Modelos de homología persistente en filogenética

T E S I S

Que para obtener el título de

Licenciado en Matemáticas

P R E S E N T A:

José María Ibarra Rodríguez

Director de Tesis:

Dr. Víctor Manuel Pérez Abreu Carrión

Universidad de Guanajuato
guanajuato, gto
Enero 2016

UNIVERSIDAD DE GUANAJUATO

En la ciudad de Guanajuato, Gto. a las 16:00 (dieciseis horas y cero minutos) del día 22 de enero del año 2016, se reunieron en el (la) Salón de Actos del Departamento de Matemáticas de la División de Ciencias Naturales y Exactas del campus Guanajuato de la Universidad de Guanajuato, los señores Doctor Miguel Nakamura Savoy, Doctor Jose Perea, Doctor Luis José Delaye Arredondo designados para verificar el examen de titulación del señor **JOSÉ MARÍA IBARRA RODRÍGUEZ**, concedido por acuerdo de la Rectoría General a efecto de obtener el título de **LICENCIADO EN MATEMÁTICAS**.

Instalado el jurado bajo la Presidencia del señor Doctor Miguel Nakamura Savoy y fungiendo como Secretario el señor Doctor Jose Perea, dio principio el examen e interrogaron sucesivamente los tres sinodales sobre diversas materias y aspectos del trabajo de titulación y, terminado el interrogatorio, se procedió a verificar la votación secreta, habiendo resultado el señor **JOSÉ MARÍA IBARRA RODRÍGUEZ**.

Aprobado por unanimidad

Para constancia se levanta la presente acta por triplicado que firman los señores sinodales.

NO ES VÁLIDO SI LLEVA RASPADURAS O ENMENDADURAS

Presidente

Dr. Miguel Nakamura Savoy

Secretario

Dr. Jose Perea

Vocal

Dr. Luis José Delaye Arredondo

UNIVERSIDAD DE GUANAJUATO
DIRECCIÓN DE ASUNTOS ACADÉMICOS

UNIVERSIDAD DE GUANAJUATO

Dirección de Asuntos Académicos

El suscrito, ***Maestro Francisco Javier Pérez Arredondo, Director de Asuntos Académicos,*** *hace constar que la presente acta constituye formalmente la realización del examen de obtención* ***de Título de LICENCIADO EN MATEMÁTICAS,*** *en los términos del Artículo 68 del Estatuto Académico de la Universidad de Guanajuato.*

UNIVERSIDAD DE GUANAJUATO
DIRECCIÓN DE ASUNTOS ACADÉMICOS

Índice general

Agradecimientos

Gracias a mi familia, a mis padres, hermanos y abuelos, porque nunca dejaron de creer en mí. A ellos les dedico este trabajo. Gracias a mis amigos por compartir la aventura de vivir en Guanajuato estos cinco años en los que aprendí y disfruté cada momento.

Gracias a la comunidad CIMAT-DEMAT, por construir un ambiente fértil para disfrutar las matemáticas. Gracias por todos los que, mediante clases, seminarios y discusiones, me acompañaron en este proceso. Especialmente gracias a Fermín Reveles por su disposición para escuchar y las sesiones de trabajo codo a codo. Así mismo, gracias a mis sinodales: Dr. Miguel Nakamura Savoy, CIMAT, Dr. Jose Perea, Michigan State University, y Dr. Luis José Delaye Arredondo, CINVESTAV Irapuato, por sus valiosos comentarios.

Gracias al Dr. Saugata Basu, Purdue University, por responder pacientemente a nuestras preguntas, por acoger nuestros comentarios e incorporar nuestras sugerencias a su trabajo.

Gracias especialmente a Víctor Pérez Abreu por brindarme la oportunidad de trabajar juntos. Gracias por la confianza y por cada consejo. Gracias por todos los retos a los que me aventaste y las aventuras académicas en las que nos embarcamos. Gracias por ser maestro y amigo.

Agradezco por último a las instituciones que me soportaron económicamente. Gracias al programa de becas de excelencia de CIMAT que me apoyó durante el periodo completo de licenciatura: sin la beca no hubiera podido completar mis estudios. Además, esta tesis fue posible de realizar gracias al apoyo, mediante ayudantías de investigación, del SNI y del proyecto de investigación de CIMAT en Análisis topológico de datos.

Introducción

La presente tesis de licenciatura aborda el estudio de modelos en filogenética basados en el análisis topológico de datos.

En biología, la reciente abundancia de datos genómicos ha motivado que aparezcan preguntas que requieren de una creciente interacción con distintas ramas de la matemáticas para ser estudiadas. Es en este contexto que en los últimos tres años, ante la necesidad de describir eventos evolutivos complejos, han aparecido esfuerzos para utilizar ideas provenientes de análisis topológico de datos en filogenética. El objetivo principal de nuestra tesis es presentar este nuevo marco teórico exponiendo a detalle dos de los trabajos pioneros en el área; presentar los elementos de topología y las nociones de filogenética necesarios para abordar el tema y, finalmente, identificar posibles direcciones hacia las cuales podrían extenderse estos modelos en el futuro.

Poder describir el proceso evolutivo de un conjunto de organismos es un problema importante dentro de la biología y está presente desde que Charles Darwin introdujo la teoría de la evolución. Dada una muestra de organismos en tiempo presente, llamamos *filogenia* de la muestra al conjunto de individuos que son parte de su historia evolutiva junto con las relaciones hereditarias que existen entre ellos.

Por la naturaleza misma del problema, resulta imposible determinar la verdadera filogenia de una muestra. No obstante, han surgido distintas estrategias que intentan describir de la manera más completa posible dichos procesos. A estas estrategias se les llama métodos de *reconstrucción* filogenética. La rama de la biología evolutiva que se encarga de reconstruir la filogenia de un conjunto de organismos se llama *filogenética*.

Las matemáticas siempre han desempeñado un rol central en la obtención de información sobre procesos evolutivos a partir de datos recabados de una muestra de organismos. Dicha información suele representarse en forma de árboles, o grafos más generales, que son estructuras matemáticas que surgen de manera natural en este contexto. Asociarle un árbol filogenético a una muestra constituye la primera y más sencilla descripción de su filogenia.

Además, para crear los métodos filogenéticos modernos, basados en comparar características genómicas de los organismos de una muestra, fue necesario desarrollar modelos matemáticos específicos que ahora forman parte de una rica teoría matemática. El artículo de Steel [60], en el número espe-

cial de Biología Matemática del *American Mathematical Monthly* de 2014, presenta algunos de los conceptos centrales de filogenética y la relevancia que han tenido para su estudio diversas ramas de las matemáticas como el álgebra, la combinatoria o la probabilidad.

En este trabajo abordaremos una de las interacciones más recientes entre filogenética y matemáticas, la cual tiene como fin caracterizar *eventos horizontales de evolución*. La base matemática de los modelos que aquí presentamos es la homología persistente, uno de los elementos más importantes del análisis topológico de datos.

El análisis topológico de datos es una rama reciente de la matemática y en los últimos años se ha constituido como una herramienta exitosa en la extracción de información a partir de datos complejos. Una característica clave de su éxito, en palabras de uno de los fundadores y promotores del área, reside en que "*los datos tienen forma, y esta forma importa*"[1] y el lenguaje adecuado para describir los patrones y la forma, dentro de matemáticas, es topología.

La homología persistente ofrece un amplia gama de aplicaciones que van desde la visión computacional [10] y el aprendizaje máquina [11] hasta la detección de un subtipo de cáncer [52], entre otras. Así mismo, los avances tecnológicos de las últimas décadas han facilitado la obtención de datos complejos en campos como la medicina y la biología motivando a su vez el desarrollo de nuevas aplicaciones en estas áreas; ver [1, 16, 42]. En la presente tesis nos enfocaremos en los modelos topológicos aplicados a la filogenética pues consideramos relevante que además de incluir la faceta usual de análisis de datos, también responden a la necesidad de modelar situaciones evolutivas complejas.

Entre las voces con mayor eco dentro de la teoría filogenética moderna están las que consideran que existen fuerzas evolutivas que poseen una incidencia más significativa de lo que anteriormente se pensaba. En particular, ahora se sabe que es necesario perfeccionar los modelos para incluir eventos de evolución horizontal [18]. Llamamos eventos de evolución horizontal a los sucesos evolutivos que requieren de la interacción entre dos organismos o especies cuando el proceso natural sólo requeriría de una. Ejemplos de tales eventos son la transferencia horizontal de genes en bacterias o el fenómeno de hibridación en plantas, entre otros. Debido a que la capacidad de representación de un árbol filogenético se ve rebasada por estos fenómenos, han surgido diferentes estrategias para modelarlos, entre ellas el enfoque que tomamos como objeto de estudio de esta tesis.

La primera vez que se utilizó la homología persistente para estudiar dichos procesos fue en el trabajo de Chan, Carlsson y Rabadan [14] publicado en la revista *Proceedings of the National Academy of Sciences* en 2013, así co-

[1] Gunnar Carlsson, 2015, http://www.ayasdi.com/blog/bigdata/why-topological-data-analysis-works/

mo en la tesis doctoral de Chan [13] de la Universidad de Columbia, Estados Unidos. El propósito central de este modelo es aprovechar las implicaciones evolutivas de las distancias genéticas para establecer un marco teórico que contemple, mediante homología persistente, los eventos verticales y horizontales en la evolución viral.

No obstante, como los autores mismos advierten, las ideas principales del modelo pueden ser fácilmente adaptables a contextos no virales. En esta tesis nos referiremos a este marco conceptual como el *modelo CCR* para evolución viral.

A partir del modelo CCR, surgió una variedad de esfuerzos que intentaron obtener el mayor provecho posible a la información obtenida mediante homología persistente [8, 23, 24, 25, 54]. Algunos plantean el modelo CCR en contextos biológicos ligeramente distintos, mientras que otros se centran en explorar diferentes variaciones del modelo matemático en sí. Entre estos trabajos resulta de especial interés la investigación de Parida *et al.* [54] publicada en 2015 en *Research in Computational Molecular Biology: 19th Annual International Conference, RECOMB 2015*. Este modelo introduce los *distintivos topológicos para mezclas de poblaciones* y se distingue de los otros trabajos porque adapta las ideas presentes en el modelo CCR a un contexto evolutivo bien definido, proporcionando además un sustento matemático adicional para tal postulación.

En esta tesis nos enfocamos en los aspectos de *modelación matemática* presentes en el trabajo pionero de Chan, Carlsson y Rabadan [14] y en la extensión del mismo al contexto de mezclas de poblaciones, realizado por Parida *et al* [54]. Debido a que ambos trabajos están publicados en medios dirigidos principalmente a lectores no matemáticos, decidimos hacer de este trabajo un compendio con una estructura y un lenguaje que le resulte familiar al lector que quiera acercarse al tema desde un punto de vista matemático y sin conocimientos previos en filogenética.

Creemos que una presentación en este sentido es relevante puesto que la construcción de estos modelos requirió de resultados matemáticos nuevos. Estos resultados, además de estar claramente inspirados en la interpretación que se hace de ellos al usarse en filogenética, resultan interesantes en sí mismos. Si bien los resultados ya están incluidos en los artículos originales [14] y [54], son mencionados discretamente en estos artículos debido a que el énfasis está puesto en la interpretación que se puede hacer de ellos. En el presente trabajo, además de desgranar las interpretaciones de los modelos, reconstruimos cuidadosamente y desarrollamos las pruebas de esos resultados matemáticos nuevos en los que se basan dichas interpretaciones.

Una de las contribuciones de la tesis es concretar las ideas detrás del modelo CCR en forma de postulado. Pensamos que esto permite comprender los indicadores que los autores proponen como los primeros intentos de obtener información filogenética de un código de barras de persistencia, pero que no están determinados de forma única por el teorema que sustenta el

modelo. Al mismo tiempo, postular el modelo de forma concreta permite entender las motivaciones de los trabajos posteriores que se basaron en él. De esta manera también se explica por qué han aparecido tantas variantes del modelo en tan poco tiempo.

Es importante resaltar que los objetivos originales de esta tesis estaban enfocados solamente en el artículo de Chan, Carlsson y Rabadan [14] pues al comenzar el proyecto, era el único trabajo publicado que empleaba análisis topológico de datos en filogenética del cual tuviéramos conocimiento. Conforme avanzó la tesis fueron apareciendo otras extensiones del modelo y el hecho de descubrirlas en tiempo real confirmó que la interacción entre homología persistente y filogenética es un área fértil y de mucha actualidad.

Como hemos dicho, consideramos que una de las variantes del modelo CCR más acertadas es precisamente el propuesto por Parida *et al.* [54]. El estudio detallado de este trabajo nos llevó a comunicarnos con uno de sus autores, el profesor Saugata Basu, de la Universidad de Purdue, Estados Unidos. Lo anterior concluyó en una reformulación del teorema principal de ese trabajo de tal manera que sustenta efectivamente la interpretación que se quiere hacer en el artículo y que nosotros presentamos en el Teorema 4.3.

En cuanto a herramientas filogenéticas se refiere, las aplicaciones que involucran homología persistente aún están en sus inicios. Sin embargo, en nuestra opinión es un área llena de oportunidades, retos y preguntas abiertas que requieren resultados matemáticos puntuales para situaciones concretas, donde puede resultar crucial la flexibilidad del análisis topológico de datos para adaptarse a distintos contextos. Todo trabajo futuro en este campo deberá tener como punto de referencia los modelos que aquí presentamos. Dicho esto, esperamos lograr en el lector la comprensión de ambos trabajos a la luz de un mismo marco teórico.

El trabajo está organizado de la siguiente manera. En el primer capítulo describimos los elementos de filogenética que se utilizarán en el resto de la tesis. Comenzamos definiendo la notación básica que vamos a usar en este trabajo para plantear los modelos básicos de árboles y redes filogenéticas. Concluimos esta parte con un breve resumen de algunas estrategias ampliamente utilizadas haciendo énfasis en las bases matemáticas que las sustentan.

En el segundo capítulo repasaremos los elementos de topología algebraica que sirven de base a los modelos. Partiremos de complejos simpliciales y homología simplicial para terminar con las nociones de homología persistente en las que se basan los modelos particulares que aquí abordamos. El Apéndice A contiene a los elementos necesarios para la prueba de un resultado particular que es parte de la demostración del teorema fundamental detrás del modelo CCR. La exposición que aquí hacemos adapta la teoría general de homología persistente a los complejos simpliciales y filtraciones necesarios para esta aplicación concreta. Para mayores detalles de homo-

logía persistente recomendamos los artículos panorámicos de Carlsson [9] de 2009, o los de Edelsbrunner [20] y Ghrist [33] de 2008 así como el libro de Edelsbrunner y Harer [21] de 2010.

En el Capítulo 3 estudiamos el modelo CCR; es decir, exponemos el modelo propuesto en [13] y [14]. Presentamos primero los detalles matemáticos y los resultados principales que le sirven de base. A continuación establecemos el modelo CCR, su interpretación, los indicadores que los autores proponen para eventos reticulares y mencionamos además las pruebas que los autores hacen mediante simulaciones y datos reales. Terminaremos el capítulo con una lista de observaciones puntuales sobre posibles direcciones a explorar partiendo de las ideas de Chan, Carlsson y Rabadan.

En el cuarto capítulo desarrollamos las ideas que aparecen en el trabajo de Parida *et al.* [54]. Establecemos los detalles del fenómeno concreto que se intenta modelar y los conceptos de topología propuestos para abstraer el fenómeno. Los resultados matemáticos que aquí presentamos difieren de los que aparecen en el artículo y son consecuencia de la comunicación con uno de los autores de [54], que amablemente respondió a nuestras preguntas. Se construyó junto con él un resultado ligeramente diferente, pero que sustenta la interpretación que se propone en el artículo de los códigos de barras de persistencia para detectar mezcla de poblaciones.

Capítulo 1

Filogenética

Introducción

El objetivo de la filogenética molecular es *explicar* las relaciones evolutivas entre diferentes especies u organismos partiendo de un análisis comparativo de secuencias moleculares. Típicamente, estas relaciones evolutivas se pueden representar apropiadamente con un grafo, siendo los *árboles filogenéticos* la herramienta más utilizada para ello.

Los análisis filogenéticos tienen como objetivo describir las características de la historia evolutiva concreta de un conjunto de organismos. En términos propios del área, se dice que la filogenética se dedica a *reconstruir* la historia evolutiva de un conjunto de organismos. Es claro que en muchos casos no es posible conocer el proceso que a lo largo de la historia originó la muestra de organismos que se pretende estudiar, pero lo que sí se puede es proponer posibles *filogenias* que expliquen la historia del conjunto de esos organismos. El rango de objetivos de la filogenética va desde tratar de estimar los diferentes ancestros de la raza humana hasta inferir el *árbol de la vida* que da cuenta de la biodiversidad que hay en la Tierra.

Ante la importancia de la teoría de la evolución en biología y la reciente abundancia de datos genómicos, se ha intensificado el uso de herramientas filogenéticas para apoyar el estudio de otras áreas como la taxonomía, la epidemiología, la evolución de familias de genes o las relaciones coevolutivas entre organismos parásitos y hospederos.

En el artículo panorámico *Mathematical Challenges from Genomics and Molecular Biology* publicado en *Notices of the American Mathematical Society* en 2002, Karp [43] incluye el problema de la reconstrucción filogenética dentro de las áreas de estudio de la biología en las que la necesidad de nuevas herramientas matemáticas y computacionales surge de manera natural. Karp enfatiza: "*A medida que el nivel de las aspiraciones [de la biología] se eleva y la cantidad de datos disponibles crece en órdenes de magnitud, el campo se vuelve cada vez más dependiente de la modelación matemática, del*

análisis matemático y de la computación".

La filogenética ha sido, y sigue siendo, un campo fértil para la interacción entre la biología y distintas ramas de las matemáticas. Por ejemplo, durante los últimos cincuenta años, los esfuerzos por modelar las distintas fuerzas que moldean el proceso evolutivo de un conjunto de organismos han dado como fruto una rica teoría estadística llamada *genética de poblaciones* que integra aspectos estocásticos y determinísticos de la evolución.

Por otro lado, el reto de desarrollar las herramientas de reconstrucción filogenética ha requerido la interacción de distintas áreas de las matemáticas, como la combinatoria, la teoría de grafos e incluso la geometría algebraica. También debe mencionarse que la magnitud y complejidad de los datos y fenómenos involucrados demanda el desarrollo de algoritmos eficientes y técnicas heurísticas avanzadas que puedan hacer frente a la inevitable presencia de problemas con complejidad NP.

En este trabajo vamos a describir de la manera más simple posible las nociones de biología que se van a modelar comentando distintas estrategias que se han desarrollado para cada caso. No pretendemos ofrecer una presentación exhaustiva de los fenómenos biológicos, ni hacer un compendio de las bases matemáticas requeridas en las herramientas filogenéticas existentes. En cambio, simplificaremo—quizá de más—los fenómenos a modelar. Asimismo, comentamos brevemente las ideas matemáticas que soportan la modelación, análisis e inferencia de las distintas herramientas de reconstrucción filogenética. Esto último, con el fin de ubicar dentro de un contexto específico los modelos de homología persistente que explicamos a detalle en los capítulos posteriores.

Este capítulo está organizado de la siguiente manera. En la primera sección presentamos la notación y los elementos de teoría de grafos sobre los que se sustentan las discusiones de las secciones siguientes. A continuación, en la Sección 1.2 presentamos los árboles filogenéticos, comentando los fenómenos biológicos que motivan su uso, su definición matemática concreta y dando un vistazo a las distintas técnicas para la reconstrucción de los mismos. Por último, en la Sección 1.3 comentamos los fenómenos evolutivos que no pueden ser representados en forma de árbol filogenético y algunos de los esfuerzos que se han hecho para lidiar con este problema. Nos interesan particularmente los eventos horizontales porque son los que vamos a modelar en los siguientes capítulos mediante homología persistente.

1.1. Elementos de teoría de grafos

La teoría de grafos constituye un lenguaje natural para describir las relaciones evolutivas entre los organismos. En esta sección presentaremos algunas definiciones que requerimos para plantear las preguntas sobre filogenética que vamos a abordar en nuestro trabajo y también estableceremos

la notación a utilizar.

Para un estudio más amplio acerca de la extensa teoría matemática relacionada con filogenética recomendamos el artículo panorámico de Steel [60] y el libro de filogenética, dirigido a matemáticos, publicado por Semple y Steel [59]. Las definiciones y notación que aquí presentamos siguen, en la medida de lo posible, la notación de este último volumen.

Empezaremos estableciendo la notación para los grafos no dirigidos describiendo las propiedades que usaremos en este trabajo. Al final de la sección consideramos también los grafos dirigidos, dando por hecho que, en general, se puede usar la misma terminología para describir ambos tipos de grafo.

DEFINICIÓN 1.1 Un *grafo no dirigido* $G(V_G, E_G)$ es un conjunto finito, no vacío, de *vértices* o *nodos* V_G y un conjunto de *aristas* E_G en el que cada arista $e \in E_G$ es de la forma $\{u, v\}$ y donde u y v son elementos de V_G.

Dado un grafo $G(V, E)$ y una arista $e = \{u, v\}$ decimos que e *conecta* u y v y que dos nodos son *adyacentes* cuando existe una arista que los conecta. Por otro lado, a los dos nodos que definen una arista $e = \{u, v\}$ los llamamos los *extremos* de e. Si $u = v$ decimos que e es un *bucle*. Además, si existe otra arista $e' \in E$ que conecte a los mismos nodos que e decimos que e y e' son aristas *paralelas*. Un grafo es *simple* si no tiene bucles ni aristas paralelas. En este trabajo, a menos que se especifique lo contrario, supondremos que todos los grafos son simples.

Dado un nodo $u \in V$ definimos el *grado* de u como el número de aristas que conectan con dicho nodo. A los vértices de grado 1 les llamamos *hojas* y si un vértice tiene grado 0 lo denominamos *aislado*.

Si tenemos los grafos $H(V_H, E_H)$ y $G(V_G, E_G)$, decimos que H es *subgrafo* de G si $V_H \subset V_G$ y $E_H \subset E_G$. Al considerar un subconjunto de vértices $V' \subset V$ del grafo G, el *grafo inducido por* V' consiste en la pareja $G'(V', E')$, donde $E' \subset E$ es el conjunto de todas las aristas de G que conectan elementos de V'. De la misma manera, dado un subconjunto de aristas $E'' \subset E$ el *grafo inducido por* E'', consiste en la pareja $G''(V'', E'')$, donde $v \in V''$ si existe una arista en E'' que lo tenga como extremo.

Respecto del grafo $G(V, E)$ decimos que el grafo inducido por un subconjunto de nodos $V' \subset V, |V'| = k$ es un *clique* en k puntos si para cada par de vértices $u, v \in V'$ la arista $\{u, v\}$ está en E'. Cuando el grafo G es un clique en todos sus nodos también se dice que es un *grafo completo*.

Un *camino* en el grafo $G(V, E)$ es un conjunto de vértices $v_0, v_1, \ldots, v_k$ tal que $\{v_i, v_{i+1}\} \in E$ para cada $i \in \{0, 1, \ldots, k-1\}$. Dados $u, v \in V$ decimos que existe un camino que va de u a v en k pasos si existe un camino con $k+1$ vértices en G tal que $v_0 = u$, y $v_k = v$. Un *ciclo* es un camino $P = \{v_0, v_1, \ldots, v_k\}$ tal que $v_0 = v_k$.

Un grafo $G(V, E)$ se denomina *conexo* si para cada par de vértices $u, v \in V$ existe un camino que va de u a v. Si un grafo conexo no tiene ciclos decimos que es un *árbol*.

Definición 1.2 Un *grafo ponderado* es un grafo $G(V, E)$ conexo y una función $\omega : E \to \mathbb{R}$ que asigna valores no negativos a cada arista. Al valor $\omega(e)$ de cada arista lo denominamos el *peso* o *largo* de la arista e.

Llamamos el peso de un camino a la suma de los pesos de las aristas que lo definen.

Nos encontramos con un *grafo aditivo* cuando todos los pesos asignados a las aristas de un grafo ponderado son no negativos. Todo grafo no dirigido puede considerarse como un grado aditivo, asignando una distancia 1 a cada una de sus aristas. Dado un $r > 0$, decimos que un camino $P = \{u_0, \ldots, u_k\}$, dentro de un grafo aditivo, es un *r-camino* si satisface que $d(u_i, u_{i+1}) \leqslant r$, para todo $i \in \{-1, \ldots, k-1\}$.

La propiedad más importante de los grafos aditivos es que nos permiten definir una distancia entre los vértices.

Lema 1.1 Si $G(V, E)$ es un grafo aditivo, con una función de pesos ω entonces el par (V, d_G) forma un espacio métrico aditivo en el cual el valor de $d_G(u, v)$ queda definido por el peso de algún camino mínimo entre ellos, es decir:

$$d_G(u, v) := \min_{P \text{ camino de } u \text{ a } v} \sum_{\{u_i, u_{i+1}\} \in P} \omega(\{u_i, u_{i+1}\}).$$

Denotaremos a (V, d_G) por $\mathcal{M}(G)$ y nos referiremos a la función distancia como la *distancia inducida por caminos mínimos sobre G*.

Observemos que aunque es posible que no exista un camino mínimo único entre dos nodos, la distancia entre ellos sí está bien definida. Es claro que para cualquier espacio métrico finito (M, d) es posible representar su métrica en un grafo aditivo. Esto se logra tomando a M como conjunto de vértices, construyendo el grafo completo sobre este conjunto y pesando la arista $\{x, y\}$ con el valor de $d(x, y)$. En filogenética resulta de especial interés la métrica que un árbol induce en sus vértices.

Definición 1.3 Decimos que un espacio métrico finito $\mathcal{M}(M, d)$ es *aditivo* si existe un árbol $T(V, E)$ aditivo y una función $\pi : M \to V$ que preserve las distancias. Esto es,

$$d(x, y) = d_T(\pi(x), \pi(y)), \qquad \forall x, y \in M.$$

Es fácil ver que si (M, d) es un espacio métrico aditivo, entonces existe un árbol T al que podemos incluir M, con una isometría, de tal manera que las hojas de T estén contenidas en la imagen de M. Los siguientes dos resultados elementales de árboles aditivos, el Lema 1.2 y la Proposición 1.1, fueron

tomados, junto con sus demostraciones, del material complementario del artículo [14].

LEMA 1.2 Si M es un espacio métrico aditivo, entonces M puede ser incluido isométricamente en $\mathcal{M}(T)$ para algún árbol aditivo T de tal manera que todas las hojas de T estén dentro de la imagen de M en el conjunto de vértices de T.

DEMOSTRACIÓN

Al ser M un espacio métrico aditivo, por definición, existe un árbol aditivo $T(V, E)$ y una isometría $\pi : M \to V_T$. Vamos a construir ahora un árbol T' con base en T que satisfaga la segunda condición. Supongamos que $e \in V_T$ es una hoja de T que no está contenida en la imagen de π, es decir $\pi(M) \subset V_t \setminus \{e\}$. Observemos primero que ningún camino mínimo entre los elementos de $\pi(M)$ pasa por e: sea $(f, e) \in E_T$ la única arista que contiene a e y sea C un camino entre dos nodos distintos a e que pasa por e, luego C debe ser de la forma

$$C = \ldots (c_i, f)(f, e)(e, f)(f, c_j) \ldots$$

y por lo tanto el camino C' que se obtiene de C al eliminar $(f, e)(e, f)$ es más corto que C. En particular esto implica que ningún camino mínimo en T entre los elementos de $\pi(M)$ pasa por e.

Así pues, como $\pi(M) \subset V_t \setminus \{e\}$ y los caminos mínimos entre los elementos de $\pi(M)$ no pasan por e podemos incluir de manera isométrica a M en el árbol T' que tiene como conjunto de vértices $V_T \setminus \{e\}$ y como conjunto de aristas a $E_T \setminus \{\{f, e\}\}$.

Si continuamos "*podando*" las hojas innecesarias en T de esta manera obtendremos el árbol que buscamos.

□

Ahora bien, es claro que si tenemos un árbol aditivo $T(V, E)$ y tomamos un subconjunto de vértices $V' \subset V$ que solo difiere de V por un conjunto de hojas, entonces el grafo inducido $T'(V', E')$ es, a su vez, un árbol aditivo. Decimos pues que T' es un *subárbol* aditivo. Veamos a continuación que T' define un espacio métrico aditivo sobre V' sin necesidad de ser considerado como subconjunto de T.

PROPOSICIÓN 1.1 Sean T un árbol aditivo y $T' \subset T$ un subárbol aditivo que se obtiene al remover un conjunto de hojas $\{e_1, \ldots, e_n\}$ de T. Se tiene entonces la siguiente igualdad

$$d_{T'} = d_T \mid_{V_{T'} \times V_{T'}} .$$

Esto es, $\mathcal{M}(T')$ es un subespacio métrico de $\mathcal{M}(T)$ ya que la función distancia en $\mathcal{M}(T')$ es una restricción de la función distancia en $\mathcal{M}(T)$.

DEMOSTRACIÓN

Para probarlo es suficiente ver que la ecuación es cierta cuando una hoja es removida y luego se puede proceder inductivamente. Es decir, todo lo que necesitamos mostrar es que dada una hoja e, el camino mínimo entre dos nodos cualquiera $v_0, v_1 \in V \setminus \{e\}$ no pasa por e y esto ya se vio en la prueba del Lema 1.2.

□

En ocasiones es útil considerar las aristas de un grafo como *flechas* que conectan un nodo con otro. Al asignarles una *dirección* a las aristas podemos diferenciar los nodos que las componen. Es decir, podemos diferenciar si las aristas *salen* de un nodo o si *apuntan*, o *llegan*, a otro.

Definición 1.4 Un *grafo dirigido* $G(V_G, E_G)$ consiste en un conjunto finito, no vacío, de *vértices* o *nodos* V_G y un conjunto de *arcos*, o aristas, E_G. Cada arco $e \in E_G$ es de un par ordenado (u, v), donde u y v son elementos de V_G.

Dado un grafo dirigido $G(V, E)$ se dice que la arista $(u, v) \in E$ *va de* u a v. En este caso, v es la *cabeza* o *punta* de la arista mientras que u es su *base* o *cola*. El *grado de entrada* de u es el número de aristas que llegan al vértice u. Asimismo, el número de aristas que salen de él constituyen el *grado de salida* de u. El grado de un nodo en un grafo dirigido no es otra cosa que la suma de sus grados de entrada y salida.

En general, la nomenclatura para grafos no dirigidos se puede adaptar de manera similar a los grafos dirigidos. Por ejemplo, un *camino dirigido* en un grafo dirigido es una secuencia distinta de vértices $v_0, \ldots, v_k$ tal que $(v_i, v_{i+1}) \in E$ para $i \in \{0, 1, \ldots, k-1\}$. Todo grafo no dirigido puede considerarse dirigido conviniendo que cada arista $\{x, y\}$ define las dos aristas dirigidas (x, y) y (y, x) involucradas.

Un árbol dirigido es *árbol binario* cuando todos los nodos tienen grado de salida menor o igual a 2. Siendo el caso, se les dice *hojas* a los nodos con grado de salida cero y *nodos interiores* a los nodos que tienen grado de entrada y de salida positivos. Los nodos interiores de un árbol binario que tienen grado tres son denominados *junturas*.

1.2. Árboles filogenéticos

En esta sección daremos un breve repaso a las ideas principales alrededor de los métodos de reconstrucción de árboles filogenéticos. Presentaremos primero una breve descripción del tipo de fenómenos biológicos que queremos representar en forma de árboles filogenéticos y a continuación los definiremos detalladamente en su calidad de objetos matemáticos. Luego daremos cuenta de los datos que sustentan los análisis de reconstrucción.

Por último, comentaremos los distintos enfoques en los que se basan dichos análisis. El contenido que aquí resumimos se fundamenta en los libros [2, 41, 45, 59], cuya lectura sugerimos para abordar a fondo este tema.

1.2.1. Motivación biológica

De acuerdo con la teoría de evolución, todas las especies provienen de un ancestro común. Así, deducimos que desde el origen de la vida en la Tierra, las nuevas especies que van apareciendo en el tiempo surgen de especies ya existentes. Diferentes factores han influido en este proceso que ha desembocado en la biodiversidad actual y que continúa moldeándose al día de hoy. Entre estos mecanismos, existen eventos de evolución molecular que son indispensables para explicar el origen de nuevas especies. Por ejemplo, fenómenos como las mutaciones, la duplicación de genes, la recombinación o los reacomodos de genomas.

Entre el conjunto de estos sucesos, el concepto de *mutaciones* es el más utilizado para inferir relaciones evolutivas entre los genes. En este trabajo entendemos por mutaciones los cambios puntuales en la información genética de un organismo que son heredados a su descendencia. Ejemplos de mutaciones puntuales en el ADN son los cambios de base, las inserciones y las deleciones.

La acumulación de estos cambios en la información genética se traduce en cambios fenotípicos que provocan que una especie dé origen a otra. En concreto, dos poblaciones diferentes de una misma especie pueden dar origen a dos especies distintas si pierden la capacidad de reproducirse entre ellas debido a la acumulación de estos cambios en el material genético lo que impide el apareamiento o que la descendencia sea fértil.

Consideramos que dos especies están relacionadas si comparten un ancestro común *reciente*. Además la especie a están *más relacionadas* con la especie b que con la c si el último ancestro común más reciente que tienen a y b es más reciente que el último ancestro común más reciente entre a y c.

Durante mucho tiempo los métodos clásicos de reconstrucción filogenética dependían de comparar aspectos morfológicos de los organismos. Ello dificultaba que los métodos de reconstrucción fueran objetivos dadas las limitantes de la información disponible y la falta de herramientas adecuadas para el análisis. Esto ha cambiado en los últimos años debido a la velocidad con la que se incrementa la obtención de información genética disponible, como cadenas de nucleótidos y aminoácidos, y los avances tecnológicos que permiten tener la capacidad de procesar dicha información. Es por ello que las técnicas de reconstrucción moleculares son las más aceptadas en la actualidad.

Investigar la evolución y la cercanía evolutiva entre genes e individuos requiere de un análisis comparativo de las secuencias genéticas de los organismos que se están estudiando. Es un área muy amplia que posee diversas aplicaciones como describir la evolución de genes concretos; reconstruir el árbol genealógico de organismos de una misma especie; establecer el proceso evolutivo de cepas de una misma familia e incluso estudiar el proceso evolutivo que originó grandes conjuntos de especies, como los mamíferos o los

reptiles. Todos ellos son esfuerzos por estudiar *pedazos*, de distinto tamaño y resolución, del árbol universal de la vida.

Desde las hipótesis simplificadas de las que partimos es viable representar los procesos evolutivos en forma de árboles filogenéticos en los cuales los vértices representan las distintas especies que son parte de la filogenia de la muestra que se encuentra dentro de las hojas del árbol. A su vez, las aristas representan la relación de *paternidad evolutiva* entre la secuencia *parental* y la secuencia *hija*. Existen diferentes maneras de expresar más información en un árbol evolutivo, es por ello que suelen utilizarse árboles enraizados con las aristas pesadas en los que la raíz del nodo representa el ancestro común más reciente de la muestra y el peso de las aristas cuantifica de alguna manera la *diferencia evolutiva* que distingue la secuencia parental de la hija.

En general, esta diferencia evolutiva trata de reflejar el número de cambios o mutaciones que separan una secuencia de otra, o bien, el tiempo estimado que se necesita para que ocurran los cambios que dan cuenta de la diferenciación. Notemos que estas mismas estructuras pueden ser adaptadas para explicar las diferentes situaciones que ya mencionamos.

1.2.2. Árboles filogenéticos

Como objeto matemático, el árbol filogenético pertenece a una familia de estructuras más amplias que permiten asociar grafos a conjuntos finitos. Este tipo de estructuras reciben el nombre de *grafos asociados a un conjunto* o *grafos definidos sobre un conjunto*. Entre dichas estructuras, los árboles filogenéticos permiten asociar un árbol a un conjunto de datos de tal manera que coincida con la interpretación biológica que se requiera de ellos.

Definición 1.5 Consideremos un conjunto finito C. Un *árbol filogenético T sobre C* es un árbol (T, ϕ) sobre C con la propiedad de que ϕ es una biyección con las hojas del árbol. Si además T es un árbol binario, decimos que T es un *árbol filogenético binario sobre C*.

De aquí en adelante, a menos que se diga lo contrario, llamaremos *muestra de organismos* al conjunto de secuencias genómicas extraídas de los organismos en cuestión. Cuando sea necesario especificaremos si nos estamos refiriendo a secuencias de ADN, de ARN, aminoácidos u otros caracteres genéticos. Partiendo de varias simplificaciones podemos suponer que una muestra de organismos evoluciona de la manera ya descrita.

Ahora bien, si el proceso evolutivo de una muestra efectivamente tiene la forma de un árbol filogenético esperaríamos que la distancia de disimilitud entre los organismos de la muestra refleje la distancia definida por caminos mínimos dentro del árbol.

Definición 1.6 Sea X un conjunto finito; una función arbitraria $\delta : X \times X \to \mathbb{R}$ es un *mapeo de disimilitud* si para todos los $x, y \in X$ se tiene que $\delta(x, x) = 0$ y $\delta(x, y) = \delta(y, x)$.

Afirmamos que un mapeo de disimilitud induce una distancia aditiva, si existe un árbol filogenético aditivo sobre X, (T, ϕ), de tal manera que

$$\delta(x, y) = d_T(\phi(x), \phi(y)).$$

En este caso, es fácil ver que el mapeo de disimilitud constituye un espacio métrico aditivo. Para asociar un árbol a un mapeo de disimilitud, es necesario asegurarnos primero de que el mapeo es en efecto aditivo, y sólo después, podremos reconstruir el árbol que le da soporte. El siguiente resultado ofrece un criterio útil para saber si un mapeo de disimilitud es aditivo. Apareció independientemente en los trabajos de Zaretskii [64] de 1965, Pereira [55] de 1969 y Buneman [7] de 1974.

Teorema 1.1 Sea δ un mapeo de disimilitud sobre X. Entonces, δ es una métrica aditiva en X si, y sólo si, para cada cuarteta de elementos, no necesariamente distintos $w, x, y, z \in X$ se tiene que

$$\delta(w, x) + \delta(y, z) \leqslant \max\{\delta(w, y) + \delta(x, z), \delta(w, z) + \delta(x, y)\}.$$

Si asociamos a la muestra un mapeo de disimilitud y éste resulta ser aditivo, nos interesará encontrar un árbol T que describa de forma coherente sus diferencias. De hecho, se puede probar que si se tiene un espacio métrico aditivo hay una sola representación en forma de árbol de tal distancia[1]. Incluso existen algoritmos para reconstruir el árbol que sustenta la métrica aditiva. Ejemplos de ello se pueden ver en Bandelt y Dress [3] o en Buneman [6].

Como veremos más adelante, las secuencias genéticas de las muestras rara vez forman un espacio aditivo *exacto*. Por lo tanto, las técnicas de reconstrucción filogenética suelen tener como objetivo llegar al árbol que mejor se *aproxime* a la matriz de disimilitud obtenida con una muestra de datos reales. Sin embargo, por la naturaleza combinatoria del problema, los algoritmos diseñados para esta tarea poseen una complejidad computacional considerable. Ver por ejemplo los trabajos de Erdös *et al.* [26, 27].

1.2.3. Datos

Como ya mencionamos, los análisis filogenéticos establecen relaciones entre genes, o fragmentos de genes, para inferir su historia común. Ello requiere suponer que las secuencias que se van a comparar son *homólogas*. En el contexto de la biología se dice que dos secuencias son homólogas cuando comparten un ancestro común. Esta acepción de *homología* se utilizará únicamente en esta sección. En el resto de nuestro trabajo siempre que nos refiramos a homología lo haremos en el sentido matemático del término.

[1] Ver el Teorema 7.1.8 del libro de Semple y Steel [59], página 148.

Por su naturaleza, las secuencias homólogas son muy parecidas; es decir, comparten los mismos nucleótidos o aminoácidos en la mayoría de los sitios. Al proceso de establecer qué regiones del genoma son homólogas, y por lo tanto susceptibles de usarse en la comparación, lo llamamos *alineamiento*. Una vez alineadas dos secuencias, la manera más sencilla de medir su grado de similitud es considerar el porcentaje de nucleótidos idénticos relativos al largo de la secuencia.

Definición 1.7 Considérense dos palabras $a = a_1a_2 \dots a_n$ y $b = b_1b_2 \dots b_n$ con la misma cantidad de letras. Sea k el número de posiciones i en las cuales a y b difieren, entonces la *distancia de Hamming* entre a y b se define como

$$d_H(a, b) = \frac{k}{n}.$$

En biología, a esta distancia se le conoce también como la *distancia observada* o incluso como la *p-distancia* entre secuencias.

A lo largo del tiempo, pueden ocurrir múltiples mutaciones por sitio y las secuencias pueden volverse aleatorias o saturadas. Como consecuencia de este proceso, la distancia observada suele subestimar la verdadera *distancia genética*, es decir, el número real de sustituciones que ocurrieron en cada sitio. Así pues, es posible que dos secuencias estén menos relacionadas de lo que aparenta su distancia observada. Existen modelos matemáticos de sustitución que *corrigen* esta distancia observada para dar una medida más realista de la distancia genética entre secuencias.

Un *modelo de sustitución* que brinda una versión más realista de la distancia genética, toma en cuenta las distintas probabilidades con las que pueden ocurrir las mutaciones, o sustituciones entre nucleótidos, a lo largo del tiempo. Es decir, incluyen en el análisis suposiciones específicas sobre la probabilidad de que si la secuencia parental tiene el nucleótido x en la posición i, éste herede el nucleótido y en la misma posición.

Estos modelos de probabilidad permiten describir las mutaciones como una cadena de Markov. Diferentes suposiciones sobre las probabilidades asociadas a las distintas mutaciones dan origen a distintas distancias. Entre los modelos más utilizados se encuentran los de *Jukes-Cantor*, *Nei-Tamura* así como el *General Time Reversible Model*. Una vez seleccionado un modelo de sustitución, existen procedimientos para inferir fácilmente la distancia genética de las secuencias a partir de una muestra. Para ahondar en el tema de los modelos de sustitución desde un punto de vista biológico recomendamos el libro de Huson *et al.* [41] y para un enfoque probabilista, el trabajo de Ewens [30].

Aunque las secuencias de ADN son las más usadas en los análisis filogenéticos, se pueden utilizar otras fuentes de información. Las distintas características que se eligen para comparar una muestra de organismos se denominan *caracteres*. Matemáticamente hablando, un carácter es una función que traslada el conjunto de organismos bajo estudio a un espacio que

permite comparar los organismos entre sí. Para ello necesitamos concretar cómo será la comparación de los *caracteres* biológicos en un lenguaje matemático.

Definición 1.8 Un *carácter*, cdualitativo o cuantitativo, sobre un conjunto X es una función χ definida sobre un subconjunto X' no vacío de X hacia un conjunto C. Definimos C como el *conjunto de estados* del carácter χ.

Entre los ejemplos de caracteres biológicos que se pueden usar para analizar una muestra existen algunos morfológicos, como la cantidad de patas de un animal o el tipo de sistema respiratorio; también pueden ser bioquímicos, fisiológicos, de comportamiento, etcétera. Es importante notar que el hecho de definir el dominio de un carácter como un subconjunto del total nos permite considerar características que no se encuentran presentes en todos los individuos de la muestra. Para un estudio matemático sobre la factibilidad de asociar árboles filogenéticos a caracteres recomendamos el libro de Semple y Steel [59].

Hemos mencionado ya a las secuencias de ADN como ejemplos de caracteres genéticos. Otros caracteres genéticos útiles están en forma palabras binarias en las cuales el valor 1 en un sitio específico representa la *presencia* de una característica genética y el estado 0 corresponde a la *ausencia* de tal característica. Las características genéticas más simples pueden representar la presencia de un sitio diferente a las otras secuencias, con lo cual corresponden a una mutación puntual. Pero se pueden considerar características genéticas más generales como la presencia/ausencia de un gen, diferencias en los codones dentro de un mismo gen, entre otras.

Debido a la naturaleza estocástica de la evolución y a las limitantes en el procesamiento de los datos es poco común obtener un espacio métrico aditivo a partir de las distancias observadas en la muestra. Comentaremos en la siguiente sección las estrategias que suelen utilizarse para asociar un árbol filogenético a una muestra de organismos representada por un conjunto de caracteres.

1.2.4. Reconstrucción de árboles filogenéticos

Como ya mencionamos, los análisis filogenéticos suelen comenzar con un alineamiento de secuencias multiespecies. A partir de ahí, existen principalmente dos tipos de análisis según el tipo de información que utilicen. Por un lado, están los basados en la matriz de distancias que se obtiene del alineamiento y por otro, los que se basan en la información de las secuencias contenidas en el mismo. Pese a la disponibilidad de distintos métodos, ninguno garantiza que el árbol filogenético inferido represente efectivamente la filogenia de la muestra.

Los métodos basados en secuencias pueden utilizarse sobre cualquier conjunto de caracteres discretos. Cuando se comparan secuencias, cada posición

en la secuencia es un carácter donde los distintos nucleótidos son conjunto de estados. Usualmente, cada posición se analiza de manera independiente, y por lo tanto, se considera cada columna del alineamiento como un proceso evolutivo independiente. Los métodos por secuencias retienen los valores originales de cada sitio y tratan de reconstruir los valores de las secuencias ancestrales a partir de ellos.

En un análisis basado en el método de distancias, el primer paso es obtener la matriz de distancias observadas—la distancia de Hamming—entre las secuencias. Estas distancias se *corrigen* al adaptarlas con un modelo de sustitución apropiado. Posteriormente, la matriz de distancias resultante se utiliza como base para distintos métodos como el UPGMA[2]. Los métodos por distancias descartan la información de los estados que toman los caracteres de los diferentes organismos. Por lo tanto, se pierde la información requerida para reconstruir los estados de los caracteres asociados a los nodos ancestrales. La principal ventaja de los métodos por distancia es que requieren poco poder de cómputo y por lo tanto, permiten realizar análisis para muestras de gran tamaño.

Otra manera de clasificar los métodos es en función del tipo de algoritmo que utilizan para realizar inferencias. Los algoritmos pueden estar basados en ideas de clustering, parsimonia, máxima verosimilitud o partiendo de un enfoque bayesiano, entre otros. Un enfoque de máxima verosimilitud, tiene como objetivo llegar a un árbol filogenético que maximice la verosimilitud de generar la muestra bajo un modelo de evolución particular. En cambio, un método basado en parsimonia elige un árbol filogenético que explique la muestra usando un número mínimo de mutaciones observables. Por su parte, los métodos bayesianos, basados en los datos iniciales, un modelo de evolución específico y una distribuición a priori sobre los árboles filogenéticos candidatos a explicar la historia de la muestra; obtienen una distribución de probabilidad entre estos mismos árboles y le asignan el que resulta con una mayor probabilidad de explicar su historia.

Información sobre el software disponible y más detalles sobre la metodología de cálculo de árboles filogenéticos se pueden encontrar en el libro de Hall [36]. Para ahondar en las bases teóricas de los métodos y los distintos modelos de evolución, recomendamos [2, 31, 40].

1.3. Eventos evolutivos horizontales

Una de las principales suposiciones que asume la mayoría de los métodos filogenéticos es que la filogenia de la muestra está determinada de manera única. Esto es, que efectivamente la historia de la muestra se puede representar con un árbol filogenético. Sin embargo, existen fenómenos evolutivos

[2]Unweighted Pair Group Method with Arithmetic Mean,UPGMA, es un método de clustering jerárquico ampliamente utilizado en filogenética y minería de datos.

que no permiten que la historia de una muestra pueda representarse como árbol. Entre estos fenómenos centraremos nuestra atención en los llamados *eventos evolutivos reticulares* o *eventos horizontales*[3].

Los eventos reticulares son fenómenos en los cuales una secuencia no tiene una única secuencia parental, sino que es producto de la interacción de dos secuencias parentales. Un suceso reticular podría ser por ejemplo el evento de la *recombinación*, la cual—para fines de este trabajo—está definida como el intercambio de información genética entre dos secuencias de nucleótidos que da como resultado una secuencia con partes de las dos secuencias parentales, y que es heredada a la descendencia. Se sabe ahora que la recombinación influye de manera importante en los procesos evolutivos. Entre otras cosas, la recombinación ayuda a explicar la compleja diversidad genética presente en las poblaciones naturales.

Los eventos de recombinación impiden que la historia de una muestra sea explicada con un único árbol filogenético ya que al mezclarse secuencias de padres diferentes, ocurre que diferentes segmentos de las secuencias demandan árboles distintos. Por lo tanto, a una secuencia resultante de un evento de recombinación no se le puede asociar una única secuencia parental, sino que se requieren dos secuencias parentales para explicar su origen.

Existe una diversidad de fenómenos que requieren de la interacción de dos secuencias parentales para originar una nueva secuencia. No entraremos en más detalle sobre las diferentes situaciones en las que este fenómeno se presenta. Al decir evento horizontal de evolución nos podemos referir a fenómenos tan dispares como la recombinación y el reacomodo de segmentos genéticos en virus o los distintos mecanismos de transferencia horizontal de genes en bacterias: conjugación, transducción y transformación, etcétera. Especificaremos en los siguientes capítulos a qué fenómeno particular nos referimos cuando hablemos de eventos horizontales sólo si es necesario.

Otro objetivo elemental será contar con criterios para decidir cuándo hubo eventos reticulares en la historia evolutiva de una determinada muestra. Pese a que podemos examinar visualmente para buscar estos fenómenos en un alineamiento múltiple de secuencias, es necesario recurrir a pruebas estadísticas con el fin de evaluar apropiadamente la hipótesis de la presencia de eventos horizontales.

Asumiendo que los eventos horizontales son necesarios para explicar la historia de la muestra, lo ideal sería contar con una extensión de los árboles filogenéticos que permita incluir tales fenómenos en una estructura gráfica sencilla. Es así como surgen las *redes filogenéticas*. Entre ellas comentaremos más a detalle los llamados *grafos de recombinación ancestral* (ARG).

Ignorar la posible presencia de eventos horizontales durante la evolución de una muestra puede alterar seriamente el resultado de los análisis filo-

[3]La notación *eventos hortizontales* no es común en el lenguaje filogenético. Gracias al Dr. Luis José Delaye por esta aclaración.

genéticos, pues estaríamos usando el método en una muestra que no cumple las hipótesis necesarias para su aplicación. Un estudio en forma sobre la influencia que tienen los eventos horizontales en los métodos filogenéticos puede encontrarse en los libros de Huson [41] y Lemey [45] que en gran parte utilizamos como base de este capítulo.

1.3.1. Detección de eventos horizontales

Resulta clara la importancia de detectar los trazos de recombinación entre las secuencias parentales de una muestra. De hecho, existe toda una variedad de métodos que se han desarrollado con este propósito. En general, estos métodos se pueden clasificar en cuatro tipos: métodos de distancias, métodos filogenéticos, métodos de compatibilidad y métodos de sustitución. Los métodos de distancias suelen ser veloces y consisten en buscar patrones invertidos de las distancias entre las secuencias al utilizar un enfoque de *ventana deslizante*. Los métodos filogenéticos comparan los árboles filogenéticos asociados a segmentos adyacentes dentro de las secuencias. Si estos segmentos originan árboles distintos se tiene un indicador de la posible presencia de eventos de recombinación.

En 2001, Posada y Crandall [56] compararon 14 de los métodos más utilizados edn aquel entonces. En este trabajo los autores muestran que los métodos de compatibilidad y de sustitución son relativamente más poderosos que los métodos más sencillos basados en distancia. Algunos de los métodos que sobresalen por su eficiencia son Genecov, Maxχ^2, RDP, Phypro y el *neighbor similarity score* (NSS) como se menciona en [5].

1.3.2. Redes filogenéticas

La representación natural de fenómenos evolutivos horizontales mediante árboles filogenéticos conduce al concepto de *redes filogenéticas*. Es por ello que estos sucesos sed denominan también *eventos reticulares*.

Definición 1.9 Una *red filogenética* es cualquier grafo usado para representar relaciones evolutivas entre los ancestros de una muestra. Esto es, la muestra etiqueta un subconjunto de nodos del grafo, usualmente con nodos de grado 1.

Notemos que esta definición más amplia que incluye a los árboles filogenéticos se realiza de manera general para incluir a los grafos que representan de manera *explícita* los eventos reticulares y también los que lo hacen de manera *abstracta*. Es decir, utilizar un grafo para representar las incompatibilidades presentes en las secuencias que impiden asociarle satisfactoriamente un árbol filogenético a la muestra, aunque este grafo no pueda sedr interpretado de la misma manera que un árbol filogenético.

En teoría, las redes filogenéticas en las que cada nodo con grado de entrada 2 corresponde a un evento horizontal específico, pueden ser utilizados

para describir explícitamente estos fenómenos. Sin embargo, calcular las estructuras es una tarea complicada y como se menciona en [41]:"*los métodos computacionales diseñados para ello no han madurado aún para ser usados como herramientas prácticas de análisis filogenéticos*".

Se puede intentar construir una red filogenética imitando los pasos que utilizan las técnicas de reconstrucción de árboles filogenéticos. Esto es, primero se dejan al descubierto, mediante el alineamiento, las incompatibilidades con una estructura de árbol. Después, se convierte este resumen evolutivo en un diagrama que represente la forma en que dichas incompatibilidades alteran el agrupamiento natural que se tendría en un árbol evolutivo. Matemáticamente hablando, este diagrama es un grafo con ciclos, o en términos filogenéticos una *red filogenética abstracta*.

Esta red puede ser utilizada para visualizar datos que no forman un espacio métrico aditivo pero desafortunadamente no pueden ser interpretados en términos de la evolución de los caracteres en los que se basan. Esto se debe a que los nodos con grado de entrada mayor a 1 representan las incompatibilidades de la muestra, y no necesariamente representan estados ancestrales de los caracteres en las secuencias de la muestra. Por lo tanto, este tipo de redes no pueden interpretarse en términos evolutivos, a menos que se realicen análisis adicionales.

Las redes filogenéticas pueden calcularse a partir de una amplia variedad de datos. Por ejemplo, desde un alineamiento múltiple de secuencias, una matriz de distancias o un conjunto de árboles filogenéticos. Hasta donde sabemos, no se ha desarrollado un concepto de red filogenética que predomine sobre los demás, por ello existe en la literatura una gran variedad de definiciones de distintos tipos de redes filogenéticas[4]. La nomenclatura para cada tipo de red está inspirada en los diferentes aspectos que pueden adoptar. Algunas se definen en términos de los algoritmos combinatorios en los que se basa su construcción; otras, en propiedades matemáticas que las distinguen. Hay redes establecidas en función de las restricciones impuestas en su definición, o incluso por los fenómenos evolutivos específicos que intentan modelar. La notación de redes explícitas y abstractas que aquí utilizamos sigue la nomenclatura de [41] y [50]. La lectura de ambos volúmenes, junto con el trabajo de Makarenko [46], es recomendable para una introducción completa al tema.

Como se menciona en estos mismos trabajos, existen varias razones por las cuales el campo de las redes filogenéticas sigue abierto. Por un lado, muchos de los algoritmos para calcular redes explícitas existen sólo en forma de prototipo y no están diseñados para ser usados en estudios reales. Diseñar tales algoritmos es difícil no sólo por los problemas computacionales inherentes que presentan sino también porque los problemas NP involucrados suelen llevar a tiempos de funcionamiento poco prácticos.

[4] Como mencionan en [41], existen al menos 20 definiciones diferentes.

Por lo anterior, es necesario considerar modelos biológicos que incluyan de manera más específica los eventos evolutivos horizontales para producir resultados más pertinentes. Por otro lado, aunque los problemas computacionales asociados ya están muy estudiados y se cuenta con algoritmos que corren a tiempos aceptables, las redes filogenéticas abstractas son difíciles de interpretar y es complicado obtener con ellas información precisa acerca del proceso evolutivo de la muestra.

En su libro *Phylogenetic Networks*, Huson *et al.* [41], recalcan que: "*A tiempo presente, ninguno de los métodos para calcular redes filogenéticas explícitas es usado de manera amplia, ni son usados como herramienta rutinaria para ayudar a entender la historia evolutiva de un conjunto dado de especies en términos de mutaciones, especiación y diferentes tipos de eventos reticulares específicos*". Asimismo, hacen énfasis en la necesidad de nuevas herramientas matemáticas para modelar estos fenómenos reticulares: "*Mientras que las redes filogenéticas explícitas son conceptualmente muy atractivas, el desarrollo de métodos apropiados para calcularlas permanece como un reto formidable*".

Capítulo 2

Homología persistente

Introducción

El concepto de *homología persistente* fue introducido de manera independiente[1] por Robins [58] en 1999 y por Edelsbrunner, Letscher y Zomorodian [22] en 2002. Este método se ha convertido en los últimos años en una de las herramientas principales del *análisis topológico de datos*. No obstante sus orígenes recientes, la homología persistente también se ha posicionado en el centro de la topología aplicada y ha sido utilizada en un amplio espectro de campos de la ciencia y la ingeniería como se mencionó en la introducción de la tesis.

Existen distintos enfoques para abordar la homología persistente, dado que tiene componentes fundamentales de topología, álgebra y computación. Más aún, las potenciales aplicaciones de este método han sido parte central de su desarrollo como herramienta para el análisis de datos. Es por ello que el crecimiento del tema se ha dado en diversas direcciones, desde intentar generalizar los conceptos fundamentales—lo que ha llevado a descubrir nuevas conexiones dentro del álgebra abstracta—hasta la creación e implementación de nuevos algoritmos o estructuras de datos. Por otro lado, el uso de la homología persistente como una herramienta versátil de análisis de datos, ha puesto al descubierto la necesidad de un lenguaje que cuantifique la incertidumbre implícita en los análisis y ello demanda incorporar elementos de probabilidad y estadística.

El presente capítulo no ofrece una exposición exhaustiva de la teoría sino que únicamente aborda los elementos necesarios para entender el modelo CCR de evolución viral así como los distintivos topológicos para mezclas de poblaciones. Por lo tanto, motivaremos las definiciones y resultados adaptándolos a las necesidades de dichos modelos.

Como su nombre lo indica, el análisis topológico de datos consiste en una familia de herramientas que intentan describir los aspectos topológicos de

[1] Según se comenta en el libro de Edelsbrunner y Harer[21].

un conjunto de datos de manera análoga a como son descritos los espacios abstractos en el lenguaje de la topología algebraica.

En otras palabras, queremos describir una *forma* de la nube de puntos lo cual incluye precisar sus componentes conexas; describir la *conectividad* de la nube en términos de sus posibles *agujeros* en distintas dimensiones y conocer las relaciones entre ellos. El objetivo es asignar a la nube de datos un conjunto de invariantes algebraicos que describan sus propiedades geométricas y topológicas. Una manera usual de hacer esta descripción para espacios topológicos abstractos es usando teoría de homología; en particular, homología persistente permite extender esta noción sobre una nube de puntos, proporcionando así información cualitativa de la naturaleza de los mismos.

Para utilizar homología persistente en un conjunto de datos es indispensable poder calcular las distancias existentes entre ellos. De aquí en adelante, definimos una *nube de puntos* como un conjunto de datos que viene acompañada de una función distancia entre sus elementos. Es común, pero no necesario, que las nubes de puntos sean un subconjunto de $\mathbb{R}^d$ para algún $d > 0$. A continuación describimos brevemente el paradigma tradicional en una aplicación de homología persistente.

Consideremos la nube de puntos como una *muestra* de un espacio topológico subyacente, aunque posiblemente con ruido. Si nos interesa estudiar la naturaleza de los datos, es importante entender la estructura del *espacio muestral* subyacente. De manera natural surgen las siguientes preguntas: ¿es posible rescatar las propiedades de tal espacio a partir de la nube de puntos? ¿Cómo se aproxima un invariante topológico?

Una primera respuesta sería asociar a la nube de puntos una estructura topológica mediante la creación de conexiones entre los puntos *cercanos* y luego describir las propiedades del objeto resultante esperando que sean una buena aproximación a las características que presenta el espacio verdadero. Es fácil ver que el objeto resultante y sus propiedades dependen fuertemente de la definición de cercanía que consideremos. Así, la calidad de nuestra aproximación quedará determinada por el criterio que utilicemos para decidir cuándo dos puntos están suficientemente cercanos. La alternativa que propone la homología persistente es considerar distintas escalas de cercanía; observar la familia de espacios topológicos que se generan en el proceso y detectar las características topológicas que *persisten* a través de las diferentes escalas.

La estructura del presente capítulo es como sigue. En la primera sección describimos las estructuras topológicas que se asocian a los datos así como el lenguaje algebraico que se utiliza para describir las propiedades de tales estructuras. En la segunda sección, ofrecemos un breve resumen de los elementos de homología persistente necesarios para describir las propiedades topológicas de las nubes de puntos que estudiamos en esta tesis.

2.1. Homología simplicial

En esta sección recopilamos el material básico de la teoría de homología; sin embargo, las definiciones que utilizamos aquí difieren ligeramente de la presentación que usualmente se hace sobre el tema ya que no consideraremos coeficientes sobre un grupo abeliano arbitrario,en la práctica se usan típicamente $\mathbb{Z}$ o $\mathbb{Q}$, sino que trabajaremos sobre el campo $\mathbb{Z}_2$. Esta diferencia es importante ya que, al simplificar los cálculos, permite diseñar algoritmos para hacer el cálculo de la homogía persistente computacionalmente. Además, esto mismo nos permite ofrecer una exposición autocontenida dirigida a lectores sin experiencia en topología.

El presente apartado se basa en los libros de Munkres [51], Edelsbrunner y Harer [21] así como en el trabajo de Espinoza [28]. El libro de Giblin [34] es un buen punto de referencia para quienes deseen estudiar teoría de homología aun sin tener formación en topología algebraica.

2.1.1. Complejos simpliciales

Los espacios topológicos a estudiar son los complejos simpliciales. En topología llamamos complejos a los espacios que se pueden definir a partir de la composición de objetos sencillos. Para entender estas construcciones se usan principalmente dos enfoques: el punto de vista geométrico y la perspectiva combinatoria. En este trabajo utilizamos el primer enfoque aunque mencionando brevemente algunas equivalencias con la definición combinatoria. Detrás de la propiedad de los complejos simpliciales que permite ambos lenguajes está la idea intuitiva de que para construir un objeto de gran tamaño debemos considerar los objetos sencillos que lo componen. Esto se traduce matemáticamente a que la definición de las piezas que conforman un complejo debe ser topológicamente simple. Además, las reglas de *pegado* deben contemplar que las intersecciones comunes entre dos piezas producen una pieza de la misma clase pero de dimensión menor.

En la construcción geométrica de un complejo es indispensable describir primero los bloques constructores que llamamos en este caso simplejos. Para cada dimensión $k \geqslant 0$, la definición del k-simplejo nos remite al subconjunto convexo más pequeño, definido en $k+1$ puntos, que no está contenido en un hiperplano de dimensión menor a k. El camino para formalizar dicha noción empieza con las siguientes definiciones.

Definición 2.1 Sea $U = \{u_0, u_1, \ldots, u_k\}$ un conjunto finito de puntos en $\mathbb{R}^k$. Decimos entonces que $x \in \mathbb{R}^k$ es una *combinación convexa* de los elementos de U cuando existen $\lambda_i \geqslant 0$ tal que

i) $x = \sum_{1=0}^{k} \lambda_i u_i$,

ii) $\sum_{1=0}^{k} \lambda_i = 1.$

Llamamos *envolvente convexa* de U al conjunto de combinaciones convexas de elementos de U.

Definición 2.2 Los puntos $u_0, u_1, \ldots, u_m$ en $\mathbb{R}^k$ son *independientes* si $\{u_i - u_0 \mid 1 \leqslant i \leqslant m\}$ es un conjunto de vectores linealmente independientes.

Notemos que para que $n+1$ puntos sean independientes no es necesario establecer entre ellos un orden en particular, con lo cual la propiedad de independencia no depende de la elección de u_0.

Definición 2.3 Un *simplejo* $\sigma(u_0, u_1, \ldots, u_k)$ es el envolvente convexo de $u_0, u_1, \ldots, u_k$ puntos independientes. Los puntos que lo generan se llaman *vértices* del simplejo, mientras que la *dimensión* del simplejo viene dada por el número de vértices que lo definen.

En este trabajo, dado un simplejo σ de dimensión k, nos referiremos a él como un k-simplejo siempre que no sea necesario mencionar los puntos que lo definen. Consideremos $\sigma(u_0, u_1, \ldots, u_k)$ y un simplejo τ definido por un subconjunto de los vértices de σ. En este caso, decimos que τ es una *cara* de σ. Además, si el subconjunto que define a τ es propio, decimos que es una cara propia.

Como se explica más a detalle en el trabajo de Espinoza [28], dos simplejos cualquiera son homeomorfos si poseen la misma dimensión. Debido a lo anterior, geométricamente podemos considerar a los simplejos como deformaciones de *simplejos estándar*. Por ello es común realizar las construcciones que vienen más adelante con base en los *k-simplejos estándar*, que no son otra cosa que los definidos por el cero y los elementos de la base estándar de $\mathbb{R}^k$, $\sigma(0, e_1, \ldots, e_k)$.

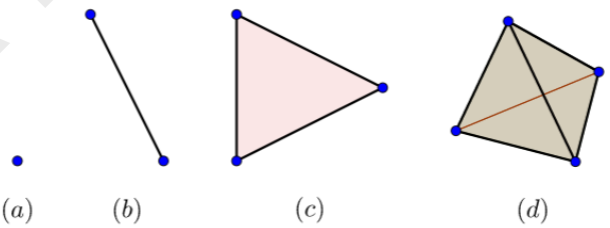


Figura 2.1: En las figuras $(a), (b), (c)$ y (d) se muestran simplejos de dimensión 0, 1, 2 y 3 respectivamente.

Las estructuras que definimos a continuación son ampliamente usadas en topología computacional, entre otras razones, porque permiten asociar

un espacio topológico a una nube de puntos dentro de un espacio métrico abstracto. Por otro lado, una definición en términos de simplejos es una manera eficiente de almacenar espacios topológicos en la computadora.

Definición 2.4 Un *complejo simplicial geométrico* K es un conjunto finito de simplejos tal que:

- Si $\sigma \in K$ y τ es una cara de σ, entonces $\tau \in K$.
- Si $\sigma_0, \sigma_1 \in K$, entonces $\sigma_0 \cap \sigma_1$ es vacía o una cara de ambos.

Para cada complejo simplicial K, su *dimensión* queda definida como la dimensión máxima que está presente en sus simplejos. Dado un complejo simplicial K si se tiene que un subconjunto $K' \subset K$ es un complejo simplicial por sí mismo, decimos que K' es un *subcomplejo* de K.

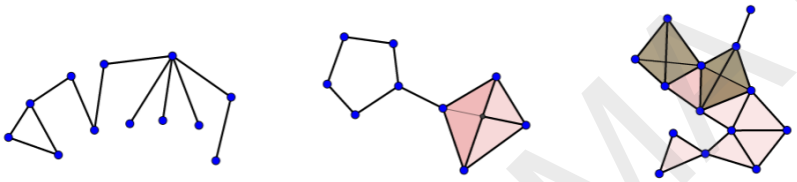


Figura 2.2: Las figuras $(a), (b)$ y (c) son representaciones de complejos simpliciales de dimensión 1, 2 y 3 respectivamente.

Entre los subcomplejos en los que centramos nuestro interés se encuentran los $j-$*esqueletos*. Dado un complejo simplicial K, se define su j-esqueleto como el conjunto de todos los simplejos de dimensión menor o igual a j; es decir, $K^{(j)} = \{\sigma \in K | \ \dim\sigma \leqslant j\}$. Al $0-$esqueleto se le denomina también conjunto de vértices del complejo.

Observemos que si $i \leqslant j$, entonces $K^{(i)}$ es subcomplejo de $K^{(j)}$. Así, la familia de los j-esqueletos nos permite *construir* el complejo K empezando con el conjunto de vértices y añadiendo en cada paso los simplejos de la dimensión correspondiente. La noción general sobre cómo podemos descomponer un complejo simplicial K en una familia de subcomplejos que nos permita *visualizar* su construcción queda formalizada de la siguiente manera.

Definición 2.5 Sea K un complejo simplicial, se define una *filtración* $\mathfrak{F}$ del complejo como una secuencia de subcomplejos $K_0, K_1, \ldots, K_n$ tal que

$$\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots \subseteq K_n = K.$$

Sin embargo, para poder estudiar las relaciones entre los complejos simpliciales como espacios topológicos, necesitamos construir funciones continuas entre ellas. Dicha tarea se simplifica si consideramos la naturaleza combinatoria que está implícita en la definición de los complejos y la cual nos permite determinar los llamados *mapeos simpliciales*. Se trata de funciones que pueden definirse únicamente en los 0-simplejos y que envían simplejos en simplejos.

Lema 2.1 Consideremos dos complejos simpliciales, K y L, así como una función $f : K^{(0)} \to L^{(0)}$ con la propiedad de que para cualquier subconjunto de vértices $V = \{v_0, \ldots, v_n\}$ que defina un simplejo en K se tiene que los vértices $f(v_0), \ldots, f(v_n)$ definen a su vez un simplejo en L. Entonces se puede extender f a una función continua $g : K \to L$ de tal manera que

$$x = \sum_{1=0}^{k} \lambda_i v_i \qquad \Rightarrow \qquad g(x) = \sum_{1=0}^{k} \lambda_i f(v_i).$$

De este modo decimos que la función g es el *mapeo simplicial* inducido por la función entre los vértices f.

Un primer ejemplo de mapeos simpliciales son las inclusiones naturales entre los subcomplejos de una filtración. Es fácil ver que la composición de mapeos simpliciales es a su vez un mapeo simplicial. Además, el mapeo simplicial inducido por f es un homeomorfismo cuando el mapeo simplicial $f : K^{(0)} \to L^{(0)}$ es una biyección entre los conjuntos de vértices de tal manera que cada subconjunto de vértices $v_0, \ldots, v_n$ en $K^{(0)}$ define un simplejo si, y sólo si, $f(v_0), \ldots, f(v_n)$ define un simplejo en L. En este caso, decimos que f induce un *homeomorfismo simplicial* o *isomorfismo* entre K y L.

Trabajar con complejos simpliciales considerándolos como una unión de conjuntos particulares dentro de $\mathbb{R}^n$ suele ser complicado. Como veremos más adelante, es posible definir un complejo simplicial en abstracto sin preocuparse por su realización geométrica ya que ésta siempre existe. Este cambio de objetos geométricos a un lenguaje combinatorio resulta muy útil para caracterizar distintos objetos como complejos simpliciales.

Definición 2.6 Un *complejo simplicial abstracto* es una colección finita de conjuntos $\mathcal{K}$ tal que para todo $\sigma \in \mathcal{K}$ y $\tau \subset \sigma$ se tiene que $\tau \in \mathcal{K}$.

Es claro que si tenemos un complejo simplicial K *geométrico* y entendemos por V el conjunto de vértices de K, podemos asociar a cada simplejo el conjunto de vértices que lo definen mientras que la colección de subconjuntos de V resultante, define a su vez un complejo simplicial *abstracto* $\mathcal{K}$. Debido a ello, los conjuntos en $\mathcal{K}$ son llamados *simplejos* y su *dimensión* queda definida como $\dim \sigma = |\sigma| - 1$. De manera análoga, la dimensión del complejo $\mathcal{K}$ es la dimensión máxima presente en sus simplejos. Decimos que $\mathcal{K}$ es un *esquema de vértices* de K mientras que K es una *realización geométrica* de $\mathcal{K}$.

No obstante, en la práctica ambas definiciones son intercambiables debido al siguiente resultado.

LEMA 2.2 Todo complejo simplicial abstracto de dimensión d tiene una realización geométrica en $\mathbb{R}^{2d+1}$.

El LEMA 2.2 forma parte de una extensa teoría que se ha desarrollado alrededor de los complejos simpliciales. Un estudio general sobre el tema puede realizarse a partir de los libros de topología algebraica de Hatcher [37] y Munkres [51]. Un abordaje más especializado puede hallarse en el libro de Koslov [44].

Recordemos que nos interesa utilizar los complejos simpliciales como una herramienta para analizar datos. Ahora bien, existen varias maneras de asociar un complejo simplicial a una nube de puntos. El artículo panorámico de Zomorodian [65] de 2012, hace un repaso de las estrategias más usadas. Entre ellas se encuentran los complejos simpliciales Vietoris-Rips, los cuales son muy populares en el análisis de datos debido a la sencillez de su definición e implementación.

Los complejos simpliciales Vietoris-Rips se utilizan en muchas aplicaciones—entre ellas las que estudiaremos en capítulos posteriores—ya que para su construcción no es necesario partir de un espacio métrico; basta con tener una función que se comporte de manera similar. Es común que la nube de datos venga acompañada de una función que permite cuantificar la *semejanza*, *cercanía* o *similitud* entre cada par de puntos. Idealmente nos gustaría que el par formado por la nube de puntos y su función constituyeran un espacio métrico. Sin embargo, dependiendo del contexto, las funciones que miden la *distancia* entre los datos no siempre constituyen una métrica en el sentido matemático. Por otro lado, para definir los complejos Vietoris-Rips necesitamos considerar los grafos de ϵ-vecindad de una nube de datos determinados por la función que mide su cercanía.

DEFINICIÓN 2.7 Consideremos un par (M, d) donde M es un conjunto finito y d es una función $d : M \times M \to \mathbb{R}$ tal que $d(x, y) \geqslant 0$ para todo $x, y \in M$ y $d(x, y) = 0$ si, y sólo si, $x = y$. Para cada $\epsilon \geqslant 0$, el *grafo de ϵ-vecindad* de M, $G_\epsilon(M, E_\epsilon)$, surge como resultado de considerar a M un conjunto de vértices definiendo el conjunto de aristas de la siguiente manera,

$$E_\epsilon = \{(x, y) \mid d(x, y) \leqslant \epsilon,\ x \neq y\}.$$

Recordemos que un clique en un grafo $G(V, E)$ es un subconjunto de vértices que inducen un subgrafo completo en G. Definimos que un clique es máximo si al añadir cualquier otro vértice al subconjunto, éste deja de ser clique. El complejo clique de un grafo—también llamado complejo bandera—se determina al definir el 0—esqueleto del complejo como el conjunto de vértices del grafo y tras añadir el correspondiente k-simplejo por cada clique

en k puntos. De hecho, los complejos Vietoris-Rips son un caso particular de los complejos cliques.

Definición 2.8 Consideremos un par (M, d) en el cual M es un conjunto finito y d es una función $d : M \times M \to \mathbb{R}$ tal que $d(x, y) \geqslant 0$ para todo $x, y \in M$ y $d(x, y) = 0$ si, y sólo si, $x = y$. Sea $r \geqslant 0$, entonces el *complejo simplicial Vietoris-Rips* $V(M, r)$ es el complejo clique del grafo de r-vecindad de M. Es decir, un subconjunto $\{u_1, \ldots, u_k\} \subset M$ define un k-simplejo en $V(M, r)$ si, y sólo si, el conjunto de nodos $\{u_1, \ldots, u_k\}$ define un clique en el grafo de r-vecindad de (M, d).

Podemos ver que si un conjunto de k nodos define un clique en un grafo, todos sus subconjuntos son a su vez cliques. Por lo tanto, los complejos Vietoris-Rips quedan bien definidos.

También podemos notar que si $r' < r$ entonces existe una inclusión $\mathcal{V}(\mathcal{M}, r') \hookrightarrow \mathcal{V}(\mathcal{M}, r)$ que es la identidad en el conjunto de vértices. Esto nos proporciona una manera natural de definir filtraciones de complejos Vietoris-Rips y son justo estas construcciones las que utilizamos en las aplicaciones del siguiente capítulo.

Las propiedades topológicas de una filtración se describen mediante *homología persistente*. Pero para entender este concepto es necesario comprender primero la descripción de las propiedades topológicas de un complejo mediante *homología simplicial*, lo cual haremos en la siguiente sección.

2.1.2. Grupos de homología

Una cuestión central en la topología es tratar de responder si dos espacios son homeomorfos. Para saber si se da el caso, basta con construir el homeomorfismo correspondiente. En cambio, determinar que dos espacios *no* son homeomorfos resulta un poco más complicado. El método usual es encontrar una propiedad topológica (sea una propiedad invariante bajo homeomorfismos) que se cumpla sólo en uno de los dos espacios. La teoría de homología provee el lenguaje matemático que permite describir los agujeros de un espacio topológico como invariantes topológicos. En particular, los grupos de homología simplicial están diseñados para caracterizar los complejos simpliciales de esta manera.

Históricamente, la homología simplicial fue uno de los primeros esfuerzos dentro de la teoría de homología y como tal, tiene alcances limitados en comparación con otras generalizaciones. En los libros de Maunder [47] y Munkres [51] podemos ver un panorama general sobre la teoría de la homología. Sin embargo, como se menciona en [66], la homología simplicial resulta ser un lenguaje adecuado para las aplicaciones de topología computacional. A continuación presentamos los elementos necesarios para poder definir la homología simplicial y, más adelante, extender dicha noción a la homología

persistente.

Definición 2.9 Sea K un complejo simplicial de dimensión n, y sea α_p el número de p-simplejos de K. Para cada $0 \leqslant p \leqslant n$, consideremos $\sigma_p^1, \sigma_p^2, \ldots, \sigma_p^{\alpha_p}$ los p-simplejos de K. Una p-cadena de K, es entonces una combinación lineal de la forma

$$\lambda_1 \sigma_p^1 + \lambda_2 \sigma_p^2 + \cdots + \lambda_{\alpha_p} \sigma_p^{\alpha_p}$$

y en la cual los coeficientes λ's son elementos de $\mathbb{Z}_2$.

Hacemos hincapié en que los coeficientes pertenecen al campo $\mathbb{Z}_2$, puesto que en general suele definirse la homología simplicial otorgando una orientación a las cadenas; esto se refleja en los coeficientes, que deben obtenerse de otros anillos conmutativos con unidad, típicamente $\mathbb{Z}$. Es por ello que en la literatura la homología simplicial viene descrita comúnmente en esos términos. Remitimos al Capítulo 3 del libro de Kozlov [44] para un enfoque más completo del tema. En cuanto a la tesis se refiere, es suficiente considerar coeficientes en $\mathbb{Z}_2$ pues, entre otras cosas, esto nos permite concebir las p-cadenas como conjuntos de p-simplejos en los cuales los p-simplejos incluidos en el subconjunto están señalados con coeficientes 1.

Por otro lado, podemos definir la suma entre cadenas de la misma forma que una suma entre polinomios, es decir, sumando los coeficientes correspondientes a los mismos simplejos. Si procedemos así, la suma de dos p-cadenas resulta ser la suma de los p-simplejos donde las dos cadenas originales difieren. De hecho, esta suma queda bien definida y le da una estructura de grupo a las p-cadenas en el complejo.

Proposición 2.1 Sea K un complejo simplicial y sea $C_p(K)$ el conjunto de p-cadenas de K. Tenemos que $C_p(K)$ es un grupo abeliano al que llamaremos el *grupo de p-cadenas de K*.

Cuando el contexto es claro, entonces $C_p(K)$ suele describirse únicamente como C_p. Dado un complejo simplicial K, para todo $p < 0$ y $p > \dim(K)$, definimos $C_p = \{0\}$. De esta manera C_p queda definido para todo $p \in \mathbb{Z}$. Este es el primer paso para «traducir» una pregunta *geométrica* al lenguaje algebraico. Recordemos que el propósito de la homología es describir cómo está conectado un espacio, es decir, poder describir sus *agujeros*. Pero no todos los agujeros son iguales. No es lo mismo la cerradura de una puerta que un globo inflado y sin embargo ambos tienen *huecos*. Surge así la necesidad de describir los huecos de manera formal y establecer sus distintas clases.

Si se tiene una superficie con un agujero basta dibujar un círculo a su alrededor para señalar "*aquí hay un agujero*". Igual, para distinguir las componentes conexas de un espacio basta considerar un punto como representante de cada una de ellas. En cambio, un punto o un círculo son insuficientes para

señalar el hueco en un globo inflado. Para englobarlo se necesita precisamente la *película* de una burbuja. Esto es porque, intuitivamente, los agujeros descritos son de diferente *dimensión* y por lo tanto requieren objetos de distinta dimensión para caracterizarlos.

Regresando a los complejos simpliciales, los huecos de dimensión p en un complejo simplicial K son los que *detectamos* y *caracterizamos* por una $(p-1)$-cadena. Para determinar formalmente cuándo una p-cadena detecta un agujero del simplejo se requieren las siguientes definiciones.

Definición 2.10 La *frontera* $\partial_p\sigma$ de un p-simplejo σ es la $(p-1)$-cadena conformada por la suma de las $(p-1)$-caras de σ. Para una p-cadena c, definimos su frontera como la suma de las fronteras de los simplejos en ella

$$\partial_p c = \sum \lambda_i \partial_p \sigma_i.$$

Esta definición induce un operador $\partial_p : C_p \to C_{p-1}$ que a cada p-cadena c le asigna su frontera; a esta función la llamamos *mapa frontera* de nivel p. Es claro que, por construcción, los mapas frontera son homomorfismos de grupos.

Definición 2.11 El conjunto de grupos de p-cadenas asociadas a un complejo simplicial K, junto con los homomorfismos frontera que los conectan, constituyen el *complejo de cadenas* de K. Usualmente estos se denotan como $\mathcal{C}(K) = (C_p(K), \partial_C)$.

Como se muestra en el Apéndice A, la construcción de los complejos de cadenas se puede realizar de manera meramente algebraica. En este caso, o cuando no es necesario señalar el complejo simplicial que la define, es común denotar una cadena compleja $\mathcal{C}$ como $\mathcal{C} = (C_p, \partial_C)$.

Las siguientes definiciones están asociadas a un complejo de cadenas $\mathcal{C} = (C_p, \partial_C)$.

Definición 2.12 Un p-ciclo es una p-cadena c tal que $\partial c = 0$ mientras que Z_p es el conjunto de p-ciclos.

Definición 2.13 Una p-frontera es una p-cadena c para la cual existe una $(p+1)$-cadena d tal que $\partial d = c$. Al conjunto de p-fronteras lo denotaremos como B_p.

Observemos que tanto Z_p como B_p son subgrupos de C_p; uno, por ser el *kernel* de ∂_p y el otro, por ser la imagen de ∂_{p+1}. La relación fundamental entre estos subgrupos queda al descubierto a través del lema fundamental de homología:

Lema 2.3 Lema fundamental de homología Para todo entero p y toda $(p+1)$-cadena d se satisface que

$$\partial_p \partial_{p+1} d = 0.$$

Esto significa que B_p es a su vez un subgrupo de Z_p, lo cual nos permite considerar el cociente entre ambos grupos y por consiguiente, definir los grupos de homología de un complejo simplicial.

Definición 2.14 El *p-ésimo grupo de homología* H_p de un complejo simplicial K es el grupo cociente del grupo de p-ciclos módulo del grupo de p-fronteras

$$H_p = Z_p/B_p.$$

El *p-ésimo número de Betti* es el rango de este grupo; $\beta_p = \text{rank}(H_p)$.

Nos referiremos a los elementos de H_p como las *clases de homología* de dimensión p. En ese sentido, decimos que dos ciclos son *homólogos* si están en la misma clase de homología.

Observación 2.1 Si se tiene un mapeo simplicial $f : K \to K'$, existe un homomorfismo inducido de manera canónica entre los grupos de homología

$$f_* : H_*(K) \to H_*(K').$$

Es importante resaltar que $(id_K)_* = id_{H_*(K)}$ para todo complejo simplicial K. Además, si se tiene otro mapeo simplicial $g : K' \to K''$, entonces $(g \circ f)_* = g_* \circ f_*$.

Para fines de este trabajo, como una introducción intuitiva a la teoría de homología recomendamos el trabajo de Espinoza [28]. En cambio, para un estudio más general del tema recomendamos los libros de Hatcher [37] y Munkres [51].

Por último, la siguiente definición distingue—en lenguaje de homología—los complejos simpliciales conexos que no presentan agujeros y que por lo tanto son los más simples en términos de conectividad.

Definición 2.15 Decimos que un complejo simplicial X es *acíclico* si

$$H_i(X) = \{0\} \qquad \forall i \geqslant 1.$$

Estos espacios nos servirán como punto de referencia para clasificar los espacios métricos en las aplicaciones del próximo capítulo. En particular, usaremos el siguiente criterio para detectar complejos simpliciales acíclicos.

Proposición 2.2 Sea X un complejo simplicial con dos subcomplejos U y V tales que $X = V \cup V$ y $U \cap V$ sea no vacía. Si U, V y $U \cap V$ son conexos y acíclicos, entonces X también es conexo y acíclico.

La demostración de esta proposición es consecuencia inmediata del teorema de sucesiones exactas Mayer-Vietoris. Tanto la demostración de la proposición como los resultados previos que se necesitan se encuentran en el Apéndice A.

2.2. Homología persistente

2.2.1. Introducción

Si seguimos el razonamiento que expusimos en la introducción del capítulo, podríamos intentar describir una nube de datos $\mathcal{M}$ asociándole un complejo $\mathcal{V}(\mathcal{M}, r)$ para alguna $r > 0$. Sin embargo, no siempre se conoce qué elección de r reflejaría de manera más apropiada las características del espacio subyacente. Como alternativa, se puede estudiar cómo cambian las propiedades topológicas al variar la distancia considerada. Esto es, definir una secuencia de números positivos $0 < r_1 < r_2 < \cdots < r_n$ y estudiar la filtración de complejos $\{\mathcal{V}(\mathcal{M}, r_i)\}_{i=1}^{n}$ resultante.

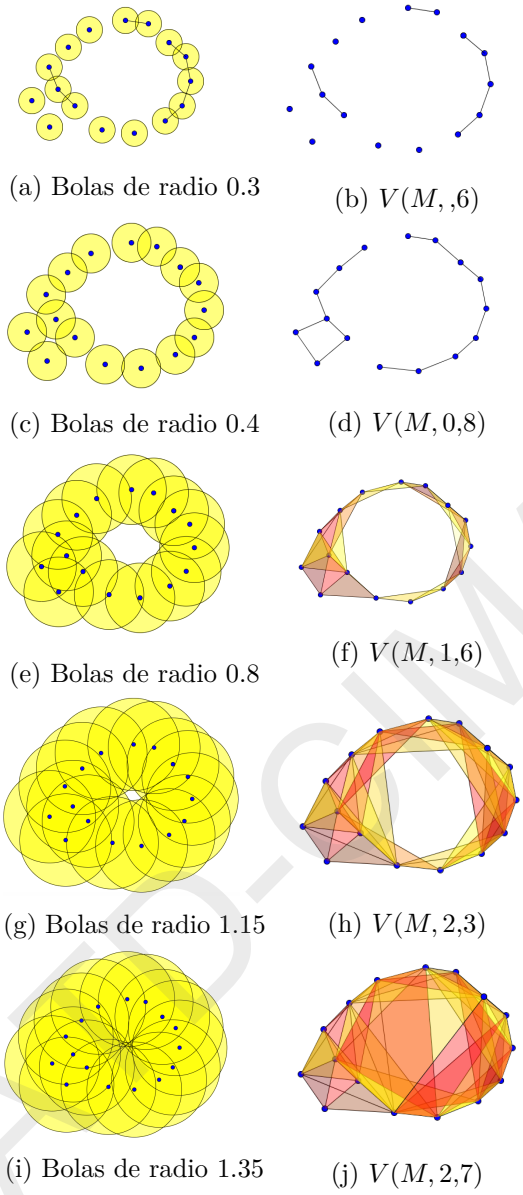
Observemos que en los distintos pasos de la filtración, mientras se añaden los simplejos, se pueden generar nuevos agujeros o incluso desaparecer los ya existentes. La homología persistente es lo que usamos para describir este proceso y con ella buscamos no sólo capturar los agujeros, sino saber cuánto tiempo *persiste* cada uno. Esto es válido tanto para las filtraciones continuas de un espacio topológico arbitrario, como para las filtraciones discretas asociadas a un espacio finito. En ese sentido, la homología persistente es para las filtraciones de un complejo lo que la homología simplicial es para los complejos simpliciales. El paradigma principal en varias aplicaciones de homología persistente es que las propiedades verdaderas del espacio subyacente son las que sobreviven más tiempo.

Tenemos ya el lenguaje de homología requerido para construir el concepto de homología persistente. Seguiremos el enfoque presentado en los libros [21] y [66] que ya mencionamos antes. Además, algunos de los elementos que aquí tratamos también se encuentran en los artículos panorámicos de Zomorodian [65], Carlsson [9] y Ghrist[33]. Recomendamos su lectura para obtener una visión general de la homología persistente.

2.2.2. Grupos de homología persistente

El concepto de homología persistente se puede definir para filtraciones de un complejo simplicial en general. Sin embargo, para los fines de este trabajo es suficiente considerar una familia particular de filtraciones sobre complejos simpliciales Vietoris-Rips, y por ello establecemos las definiciones adaptándolas a nuestras necesidades. A menos que se indique lo contrario, todas las filtraciones que consideramos se construyen tomando un espacio métrico $\mathcal{M}$ junto con una secuencia de números positivos $0 < r_1 < r_2 < \cdots < r_n$ que definen una filtración $\mathcal{F} = \{K_i\}_{1=0}^{n}$ donde $K_i = \mathcal{V}(\mathcal{M}, r_i)$. Diremos así que el *nivel* K_i de la filtración queda definida por el *tiempo* r_i.

Consideremos pues una filtración $\mathcal{F}$ de complejos $\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots \subseteq K_n = K$, así como sus correspondientes grupos de homología $H_p(K_0)$, $H_p(K_1), \ldots, H_p(K_n)$ para cada dimensión p. Recordemos que si $i \leqslant j$, entonces $K_i \subseteq K_j$ y por lo tanto la inclusión natural induce una



(a) Bolas de radio 0.3 (b) $V(M,,6)$

(c) Bolas de radio 0.4 (d) $V(M,0{,}8)$

(e) Bolas de radio 0.8 (f) $V(M,1{,}6)$

(g) Bolas de radio 1.15 (h) $V(M,2{,}3)$

(i) Bolas de radio 1.35 (j) $V(M,2{,}7)$

Figura 2.3: Filtración de complejos Vietoris-Rips sobre una nube de puntos. Por claridad, incluimos la nube de puntos dibujando bolas de un radio dado alrededor de cada punto. Observemos que a cada complejo $V(M,r)$, le asociamos las bolas con radio $r/2$, lo cual facilita distinguir si un segmento entre dos puntos debe dibujarse en el complejo, ya que corresponde a que las bolas entradas en tales puntos se intersectan.

función entre los grupos de homología correspondientes,

$$f_p^{i,j} : H_p(K_i) \to H_p(K_j) \text{ para cada } p \in \mathbb{Z}.$$

Esto es todo lo que necesitamos para definir los grupos de homología persistente.

Definición 2.16 Sea $\mathcal{F}$ una filtración. El (i,j)-*ésimo grupo de homología persistente de nivel* p de la filtración es la imagen ddel homomorfismo inducido por la inclusión entre los grupos de homología correspondientes,

$$H_p^{i,j} = \text{Im}(f_p^{i,j}).$$

Ahora bien, el correspondiente *número de Betti persistente* queda definido como $\beta_p^{i,j} = \dim(H_p^{i,j})$.

Esta definición podría parecer oscura pero provee un lenguaje apropiado para describir cómo van cambiando las propiedades topológicas del complejo K conforme se produce su *construcción* mediante filtración. En particular, para cada dimensión p, será útil poner atención en la siguiente secuencia:

$$0 = H_p(K_0) \to H_p(K_1) \to \cdots \to H_p(K_n) = H_p(K).$$

En esta secuencia podemos observar cómo las clases de homología van pasando de un nivel de filtración al siguiente con las inclusiones que abarcan desde el nivel i hasta el $i+1$. Nos interesa descubrir cuándo una clase aparece, cuánto tiempo sobrevive y cuándo desaparece durante el proceso. Diremos que una clase aparece en el tiempo i si el agujero que representa aparece en el complejo por primera vez en ese tiempo i. Si en algún nivel j posterior de la filtración tapamos dicho agujero con simplejos, éste desaparece en el nivel j. Formalizamos estas intuiciones en la siguiente definición.

Definición 2.17 Sea $(K_i)_{i=0}^n$ una filtración y $f_p^{i,i+1} : H_p(K_i) \to H_p(K_{i+1})$ para cada $p \in \mathbb{Z}$ la función inducida por la inclusión natural entre los complejos. Consideremos además una clase de homología $[\alpha] \in H_p(K_i)$ que no sea $[0]$, se tienen también las siguientes definiciones:

- Decimos que $[\alpha]$ *nace* en el tiempo r_i, o bien en el nivel i de la filtración, siempre que $[\alpha] \notin H_p^{i-1,i}$.
- $[\alpha]$ *muere* en el tiempo r_j,o bien en el nivel j de la filtración para algún $j \geqslant i$, si
$$f_p^{j-1,j}([\alpha]) \in H_p^{i-1,j}.$$
- Finalmente, si $[\alpha]$ muere al tiempo j, para algún $j \geqslant i$, definimos su *persistencia* como $r_j - r_i$ mientras que su *índice de persistencia* es la diferencia $j - i$. Para las clases que no mueren, la persistencia queda definida como ∞.

La definición del tiempo de muerte de una clase de homología contempla no sólo el caso en el que un agujero desaparece del complejo; también ofrece un criterio para unificar las clases de homología que representan agujeros distintos que se unen en un cierto nivel de la filtración. Si se presentan dos clases de homología $[\alpha], [\beta]$ distintas al tiempo $j-1$ tales que

$$f_p^{j-1,j}([\alpha]) = f_p^{j-1,j}([\beta]),$$

entonces es claro que ambas representaron todo el tiempo al mismo agujero del complejo K. En esta situación, una de las clases muere, pero deja a la más antigua como representante. Esta *regla de antigüedad* es necesaria para que la persistencia de una clase quede bien definida. A las clases de homología $[\alpha]$ que en su tiempo de muerte ésta ocurre porque $f_p^{j-1,j}([\alpha]) = [0]$ y a las que nunca mueren las denominamos clases *independientes*. Esta etiqueta refleja el hecho de que los agujeros que son detectados por clases no independientes son también detectados por una, y sólo una, clase independiente. Una explicación detallada sobre estas convenciones y observaciones se encuentran en el trabajo de Espinoza [28].

De esta manera podemos observar el crecimiento de un complejo simplicial a través de una filtración, y mediante homología persistente, estudiar la *evolución* de las clases de homología independientes que aparecen durante el proceso. Enseguida describimos las estrategias que se utilizan para presentar de una manera gráfica la información que se obtiene al utilizar la homología persistente.

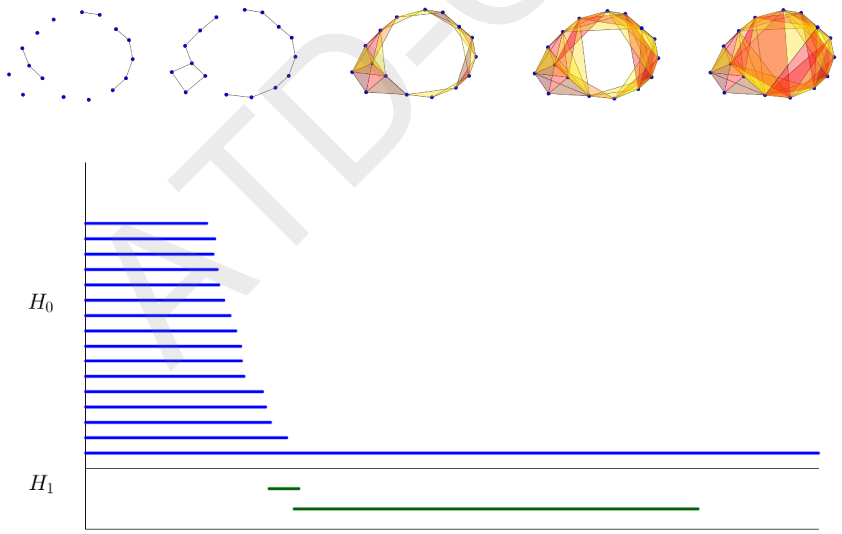


Figura 2.4: Código de barras de persistencia asociado a la filtración que se muestra en la Figura 2.3.

2.2.3. Códigos de barras y diagramas de persistencia

Notemos que a todas las clases de homología linealmente independientes, denotemos por $[\alpha]$ a una de ellas, le podemos asignar un par de puntos ordenados $(t_1(\alpha), t_2(\alpha))$ de manera tal que $t_1(\alpha)$ sea el tiempo de nacimiento de la clase y $t_2(\alpha)$ el tiempo de muerte que puede ser ∞ [2]. Así, para cada clase de nivel p se puede dibujar un segmento de línea horizontal en el cuadrante positivo que empiece en $t_1(\alpha)$ y termine en $t_2(\alpha)$. Cuando el segmento de recta vaya hasta a infinito esto se indica con una línea más larga que el valor asociado al último nivel de la filtración. Al dibujo resultante lo llamaremos el *código de barras de persistencia* de nivel p.

Si queremos obtener el número de Betti de persistencia $\beta_p^{i,j}$ basta contar las líneas que cubren el segmento que va de i a j en el correspondiente código de barras. Existe otra manera de representar la información contenida en un código de barras. Esta alternativa consiste en un gráfico contenido en el primer cuadrante de $\mathbb{R}^2$ y se llama *diagrama de persistencia*.

El número de clases que nacen exactamente al tiempo i y mueren al tiempo j queda representado por $\mu_p^{i,j}$. Nuevamente, para las clases que no mueren, hacemos $j = \infty$. El diagrama de persistencia de nivel p se construye como sigue: tomamos el cuadrante positivo en $\mathbb{R} \times (\mathbb{R} \cup \infty)$; después seleccionamos los puntos (i, j) tales que $\mu_p^{i,j} > 0$ y los marcamos asignándoles una multiplicidad $\mu_p^{i,j}$. Por último, añadimos todos los puntos en la diagonal pero asignándoles multiplicidad infinita. Este último detalle técnico permite definir una distancia entre diagramas de persistencia, punto que abordaremos más adelante.

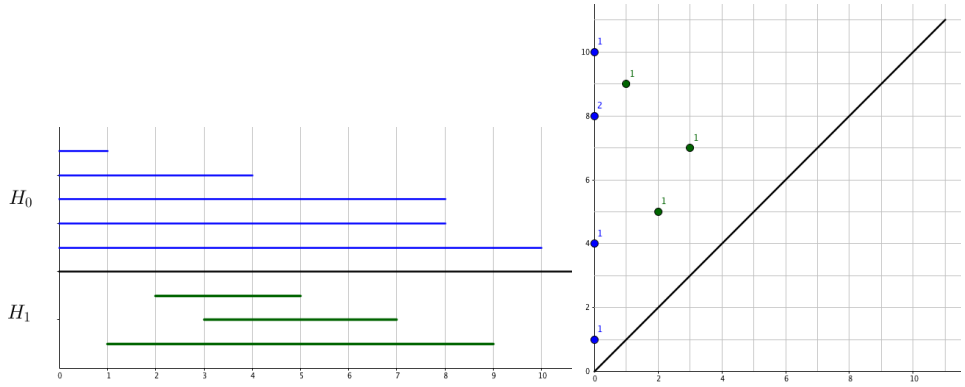
Para recuperar $\beta_p^{i,j}$ del diagrama de persistencia de nivel p, basta contar los puntos, con sus multiplicidades, en un determinado cuadrante. Específicamente

$$\beta_p^{i,j} = \sum_{k \leqslant i} \sum_{l > j} \mu_p^{k,l}.$$

Esta afirmación se basa en el siguiente lema, el cual justifica la construcción de dichos gráficos y nos dice que un diagrama de persistencia codifica toda la información necesaria para recuperar los grupos de persistencia de la filtración.

Lema 2.4 Lema Fundamental de Homología Persistente Sea $\emptyset = K_0 \subseteq K_1 \subseteq K_2 \subseteq K_3 \subseteq \cdots \subseteq K_n = K$ una filtración. Para cada par de índices $0 \leqslant i \leqslant j \leqslant n$ y cada dimensión p, el p-ésimo número de Betti

[2] Esta notación refleja que se puede definir homología persistente sobre filtraciones infinitas que se estabilizan. Esto es, que a partir de un valor finito la homología de la filtración no cambia, a las clases de homología que persisten de esta manera son a las que se les da un tiempo de muerte ∞. Teniendo esto presente, en este trabajo consideraremos que una clase tiene tiempo de muerte ∞ si ésta persiste, sin morir, en el último nivel de la filtración.



(a) Código de barras (b) Diagrama de persistencia.

Figura 2.5: Codigo de barras y diagrama de persistencia equivalentes. El número en cada punto del diagrama de persistencia indica la multiplicidad del mismo.

satisface la siguiente ecuación:

$$\beta_p^{i,j} = \sum_{k \leqslant i} \sum_{l > j} \mu_p^{k,l}.$$

Observemos que dado un diagrama de persistencia, a éste le corresponde un código de barras de manera única y viceversa. No entraremos en los detalles computacionales acerca del cálculo de homología persistente en casos concretos; basta decir que se han desarrollado algoritmos eficientes y robustos que reciben como entrada un espacio métrico finito, o una filtración, y a partir de ahí producen el diagrama de persistencia o código de barras respectivo. Más información acerca de los aspectos teóricos de estos algoritmos se puede consultar en [21] y en [66]. Existen distintas opciones de software para el cálculo de homologías basado en tales algoritmos. Entre las más populares se encuentran Javaplex [61], implementado en Matlab, o también Perseus [49] que fue desarrollado en C++.

2.2.4. Estabilidad

Como herramienta de análisis de datos, esperaríamos que la persistencia arrojara resultados "parecidos" cuando los datos de entrada son "similares". Por otro lado, nos interesaría que el método fuera estable frente a la presencia de ruido estadístico en los datos. Es decir, si una muestra proviene de datos con ruido, nos gustaría que dicho ruido no afectara la estructura general de los diagramas de persistencia asociados. Por eso buscamos que las clases de homología que persisten provengan de la estructura general de los datos y que no sean un producto del ruido. Asimismo, esperamos que pequeñas

diferencias en los datos impliquen sólo cambios pequeños en los diagramas de persistencia respectivos.

Es razonable pedir estas propiedades a la homología persistente puesto que el ruido en la muestra no debería alterar la estructura global de la nube de datos. El efecto que el ruido pueda tener sobre el tiempo de persistencia de las clases de homología debería ser pequeño dado que únicamente modifica cada punto de manera local. Las clases de homología que provienen del ruido persisten sólo localmente y por lo tanto, al observar un diagrama de persistencia, las clases que persisten durante un intervalo considerable de tiempo son en realidad un reflejo de la estructura de los datos. El enfoque de incluir el *ruido topológico* entre los puntos cercanos a la diagonal es común en las aplicaciones de homología persistente, como se muestra en [17, 32].

Todavía no especificamos a qué nos referimos cuando decimos que cambios pequeños en los datos resultan en cambios ligeros en los diagramas de persistencia. A fin de precisar esta propiedad, primero necesitamos formalizar el significado de las *nubes de datos similares* y los *diagramas de persistencia parecidos*. La noción de similaridad en ambos casos se establece formalmente definiendo una métrica sobre el espacio de diagramas de persistencia y otra sobre el espacio de espacios métricos finitos.

Ahora bien, existen varias formas de definir una métrica en el espacio de diagramas de persistencia. Para los fines de este trabajo es suficiente considerar la distancia llamada *cuello de botella*. Al establecer una métrica entre los diagramas de persistencia, nos gustaría que aquellos que son *visualmente parecidos* resulten también *matemáticamente cercanos*. Con esta motivación, buscamos cuantificar, intuitivamente, el esfuerzo requerido para transformar un diagrama en otro.

Recordemos que en la definición de diagrama de persistencia añadimos todos los puntos de la diagonal con multiplicidad infinita; por tanto, dos diagramas de persistencia cualquiera poseen la misma cardinalidad. Esto último nos permite considerar biyecciones entre cada par de diagramas. Dada una biyección ψ de un diagrama a otro, observamos cuánto *movemos* cada punto. Es decir, nos fijamos en $||x - \psi(x)||_\infty$. El valor máximo que alcanza esta expresión nos ofrece un modo de medir el efecto de ψ. Para definir la distancia cuello de botella tomamos el ínfimo de esta medida sobre todas las biyecciones que van de un diagrama a otro.

Definición 2.18 Sean X y Y dos diagramas de persistencia. Consideremos todas las biyecciones $f : X \to Y$ y observemos para cada una de ellas la diferencia máxima entre un punto así como su imagen mediante la norma L_∞. Al considerar el ínfimo de estas diferencias definimos la distancia de cuello de botella entre X y Y, $d_B(X, Y)$, como sigue

$$d_B(X, Y) = inf_{f:X\to Y}\{sup_{x\in X}||x - f(x)||_\infty\}.$$

Si M es el conjunto de los espacios métricos compactos módulo isometría, este conjunto forma un espacio métrico (M, d_{GH}) al definir la distancia *Gromov-Hausdorff*, d_{GH}, entre sus elementos. Por lo tanto, el conjunto de los espacios métricos finitos M^{fin} hereda esta estructura. Para definir dicha distancia recurrimos a la siguiente noción de correspondencia.

Definición 2.19 Sean (X, d_X) y (Y, d_Y) dos espacios métricos compactos. Dado un $\epsilon > 0$, una *ϵ-correspondencia* entre X y Y es un subconjunto $C \subset X \times Y$ tal que

i) Para todo $x \in X$ existe $y \in Y$ tal que $(x, y) \in C$.

ii) Para todo $y \in Y$ existe $x \in X$ tal que $(x, y) \in C$.

iii) Para todo $(x, y), (x', y') \in C$ se tiene que $|d_X(x, x') - d_Y(y, y')| \leqslant \epsilon$.

Definición 2.20 Dados los espacios métricos compactos (X, d_X) y (Y, d_Y) definimos la *distancia de Gromov-Hausdorff* entre ellos como sigue:

$$d_{GH}(X, Y) = \frac{1}{2} inf\{\epsilon \geqslant 0\text{: existe una } \epsilon\text{-correspondencia entre } X \text{ y } Y\}.$$

Para establecer el teorema que deja al descubierto la estabilidad de los diagramas de persistencia requiere que establezcamos un poco más de notación. Sabemos ya que a cada espacio métrico finito le podemos asignar de forma única un diagrama de persistencia de cada dimensión. Denotando por P a la familia de los diagramas de persistencia definamos pues a $F_k : \mathsf{M}^{fin} \to P$ como la función que asigna a cada espacio su diagrama de persistencia de nivel k.

Teorema 2.1 Estabilidad de los diagramas de persistencia Para dos espacios métricos finitos X y Y cualquiera, se tiene que

$$d_B(F_k(X), F_k(Y)) \leqslant d_{GH}(X, Y) \text{ para cada dimensión } k.$$

Para la demostración de este teorema—que aparece en el artículo [15] de 2009—es necesario considerar la *filtración estándar* en ambos espacios. La filtración estándar queda definida al considerar la familia de complejos $\mathcal{V}(\mathcal{M}, r_i)$ para cada $i \in (0, \infty^+)$. Los diagramas de persistencia de esta filtración son idénticos a la filtración discreta donde la distancia entre cada par de puntos del espacio $\mathcal{M}$ define un nivel.

Como ya mencionamos, las construcciones que aquí realizamos se pueden efectuar de manera más general. Es por eso que existen distintos enfoques para hablar de estabilidad en homología persistente. Recomendamos el artículo panorámico de Vejdemo-Johansson [62] para un estudio a fondo de este tema.

Capítulo 3

Modelo CCR para evolución viral

Introducción

Como se mencionó en la introducción de la tesis, la primera vez que apareció publicado un modelo de homología persistente aplicado a filogenética fue en 2013 con el artículo de Chan, Carlsson y Rabadan [14] y también en la tesis doctoral de Chan [13]. El objetivo central de su trabajo es analizar muestras de secuencias genéticas mediante homología persistente para obtener información sobre su filogenia. Con este mismo propósito, los autores proponen un marco conceptual, que aquí denominamos modelo CCR y el cual concretamos en el Postulado 3.1, para describir en un solo escenario los eventos evolutivos verticales y horizontales presentes en la evolución de la muestra viral estudiada.

Los autores explican que la elección de situar el modelo dentro de un contexto viral se debió a que se sabe que la presencia de eventos reticulares es común en algunas cepas virales como la influenza, además de que los genomas virales se encuentran ampliamente secuenciados y anotados. Esto último nos permite plantear las hipótesis principales del modelo dentro de un marco teórico específico. Sin embargo, los autores también prevén que el modelo puede ser utilizado para otros organismos como los eucariotas o las bacterias.

Dividimos la presentación de este capítulo en tres partes. En la primera, detallamos los preliminares matemáticos y el teorema que describe el resultado de analizar, mediante homología persistente, una muestra proveniente de una historia que no requiere eventos horizontales para ser explicada. En la segunda, describimos el modelo postulado junto con las estrategias que los autores sugieren para relacionar la información contenida en los códigos de barras con los eventos retículares presentes en la historia ded la muestra. En particular, presentamos los indicadores propuestos para resumir las carac-

terísticas más importantes del proceso evolutivo. Finalmente, en la tercera sección, hacemos comentarios puntuales sobre las características del modelo CCR que creemos necesario resaltar.

3.1. Homología persistente y espacios métricos aditivos

El objetivo de esta sección es describir el resultado que se obtiene al estudiar un espacio métrico aditivo mediante homología persistente. La exposición que aquí realizamos incluye los resultados de la sección titulada *Teoremas acerca de obstrucción topológica de filogenia* del material complementario contenido en el artículo [14]. La presentación de los autores culmina con la demostración del Teorema 3.1, cuya demostración difiere de la que aquí presentamos en detalles menores además de que exponemos una versión más detallada.

Abandonaremos por un momento el lenguaje de nubes de datos para considerarlas en su lugar como espacios métricos finitos. De esta forma podremos escribir de forma concisa las afirmaciones matemáticas que necesitamos sobre los complejos simpliciales asociados a estos espacios. La siguiente proposición resulta importante porque nos permite realizar tales afirmaciones mediante inducción sobre la cardinalidad de los espacios métricos.

Proposición 3.1 Sea M un espacio métrico finito y sean $m_0, m_1 \in M$ dos puntos en los que se alcanza la distancia máxima, es decir

$$d(m_0, m_1) \geqslant d(m, m') \qquad \forall m, m' \in M.$$

Definamos ahora $M_0 = M \backslash \{m_0\}$ y $M_1 = M \backslash \{m_1\}$ considerándolos espacios métricos. Entonces, para cada $r < d(m_0, m_1)$ se tiene que

$$V(M, r) = V(M_0, r) \cup V(M_1, r).$$

Además, $V(M_0, r) \cap V(M_1, r) = V(M_0 \cap M_1, r)$, donde $M_0 \cap M_1 = M \setminus \{m_0, m_1\}$.

Demostración

El resultado deriva fácilmente de la observación de que ningún simplejo en $V(M, r)$ contiene a la vez a m_0 y m_1 para todo $r < d(m_0, m_1)$.

□

Por otro lado, es fácil ver que si se tiene un espacio métrico finito M y una distancia $r > 0$ se puede definir una relación de equivalencia entre los puntos de M si éstos se encuentran en la misma componente conexa del grafo de r-vecindad de M. Esta relación se puede definir también sobre el grafo completo inducido por la matriz de distancias, en este caso se dice que dos puntos son equivalentes si existe un r-camino entre ellos. Utilizaremos esta

misma notación para refererirnos a la relación de equivalencia definida por r-caminos entre los nodos de $V(M,r)$ que corresponde, a su vez, a los nodos de las distintas componentes conexas. Es decir, la partición del espacio métrico M definida por la relación de equivalencia de r-caminos $\simeq_r$, $M = \bigsqcup_\alpha M_\alpha$, nos permite expresar el complejo $V(M,r)$ en términos de los complejos que constituyen cada una de sus componentes conexas:

$$V(M,r) = \coprod_\alpha V((M_\alpha, d_M|_{M_\alpha \times M_\alpha}), r).$$

Por otro lado, el siguiente teorema nos permite detectar las consecuencias de la aditividad de un espacio métrico M en el complejo $V(M,r)$ asociado.

Teorema 3.1 Sea M un espacio métrico aditivo y $r \geqslant 0$. Entonces el complejo simplicial $V(M,r)$ es una unión disjunta de complejos acíclicos y, por lo tanto,

$$H_i(M) = \{0\} \qquad \forall i \geqslant 1.$$

Demostración

Es claro que el resultado se cumple para todo r mayor o igual que el máximo de las distancias en M, pues el complejo resultante es homotópico a una bola de dimensión $|M| - 1$.

Para el caso en el que r sea menor que la distancia mayor, entonces procedemos por inducción sobre la cardinalidad de los espacios métricos. Consideraremos como hipótesis de inducción que la proposición es cierta para todo espacio métrico con cardinalidad menor a n. Para el caso base, $|M| = 2$, podemos ver que el resultado se cumple ya que en el complejo Vietoris-Rips $V(M,r)$ cada uno de los puntos constituye una componente conexa.

Vimos también que una manera de observar las componentes conexas de $V(M,r)$ es considerando la relación de equivalencia $\simeq_r$ dada por los r-caminos. De esta manera y utilizando la misma notación, si $\mathbf{C} = \{M_\alpha\}_{\alpha\in\Lambda}$ es la familia de clases de equivalencia, entonces tenemos que

$$V(M,r) = \coprod_{\alpha\in\Lambda} V((M_\alpha, d_M\,|_{M_\alpha \times M_\alpha}), r).$$

Paso 1. Inducción

Ahora bien, si $V(M,r)$ no es conexo, podemos afirmar que cada subconjunto en la partición tiene cardinalidad menor a la de M. Por ello, y debido a la Proposición 1.1, sabemos que cada uno de estos subespacios es a su vez un espacio métrico aditivo y, por lo tanto, el resultado se cumple mediante la hipótesis de inducción.

Supongamos entonces que $V(M,r)$ es conexo. Sean $m_0, m_1 \in M$ dos puntos que maximicen la distancia en M y consideremos como espacios

métricos a $M_0 = M \setminus \{m_0\}$ y $M_1 = M \setminus \{m_1\}$. Por la Proposición 3.1 sabemos que

$$V(M, r) = V(M_0, r) \cup V(M_1, r),$$

y que $V(M_0, r) \cap V(M_1, r) = V(M_0 \cap M_1, r)$, donde $M_0 \cap M_1 = M \setminus \{m_0, m_1\}$.

Aquí podemos ver que tanto M_0, M_1 como su intersección son espacios métricos aditivos con cardinalidad menor a la de M. Así, por la hipótesis de inducción, tenemos que $V(M_0, r), V(M_1, r)$ y $C(M_0 \cap M_1, r)$ constituyen una unión disjunta de complejos acíclicos. En conclusión, y partiendo de la Proposición 2.2, resulta suficiente si probamos que $V(M_0, r), V(M_1, r)$ y $V(M_0 \cap M_1, r)$ son conexos.

Paso 2. M_0 es conexo: notación adicional

Para probar que $V(M_0, r)$ es conexo basta con demostrar que existe un r-camino entre dos elementos cualquiera de M_0. Por el Lema 1.2 podemos suponer que M está incluido en $\mathcal{M}(T)$ para algún árbol aditivo T de tal manera que las hojas de T estén contenidas en la imagen de M. Aprovecharemos esta estructura para probar lo que queremos aunque ello requerirá introducir notación adicional.

Empecemos viendo que para cada hoja e en el árbol existe una única juntura $j(e)$ más cercana a ella, a menos que no existan junturas en lo absoluto. En este último caso, el árbol es una línea con nodos intermedios cuyo resultado se obtiene inmediatamente. Supongamos pues que existe al menos una juntura en T.

Para cada par (v, e) donde $v \in V_T$ y donde e es una arista de T que contiene a v, $Br(e, v)$ denotará al subárbol formado por los vértices $v' \in V_T$ tales que el camino mínimo de v a v' pase por e; además nos referiremos a él como la *rama de T que pasa por v y e*.

Consideremos ahora los nodos unidos a la juntura más cercana de m_0. Esto es, tomemos los nodos $\{v_0, v_1, \ldots, v_n\}$ tales que $(j(m_0), v_i) \in E_T$ y hagamos $\mathcal{B}_i = Br(j(m_0), (j(m_0), v_i))$ para cada uno de ellos. Sin pérdida de generalidad podemos suponer que $m_0 \in \mathcal{B}_0$ y que $m_1 \in \mathcal{B}_1$.

Observemos asimismo que:

$$d(v, j(m_0)) \leqslant d(m_0, j(m_0)) \qquad \forall v \in \mathcal{B}_i, \qquad i \geqslant 2.$$

De otro modo, se contradice la maximalidad de la distancia de m_0 a m_1. Como consecuencia de esta desigualdad tenemos que

$$d(w, v) \leqslant d(w, m_0) \qquad \text{donde } v \in \mathcal{B}_i, \quad i \geqslant 2 \text{ y } w \in \mathcal{B}_j, \quad j \geqslant 1.$$

En efecto, sean w y v tal y como los describimos antes y consideremos además el camino por T que va de w a v pasando por $j(m_0)$ y el cual se obtiene al concatenar los caminos mínimos que van de w a $j(m_0)$ y de $J(m_0)$ a v. Debido a que la distancia entre u y w queda definida por el largo del

camino mínimo entre ellos, es posible que el largo del camino que acabamos de construir sea mayor a esta distancia; esto es,

$$d(u, w) \leqslant d(w, j(m_0)) + d(j(m_0), v). \tag{3.1}$$

Utilizando la desigualdad anterior podemos concluir entonces que

$$d(u, w) \leqslant d(w, j(m_0)) + d(j(m_0), v) \leqslant d(w, j(m_0)) + d(j(m_0), m_0) = d(w, m_0).$$

Paso 3. M_0 es conexo por r-caminos

Probemos, ahora sí, que $V(M_0, r)$ es conexo por r-caminos. Sean $m, m' \in M_0$, all ser M r-conexo existe un r-camino $(x_0, x_1, \ldots, x_n)$ que los conecta. Si m_0 no se encuentra en dicho camino, entonces los puntos están conectados por un camino en $V(M_0, r)$. Supongamos pues que $x_i = m_0$ para algún i; a fin de simplificar el proceso podemos suponer que el camino pasa exactamente una vez por m_0. En particular tenemos que $d(x_{i-1}, m_0) \leqslant r$ y que $d(m_0, x_{i+1}) \leqslant r$. Ahora procederemos por casos.

Primero, supongamos que $M \cap (\mathcal{B}_0 \setminus \{m_0\})$ es no vacío. Sea $\overline{m}$ el punto más cercano a m_0, es claro entonces que reemplazando el segmento v_{i-1}, m_0, v_{i+1} por $v_{i-1}, \overline{m}, v_{i+1}$ obtenemos un r-camino que no pasa por m_0, que es precisamente lo que buscábamos.

En caso contrario, cuando $M \cap (\mathcal{B}_0 \setminus \{m_0\})$ es vacío, seleccionamos cualquier hoja $m^* \in \mathcal{B}_i$ para algún $i \geqslant 2$ y por la desigualdad 3.1 vemos que podemos reemplazar el segmento v_{i-1}, m_0, v_{i+1} por v_{i-1}, m^*, v_{i+1} obteniendo de este modo un r-camino completamente contenido en $V(M_0, r)$. Es así como concluimos que $V(M_0, r)$ es conexo por r-caminos.

El resultado para M_1 y $M_0 \cap M_1$ se consigue de igual forma puesto que podemos quitar las ocurrencias de m_0 y m_1 en los r-caminos de manera independiente, lo cual concluye la demostración.

□

La demostración que aquí presentamos sigue la lógica general del artículo, pero además llena los huecos que se dejan al lector y presenta un par de diferencias con respecto de la original. Específicamente, estas diferencias están en el planteamiento de la inducción y en la simplificación de los caminos que contienen a m_0.

Como se puede intuir naturalmente, homología persistente nos permite extender este resultado para describir las consecuencias de la aditividad de M en los complejos simpliciales que se le asocian, pero sin limitarnos a una r en particular.

Corolario 3.1 Dado un espacio métrico aditivo M, el diagrama de persistencia de nivel p de $V(M, r)$ es vacío para todo $p > 0$.

Es así como homología persistente nos permite discernir cuándo podemos utilizar un árbol filogenético para explicar el proceso evolutivo de una muestra. Cuando esto no es posible, el modelo que los autores postulan, y

que presentamos en la siguiente sección, utiliza la información contenida en los diagramas de persistencia para representar simultáneamente la evolución vertical y horizontal.

3.2. Postulación del modelo

El planteamiento del modelo consiste en utilizar homología persistente para estudiar la evolución de una muestra. Si al observar el diagrama de persistencia de algún nivel no cero, este resulta ser no vacío, entonces no le podemos asociar un árbol filogenético a la muestra. La hipótesis principal consiste en que la presencia de agujeros en la filtración es consecuencia directa de los eventos horizontales durante la evolución de la muestra.

Recordemos que podemos interpretar la distancia entre dos secuencias para obtener información acerca de qué tan cercanas son evolutivamente. Es por ello que decimos que las distancias tienen una dimensión temporal. Intuitivamente, secuencias cercanas corresponden a elementos de la muestra cercanos evolutivamente. Esto significa que el tiempo necesario para encontrar su ancestro común más reciente es relativamente corto. Teniendo de fondo el significado evolutivo de las distancias pretendemos capturar las características del proceso evolutivo de la muestra mediante homología persistente. En esta situación, el observar la homología de la filtración conforme crece el radio de vecindad, debería darnos una idea sobre cómo se ve el proceso al recorrer el tiempo hacia el pasado.

El objetivo del modelo no se limita a representar la evolución vertical y los eventos reticulares, sino que mediante homología persistente también buscamos determinar la tasa a la que estos eventos reticulares ocurren, detectar eventos complejos que involucran más de dos organismos e incluso dejar al descubierto patrones complejos de cosegregación. Para lograr esto debemos refinar la hipótesis inicial. Establezcamos nuevamente, ahora de manera formal, las hipótesis del modelo.

Postulado 3.1 (Modelo CCR en evolución viral)

Es posible estudiar las relaciones evolutivas entre los organismos de una muestra utilizando homología persistente en el espacio métrico que forma dicha muestra. El modelo postulado se basa en las siguientes hipótesis:

1. Evolución vertical y componentes conexas. Las componentes conexas del complejo en un nivel particular ϵ corresponden a las diferentes familias, o clases, presentes al tiempo representado en esa escala.

2. Evolución horizontal y agujeros de dimensión uno. Cada clase de homología de dimensión 1 corresponde a un evento reticular del proceso evolutivo.

3. Eventos reticulares complejos. Las clases de homología de dimensión mayor a 1 son el resultado de múltiples intercambios horizontales o eventos reticulares complejos que involucran múltiples secuencias parentales.

Partiendo de estas hipótesis, traduciremos el resumen topológico del proceso de persistencia en un resumen de las propiedades del proceso evolutivo. Esta es la principal diferencia con los métodos filogenéticos usuales, puesto que típicamente las estrategias que se usan, como las que mencionamos en el primer capítulo, proponen una única estructura candidata a explicar la historia de la muestra, sea árbol o red. En cambio, con homología persistente analizamos las características topológicas invariantes de las historias que podrían explicar el proceso evolutivo de la muestra en un intervalo de tiempo, es decir, un intervalo de distancias genéticas.

Es importante notar que este análisis se realiza sin imponer a los datos una estructura preconcebida o con características específicas, como podrían ocurrir con los árboles o las redes. Asimismo, por la naturaleza topológica del modelo, el análisis se comporta de manera estable ante pequeñas fluctuaciones en los datos.

Homología persistente	Evolución viral
Valor ϵ de la filtración.	Distancia genética (escala evolutiva).
Número de Betti de dimensión cero a un valor de la filtración ϵ.	Número de *clusters* a escala ϵ.
Generadores de las 0-clases de homología a una escala ϵ.	Elementos representativos de cada clase.
Relaciones jerárquicas entre los generadores de las 0-clases de homología a lo largo de la filtración.	*Clustering* jerárquico.
Número de Betti de dimensión 1.	Número de eventos horizontales (recombinación y reacomodo).
Homología de dimensión > 0 no trivial (TOP).	No existe un árbol filogenético que represente la muestra.
Número de clases de homología, de dimensión no cero, dividido por el tiempo.	Cota inferior a la tasa de eventos reticulares.

Cuadro 3.1: Diccionario de homología persistente-filogenética.

Por otro lado, los conceptos que surgen en el código de barras de persistencia de un conjunto de secuencias genéticas pueden ser utilizados para describir las características evolutivas de la historia de la muestra. En el Cua-

dro 3.1[1] se resumen las interpretaciones del código de barras de persistencia que el postulado implica en forma de diccionario. Entre las consecuencias del modelo, resultan de especial interés dos indicadores: la *obstrucción topológica de filogenia* (TOP) y la *tasa de ciclos irreducibles* (ICR). A continuación, describimos a detalle cada uno de ellos así como el resto de la terminología del modelo.

3.2.1. Obstrucción topológica de filogenia (TOP)

Como hemos dicho, el primer paso importante del modelo es usar la información de los diagramas de persistencia como criterio para discernir si podemos explicar la muestra sin eventos horizontales, es decir mediante un árbol; o si es necesario considerar intercambios horizontales para entender su historia. Sabemos que cuando las distancias entre los datos forman una *filogenia perfecta*[2] entonces el Teorema 3.1 nos permite asegurar la existencia de un árbol filogenético asociado a nuestra muestra.

Es poco común encontrar que los datos formen un espacio aditivo de manera exacta debido a la presencia de ruido proveniente de errores en la secuenciación, muestreos incompletos, etcétera. Puesto que los espacios aditivos generan filtraciones acíclicas, pequeñas perturbaciones en la aditividad generan únicamente barras pequeñas en el código de barras de persistencia. Por ello, cuando la muestra proviene de un espacio aditivo pero el ruido provoca la pérdida de la aditividad, la estabilidad que ofrece el Teorema 2.1 nos dice que el diagrama de persistencia resultante está muy cercano al diagrama de persistencia vacío. Con el objeto de separar los efectos artificiales creados durante la recolección y tratamiento de datos en una muestra proveniente de un proceso sin eventos reticulares, y al mismo tiempo medir qué tan aditivo es un espacio, los autores proponen el siguiente indicador.

Definición 3.1 Dado el código de barras de persistencia correspondiente a una muestra de secuencias genéticas definimos la *obstrucción topológica de filogenia* (TOP) de la muestra como el largo de la barra más grande correspondiente a una clase de homología de dimensión mayor a cero.

Con este indicador se pretende cuantificar qué tan cercano se encuentra de ser aditivo el espacio en cuestión. En otras palabras, el TOP correspondiente a una muestra que proviene de un espacio aditivo se puede expresar en términos de la distancia de cuello de botella entre el espacio original y el espacio que se generó al añadir el ruido.

Ahora bien, supongamos que tenemos una muestra M' que proviene de un espacio aditivo M pero que debido al ruido perdió la aditividad, con lo cual $TOP(M') > 0$. Sea (a_x, a_y) el punto en el diagrama de persistencia que

[1] Tabla 1 del artículo [14].

[2] Recordemos que una muestra forma una filogenia perfecta si la nube de datos constituye exactamente un espacio métrico aditivo.

define a $TOP(M')$. Es decir,

$$TOP(M') = a_y - a_x.$$

Es fácil ver que este mismo punto define la distancia de cuello de botella entre M y M' y que,

$$\begin{aligned} D_B(M, M') &= ||(a_x, a_y), (\frac{a_x + a_y}{2}, \frac{a_x + a_y}{2})||_\infty \\ &= \frac{a_y - a_x}{2} \\ &= \frac{TOP(M')}{2}. \end{aligned}$$

Así, por el Teorema 2.1 tenemos que

$$TOP(M') \leqslant 2D_{GH}(M, M').$$

Así, la obstrucción topológica de filogenia resulta ser una medida estable para las desviaciones de la aditividad. Los autores muestran la utilidad de este indicador mediante simulaciones y analizando datos reales provenientes de distintas familias de virus. Sustentado en simulaciones, Chan afirma que un $TOP > 4$, en distancia genética, es suficiente para asegurar estadísticamente la presencia de eventos horizontales en la muestra[3]. Al final de la siguiente sección mencionamos los resultados de las simulaciones que hicieron en [14] para mostrar el desempeño de los indicadores topológicos.

3.2.2. Estimadores topológicos de tasas de eventos reticulares

Una vez que se detectó la presencia de eventos reticulares mediante TOP y observando el código de barras de persistencia, podemos utilizar la estructura de dicho código para obtener más información acerca de la estructura de los eventos reticulares. En concreto, si utilizamos la dimensión temporal que representa la distancia, podemos dar un indicador del número mínimo de eventos reticulares que se necesitan para explicar la muestra.

Definición 3.2 La tasa de ciclos irreducibles de dimensión d, (ICR_d), de una muestra de secuencias genéticas es el número de ciclos independientes de dimensión d que aparecen en el código de barras normalizado por el tamaño de la ventana de tiempo observada. Esto es,

$$ICR_d = \frac{\text{Número total de barras de dimensión 1 para cualquier } \epsilon}{\text{Ventana de tiempo}}.$$

La ventana de tiempo que se considera en este caso está entre el 5 % de las secuencias más recientes y el 5 % de las secuencias más antiguas que se encuentran presentes en el conjunto de datos.

[3] Consultar las páginas 98, 116 y 217 en Chan [13].

Para evaluar la sensibilidad del método, los autores realizaron simulaciones que contemplaron cuatro escenarios diferentes: *i)* evolución sin eventos horizontales; *ii)* una población con un solo evento reticular y evolución con eventos horizontales; *iii)* un caso considerando reacomodos de secuencias y *iv)* otro con recombinación homóloga. Los detalles técnicos se encuentran en el apéndice de simulaciones del material complementario de [14]. Entre las observaciones importantes que los autores hacen destacan las siguientes:

i) Las clases de homología de dimensión mayor a cero aparecen cuando la tasa de eventos horizontales es mayor a cero.

ii) ICR_1 se incrementa de manera proporcional al aumento que se hace en la tasa de eventos horizontales.

iii) La presencia de eventos horizontales múltiples, sean reacomodos o recombinaciones, puede producir clases de homología de dimensión 2.

En la Figura 3.1[4] se resume el comportamiento de los indicadores topológicos ICR y TOP al analizar poblaciones simuladas bajo un modelo Wright-Fisher[5] con eventos horizontales así como con distintos valores en los parámetros de mutación y recombinación/ reacomodos. En este caso, los autores resaltan la estabilidad de TOP cuando el parámetro de eventos horizontales es 0.

En la Figura 3.2[6], los autores muestran el resultado de las simulaciones que llevaron al cabo para calibrar el desempeño del ICR y del TOP frente a otros indicadores de presencia de eventos horizontales usuales, los cuales se mencionaron en el Capítulo 1. En esta figura observamos cómo los estimadores propuestos por los autores superan notoriamente el desempeño de los otros que fueron considerados en el estudio.

Por último, en la Figura 3.3[7], se presenta el comportamiento de los estimadores topológicos de recombinación al ser utilizados en distintas cepas virales. Aquí queda al descubierto el contraste entre las familias virales conocidas por su abundante presencia de evolución horizontal y las cepas que se conocen por ser principalmente clonales.

[4]Imagen S5 del material complementario de [14].

[5]Los modelos Wright-Fisher son ampliamente usados para realizar simulaciones de evolución de poblaciones. Destacan entre los modelos de genética de poblaciones por su simplicidad y por estar estrechamente relacionados con la teoría de coalescencia. Para un estudio detallado de estos modelos recomendamos la lectura de [31, 30, 63]; para un ejemplo de cómo son utilizados para simulación, recomendamos el trabajo de Hudson [39].

[6]Imagen S7 del material complementario de [14].

[7]Imagen S10 del material complementario de [14].

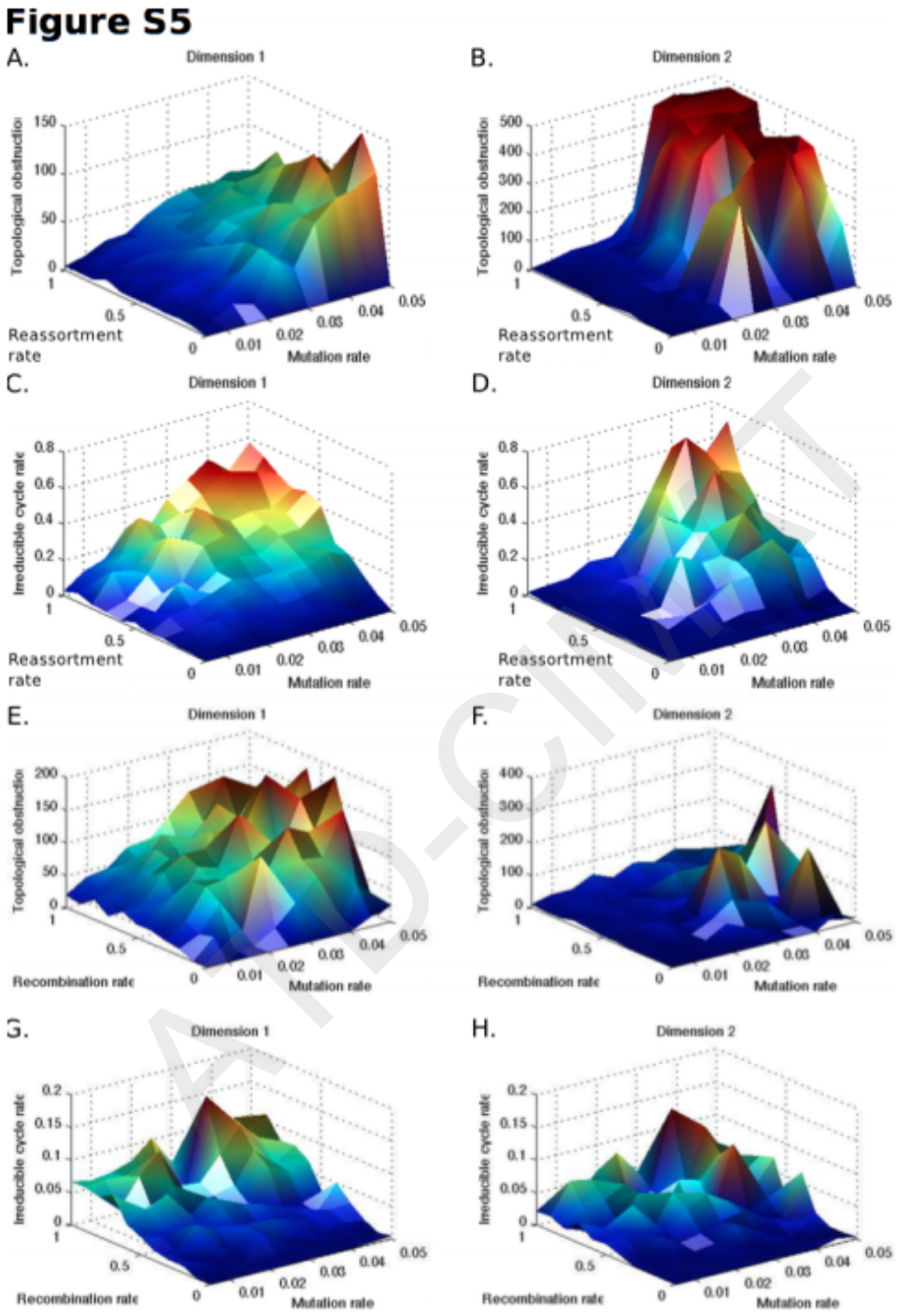


Figura 3.1: Comportamiento de TOP e ICR en simulaciones de poblaciones Wright-Fisher con recombinación al considerar diferentes valores en los parámetros de mutación y recombinación.

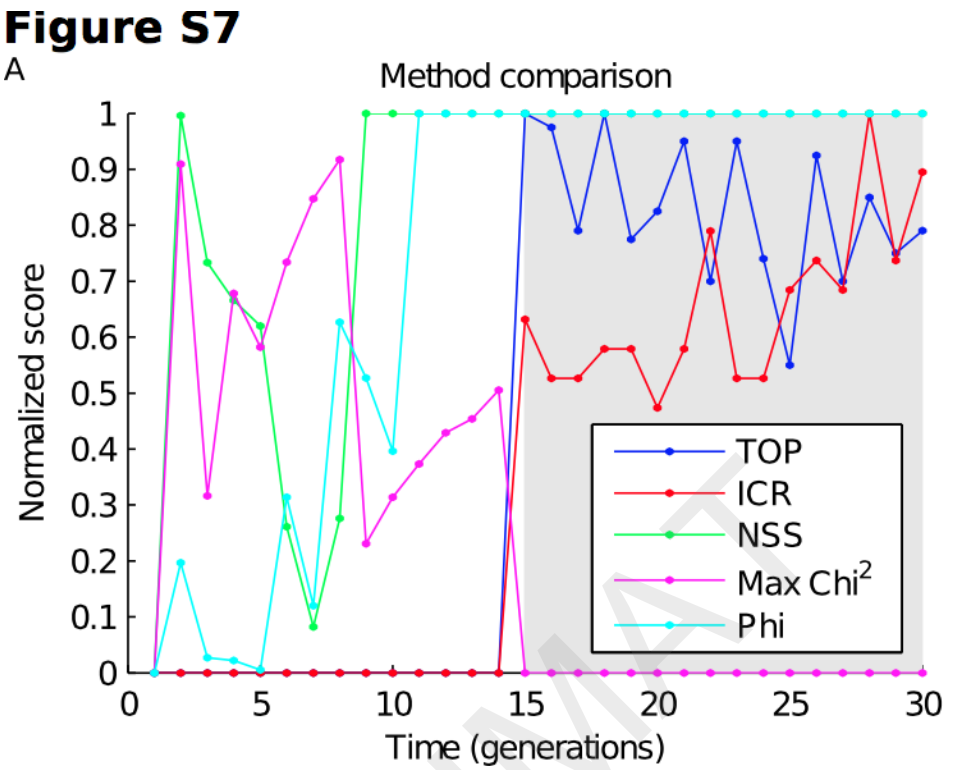


Figura 3.2: Desempeño de TOP e ICR frente a algunos indicadores usuales de presencia de eventos horizontales. Las simulaciones consisten en poblaciones Wright-Fisher con un único evento horizontal en la generación 15.

3.2.3. Eventos reticulares complejos y patrones de cosegregación

Recordemos que la cosegregación entre dos genes es la tendencia a ser intercambiados, o heredados, juntos. Los autores destacan que a pesar de que existen estudios filogenéticos previos que confirman un alta tasa de reacomodo entre las cepas de influenza aviar, ninguno ha identificado un patrón claro de segmentos de genes asociados y remiten a [19] para un contexto más amplio del problema. Asimismo, proponen un procedimiento para detectar cuándo dos segmentos tienden a cosegregar utilizando homología persistente.

Primero consideran todos los pares de segmentos concatenados para obtener un primer estimador del número de reacomodos dado por $b_1 = \beta_1$. Enseguida estiman el número de eventos horizontales considerando el $b_{1,i,j} = \beta_1'$, donde β_1' es el número de Betti persistente de dimensión 1 obtenido al comparar los segmentos i y j. Después, se estima la importancia de observar un número dado de reacomodos entre cada par de secuencias mediante un test binomial.

Denotemos por $p_{i,j}$ la probabilidad de que los segmentos i y j sean here-

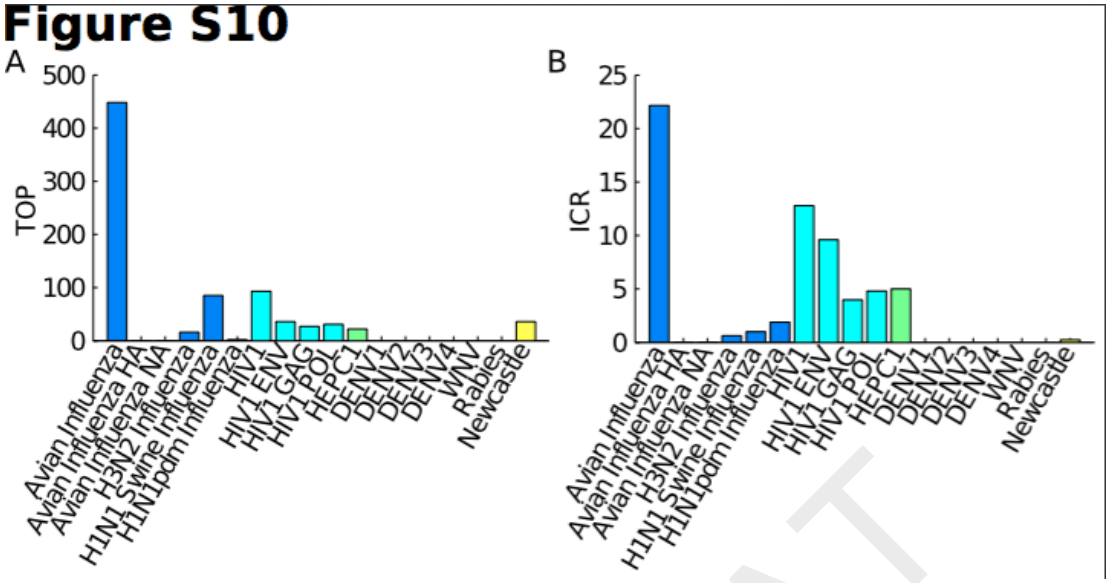


Figura 3.3: Estimadores topológicos de eventos horizontales de diferentes cepas virales

dados juntos en un evento horizontal. La hipótesis de que no existe cosegregación implicaría que $p_{i,j} = 1/2$. Ahora probamos la hipótesis, considerando que de un total de b_1 eventos horizontales los segmentos se heredaron juntos $b_{1,i,j}$ veces.

Los resultados en cepas de influenza aviar sugieren que hay cuatro segmentos concretos—PB2, PB1,PA y NP—que tienden a cosegregar juntos. Resumimos estas interacciones en la Figura 3.4 (D) en la cual el ancho de las aristas queda determinado por la probabilidad de los segmentos a cosegregar juntos.

Hasta ahora hemos mencionado únicamente las consecuencias relacionadas con las primeras dos hipótesis del Postulado 3.1. Para confrontar la tercera hipótesis frente a datos reales, los autores del artículo remiten a la cepa de influenza aviar H7N9 que ocasionó un brote epidémico en la provincia china de Jiangsu en abril de 2013, según se informa en [38]. Construyendo árboles asociados a cada gen y observando los conflictos entre las estructuras, Gao *et al.* [29] determinaron que este nuevo virus era producto de un triple reacomodo. También pudieron establecer que las secuencias parentales corresponden a cepas pertenecientes a los siguientes linajes:

- H7N3 A/ *duck* / Zhejiang /12 /2011,
- H7N9 A/ *wild bird* / Corea / A14 /2011,
- H9N2 A/ *brambling* /Beijing/ 16 / 2012.

Al observar el código de barras de los genomas concatenados, en una muestra que incluía las cepas involucradas, los autores identificaron una 2-clase de homología irreducible en la que participaban las cepas involucradas. Recordemos que las 2-clases de homología detectan las *cavidades* de dimensión 3. Así, las cepas involucradas corresponden a los vértices de un poliedro que persiste hueco un tiempo considerable durante el análisis; ver la imagen E de la Figura 3.4. De esta manera, homología persistente provee una representación gráfica que es a la vez sencilla e interpretable para un evento horizontal complejo.

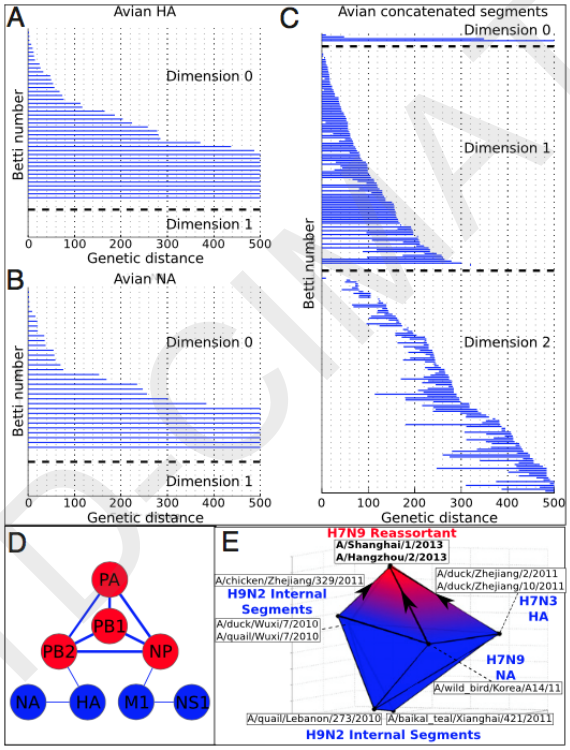


Figura 3.4: A y B son los códigos de barras obtenidos de los segmentos HA y NA por separado. C muestra los códigos de barras resultantes tras utilizar los genomas concatenados de cepas de influenza. D es el grafo que indica la predisposición a cosegregar entre segmentos específicos del genoma de cepas de influenza, mientras que E muestra la representación gráfica de los representantes del ciclo de homología irreducible en tres dimensiones asociado al evento reticular complejo de la cepa H7N9.

3.3. Comentarios finales

En resumen, la homología persistente permite modelar la historia evolutiva de una muestra en un marco de referencia que contempla tanto los eventos evolutivos verticales como los horizontales. Este enfoque permite además estudiar la escala y la estructura temporal de tales eventos. Como ya mostramos, se puede utilizar el diccionario del Cuadro 3.1 para traducir e intentar responder distintas preguntas evolutivas con lenguaje topológico. De esta manera, además de incluir conceptos estándar en filogenética, como la tasa de eventos horizontales, podemos utilizar este lenguaje para plantear una gama más amplia de preguntas. Ejemplo de esto último, es que los autores estudian distintas cepas de influenza para detectar patrones evolutivos como la cosegregación o eventos horizontales complejos.

El trabajo de Chan, Carlsson y Rabadan es una propuesta innovadora en filogenética ya que ofrece un enfoque alternativo a las estrategias usuales en el área. Creemos que la interacción entre análisis topológico de datos y filogenética en la que se basa la propuesta nos introduce a un nuevo marco teórico en el que se conjuntan temas de actualidad tanto de matemáticas como de biología. En nuestra opinión, el Teorema 3.1 y el Postulado 3.1, más que satisfacer la necesidad de modelar un fenómeno complejo, muestran el potencial que tiene dicho marco teórico para ser adaptado a modelar situaciones más concretas en filogenética. Además, eso requiere desarrollar herramientas matemáticas específicas. Como prueba de esto, durante los últimos dos años se ha publicado una variedad de artículos [8, 23, 24, 25, 54] que exploran diferentes vertientes del modelo CCR. Como dijimos en la introducción, el siguiente capítulo estudia a detalle uno de estos trabajos.

A continuación, enlistamos una serie de posibles líneas de investigación adicionales que se apoyan en las ideas presentes en el trabajo de Chan, Carlsson y Rabadan. Creemos que pueden ser un buen punto de partida para realizar estudios futuros del tema.

Comentario 3.1

Una pregunta natural es si el inverso del Teorema 3.1 es cierto y la respuesta resulta ser negativa. En discusiones sobre el tema con el profesor Frédéric Chazal, del INRIA-Francia, surgió el ejemplo de un espacio métrico no aditivo cuyos códigos de barras de persistencia son vacíos (ver la Figura 3.5). Si la aditividad no caracteriza los espacios métricos que dan origen a códigos de barras de persistencia vacíos, entonces ¿qué tienen en común estos espacios? ¿Qué propiedades comparten que los distinguen y que provoca que originen filtraciones acíclicas de complejos Vietoris-Rips?

Estas preguntas se relacionan con el estudio de los espacios métricos compactos en general. Pero además es necesario darles respuesta si se quiere utilizar el Teorema 3.1 como criterio de aditividad ya que es importante saber en qué casos el teorema arroja falsos positivos. Para el caso concreto de

esta aplicación que estamos estudiando, resulta importante conocer si estas propiedades tienen consecuencias puntuales en la interpretación evolutiva asociada a la muestra. De ser así, debemos ver si es aceptable concluir que no se necesitan eventos reticulares para explicar la historia de una muestra en tales casos.

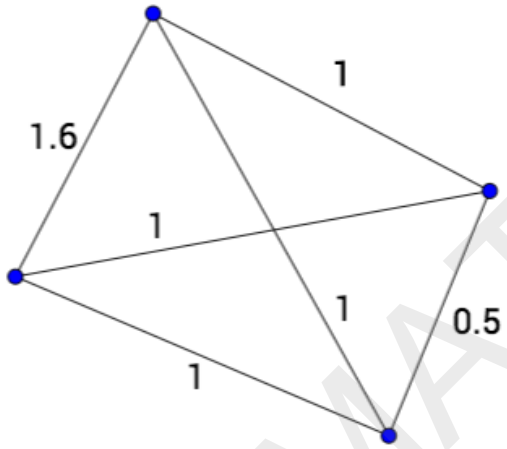


Figura 3.5: Espacio métrico no aditivo cuyos diagramas de persistencia de dimensión mayor a cero son vacíos.

Comentario 3.2

Las hipótesis del Postulado 3.1 aunque se inspiran en el Teorema 3.1 no están *unívocamente* determinadas por él. Es claro que se hace el postulado buscando aprovechar la mayor cantidad de información posible del código de barras de persistencia y con el fin de poder proponer un *diccionario* como el Cuadro 3.1. Sin embargo, queda fuera de nuestro alcance comprender si un enfoque así de general resulta ser el más adecuado. Es posible que con un modelo más puntual, aunque menos amplio, podamos responder preguntas particulares con mayor precisión. También es posible que sea necesario desarrollar más resultados matemáticos que acompañen al Teorema 3.1 y nos permitan precisar las interpretaciones que se pueden hacer del código de barras cuando éste resulta ser no vacío.

Comentario 3.3

Como los autores mencionan, el Postulado 3.1 no está diseñado solamente para modelar evolución viral. ¿Qué otras poblaciones pueden ser modeladas así? Un punto de partida importante de investigación es estudiar la manera de adaptar estas hipótesis a contextos no virales. Hasta donde sabemos, la tercera hipótesis del postulado, que se refiere a eventos reticulares

complejos, parece ser necesaria sólo en el caso viral. En cambio, las primeras dos hipótesis pueden ser fácilmente utilizables con otro tipo de organismos.

Creemos que es necesario un estudio detallado de modelación matemática sobre los distintos eventos reticulares para saber si es necesario modificar las hipótesis ya consideradas o incluso plantear modelos con hipótesis adicionales.

Asimismo, con el planteamiento general del Postulado 3.1 podemos, en principio, adaptar las ideas para estudiar la evolución de organismos a distintas *escalas*. Es decir, el análisis puede realizarse con organismos dentro de una misma especie para estudiar eventos reticulares concretos o bien, eventos de mezcla de poblaciones sobre una muestra de distintas especies dentro de una familia particular con el fin de inferir las relaciones entre ellos; más aún, podría ser utilizado para estudiar el árbol universal de la vida.

Comentario 3.4

Como se mencionó en el primer capítulo, existen distintas distancias genéticas que suponen diferentes escenarios evolutivos. En su tesis, Chan [13] explica brevemente que el desempeño del método no cambia sustancialmente ante las diferentes distancias genéticas. Entre otras cosas, queda fuera del alcance de esta tesis de licenciatura determinar cuáles criterios deberíamos usar para elegir con qué métrica queremos aplicar homología persistente.

Creemos que esta decisión debe tomarse cuidadosamente pues hasta donde sabemos la escala de tiempo puede variar detrás de cada distancia genética. Por lo tanto, si hay que proponer indicadores que incluyan una dimensión temporal, como el *ICR*, la interpretación *temporal* que se quiera hacer de los códigos de barras no será necesariamente igual en cada caso.

Por último, ademas de que se requiere precisar cómo deben elegirse las ventanas de tiempo que se están representando es importante precisar el significado de la topología de la muestra en escalas concretas. Como ejemplos, ¿cómo interpretar la distancia genética en la que la filtración se vuelve conexa?o ¿qué significa que la filtración se vuelva conexa pero no acíclica?

Comentario 3.5

Requeriríamos más formación en filogenética para determinar si el método puede ser utilizado de la misma manera con una muestra que consta de organismos *contemporáneos* o si incluso puede incluir organismos de diferentes generaciones sin necesidad de cambios en la metodología.

Si todos los organismos pertenecen a un mismo tiempo, es claro que la observación de cómo cambia la estructura de la filtración se puede interpretar como una observación del *tiempo hacia atrás*. Lo que no es claro es si la misma interpretación resulta admisible o si es necesario hacer consideraciones adicionales en el caso de que la muestra inicial contenga organismos

de distintos tiempos -pueden ser ancestros de otros organismos de la muestra o simplemente muestras de organismos que no pertenecen a una misma escala de tiempo en términos de generaciones. Entre otras cosas, creemos que es necesario precisar esto para poder definir de manera concreta cuándo podemos obtener indicadores con una dimensión temporal, como el *ICR*.

Comentario 3.6

Como herramienta filogenética, la homología persistente podría clasificarse dentro de los métodos por distancias dado que la entrada del algoritmo es precisamente una matriz de disimilitud entre las secuencias. Los datos, al ser secuencias de caracteres, tienen propiedades intrínsecas que también pueden aportar información sobre la historia evolutiva de la muestra. Sería interesante ver si es posible aprovechar la estructura propia de cada uno de los datos para obtener información sobre la filogenia utilizando análisis topológico de datos.

En ese mismo sentido, la aplicación que aquí estudiamos se basa en complejos simpliciales Vietoris-Rips, que es una técnica usual en análisis topológico de datos pero existen otras herramientas dentro del área que pueden ser utilizadas. Resultaría interesante estudiar, por ejemplo, cómo asociar complejos simpliciales a datos genéticos de tal manera que se aproveche su estructura de secuencias de caracteres. Esto requeriría plantear nuevamente el Teorema 3.1 en términos de los complejos simpliciales correspondientes.

Comentario 3.7

El uso que aquí se hace de la homología persistente como herramienta filogenética es un esfuerzo aislado en el área en el sentido de que no incorpora técnicas estudiadas de manera general en filogenética. ¿Cómo tender puentes entre homología persistente y el lenguaje de filogenética tradicional?

Creemos que sería interesante ver cómo pueden explotarse las ventajas de homología persistente y de los métodos filogenéticos usuales de manera conjunta. Por ejemplo, el método que aquí estudiamos no se enfoca en proporcionar una estructura candidata para explicar la filogenia de la muestra; sin embargo, sería útil saber bajo qué condiciones se puede aprovechar la filtración que se construye durante el análisis para proporcionar una red filogenética que explique los datos y que sea compatible con la descripción que se hace de su historia en el código de barras de persistencia.

Capítulo 4

Distintivos topológicos para mezclas de poblaciones

Introducción

En este capítulo estudiamos una estrategia diseñada para caracterizar eventos de mezclas de poblaciones mediante ciertos *distintivos topológicos* que se obtienen al utilizar homología persistente. El material que aquí exponemos sigue las ideas contenidas en el artículo de 2015 de Parida *et al.* [54]. En su trabajo—que se basa en las ideas de Chan, Carlsson y Rabadan y que nosotros resumimos ya en el Postulado 3.1—los autores proponen un modelo que permite considerar los eventos horizontales entre individuos y además representar las interacciones horizontales a nivel de poblaciones; todo esto, utilizando homología persistente.

A diferencia de otros trabajos que también parten del modelo CCR y que mencionamos ya en la introducción general, Parida *et al.* [54] adecuan el modelo para proponer una interpretación más sutil de los códigos de barras sustentando esta modificación en resultados matemáticos concretos. Sin embargo, los resultados matemáticos que aquí presentamos para fundamentar tal interpretación difieren de los incluidos en el artículo [54]. Como se explica al inicio de la Sección 4.2, estos cambios se hicieron como consecuencia del intercambio de ideas con uno de sus autores.

Por otro lado, la pregunta que estudiamos aquí difiere principalmente de lo presentado en el capítulo anterior, desde el punto de vista biológico, en que ahora vamos a estudiar individuos diploides; esto es, individuos con dos ancestros directos. Por lo tanto, los eventos horizontales corresponden a sucesos en los que el material genético heredado de ambos padres se combina durante el proceso. En concreto, vamos a enfocarnos en el caso de que la muestra esté conformada por individuos que provienen de distintas poblaciones. Esto es, contamos con información adicional que nos permite agrupar los individuos en *clusters* o poblaciones y estamos interesados en describir

las relaciones que han existido entre estas poblaciones a lo largo del tiempo.

Recordemos que al considerar una muestra de individuos, podemos asociarle una red filogenética explícita que explique su historia. En este capítulo trabajaremos en particular con el concepto de *grafo de recombinación ancestral* o ARG por sus siglas en inglés. En el caso de reproducción asexual, los ARG son grafos dirigidos en los que cada nodo corresponde a un organismo de la filogenia y la dirección indica la relación de parentesco entre ellos. En un ARG, los nodos que tienen grado de entrada 2, representan organismos producto de un evento reticular entre los dos nodos parentales. En el caso de reproducción sexual, seguiremos la definición que se da en [53] sobre un ARG en este contexto. El concepto de ARG apareció por primera vez en el trabajo de Griffiths y Marjoram [35] y es parte fundamental de la teoría de coalescencia con recombinación. A su vez, esta teoría es uno de los pilares fundamentales de la genética de poblaciones. Para un estudio detallado del tema recomendamos el libro de Wakeley [63].

Dentro de este ámbito en particular, existen diferentes técnicas para asignar un ARG a una muestra de manera que explique su historia. Ejemplos de ello se pueden encontrar en [4, 57]. Pero si además los individuos de la muestra provienen de distintas poblaciones, nos interesaría conocer cómo fue la evolución de las poblaciones a lo largo del tiempo. Debido a que las relaciones evolutivas entre poblaciones corresponden a interacciones entre individuos concretos, es natural preguntar si es posible identificar los eventos evolutivos a nivel de poblaciones observando los eventos evolutivos a nivel de individuos; es decir, en el ARG asociado a la muestra.

A su vez, podemos clasificar las relaciones evolutivas en poblaciones en verticales u horizontales. Decimos que sucedió un *evento vertical* a nivel de poblaciones si por alguna razón, una población se divide en dos subconjuntos que continúan evolucionando por separado. Asimismo, un *evento horizontal* entre poblaciones corresponde a un evento de mezcla de dos poblaciones pudiéndose tratar de una mezcla completa o una migración parcial de una población a otra donde se genera descendencia conjunta.

Para estudiar las interacciones entre poblaciones a lo largo del tiempo nos gustaría utilizar un grafo que explique la historia evolutiva de las poblaciones de manera análoga a como un ARG explica la historia evolutiva de una muestra de individuos. En el trabajo que estudiamos en este capítulo se considera que los *armazones por poblaciones de un ARG* son estructuras que cumplen este rol. Ahora bien, el objetivo principal sería describir las propiedades topológicas de tales armazones sin tener que construir los ARG de manera explícita.

Dividimos el capítulo en tres partes. Primero, en la Sección 4.1, presentamos las definiciones y el lenguaje requerido para plantear formalmente la situación que estamos estudiando. A continuación, en la Sección 4.2 damos las definiciones necesarias para concluir con un teorema que nos permite interpretar el código de barras relacionándolo con los eventos horizontales

entre poblaciones. Por último, en la Sección 4.3 mostramos la interpretación que los autores hacen de los códigos de persistencia y la cual se refiere precisamente a los *distintivos topológicos para mezcla de poblaciones*.

4.1. Mezcla de poblaciones y ARG

Como mencionamos en la introducción, supondremos que las organismos que vamos a estudiar son diploides y que es posible describir su filogenia mediante un ARG. Simplificando un poco, plantearemos el problema en términos del ARG de la muestra; esto es, supondremos por un momento que conocemos el proceso evolutivo de la muestra y por ende conocemos también el ARG que describe las relaciones hereditarias verdaderas entre las generaciones.

Precisemos ahora algo de notación adicional sobre ARG pues la necesitaremos más adelante. En un ARG, tanto las aristas como los nodos parentales se pueden etiquetar con la porción del material genético que transmiten. Vamos a concretar cuál *distancia evolutiva* queremos representar con la matriz de disimilitud de la muestra. Numeremos la *profundidad* de las generaciones considerando el tiempo hacia atrás de tal manera que la generación al tiempo actual, es decir de la que tomamos la muestra, sea la generación 0, mientras que la generación de sus padres tiene profundidad 1 y así sucesivamente. Ahora podemos definir una distancia entre dos elementos de la generación cero, como la profundidad de la generación en la que encontramos su primer ancestro común.

Esta distancia entre individuos nos servirá más adelante para inducir una noción de distancia evolutiva en el conjunto de las *etiquetas de poblaciones*. Aunque el concepto de etiquetar por poblaciones sea claro, usaremos la siguiente definición para establecer la notación que vamos a utilizar más adelante.

Definición 4.1 Sea $f : C \to P$ una función entre dos conjuntos finitos; si f es sobreyectiva, decimos que es una *función de etiquetado* de C. Vamos a denominar a P como el *espacio de etiquetas* de C definido por f. Decimos también que C es un conjunto etiquetado en $|P|$ poblaciones.

Notemos que si una muestra está etiquetada en m poblaciones, esperaríamos que las distancias entre elementos de una misma población fueran pequeñas en comparación con las de los elementos etiquetados en poblaciones diferentes. Así, al observar las distancias entre los individuos, éstas tendrían que estar separadas en *dos escalas* de forma coherente con el etiquetado. Estableceremos más adelante a qué nos referimos cuando decimos que un etiquetado satisface estas propiedades.

4.1.1. Armazón de un ARG estructurado en poblaciones

Ahora bien, si sabemos que los elementos de la muestra están compuestos por organismos de distintas poblaciones y además suponemos que durante el proceso evolutivo, la muestra estuvo estructurada en poblaciones bien definidas, es natural preguntarse cómo fueron las relaciones entre las poblaciones a lo largo del tiempo. Dicho de otro modo, si M' es un espacio de etiquetas de una muestra M, nos gustaría asociar a M' una estructura que describa su proceso evolutivo de la misma manera que el ARG explica la evolución de M. Con esta motivación definimos el armazón de un ARG estructurado en poblaciones.

Definición 4.2

Sea M una muestra de individuos etiquetados en m poblaciones y sea P un ARG que explica la historia de la muestra.

Sea P' un grafo dirigido con m hojas en el cual todos los nodos internos son parte de exactamente tres aristas. A cada arista $e \in E_{P'}$ le asignamos las siguientes cantidades:

- Largo de incubación $len(e)$.
- Número de linajes en la base de la arista $l_b(e)$.
- Número de linajes al final de la arista $l_f(e)$.

Aquí, el largo de una arista es una unidad de tiempo que quedará definida por el número de generaciones. Pero la asignación de estos números a cada arista debe satisfacer las siguientes propiedades:

(a) Si un nodo $v \in V_{P'}$ tiene grado de entrada 2, supongamos que las aristas e_1 y e_2 son las que llegan a él y sea e_3 la arista que sale de él. Entonces, debe satisfacerse que:

$$l_b(e_3) \leqslant l_f(e_1) + l_f(e_2).$$

(b) Si un nodo $v \in V_{P'}$ tiene grado de salida 2, supongamos que las aristas e_1 y e_2 salen de él, y sea e_3 la arista que llega a él. Entonces, debe satisfacerse que

$$l_f(e_3) \leqslant l_b(e_1) + l_b(e_2).$$

A cada arista $e \in E_{p'}$ le asignamos la evolución de una población Wright-Fisher P_e durante $len(e)$. La población P_e debe empezar con $l_b(e)$ linajes, y terminar con $l_f(e)$ linajes.

Si podemos *pegar* las poblaciones correspondientes en los extremos de las aristas de tal manera que

$$P = \cup_{e \in E_{P'}} P_e,$$

entonces decimos que P' es un *armazón por poblaciones* de P.

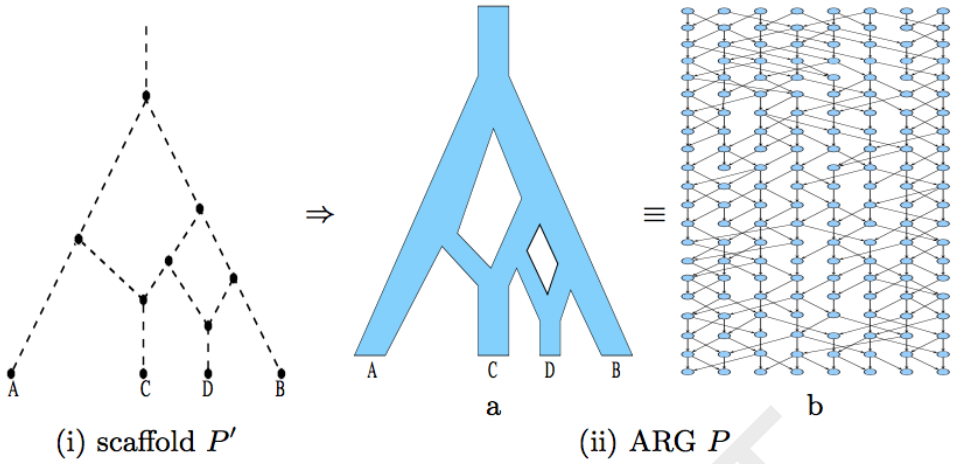


Figura 4.1: Ejemplo de un armazón en 4 poblaciones de un ARG tomado del artículo original [54].

Consideremos M como una muestra de individuos etiquetados en m poblaciones cuya filogenia se explica en un ARG P y un armazón P'. Si se observa el proceso evolutivo de M, no sólo se está viendo a P sino que se está observando a P *montado* sobre el armazón P'.

Es importante resaltar que si se encuentra un camino cerrado en P', éste corresponde a un evento de mezcla de poblaciones. En cambio, los nodos en P' con grado de salida 2 indican que una población se separó en dos grupos que continuaron evolucionando por separado. Es justo para lograr esta interpretación que se piden las desigualdades en (a) y (b) que acabamos de mostrar en la definición anterior.

4.1.2. Planteamiento del problema

Cuando desconocemos el proceso evolutivo de una muestra de individuos, lo más que podemos hacer es asociarle un ARG que explique su historia. De la misma manera, si no conocemos la historia de una muestra estructurada en poblaciones, nos gustaría asociarle un ARG—y un armazón por poblaciones a este último—para poder explicar satisfactoriamente tanto la historia de la muestra como la de las poblaciones.

Observemos que dado un ARG P asociado a una muestra que evolucionó estructurada en poblaciones, el armazón P' no es calculable de forma inmediata, debido a que los caminos cerrados en P significan eventos horizontales entre individuos, pero éstos no inducen necesariamente un camino cerrado en P'. Basados en la interpretación del código de barras que hacen Chan, Carlsson y Rabadan [14], Parida *et al.* [54] sugieren que es posible

obtener información sobre los caminos cerrados en P' mediante homología persistente.

4.2. Mezcla de poblaciones y homología persistente

En esta sección presentaremos una manera de aterrizar a lenguaje matemático las ideas que se quieren estudiar para ver cómo es posible relacionar los códigos de barras de persistencia con el fenómeno que buscamos modelar.

Los TEOREMAS 4.2 y 4.3, con los que concluye esta sección, constituyen una reformulación del resultado principal del artículo [54] la cual surgió como consecuencia de una comunicación con uno de sus autores: el profesor Saugata Basu de la Universidad de Purdue, Estados Unidos. Así, reemplazamos el teorema principal de [54] por los resultados discutidos con el profesor Basu ya que estos se pueden interpretar en el sentido original que pretendía el artículo [54].

Es claro que estructurar una muestra en poblaciones equivale a considerar una partición del conjunto de datos dado. Una manera sencilla de representar esto se obtiene al considerar las preimágenes de una función sobreyectiva a un "*espacio de poblaciones*".

DEFINICIÓN 4.3 Sean $\mathcal{M}(M,d)$ un espacio métrico finito, $\tilde{M}$ un conjunto finito y $\phi : M \to \tilde{M}$ una función sobreyectiva. Definimos $\tilde{d} : \tilde{M} \times \tilde{M} \to \mathbb{R}$ como sigue,

$$\tilde{d}(A,B) = \min_{u\in M,\phi(u)=A,u'\in M,\phi(u')=B} d(u,u').$$

Decimos que $\tilde{M}$ *etiqueta* a M mediante ϕ y a $\tilde{M}$ lo llamamos el *espacio de etiquetas* de M.

Como veremos más adelante, el espacio de etiquetas $(\tilde{M},\tilde{d})$ de un espacio (M,d) mediante una función ϕ no siempre es un espacio métrico. Sin embargo, dado un $r \geqslant 0$ es posible construir el complejo Vietoris-Rips $V(\tilde{M},r)$ asociado ya que $\tilde{d}$ sí es una función de disimilitud.

OBSERVACIÓN 4.1 Se sigue de la definición de $\tilde{d}$ que $\tilde{d}(\phi(u),\phi(v)) \leqslant d(u,v)$ para cualesquiera $u,v \in M$.

En particular, esta última observación nos permite construir mapeos simpliciales entre los complejos $V(M,r)$ y los complejos asociados a las etiquetas.

LEMA 4.1 Sea $\mathcal{M}(M,d)$ un espacio métrico finito y $\tilde{\mathcal{M}}(\tilde{M},\tilde{d})$ un espacio de etiquetas de M mediante ϕ. Entonces, para cada par $0 \leqslant a \leqslant b$, ϕ induce un mapeo simplicial

$$\phi^{a,b} : V(M,a) \to V(\tilde{M},b).$$

Si $a = b$ denotamos a $\phi^{a,b}$ por ϕ^a.

Demostración
Consideremos un subconjunto $U \subset M$ que define un simplejo en $V(M, a)$, esto es, $d(u, v) \leqslant a$ para todos los $u, v \in U$. De la Observación 4.1 se sigue que $\tilde{d}(\phi(u), \phi(v)) \leqslant a \leqslant b$. Por lo tanto, $\phi(U)$ define un simplejo en $V(\tilde{M}, b)$. □

Se pueden definir etiquetados de un espacio de manera arbitrariad, desde darle a todos los datos una misma etiqueta hasta definir una etiqueta diferente por cada punto. Nuestro objetivo es estudiar etiquetados de una nube de puntos que reflejen cómo estos puntos se encuentran agrupados, por ello necesitamos restringirnos a las nubes de datos que admiten un etiquetado de tal especie. Esto incluye que la nube de datos tenga distancias en dos *escalas*, es decir, que los puntos en una misma población estén sustancialmente más cercanos que los puntos en poblaciones diferentes. Queremos pues, definir cuáles etiquetados *reflejan* esta estructura en los datos.

Definición 4.4 Sea $\tilde{M}$ un espacio de etiquetas de M mediante ϕ y $c > 0$. Decimos que $\tilde{M}$ es un espacio de *etiquetas de nivel* c si para todos los u, v en M se satisfacen las siguientes propiedades:

(a) Si $\phi(u) = \phi(v)$, entonces $d(u, v) < c$.

(b) Si $\phi(u) \neq \phi(v)$, entonces $d(u, v) > 2c$.

(c) Si existen $u', v' \in M$ tal que $\phi(u) = \phi(u')$ y $\phi(v) = \phi(v')$, entonces

$$|d(u, v) - d(u', v')| < c.$$

Notemos que la definición anterior incluye propiedades intrínsecas de la métrica de la nube de puntos así como características necesarias en la función de etiquetado ϕ. En particular, observemos que las distancias entre las etiquetas deben estar claramente separadas de las distancias entre individuos de una misma población.

Observación 4.2 Si $\tilde{\mathcal{M}}$ es un espacio de etiquetas de $\mathcal{M}$ de nivel c, entonces se tiene que $\tilde{d}(A, B) \geqslant 2c$ para cualesquiera $A, B \in \tilde{M}$.

Recordemos que no es necesario que una función de disimilitud sea una métrica exacta para poder definir los complejos simpliciales Vietoris-Rips sobre la nube de puntos. Esto resulta importante porque como podemos ver en la Figura 4.2, no siempre ocurre que un espacio de etiquetas sea un espacio métrico, incluso si se satisfacen las hipótesis de la Definición 4.4 o si el espacio de individuos es un espacio métrico.

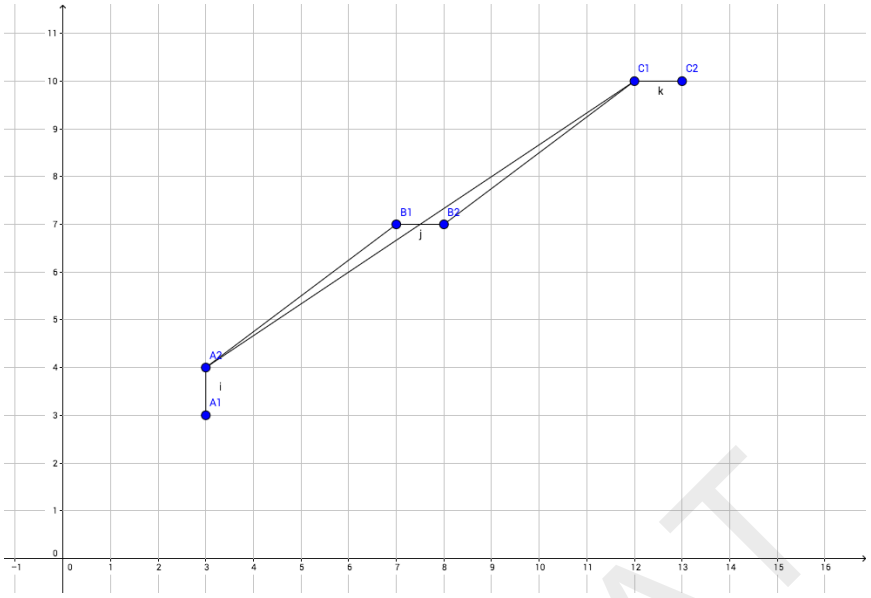


Figura 4.2: Ejemplo de un etiquetado que no forma un espacio métrico a partir de un espacio de individuos formado por seis puntos en $\mathbb{R}^2$. El espacio de etiquetas tiene tres elementos, A, B y C, mientras que los puntos denotan la etiqueta que se les asignó. Es importante destacar que aunque este etiquetado satisface las hipótesis de la Definición 4.4 sigue sin ser un espacio métrico.

Observemos también que el inciso (c) de la Definición 4.4 equivale a pedir que $d(x, y) \leqslant \tilde{d}(\phi(x), \phi(y))$ para cualesquiera $x, y \in M$. Esta observación nos permite definir ciertos mapeos simpliciales que van de los complejos basados en las etiquetas a los complejos definidos sobre los individuos.

Lema 4.2 Sea $\mathcal{M}(M, d)$ un espacio métrico finito y $c > 0$. Sea $\tilde{\mathcal{M}}(\tilde{M}, \tilde{d})$ un espacio de etiquetas sobre M de nivel c definido por una función ϕ. Sean $a, b \geqslant 0$ tal que $a + c \geqslant b$, entonces existe un mapeo simplicial,

$$\psi^{a,b} : V(\tilde{M}, a) \to V(M, b).$$

Demostración

Sea $\tilde{U} \subset \tilde{M}$ un subconjunto que define un simplejo en $V(\tilde{M}, a)$ y sea $U = \phi^{-1}(\tilde{U})$. Consideremos $\tilde{u}, \tilde{v} \in \tilde{U}$, tenemos pues que $\tilde{d}(\tilde{u}, \tilde{v}) \leqslant a$ y por lo tanto sabemos ya que para todos $u, v \in U$ tal que $\phi(u) = \tilde{u}$ y $\phi(v) = \tilde{v}$ se tiene que $d(u, v) \leqslant \tilde{d}(\tilde{u}, \tilde{v}) + c$ por la propiedad (c) de la Definición 4.4. Por lo tanto,

$$d(u, v) \leqslant a + c \leqslant b,$$

lo cual concluye la demostración.

□

A fin de establecer los teoremas principales de esta sección, además de las definiciones y las observaciones que hemos hecho, es necesario incluir el siguiente resultado general puesto que juega un papel importante en la demostración de dichos teoremas. Este teorema, que tomamos de [48], es una versión particular de un resultado clásico en topología algebraica.

Teorema 4.1 Mapeo de Vietoris-Begle Sean X y Y espacios topológicos compactos y sea $g : X \to Y$ una función propia tal que $g^{-1}(y)$ es acíclico para cada $y \in Y$. Se tiene entonces que el mapeo inducido $g_* : H_*(X) \to H_*(Y)$ es un isomorfismo.

Estamos ya preparados para presentar los teoremas principales de esta sección.

Teorema 4.2 Sean $\mathcal{M}(M, d)$ un espacio métrico finito y $c > 0$. Sea $\tilde{\mathcal{M}}(\tilde{M}, \tilde{d})$ un espacio de etiquetas sobre M de nivel c definido por una función ϕ. Se tienen entonces las siguientes afirmaciones.

(1) $H_p(V(\tilde{M}, r)) = 0$ para todo $p > 0$ y todo $r < 2c$.

(2) Para todos $r, r' \geqslant 0$, tal que $r' - r > c$, se tiene que

$$H^{r,r'}(\tilde{M}) \subset Im(\phi_*^{r'}|_{H^{r+c,r}(M)}) \subset H^{r+c,r'}(\tilde{M}),$$

donde ϕ^r es el mapeo inducido, como en el Lema 4.1, por la función de etiquetado. Si se tiene además que $\tilde{i}_*^{r,r'}$ es un isomorfismo, entonces

$$H(\tilde{M}, r) \cong H^{r,r+c}(\tilde{M}) \cong H^{r,r'}(\tilde{M}),$$

y además:

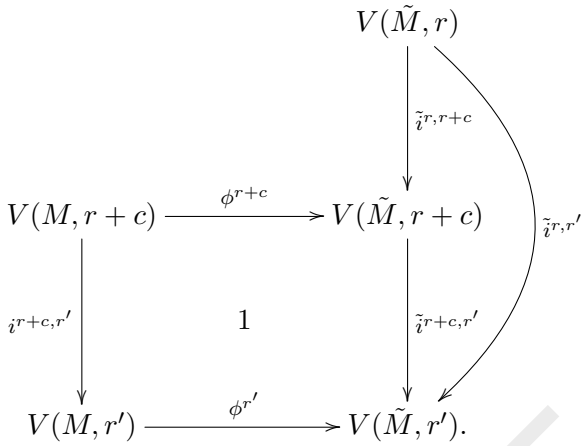
a) $\phi_*|_{H^{r+c,r'}(M)}$ es suprayectiva sobre $H^{r,r'}(\tilde{M})$,

b) $\phi_*|_{H^{r,r'}(M)}$ es una función inyectiva a $H^{r,r'}(\tilde{M})$.

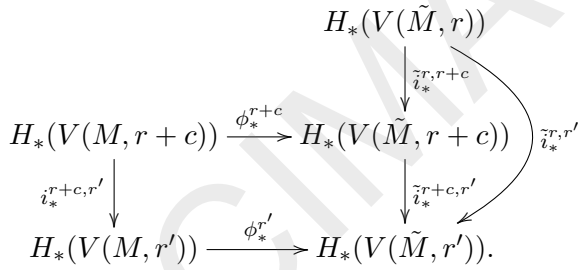
Demostración

Es fácil ver que (1) es consecuencia directa de la Observación 4.2.

Ahora bien, para probar (2) empecemos observando el siguiente diagrama resultado de considerar las funciones $\phi^{r+c}, \phi^{r'}$ dadas por el Lema 4.1 y las inclusiones naturales entre los complejos simpliciales involucrados.



Es fácil ver que el diagrama conmuta. Como consecuencia, tenemos que el siguiente diagrama inducido en los grupos de homología también conmuta:



Recordemos que por definición $H_*^{r+c,r'}(M) = \text{Im}(i_*^{r+c,r'}) \subset H_*(V(M, r'))$.

Se tiene directamente que la imagen de $\phi_*^{r'}|_{H^{r+c,r}(M)}$ está contenida en $\text{Im}(\tilde{i}_*^{r+c,r'})$, puesto que es una consecuencia directa de la conmutatividad del cuadrado 1 en el diagrama entre grupos de homología.

Por otro lado, también se tiene que $H_*^{r,r'}(M) \subset \text{Im}(\phi_*^{r'}|_{H^{r+c,r}(M)})$, ya que es una consecuencia directa de la conmutatividad en las inclusiones entre grupos de homología de la filtración sobre el espacio de etiquetas $\tilde{i}_* r, r' = \tilde{i}_*^{r+c,r'} \circ \tilde{i}_*^{r,r+c}$. Con lo anterior tenemos ya la primera parte del segundo inciso:

$$H^{r,r'}(\tilde{M}) \subset \text{Im}(\phi_*^{r'}|_{H^{r+c,r}(M)}) \subset H^{r+c,r'}(\tilde{M}). \tag{4.1}$$

Ahora bien, supongamos que $\tilde{i}_*^{r,r'}$ es un isomorfismo, la primera consecuencia directa es que las contenciones de la ecuación (4.1) son ahora isomorfismos:

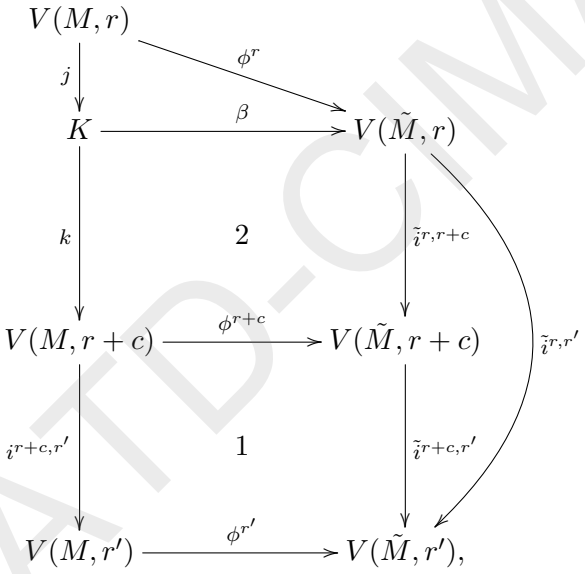
$$H^{r,r'}(\tilde{M}) \cong Im(\phi_*^{r'}|_{H^{r+c,r}(M)}) \cong H^{r+c,r'}(\tilde{M}), \tag{4.2}$$

y por lo tanto, podemos ver a $\phi_*^{r'}|_{H^{r+c,r}(M)}$ como una función que va a $H^{r,r'}(\tilde{M})$.

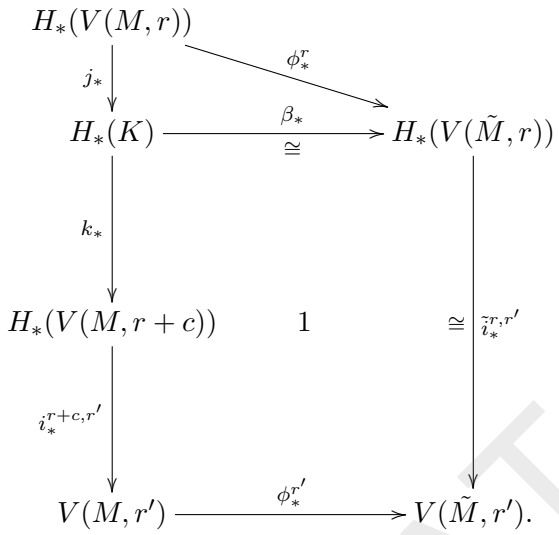
Para proceder con la prueba necesitamos definir el siguiente conjunto:

$$K = \{\sigma \in V(M, r+c) | \exists\ \tilde{u} \in V(\tilde{M}, r) \text{ tal que } f(\sigma) = \tilde{i}^{r,r'}(\tilde{u})\}.$$

Es fácil ver que, debido al mapeo ϕ^r, se tiene que $V(M,r) \subset K$. Además, podemos definir un mapeo simplicial $\beta : K \to V(\tilde{M}, r)$ que le asigna a cada elemento en K el elemento en $V(\tilde{M}, r)$ que, por definición de K, le corresponde. Así, se puede extender ahora el diagrama entre complejos simpliciales de la siguiente manera:



donde j y k son las inclusiones naturales entre los complejos. Afirmamos que β satisface las hipótesis del Teorema 4.1. En efecto, la función es propia y el mapeo simplicial $\psi^{r,r+c}$, que se obtiene debido al Lema 4.2, garantiza que las fibras $\beta^{-1}(y)$ sean acíclicas para cada $y \in Y$. Por lo tanto, β_* es un isomorfismo y tenemos el siguiente diagrama, en homologías, el cual conmuta:



Como el diagrama conmuta, tenemos que la composición $(\phi_*^{r'} \circ i_*^{r+c,r'}) \circ k_*$ es un isomorfismo, y por lo tanto, $\phi_*^{r'} \circ i_*^{r+c,r'}$ es un epimorfismo. En otras palabras, $\phi_*|_{H^{d+c,d'}(M)}$ es suprayectiva sobre $H^{d,d'}(\tilde{M})$. Con ello queda demostrado el inciso a).

Dado que $i_*^{r,r'} = i_*^{r+c,r'} \circ k_* \circ j_*$, por la conmutatividad del diagrama tenemos que $\phi_*^{r'} \circ i_*^{r,r'}$ debe ser inyectiva, lo cual prueba b) y concluye la demostración del teorema.

□

En este último resultado se intuye claramente que los códigos de barras de persistencia asociados a los organismos y a las poblaciones están estrechamente relacionados. El siguiente teorema permite, bajo ciertas condiciones, extraer información acerca de esta relación cuando se observa tan solo uno de los dos códigos.

Teorema 4.3 Sean $\mathcal{M}(M,d)$ un espacio métrico finito y $c > 0$. Sea $\tilde{\mathcal{M}}(\tilde{M},\tilde{d})$ un espacio de etiquetas sobre M de nivel c definido por una función ϕ. Para todos $r, r' \geqslant 0$, tales que $r' - r > 2c$ y que satisfagan que $\tilde{i}_*^{r,r'}$ sea un isomorfismo, se tiene que $H_*^{r+c,r'}(M)$ es isomorfo a $H_*^{r,r'}(\tilde{M})$.

Demostración

Sean $r_1, r_2 > 0$ tal que $r_2 - r_1 > 2c$. Supongamos que $\tilde{i}_*^{r,r'}$ es un isomorfismo. Utilizando el Teorema 4.2 dos veces, una con $r = r_1$ y la segunda con $r = r_1 + c$, obtenemos que

i) $\phi_*|_{H^{r_1+c,r_2}(M)}$ es suprayectiva sobre $H^{r_1,r_2}(\tilde{M})$,

ii) $\phi_*|_{H^{r_1+c,r_2}(M)}$ es una función inyectiva a $H^{r_1+c,r2}(\tilde{M})$.

Como $H_*^{r_1+c,r_2} \cong H_*^{r_1,r_2}(\tilde{M})$ por hipótesis, concluimos que $\phi_*|_{H^{r_1+c,r_2}(M)}$ es un isomorfismo con $H_*^{r_1,r_2}(\tilde{M})$.

□

4.3. Distintivos topológicos para mezclas de poblaciones

El Teorema 4.3 es aplicable al contexto de mezcla de poblaciones de la siguiente manera. Supongamos que tenemos una muestra que está bien estructurada en poblaciones. Esto es, que el espacio métrico formado por las secuencias de los organismos y la función de disimilitud satisface las propiedades de la Definición 4.4 con el etiquetado que se tiene dado, para algún $c > 0$. En este caso, por el Teorema 4.3, sabemos ya que los ciclos irreducibles presentes en el código de barras de persistencia asociado al espacio de poblaciones aparecen después del nivel de persistencia $2c$. Más aún, si existe un intervalo de tamaño mayor que $2c$ en el código de barras de las poblaciones tal que la homología no cambie a lo largo de él—esto es, en el que no aparezcan ni desaparezcan ciclos—podemos relacionar cada uno de estos ciclos con ciclos irreducibles en cierto intervalo del código de barras de los individuos.

Por lo tanto, siguiendo la interpretación de los códigos de barra propuesta en el Postulado 3.1, se tiene que los eventos reticulares entre poblaciones, las mezclas de poblaciones, se pueden detectar desde los eventos reticulares de los individuos que los causan. Las dos escalas de distancias presentes en la muestra se reflejan en dos escalas en los ciclos del código de barras de los organismos mientras que en los ciclos de la escala más grande se puede obtener información de los eventos horizontales a nivel de poblaciones, es decir de las mezclas de poblaciones.

De esta manera, el Teorema 4.3 puede *predecir* la presencia de eventos horizontales a nivel de poblaciones ya que estos pueden ser detectados a partir de los códigos de barras de persistencia asociados a los organismos de la muestra. Es importante señalar que esta predicción no requiere conocimiento previo, ni de la manera en que se *etiquetan* los organismos, ni de las distancias entre estas etiquetas, ya que está basada en las características del código de persistencia de los organismos. Es así, que la misma *estructura* del código de barras de los organismos es lo que constituye los *distintivos topológicos para la mezcla de poblaciones*.

Como los autores del artículo [54] mencionan, ya en la práctica los datos no necesariamente satisfacen *de manera exacta* los requerimientos de la Definición 4.4 y sin embargo el comportamiento de los códigos de barras resultantes es similar a la estructura teórica esperada. Esto es lo que se observa en los códigos de persistencia que obtuvieron al realizar simulaciones usando software propio, SimRA [12], como se puede ver en la Figura 4.3,

y también al probar el método con datos reales provenientes de distintas poblaciones de aguacates, Figura 4.4.

Los mismos autores prevén que debido al ruido en los datos, y posiblemente a factores aún desconocidos, será necesario *refinar* cuidadosamente la manera de definir estos *distintivos topológicos* comenzando por determinar en concreto la mejor manera de elegir el umbral $c > 0$ que separa las distancias entre individuos y las distancias entre poblaciones.

Por último, y a manera de conclusión, queremos terminar citando a los autores: "*Este trabajo preliminar es promisorio y en trabajos futuros, planeamos estudiar modelos más complejos de mezcla, tanto en términos de la topología de los armazones de ARG como en términos de caracterizaciones más complejas de los eventos reticulares a nivel de poblaciones*". Creemos por nuestra parte, que los distintivos topológicos tienen el potencial no sólo de detectar, sino también de distinguir las mezclas de poblaciones recientes de las antiguas.

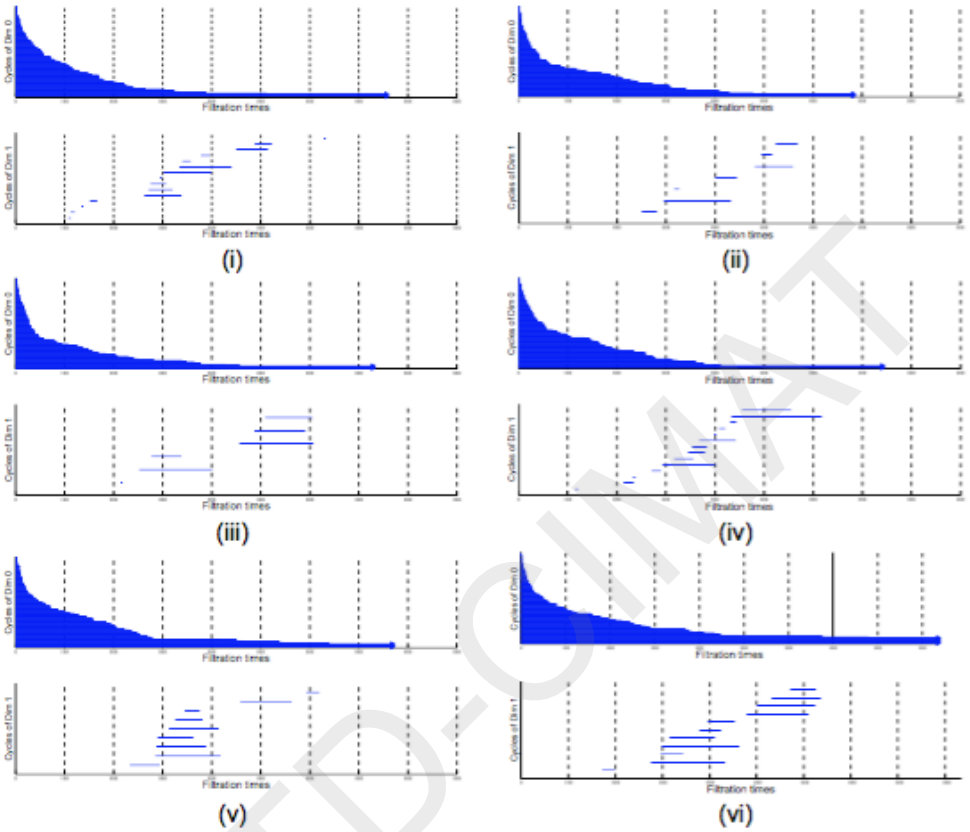


Fig. 7. Six simulations, each with effective population size $N = 10K$; with recombination ($r = 0.3 \times 10^{-8}$) as well as admixture to show that stochasticity does not affect the topological signature, i.e. the separation of the persistent cycles into roughly two groups.

Figura 4.3: Códigos de persistencia resultado de realizar simulaciones de poblaciones Wright-Fisher con distintos parámetros. Figura tomada del artículo original [54], donde se pueden encontrar más detalles sobre las simulaciones.

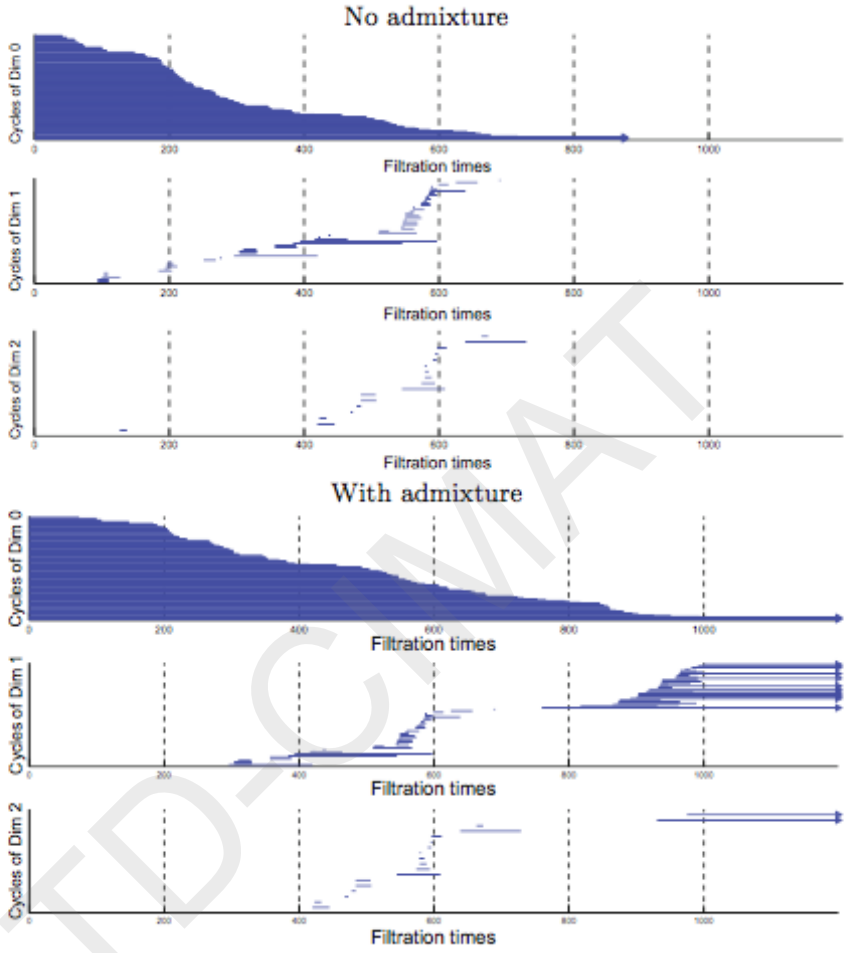


Fig. 4. Haplotypes from three groups of avocado germplasm data: West Indian (W), Mexican (M), and Guatemalan (G). The top plot corresponds to the populations with no admixture, while the bottom admits admixture in the populations. Notice the separation of the persistent cycles in both dimension 1 and 2 for the latter scenario, while the former shows no clear separation.

Figura 4.4: Códigos de persistencia obtenidos a partir de secuencias de diferentes poblaciones de aguacates. Figura tomada del artículo original [54], donde se pueden encontrar más detalles sobre el procesamiento de los datos.

Apéndice A

Sucesiones exactas

Introducción

El objetivo de este apéndice es ofrecer una demostración completa de la Proposición 2.2, la cual usamos como criterio para detectar complejos simpliciales acíclicos. Este resultado es una consecuencia inmediata del teorema de las sucesiones exactas de Mayer-Vietoris cuyos preliminares se incluyen.

El material que exponemos aquí constituye una breve recopilación de resultados estándar de álgebra que se usan en homología. En gran parte de la literatura se encuentran enunciados para grupos de homología con coeficientes en $\mathbb{Z}$ pero a lo largo de este trabajo consideramos únicamente coeficientes sobre $\mathbb{Z}_2$. Por lo tanto, teniendo en mente al lector sin experiencia previa en homología, enunciamos los resultados correspondientes para homología con coeficientes en $\mathbb{Z}_2$.

El contenido de este apéndice se basa en el libro de Edelsbrunner y Harer [21], adaptando parte de las ideas y notación presentes en el trabajo de Munkres [51]. La Proposición 2.2 se tomó del artículo de Chan, Carlsson y Rabadan [14] donde aparece sin demostración.

Lema de la Serpiente

Muchas de las definiciones que hemos hecho en el contexto de complejos simpliciales también se presentan en situaciones más generales. Vamos a considerar los grupos de cadenas asociados a un complejo simplicial como construcciones algebraicas en abstracto para estudiar las relaciones que existen entre sus grupos de homología. Para ello empezaremos estableciendo notación. Todos los grupos que vamos a considerar en este apéndice son abelianos.

Dadas dos sucesiones de grupos que comparten el mismo conjunto de

índices

$$\mathcal{G} : \cdots G_1 \to G_2 \to G_3 \to \cdots ,$$
$$\mathcal{H} : \cdots H_1 \to H_2 \to H_3 \to \cdots ,$$

podemos definir un mapeo lineal entre ellas de la siguiente manera.

Definición A.1 Sean $\mathcal{G}$ y $\mathcal{H}$ sucesiones de grupos como describimos arriba. Un *homomorfismo* α de $\mathcal{G}$ a $\mathcal{H}$, es una familia de homomorfismos $\alpha_i : G_i \to H_i$ en la cual el siguiente diagrama conmuta para cualquier i:

$$\begin{array}{ccc} G_i & \longrightarrow & G_{i+1} \\ {\scriptstyle \alpha_i}\downarrow & & \downarrow{\scriptstyle \alpha_{i+1}} \\ H_i & \longrightarrow & H_{i+1} \end{array} .$$

Decimos que α es un *isomorfismo* de sucesiones si cada α_i es un isomorfismo.

Definición A.2 Sea $\mathcal{G}$ una familia de grupos conectados por homomorfismos

$$\mathcal{G} : \ldots G_1 \xrightarrow{\phi_1} G_2 \xrightarrow{\phi_2} G_3 \xrightarrow{\phi_3} \ldots .$$

- Si $\mathrm{Im}(\phi_1) = \mathrm{Ker}(\phi_2)$, decimos que la secuencia es *exacta* en G_2.
- De la misma manera, la *sucesión es exacta* si es exacta en cada uno de sus grupos.

Análogamente, decimos que una secuencia finita de grupos es exacta si es exacta en todos sus grupos a excepción del primero y del último. Cuando una sucesión de grupos con conjunto de índices $\mathbb{Z}$ es exacta, la llamaremos *sucesión exacta larga* de grupos.

Por simplicidad, en la descripción de una sucesión de grupos intentaremos escribir la información indispensable. Sin embargo, evitaremos escribir los subíndices de los homomorfismos o incluso omitiremos el nombre de los homomorfismos mismos cuando el contexto lo permita.

La siguiente proposición resume las propiedades básicas de las secuencias exactas que vamos a utilizar.

Proposición A.1

(a) $G_1 \xrightarrow{\phi_1} G_2 \to 0$ es exacta si, y sólo si, ϕ_1 es un epimorfismo.

(b) $0 \to G_1 \xrightarrow{\phi_1} G_2$ es exacta si, y sólo si, ϕ_1 es un monomorfismo.

Las secuencias exactas de grupos nos servirán más adelante para expresar de forma simple las relaciones que existen entre los grupos de homología de un complejo simplicial. La siguiente definición nos permite olvidarnos de

los complejos simpliciales y trabajar con estructuras algebraicas que se comportan de manera idéntica a las que asociamos a los complejos simpliciales para definir su homología.

Definición A.3 Un *complejo de cadenas* $\mathcal{C}$ es una familia $(C_p, \partial_p)_{p\in\mathbb{Z}}$ de grupos y homomorfismos

$$\partial_p : C_p \to C_{p-1}$$

de tal manera que $\partial_p \circ \partial_{p+1} = 0$ para todo $p \in \mathbb{Z}$.

Al cociente

$$H_p(\mathcal{C}) = \operatorname{Ker}(\partial_p)/\operatorname{Im}(\partial_{p+1})$$

lo llamamos el *p-ésimo grupo de homología* de $\mathcal{C}$.

Claramente la estructura algebraica que se usa para calcular la homología de un complejo simplicial es el primer ejemplo de un complejo de cadenas. Observemos que esta definición tiene como objetivo principal imitar la propiedad del operador frontera establecido por el Lema Fundamental de Homología (Lema 2.3).

Es claro también que podemos etiquetar los elementos de los grupos en un complejo de cadenas como *fronteras* y *ciclos* siguiendo la misma lógica con la que etiquetamos las cadenas de simplejos. El resto de la terminología de los complejos de cadenas algebraicos también se hereda de los términos en homología simplicial. Por ejemplo, a ∂_p lo denominamos *operador frontera* y a los elementos de $\operatorname{Im}(\delta_p)$ se les llama *fronteras* de dimensión p. Asimismo, si un complejo de cadenas $\mathcal{D}$ satisface que $H_p(\mathcal{D}) = 0$ para todo p, decimos que $\mathcal{D}$ es *acíclico*. Por último, así como $H_p(K)$ denotaba el p-ésimo grupo de homología del complejo simplicial K, $H_p(C)$ denotará el p-ésimo grupo de homología de un complejo de cadena $\mathcal{C}$. Consecuentemente, a la sucesión $H_p(C)$ la llamaremos la *homología* de $\mathcal{C}$.

Si $\mathcal{C}$ y $\mathcal{D}$ son complejos de cadenas, y $\alpha : \mathcal{C} \to \mathcal{D}$ es un homomorfismo entre ellos, decimos que α es un *mapeo de cadenas*. Observemos que debido a la conmutatividad en todos los grupos, α induce un homomorfismo α_* entre las homologías dado por

$$(\alpha_p)_* : H_p(C) \to H_p(D)$$

para cada dimensión p.

La siguiente observación será utilizada más adelante.

Observación A.1 Notemos que si tenemos dos complejos de cadenas $\mathcal{C}$ y $\mathcal{D}$, podemos definir un tercer complejo de cadenas $\mathcal{E}$ donde cada grupo es de la forma

$$E_p = C_p \oplus D_p$$

y definimos el operador frontera por componentes $\partial_E = (\partial_C, \partial_D)$. Es fácil ver que para cada p tenemos que

$$H_p(E) = \frac{\operatorname{Ker}(\partial_E)}{\operatorname{Im}(\partial_E)} = \frac{\operatorname{Ker}(\partial_C) \oplus \operatorname{Ker}(\partial_D)}{\operatorname{Im}(\partial_C) \oplus \operatorname{Im}(\partial_D)} \simeq H_p(C) \oplus H_p(D).$$

Por último, en la siguiente definición extendemos la noción de homomorfismos al contexto de complejos de cadenas.

Definición A.4 Sean $\mathcal{C}$, $\mathcal{D}$ y $\mathcal{E}$ complejos de cadenas, y sean $\phi : \mathcal{C} \to \mathcal{D}$, $\psi : \mathcal{D} \to \mathcal{E}$ homomorfismos. Decimos que

$$0 \to \mathcal{C} \xrightarrow{\phi} \mathcal{D} \xrightarrow{\psi} \mathcal{E} \to 0$$

es una *sucesión exacta corta* de complejos, si para cada p la sucesión

$$0 \to C_p \xrightarrow{\phi_p} D_p \xrightarrow{\psi_p} E_p \to 0$$

es una sucesión exacta de grupos.

El siguiente lema establece el método general para construir sucesiones exactas largas en términos de los grupos de homología correspondientes y a partir de complejos de cadenas que forman una sucesión exacta corta. Nos referiremos al siguiente resultado como el lema Zigzag o lema de la Serpiente.

Lema A.1 Zigzag o de la Serpiente Sea $0 \to \mathcal{A} \xrightarrow{\phi} \mathcal{B} \xrightarrow{\psi} \mathcal{C} \to 0$ una secuencia exacta corta de complejos de cadena. Existe un mapeo bien definido $c : H_p(C) \to H_{p-1}(A)$ llamado el *homomorfismo conector* tal que

$$\cdots \to H_p(A) \to H_p(B) \to H_p(C) \xrightarrow{c} H_{p-1}(A) \to \ldots$$

es una secuencia exacta larga de grupos.

Demostración

Para construir la secuencia vamos a utilizar los homomorfismos inducidos en los grupos de homología a partir de los homomorfismos entre los complejos

$$\phi_* : H_p(A)_p \to H_p(B) \ , \ \psi_* : H_p(B) \to H_p(C);$$

tras construir el homomorfismo c, veremos que está bien definido y, por último, comprobaremos la exactitud de la secuencia.

La demostración está basada en el siguiente diagrama, que representa una sección de la secuencia exacta corta dada:

$$\begin{array}{ccccccccc}
0 & \longrightarrow & \mathcal{A} & \xrightarrow{\phi} & \mathcal{B} & \xrightarrow{\psi} & \mathcal{C} & \longrightarrow & 0 \\
 & & & & & & & & \\
 & & & & B_{p+1} & \longrightarrow & C_{p+1} & \longrightarrow & 0 \\
 & & & & \downarrow & & \downarrow & & \downarrow \\
 & & & & & 4 & & & \\
0 & \longrightarrow & A_p & \longrightarrow & B_p & \longrightarrow & C_p & \longrightarrow & 0 \\
\downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\
 & & & 3 & & 1 & & & \\
0 & \longrightarrow & A_{p-1} & \longrightarrow & B_{p-1} & \longrightarrow & C_{p-1} & \longrightarrow & 0 \\
\downarrow & & \downarrow & & \downarrow & & & & \\
 & & & 2 & & & & & \\
0 & \longrightarrow & A_{p-2} & \longrightarrow & B_{p-2} & & & & .
\end{array}$$

Las flechas verticales indican los correspondientes operadores frontera y las horizontales, los homomorfismos definidos en cada nivel por los homomorfismos entre complejos. Observemos que los cuadrados numerados son conmutativos por construcción.

Paso 1. Definir el homomorfismo c

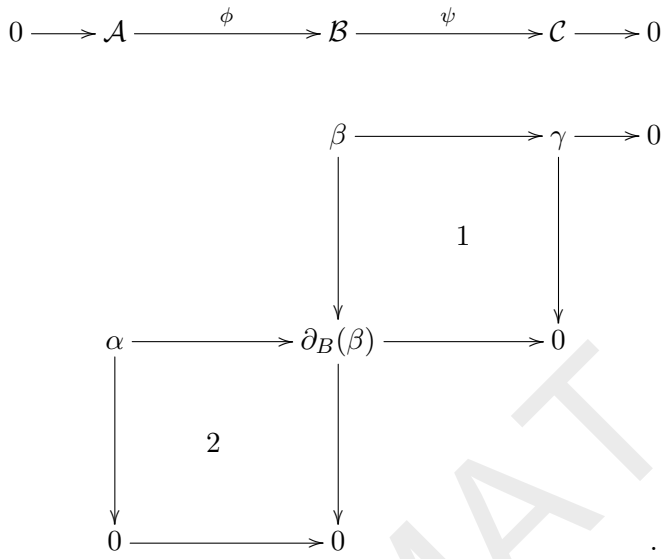
Tomemos un ciclo $\gamma \in C_p$ como representante de una clase en $H_p(C)$. Dado que ψ es sobreyectiva, existe una cadena $\beta \in B_p$ tal que $\psi(\beta) = \gamma$. Por ser γ un ciclo, su frontera es cero, lo cual aunado a la conmutatividad del diagrama en 1 da que

$$\psi(\partial_B(\beta)) = \partial_C(\psi(\beta))\partial_C(\alpha) = 0.$$

Dicho de otro modo $\partial_B(\beta) \in \text{Ker}(\psi)$ y por la exactitud de B_{p-1} existe una cadena $\alpha \in A_{p-1}$ tal que $\phi(\alpha) = \partial_B(\beta)$. Ahora bien, por la conmutatividad en 2 y la propiedad fundamental de los operadores frontera, resulta que

$$\phi(\partial_A(\alpha)) = \partial_B(\phi(\alpha)) = \partial_B(\partial_B(\beta)) = 0.$$

Resumiendo, tenemos que



Ahora bien, por la conmutatividad en 2 y la propiedad fundamental de los operadores frontera tenemos que

$$\phi(\partial_A(\alpha)) = \partial_B(\phi(\alpha)) = \partial_B(\partial_B(\beta)) = 0.$$

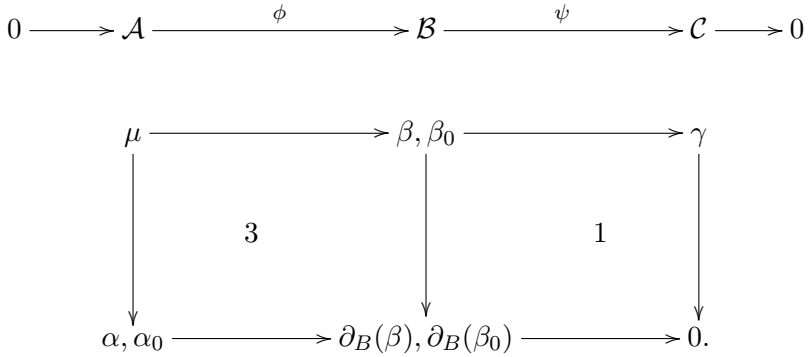
Más aún, como ϕ es inyectiva, tenemos que $\partial_A(\alpha) = 0$ con lo cual α es un ciclo y por lo tanto representante de una clase en $H_{p-1}(A)$. Definimos $c([\gamma]) = [\alpha]$.

Paso 2. El homomorfismo conector está bien definido

Para ver que efectivamente tratamos con una función, debemos ver que no exista dependencia respecto de los representantes o las elecciones de β y de γ que hicimos en el proceso. Veamos primero que no depende de la elección de β.

Supongamos que existe otro $\beta_0 \in B_p$ tal que $\psi(\beta_0) = \gamma$ y sea α_0 el único elemento en A_{p-1} tal que $\phi(\alpha_0) = \partial_B\beta_0$. Como $\phi(\beta) = \phi(\beta_0)$ tenemos que $\beta + \beta_0 \in \text{ Ker}(\psi) = \text{ Im}(\phi)$, esto último por la exactitud de B_p. Así pues, existe un elemento $\mu \in A_p$ tal que $\phi(\mu) = \beta + \beta_0$. Observemos el diagrama

resultante

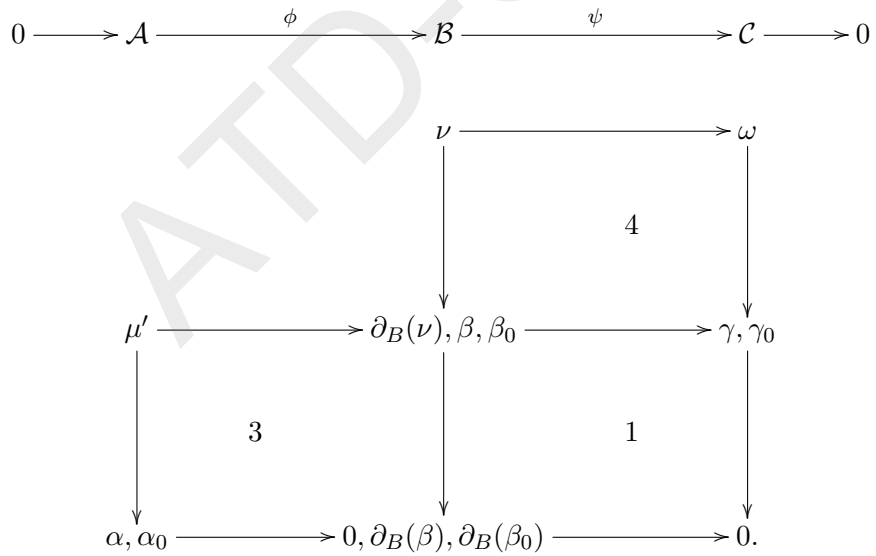


Como el cuadrado 3 conmuta, tenemos que

$$\phi(\partial_A(\mu)) = \partial_B(\phi(\mu)) = \partial_B(\beta+\beta_0) = \partial_B(\beta)+\partial_B(\beta_0) = \phi(\alpha)+\phi(\alpha_0) = \phi(\alpha+\alpha_0).$$

Además, debido a que ϕ es inyectiva se sigue que $\partial_A(\mu) = \alpha + \alpha_0$. En otras palabras, α y α_0 difieren sólo por una frontera y por lo tanto pertenecen a la misma clase de homología en $H_{p-1}(A)$. Por eso es que $c(\gamma)$ no depende de la elección de β.

Por otro lado, si γ_0 es otro representante de la misma clase de homología que γ, escojamos β_0 y α_0 a partir de γ_0 de la misma manera que hicimos con γ. Dado que γ y γ_0 son homólogas en $H_p(C)$, existe una cadena $\omega \in C_{p+1}$ tal que $\partial_C(\omega) = \gamma + \gamma_0$. Más aún, como ψ es sobreyectiva, existe un $\nu \in B_{p+1}$ tal que $\psi(\nu) = \omega$. La situación se encuentra resumida en la parte superior del siguiente diagrama.



Por la conmutatividad en 4 tenemos que

$$\psi(\partial_B(\nu)) = \partial_C(\psi(\nu)) = \partial_C(\omega) = \gamma + \gamma_0 = \psi(\beta) + \psi(\beta_0) = \psi(\beta + \beta_0)$$

con lo cual $\partial_B(\nu) + \beta + \beta_0 \in \text{Ker}(\psi)$. Por exactitud de B_p, existe $\mu' \in A_p$ tal que $\phi(\mu') = \partial_B(\nu) + \beta + \beta_0$. Por otra parte, usando conmutatividad en 3 y la propiedad fundamental de los operadores frontera se sigue que

$$\phi(\partial_A(\mu') = \partial_B(\phi(\mu')) = \partial_B(\partial_B(\omega) + \beta + \beta_0) = \partial_B(\beta + \beta_0).$$

La inyectividad de ϕ implica que $\partial_A(\mu') = \alpha + \alpha_0$. Esto es α y α_0 pertenecen a la misma clase de homología pues difieren sólo por una frontera. Con ello c está bien definida.

Paso 3. La sucesión es exacta

Falta únicamente ver que la sucesión es exacta en $H_p(C)$ y en $H_{p-1}(A)$.

Sea γ_p un representante de una clase en $H_p(C)$. Escojamos nuevamente β_p tal que $\psi(\beta_p) = \gamma_p$ y $\alpha_{p-1} \in A_{p-1}$ tal que $\phi(\alpha_{p-1}) = \partial_B\beta_p$. Por definición tenemos que $c([\gamma_p]) = [\alpha_{p-1}]$.

Si $[\gamma_p] \in \text{Im}(\psi_*)$, como $[\gamma_p] = [\psi\beta_p]$, tenemos que β_p es un ciclo en B_p. Con lo cual su frontera es vacía, y dada la inyectividad de ϕ, tenemos que $\alpha_{p-1} = 0$. Así, $[\gamma_p] \in \text{Ker}(c)$.

Por otro lado si $[\gamma_p] \in \text{Ker}(c)$, entonces $\alpha_{p-1} = \partial_A\alpha_p$ para alguna frontera $\alpha_p \in A_p$. Notemos que $\beta_p - \phi(\alpha_p)$ es un ciclo en B_p pues

$$\partial_B(\beta_p - \phi(\alpha_p)) = \partial_B\beta_p - \phi(\partial_A\alpha_p) = \phi(\alpha_{p-1}) - \phi(\alpha_{p-1}) = 0.$$

Más aún, por exactitud de la secuencia corta tenemos que

$$\psi_*([\beta_p - \phi(\alpha_p)]) = [\psi(\beta_p - \phi(\alpha_p))] = [\psi(\beta_p) - \psi(\phi(\alpha_p))] = [\gamma_p - 0] = [\gamma_p],$$

y por lo tanto $[\alpha_p] \in \text{Im}(\psi_*)$. De esta manera, la secuencia es exacta en $H_p(C)$. Veamos que también lo es en $H_{p-1}(A)$.

Sea $[\alpha_{p-1}] \in Im(c)$, esto es, existe $[\gamma_p] \in H_p(C)$ tal que $c([\gamma_p]) = [\alpha_{p-1}]$. Por definición tenemos que $\phi(\alpha_{p-1}) = \partial_B\beta_p$ para algún $\beta_p \in B_p$. Así

$$\phi_*([\alpha_{p-1}]) = [\phi(\alpha_{p-1})] = [\partial_B\beta_p] = 0,$$

y por ello $[\alpha_{p-1}] \in \text{Ker}(\phi_*)$.

Consideremos ahora $[\alpha_{p-1}] \in \text{Ker}(\phi_*)$; por definición tenemos que $[\phi(\alpha_{p-1})] = 0$. Esto es, existe un ciclo $\beta_p \in B_p$ tal que $\phi(\alpha_{p-1}) = \partial_B\beta_p$. Definamos $\gamma_p = \psi(\beta_p)$ y observemos que la exactitud de la secuencia corta implica que

$$\partial_C(\gamma_p) = \psi(\partial_B\beta_p) = \psi(\phi(\alpha_{p-1})) = 0.$$

De esta manera γ_p es un ciclo, y por construcción de c tenemos que $c([\gamma_p]) = [\alpha_{p-1}]$; es decir, $[\alpha_{p-1}] \in \text{Im(c)}$, lo cual concluye la exactitud en $H_{p-1}(A)$ y, por lo tanto, la exactitud de la cadena completa.

□

Sucesiones de Mayer-Vietoris

El Teorema de Sucesiones de Mayer-Vietoris es una de las herramientas principales para calcular la homología de un complejo simplicial al relacionar la homología de dos subcomplejos con la homología de su unión e intersección. Como veremos, su demostración es una consecuencia del lema de la Serpiente.

Teorema A.1 Sucesiones exactas de Mayer-Vietoris Sea K un complejo simplicial y K', K'' subcomplejos de K tales que $K = K' \cup K''$. Sea $A = K' \cap K''$, entonces existe una sucesión exacta

$$\cdots \to H_p(A) \to H_p(K') \oplus H_p(K'') \to H_p(K) \to H_{p-1}(A) \to \cdots .$$

Demostración

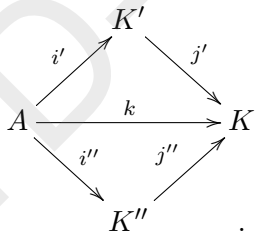
La prueba consiste en construir una secuencia exacta corta de cadenas complejas de la forma

$$0 \to \mathcal{C}(A) \xrightarrow{\phi} \mathcal{C}(K') \oplus \mathcal{C}(K'') \xrightarrow{\psi} \mathcal{C}(K) \to 0. \tag{A.1}$$

El resultado se sigue de utilizar el lema de la Serpiente sobre la secuencia construida.

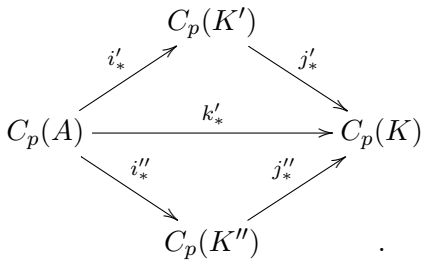
Observemos primero que tanto $\mathcal{C}(A)$ como $\mathcal{C}(K') \oplus \mathcal{C}(K'')$ son cadenas complejas bien definidas. Para la primera, basta observar que A es un subcomplejo simplicial, por ser intersección de complejos simpliciales. Para obtener la segunda, basta definir el operador frontera por componentes.

El siguiente diagrama establece la notación para las inclusiones estándar entre los complejos y es claro que el siguiente diagrama conmuta.



.

Como $C_p(A)$ es subgrupo tanto de $C_p(K')$ como de $C_p(K'')$ para cada p, obtenemos el siguiente diagrama, el cual conmuta, al considerar los homomorfismos inducidos por las inclusiones:



.

Omitimos el subíndice p en cada homomorfismo por comodidad, pero en el resto de la demostración es claro ver cuál es el nivel p al que nos estaremos refiriendo en cada caso. Definamos los homomorfismos

$$\phi : \mathcal{C}(A) \to \mathcal{C}(K') \oplus \mathcal{C}(K''), \quad \psi : \mathcal{C}(K') \oplus \mathcal{C}(K'') \to \mathcal{C}(K)$$

dados por

$$\phi_p(c) = (i'_*(c), i''_*(c)) \quad \forall c \in C_p(A) \text{ ,y}$$
$$\psi_p(d, e) = j'_*(d) + j''_*(e) \forall d \in C_p(K') \text{ y } e \in C_p(K'').$$

De inmediato vemos que ϕ y ψ son mapeos de cadenas. Falta únicamente ver que forman una sucesión exacta. Observemos que ϕ es inyectiva pues es una inclusión. Notemos también, que dado $d \in C_p(K)$ si $d' = d \cap K'$ entonces es claro que $d - d' \in K''$ y por lo tanto $\psi(d', d - d') = d$. Esto es, ψ es sobreyectiva. Por la Proposición A.1 sólo falta checar la exactitud del término que se encuentra en medio de la ecuación (A.1).

Consideremos $c \in C_p(A)$ y observemos que

$$\begin{aligned} \psi(\phi(c)) &= (\psi(i'_*(c), i''_*(c)) \\ &= j'_*(i'_*(c)) + j''_*(i''_*(c)) \\ &= k_*(c) + k_*(c) \\ &= 0. \end{aligned}$$

Por lo cual $\phi(c) \in \text{Ker}(\psi)$ y por lo tanto $\text{Im}(\phi) \subset \text{Ker}(\psi)$. Del mismo modo, si tomamos $(d, e) \in \text{Ker}(\psi)$ tenemos que:

$$\begin{aligned} & \phi(d, e) && = 0 \\ \Leftrightarrow \quad & j'_*(d) + j''_*(e) && = 0 \\ \Leftrightarrow \quad & d + e && = 0 \\ \Leftrightarrow \quad & d && = e. \end{aligned}$$

Como por hipótesis $d \in K'$ y $e \in K''$, tenemos que $d \in K' \cap K''$ y se sigue que $\text{Ker}(\psi) \subset \text{Im}(\phi)$. Así, la secuencia es exacta en $\mathcal{C}(K') \oplus \mathcal{C}(K'')$. Notemos que la secuencia larga que se obtiene al utilizar el lema de la Serpiente es

$$\cdots \to H_p(A) \to H_p(K' \oplus K'') \to H_p(K).$$

Debido a la definición del operador frontera, es sencillo ver que $H_p(K' \oplus K'') \simeq H_p(K') \oplus H_p(K'')$. Con esto obtenemos la cadena que buscábamos y así concluye la prueba.

□

Finalmente, se presenta la demostración de la Proposición 2.2.

Proposición 2.2 Sea X un complejo simplicial con dos subcomplejos U y V tales que $X = V \cup V$ y $U \cap V$ sea no vacía. Si U, Y y $U \cap V$ son conexos y acíclicos, entonces X también es conexo y acíclico.

Demostración

Tomemos la sucesión larga exacta de grupos que resulta al considerar el teorema de sucesiones exactas Mayer-Vietoris en las hipótesis de la proposición. Esta sucesión es

$$\cdots \to \{0\} \to \{0\} \oplus H_p\{0\} \to H_p(K) \to \{0\} \to \cdots,$$

lo cual implica que la sucesión

$$\cdots \to \{0\} \to H_p(K) \to \{0\} \to \cdots$$

es exacta para cada p. Así, por la Proposición A.1 concluimos que K es acíclico. □

Bibliografía

[1] J. Arsuaga, T. Borrman, R. Cavalcante, G. Gonzalez, and C. Park. Identification of copy number aberrations in breast cancer subtypes using persistence topology. *Microarrays*, 4(3):339–369, 2015.

[2] D. J. Balding, M. Bishop, and C. Cannings. *Handbook of Statistical Genetics*, volume 1. John Wiley & Sons, 2008.

[3] H.-J. Bandelt and A. Dress. Reconstructing the shape of a tree from observed dissimilarity data. *Advances in Applied Mathematics*, 7(3):309–343, 1986.

[4] M. Birkner, J. Blath, and B. Eldon. An ancestral recombination graph for diploid populations with skewed offspring distribution. *Genetics*, 193(1):255–290, 2013.

[5] T. C. Bruen, H. Philippe, and D. Bryant. A simple and robust statistical test for detecting the presence of recombination. *Genetics*, 172(4):2665–2681, 2006.

[6] P. Buneman. The recovery of trees from measures of dissimilarity. *Mathematics in the Archaeological and Historical Sciences*, 1971.

[7] P. Buneman. A note on the metric properties of trees. *Journal of Combinatorial Theory, Series B*, 17(1):48–50, 1974.

[8] P. G. Camara, A. J. Levine, and R. Rabadan. Inference of ancestral recombination graphs through topological data analysis. *arXiv preprint arXiv:1505.05815*, 2015.

[9] G. Carlsson. Topology and data. *Bulletin of the American Mathematical Society*, 46(2):255–308, 2009.

[10] G. Carlsson, T. Ishkhanov, V. De Silva, and A. Zomorodian. On the local behavior of spaces of natural images. *International Journal of Computer Vision*, 76(1):1–12, 2008.

[11] G. Carlsson, R. Jardine, D. Feichtner-Kozlov, D. Morozov, F. Chazal, V. de Silva, B. Fasy, J. Johnson, M. Kahle, G. Lerman, et al. Topological data analysis and machine learning theory. 2012.

[12] A. Carrieri, F. Utro, and L. Parida. Accurate and efficient sampling of underlying arg of multiple populations under subdivision and admixture. *Manuscript Under Submission*, 2015.

[13] J. Chan. *Network and Algebraic Topology of Influenza Evolution Phd Dissertation*. Columbia University, 2013.

[14] J. M. Chan, G. Carlsson, and R. Rabadan. Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571, 2013.

[15] F. Chazal, D. Cohen-Steiner, L. J. Guibas, F. Mémoli, and S. Y. Oudot. Gromov-hausdorff stable signatures for shapes using persistence. In *Computer Graphics Forum*, volume 28, pages 1393–1403. Wiley Online Library, 2009.

[16] M. K. Chung, P. Bubenik, and P. T. Kim. Persistence diagrams of cortical surface data. In *Information Processing in Medical Imaging*, pages 386–397. Springer, 2009.

[17] A. Collins, A. Zomorodian, G. Carlsson, and L. J. Guibas. A barcode shape descriptor for curve point cloud data. *Computers & Graphics*, 28(6):881–894, 2004.

[18] W. F. Doolittle. Phylogenetic classification and the universal tree. *Science*, 284(5423):2124–2128, 1999.

[19] V. G. Dugan, R. Chen, D. J. Spiro, N. Sengamalay, J. Zaborsky, E. Ghedin, J. Nolting, D. E. Swayne, J. A. Runstadler, G. M. Happ, et al. The evolutionary genetics and emergence of avian influenza viruses in wild birds. *PLoS Pathog*, 4(5):e1000076, 2008.

[20] H. Edelsbrunner and J. Harer. Persistent homology-a survey. *Contemporary Mathematics*, 453:257–282, 2008.

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

[22] H. Edelsbrunner, D. Letscher, and A. Zomorodian. Topological persistence and simplification. *Discrete and Computational Geometry*, 28(4):511–533, 2002.

[23] K. Emmett and R. Rabadan. Quantifying reticulation in phylogenetic complexes using homology. *arXiv preprint arXiv:1511.01429*, 2015.

[24] K. Emmett, D. Rosenbloom, P. Camara, and R. Rabadan. Parametric inference using persistence diagrams: A case study in population genetics. *arXiv preprint arXiv:1406.4582*, 2014.

[25] K. J. Emmett and R. Rabadan. Characterizing scales of genetic recombination and antibiotic resistance in pathogenic bacteria using topological data analysis. In *Brain Informatics and Health*, pages 540–551. Springer, 2014.

[26] P. L. Erdös, M. A. Steel, L. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees: Part ii. *Theoretical Computer Science*, 221(1):77–118, 1999.

[27] P. L. Erdos, M. A. Steel, L. A. Székely, and T. J. Warnow. A few logs suffice to build (almost) all trees (i). *Random Structures and Algorithms*, 14(2):153–184, 1999.

[28] M. E. Espinoza Lara. *Notas en Homología Persistente*. CIMAT., 2015.

[29] G. et al. Human infection with a novel avian-origin influenza a (h7n9) virus. *New England Journal of Medicine*, 368(20):1888–1897, 2013. PMID: 23577628.

[30] W. J. Ewens. *Mathematical population genetics. I. Theoretical introduction. Interdisciplinary Applied Mathematics, 27*. Springer-Verlag, New York, 2004.

[31] W. J. Ewens and G. R. Grant. *Statistical methods in bioinformatics: an introduction*. Springer Science & Business Media, 2006.

[32] B. T. Fasy, F. Lecci, A. Rinaldo, L. Wasserman, S. Balakrishnan, and A. Singh. Confidence sets for persistence diagrams. *The Annals of Statistics*, 42(6):2301–2339, 2014.

[33] R. Ghrist. Barcodes: the persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.

[34] P. Giblin. *Graphs, surfaces and homology: An introduction to algebraic topology*. Springer Science & Business Media, 2013.

[35] R. C. Griffiths and P. Marjoram. An ancestral recombination graph. *Institute for Mathematics and its Applications*, 87:257, 1997.

[36] B. G. Hall. *Phylogenetic trees made easy: a how-to manual*, volume 547. Sinauer Associates Sunderland, 2004.

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

[38] D. Holmes. The world waits for h7n9 to yield up its secrets. *The Lancet Infectious Diseases*, 13(6):477–478, 2015/11/03.

[39] R. R. Hudson. Generating samples under a wright–fisher neutral model of genetic variation. *Bioinformatics*, 18(2):337–338, 2002.

[40] D. Husmeier, R. Dybowski, and S. Roberts. *Probabilistic modeling in bioinformatics and medical informatics*. Springer Science & Business Media, 2006.

[41] D. H. Huson, R. Rupp, and C. Scornavacca. *Phylogenetic networks: concepts, algorithms and applications*. Cambridge University Press, 2010.

[42] A. M. Ibekwe, J. Ma, D. E. Crowley, C.-H. Yang, A. M. Johnson, T. C. Petrossian4, and P. Y. Lum. Topological data analysis of escherichia coli o157:h7 and non-o157 survival in soils. *Frontiers in Cellular and Infection Microbiology*, 4(122), 2014.

[43] R. M. Karp. Mathematical challenges from genomics and molecular biology. *Notices of the American Mathematical Society*, 49(5):544–553, 2002.

[44] D. Kozlov. *Combinatorial algebraic topology*, volume 21. Springer Science & Business Media, 2007.

[45] P. Lemey. *The phylogenetic handbook: a practical approach to phylogenetic analysis and hypothesis testing*. Cambridge University Press, 2009.

[46] V. Makarenkov and P. Legendre. From a phylogenetic tree to a reticulated network. *Journal of Computational Biology*, 11(1):195–212, 2004.

[47] C. R. F. Maunder. *Algebraic topology*. Courier Corporation, 1996.

[48] K. Mischaikow, M. Mrozek, and P. Pilarczyk. Graph approach to the computation of the homology of continuous maps. *Foundations of Computational Mathematics*, 5(2):199–229, 2005.

[49] K. Mischaikow and V. Nanda. Morse theory for filtrations and efficient computation of persistent homology. *Discrete & Computational Geometry*, 50(2):330–353, 2013.

[50] D. A. Morrison. Networks in phylogenetic analysis: new tools for population biology. *International Journal for Parasitology*, 35(5):567–582, 2005.

[51] J. R. Munkres. *Elements of algebraic topology*, volume 2. Addison-Wesley Reading, 1984.

[52] M. Nicolau, A. J. Levine, and G. Carlsson. Topology based data analysis identifies a subgroup of breast cancers with a unique mutational profile and excellent survival. *Proceedings of the National Academy of Sciences*, 108(17):7265–7270, 2011.

[53] L. Parida. Ancestral recombinations graph: a reconstructability perspective using random-graphs framework. *Journal of Computational Biology*, 17(10):1345–1370, 2010.

[54] L. Parida, F. Utro, D. Yorukoglu, A. P. Carrieri, D. Kuhn, and S. Basu. Topological signatures for population admixture. In *Research in Computational Molecular Biology*, pages 261–275. Springer, 2015.

[55] J. S. Pereira. A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6(3):303–310, 1969.

[56] D. Posada and K. A. Crandall. Evaluation of methods for detecting recombination from dna sequences: computer simulations. *Proceedings of the National Academy of Sciences*, 98(24):13757–13762, 2001.

[57] M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. 2014.

[58] V. Robins. Towards computing homology from finite approximations. In *Topology Proceedings*, volume 24, pages 503–532, 1999.

[59] C. Semple and M. A. Steel. *Phylogenetics*, volume 24. Oxford University Press, 2003.

[60] M. Steel. Tracing evolutionary links between species. *The American Mathematical Monthly*, 121(9):771–792, 2014.

[61] A. Tausz, M. Vejdemo-Johansson, and H. Adams. JavaPlex: A research software package for persistent (co)homology. In H. Hong and C. Yap, editors, *Proceedings of ICMS 2014*, Lecture Notes in Computer Science 8592, pages 129–136, 2014. Software available at http://appliedtopology.github.io/javaplex/.

[62] M. Vejdemo-Johansson. Sketches of a platypus: persistent homology and its algebraic foundations. arxiv. *Contemporary Mathematics*, 620, 2014.

[63] J. Wakeley. *Coalescent theory: an introduction*, volume 1. Roberts & Company Publishers Greenwood Village, Colorado, 2009.

[64] K. Zaretskii. Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6):90–92, 1965.

[65] A. Zomorodian. Topological data analysis. *Advances in Applied and Computational Topology: American Mathematical Society Short Course on Computational Topology, January 4-5, 2011, New Orleans, Louisiana*, 70:1–40, 2012.

[66] A. J. Zomorodian. *Topology for computing*, volume 16. Cambridge University Press, 2005.