para la Influenza y el VIH, sino que es necesario emplear un retículo: los virus intercambian información genética entre ellos y alteran su genoma a través del contacto entre otros virus (transferencia horizontal de genes), y no solo al heredar mutaciones.
Poder conocer, entender, y predecir este tipo de eventos tiene aplicaciones importantes: En el

[1]Henri Poincaré escribió en el capítulo 2 de *Ciencia e Hipótesis*(1902) algo muy parecido: "Las matemáticas no estudian los objetos, sino las relaciones entre objetos"

caso de la Influenza, nos ayudaría a predecir y prepararnos para futuras cepas de gripe estacional, ya que cada vez es más probable que aparezcan cepas virulentas con una alta tasa de mortalidad, y es necesario desarrollar vacunas específicas para cada cepa.
Para el VIH se ha observado que una alta tasa de recombinación del genoma en VIH de larga duración en una persona puede tener una fuerte relación con que se desarrolle demencia. Esta demencia es tratable, pero no reversible, por lo que entender cuándo y por qué va a desarrollarse, lo que podría evitar su aparición y desarrollo.

Así con todo, el uso del análisis topológico de datos tiene muchas más aplicaciones en la biología de las aquí expuestas, siendo muy destacable su aplicación en distintos tipo de cáncer como método de detección y prevención. Dicho estudio pormenorizado se encuentra en la bibliografía de referencia para este trabajo, *Topological Data Analysis for Genomics and Evolution*, de Raúl Rabadán y Andrew J. Blumberg, donde se profundiza tanto en la matemática detrás (la topología y la estadística) como en todo el campo de estudio que existe ahora mismo en la genética y la evolución gracias a estas nuevas herramientas.

Esta memoria se divide en 3 capítulos, a lo largo de los cuales se irá viendo la construcción matemática y su aplicación en la biología: en el primer capítulo se tratará de la homología clásica y tradicional mediante de la homología simplicial; en el segundo se expandirán estos conceptos a través de la homología persistente y la homología zigzag; en el último capítulo se aplicarán estos conceptos de homología en la biología, primero en los árboles filogenéticos y luego en los casos particulares del virus de la Influenza y del VIH.

Capítulo 1

Complejos simpliciales y homología simplicial

En este capítulo, comenzaremos con un repaso sobre ciertos conceptos de álgebra, sobre todo aquellos enfocados en grupos. También daremos una pequeña introducción a la teoría de categorías, ya que será una herramienta útil para poder conectar todos los conceptos que se irán explicando a lo largo del trabajo. Esto permitirá relacionar estructuras entre los distintos conceptos topológicos que se tratarán en las secciones sucesivas.
Una vez se sienten las bases algebraicas necesarias, pasaremos a explicar los conceptos de *grupo de homología* y los *complejos de cadenas*, en los cuales se basa el análisis topológico de datos para la biología, como veremos en el último capítulo.
Las principales fuentes consultadas a lo largo del capítulo son [11] y [12]. En cada sección correspondiente se especificará si hay otras.

1.1. Preliminares: grupos abelianos y teoría de categorías

Para poder definir los grupos de homología a partir de los complejos simpliciales abstractos, primero necesitamos repasar algunos conceptos de álgebra vistos durante la carrera, y expandirlos a través de la teoría de categorías. Esta base algebraica nos dará herramientas para poder computar y emplear el grupo de homología eficazmente a través de su estructura algebraica subyacente.

1.1.1. Grupos

Usaremos como referencia [9] y [3] para los contenidos de esta sección.

Definición 1.1 (Grupo)**.** Un grupo se define como un par $(G, \cdot)$, donde G es un conjunto y $\cdot$ es una operación interna o aplicación,

$$\begin{aligned} \cdot : G \times G &\longrightarrow G \\ (x, y) &\longmapsto x \cdot y \end{aligned}$$

que verifica las siguientes propiedades:

- **GR1** Para cualquier $x, y \in G$, la aplicación es asociativa: $x \cdot (y \cdot z) = (x \cdot y) \cdot z$
- **GR2** Existe un elemento e de G tal que: $e \cdot x = x \cdot e = e, \quad \forall x \in G$
- **GR3** Si x es un elemento de G, entonces existirá un y de G tal que: $x \cdot y = y \cdot x = e$

Es decir, un grupo es un conjunto de elementos, dotado con una operación interna que es asociativa entre sus elementos, que tiene neutro, y tal que todo elemento tiene un **inverso**.

Con esta notación, G sería un *grupo multiplicativo*. De sustituir $\cdot$ por $+$, tendríamos notación para un *grupo aditivo*.

Un grupo $(G, +)$ se dice que es *abeliano* o *conmutativo* si, además, verifica la **propiedad conmutativa**:

$$x + y = y + x, \quad \forall x, y \in G.$$

Definición 1.2. Definimos la *base* en un grupo como la familia $\{g_i\}_{i \in J}$ de elementos de G tales que, cada elemento g de G se pueda escribir de manera única como la suma finita:

$$g = \sum_i c_i g_i$$

donde c_i son enteros pertenecientes a $\mathbb{Z}$. El número de elementos $g_1, ..., g_n$ de una base, es su *rango*, en este caso n. Dicho rango puede o no ser finito.

Alternativamente, si tenemos un conjunto de elementos finito (no necesariamente únicos) que generan a G, se dice que G está *finitamente generado*.

Definición 1.3. Un grupo abeliano G es un *grupo libre* cuando tiene una **base**.

Definición 1.4. Sea G un grupo abeliano. Un elemento g de G se dice de *torsión* o que tiene *orden finito* si, para algún entero positivo n, tenemos que $ng = 0$.

El conjunto de elementos de torsión de un grupo G es un subgrupo T, llamado *subgrupo de torsión*. Cuanto $T = 0$ se dice que G es un grupo *libre de torsión*.

Proposición 1.5. *Un grupo abeliano G finitamente generado es libre si, y solo si, es libre de torsión*

Como consecuencia directa de esta proposición, podemos expresar cualquier grupo abeliano finitamente generado como la suma directa de un grupo finito y un grupo libre. De hecho:

Teorema 1.6. *Sea G un grupo abeliano finitamente generado, y sea T su grupo de torsión. Entonces G/T es un grupo abeliano libre, y además G es la suma directa*

$$G = T \oplus F$$

donde F es libre.

Teorema 1.7. ***(Teorema fundamental de grupos abelianos finitamente generados).*** *Sea G un grupo abeliano finitamente generado. Sea T su subgrupo de torsión.*

(a) Existe un subgrupo abeliano libre, H de G, con rango finito β y tal que $G = H \oplus T$.

(b) Existen grupos cíclicos finitos $T_1, ..., T_k$, tal que T_i tiene orden $t_i > 1$, cumpliéndose que $t_1|t_2...|t_k$ y

$$T = T_1 \oplus T_2 \oplus ... \oplus T_k$$

.

(c) Los números β y $t_1, ..., t_k$ están unívocamente determinados por G.

El número β es el *número de Betti* de G, y que además es el rango del grupo abeliano libre $G/T \cong H$, y los números $t_1, ..., t_k$ son los *coeficientes de torsión* de G

Demostración. La prueba se puede encontrar en el Munkres [11], Ch.1 §4. □

Teorema 1.8. *Sea G un grupo abeliano finitamente generado. Entonces, existe un isomorfismo*

$$G \cong \underbrace{\mathbb{Z} \times \mathbb{Z} \times ... \mathbb{Z}}_{k} \times \mathbb{Z}/p_1^{n_1} \times \mathbb{Z}/p_2^{n_2} ... \times \mathbb{Z}/p_m^{n_m}$$

donde los p_i son primos, no necesariamente distintos.

Aquí, el *rango* de G sería el número k de factores $\mathbb{Z}$, y la parte de G de el anillo cociente es la *torsión*. Los números p_i son los *factores invariantes* de G.

Acabamos la sección con la definición de anillo y, sobretodo, cuerpo, que será relevante al trabajar con la homología persistente.

Definición 1.9 (Anillo)**.** Un anillo R es un conjunto de elementos con dos operaciones internas $+$ y $\cdot$, tales que:

- **RI1** El par $(R, +)$ es un grupo abeliano con elemento neutro 0.
- $(R, \cdot)$ es un monoide:

  RI2 La operación es asociativa: $(x \cdot y) \cdot z = x \cdot (y \cdot z), \quad \forall x, y, z \in R.$

  RI3 Existe elemento neutro $1 \in R$: $1 \cdot x = x = x \cdot 1, \quad \forall x \in R.$
- **RI4** Se cumple la propiedad distributiva de $\cdot$ respecto a $+$:

$$x \cdot (y + z) = x \cdot y + x \cdot z$$
$$(y + z) \cdot x = y \cdot x + z \cdot x$$

$\forall x, y, z \in R.$

A pesar de que pueda existir elemento inverso multiplicativo para algunos elementos, no lo hay para todos. Se denota como $-x$ como el inverso aditivo y x^{-1} como el inverso multiplicativo, cuando lo haya. Aquellos elementos no nulos del anillo que tengan inverso para la multiplicación son llamados *unidades*.

Un anillo se dice *conmutativo* cuando el producto también cumpla la propiedad conmutativa.

Definición 1.10. Un *cuerpo* es un anillo conmutativo cuyos elementos son todos unidades, es decir, para todo $x \in R,\ x \neq 0$ existe un x^{-1} tal que $xx^{-1} = x^{-1}x = 1$.

1.1.2. Teoría de categorías

Esta sección toma como referencia una de las mejores introducciones para la teoría de categorías, que es el libro escrito por Emily Riehl ([13]).

Definición 1.11 (Categoría)**.** Una *categoría* $\mathcal{C}$ es una colección de objetos $ob(\mathcal{C})$ de la forma $A, B, C...$ y *flechas* o *morfismos* $f, g, h...$ cuyo dominio y codominio son objetos de la categoría,

$$f : A \to B,\ f \in Hom_{\mathcal{C}}(A, B).$$

Además, ha de existir un *morfismo identidad* para cada objeto, denotado $id_A : A \to A$, y para cualesquiera dos morfismos f, g de $\mathcal{C}$ tales que $f : A \to B, \quad g : B \to C$, existirá un *morfismo composición* $g \circ f \equiv gf$:

$$A \xrightarrow{f} B \xrightarrow{g} C, \qquad A \xrightarrow{gf} C$$

Además, esta composición debe de ser unitaria y asociativa, es decir, verificar los siguientes **axiomas categóricos**:

1. Para cualquier morfismo $f : A \to B$, se tiene que $id_B f = f id_A = f$.
2. Para cualquier triple composición de los morfismos f, g, h tales que

$$f : A \to B, \quad g : B \to C, \quad h : C \to D,$$

la composición será asociativa, $(hg)f = h(gf)$, y se denotará directamente como $hgf : A \to D$.

Ejemplo 1.12. Prácticamente todas las ramas de las matemáticas son, o forman parte de, una categoría. Algunos ejemplos podrían ser:

1. La categoría **Set**, probablemente la más famosa y la más importante. Sus objetos son los conjuntos, y los morfismos son las aplicaciones que van entre conjuntos.
2. La categoría **Top**, donde los objetos son los espacios topológicos, y los morfismos son las aplicaciones continuas.
3. La categoría **Grp**, que tiene como objetos los grupos y los morfimos son los homomorfismos entre ellos.

Definición 1.13. Una *subcategoría* $\mathcal{D}$ de una categoría $\mathcal{C}$ es una subcolección de objetos y una subcolección de morfismos de $\mathcal{C}$, tal que para todo $A, B \in ob(\mathcal{C})$ tengamos

$$Hom_{\mathcal{D}}(A, B) \subseteq Hom_{\mathcal{C}}(A, B).$$

Cuando esto es una igualdad, es decir, cuando para cualquier par de objetos A, B contenidos en $ob(\mathcal{D})$ tengamos todos los morfismos de la categoría, $Hom_{\mathcal{C}}(A, B)$, se dice que la subcategoría es *plena*. Si en cambio la subcategoría contiene todos los objetos de $\mathcal{C}$, independientemente de los morfismos, es una subcategoría *amplia*.

Definición 1.14. En una categoría $\mathcal{C}$, un *isomorfismo* es un morfismo $f : A \to B$ para el cual existe otro morfismo en la categoría, $g : B \to A$, tal que

$$fg = id_B \in Hom_{\mathcal{C}}(B, B), \quad \text{y} \quad gf = id_A \in Hom_{\mathcal{C}}(A, A).$$

Cuando existe un isomorfismo entre dos objetos, se dice que son *isomorfos* y lo denotamos como $A \cong B$.

Definición 1.15 (Functor)**.** Sean dos categorías $\mathcal{C}$ y $\mathcal{D}$. Un *functor* $F : \mathcal{C} \to \mathcal{D}$ está definido por:

- Unos objetos $F(A) \equiv FA \in ob(\mathcal{D})$, para cada $A \in ob(\mathcal{C})$.
- Unos morfismos $F(f) \equiv Ff : FA \to FB$, $Ff \in Hom_{\mathcal{D}}(FA, FB)$ para cada morfismo $f : A \to B$ de $Hom_{\mathcal{C}}(A, B)$.

que cumplen los **axiomas functoriales**:

1. Para cualquier par de morfismos f, g de $\mathcal{C}$ que puedan ser compuestos, entonces $FgFf = F(gf)$.
2. Para cualquier objeto A de $\mathcal{C}$, $Fid_A = id_{FA}$.

Es decir, los functores además de llevar los objetos y morfismos de una categoría a otra, preservan la identidad y son compatibles con la composición.

Todos los factores invariantes que estudiaremos son, en realidad, functores que van de categorías geométricas a algebraicas. Por lo tanto, el concepto del functor justifica la introducción de las categorías en este trabajo.

Un functor permite abstraerse más aun de los objetos y centrarse sólo en las interacciones entre estos dentro de una misma categoría, siendo capaz de "traducir" estas interacciones, total o parcialmente, hacia otro contexto matemático distinto, es decir, una categoría distinta.

Ejemplo 1.16. Ilustramos el functor con un par de ejemplos:

1. La asignación de un espacio topológico con sus componentes por caminos es un functor que va de la categoría **Top** a la categoría **Set**.
2. Existe un functor que va **Grp** a **Set** que conserva todos los elementos, pero "olvida" la estructura de grupo entre dichos elementos.
3. Análogamente, se puede definir un functor de **Set** a **Grp** que lleve un conjunto a los generadores de un grupo libre, donde los elementos del conjunto de **Set** son, precisamente, esos generadores.

Definición 1.17 (Transformación natural)**.** Sean F y G functores de $\mathcal{C}$ a $\mathcal{D}$. Una *transformación natural* $\tau : F \Rightarrow G$ viene dada por:

- Un morfismo $\tau_A : FA \to GA$ para cada objeto $A \in ob(\mathcal{C})$, la cual formará parte de la colección de **componentes** de la transformación natural.

- El diagrama conmutativo

$$\begin{array}{ccc} FA & \xrightarrow{Ff} & FA \\ {\scriptstyle \tau_A}\downarrow & & \downarrow{\scriptstyle \tau_B} \\ GA & \xrightarrow{Gf} & GB \end{array}$$

  para cada morfismo $f : A \to B$ en $Hom_{\mathcal{C}}(A, B)$.

El concepto de transformación natural cobrará relevancia a la hora de definir los complejos de cadenas. Tomamos la categoría de los números naturales $\mathbb{N}$, con objetos los propios elementos de $\mathbb{N}$ y morfismos de n a m, cuando $n < m$. Teniendo una categoría $\mathcal{C}$ y un functor $F : \mathbb{N} \to \mathcal{C}$ dado por

$$F0 \to F1 \to F2 \to ...$$

tenemos una transformación natural $\tau : F \to G$ determinada por el diagrama conmutativo

$$\begin{array}{ccccccc} F0 & \longrightarrow & F1 & \longrightarrow & F2 & \longrightarrow & ... \\ {\scriptstyle \tau_0}\downarrow & & {\scriptstyle \tau_1}\downarrow & & {\scriptstyle \tau_2}\downarrow & & \\ G0 & \longrightarrow & G1 & \longrightarrow & G2 & \longrightarrow & ... \end{array}$$

1.2. Complejos simpliciales

Tras haber dado una base sobre la conceptos de fuera de la topología que estarán presentes a lo largo del trabajo, procedemos a exponer la base topológica sobre la que trabajaremos, empezando por los complejos simpliciales.

Definición 1.18. Dado un conjunto de puntos $\{a_0, ..., a_n\} \in \mathbb{R}^n$, se dice que este es *geométricamente independiente* si, para escalares cualesquiera $t_i \in \mathbb{R}$,

$$\sum_{i=0}^{n} t_i = 0 \quad \text{y} \quad \sum_{i=0}^{n} t_i a_i = \mathbf{0},$$

implica que $t_1 = t_2 = ... = t_n = 0$.

Observación 1.19. Se puede dar una interpretación esta definición en el sentido tradicional de "linealmente independientes". Para ello, tomamos el primer punto a_0, y vemos que un conjunto es geométricamente independiente si y solo si el conjunto

$$\{a_1 - a_0, a_2 - a_0, ..., a_n - a_0\}$$

es linealmente independiente.
Tenemos además, por definición, que un conjunto con un solo punto siempre es geométricamente independiente; un conjunto de dos puntos lo es cuando estos son distintos, un conjunto de 3 puntos es geométricamente independiente cuando no están los 3 en la misma recta, etc.

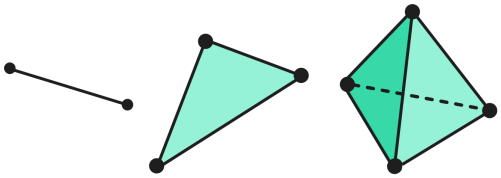


Figura 1.1: Ejemplo de símplices de 0, 1, y 2 dimensiones

Definición 1.20. El *n-símplice* σ generado por los puntos del conjunto geométricamente independiente $\{a_0, a_1, ..., a_n\} \in \mathbb{R}^n$ será el conjunto de todos los puntos $x \in \mathbb{R}^n$ tal que

$$x = \sum_{i=0}^{n} t_i a_i,$$

con $\sum_{i=0}^{n} t_i = 1$ y $t_i \geq 0, \forall i$.
Los números t_i están unívocamente determinados por x, y son las llamadas *coordenadas baricéntricas* del punto x de σ respecto a $a_0, ..., a_n$.

Observación 1.21. Un *n-símplice* es el análogo n-dimensional de un triángulo. Esto es, el menor conjunto convexo en $\mathbb{R}^n$ que contiene $(n + 1)$ puntos distintos que no se encuentran en un hiperplano de dimensión menor a n, por tanto la condición de que sean geométricamente independientes. Estos puntos son los que generarán el n-símplice. El objetivo es, por supuesto, generalizar construcciones geométricas a cualquier dimensión.

Definición 1.22. Los puntos $a_0, ..., a_n$ geométricamente independientes que generan σ son llamados *vértices* de σ, n es la *dimensión*, y cualquier símplice generado por un subconjunto de $\{a_0, ..., a_n\}$ se le llama *cara* de σ.

Particularmente, la cara de σ generada por $a_1, a_2..., a_n$ es la *cara opuesta* a a_0, y cualquier cara de σ distinta del propio σ son las *caras propias*. La unión de las caras propias es la *frontera*, denotada como $Bd(\sigma)$, es decir, la unión de todas las caras generadas a partir de subconjuntos distintos del total de vértices que generan el símplice.

El *interior* del n-símplice σ se define como $Int(\sigma) = \sigma - Bd(\sigma)$. Se puede caracterizar el interior como el subconjunto de puntos x tales que, para todas las coordenadas baricéntricas t_i, $t_i > 0$.

Ahora que hemos visto cómo se generan las estructuras básicas, los símplices, definimos cómo se agrupan para formar una estructura superior.

Definición 1.23 (Complejo simplicial)**.** Un *complejo simplicial* K de $\mathbb{R}^N$ es una colección de símplices en $\mathbb{R}^N$ tal que:

1. Toda cara de un símplice de K está en K.

2. La intersección de cualesquiera dos símplices de K será una cara en ambos símplices.

En un complejo simplicial, los 0-símplices son los *vértices*. Más generalmente, la colección de símplices de, como máximo, dimensión n, se llamará el *n-esqueleto* del complejo simplicial, y se denota $K^{(n)}$. Por tanto, los vértices son los aquellos puntos que conforman $K^{(0)}$.
Vemos un ejemplo en la Figura 1.2.

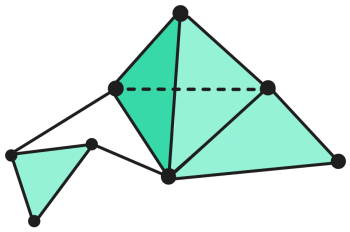


Figura 1.2: Ejemplo de un complejo simplicial

Definición 1.24. Sea $|K|$ el subconjunto de $\mathbb{R}^N$ dado por la unión de símplices de K, complejo simplicial. Si cada símplice σ de K hereda la topología usual de $\mathbb{R}^N$, entonces se puede establecer una topología en $|K|$ tomando un subconjunto A de $|K|$ como cerrado en $|K|$ si, y solo si, $A \cap \sigma$ es un cerrado en σ, para cada σ. Esto define una topología en $|K|$ (la *topología débil*), ya que se puede tomar uniones finitas e intersecciones arbitrarias, y es relevante cuando tomemos complejos simpliciales infinitos, ya que para los complejos finitos la topología débil coincide con la topología usual heredada.
El espacio $|K|$ se llama *espacio subyacente* de K, o *realización geométrica* de K.

Observación 1.25. No ahondaremos en aquellos complejos simpliciales infinitos, si no que nos restringiremos a trabajar con aquellos con un número finito de símplices, es decir, los *complejos simpliciales finitos*.

Observación 1.26. La realización geométrica de un complejo simplicial se le puede dar la estructura de un CW-complejo (o complejo de celdas), que es un espacio topológico de construcción similar a la de los complejos simpliciales pero a partir de esferas y discos, restringido al espacio topológico de $\mathbb{R}^n$. Esta construcción se obviará en este trabajo, debido a la dificultad que existe para computarlos, y que están bien generalizados en los complejos simpliciales.

1.2.1. Aplicaciones simpliciales

Definición 1.27 (Aplicación simplicial)**.** Sean K e L dos complejos simpliciales, y $f^{(0)} : K^{(0)} \to L^{(0)}$ una aplicación. Suponemos que, para todo conjunto de vértices $a_0, ..., a_n$ de K que generan

un símplice σ de K, sus imágenes $f^{(0)}(a_0), ..., f^{(0)}(a_n)$ son vértices de un símplice de L. Entonces, $f^{(0)}$ puede ser extendido a una aplicación continua $|f| : |K| \to |L|$ tal que

$$x = \sum_{i=0}^{n} t_i a_i \quad \longmapsto \quad |f(x)| = \sum_{i=0}^{n} t_i f^{(0)}(a_i).$$

A la aplicación $|f|$ se le llama *aplicación simplicial* inducida por la aplicación de vértices $f^{(0)}$.

Observación 1.28. Tomando como objetos los complejos simpliciales, y como morfismos (o flechas) las aplicaciones simpliciales, podemos formar una categoría de complejos simpliciales. De hecho, la realización geométrica es un functor que va de la categoría de los complejos simpliciales y aplicaciones simpliciales, a la categoría de espacios topológicos y aplicaciones continuas.

Habiendo definido las aplicaciones que podemos definir entre complejos simpliciales a partir de sus vértices, vemos ahora cómo definir isomorfismos entre complejos.

Definición 1.29. Sean K y L complejos simpliciales. Suponemos que $f^{(0)} : K^{(0)} \to L^{(0)}$ es una correspondencia biyectiva de manera que los vértices $a_0, ..., a_n$ de K generan un símplice de K sí y solo sí $f^{(0)}(a_0), ..., f^{(0)}(a_n)$ generan un símplice de L. Entonces, la aplicación simplicial inducida $|f| : |K| \to |L|$ es un homeomorfismo llamado *homeomorfismo simplicial*, y, particularmente, es un **isomorfismo** entre K y L.

Corolario 1.30. *Sea Δ^N el complejo simplicial compuesto por un N-símplice y sus caras. Si K es un complejo finito, entonces K es isomorfo a un subcomplejo de Δ^N, para algún N.*

Demostración. Para cada símplice σ de K, existe una aplicación $|g|$ que va de σ a otro símplice de misma dimensión, τ de L. Debemos demostrar que la aplicación lineal $|h| : \tau \to \sigma$ inducida por la correspondencia de vértices, $(f^{(0)})^{-1}$ es la inversa de la aplicación $|g| : \sigma \to \tau$. Para esto, denotamos como $x = \sum_{i=0}^{n} t_i a_i$. Por definición tenemos que $|g(x)| = \sum_{i=0}^{n} t_i f^{(0)}(a_i)$, por lo tanto

$$|h(|g(x)|)| = |h(\sum_{i=0}^{n} t_i f^{(0)}(a_i))| = \sum_{i=0}^{n} t_i (f^{(0)})^{-1}(f^{(0)}(a_i)) = \sum_{i=0}^{n} t_i(a_i) = x.$$

□

La información que nos da los complejos simpliciales vistos hasta ahora es dependiente de cuál es el embebimiento hecho en el espacio euclídeo $\mathbb{R}^n$, donde estamos aun trabajando. Como el objetivo es poder analizar un conjunto de datos, podemos abstraernos del espacio topológico subyacente fijándonos únicamente en cuántos símplices hay, y qué caras están pegadas entre sí.

Para ello, introducimos la noción de complejo simplicial abstracto.

1.2.2. Complejos simpliciales abstractos

Definición 1.31. Un *complejo simplicial abstracto* es una colección $\mathcal{S}$ de conjuntos finitos y no vacíos, tales que si A es un elemento de $\mathcal{S}$, entonces cualquier subconjunto no vacío de A también serán elementos de $\mathcal{S}$.

De manera análoga a los complejos simpliciales descritos previamente (que serán los *complejos simpliciales geométricos*), tenemos los siguientes resultados y definiciones inmediatos:

1. Un elemento A de $\mathcal{S}$ es un *símplice*. Nos referimos a los elementos de $\mathcal{S}$ como *símplices abstractos*. Su dimensión será una unidad inferior al número de elementos ($|A| - 1$, donde $|-|$ denota el cardinal).

2. La dimensión de $\mathcal{S}$ es la mayor dimensión de uno de sus símplices. Si no la hay, es decir, si no existe un elemento maximal en $\mathcal{S}$ para un $\mathcal{S}$ no vacío, entonces tendrá dimensión infinita.

3. Cualquier subconjunto no vacío de un símplice A es una cara de A.

4. El *conjunto de vértices* V de $\mathcal{S}$ es la unión de los conjuntos de un solo elementos. Cualquier símplice será la unión de vértices.

5. Denotaremos el subconjunto de $\mathcal{S}$ de conjuntos con dimensión $\leq k$ como $\mathcal{S}_k$, el k-esqueleto.

Análogamente, también podemos definir las aplicaciones simpliciales y los isomorfismos aplicados a los complejos simpliciales abstractos.

Definición 1.32. Dos complejos simpliciales abstractos $\mathcal{S}$ y $\mathcal{T}$ se dice que son *isomorfos* cuando existe una correspondencia biyectiva que lleva el conjunto de vértices de $\mathcal{S}$ al conjunto de vértices de $\mathcal{T}$.

Definición 1.33. Sea K un complejo simplicial, y V el conjunto de vértices de K. Sea $\mathcal{K}$ la colección de todos los subconjuntos $\{a_0, ..., a_n\}$ de V tal que los vértices $a_0, ..., a_n$ generan un símplice de K. La colección $\mathcal{K}$ se llama *esquema de vértices* de K.

Ahora que definimos el esquema de vértices, podemos explicar la relación entre los complejos simpliciales abstractos y los geométricos.

Teorema 1.34. *Cualquier complejo simplicial abstracto* $\mathcal{S}$ *es isomorfo al esquema de vértices de algún complejo simplicial geométrico* K*.*

Cabe destacar que cada complejo simplicial abstracto se asocia de manera **única** salvo isomorfismos a su complejo simplicial geométrico asociado. De hecho, dos complejos simpliciales geométricos son isomorfos si, y solo si, son isomorfos sus complejos simpliciales abstractos asociados. Por tanto, la relación entre ambos conceptos es clara.

Observación 1.35. Como antes mencionamos, existe un functor (la realización geométrica), que lleva la categoría de objetos complejos simpliciales geométricos y morfismos aplicaciones simpliciales a la categoría de espacios topológicos y aplicaciones continuas. De manera análoga, teniendo en cuenta la asociación unívoca entre un complejo simplicial geométrico y el complejo simplicial abstracto con el cual existe un isomorfismo a través de su esquema de vértices, podemos establecer una vez más un functor (de realización geométrica) entre la categoría de objetos complejos simpliciales abstractos y morfismos aplicaciones simpliciales, a la categoría de espacios topológicos y aplicaciones continuas.

Definición 1.36. Si el complejo simplicial abstracto $\mathcal{S}$ es isomorfo al esquema de vértices del complejo simplicial geométrico K, entonces K es la *realización geométrica* de $\mathcal{S}$. Está unívocamente determinada, salvo isomorfismos.

Esto es completamente coherente con la definición anterior, pues la motivación para definir los complejos simpliciales abstractos en contraposición a los geométricos es, precisamente, la abstracción del espacio topológico subyacente.

Aun así, se puede inferir la identificación con un espacio topológico de un complejo simplicial abstracto cuando exista la correspondencia con el esquema de vértices con un complejo simplicial geométrico. Esto será, por tanto, su realización geométrica.

1.3. Homología simplicial

1.3.1. Complejos de cadenas

Para esta sección seguimos las principales referencias [7], [11] y [12]. Cabe notar, que por el tratamiento a través de grupos y categorías, nos hemos apoyado también en [14].

Antes de comenzar, es importante recalcar que toda la construcción de homología se hará sobre la noción de **grupos**, en base a las definiciones dadas. Aunque para la parte teórica se puede explicar desde esta perspectiva, es habitual ver las siguientes definiciones tomando un espacio vectorial de dimensión finita V sobre un cuerpo arbitrario $\mathbb{F}$, y tomar transformaciones lineales en vez de homomorfimos, ya que con estas aplicaciones lineales es como se llevará luego a la práctica. Esto no es más que una particularización de las definiciones sobre grupos, si interpretamos el grupo como un $\mathbb{Z}$-módulo.

Partimos pues de la base de que tenemos un complejo simplicial abstracto, que denotamos K, el cual se va construyendo a partir de símplices de $0, 1, ..., (n-1)$ dimensiones.
Los grupos de homología, intuitivamente, dan información de cómo están "pegados" los símplices a medida que se va aumentando la dimensión. Esto implica ver qué agujeros existían antes de que un símplice de determinada dimensión fuese añadido, lo cual se podrá estudiar gracias al grupo cociente.

Para poder hacer este estudio, es importante primero definir la orientación de los símplices:

Definición 1.37. Sea σ un n-símplice, tanto geométrico como abstracto, y asumimos que $n > 0$ (para los vértices, sólo existe una orientación).
La *orientación* de los vértices de σ es una **clase de equivalencia** de ordenaciones de los vértices, bajo la relación de equivalencia de que dos ordenaciones son la misma si se diferencian por una permutación par.
Un n-símplice orientado con vértices $\{a_0, ..., a_n\}$, se denota como $[a_0, ..., a_n]$.

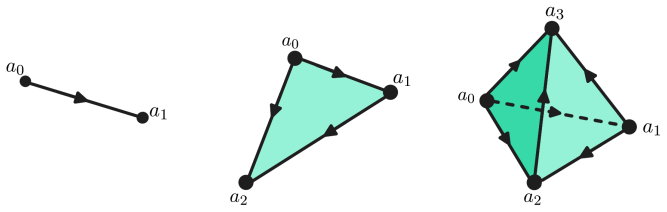


Figura 1.3: Ejemplo de 0, 1 e 2-símplices orientados

El construir la homología con grupos de cadenas y homomorfismo de bordes permite dar la información combinatoria de un complejo simplicial a través de herramientas algebraicas.

Definición 1.38 (Cadenas)**.** El *n-ésimo grupo de cadenas* C_n (o, simplemente, *las cadenas*) del complejo simplicial K, es el grupo abeliano del conjunto de n-símplices orientados. Tomando $\sigma, \tau \in K$ tenemos que $\sigma = -\tau$ si σ y τ son iguales salvo orientación. Un elemento

$$c = \sum_{i=0}^{n} k_i \sigma_i$$

para $\sigma_i \in K$ y coeficientes $k_i \in \mathbb{Z}$, es una *n-cadena* c, $c \in C_n(K, \mathbb{Z})$.

Definición 1.39. Para un $n \in \mathbb{Z}$ fijado, la transformación lineal

$$\partial_n : C_n(K; \mathbb{F}) \to C_{n-1}(K; \mathbb{Z})$$

es la llamada *operación borde*. Sea $\sigma = [w_0, w_1, ..., w_n]$ y $n > 0$, tenemos que

$$\partial_n(\sigma) = \partial_n([w_0, w_1, ..., w_n]) \mapsto \sum_{i=0}^{n} (-1)^i [w_0, ..., \hat{w}_i, ..., w_n],$$

donde $\hat{w}_i$ denota un vértice que ha sido eliminado.

Se define un homomorfismo al extender linealmente esta aplicación a todo el grupo $C_n(K;\mathbb{Z})$, y la orientación de la imagen queda determinada por los vértices. Además, como para $n < 0$ el grupo $C_n(K;\mathbb{Z})$ es el grupo trivial, se tiene que el operador ∂_n es el homomorfismo trivial para $n \leq 0$.

Esta definición se puede ver de manera geométrica: Aplicado a un símplice, la operación borde es la suma alterna de las caras que componen la frontera de dicho símplice. Tomemos como ejemplo la Figura 1.4.

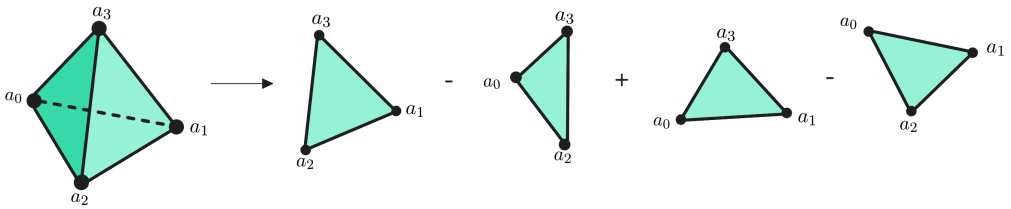


Figura 1.4: Veamos el borde de un 2-símplice. Sea $K = [a_0, a_1, a_2, a_3]$, le aplicamos el operador borde y obtenemos $K = [a_1, a_2, a_3] - [a_0, a_2, a_3] + [a_0, a_1, a_3] - [a_0, a_1, a_2]$.

Lema 1.40. *$C_n(K;\mathbb{Z})$ es un grupo abeliano libre.*

La operación borde tiene una propiedad característica por la cual es tan relevante, que viene siendo "la frontera de la frontera es 0". Lo formalizamos en el siguiente resultado.

Lema 1.41. *La composición $\partial_{n-1} \circ \partial_n = 0$.*

Como corolario inmediato de este resultado, tenemos:

Corolario 1.42. *Para cualquier complejo simplicial K, y $n \in \mathbb{N}$, tenemos:*

$$Im(\partial_{n+1}) \subseteq Ker(\partial_n)).$$

Definimos ahora, para grupos abelianos, el complejo de cadenas:

Definición 1.43 (Complejo de cadenas)**.** Sean $C_n = C_n(K,\mathbb{Z})$, $n \in \mathbb{N}$ las $n-$cadenas de un símplice orientado K. El *complejo de cadenas* es la sucesión de homomorfismos:

$$\ldots \xrightarrow{\partial_{n+2}} C_{n+1} \xrightarrow{\partial_{n+1}} C_n \xrightarrow{\partial_n} C_{n-1} \xrightarrow{\partial_{n-1}} \ldots \xrightarrow{\partial_1} C_0 \xrightarrow{\partial_0} 0 \xrightarrow{\partial_0} 0 \xrightarrow{\partial_0} \ldots$$

Definimos una categoría de objetos los enteros $\mathbb{Z}$, y los morfismos de n a m, cuando $n < m$. Podemos tomar la categoría opuesta (o dual), que consistiría en invertir las flechas: en vez de ir

$n \to m, n < m$, tomamos las flechas $m \to n, m > n$, por lo que tenemos una sucesión de flechas $... \to 5 \to 4 \to 3 \to ....$
Así, podemos ver el complejo de cadenas como un functor $(\mathbb{Z}, \leq)^{op} \to \mathbf{Ab}$, categoría de grupos abelianos, tal que se satisfaga la condición para la doble composición $\partial_{n-1} \circ \partial_n = 0$.

Definición 1.44. Para un complejo simplicial K orientado, y la operación borde asociada al mismo, tenemos las siguientes definiciones:

- El núcleo de $\partial_n : C_n(K;\mathbb{Z}) \to C_{n-1}(K;\mathbb{Z})$ es el *grupo de n-ciclos*, y se denota como $Z_n(K;\mathbb{Z})$.
- La imagen de $\partial_{n+1} : C_{n+1}(K;\mathbb{Z}) \to C_n(K;\mathbb{Z})$ es el *grupo de n-bordes*, y se denota como $B_n(K;\mathbb{Z})$.

Podemos reformular el Resultado 1.42 como:

Corolario 1.45. *Para cualquier complejo simplicial K, y $n \in \mathbb{N}$, tenemos*

$$B_n(K;\mathbb{Z}) \subseteq Z_n(K;\mathbb{Z}).$$

Esto tiene sentido, ya que el borde de un complejo simplicial es también considerado un ciclo, pero esto no incluye a todos los ciclos sino solo aquellos que encierran un símplice dentro.

Ahora que tenemos esta inclusión, podemos definir ya los grupos de homología. La idea es la siguiente: Tenemos una inclusión de los bordes de un complejo simplicial en sus ciclos, es decir, los "recorridos cerrados" que se puedan hacer recorriendo los símplices. Lo que queremos calcular, o ver, es lo precisa que es esta inclusión: Cuando es perfecta, quiere decir que a medida que se desgranan las dimensiones de los complejos (es decir, se estudian progresivamente los símplices del complejo, empezando por la mayor dimensión hasta llegar a los 0-símplices), dimensión a dimensión, donde había un grupo de símplices queda su borde como otros símplices. Si, en cambio, la inclusión no es perfecta, quiere decir que debajo de un n-símplice **no** hay otros $(n-1)$-símplices subyacentes que hacían de frontera, si no que hay un "agujeros". Esta información sobre los "agujeros" que hay, el desfase en la inclusión de los ciclos y los bordes subyacentes, es lo que da origen al **grupo de homología**.

Definición 1.46. Definimos el *n-ésimo grupo de homología* de K con coeficientes en $\mathbb{Z}$ como el cociente:

$$H_n(K;\mathbb{Z}) = Z_n(K)/B_n(K) = Ker(\partial_n)/Im(\partial_{n+1}).$$

Además, el *número de Betti* descrito en el Resultado 1.7 se encuentra aquí también para un K con número de símplices finitos, siendo $n \in \mathbb{N}$, tenemos:

$$\beta_n = rango(H_n(K;\mathbb{Z})).$$

Añadimos unas definición sobre cadenas que impondrá una relación de equivalencia:

Definición 1.47. Dos n-cadenas son *homólogas* si $c_1 - c_2 = \partial_{n+1} c_3$ para una $(n+1)$-cadena c_3, es decir, si $c_1 - c_2$ es un elemento de $Im(\partial_{n+1})$. En particular, si $c_1 = \partial_{n+1} c_3$ decimos que c_1 es *homóloga a cero.*

Del $0-$grupo de homología hay una interpretación directa que merece la pena resaltar:

Proposición 1.48. *Sea K un complejo simplicial abstracto no vacío. El grupo de homología $H_0(K;\mathbb{Z})$ es una suma directa de copias de $\mathbb{Z}$, una por cada componente por caminos de K. Esto es, $H_0(K;\mathbb{Z})$ es un espacio vectorial cuyos generadores son biyectivos a las componentes por caminos de K.*

Generalmente, y subiendo de dimensión, dado un complejo simplicial K que contenga el borde de un n-símplice pero no el n-símplice en sí, es decir, sólo su frontera con un agujero en medio, el grupo de homología $H_{n-1}(K;\mathbb{Z})$ va a tener una clase que represente ese agujero ya que la propia idea del grupo de homología es poder detectarlos.

Ejemplo 1.49. Pasamos a un ejemplo de cómo calcular los grupos de homología para un complejo simplicial orientado. Tomamos el complejo simplicial de la Fig. 1.5, que vemos que

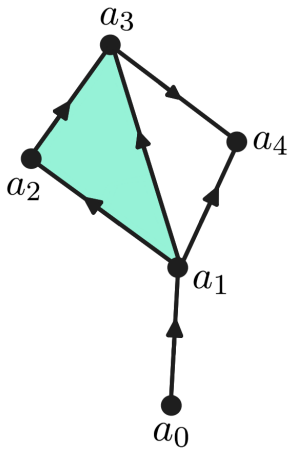


Figura 1.5: Complejo simplicial orientado

tiene como vértices a_0, a_1, a_2, a_3 y a_4. Definimos sus cadenas no nulas para poder realizar el cálculo:

$$e_1 = [a_0, a_1]\,, \qquad e_2 = [a_1, a_2]\,, \qquad e_3 = [a_1, a_3]\,,$$
$$e_4 = [a_1, a_4]\,, \qquad e_5 = [a_2, a_3]\,, \qquad e_6 = [a_3, a_4]\,.$$
$$\sigma = [a_1, a_2, a_3]\,.$$

Para calcular los grupos de homología, H_0 y H_1 del complejo simplicial, debemos calcular primero sus imágenes y núcleos del operador borde aplicado en cada dimensión.
Empezamos por la única 2-cadena que tenemos:

$$\partial_2(\sigma) = \partial_2 [a_1, a_2, a_3] = [a_2, a_3] - [a_1, a_3] + [a_1, a_2] = e_5 - e_3 + e_2.$$

Y para las 1-cadenas:

$$\begin{aligned}\partial_1(\sum_{i=1}^{6} k_i e_i) &= k_1(a_1 - a_0) + k_2(a_2 - a_1) + k_3(a_3 - a_1) + \\ &+ k_4(a_4 - a_1) + k_5(a_3 - a_2) + k_6(a_4 - a_3) = \\ &= a_0(-k_1) + a_1(k_1 - k_2 - k_3 - k_4) + a_2(k_2 - k_5) + a_3(k_3 + k_5 - k_6) + a_4(k_4 + k_5).\end{aligned}$$

Calculamos ahora los grupos de homología:

$$H_0(K) = Ker(0)/B_0(K) = \frac{\langle a_0, a_1, a_2, a_3, a_4 \rangle}{\langle a_1 - a_0, a_2 - a_1, a_3 - a_1, a_4 - a_3 \rangle} \cong \mathbb{Z}.$$

$$H_1(K) = Z_1(K)/B_1(K) = \frac{\langle e_2 - e_3 + e_5, e_3 - e_4 + e_6 \rangle}{\langle e_2 - e_3 + e_5 \rangle} \cong \mathbb{Z}.$$

Para H_1 tomamos los 1-símplices que generan los dos "triángulos" del complejo, viendo que hay uno de ellos que se desvanecerá en la homología. Esto se puede apreciar al observar el ejemplo, pero es cierto que la homología no se puede calcular normalmente a simple vista, para lo cual existen algoritmos específicos de cálculo, que no trataremos en este trabajo, y que trabajan normalmente en espacios vectoriales para simplificar las cuentas necesarias.

En resumen, el grupo de homología n-dimensional, $H_n(K;\mathbb{Z})$ es una medida de las características geométricas n-dimensionales de K, específicamente el número de "agujeros" n-dimensionales que tiene K.

Una de las grandes ventajas de la homología simplicial es que se puede calcular con un ordenador (es computacionalmente viable) para complejos simpliciales abstractos.

1.3.2. Homología de complejos de cadenas

Es destacable ahora la intuición de que los grupos de homología H_n sean invariantes homotópicos. Es útil, a la hora de analizar esta posibilidad, tomar los grupos de homología como functores. El hecho de que sean functoriales es, precisamente, una de las propiedades más relevantes de los invariantes topológicos. Por la propia definición de homología, podemos ver que es un functor:

$$H_n : \mathbf{Simp} \to \mathbf{Ab},$$

donde **Simp** es la categoría cuyos objetos son los complejos simpliciales, y los morfismos son las aplicaciones simpliciales.

Observación 1.50. Como hemos mencionado al principio de la sección y por la construcción que estamos haciendo, estudiaremos el complejo de cadenas a través de los grupos abelianos, pero es posible extender esta noción a los espacios vectoriales sobre un cuerpo, $\mathrm{Vect}_{\mathbb{F}}$, ya que los grupos abelianos no dejan de ser $\mathbb{Z}$-módulos. Por no recargar este trabajo de más, los obviaremos.

Para las definiciones y resultados de esta sección, nos apoyaremos en [8]. Hemos definido en 1.43 los complejos de cadenas. Estos, serán los objetos de la categoría de Complejos de cadenas de grupos abelianos, tomando coeficientes en $\mathbb{Z}$.

Ahora, vamos a ver cuáles serían los morfismos de esta categoría:

Definición 1.51 (Aplicaciones de complejos de cadenas)**.** Una *aplicación de complejos de cadenas* $f : A_\bullet \to B_\bullet$ es una colección de transformaciones lineales $f_k : A_k \to B_k$ para cada $k \in \mathbb{Z}$ tal que la aplicación conmute con la operación borde: $f_{k-1} \circ \delta_k^A = \delta_k^B \circ f_k$.

Es decir, debe conmutar el diagrama:

$$\begin{array}{ccc}
\vdots & & \vdots \\
\downarrow{\scriptstyle \partial^A_{k+1}} & & \downarrow{\scriptstyle \partial^B_{k+1}} \\
A_k & \xrightarrow{f_k} & B_k \\
\downarrow{\scriptstyle \partial^A_k} & & \downarrow{\scriptstyle \partial^B_k} \\
A_{k-1} & \xrightarrow{f_{k-1}} & B_{k-1} \\
\downarrow{\scriptstyle \partial^A_{k-1}} & & \downarrow{\scriptstyle \partial^B_{k-1}} \\
\vdots & & \vdots
\end{array}$$

Con estas aplicaciones como morfismos de la categoría **Ch(Ab)**, estaría ya bien definida.

Además, tenemos un functor que surge de manera natural **Ab** $\to$ **Ch(Ab)** que lleva un objeto de la categoría de grupos abelianos al complejo de cadenas:

$$A_\bullet = 0 \to ... \to 0 \to A \to 0 \to 0 \to ...$$

Donde el único elemento no nulo de la cadena es el propio grupo abeliano.

Análogamente a la Definición 1.46, pero para complejos de cadenas:

Definición 1.52. Para un complejo de cadenas $B_\bullet$, y $k \in \mathbb{Z}$, el k-ésimo grupo de homología H_k es

$$H_k = Ker(\delta_k)/Im(\delta_{k+1}).$$

Construímos la homología de manera functorial:

Lema 1.53. *Una aplicación $f : A_\bullet \to B_\bullet$ de complejos de cadenas induce una aplicación $H_k(f) : H_k(A_\bullet) \to H_k(B_\bullet)$, ya que*

$$f_k(Ker(\delta_k : A_k \to A_{k-1})) \subset Ker(\delta_k : B_k \to B_{k-1}),$$

*y de manera análoga para $Im(\delta_{i+1})$. Por tanto, obtenemos un functor H_k : **Ch(Ab)** $\to$ **Ab**.*

Definición 1.54 (Quasi-isomorfismo)**.** Una aplicación $f : A_\bullet \to B_\bullet$ es un *quasi-isomorfismo* si la aplicación inducida $H_k(f) : H_k(A_\bullet) \to H_k(B_\bullet)$ es un isomorfismo para cada $k \in \mathbb{Z}$.

Es importante destacar que si cada f_k es un isomorfismo, no necesariamente es f un isomorfismo, pero sí es un quasi-isomorfismo en base a la definición dada, de ahí su importancia.
Añadimos una definición adicional por su presencia en la literatura consultada:

Definición 1.55. Un complejo de cadenas $A_\bullet$ es *acíclico* si todos los objetos del grupo de homología $H_i(A_\bullet)$, son 0.

Lema 1.56. *Una aplicación de complejos simpliciales $K \to K'$ determina una aplicación de cadenas $C_\bullet(K;\mathbb{Z}) \to C_\bullet(K';\mathbb{Z})$. Es decir, el paso a las cadenas simpliciales induce un functor*

$$C_\bullet(-,\mathbb{Z}) : \mathbf{Simp} \longrightarrow \mathbf{Ch(Ab)}.$$

Y aplicando homología, tenemos el functor compuesto:

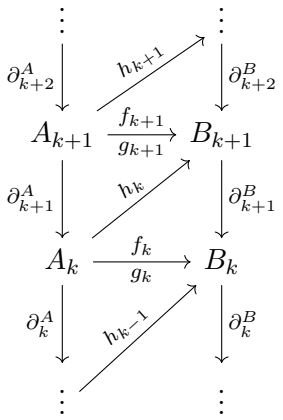
$$H_n(-,\mathbb{Z}) : \mathbf{Simp} \longrightarrow \mathbf{Ch(Ab)} \longrightarrow \mathbf{Ab}.$$

De esto se puede deducir que la homología es functorial sobre las cadenas. Vemos ahora cuándo dos morfismos dentro de un complejo de cadenas definen la misma aplicación en la homología.

Definición 1.57 (Homotopía por cadenas)**.** Dos aplicaciones en los complejos de cadenas, $f, g : A_\bullet \to B_\bullet$ tienen la misma *homotopía por cadenas* si existen aplicaciones $h_k : A_k \to B_{k+1}$ tales que

$$f_k - g_k = \partial^B_{k+1} \circ h_k - h_{k-1} \circ \partial^A_k.$$

Es decir, de forma que el siguiente diagrama conmute:



La familia $h = \{h_n\}$ se llama *homotopía por cadenas*, y diremos que f tiene la misma *homotopía por cadenas* a g.

Definición 1.58. Si $f, g : X \to Y$ son aplicaciones de complejos simpliciales abstractos tales que $|f|, |g| : |K| \to |S|$ tienen la misma homotopía, entonces las aplicaciones inducidas $f, g : C_\bullet(K, \mathbb{Z}) \to C_\bullet(S; \mathbb{Z})$, tienen la misma homotopía por cadenas.

Llegamos a los dos resultados más importantes de la sección:

Proposición 1.59. *Si dos aplicaciones $f, g : A_\bullet \to B_\bullet$ de complejos de cadenas tienen la misma homotopía por cadenas, entonces inducen la misma aplicación en homología.*

Lo que esto viene a decir es que la diferencia entre f_k y g_k es, simplemente, la frontera:

$$\partial^B_k \circ (f_k - g_k) = \partial^B_k \circ \partial^B_{k+1} \circ h_k - \partial^B_k \circ h_{k-1} \circ \partial^A_k = \partial^B_k \circ h_{k-1} \circ \partial^A_k.$$

Corolario 1.60. *Si $f : K \to S$ es una aplicación simplicial de complejos simpliciales abstractos, tales que $|f| : |K| \to |S|$ es una equivalencia por homotopías, entonces f induce un isomorfismo en la homología.*

Esto viene a decir que tenemos un functor $H_k : \mathbf{Ho(Simp)} \to \mathbf{Ab}$, donde la categoría **Ho(Simp)** es aquella cuyos objetos son los complejos simpliciales abstractos y los morfismos de K a S vienen dados por las clases de homotopía de las aplicaciones $|K| \to |S|$.

Tras haber visto todos los resultados relevantes relacionados con la homología para complejos de cadenas, veamos un último resultado relacionado con el grupo fundamental, antes de pasar a la siguiente sección.

Teorema 1.61. *Sea K un complejo simplicial abstracto conexo por caminos. El grupo de homología $H_1(K;\mathbb{Z})$ es la abelianización del grupo fundamental $\pi_1(K,x_0)$. Dicha abelianización consiste en cocientar el grupo por el subgrupo conmutador, es decir por aquellos elementos de la forma $xyx^{-1}y^{-1}$, con $x,y \in \pi_1(K,x_0)$.*

Demostración. Teorema 2A.1, Sección A de [7] □

Cerramos el capítulo con un comentario sobre la viabilidad computacional de estos cálculos, ya que el objetivo en mente es siempre poder calcularlos o computarlos de alguna manera. Tenemos, por tanto:

Teorema 1.62. *Dado un complejo simplicial, existe un algoritmo para computar $H_k(-;\mathbb{Z})$ cuyo tiempo de ejecución es el cubo del total de $(k+1)$-símplices, k-símplices y $(k-1)$-símplices.*

Hemos descrito los complejos simpliciales, que aportan una manera discreta y un marco de trabajo para poder estudiar los espacios topológicos de manera combinatoria. Esta naturaleza combinatoria desencadena en la homología simplicial, donde el n-ésimo grupo de homología mide la diferencia entre $C_n(K)$, el grupo de combinaciones lineales de n-símplices orientados, y el conjunto de bordes de los elementos de $C_{n+1}(K)$, es decir, el grupo de homología da información de los agujeros n-dimensionales de K. Pueden ser calculados computacionalmente usando álgebra lineal.

Ahora, orientaremos estas herramientas a estudiar cómo se comportan los grupos de homología para un conjunto arbitrario de puntos, que representen nuestros datos. Esto se hará usando la **homología persistente**.

Capítulo 2

Homología persistente: análisis topológico de datos

Para esta sección nos apoyaremos en [6] y, por supuesto, en [12].
En el análisis topológico de datos existe una problemática fundamental: ser capaz de inferir una estructura de dimensión elevada a través de una representación a baja dimensión y, a partir de ahí, cómo relacionar el conjunto discreto de puntos para generar una estructura global, y continua.

Un ejemplo muy trivial de esto, podría ser el firmamento. Hay miles de estrellas en el cielo observables (que podrían ser nuestros 0-símplices), pero el ser humano las ha agrupado de forma que formen dibujos, denominadas constelaciones (los 1-símplices) e ilustrándolas en 3D haciendo figuras de ellas (lo que podría ser un 2-símplice). Es cierto que en este ejemplo se puede tomar una arbitrariedad a la hora de unir los puntos, ya que no necesariamente son los que más cerca están entre sí si no los que generan una estructura base de fondo: la estructura algebraica. Esta búsqueda de la estructura subyacente de un conjunto discreto de puntos será el objetivo final del análisis topológico de datos.

Con este fin, buscamos encontrar un flujo de información a partir de el conjunto de puntos disponibles:

$$\{\text{datos}\} \to \{\text{complejos simpliciales}\} \to \{\text{invariantes topológicos}\}$$

Como mencionamos, es necesario construir un conjunto discreto de datos a un espacio topológico para poder estudiar la forma de los datos y poder calcular sus invariantes topológicos. Para esto es importante calcular su escala de características (*feature scale*, en inglés), valor variable que determina qué características de los datos (grupos de homología, por ejemplo) aparecen o desaparecen según la distancia entre los puntos de un conjunto. Para una alta tolerancia

(un mayor umbral) de escala, que permite grandes distancias entre puntos, quizás encontremos demasiadas relaciones y se emborrone la imagen subyacente; para una tolerancia muy baja (permitiendo sólo las distancias más inmediatas) quizás no se llegue a capturar ninguna estructura. Por tanto, es importante obtener la información para todos los valores de la escala en los que se pueden encontrar relaciones interesantes y aportar información útil.

Esto se puede resolver a través de la *homología persistente*, herramienta relativamente reciente muy útil para el análisis de datos a través de la topología.

2.1. Datos a través de los complejos simpliciales

Comenzamos estableciendo dónde vamos a trabajar. Como estaremos "midiendo" distancias entre puntos, partimos de un espacio métrico, por ejemplo, $\mathbb{R}^n$ con una métrica arbitraria ∂_X. Dentro de un espacio métrico, podemos definir una *nube de puntos* que será una colección finita de puntos en dicho espacio, $\{x_i\}_{i\in I}$, $I = \{0, ...n\}$, que conformará un subespacio finito X. Con esto, buscamos crear estructuras que relacionen estos puntos. Empezamos definiendo un complejo de Čech.

Definición 2.1 (Complejo de Čech)**.** Sea $X \subset \mathbb{R}^n$ un subespacio finito, y un $\varepsilon > 0$ fijado. El *complejo de Čech*, C_ε, o $C_\varepsilon(X, \partial_X)$, es el complejo simplicial abstracto:

1. Cuyos vértices son los puntos x_i de X.
2. Cada k-símplice viene dado por $(k+1)$-tupla de puntos, $\{x_0, ..., x_k\} \in X$ para los cuales toda bola cerrada centrada en dichos puntos, y radio $\varepsilon/2$ tiene, por lo menos, un punto común en cada intersección. Es decir, $\bigcap_{i<k} \overline{B(x_i, \varepsilon/2)} \neq \emptyset$.

El complejo de Čech permite asignar un complejo simplicial a un espacio métrico finito embebido en $\mathbb{R}^n$. De esta manera, se puede también asociar el complejo simplicial con el recubrimiento de un espacio. Dado un recubrimiento $\{U_i\}_{i\in I}$ del espacio finito X, definimos:

Definición 2.2. El *nervio* $N(\{U_i\})$ de un recubrimientos $\{U_i\}_{i\in I}$ de X es el complejo simplicial:

1. Cuyos vértices corresponden a los conjuntos de $\{U_i\}_{i\in I}$.
2. Para j_i índices del recubrimiento, tiene un k-símplice $[j_0, ..., j_k]$ para cada intersección no vacía $U_{j_0} \cap U_{j_1} \cap ... \cap U_{j_k} \neq \emptyset$.

Tenemos ahora un resultado relevante relacionado con la realización geométrica, que era el functor que llevaba la categoría de complejos simpliciales abstractos y sus aplicaciones a los espacios topológicos y aplicaciones continuas, usando el esquema de vértices.

Teorema 2.3 (Teorema de Čech)**.** *Sea X un espacio topológico. Sea $\{U_i\}_{i\in I}$ una cobertura abierta de X tal que todas las intersecciones finitas no vacías,*

$$U_{j_0} \cap U_{j_1} \cap ... \cap U_{j_k}$$

son contráctiles, es decir, equivalentes homotópicamente a un punto. Entonces, la realización geométrica $|N(\{U_i\})|$ tiene el mismo tipo de homotopía que X.

Existe un corolario que relaciona el nervio de Čech geométrico con la realización geométrica de un complejo de Čech, que queda reflejada aquí:

Corolario 2.4. *Sea $X \subset \mathbb{R}^n$ el subespacio finito generado a partir de la colección finita de puntos $\{x_i\}_{i\in I}$, y un $\varepsilon > 0$ fijado. Entonces, existe un homeomorfismo entre la unión de las bolas y la realización geométrica de un complejo de Čech: $\bigcup_{x\in X} B(x,\varepsilon) \cong |C_\varepsilon(X, \partial_X)|$.*

Aunque la construcción de los complejos de Čech tiene su interés gracias el Teorema 2.3, es cierto que ni podemos contar siempre (ni nos interesa) con la métrica euclídea, ni es eficiente exigir que la intersección de todas las bolas B_ε sea no vacía una vez aumentamos la dimensión. Por tanto, se definen los *complejos de Vietoris-Rips*, que serán una construcción menos exigente que los complejos de Čech.

Definición 2.5 (Complejo de Vietoris-Rips)**.** Para una colección de puntos $X = \{x_i\}_{i\in I}$ en un espacio métrico finito, con la métrica euclídea ∂_X, y un $\varepsilon > 0$ fijado. El *complejo de Vietoris-Rips* $VR_\varepsilon(X, \partial_X)$ es el complejo simplicial abstracto tal que:

1. Cuyos vértices son los puntos x_i de X.
2. Tiene un k-símplice orientado $[x_0, x_1, ..., x_k]$ dado por una $(k+1)$-tupla de puntos, tal que distan menos de ε dos a dos, es decir:

$$\partial_X(x_i, x_j) \leq \varepsilon, \quad 0 \leq i, j \leq k$$

.

Observación 2.6. Los complejos de Vietoris-Rips pueden ser generalizados a través de los *complejos de clique* (o complejos de pandillas). Estos son aquellos complejos cuyos puntos están todos conectados a través de 1-símplices, es decir, existen caminos entre cualesquiera 2 puntos independientemente de la distancia entre ellos. Esto es relevante, sobre todo, en la teoría de grafos.

Los complejos de Vietoris-Rips y los de Čech, a pesar de ser muy parecidos, **no** son iguales, como se puede observar en la Figura 2.1. A nivel computacional, a partir de cierta dimensión

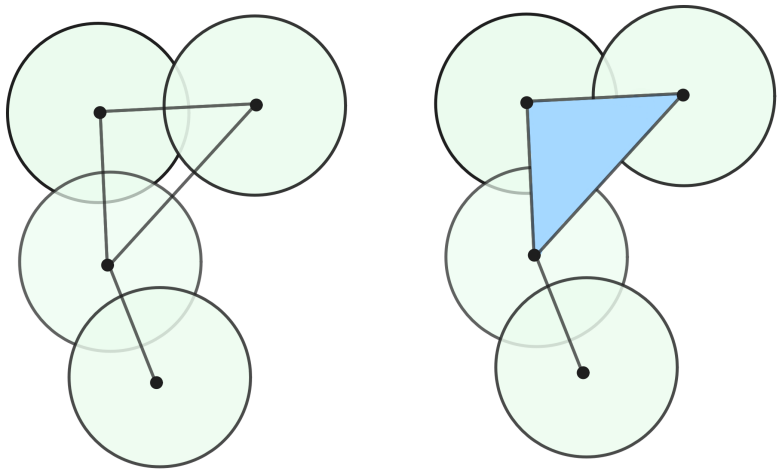


Figura 2.1: Ejemplo de un complejo de Čech (izq.) y un complejo de Vietoris-Rips, para el mismo ε fijado.

sólo es viable calcular los de Vietoris-Rips, pero en cambio éstos no dan una equivalencia por homotopías y los complejos de Čech sí, por lo que (aun siendo los primeros una generalización de los segundos), merece la pena definir y conocer ambos.

Marcamos ahora exactamente cuál es la relación entre ambos complejos:

Lema 2.7. *Sea $X \subset \mathbb{R}^n$ sea un subespacio finito, y un $\varepsilon > 0$ fijado. Tenemos las inclusiones:*

$$C_\varepsilon(X, \partial_X) \subseteq VR_\varepsilon(X, \partial_X) \subseteq C_{2\varepsilon}(X, \partial_X).$$

Acabamos esta sección con una de las propiedades más importantes de los complejos de Vietoris-Rips, aunque sería análogo para los complejos de Čech, que es su functorialidad:

Sea el $\varepsilon > 0$ fijado, tomamos un $\hat{\varepsilon}$ tal que $\varepsilon < \hat{\varepsilon}$, y cualquier espacio métrico (X, ∂_X), tenemos una aplicación simplicial dada por la inclusión:

$$VR_\varepsilon(X, \partial_X) \to VR_{\hat{\varepsilon}}(X, \partial_X),$$

ya que aumentar el umbral únicamente añade más símplices, por lo que tenemos un functor de $(\mathbb{R}_{>0}, \leq)$ a **Simp** para la construcción $\mathrm{VR}_{(-)}(X, \partial_X)$, donde $(\mathbb{R}_{>0}, \leq)$ es la categoría de objetos los reales mayores a 0 y con flechas que respetan la desigualdad $\leq$.

Esto se puede hacer no sólo variando el parámetro del umbral, si no también cambiando de espacio métrico. Para ello, recordamos que una aplicación $f : X \to Y$ entre dos espacios métricos (X, ∂_X) y (Y, ∂_Y) es lipschitziana con constante k, si $\partial_Y(f(x_1), f(x_2)) \leq k\partial_X(x_1, x_2),\ 0 < k \leq 1$. Para una función lipschitziana de constante de Lipschitz k, $f : X \to Y$ hay un mapa simplicial inducido

$$f : VR_\varepsilon(X, \partial_X) \to VR_{k\varepsilon}(Y, \partial_Y).$$

Así, existe un functor para la construcción $\mathrm{VR}_\varepsilon(-)$ entre la categoría de los espacios métricos finitos, y aplicaciones lipschitzianas con $k = 1$, y **Simp**.

Es decir, para dos conjuntos finitos de puntos (o datos) X, Y y dos umbrales fijados, $0 < \varepsilon < \hat{\varepsilon}$, tenemos que el diagrama

$$\begin{array}{ccc} VR_\varepsilon(X, \partial_X) & \longrightarrow & VR_\varepsilon(Y, \partial_Y) \\ \downarrow & & \downarrow \\ VR_{\hat{\varepsilon}}(X, \partial_X) & \longrightarrow & VR_{\hat{\varepsilon}}(Y, \partial_Y) \end{array}$$

conmuta.

2.2. Homología persistente

Como comentamos al principio de la sección, según el cierta escala de características, podemos obtener un resultado muy distinto de otro al estudiar la homología. En este caso, el umbral viene determinado por ε, pero con sólo escoger uno no es suficiente. Los invariantes topológicos de los complejos de Vietoris-Rips son muy sensibles al ruido en el dato, y al variar la muestra puede dar resultados muy dispares en su homología. La homología persistente busca, una vez se recoja toda la información posible, discernir qué "agujeros" son los relevantes: cuáles siguen existiendo al cambiar el parámetro y aumentar la dimensión, y cuáles son los que desaparecen o sólo aparecen fugazmente y que deberán ser ignorados.

Tendremos, pues, como filosofía general la noción de que es necesario estudiar varios umbrales para poder elegir las características homológicas que son estables y existen para un rango de ε distintos, y sí reflejan con veracidad las características del verdadero espacio topológico subyacente.

A la hora de poder calcular computacionalmente los distintos grupos de homología, debemos escoger previamente un conjunto finito de posibles $\{\varepsilon_i\}$, preferiblemente aquellos que supongan un "punto de inflexión" para el complejo de Vietoris-Rips: valores donde va a haber un cambio significativo. Explicamos este concepto en el siguiente lema:

Lema 2.8. *Sea (X, ∂_X) un espacio métrico finito. Entonces existen, como máximo, un número finito de valores $\{\varepsilon_i\}$ a partir de los cuales $VR_{\varepsilon_i}(X, \partial_X)$ cambia. Es decir, para que tomando unos δ suficientemente pequeños:*

$$\begin{cases} VR_\varepsilon(X, \partial_X) = Z, & \varepsilon \in [\varepsilon_i - \delta,\ \varepsilon). \\ VR_\varepsilon(X, \partial_X) = Z', & \varepsilon \in [\varepsilon_i,\ \varepsilon_i + \delta.] \end{cases}$$

para $Z \neq Z'$.

Aun así, aun después de escoger los ε_i que se encuentran en estos puntos de inflexión, seguimos necesitando poder comparar los valores distintos del grupo de homología. Aquí es donde entra el concepto de la **persistencia**, que realizará esta comparación. Para esto es importante recordar que, como $VR_{(-)}(X, \partial_X)$ es functorial en ε para $\varepsilon < \hat{\varepsilon}$, tenemos la aplicación de complejos simpliciales

$$VR_{\varepsilon}(X, \partial_X) \to VR_{\hat{\varepsilon}}(X, \partial_X).$$

De hecho, tomando una colección de $\{\varepsilon_i\}$ tal que $\varepsilon_1 < \varepsilon_2 < ... < \varepsilon_l$, obtenemos la secuencia de aplicaciones simpliciales

$$VR_{\varepsilon_1}(X, \partial_X) \to VR_{\varepsilon_2}(X, \partial_X) \to ... \to VR_{\varepsilon_l}(X, \partial_X).$$

Y como el grupo de homología H_k también es un functor, tenemos los homomorfismos de grupos abelianos:

$$H_k(VR_{\varepsilon_1}(X, \partial_X)) \to H_k(VR_{\varepsilon_2}(X, \partial_X)) \to ... \to H_k(VR_{\varepsilon_l}(X, \partial_X)).$$

Formalizándolo con categorías, tenemos:

Definición 2.9. Dado un espacio métrico finito fijado, (X, ∂_X), el complejo de Vietoris-Rips induce un functor

$$VR_{(-)}(X, \partial_X) : (\mathbb{R}_{>0}, \leq) \to \mathbf{Simp}$$

que va de los números reales (en la selección de ε_i), tomado como la categoría asociada a un conjunto parcialmente ordenado a la categoría de los complejos simpliciales. Podemos componer con el functor del k-ésimo grupo de homología y que da lugar al functor

$$H_k(VR_{(-)}(X, \partial_X)) : (\mathbb{R}_{>0}, \leq) \to \mathbf{Ab}.$$

Esta es, al final, la estructura que sigue la construcción de la homología persistente. Damos un par de definiciones más para completar el significado de esto, siguiendo [6] y, sobretodo [4].

Asumimos que, para una sucesión creciente de valores $\{\varepsilon_i\}_{i=1}^n$ y una sucesión de complejos de Vietoris-Rips asociados a una nube de puntos, $\mathcal{VR} = \{VR_i\}_{i=1}^n = \{VR_{\varepsilon_i}(X, \partial_X)\}_{i=1}^n$, hay una serie de inclusiones naturales

$$VR_1 \overset{\iota}{\hookrightarrow} VR_2 \overset{\iota}{\hookrightarrow} ... \overset{\iota}{\hookrightarrow} VR_n.$$

Formalizamos esto con una definición:

Definición 2.10 (Sistema filtrado)**.** La familia de los complejos simpliciales descritos, $\mathcal{VR} = \{VR_{\varepsilon_i}(X, \partial_X)\}_{i=1}^n$, con sus inclusiones, $VR_1 \subseteq VR_2 \subseteq ... \subseteq VR_n$, es el *complejo simplicial filtrado de Vietoris-Rips.*

Como hemos visto en el capítulo anterior, una aplicación simplicial entre complejos (en este caso, de Vietoris-Rips) va a inducir un homomorfismo entre los grupos de homología asociados. En este caso, la aplicación inclusión para la familia $\mathcal{VR}$, esto induce un homomorfismo de inclusión también en sus grupos de homología para una dimensión fija q:

$$0 = H_q(VR_1) \xrightarrow{\tau} H_q(VR_2) \xrightarrow{\tau} ... \xrightarrow{\tau} H_q(VR_i) \xrightarrow{\tau} H_q(VR_k) \xrightarrow{\tau} ... \xrightarrow{\tau} H_q(VR_q)$$

para $i \leq k$.

En estas circunstancias, definimos el homomorfismo inducido como $\tau_q^{i,j} : H_q(VR_i) \to H_q(VR_k)$.

Al ir aumentando de dimensión en los complejos, se van añadiendo y desvaneciendo clases en el grupo de homología. Según unos umbrales marcados, recogemos las clases de homología que se crearon antes del umbral, y que se pierden tras él. Así, definimos:

Definición 2.11 (Homología persistente)**.** El q-ésimo grupo de homología persistente son las imágenes de los homomorfismos inducidos por la inclusión τ, es decir:

$$H_q^{i,j} = \text{im}\tau_q^{i,j}, \quad 0 \leq i \leq j \leq n,$$

y para cada dimensión del grupo q, el q-ésimo número de Betti persistente es, análogamente, el rango de cada grupo de homología persistente

$$\beta_q^{k,l} = \text{rango}(H_q^{i,j}).$$

Exactamente la información que nos aporta grupo de homología persistente entre un i y un j son las clases de homología que "persisten" de un grupo a otro, es decir, aquellas que existen en VR_i y que también están en VR_j, y no desaparecieron sino que se mantienen.

Esta diferencia de los índices, $j - i$, se llama *índice de persistencia* e indica en cuántos grupos de homología persiste una clase. El máximo índice de persistencia para cada clase es lo que indicará la significancia y estabilidad de dicha característica geométrica. Si la clase nunca muere, su índice de persistencia será infinito.

2.2.1. Códigos de barras

Los códigos de barras son una manera visual de ver la persistencia de una clase de homología a medida que se va aumentando el umbral, ε, en las distintas dimensiones, H_i.

Definición 2.12. Un *código de barras* (o *barcode*, en inglés), es un conjunto de intervalo no vacíos de la forma $[x, y) \subset \mathbb{R}$ o $[x, \infty)$, que describe la homología de una familia de complejos simpliciales filtrados a medida que varía su umbral, ε.

Este código de barras representa la vida de un elemento particular de la homología, mediante una representación en $\mathbb{R}^2$, llamado *diagrama de persistencia*. Cuando el intervalo es amplio, se infiere que el el elemento en homología (el "agujero") que representa es muy relevante, y cuando es un intervalo pequeño, se interpreta que sólo era ruido o alguna característica transitoria.

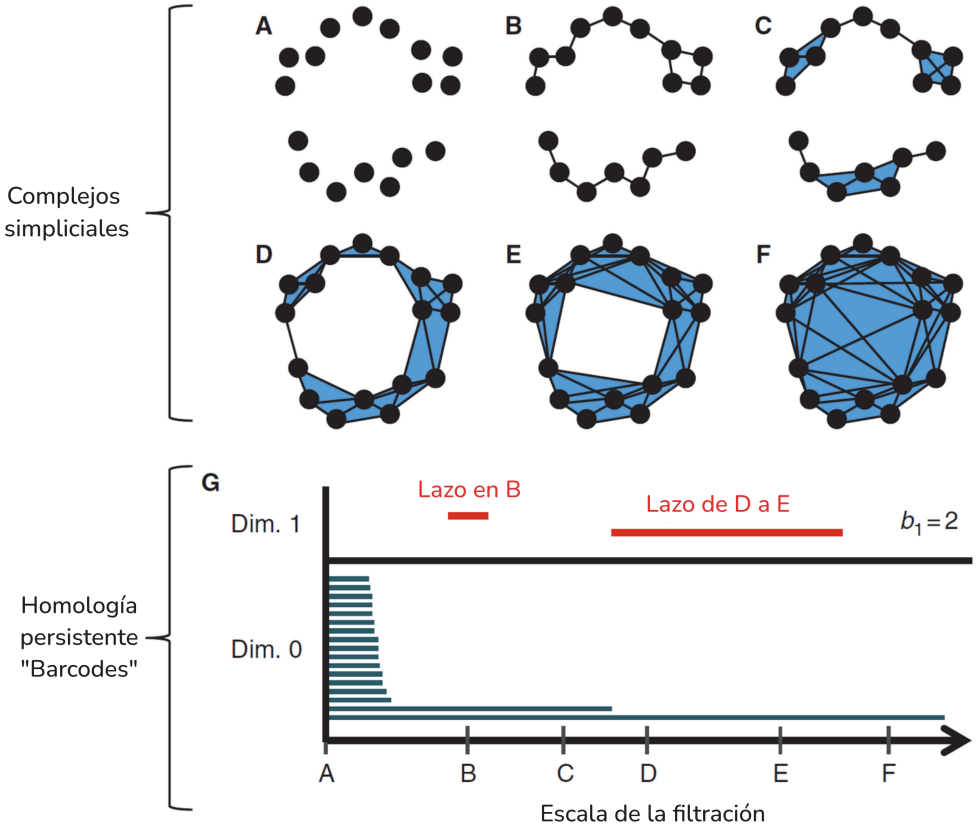


Figura 2.2: Figura sacada de [12].

Ejemplo 2.13. Vemos en la Figura 2.2 el diagrama superior con complejos simpliciales, y el inferior con su código de barras (o "barcodes") correspondiente. A medida que se aumenta la escala de la filtración, es decir, el umbral ε se empiezan a unir los vértices y aparecen clases d homología no triviales. En particular y de interés en este caso son las 2 clases de homología de dimensión 1 que aparecen, primero en la figura B (arriba a la derecha), y más tarde aparece un lazo que se mantiene de la figura D a la E, a lo largo de 2 umbrales distintos. Estos vienen representados en el código de barras del diagrama inmediatamente inferior. Se puede observar, también, cómo entre C y D se desvanece la clase homología de dimensión 0 y aparece una de dimensión 1 cuando se "une" la parte de arriba con la de abajo, formando un agujero en medio, que aparece como barra en la dimensión superior.

Por tanto y hasta este momento, en base a una nube de puntos, tenemos el siguiente flujo

de información:

{Espacios métricos finitos} → {Complejos simpliciales filtrados} → {Diagramas de persistencia}.

2.3. Persistencia zigzag

Aunque se defina la homología persistente en base al sistema de filtraciones de complejo recién expuesto, existen otros tipos de filtración que se pueden dar, y pueden dar una visión más general de la homología. Es así que surge la *persistencia en zigzag*, donde las inclusiones no solo van en una dirección constantemente, sino que pueden ir variando en sentido a lo largo del diagrama. Es decir, para nuestro conjunto de muestras X_i y un espacio métrico fijado (X, ∂_X), puede darse el siguiente diagrama:

$$X_1 \longrightarrow X_1 \cup X_2 \longleftarrow X_2 \longrightarrow X_2 \cup X_3 \cup X_4 \longleftarrow X_2 \cup X_3$$

En este caso, siendo las flechas inclusiones, podemos también componerlo con el grupo de homología para cada elemento, obteniendo así el siguiente diagrama:

$$\begin{array}{ccccc} H_k(VR_\varepsilon(X_1)) & \longrightarrow & H_k(VR_\varepsilon(X_1 \cup X_2)) & \longleftarrow & H_k(VR_\varepsilon(X_2)) \\ & & & & \downarrow \\ & & & & H_k(VR_\varepsilon(X_2 \cup X_3 \cup X_4)) \end{array} \longleftarrow H_k(VR_\varepsilon(X_2 \cup X_3)).$$

Es decir, se puede dar tanto $VR_i \to VR_{i+1}$ como $VR_i \leftarrow VR_{i+1}$, ya que el índice no representa una construcción más general.

Para poder trabajar y definir estas nuevas filtraciones, denotamos el *diagrama en zigzag* como S, que es una secuencia de letras L o R, según si las flechas avanzan hacia la izquierda o a la derecha, respectivamente. Se puede interpretar cuando va a la derecha como que el objeto matemático "crece" (el zig), y la flecha a la izquierda como que se rompe o se "descompone" (el zag).

Observación 2.14. En consistencia con lo visto hasta ahora, tomando la homología como un grupo abeliano, que se puede interpretar como un $\mathbb{Z}$-módulo, expandimos ahora las siguientes definiciones a un espacio vectorial de dimensión finita, V, sobre un cuerpo arbitrario $\mathbb{F}$.

Definición 2.15 (Diagrama zigzag)**.** Para una secuencia de espacios vectoriales y aplicaciones lineales de tamaño n, denotada como $\mathbb{V}$, tenemos el diagrama

$$V_1 \xrightarrow{b_1} V_2 \xrightarrow{b_2} \dots \xrightarrow{b_{n-1}} V_n$$

donde los b_i indican flechas que pueden ir tanto a la izquierda, que denotamos como l_i, como a la derecha, r_i. Esta colección es el *diagrama zigzag* o *módulo zigzag*.

La secuencia de las aplicaciones en el diagrama es el *tipo* del diagrama. Por ejemplo, para un diagrama de tipo $\tau = rrl$ sería de la forma

$$V_1 \xrightarrow{r_1} V_2 \xrightarrow{r_2} V_3 \xleftarrow{l_3} V_4.$$

El *largo* del tipo τ es el número de objetos sobre los que está el diagrama (en este caso 4) y normalmente consideraremos diagramas con un tipo fijado n. Dichos diagramas se llaman *τ-módulos*.

Es de interés, para poder trabajar con los diagramas zigzag, poder caracterizar sus invariantes homológicos a través de un invariante numérico, como los códigos de barras. Para ello, los descomponemos primero.

Definición 2.16. Un *submódulo zigzag* $\mathbb{W}$ de un τ-módulo zigzag $\mathbb{V}$ es un módulo zigzag de tipo τ, tal que cada W_i es un subespacio de V_i, y las aplicaciones son las restricciones de b_i. Se denota como $\mathbb{W} \leq \mathbb{V}$.

Decimos que el τ-módulo zigzag $\mathbb{V}$ se puede *descomponer* si puede ser escrito como una suma directa de submódulos $\mathbb{W}_j$ no triviales, es decir, podemos escribir

$$\mathbb{V} = \bigoplus_j \mathbb{W}_j.$$

Si no existen submódulos no triviales, se dice que $\mathbb{V}$ *no admite descomposición*.

Tenemos, además, que cualquier módulo puede ser expresado como una suma directa de submódulos que no admiten descomposición, y de manera única salvo permutación.

Definición 2.17. Sea un tipo τ de longitud n, y consideramos las desigualdades $1 \leq b \leq d \leq n$ para dos enteros b, d. Tenemos que el *intervalo* τ-módulo zigzag $\mathbb{I}_\tau(b, d)$, con nacimiento en b y muerte en d está definido como la secuencia de espacios

$$I_i = \begin{cases} \mathbb{F}, & b \leq i \leq d, \\ 0 & \text{en otro caso}, \end{cases}$$

donde las aplicaciones entre las copias de $\mathbb{F}$ son la identidad, y en otro caso son 0, donde $\mathbb{F}$ es el cuerpo arbitrario sobre el que se construye el espacio vectorial.

Proposición 2.18. *Los intervalos τ-módulos no admiten descomposición.*

Ahora, presentamos el teorema que le da sentido a la construcción de los códigos de barras zigzag, en contraposición a los códigos de barras tradicionales.

Teorema 2.19. *Los τ-módulos zigzag que no se pueden descomponer son, precisamente, los intervalos τ-módulos zigzag $\mathbb{I}(b,d)$, tales que $1 \leq b \leq d \leq n = len(\tau)$. Equivalentemente, cada τ-módulo puede ser escrito como una suma directa de intervalos.*

Esto implica que cualquier módulo puede ser expresado (salvo isomorfismos) como una lista de intervalos $[b,d]$ que correspondan con los sumandos del módulo.

Como consecuencia, podemos obtener el conjunto del código de barras, que será la *persistencia en zigzag* y se representa de manera análoga a la persistencia vista en la anterior sección.

Observación 2.20. Se puede encontrar una explicación sobre el algoritmo computacional para calcular la persistencia zigzag en la sección 5 de [1].

Finalizamos ahora la sección aterrizando la persistencia zigzag sobre los complejos de Vietoris-Rips, de manera que sea cohesivo con el resto del capítulo. Sea (X, ∂_X) un espacio métrico finito, y tomamos un conjunto ordenado de puntos $X = \{x_1, x_2, x_3...\}$, y denotamos $X_k \subseteq X$ como el subconjunto que tiene los primeros k elementos ordenados del conjunto, es decir

$$X_1 = \{x_1\}, \quad X_2\{x_1, x_2\}, \quad X_3 = \{x_1, x_2, x_3\}, \quad \dots$$

Podemos definir una serie de umbrales distintos, usando la *distancia de Hausdorff*, d_H, que para dos subconjuntos no vacíos A, B de un espacio métrico (X, ∂_X), se define como

$$d_H(A,B) = \text{máx}(\sup_{a\in A} \inf_{b\in B} \partial_X(a,b),\ \sup_{b\in B} \inf_{a\in A} \partial_X(a,b)),$$

o, alternativamente

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

donde A_ε y B_ε denotan los conjuntos de todos los puntos a una distancia ε de A o B.

Así, para $\varepsilon_i = d_H(X_k, X)$, tenemos que si $\varepsilon_i \geq \varepsilon_{i+1}$, entonces X_{i+1} estará siempre igual o más cerca de X que X_i según la distancia de Hausdorff. Así, llegamos a la última definición del capítulo:

Definición 2.21. Tomamos dos números reales, $\alpha > \beta > 0$. El *zigzag de Rips* consiste en el módulo zigzag especificado en el diagrama de complejos simpliciales.

$$\begin{array}{l} \ldots \longleftarrow VR_{\beta_{\varepsilon_{i-1}}}(X_{i-1}) \longrightarrow VR_{\alpha_{\varepsilon_{i-1}}}(X_i) \longleftarrow VR_{\beta_{\varepsilon_i}}(X_i) \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \downarrow \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad VR_{\alpha_{\varepsilon_i}}(X_{i+1}) \longleftarrow VR_{\beta_{\varepsilon_{i+1}}}(X_i) \longrightarrow \ldots \end{array}$$

A través de los parámetros α y β podemos limitar el tamaño de los complejos en el módulo zigzag, lo cual es eficiente a nivel computacional.

Ahora que hemos definido todas las herramientas necesarias para poder estudiar los datos que tendremos, pasamos al siguiente y último capítulo.

Capítulo 3

Evolución de los virus desde la topología algebraica

Las referencias principales para este capítulo son aquellas que dieron pie al trabajo, [12] y [2].

3.1. Árboles filogenéticos

Desde la edad antigua se ha intentado organizar los seres vivos de manera ordenada, con el objetivo de representar y diferenciar los organismos que pueblan la Tierra. Este trabajo empieza con Aristóteles, y continúa a día de hoy.
La taxonomía moderna se funda en el siglo XVIII, de la mano de Carl Linnaeus, que introdujo la nomenclatura binomial que se usa en la actualidad (género y especie, escrito en latín), y dividió en especie, género, familia, clase y reino, clasificando así más de 14.000 especies, tanto plantas como animales.

Aun así, no es hasta el 1859 con la publicación de *El origen de las especies*, de Charles Darwin, que no se interpreta la evolución a través de la teoría de la selección natural: mutaciones aleatorias que aparecen entre un progenitor y sus descendientes, que pueden mejorar su supervivencia en el ambiente. Dichas mutaciones son transmisibles de futuros progenitores a sus descendientes, formando parte permanentemente del árbol genético. Esta acumulación de mutaciones, a lo largo del tiempo, será el mecanismo principal que de lugar a la aparición de nuevas especies.

Esta teoría replanteó la manera en la que se estudian los organismos, ya que se puede rehacer su historia y su pasado genético a parir del estudio de su genoma, y las distintas mutaciones sufridas. Para ello, una representación adecuada de este pasado o linaje, podrá ser un árbol

filogenético (Fig. 3.1).

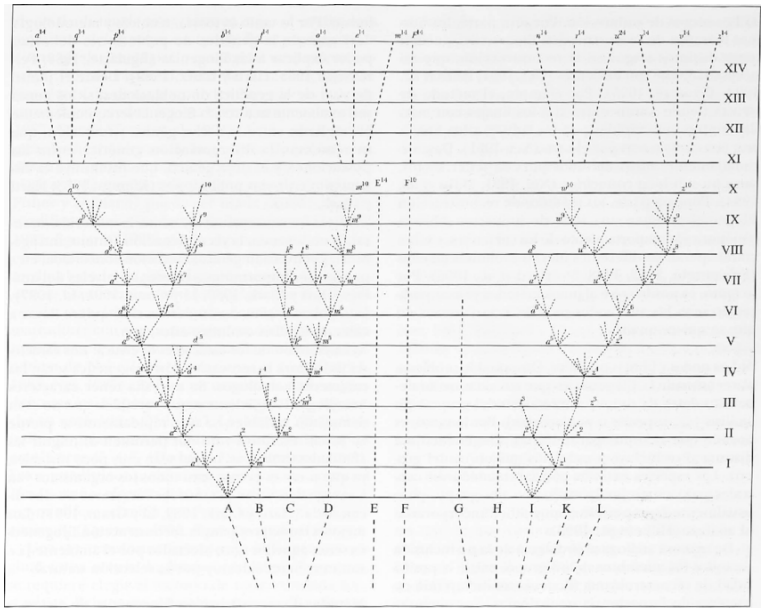


Figura 3.1: Diagrama de árbol donde se expresa la evolución de las especies a lo largo del tiempo, en este caso, subiendo por el árbol. La raíz es una especie progenitora desde donde se irá divergiendo para generar nuevas especies. Imagen sacada de *El origen de las especies* de Charles Darwin, 1859.

Aunque con el paso de los años esta clasificación se fue refinando y ampliando, la base de la evolución se estudia con diagramas de árbol. Esto, a la hora de estudiar organismos como las bacterias y los virus, presenta problemas: la información genética no sólo se transforma al pasar de un progenitor a un descendiente, sino que puede haber eventos de intercambio genético entre dos organismos, sin que produzcan descendencia. Esto es la **evolución horizontal o reticular**, que impide poder realizar un linaje claro hasta el origen de estos organismos. Por ello es importante desarrollar nuevas herramientas que ilustren bien este tipo de intercambios, evolución, y parentesco, para poder entender el origen de las epidemias, o incluso de algunos tipos de cáncer.

Daremos una base matemática sobre la que construir los diagramas filogenéticos que permitan representar este tipo de eventos. Para poder interpretar los datos que nos aportan las distintas muestras de genoma, y por tanto definir un linaje que las relacione a través de una estructura de árbol filogenético, debemos tomar, como hasta ahora, un espacio métrico finito sobre el que trabajar (X, ∂_X).

Definición 3.1. Definimos un *árbol* como un complejo simplicial 1-dimensional, conexo por caminos y finito, que no tenga ningún lazo. Un *árbol aditivo* se define como un árbol con una

función real en los vértices (1-símplices), llamada *peso*, que toma valores reales positivos.
Un *árbol filogenético* es un árbol aditivo con m vértices etiquetados del $\{1, ..., m\}$, denominados como *hojas* del árbol. La distancia en un árbol ∂_X se denomina *métrica de árbol* y se corresponde con la menor longitud (dada por los pesos) necesaria para ir de una hoja a otra, es decir, el camino más corto.

Para un espacio métrico arbitrario (X, ∂_X) podemos identificar cuándo la métrica es una métrica de árbol, empleando el siguiente lema conocido como la *condición de los 4 puntos*:

Lema 3.2. *Un espacio métrico (X, ∂_X) es isométrico a una espacio con una métrica de árbol si, y solo si, para cualquier $a, b, c, d \in X$, y las siguientes sumas*

$$\partial_X(a,b) + \partial_X(c,d), \quad \partial_X(a,c) + \partial_X(b,d), \quad \partial_X(a,d) + \partial_X(b,c),$$

dos de las tres sumas deben ser iguales, y ser mayores a la tercera suma.

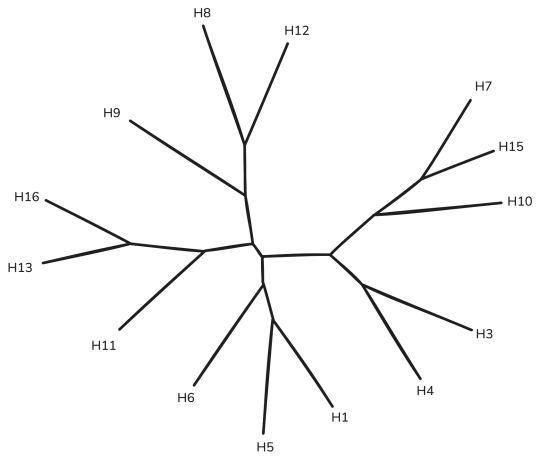


Figura 3.2: Árbol filogenético que relaciona evolutivamente los distintos tipos de hemaglutinina, proteína presente en la envoltura vírica de la Influenza A. Datos obtenidos de [2].

Para nuestros objetivos, podemos asumir que el genoma de unos organismos cuya población está en crecimiento, forma el espacio métrico (X, ∂_X) que nunca se podrá observar directamente. En cambio, podemos ver una muestra, una nube de puntos, de X. Si restringimos la métrica a dicha muestra tenemos la distancia entre puntos, y tomando los genomas como una serie de letras podemos tomar la distancia de Hamming (número de posiciones en las cuales dos series de letras son distintas). Así, tenemos un espacio métrico finito generado por las secuencias de genoma que se separan en una distancia genética, que se puede estudiar con técnicas del análisis topológico de datos.

3.1.1. Topología de la evolución

Hay dos maneras principales en las que un organismo adquiere información genética nueva:

1. **Evolución clonal**: Evolución vertical donde hay una transmisión de la información genética de un progenitor a su descendiente directamente. Así, cualquier información genética nueva será un resultado de una mutación. Esto se representa bien a través de un árbol filogenético. Los nuevos linajes desde un ancestro común se llaman *clados*.
2. **Evolución reticular u horizonal**: Distintos clados se fusionan, formando un nuevo linaje híbrido. Esto ocurre en todos los distintos dominios, particularmente en virus (como la gripe -influenza- o el VIH). Los árboles filogenéticos no son capaces de recoger estos eventos, por los que hace falta una nueva estructura llamada *red reticular o red filogenética* en las cuales las ramas pueden dividirse o unirse arbitrariamente de manera horizontal o diagonal, y no sólo vertical. Esto resulta en una red con varios árboles y topologías fusionándose. El resultado de estas fusiones entre ramas son ciclos o lazos, detectables a través de la homología persistente.

Tradicionalmente, para poder detectar eventos reticulares, es necesario construir un árbol para cada gen en el genoma, y luego compararlos entre cada par de genes para encontrar los conflictos en la línea temporal. Estos conflictos vendrán representados a través de topologías distintas entre los árbol filogenéticos de dos genes distintos del mismo genoma para la misma temporalidad. Esto, evidentemente, es un trabajo tedioso e inviable para un organismo mínimamente complejos. En este punto es donde se puede emplear la homología persistente para detectar eventos de evolución horizontal, en base al genoma observado.

Tomando, por ejemplo, la Figura 3.2, vemos que no existen ciclos y es contráctil a un punto. Esto se representa directamente en clases de homología únicamente de dimensión 0, es decir, para una evolución vertical representable en un árbol filogenético, no existirán clases de homología distintas a la trivial para dimensiones ≥ 1. Formalizamos esto con un teorema.

Teorema 3.3. *Sea (M, ∂_M) cualquier espacio métrico finito isométrico a un árbol (es decir, que satisfaga la condición de los 4 puntos 3.2), y sea $\varepsilon \geq 0$. Entonces, el complejo de Vietoris-Rips $VR_\varepsilon(M, \partial_M)$ es la unión disjunta de complejos acíclicos, es decir, que no existen ciclos o lazos en él. En particular, $H_i(VR_\varepsilon(M, \partial_M)) = \{0\}$, para $i \geq 1$.*

Esto quiere decir que si existe homología de una dimensión mayor a cero (ya que los árboles filogenéticos son contráctiles), indican que el espacio métrico **no** satisface propiedades métricas similares a un árbol. Identificando qué genomas son los generadores de estas clases de homología, podemos encontrar dónde hubo procesos de evolución horizontal, que no podrían pertenecer a un árbol.

Como describimos en el capítulo anterior, un código de barras o "barcode" representa ciclos k-dimensionales independientes que generan clases de homología distintas a las triviales. Cuando hay clases no triviales para $k \geq 1$, hay una desviación de la métrica de árbol, y queremos definir una cantidad que refleje cuánta desviación hay. Para ello consideramos la distribución B_k de la longitud de las barras de ciclos k-dimensionales para $k > 0$. En particular, una medida natural de la desviación sobre la métrica de árbol es algún recuento de las barras dimensión 1. Definimos la *obstrucción topológica a la filogenia*, o *TOP* por sus siglas en inglés, como la norma L^∞ sobre las barras, es decir, su mayor longitud. Como se establece en [2], una filtración con TOP distinto de cero implica que un espacio métrico finito **no** isométrico a un árbol. Otra posible medida es la norma L^1 (la suma de la longitud de las barras). En simulaciones con una tasa de evolución horizontal h, se descubrió que de todas las normas L^p, las que mejor se corresponden con la tasa h es la norma L^1 o, en menor medida, la norma L^0 que representa el número de barras.
Si normalizamos estas normas, L^0 y L^1 por el tiempo, definimos la *tasa de ciclo irreducible* como esta normalización. Esta, aporta un umbral inferior estimado para la tasa de reagrupación.

Con todo esto, sumado a algunos resultados, se puede establecer una conexión entre los invariantes de la topología algebraica y eventos de intercambio genético (Fig. 3.3).

Homología persistente	**Evolución vírica**
Valor de filtración ε.	Escala de la distancia genética (evolutiva).
Número de Betti en dimensión 0 para valor de filtración ε.	Número de clusters a la escala ε.
Número de Betti de dimensión 1.	Número de eventos irreducibles de recombinación/reagrupación.
Generadores de la homología 1-dimensional.	Eventos de recombinación/reagrupación.
Generadores de la homología 2-dimensional.	Intercambio genético horizontal complejo.
Número de generadores en dimensiones más altas en un segmento de tiempo.	Umbral mínimo para la tasa de recombinación/reagrupación.
Homología con dimensión distinta a cero(obstrucción topológica a la filogenia).	Representación a través de una red reticular.

Figura 3.3: Correspondencia entre conceptos de la homología persistente y su contrapartida evolutiva. Adaptado de [2].

Esto implica que podemos inferir o descubrir el proceso biológico ocurrido a través de la homología persistente de manera concreta. Exploramos ahora esta relación más en detalle, a través de ejemplos en virus.

Observación 3.4. En algunos casos, esta relación entre evolución e invariantes topológicos se puede explicitar aun más, a través de los *árboles de agallas* ("galled tree" en inglés, de donde

"gall" hacer referencia a deformaciones de un árbol debido al efecto de un patógeno, llamadas "agallas" en castellano). Esta construcción se puede encontrar en [10].

3.2. Evolución de la Influenza A

Para esta sección se ha seguido los dos principales trabajos sobre los que se basa todo este documento, [12] y [2].

El virus de la influenza, conocido como la gripe común, es un virus de la familia de los Orthomyxoviridae cuyo genoma se compone de una única cadena de ARN (monocatenario), y puede infectar a varias especies. Su mayor diversidad genética y principal población son las aves, pero también infectan humanos, equinos, o incluso focas. Su principal hábitat son los Anseriformes (patos, ocas, gansos...). El Influenzavirus A es muy relevante en los países asiáticos por la particular virulencia de sus cepas, por lo que siempre está siendo controlado y estudiado.

La manera de clasificar el virus de la influenzia ha sido siempre a través de dos de sus componentes que se encuentran en la envoltura proteica del virus: La hemaglutinina (HA), que se clasifica de H1 a H16, y la neuraminidasa (NA), que toma valores entre N1 y N9, aunque actualmente se sigan explorando y ampliando estos rangos.

Aproximadamente medio millón de muertes anuales se asocian al virus de la Influenza, donde en humanos suele infectar el tracto respiratorio superior, y dar síntomas de fiebre, dolor de garganta, mucosidad... son además infecciones típicas de invierno, ya que la tasa de infección mejora a baja humedad y temperatura. Aunque en mamíferos suela cursar síntomas, rara vez lo hace en aves, a pesar de ser portadores y transmisores. Cuando, por mutación, sí es patogénico, causa un fallo multi-orgánico y mata al portador. Esta mutación de virus en aves se controlan con frecuencia para evitar su paso a contagio humano.

El paso del virus de ave a humano, muchas veces, se realiza a través del cerdo. Esto es debido a que para la infección, la hemaglutinina del virus debe reconocer residuos de monosacáridos en las células epiteliales, los cuales son distintos en aves y humanos, pero el cerdo tiene ambas variedades, lo que lo haría efectivo como puente entre ambas especies. Aun así, la transmisión de ave a humano no siempre ha tenido intermediarios, si no que por mutación ha ocurrido directamente.

Además, los monosacáridos que tienen las aves, sí se encuentran en el humano, pero en baja concentración y en la parte baja del tracto respiratorio. Esto implica que, de contagio directo, el virus causaría neumonía y síntomas mucho más agresivos, lo cual aumenta la tasa de mortalidad. Un ejemplo de esto es el virus H5N1, cuya tasa de mortalidad fue del 60 % en 2019, pero de momento la transmisión entre humanos se mantuvo baja. Este tipo de virus son muy peligrosos

de convertirse en gripe estacional.

El genoma del virus de Influenza tiene 13.000 bases, y se compone de 8 segmentos (análogos a nuestros cromosomas) de una única cadena de ARN, donde cada segmento codifica uno, o dos genes víricos. El ARN es antisentido, por lo que las hebras de sentido positivo (es decir, aquellas que se traducen directamente como una proteína) toman como plantilla el ARN original, codifican el complementario (sentido positivo), y se debe volver a codificar sobre este para obtener una réplica del ARN original, y así crear viriones (partículas infecciosas que transmiten el virus fuera de la célula huésped).

Así, el Influenzavirus evoluciona acumulando mutraciones a una tasa muy alta, alrededor de unos 10^{-3} nucleótidos anuales, y cada año se realizan nuevas vacunas para poder seguir el ritmo de las mutaciones. La Organización Mundial de la Salud (OMS) actualiza cada año la composición de la vacuna para que siga reflejando las nuevas cepas, y por ende existen 10.000 genomas en la base de datos sobre el virus de la Influenza.

Cada pandemia comienza con solo huéspedes animales, y que luego incorpora genes que la hacen compatible con el contagio humano a través de la reagrupación. Estas mutaciones y reagrupaciones cambian las propiedades antígenas de las cepas, de manera que las vacunas útiles para su anterior genoma resultan inútiles. El mayor ejemplo de una pandemia de Influenza es, por supuesto, la Gripe Española de 1918, con la cepa H1N1. Fue también esta misma cepa, H1N1, la causante de la pandemia mundial del 2009, presente en España con el nombre de “gripe A”.

Esta reagrupación no es completamente aleatoria, si no que se observan ciertos patrones o preferencias por qué segmento es más probable que sufra un evento de evolución horizontal. Aun así, no se conoce exactamente el mecanismo por el que esta decisión es tomada, lo cual sería de gran interés para poder limitar el número de cepas posibles, y con suerte, prepararse para ellas. Esta evolución a través de la reagrupación se puede estudiar usando la homología persistente.

La homología persistente relacionada a los virus surge, realmente, en [2]. El objetivo del mismo era presentar la homología persistente coma técnica novedosa y eficiente para poder identificar patrones en el genoma de los virus, que representen tanto una evolución horizontal como una vertical. Además, para una evolución horizontal compleja, entre varias cepas progenitoras, se pueden descubrir patrones reticulares como la segregación de los segmentos durante la reagrupación.

Empezamos analizando una de las dos proteínas mencionadas de la envoltura, que le dan parte del nombre a la cepa: la hemaglutinina, que tiene representa de los ocho segmentos del ARN. Para poder emplear la homología persistente, computamos las distancias por pares entre 2 cepas distintas y generamos un espacio métrico finito, con puntos que representan las distintas secuencias. La mayoría de las barras de homología serán de dimensión cero, con alguna excepción

de dimensión 1.

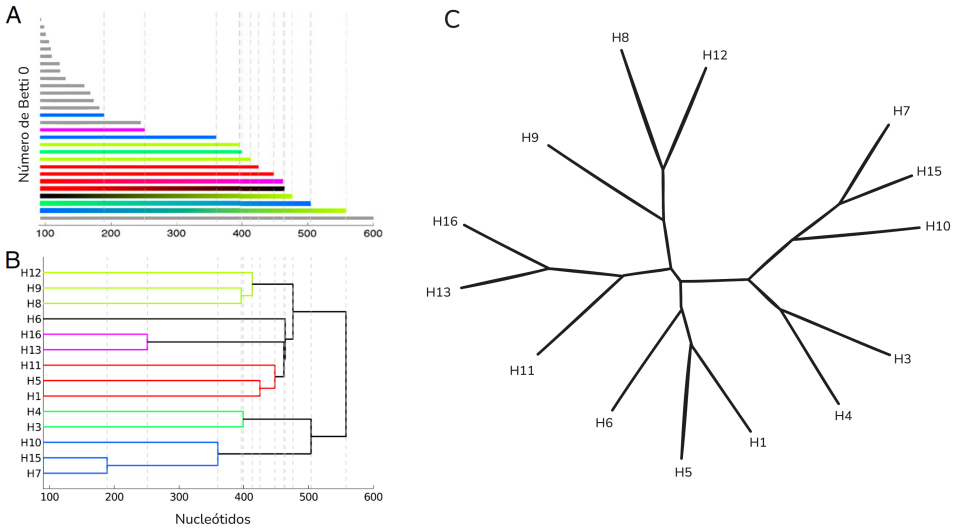


Figura 3.4: Cuando la dimensión del grupo de homología más alta es 0, esta es representada con árboles. En este caso, se toma sólo la proteína de la hemaglutinina de la Influenza A, y vemos en **A** que sólo barras para número de Betti 0. La evolución para cada subtipo de hemaglutinina se puede ver en **B**, y la relación entre los subtipos en **C**. Fuente: [2].

Al observar los generadores de las clases de dimensión cero, se puede reconstruir algo similar a un árbol filogenético a través de técnicas clásicas, no relacionadas con la homología. Si estudiamos los segmentos aislados, no se encuentra información relevante en la dimensión 1 o más de homología (Fig. 3.4). En cambio, al estudiar la homología persistente para varios genes a la vez, sí empiezan a aparecer clases de homología de mayor dimensión que cero, que harían referencia a reagrupaciones, y estos ciclos de dimensión 1 o más en la homología son resultado de procesos que no cumplen el árbol (Fig. 3.5): recombinaciones, reagrupaciones u homoplasias (evolución paralela e independiente que resulta en el mismo rasgo). Esto nos permite estimar, por ejemplo, cuán frecuente es que combinaciones distintas de los 8 segmentos sufran cosegregación (que se hereden 2 genes juntos por estar muy cerca físicamente). Aplicado a la Influenza A, podemos estimar que es muy poco frecuente que el complejo polimerasa PA, PB1, PB2 y NP no se herede de manera conjunta, lo cual es coherente con el hecho de que trabajen juntas en el virus, por lo que iría en contra de la selección natural que sufriesen alteraciones independientes y dejasen de funcionar de manera grupal.

Otra información que se puede obtener a través de la homología proviene de cuándo nace y muere una clase, por ejemplo, que respondería a la distancia genética de los virus progenitores. El tamaño de las barras de la clase de homología no trivial indica también el tipo de reagrupación sufrida. Así, las barras cortas revelan una mezcla de virus con una relación genética cercana,

seguramente del mismo subtipo (dos cepas de H5N1, por ejemplo), y las barras más largas son resultado de mezclar material genético de virus más distantes, como sería el H5N1 y el H7N2.

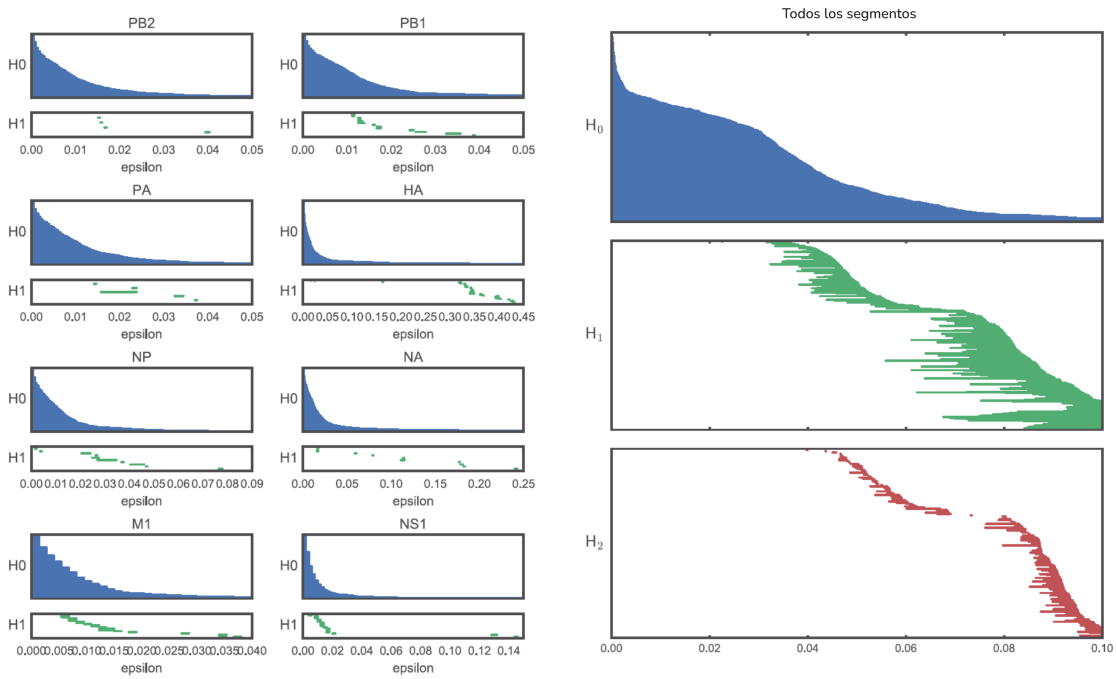


Figura 3.5: Al examinar los segmentos de la Influenza por separado (izq), la mayoría de la homología es exclusivamente de dimensión 0. En cambio, al concatenar segmentos y estudiarlos en conjunto (dcha), aparece gran presencia de homología de dimensión 1 y 2. Esto es debido a la evolución horizontal, con eventos de agrupamiento, que se ven reflejados en los códigos de barras. Fuente: [12].

3.3. Evolución del VIH

El Virus de la Inmunodeficiencia Humana, o VIH, es una de las enfermedades infecciosas más famosas desde su aparición en los años 80.

Se estima que alrededor más de medio millón de personas sigue muriendo al año de sida, y que casi 40 millones de personas son portadoras del virus ([15], p30-31). La forma de ataque del VIH es destruyendo células del sistema inmunitario, las células T (CD4+). Estas forman gran parte de la respuesta inmunitaria frente a otra infección, coordinando al organismo, creando anticuerpos, activando otras células que puedan atacar a la infección... Por ello, cuando las células T CD4+ mueren, la respuesta inmunitaria que el cuerpo puede dar es mucho más deficiente, y una persona seropositiva puede morir por un patógeno que en una persona sana, no tendría ni síntomas. Esta

bajada de células T tarda años en ser lo suficiente relevante como para ser detectada a pesar de seguir transmitiéndose, y por este motivo, el tiempo medio de muerte para una persona es de 9 a 11 años desde.

El Virus de Inmunodeficiencia Humana es de la familia de los *retrovirus*, aquellos que realizan un proceso "inverso" que el de las células humanas: Mientras que en una célula normal se crean las cadenas de ARN usando el ADN como patrón, en los retrovirus se crea una cadena de ADN a partir de las cadenas de ARN presentes.
Así, igual que la Influenza A, el genoma del VIH consiste una cadena de ARN (monocatenario), pero de sentido positivo. De hecho, cada virus tiene dos copias idéntidas de ARN, cada una de las cuales cuenta con 10.000 bases. Hay 3 genes principales en todos los retrovirus:

- El gen *gag*, que codifica las proteinas que conforman el cápside, que es una capa que protege el genoma.
- El gen *pol* codifica las encimas necesarias para la replicación y transcripción inversa, para integrar el ADN vírico en el huésped, y para escindir poliproteínas víricas y activarlas.

  El gen *env*, que codifica las glicoproteínas que actúan sobre las células T, uniéndose a ellas, y permiten la infección.

Hay otras 6 proteínas en el genoma el VIH, pero menos relevantes.

La primera vez que fue identificado fue en 1981, y fue estudiado formalmente en 1983 por Barré-Sinnousi y Montagnier, por lo que recibieron el Nobel en 2008. Se pudo identificar por la presencia en los primeros pacientes de muerte por infecciones oportunistas (sobretodo, neumonías, como *Pneumocystis jirovecii*) y la aparición de un tipo de cáncer de piel, sarcoma de Kaposi, el cual es causado en sí mismo por una infección. Hay dos tipos de retrovirus que causan la enfermedad de SIDA, el VIH-1 y el VIH-2. El segundo se identificó en África, en el 1986. Es precisamente en África central donde se encuentra mayor diversidad genética del virus, lo que sugiere que el virus apareció originalmente ahí. De hecho, se han realizado estudios que asocian la secuencia del VIH a un ancestro común de las cepas tipo M (M por mayoritarias, ya que representan el 90 % del VIH) de 1920, en República Democrática del Congo (Faria et al., 2014).

El VIH es un virus particularmente mutágeno, con gran diversidad y que sufre recombinaciones muy a menudo, con una tasa de mutación del orden de $10^{-}3$, aunque muchas de las mutaciones causan la muerte del virus. La mayor causa de la mutación es en la transcripción inversa (el paso de ARN a ADN). A diferencia de la Influenza, el genoma del VIH no está divido en segmentos diferenciados, por lo cual no se dan eventos de reagrupamiento como los estudiados en la anterior sección, y el mutágeno principal en el VIH será la recombinación.

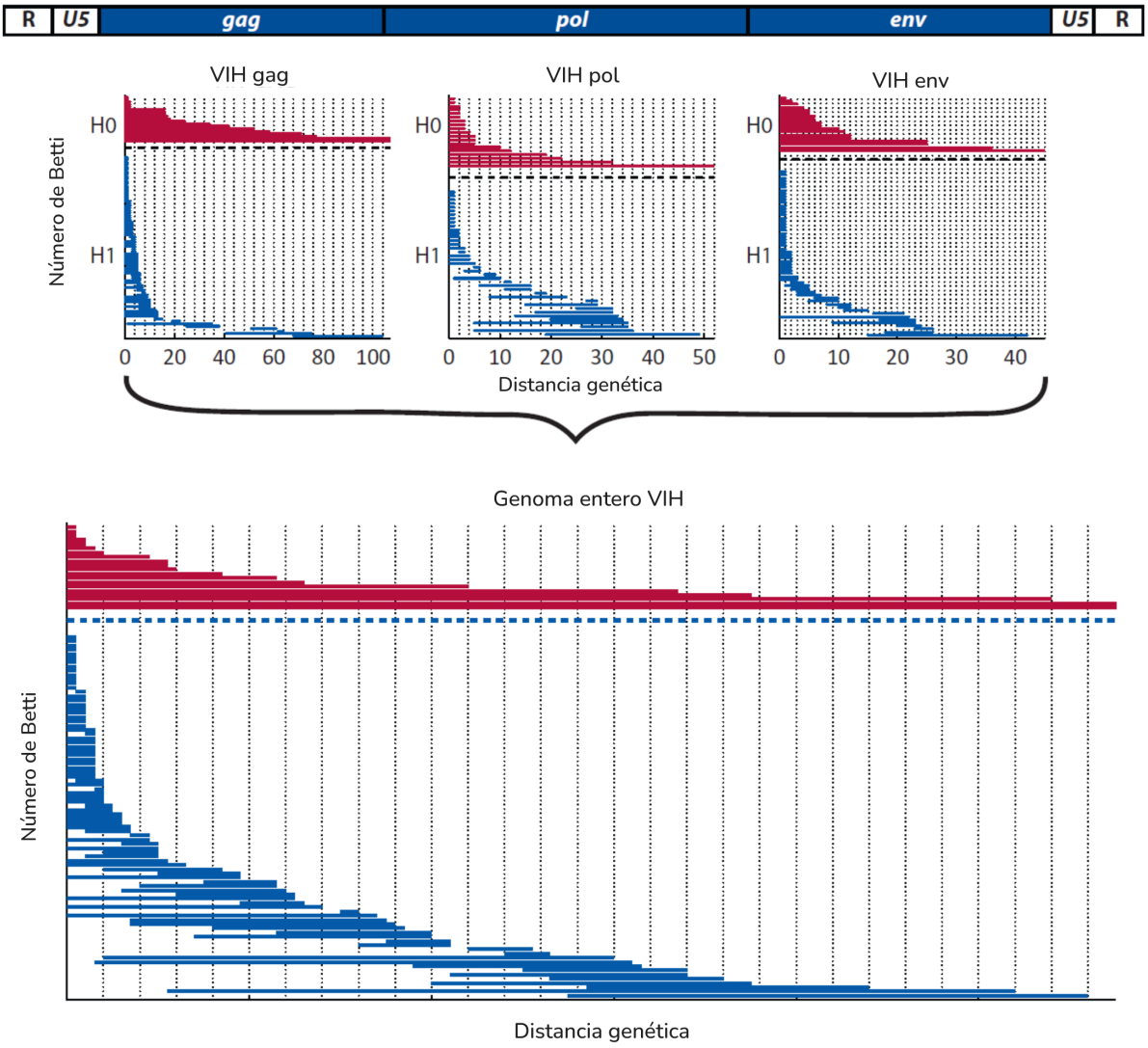


Figura 3.6: Ejemplo de homología persistente aplicado al genoma del VIH. A diferencia de lo que pasaba con la Influenza, sí existe clase de homología de dimensión 1 para cada gen separado, pero al concatenarlos y estudiarlos juntos (abajo), aparecen muchas más barras. Fuente: [2].

El VIH tiene, al menos, 2 copias del genoma en una cadena de ARN. A los viriones (partículas víricas, o el propio virus cuando aun no tiene huésped) pueden llevar, en vez de 2 copias del mismo virus, 1 copia de ARN de 2 virus de VIH distintos cuando estos co-infectan una célula. Al realizar la transcripción inversa, y generar el ADN a partir del ARN, la proteína que realiza esta acción (polimerasa) puede saltar de una cadena a la otra, mezclando ambos genomas distintos en un nuevo virus. Este proceso es algo común, y la recombinación puede convertirse en el genoma dominante dentro de un huésped.

Las formas recombinantes, CRFs (por sus siglas en inglés, “Common Recombinant Forms”)

son los resultados más comunes de la recombinzación entre los subtipos de virus. No son fáciles de clasificar por todas las posibilidades que ofrece la recombinación, sin que haya tampoco puntos definidos en el genoma donde se pasa de una hebra a la otra (y de hecho, esos puntos de alternancia entre hebras pueden ser varios). Es imposible, por ello, dibujar un árbol evolutivo para cada gen. Al aplicar la homología persistente al VIH aparecen naturalmente muchas clases de dimensiones distintas a la trivial, que indican un histórico de eventos recombinatorios. Al concatenar varios genes y no sólo estudiarlos individualmente, se pueden observar grandes eventos recombinatorios entre varias cepas de subtipo A (dentro del subtipo mayoritario, los subtipos del VIH-1 van de la A hasta la L, según su genoma) (Fig. 3.6).

3.3.1. Recombinación viral para el VIH de larga duración

Cuando una persona se infecta con VIH, los primeros síntomas serán relacionados con su sistema inmunitario, pero no son los únicos síntomas posibles. Para un paciente con una infección de VIH durante mucho tiempo, puede llegar a sufrir demencia asociada al VIH (HAD por sus siglas en inglés, *HIV-associated dementia*), el peor de los trastornos neurocognitivos relacionados con el VIH, que se asocia a una alta exposición cerebral al virus en un momento dado de la infección. Aunque es evitable de recibir tratamiento al principio de la infección, una vez aparecen síntomas no hay manera de revertirlos, por lo que también es interesante conocer la naturaleza de la población viral del cerebro.

Al muestrear el líquido cefalorraquídeo, presente en la médula espinal y el cerebro, y comparar su población vírica con la de la sangre en otras zonas del cuerpo, suelen ser distintas genéticamente, lo que puede indicar que una replicación del virus en el cerebro puede causar directamente el HAD. Además es más frecuente encontrar recombinación en aquellas poblaciones de virus en el cerebro de pacientes con HAD, frente a otros pacientes seropositivos, lo cual vuelve a indicar la replicación vírica descontrolada como causa de la demencia.

Explicaremos ahora cómo usar las herramientas de la homología persistente para caracterizar, y estudiar, la recombinación viral dentro de un huésped del VIH, particularmente en aquellos pacientes que han sido seropositivos un largo periodo. Para ello, será interesante entender cómo el virus recombinado se expande por los distintos tejidos, lo cual se realiza comparando las secuencias genéticas del sistema nervioso central frente a aquellas en otros tejidos. Usaremos la **homología zigzag** para formalizar estas ideas, y así comparar eventos entre distintas poblaciones. La muestra serán 11 pacientes fallecidos por SIDA, de los cuales 5 tienen la secuenciación de varios tejidos y el Sistema Nervioso Central, y otro sin el sistema nervioso central pero sí con distintos tejidos. La referencia a esta muestra se puede encontrar en [12].

Pacientes	Secuencias solo del SNC (secuencias únicas)	Secuencia fuera del SNC (secuencias únicas)
AZ	35 (33)	52(48)
DY	107 (99)	59 (54)
BW	103 (99)	18(18)
CX	162 (152)	47 (43)
GA	75 (73)	57 (55)

Figura 3.7: Resumen de la información de cada paciente. La primera columna identifica al paciente, y en la segunda y tercera columnas están las secuencias obtenidas del sistema nervioso central. Se puede encontrar estos datos en el GenBank (referencia [12]).

Para la muestra de secuencias de dos subpoblaciones relacionadas, podemos dividir los eventos en 4 clases:

1. Un evento que ocurre en la primera población, pero no en la segunda
2. Un evento que ocurre en la segunda población, pero no en la primera
3. Cualquier evento que se pueda detectar en cualquiera de las dos poblaciones, pero que pase exclusivamente en solo una de ellas (típico de poblaciones muy relacionadas entre sí).
4. Eventos que acontecen en ambas poblaciones y que sólo se pueden detectar uniendo secuencias de ambas, y no estudiando cada población individualmente. Esta clase representa representa el caso de flujo genético entre dos poblaciones genéticamente distintas.

En la Fig. 3.8 se puede ver de color verde una población, y azul la otra población (la diferencia puede ser, por ejemplo, una región geográfica distinta de los pacientes o que las muestras sean de un tejido distinto).

Viendo ya la primera representación, **B**, se observa un lazo, que al añadirle más datos (el "zig" de las flechas r_i en la Definición 2.15), aparece un evento de recombinación en **C** el cual proviene de una recombinación individual que ocurre en ambas poblaciones, y se puede calcular computacionalmente. Al proseguir con un "zag" (las flechas l_i), se obtiene la población 2 aislada, donde también hubo un evento de recombinación. Así, tanto en **B** como en **D** se puede ver la misma clase de homología para ambas poblaciones, lo cual responde a un evento de clase 3, en base a lo descrito anteriormente.

En el árbol inferior, **E**, tenemos un evento recombinatorio entre las dos poblaciones que las une. Cuando estudiamos la homología por separado, solo observamos una clase de homología

distinta a la trivial en **G**, representando ambas poblaciones a la vez, lo cual sería un evento de recombinación de clase 4.

Es decir, usando la homología persistente se pueden identificar supuestos eventos recombinatorios entre poblaciones. Para los datos en los que están disponibles tanto el genoma del SNC como de fuera del mismo (que constituirían las 2 poblaciones descritas), se utilizó la persistencia zigzag para clasificar los eventos de recombinación. En este caso, tenemos:

Pacientes	**Situación del HAD**	**Grado de neuropatía**	**Nº eventos, SNC**	**Nº eventos, sitios cruzados**	**Nº eventos, No-SNC**
AZ	Ninguno	3	0	1	2
DY	Agudo	1	2	5	1
BW	Progresivo	2	3	0	0
CX	Progresivo	5	8	0	1
GA	Progresivo	5	5	7	8

Cuadro 3.1: Resumen por paciente y supuestos eventos recombinatorios detectados a través de la homología persistente.

Nos centramos ahora en los pacientes que tienen HAD progresivo, que también tienen la neuropatía (enfermedad del sistema nervioso) más severa, CX y GA. Ambos pacientes tienen el mayor número de eventos recombinatorios en el SNC, lo cual puede sugerir que existe una relación entre recombinaciones frecuentes de los virus y la demencia asociada al VIH. Al margen de esta similitud en particular, el resto de sus datos son muy distintos: Mientras que CX no

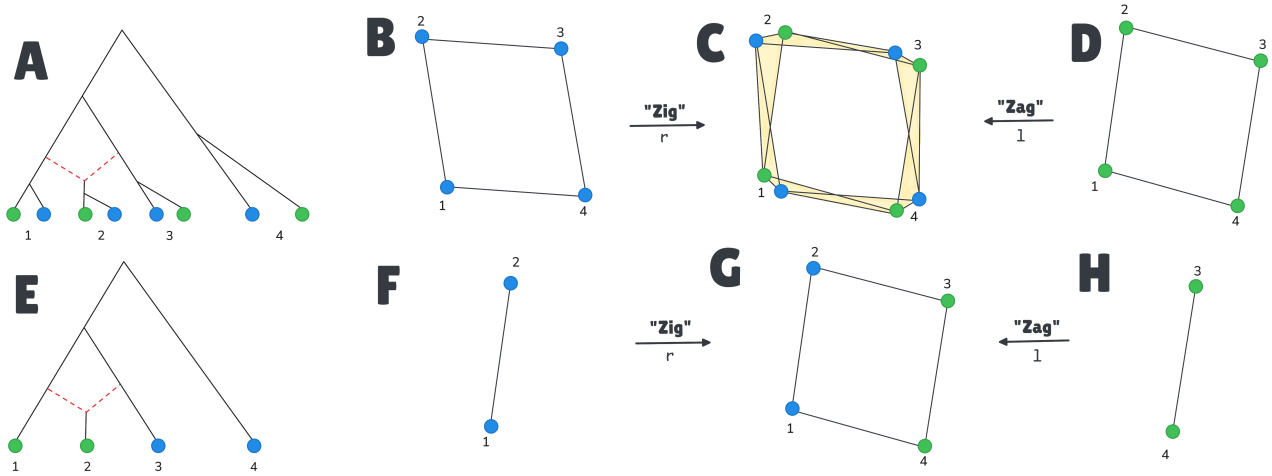


Figura 3.8: Esquema donde se interpreta los eventos de recombinación entre poblaciones, a través de la persistencia zigzag.

tiene eventos cruzados, y casi todos los eventos recombinatorios ocurren en el SNC, el paciente GA tiene entrelazados los eventos recombinatorios entre secuencias de dentro, y fuera, del SNC. Mostramos ahora una figura que representa la red filogenética de ambos pacientes, con sus diferentes estructuras (Fig.3.9).

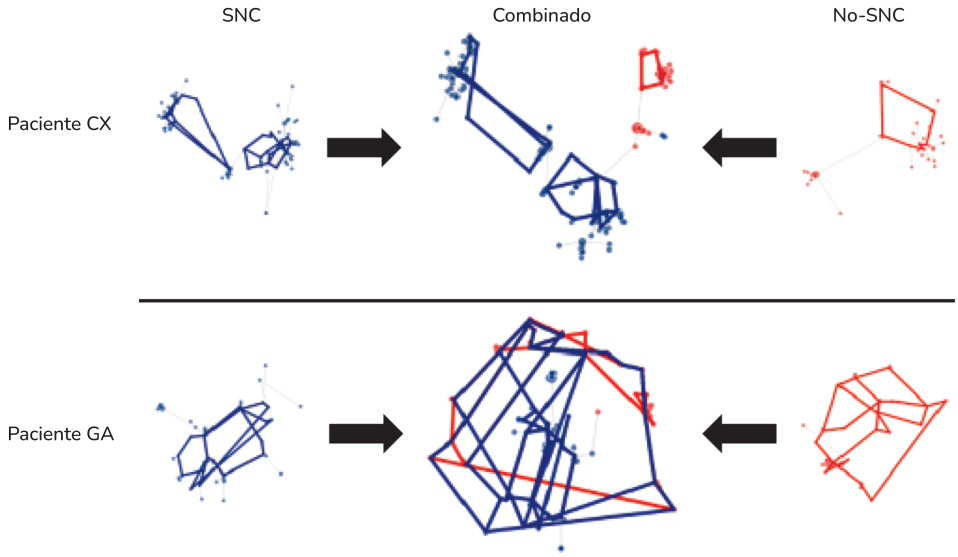


Figura 3.9: Red filogenética del VIH-1 obtenida de los pacientes CX y GA, donde cada nodo representa una secuencia de genoma distinta. Los nodos azules provienen de las muestras del SNC, y las rojas del resto del cuerpo. La posición viene determinada por la distancia de Hamming (número de letras distintas entre secuencias). Las aristas azules y rojas son los generadores de los ciclos identificados a través de la homología persistente. Mientras que los ciclos azules denotan secuencias relacionadas con el SNC, las rojas son de aquellas secuencias enteramente fuera del SNC.

Tomando un modelo de coalescencia para calcular un estimador de la tasa de recombinación, descrito en [12]§5.7, obtenemos que los pacientes con una mayor tasa de recombinación en el SNC son, otra vez, los de mayor severidad neuropática, CX y GA, lo cual indica que no es un resultado sólo sobre la muestra (Tabla 3.2).

Cabe notar que la suma de la estimación para las poblaciones de SNC y no-SNC no tiene por qué ser el mismo resultado que la estimación de todas las secuencias sumadas. Esto viene, como mencionamos antes, por la entremezcla de ambas poblaciones al sufrir eventos recombinatorios y un ancestro común. Precisamente por esto es destacable el caso del paciente DY, cuya estimación de ambas poblaciones es mayor que la suma de estimaciones. Este paciente era, también, el de mayor tasa de recombinación en sitios cruzados, y está recogido que sufrió otra infección vírica reciente que explicaría esta anomalía.

Pacientes	$\hat{\rho}_{PH}$, **secuencia SNC**	$\hat{\rho}_{PH}$, **secuencia no-SNC**	**Suma de ambos estimadores**	$\hat{\rho}_{PH}$ **de todas las secuencias**
AZ	0	6.7	6.7	4.4
DY	4.0	3.0	6.9	11.0
BW	5.9	0	5.9	5.3
CX	12.7	3.7	16.3	12.5
GA	12.5	27.4	39.9	35.2

Cuadro 3.2: Estimación sobre la tasa de recombinación en base a los datos obtenidos. Fuente: [12]

3.4. Conclusiones

El método de la homología persistente, usado en este trabajo sobre la reagrupación del virus de la Influenza A y la recombinación en el VIH, es aplicable a otros virus. Particularmente, la familia a la que pertenece la Hepatitis C, que tiene algunas cepas recombinantes, o el virus del denge y la zika. Estos virus pertenecen al género *Flaviviridae*, sobre el cual se pueden encontrar resultados usando la homología persistente. Eso sí, el mayor resultado se obtiene con las dos familias vistas aquí, debido a su gran tasa de eventos de evolución horizontal. Por ende, tanto para prevenir y conocer las posibles futuras cepas de la gripe, sobre todo aquellas más peligrosas, y conocer y controlar los mecanismos que causan la demencia por el VIH, es importante seguir investigando y trabajando con métodos matemáticos, y apostar por la ciencia.

Bibliografía

[1] Carlsson, G. & De Silva, V. (2018).*Zigzag persistence*. arXiv.org. https://arxiv.org/pdf/0812.0197.

[2] Chan, J. M., Carlsson, G., & Rabadan, R. (2013). Topology of viral evolution. *Proceedings Of The National Academy Of Sciences, 110*(46), 18566-18571. https://doi.org/10.1073/pnas.1313480110.

[3] Cohn, P. M. (1974). *Algebra* (Vol. 1). John Wiley & Sons.

[4] Edelsbrunner, H., & Harer, J. L. (2022). *Computational topology: An Introduction*. American Mathematical Society.

[5] Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., Chen, J., Jie, Z., Qiu, H., Xu, K., Xu, X., Lu, H., Zhu, W., Gao, Z., Xiang, N., Shen, Y., He, Z., Gu, Y., Zhang, Z., . . . Shu, Y. (2013). Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus. *New England Journal Of Medicine, 368*(20), 1888-1897. https://doi.org/10.1056/nejmoa1304459

[6] Ghrist, R. (2007). Barcodes: The persistent topology of data. *Bulletin Of The American Mathematical Society, 45(01)*, 61-76. https://doi.org/10.1090/s0273-0979-07-01191-3.

[7] Hatcher, A. (2002). *Algebraic topology*. Cambridge University Press.

[8] de Jong, A.J. (2018). *The Stacks Project. Section 12.13 (010V): Complexes*. https://stacks.math.columbia.edu/tag/010V.

[9] Lang, S. (2006). *Undergraduate Algebra*. Springer Science & Business Media.

[10] Lesnick, M., Rabadán, R., & Rosenbloom, D. I. S. (2018, 3 abril). *Quantifying Genetic Innovation: Mathematical Foundations for the Topological Study of Reticulate Evolution*. arXiv.org. https://arxiv.org/abs/1804.01398.

[11] Munkres, J. R. (1984). *Elements of algebraic topology*. Addison-Wesley.

[12] Rabadan, R., & Blumberg, A. J. (2019). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press.

[13] Riehl, E. (2014). *Category Theory in Context*. Dover Publications.

[14] Zomorodian, A. J. (2005). *Topology for Computing*. Cambridge: Cambridge University Press.

[15] UNAIDS.(s. f.). *2024 global AIDS report — The Urgency of Now: AIDS at a Crossroads.* `https://www.unaids.org/en/resources/documents/2024/global-aids-update-2024`.