



Universidad Nacional Autónoma de
México y Universidad Michoacana
de San Nicolás de Hidalgo



Posgrado Conjunto en Ciencias Matemáticas
UMSNH–UNAM

Análisis Topológico de Datos para la Caracterización de Patrones Celulares en Tejidos Epiteliales

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Ciencias Matemáticas

PRESENTA:

Luis Raúl Figueroa Martínez

ASESORES:

Dr. Eugenio Martín Azpeitia Espinosa
Centro de Ciencias Matemáticas, UNAM

Dra. Adriana Haydeé Contreras Peruyero
Centro de Ciencias Matemáticas, UNAM

Morelia, Michoacán, México
Enero, 2026

Agradecimientos

Quiero expresar mi más sincero agradecimiento al **Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT)** de la **Universidad Nacional Autónoma de México (UNAM)**, por el apoyo otorgado a través de los proyectos **IN114323**, “*Grupos modulares de superficies y aplicaciones de topología a análisis de datos*”, y **IA204624**, “*Desarrollo de un modelo matemático computacional para el estudio del destino celular mediado por p53 y asociado a la cantidad de daño en el DNA*”.

Agradezco profundamente a la **Dra. Nelly Selem Mojica** por permitirme hacer uso del servidor *Elisa*, en el cual procesé una parte importante de mis datos. Asimismo, extiendo mi reconocimiento al **M. en C. Luis Gerardo Tejero Gómez** y al **M. en C. Miguel Ángel Magaña Lemus**, responsables de la administración y mantenimiento del servidor, por su apoyo y disposición durante todo el proceso.

Mi agradecimiento también a **SECIHTI** por el apoyo otorgado mediante la beca durante mis estudios de maestría, que me permitió dedicarme plenamente a la investigación a lo largo de los dos años del programa.

Deseo expresar mi más profundo agradecimiento a mis asesores, el **Dr. Eugenio Martín Azpeitia Espinosa** y la **Dra. Adriana Haydeé Contreras Peruyero**, quienes desde el principio de mi maestría me dieron su apoyo incondicional. Su guía, **PACIENCIA** y acompañamiento constante hicieron posible la realización de esta tesis. Sin ninguno de los dos, estoy seguro de que no hubiera sido posible completar este trabajo. Siempre estaré agradecido por todo lo que han hecho por mí.

Quiero expresar un agradecimiento especial a la **Lic. en Mat. Lizbeth Karina Guzmán Pedroza**, quien me motivó a realizar la maestría y me brindó un apoyo constante durante todo este proceso, gracias a ella pude hacer el examen de admisión y el resto es historia.

A toda mi familia, por su apoyo incondicional, cariño y confianza. Gracias a ustedes estoy donde estoy, y por ustedes he logrado avanzar en cada etapa de mi formación académica y personal.

Índice general

Resumen	iv
Abstract	v
1. Cuerpo de la Revisión	1
1.1. Métodos tradicionales en el análisis de imágenes biológicas	3
1.2. Introducción al Análisis Topológico de Datos	6
1.2.1. Complejo de Vietoris-Rips	10
1.2.2. Homología persistente	11
1.2.3. Aplicaciones del TDA en imágenes biológicas	22
1.2.4. Ventajas del enfoque topológico	23
1.3. La anemia de Fanconi como modelo de estudio del desarrollo de carcinomas	24
1.4. Resumen y planteamiento de la pregunta de investigación	26
2. Métodos	28
2.1. Descripción general del enfoque metodológico	28
2.2. Datos: imágenes <i>t</i> -CyCIF	31
2.3. Preprocesamiento de los datos	35
2.3.1. Limpieza y estandarización de coordenadas	36
2.3.2. Visualización espacial de las células	36
2.4. Construcción de complejos simpliciales y obtención de diagramas de per-	
sistencia	37
2.4.1. Cota máxima del parámetro de filtración (ε)	38
2.4.2. Dimensiones de los grupos de homología analizados	38
2.4.3. Ejemplo de los gráficos de salida	39
2.5. Cálculo de distancias entre diagramas de persistencia	41
2.5.1. Distancia de Wasserstein y cálculo computacional	41
2.5.2. Justificación del uso de la distancia de Wasserstein	41
2.6. Agrupamiento jerárquico y visualización	42

2.6.1. Procedimiento de agrupamiento	42
2.6.2. Generación de <i>clustermaps</i>	44
2.7. Análisis de diferencias entre grupos biológicos mediante TopKAT	44
3. Resultados	48
3.1. Análisis global de patrones topológicos	49
3.1.1. Construcción de complejos de Rips y diagramas de persistencia . . .	49
3.1.2. Cálculo de matrices de distancia	50
3.1.3. Clustering jerárquico	55
3.2. Análisis por grupos celulares	58
3.2.1. Procedimiento topológico por grupos	58
3.3. Análisis complementario con TopKAT	64
4. Discusión	71
Bibliografía	77

Resumen

El análisis de imágenes es una herramienta fundamental en biología, ya que permite estudiar propiedades celulares y tisulares relevantes para comprender el estado y funcionamiento de los sistemas biológicos. En particular, ha sido ampliamente utilizado para la identificación de biomarcadores, la segmentación celular y la clasificación de tejidos. Sin embargo, los métodos tradicionales presentan limitaciones cuando se enfrentan a tejidos con alta densidad celular, formas irregulares o disposiciones espaciales complejas, como ocurre en tumores, tejidos neuronales y datos derivados de ómicas espaciales. Estas técnicas generan imágenes de alta dimensionalidad que requieren enfoques capaces de capturar no solo características individuales de las células, sino también su organización espacial y conectividad.

En este contexto, el Análisis Topológico de Datos (TDA, por sus siglas en inglés) surge como una alternativa prometedora para el estudio de patrones celulares complejos. A diferencia de los métodos convencionales, el TDA permite analizar la estructura global de los datos, capturando relaciones espaciales y patrones de conectividad mediante herramientas de la topología algebraica, como la homología persistente y los complejos de Vietoris–Rips. Estas características hacen del TDA una herramienta adecuada para estudiar tejidos con alta heterogeneidad y organización espacial intrincada.

El objetivo de esta investigación es evaluar la utilidad del TDA en la clasificación de patrones celulares en datos de ómicas espaciales, con especial énfasis en tejidos afectados por Anemia de Fanconi, una enfermedad genética asociada a alteraciones en la reparación del ADN y a una mayor predisposición al cáncer. Se busca determinar si las representaciones topológicas permiten identificar diferencias estructurales relevantes entre tejidos normales y afectados, contribuyendo al desarrollo de enfoques más robustos para el análisis automatizado de imágenes biológicas en contextos complejos.

Palabras clave: topología algebraica, homología persistente, complejos simpliciales, análisis de datos espaciales, Anemia de Fanconi.

Abstract

Image analysis is a fundamental tool in biology, as it enables the study of cellular and tissue properties that are essential for understanding the state and function of biological systems. It has been widely applied to tasks such as biomarker identification, cell segmentation, and tissue classification. However, traditional methods face important limitations when dealing with tissues exhibiting high cellular density, irregular shapes, or complex spatial arrangements, as observed in tumors, neuronal tissues, and data derived from spatial omics. These techniques generate high-dimensional images that require approaches capable of capturing not only individual cellular features but also their spatial organization and connectivity.

In this context, Topological Data Analysis (TDA) has emerged as a promising alternative for the study of complex cellular patterns. Unlike conventional methods, TDA focuses on the global structure of the data, capturing spatial relationships and connectivity patterns through tools from algebraic topology, such as persistent homology and Vietoris–Rips complexes. These properties make TDA particularly suitable for the analysis of highly heterogeneous tissues with intricate spatial organization.

The aim of this research is to evaluate the usefulness of TDA for the classification of cellular patterns in spatial omics data, with a particular focus on tissues affected by Fanconi Anemia, a genetic disorder associated with defects in DNA repair and increased cancer susceptibility. This work seeks to determine whether topological representations can identify relevant structural differences between normal and affected tissues, thereby contributing to the development of more robust approaches for the automated analysis of biological images in complex contexts.

Keywords: algebraic topology, persistent homology, simplicial complexes, spatial data analysis, Fanconi anemia.

Capítulo 1

Cuerpo de la Revisión

El análisis de imágenes biológicas se ha consolidado como una herramienta fundamental en la investigación biomédica actual, al permitir la caracterización a múltiples escalas espaciales de sistemas celulares *complejos*. Con *tejidos complejos* nos referimos a tejidos con una alta densidad celular, con formas celulares irregulares y compuestos por una gran diversidad de tipos celulares sin un patrón evidente [26]. Particularmente, el análisis de imágenes de tejidos ha permitido profundizar en el estudio de la organización espacial de células en tejidos sanos y patológicos, brindando avances importantes en la comprensión de los mecanismos fisiopatológicos, la progresión de enfermedades y la respuesta a terapias [17]. Si bien los métodos tradicionales de análisis de imágenes, como la morfometría clásica, los algoritmos de segmentación y los enfoques basados en aprendizaje profundo, han sido fundamentales para el desarrollo del análisis de imágenes, presentan ciertas limitaciones cuando se aplican a *tejidos complejos* o a datos con alta dimensionalidad [44].

Estas limitaciones técnicas adquieren mayor relevancia en el contexto de la medicina de precisión o personalizada, donde la capacidad para cuantificar y clasificar patrones multicelulares podría transformar los paradigmas diagnósticos y pronósticos [23]. En especial, en los últimos años, la medicina de precisión ha desarrollado un conjunto de nuevas técnicas conocidas como *ómicas espaciales*. Las *ómicas espaciales* generan imágenes de tejidos con los perfiles proteómicos o transcriptómicos asociados, con resolución unicelular, manteniendo la organización espacial de las células en el tejido. De esta forma, estas técnicas han permitido realizar caracterizaciones de tejidos extremadamente detalladas, que incluyen, además de la organización celular, los tipos celulares que las componen y la variabilidad en la expresión genética entre las células que conforman el tejido. Al contener tal cantidad de detalles, las imágenes generadas por las *ómicas espaciales* contienen enormes volúmenes de información. Aunado a esto, estas técnicas son comúnmente empleadas en *tejidos complejos*. Por estos motivos, los métodos tradicionales de análisis de imágenes

se han visto limitados al tratar de analizar este tipo de datos [24, 45, 19, 28].

Es en este escenario donde el TDA emerge como un marco matemático innovador, capaz de capturar características globales de la organización tisular, entendida como la disposición espacial de las células, sus relaciones de vecindad y la arquitectura colectiva que conforman en el tejido. Estas propiedades trascienden las mediciones geométricas convencionales y permiten superar varias limitaciones de los métodos tradicionales, ya que el TDA puede lidiar con grandes cantidades de información, alta densidad celular y formas celulares irregulares [8]. El objetivo de este trabajo es explorar la capacidad del TDA para estudiar y analizar imágenes de tejidos generadas mediante una técnica de ómica espacial llamada inmunofluorescencia cíclica en tejido (*t*-CyCIF por sus siglas en inglés).

Esta revisión se organiza en cuatro secciones principales que reflejan los ejes conceptuales de la investigación:

1. La primera sección examina las limitaciones fundamentales de los métodos tradicionales de análisis de imágenes biológicas, con énfasis en su incapacidad para caracterizar la topología de sistemas multicelulares “complejos”.
2. La segunda sección introduce los fundamentos matemáticos del TDA y sus aplicaciones pioneras en biomedicina. Se discuten conceptos como homología persistente y complejos de Vietoris-Rips, ilustrando su implementación en problemas de clasificación de tejidos y análisis de imágenes médicas.
3. En la tercera sección hablaremos acerca de estudios recientes que emplean TDA para clasificación de patrones celulares. Por ejemplo en tejidos generados *in silico* [5], estudios en cáncer [46] entre otros.
4. La cuarta sección explora los entornos específicos donde se aplicará el TDA en esta investigación. En particular, se detallan los análisis basados en datos derivados de tecnologías de ómicas espaciales de nueva generación (como *t*-CyCIF), las cuales generan imágenes con resolución unicelular que preservan la organización espacial de los tejidos. Estas imágenes representan un escenario realista y desafiante para evaluar el poder del TDA en contextos biomédicos complejos.

En conjunto, estas secciones proporcionan el marco teórico necesario para comprender la relevancia del TDA como herramienta de extracción de características topológicas en imágenes biológicas complejas, estableciendo así las bases conceptuales del presente estudio.

1.1. Métodos tradicionales en el análisis de imágenes biológicas

Los métodos convencionales para el análisis de imágenes de tejidos biológicos se han basado principalmente en técnicas de morfometría, que, aunque han sido ampliamente utilizadas, presentan importantes limitaciones al estudiar *tejidos complejos* [17]. Estos enfoques se centran en medir características geométricas básicas, como el tamaño nuclear, la relación núcleo-citoplasma y la forma celular, así como parámetros de textura basados en la intensidad de los píxeles [20]. En aplicaciones clínicas rutinarias, como el análisis de muestras de sangre o la evaluación de biopsias simples, estos métodos han demostrado ser útiles, alcanzando precisiones del 85–92 % en la clasificación de tipos celulares [20]. Sin embargo, cuando se enfrentan a tejidos con alta densidad celular o estructuras complejas, su eficacia disminuye considerablemente [27]. Por ejemplo, en tumores sólidos como los glioblastomas, los errores en la caracterización celular pueden alcanzar entre el 35 % y el 60 % [6], debido principalmente a la superposición de núcleos y a la pérdida de bordes celulares en zonas de contacto estrecho [27]. Las limitaciones se vuelven aún más evidentes al estudiar células con morfologías irregulares, como las células metastásicas o las neuronas con arborizaciones dendríticas. En estos casos, los descriptores geométricos tradicionales resultan insuficientes para caracterizar adecuadamente la estructura celular y tisular [9]. Además, estos métodos muestran una marcada incapacidad para analizar patrones arquitecturales complejos, como los que se observan en los microambientes tumorales o en las redes vasculares tridimensionales [22].

A éstas limitaciones se suma la dificultad de trabajar con datos de alta dimensionalidad. Por ejemplo, en imágenes biológicas modernas, especialmente aquellas generadas mediante *ómicas espaciales*, es posible caracterizar varias propiedades de cada célula de forma simultánea, incluyendo la concentración de distintas proteínas (marcadores proteicos) o la expresión de diferentes genes (información transcriptómica). La concentración de cada proteína o el nivel de expresión de cada gen se puede considerar como una dimensión adicional en un espacio de características que describe a cada una de las células que componen el tejido. Así, aunque geoméricamente una imagen sea bidimensional, cada célula tiene asociada un vector de decenas o incluso cientos de dimensiones. Esta alta dimensionalidad añade complejidad al análisis computacional y pone en evidencia la necesidad de métodos capaces de integrar información multivariada.

Más allá de sus limitaciones geométricas y dimensionales, los métodos tradicionales no capturan relaciones multivariantes ni dependencias espaciales entre células. Esto provoca una pérdida de información sobre la organización global del tejido, dificultando la identificación de patrones multicelulares que podrían ser de relevancia clínica [26, 24]. Esta

limitación cobra especial importancia cuando se busca explotar la riqueza de los datos de *ómicas espaciales*, que combinan alta resolución espacial con múltiples propiedades celulares, generando conjuntos de datos extremadamente grandes para los cuales los métodos tradicionales de análisis de imágenes no están diseñados.

Además de las limitaciones estructurales y de complejidad celular, los métodos tradicionales enfrentan dificultades significativas al manejar grandes cantidades de información. En estudios modernos de tejidos, cada muestra puede contener miles de células y múltiples regiones de interés, e incluso se analizan decenas o cientos de imágenes por paciente. Si bien varios enfoques convencionales pueden procesar estos datos a nivel local, no están diseñados para integrar y resumir de manera efectiva la organización espacial global de los tejidos, lo que dificulta la identificación de patrones multicelulares relevantes y biomarcadores espaciales críticos [26, 24, 19].

Frente a estas limitaciones, ha surgido la necesidad de desarrollar nuevos enfoques analíticos que permitan caracterizar la estructura espacial y topológica de los sistemas biológicos, capturando relaciones multiescala que no son evidentes mediante descriptores locales.

Como evolución natural de los métodos tradicionales, las técnicas de aprendizaje automático y en particular las redes neuronales convolucionales (CNNs) han emergido como herramientas poderosas para el análisis de imágenes biológicas. Estos enfoques superan muchas de las limitaciones de la morfometría clásica al aprender patrones directamente de los datos, sin depender exclusivamente de características predefinidas. Las CNNs han demostrado un notable éxito en aplicaciones como la clasificación automatizada de tipos celulares y la detección de anomalías morfológicas, alcanzando precisiones que superan el 90 % en condiciones controladas. Su capacidad para capturar patrones complejos en múltiples escalas las hace particularmente adecuadas para analizar tejidos con estructuras irregulares, donde los métodos tradicionales fallan. Por ejemplo, en el análisis de tumores cerebrales, las CNNs pueden identificar subtipos histológicos con una sensibilidad del 85-92 %, superando significativamente los enfoques basados en morfometría convencional.

Sin embargo, estos métodos presentan sus propias limitaciones. Requieren grandes conjuntos de datos etiquetados para entrenamiento, son vulnerables a sesgos en los datos de entrada y funcionan como “cajas negras”, proporcionando poca información sobre qué características específicas están utilizando para sus clasificaciones. Además, aunque pueden detectar patrones complejos, no cuantifican explícitamente las relaciones en que las células están distribuidas y organizadas unas respecto a otras, más allá de su forma individual. En otras palabras, las CNN no pueden cuantificar las propiedades topológicas de los tejidos, las cuales son cruciales para entender la organización tisular. Esto incluye, por ejemplo, conjuntos de células que se agrupan o se encuentran próximas, la formación

de bucles o cavidades (zonas vacías rodeadas por células), y patrones generales que se repiten a lo largo del tejido. Analizar las relaciones topológicas permite caracterizar la estructura global del tejido, algo que los descriptores locales o incluso las CNNs no capturan directamente. Es precisamente en esta última limitación donde los métodos basados en topología ofrecen una alternativa complementaria. Mientras que las CNNs se enfocan en patrones locales de intensidad y textura, el análisis topológico proporciona herramientas matemáticas para cuantificar propiedades globales como conectividad, agrupamiento y patrones espaciales a gran escala. Esta diferencia fundamental sugiere que la integración de ambos enfoques podría ofrecer una solución más completa, combinando la potencia del aprendizaje profundo para el reconocimiento de patrones con la capacidad de la topología para caracterizar la organización estructural de los sistemas biológicos.

Este marco conceptual nos lleva a explorar el análisis topológico de datos, un área de las matemáticas que combina conceptos de topología algebraica y geometría para estudiar la estructura intrínseca de los datos. El TDA permite cuantificar de manera rigurosa la conectividad, los agrupamientos y los patrones espaciales de las células en un tejido, superando las limitaciones de los métodos tradicionales y complementando la información extraída por las redes neuronales convolucionales (CNNs). A continuación, se describe en qué consiste el análisis topológico de datos.

1.2. Introducción al Análisis Topológico de Datos

Como se describió en la sección previa, el surgimiento del TDA para el análisis de imágenes de tejidos biológicos representa una respuesta a las limitaciones de los enfoques tradicionales que existen en el área. Frente a la dificultad de caracterizar la estructura de grandes bases de datos y de formas (celulares y tisulares) complejas mediante herramientas métricas o geométricas convencionales, el TDA ofrece un marco matemático que permite capturar la “forma” de los datos, entendida como su estructura global de conectividad y organización.

En esta sección se presentan los fundamentos teóricos del TDA y se revisan ejemplos de implementación de estas herramientas en múltiples entornos. Para las definiciones, teoremas y resultados formales que se emplearán, así como para una exposición más detallada del marco teórico, se puede consultar [8, 36].

Supongamos que se tiene un conjunto de puntos $\{x_0, \dots, x_k\}$ en $\mathbb{R}^n$. Se asume que estos puntos satisfacen la condición de que el conjunto de vectores en $\mathbb{R}^n$ representados por las diferencias $\{x_1 - x_0, \dots, x_k - x_0\}$ son linealmente independientes.

Definición 1. El k -símplice generado por los puntos $\{x_0, \dots, x_k\}$ es el conjunto de todos los puntos $z = \sum_{i=0}^k a_i x_i$ donde $a_i \geq 0$ y $\sum_{i=0}^k a_i = 1$. Para un z dado, nos referimos a a_i como la i -ésima coordenada baricéntrica. El k -símplice generado por los puntos $\{x_0, \dots, x_k\}$ se denota por $\langle x_0, \dots, x_k \rangle$. La *dimensión* de un k -símplice se define como k , es decir, el número de vértices que lo generan menos uno. En la Figura 1.1 se muestran ejemplos de símlices en distintas dimensiones.

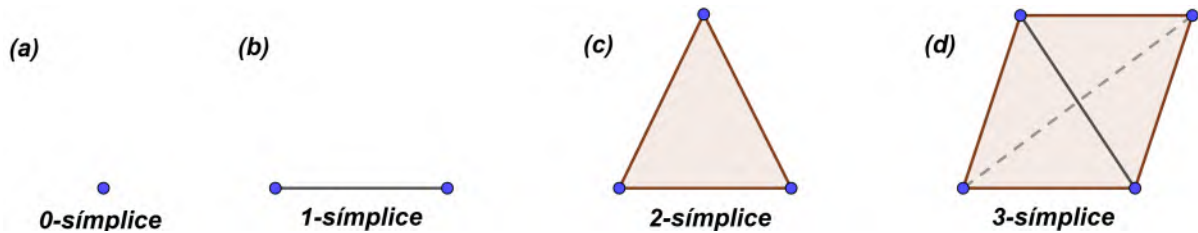


Figura 1.1: Ejemplos de símlices en distintas dimensiones: (a) 0-símplice (vértice), (b) 1-símplice (arista), (c) 2-símplice (triángulo) y (d) 3-símplice (tetraedro). Cada símplex es la envolvente convexa de sus vértices.

Definición 2. El *interior* de un símplex S generado por los puntos $x_0, \dots, x_k$, denotado por $\text{int}(S)$, es el conjunto de todos los puntos cuyas coordenadas baricéntricas a_i son

estrictamente positivas, es decir,

$$\text{int}(S) = \left\{ \sum_{i=0}^k a_i x_i \mid a_i > 0, \sum_{i=0}^k a_i = 1 \right\}.$$

La *frontera* del símple, denotada por ∂S , está formada por todos los puntos de S para los cuales al menos una de sus coordenadas baricéntricas es cero, es decir,

$$\partial S = \left\{ \sum_{i=0}^k a_i x_i \mid a_i \geq 0, \sum_{i=0}^k a_i = 1, \exists i \text{ tal que } a_i = 0 \right\}.$$

La Figura 1.2 ilustra un ejemplo de un símple y la distinción entre su interior y su frontera.

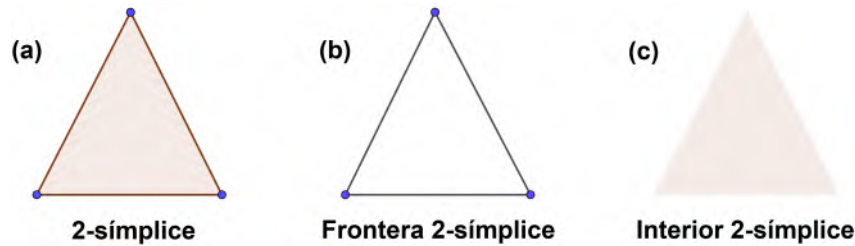


Figura 1.2: Ejemplo de un 2-símple (triángulo), su interior y frontera: (a) símple completo, que incluye interior y frontera; (b) frontera del símple (∂S), correspondiente a los lados y vértices; (c) interior del símple ($\text{int}(S)$), formado por los puntos cuyas coordenadas baricéntricas son estrictamente positivas.

Definición 3. Para un símple S generado por los puntos $P = \{x_0, \dots, x_k\}$, una *cara* de S se refiere a cualquier símple generado por un subconjunto no vacío de P .

Definición 4. Un *complejo simplicial* X en $\mathbb{R}^n$ es un conjunto de símlices en $\mathbb{R}^n$ tal que:

1. Cada cara de un símple en X también pertenece a X .
2. La intersección de dos símlices cualesquiera en X , si no es vacía, es una cara común de ambos.

La *dimensión* de un complejo simplicial X se define como la dimensión máxima de los símlices que lo componen.

Ejemplo 5. A continuación se muestra un ejemplo de un complejo simplicial en $\mathbb{R}^2$. Sea X el complejo simplicial definido como:

$$X = \{\langle P \rangle, \langle S \rangle, \langle N \rangle, \langle O \rangle, \langle Q \rangle, \langle T \rangle, \langle R \rangle, \langle P, S \rangle, \langle P, N \rangle, \langle P, Q \rangle, \langle P, T \rangle, \langle S, N \rangle, \langle N, T \rangle, \langle N, O \rangle, \langle O, Q \rangle, \langle O, T \rangle, \langle O, R \rangle, \langle Q, T \rangle, \langle Q, R \rangle, \langle N, O, T \rangle, \langle O, Q, T \rangle\}.$$

La Figura 1.3 muestra la representación geométrica del complejo, donde se observan los vértices (símplices de dimensión 0), las aristas (símplices de dimensión 1) y los triángulos (símplices de dimensión 2). En total, el complejo está formado por 7 símplexes de dimensión 0, 12 símplexes de dimensión 1 y 2 símplexes de dimensión 2.

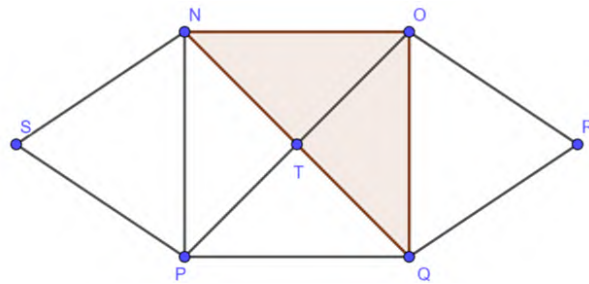


Figura 1.3: Ejemplo de un complejo simplicial en $\mathbb{R}^2$. Se muestran los símplexes de dimensión 0, 1 y 2 que conforman el complejo, correspondientes a vértices, aristas y triángulos, respectivamente.

Definición 6. Sea X un conjunto, un *recubrimiento* de X es una colección de conjuntos no vacíos $\mathcal{C} = \{C_\alpha\}$ tal que la unión de todos ellos contiene completamente a X , es decir, $X \subset \bigcup_\alpha C_\alpha$.

Definición 7. Sea X un conjunto y $\mathcal{C} = \{C_\alpha\}$ un recubrimiento de X . El *nervio* de $\mathcal{C}$, denotado $K(\mathcal{C})$, es el complejo simplicial cuyos símplexes son todos los subconjuntos finitos no vacíos de $\mathcal{C}$ cuya intersección no es vacía.

Más precisamente, un símplex $s \in K(\mathcal{C})$ tiene la forma

$$s = \{C_{i_1}, C_{i_2}, \dots, C_{i_n}\} \subseteq \mathcal{C}, \quad \text{con} \quad \bigcap_{k=1}^n C_{i_k} \neq \emptyset.$$

En otras palabras, la construcción del nervio se realiza de la siguiente manera: primero, cada conjunto del recubrimiento representa un vértice (0-símplex). Cuando dos conjuntos del recubrimiento se intersectan, se crea una arista (1-símplex) entre sus vértices correspondientes. Si tres conjuntos tienen una intersección común, sus vértices forman un triángulo (2-símplex). Este proceso continúa de manera similar para intersecciones de más conjuntos, generando símplexes de dimensiones superiores.

Ejemplo 8. Consideremos $X = \{A, B, C, D, E\}$ un conjunto de puntos con un recubrimiento $\mathcal{C} = \{a, b, c, d, e\}$, como se muestra en la Figura 1.4. El nervio del complejo simplicial asociado es:

$$\mathcal{N}(\mathcal{C}) = \{\langle A \rangle, \langle B \rangle, \langle C \rangle, \langle D \rangle, \langle E \rangle, \langle C, E \rangle, \langle C, B \rangle, \langle B, D \rangle, \langle E, B \rangle, \langle A, B \rangle, \langle B, C, E \rangle\}$$

Como se observa en la figura, el nervio refleja la estructura de intersecciones del recubrimiento.

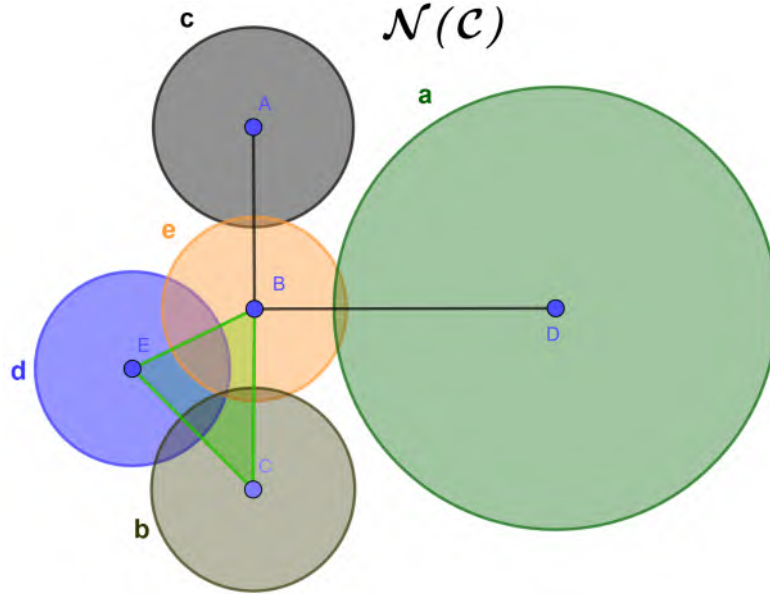


Figura 1.4: Ejemplo de construcción del nervio a partir de un recubrimiento. Cada conjunto del recubrimiento corresponde a un vértice (0-símplice). Si dos conjuntos se intersectan, se genera una arista (1-símplice) entre los vértices correspondientes. Cuando tres conjuntos comparten intersección, forman un triángulo (2-símplice).

Teorema 9 (Teorema del Nervio [21, Corolario 4G.3]). Sea X un espacio topológico paracompacto y $\mathcal{U} = \{U_i\}_{i \in A}$ un recubrimiento por abiertos de X , donde A es un conjunto finito o numerable. Si para toda familia finita no vacía $S \subseteq A$, la intersección $\bigcap_{s \in S} U_s$ es contráctil (o vacía), entonces el nervio $N(\mathcal{U})$ es homotópicamente equivalente a X .

El Teorema anterior garantiza que, bajo ciertas condiciones, el nervio de un recubrimiento abierto de un espacio topológico es homotópicamente equivalente al espacio original. Esto es relevante en este trabajo, ya que permite modelar los patrones celulares mediante complejos simpliciales contruidos a partir de centroides y relaciones de proximidad, asegurando que la información topológica esencial del espacio original se preserve en los análisis de persistencia.

Con las definiciones anteriores podemos definir la construcción de uno de los complejos simpliciales más utilizados dentro de lo que es el análisis de datos que usaremos a lo largo de esta tesis.

1.2.1. Complejo de Vietoris-Rips

Definición 10 (Complejo de Vietoris-Rips). Sea (X, d) un espacio métrico finito y $\varepsilon > 0$ un parámetro. El complejo de Vietoris-Rips con parámetro ε , denotado por $VR_\varepsilon(X)$, es el complejo simplicial que tiene como vértices los puntos de X , y donde un subconjunto finito $\{x_0, x_1, \dots, x_k\} \subseteq X$ genera un k -símplice si la distancia entre cada par de sus puntos satisface

$$d(x_i, x_j) \leq 2\varepsilon, \quad \text{para todo } 0 \leq i, j \leq k.$$

Intuitivamente, el complejo de Vietoris-Rips conecta todos los puntos que están a una distancia menor o igual que 2ε , y añade símlices cuando hay conjuntos de puntos que están mutuamente conectados por esa distancia.

En la Figura 1.5 se muestra un ejemplo con ocho puntos, a cada uno de los cuales se le asignó un radio fijo. Como resultado, se generan nueve 1-símlices (aristas) y dos 2-símlices (triángulos) que ilustran cómo surge la estructura simplicial a partir de las relaciones de proximidad.

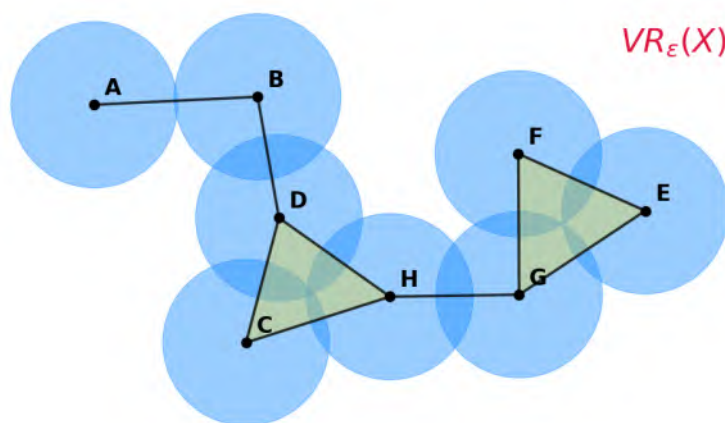


Figura 1.5: Ejemplo de construcción de un complejo de Vietoris-Rips con $\varepsilon = 2.1$ sobre un conjunto de ocho puntos. Se forman nueve 1-símlices (aristas) y dos 2-símlices (triángulos).

1.2.2. Homología persistente

La homología persistente es un método del Análisis Topológico de Datos (TDA) que permite estudiar las características cualitativas de los datos que persisten a través de múltiples escalas. Este enfoque proporciona una representación compacta y robusta de las propiedades topológicas de un conjunto de datos, siendo insensible a perturbaciones, dimensiones o coordenadas específicas. Además, la homología persistente constituye un área activa de investigación, tanto en su formulación teórica como en el desarrollo de algoritmos y software especializados para su cómputo [31, 8, 12]. En este trabajo no se profundiza en los aspectos algebraicos formales de la homología persistente, centrándose en su interpretación geométrica y en su aplicación al análisis de datos espaciales. Para una exposición detallada de la teoría subyacente y de sus propiedades, se remite al lector a [16, 14].

Definición 11. Sea S un complejo simplicial. Una *filtración* $\mathcal{F}$ de S es una familia de complejos simpliciales $\{S_i\}_{i \in I}$, indexada por un conjunto totalmente ordenado I , tal que

$$S_i \subseteq S_j \quad \text{si } i \leq j.$$

Un modo de obtener una filtración a partir de un conjunto de puntos es considerar, para cada $\varepsilon \geq 0$, el complejo de Vietoris–Rips $VR_\varepsilon(X)$. Dado un conjunto de puntos (X, d) y un parámetro $\varepsilon > 0$, el complejo $VR_\varepsilon(X)$ tiene como vértices los puntos de X y contiene un k -símplice $\{x_0, \dots, x_k\}$ siempre que $d(x_i, x_j) \leq 2\varepsilon$ para todo i, j . Si $\varepsilon \leq \varepsilon'$, entonces

$$VR_\varepsilon(X) \subseteq VR_{\varepsilon'}(X),$$

por lo que la familia $\{VR_\varepsilon(X)\}_{\varepsilon \geq 0}$ constituye una filtración. En la práctica, esta filtración continua suele discretizarse mediante una sucesión finita de valores $\varepsilon_1 < \varepsilon_2 < \dots < \varepsilon_n$, lo que permite identificarla con una filtración indexada por enteros del tipo $\{S_i\}_{i \in \mathbb{N}}$.

En la Figura 1.6 se ilustra este proceso tomando como referencia el conjunto de puntos mostrado en la Figura 1.5. Los vértices (los puntos) están presentes desde $\varepsilon = 0$; al incrementar el radio ε aparecen aristas entre pares de puntos cuya distancia satisface la condición, y para valores mayores de ε se forman símlices de dimensión superior (triángulos, etc.).

A cada nivel de la filtración se le puede asociar un grupo de homología $H_k(S_i)$, que describe las clases de k -agujeros presentes en dicha escala. Gracias a la inclusión $S_i \subseteq S_j$ para $i \leq j$, existe un morfismo inducido en homología

$$f_{i,j}^k : H_k(S_i) \rightarrow H_k(S_j),$$

el cual permite seguir la evolución de una clase de homología a lo largo de la filtración.

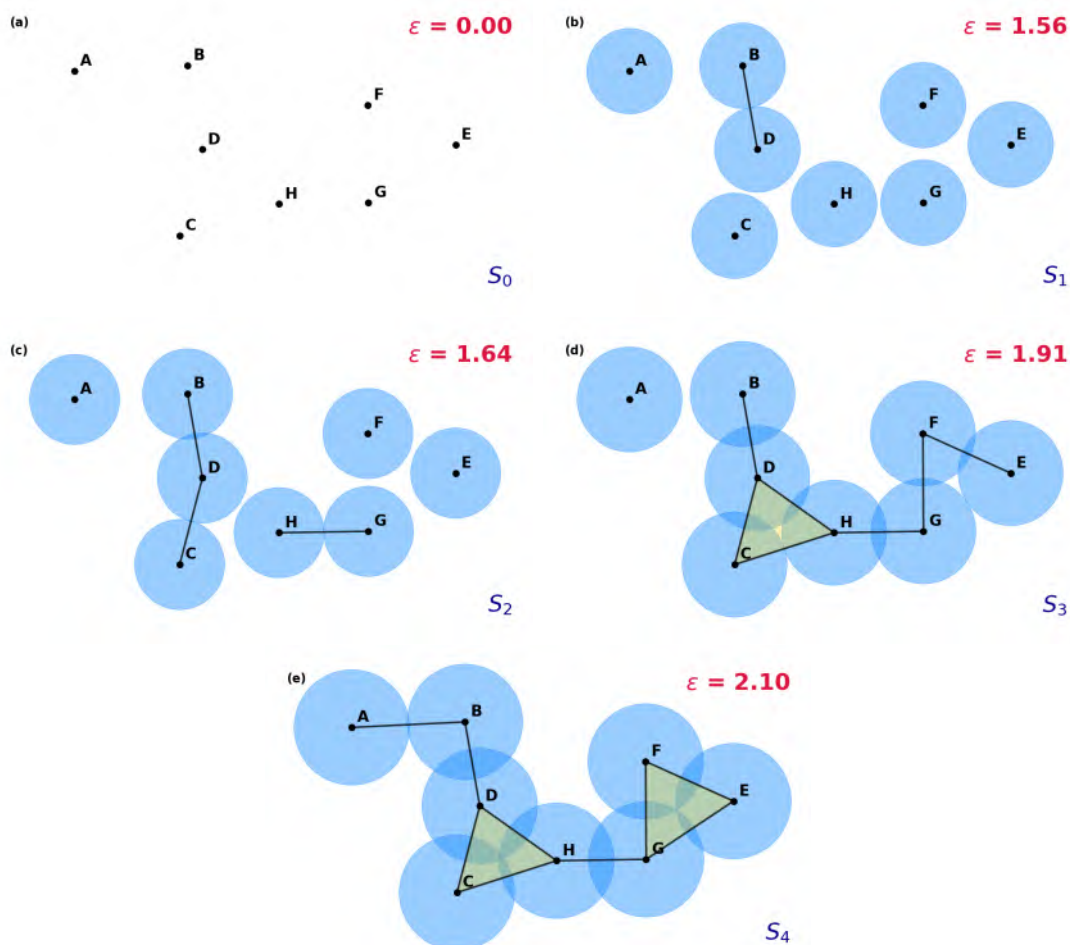


Figura 1.6: Filtración de un complejo de Vietoris–Rips construida a partir del conjunto de ocho puntos mostrado en la Figura 1.5. Las subfiguras (a)–(e) muestran la evolución del complejo al incrementar el parámetro $\varepsilon = 0, 1.56, 1.64, 1.91$, y 2.10 , respectivamente. A medida que aumenta ε , aparecen progresivamente aristas (1-símplices) y triángulos (2-símplices), generando la secuencia creciente de complejos $\emptyset = S_0 \subset S_1 \subset \cdots \subset S_4$.

Definición 12. Sea $\mathcal{F} : S_0 \subset S_1 \subset \cdots \subset S_n$ una filtración de complejos simpliciales y $\alpha \in H_k(S_i)$ una clase de homología k -ésima no trivial. Decimos que α *nace* en S_i si no está en la imagen del morfismo inducido $f_{i-1,i}^k : H_k(S_{i-1}) \rightarrow H_k(S_i)$. La clase α *mueve* en S_j (con $j > i$) si su imagen bajo $f_{i,j}^k : H_k(S_i) \rightarrow H_k(S_j)$ es trivial, es decir, $f_{i,j}^k(\alpha) = 0$, pero $f_{i,j-1}^k(\alpha) \neq 0$.

Intuitivamente, una clase de homología nace cuando aparece una nueva característica topológica independiente, como un componente conexo o un agujero. Dicha clase muere cuando, al aumentar el parámetro de filtración, la estructura que representa se rellena

(por ejemplo, un ciclo que pasa a ser la frontera de un s mplice de dimensi n superior) o colapsa (como una componente que se fusiona con otra), volvi ndose homol gicamente trivial.

Geom tricamente, una clase puede representar:

- Una **componente conexa** (dimensi n 0): nace cuando aparece una nueva regi n desconectada del resto del complejo y muere cuando esta regi n se conecta con otra mediante la adici n de s mplices.
- Un **ciclo o agujero** (dimensi n 1): nace cuando se forma un lazo cerrado de aristas y muere cuando dicho lazo queda relleno por s mplices de dimensi n superior.
- Una **cavidad** (dimensi n 2): nace cuando se forma una superficie cerrada delimitada por 2-s mplices y muere cuando esa superficie se rellena con s mplices de dimensi n superior.

Ejemplo 13. Considere un conjunto de puntos distribuidos en forma circular. Al construir el complejo de Vietoris-Rips, inicialmente s lo se observan componentes aisladas (caracter sticas de homolog a de dimensi n 0). Conforme se aumenta el par metro de filtraci n ε , se crean conexiones (aristas) y eventualmente se forma un ciclo 1-dimensional que representa el agujero en el centro del c rculo. Este ciclo nace en un valor $\varepsilon = \varepsilon_{\text{nacimiento}}$ y muere en $\varepsilon = \varepsilon_{\text{muerte}}$, cuando suficientes conexiones hacen que el ciclo quede relleno. El intervalo $[\varepsilon_{\text{nacimiento}}, \varepsilon_{\text{muerte}})$ mide la *persistencia* del agujero.

El conjunto de clases de homolog a y sus intervalos de persistencia puede representarse de manera visual mediante dos construcciones equivalentes:

- **C digo de barras:** cada clase de homolog a se representa mediante un segmento horizontal que inicia en su punto de nacimiento y termina en su punto de muerte, reflejando as  su intervalo de persistencia.
- **Diagrama de persistencia:** cada clase se representa como un punto en el plano con coordenadas (nacimiento, muerte), donde los puntos m s alejados de la diagonal indican caracter sticas topol gicas m s persistentes.

Ambas representaciones est n fundamentadas en una construcci n algebraica llamada *m dulo de persistencia*, que formaliza c mo evolucionan las clases de homolog a a lo largo de una filtraci n [14, 32]. A continuaci n se presentan las definiciones formales del *c digo de barras* y del *diagrama de persistencia*, que permiten no solo visualizar las clases persistentes, sino tambi n analizarlas de manera cuantitativa.

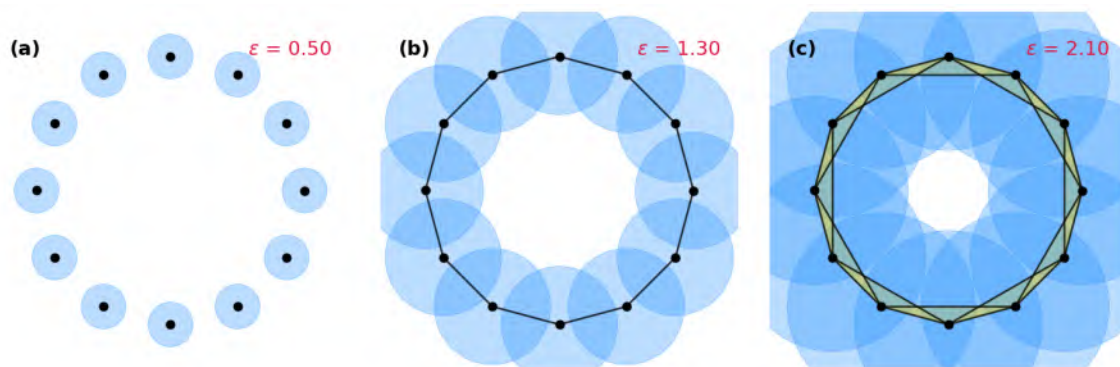


Figura 1.7: Filtración del complejo de Vietoris–Rips construida sobre un conjunto de puntos distribuidos en forma circular. (a) Para valores pequeños de ϵ , sólo existen componentes conexas aisladas (0-símplices). (b) Al incrementar ϵ , aparecen aristas (1-símplices) que forman un ciclo 1-dimensional, el cual representa el agujero central. (c) Para valores aún mayores de ϵ , se forman algunos triángulos (2-símplices), rellenando parcialmente el ciclo; sin embargo, la clase de homología 1 asociada al agujero central todavía persiste. El intervalo de persistencia $[\epsilon_{\text{nacimiento}}, \epsilon_{\text{muerte}})$ cuantifica la duración de este agujero.

Definición 14 (Código de barras). Sea $\mathcal{F} = \{S_i\}_{i \in \mathbb{N}}$ una filtración de complejos simpliciales y $H_k(S_i)$ el grupo de homología en dimensión k . El *código de barras* es una colección de intervalos $[b_i, d_i)$ que describen el tiempo de vida de las clases de homología: b_i corresponde al nivel (o valor del parámetro de la filtración) en que una clase nace, y d_i al nivel en que muere dentro de la filtración.

Formalmente, el código de barras resume la evolución de las clases de homologías a lo largo de la filtración, indicando en qué valores del parámetro de filtración cada clase aparece y desaparece. Para una exposición más detallada sobre su construcción, véase [47, 14]. Un ejemplo de esta representación puede observarse en la Figura 1.8.

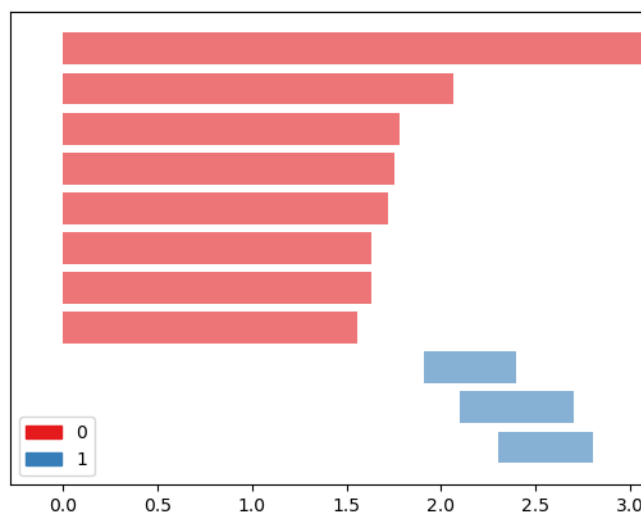


Figura 1.8: **Código de barras.** Cada barra horizontal representa una clase de homología: las barras rojas corresponden a clases en H_0 (componentes conexas), mientras que las barras azules representan clases en H_1 (agujeros o ciclos). El eje horizontal indica el parámetro de filtración, donde el inicio de cada barra representa el *nacimiento* de una clase y el final su *muerte*. La longitud de la barra refleja la persistencia de la característica topológica.

Definición 15 (Diagrama de persistencia). Dado un código de barras $\{[b_i, d_i]\}_{i=1}^n$, el *diagrama de persistencia* es un multiconjunto de puntos en el plano:

$$\text{Dgm} = \{(b_i, d_i) \in \mathbb{R}^2 \mid [b_i, d_i] \text{ es un intervalo en el código de barras}\}.$$

Cada punto (b_i, d_i) indica que una clase de homología aparece en el *nivel* (o *valor del parámetro de la filtración*) b_i y desaparece en d_i . Si una clase persiste indefinidamente, se le asigna $d_i = \infty$. Por convención, se incluye la diagonal $b = d$ con multiplicidad infinita, lo que significa que se considera disponible un número ilimitado de puntos sobre la diagonal para emparejar clases que aparecen y desaparecen inmediatamente al calcular distancias entre diagramas, permitiendo así definir métricas estables entre diagramas de persistencia [12, 14, 32]. Un ejemplo de diagrama de persistencia correspondiente al código de barras de la Figura 1.8 se muestra en la Figura 1.9.

Tanto el código de barras como el diagrama de persistencia representan la misma información topológica: cada intervalo $[b_i, d_i]$ del código de barras se corresponde con un punto (b_i, d_i) en el diagrama de persistencia. La diferencia principal está en la visualización y el análisis: el código de barras destaca la duración o persistencia de cada clase de homología a lo largo de la filtración, mientras que el diagrama de persistencia facilita la comparación entre conjuntos de datos y el análisis de estabilidad mediante el uso de distancias métricas entre diagramas [8, 31].

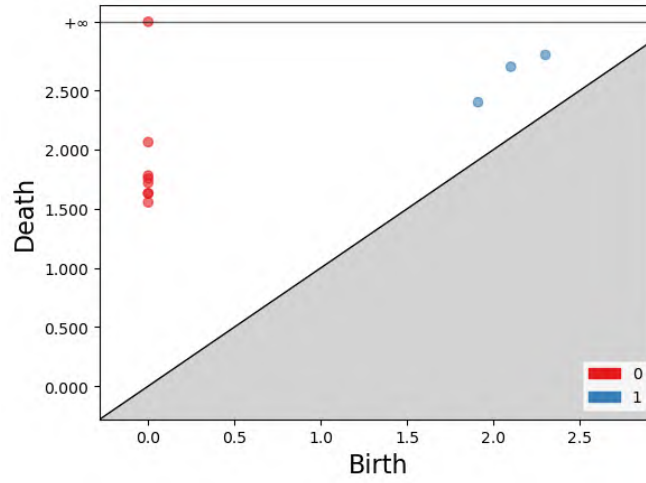


Figura 1.9: **Diagrama de persistencia.** Cada punto (b_i, d_i) representa una clase de homología: los puntos rojos corresponden a clases en H_0 y los azules a clases en H_1 . El eje x indica el *nacimiento* de la clase y el eje y su *muerte*. La distancia a la diagonal refleja la persistencia: los puntos más alejados de la diagonal representan características topológicas más estables.

Una vez obtenidos los diagramas de persistencia, surge la necesidad de compararlos entre sí, especialmente al analizar múltiples conjuntos de datos, como distintas imágenes celulares o tejidos. Estas comparaciones permiten identificar similitudes o diferencias en las estructuras topológicas subyacentes. Para ello, se han desarrollado métricas que cuantifican la distancia entre diagramas de persistencia, como la *distancia de cuello de botella* y la *distancia de Wasserstein*. Estas métricas proporcionan una base matemática sólida para evaluar la estabilidad de los invariantes topológicos y para construir modelos de clasificación, agrupamiento o regresión sobre datos biológicos complejos [12, 32].

Definición 16 (Distancia cuello de botella). Sean D_1 y D_2 dos diagramas de persistencia. La *distancia cuello de botella* entre ellos se define como

$$d_B(D_1, D_2) = \inf_{\gamma} \sup_{x \in D_1} \|x - \gamma(x)\|_{\infty},$$

donde γ recorre todos los emparejamientos (biyecciones parciales) entre los puntos de D_1 y D_2 . En estos emparejamientos se permite asociar puntos de un diagrama con la diagonal $b = d$, lo que corresponde a considerar clases de homología que aparecen y desaparecen inmediatamente. La norma $\|\cdot\|_{\infty}$ denota la *norma del supremo* en $\mathbb{R}^2$, definida para un punto $(x_1, x_2) \in \mathbb{R}^2$ como

$$\|(x_1, x_2)\|_{\infty} = \max\{|x_1|, |x_2|\}.$$

Esta distancia mide el mayor desplazamiento necesario para transformar un diagrama de persistencia en otro, capturando la diferencia máxima entre sus características topológicas más persistentes.

En la definición general, los puntos con coordenada de muerte infinita se consideran emparejados entre sí o con la diagonal. En nuestro caso particular, como se explica en la Sección 2.3, los diagramas de persistencia se generan a partir del complejo de Vietoris–Rips usando un radio lo suficientemente grande, por lo que en cada diagrama aparece exactamente un punto que persiste hasta el infinito. Al comparar diagramas, estos puntos se ignoran de manera explícita, ya que, al haber solo uno por diagrama, no aportan información discriminante.

Ejemplo 17 (Cálculo paso a paso de la distancia de cuello de botella). Para ilustrar el cálculo de la distancia de cuello de botella entre dos diagramas de persistencia, consideremos los diagramas D_1 y D_2 mostrados en la Figura 1.10. Cada diagrama representa características topológicas obtenidas a partir del complejo de Vietoris–Rips, donde los círculos corresponden a componentes conexas (clases en H_0) y los triángulos a ciclos (clases en H_1).

La distancia de cuello de botella $d_B(D_1, D_2)$ mide la máxima discrepancia entre las clases de homología de ambos diagramas, de acuerdo con la correspondencia óptima entre sus puntos. El cálculo se realiza de la siguiente manera:

1. **Visualización de los diagramas:** se consideran los dos diagramas de persistencia D_1 y D_2 , donde cada punto (b, d) representa el nacimiento y muerte de una clase de homología.
2. **Superposición de diagramas:** los diagramas D_1 y D_2 se superponen en un mismo plano para identificar visualmente las posibles correspondencias entre puntos, con esto podemos apreciar las diferencias globales entre las estructuras topológicas representadas en ambos diagramas.
3. **Emparejamiento óptimo de puntos:** se busca una correspondencia $\gamma : D_1 \rightarrow D_2$ que minimice la distancia máxima entre los puntos emparejados. Para un punto $x = (b_1, d_1) \in D_1$ y su imagen $\gamma(x) = (b_2, d_2) \in D_2$, la distancia entre ambos se mide utilizando la norma infinito

$$\|x - \gamma(x)\|_\infty = \max\{|b_1 - b_2|, |d_1 - d_2|\}.$$

El valor óptimo de esta distancia máxima se denomina *distancia de cuello de botella*.

Este procedimiento permite cuantificar la similitud entre diagramas de persistencia de forma robusta, capturando la mayor discrepancia entre sus características topológicas dominantes.

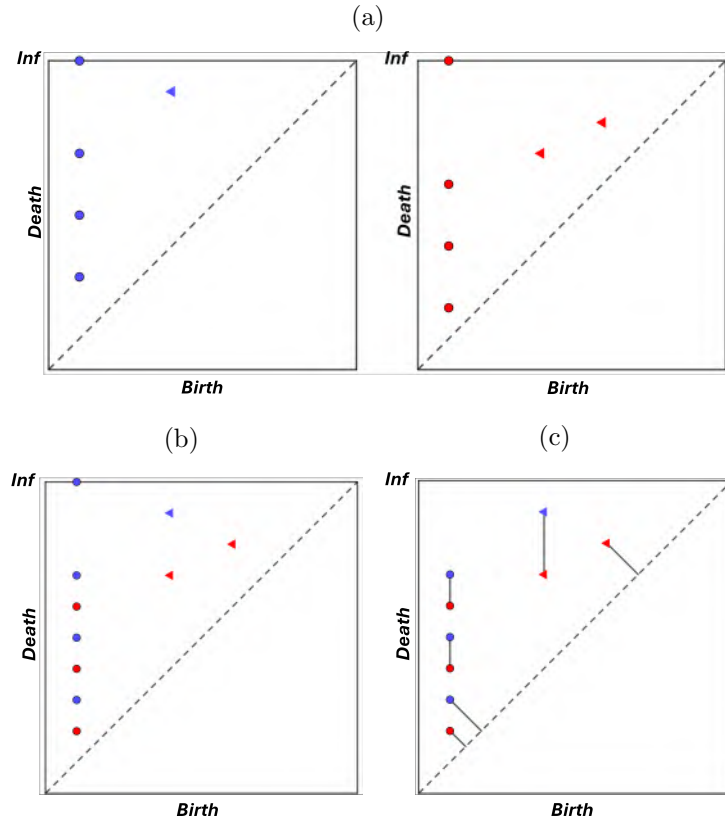


Figura 1.10: **Distancia de cuello de botella.** (a) Diagramas de persistencia D_1 y D_2 . (b) Superposición de ambos diagramas. (c) Emparejamiento óptimo; la línea más larga define $d_B(D_1, D_2)$. Esta métrica se define por la mayor distancia entre los puntos emparejados, enfatizando las diferencias más significativas entre diagramas.

Definición 18 (Distancia de Wasserstein). Dado $p \geq 1$, la distancia de Wasserstein de orden p entre dos diagramas de persistencia D_1 y D_2 se define como

$$d_{W_p}(D_1, D_2) = \left(\inf_{\vartheta} \sum_{x \in D_1} \|x - \vartheta(x)\|_{\infty}^p \right)^{1/p},$$

donde ϑ recorre todos los emparejamientos admisibles entre D_1 y D_2 , y $\|\cdot\|_{\infty}$ denota la norma infinito.

A diferencia de la distancia de cuello de botella, esta métrica considera el desplazamiento total acumulado (elevado a la potencia p y promediado), por lo que resulta más sensible a diferencias pequeñas distribuidas a lo largo del diagrama.

De manera análoga al caso del cuello de botella, los puntos con muerte infinita suelen emparejarse entre sí o con la diagonal. Sin embargo, en este análisis se ignoran dichos puntos, ya que la construcción mediante complejos de Vietoris-Rips garantiza que únicamente aparece un punto infinito por diagrama, el cual es común en todos los casos y no afecta la comparación entre muestras.

Ejemplo 19 (Cálculo paso a paso de la distancia de Wasserstein). Para ilustrar el cálculo de la distancia de Wasserstein entre dos diagramas de persistencia, consideremos los diagramas D_1 y D_2 mostrados en la Figura 1.11. Cada diagrama representa características topológicas obtenidas a partir del complejo de Vietoris-Rips, donde los círculos corresponden a componentes conexas (clases en H_0) y los triángulos a ciclos (clases en H_1).

El cálculo se realiza siguiendo los pasos que se describen a continuación, haciendo referencia a la definición de la distancia de Wasserstein $d_{W_p}(D_1, D_2)$ ya presentada:

1. **Proyección de puntos a la diagonal:** se proyectan los puntos fuera de la diagonal sobre la diagonal. Los puntos cercanos a la diagonal representan características topológicas de corta duración y se consideran poco significativos, pudiendo ignorarse temporalmente. Los puntos con *muerte infinita* también se ignoran en este paso, ya que aparecen de manera consistente en todos los diagramas y no afectan la comparación.
2. **Intercambio de proyecciones entre diagramas:** si los diagramas contienen diferente número de puntos fuera de la diagonal, las proyecciones se intercambian para equilibrar la cantidad de puntos y permitir un emparejamiento completo.
3. **Determinación de la biyección óptima:** se define la correspondencia $s : D_1 \rightarrow D_2$ que minimiza la suma total de las distancias entre los puntos emparejados. En la práctica, cada elemento de una clase de homología en dimensión 0 (círculo) se empareja únicamente con otro elemento de dimensión 0, y cada clase en dimensión 1 (triángulo) con otra de dimensión 1, de modo que la biyección preserve la dimensión homológica.
4. **Cálculo de la distancia:** una vez obtenida la biyección óptima s , se calcula la distancia de Wasserstein $W_p(D_1, D_2)$ como la suma acumulada de las distancias de los puntos emparejados, elevada a la potencia p y posteriormente se extrae la raíz p -ésima.

Este procedimiento permite comparar diagramas de persistencia de manera rigurosa, incluso cuando los diagramas contienen diferente número de características, algunas desaparecen rápidamente o existen puntos con muerte infinita. En estudios de simulación, se utiliza para cuantificar la similitud global entre configuraciones de partículas y, en nuestro caso, entre patrones de organización celular.

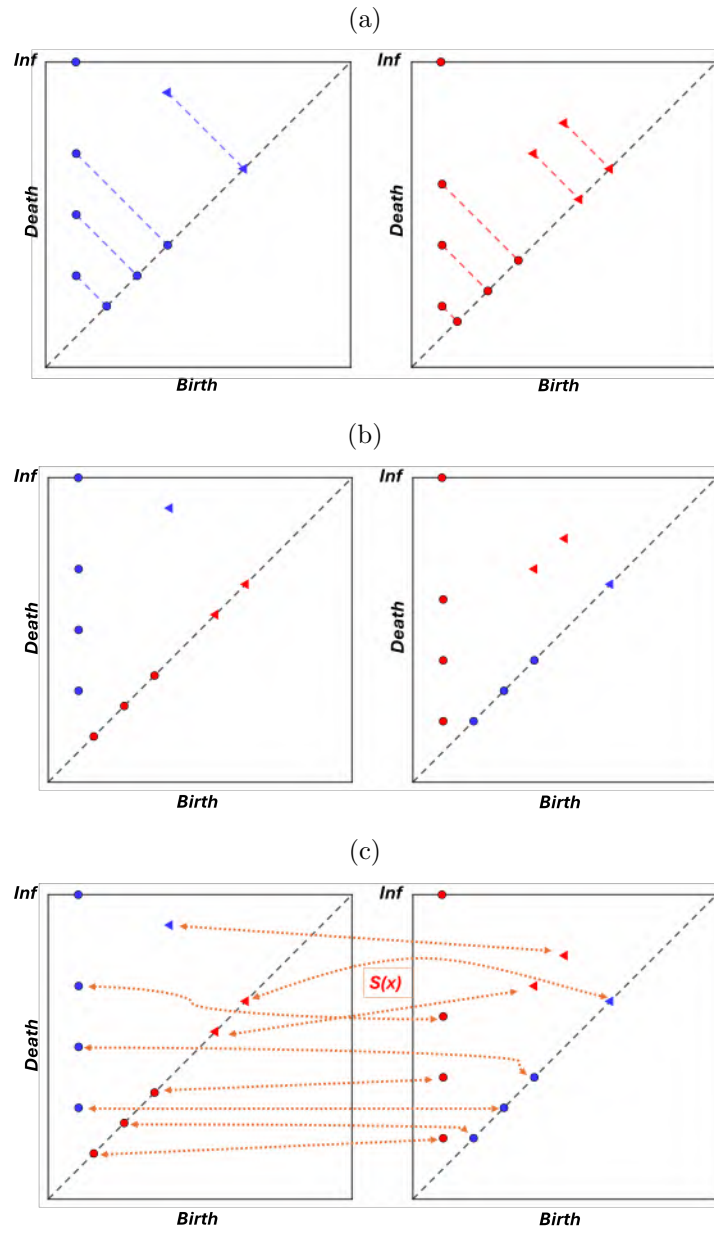


Figura 1.11: **Distancia de Wasserstein.** (a) Proyección ortogonal sobre la diagonal. (b) Intercambio de proyecciones entre diagramas. (c) Emparejamiento óptimo mediante la biyección S . La distancia de Wasserstein evalúa la similitud entre diagramas considerando la suma total de las distancias entre los puntos emparejados.

1.2.3. Aplicaciones del TDA en imágenes biológicas

La homología persistente, como herramienta computacional, ha sido implementada con éxito en estudios de imágenes biomédicas, revelando patrones que pasan desapercibidos para los métodos tradicionales basados en segmentación o geometría local [44]. Gracias a su naturaleza multiescala y a su invariancia frente a transformaciones geométricas, el TDA permite capturar la estructura global de los datos, incluyendo relaciones espaciales y organizaciones celulares no triviales.

La aplicación del TDA para la clasificación de patrones celulares ha sido previamente demostrada. Por ejemplo, en el trabajo de Bhaskar et al. (2023) [5], el TDA se empleó para analizar la distribución espacial de células con diferentes niveles de adhesión, utilizando la homología persistente para agrupar patrones celulares basándose en su conectividad topológica, sin necesidad de segmentación explícita. Este enfoque permitió distinguir subpoblaciones celulares en ambientes heterogéneos u homogéneos, revelando estructuras organizacionales relacionadas con estados funcionales. El trabajo previo se enfocaba en la clasificación de patrones celulares simples con morfologías celulares sencillas. Sin embargo, una de las aplicaciones más destacadas ha sido en el análisis de imágenes celulares obtenidas mediante tecnologías de multiplexado o microscopía de célula única, donde la alta densidad y variabilidad morfológica dificultan el uso de herramientas convencionales. Por ejemplo, el estudio de Aukerman et al. (2020) [3], utilizó homología persistente para caracterizar la organización espacial de células inmunes y tumorales en el microambiente del cáncer de mama, a partir de imágenes obtenidas mediante inmunofluorescencia multiplexada. Los diagramas de persistencia extraídos de las coordenadas celulares mostraron capacidad para diferenciar subtipos de cáncer y predecir la supervivencia general de los pacientes, incluso superando a métodos tradicionales de análisis espacial. Estos resultados sugieren que los descriptores topológicos derivados del TDA pueden actuar como biomarcadores robustos y clínicamente relevantes.

Por su parte, Samorodnitsky et al. (2023) [41] introdujeron TopKAT (Topological Kernel Association Test), un método estadístico que combina la información de homología persistente con pruebas de asociación basadas en kernels. Un kernel es una función matemática que mide el grado de semejanza entre dos objetos (en este caso, entre dos diagramas de persistencia), permitiendo representar relaciones complejas en un espacio de mayor dimensión sin necesidad de calcular explícitamente esa transformación. La idea central es transformar los diagramas de persistencia en funciones de kernel que capturan la similitud topológica entre muestras, y luego evaluar si dichas estructuras topológicas se asocian con variables de interés (clínicas o experimentales), que pueden ser continuas, binarias o de supervivencia. Este enfoque no solo extiende el alcance del TDA hacia hipótesis estadísticas formales, sino que también permite integrar covariables y ajustar por factores

de confusión. En el contexto del cáncer de mama triple negativo (TNBC), un subtipo de cáncer de mama más agresivo que no expresa ciertos receptores celulares, TopKAT identificó patrones celulares con organización tipo “anillo” que se correlacionaron con malos pronósticos clínicos. Estos resultados superaron en potencia a métodos tradicionales basados en procesos puntuales espaciales, demostrando la utilidad de la homología persistente para capturar estructuras relevantes en datos complejos. La herramienta está disponible como paquete de R [40], lo que facilita su aplicación.

La poca cantidad de trabajos empleando TDA para el análisis de tejidos biológicos muestra que es un campo en desarrollo. Sin embargo, a nuestro parecer, en conjunto, estos estudios demuestran el potencial del TDA no solo para representar la geometría de los tejidos, sino también para derivar características que puedan alimentar modelos de clasificación, agrupamiento o predicción clínica. Además, validan su utilidad en entornos reales con alta variabilidad biológica, como el microambiente tumoral o los tejidos inmunológicos. Es por estos motivos que, en nuestro trabajo, aplicaremos este enfoque para analizar la organización celular en imágenes de tejidos, evaluando si las características topológicas se asocian con diferentes propiedades de las muestras, ya sean tumorales o no tumorales, que puedan servir como posibles blancos terapéuticos para el tratamiento de tumores. Para este propósito, emplearemos una patología conocida como anemia de Fanconi (AF) como modelo de estudio. A continuación, se describe la anemia de Fanconi y se explica por qué sirve como un buen modelo para la búsqueda de nuevas terapias tumorales, como se detalla en la Sección 1.3.

1.2.4. Ventajas del enfoque topológico

El TDA y, en particular, la homología persistente, no solo permiten describir estructuras espaciales complejas, sino que también ofrecen una base sólida para la clasificación de patrones multicelulares a partir de características topológicas extraídas directamente de datos geométricos. Estas características, comúnmente representadas mediante diagramas de persistencia o códigos de barras, pueden ser vectorizadas y utilizadas como variables de entrada en modelos de aprendizaje supervisado o no supervisado.

Una de las principales ventajas del enfoque topológico es su capacidad para capturar la conectividad y la organización espacial de las células en diferentes escalas, sin depender de supuestos específicos sobre la forma, el tamaño o la orientación de los objetos en la imagen. Esta invariancia geométrica le confiere robustez frente a ruido, rotaciones, traslaciones y deformaciones locales [10]. De esta manera, el TDA proporciona descriptores más estables y globales que complementan a los enfoques tradicionales basados en textura o morfometría.

Además, el TDA puede integrarse fácilmente con técnicas de análisis de imágenes existentes. Por ejemplo, las coordenadas de los centroides celulares, obtenidas mediante segmentación automática, pueden servir como entrada para la construcción de complejos simpliciales como el de Vietoris-Rips. A partir de estos, se genera una filtración cuya homología persistente describe las características topológicas del patrón multicelular. Posteriormente, los diagramas de persistencia pueden compararse entre muestras utilizando distancias como la de Wasserstein o cuello de botella, o bien convertirse en vectores mediante funciones resumen (persistence landscapes, persistence images, etc.) [37], habilitando así su uso en clasificadores estándar como máquinas de soporte vectorial (SVM), redes neuronales o métodos de árboles.

Por su flexibilidad y capacidad para modelar relaciones espaciales no triviales, el TDA se ha consolidado como una herramienta complementaria en el análisis de imágenes biomédicas, especialmente útil en contextos donde los patrones celulares tienen implicaciones diagnósticas o pronósticas, como ocurre en el estudio de tejidos tumorales, tejidos inmunes o cultivos celulares heterogéneos.

1.3. La anemia de Fanconi como modelo de estudio del desarrollo de carcinomas

La anemia de Fanconi (AF) es una enfermedad congénita causada por mutaciones bialélicas (es decir, mutaciones en ambas copias de un gen, tanto en el cromosoma paterno como en el materno) en cualquiera de los 22 genes que conforman la vía de reparación del ADN conocida como vía FA/BRCA [38]. Esta vía funciona como un sistema molecular especializado en detectar y corregir roturas en el ADN. Como consecuencia, los pacientes con AF presentan *inestabilidad genómica*, es decir, una tendencia a acumular mutaciones y reorganizaciones en su ADN generando inestabilidad estructural del genoma [7, 38]. Por estos motivos, la AF se clasifica dentro de los llamados *síndromes de inestabilidad genómica*, un grupo de enfermedades caracterizadas por fallos en la preservación del ADN que aumentan el riesgo de transformación tumoral.

Al afectar la correcta reparación del ADN, fallas en la vía FA/BRCA termina comprometiendo múltiples procesos biológicos esenciales, generando un efecto sistémico en el organismo. De esta forma, los pacientes con AF suelen presentar malformaciones congénitas (defectos estructurales presentes desde el nacimiento), retraso en el crecimiento, talla baja, insuficiencia medular (fallo en la producción normal de células sanguíneas), predisposición a diabetes, envejecimiento prematuro y, de manera destacada, una elevada susceptibilidad al desarrollo de carcinomas.

Gracias a los avances en los trasplantes de médula ósea, la esperanza de vida de los pacientes con AF ha aumentado significativamente. Sin embargo, este incremento conlleva un mayor riesgo de desarrollar carcinomas agresivos, asociados a la acumulación progresiva de mutaciones derivadas del daño genómico no reparado [1]. El tratamiento convencional de tumores se realiza por medio de quimioterapias y radioterapias, las cuales buscan detener el crecimiento de las células tumorales dañando su ADN. Sin embargo, debido a que los pacientes con AF presentan inestabilidad genómica, las terapias convencionales resultan extremadamente tóxicas, por lo que el tratamiento de tumores representa un desafío clínico en estos pacientes [25].

Ante este panorama, se ha propuesto enfocar la atención en el **microambiente tumoral** (MAT) como posible fuente de blancos terapéuticos alternativos para pacientes con AF. El MAT puede entenderse como un *ecosistema celular* donde coexisten y se comunican distintos tipos de células: malignas (tumorales), inmunes, endoteliales (que forman vasos sanguíneos) y estromales (de soporte estructural). Estas interacciones conforman una red dinámica que influye de manera decisiva en la progresión tumoral y en la respuesta a los tratamientos [2]. Por ejemplo, estudios del MAT han mostrado que las células inmunes infiltrantes —aquellas que penetran el tejido tumoral— desempeñan funciones duales, tanto proinflamatorias como antiinflamatorias, que son esenciales para mantener el equilibrio tisular. Sin embargo, dentro del entorno tumoral, estos mecanismos pueden invertirse y favorecer la progresión del cáncer. Por ello, dichas células representan blancos prometedores para el desarrollo de nuevas estrategias terapéuticas [4]. No obstante, el MAT en los carcinomas asociados a AF sigue siendo escasamente caracterizado, a pesar de su relevancia para comprender los mecanismos que subyacen a la alta agresividad tumoral observada en estos pacientes.

En este contexto, el grupo de investigación del Dr. Alfredo de Jesús Rodríguez Gómez del Instituto de Investigaciones Biomédicas ha comenzado a generar imágenes del MAT de carcinomas provenientes de pacientes con y sin AF mediante la técnica de *inmunofluorescencia cíclica en tejido* (*t-CycIF*). Esta tecnología, perteneciente al campo de la *ómica espacial*, permite cuantificar la expresión de múltiples proteínas a nivel de célula única sin perder la organización espacial del tejido, ofreciendo así una representación integral del microambiente tumoral [33]. Es por estos motivos, que estos datos brindan una oportunidad única para la caracterización y estudio de carcinomas agresivos. El hecho de que los pacientes AF sean sensibles a las terapias tradicionales como la quimioterapia, permite buscar en ellos blancos potenciales para nuevos tratamientos. Sin embargo, tal como se describió previamente, los métodos tradicionales del análisis de imágenes de tejidos presentan varias limitaciones para el estudio de este tipo de imágenes, por lo que es necesario buscar nuevos métodos para su análisis, como podría ser el TDA.

1.4. Resumen y planteamiento de la pregunta de investigación

En resumen, la literatura revisada respalda la utilidad del TDA para caracterizar patrones multicelulares a partir de la estructura espacial de las células. Se ha demostrado que el TDA permite identificar relaciones topológicas no evidentes mediante herramientas como la homología persistente, lo que le otorga ventajas frente a métodos tradicionales de análisis de imágenes, que suelen centrarse únicamente en aspectos morfológicos o métricos locales.

Además, el TDA ha mostrado ser una herramienta complementaria en entornos biomédicos complejos, con aplicaciones que van desde la diferenciación de subtipos tumorales hasta la predicción de desenlaces clínicos. Estos hallazgos abren nuevas perspectivas para su uso en la clasificación automatizada de tejidos, especialmente cuando se combina con modelos de aprendizaje automático o se emplea como herramienta de extracción de características.

En el presente trabajo, se analiza un conjunto de datos reales obtenidos gracias a una colaboración con el Dr. Alfredo de Jesús Rodríguez Gómez, quien junto con sus estudiantes, en especial el candidato a doctor en Ciencias Biomédicas Pablo Siliceo Portugal, obtuvieron muestras de tejidos tumorales mediante *t*-CycIF (inmunofluorescencia cíclica en tejido), técnica que posibilita la caracterización proteica con resolución unicelular. Las muestras corresponden a tejidos tumorales de pacientes con y sin anemia de Fanconi (AF), un síndrome de falla medular congénito que predispone a los pacientes a desarrollar tumores [39].

En este contexto, utilizamos dos niveles de comparación:

- La **condición histológica**, que se refiere al estado del tejido según su arquitectura y composición celular, por ejemplo, carcinoma, displasia o estroma.
- La **condición clínica**, que alude a las características del paciente o su contexto biológico, como la presencia o ausencia de anemia de Fanconi.

Este conjunto de datos permite explorar si las diferencias topológicas capturadas por el TDA pueden utilizarse para clasificar tejidos según su condición histológica (por ejemplo, carcinoma, displasia o estroma) y su condición clínica (pacientes con o sin anemia de Fanconi). De esta forma, se evalúa sistemáticamente la utilidad del TDA como herramienta de extracción de características espaciales para la clasificación de patrones multicelulares en aplicaciones biomédicas reales.

A partir de esta revisión, surgen las siguientes preguntas que orientan el presente trabajo:

- ¿Puede el Análisis Topológico de Datos capturar diferencias topológicas relevantes en la organización espacial celular que permitan clasificar tejidos según su condición histológica (carcinoma, displasia y estroma)?
- ¿Permite el Análisis Topológico de Datos identificar diferencias topológicas asociadas a la condición clínica de los tejidos, en particular entre pacientes con y sin anemia de Fanconi?

Estas preguntas articulan el desarrollo del trabajo: la primera evalúa la capacidad del Análisis Topológico de Datos para distinguir patrones espaciales celulares, mientras que la segunda explora si las diferencias topológicas identificadas se asocian con variables biológicas y clínicas relevantes.

Capítulo 2

Métodos

2.1. Descripción general del enfoque metodológico

Mediante técnicas de TDA, este estudio analiza la organización espacial celular de imágenes de tejido tumoral obtenido por medio de biopsias. A continuación, se presenta un resumen del protocolo para la obtención de imágenes:

- **Fuentes de datos:** Se utilizaron imágenes de inmunofluorescencia cíclica en tejido (*t*-CycIF) provenientes de muestras tumorales de pacientes con y sin Anemia de Fanconi.
- **Extracción de tipos celulares:** La inmunofluorescencia fue diseñada para detectar a nivel de proteína marcadores celulares. De esta forma, esta información permitió definir el tipo celular de cada una de las células que conformaban cada tejido.
- **Clasificación de las células por funcionalidad:** Con base en el tipo celular, las células fueron agrupadas en cuatro categorías principales las cuales representan diferentes funcionalidades biológicas: células tumorales, células linfoides, células mieloides y **otro tipo de células** (ver Tabla 2.1).
- **Clasificación de las zonas del tejido:** El tejido podía contener diferentes zonas, las cuales fueron clasificadas en dos categorías distintas que aquí llamaremos zonas histológicas: (1) estroma, las cuales son zonas que no tienen o tienen una muy baja presencia de células tumorales, y (2) displasia y carcinoma, las cuales son zonas con una alta presencia de células tumorales.
- **Extracción de coordenadas celulares:** A partir de las imágenes segmentadas, se obtuvieron las coordenadas (x, y) de los centroides de cada una de las células que conformaban cada tejido.

Tabla 2.1: Clasificación de los tipos celulares en cada una de las cuatro categorías funcionales.

Categoría	Tipos de células
Células tumorales	<i>tumor cells, Ki67+ tumor cells, IBA1+ tumor cells</i>
Células linfoides	<i>NK, B cells, effector CD8+ T cells, memory CD8+ T cells, CD4+ T cells, regulatory T cells, memory CD4+ T cells</i>
Células mieloides	<i>neutrophils, other APCs, dendritic cells, M1/M0 macrophages, M2 macrophages</i>
No tumorales	<i>endothelial cells, stromal cells</i>

Toda esta información fue proporcionada por el estudiante de doctorado Pablo Siliceo Portugal y el estudiante de biología Cruz Francisco Osuna Aguirre, quienes realizaron su tesis bajo la supervisión del Dr. Alfredo de Jesús Rodríguez Gómez del Instituto de Investigaciones Biomédicas de la UNAM. En total, se proporcionaron 288 muestras en archivos `csv`, correspondientes a muestras individuales. La Tabla 2.2 resume la distribución de estas muestras de acuerdo con la condición clínica de los pacientes (con y sin anemia de Fanconi) y su clasificación histológica, incluyendo carcinoma, displasia y sus respectivos estromas. Esta tabla permite visualizar el balance del conjunto de datos y proporciona un contexto cuantitativo para los análisis posteriores de clasificación basados en TDA. Cada muestra estaba compuesta por aproximadamente 2,000 células. El análisis de las muestras mediante TDA se realizó en tres etapas principales:

Tabla 2.2: Número de muestras por presencia/ausencia de FA y condición histológica

Tipo histológico	Paciente FA	Paciente no-FA	Total
Carcinoma	40	40	80
Displasia	32	8	40
Estroma adyacente a displasia	74	5	79
Estroma adyacente a carcinoma	39	50	89
Total general	185	103	288

- Construcción de complejos simpliciales y obtención de diagramas de per-

sistencia: A partir de las coordenadas de los centroides celulares, se construyeron complejos de Vietoris–Rips para modelar la conectividad espacial entre las células en cada muestra. Posteriormente, a lo largo de un rango de filtraciones ε se calcularon las dimensiones de los grupos de homología, con la finalidad de capturar propiedades topológicas asociadas tanto a la conectividad (componentes conexas) como a la aparición de cavidades o patrones espaciales (ciclos), respectivamente. Este análisis se efectuó de dos maneras:

1. Considerando todas las células de cada muestra (análisis global), sin distinguir por tipo celular;
 2. Separando las células por grupos (análisis por grupos) según su categoría (tumores, inmunes, mieloides u otro tipo de células) (Tabla [2.1](#)).
- **Generación de diagramas de persistencia:** A partir de los complejos y de las dimensiones de los grupos de homología calculadas en el paso anterior, se obtuvieron los diagramas de persistencia correspondientes a cada muestra. Estos diagramas resumen, de manera compacta, la aparición y desaparición de las estructuras topológicas a lo largo del proceso de filtración. Se generó un diagrama independiente para cada grupo de homología considerado: uno para H_0 y otro para H_1 .
 - **Extracción de características topológicas y análisis comparativo:** Los diagramas de persistencia fueron comparados entre sí mediante el cálculo de distancias topológicas (Wasserstein y Cuello de botella) entre pares de muestras.
 - **Análisis de asociación clínica:** Para evaluar si las diferencias topológicas entre las muestras se relacionan con sus condiciones clínicas y características histológicas, se aplicó TopKAT. Este método combina la información de los diagramas de persistencia con pruebas de asociación basadas en kernels, permitiendo determinar si las muestras con estructuras topológicas similares también comparten atributos clínicos o biológicos semejantes (por ejemplo, presencia o ausencia de anemia de Fanconi, o tipo histológico del tejido).

En la Figura [2.1](#) se muestra un diagrama de flujo general del pipeline metodológico seguido. A continuación, se describe con más detalle la metodología empleada. La Sección 2.2 explica el trabajo realizado por el grupo del laboratorio del Dr. Rodríguez; una descripción más completa del procesamiento de estas imágenes se encuentra en la tesis de licenciatura del estudiante Pablo Siliceo [\[43\]](#). A partir de la Sección 2.3 se detalla el trabajo realizado por mí para el desarrollo de esta tesis.

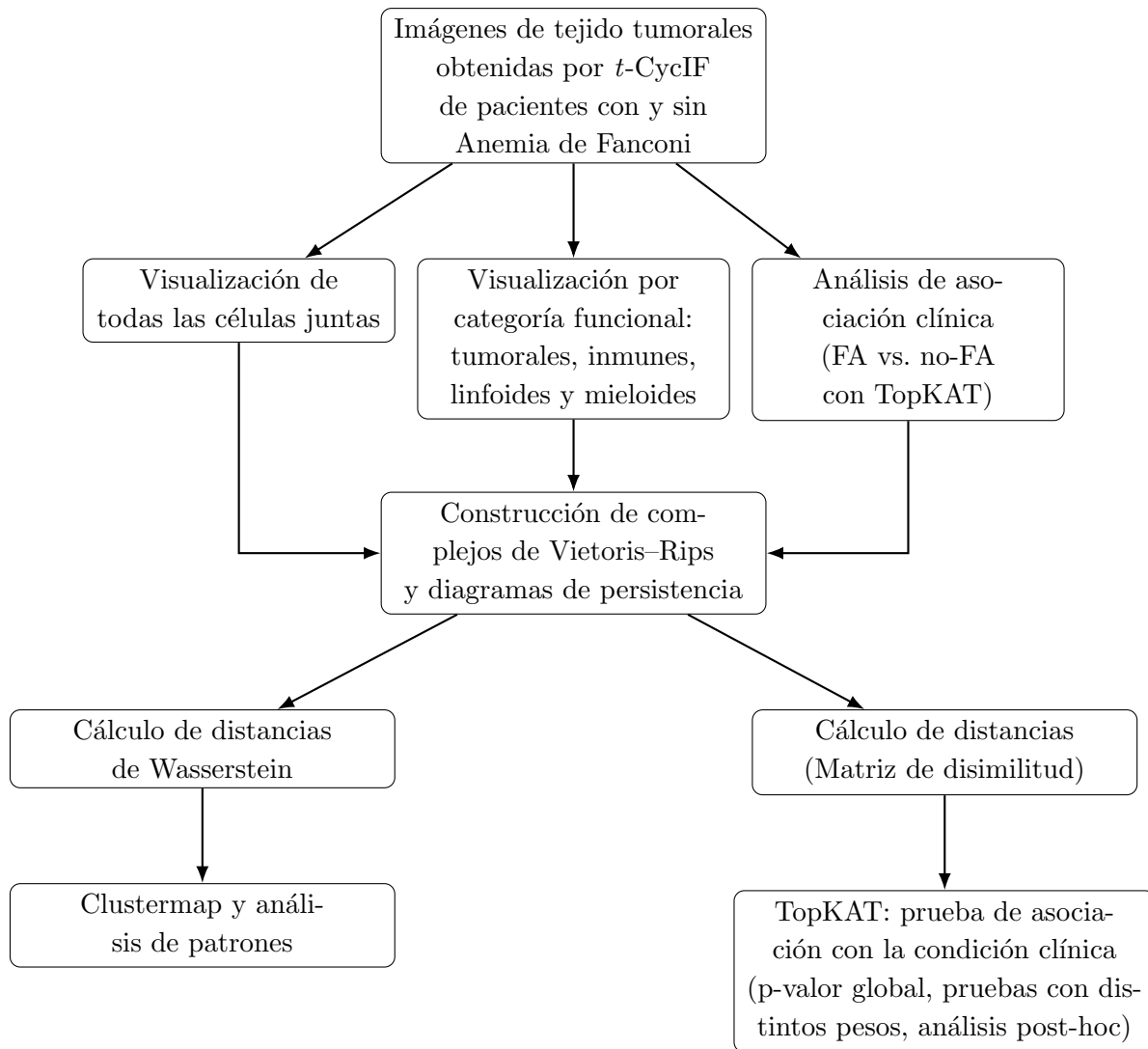


Figura 2.1: Diagrama de flujo del pipeline metodológico seguido para el análisis topológico de imágenes celulares. El análisis parte de imágenes obtenidas mediante *t*-CycIF y se realiza en tres niveles: (1) considerando todas las células juntas, (2) separadas por tipo celular, y (3) mediante evaluación formal de asociación con la condición clínica (FA/no-FA). Todos los análisis convergen en la construcción de complejos de Vietoris–Rips y diagramas de persistencia, a partir de los cuales se calculan distancias topológicas o matrices de disimilitud, seguidas de análisis de agrupamiento o pruebas estadísticas mediante TopKAT.

2.2. Datos: imágenes *t*-CyCIF

Para el análisis del microambiente tumoral se empleó la técnica de inmunofluorescencia cíclica en tejido (*t*-CyCIF), la cual permite obtener imágenes multiparamétricas con reso-

lución unicelular, preservando la localización espacial de cada célula y su perfil proteico. Esta metodología posibilita la observación simultánea de distintas proteínas (comúnmente denominadas marcadores) mediante ciclos secuenciales de tinción, adquisición de imagen y eliminación de fluoróforos, lo que resulta en una reconstrucción espacial precisa del tejido.

Las muestras fueron procesadas mediante un protocolo de siete ciclos de tinción que incluyó marcadores de distintos tipos celulares y funciones: inmunes (CD4, CD8a, FOXP3, CD20), tumorales (SOX2, PAN-CK, E-cadherin) y funcionales (p-S6, gH2AX, PD-1, PD-L1, TIM3). El escaneo se realizó con un microscopio *RareCyte Cytefinder II*, generando imágenes por segmentos (*tiles*) que posteriormente fueron ensambladas y registradas en una imagen multidimensional por muestra, donde cada dimensión corresponde a un canal de fluorescencia asociado a un marcador específico.

Los principales pasos de preprocesamiento e integración fueron:

- **Corrección de iluminación:** realizada con el software *BaSiC* [34] para eliminar artefactos ópticos y homogeneizar el fondo de las imágenes.
- **Reconstrucción de imágenes:** mediante *ASHLAR* [18], se ensamblaron los *tiles* y se registraron los ciclos de tinción para generar una imagen multicanal única por muestra.

En particular, el canal de *Hoechst* marca el ADN nuclear y permite identificar la posición de cada célula, mientras que el marcador *PAN-CK* resalta las células epiteliales tumorales, facilitando la distinción entre regiones con diferente densidad tumoral. Un ejemplo representativo de un corte procesado mediante *t-CyCIF* se muestra en la Figura 2.2.

Procedencia de los datos

Los cortes de tejido analizados corresponden a carcinomas de células escamosas obtenidos de dos grupos de pacientes provenientes de instituciones colaboradoras. Los tejidos de pacientes con anemia de Fanconi (AF) —seis de cabeza y cuello y cuatro anogenitales— fueron proporcionados por la Dra. Eunike Velleuer (Universidad de Düsseldorf, Alemania), mientras que los tejidos de pacientes sin AF fueron proporcionados por la Dra. Azucena Ocampo Bárcenas (Instituto Estatal de Cancerología, Guerrero, México).

Posterior a su procesamiento mediante la técnica *t-CyCIF*, patólogos especialistas en imágenes tumorales identificaron y clasificaron las zonas histológicas de interés: *dysplasia*, *carcinoma* y *stroma adyacente a carcinoma* o *dysplasia (stroma_ad_carcinoma* o

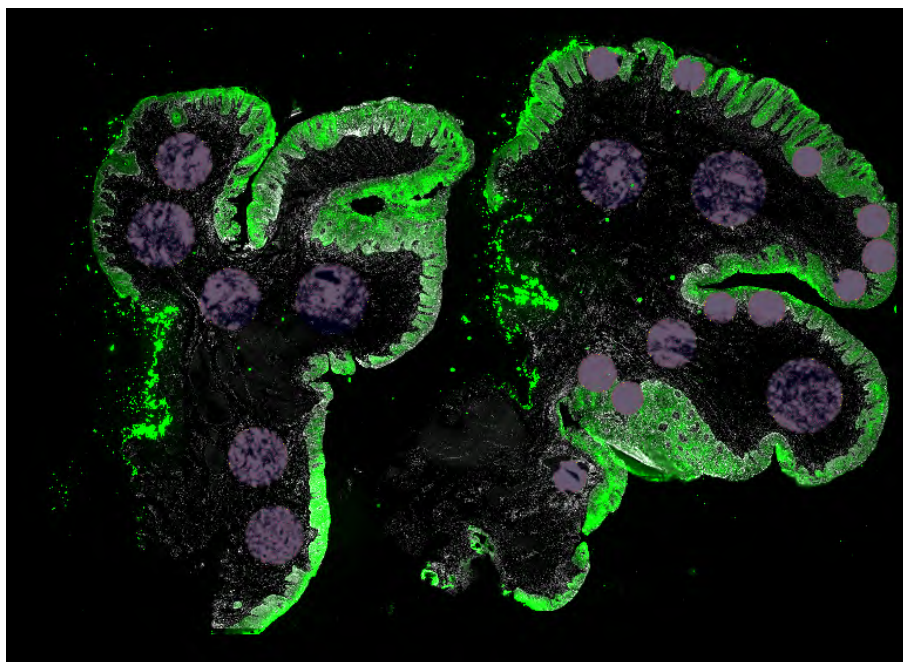


Figura 2.2: Imagen de una muestra de tumor anogenital de un paciente con anemia de Fanconi procesada mediante *t*-CyCIF. El canal de *Hoechst* (gris) es un colorante fluorescente que marca el ADN nuclear, permitiendo identificar la posición de cada núcleo y, por tanto, de cada célula en el tejido. El marcador *PAN-CK* (verde) identifica las células epiteliales tumorales, facilitando la delimitación de regiones con alta densidad tumoral.

stroma_ad_dysplasia). A partir de estas anotaciones, se extrajeron *subzonas* del tejido (en adelante denominadas *muestras*), que constituyen las unidades de análisis utilizadas en los procedimientos computacionales de esta tesis. Estas subzonas corresponden a áreas circulares delimitadas dentro de las zonas histológicas clasificadas, seleccionadas de manera que preservaran la organización celular local y el número de células. Ver Figura [2.3](#).

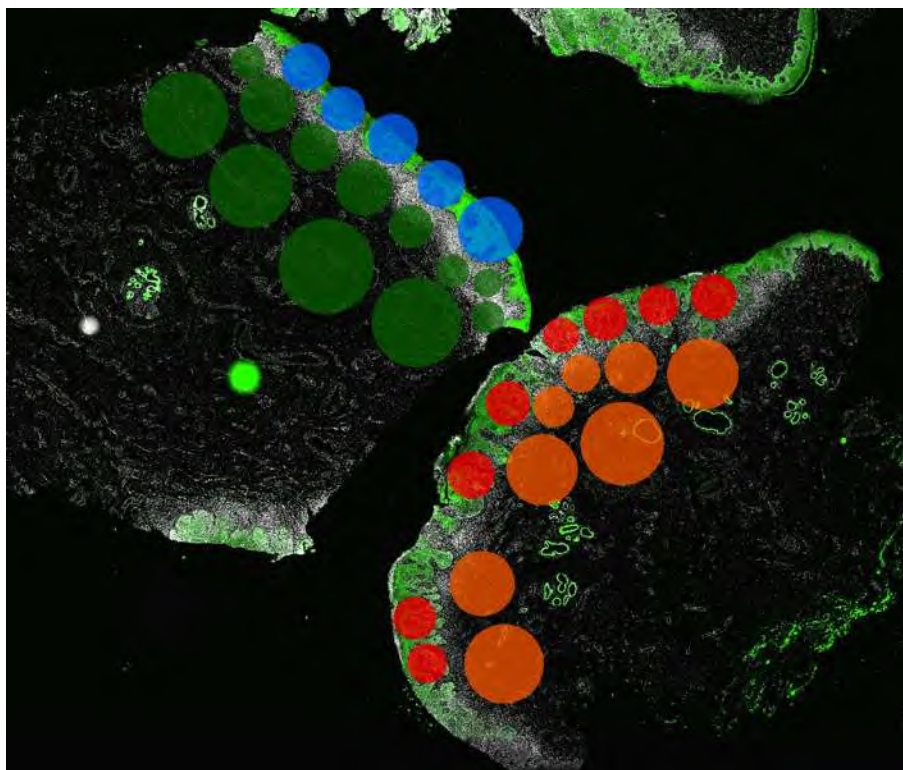


Figura 2.3: Imagen procesada mediante *t*-CyCIF. Las zonas delineadas por colores corresponden a regiones histológicas clasificadas por especialistas: **azul**, áreas de *displasia*; **verde**, estromas adyacentes a displasia; **rojo**, regiones de *carcinoma*; y **naranja**, estromas adyacentes a carcinoma.

Obtención de las coordenadas celulares

Las imágenes finales fueron segmentadas utilizando el algoritmo **StarDist** [42], que permite detectar núcleos y delimitar células individuales a partir de la señal de ADN. A partir de las máscaras de segmentación generadas, se extrajeron las coordenadas del centroide de cada célula, así como su área y forma aproximada.

Estos centroides, representados como coordenadas bidimensionales (x, y) , constituyen la base para los análisis espaciales posteriores. Además de la posición espacial, se asignó a cada célula un tipo celular específico en función de su perfil de marcadores inmunofenotípicos, utilizando la paquetería **SCIMAP** [35]. Finalmente, para cada muestra se generó un archivo en formato csv, donde se registraron las coordenadas de los centroides y el tipo celular correspondiente de cada célula, junto con información adicional. Una muestra de este tipo de imágenes segmentadas se muestra en la Figura 2.4a, donde se puede observar la subzona extraída para el análisis topológico. Asimismo, en la Figura 2.4b se muestra la representación de dicha subzona obtenida en Python a partir de los centroides y su tipo

celular.

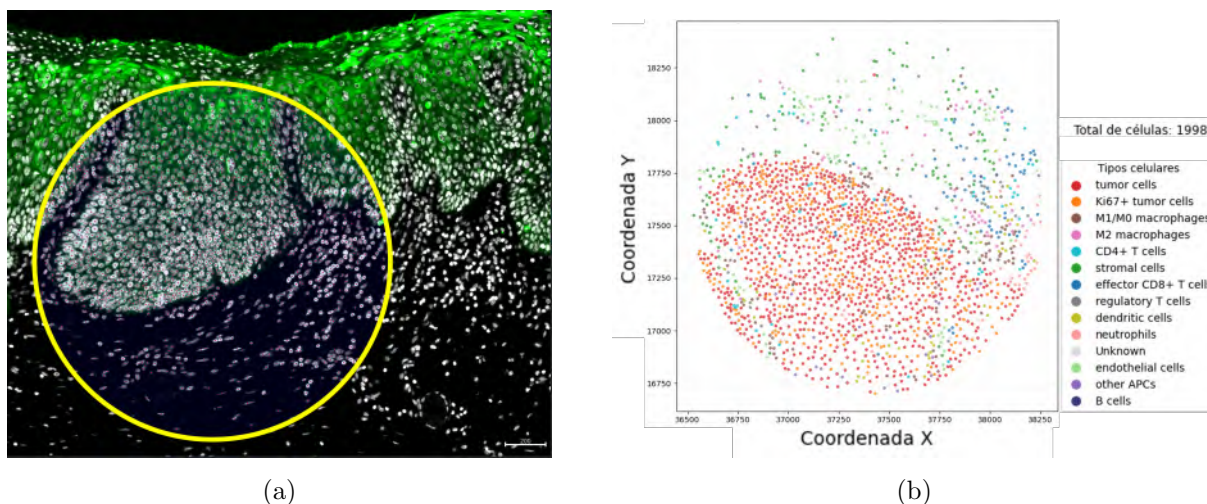


Figura 2.4: Ejemplo de una subzona de Displasia extraída de la muestra para el análisis topológico. En (a) se muestra la subzona seleccionada, delimitada por un círculo, donde se observan los núcleos celulares identificados a partir de los marcadores de inmunofluorescencia. En (b) se presenta su representación computacional generada en *Python* (esta representación puede encontrarse rotada con respecto a (a), sin alterar las relaciones espaciales entre las células), donde cada punto corresponde al centroide de una célula y el color indica su tipo celular. Los tipos celulares se detallan en el recuadro de la derecha.

Para estandarizar las visualizaciones y los análisis topológicos, se seleccionaron subzonas aproximadamente circulares de cada muestra. Esta elección elimina sesgos por geometría irregular, permite comparaciones consistentes entre muestras y facilita la interpretación de características topológicas [8].

Una descripción más detallada del procesamiento de estas imágenes se encuentra en la tesis de licenciatura del estudiante Pablo Siliceo [43].

2.3. Preprocesamiento de los datos

El preprocesamiento de los datos consistió en una serie de pasos orientados a garantizar la homogeneidad, calidad y utilidad de las coordenadas celulares para los análisis espaciales posteriores. Este procedimiento se aplicó a cada una de las muestras derivadas de las imágenes *t-CyCIF* en formato csv, asegurando que todas se encontraran en condiciones comparables y adecuadamente normalizadas para los análisis computacionales subsecuentes. Todos los códigos empleados en las siguientes secciones se encuentran disponibles en el repositorio de GitHub: <https://github.com/Luis0815/tda-histopathology-classifier>.

2.3.1. Limpieza y estandarización de coordenadas

En primer lugar, se verificó la estructura de cada archivo para identificar y unificar las columnas relevantes: identificador celular (llamado dentro del archivo csv `CellID`), coordenadas de centroides (`X_centroid`, `Y_centroid`) y tipo celular (`phenotype_key`).

Dado que la nomenclatura de las columnas variaba entre archivos, se implementó una detección condicional para garantizar la correcta asignación de nombres. Este primer paso dio como resultado una base de datos homogénea, en la que cada muestra presenta la misma estructura de variables, permitiendo su integración y posterior análisis espacial sin inconsistencias.

Se realizó un conteo del número de identificadores únicos (`CellID`) en cada archivo con el fin de comprobar que todas las muestras tuvieran una cantidad de células comparable. En todos los casos, el número de células se mantuvo dentro de un rango estrecho —entre aproximadamente 1,990 y 2,010 células por muestra—, lo que corresponde a una variación menor al 1 % entre los archivos. Este control fue esencial, ya que en análisis posteriores se requiere que las muestras tengan tamaños celulares comparables para evitar sesgos en las métricas espaciales y topológicas ocasionados por cantidades muy diferentes en el número de células entre muestras.

Cabe mencionar que en etapas iniciales del trabajo se realizaron pruebas con conjuntos de datos que presentaban variaciones más amplias en el número de células, lo que generaba sesgos significativos: las diferencias observadas entre muestras estaban dominadas por la cantidad total de células más que por la organización espacial o las propiedades topológicas de las mismas. Por este motivo, se estableció como criterio mantener un número de células prácticamente constante entre muestras antes de proceder con el análisis comparativo. Durante esta etapa, también se revisaron los tipos celulares presentes en cada muestra, asegurando que no existieran valores nulos o células no identificadas. En caso de encontrarse células sin un tipo asignado, estas se etiquetaron como `Unknown` para su correcta diferenciación visual.

2.3.2. Visualización espacial de las células

Para facilitar la interpretación de los datos, se implementaron representaciones gráficas de la disposición espacial de las células, utilizando dos esquemas de coloración complementarios:

- **Coloración por tipo celular:** cada tipo celular se representó con un color único, lo que permitió observar en detalle la distribución espacial de cada tipo celular (Figura 2.4b).

- **Coloración por agrupamiento funcional:** los diferentes tipos celulares se representaron dentro de cuatro grupos funcionales definidos previamente cada uno de los cuales contiene células con funcionalidades biológicas similares (tabla 2.1). Este esquema permitió identificar patrones generales en la organización del tejido (Figura 2.5).

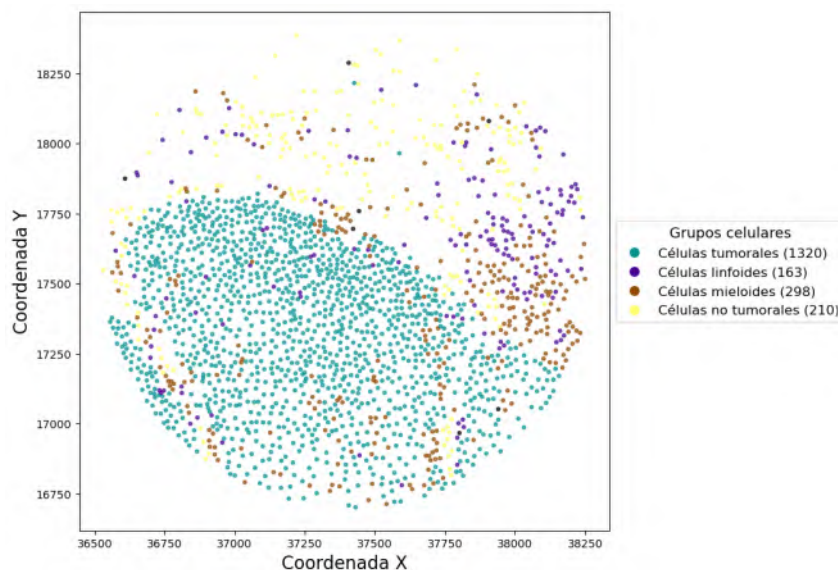


Figura 2.5: Representación espacial de la misma subzona mostrada en la Figura 2.4b donde los tipos celulares se agrupan en las cuatro categorías funcionales principales definidas en la Tabla 2.1.

Es importante resaltar que estas visualizaciones se emplearon como herramienta exploratoria visual para resaltar lo obtenido a partir de la extracción de tipo celular y su clasificación funcional. Estas representaciones proporcionaron una primera intuición de las diferencias estructurales entre las muestras.

2.4. Construcción de complejos simpliciales y obtención de diagramas de persistencia

Para capturar la topología de los datos espaciales celulares, se emplearon complejos de Vietoris–Rips. Como vimos en la Subsección 1.2.1, este tipo de complejo se construye a partir de un conjunto de puntos en el plano (en este caso, las coordenadas de los centroides celulares) y un parámetro de filtración ε , conectando puntos cuya distancia euclidiana es menor o igual a 2ε . El complejo describe la conectividad espacial entre las células y

permite estudiar sus características topológicas (componentes conexas, ciclos, cavidades) a diferentes escalas de resolución conforme se incrementa el valor de ε . A partir de estos complejos, se calcularon las dimensiones de cada uno de los grupos de homología y se obtuvieron los correspondientes diagramas de persistencia. Todo el procedimiento se implementó con la biblioteca GUDHI 3.8.0 en Python 3.7.12 [29], que permite construir complejos de Vietoris–Rips y extraer los pares de nacimiento y muerte (*birth–death*) asociados a cada clase topológica.

2.4.1. Cota máxima del parámetro de filtración (ε)

La construcción de un complejo de Vietoris–Rips depende de un parámetro de filtración ε , que define la distancia máxima a la cual se consideran conectados dos puntos del conjunto de datos. Aunque la distancia máxima entre puntos varía entre muestras, en el conjunto completo de datos la distancia euclidiana máxima observada entre centroides celulares es aproximadamente de 2000 unidades. Con base en esta escala, se utilizaron dos cotas máximas distintas para ε , dependiendo del tipo de análisis.

- **Análisis global (todas las células):** $\varepsilon = 1000$ unidades.
- **Análisis por grupos celulares:** $\varepsilon = 2000$ unidades.

La elección de estos valores se realizó a partir de pruebas exploratorias sobre las distancias intercelulares. Se observó que un valor de $\varepsilon = 1000$ permitía capturar adecuadamente la conectividad básica entre células sin generar un exceso de símlices, mientras que $\varepsilon = 2000$ resultaba más apropiado para los subconjuntos celulares, compensando la menor densidad espacial que presentan algunos grupos. De este modo, ambos valores proporcionaron un balance entre preservar la estructura topológica relevante y evitar la aparición de ruido. Finalmente, ambos valores fueron lo suficientemente grandes como para garantizar que en los diagramas de persistencia se observara únicamente un punto que tiende a infinito en la dimensión 0, correspondiente a la componente conexa principal del sistema celular.

2.4.2. Dimensiones de los grupos de homología analizados

La homología se estudió en las siguientes dimensiones:

- Dimensión 0: corresponde a las componentes conexas del espacio.
- Dimensión 1: representa ciclos o agujeros unidimensionales.

- **Dimensión 2:** identifica cavidades bidimensionales.

A partir de la filtración inducida por el parámetro ε , la biblioteca GUDHI 3.8.0 calculó las clases topológicas correspondientes a cada dimensión de homología y obtuvo la *homología persistente*, identificando los pares de *nacimiento* y *muerte* asociados a cada clase, generando como salida principal:

- **Diagramas de persistencia:** representaciones en las que cada punto (b, d) indica el valor de filtración en el que una característica topológica aparece (*nace*) y desaparece (*muere*).

En la mayoría de los casos, los diagramas de persistencia estuvieron dominados por un punto con valor de muerte infinito, correspondiente a la componente conexa que persiste a lo largo de toda la filtración y que representa la conectividad global del conjunto celular. El resto de los puntos representan características con vida finita, asociadas a estructuras locales o transitorias.

2.4.3. Ejemplo de los gráficos de salida

Para la visualización de los complejos de Vietoris-Rips y los diagramas de persistencia se desarrollaron dos scripts que automatizan el proceso [15]. Estos scripts son:

- **rips.py:** este script genera el complejo de Vietoris–Rips y su correspondiente diagrama de persistencia para todos los archivos de las muestras considerando todos los centroides celulares.
- **rips_grupos.py:** aplica el mismo procedimiento, pero segmentando previamente las células en los cuatro grupos funcionales (tumores, linfoides, mieloides y no tumorales), lo que permite explorar patrones topológicos específicos por grupos funcionales.

Las visualizaciones generadas por los scripts son:

- **Complejo de Rips:** representación bidimensional en la que los puntos negros indican los centroides celulares y las aristas grises conectan células separadas por una distancia menor o igual que 2ε . A modo de ejemplo, la Figura 2.6a muestra el complejo de Rips correspondiente a la subzona presentada en la Figura 2.4a.
- **Diagrama de persistencia:** cada punto en este gráfico representa un par *birth–death*, que indica cuándo aparece y desaparece una característica topológica. Los puntos cercanos a la diagonal suelen representar ruido topológico, mientras que los alejados reflejan estructuras persistentes y relevantes (Figura 2.6c).

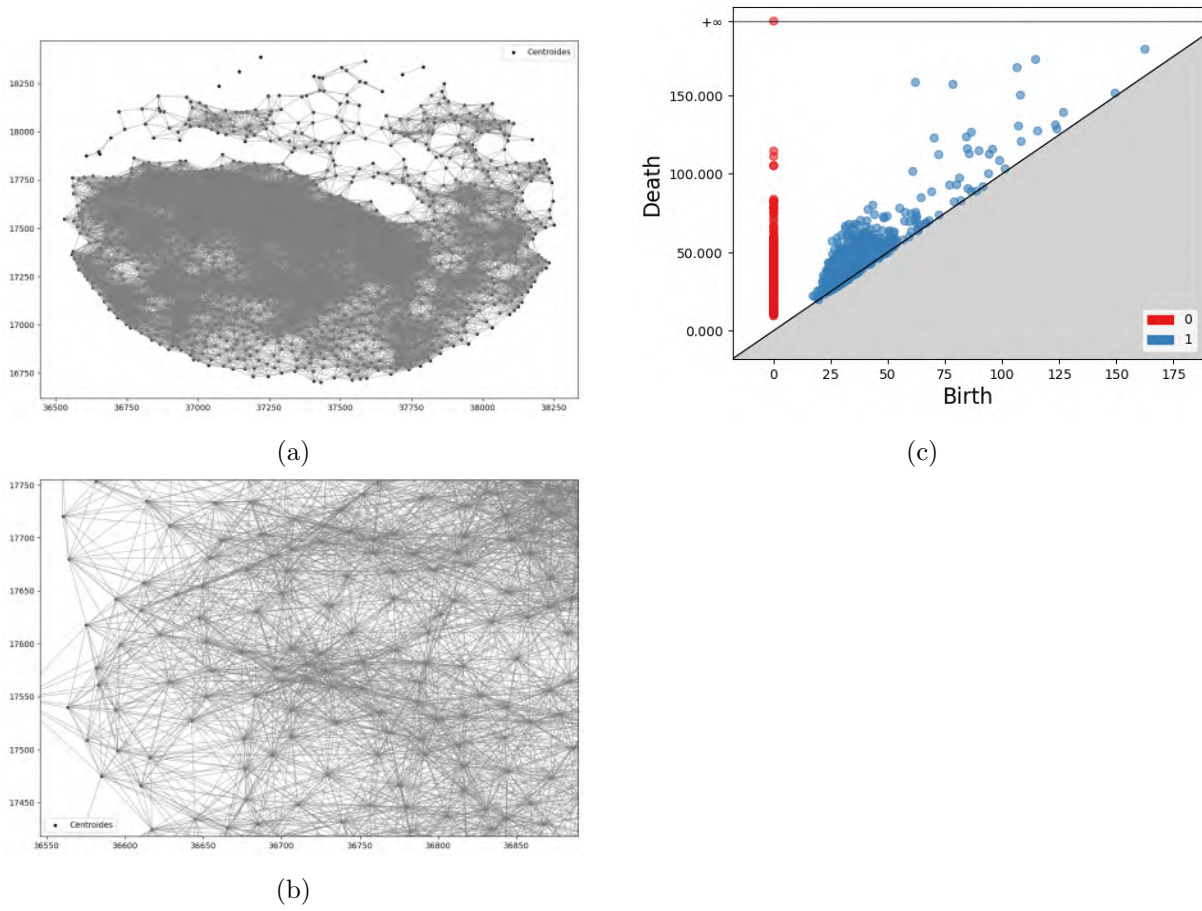


Figura 2.6: Resultados generados por los scripts del análisis topológico. (a) Complejo de Vietoris-Rips generado a partir de la muestra, mostrando los centroides celulares (puntos negros) y las conexiones entre células (aristas grises). (b) Ampliación del complejo para visualizar la formación de los símlices. (c) Diagrama de persistencia correspondiente, donde cada punto representa un par (*birth-death*).

Estas visualizaciones se generaron tanto para el conjunto completo de todas las células que conforman cada muestra como para cada uno de los grupos funcionales celulares por separado.

2.5. Cálculo de distancias entre diagramas de persistencia

2.5.1. Distancia de Wasserstein y cálculo computacional

Una vez obtenidos los diagramas de persistencia para cada muestra, el siguiente paso consistió en cuantificar la similitud entre los diagramas. En este estudio se utilizó principalmente la **distancia de Wasserstein** (ver Definición [18](#)), la cual mide el coste mínimo de transformar un diagrama en otro considerando el emparejamiento óptimo entre sus puntos (b, d) , que representan los tiempos de *nacimiento* y *muerte* de las características topológicas. En particular, se empleó la **distancia de Wasserstein de orden $p = 2$** , ya que ofrece un equilibrio adecuado entre la sensibilidad a diferencias topológicas relevantes y la robustez frente al ruido, como los puntos cercanos a la diagonal [\[13\]](#). El cálculo computacional se realizó utilizando la función `wasserstein_distance` de la biblioteca GUDHI 3.8.0 en Python 3.7.12. Para automatizar el proceso y permitir el análisis comparativo entre múltiples muestras, se desarrollaron los siguientes scripts:

- `distancias_wasserstein.py`: calcula matrices de distancias (Wasserstein y, opcionalmente, Cuello de botella) entre todos los diagramas de persistencia de un conjunto de muestras, almacenando los resultados en archivos csv.
- `distancias_grupos.py`: aplica el mismo procedimiento segmentando las comparaciones según los cuatro grupos celulares definidos (tumores, linfoides, mieloides y no tumores), generando matrices independientes para cada grupo.

2.5.2. Justificación del uso de la distancia de Wasserstein

La elección de la distancia de Wasserstein como métrica principal para comparar diagramas de persistencia se basó en criterios teóricos y prácticos:

- **Estabilidad matemática:** la distancia de Wasserstein es estable bajo pequeñas perturbaciones de los datos originales, lo que significa que cambios menores en la posición de los puntos del espacio no generan variaciones desproporcionadas en la métrica [\[12, 13\]](#).
- **Sensibilidad controlada al ruido:** a diferencia de la distancia de cuello de botella, que se enfoca únicamente en el punto más diferente entre dos diagramas, la de Wasserstein considera todas las correspondencias entre puntos, penalizando de forma proporcional según su persistencia. Esto reduce el impacto de ruido topológico

(puntos cercanos a la diagonal) y da mayor relevancia a las características persistentes [11].

- **Interpretabilidad:** su formulación se basa en el problema de transporte óptimo, lo que permite interpretarla como el *coste mínimo* necesario para transformar un diagrama en otro, ya sea moviendo puntos o enviándolos a la diagonal [30].
- **Versatilidad:** permite trabajar con distintos órdenes p (en este trabajo $p = 2$), ajustando la sensibilidad de la métrica a grandes o pequeñas discrepancias entre diagramas.
- **Eficiencia computacional:** existen implementaciones optimizadas en bibliotecas como GUDHI y Hera, lo que permite procesar grandes volúmenes de datos en tiempos razonables incluso en análisis masivos.

En conclusión, aunque en este trabajo también se calculó la distancia de cuello de botella como referencia [15], se priorizó la de Wasserstein por su capacidad para capturar diferencias topológicas de forma más gradual y robusta, facilitando un análisis cuantitativo más representativo en contextos biológicos con alta variabilidad.

2.6. Agrupamiento jerárquico y visualización

2.6.1. Procedimiento de agrupamiento

Una vez obtenidas las matrices de distancia (tanto para dimensión 0 como para dimensión 1), se aplicó un análisis de agrupamiento jerárquico con el objetivo de identificar patrones de similitud entre las muestras y evaluar la coherencia de la clasificación en relación con las variables biológicas de interés. El agrupamiento jerárquico es un método de análisis de datos que organiza los elementos en una estructura en forma de árbol (*dendrograma*), de manera que los objetos más similares se agrupan primero y, progresivamente, se van uniendo a otros grupos hasta formar un único clúster. Este procedimiento no requiere conocer de antemano el número de grupos. Esta propiedad lo hace especialmente útil en contextos exploratorios como el presente, donde el objetivo es descubrir relaciones intrínsecas entre muestras.

Para aplicar el análisis de agrupamiento jerárquico es necesario definir un *método de enlace* (*linkage*), el cual determina cómo se calcula la distancia entre dos grupos (o clústeres) en cada etapa del proceso de fusión. En otras palabras, el método de enlace especifica la forma en que se mide la similitud entre conjuntos de elementos ya agrupados. Existen diversos criterios para ello:

- **Single linkage** (enlace simple): la distancia entre dos clústeres se define como el mínimo de las distancias entre todos los pares de puntos, donde cada par está formado por un punto de cada clúster. Este método tiende a generar clústeres alargados o con estructura encadenada.
- **Complete linkage** (enlace completo): la distancia entre dos clústeres se define como el máximo de las distancias entre todos los pares de puntos, donde cada par está formado por un punto de cada clúster. Este método favorece la formación de grupos compactos y es menos sensible a valores atípicos.
- **Average linkage** (enlace promedio): considera el promedio de todas las distancias entre los elementos de los dos clústeres. Es un método intermedio entre los dos anteriores y suele producir agrupamientos equilibrados.
- **Ward**: minimiza la varianza dentro de los grupos en cada paso de la fusión, buscando maximizar la homogeneidad interna de los clústeres. Es especialmente útil cuando se espera que los grupos sean relativamente esféricos y de tamaño similar.

Para seleccionar el método de enlace más adecuado, se evaluaron las distintas opciones mediante el **coeficiente de correlación cophenética**, el cual cuantifica en qué medida las distancias entre las muestras en el dendrograma preservan las distancias originales de la matriz de entrada. Valores cercanos a 1 indican que la estructura jerárquica obtenida representa fielmente las relaciones de distancia originales. Los valores obtenidos para cada método de enlace se muestran en la Tabla 2.3. En el caso de las distancias de Wasserstein con clase de dimensión 0, el método average linkage presentó el coeficiente más alto (0.872), mientras que para las clases de dimensión 1, también fue el método average linkage el que alcanzó el mejor ajuste (0.879).

Tabla 2.3: Coeficientes de correlación cophenética para los distintos métodos de enlace.

	Ward	Single	Complete	Average
wasserstein_dim0	0.833	0.839	0.866	0.872
wasserstein_dim1	0.699	0.578	0.871	0.879

En función de estos resultados, se seleccionó el método average linkage para la construcción de los dendrogramas y clustermaps, dado que ofrece la mayor consistencia con las distancias topológicas originales. Este método calcula la distancia entre dos clústeres como el promedio de todas las distancias entre pares de elementos de cada clúster, ofreciendo un balance adecuado entre:

- el **efecto de encadenamiento** (*chaining effect*) característico del método de enlace *single*, el cual tiende a formar cadenas alargadas de elementos conectados por distancias mínimas. Este comportamiento puede producir agrupamientos poco compactos, especialmente cuando existen puntos intermedios que conectan regiones separadas del espacio de datos;
- y la alta sensibilidad a valores extremos del método de enlace *complete*, que define la distancia entre clústeres como la mayor distancia entre sus elementos. Este enfoque puede sobreestimar las diferencias entre grupos cuando existen pares de puntos atípicamente distantes.

De esta manera, el método de enlace *average* mantiene una estructura interpretable y estable en contextos biológicos con alta variabilidad morfológica. Este comportamiento coincide con reportes previos, en los cuales dicho método ha sido utilizado exitosamente para agrupar datos basados en distancias de Wasserstein, gracias a su capacidad de suavizar el efecto de variaciones extremas sin perder la estructura global de los datos [11, 30].

2.6.2. Generación de *clustermaps*

El agrupamiento jerárquico se representó mediante *clustermaps* generados con la biblioteca `seaborn` 0.12.2 en `Python` 3.7.12, utilizando como entrada las matrices de distancia calculadas previamente. Se realizaron visualizaciones independientes para las dimensiones de homología 0 y 1. En cada *clustermmap*, las muestras fueron anotadas con información categórica adicional, como tipo de muestra, estado Fanconi y origen anatómico, codificada mediante esquemas de color consistentes. Esto permitió observar la estructura jerárquica del agrupamiento y explorar visualmente posibles asociaciones con variables clínicas o experimentales. El resultado de este análisis proporciona una representación integrada de la similitud topológica entre muestras, facilitando la interpretación de las relaciones detectadas por las métricas de persistencia.

2.7. Análisis de diferencias entre grupos biológicos mediante TopKAT

El objetivo de este análisis fue evaluar si la organización geométrica de las células en el tejido, descrita a través de su estructura topológica, se asocia con las condiciones clínicas de las muestras, para ello, se utilizó `TopKAT`, una herramienta que combina la técnica de TDA basada en homología persistente con pruebas de asociación mediante kernels [40].

En términos generales, **TopKAT** emplea la homología persistente para caracterizar la cantidad y el tamaño de las estructuras topológicas presentes en las muestras (principalmente componentes conexas y ciclos), lo que proporciona una representación robusta de la organización espacial celular. Posteriormente, el método evalúa si las muestras que presentan estructuras topológicas similares también comparten características clínicas, como en este caso la condición de anemia de Fanconi (AF vs. no AF). Esta comparación se realiza mediante un enfoque de prueba no paramétrica basado en kernels, que permite medir la similitud entre diagramas de persistencia y realizar inferencias estadísticas sobre sus asociaciones con variables como las condiciones histológicas (por ejemplo, carcinoma, displasia o estroma) y clínicas (presencia o ausencia de anemia de Fanconi) de las muestras.

Una de las principales ventajas de **TopKAT** es que integra la información geométrica multiescala contenida en los diagramas de persistencia, permitiendo comparar muestras en función de la presencia, ausencia y tamaño de las estructuras topológicas observadas. Además, la homología persistente confiere al análisis un efecto de suavizado frente al ruido o variabilidad técnica, ofreciendo una descripción más estable y representativa de la arquitectura celular. En conjunto, esta herramienta proporciona un marco metodológico que además de describir patrones espaciales complejos, también permite contrastar formalmente si dichas estructuras se asocian con condiciones biológicas o clínicas específicas.

TopKAT extiende el marco de las pruebas de asociación de kernels (Kernel Association Tests, KAT) a datos topológicos, lo que permite:

- Comparar grupos sin suponer distribuciones paramétricas de los datos.
- Integrar información de todos los grupos de homología (H_0 , H_1 , etc.) de manera simultánea.
- Evaluar la significancia tanto de forma global como por grupo de homología.
- Detectar el valor de ε que maximizan la discriminación entre grupos.

En este estudio, se analizaron comparaciones binarias entre muestras provenientes de pacientes con anemia de Fanconi (FA) y sin anemia de Fanconi (no FA) para tres categorías histológicas: carcinoma, estroma y displasia.

A diferencia de las distancias clásicas empleadas en análisis de homología persistente (como las distancias de Wasserstein o cuello de botella), **TopKAT** define una medida de disimilitud específica denominada *total dissimilarity*, la cual enfatiza las diferencias tanto en la cantidad como en la duración (*lifespan*) de las clases de homología persistentes entre

pares de diagramas. Para dos diagramas de persistencia Z_i y Z_j , la disimilitud de grado- k se define como:

$$D_k(Z_i, Z_j) = \sum_{p=1}^{n_k} |\ell(z_i^k(p)) - \ell(z_j^k(p))| \quad (2.1)$$

donde n_k es el número máximo de clases de homología persistentes de grado k entre ambos diagramas, y $\ell(z^k(p))$ representa la duración (*lifespan*) de la p -ésima clase de homología más persistente en el diagrama Z .

Las diferencias en el número de características se completan con ceros, y la suma de desviaciones se utiliza como una métrica de disimilitud. Posteriormente, esta matriz de disimilitud se transforma en una matriz kernel $\mathbf{K}$ que refleja la similitud topológica entre las muestras.

El procedimiento completo se resume a continuación:

1. **Construcción de complejos de Vietoris-Rips y diagramas de persistencia:** Para cada imagen, se generó la filtración hasta un umbral máximo predefinido y se identificaron las clases de homología persistentes en dimensiones 0 y 1.
2. **Cálculo de matrices de similitud:** Usando la función `rips_similarity_matrix` de TopKAT, se calculó la matriz kernel $\mathbf{K}$ a partir de la *total dissimilarity* definida sobre las clases de homología persistentes en dimensiones 0 y 1, descrita anteriormente.
3. **Pruebas de asociación TopKAT:** Se aplicó la función `TopKAT()` con variable de respuesta binaria (1 = FA, 0 = no FA) para evaluar la significancia estadística global y por dimensión homológica.
4. **Identificación de la escala topológica más discriminante:** Se utilizó la función `scale_importance()` para explorar un rango de umbrales ε , determinando el valor que maximiza las diferencias topológicas entre grupos. Este umbral se interpretó como la escala espacial en la que se evidencian mayores discrepancias en la conectividad celular.
5. **Generación de matrices de conectividad:** A partir del ε óptimo, se construyeron matrices de adyacencia que representan la conectividad entre células, lo que permitió visualizar y comparar patrones estructurales característicos de cada grupo.

Este enfoque permitió detectar diferencias significativas en la topología global de las muestras entre FA y no FA e identificar la escala espacial específica en la que dichas diferencias son más pronunciadas, lo cual puede tener implicaciones biológicas relevantes en

la caracterización del microambiente tumoral.

Con este conjunto de herramientas y procedimientos implementados, se estableció un flujo completo desde la extracción de coordenadas celulares hasta el análisis topológico cuantitativo y la evaluación de diferencias entre grupos biológicos. A continuación, se presentan los resultados obtenidos, mostrando cómo las métricas de persistencia y los análisis de agrupamiento permitieron identificar patrones espaciales característicos en el microambiente tumoral de pacientes con y sin anemia de Fanconi.

Capítulo 3

Resultados

Este capítulo presenta los resultados del análisis topológico de las distribuciones espaciales observadas en las muestras de tejido tumoral de paciente con y sin AF obtenidas mediante t -CyCIF. El objetivo principal fue evaluar si la homología persistente permite distinguir entre zonas tisulares correspondientes a distintas condiciones histológicas (carcinoma, displasia y estroma). Posteriormente, se exploró si existen diferencias topológicas entre pacientes con anemia de Fanconi (FA) y sin anemia de Fanconi (no FA).

Como se describe en la sección previa, los análisis de las propiedades topológicas se llevaron a cabo siguiendo la siguiente secuencia metodológica:

1. Construcción de complejos de Vietoris–Rips y extracción de características topológicas.
2. Representación de los resultados mediante diagramas de persistencia.
3. Cálculo de matrices de distancia de Wasserstein.
4. Análisis de agrupamiento jerárquico y visualización mediante *clustermaps*.
5. Comparación entre grupos FA y no FA a través del método TopKAT.

Los análisis se realizaron considerando dos enfoques complementarios: (i) un **análisis global**, en el que se emplearon los centroides de todas las células presentes en cada muestra, sin distinguir por tipo celular; y (ii) un **análisis estratificado**, en el que se consideraron de forma individual los centroides correspondientes a cuatro grupos funcionales definidos previamente: tumorales, linfoides, mieloides y no tumorales (Tabla [2.1](#)).

3.1. Análisis global de patrones topológicos

Como primer paso, se realizó un análisis topológico global considerando todas las células en conjunto. El objetivo de este primer análisis fue evaluar si las distribuciones espaciales globales de los centroides celulares son lo suficientemente informativas como para distinguir entre diferentes condiciones histológicas, específicamente entre estroma contra displasia y carcinoma.

Este análisis inicial permite validar la pertinencia de utilizar los centroides celulares como representación de la organización espacial del microambiente tisular, y verificar si las diferencias topológicas entre condiciones histológicas son detectables antes de introducir distinciones más finas, como la presencia o ausencia de anemia de Fanconi.

3.1.1. Construcción de complejos de Rips y diagramas de persistencia

A partir de las coordenadas de los centroides celulares de cada muestra, se construyeron los complejos de Vietoris-Rips. Para el análisis global, se empleó un valor máximo de filtración de $\varepsilon = 1000$ unidades, lo cual aseguró una conectividad suficiente sin introducir ruido topológico excesivo.

En la Figura 3.1 se observa cómo el incremento del parámetro de filtración ε provoca que las componentes conexas inicialmente aisladas comiencen a fusionarse, hasta que finalmente toda la nube de puntos se conecta en una única componente. Este comportamiento se registra mediante los diagramas de persistencia, que muestran los intervalos de existencia de las clases de homología: clases de dimensión 0 asociadas a componentes conexas y clases de dimensión 1 asociadas a ciclos unidimensionales, a lo largo de la filtración.

En la Figura 3.2 se muestran ejemplos de los diagramas de persistencia obtenidos para muestras correspondientes a las cuatro condiciones histológicas analizadas: carcinoma, displasia, estroma adyacente a carcinoma y estroma adyacente a displasia. En todos los casos aparece una clase de homología de dimensión 0 que persiste hasta infinito, lo que refleja que el valor máximo de filtración ε genera una única componente conexa principal, tal como se describió en la sección de Métodos. Por otra parte, el rango de nacimiento y muerte de las clases de homología de dimensión 0, así como la dispersión de las clases de homología de dimensión 1, parecen variar según la condición histológica. Estas diferencias, detectables de forma visual, sugieren la existencia de variaciones topológicas entre las distintas condiciones. Para corroborar estas observaciones de manera sistemática y cuantitativa, se calcularon las distancias de Wasserstein entre todos los pares de diagramas de persistencia correspondientes a las muestras analizadas, y dichas distancias se organizaron en matrices de distancia.

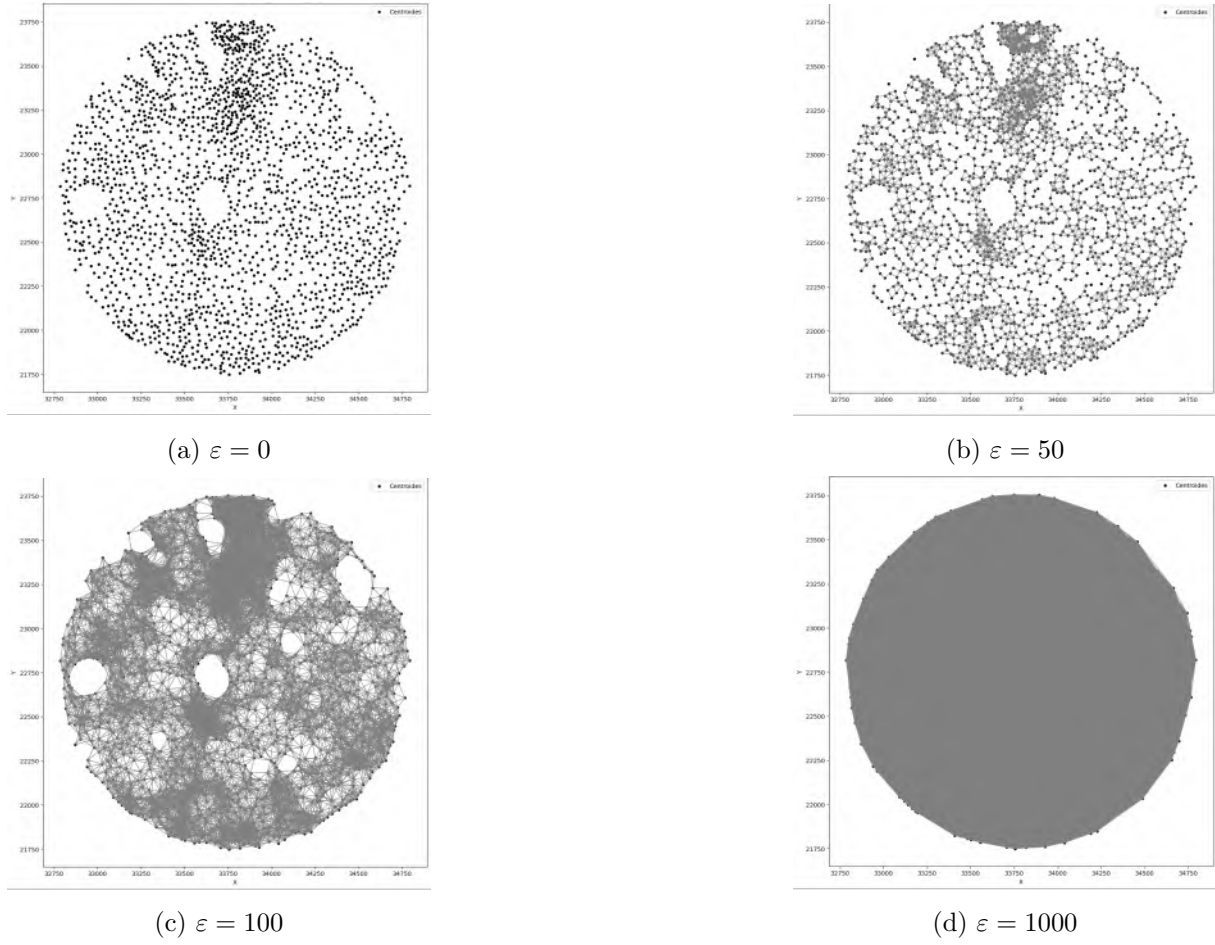


Figura 3.1: Ejemplo de complejos de Vietoris-Rips en una muestra de Displasia, construidos con diferentes valores del parámetro de filtración ε , el cual representa el radio de vecindad empleado para conectar puntos (centroides celulares) que se encuentran a una distancia menor o igual a 2ε . A medida que ε aumenta, desde (a) $\varepsilon = 0$ hasta (d) $\varepsilon = 1000$, los centroides se conectan progresivamente, fusionando componentes conexas y dando lugar a la formación de ciclos topológicos.

3.1.2. Cálculo de matrices de distancia

Una vez obtenidos los diagramas de persistencia para cada muestra, el siguiente paso fue cuantificar su grado de similitud mediante distancias. Para este propósito se utilizó la **distancia de Wasserstein de orden 2**, ampliamente empleada en el TDA por su capacidad para comparar distribuciones en el plano (nacimiento–muerte) de manera robusta frente a variaciones locales.

El procedimiento consistió en construir, de forma independiente para las dimensiones de homología 0 y 1, una matriz de distancia cuadrada y simétrica, donde la entrada D_{ij}

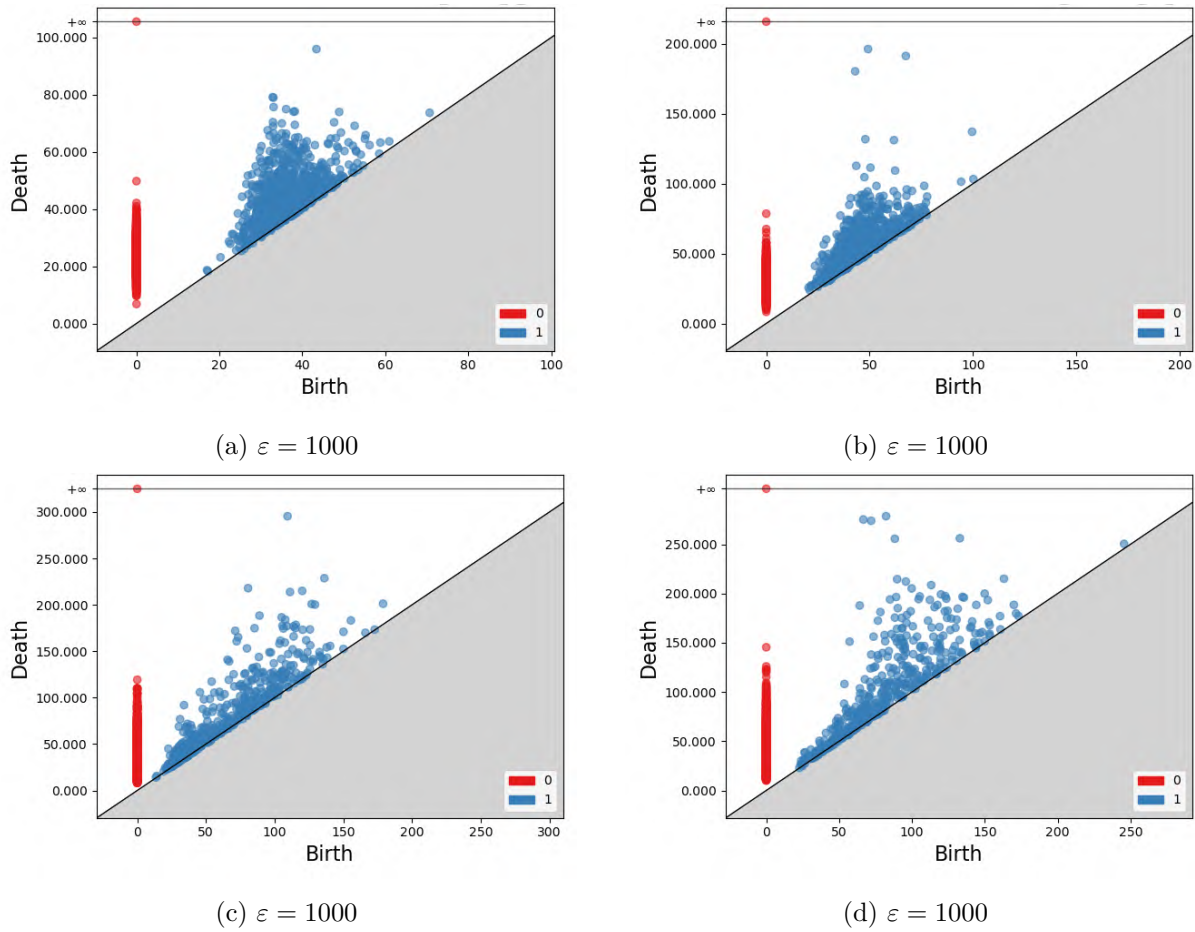


Figura 3.2: Diagramas de persistencia correspondientes a las clases de homología de dimensión 0 y 1 para las cuatro condiciones: (a) carcinoma, (b) displasia, (c) estroma adyacente a carcinoma y (d) estroma adyacente a displasia, generados con un valor de $\varepsilon = 1000$. En las clases de homología de dimensión 0 se observa una única clase que persiste indefinidamente, mientras que las diferencias más marcadas entre condiciones se reflejan en la homología de dimensión 1, mostrando variaciones en la aparición y duración de ciclos.

representa la distancia entre los diagramas de persistencia correspondientes a las muestras i y j . De este modo, se obtuvieron dos matrices, una para cada tipo de clase de homología considerada.

Tabla 3.1: Muestra de parte de la matriz de distancia obtenida mediante la métrica de Wasserstein para la dimensión de homología 0, asociada a la dinámica de las componentes conexas.

	D1	C1	S1	C2	S2	S3	D2	C3
Dysplasia 1 (D1)	0.0	20464.5	13734.6	13693.4	26725.2	35421.3	14926.1	21342.0
Carcinoma 1 (C1)	20464.5	0.0	10663.0	6952.6	6541.8	43913.8	5630.5	970.1
Stroma ady. Carc. 1 (S1)	13734.6	10663.0	0.0	9826.1	15416.1	33606.6	8687.6	10970.7
Carcinoma 2 (C2)	13693.4	6952.6	9826.1	0.0	13465.0	41378.1	1900.2	7913.4
Stroma ady. Dysp. 1 (S2)	26725.2	6541.8	15416.1	13465.0	0.0	47265.1	12101.4	5583.4
Stroma 3 (S3)	35421.3	43913.8	33606.6	41378.1	47265.1	0.0	41126.7	44140.5
Dysplasia 2 (D2)	14926.1	5630.5	8687.6	1900.2	12101.4	41126.7	0.0	6572.3
Carcinoma 3 (C3)	21342.0	970.1	10970.7	7913.4	5583.4	44140.5	6572.3	0.0

Tabla 3.2: Muestra de parte de la matriz de distancia obtenida mediante la métrica de Wasserstein para la dimensión de homología 1, asociada a la dinámica de los ciclos.

	D1	C1	S1	C2	S2	S3	D2	C3
Dysplasia 1 (D1)	0.0	5457.3	3518.3	3746.9	6803.5	9423.7	4091.2	5581.7
Carcinoma 1 (C1)	5457.3	0.0	4578.2	2547.6	3385.4	9023.8	1693.7	621.1
Stroma ady. Carc. 1 (S1)	3518.3	4578.2	0.0	4145.0	5105.9	6813.6	3877.0	4588.7
Carcinoma 2 (C2)	3746.9	2547.6	4145.0	0.0	5145.4	9360.4	1327.0	2815.3
Stroma ady. Dysp. 1 (S2)	6803.5	3385.4	5105.9	5145.4	0.0	8792.8	4379.6	3042.2
Stroma 3 (S3)	9423.7	9023.8	6813.6	9360.4	8792.8	0.0	9045.2	9015.2
Dysplasia 2 (D2)	4091.2	1693.7	3877.0	1327.0	4379.6	9045.2	0.0	1915.8
Carcinoma 3 (C3)	5581.7	621.1	4588.7	2815.3	3042.2	9015.2	1915.8	0.0

Las matrices de distancia mostradas en las Tablas [3.1](#) y [3.2](#) son una muestra de las matrices generadas en este trabajo, las cuales representan el grado de disimilitud topológica entre pares de diagramas de persistencia : **cuanto mayor es el valor, mayor es la diferencia entre los diagramas de persistencia correspondientes**, mientras que valores cercanos a cero indican alta similitud. Es por este motivo que la diagonal de la matrices es cero, ya que representa la distancia entre un diagrama y el mismo.

Tabla 3.3: Resumen de distancias de Wasserstein en dimensión H_0 .

Par de muestras	Rango	Media	Desviación estándar
Carcinoma vs Carcinoma	0 – 27052	6410	4412
Displasia vs Displasia	0 – 16357	6064	4199
Displasia vs Carcinoma	511 – 28546	7655	5239
Carcinoma vs Estroma adyacente al carcinoma	501 – 51160	12137	8133
Displasia vs Estroma adyacente a la displasia	677 – 41936	14795	9830
Estroma adyacente al carcinoma vs Estroma adyacente a la displasia	312 – 51567	17269	11145

Tabla 3.4: Resumen de distancias de Wasserstein en dimensión H_1 .

Par de muestras	Rango	Media	Desviación estándar
Carcinoma vs Carcinoma	0 – 6040	2327	1115
Displasia vs Displasia	0 – 4359	2094	1080
Displasia vs Carcinoma	685 – 5723	2528	1101
Carcinoma vs Estroma adyacente al carcinoma	711 – 9092	3778	1752
Displasia vs Estroma adyacente a la displasia	823 – 9424	4360	2204
Estroma adyacente al carcinoma vs Estroma adyacente a la displasia	494 – 8979	4430	2003

En general, considerando las clases de homología de dimensión 0, las muestras de **carcinoma** presentan distancias relativamente bajas entre sí (rango: 0 – 27,052; media: 6,410; desviación estándar: 4,412). Esto indica que la organización de los conglomerados celulares en estas muestras es bastante consistente, es decir, los patrones de conectividad de las células tumorales son similares entre diferentes muestras de carcinoma. Al comparar **carcinoma** con **estroma adyacente al carcinoma**, las distancias son considerablemente mayores (rango: 501 – 51,160; media: 12,137; desviación estándar: 8,133), lo que refleja diferencias importantes en la conectividad: la distribución y agrupamiento de las células en el estroma adyacente difiere notablemente de la del tejido tumoral. Las distancias entre **displasia** y **carcinoma** (rango: 511 – 28,546; media: 7,655; desviación estándar: 5,239) muestran valores intermedios, lo que sugiere que algunas muestras de displasia presentan patrones de conectividad parcialmente similares al carcinoma. En contraste, las distancias entre **displasia** y **estroma adyacente a la displasia** (rango: 677 – 41,936; media: 14,795; desviación estándar: 9,830) son elevadas y más variables, indicando que la disposición de los conglomerados celulares en el estroma adyacente a displasia es más diversa y menos comparable con la propia displasia. Además, al comparar los dos tipos de estroma, se observa que **estroma adyacente al carcinoma** vs **estroma adyacente a la displasia** presentan distancias elevadas y mayor variabilidad (rango: 312 – 51,567; media: 17,269; desviación estándar: 11,145), lo que indica que la organización de los conglomerados celulares en los estromas depende de su proximidad a la lesión y difiere significativamente entre sí, ver Figura 2.3. En resumen, los resultados obtenidos a partir de las clases de homología de dimensión 0 indican que las muestras de **carcinoma** presentan alta similitud entre sí, al igual que las muestras de **displasia**. Las comparaciones cruzadas muestran que **displasia** y **carcinoma** mantienen similitudes intermedias, mientras que las muestras de estroma presentan diferencias más pronunciadas tanto con respecto a las otras zonas como entre ellas mismas. En particular, **estroma adyacente al carcinoma** y **estroma adyacente a displasia** exhiben distancias elevadas y mayor variabilidad. En conjunto, estos resultados sugieren que podrían identificarse patrones to-

pológicos característicos para los distintos tipos de tejido, aunque se requiere un análisis más detallado para confirmarlo.

En las clases de homología de dimensión 1 se observa un patrón general similar, aunque las distancias tienden a ser menores que en H_0 . Las muestras de **carcinoma** presentan distancias bajas entre sí (rango: 0 – 6,040; media: 2,327; desviación estándar: 1,115), lo que indica que la presencia y disposición de los ciclos unidimensionales es consistente entre estas muestras. Las distancias entre **carcinoma** y **estroma adyacente al carcinoma** (rango: 711 – 9,092; media: 3,778; desviación estándar: 1,752) y entre **displasia** y **estroma adyacente a la displasia** (rango: 823 – 9,424; media: 4,360; desviación estándar: 2,204) muestran diferencias moderadas. Por otra parte, la comparación entre **estroma adyacente al carcinoma** y **estroma adyacente a la displasia** (rango: 494 – 8,980; media: 4,430; desviación estándar: 2,003) revela que, aunque ambos tipos de estroma presentan cierta consistencia interna, existen diferencias en la organización de los ciclos unidimensionales entre estos dos tipos de estroma. Esto sugiere que, aunque la organización espacial general de los tejidos varía entre tipos de tejido, la formación de ciclos dentro de cada conglomerado es relativamente consistente.

Las tablas 3.3 y 3.4 resumen los resultados de las distancias entre diagramas de persistencia para las clases de homología de dimensión 0 y 1, respectivamente. En general, las diferencias topológicas más evidentes entre los tipos de tejido se manifiestan principalmente en las clases de H_0 , correspondientes a componentes conexas, lo que indica que la distribución global de los conglomerados celulares varía significativamente entre Carcinoma, Displasia y los diferentes tipos de estroma. Por el contrario, la presencia de ciclos unidimensionales es relativamente más consistente entre los tejidos, reflejando que los patrones locales de vacíos o ciclos dentro de los conglomerados celulares son similares incluso entre tipos de tejido distintos. Estos resultados sugieren que, para diferenciar tipos de tejido a partir de la organización espacial celular, las métricas derivadas de las clases de H_0 son más informativas, mientras que las clases de H_1 aportan información complementaria sobre la topología local de los conglomerados.

Estos resultados muestran que, aunque algunas muestras de diferentes tipos de tejido pueden compartir similitudes topológicas parciales, los tejidos tienden a diferenciarse en términos de conectividad y presencia de ciclos. Esto sugiere que la información topológica capturada mediante las distancias de Wasserstein puede resultar útil para identificar patrones de similitud entre muestras con respecto a su condición histológica y puede orientar procesos de clasificación. Técnicas de clustering jerárquico sobre estas matrices permitirían visualizar si existen agrupamientos naturales determinados por la condición histológica.

3.1.3. Clustering jerárquico

Con el propósito de evaluar si las distancias topológicas reflejan la organización biológica de las muestras según su **condición histológica** (carcinoma, displasia y estroma), se aplicó *clustering* jerárquico a las matrices de distancia construidas a partir de las clases de homología de dimensión 0 y 1. Este procedimiento permite representar gráficamente la similitud entre muestras mediante dendrogramas y mapas de calor, facilitando la identificación de agrupamientos naturales basados en su estructura topológica.

En la Figura 3.3 se muestra el análisis correspondiente a la homología en dimensión 0. Se observa que las muestras clasificadas como carcinoma y displasia tienden a agruparse de forma diferenciada respecto a los estromas adyacentes, validando los resultados estadísticos mostrados en la Tabla 3.3. Esta separación indica que las características topológicas asociadas con las conexiones entre conglomerados celulares son informativas para distinguir entre carcinomas y displasias frente a estromas. No obstante, se observan algunas excepciones en las que ciertas muestras no se agrupan con otras de la misma condición histológica. Estas desviaciones pueden explicarse por al menos tres factores:

1. **Zonas de transición histológica:** algunas muestras incluyen regiones mixtas, por ejemplo, *displasia–estroma*, lo que puede hacer que se clasifiquen según la condición con mayor similitud topológica. Como se observa en la Figura 2.5, la mayoría de las células pertenecen al grupo tumoral, por lo que resulta aceptable que estas muestras (color negro en el dendrograma) estén más relacionadas con displasia o carcinoma.
2. **Grado de displasia:** las displasias de bajo grado (LG) presentan una organización celular más cercana al estroma adyacente que las de alto grado, lo que puede provocar su agrupamiento con muestras estromales.
3. **Error de delimitación manual:** dado que la selección de regiones fue realizada por un patólogo, no se descarta cierta variabilidad o error humano en la segmentación.

En la Figura 3.3 se observa que algunas regiones incluyen mezclas entre displasias de bajo grado, zonas de transición y estroma adyacente, o regiones mixtas denominadas **displasia–estroma**, donde el límite entre el epitelio displásico y el estroma no es nítido. Estas transiciones topológicas se reflejan en valores intermedios de la matriz de distancias de Wasserstein, lo que dificulta su separación completa mediante clustering jerárquico. Por tanto, las mezclas puntuales observadas en el dendrograma no invalidan la separación global por condición histológica; más bien, señalan la existencia de casos intermedios o con ruido biológico que merecen inspección adicional. En conjunto, los resultados obtenidos a partir de las clases de homología de dimensión 0 indican que las diferencias en

la conectividad celular constituyen un descriptor topológico robusto para distinguir entre regiones epiteliales y estromales.

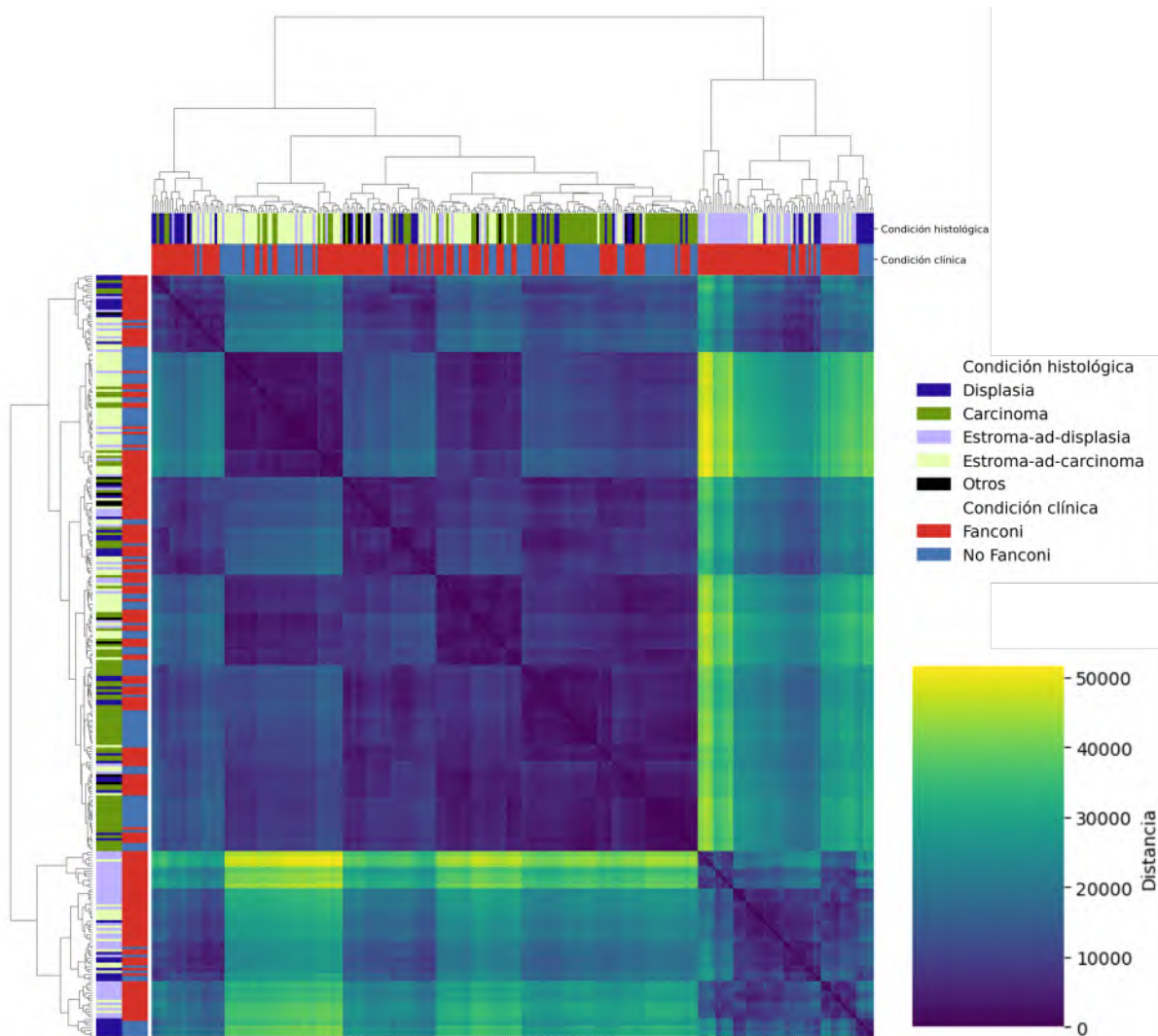


Figura 3.3: Clustermap jerárquico basado en las distancias de Wasserstein para las clases de homología de dimensión 0. El dendrograma muestra la relación jerárquica entre las muestras. El mapa de calor representa la magnitud de la distancia: los colores azul indican mayor similitud topológica entre muestras, mientras que los tonos verdes reflejan menor similitud. Las barras de color superiores indican la condición histológica (displasia – azul oscuro, carcinoma – verde, estroma adyacente a displasia – lila, estroma adyacente a carcinoma – verde claro, regiones mixtas – negro) y la condición clínica (Fanconi – rojo, No Fanconi – azul).

De manera análoga, en la Figura 3.4 se muestran los resultados del clustermap para las características topológicas persistentes de dimensión 1. Se destacan varios agrupamientos

señalados con cuadros de color:

- **Cuadros naranjas:** similar a lo observado en H_0 , se aprecia un agrupamiento de condiciones histológicas de carcinoma y displasia, reflejando consistencia en la organización de ciclos unidimensionales entre estas muestras.
- **Cuadro negro:** se observan zonas de carcinoma–estroma agrupadas junto con muestras de estroma adyacente al carcinoma y otras relacionadas con carcinoma. La variación en la agrupación se debe a la proporción de células tumorales presentes, la cual está influida por la proximidad de la región con la zona tumoral.
- **Cuadro amarillo:** se agrupan zonas de estroma adyacente con muestras de displasia de bajo grado, indicando que las configuraciones locales de ciclos presentan similitudes topológicas entre estas regiones.

En conjunto, estos resultados muestran que, aunque las clases de homología persistente de dimensión 1 reflejan principalmente la topología local de los ciclos, es posible identificar agrupamientos coherentes con la organización histológica de las muestras.

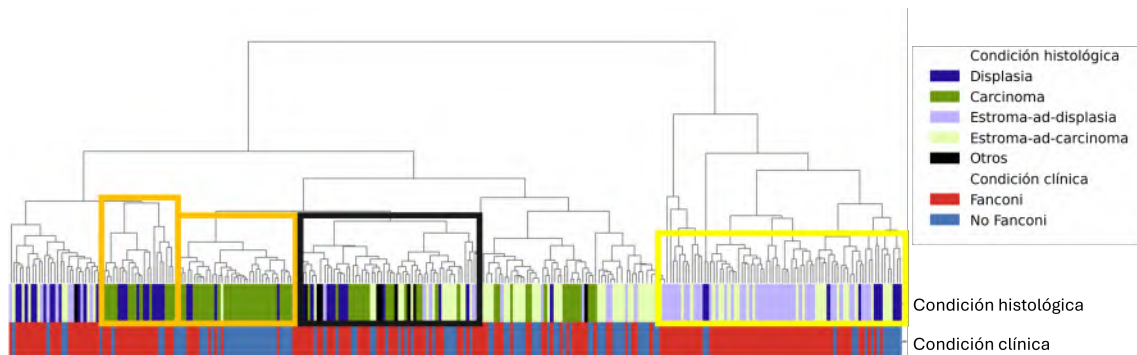


Figura 3.4: Resultados del análisis de clustering jerárquico basado en las distancias de Wasserstein para la homología en dimensión 1. La figura muestra la estructura jerárquica de similitud topológica entre las muestras, junto con las barras de color que representan la condición histológica y la condición clínica.

Los resultados obtenidos a partir de los análisis de clustering jerárquico basados en las distancias de Wasserstein evidencian que la información topológica capturada por las clases persistentes de dimensión 0 y 1 permite diferenciar de forma consistente los tejidos tumorales (carcinoma y displasia) de los estromales, reflejando variaciones estructurales relevantes en su organización celular. No obstante, al considerar la condición clínica de los pacientes (FA vs. no FA), no se observaron agrupamientos claros que indiquen una separación topológica consistente entre ambas categorías, como se aprecia en la barra de estado

“Fanconi” de la Figura 3.4. Esto sugiere que, con base en las características topológicas analizadas, la condición clínica (FA o no FA) no se refleja en diferencias detectables en la organización global del tejido. Sin embargo, estas propiedades podrían variar al considerar niveles más específicos de organización. Por ello, en la siguiente sección se presenta un análisis por grupos funcionales, con el objetivo de evaluar si las características topológicas permiten distinguir entre distintos tipos de tejido y condiciones biológicas.

3.2. Análisis por grupos celulares

Los análisis previos mostraron que, aun sin distinguir a las células según su funcionalidad, el clustering jerárquico basado en las distancias de los diagramas de persistencia tiende a agrupar correctamente las muestras tumorales entre sí y las no tumorales entre sí. Sin embargo, diferenciar a las células por su funcionalidad, podría mejorar el agrupamiento, ya que es probable que dada su funcionalidad las células tengan distintas disposiciones dentro del tejido. Por ello, se realizó nuevamente el análisis topológico considerando grupos celulares definidos según su función biológica, tal como se describe en la Tabla 2.1. Esta agrupación permitió evaluar si las relaciones topológicas varían de acuerdo a la principales funcionalidades celulares y si dichas variaciones contribuyen a distinguir entre las distintas condiciones histológicas.

3.2.1. Procedimiento topológico por grupos

A partir de las muestras analizadas, se generaron subconjuntos de datos correspondientes a los grupos celulares definidos en la Tabla 2.1. Para cada grupo se construyó un complejo de Vietoris-Rips y se calcularon los diagramas de persistencia correspondientes a las clases topológicas de dimensión 0 y 1. En la Figura 3.5 se muestran ejemplos de los complejos generados, y en la Figura 3.6 los diagramas de persistencia correspondientes. A partir de estos diagramas se calcularon las distancias de Wasserstein y se construyeron las matrices de distancia para cada grupo celular, sobre las cuales se aplicó clustering jerárquico para evaluar la similitud topológica entre muestras.

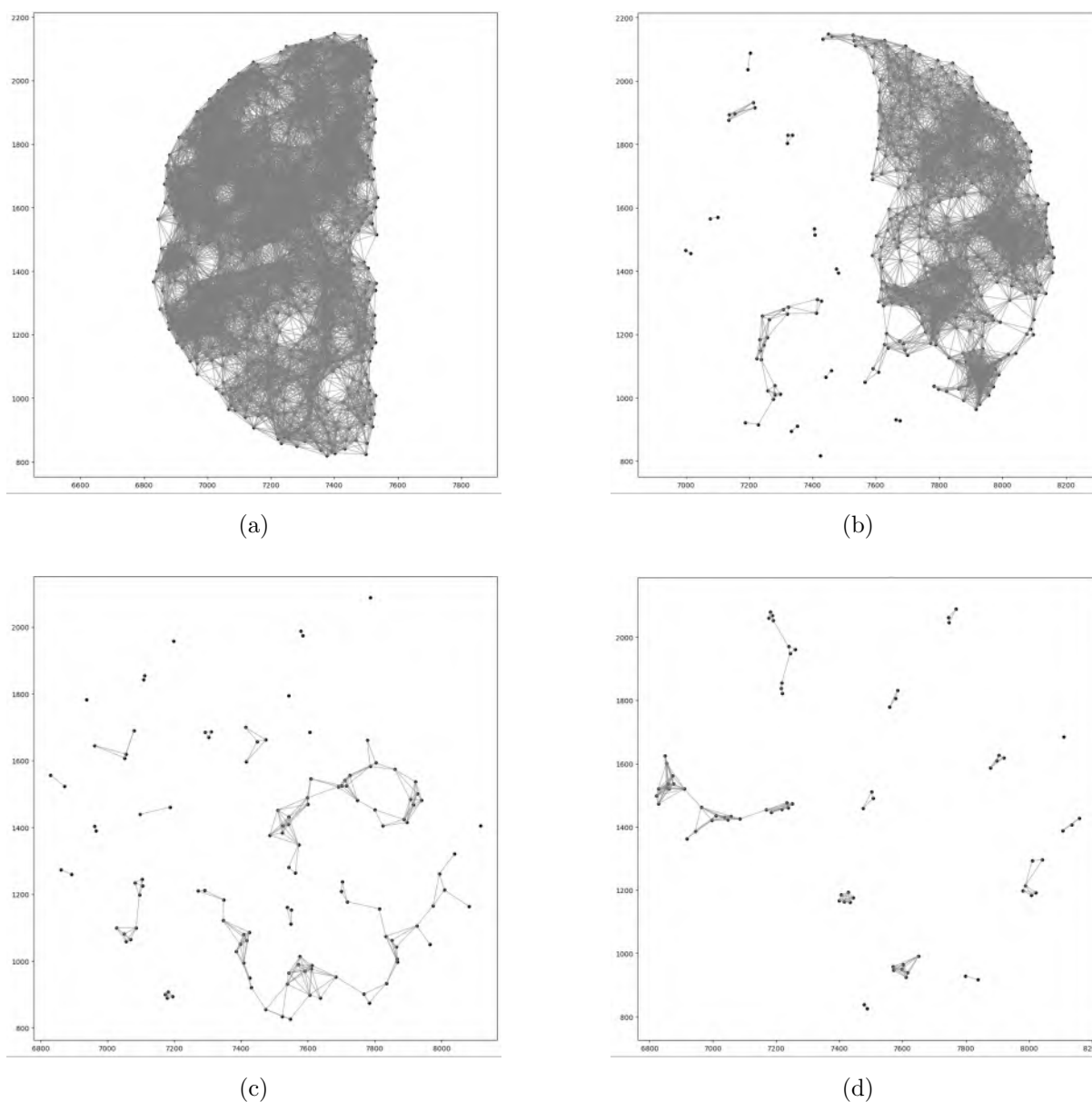
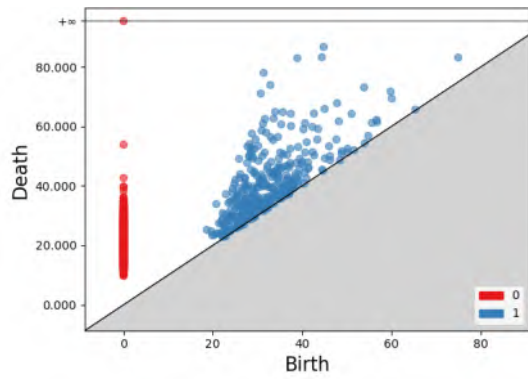
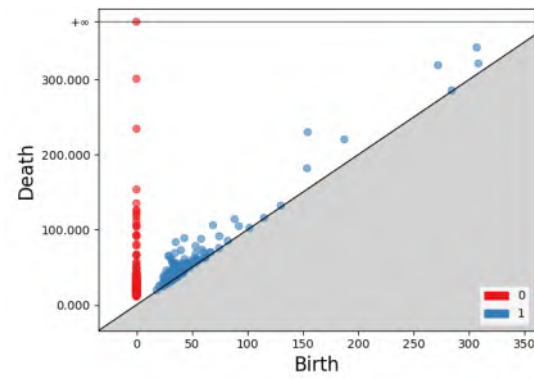


Figura 3.5: Ejemplos de complejos de Vietoris–Rips contruidos para los cuatro grupos funcionales de células en una muestra de carcinoma: **(a)** células tumorales, **(b)** células linfoides, **(c)** células mieloides, y **(d)** otros tipos celulares. Se observa que la distribución y densidad de los centroides varía entre grupos, lo que se refleja en la conectividad del complejo. Los complejos correspondientes a las células tumorales presentan conglomerados más compactos, mientras que los de células linfoides o mieloides muestran una disposición más dispersa y conexiones menos densas.

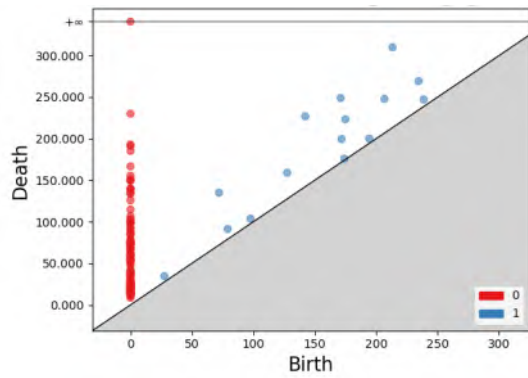
Al realizar el agrupamiento jerárquico en el grupo de células tumorales (Figura [3.7](#)), observamos que el agrupamiento generado por los clústeres mejora en comparación con



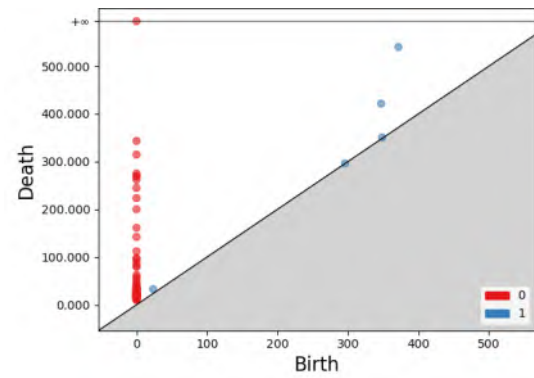
(a) Células tumorales



(b) Células linfoides



(c) Células mieloides



(d) Células no tumorales

Figura 3.6: Diagramas de persistencia generados a partir de los complejos de Vietoris-Rips para los cuatro grupos celulares en una muestra de carcinoma.

el análisis global. En este caso, se logra una separación más clara entre estromas, por un lado, y carcinomas/displasias, por otro. Este resultado corresponde a las distancias de Wasserstein calculadas sobre los diagramas de persistencia de clases topológicas de dimensión 1.

Asimismo, aunque la separación general entre las condiciones histológicas es consistente, se identifica una región del dendrograma que muestra una mezcla de muestras pertenecientes a distintas condiciones histológicas (zona resaltada en amarillo en la Figura 3.7).

De manera análoga a lo observado en el análisis global, este comportamiento se debe a que dichas muestras corresponden a carcinomas, displasias de alto grado y regiones de transición que incluyen tanto áreas de displasia como componentes estromales. A pesar de contener una fracción de estroma, la topología global de estas zonas de transición se

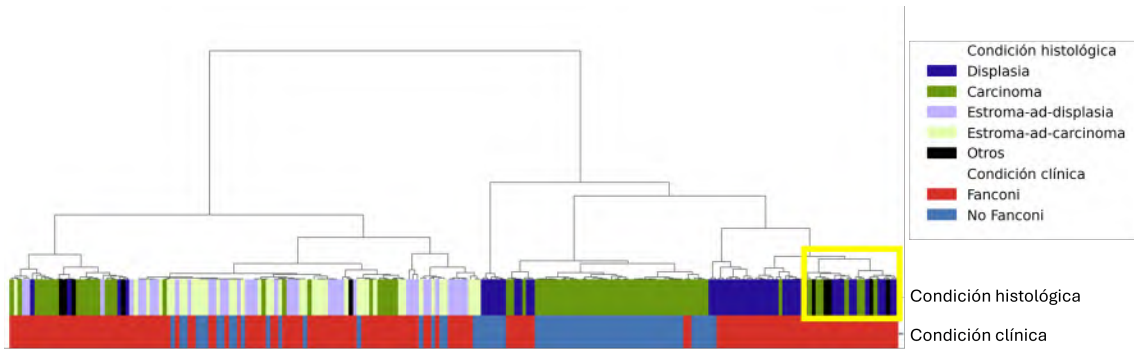


Figura 3.7: Dendrograma jerárquico basado en las distancias de Wasserstein correspondientes a la homología de dimensión H_0 para el grupo de células tumorales. La longitud de las ramas refleja la distancia topológica entre muestras. Las barras de color inferiores indican la **condición histológica** y la **condición clínica**. El cuadro amarillo resalta una región del dendrograma donde se observa una mezcla de muestras pertenecientes a distintas condiciones histológicas relacionadas con carcinomas, displasias de alto grado y zonas de displasia o carcinoma con estroma.

asemeja más a la de los tejidos con condición histológica de carcinoma y displasia, como se observa en las regiones (a) y (b) de la Figura 3.8. Sin embargo, pueden existir casos en los que la fracción estromal sea mayor, mostrando una organización más similar a la del estroma adyacente, como se aprecia en la región (c) de la misma figura. Esto sugiere que la organización espacial de las células en estas regiones mixtas conserva características topológicas propias del tejido tumoral, incluso cuando coexisten con componentes estromales.

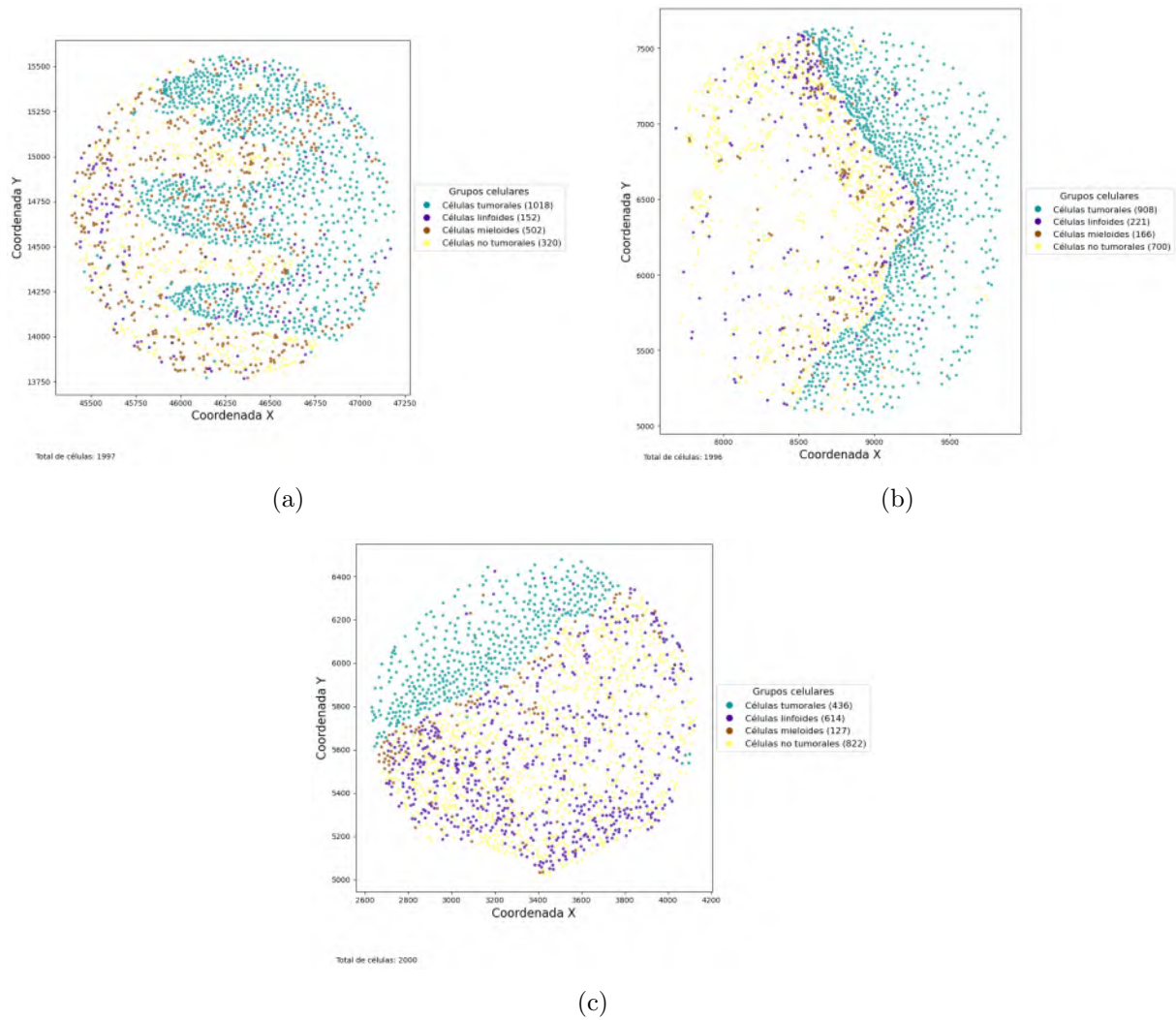
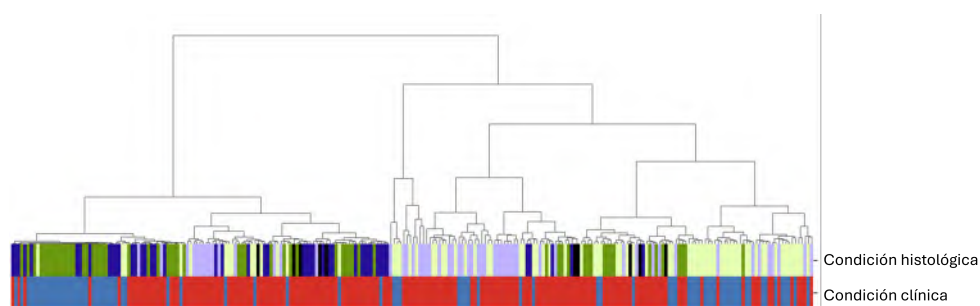


Figura 3.8: Representación espacial de cuatro regiones mixtas de displasia–estroma con diferentes proporciones de células tumorales. En (a) y (b) se observa una mayor presencia de células tumorales, mientras que en (c) la proporción de células tumorales es menor, manteniéndose la presencia de los otros tres grupos celulares.

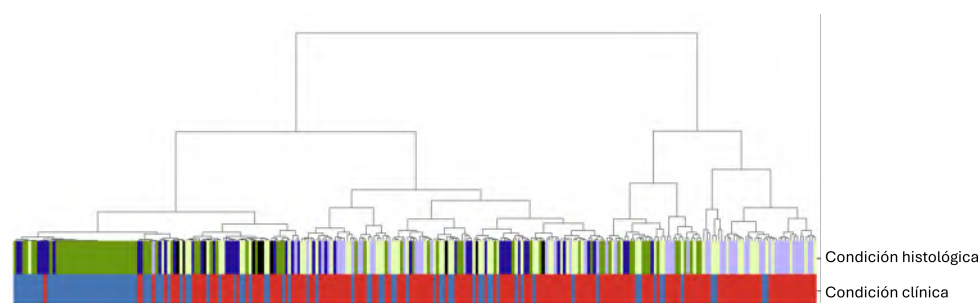
Por otro lado, al incorporar la condición clínica (FA vs. no FA), no se observa una separación clara entre ambas categorías dentro del grupo de células tumorales, lo que indica que las diferencias topológicas detectadas en esta escala se asocian principalmente con la condición histológica del tejido y no con la presencia de anemia de Fanconi (ver Figura 3.7). Este patrón se mantiene al considerar las distancias de Wasserstein calculadas sobre los diagramas de persistencia de clases topológicas de dimensión 0, donde los agrupamientos siguen la misma tendencia general observada en el análisis jerárquico.

En la Figura 3.9, se presentan los dendogramas correspondientes a los grupos de células **linfoides**, **mieloides** y **no tumorales**, contruidos a partir de las distancias de Wasserstein para la homología de dimensión 1. En comparación con las células tumorales, la agrupación por condición histológica en estos casos resulta menos definida; sin embargo, se observa que las muestras tienden a preservar ciertos patrones de similitud topológica, consistentes con las tendencias descritas tanto en el análisis de células tumorales como en el análisis global. Esto sugiere que, aunque las diferencias más pronunciadas emergen en la organización espacial de las células tumorales, los demás tipos celulares también reflejan, en menor medida, la estructura jerárquica y las relaciones espaciales características de cada tipo de tejido.

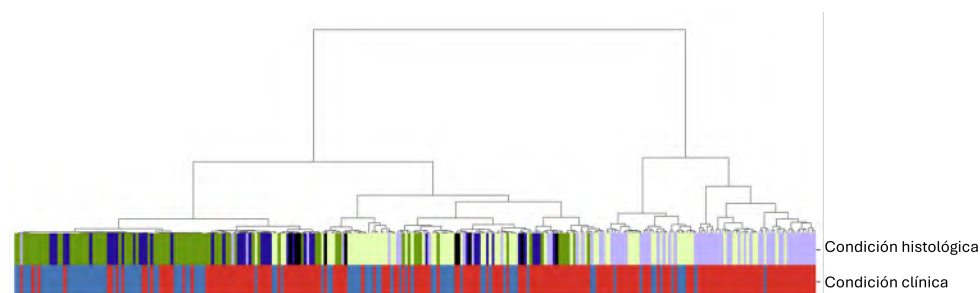
En conjunto, los resultados, tanto del análisis global como del análisis por grupos, sugieren que la estructura topológica captura de manera efectiva las diferencias espaciales entre condiciones histológicas, pero no permite distinguir con claridad entre las condiciones clínicas (AF vs. no AF). Con el fin de profundizar en este aspecto y evaluar si la topología de las interacciones celulares puede revelar diferencias asociadas a la condición genética, se empleó el método TopKAT. Esta herramienta permite incorporar información clínica durante el análisis topológico, facilitando la identificación de patrones estructurales relacionados con la condición de interés.



(a) Células linfoides



(b) Células mieloides



(c) Otro tipo de células

Figura 3.9: Dendogramas obtenidos a partir de las distancias de Wasserstein para la homología de dimensión 1 en distintos grupos celulares. Aunque la agrupación por condición histológica es menos definida que en el caso de las células tumorales, se conservan ciertas características topológicas comunes que reflejan las tendencias observadas en los análisis previos.

3.3. Análisis complementario con TopKAT

El análisis topológico global y posteriormente el realizado por grupos celulares mostraron que es posible identificar diferencias entre las zonas tumorales (carcinoma y displasia) y los estromas adyacentes. Sin embargo, esta aproximación no fue suficiente para responder otra pregunta central de este trabajo: determinar si la condición clínica de los

pacientes (anemia de Fanconi, FA, frente a no-FA) se refleja en la organización espacial celular.

Para abordar este punto se recurrió a **TopKAT** (*Topological Kernel Association Test*), una herramienta estadística específicamente diseñada para evaluar si los diagramas de persistencia presentan diferencias significativas entre grupos predefinidos [41]. El uso de **TopKAT** permitió complementar el análisis previo, ya que proporciona un marco formal para poner a prueba la hipótesis de que existen diferencias topológicas asociadas con la condición FA/no-FA. Entonces, se aplicó **TopKAT** restringiendo el análisis únicamente a las muestras de carcinoma, con el objetivo de identificar diferencias topológicas entre los carcinomas provenientes de pacientes con anemia de Fanconi (FA) y aquellos sin dicha condición (no-FA).

Los datos de entrada para este análisis fueron los mismos utilizados en la sección anterior, es decir, las coordenadas espaciales y los tipos celulares obtenidos a partir de las imágenes de tejido. A diferencia del análisis exploratorio, en este caso se incorporó una variable clínica binaria que codifica la condición de cada paciente: *FA* o *no-FA*. De esta forma, cada imagen se asoció no solo con su estructura celular, sino también con la condición clínica correspondiente, lo que permitió evaluar si los patrones topológicos de organización celular difieren significativamente entre ambos grupos de pacientes.

Paso 1: Complejo de Vietoris-Rips y diagrama de persistencia

Al igual que en las secciones anteriores, el punto de partida es la construcción de complejos de Vietoris–Rips a partir de la distribución espacial de las células en cada muestra. En la Figura 3.1 y la Figura 3.2 se muestra un ejemplo ilustrativo: una representación del complejo de Vietoris–Rips y su correspondiente diagrama de persistencia. Este esquema resume el flujo de trabajo aplicado de manera consistente en todos los conjuntos de datos y proporciona una base visual para comprender cómo la organización espacial de las células se traduce en características topológicas cuantificables.

Paso 2: Similitud topológica entre imágenes

A partir de los diagramas de persistencia se calcularon las matrices de disimilitud topológica, empleando la métrica definida en la Ecuación 2.1. Cada entrada de la matriz $D_k(Z_i, Z_j)$ cuantifica la diferencia topológica entre dos muestras i y j , para la homología de dimensión k . En este contexto, una **matriz de disimilitud** es una matriz cuadrada y simétrica donde cada elemento (i, j) representa la distancia o diferencia topológica entre dos diagramas de persistencia, de modo que valores pequeños indican alta similitud topológica y valores grandes, mayor divergencia entre las imágenes y constituyen el insumo

principal para el test estadístico TopKAT, que evalúa la asociación con la condición clínica FA/no-FA

Estas matrices se generaron mediante la función `rips-similarity-matrix`, obteniendo una matriz independiente para cada dimensión homológica.

Paso 3: Test estadístico (TopKAT)

Una vez obtenidas las matrices de disimilitud topológica para cada dimensión de homología, se aplicó el test estadístico **TopKAT** para evaluar si la organización espacial de las células se asocia con la condición clínica FA/no-FA.

Se plantearon las hipótesis estadísticas:

- H_{null} : No existe asociación entre la topología celular y la condición clínica.
- H_{alt} : Existe asociación entre la topología celular y la condición clínica.

TopKAT utiliza un enfoque tipo *kernel*, que consiste en representar las relaciones entre imágenes mediante matrices de similitud y combinar estas matrices linealmente con distintos pesos ($\omega = 0, 0.5, 1$). De esta manera, se evalúa simultáneamente la contribución de diferentes dimensiones de homología a la asociación con la variable de interés. El test resultante se denomina *omnibus*, porque integra información de múltiples kernels en un solo valor p global que resume la evidencia de asociación.

El resultado del test corresponde a un **valor p global**, el cual indica si existe evidencia estadística de diferencias topológicas significativas entre pacientes FA y no-FA. En este estudio se adoptó un umbral de significancia de $p < 0.05$, siguiendo la práctica estándar. Por lo tanto, valores menores a este umbral permiten rechazar la hipótesis nula (H_{null}) y sugieren que la organización espacial de las células difiere según la condición clínica de los pacientes. En este caso, el valor obtenido fue $p = 0.0002$, lo que es suficientemente bajo para indicar una diferencia topológica significativa entre ambos grupos.

Paso 4: Análisis post-hoc (análisis descriptivos)

Una vez obtenido un valor p global mediante el test de asociación (**TopKAT**), el siguiente paso consiste en realizar un análisis *post-hoc*. En términos generales, este tipo de análisis se aplica tras un test global para identificar qué aspectos específicos del fenómeno contribuyen a la asociación detectada. En este contexto, el objetivo es determinar en qué **escala de distancia** —definida por el parámetro ε utilizado en la construcción del complejo de Vietoris–Rips— la relación entre la topología celular y el *outcome* clínico es más fuerte. La **escala de distancia** representa el radio máximo de conexión entre células, es decir, la resolución espacial a la que emergen las estructuras topológicas. Por su parte, el *outcome* corresponde a la condición clínica de los pacientes, codificada como 1 para Anemia de Fanconi (FA) y 0 para no FA. Este análisis se lleva a cabo mediante la función `scale_importance`, que estima la contribución de cada escala de filtración a la señal estadística global.

El procedimiento puede describirse de la siguiente manera:

1. A partir de los diagramas de persistencia obtenidos en pasos previos, se consideró un conjunto discreto de 50 valores de la escala de filtración ε (radios en los complejos de Vietoris–Rips), uniformemente espaciados en el intervalo $[0, \varepsilon_{\text{máx}}]$, donde $\varepsilon_{\text{máx}}$ corresponde a la cota máxima utilizada en la construcción del complejo. Ver Figura [3.10](#).
2. Para cada valor de ε , los diagramas de persistencia se truncaron eliminando las clases topológicas con nacimiento o muerte mayores que dicha escala, y posteriormente se calculó un estadístico global de asociación mediante TopKAT, análogo a un valor p , que cuantifica las diferencias topológicas entre pacientes con anemia de Fanconi (FA) y sin anemia de Fanconi (no-FA).
3. Los resultados se resumieron en una curva que muestra el valor del estadístico de asociación en función de ε , lo que permite identificar la escala en la que dicho valor alcanza su mínimo, correspondiente a la escala espacial donde las diferencias topológicas entre ambos grupos son más marcadas.

En este caso, el análisis reveló que la **escala óptima de asociación** se encontró en `res_scale` = 30.61224, lo cual sugiere que, a dicha resolución espacial, la organización de las células tumorales refleja con mayor claridad las diferencias entre ambas condiciones clínicas. La Figura [3.11](#) muestra el resultado de este análisis, destacando el punto en el que el valor p alcanza su mínimo al considerar exclusivamente las muestras de carcinoma.

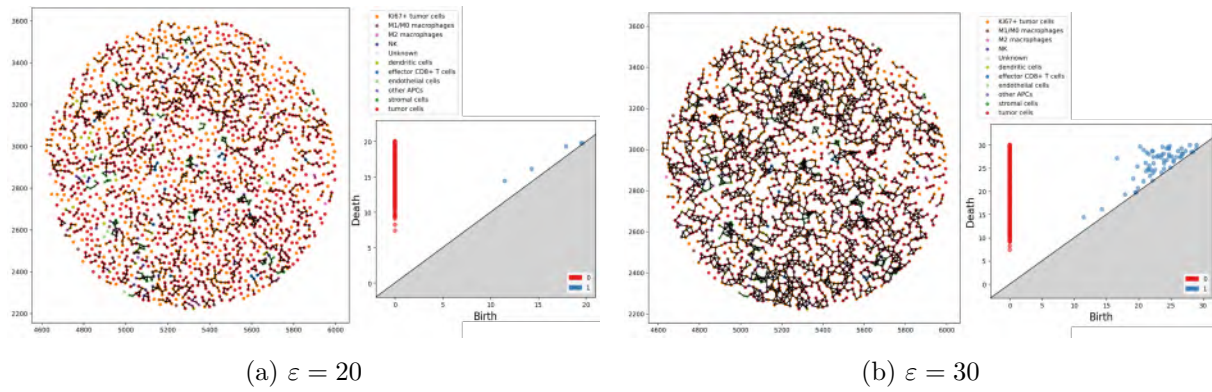


Figura 3.10: Diagramas de persistencia y sus complejos de Vietoris–Rips para distintos valores ε . Estos ejemplos ilustran cómo las propiedades cambian a medida que aumenta el radio, lo cual permite identificar las escalas en las que emergen diferencias entre pacientes FA y no-FA.

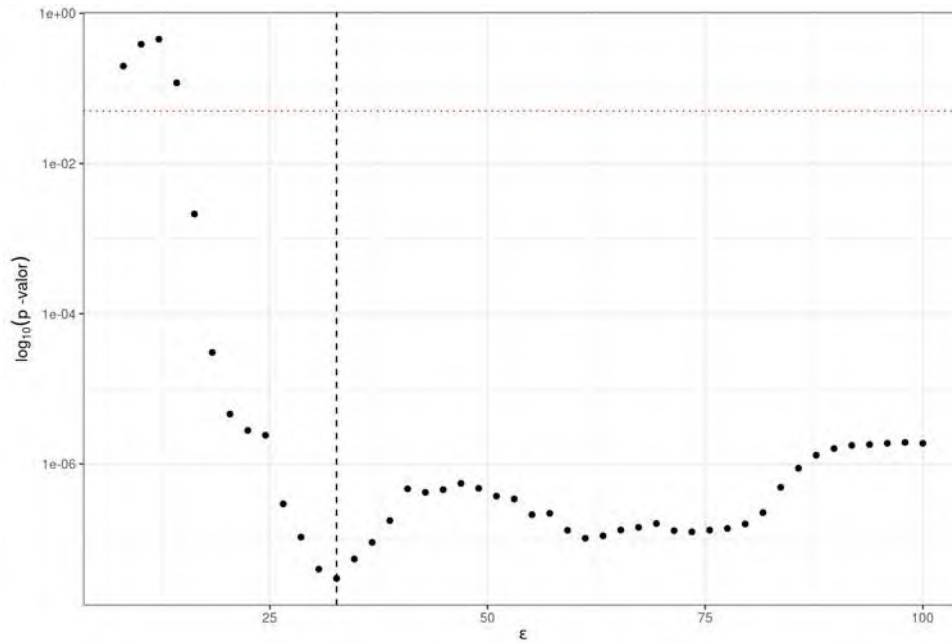
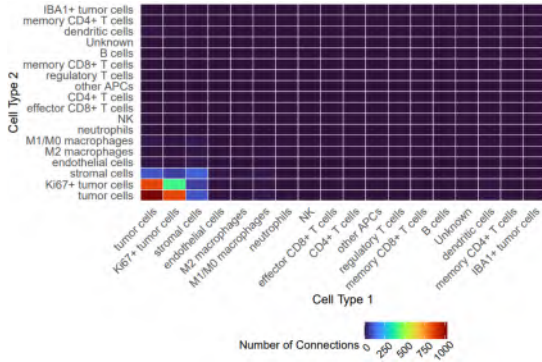


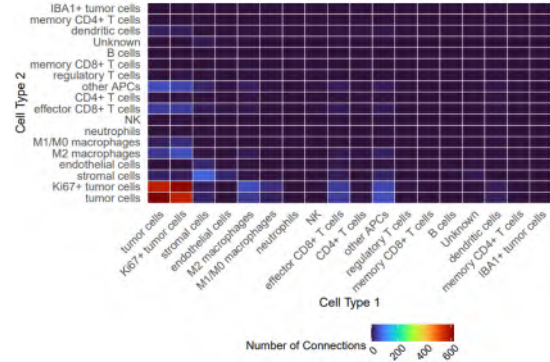
Figura 3.11: Resultado del análisis iterativo, en el que se identifica la escala óptima de asociación topológica ($\text{res_scale} = 30.61224$), correspondiente al mínimo del estadístico de asociación entre las condiciones (FA y no-FA). El eje vertical se muestra en escala logarítmica y representa el valor del estadístico de asociación (análogo a un valor p), con el fin de mejorar la visualización. La línea horizontal indica el umbral de significancia estadística ($p = 0.05$), permitiendo identificar las escalas con asociación significativa, mientras que la línea vertical negra marca la escala de filtración asociada al valor mínimo del estadístico.

A partir de la escala óptima definida, se construyeron matrices de conectividad para las muestras de carcinoma, las cuales resumen las interacciones más frecuentes entre tipos celulares dentro del complejo simplicial.

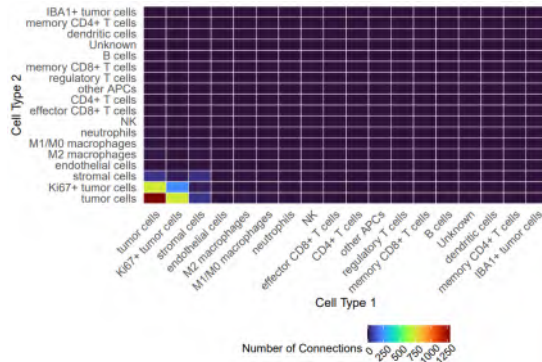
En el caso de las muestras no-FA, se observa que las conexiones predominan entre células tumorales, lo que sugiere una organización más homogénea del tejido neoplásico. En contraste, en las muestras FA comienzan a aparecer interacciones relevantes entre células no tumorales, lo que podría indicar una mayor heterogeneidad en la microarquitectura tumoral. En la Figura 3.12 podemos observar 2 ejemplos representativos tanto de muestras FA como de muestras no-FA donde se pueden observar estas diferencias.



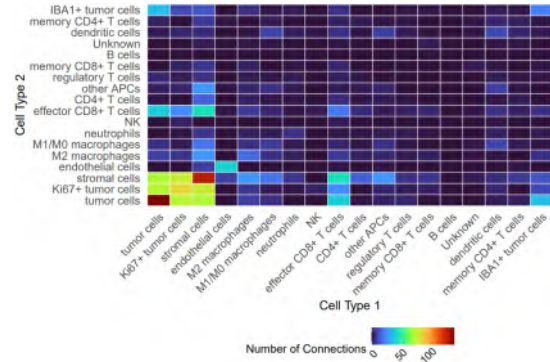
(a) Matriz de conexiones Carcinoma no-FA



(b) Matriz de conexiones Carcinoma FA



(c) Matriz de conexiones Carcinoma no-FA



(d) Matriz de conexiones Carcinoma FA

Figura 3.12: Matrices de conectividad obtenidas con `res_scale = 30.61224` en muestras de carcinoma. (a) y (c) Son ejemplos de muestras no-FA donde las conexiones predominantes ocurren entre células tumorales. (b) y (d) Son ejemplos de muestras FA en la que se observan interacciones relevantes entre células no tumorales.

De manera complementaria, se generaron matrices de conectividad promedio para cada condición clínica (FA y no-FA), promediando las conexiones obtenidas a partir de todas las

muestras pertenecientes a cada grupo. Estas matrices permiten visualizar las tendencias globales en la organización celular, eliminando las variaciones individuales entre muestras.

En la Figura 3.13 se observa que en el grupo no-FA las conexiones más fuertes se mantienen principalmente entre células tumorales, reflejando una red interna densa y cohesiva. En cambio, en el grupo FA se evidencia un aumento en la conectividad entre células no tumorales y entre tipos celulares mixtos, lo que sugiere una desorganización estructural y posibles alteraciones en la interacción tumoral-estromal asociadas a la condición clínica.

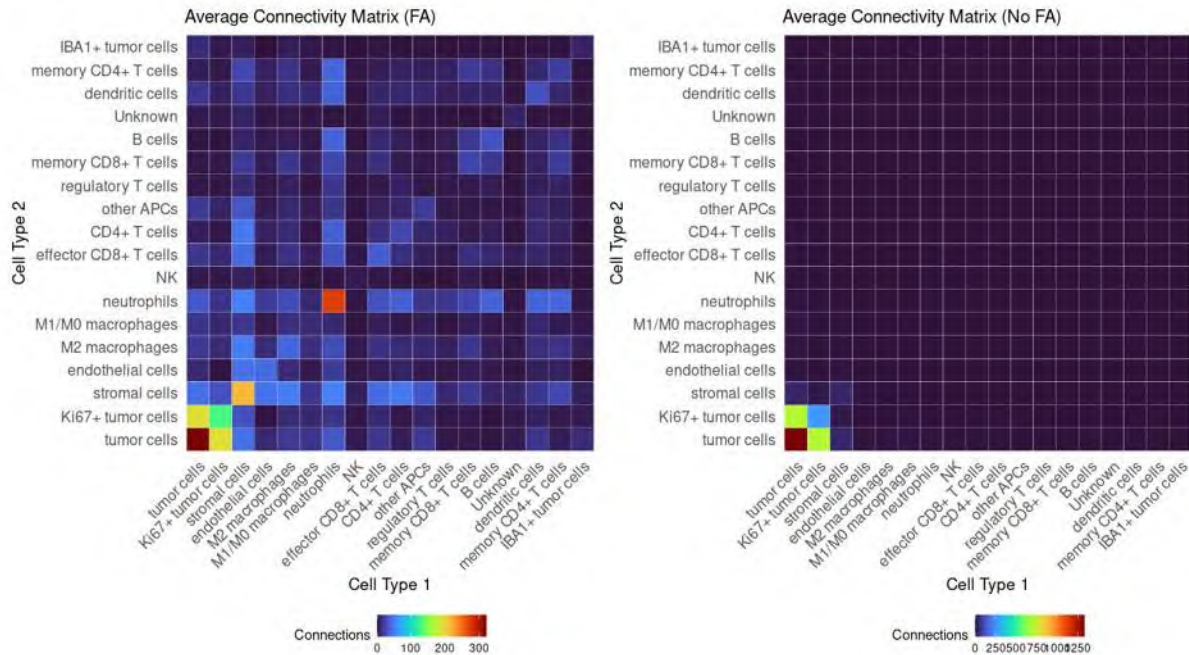


Figura 3.13: Matrices de conectividad promedio calculadas para las condiciones clínicas FA y no-FA. La matriz izquierda corresponde al grupo FA, donde las conexiones predominan entre células tumorales, indicando una organización interna más homogénea. La matriz derecha corresponde al grupo no-FA, mostrando una mayor heterogeneidad estructural con conexiones distribuidas entre distintos tipos celulares.

En conjunto, estos resultados sugieren que las características topológicas detectadas mediante el método TopKAT capturan diferencias significativas en la organización celular entre las condiciones FA y no-FA. Estas diferencias se pueden observar en la mayoría de las muestras, lo cual sugiere que la arquitectura tisular de los carcinomas asociados a FA presenta una mayor complejidad en comparación con los casos no-FA.

Capítulo 4

Discusión

Introducción

El presente estudio tuvo como objetivo evaluar el potencial del Análisis Topológico de Datos como herramienta para caracterizar la organización celular espacial de tejidos tumorales en imágenes obtenidas mediante inmunofluorescencia cíclica en tejido (*t*-CycIF). A través de la homología persistente y del uso de distancias topológicas, se buscó determinar si las propiedades topológicas de las distribuciones celulares permiten distinguir entre diferentes condiciones histológicas y clínicas.

Los resultados obtenidos muestran que los descriptores topológicos derivados de los diagramas de persistencia capturan diferencias estructurales en la disposición celular entre zonas tisulares con distintas condiciones histológicas, como carcinoma, displasia y estroma adyacente. Además, mediante el uso de TopKAT se identificaron variaciones topológicas asociadas a la condición de anemia de Fanconi, lo que sugiere que las diferencias en la arquitectura celular entre muestras FA y no-FA presentan una base topológica detectable. En consecuencia, los análisis realizados evidencian que los descriptores topológicos permiten cuantificar diferencias significativas en la organización espacial de las células dentro del tejido. Las zonas tisulares provenientes de distintas condiciones histológicas presentan firmas topológicas diferenciables, indicando que la estructura celular subyacente tiene una correlación cuantificable con el estado biológico del tejido. El uso de TopKAT permitió identificar asociaciones significativas entre las características topológicas y la presencia o ausencia de anemia de Fanconi, evidenciando que estas alteraciones genéticas o funcionales se reflejan en la arquitectura celular y pueden ser detectadas mediante enfoques topológicos. Estos hallazgos establecen un puente directo entre los métodos topológicos aplicados y la interpretación biológica de la organización del microambiente tumoral.

Los hallazgos se alinean con Bhaskar et al. (2020) y Aukerman et al. (2024), quienes

demonstraron que el TDA es capaz de extraer información espacial significativa en poblaciones celulares heterogéneas. Samorodnitsky et al. (2023) introdujeron TopKAT como un método estadístico que combina la información de homología persistente con pruebas de asociación basadas en kernels; nuestros resultados confirman que esta combinación es efectiva para detectar diferencias asociadas a condiciones clínicas, como FA vs no-FA, en contextos tisulares complejos. De manera complementaria, Keren et al. (2023) mostraron que el análisis espacial detallado de microambientes tumor-inmunes mediante técnicas multiplexadas permite identificar estructuras celulares organizadas correlacionadas con fenotipos clínicos, reforzando la relevancia de aplicar enfoques topológicos y estadísticos para capturar la heterogeneidad y las interacciones celulares. De esta forma, este trabajo fortalece la propuesta de TDA como una herramienta efectiva para estudiar tanto datos provenientes de ómicas espaciales como de tejidos complejos.

Estos resultados tienen varias implicaciones para la investigación en biología celular y medicina. Primero, demuestran que el TDA puede complementar los análisis espaciales tradicionales, ofreciendo un nuevo punto de vista de la organización multicelular. Segundo, la detección de diferencias topológicas entre muestras FA y no-FA sugiere que estas herramientas podrían ser útiles para estudios de biomarcadores espaciales o para caracterizar microambientes tumorales asociados a alteraciones genéticas. Además, desde un enfoque metodológico y matemático, este trabajo destaca el potencial del TDA como un área en desarrollo con aplicaciones amplias; la integración de herramientas topológicas con métodos estadísticos basados en kernels, como TopKAT, establece un marco robusto y reproducible que puede inspirar nuevas métricas, algoritmos y enfoques para analizar datos espaciales complejos en diversos contextos. Finalmente, estas estrategias ofrecen un camino para formalizar y cuantificar patrones estructurales de manera rigurosa, abriendo oportunidades tanto para la investigación fundamental como para aplicaciones clínicas.

Limitaciones

Si bien los resultados muestran el potencial del TDA para caracterizar la organización espacial celular, este estudio presenta algunas limitaciones que deben considerarse al interpretar los hallazgos.

Primero, la cantidad de muestras disponibles fue limitada, apenas suficiente para realizar los análisis. Con un mayor número de muestras, las firmas topológicas detectadas podrían variar, afectando la generalización y robustez de los resultados.

Segundo, los estudios de imágenes mediante TDA aún no se encuentran estandarizados, por lo que se requiere un largo y arduo trabajo antes de que esta metodología pueda ser empleada de forma extendida por la comunidad científica y médica. De hecho, todavía

existe falta de claridad en aspectos fundamentales, como la definición del rango de filtración, la elección de la métrica para la construcción de los complejos de Vietoris–Rips o los criterios de preprocesamiento y estandarización de los datos, así como en la aplicación e interpretación de métodos estadísticos como TopKAT. Estas decisiones metodológicas pueden influir de manera significativa en los resultados, ya que la homología persistente depende de la escala de filtración y de la resolución espacial de los datos. Diferentes configuraciones podrían conducir a variaciones en las distancias topológicas y en la detección de asociaciones estadísticas. Por estos motivos, será fundamental avanzar hacia una mayor estandarización del análisis de imágenes basado en TDA.

Tercero, aunque el TDA permite capturar patrones abstractos de organización, su interpretación biológica requiere cautela. Las diferencias topológicas identificadas no siempre implican causalidad o relevancia funcional directa, por lo que deben complementarse con análisis adicionales o validaciones experimentales.

Cuarto, el análisis topológico implica un costo computacional considerable. El cálculo de los complejos de Vietoris–Rips y la generación de los diagramas de persistencia requieren una cantidad significativa de memoria y tiempo de procesamiento. En este estudio, el procesamiento se realizó utilizando los servidores disponibles, cuyas características son las siguientes:

- **338-CGRJ:** Intel Xeon W5-3423 con 12 núcleos y 24 subprocesos, 30 MB de caché, frecuencia base de 2,1 GHz y hasta 4,2 GHz con turbo, con un consumo de 220 W.
- **370-AHHP:** 512 GB de memoria RDIMM DDR5 a 4800 MHz con ECC, distribuida en 8 unidades de 64 GB, expandida posteriormente a 755 GB.

El uso de cinco núcleos del servidor demandó aproximadamente 400 GB de memoria RAM por ejecución, y el procesamiento de alrededor de 100 muestras requirió cerca de 20 horas. Este elevado costo computacional se debe a la naturaleza combinatoria del complejo de Vietoris–Rips, cuya complejidad crece rápidamente con el número de puntos (células) en la muestra. Además, la variación en el número de células por imagen puede influir directamente en los resultados topológicos. Comparar, por ejemplo, imágenes con 1000 células frente a otras con 3500 puede generar diferencias en las distancias topológicas que no reflejan cambios biológicos reales, sino el tamaño desigual de los conjuntos de puntos. Por ello, es fundamental estandarizar o seleccionar subconjuntos con un número comparable de células antes del análisis, garantizando que las diferencias observadas sean atribuibles a la organización espacial y no a la cantidad de elementos analizados.

Por último, el estudio se centró en datos de *t*-CycIF de un tipo específico de tejido y condiciones clínicas, lo que puede limitar la extrapolación de los hallazgos a otros contextos o tipos de cáncer.

Recomendaciones para futuras investigaciones

Primero, sería fundamental aumentar el número y la diversidad de muestras analizadas para evaluar la robustez del estudio. De esta manera, se podría lograr un equilibrio entre las diferentes condiciones histológicas y clínicas, lo cual sería vital para corroborar los resultados obtenidos.

Segundo, se recomienda explorar otros métodos más allá del complejo de Vietoris-Rips, por ejemplo el complejo Alfa, así como el uso de imágenes de persistencia u otras transformaciones de los diagramas de persistencia. Estas alternativas podrían ofrecer perspectivas complementarias de la organización celular y capturar patrones que podrían pasar desapercibidos con un único enfoque.

Tercero, la integración de estas representaciones topológicas con técnicas de aprendizaje automático, incluyendo redes neuronales convolucionales o modelos híbridos, podría mejorar la capacidad de clasificación y predicción de condiciones histológicas y condiciones clínicas, aprovechando al máximo la información estructural contenida en las distribuciones celulares.

Finalmente, se sugiere realizar estudios complementarios que validen biológicamente las diferencias topológicas identificadas. Por ejemplo, se podría incorporar información adicional disponible, como variables de tiempo o supervivencia de las muestras para fortalecer la interpretación biológica de los resultados.

Desafíos y oportunidades en el análisis topológico

A pesar de sus ventajas teóricas y aplicaciones exitosas, la adopción generalizada del Análisis Topológico de Datos en el ámbito biomédico enfrenta ciertos desafíos técnicos y metodológicos. Uno de los principales obstáculos es la complejidad computacional asociada al cálculo de la homología persistente, especialmente cuando se trabaja con grandes volúmenes de datos o en espacios de alta dimensión. La construcción de complejos simpliciales como el de Vietoris-Rips tiene un crecimiento exponencial en el número de símlices respecto al número de puntos, lo que puede limitar su escalabilidad en conjuntos de datos masivos. A esto se suma el hecho de que actualmente no existe un procedimiento estandarizado para llevar a cabo análisis topológicos en datos biomédicos, lo que genera una alta dependencia del diseño experimental, las decisiones computacionales y la experiencia del usuario. Esta falta de estandarización representa una barrera para la reproducibilidad y comparabilidad entre estudios, y limita su adopción en contextos clínicos.

Además, aunque existen bibliotecas eficientes como **Ripser** y **GUDHI**, muchas de ellas requieren conocimientos técnicos avanzados, por ejemplo, en topología algebraica y en el manejo computacional de complejos simpliciales de alta dimensión. La correcta im-

plementación de estos métodos demanda comprender cómo se representan y procesan los complejos de Vietoris–Rips, así como la interpretación adecuada de los diagramas de persistencia y de las métricas empleadas para comparar su estructura. Estas exigencias pueden representar una barrera para usuarios no especializados y limitan, por ahora, la adopción generalizada de estas herramientas en el análisis de imágenes biológicas.

Por otro lado, una de las áreas de oportunidad más prometedoras es la integración del TDA con modelos de aprendizaje profundo. Esta combinación permite aprovechar el poder expresivo de las redes neuronales para aprender representaciones latentes que capturen patrones topológicos relevantes. Por ejemplo, se han desarrollado arquitecturas que incorporan información de persistencia como capas dentro de redes convolucionales, o que utilizan diagramas de persistencia como entradas para modelos híbridos (Hofer et al., 2019). Estas estrategias abren la puerta a aplicaciones más precisas y robustas en clasificación, segmentación o predicción clínica basada en imágenes.

Finalmente, se reconoce la necesidad de diseñar pipelines automatizados que integren TDA con procesos estándar de análisis de imágenes médicas, desde la adquisición y preprocesamiento de datos hasta la interpretación clínica. En este sentido, el presente trabajo busca contribuir con una aplicación práctica del TDA como herramienta de extracción de características espaciales para la clasificación de patrones multicelulares, evaluando su desempeño en datos reales.

Conclusión

En conclusión, los resultados de este estudio sugieren que el Análisis Topológico de Datos, en combinación con métodos estadísticos basados en kernels como TopKAT, puede capturar diferencias significativas en la organización espacial de poblaciones celulares y ayudar a discriminar entre condiciones histológicas y clínicas relevantes. El TDA ofrece una perspectiva complementaria a los análisis espaciales tradicionales, proporcionando una representación abstracta y cuantitativa de la arquitectura celular que puede integrarse con información biológica y clínica.

En conjunto, este trabajo indica que la homología persistente y las herramientas de asociación topológica constituyen un marco metodológico sólido y con gran potencial para el estudio de tejidos complejos. Si bien su aplicación aún enfrenta desafíos relacionados con la estandarización y la reproducibilidad, los resultados obtenidos respaldan su valor como enfoque emergente para el análisis cuantitativo del microambiente tumoral y sientan las bases para desarrollos futuros que consoliden su uso en investigación biomédica.

Estos hallazgos responden directamente a las preguntas de investigación planteadas:

- **¿Puede el Análisis Topológico de Datos utilizarse para clasificar patrones**

celulares a partir de su organización espacial?

Los resultados sugieren que los descriptores topológicos derivados de los diagramas de persistencia podrían permitir diferenciar patrones celulares y caracterizar su organización dentro del tejido.

- **¿Permite el TDA identificar diferencias topológicas que reflejen propiedades biológicas y clínicas relevantes, como la condición histológica o clínica de los tejidos?**

Los análisis indican que las diferencias topológicas capturadas por el TDA y evaluadas mediante TopKAT podrían reflejar variaciones asociadas a la condición clínica, como la presencia o ausencia de anemia de Fanconi.

Bibliografía

- [1] B. Altintas, N. Giri, L. J. McReynolds, A. Best, and B. P. Alter. Genotype-phenotype and outcome associations in patients with fanconi anemia: the national cancer institute cohort. *Haematologica*, 108(1):69–80, 2022. doi: 10.3324/haematol.2021.279462.
- [2] B. Arneth. Tumor microenvironment. *Medicina*, 56(1):15, 2019. doi: 10.3390/medicina56010015.
- [3] A. Aukerman, M. Carrière, C. Chen, K. Gardner, R. Rabadán, and R. Vanguri. Persistent homology based characterization of the breast cancer immune microenvironment: A feasibility study. In *Proceedings of the 36th International Symposium on Computational Geometry (SoCG 2020)*, pages 11:1–11:20. Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2020. doi: 10.4230/LIPIcs.SoCG.2020.11. URL <https://doi.org/10.4230/LIPIcs.SoCG.2020.11>.
- [4] S. T. Barry, D. I. Gabrilovich, O. J. Sansom, A. D. Campbell, and J. P. Morton. Therapeutic targeting of tumour myeloid cells. *Nature Reviews Cancer*, 23(4):216–237, 2023. doi: 10.1038/s41568-023-00557-4.
- [5] D. Bhaskar, W. Y. Zhang, A. Volkening, B. Sandstede, and I. Y. Wong. Topological data analysis of spatial patterning in heterogeneous cell populations: clustering and sorting with varying cell-cell adhesion. *Nature Communications*, 2021. URL <https://www.nature.com/articles/s41467-021-24576-0>.
- [6] V. Brancato, C. Cavaliere, N. Garbino, F. Isgrò, M. Salvatore, and M. Aiello. The relationship between radiomics and pathomics in glioblastoma patients: Preliminary results from a cross-scale association study. *Frontiers in Oncology*, 12:1005805, 2022. doi: 10.3389/fonc.2022.1005805. URL <https://www.frontiersin.org/articles/10.3389/fonc.2022.1005805/full>.
- [7] M. Carbone, S. T. Arron, B. Beutler, A. Bononi, W. K. Cavenee, J. E. Cleaver, others, and H. Yang. Tumour predisposition and cancer syndromes as models to

- study gene–environment interactions. *Nature Reviews Cancer*, 20(9):533–549, 2020. doi: 10.1038/s41568-020-0282-5.
- [8] G. Carlsson. Topology and data. *Bulletin of The American Mathematical Society - BULL AMER MATH SOC*, 46:255–308, 04 2009. doi: 10.1090/S0273-0979-09-01249-X.
- [9] A. E. Carpenter, T. R. Jones, M. R. Lamprecht, C. Clarke, I. H. Kang, O. Friman, D. A. Guertin, J. H. Chang, R. A. Lindquist, J. Moffat, P. Golland, and D. M. Sabatini. Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology*, 7(10):R100, 2006. doi: 10.1186/gb-2006-7-10-r100. URL <https://genomebiology.biomedcentral.com/articles/10.1186/gb-2006-7-10-r100>.
- [10] F. Chazal and B. Michel. An introduction to topological data analysis: Fundamental and practical aspects for data scientists. *Frontiers in Artificial Intelligence*, 4: 667963, 2021. doi: 10.3389/frai.2021.667963. URL <https://www.frontiersin.org/articles/10.3389/frai.2021.667963/full>.
- [11] F. Chazal, V. De Silva, M. Glisse, and S. Oudot. The structure and stability of persistence modules. *SpringerBriefs in Mathematics*, 2016.
- [12] D. Cohen-Steiner, H. Edelsbrunner, and J. Harer. Stability of persistence diagrams. *Discrete & Computational Geometry*, 37(1):103–120, 2007. doi: 10.1007/s00454-006-1276-5.
- [13] D. Cohen-Steiner, H. Edelsbrunner, J. Harer, and Y. Mileyko. Lipschitz functions have l_p -stable persistence. *Foundations of Computational Mathematics*, 10:127–139, 2010.
- [14] H. Edelsbrunner and J. Harer. *Computational Topology: An Introduction*. American Mathematical Society, 2010.
- [15] L. Figueroa. tda-histopathology-classifier, 2025. URL <https://github.com/Luis0815/tda-histopathology-classifier>. [En línea; consulta: 22 octubre 2025].
- [16] R. Ghrist. Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society*, 45(1):61–75, 2008.
- [17] R. J. Gillies, P. E. Kinahan, and H. Hricak. Radiomics: Images are more than pictures, they are data. *Radiology*, 278(2):563–577, 2016. doi: 10.1148/radiol.2015151169.

- [18] D. Goldenberg, K. McLaughlin, M. Serafin, J. Chen, R. Murphy, and L. Keren. Ashlar: Efficient stitching and registration of highly multiplexed whole-slide images. *bioRxiv*, 2021. doi: 10.1101/2021.06.04.447059.
- [19] Y. Guo, Y. Li, J. Song, G. Zhu, S. Gao, and W. Wan. The promise and challenge of spatial omics in dissecting tumour microenvironment and the role of ai. *Cancer Letters*, 558:216077, 2023. doi: 10.1016/j.canlet.2023.216077.
- [20] R. Gupta, T. Kurc, A. Sharma, J. S. Almeida, and J. Saltz. The emergence of pathomics. *Current Pathobiology Reports*, 7(3):73–84, 2019. doi: 10.1007/s40139-019-00200-x. URL <https://link.springer.com/article/10.1007/s40139-019-00200-x>.
- [21] A. Hatcher. *Algebraic Topology*. Cambridge University Press, 2002. See Section 4G, The Nerve Theorem.
- [22] A. Heindl, S. Nawaz, Y. Yuan, and Y. Yuan. Mapping spatial heterogeneity in the tumor microenvironment: a new era for digital pathology. *Lab Investigation*, 95: 377–384, 2015. doi: 10.1038/labinvest.2014.157. URL <https://www.nature.com/articles/labinvest2014157>.
- [23] P. Lambin, R. T. H. Leijenaar, T. M. Deist, J. Peerlings, E. E. C. de Jong, J. van Timmeren, S. Sanduleanu, R. T. H. M. Larue, A. J. G. Even, A. Jochems, Y. van Wijk, H. Woodruff, J. van Soest, T. Lustberg, E. Roelofs, W. van Elmpt, A. Dekker, F. M. Mottaghy, J. E. Wildberger, and S. Walsh. Radiomics: the bridge between medical imaging and personalized medicine. *Nature Reviews Clinical Oncology*, 14 (12):749–762, 2017. doi: 10.1038/nrclinonc.2017.141.
- [24] L. Larsson, J. Frisén, and J. Lundeberg. Enablers and challenges of spatial omics, a melting pot of technologies. *Molecular Systems Biology*, 18(10):e10571, 2022. doi: 10.15252/msb.202110571.
- [25] R. H. Lee, H. Kang, S. S. Yom, A. Smogorzewska, D. E. Johnson, and J. R. Grandis. Treatment of fanconi anemia-associated head and neck cancer: opportunities to improve outcomes. *Clinical Cancer Research*, 27(19):5168–5187, 2021. doi: 10.1158/1078-0432.CCR-21-0302.
- [26] R. Levenson, Y. Singh, B. Rieck, Q. Hathaway, C. Farrelly, J. Rozenblit, P. Prasanna, B. Erickson, A. Choudhary, G. Carlsson, and D. Sarkar. Advancing precision medicine: Algebraic topology and differential geometry in radiology and

- computational pathology. *Laboratory Investigation*, 104(6):102060, June 2024. doi: 10.1016/j.labinv.2024.102060.
- [27] Y. Liu et al. Limitations of conventional image analysis in high-density tumor regions. *Medical Image Analysis*, 55:136–147, 2019. doi: 10.1016/j.media.2019.02.004. URL <https://doi.org/10.1016/j.media.2019.02.004>.
- [28] M. Lyu and A. Diefenbach. Tissue niches and immunopathology through the lens of spatial tissue profiling techniques. *European Journal of Immunology*, 53(11): e2350484, 2023. doi: 10.1002/eji.202350484.
- [29] C. Maria and T. G. P. Team. Gudhi user and reference manual. *GUDHI Editorial Board*, 2021. <https://gudhi.inria.fr>.
- [30] Y. Mileyko, S. Mukherjee, and J. Harer. Probability measures on the space of persistence diagrams. *Inverse Problems*, 27(12):124007, 2011.
- [31] N. Otter, M. A. Porter, U. Tillmann, P. Grindrod, and H. A. Harrington. A roadmap for the computation of persistent homology. *EPJ Data Science*, 6(1): 17, 2017. doi: 10.1140/epjds/s13688-017-0109-5. URL <https://doi.org/10.1140/epjds/s13688-017-0109-5>.
- [32] S. Y. Oudot. *Persistence Theory: From Quiver Representations to Data Analysis*, volume 209 of *Mathematical Surveys and Monographs*. American Mathematical Society, 2015.
- [33] J. Park, J. Kim, T. Lewy, C. M. Rice, O. Elemento, A. F. Rendeiro, and C. E. Mason. Spatial omics technologies at multimodal and single cell/subcellular level. *Genome Biology*, 23(1):256, 2022. doi: 10.1186/s13059-022-02811-w.
- [34] T. Peng, K. Thorn, T. Schroeder, L. Wang, F. J. Theis, C. Marr, and N. Navab. Basic: Background and shading correction for optical microscopy. *Nature Methods*, 14:925–926, 2017.
- [35] F. Pérez et al. Scimap: Single-cell analysis for multiplex imaging. *Bioinformatics*, 35 (22):4717–4719, 2019.
- [36] R. Rabadan and A. J. Blumberg. *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, 2019.

- [37] B. Rieck. Persistent homology for multivariate data visualization. In *Proceedings of the 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 1–10, 2016. doi: 10.1109/DSAA.2016.33. URL <https://doi.org/10.1109/DSAA.2016.33>.
- [38] A. Rodríguez and A. D’Andrea. Fanconi anemia pathway. *Current Biology*, 27(18): R986–R988, 2017. doi: 10.1016/j.cub.2017.07.043.
- [39] P. S. Rosenberg, B. P. Alter, W. Ebell, and L. C. Brody. Cancer risks in Fanconi Anemia: Findings from the german fanconi anemia registry. *Haematologica*, 93(4): 511–517, 2008. doi: 10.3324/haematol.11111.
- [40] S. Samorodnitsky. *TopKAT: Topological Kernel Association Test (TopKAT)*, 2025. URL <https://sarahsamorodnitsky.github.io/TopKAT/>. R package version 0.0.0.9000.
- [41] S. Samorodnitsky et al. Detecting clinically relevant topological structures in multiplexed spatial proteomics imaging using topkat. *bioRxiv*, 2024. doi: 10.1101/2024.12.18.628976v1.
- [42] U. Schmidt, M. Weigert, C. Broaddus, and G. Myers. Cell detection with star-convex polygons. *International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI)*, pages 265–273, 2018.
- [43] P. Siliceo Portugal. Análisis fenotípico y espacial de carcinomas mediante t-cycif, 2021. Disponible en: <https://ru.dgb.unam.mx/items/75125315-31cb-4a43-b867-e2c8f257ac77>.
- [44] Y. Singh, C. M. Farrelly, Q. A. Hathaway, T. Leiner, J. Jagtap, G. E. Carlsson, and B. J. Erickson. Topological data analysis in medical imaging: current state of the art. *Insights into Imaging*, 14(1):58, 2023. doi: 10.1186/s13244-023-01413-w.
- [45] S. H. Støy, G. K. Sandve, and O. Myklebost. Spatial transcriptomics: Technical aspects of recent developments and their applications in neuroscience and cancer research. *International Journal of Molecular Sciences*, 24(10):8780, 2023. doi: 10.3390/ijms24108780.
- [46] C. Yan, K. Nakane, X. Wang, Y. Fu, H. Lu, X. Fan, M. D. Feldman, A. Madabhushi, and J. Xu. Automated gleason grading on prostate biopsy slides by statistical representations of homology profile. *Computer Methods and Programs in Biomedicine*, 194:105528, Oct 2020. doi: 10.1016/j.cmpb.2020.105528. URL <https://pubmed.ncbi.nlm.nih.gov/32470903/>.

-
- [47] A. Zomorodian and G. Carlsson. Computing persistent homology. *Discrete & Computational Geometry*, 33(2):249–274, 2005. doi: 10.1007/s00454-004-1146-y.

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo



A quien corresponda,

Por este medio, quien abajo firma, bajo protesta de decir verdad, declara lo siguiente:

- Que presenta para revisión de originalidad el manuscrito cuyos detalles se especifican abajo.
- Que todas las fuentes consultadas para la elaboración del manuscrito están debidamente identificadas dentro del cuerpo del texto, e incluidas en la lista de referencias.
- Que, en caso de haber usado un sistema de inteligencia artificial, en cualquier etapa del desarrollo de su trabajo, lo ha especificado en la tabla que se encuentra en este documento.
- Que conoce la normativa de la Universidad Michoacana de San Nicolás de Hidalgo, en particular los Incisos IX y XII del artículo 85, y los artículos 88 y 101 del Estatuto Universitario de la UMSNH, además del transitorio tercero del Reglamento General para los Estudios de Posgrado de la UMSNH.

Datos del manuscrito que se presenta a revisión		
Programa educativo	Maestría en Ciencias Matemáticas	
Título del trabajo	Análisis Topológico de Datos para la Caracterización de Patrones Celulares en Tejidos Epiteliales	
	Nombre	Correo electrónico
Autor/es	Luis Raúl Figueroa Martínez	
Director	Dr. Eugenio Martín Azpeitia Espinosa	eazpeitia@matmor.unam.mx
Codirector	Dra. Adriana Haydeé Contreras Peruyero	haydeeperuyero@matmor.unam.mx
Coordinador del programa	Dr. Elmar Wagner	elmar.wagner@umich.mx

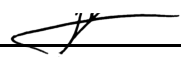
Uso de Inteligencia Artificial		
Rubro	Uso (sí/no)	Descripción
Asistencia en la redacción	no	

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo



Uso de Inteligencia Artificial		
Rubro	Uso (sí/no)	Descripción
Traducción al español	no	
Traducción a otra lengua	no	
Revisión y corrección de estilo	si	ayuda para poder mejorar el estilo de redacción en algunos parrafos
Análisis de datos	no	
Búsqueda y organización de información	si	Buscar articulos relacionados con mi trabajo de investigación
Formateo de las referencias bibliográficas	no	
Generación de contenido multimedia	no	
Otro	no	

Datos del solicitante	
Nombre y firma	Luis Raúl Figueroa Martínez 
Lugar y fecha	Morelia, Mich., 28 de octubre de 2025 

Luis Raúl Figueroa Martínez

Análisis Topológico de Datos para la Caracterización de Patrones Celulares en Tejidos Epiteliales.pdf

 Universidad Michoacana de San Nicolás de Hidalgo

Detalles del documento

Identificador de la entrega

trn:oid:::3117:519909994

Fecha de entrega

29 oct 2025, 7:48 a.m. GMT-6

Fecha de descarga

29 oct 2025, 7:59 a.m. GMT-6

Nombre del archivo

Análisis Topológico de Datos para la Caracterización de Patrones Celulares en Tejidos Epiteliales.pdf

Tamaño del archivo

33.5 MB

88 páginas

25.263 palabras

145.236 caracteres

6% Similitud general

El total combinado de todas las coincidencias, incluidas las fuentes superpuestas, para ca...

Filtrado desde el informe

- Texto citado
- Texto mencionado
- Coincidencias menores (menos de 10 palabras)

Fuentes principales

- 6%  Fuentes de Internet
- 4%  Publicaciones
- 0%  Trabajos entregados (trabajos del estudiante)

Marcas de integridad

N.º de alerta de integridad para revisión

-  **Caracteres reemplazados**
16 caracteres sospechosos en N.º de páginas
Las letras son intercambiadas por caracteres similares de otro alfabeto.

Los algoritmos de nuestro sistema analizan un documento en profundidad para buscar inconsistencias que permitirían distinguirlo de una entrega normal. Si advertimos algo extraño, lo marcamos como una alerta para que pueda revisarlo.

Una marca de alerta no es necesariamente un indicador de problemas. Sin embargo, recomendamos que preste atención y la revise.