Universidad Nacional Autónoma de México

Facultad de Algo Seguramente Muy Importante

Título de la tesis

T E S I S

QUE PARA OPTAR POR EL GRADO DE:

Médic@/Ingenier@/Licenciad@

PRESENTA:

Nombre Apellido1 Apellido 2

TUTOR:

Nombre Director

Ciudad Universitaria, CDMX, 2020

A la Facultad de Ingeniería y a la Universidad, por la formación que me han dado.
Es gracias a ustedes que es posible el presente trabajo.
En verdad, gracias.
Yo.

Reconocimientos

También quisiera reconocer a ... por ...CONACYT, PAPIIT / etc. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Declaración de autenticidad

Por la presente declaro que, salvo cuando se haga referencia específica al trabajo de otras personas, el contenido de esta tesis es original y no se ha presentado total o parcialmente para su consideración para cualquier otro título o grado en esta o cualquier otra Universidad. Esta tesis es resultado de mi propio trabajo y no incluye nada que sea el resultado de algún trabajo realizado en colaboración, salvo que se indique específicamente en el texto.

Nombre Apellido1 Apellido 2. Ciudad Universitaria, CDMX, 2020

Resumen

This is where you write your abstract ... Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Presentación

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.2. Objetivo

Este trabajo tiene por objetivo ... Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.3. Motivación

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.4. Planteamiento del problema

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.5. Metodología

Se tiene un objetivo principal, y para llegar a él Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.6. Contribuciones

La principal contribución de este trabajo es Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

1.7. Estructura de la tesis

Este trabajo está dividido en XX capítulos. Al principio se encuentra

Finalmente se encuentra la parte de

Capítulo 2

Fundamentos biológicos

2.1. Virus

Un virus es un parásito intracelular infeccioso, cuyo genoma está compuesto por ADN o ARN. Estos utilizan la maquinaria de las células huésped para dirigir la síntesis de sus componentes y ensamblar nuevas partículas virales infecciosas (viriones) que pueden transmitir el genoma viral e iniciar nuevos ciclos de infección en otros huéspedes.(8)

2.2. Genoma

El genoma es el conjunto completo de instrucciones que se encuentran en un organismo, en otras palabras contiene toda la información necesaria para que un organismo se desarrolle y funcione. En la gran mayoría de organismos está compuesto por moléculas de ADN, sin embargo en algunos virus, su material genético está compuesto por moléculas de ARN en lugar de ADN. Ejemplo de esto son el virus del SIDA (VIH) y el virus de la influenza.

En este trabajo se analizarán genomas de virus, específicamente de influenza y SARS-CoV-2. El genoma viral contiene toda la información genética necesaria para que el virus crezca, se replique y pueda continuar transmitiendose de célula a célula.

Existen varios repositorios públicos que se encargan de recolectar, mantener y preservar las secuencias genómicas de virus para su estudio por parte de la comunidad científica. Dos de los principales repositorios y que se estarán ocupando en este trabajo son:

- **NCBI Influenza Virus Database**: Esta base de datos del Centro Nacional para la Información Biotecnológica (NCBI) de Estados Unidos contiene más de 500,000 secuencias genómicas de virus influenza de diferentes tipos, subtipos y hospedadores. Permite hacer búsquedas y descargas de secuencias.(23)
- **GISAID**:

> La Iniciativa Global para Compartir Todos los Datos sobre Influenza (GISAID) es una plataforma que promueve el intercambio internacional de datos genómicos y clínicos sobre los virus influenza. Durante la pandemia de COVID-19, GISAID también se ha convertido en el principal repositorio de secuencias genómicas del SARS-CoV-2 compartidas por científicos de todo el mundo. A mayo de 2023, GISAID alberga más de 14 millones de secuencias genómicas de SARS-CoV-2.(12)

El uso de repositorios como NCBI Influenza Virus Database y GISAID permite acceder a una gran cantidad de datos genómicos que representan la diversidad de virus influenza y SARS-CoV-2 circulantes. Esto posibilita estudios comparativos, evolutivos y de vigilancia epidemiológica.

2.3. Genoma de referencia

Una secuencia de referencia es una representación aceptada que es utilizada por los investigadores como estándar para la comparación con otras secuencias generadas.(11) Esta secuencia de referencia es obtenida mediante el mapeo de miles de muestras individuales. En el caso del virus del SARS-CoV-2 el primer genoma de referencia obtenido fue reportado en enero 2020 por Zhu et al, sobre la base de muestras iniciales del brote en Wuhan. Este se ha venido actualizando con la incorporación de nuevas variantes (39)

2.4. El virus de influenza

Los virus influenza son un tipo de virus segmentado perteneciente a la familia Orthomyxoviridae (15). Hay cuatro tipos de virus de la influenza: A, B, C y D. Solo los virus A y B son los causantes de epidemias estacionales de la enfermedad en las personas.

El genoma del virus influenza A consta de 8 segmentos de ARN de sentido negativo que codifican para 10 proteínas virales. Entre estas proteínas se incluyen glicoproteínas de superficie, proteínas estructurales y proteínas implicadas en la replicación viral (5).

Existen 18 subtipos diferentes de hemaglutinina (HA) y 11 subtipos de neuraminidasa (NA). Los subtipos son designados mediante la combinación de las cantidades de H y N, esto permite categorizarlos y estudiar sus características específicas (2).

La distribución de los subtipos de influenza varía según el hospedador. Los virus humanos poseen las combinaciones de H1, H2, H3 y N1 y N2, las cuales se comparten con los cerdos. Otros subtipos como H5, H6, H7, H9 y N1, N2, N3, N7, solo circulan en aves de corral, mientras que las ballenas y gaviotas tienen H13, N2, N6. Todos los tipos, subtipos y variantes se encuentran en las aves migratorias y silvestres.

Aunque se han identificado más de 130 combinaciones de subtipos de influenza A, es posible que existan muchas más combinaciones gracias a una caracteristica compartida

de los virus de ARN segmentados conocido como reordenamiento (35), cuya definición se abordará más adelante.

2.4.1. Principales proteínas de estudio

- **Hemaglutinina (HA)**: Esta glicoproteína de superficie se une a los receptores de ácido siálico en las células huésped, facilitando la entrada del virus. Es el principal objetivo de los anticuerpos neutralizantes.
- **Neuraminidasa (NA)**: Esta enzima de superficie ayuda a liberar las nuevas partículas virales de las células infectadas cortando los residuos de ácido siálico.
- **Proteína de la matriz (M1)**: Forma la capa estructural debajo de la envoltura viral y juega un papel crucial en el ensamblaje y la liberación del virus.
- **Proteína de canal iónico (M2)**: Forma canales de protones en la envoltura viral, esenciales para la desencapsidación del virus después de la entrada.
- **Nucleoproteína (NP)**: Se une al ARN viral y es esencial para la transcripción, replicación y empaquetamiento del genoma viral.
- **Proteínas del complejo de la polimerasa (PA, PB1, PB2)**: Forman el complejo de la ARN polimerasa viral, responsable de la transcripción y replicación del genoma viral.

(5)

Estas proteínas trabajan en conjunto para permitir la entrada del virus en las células huésped, su replicación y la liberación de nuevas partículas virales. La hemaglutinina y la neuraminidasa, en particular, son las principales determinantes antigénicas y juegan un papel crucial en la interacción del virus con el sistema inmune del huésped (5).

2.5. Reordenamiento

El reordenamiento es el proceso mediante el cual dos o más virus intercambian segmentos de genes, esto ocurre cuando dos virus parcialmente diferentes coinfectan una célula (31). Decimos parcialmente diferentes pues un reordenamiento exitoso entre dos cepas parentales requiere de un alto grado de compatibilidad genética. De hecho, no hay registrado algún caso en el que haya ocurrido reordenamiento entre virus de ARN segmentados que pertenezcan a diferentes familias.(20)

Durante la replicación viral dentro de la célula, los segmentos genómicos de los diferentes virus parentales se mezclan y se empaquetan al azar en las nuevas partículas virales que se están formando. Como resultado, se genera una descendencia viral híbrida que puede contener una combinación única de segmentos genéticos derivados de los virus originales que co-infectaron la célula. (20)

Este proceso al darle a un virus contenido genético derivados de más de un progenitor, confiere potencialmente importantes ventajas o desventajas de aptitud a la progenie viral (18) lo que promueve la evolución de los virus de influenza, llegando a causar efectos como nuevas cepas con la capacidad de evadir la respuesta inmune del huésped ,adquirir resistencia a antivirales e inclusive puede permitir a los virus infectar nuevas especies hospedadoras. (22). (10)

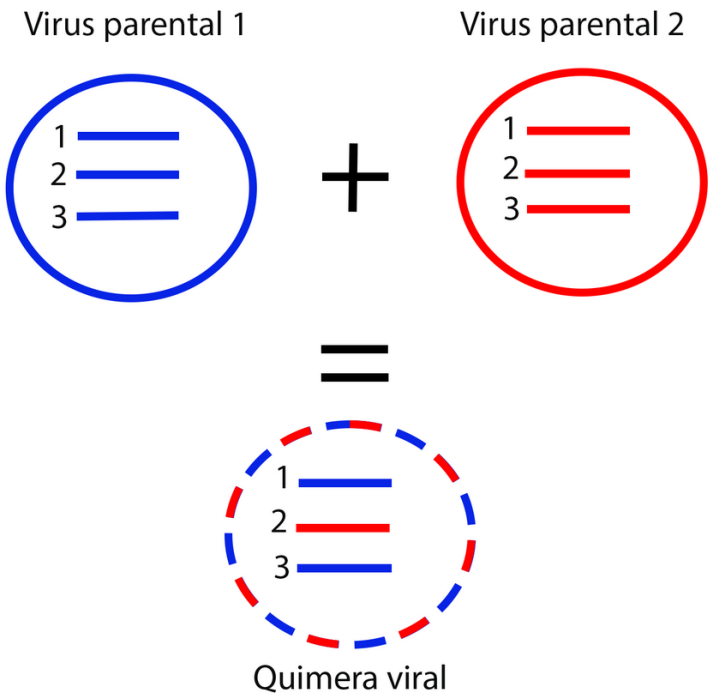

Figura 2.1: Ejemplo de reordenamiento en virus con genoma segmentado.

2.6. El virus del Sars Cov2

El SARS-CoV-2 es un nuevo coronavirus que emergió a finales de 2019 en la ciudad de Wuhan, China, y es el agente causal de la enfermedad por coronavirus 2019 (COVID-19). Este virus altamente transmisible y patógeno generó una pandemia que amenazó la salud humana y la seguridad pública a nivel global.La Organización Mundial de la Salud declaró oficialmente pandemia de COVID-19 el 11 de marzo de 2020 (14).

Pertenece a la familia Coronaviridae, específicamente a la subfamilia Orthocoronavirinae dentro del género Betacoronavirus. Está estrechamente relacionado con el coronavirus del SARS (SARS-CoV) que causó la epidemia de SARS en 2002-2003,

compartiendo aproximadamente un 79 % de genoma (14, 40).

Aunque la evidencia genética sugiere que el SARS-CoV-2 es un virus de origen natural que probablemente se originó en animales, aún no hay una conclusión sobre cuándo y dónde entró por primera vez en los humanos (40).

De manera similar al proceso de reordenamiento en la influenza,en el virus del Sars Cov2 existe un proceso llamado recombinación.

2.6.1. Principales proteínas de estudio

- **Proteína S (spike)**: Esta proteína forma los picos que sobresalen de la superficie del virus. Permite que el virus se una a las células humanas para infectarlas. Coordenadas de Secuencia de Codificación de Proteínas (21563-25384)
- **Proteína M (membrana)**: Esta proteína forma la envoltura que rodea al material genético del virus. Desempeña un papel importante en el ensamblaje y liberación de partículas virales. Coordenadas de Secuencia de Codificación de Proteínas (26523-27191)
- **Proteína E (envoltura)**: Es una pequeña proteína de membrana involucrada en el ensamblaje y liberación de nuevas partículas virales. Coordenadas de Secuencia de Codificación de Proteínas (26245-26472)
- **Proteína N (nucleocápside)**: Se une al material genético viral (ARN) y lo protege dentro de la nucleocápside. Coordenadas de Secuencia de Codificación de Proteínas (28274-29533)

(25) (17)

2.7. Recombinación

La recombinación, al igual que el reordenamiento, es un proceso fundamental en la evolución viral que permite el intercambio de material genético entre diferentes genomas virales (26). Este fenómeno ocurre cuando al menos dos genomas virales co-infectan la misma célula huésped e intercambian partes de genes o secuencias del mismo segmento (26, 34). A diferencia del reordenamiento, que solo ocurre en virus con genoma segmentado, la recombinación puede darse tanto en virus con genoma segmentado como en aquellos con genoma no segmentado.

La recombinación está íntimamente ligada a la replicación viral. En virus de ADN, ocurre durante la replicación del ADN e implica tanto proteínas virales como celulares (34). En virus de ARN, está generalmente asociada con el proceso de síntesis de ARN viral. El modelo más aceptado es el de .elección de copia", donde la ARN polimerasa viral cambia de molde durante la síntesis (16, 37). Este proceso requiere suficiente

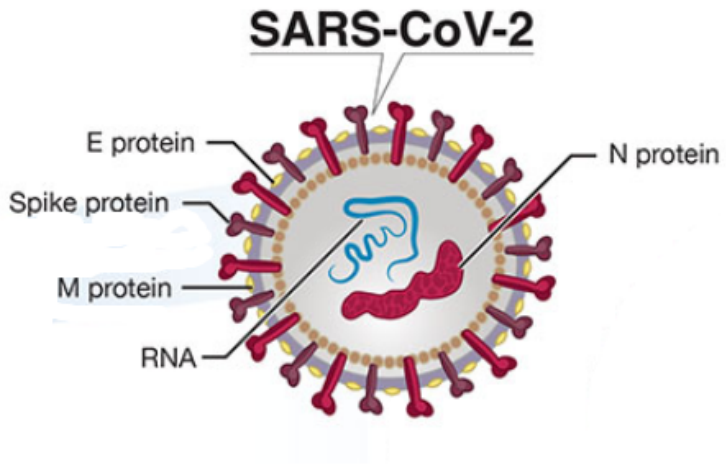

Figura 2.2: Estructura del SARS-CoV-2 mostrando sus principales componentes. En la parte central del virus se observa el ARN viral asociado a la proteína de la nucleocápside (N). En la periferia se encuentran los trímeros de la glicoproteína S (spike) que forman los característicos picos, junto con las proteínas de membrana (M) y las proteínas de envoltura (E). Fuente:"Fphar-11-00937-g001"por Colin D. Funk, Craig Laferrière, y Ali Ardakani. Gráfico por Ian Dennis - http://www.iandennisgraphics.com, CC BY 4.0, via Wikimedia Commons, https://commons.wikimedia.org/wiki/File:Fphar-11-00937-g001.jpg. Imagen recortada.

similitud de secuencia entre los genomas parentales para permitir el apareamiento de bases y la continuación de la síntesis (37).

La frecuencia de recombinación varía ampliamente entre los virus. Por ejemplo, el virus SARS-CoV-2 tiene tasas de recombinación moderadamente altas, aunque no tan elevadas como las de los alfaherpesvirus (17, 26, 37).

La recombinación tiene importantes implicaciones evolutivas y epidemiológicas. Puede generar nuevas variantes virales con propiedades alteradas, como cambios en el tropismo celular, la virulencia o la capacidad de evadir la respuesta inmune del huésped (26, 37). También puede facilitar la adaptación viral a nuevos huéspedes y ambientes (26).

Capítulo 3

Fundamentos matemáticos

3.1. Análisis Topológico de Datos

El Análisis Topológico de Datos (TDA, por sus siglas en inglés) es una colección de métodos que buscan encontrar y cuantificar estructura en los datos utilizando conceptos de la topología. El TDA se basa en la idea de que los datos tienen una "forma"que puede ser analizada rigurosamente para obtener información valiosa (36).

Los métodos de TDA incluyen:

1. Agrupamiento (clustering): Quizás la forma más simple y antigua de TDA, el agrupamiento busca identificar grupos naturales en los datos. Los métodos basados en densidad, como el agrupamiento por nivel de densidad, tienen una clara conexión con la topología (36).

2. Estimación de variedades (manifold estimation): Estos métodos asumen que los datos de alta dimensión yacen cerca de una variedad de menor dimensión, y buscan estimar esta estructura subyacente (36).

3. Reducción no lineal de dimensionalidad: Técnicas como Isomap o mapas de difusión que intentan preservar la estructura topológica de los datos al proyectarlos en un espacio de menor dimensión (21).

4. Estimación de modos y crestas: Métodos que buscan identificar características de alta densidad en los datos, como modos (máximos locales de densidad) o crestas (curvas o superficies de alta densidad) (36).

5. Homología persistente: Una técnica central en TDA que estudia cómo evolucionan las características topológicas (como componentes conectados, ciclos, o cavidades) a diferentes escalas (21).

El TDA ofrece varias ventajas sobre los métodos tradicionales de análisis de datos. Es insensible a la elección específica de coordenadas y puede manejar datos con rui-

do. Además, proporciona una forma de resumir y visualizar datos complejos de alta dimensión (21).

Las técnicas de TDA han demostrado ser herramientas muy útiles. En la actualidad se han encontrado varias aplicaciones en diversos campos, desde la biología y la neurociencia hasta la astrofísica y el análisis de imágenes. (1, 13, 21, 30, 36).

En la siguientes sección, profundizaremos en la técnica clave de este proyecto, la homología persistente, una de las herramientas centrales del TDA.

3.2. Homología persistente

La homología persistente es una herramienta fundamental del análisis topológico de datos que permite estudiar las características topológicas de un objeto a múltiples escalas.(27)A diferencia de la homología clásica, la homología persistente captura los cambios en la homología a medida que el objeto evoluciona con respecto a un parámetro.(9)

La homología persistente se construye sobre la noción de complejos simpliciales, que son generalizaciones de grafos que incluyen componentes de dimensiones superiores. Un simplejo es el bloque constructor básico de un complejo simplicial. Formalmente, un k-simplejo es la envolvente convexa de k+1 puntos afinmente independientes en un espacio euclídeo (6). Por ejemplo, un 0-simplejo es un punto, un 1-simplejo es una arista que conecta dos puntos, un 2-simplejo es un triángulo lleno, un 3-simplejo es un tetraedro sólido, y así sucesivamente.

Dada una filtración, es decir, una secuencia finita de complejos simpliciales anidados, la homología persistente rastrea las clases de homología que aparecen y desaparecen a lo largo de la filtración.(9).

Cada clase de homología tiene un tiempo de nacimiento (cuando aparece por primera vez), y un tiempo de muerte (cuando desaparece). La persistencia de una clase de homología, dada por la diferencia entre su tiempo de muerte y nacimiento, mide su relevancia dentro de la estructura topológica global del objeto.

La información capturada por la homología persistente se puede visualizar mediante códigos de barras (barcodes) o diagramas de persistencia. En un código de barras, cada barra representa una característica topológica, y la posición y longitud de la barra indican el rango de escalas en el que la característica está presente. En un diagrama de persistencia, cada punto representa una característica topológica, donde la coordenada x es el tiempo de nacimiento y la coordenada y es el tiempo de muerte. En ambos casos, las características con mayor persistencia se consideran más relevantes.(27)

(9) En el contexto del análisis de datos genómicos, la homología persistente provee un enfoque para estudiar la evolución horizontal. (28)(4)(29). El espacio de posibles secuencias genéticas se puede ver como un objeto topológico de alta dimensión, donde cada secuencia es un punto. Bajo la suposición de que cada sitio genético muta a lo más una vez en la historia evolutiva de la muestra, la única forma de crear ciclos en este espacio es a través de eventos de recombinación que combinan mutaciones provenientes

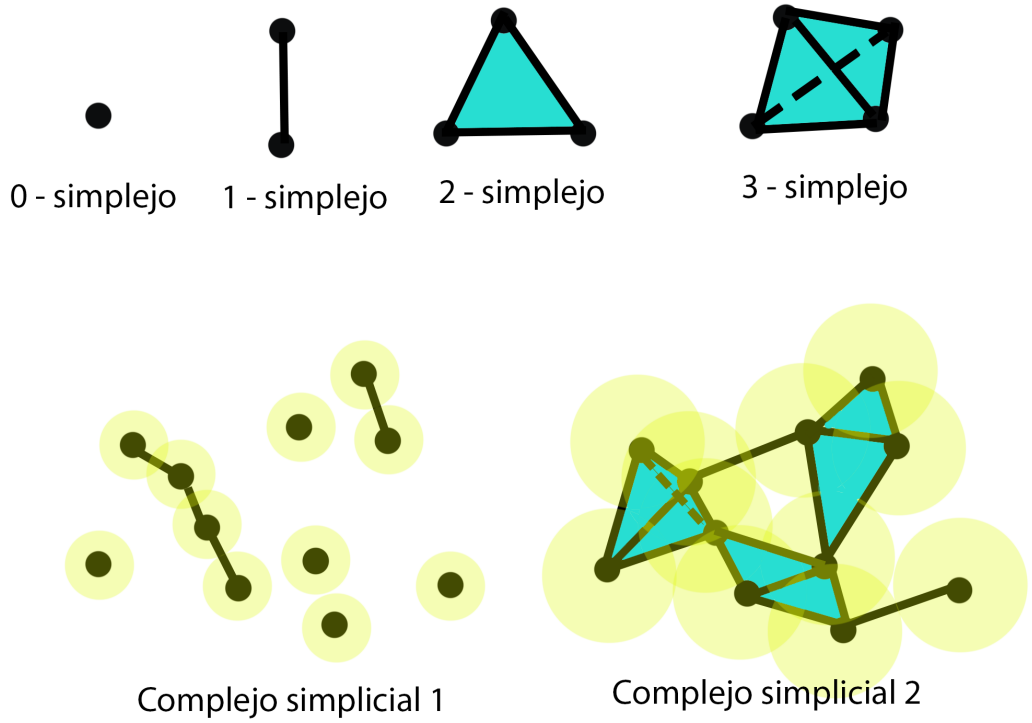

Figura 3.1: Simplejos y complejos simpliciales. La parte superior muestra ejemplos de k-simplejos, desde un 0-simplejo (punto) hasta un 3-simplejo (tetraedro). La parte inferior ilustra dos complejos simpliciales construidos, podemos notar como dependiendo del radio r de los circulos amarillos va variando el complejo simplicial que obtendremos

de distintos linajes. Por lo tanto, los ciclos detectados por homología persistente a diferentes resoluciones capturan la presencia de recombinación ancestral en la muestra (4).

Así, cada objeto de dimensión uno encontrado en los códigos de barras o en los diagramas de persistencia nos representa un evento de recombinación, donde los tiempos de nacimiento y muerte indican las distancias genéticas entre las secuencias parentales involucradas (28).

Los números de Betti, denotados como b_k cuentan el número de ciclos independientes de dimensión k. Intuitivamente, b_0 cuenta el número de componentes conectados, b_1 cuenta los ciclos o "agujeros" unidimensionales, b_2 cuenta las cavidades o "vacíos"bidimensionales, y así sucesivamente (27). En el análisis de datos genómicos, los números de Betti tienen una interpretación biológica. b_0 captura el número de linajes o componentes distintos en la muestra. b_1 y números de Betti de orden superior cuentan eventos de recombinación ancestral, donde secuencias de diferentes linajes se

han mezclado (28)(4)(29). Específicamente, b_1 cuenta eventos de recombinación que involucran dos secuencias parentales, b_2 cuenta eventos que involucran tres secuencias parentales, y así sucesivamente (4)(29). De esta manera, los números de Betti proveen una descripción topológica de la complejidad de la historia evolutiva de la muestra.

Capítulo 4

Metodología de análisis topológico para detección de transferencia genética horizontal

4.1. Aplicación de TDA para detección de reordenamiento triple en H7N9

Para detectar el evento de reordenamiento triple en el virus de influenza H7N9, se aplicó un enfoque basado en Análisis Topológico de Datos (TDA). El procedimiento consistió en los siguientes pasos:

1. **Obtención de datos:** Se obtuvieron las secuencias genómicas de interés de la base de datos NCBI (23). Cada genoma se descargó en ocho segmentos separados, correspondientes a los ocho genes que componen el genoma de los virus de influenza A: PB2, PB1, PA, HA, NP, NA, M y NS. La descarga de las secuencias se realizó en marzo de 2024.

   Las secuencias utilizadas fueron las siguientes:

   - A/chicken/Zhejiang/329/2011(H9N2)
   - A/brambling/Beijing/16/2012(H9N2)
   - A/Shanghai/02/2013(H7N9)
   - A/wild bird/Korea/A14/2011(H7N9)
   - A/quail/Lebanon/273/2010(H9N2)
   - A/duck/Wuxi/7/2010(H9N2)
   - A/quail/Wuxi/7/2010(H9N2)

- A/baikal teal/Xianghai/421/2011(H9N2)
- A/duck/Zhejiang/2/2011(H7N3)
- A/duck/Zhejiang/10/2011(H7N3)

2. **Separación de cada segmento:** Se separó cada segmento de cada secuencia utilizando un programa en python de modo que todos los segmentos correspondientes a una misma región estuvieran en un solo archivo fasta.

3. **Alineamiento de secuencias:** Cada uno de los ocho archivos que contenían a nuestros segmentos fueron alineados. Estos se alinearon utilizando el software Clustal Omega (19)(7), con los siguientes parámetros: formato de salida Pearson/FASTA, sin input DEALIGN, agrupamiento MBED-like, con un máximo de iteraciones guiadas por árbol por defecto y orden alineado. El acceso al software Clustal Omega se realizó en marzo de 2024.
   Clustal Omega comienza realizando alineamientos por pares de todas las secuencias ingresadas. Utiliza el método de k-meros para una rápida comparación inicial.

4. **Concatenación de los segmentos:** Utilizando un programa creado en python se concatenaron cada uno de los segmentos debidamente alineados de cada una de nuestras secuencias de manera que para cada secuencia el segmento 1 fuera el primero, luego se le concatenó el 2, luego el 3 y así sucesivamente.
   Todas las secuencias concatenadas fueron colocadas en un archivo fasta.

5. **Obtención de la matriz de distancias:** Utilizando un programa creado en python, al mandar como parámetro de entrada el archivo fasta anteriormente creado se obtuvo la matriz de distancia.
   La distancia se basó en el número de diferencias (mutaciones) entre pares de secuencias en los rangos especificados.
   Se excluyeron las posiciones con nucleótidos indeterminados ('N') del cálculo. Para cada par de secuencias, solo se contó la cantidad de diferencias de la posición 0 a la 6820 y de la posición 7636 hasta la última (13498). Esto debido a que la secuencia A/wild bird/Korea/A14/2011(H7N9) en el segmento 4 correspondiente a la hemaglutinina (HA) después de ser alineada contenía muchas deleciones, lo cual impedia trabajar con ella de manera correcta.

6. **Aplicación del método:** Se trabajó en python con 2 librerías similares para verificar resultados. Estas librerias fueron GUDHI(33) y giotto-tda (32). Primeramente se creó un mapa de calor, proporcionando una representación visual de la variabilidad genética en cada región, a continuación se construyó el complejo de Rips a partir de la matriz de distancia.
   Finalmente se calculó el diagrama de persistencia y código de barras, capturando las características topológicas de los datos.
   Al hacer este procedimiento en ambas librerias con sus respectivas funciones pudimos comprobar que el resultado era el mismo en ambas.

4.2. Aplicación de TDA para detección de recombinación en Sars-Cov-2

Para detectar el evento de recombinación en el virus de Sars-Cov-2, al igual que en el experimento anterior nos basamos en Análisis Topológico de Datos (TDA). El procedimiento consistió en los siguientes pasos:

1. **Obtención de datos:** Se obtuvieron las secuencias genómicas de interés completas de la base de datos Gisaid/Epicov (12)

   Las secuencias utilizadas fueron las siguientes:

   - hCoV-19/Mexico/QUE-LANGEBIO_IMSS_14092/2023(BA.2.10.1)
   - hCoV-19/Mexico/CMX_IBT_IMSS_9908/2022(BA.2.75.2)
   - hCoV-19/Mexico/HID-INMEGEN-70-132/2022(BA.2.10.1)
   - hCoV-19/Mexico/MEX-INMEGEN-92x-18/2022(BA.2.75.2)
   - hCoV-19/Mexico/BCN-CIAD-IMSS_20220105091312022(BA.2)
   - hCoV-19/Mexico/MEX_InDRE_FB13911_E_S13150/2022(BA.2)
   - hCoV-19/Mexico/QRO-CIAD-VZT_HJ6105/2022(BA.2)
   - hCoV-19/Mexico/TAM_LANGEBIO_IMSS_13849/2022(XBB.1)
   - hCoV-19/Mexico/TAM_LANGEBIO_IMSS_13854/2022(XBB.1)
   - hCoV-19/Mexico/CIAD-VZT_HJ6104/2022(XBB.1)

2. **Alineación de secuencias:** Las secuencias fueron alineadas utilizando la interfaz web de la herramienta NextClade v3.7.4 (24) utilizando como secuencia de referencia del genoma la secuencia SARS-CoV-2 (XBB) que viene incluida en su base de datos de referencia.
   Nextclade utiliza un algoritmo de alineamiento basado en el algoritmo de Needleman-Wunsch modificado. Este algoritmo compara cada secuencia ingresada con la secuencia de referencia.

3. **Obtención de la matriz de distancias:** Se utilizó el mismo programa que en el experimento anterior para obtener nuestra matriz de distancias pero esta vez no hubo una exclusión de alguna posición de nuestros genomas.

4. **Aplicación del método:** Por último se utilizó la libreria de GUDHI y giotto-tda en python. De manera análoga al experimento anterior se creó un mapa de calor para la matriz de distancia, a continuación se construyó el complejo de Rips a partir de la matriz de distancia.
   Finalmente se calculó el diagrama de persistencia y código de barras.

4.3. Visualización tridimensional de complejos simpliciales

Para facilitar la interpretación de los resultados del análisis topológico de datos, se desarrollaron dos métodos de visualización tridimensional de los complejos simpliciales:

4.3.1. Visualización detallada para el virus de la influenza

Se implementó una visualización tridimensional interactiva y detallada específicamente para el análisis del virus de la influenza, utilizando la biblioteca Plotly en Python. Esta visualización permite una exploración dinámica de la formación de los complejos simpliciales a medida que cambia el valor de filtración, para esto se puede escribir el valor de filtración que específicamente se busca visualizar o si se desea se puede deslizar una barra que nos permite avanzar o retroceder el tiempo de filtración. Las características principales de esta visualización son:

- Representación de nodos (0-simplejos) como vértices, con colores que indican la cepa viral (H9N2 en azul, H7N9 en naranja, H7N3 en verde).
- Visualización de aristas (1-simplejos) como líneas que conectan los nodos.
- Representación de triángulos (2-simplejos) como superficies semitransparentes de color rojo.
- Visualización de tetraedros (3-simplejos) como volúmenes semitransparentes de color verde.
- Se destacó el "agujero" topológico tridimensional encontrado mediante un color amarillo distintivo.
- Controles interactivo que nos permiten modificar el valor de filtración, permitiendo observar la evolución del complejo simplicial.

4.3.2. Visualización simplificada

Adicionalmente, se desarrolló una visualización tridimensional más simple y generalizada, aplicable tanto al análisis del virus de la influenza como al SARS-CoV-2. Esta visualización solo nos permite visualizar un valor de filtración a la vez. Aunque menos detallada, requiere de mucho menos recursos y permite una comparación directa entre los resultados. Las características de esta visualización son las mismas que la visualización anteriormente mencionada con excepción de los controles interactivos, ya que es una visualización estática.

Capítulo 5

Análisis de Resultados

5.1. Aplicación de TDA para detección de reordenamiento triple en H7N9

Luego de aplicar homología persistente de manera satisfactoria a nuestras secuencias obtuvimos su gráfica de persistencia (Figura 4.1) junto con su respectivo código de barras de persistencia (Figura 4.2). En ambas gráficas podemos visualizar que se encontraron varios objetos entre los que destacan dos objetos de dimensión 1 (H1) y un objeto de dimensión 2 (H2).

El primer objeto que se obtuvo de dimensión 1 nació en el tiempo 1319 y murió en el tiempo 1383, nuestro segundo objeto de dimensión 1 nació en el tiempo 1673 y murió en el tiempo 2020. Y por último el objeto de dimensión 3 nació en el tiempo 2057 y murió en el tiempo 2080.

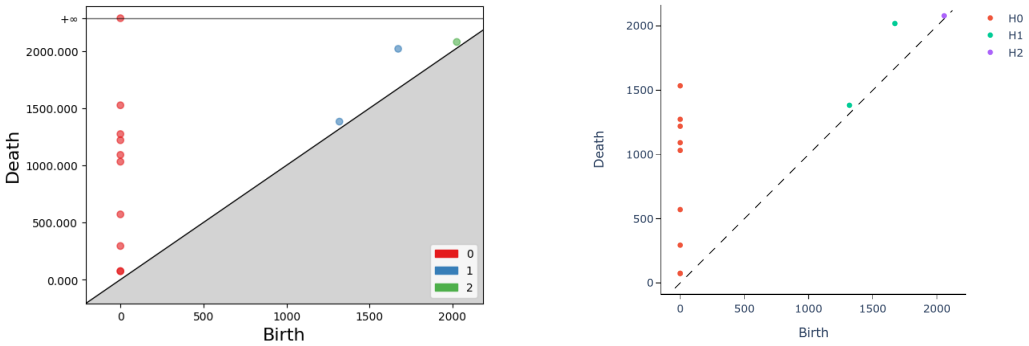

Figura 5.1: Diagramas de persistencia obtenidos como resultado de aplicar TDA a la matriz de distancia de nuestras secuencias de influenza, se utilizaron dos librerias para el procesamiento de TDA ambas con el mismo resultado, GUDHI (izquierda) y GIOTTO-TDA (derecha)

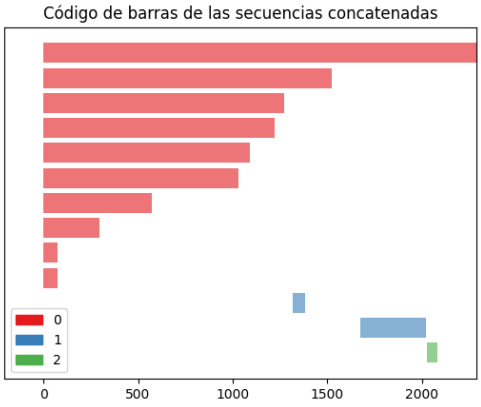

Figura 5.2: Código de barras de persistencia de nuestras secuencias de Influenza A obtenido mediante la librería GUDHI

Así mismo se obtuvo un mapa de calor de nuestra matriz de distancias (Figura 4.3), este nos ayuda a mostrar las relaciones o distancias genéticas entre las secuencias utilizadas. Las celdas más oscuras indican una mayor distancia o diferencia genética entre las cepas correspondientes, mientras que las celdas más claras sugieren una mayor similitud genética. Si nos fijamos en el renglón de la secuencia A/Shanghai/02/2013(H7N9) podemos notar que hay 3 cuadros que son de un color anaranjado, lo cual nos representa una lejania filogenética con respecto a las demás secuencias. Esto puede ser un indicio de que las secuencias A/brambling/Beijing/16/2012(H9N2), A/chicken/Zhe-

jiang/329/2011(H9N2) y A/wild bird/Korea/A14/2011(H7N9) estén implicadas en el reordenamiento triple que formó a la secuencia A/Shanghai/02/2013(H7N9).

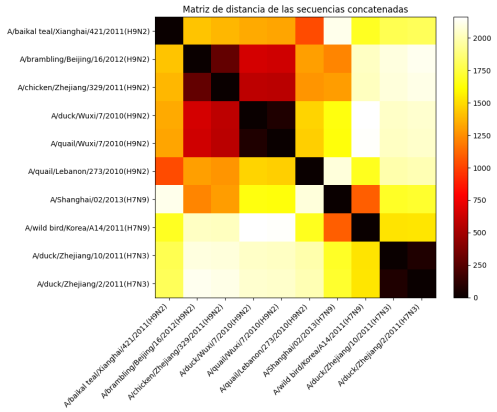

Figura 5.3: Mapa de calor de la matriz de distancias de nuestras secuencias de Influenza A

Podemos comparar los resultados obtenidos con los resultados obtenidos en los artículos "Human Infection with a Novel Avian-Origin Influenza A (H7N9) Virus"(10) y "Sequential Reassortments Underlie Diverse Influenza H7N9 Genotypes in China"(38) donde utilizaron otras técnicas para lograr encontrar los genes relacionados a la cepa H7N9 y llegaron a resultados similares a los nuestros.

En el artículo de Gao et al., los autores realizaron un análisis filogenético de los genes de las cepas A/Shanghai/1/2013, A/Shanghai/2/2013 y A/Anhui/1/2013. Encontraron que el gen de la neuraminidasa (NA) estaba más cercano a A/wild bird/Korea/A14/2011(H7N9). Además, los seis genes internos compartían la mayor similitud con virus A/brambling/Beijing/16/2012-like(H9N2).(10)

Por otro lado, Wu et al. realizaron un análisis evolutivo en profundidad de 45 virus H7N9 y 42 virus H9N2. Sugieren que los virus H7N9 se generaron mediante al menos dos pasos de reordenamientos secuenciales que involucran distintos virus H9N2 donantes en diferentes hospederos (Lo cual podría explicar por qué encontramos los objetos de dimensión 1 al aplicar homología persistente a nuestros datos). El primer reordenamiento probablemente ocurrió en aves silvestres, dando origen a virus similares a A/Shanghai/1/2013(H7N9) como es el caso de A/Shanghai/2/2013(H7N9). Estos virus se transmitieron a aves domésticas donde sufrieron un segundo reordenamiento con virus H9N2 que circulaban en aves de corral en el este de China, dando lugar a los diversos genotipos H7N9.(38)

Además se aplicó homología persistente a cada uno de los segmentos de nuestras secuencias por separado, pudiendo resaltar que en ninguno de sus diagramas de persistencia se encontraron objetos de dimensión mayor a 0 (Figura 4.4). La explicación más probable a que los segmentos por separado no contengan objetos de dimensión mayor a cero mientras que al concatenarlos sí, es que haya ocurrido un reordenamiento génico

completo, en el que un segmento entero de ARN se transfirió de un virus a otro.

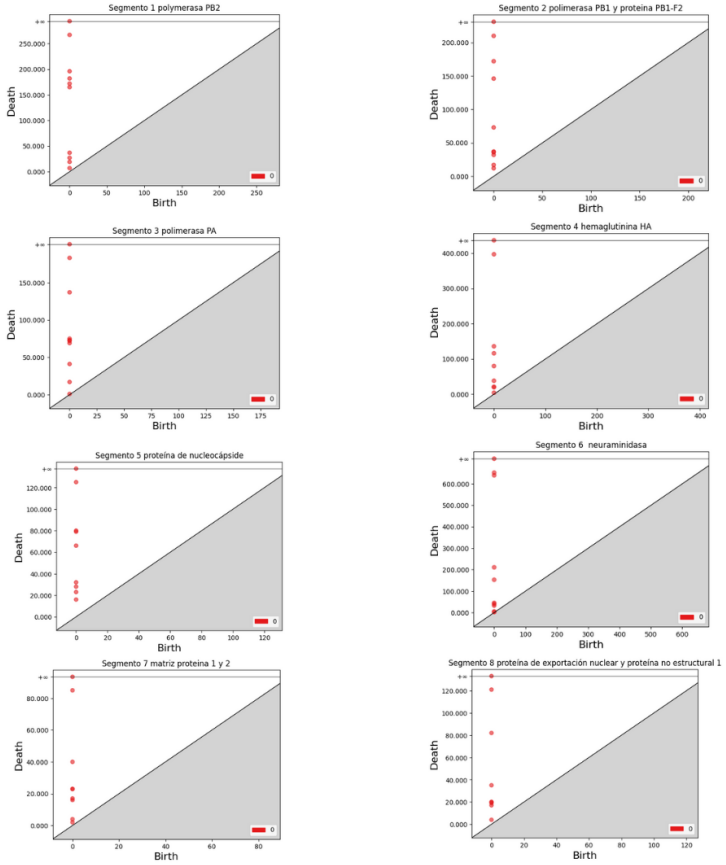

Figura 5.4: Diagrama de persistencia de cada uno de los segmentos de nuestras secuencias de Influenza A

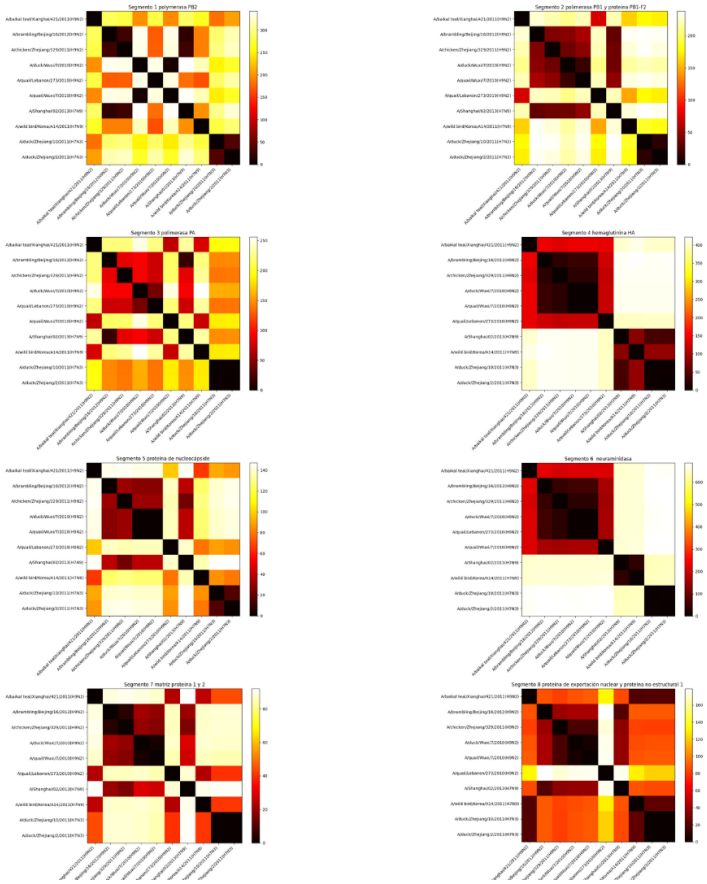

Figura 5.5: Mapa de calor de cada uno de los segmentos de nuestras secuencias de Influenza A

5.1.1. Resultados de la visualización tridimensional para el virus de la influenza

La visualización tridimensional de los complejos simpliciales permitió identificar claramente la formación de agujeros topológicamente significativos:

- Dos agujeros unidimensionales (1-dimensional holes) que aparecen en los intervalos de filtración [1319, 1383] y [1673, 2020].
- Un agujero bidimensional (2-dimensional hole) que se forma en el intervalo [2026, 2080].

La aparición de estos agujeros corresponde a eventos de reordenamiento genético entre las diferentes cepas virales analizadas. Específicamente:

- Los agujeros unidimensionales sugieren la presencia de dos eventos de reordenamiento distintos que involucran pares de cepas virales. Estos podrían corresponder a los pasos secuenciales de reordenamiento propuestos por Wu et al.(38)
- El agujero bidimensional indica un evento de reordenamiento más complejo que involucra al menos tres cepas virales diferentes, lo cual es consistente con la hipótesis de un reordenamiento triple en el virus H7N9.

La visualización tridimensional reveló cómo la estructura del complejo simplicial evoluciona a medida que aumenta el valor de filtración. Se observó que:

- A valores de filtración bajos, los nodos (que representan las diferentes cepas virales) aparecen aislados.
- A medida que aumenta el valor de filtración, se forman conexiones entre los nodos, primero creando aristas y luego triángulos.
- Los agujeros unidimensionales se forman cuando se cierran ciclos entre tres o más nodos, indicando relaciones evolutivas más complejas.
- El agujero bidimensional aparece cuando se forma una cavidad tridimensional en el complejo, sugiriendo un evento de reordenamiento que involucra múltiples cepas simultáneamente.

Esta representación visual proporciona una intuición clara de cómo las diferentes cepas virales están relacionadas genéticamente y cómo los eventos de reordenamiento han dado forma a la evolución del virus H7N9.

5.2. Aplicación de TDA para detección de recombinación en Sars-Cov-2

Luego de aplicar homología persistente de manera satisfactoria a nuestras secuencias obtuvimos su gráfica de persistencia (Figura 4.6) junto con su respectivo código de barras de persistencia (Figura 4.7). En ambas gráficas podemos visualizar que se encontró un objetos de dimensión 1 (H1). El objeto detectado nació en el tiempo 58 y murió en el tiempo 74.

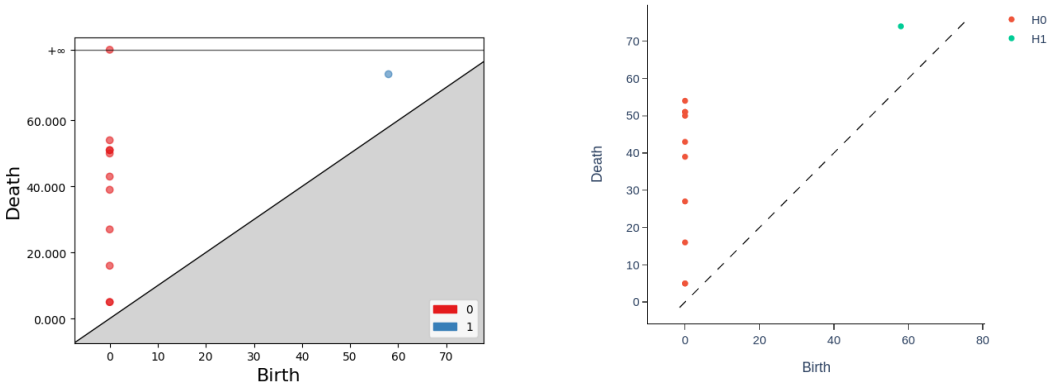

Figura 5.6: Diagramas de persistencia obtenidos como resultado de aplicar TDA a la matriz de distancia de nuestras secuencias de Sars-Cov-2, se utilizaron dos librerias para el procesamiento de TDA ambas con el mismo resultado, GUDHI (izquierda) y GIOTTO-TDA (derecha)

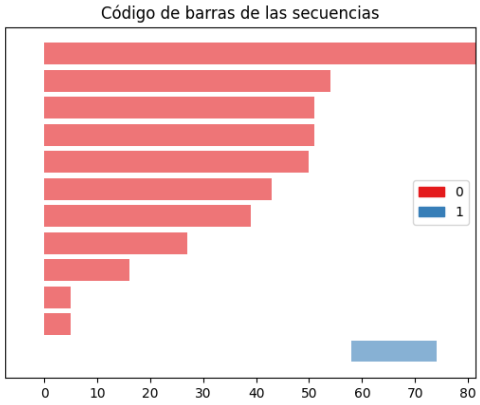

Figura 5.7: Código de barras de persistencia de nuestras secuencias de Sars-Cov-2 obtenido mediante la librería GUDHI

Por otro lado, obtuvimos el mapa de calor de nuestra matriz de distancia (Figura 4.8), por cuestiones de espacio no fue posible agregar una etiqueta a un lado de la fila y debajo de la columna con el nombre de la secuencia representada, por lo que sustituimos los nombres por números del modo que se describe a continuación.

0. BA.2.10.1_hCoV-19/Mexico/QUE-LANGEBIO_IMSS_14092/2023

1. BA.2.75.2_hCoV-19/Mexico/CMX_IBT_IMSS_9908/2022
2. BA.2.10.1_hCoV-19/Mexico/HID-INMEGEN-70-132/2022
3. BA.2.75.2_hCoV-19/Mexico/CMX-INMEGEN-95-310/2022
4. BA.2.75.2_hCoV-19/Mexico/MEX-INMEGEN-92x-18/2022
5. BA.2_hCoV-19/Mexico/BCN_CIAD-IMSS_202201050913/2022
6. BA.2_hCoV-19/Mexico/MEX_InDRE_FB13911_E_S13150/2022
7. BA.2_hCoV-19/Mexico/SIN_CIAD-MZT_HJ6105/2022
8. XBB.1_hCoV-19/Mexico/TAM_LANGEBIO_IMSS_13849/2022
9. XBB.1_hCoV-19/Mexico/TAM_LANGEBIO_IMSS_13854/2022
10. XBB.1_hCoV-19/Mexico/VER-LANGEBIO_IMSS_13979/2023

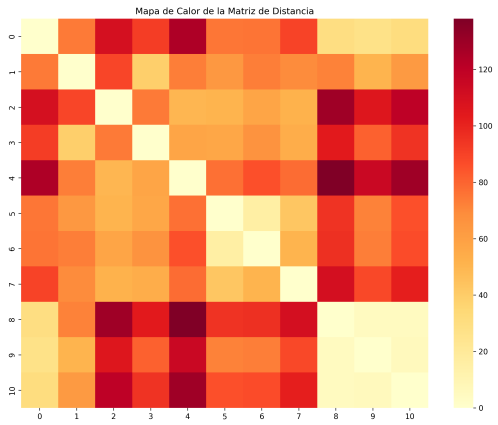

Figura 5.8: Mapa de calor de la matriz de distancias de nuestras secuencias de Sars-Cov-2

Para obtener más información sobre la recombinación que involucra a nuestras secuencias, se segmentó el genoma de cada muestra de modo que cada parte correspondiente a una proteína del virus SARS-CoV-2 quedara separada. Posteriormente, se comparó cada segmento por separado para obtener una matriz de distancia de cada proteína y, finalmente, se aplicó homología persistente a cada una de ellas. En la Figura 4.9 podemos visualizar el diagrama de persistencia de cada proteína, resaltando que únicamente en la proteína Spike (S) se obtuvo un objeto de dimensión 1. Esto sugiere que la transferencia horizontal ocurrió específicamente en esta proteína. Este resultado es respaldado por el artículo "The SARS-CoV-2 Omicron recombinant subvariants XBB, XBB.1, and XBB.1.5 are expanding rapidly with unique mutations, antibody evasion, and immune escape properties - an alarming global threat of a surge in COVID-19 cases again?"(3).

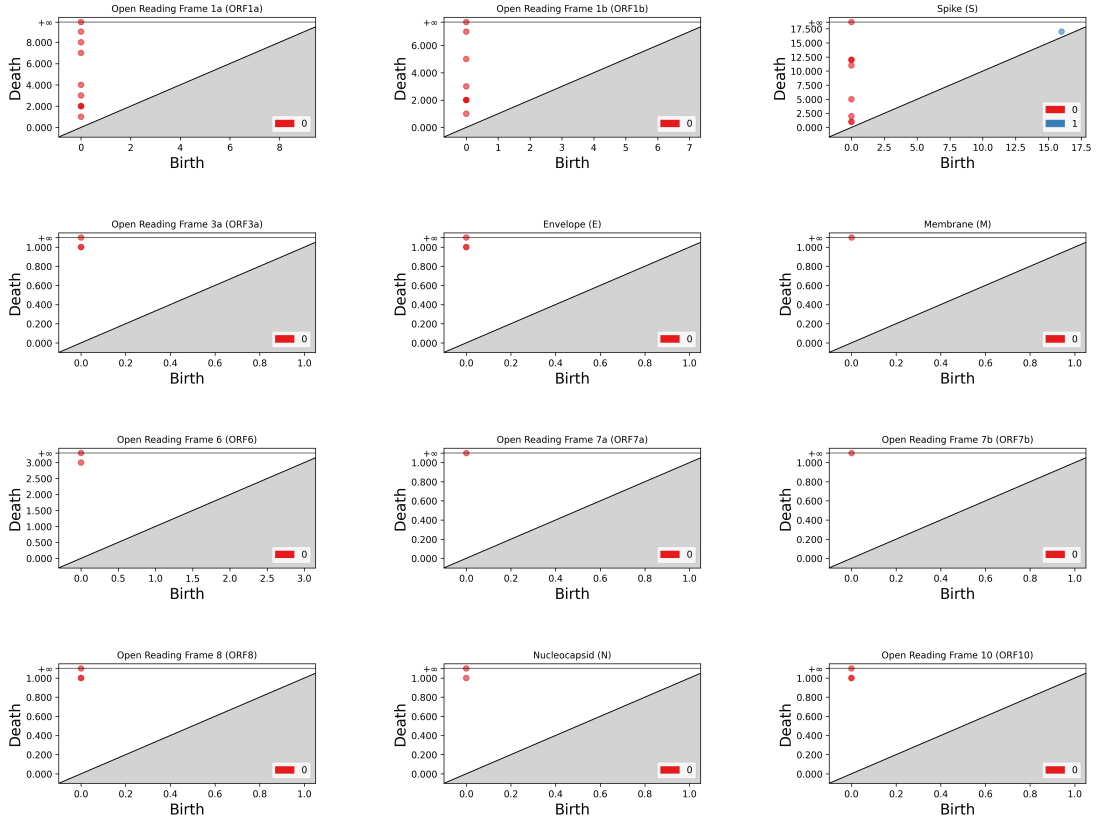

Figura 5.9: Diagrama de persistencia de cada uno de los segmentos que codifican proteínas en el virus Sars-Cov-2

Finalmente, se obtuvo el mapa de calor de la matriz de distancia de cada uno de los segmentos (Figura 4.10), permitiéndonos visualizar algunos datos interesantes. En primer lugar, podemos observar que las proteínas M, ORF7a y ORF7b son idénticas en todas nuestras muestras. Por otro lado, ORF6 solo difiere en la muestra BA.2.10.1_hCoV-19/Mexico/QUE-LANGEBIO_IM

Capítulo 6

Conclusiones

Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Etiam lobortis facilisis sem. Nullam nec mi et neque pharetra sollicitudin. Praesent imperdiet mi nec ante. Donec ullamcorper, felis non sodales commodo, lectus velit ultrices augue, a dignissim nibh lectus placerat pede. Vivamus nunc nunc, molestie ut, ultricies vel, semper in, velit. Ut porttitor. Praesent in sapien. Lorem ipsum dolor sit amet, consectetuer adipiscing elit. Duis fringilla tristique neque. Sed interdum libero ut metus. Pellentesque placerat. Nam rutrum augue a leo. Morbi sed elit sit amet ante lobortis sollicitudin. Praesent blandit blandit mauris. Praesent lectus tellus, aliquet aliquam, luctus a, egestas a, turpis. Mauris lacinia lorem sit amet ipsum. Nunc quis urna dictum turpis accumsan semper.

Apéndice A

Código/Manuales/Publicaciones

A.1. Apéndice

Apéndice

Bibliografía

[1] Asaad, A. and Jassim, S. (2017). Topological data analysis for image tampering detection. In Kraetzer, C., Shi, Y., Dittmann, J., and Kim, H., editors, *Digital Forensics and Watermarking*, volume 10431 of *Lecture Notes in Computer Science*, Cham. Springer. 12

[2] Centers for Disease Control and Prevention (2023). Tipos de virus de influenza. https://espanol.cdc.gov/flu/about/viruses/types.htm. Accessed: 28 February 2024. 6

[3] Chakraborty, C., Bhattacharya, M., Chopra, H., Islam, M. A., Saikumar, G., and Dhama, K. (2023). The sars-cov-2 omicron recombinant subvariants xbb, xbb.1, and xbb.1.5 are expanding rapidly with unique mutations, antibody evasion, and immune escape properties – an alarming global threat of a surge in covid-19 cases again? *International Journal of Surgery*, 109:1041–1043. 26

[4] Chan, J. M., Carlsson, G., and Rabadan, R. (2013). Topology of viral evolution. *Proceedings of the National Academy of Sciences*, 110(46):18566–18571. 12, 13, 14

[5] Dangi, T. and Jain, A. (2012). Influenza virus: A brief overview. *Proceedings of the National Academy of Sciences, India Section B: Biological Sciences*, 82(1):111–121. 6, 7

[6] Edelsbrunner, H. and Harer, J. (2010). *Computational Topology: An Introduction*. American Mathematical Society, Providence, Rhode Island. 12

[7] European Bioinformatics Institute (2023). Clustal omega - multiple sequence alignment (msa). https://www.ebi.ac.uk/jdispatcher/msa/clustalo. 16

[8] Flint, S. J., Enquist, L. W., Racaniello, V. R., Rall, G. F., and Skalka, A. M. (2015). *Principles of Virology*. ASM Press, Washington, DC, 4th edition. 2 Volume Set, Chapter 1. 5

[9] Fugacci, U., Scaramuccia, S., Iuricich, F., and De Floriani, L. (2016). Persistent homology: a step-by-step introduction for newcomers. *Smart Tools and Apps in computer Graphics (STAG)*. 12

[10] Gao, R., Cao, B., Hu, Y., Feng, Z., Wang, D., Hu, W., Chen, J., Jie, Z., Qiu, H., Xu, K., Xu, X., Lu, H., Zhu, W., Gao, Z., Xiang, N., Shen, Y., He, Z., Gu, Y., Zhang, Z., Yang, Y., Zhao, X., Zhou, L., Li, X., Zou, S., Zhang, Y., Li, X., Yang, L., Guo, J., Dong, J., Li, Q., Dong, L., Zhu, Y., Bai, T., Wang, S., Hao, P., Yang, W., Zhang, Y., Han, J., Yu, H., Li, D., Gao, G. F., Wu, G., Wang, Y., Yuan, Z., and Shu, Y. (2013). Human infection with a novel avian-origin influenza A (H7N9) virus. *New England Journal of Medicine*, 368(20):1888–1897. 8, 21

[11] Genome.gov (2024). Human genome reference sequence. https://www.genome.gov/genetics-glossary/Human-Genome-Reference-Sequence. Accessed: 28 February 2024. 6

[12] GISAID Initiative (2023). Gisaid - global initiative on sharing all influenza data. https://www.gisaid.org/. 6, 17

[13] Heydenreich, S., Brück, B., and Harnois-Déraps, J. (2021). Persistent homology in cosmic shear: Constraining parameters with topological data analysis. *A&A*, 648:A74. 12

[14] Hu, B., Guo, H., Zhou, P., and Shi, Z.-L. (2021). Characteristics of sars-cov-2 and covid-19. *Nature Reviews Microbiology*, 19(3):141–154. 8, 9

[15] ICTV (2011). Orthomyxoviridae. https://ictv.global/report_9th/RNAneg/Orthomyxoviridae. Accessed: 28 February 2024. 6

[16] Lai, M. M. (1992). Rna recombination in animal and plant viruses. *Microbiological Reviews*, 56(1):61–79. 9

[17] Li, F. (2016). Structure, function, and evolution of coronavirus spike proteins. *Annual Review of Virology*, 3(1):237–261. 9, 10

[18] Lowen, A. C. (2018). It's in the mix: Reassortment of segmented viral genomes. *PLoS Pathogens*, 14(7):e1007200. 8

[19] Madeira, F., Pearce, M., Tivey, A. R., et al. (2022). Search and sequence analysis tools services from embl-ebi in 2022. *Nucleic Acids Research*, 50(W1):W276–W279. 16

[20] McDonald, S. M., Nelson, M. I., Turner, P. E., and Patton, J. T. (2016). Reassortment in segmented rna viruses: mechanisms and outcomes. *Nature Reviews Microbiology*, 14(7):448–460. 7

[21] Munch, E. (2017). A user's guide to topological data analysis. *Journal of Learning Analytics*, 4(2):47–61. 11, 12

[22] Murphy, B. R. and Webster, R. G. (1996). *Orthomyxovirus*, pages 1397–1445. Lippincott-Raven Publishers. 8

[23] National Center for Biotechnology Information (2023). Ncbi. https://www.ncbi.nlm.nih.gov/. 5, 15

[24] Nextclade team (2024). Nextclade: clade assignment, mutation calling, and quality control for viral genomes. https://clades.nextstrain.org. Accesed: 14 July 2024. 17

[25] Nextstrain team (2024). Genomic epidemiology of sars-cov-2 with subsampling focused globally over the past 6 months. https://nextstrain.org/ncov/global. Accesed: 28 February 2024. 9

[26] P'erez-Losada, M., Arenas, M., Gal'an, J. C., Palero, F., and Gonz'alez-Candelas, F. (2015). Recombination in viruses: mechanisms, methods of study, and evolutionary consequences. *Infection, Genetics and Evolution*, 30:296–307. 9, 10

[27] Postol., M. S. (2023). *Algebraic Topology for Data Scientists*. The MITRE Corporation. 12, 13

[28] Rabadan, R. (2016). Inference of ancestral recombination graphs through topological data analysis. *PLOS Computational Biology*, 12(8):e1005071. 12, 13, 14

[29] Rabadan, R. and Blumberg, A. J. (2020). *Topological Data Analysis for Genomics and Evolution: Topology in Biology*. Cambridge University Press, Cambridge, United Kingdom. 12, 14

[30] Singh, Y., Farrelly, C., Hathaway, Q., et al. (2023). Topological data analysis in medical imaging: current state of the art. *Insights Imaging*, 14:58. 12

[31] Steel, J. and Lowen, A. C. (2014). Influenza a virus reassortment. *Current Topics in Microbiology and Immunology*, 385:377–401. 7

[32] Tauzin, G., Lupo, U., Tunstall, L., Pérez, J. B., Caorsi, M., Medina-Mardones, A. M., Dassatti, A., and Hess, K. (2021). giotto-tda: A topological data analysis toolkit for machine learning and data exploration. *Journal of Machine Learning Research*, 22(39):1–6. 16

[33] The GUDHI Project (2015). *GUDHI User and Reference Manual*. GUDHI Editorial Board. 16

[34] Thiry, E., Meurens, F., Muylkens, B., McVoy, M., Gogev, S., Thiry, J., Vanderplasschen, A., Epstein, A., Keil, G., and Schynts, F. (2005). Recombination in alphaherpesviruses. *Reviews in Medical Virology*, 15(2):89–103. 9

[35] Uribe, H. (2008). El virus influenza y la gripe aviar. 50(1). 7

[36] Wasserman, L. (2018). Topological data analysis. *Annual Review of Statistics and Its Application*, 5:501–532. 11, 12

[37] Worobey, M. and Holmes, E. C. (2011). Evolutionary aspects of recombination in rna viruses. *Journal of General Virology*, 92(4):751–761. 9, 10

[38] Wu, A., Su, C., Wang, D., Peng, Y., Liu, M., Hua, S., Li, T., Gao, G. F., Tang, H., Chen, J., Liu, X., Shu, Y., Peng, D., and Jiang, T. (2013). Sequential Reassortments Underlie Diverse Influenza H7N9 Genotypes in China. *Cell Host Microbe*, 14(4):446–452. 21, 24

[39] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., Tan, W., and China Novel Coronavirus Investigating and Research Team (2020a). A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 382(8):727–733. 6

[40] Zhu, N., Zhang, D., Wang, W., Li, X., Yang, B., Song, J., Zhao, X., Huang, B., Shi, W., Lu, R., Niu, P., Zhan, F., Ma, X., Wang, D., Xu, W., Wu, G., Gao, G. F., Tan, W., and China Novel Coronavirus Investigating and Research Team (2020b). A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*, 382(8):727–733. 9