Análisis Topológico de Datos: Homología Persistente en Imágenes Digitales de Patología

Lic. Anselmo Daniel Suarez Muñoz, Ing. Luciano Andres Juárez López
Centro de Investigación en Matemáticas. Unidad Monterrey
anselmo.suarez@cimat.mx, luciano.juarez@cimat.mx

***Resumen*—En este proyecto final se explora el uso del Análisis Topológico de Datos (TDA) y la homología persistente para analizar imágenes histológicas de melanoma en zona tumoral y zona no tumoral.**

I. Introducción

La topología es una rama fundamental de las matemáticas que estudia las propiedades de los espacios que se mantienen inalteradas bajo deformaciones continuas, tales como el estiramiento, la torsión y la contracción, pero no el desgarramiento ni el pegado. Esta disciplina se centra en entender conceptos de proximidad, continuidad y límites, sin considerar las medidas de distancia o ángulos, lo que la diferencia de la geometría. Los objetos de estudio en topología incluyen conjuntos abiertos, cerrados, compactos, conexos y muchos otros, los cuales son analizados mediante herramientas como la teoría de homotopía y la teoría de homología. La topología tiene aplicaciones en diversas áreas, desde la física teórica hasta la biología y las ciencias de la computación, ofreciendo una perspectiva única y abstracta sobre la estructura y la naturaleza de los espacios.

El análisis topológico de datos (TDA, por sus siglas en inglés) es una metodología emergente dentro del campo del análisis de datos que utiliza conceptos y herramientas de la topología para entender la estructura y las relaciones subyacentes en conjuntos de datos complejos. Esta técnica se basa en la idea de que la forma de los datos puede revelar información esencial sobre el fenómeno que se está estudiando. Uno de los métodos más conocidos dentro del TDA es la homología persistente, que permite identificar y analizar características topológicas a diferentes escalas, proporcionando una visión multi-resolutiva de los datos.

El TDA ha demostrado ser particularmente útil en áreas donde los datos son intrínsecamente de alta dimensionalidad o cuando la estructura de los datos no se puede captar adecuadamente mediante técnicas tradicionales de análisis. Aplicaciones del TDA se encuentran en campos tan diversos como la biología, donde se utiliza para el estudio de estructuras biomoleculares, la medicina, para la identificación de patrones en imágenes médicas, y en el aprendizaje automático, donde ayuda a mejorar la interpretación de modelos complejos.

La combinación de topología y análisis de datos ofrece un enfoque poderoso para desentrañar las complejidades y las sutilezas de los datos modernos, permitiendo descubrir patrones y relaciones que de otro modo permanecerían ocultos.

II. Conceptos Básicos de Homología persistente

El método más destacado es la homología persistente, que proporciona la cuantificación de la forma de los datos a través de características topológicas como componentes conexas, esto es, conjuntos de punto de datos, así como de ciclos, que están formados por puntos de datos que rodean un agujero vacío.

En esta sección vamos a introducir conceptos topológicos necesarios para poder realizar nuestro análisis. Los conceptos teóricos usados son tomados de [1]

II-A. Complejos simpliciales

El espacio topológico se puede aproximar mediante estructuras llamadas complejos simpliciales. Para aplicar la homología persistente, es necesario construir una filtración de complejos simpliciales a partir de los datos de la nube de puntos, es decir, una secuencia de estructuras similares a gráficos anidados que incluyen nodos y aristas, así como conexiones de orden superior como triángulos y tetraedros. Haremos un resumen de los conceptos básicos que se basan en [4]

Tomamos una nube de puntos de datos como coordenadas en el espacio. Esta nube de puntos se dota de una estructura matemática.

Definición 3.1 Sea $V = \{v_0, v_1, \ldots, v_k\}$ puntos en $\mathbb{R}^d$. Un punto $x \in \mathbb{R}^d$ es una combinación afín de los puntos $v_i \in V$ con $i \in \{0, \ldots, k\}$ si existe $\lambda_i \in \mathbb{R}$ tal que:

1. $x = \sum_{i=0}^{k} \lambda_i v_i$,
2. $\sum_{i=0}^{k} \lambda_i = 1$.

El conjunto de todas las combinaciones lineales de V se llama envoltura afín de V. La envoltura afín de V es el espacio afín más pequeño que contiene a V y, por tanto, el espacio afín más pequeño que contiene una nube de puntos.

Los puntos han de ser afínmente independientes, ya que los vectores formados por dichos puntos son linealmente independientes (si los vectores fuesen linealmente dependientes no podrían formar una base en el espacio).

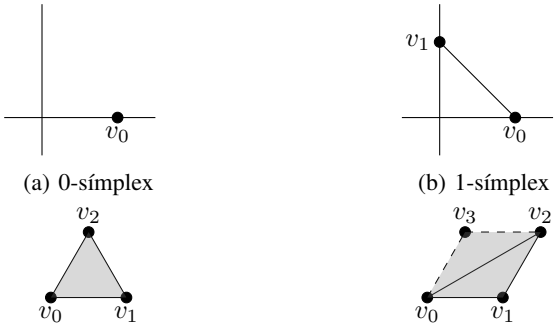
Definición 3.2 Sea $V = \{v_0, v_1, \ldots, v_k\}$ puntos en $\mathbb{R}^d$. Se dice que los $k+1$ puntos en V son afínmente independientes si los vectores $\{v_i - v_0 : i \in \{0, \ldots, k\}\}$ son linealmente independientes.

El polígono de menor área que contiene la nube de puntos en su interior se denomina envoltura convexa.

Definición 3.3 Dada una combinación afín $x = \sum_{i=0}^{k} \lambda_i v_i$, tenemos que es una combinación convexa si $\lambda_i \geq 0$ para

todo $i \in \{0, \ldots, k\}$. El conjunto de todas las combinaciones convexas de puntos en V es llamado envoltura convexa de V.

Definición 3.4 Un k-símplice es la envoltura convexa de $k+1$ puntos afínmente independientes $v \in \mathbb{R}^d$ y es denotado como: $\Delta = [v_0, v_1, \ldots, v_k]$.



(a) 0-símplex (b) 1-símplex

(c) 2-símplex (d) 3-símplex

Figura 1: Símplices. Podemos observar por la definición de n-símplex que un 0-símplex es un único punto, 1-símplex es un intervalo (dos puntos unidos por una arista), 2-símplex es una región triangular (un triángulo con su superficie) y 3-símplex un tetraedro sólido.

Los símplices de dimensiones superiores están limitados por los símplices de dimensiones inferiores. Por ejemplo, un triángulo está delimitado por tres aristas y una arista está delimitada por dos puntos (Figura 3.1).

Definición 3.5 Dado un n-símplex $[v_0, v_1, \ldots, v_n]$, cada m-símplex $[v_{i_0}, \ldots, v_{i_m}] \subseteq [v_0, v_1, \ldots, v_n]$ con $0 \leq i_k \leq n$ es una m-cara.

Un complejo simplicial es un espacio que se construye a partir de la unión de puntos, aristas, triángulos, tetraedros y politopos de dimensiones superiores, es decir, se construye "pegando" diferentes símplices a lo largo de un símplex de dimensiones inferiores.

Definición 3.6 Un complejo simplicial, K, es una colección de símplices que satisface las siguientes propiedades:

- Si un n-símplex $[v_0, v_1, \ldots, v_n] \in K$ entonces toda m-cara $[v_{i_0}, \ldots, v_{i_m}] \in K$, es decir, cada cara de un elemento de K también está en K.
- Si hay símplices que intersecan, la intersección es una m-cara de cada uno de ellos.

Podemos ver ejemplos de complejos simpliciales en la Figura

Homología y números de Betti La homología simplicial proporciona técnicas computacionales para estudiar espacios topológicos que se representan como complejos simpliciales. En homología simplicial queremos identificar agujeros en espacios topológicos; por ejemplo, podemos distinguir una esfera de un toro.

A través de la homología simplicial queremos identificar ciclos en un objeto dado y queremos saber si un ciclo dado limita una colección de símplices de dimensiones superiores

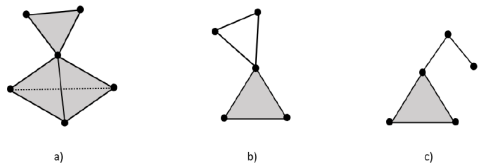


Figura 2: **Complejos simpliciales**: a) Un 3-complejo simplicial creado al conectar un 3-símplex (tetraedro) y un 2-símplex (triángulo). Este complejo simplicial tiene dimensión 3. b) Un 2-complejo simplicial compuesto de un 2-símplex y tres 1-símplex. Este complejo simplicial tiene dimensión 2. c) Un 2-complejo simplicial producido por un 2-símplex y dos 1-símplex. Este complejo simplicial tiene dimensión 2.

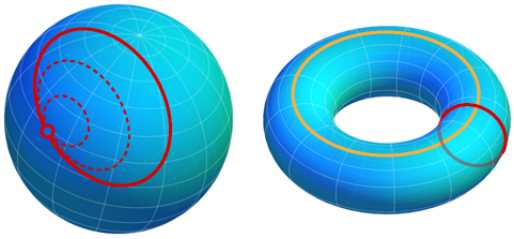
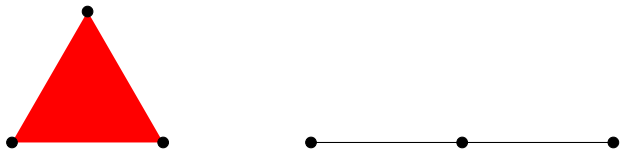


Figura 3: Esfera y toro

(si no lo hace entonces el ciclo forma un agujero). Un ejemplo de este concepto podemos verlo en la Figura.



a) b)

Figura 4: Diferencia entre dos 1-complejos simpliciales. Tanto en a) como en b) tenemmos un 1-complejo con la diferencia de que en a) el 1-s´ımplice delimita un espacio vac´ıo que constituye un agujero mientras que en b) el complejo representa una l´ınea sin agujeros.

Los siguientes conceptos nos permitir´an identificar la presencia de agujeros y ciclos.

Definición 3.7 Una k-cadena simplicial es una suma ponderada finita definida en todos los k-símplices dentro de un complejo K:

$$\sum_{i \in I} c_i \langle \sigma_i \rangle;$$

donde $c_i \in \mathbb{Z}$ y $\langle \sigma_i \rangle$ son k-símplices, I es el conjunto de índices.

Cuando trabajamos con coeficientes en $\mathbb{Z}_2$ (véase el Apéndice A), podemos observar que la suma de dos k-cadenas resulta en la suma de todos los $(k-1)$-símplices en los que difieren las $(k-1)$-cadenas originales. Los $(k-1)$-símplices que las

dos $(k-1)$-cadenas tienen en común estarán presentes en la suma dos veces y, por tanto, desaparecen por las propiedades de la adición en $\mathbb{Z}_2$. A partir de ahora, supondremos que los coeficientes de nuestras $(k-1)$-cadenas se toman en $\mathbb{Z}_2$, lo que simplificará algunas de las siguientes definiciones.

Definición 3.8 El operador borde es una aplicación de las k-cadenas de un complejo a sus bordes. Para un conjunto de k-cadenas (denotadas como C_k), definimos el operador borde ∂ como:

$$\partial : C_k \to C_{k-1}.$$

La operación de borde o frontera en un símplice general σ con vértices $[v_0, v_1, \ldots, v_k]$ se muestra a continuación, donde el vértice v_i se elimina del conjunto de vértices en la suma. La operación de borde en un k-símplice asigna el símplice a una suma de sus $(k-1)$ caras (véase la Figura 3.5):

$$\partial([v_0; v_1; \ldots; v_k]) = \sum_{i=0}^{k} (-1)^i [v_0; \ldots; \hat{v_i}; \ldots; v_k]$$

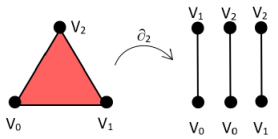


Figura 5: Operador borde: Aplicación del operador borde a un 2-símplice.

Usamos la notación abreviada $\partial_k = \partial(C_k)$ para representar una operación de frontera. Además, la k-cadena para dimensiones mayores que k y menores que 0 son cero y por tanto, ∂_{k+1} y ∂_0 son 0.

$$0 \xrightarrow{\partial_{k+1}} C_k(K) \xrightarrow{\partial_k} C_{k-1}(K) \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_1} C_0(K) \xrightarrow{\partial_0} 0.$$

Teorema 3.2 Sea $d \in C_{k+1}$, tenemos que

$$(\partial_k \circ \partial_{k+1})(d) = 0.$$

En homología simplicial, un k-ciclo se define como una k-cadena que se asigna al espacio nulo o núcleo del operador borde (denotado como $\ker(\partial_k)$) y que es cero.

Definición 3.9 Los k-ciclos dimensionales están dados por

$$Z_k = \ker(\partial_k),$$

donde $\ker(\partial_k)$ es el núcleo del operador ∂_k.

Definición 3.10 Los bordes vienen dados por

$$B_k = \mathrm{im}(\partial_{k+1}),$$

donde $\mathrm{im}(\partial_{k+1})$ es la imagen del operador ∂_{k+1}.

En definitiva, la información que abarca Z_k nos proporciona los ciclos de dimensión k dentro de un complejo dado y la información en B_k nos manifiesta si un ciclo es o no el borde de una colección de símplex de dimensiones superiores.

Además, es importante tener en cuenta que B_k es un subgrupo de Z_k, es decir, $B_k \subseteq Z_k$ ("un borde no tiene borde"), esto es consecuencia del Teorema 3.2.

Asimismo, si un ciclo no es un borde, entonces se conoce como agujero. Esta información de cualquier complejo la podemos obtener definiendo el k-grupo de homologías H_k y el número de Betti β_k. En términos simples, el grupo H_k contiene el único k-agujero dentro de un complejo y β_k cuenta el número de k-agujeros únicos en un complejo.

Definición 3.11 El k-grupo de Homología H_k viene dado por el grupo de cocientes.

Definición 3.13 El k-ésimo Número Betti β_k es el rango de H_k, y viene dado por:

$$\beta_k = \mathrm{rank}(H_k) = \mathrm{rank}(Z_k) - \mathrm{rank}(B_k).$$

Vemos que el grupo H_1 identifica los agujeros unidimensionales dentro del complejo y β_1 cuenta el número de agujeros unidimensionales en el complejo. Lo mismo es cierto para todas las demás dimensiones $k \geq 0$ siempre y cuando $\dim(K) \geq k$ ya que $H_k = 0$ para todo $k > \dim(K)$.

Los 0-ésimos grupos de Homología, H_0, juegan un papel importante en el análisis topológico. H_0 es la medida del número de componentes conectados en un complejo.

II-B. Homología Persistente

La homología persistente es una metodología que se ha desarrollado para extraer y cuantificar información topológica de datos, proporciona una visión profunda de la estructura de los datos y las capacidades de cuantificación de sus características geométricas.

El cálculo directo de la homología de un conjunto arbitrario definido por el espacio $X \subseteq \mathbb{R}^n$ es una tarea compleja, para simplificarla, usamos la homología simplicial; identificamos un complejo simplicial K tal que su homología sea la misma o similar a la de X. Podemos definir tal complejo creando un objeto geométrico conocido como la cubierta ($\mathcal{U}$) de X.

Definición 3.14 $\mathcal{U} = \{U_i\}_{i \in I}$ es una cubierta de un espacio métrico X si $X \subseteq \bigcup_{i \in I} U_i$.

Si definimos nuestro espacio métrico como un conjunto finito de puntos entonces podemos imaginar cada conjunto U como una bola centrada alrededor de cada punto. Usamos esta cubierta para desarrollar un complejo simplicial conocido como complejo de Čech.

Definición 3.15 El complejo de Čech es un complejo simplicial construido a partir de k-símplex que son la intersección no vacía de $k+1$ conjuntos de la cubierta $\mathcal{U}$.

A medida que ajustamos el radio r de la cubierta U, obtenemos diferentes complejos de Čech, cada uno con una homología diferente. Como queremos asegurarnos que la homología capture las características más interesantes de los datos, caracterizamos el conjunto de datos para múltiples valores de r. Esta información se ve reflejada en un complejo simplicial filtrado.

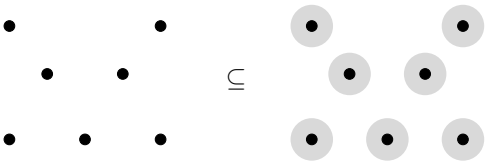


Figura 6: Cubierta de una nube de puntos. Podemos observar a la izquierda un conjunto de puntos y a la derecha una cubierta de puntos definida por un conjunto de bolas $B(x_i, r)$ expandidas alrededor de cada punto.

Definición 3.16 Un complejo simplicial filtrado $K \subseteq \mathbb{R}^m$ es un complejo simplicial para el que hay una serie de subcomplejos simpliciales anidados $K_r \subseteq \mathbb{R}^m$ tal que:

$$K_{r_0} \subseteq K_{r_1} \subseteq \cdots \subseteq K_{r_n}$$

donde $r_0 < r_1 < \cdots < r_n$.

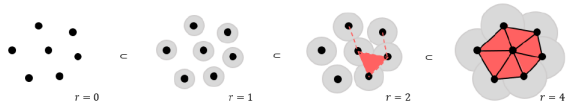


Figura 7: Operador borde: Aplicación del operador borde a un 2-símplice.

Dada una característica topológica presente en una filtración, podemos identificar el valor de r donde el rasgo nace (aparece) y muere (desaparece). Definición 3.17 Para un complejo filtrado K y subcomplejos K_i; K_j donde $i < j$, una característica topológica $x \in H_p(K_j)$ nace en j si $x \neq 0 \in H_p(K_i)$:

Definición 3.18 Para un complejo filtrado K y subcomplejos K_i; K_j donde $i < j$, una característica topológica $x \in H_p(K_i)$ muere en j si $x = 0 \in H_p(K_j)$. Una característica también morirá si se fusiona con una característica nacida antes en la filtración; esto se conoce como la regla mayor.

Definición 3.19 Para una característica topológica determinada x, con un punto de nacimiento i y punto de muerte j, el intervalo de persistencia (Int) de la característica viene dado por:

$$\text{Int} = [i; j) : i \in \mathbb{R}^+ \cup \{0\}; j \in \mathbb{R}^+ \cup \{0, 1\}; j \geq i.$$

Por tanto, podemos definir la persistencia (p) como:

$$p = j - i.$$

Si $j = 1$ entonces la componente no muere durante la filtración (persiste para siempre).

III. Implementación

Se accedió al conjunto de datos CMB-MEL en el repositorio Cancer Imaging Archive (TCIA) a través del enlace [3].

Dentro del conjunto de datos, se seleccionó un archivo SVS para su análisis. Este archivo fue elegido por su relevancia para el estudio del melanoma y su calidad de imagen.

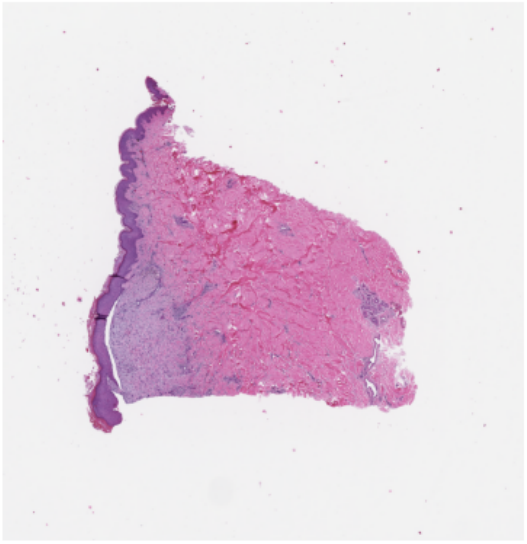


Figura 8: Imagen del archivo SVS seleccionado.

La Figura 8 muestra la imagen SVS seleccionada del archivo. Esta imagen servirá como base para el análisis posterior.

Se extrajo dos secciones de la imagen (parches) de 400x400 píxeles de la imagen SVS seleccionada. Estos parches fueron seleccionado cuidadosamente para incluir áreas relevantes en zona con tumor y no tumor.

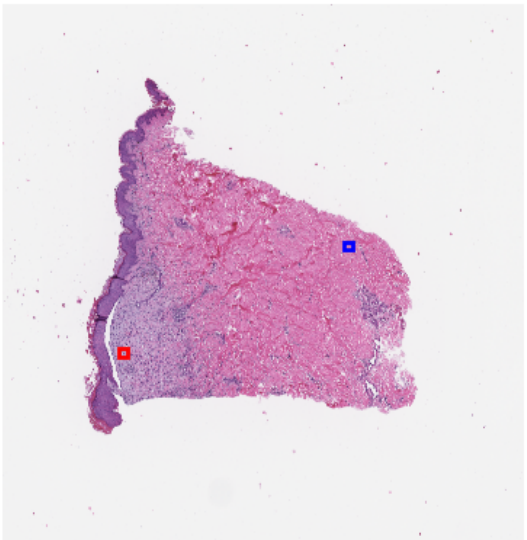
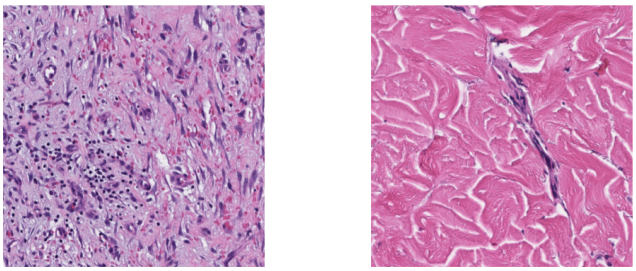


Figura 9: Aqui se muestran los parches obtenidos, donde el de color azul se encuentra en zona no tumoral y la roja en zona tumoral



(a) Zona tumoral (b) Zona no tumoral

Figura 10

Para esta implementación ocuparemos primero la imagen de zona tumoral (parche 1), convirtiendola a escala de grises, esta conversión es beneficiosa de esta forma se reduce la complejidad al tratar solo con un canal y nos ayuda a resaltar las características estructurales de la muestra.

Aplicando el algoritmo de homología persistente, que se describe en el *International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 de julio de 2016, Loughborough, Reino Unido* [2]. Se toma una imagen en escala de grises M de tamaño $m \times n$, donde las intensidades de gris son valores enteros entre 0 y 255, y B es un rectángulo cerrado de tamaño $m \times n$. Para cada valor del umbral t, $B(t) \subseteq B$ es la unión de píxeles con intensidad menor o igual a t. Se binarizan los valores de intensidad en M, reemplazando cualquier valor menor o igual a t por 0 como también los pixeles alrededor de ellos, si el píxel es mayor a t el píxel es remplazado por 1. La matriz resultante $M(t)$ nos da una imagen en blanco y negro.

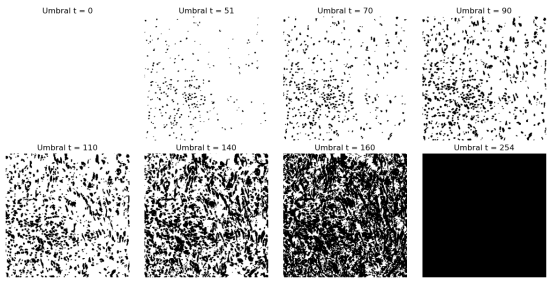


Figura 11: Binarización en el parche, utilizando diferentes umbrales t

El algoritmo busca determinar el número de componentes conexos y el número de agujeros en la imagen binarizada para cada valor de umbral t. Esto se logra mediante el cálculo de los números de Betti cero (β_0) y uno (β_1), los cuales representan el número de componentes y el número de agujeros respectivamente.

En la implementación, calculamos los números de Betti uno (β_1) para cada matriz $M(t)$ que se genera. Esto nos permite obtener información sobre la topología de la imagen y cómo cambia a medida que varía el umbral.

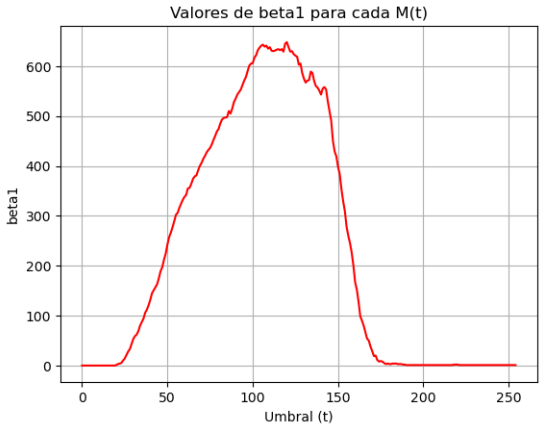
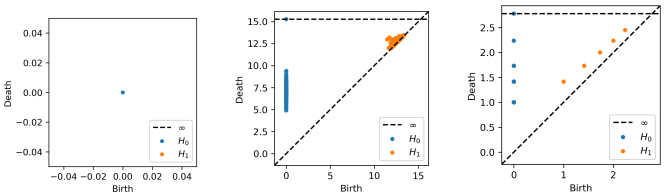


Figura 12: Números de Betti obtenidos por $M(t)$.

La Figura 12 muestra los números de Betti obtenidos para cada $M(t)$.

Para obtener los β_1 de cada $M(t)$ se hizo uso de la función 'label' de la libreria 'scipy.ndimage' para etiquetar regiones conectadas en una matriz de entrada, la cual toma como parametros una matriz binaria y una estructura de conectividad, la cual nos retorna un arreglo de etiquetas del mismo tamaño que el de entrada, donde cada componente conectado tiene un valor de etiqueta único y el número total de componentes conectados encontrados en la imagen, el cual tomaremos como β_1.

Para verificar que se obtuvieron los β_1 correctamente, se utilizó la biblioteca 'ripser' de Python, que proporciona herramientas para calcular la homología persistente de una variedad. En particular, se importó la función 'ripser' y la clase 'Rips' de esta biblioteca.



(a) Para M(10) (b) Para M(150) (c) Para M(230)

Figura 13: Diagramas de persistencia para $t = 10$, 150 y 230

En la figura 13 se muestran los diagramas de persistencia obtenidos con la biblioteca 'ripser', los cuales comparamos a los $M(t)$ para $t = 10$, 150 y 230 que se muestran en la figura 14.

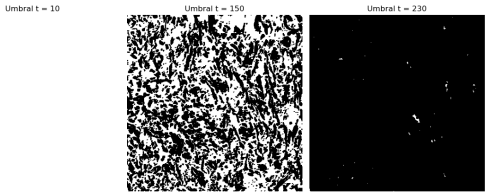


Figura 14: $M(t)$ para $t = 10$, 150 y 230

Se llevo a cabo la misma implementación para el parche que se obtuvo de la zona de no tumor, y se compararón las graficas de cada parche como se muestra a continuación.

Figura 15: β_1 para todo $M(t)$ de los parches en zona tumoral y no tumoral

Por lo que todo este proceso produce un **perfil de homología persistente** para cada parche (en zona tumoral y zona no tumoral), los cuales son distinguibles en términos de sus características cualitativas como se muestra en la figura 15

IV. Conclusión

La homología persistente ha demostrado ser una herramienta efectiva y poderosa para la segmentación de imágenes tumorales en escaneos de diapositivas completas (WSI). Al capturar y cuantificar características topológicas en diferentes escalas, esta metodología permite identificar patrones y estructuras subyacentes en las imágenes que son difíciles de detectar mediante técnicas tradicionales. En las regiones tumorales, donde los núcleos presentan características atípicas y están más densamente agrupados, la homología persistente facilita la segmentación precisa al distinguir estas áreas de las regiones normales, que muestran una mayor variabilidad topológica.

La relevancia de la topología algebraica para la homología persistente radica en su capacidad para proporcionar una representación matemática robusta de las características topológicas de los datos. El análisis topológico de datos, y en particular la homología persistente, es crucial en el contexto de la patología digital y el diagnóstico médico. Proporciona una visión profunda de la estructura y organización de los datos, permitiendo una mejor comprensión y clasificación de las imágenes histológicas. Este enfoque no solo mejora la precisión de la segmentación, sino que también aporta robustez frente a las variaciones en la calidad de las imágenes, ofreciendo un método fiable para el análisis y diagnóstico en entornos clínicos.

Referencias

[1] Munkres, J. R. (2000). *Topología* (2.a ed.). Massachusetts Institute of Technology.

[2] Qaiser, T., Sirinukunwattana, K., Nakane, K., Tsang, Y.-W., Epstein, D., & Rajpoot, N. (2016). Persistent Homology for Fast Tumor Segmentation in Whole Slide Histology Images. En *International Conference On Medical Imaging Understanding and Analysis 2016, MIUA 2016, 6-8 de julio de 2016, Loughborough, Reino Unido*. Department of Computer Science, University of Warwick, Coventry, CV4 7AL, Reino Unido.

[3] Cancer Moonshot Biobank. (2022). *Cancer Moonshot Biobank – Melanoma Collection (CMB-MEL) (Version 5)* [dataset]. The Cancer Imaging Archive. https://doi.org/10.7937/GWSP-WH72

[4] Carlsson, G. (2009). Topology and Data. *Bulletin of the American Mathematical Society*. https://doi.org/10.1090/S0273-0979-09-01249-X