UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
POSGRADO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

DETECCIÓN DE GENOMAS ANÓMALOS DE SARS-COV-2 POR MEDIO DE ALGORITMOS DE APRENDIZAJE NO SUPERVISADOS

TESIS
QUE PARA OPTAR POR EL GRADO DE
MAESTRO EN CIENCIA E INGENIERÍA DE LA COMPUTACIÓN

PRESENTA:
SERGIO ADRIAN MARTÍNEZ TENA

TUTOR O TUTORES PRINCIPALES
DR. JOSÉ ANTONIO NEME CASTILLO
INSTITUTO DE INVESTIGACIONES EN MATEMÁTICAS APLICADAS Y EN SISTEMAS

CIUDAD UNIVERSITARIA, CIUDAD DE MÉXICO. NOVIEMBRE, 2023

Resumen

En los últimos años, la pandemia de COVID-19 ha ocasionado una multitud de desafíos a nivel global. Hasta la fecha se han realizado numerosas investigaciones con el objetivo de caracterizar y comprender con mayor detalle el virus SARS-CoV-2, todo con la finalidad de desarrollar nuevas estrategias para enfrentar posibles infecciones futuras. Este esfuerzo ha dado lugar a la creación de bases de datos que contienen cientos de miles de genomas secuenciados del virus.

Asimismo, diversas herramientas computacionales han tenido un impacto significativo en la investigación de este virus, ya que facilitan el análisis de cientos de miles de secuencias virales. En este proceso, el aprendizaje computacional, el análisis de datos y la visualización de datos han desempeñado roles esenciales.

En particular, debido a la vasta cantidad de secuencias disponibles, identificar los genomas más anómalos del virus ha adquirido una gran importancia. Si bien estos genomas podrían indicar problemas en su almacenamiento o incluso errores en su secuenciación, desde una perspectiva biológica, estas anomalías también podrían estar relacionadas con mutaciones en el virus. Mutaciones que podrían ser causantes de acelerar el ritmo de transmisión y/o gravedad de la infección.

Este trabajo tiene como objetivo emplear diversos métodos computacionales con el fin de detectar los genomas más inusuales secuenciados en México desde el 1 de enero de 2020 hasta el 16 de junio de 2022. Para ello, se emplean herramientas computacionales como el cálculo de distancias, el cálculo de la entropía, la reducción de dimensionalidad y diversos algoritmos de detección de anomalías.

Con todo esto, lo que se pretende es proporcionar a los especialistas una herramienta adicional que les permita realizar análisis posteriores más detallados sobre el virus SARS-CoV-2.

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

En los últimos años, el uso de herramientas de aprendizaje computacional y del análisis de datos han experimentado un aumento significativo. Una de las principales razones de este incremento se debe a la gran cantidad de datos que se generan a diario en diversos campos de estudio [1], lo cual ha motivado a numerosos procesos de investigación a adoptar este tipo de herramientas con el propósito de obtener una mejor comprensión del objeto de estudio.

El uso de estas herramientas, que incluyen algoritmos, técnicas y descripciones, han permitido obtener descubrimientos o conclusiones que se derivan únicamente del conjunto de datos bajo análisis. Uno de los enfoques del aprendizaje computacional que ha tomado gran relevancia en estos procesos de investigación es el aprendizaje no supervisado. Esta familia de algoritmos son capaces de encontrar patrones inherentes en los conjuntos de datos, mostrando ser de gran utilidad en el análisis de estos. Entre las aplicaciones más destacadas de este tipo de algoritmos se incluyen la reducción de dimensionalidad, el agrupamiento de datos y la detección de elementos anómalos.

Del mismo modo, el interés por la detección de anomalías ha ido en aumento, debido a que detectar este tipo de instancias tiene mucha utilidad en diferentes campos de estudio, tal como en medicina [2], finanzas [3], ciberseguridad [4], entre otros [5, p. 7–18]. Con frecuencia, un valor atípico en los datos puede ser un indicador de un error o de algún comportamiento problemático. Por ejemplo, en datos recolectados de signos vitales de una persona, encontrar instancias anómalas podría señalar algún padecimiento médico. Sin embargo, a menudo estos valores solo indican un fenómeno que no es entendido del todo, hecho que llama a expertos en el campo de estudio en cuestión a tratar de entender estos sucesos.

Uno de los muchos campos de estudio que se ha beneficiado con diversas técnicas de detección de anomalías y del análisis de datos es la biología, específicamente la genómica [6, 7]. La capacidad de identificar regiones atípicas dentro de un genoma o genomas anómalos en un conjunto de múltiples genomas se ha

convertido en una herramienta de gran utilidad en esta área.

Una vez establecido lo anterior, este documento explora la aplicación de diversos algoritmos y técnicas de detección de anomalías en un conjunto de genomas secuenciados del virus SARS-CoV-2. A lo largo de este texto, mostraremos cómo fue posible identificar aquellos genomas con mayores discrepancias respecto al resto. Del mismo modo, se emplearán diversas herramientas computacionales con el fin de analizar la evolución de ciertas características del genoma del virus a lo largo del tiempo, así como las diferencias encontradas en distintas zonas dentro de México. Con los métodos utilizados en este trabajo y los resultados que describimos, brindamos a los especialistas una herramienta adicional que les permitirá llevar a cabo análisis posteriores y más detallados sobre el tema.

1.1. Planteamiento del problema

En la actualidad, gracias al desarrollo de nuevas tecnología y al creciente interés sobre el tema en las últimas décadas, obtener secuencias de genomas completos para diferentes organismos se ha vuelto mucho más rápido y económico [8]. Esto ha permitido que la cantidad de genomas secuenciados este aumentado de manera muy acelerada, a tal grado que a día de hoy existen bases de datos con cientos de miles de genomas secuenciados para diferentes organismos, incluyendo virus. Por otro lado, examinar estos grandes conjuntos de datos se ha facilitado gracias a muchas técnicas de aprendizaje computacional y del análisis de datos.

¿Pero qué es un genoma? Un genoma es toda la información genética contenida en el ADN de un organismo (ARN en algunos virus). El ADN consiste en una secuencia de nucleótidos (*adenina*, *guanina*, *citosina* y *tiamina* o *uracilo* en el ARN), los cuales contienen instrucciones para el funcionamiento, desarrollo y reproducción de un organismo o virus [9, 10]. Muchos análisis realizados sobre conjuntos de estas secuencias han contribuido en el entendimiento de los organismos. Por ejemplo, encontrando una gran correlación entre la composición genómica del algunos virus con su respectivo anfitrión [11]. Sin embargo, dentro de una especie, la mayor parte del genoma es similar entre individuos, por lo cual para lograr comprender su diversidad genética y evolución es necesario obtener y analizar secuencias completas de varios individuos a través del tiempo.

Durante la pandemia causada por el COVID-19, se han secuenciado una gran cantidad de genomas, lo que ha dado lugar a la generación de conjuntos de datos con centenas de miles de secuencias completas del virus SARS-CoV-2. En particular, la iniciativa GISAID [12] proporciona repositorios con datos públicos sobre estas secuencias genéticas. El conjunto de estos datos, además de contener las secuencias completas del virus contienen detalles clínicos (fecha de secuenciación,

estado de procedencia, institución que realizó la secuenciación, entre otros), hecho que permite realizar un análisis aún más detallado sobre el conjunto de datos.

Una de las representaciones del genoma que ha sido de gran utilidad para realizar diferentes análisis es el uso de codones. Esta representación nos permite visualizar el genoma completo como la frecuencia de ocurrencia de codones en lugar de toda la cadena de nucleótidos. Un codón es una secuencia de 3 nucleótidos consecutivos, los cuales codifican a un aminoácido (componente básico de las proteínas) o marcan el término en la producción de una proteína. Existen 64 codones (formados por las combinaciones posibles de nucleótidos), 61 de estos codifican aminoácidos y 3 son codones de paro [13]. De esta forma, el uso de codones nos permite caracterizar los genomas completos con solo 64 variables.

Así mismo, hoy en día existen muchas técnicas computacionales para la detección de anomalías que se han desarrollado a través de los años. Estas han probado ser de gran utilidad en diversas ramas de la ciencia, no solo con el fin de remover valores atípicos del conjunto de datos a trabajar, sino que también proporcionan la capacidad de identificar instancias que se pueden relacionar con sucesos sospechosos o significativos. Concretamente, encontrar secuencias atípicas dentro del conjunto de genomas del SARS-CoV-2 podría proporcionar información valiosa, como relacionarse con algún error durante la secuenciación, ser un indicativo de alguna mutación en el virus, u otros muchos factores.

En líneas generales, se tiene como objetivo construir diferentes modelos capaces de asignar cierto puntaje sobre el nivel de anomalía a cada elemento, luego las anomalías se detectan de acuerdo a un umbral previamente definido. Aunque existen algoritmos de aprendizaje supervisado y semi-supervisado, en esta tesis se busca aplicar la detección de anomalías a datos sin etiquetar (al conjunto de secuencias de SARS-CoV-2), por lo que se recurre a algoritmos de aprendizaje no supervisado. Aunado a esto, la aplicación de diferentes técnicas de visualización y de reducción de la dimensionalidad nos permiten tener una perspectiva mas general sobre la variación de todas las secuencias en el conjunto de datos.

1.2. Objetivo General

El objetivo de este trabajo consiste en emplear diversos métodos computacionales de aprendizaje no supervisado, capaces de determinar el grado de anomalía presente en cada una de las representaciones de los genomas del virus SARS-CoV-2 secuenciados en México, desde el 1 de enero de 2020 hasta el 16 de junio de 2022. Asimismo, utilizar distintas técnicas de visualización con el fin de observar las variaciones en este grado de anomalía. Todo esto con el propósito de detectar aquellos genomas que presenten un mayor nivel de anomalía.

1.3. Objetivos Específicos

- Obtener, limpiar, pre-procesar y extraer el uso de codones de los genomas secuenciados del SARS-CoV-2 junto con sus metadatos.
- Comparar cada representación del virus según diferentes herramientas elegidas, por región y por fecha.
- Aplicar diferentes algoritmos de aprendizaje no supervisados para la detección de anomalías a todo el conjunto de datos.
- Utilizar diferentes técnicas de visualización a fin obtener una perspectiva general de los resultados.

1.4. Hipótesis

Es posible utilizar métodos del aprendizaje computacional no supervisado con la finalidad de diferenciar aquellos genomas que divergen más en un conglomerado de genomas secuenciados para un mismo virus, considerando sus usos de codones.

1.5. Antecedentes

Un análisis detallado sobre la representación del uso de codones del SARS-CoV-2 se muestra en *Dilucca et al.* [14]. En este trabajo se tiene como propósito analizar la evolución del virus y compararlo con otros virus pertenecientes a la misma subfamilia. Para ello emplearon varias medidas del *sesgo en el uso de codones* (como, el uso relativo de codones sinónimos, el índice de adaptación de codones, entre otros) a fin de caracterizar cada uno de los genomas utilizados en el estudio. Asimismo, emplearon diferentes métodos estadísticos (como la unidad tipificada y la regresión lineal) y de visualización de datos (como mapas de calor y gráficos de violín) para llevar a cabo todo el análisis. No obstante, esta investigación se realizó durante los primeros meses de la pandemia, por lo cual algunos resultados que se presentan difieren de lo obtenido en trabajos posteriores.

En *Hou* [15] se presenta otro estudio realizado al SARS-CoV-2. En este buscan identificar la arquitectura usual del uso de codones del virus, así como también compararlo con otro grupo de coronavirus. La investigación también emplea diferentes medidas para caracterizar el uso de codones de cada virus a estudiar. Los

métodos estadísticos y de visualización utilizados son semejantes a los de *Dilucca et al.* [14]. Este trabajo revela que el patrón del uso de codones del SARS-CoV-2 presenta pequeñas variaciones según la región geográfica de donde proviene.

En *Maldonado et al.* [16] comparan los genes altamente expresados en el tejido pulmonar humano y algunos genes específicos de varios virus, incluido el SARS-CoV-2. En la investigación encuentran una similitud entre los genes altamente expresados del tejido y los genes de los virus que infectan a humanos. Por lo cual, proponen que este resultado podría ser un indicativo de como el virus SARS-CoV-2 mutó para adaptarse al uso de codones del humano. El trabajo emplea métodos como el análisis de agrupamientos, el análisis de componentes principales y el cálculo de diferentes índices (medidas) del uso de codones (de manera similar a *Dilucca et al.* [14] y *Hou* [15]).

La investigación que se presenta en *Simón et al.* [11] tiene como finalidad obtener una visión general de la composición de alrededor de 10,000 especies de virus, para este fin hacen uso de diferentes representaciones del genoma. Algunas de las representaciones utilizadas son, por ejemplo, la composición completa de nucleótidos, el contenido de GC (guanina y citosina) y el uso de codones. Teniendo como finalidad emplear diferentes técnicas estadísticas sobre estas representaciones para comparar las diferentes especies de virus entre sí. Adicionalmente, calculan la correlación que existe entre algunos virus y su anfitrión. Encontrando que el sesgo en uso de codones de algunos virus está muy correlacionado al de sus anfitriones.

Un trabajo que se ha llevado a cabo sobre un gran número de secuencias del SARS-CoV-2 es el de *Posani et al.* [17]. En este analizan la evolución del uso de codones del virus. Sin embargo, en lugar de utilizar el genoma completo del virus para el análisis, en el trabajo solo emplean específicamente la composición de seis genes, cuatro de los cuales codifican proteínas estructurales del virus. Esta investigación, de manera similar a *Dilucca et al.* [14], *Hou* [15] y *Maldonado et al.* [16], emplea diferentes medidas del sesgo del uso de codones a fin de caracterizar cada gen del conjunto total. Donde como resultado, se muestra como el genoma acumula continuamente mutaciones con respecto al primer virus secuenciado y como con el paso del tiempo el uso de codones de estos genes se hace menos eficiente con respecto al uso de codones del humano.

En *Hahn et al.* [18], se ha investigado la relación entre la detección de elementos atípicos y ciertos genomas de interés. En este estudio, se propone el uso de diferentes herramientas computacionales para identificar variantes emergentes del SARS-CoV-2. Para ello, obtienen un índice de similitud entre las secuencias de nucleótidos completas y, empleando el análisis de componentes principales, localizan aquellas instancias más inconsistentes con el resto, según diferentes criterios. El estudio revela que la mayoría de las anomalías detectadas corresponden a genomas de variantes comunes del virus, como *alpha*, *delta*, *ómicron*, entre otras.

Finalmente, es importante destacar que tanto *Dilucca et al.* [14], *Posani et*

al. [17] y *Hahn et al.* [18] han empleado los datos recopilados por la iniciativa GISAID [12] para obtener sus resultados. Esto hace que GISAID sea una opción viable para adquirir los datos a analizar.

1.6. Metodología

Las secuencias utilizadas para el análisis se obtuvieron de las bases de datos proporcionadas por GISAID [12]. Específicamente, se adquirieron secuencias únicamente de la República Mexicana durante el período del 1 de enero de 2020 al 16 de junio de 2022. Estas secuencias se preprocesaron y limpiaron, obteniendo los genomas completos de cada virus como su uso de codones.

Como parte integral del análisis, se comparó el uso de codones de cada virus con respecto al uso de codones humano, empleando diferentes medidas. La información sobre el uso de codones humano se obtuvo de *Nakamura et al.* [19].

Inicialmente se calcularon las siguientes medidas para cada genoma del virus: su *distancia euclidiana* al genoma humano, su *distancia de Wasserstein* al genoma humano, la *entropía* de su uso de codones y su *distancia log-cociente esperada*. Luego, se identificaron los genomas que mostraron una mayor divergencia en estas medidas. Agrupando los resultados por estado, se generaron algunas gráficas de dispersión que permitieron identificar aquellos estados donde se encuentran los genomas virales más divergentes. Estas medidas también se utilizaron para analizar la evolución del virus mediante la creación de series de tiempo.

Posteriormente, se aplicaron dos técnicas de reducción de la dimensionalidad con dos enfoques diferentes: el *análisis de componentes principales* (*PCA*) y *el mapeo isométrico* (*Isomap*). A partir de estas proyecciones obtenidas, se calculó la distancia promedio entre cada genoma viral y los demás, lo que permitió identificar aquellos elementos que más divergen en estos nuevos subespacios.

Los algoritmos empleados para la detección de anomalías fueron los siguientes: el algoritmo *DBSCAN*, el algoritmo de *k-medias* para la detección de anomalías, el algoritmo del *valor atípico local* (*LOF*) y el algoritmo del *bosque de aislamiento*. En este caso, todos los algoritmos, a excepción de *DBSCAN*, calculan puntuaciones de anomalías. Por lo tanto, se define un umbral para determinar cuáles elementos son los más anómalos. Una vez obtenidas estas instancias atípicas, se visualizan los resultados según su estado y fecha de registro.

El algoritmo de *k-medias* para la detección de anomalías no se aplicó al conjunto de datos completo, sino que se aplicó a varios subconjuntos del mismo. Específicamente, se utilizaron dos criterios para crear los subconjuntos a partir del conjunto de datos completo: el estado de registro de los genomas y los periodos de tiempo en los que estos se registraron.

Finalmente, los genomas anómalos identificados por todos los criterios mencionados se presentaron en una única figura, en la cual se muestra su estado y fecha de registro. Asimismo, se incluye una tabla con la información de los genomas identificados con mayor frecuencia como anomalías.

1.7. Contribución y relevancia

Una de las razones principales para desarrollar un software destinado a detectar anomalías en este conjunto de genomas, es la de encontrar alguna relación entre estas instancias atípicas y la presencia de eventos interesantes, peculiares o sospechosos. Por ejemplo, desde un enfoque informático, este tipo de instancias podrían señalar algún error en el almacenamiento de los datos, o hasta un falló durante la secuenciación del genoma. Igualmente, desde la perspectiva de la biología, estos elementos incluso podría estar asociados con alguna mutación en el virus, un evento muy relevante ya que con frecuencia estas mutaciones provocan que los virus desarrollen nuevas adaptaciones a distintos entornos y condiciones. Un ejemplo concreto se muestra en *Hahn et al.* [18], donde se relaciona la detección de estas instancias con el surgimiento de nuevas variantes del virus SARS-CoV-2.

Además, un software con la capacidad de analizar la presencia de anomalías a medida que el virus evoluciona y de diferenciar los elementos por regiones geográficas, apoyándose también en diferentes técnicas de visualización de datos, podría ser una herramienta de gran ayuda para los especialistas, contribuyendo así a una mejor comprensión del comportamiento del virus SARS-CoV-2.

1.8. Estructura de la tesis

En el Capítulo 1, se aborda la introducción del tema que se tratará en la tesis. Se presenta el planteamiento del problema, el objetivo principal y los objetivos específicos que se buscan alcanzar con este trabajo. Además, se analizan estudios previos que trataron temas similares, se describe brevemente la metodología que se empleará y se resalta la principal contribución de este trabajo.

El Capítulo 2 presenta algunos aspectos generales del virus SARS-CoV-2, como la composición de su genoma viral y la descripción de algunos de sus genes principales. Asimismo, se aborda de manera concisa la estructura del virus, el proceso de replicación y su variabilidad genómica. Este capítulo también introduce el concepto del *uso de codones*, que será la representación empleada para analizar los genomas del virus a lo largo de toda la tesis.

En el Capítulo 3, se presentan algunos conceptos, herramientas y algoritmos empleados para analizar los *usos de codones* de los genomas del SARS-CoV-2. En este capítulo se detallan algunas características de los *datos composicionales*, las métricas de distancia utilizadas, el algoritmo de *k-medias*, el concepto de *entropía* en una distribución de probabilidad y las técnicas de reducción de la dimensionalidad aplicadas: *PCA* e *Isomap*.

El Capítulo 4 describe el problema de la detección de anomalías. En este capítulo, se ofrece una breve descripción de los diversos tipos de anomalías que pueden surgir, así como una exposición de las categorías de algoritmos de aprendizaje no supervisado para la detección. A continuación, se explican los algoritmos que se utilizan en esta tesis para detectar los genomas anómalos del SARS-CoV-2. Estos incluyen el algoritmo *DBSCAN*, el algoritmo *k-medias* para la detección de anomalías, el algoritmo del *valor atípico local* (*LOF*) y el algoritmo del *bosque de aislamiento*. Adicionalmente, se mencionan otras estrategias que emplean algunas de las herramientas presentadas en el Capítulo 3 para detectar anomalías.

En el Capítulo 5 se presentan los resultados obtenidos mediante la metodología descrita previamente. Se examina en detalle el procedimiento utilizado para adquirir, preprocesar y limpiar el conjunto de datos, así como los criterios empleados para identificar las instancias más anómalas dentro de dicho conjunto. Por último, se muestra la descripción de los genomas que fueron identificados con mayor frecuencia como atípicos según los diferentes criterios seleccionados.

Finalmente en el Capítulo 6 se dan las conclusiones generales de este trabajo de tesis, además de abordar de manera concisa el posible trabajo a futuro.

Capítulo 2

SARS-CoV-2

Los coronavirus son una subfamilia de virus que infecta tanto animales como a humanos, causando múltiples enfermedades con amplia variedad de síntomas. Según su relación genética y estructura genómica, se clasifican en cuatro géneros: *Alphacoronavirus*, *Betacoronavirus*, *Gammacoronavirus* y *Deltacoronavirus*. Los primeros dos géneros infectan sólo a mamíferos, mientras que los últimos dos pueden infectar tanto a aves como a mamíferos [20, 21].

Desde la aparición de cepas como el SARS-CoV (China 2002) y el MERS-CoV (Medio Oriente 2012), se han llevado a cabo extensas investigaciones acerca de los diferentes tipos de coronavirus que afectan a los seres humanos. Encontrando, por ejemplo, que muchos de estos virus probablemente se originaron en murciélagos y roedores, y que su transmisión a los humanos ocurre posiblemente a través de un hospedero intermediario (otro animal) [20].

A finales del 2019, en China, fue identificada una nueva cepa de coronavirus, el cual ahora conocemos como SARS-CoV-2. El virus desde entonces se ha propagado rápidamente por todo el mundo y en marzo del 2020 la enfermedad fue declarada como pandemia por la Organización Mundial de la Salud (OMS) [22], a la fecha de este documento esta enfermedad continúa afectando a nivel global. Este virus ha presentado desafíos muy complejos para toda la población, desde ser capaz de causar que los individuos enfermen de gravedad, hasta provocar todo tipo de restricciones sociales y medidas preventivas, debido a su rápida propagación. Estos hechos han cambiado de manera significativa la vida de millones de personas en todo el mundo.

Asimismo, debido al surgimiento de este virus, se han llevado a cabo numerosas investigaciones con el propósito de caracterizarlo y comprenderlo con más detalle, a fin de desarrollar nuevas estrategias para combatir las posibles infecciones. Por ejemplo, se ha estudiando su composición genómica [15, 23], sus propiedades estructurales y ciclo de replicación [21], la evolución del genoma [17], la emergencia de diferentes variantes [24], entre otros aspectos. Gran parte de es-

tos estudios han sido posibles gracias a diferentes técnicas en biología, genómica y bioinformática. Igualmente, el uso de diferentes técnicas de análisis de datos ha probado ser fundamental gracias a los miles de genomas secuenciados del virus desde su aparición.

En este capítulo se explora con mayor detalle el virus SARS-CoV-2. En primer lugar, se aborda su genoma, describiendo su composición genética característica, su uso de codones, la organización del genoma y algunos de sus genes cruciales. Posteriormente, se analiza brevemente la estructura del virus, el funcionamiento de las proteínas esenciales, el ciclo de replicación y algunas de las mutaciones que pueden llegar a ocurrir. Por último, se considera la variabilidad genómica, explorando lo que es una variante del virus, así como su evolución y adaptación.

2.1. Genoma del virus

El genoma es toda la información genética que posee un organismo o virus. En organismos multicelulares, esta información consiste en secuencias de ADN, mientras que en algunos virus este se compone por secuencias de ARN [9]. Estas secuencias, a su vez, están formadas por cuatro nucleótidos: *adenina* (A), *guanina* (G), *citosina* (C) y *tiamina* (T) o *uracilo* (U) en el caso del ARN.

En general, el genoma de los virus es mucho más reducido y simple que el de las células procariotas y eucariotas [25]. Donde normalmente el genoma viral consiste solo de un pequeño conjunto de genes con la finalidad de sintetizar proteínas esenciales para su replicación. Debido a que los virus carecen de la capacidad de replicarse por sí mismos, éstos necesitan infectar una célula huésped y utilizar la maquinaria celular de la misma para replicarse [26].

Existen diferentes tipos de genomas virales que se clasifican según su composición de ácido nucleico (ARN o ADN), su tipo de cadena y el sentido de la misma. Cada tipo de genoma viral tiene características únicas que afectan su forma de replicación y su interacción con las células del huésped [27].

La composición genómica típica de los coronavirus consiste en ARN monocatenario positivo ($ARNmc+$), con una longitud que oscila entre 27 mil y 30 mil nucleótidos [20]. Alrededor de dos tercios del genoma codifican proteínas no estructurales, esenciales para la replicación y transcripción del ARN viral dentro del huésped, mientras que el tercio restante del genoma codifica a las proteínas estructurales, responsables de la forma característica de este tipo de virus, y a las proteínas accesorias, importantes durante la infección inicial [28].

Específicamente, el virus SARS-CoV-2 pertenece al género Betacoronavirus. Su genoma consiste de alrededor de 30 mil nucleótidos, donde el valor composicional más alto es el del *uracilo* ($\sim 32.2\,\%$), seguido por la *adenina* ($\sim 29.9\,\%$),

citosina ($\sim 19.6\,\%$) y *guanina* ($\sim 18.3\,\%$) [15]. Por lo que el contenido de GC es de alrededor del 38 %, similar a otros coronavirus del mismo género.

Una de las características fundamentales de cualquier secuencia genética es su uso de codones. Como se sabe, los codones son secuencias de tres nucleótidos que codifican un aminoácido específico durante la síntesis de proteínas. Gracias a la existencia de cuatro nucleótidos diferentes, se puede generar un total de 64 codones. De estos, 3 son codones de paro, los cuales marcan el fin de la síntesis de una proteína, mientras que los otros 61 codifican a 20 aminoácidos (Tabla 2.1) [13, 29]. La Figura 2.1 muestra el uso de codones característico del virus SARS-CoV-2, obtenido a partir de la secuencia de referencia de GISAID [12].

Primer nucleótido de codón	Segundo nucleótido de codón: U		C		A		G		Tercer nucleótido de codón
U	**UUU**	*Phe*	**UCU**	*Ser*	**UAU**	*Tyr*	**UGU**	*Cys*	**U**
	UUC	*Phe*	**UCC**	*Ser*	**UAC**	*Tyr*	**UGC**	*Cys*	**C**
	UUA	*Leu*	**UCA**	*Ser*	**UAA**	*PARO*	**UGA**	*PARO*	**A**
	UUG	*Leu*	**UCG**	*Ser*	**UAG**	*PARO*	**UGG**	*Trp*	**G**
C	**CUU**	*Leu*	**CCU**	*Pro*	**CAU**	*His*	**CGU**	*Arg*	**U**
	CUC	*Leu*	**CCC**	*Pro*	**CAC**	*His*	**CGC**	*Arg*	**C**
	CUA	*Leu*	**CCA**	*Pro*	**CAA**	*Gln*	**CGA**	*Arg*	**A**
	CUG	*Leu*	**CCG**	*Pro*	**CAG**	*Gln*	**CGG**	*Arg*	**G**
A	**AUU**	*Ile*	**ACU**	*Thr*	**AAU**	*Asn*	**AGU**	*Ser*	**U**
	AUC	*Ile*	**ACC**	*Thr*	**AAC**	*Asn*	**AGC**	*Ser*	**C**
	AUA	*Ile*	**ACA**	*Thr*	**AAA**	*Lys*	**AGA**	*Arg*	**A**
	AUG	*Met*	**ACG**	*Thr*	**AAG**	*Lys*	**AGG**	*Arg*	**G**
G	**GUU**	*Val*	**GCU**	*Ala*	**GAU**	*Asp*	**GGU**	*Gly*	**U**
	GUC	*Val*	**GCC**	*Ala*	**GAC**	*Asp*	**GGC**	*Gly*	**C**
	GUA	*Val*	**GCA**	*Ala*	**GAA**	*Glu*	**GGA**	*Gly*	**A**
	GUG	*Val*	**GCG**	*Ala*	**GAG**	*Glu*	**GGG**	*Gly*	**G**

Tabla 2.1: Traducción de cada triplete o codón a su aminoácido respectivo [13].

Debido a que existen más codones que aminoácidos, múltiples codones pueden codificar a un mismo aminoácidos, a excepción de la *metionina* (*Met*) y el *triptófano* (Trp), que son codificados por un solo codón. La diferencia en la frecuencia de ocurrencia de estos codones sinónimos se conoce como *sesgo en el uso de codones*. Es decir, algunos codones sinónimos son utilizados con mayor frecuencia que otros para codificar los mismos aminoácidos [30]. La Figura 2.2 presenta la frecuencia de ocurrencia de los diferentes codones sinónimos para la secuencia del virus SARS-CoV-2 de GISAID [12].

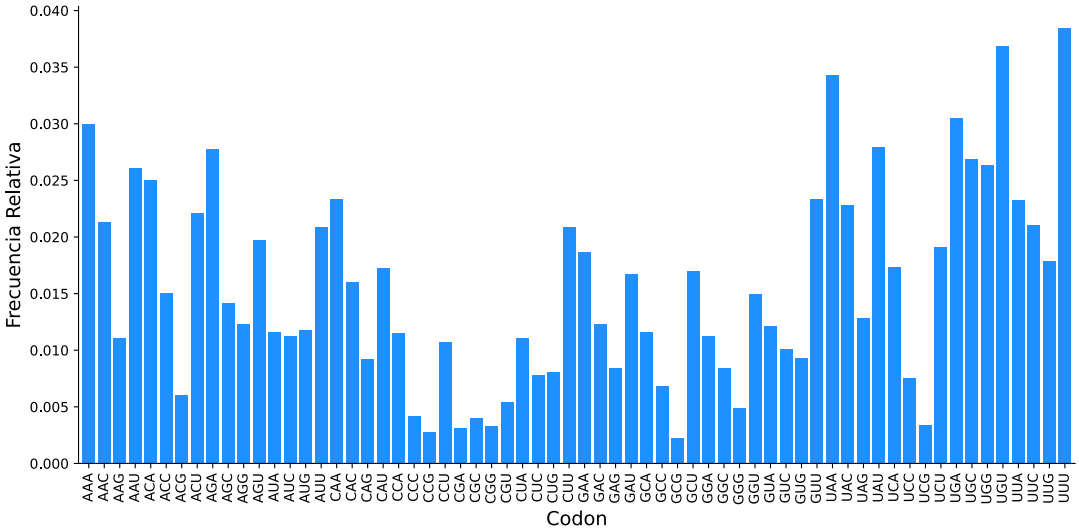

Figura 2.1: Uso de codones característico del SARS-CoV-2, representado como frecuencia relativa. Obtenido a partir del genoma de referencia de GISAID [12].

Análisis más detallados sobre uso de codones del virus revelan una preferencia para los codones sinónimos que terminan en A/U. Igualmente, muestran que el codón con mayor frecuencia y menor frecuencia, en comparación a sus codones sinónimos, es el codón *AGA* (para el aminoácido *Arginina*) y el codón UCG (para el aminoácido *Serina*). A pesar de esto, el virus presenta un *sesgo en el uso de codones* relativamente bajo, lo que significa que en promedio no hay mucha preferencia por un codón sinónimo en particular para codificar un aminoácido. Por otro lado, también se ha encontrado que el uso general de codones del SARS-CoV-2 es similar a otros coronavirus no humanos cercanos filogenéticamenete, lo que podría indicar una relación evolutiva entre ellos [15].

El genoma del SARS-CoV-2 está compuesto por diversas secciones que contienen información para la síntesis de proteínas específicas (Fig. 2.3). En el extremo izquierdo de la secuencia se encuentra el fragmento más grande (*1ab*), ocupando alrededor de 20 mil nucleótidos del genoma. Este codifica 16 proteínas no estructurales, encargadas de la replicación, transcripción y corrección del ARN viral. El resto del genoma (alrededor de 10 mil nucleótidos) codifica a cuatro proteínas estructurales y a un conjunto de proteínas accesorias. Las proteínas estructurales, al igual que en muchos otros coronavirus, son la espiga (S), membrana (M), envoltura(E) y nucleocápside (N). Las proteínas accesorias, las cuales contribuyen a modular la respuesta del huésped a la infección, son codificadas por las secciones *3a, 3b, 6, 7a, 7b, 8, 9a y 9b*. Este conjunto de proteínas accesorias es sumamente diverso según el tipo de coronavirus [21, 23].

Figura 2.2: Uso de codones sinónimos del SARS-CoV-2. Las barras de color verde marcan el codón más utilizado para codificar el aminoácido correspondiente. No se muestra los codones de paro ni los aminoácidos *metionina* y *triptófano*.

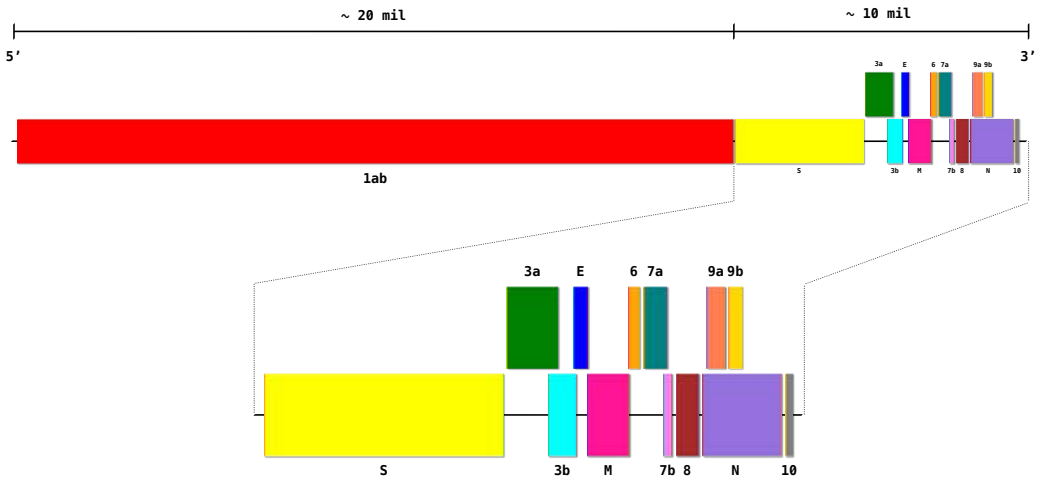

Figura 2.3: Organización del genoma del SARS-CoV-2. Cada sección codificante del genoma está representada por un recuadro de color, el fragmento *1ab* marca el conjunto proteínas no estructurales, las proteínas estructurales son codificadas por las secciones *S, E, M, y N*. Por último, las proteínas accesorias son codificadas por las secciones *3a, 3b, 6, 7a, 7b, 8, 9a y 9b*.

En *V'kovski et al.* [21] obtienen la similitud entre cada una de las secciones mencionadas anteriormente y sus homólogas en los virus RaTG13, RmYN02, MP789, P1E, ZC45, ZXC21 y SARS-CoV (coronavirus representativos relacionado con el síndrome respiratorio agudo grave). Esto revela que el gen que codifica la proteína espiga (S), es el que muestra una menor similitud entre el SARS-CoV-2 y los otros coronavirus. Este hecho es significativo ya que este gen es uno de los más críticos en la transmisión del virus de animales a humanos [21].

2.2. Estructura del virus

La mayoría de los virus comparten ciertos aspectos en común. Por ejemplo, tanto los virus de *ADN* como los de *ARN*, presentan una capa proteica que protege su material genético. Igualmente, algunos virus tienen una envoltura alrededor de dicha capa [26]. Sin embargo, la estructura característica de cada virus puede variar en un amplio rango, desde su tamaño y forma, hasta incluir elementos únicos, como la integración de proteínas específicas dentro del virión (partícula

completa del virus). Esta amplia variedad de diferencias impacta en la capacidad del virus al momento de infectar a las células del huésped y en la eficacia durante la replicación viral una vez dentro de estas células [27]. Por esta razón, comprender la estructura viral es esencial para entender el ciclo de vida del virus.

La estructura del virus SARS-CoV-2 es muy similar a otros coronavirus, donde el tamaño de la partícula llega a ser entre 60 y 140 nm de diámetro [31]. El virión se compone de cuatro proteínas estructurales: la proteína espiga (S), envoltura (E), membrana (M) y nucleocápside (N). La Figura 2.4 muestra cómo se organizan estas proteínas para formar el virión. En general, como en el caso de muchos coronavirus, la proteína N encapsula y protege el ARN, mientras que las proteínas S, M y E forman la envoltura del virus. Más específicamente, las proteínas M y E son las encargadas de integrar la partícula viral en el ensamblaje de nuevos viriones (durante el proceso de replicación viral). La proteína S, además de dar la apariencia de una corona solar a las partículas, es la que permite al virus enlazarse con los receptores de entrada en las células del huésped [21].

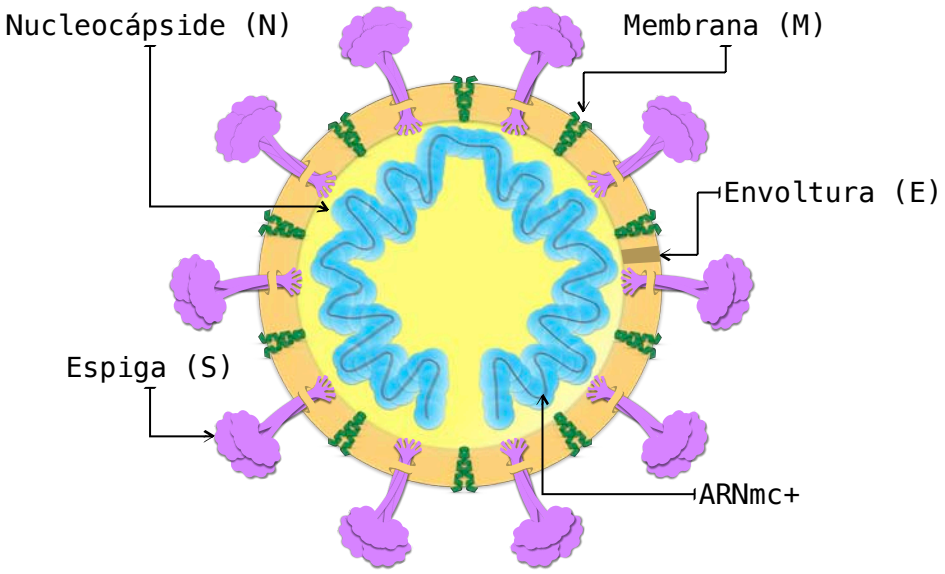

Figura 2.4: Estructura del SARS-CoV-2. La proteína N encapsula y protege el *ARNmc+*, mientras que las proteínas S, M y E forman la envoltura del virus

Después de que la espiga del virus se enlaza a la célula, éste penetra en ella. Una vez dentro, el virus inicia su proceso de replicación. En pocas palabras, primero se libera su material genético en el interior de la célula, luego se comienza a sintetizar las proteínas codificadas en este genoma, es decir, se expresa el ARN

del virus dentro de la célula. Posteriormente, éste conjunto de proteínas, tanto estructurales como no estructurales, se encarga de replicar el genoma viral e incorporarlo en nuevas partículas virales. Finalmente, estos nuevos viriones son expulsados de la célula para repetir el ciclo descrito en otras células [21].

Durante el proceso de replicación del genoma viral, pueden producirse alteraciones en las secuencias de nucleótidos de las nuevas copias, lo que se conoce como mutaciones [32]. Con el paso del tiempo, estas mutaciones pueden causar cambios en los genes que sintetizan las proteínas estructurales del virus. Estas proteínas, a su vez, pueden afectar la forma en que el virus interactúa con las células del huésped, lo que puede disminuir o aumentar su capacidad para infectarlas [24].

Adicionalmente, es importante señalar que mutaciones en el genoma de algunos virus pueden tener otras consecuencias importantes, como provocar que estos adapten su uso de codones al del huésped. Básicamente, en un organismo, la maquinaria celular funciona de manera óptima bajo su respectivo uso de codones. Por lo tanto, esta adaptación consiste en que el *sesgo en el uso de codones* del virus sea lo más similar posible al del huésped, lo que hace que el virus pueda replicarse más rápidamente y con mayor eficacia dentro de las células [30].

2.3. Variabilidad genómica del virus

Desde el surgimiento del virus SARS-CoV-2 a finales de 2019, este ha acumulado mutaciones, lo que significa que ha evolucionado con el tiempo. Una vez que el virus tiene una o más mutaciones significativas, se vuelve genéticamente distinto a la cepa inicial, generando una nueva variante del virus [24].

Actualmente, se han identificado y clasificado diferentes variantes del SARS-CoV-2 en todo el mundo. Aunque se han encontrado variantes con varias mutaciones en proteínas distintas a la espiga, la mayoría de las variantes detectadas hasta la fecha están relacionadas con mutaciones en esta proteína [24].

Como se mencionó anteriormente, la proteína espiga permite que el virus se enlace con los receptores de entrada de las células, por lo que estas mutaciones podrían mejorar dicho enlace y, por ende, facilitar la entrada del virus en las células. Este tipo mutaciones pueden afectar el ritmo de transmisión y la gravedad de la infección. No obstante, cada nueva variante emergente representa un peligro potencial en la efectividad de las vacunas existentes [24].

A partir del año 2020, la Organización Mundial de la Salud (OMS) comenzó a caracterizar algunas de estas variantes, lo que ha permitido establecer prioridades en el seguimiento e investigación sobre ellas [33]. A medida que pasa el tiempo, se han caracterizado diversas variantes del virus, entre las que se incluyen la *Alpha* (B.1.1.7), *Beta* (B.1.351), *Gamma* (P.1), *Delta* (B.1.617.2), *Omicron* (B.1.1.529),

entre otras. Algunas de estas variantes han generado preocupación debido a su alta transmisibilidad y capacidad para evadir el sistema inmunológico [24].

La vigilancia continua en la variación del genoma viral es fundamental para la identificación temprana de nuevas variantes del SARS-CoV-2. En este sentido, existen trabajos como el de *Hahn et al.* [18] que utilizan métodos de detección de anomalías por medio de aprendizaje no supervisado con la finalidad de detectar variantes del virus en tiempo real. Esto se asemeja a un sistema de alarma, ya que al identificar un aumento en el número de anomalías en los resultados, podría ser un indicador de la aparición de nuevas variantes del virus.

Investigaciones más detalladas sobre un gran conjunto de genomas del virus, como la realizada por *Posani et al.* [17], han analizado el uso de codones de cada gen del virus a través del tiempo a fin de describir su evolución. Gracias a ello, se observó que los genes que codifican la espiga (S) y el nucleócapside (N) son los que han mostrado mayores divergencias, acumulando más variaciones genéticas. Sin embargo, es interesante destacar que en general todos los genes del virus muestran una desoptimización en su uso de codones, con respecto al uso de codones del humano. El virus tiende a utilizar un conjunto más uniforme de codones sinónimos, es decir, tiene menos preferencia por codones específicos.

2.4. Resumen del capítulo

En este capítulo se hizo una breve revisión de las características biológicas del virus SARS-CoV-2. Se abordaron aspectos como la composición y organización de su genoma, así como la estructura distintiva de la partícula viral. Se destacó también cómo las mutaciones en el genoma pueden repercutir en la estructura de este, lo que puede influir en la capacidad de infectar y replicarse en las células del huésped. Además, se presentó el concepto de variante, que se refiere a una forma ligeramente diferente del virus y cómo cada una de ellas representa un peligro latente para los tratamientos existentes destinados a controlar la infección.

Dentro de este capituló se abordó el uso de codones, uno de los conceptos cruciales para el desarrollo de este trabajo. A partir del genoma viral completo, es decir, la secuencia de nucleótidos, también es posible obtener una representación mediante el uso de 64 variables, las cuales indican la frecuencia de uso de cada codón. De esta manera es posible caracterizar todo el conjunto de datos como distribuciones de probabilidad o dentro de un espacio geométrico, donde posteriormente es posible aplicar diferentes métodos computacionales de interés.

Una vez descrita la representación del conjunto de datos que se utilizará, en los siguientes dos capítulos se discutirán diversas herramientas necesarias para llevar a cabo el análisis propuesto en la introducción. En el Capítulo 3, se describen los

algoritmos de agrupamiento y reducción de dimensionalidad utilizados, así como algunas herramientas y conceptos complementarios. El Capítulo 4 presenta los algoritmos empleados para la detección de anomalías. Finalmente, el Capítulo 5 y el Capítulo 6 presentan los resultados obtenidos, así como la conclusión de la tesis.

Capítulo 3

Herramientas para analizar el uso de codones del SARS-CoV-2

En el capítulo anterior, se discutió el fenómeno del uso de codones. Este hecho posibilita una manera de representar una secuencia completa de nucleótidos con solo 64 variables. El cálculo de la frecuencia de aparición de cada codón permite generar un histograma específico para cada genoma del conjunto de datos. De igual forma, estas representaciones se pueden considerar en un espacio geométrico, donde cada codón representa una dimensión en un sistema de coordenadas.

Aunque representar el genoma completo mediante el uso de codones implica una pérdida de información, esta representación permite la aplicación de un conjunto de herramientas y técnicas existentes para el análisis de datos. Además, tal como se discute en los estudios de *Simón et al.* [11] y *Posani et al.* [17], emplear esta representación permite comparar genomas de diferentes organismos sin importar el tamaño de sus secuencia. Del mismo modo, permite analizar la evolución en el uso de codones del virus SARS-CoV-2, el cual es analizado en esta tesis.

En este capítulo se presentan algunos de los conceptos, herramientas y algoritmos utilizados para llevar a cabo el análisis exploratorio de datos. En primer lugar, debido a que el uso de codones se puede considerar como una instancia de datos composicionales, la primera sección aborda las características principales de este tipo de datos. Después, se mencionan las medidas de distancia empleadas para cuantificar la similitud o diferencia entre el uso de codones de las diversas secuencias utilizadas. Luego, se describe el algoritmo de *k-medias*, fundamental en cierta parte de la metodología utilizada. Posteriormente, se menciona brevemente la *entropía* como medida, la cual permite cuantificar el grado de incertidumbre en las diferentes distribuciones de uso de codones. Finalmente, se abordan algunos algoritmos de reducción de la dimensionalidad, los cuales resultan útiles para simplificar y visualizar conjuntos de datos descritos por muchos atributos.

El desarrollo de gran parte de la tesis se basa en la combinación de estas herramientas y algoritmos mencionados, así como en diversas técnicas de visualización de datos, tales como gráficas de dispersión y de barras.

3.1. Datos composicionales

Un conjunto de datos se considera composicional si la suma de los atributos o componentes de cada una de sus instancias siempre es igual a una constante. En contraste con otros tipos de datos, donde cada atributo puede tomar cualquier valor dentro de un rango definido, en los datos composicionales cada componente está determinado por una combinación específica de los demás [34, p. 1–2]. Los histogramas normalizados son un ejemplo de este tipo de datos, donde la suma de las diferentes frecuencias relativas de cada componente siempre es igual a uno.

Entonces, dentro de un conjunto de este tipo de datos, cada instancia $\mathbf{x}_i$ se puede representar definiendo sus J partes composicionales:

$$\mathbf{x}_i = \left\{ (x_{i,1}, x_{i,2}, x_{i,3}, \ldots, x_{i,J}) \mid x_{i,j} > 0; j = 1, 2, \ldots, J; \sum_{j=1}^{J} x_{i,j} = k \right\} \tag{3.1}$$

donde $x_{i,j}$ representa la proporción del componente j de la instancia i, proporciones que siempre son positivas. Además, como se definió anteriormente, las instancias $\mathbf{x}_i$ cumplen con la propiedad de *cerradura*, lo que implica que la suma de las partes composicionales es igual a una constante k. Generalmente $k = 1$, para lo cual es necesario aplicar la operación de cierre, ésta consiste en dividir la proporción de cada componente entre el agregado total de éstos [34, p. 3–5].

Existen otras características importantes en este tipo de datos, como la invariancia de escala y la invariancia de permutación. La primera se refiere a cómo al multiplicar los datos por una constante cualquiera, no hay efecto alguno en ellos, siempre y cuando se vuelva a aplicar la operación de cierre. Por otro lado, la invariancia de permutación implica que los resultados de las operaciones no dependen del orden en que se presenten los componentes [34, p. 5].

Una de las formas de visualizar de este tipo de datos es a través del espacio geométrico conocido como *símplex de probabilidad*, el cual es el resultado de la restricción mencionada anteriormente, donde la suma de las partes composicionales es constante [34, p. 12–13]. Un ejemplo de un *símplex* en dos dimensiones se obtiene a partir de datos composicionales con tres componentes ($J = 3$). La Figura 3.1 muestra un conjunto de datos con estas características, donde todas las instancias se encuentran dentro de un triángulo embebido en 3 dimensiones (puntos azules), cada una de estas instancias representa una composición en específico

de los tres componentes (x, y, z). La proporción de los componentes es homogénea en las instancias ubicadas en el centro del triángulo, mientras que aquellas cerca de los vértices presentan una proporción mayor en uno o dos componentes. Los histogramas de la Figura 3.1 presentan algunos ejemplos de la composición de las instancias, según su posición en el *símplex*.

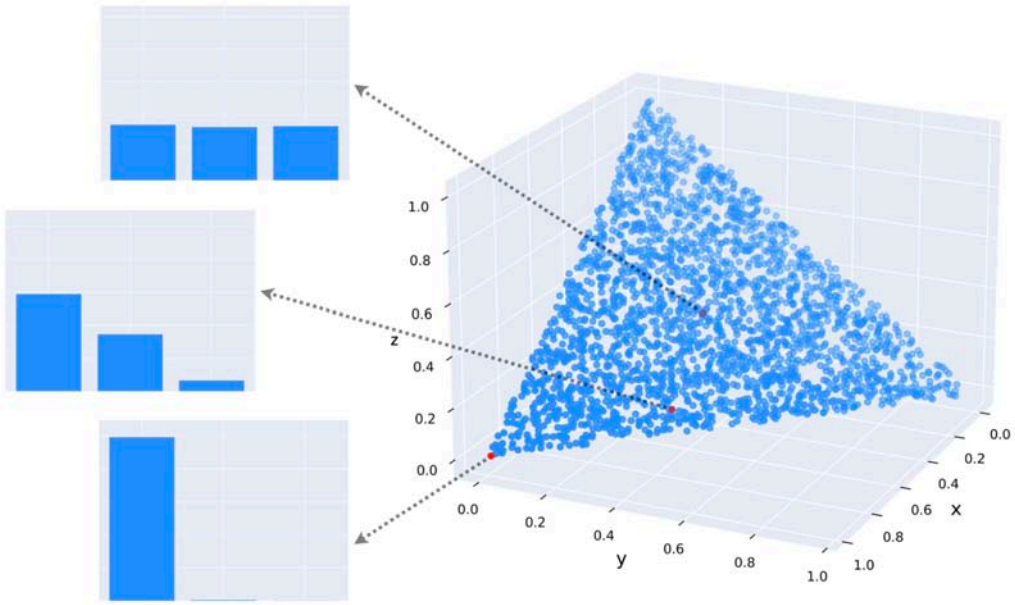

Figura 3.1: Símplex de datos composicionales con tres componentes. La gráfica tridimensional muestra el símplex en dos dimensiones, mientras que los histogramas de la izquierda muestran instancias específicas dentro del conjunto de datos.

De manera similar, los datos composicionales que constan de cuatro componentes son representados por un tetraedro, siendo este un *símplex* en tres dimensiones. Si los datos están compuestos por más componentes, su representación se dará a través de un *símplex* de mayor dimensión. La dimensionalidad de este espacio está determinada por el número total de componentes de los datos menos uno. En otras palabras, si los datos tienen J componentes, su representación se dará en un *símplex* de $J - 1$ dimensiones [34, p. 13].

En el análisis de datos composicionales, suele ser común trasladar las instancias del espacio simplicial al espacio euclidiano. Esta técnica permite estudiar la estructura de los datos sin la restricción de la suma constante, al mismo tiempo que conserva la información relativa de cada composición. Esto facilita la aplicación de diversas técnicas estadísticas y el cálculo de una *distancia euclidiana* más intuitiva. Para llevar a cabo este procedimiento, se suelen utilizar algunas

transformaciones llamadas *log-cocientes*. Estas transformaciones trasladan las instancias a un espacio real descrito por $J-1$ atributos independientes y, al mismo tiempo, eliminan la propiedad de *cerradura* [34, p. 17–24].

En este trabajo, se utiliza la transformación del *log-cociente centrado*, la cual obtiene un nuevo conjunto de atributos para cada instancia i. Ésta se calcula como el logaritmo del cociente de cada componente j sobre la media geométrica del total de J componentes:

$$y_{i,j} = \log\left(\frac{x_{i,j}}{\prod_{j=1}^{J} x_{i,j}}\right) = \log(x_{i,j}) - \log(\prod_{j=1}^{J} x_{i,j}) \quad j = 1, 2, \ldots, J \tag{3.2}$$

donde $y_{i,j}$ es el nuevo atributo y el conjunto de $(y_{i,1}, y_{i,2}, \ldots, y_{i,J})$ atributos describen a la instancia en el nuevo espacio geométrico [34, p. 18]. La Figura 3.2 muestra esta transformación aplicada al conjunto de datos composicionales con tres componentes, mencionado anteriormente.

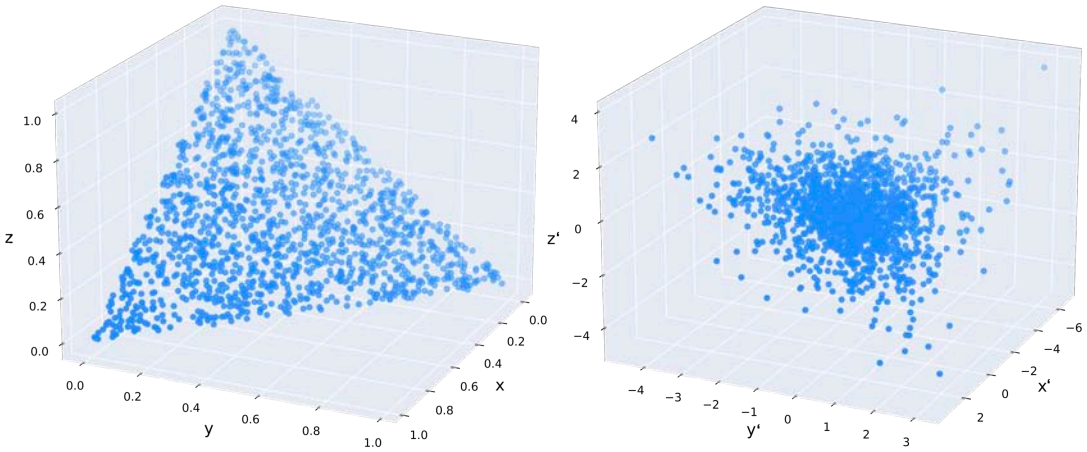

Figura 3.2: Transformación *log-cociente centrado* en datos composicionales. Izquierda: Símplex generado a partir de instancias con tres componentes (x, y, z). Derecha: Espacio con los componentes transformados (x', y', z').

El conjunto de datos a considerar en esta tesis es composicional. Cada instancia es un uso de codones característico en el que, después de aplicar la operación de cierre, la suma de sus componentes (64 codones) es igual a uno. La representación geométrica de estas composiciones esta dada por un símplex de alta dimensionalidad ($d = 63$). En este espacio, el valor numérico de cada dimensión d indica la frecuencia relativa de uso de cada uno de los codones. Por lo que a fin de emplear algunas herramientas abordadas en este documento, fue necesario aplicar la transformación *log-cociente centrado* descrita.

3.2. Distancias

Las métricas de distancia son una medida cuantitativa que nos permite saber qué tan cercanos o distantes se encuentran dos objetos en el espacio de atributos. El valor numérico de éstas disminuye a medida que los objetos se vuelven más similares. Una medida se considera como métrica de distancia sólo si cumple con cuatro propiedades fundamentales [35, p. 304]:

- La propiedad de *positividad*: La cual establece que la distancia entre dos elementos siempre es mayor o igual a cero.

$$d(x, y) \geq 0 \quad \text{para todo } x \text{ y } y$$

- La propiedad de *identidad*: Ésta dicta que la distancia entre dos elementos es cero sí y sólo sí éstos son idénticos.

$$d(x, y) = 0 \quad \text{sí y sólo sí } x = y$$

- La propiedad de *simetría*: Ésta establece que la distancia de x a y es igual que la distancia de y a x.

$$d(x, y) = d(y, x) \quad \text{para todo } x \text{ y } y$$

- La propiedad de *desigualdad triangular*: La cual dicta que la distancia directa entre dos puntos es igual o menor a la suma de las distancias a través de un tercer punto.

$$d(x, y) \leq d(x, z) + d(z, y) \quad \text{para todo } x, y \text{ y } z$$

Existen diversas medidas de distancias que se han empleado en distintos campos de estudio, donde en muchas ocasiones la selección de éstas depende de la naturaleza y representación de los datos a examinar. Un trabajo que ha presentado algunas medidas relevantes es el de *Sung-Hyuk Cha* [36], en este se aborda una amplia variedad de distancias con el fin de comparar funciones de densidad de probabilidad, ya sea mediante un enfoque vectorial o probabilístico.

En esta tesis se emplearon dos métricas de distancia para comparar el uso de codones en diferentes genomas: la *distancia euclidiana*, que considera las instancias en un espacio euclidiano para calcular la medida, y la *distancia de Wasserstein*, que considera cada uso de codones como una distribución de probabilidad para obtener la distancia.

3.2.1. Distancia euclidiana

La *distancia euclidiana* es una medida ampliamente utilizada en matemáticas para calcular la separación que existe entre dos puntos en un espacio euclidiano en d dimensiones. Se define como:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2} \tag{3.3}$$

En esta tesis, tanto $\mathbf{x}$ como $\mathbf{y}$ representan un uso de codones característico. Estas instancias son composicionales y, como se mencionó en la Sección 3.1, se encuentran en un espacio simplicial multidimensional. Por esta razón, aunque es posible aplicar la *distancia euclidiana* sin tener en cuenta las propiedades de este espacio geométrico, es preferible considerarlo mediante la aplicación de la transformación *log-cociente centrado* antes de calcular la distancia [34, p. 85]. El procedimiento en conjunto para calcular esta medida también es referido como la *distancia de Aitchison* o *distancia log-cociente* [34, p. 116].

La Figura 3.3 muestra, a modo de ejemplo, como cambia la distancia esperada en un espacio simplicial en dos dimensiones después de aplicar la transformación mencionada.

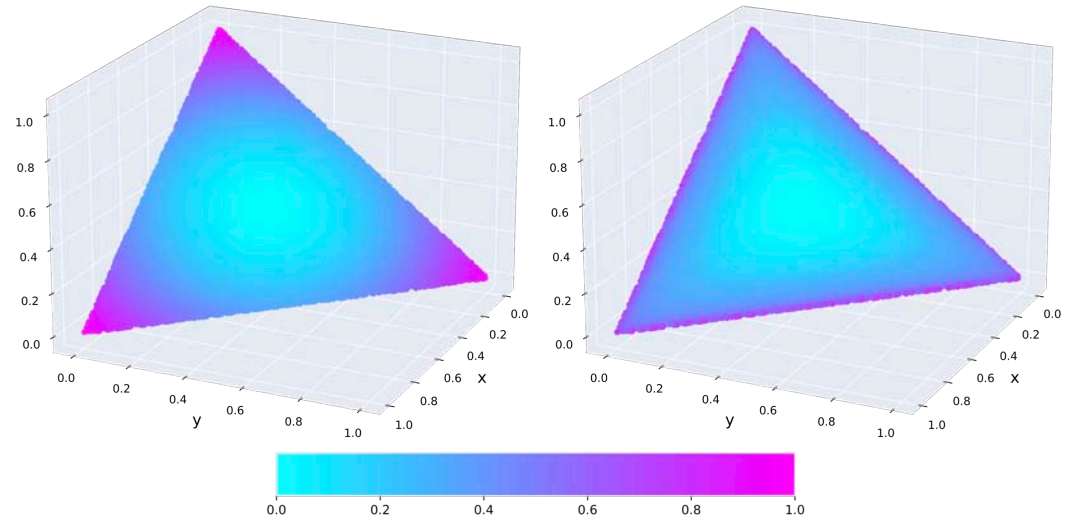

Figura 3.3: *Distancia euclidiana* esperada (normalizada) en datos composicionales. Izquierda: Distancia aplicada directamente en el espacio simplicial. Derecha: Distancia aplicada a la transformación *log-cociente centrado* de los datos.

3.2.2. Distancia de Wasserstein

La *distancia de Wasserstein*, también denominada *distancia de acarreo de arena*, se distingue de la *distancia euclidiana* al considerar la forma de la distribución de probabilidad de las variables involucradas en el cálculo. Esta métrica calcula el *trabajo* mínimo necesario para transformar una distribución en otra, donde el *trabajo* se define como la cantidad de masa de distribución que se debe mover multiplicada por la distancia al punto donde se tiene que transportar. Para ello, es necesario resolver el problema del transporte óptimo, que consiste en encontrar un plan de transporte τ entre todos los posibles planes de transporte T, que minimice este *trabajo* de transformación [37].

En esta tesis, se emplea la *distancia de Wasserstein* para distribuciones discretas, que se define como:

$$W(P,Q) = \min_{\tau \in T} \sum_{i=1}^{N} \sum_{j=1}^{M} \tau(x_i, y_j) \cdot |x_i - y_j| \tag{3.4}$$

donde el plan de transporte τ debe cumplir con ciertas restricciones:

$$\sum_{i=1}^{N} \tau(x_i, y_j) = q(y_j)$$

$$\sum_{j=1}^{M} \tau(x_i, y_j) = p(x_i) \quad \tau(x_i, y_j) \geq 0$$

En esta formulación, P y Q representan dos variables aleatorias, donde $p(x_i)$ y $q(y_j)$ son sus respectivas funciones de probabilidad. N y M corresponden al total de estados posibles de las variables aleatorias, mientras que $|x_i - y_j|$ representan la distancia entre dos estados particulares de cada variable aleatoria.

La Figura 3.4 ejemplifica de manera visual cómo se calcula esta distancia. En la parte superior se presentan dos variables aleatorias, P y Q, con sus respectivas distribuciones de probabilidad. En la parte inferior de ésta, se muestra el transporte óptimo τ, el cual define las cantidades de masa de la distribución de P que deben moverse para construir la distribución de Q. Estas cantidades de masa se multiplican por la distancia a la que deben ser transportadas ($|x_i - y_j|$). Finalmente, la suma de todas estas multiplicaciones da como resultado la *distancia de Wasserstein* entre P y Q, o dicho de otra manera, el mínimo *trabajo* necesario para transformar la distribución de P en la distribución de Q.

En este trabajo, con el objetivo de calcular la *distancia de Wasserstein*, cada uso de codones que se analiza se considera como una distribución de probabilidad. En estas distribuciones, la probabilidad de cada codón es su frecuencia relativa de aparición en la secuencia bajo estudio.

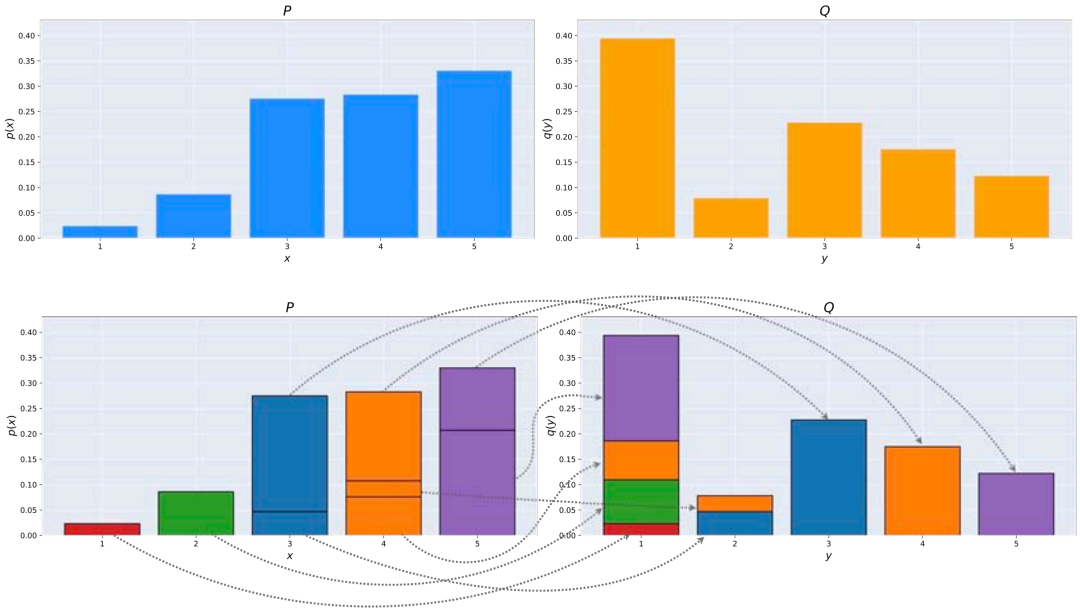

Figura 3.4: Cálculo de la *distancia de Wasserstein* entre dos distribuciones. Arriba: Distribuciones discretas de probabilidad de P y Q. Abajo: Transporte óptimo para transformar la distribución de P en la distribución de Q.

3.3. Método de agrupamiento k-medias

El objetivo de este método es segmentar un conjunto de datos en k grupos, donde los datos contenidos en un grupo particular presentan una mayor similitud entre sí que con datos no contenidos en éste. Para lograr esta segmentación, el algoritmo calcula en cada iteración el centroide de los k diferentes grupos formados y asigna todos los elementos del conjunto de datos al grupo cuyo centroide se encuentre más cercano a ellos, según una función de distancia elegida. Este proceso se repite hasta que no haya (o haya muy pocos) cambios en la asignación de nuevas agrupaciones. Para lo cual, es importante establecer previamente el número de k grupos que se desean formar en el conjunto de datos [38, p. 43–45]. El Algoritmo 1 presenta el procedimiento básico necesario para obtener los centroides representativos de un conjunto de datos.

En esta tesis, se emplea este algoritmo para encontrar un grupo representativo de diferentes distribuciones de uso de codones en un conjunto específico de genomas seleccionados. De esta manera, cada conjunto de secuencias se representa por

k centroides, donde cada centroide es un uso de codones esperado para todos los genomas agrupados en torno a él. Esto permite reducir la cantidad de elementos del conjunto de datos completo. Además, este algoritmo, junto con el cálculo de distancias, permite visualizar la variación de todos los elementos del conjunto de datos mediante un enfoque diferente.

Algoritmo 1: Agrupamiento k-medias

Entrada: Conjunto de datos $X = \{x_i\}_{i=0}^{n}$, Número de agrupaciones k

Salida: Conjunto de centroides $C = \{c_j\}_{j=0}^{k}$

Se seleccionan aleatoriamente k instancias x_i como los centroides en C;

repetir

  para cada $x_i \in X$ **hacer**
    Asignar x_i al grupo j, según el c_j más cercano.
  fin
  para cada $c_j \in C$ **hacer**
    Actualizar c_j computando la media de los x_i en el grupo j;
  fin

hasta que *Las agrupaciones ya no cambien*;

3.4. Entropía

La *entropía* es un concepto fundamental en la teoría de la información, ya que esta se puede entender como el grado de sorpresa, incertidumbre o aleatoriedad implícita en una distribución de probabilidad. Por ejemplo, en una distribución uniforme, la *entropía* es máxima. Esto se traduce en que anticipar el estado en el que se encontrará el sistema o el valor que tomará la variable aleatoria resulta imposible, considerando como única información disponible esta distribución. Por otro lado, si la distribución es una delta de Dirac, es decir, todas las probabilidades se concentran en un único valor, la *entropía* será mínima, lo que indica que la distribución es totalmente determinista [39, p. 201–203].

La *entropía*, entonces, es el valor esperado de sorpresa de una variable aleatoria. La medida de sorpresa se expresa mediante la función $log_2(\frac{1}{p(x_i)})$, donde $p(x_i)$ representa la probabilidad de que ocurra un estado en particular. Cuando

un estado tiene una alta probabilidad, su ocurrencia no genera mucha sorpresa. Sin embargo, cuando un estado es altamente improbable, su ocurrencia resulta sumamente sorprendente. Por lo tanto, la *entropía* H de una variable aleatoria discreta P se define como:

$$H(P) = \sum_{i=1}^{N} p(x_i) \cdot \log_2(\frac{1}{p(x_i)}) \tag{3.5}$$

donde $p(x_i)$ es su función de probabilidad, N son los estados posibles de esta variable aleatoria y x_i es un estado en particular [40, p. 32].

Como se ha mencionado, las variables aleatorias a considerar son los usos de codones de diferentes genomas. De esta manera, es posible comparar la *entropía* de cada uno de estos, donde su valor es un indicador del grado de *sesgo en el uso de codones*, mencionado en el Capítulo 2. Si el *sesgo en el uso de codones* es alto, se espera que el valor de la *entropía* sea bajo. Por el contrario, si el valor de la *entropía* es alto, se espera que este sesgo sea relativamente bajo.

La Figura 3.5 muestra el valor de la *entropía* en dos genomas con diferente *sesgo en el uso de codones*. La parte superior se presenta la *entropía* del uso de codones característico del virus SARS-CoV-2 de referencia, mencionado en el Capítulo 2. El uso de codones en la parte inferior es sintético, generado a partir del mismo SARS-CoV-2 de referencia. En este, todos los codones sinónimos (secuencias que codifican el mismo aminoácido) tienen una misma frecuencia relativa de uso, lo que en consecuencia reduce el grado en el *sesgo del uso de codones* y aumenta el valor de la *entropía*. En este ejemplo, se puede observar claramente cómo a medida que el uso de codones (la distribución) se vuelve más uniforme, el valor de la *entropía* aumenta.

3.5. Reducción de la dimensionalidad

Las técnicas de visualización son herramientas fundamentales para llevar a cabo el análisis exploratorio de datos, ya que nos ayudan a comprender mejor nuestro conjunto de datos. Por ejemplo, al emplear estas técnicas, podemos observar las variaciones de los elementos, así como detectar patrones y tendencias que no serían evidentes con un simple resumen estadístico [35, p. 157]. No obstante, existen conjuntos de datos en alta dimensionalidad, lo que limita las técnicas de visualización que se pueden utilizar para mostrar todo el espacio de atributos.

En este trabajo, se aborda un conjunto de datos en alta dimensionalidad, donde cada instancia es descrita por 64 atributos, debido a esto, la dimensión del espacio de atributos es de 64. Aunque esta representación es más concisa que

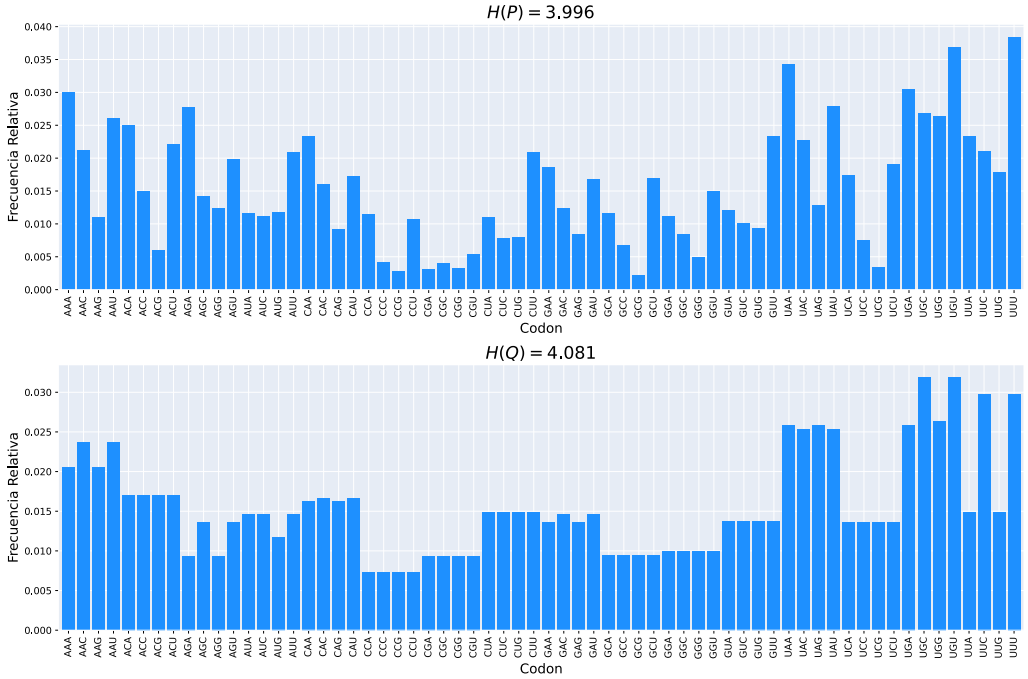

Figura 3.5: *Entropía* en el uso de codones. Arriba: *Entropía* en el uso de codones del SARS-CoV-2 de referencia obtenido de GISAID [12]. Abajo: *Entropía* en un uso de codones sintético (codones sinónimos presentan la misma frecuencia relativa).

la secuencia completa de nucleótidos, sigue siendo imposible visualizar todo el espacio de atributos en una sola gráfica de dispersión. Por lo tanto, es necesario recurrir a métodos de reducción de la dimensionalidad.

El objetivo de los métodos de reducción de la dimensionalidad es obtener una representación de los datos con un menor número de atributos, manteniendo la mayor cantidad de información posible. En otras palabras, estos métodos transforman los datos de un espacio de alta dimensionalidad $\mathbf{x} \in \mathbb{R}^d$ a uno de baja dimensionalidad $\mathbf{y} \in \mathbb{R}^l$, donde $l < d$ [39, p. 651].

La proyección lineal y el aprendizaje de variedad topológica son dos de los enfoques principales de este tipo de métodos [41, p. 205-206]. El primero consiste en proyectar el conjunto de datos perpendicularmente sobre un hiperplano de menor dimensión. En esta sección, se abordará el *análisis de componentes principales* (*PCA*), que emplea la proyección lineal para reducir la dimensionalidad de los datos. Por otro lado, el aprendizaje de variedad topológica se basa en la idea de que conjuntos de datos en alta dimensionalidad están contenidos en estructuras de baja dimensión, como superficies o curvas. El método bajo este enfoque que se abordará en esta sección es el *mapeo isométrico* (*Isomap*).

3.5.1. Análisis de componentes principales (PCA)

El *análisis de componentes principales* es uno de los métodos más populares y sencillos para reducir la dimensionalidad de un conjunto de datos. Como se mencionó anteriormente, el enfoque de este método se basa en la proyección lineal, lo que implica realizar una proyección ortogonal de los datos desde un espacio de alta dimensionalidad $\mathbf{x} \in \mathbb{R}^d$ hacia un subespacio de baja dimensionalidad $\mathbf{y} \in \mathbb{R}^l$, donde este subespacio busca conservar la mayor información posible del espacio original [39, p. 651]. El método de *PCA* crea este subespacio eligiendo un subconjunto de vectores base o *componentes principales* que capturen las direcciones de la máxima varianza del conjunto de datos original.

En concreto, el método determina los d *componentes principales*, ortogonales entre sí, que expliquen la mayor varianza posible de un conjunto de datos en d dimensiones. El primer componente se encuentra en la dirección que explica la máxima varianza posible de todos los datos. El segundo componente explica la mayor varianza posible de la varianza restante y así sucesivamente para cada componente principal. Finalmente, se seleccionan los primeros l *componentes principales* en los que se proyectarán todos los elementos del conjunto de datos, reduciendo así las atributos que describen a cada instancia [41, p. 212].

Una de las formas de obtener los *componentes principales* es a través de la *descomposición de valores singulares*. Esta técnica consiste en factorizar una matriz, o en este caso, un conjunto de datos con n instancias y d atributos, $\mathbf{X}_{n\times d}$, en tres matrices diferentes: $\mathbf{U}_{n\times n}$, $\mathbf{\Sigma}_{n\times d}$ y $\mathbf{V}_{d\times d}$. De manera que:

$$\mathbf{X} = \mathbf{U} \cdot \mathbf{\Sigma} \cdot \mathbf{V}^T \tag{3.6}$$

Las columnas de la matriz $\mathbf{V}$ definen los *componentes principales* mencionados, mientras que los valores singulares contenidos en $\mathbf{\Sigma}$ pueden ser utilizados para determinar el porcentaje de contribución de cada *componente principal* a la explicación de la varianza total del conjunto de datos [35, p. 258–259].

Una vez obtenida la matriz $\mathbf{V}_{d\times d}$ se seleccionan las l columnas (*componentes principales*) para formar la matriz de proyección $\mathbf{W}_{d\times l}$. Por último, se obtiene el conjunto de datos proyectado $\mathbf{Y}_{n\times l}$, mediante la transformación mostrada en la ecuación 3.7 [41, p. 2014].

$$\mathbf{Y} = \mathbf{X} \cdot \mathbf{W} \tag{3.7}$$

Un ejemplo sencillo de la aplicación de este método se muestra en la Figura 3.6. En el lado izquierdo de ésta, se presenta un conjunto de datos descrito por 2 atributos (x y y), así como la dirección de sus 2 *componentes principales* (CP1 y CP2). En el lado derecho de la figura, parte superior, se muestra el conjunto de datos con los *componentes principales* como vectores base. En el lado derecho, parte central, se presenta la proyección de los datos sobre cada componente.

Finalmente, en la parte inferior derecha, se muestra el porcentaje de contribución de cada componente a la explicación de la varianza total del conjunto de datos, donde CP1 explica la mayor parte de esta varianza ($\sim 95\,\%$).

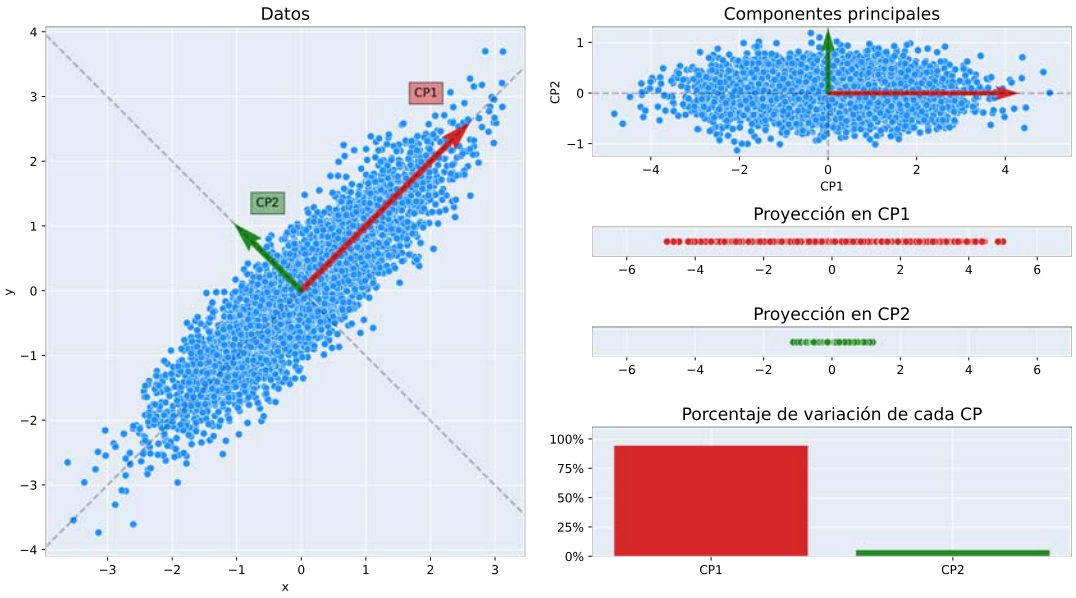

Figura 3.6: Aplicación de *análisis de componentes principales*. Izquierda: Conjunto de datos original y dirección de *componentes principales*. Derecha: Proyección de los datos sobre los componentes y su porcentaje de explicación de varianza.

Las proyecciones mostradas en la Figura 3.6 definen cada instancia con un solo atributo. Se puede observar que la proyección sobre el primer *componente principal* (CP1) es la más conveniente para realizar la reducción de dimensionalidad, ya que es la que conserva la mayor parte de la varianza de los datos.

Una de las características que asume este método es que los datos se encuentran en un subespacio real, donde existe una relación lineal entre los atributos. De esta manera, los *componentes principales* pueden representar combinaciones lineales de los atributos originales (Fig. 3.6 parte superior derecha). Sin embargo, como se mencionó anteriormente, el conjunto de datos a utilizar en esta tesis es composicional, lo cual implica que el espacio en el que existen es simplicial, con propiedades mencionadas en la Sección 3.1.

Por lo tanto, en este trabajo, en lugar de aplicar *PCA* directamente sobre los datos originales, se prefiere aplicarlo a su transformación *log-cociente*. Esta transformación, también abordada en la Sección 3.1, traslada los datos composicionales a un espacio real, donde posteriormente se aplica *PCA*. A este procedimiento conjunto también se le denomina *análisis log-cociente* [34, p. 41].

Algo importante a tener en cuenta es que, aunque el método de *PCA* es eficaz para reducir la dimensionalidad de conjuntos de datos que se encuentran cerca o sobre un subespacio lineal, existen conjuntos de datos que se encuentran en un subespacio no lineal, donde aplicar *PCA* no es lo más adecuado. Por ejemplo, para el conjunto de datos mostrado en la Figura 3.7 el método de *PCA* no es capaz de encontrar un subespacio que capture la estructura no lineal inherente de este conjunto de datos (Fig. 3.7 izquierda), donde instancias muy distintas dentro de la estructura original terminan en regiones cercanas en el subespacio generado por *PCA* [41, p. 209].

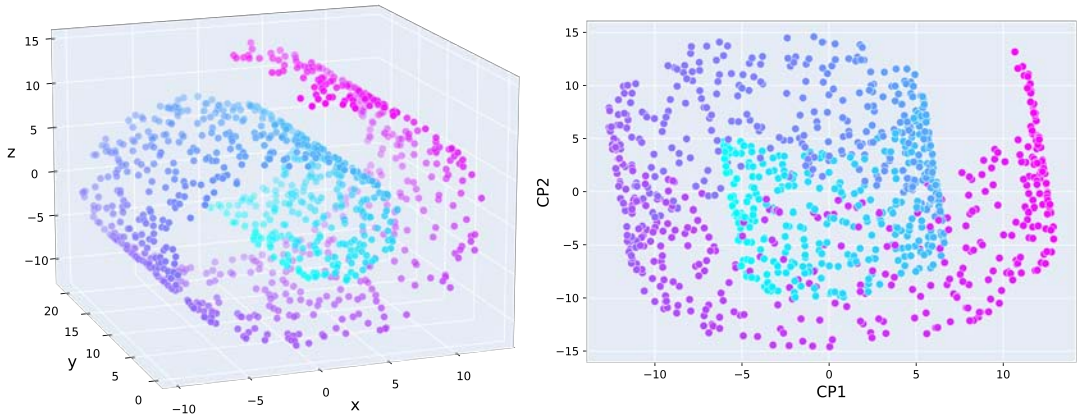

Figura 3.7: Aplicación de *PCA* sobre un conjunto de datos no lineal. Izquierda: Conjunto de datos original. Derecha: Proyección de los datos sobre sus primeros *componentes principales*. El color marca la cercanía dentro de la estructura original.

3.5.2. Mapeo Isométrico (Isomap)

A diferencia del método de *análisis de componentes principales*, los métodos de aprendizaje de variedad topológica tienen la capacidad de descubrir estructuras no lineales que están presentes en muchos conjuntos de datos de alta dimensionalidad. En estos se asume que dichas estructuras corresponden a una variedad topológica. Este concepto se refiere a un espacio en el cual las regiones locales alrededor de cada instancia pueden ser consideradas como espacios euclidianos [39, p. 683]. Un ejemplo se puede apreciar en el conjunto de datos mostrado en la Figura 3.8, donde al considerar las regiones locales a cada instancia, se observa como los datos están sobre o muy cerca de una superficie bidimensional embebida en el espacio tridimensional.

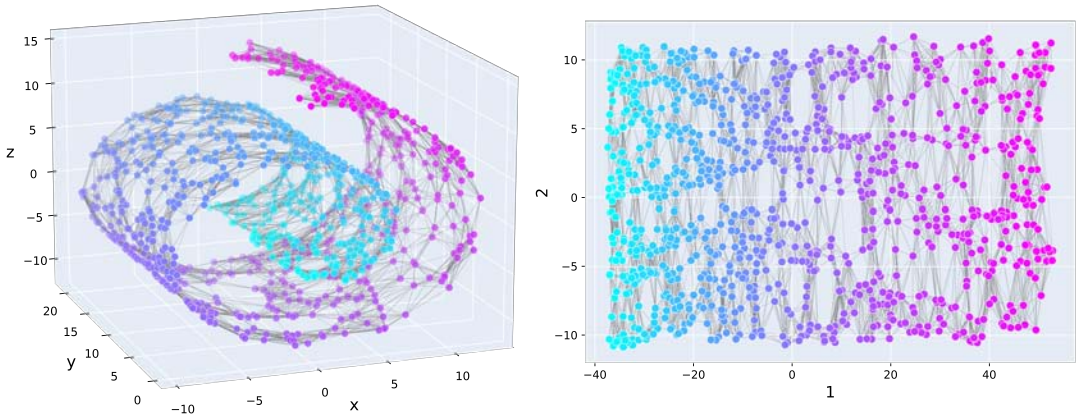

Figura 3.8: Aplicación de *mapeo isométrico* a un conjunto de datos no lineal. Izquierda: Conjunto de datos original y grafo generado apartir de los k vecinos. Derecha: Estructura bidimensional aproximada por *Isomap*. El color marca la cercanía dentro de la estructura original.

El método bajo este enfoque utilizado en esta tesis es el *mapeo isométrico*. En pocas palabras, este método crea un grafo conectando los k vecinos más cercanos a cada punto (instancia) y luego trata de aproximar la geometría de la variedad topológica, de forma que se conserve mejor la distancia geodésica (longitud del camino más corto) entre cada par de puntos dentro de este grafo generado [42].

En concreto, para un conjunto de datos $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^d\}_{i=1}^n$, el método inicialmente crea un grafo conectando los k vecinos más cercanos de cada $\mathbf{x}_i$. Luego, produce una matriz $\mathbf{D}_G$, de tamaño $n \times n$, con las distancias geodésicas $d_{i,j}$ entre cada par de puntos $(\mathbf{x}_i, \mathbf{x}_j)$ contenidos en el grafo. Para puntos vecinos, la distancia geodésica es igual a su distancia en el espacio multidimensional original, mientras que para puntos más lejanos, esta distancia se estima con la longitud del camino más corto entre ellos. Dicho camino puede obtenerse mediante el algoritmo de *Dijkstra* [39, p. 688].

Una vez obtenida la matriz $\mathbf{D}_G$, *Isomap* busca encontrar un conjunto de vectores en baja dimensionalidad $\mathbf{Y} = \{\mathbf{y}_i \in \mathbb{R}^l\}_{i=1}^n$, donde $l < d$, los cuales preserven lo mejor posible la distancia geodésica $d_{i,j}$. Una forma de lograr esto es formularlo como un problema de optimización, donde los vectores $\mathbf{y}$ elegidos son aquellos que minimizan una función de costo C [42], por ejemplo:

$$C = \sum_{i,j=1}^{n} (d_{j,i} - ||\mathbf{y}_i - \mathbf{y}_j||)^2 \tag{3.8}$$

La Figura 3.8 presenta la aplicación del método de *mapeo isométrico* al mismo conjunto de datos no lineal de la Figura 3.7. El conjunto de datos en el espacio original ($d = 3$) esta en la parte izquierda de la figura, donde también se aprecia el grafo generado al conectar los k vecinos más cercanos de cada punto. La parte derecha de la figura muestra el nuevo conjunto de vectores en baja dimensión ($l = 2$) calculados por el método, los cuales preservan mejor las distancias geodésicas del grafo generado del conjunto de datos original.

En esta tesis, se emplearon tanto *PCA* como *Isomap* para reducir la dimensionalidad del conjunto de usos de codones a estudiar. La elección de ambas técnicas se basa en el hecho de que cada una asume diferentes características sobre el espacio en el que se encuentra el conjunto de datos. Es importante tener en cuenta que se desconoce la estructura real de este conjunto de datos en su espacio multidimensional original.

3.6. Resumen del capítulo

En este capítulo se abordaron las herramientas necesarias para llevar a cabo cierta parte de la metodología propuesta en el Capítulo 1. Estas herramientas se describieron teniendo en cuenta el tipo de datos a analizar en la tesis, los cuales consisten en un conjunto de diferentes genomas del SARS-CoV-2 representados por su uso de codones.

En la Sección 3.1, se abordaron las características principales de los datos composicionales, centrándose en las limitaciones del espacio simplicial y la transformación necesaria antes de aplicar ciertas herramientas. La Sección 3.2 describió las distancias *euclideana* y de *Wasserstein*, destacando cómo estas interpretan a cada instancia para calcular la medida. Posteriormente, la Sección 3.3 realizó una breve introducción al método de agrupamiento *k-medias*. En Sección 3.4 se exploró la *entropía* como una medida del grado de *sesgo en el uso de codones*. Por último, la Sección 3.5 presentó dos métodos de reducción de dimensionalidad, *PCA* e *Isomap*, cada uno con un enfoque distinto.

El siguiente capítulo describe la tarea de detección de anomalías, junto con los algoritmos de aprendizaje no supervisado que se emplean en la metodología para la detección, algunos de los cuales hacen uso de las herramientas mencionadas en este capítulo. Posteriormente, el Capítulo 5 desarrolla la metodología mencionada en la introducción. Finalmente, el Capítulo 6 presenta las conclusiones de este trabajo.

Capítulo 4

Detección de anomalías

La detección de anomalías se puede definir como el proceso que trata de identificar elementos en un conjunto de datos que no se ajustan al comportamiento esperado, que parecen inconsistentes con respecto al resto de los datos [43]. Uno de los desafíos cruciales de esta tarea es que no existe una única forma de evaluar qué tan inconsistente es una instancia en comparación con las demás. Debido a este aspecto, a lo largo de los años se han propuesto numerosos enfoques para abordar la detección de estos elementos.

En el pasado, la principal motivación para encontrar este tipo de instancias anómalas era con el fin de removerlas del conjunto de datos a trabajar, ya que, por lo regular, muchos algoritmos de aprendizaje computacional son altamente sensibles a dichas instancias [44]. Sin embargo, en la actualidad, existe un gran interés en las anomalías en sí mismas, ya que con frecuencia estas se asocian con sucesos interesantes o sospechosos. Por lo tanto, este tipo de instancias suelen proporcionar información útil sobre el fenómeno que se está estudiando.

La detección de anomalías es de suma importancia en una gran variedad de campos de estudio [45, p. 1-2]. Por ejemplo, en el diagnóstico médico, este tipo de técnicas se han aplicado exitosamente para identificar patrones inusuales en datos recopilados por todo tipo de escáneres médicos [2]. En el ámbito financiero, su detección juega un papel fundamental para descubrir fraudes en transacciones bancarias [3]. En el campo de la ciberseguridad, la tarea es esencial para identificar intrusiones en sistemas computacionales [4]. De la misma forma, en muchos otros campos la detección de anomalías juega un papel significativo.

Específicamente, en el campo de la genómica y la epidemiología, la detección de estos elementos también ha adquirido gran relevancia, pues comprender las mutaciones y la evolución de bacterias y virus resulta crucial para combatir las enfermedades que estos pueden causar [5, p. 12]. En este sentido, contar con técnicas de detección de anomalías resulta extremadamente útil para localizar los virus o bacterias más irregulares en grandes conjuntos de datos, lo que permite a

expertos en el campo analizar en detalle estos resultados en etapas posteriores.

Durante situaciones como la pandemia del COVID-19, este tipo de herramientas posibilita identificar aquellos virus más atípicos, los cuales podrían ser causantes de acelerar el ritmo de transmisión y/o gravedad de la infección. Un ejemplo de esto es el estudio realizado por *Hahn et al.* [18], mencionado previamente en el Capítulo 1. En este se establece una relación entre el aumento de elementos atípicos y la aparición de algunas nuevas variantes del virus SARS-CoV-2.

En este capítulo se abordan algunas características generales de la detección de anomalías y se describen los algoritmos empleados en la metodología propuesta. En primer lugar, se aborda la clasificación general de las técnicas de detección de anomalías. Luego, se define los tipos de anomalías existentes (*anomalías puntuales*, *anomalías contextuales* y *anomalías colectivas*). Posteriormente, se presenta de manera concisa la clasificación de diversos algoritmos de aprendizaje no supervisado para la detección y se detalla el funcionamiento de los algoritmos empleados en este trabajo. Estos incluyen *DBSCAN*, *detección mediante k-medias*, *valor atípico local* (*LOF*), *bosque de aislamiento* y algunas otras estrategias que se diseñaron con las herramientas presentadas en el Capítulo 3.

4.1. Clasificación de técnicas de detección

Al día de hoy existen diferentes técnicas de detección de anomalías que se basan en el aprendizaje supervisado, semi-supervisado y no supervisado [44]. A continuación, se describe brevemente la idea principal de cada una de éstas:

- Las técnicas de aprendizaje supervisado emplean datos de entrenamiento completamente etiquetados, es decir, se distingue entre las instancias regulares y las atípicas (frecuentemente ambas clases muy desbalanceadas). Por esta razón, usualmente estas técnicas se apoyan de generar modelos de clasificación capaces de detectar anomalías en datos que no han sido vistos previamente [46].

- Las técnicas de aprendizaje semi-supervisado emplean datos de entrenamiento que solo poseen etiquetas para las instancias regulares o "normales". Por lo tanto, en estos casos, el enfoque usualmente es modelar este comportamiento "normal" y, posteriormente, clasificar como anómalas aquellas nuevas instancias que se desvíen significativamente de este modelo [46].

- Las técnicas de aprendizaje no supervisado no emplean datos de entrenamiento, es decir, no dependen de un modelo pre-entrenado para evaluar el grado de anomalía de los datos. En cambio, logran esta tarea empleando

únicamente las propiedades intrínsecas presentes en el conjunto de datos bajo análisis, por ejemplo, las distancias entre elementos o la densidad de elementos en el espacio [46]. Estas técnicas suelen emplearse comúnmente en el análisis exploratorio de datos, permitiendo detectar anomalías que luego se someten a un análisis más detallado [45, p. 4].

Debido a que el conjunto de genomas a analizar en la metodología carece de etiquetas, es decir, no se tiene una noción previa de las particularidades que distinguen a un genoma anómalo. Lo que se pretende hacer es utilizar únicamente las características propias presentes en este conjunto de datos para identificar aquellos genomas más atípicos del virus SARS-CoV-2. Por lo tanto, en esta tesis se emplean algunos algoritmos de detección de anomalías basados en el aprendizaje no supervisado. La Sección 4.3 aborda en detalle los algoritmos que se utilizarán.

4.2. Tipos de anomalías

Otra de las características importantes que debe tenerse en cuenta para elegir la técnica de detección de anomalías más adecuada para nuestro conjunto de datos es el tipo de anomalías que se pretenden identificar. Siguiendo la clasificación propuesta en *Goldstein y Uchida* [44] y en *Chandola et al.*[46], existen tres categorías principales de anomalías: *anomalías puntuales*, *anomalías contextuales* y *anomalías colectivas*. En pocas palabras, estas se definen de la siguiente manera:

- Las *anomalías puntuales* son aquellas instancias individuales que se consideran anómalas con respecto al resto del conjunto de datos. La mayoría de las técnicas de detección se enfocan en este tipo de anomalías [46]. Un ejemplo de *anomalías puntuales* son las instancias A_0, A_1, A_2, A_3 y A_4, mostradas en la Figura 4.1, donde claramente se observa que estas instancias no concuerdan con el valor de los atributos espaciales del resto de datos.

- Las *anomalías contextuales* son aquellas instancias donde su grado de anomalía se determina en función de un contexto específico. Estas se trabajan comúnmente en series de tiempo, donde el contexto es temporal [46]. Por ejemplo, una temperatura baja en los meses de invierno se considera un comportamiento habitual. Sin embargo, la misma temperatura registrada en los meses de verano se podría considerar como una anomalía.

- Las *anomalías colectivas* se refieren a un conjunto de instancias que se identifican como un evento anómalo. En este caso, no se consideran necesariamente anómalas las instancias individuales, sino que la anomalía consiste en la ocurrencia conjunta de todas ellas [46]. Por ejemplo, en el campo de la

ciberseguridad, la ocurrencia de varias acciones en conjunto podría indicar alguna intrusión en un sistema y, por consiguiente, ser una anomalía.

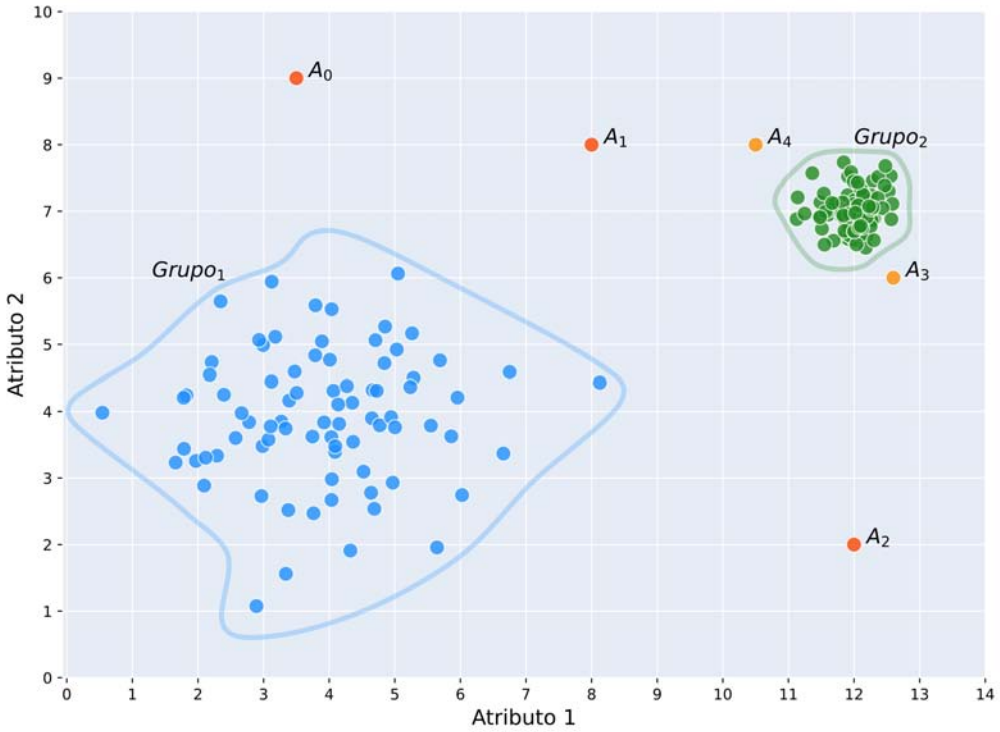

Figura 4.1: Ejemplo de *anomalías puntuales* en dos dimensiones. Las instancias A_0, A_1 y A_2 se categorizan como *anomalías globales*, mientras que las instancias A_3 y A_4 se consideran *anomalías locales*.

Aunque se ha señalado que las *anomalías puntuales* son las instancias individuales que presentan las mayores desviaciones con respecto al resto de los datos, en muchas ocasiones, identificar una anomalía bajo esta descripción resulta no ser del todo clara. Por consiguiente, se hace una distinción entre *anomalías globales* y *anomalías locales* [44]. Por ejemplo, en la Figura 4.1, las instancias A_0, A_1 y A_2 se categorizan como *anomalías globales* por ser muy diferentes respecto a los dos grupos principales. En cambio, las instancias A_3 y A_4 se categorizan como *anomalías locales* al ser anómalas con respecto a sus vecinos más cercanos.

En este trabajo, los genomas del virus SARS-CoV-2 que se emplean se representan como su uso de codones. Por lo tanto, cada genoma puede considerarse en un espacio numérico multidimensional definido por cada uno de los codones. Específicamente, como se menciona en el Capítulo 3, todos los genomas bajo esta representación se encuentran en un espacio simplicial de 63 dimensiones.

El objetivo principal radica en identificar los genomas del virus que presentan una mayor divergencia dentro de este espacio simplicial, en comparación con los demás genomas. Por esta razón, en la tesis se opta por emplear técnicas para detectar *anomalías puntuales*, eligiendo algunas que muestran un mejor rendimiento en la detección de *anomalías globales* y otras en el caso de *anomalías locales*.

4.3. Algoritmos de aprendizaje no supervisado

Los algoritmos de detección de anomalías no supervisados parten del supuesto de que en un conjunto de datos el número de instancias no anómalas es considerablemente mayor que el de instancias anómalas [46]. Tomando en consideración esta premisa, a día de hoy se han desarrollado diversos algoritmos con distintas estrategias computacionales, los cuales son capaces de llevar a cabo la tarea de detección. Asimismo, aunque puede resultar complejo organizar todos los algoritmos que han sido propuestos en grupos específicos, [44] propone algunas categorías potenciales para su clasificación. Esta clasificación se fundamenta en la estrategia computacional implementada por el algoritmo en cuestión.

Por consiguiente, es posible clasificar inicialmente los algoritmos de detección de anomalías no supervisados en cuatro categorías (Fig. 4.2):

- Los *basados en la agrupación*: Este tipo de algoritmos primero generan agrupaciones con las instancias más similares entre sí. Luego, consideran como anomalías aquellas instancias que no forman parte de ninguna agrupación o que se encuentran más distantes del centroide de su agrupación correspondiente [46].

- Los *basados en los vecinos más cercanos*: Este tipo de algoritmos asumen que las instancias regulares tienden a presentarse en vecindarios con una alta densidad, mientras que las anomalías son aquellas instancias que se encuentran más alejadas de sus instancias vecinas más cercanas [46].

- Los *basados en técnicas estadísticas*: Estos algoritmos se caracterizan por ajustar en primer lugar un modelo estadístico a los datos. Luego, asumen que las instancias anómalas tienden a presentarse en regiones de baja probabilidad dentro del modelo, mientras que las instancias regulares (no anómalas) tienden a presentarse en regiones de alta probabilidad [46].

- Los *basados en técnicas subespaciales*: En principio, estos algoritmos crean una representación de los datos en un subespacio de baja dimensionalidad, en donde luego asumen que las instancias normales se distinguen de manera más clara de las instancias anómalas [46].

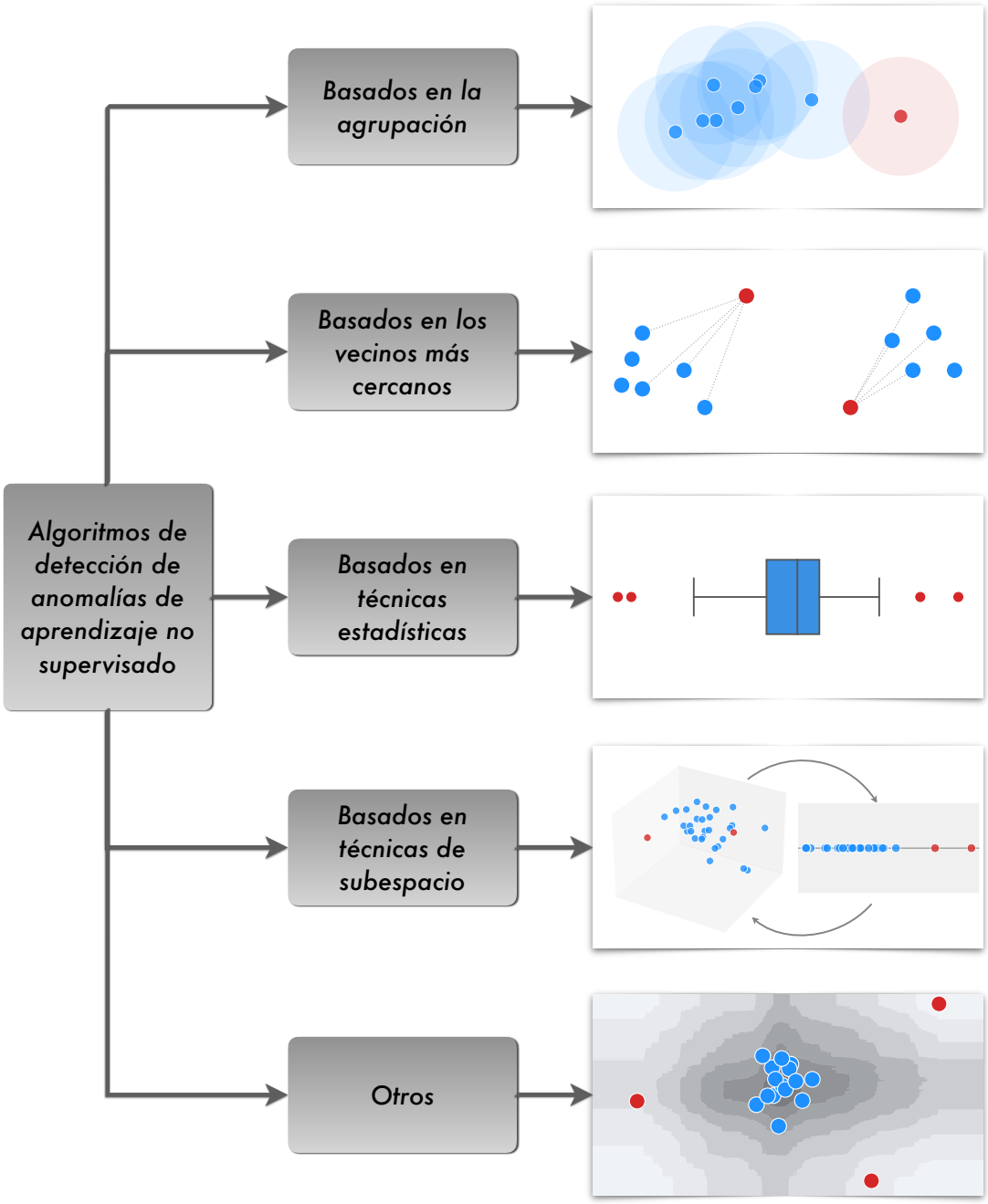

Figura 4.2: Clasificación de los algoritmos de detección de anomalías de aprendizaje no supervisado, junto con ilustraciones representativas de cada categoría. Los elementos de color rojo representan instancias anómalas, mientras que los de color azul representan instancias regulares o "normales".

Es importante señalar que existen algoritmos que no encajan directamente en las categorías mencionadas. Esto aplica al algoritmo del *bosque de aislamiento* (ilustrado en la Figura 4.2 como ejemplo de *otros*), así como a aquellos algoritmos que hacen uso de redes neuronales o máquinas de soporte vectorial [44].

Al igual que en otros tipos de técnicas de detección de anomalías, algunos algoritmos de aprendizaje no supervisado tienen la capacidad de etiquetar directamente las instancias anómalas. Es decir, pueden realizar una distinción binaria entre los elementos regulares y los atípicos. Sin embargo, la mayor parte de los algoritmos de aprendizaje no supervisado asigna un grado o puntaje que indica el nivel de anomalía de cada instancia [45, p. 2].

Cuando el algoritmo solo proporciona un puntaje de anomalía, le corresponde al usuario evaluar y seleccionar aquellos elementos con la puntuación más alta utilizando un umbral adecuado [43]. Este último hecho implica un desafío significativo, ya que suele ser complicado llevar a cabo una evaluación rigurosa para determinar si el número de anomalías seleccionadas es el más apropiado [45].

En el resto de esta sección se describe el funcionamiento de los algoritmos empleados en la metodología de este trabajo. En primer lugar, se aborda el algoritmo *DBSCAN* y el algoritmo *k-medias* para la detección de anomalías, ambos *basados en la agrupación* de datos. Luego, se presenta el algoritmo del *valor atípico local* (*LOF*), el cual adopta un enfoque *basado en los vecinos más cercanos*. Posteriormente, se analiza el algoritmo del *bosque de aislamiento*, cuya estrategia computacional difiere de las mencionadas. Finalmente, se abordan otras estrategias que emplean algunas de las herramientas mencionadas en el Capítulo 3.

4.3.1. DBSCAN

El propósito principal del algoritmo de *agrupamiento espacial basado en densidad para aplicaciones con ruido* (*DBSCAN*, por sus siglas en inglés) consiste en descubrir agrupaciones dentro de un conjunto de datos, de manera similar al algoritmo *k-medias* mencionado en el Capítulo 3. Sin embargo, a diferencia de *k-medias*, *DBSCAN* no limita que las agrupaciones estén contenidas en una celda de voronoi, sino que permite que adopten formas arbitrarias en el espacio [47].

Este algoritmo asume que las agrupaciones se conforman de una densidad característica de puntos (instancias) significativamente mayor que fuera de dichas agrupaciones. Donde al mismo tiempo, los puntos que se localizan en las áreas de baja densidad (es decir, fuera de las agrupaciones) se identifican como *ruido* [47].

Para los propósitos de esta tesis, nos interesamos particularmente en este *ruido*, ya que estas instancias también pueden ser clasificadas como anomalías al no formar parte de ninguna agrupación, siguiendo la premisa de los algoritmos de detección de anomalías no supervisado *basados en la agrupación*.

La idea de este algoritmo consiste en que cada punto dentro de una agrupación debe tener un número mínimo de puntos (*MinPts*) en su vecindario. Este vecindario se define mediante un radio *eps* alrededor del punto en cuestión (Fig. 4.3) y su forma se determina por la función de distancia elegida [47]. Por ejemplo, en la Figura 4.3 se emplea la *distancia euclidiana* para definir esta forma.

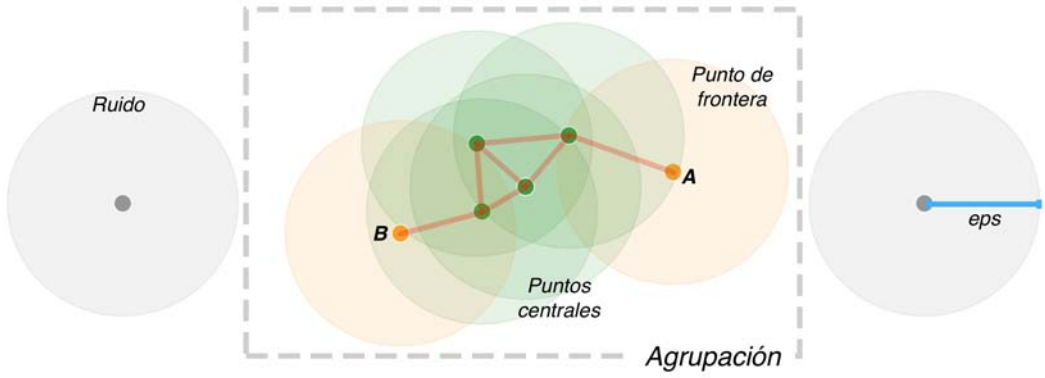

Figura 4.3: Conceptos generales del algoritmo *DBSCAN*. Utilizando la *distancia euclidiana*, un radio *eps* y un número mínimo de puntos vecinos *MinPts*. Los *puntos centrales* se identifican en color verde, los *puntos de frontera* en color naranja, mientras que el *ruido* está representado por aquellos puntos de color gris.

No obstante, para *DBSCAN*, cada agrupación se compone de dos tipos de puntos: los *puntos centrales* y los *puntos de frontera* (Fig. 4.3). Los *puntos centrales* se identifican como aquellos puntos dentro de la agrupación que deben cumplir con tener un número mínimo de puntos en su vecindario, como se mencionó previamente. En cambio, los *puntos de frontera* se distinguen al encontrarse en la frontera de las agrupaciones, donde la densidad de puntos vecinos es considerablemente menor. Por lo tanto, este tipo de puntos solo requieren estar en la vecindad de otro de los puntos de la agrupación, sin necesidad de cumplir con un número mínimo de puntos en su vecindario [47].

Para formar una agrupación, es necesario que todos los puntos que la componen estén conectados entre sí. En otras palabras, debe existir al menos una secuencia de puntos que estén directamente conectados entre cualquier par de puntos que no estén directamente conectados. Se considera que un punto está directamente conectado a otro si este se encuentra en su vecindario [47]. Por ejemplo, en la agrupación mostrada en la Figura 4.3, es posible identificar una secuencia de puntos directamente conectados entre dos puntos que no lo están, como sucede en el caso de los puntos A y B.

De esta manera, se puede decir que *DBSCAN* genera un grafo por cada agrupación identificada, conectando las instancias que conforman a éstas. Aquellas instancias que no pudieron ser conectadas a ninguna de las agrupaciones generadas son etiquetadas como *ruido* (Fig. 4.3), o en nuestro caso, anomalías. Por lo tanto, este algoritmo proporciona directamente una etiqueta para identificar las instancias atípicas, sin requerir una mayor intervención por parte del usuario.

Un ejemplo concreto del funcionamiento del algoritmo *DBSCAN* se muestra en la Figura 4.4. En esta figura se presenta un conjunto de datos bidimensional, junto con las etiquetas que el algoritmo asignó a cada instancia. Además, se muestra el radio *eps* alrededor de cada punto. Las etiquetas se obtuvieron empleando un radio de vecindad ($eps = 0.8$) y un número mínimo de puntos vecinos ($MinPts = 2$) específicos. Los resultados muestran cómo el algoritmo identificó tres grupos distintos y etiquetó cinco instancias como *ruido*.

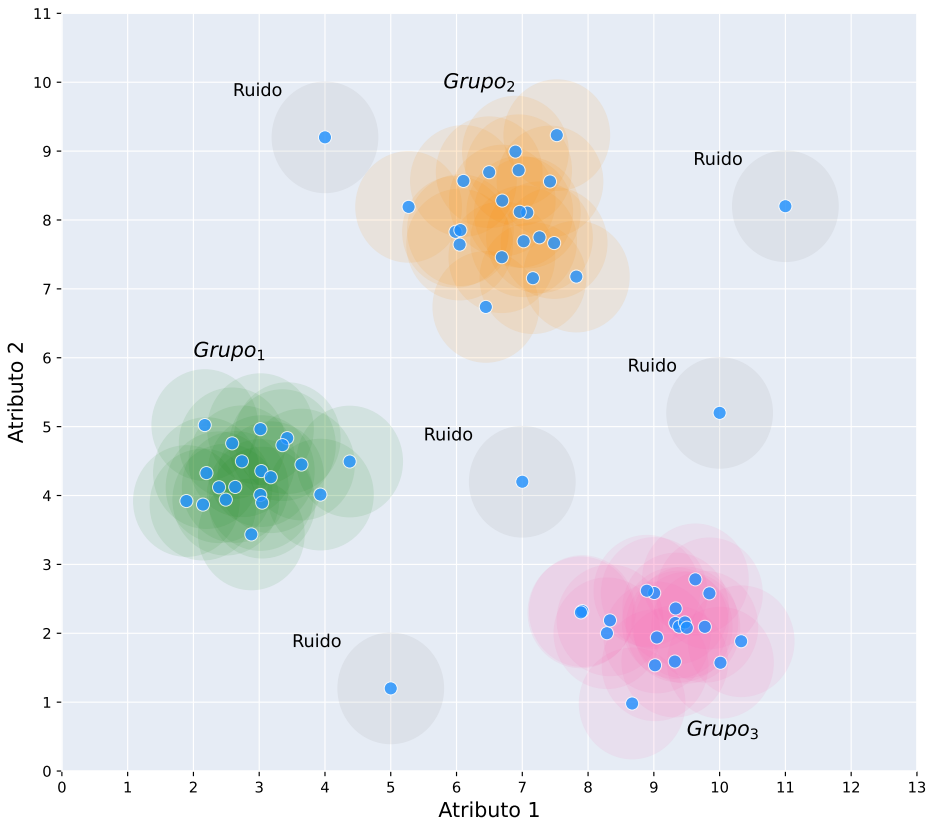

Figura 4.4: Algoritmo *DBSCAN* aplicado a un conjunto de datos bidimensional. Se muestran los grupos y ruido identificados por el algoritmo con un valor de *eps* igual a 0.8 y *MinPts* igual a 2.

4.3.2. Detección con el algoritmo de k-medias

Al igual que el algoritmo *DBSCAN*, la detección de anomalías mediante el algoritmo *k-medias* se fundamenta en la agrupación de datos. No obstante, este método difiere en su enfoque. En lugar de considerar como anomalías aquellas instancias que no están asignadas a ninguna agrupación, éste evalúa la proximidad de cada instancia a su centroide más cercano. Así, asigna un nivel de anomalía a cada una de estas instancias [5, p. 49-51].

En concreto, este método consiste en emplear el algoritmo de *k-medias*, mencionado en el Capítulo 3, para identificar la agrupación respectiva de cada elemento. Posteriormente, se calcula la distancia entre cada uno de estos elementos y su centroide respectivo. Esta distancia se interpreta como un puntaje de anomalía. Cuanto mayor sea el valor, es decir, mientras más alejado esté el elemento de su centroide, más anómalo será considerado éste. Finalmente, es necesario elegir un umbral adecuado para poder identificar las instancias anómalas [46].

La Figura 4.5 muestra la implementación de este método en un conjunto de datos bidimensional, donde se pueden apreciar claramente cuatro agrupaciones. El puntaje de anomalía de cada instancia bajo esta estrategia depende del centroide de la agrupación a la cual pertenece (estrellas amarillas). Esto permite identificar de manera más eficiente las *anomalías locales* dentro del conjunto de datos.

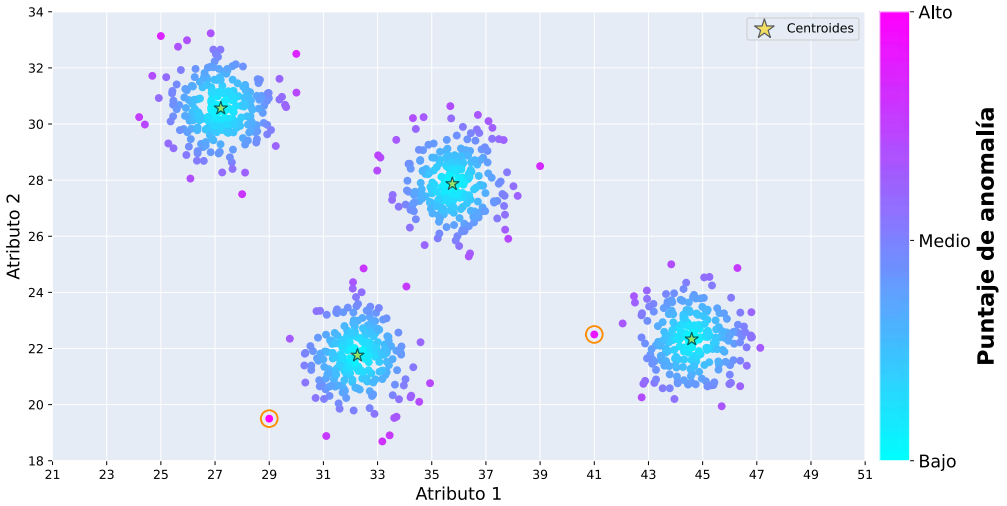

Figura 4.5: Detección de anomalías con el algoritmo de *k-medias* en un conjunto de datos bidimensional. El color de cada instancia indica su puntaje de anomalía, mientras que las instancias anómalas, clasificadas por un umbral, se marcan con un círculo naranja alrededor de ellas.

La Figura 4.6 muestra el histograma de todos los puntajes de anomalías obtenidos. En este caso, se ha seleccionado un umbral de 3.5, lo que resulta en la clasificación de las 2 instancias anómalas, mostradas en la Figura 4.5.

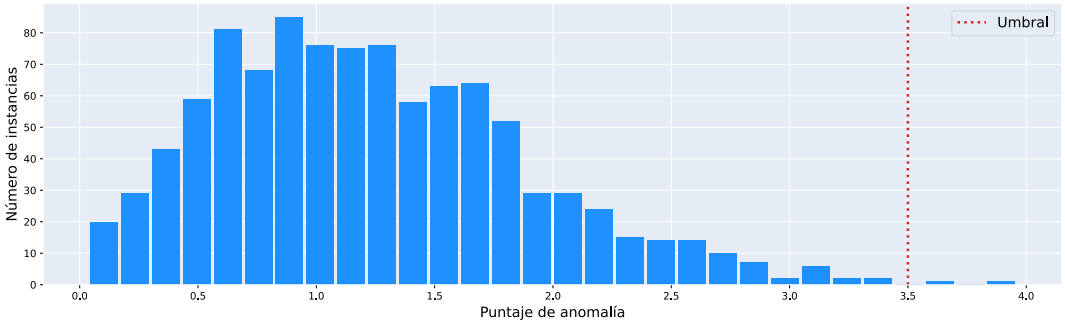

Figura 4.6: Histograma de los puntajes de anomalía mostrados en la Figura 4.5, junto con el umbral seleccionado para clasificar las instancias anómalas.

4.3.3. Valor atípico local (LOF)

El *valor atípico local* (LOF, por sus siglas en inglés) es un algoritmo de aprendizaje no supervisado *basado en los vecinos más cercanos*. El valor atípico computado por el algoritmo se considera local, ya que se calcula tomando en cuenta únicamente un vecindario restringido alrededor de cada elemento.

En esencia, el algoritmo *LOF* inicia calculando una densidad local para cada elemento del conjunto de datos. Esta densidad se obtiene empleando las distancias entre el elemento en cuestión y sus k vecinos más cercanos. Posteriormente, el *valor atípico local* o puntaje de anomalía de cada elemento se obtiene al comparar su densidad local con las densidades locales de sus k vecinos más cercanos. La idea principal es que los elementos regulares o "normales" cuentan una densidad local similar a la de sus vecinos más cercanos, mientras que la densidad local de los elementos anómalos es inferior a la de estos vecinos [46].

En concreto, la metodología que sigue *LOF*, propuesto por *Breunig et al.* [48], consiste en primero obtener el conjunto de los k vecinos más cercanos al elemento p, denominado $N_k(p)$. Luego, el algoritmo emplea cada uno de estos vecinos para calcular la densidad local del elemento p. A esta medida se le conoce como $lrd_k(p)$ o *densidad local de alcance* de p, la cual se define de la siguiente manera:

$$lrd_k(p) = 1/\left(\frac{\sum_{q \in N_k(p)} dist\text{-}alcance_k(p,q)}{|N_k(p)|}\right) \tag{4.1}$$

Esta *densidad local de alcance* es el inverso del promedio de las *distancias de alcance*. Donde la *distancia de alcance* entre un elemento p y un elemento q es simplemente la distancia entre ellos. Sin embargo, si estos elementos se encuentran muy cercanos, su *distancia de alcance* será la distancia entre q y el k vecino más cercano de q, distancia denominada k-dist(q). Esta operación se realiza con el propósito de reducir las fluctuaciones estadísticas de todas las distancias entre los elementos p cercanos a q. Por lo tanto, la *distancia de alcance* se define como:

$$dist\text{-}alcance_k(p,q) = max\{k\text{-}dist(q), d(p,q)\} \tag{4.2}$$

Finalmente, el *valor atípico local* del elemento p es el promedio de las *densidades locales de alcance* de sus k vecinos más cercanos entre su *densidad local de alcance*:

$$LOF_k(p) = \left(\frac{\sum_{q \in N_k(p)} \frac{ldr_k(q)}{ldr_k(p)}}{|N_k(p)|} \right) \tag{4.3}$$

Examinando la ecuación 4.3 se puede observar que elementos con una *densidad local de alcance* similar a la de sus vecinos obtienen un puntaje de *LOF* de alrededor de 1, mientras que elementos con una *densidad local de alcance* inferior a las de estos vecinos obtienen un puntaje más alto [44].

Un ejemplo de la aplicación de este algoritmo se ilustra en la Figura 4.7. En este caso, se emplea el conjunto de datos previamente mostrado en la Sección 4.2, y el algoritmo *LOF* se aplica con un valor de k vecinos igual a 5. En la figura el *valor atípico local* obtenido para cada elemento se representa por el tamaño de los círculos de color naranja alrededor de cada punto. Asimismo, se presentan los valores atípicos exactos para algunos elementos sobresalientes.

Al analizar la Figura 4.7, se observa cómo el algoritmo es capaz de identificar tanto *anomalías globales* como *anomalías locales* dentro del conjunto de datos. Estos elementos anómalos son evaluados por el algoritmo con puntajes de *LOF* más altos en comparación con el resto, mientras que los elementos "normales", localizados en regiones densas del espacio, obtienen un puntaje de *LOF* aproximadamente de 1, como se aprecia en los elementos de color negro de la figura.

La Figura 4.8 presenta el histograma de todos los *valores atípicos locales* calculados, junto con un umbral seleccionado para clasificar los elementos anómalos del conjunto. En este caso, el umbral elegido tiene un valor de 3, lo que resulta en la clasificación de 5 elementos anómalos. Estos elementos se presentan de color naranja en la Figura 4.7 junto con sus puntajes de *LOF* correspondientes.

Como se aprecia en la Figura 4.7, este algoritmo resulta eficiente para detectar anomalías en datos con regiones de diferentes densidades. Esto se debe a que, como se mencionó previamente, el *valor atípico* calculado depende de la densidad local del elemento, que a su vez se basa en los vecinos más cercanos de éste.

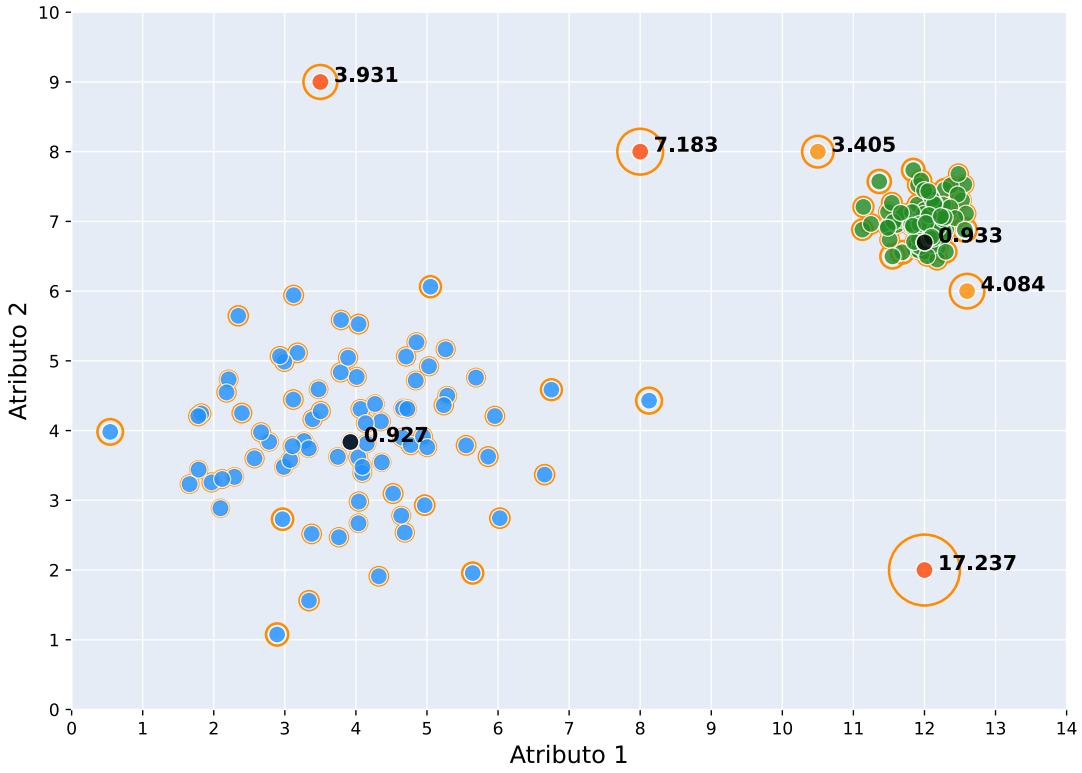

Figura 4.7: Algoritmo *LOF* aplicado al conjunto de datos mostrado en la Sección 4.2. El tamaño de los círculos alrededor de las instancias representa su puntaje. Los puntajes de *LOF* exactos se muestran para algunos elementos sobresalientes.

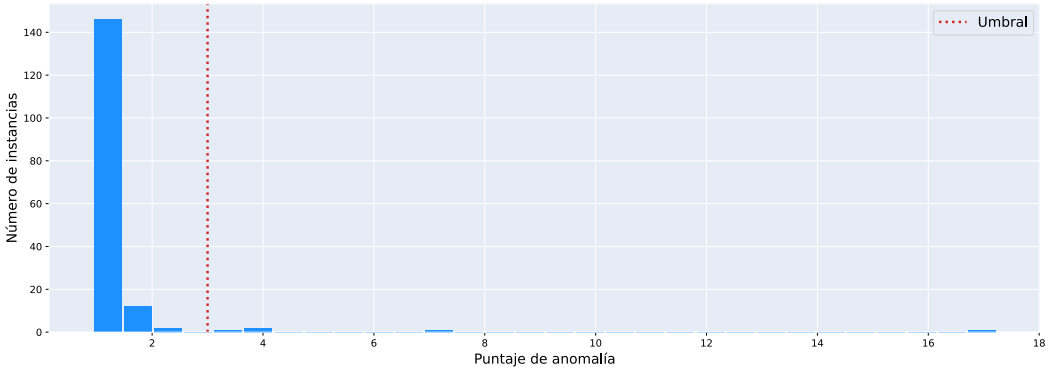

Figura 4.8: Histograma de los puntajes de *LOF* mostrados en la Figura 4.5, junto con el umbral seleccionado para clasificar las instancias anómalas.

4.3.4. Bosque de aislamiento

Este algoritmo, propuesto por *Liu et al.* [49], emplea una estrategia computacional diferente a las mencionadas al inicio de esta sección (Fig. 4.2). Su objetivo principal es aislar las instancias anómalas del resto de las instancias. Donde a diferencia de los algoritmos discutidos previamente, este no se fundamenta en el cálculo de distancias ni en medidas de densidad para lograrlo.

Este algoritmo utiliza un conjunto de árboles binarios para calcular el puntaje de anomalía de cada instancia del conjunto de datos. Cada árbol binario, también conocido como *árbol de aislamiento*, se construye generando particiones aleatorias en el espacio de atributos de forma recursiva hasta lograr aislar cada una de estas instancias. El número de particiones necesarias para aislar a una instancia en particular es equivalente al puntaje de anomalía que recibe ésta [49].

Específicamente, dado un conjunto de datos $\mathbf{X} = \{\mathbf{x}_i \in \mathbb{R}^Q\}_{i=1}^n$, primeramente se selecciona de forma aleatoria un atributo q de los Q atributos del conjunto y un valor aleatorio p dentro del rango de valores de este atributo. El *árbol de aislamiento*, se inicia en un nodo raíz, donde los datos se dividen en dos grupos según la condición $q \leq p$. Para cada una de estas nuevas divisiones, se crea un nuevo nodo en el árbol, donde se selecciona un nuevo valor de q y p para dividir nuevamente estos grupos. Este proceso se repite hasta que las divisiones generadas solo contengan una única instancia. Por lo tanto, para un conjunto con n instancias distintas, se requieren $n-1$ nodos para aislar cada una de ellas [49].

Dado que se asume que las anomalías son poco comunes y diferentes del resto de datos, se espera que estas sean aisladas en pocas particiones, lo que equivale a una longitud de camino corta dentro del *árbol de aislamiento*. Por otro lado, las instancias "normales", que requieren más particiones, se aíslan en los niveles más profundos del árbol, lo que implica una longitud de camino más larga.

Un ejemplo de un *árbol de aislamiento* parcial se presenta en la Figura 4.9. La parte derecha de la figura muestra el árbol generado para aislar los datos mostrado en la parte izquierda. En esta figura, se ejemplifica una instancia anómala frente a una instancia "normal". La instancia anómala (nodo amarillo) está aislada más cerca del nodo raíz del árbol, con una longitud de camino igual a 2, mientras que la instancia normal (nodo rosa) tiene una longitud de camino igual a 9.

El algoritmo del *bosque de aislamiento* emplea múltiples *árboles de aislamiento* con el propósito de calcular una longitud de camino promedio para cada instancia del conjunto de datos. Si una instancia constantemente produce una longitud de camino corta, es muy probable que se trate de una anomalía [49].

En la Figura 4.10, se observa como la longitud de camino promedio para aislar la instancia anómala y la instancia "normal", de la Figura 4.9, converge a 4.59 y 8.89, respectivamente, a medida que aumenta el número de árboles utilizados.

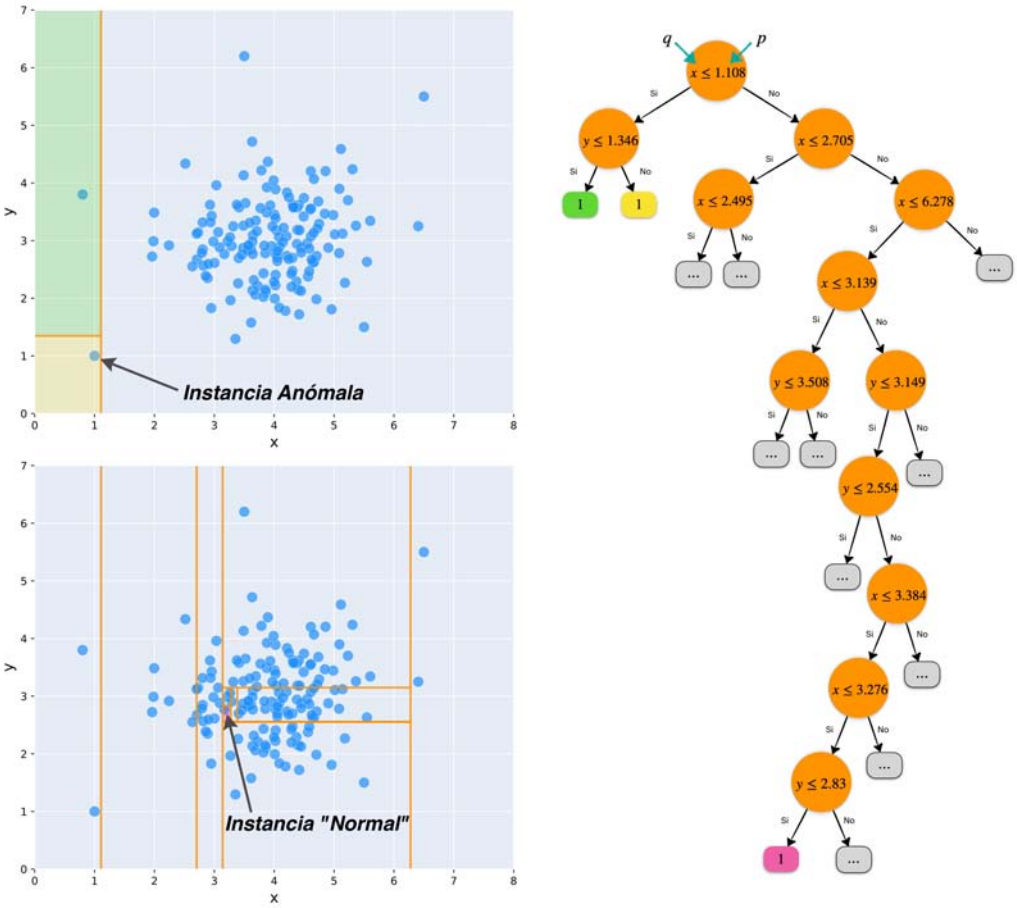

Figura 4.9: *Árbol de aislamiento* construido a partir de un conjunto de datos bidimensional. Izquierda: Particiones generadas en el espacio de atributos para aislar una instancia anómala y una instancia "normal". Derecha: Representación parcial del *árbol de aislamiento*.

La detección de anomalías mediante el algoritmo del *bosque de aislamiento* se divide en dos etapas. La primera etapa implica la construcción de los *árboles de aislamiento* utilizando los datos disponibles, tal como se explicó previamente. La segunda etapa se centra en obtener un puntaje de anomalía para cada instancia al pasar los datos a través de todos estos *árboles de aislamiento* [49].

Para calcular el puntaje de anomalía s de una instancia x mediante el algoritmo del *bosque de aislamiento*, [49] define la siguiente formula:

$$s(x, n) = 2^{-\frac{E(h(x))}{c(n)}} \tag{4.4}$$

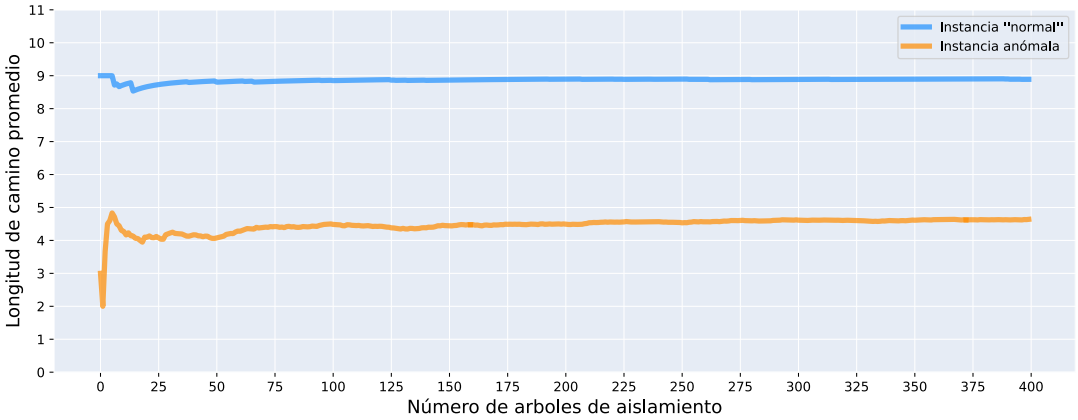

Figura 4.10: Longitud de camino promedio para aislar la instancia anómala y la instancia "normal", de la Figura 4.9, conforme se emplean más árboles.

donde $h(x)$ representa la longitud de camino en un árbol para aislar una instancia x en particular, mientras que $E(h(x))$ corresponde a la longitud de camino promedio para aislar a una instancia x empleando toda la colección de árboles.

El término $c(n)$ estima la longitud de camino promedio de un solo árbol para aislar a cualquier instancia x dado un conjunto de n instancias. Este valor se utiliza para normalizar $h(x)$. El origen de este término se encuentra en la equivalencia entre los *árboles de aislamiento* y los árboles binarios [49], y se define como:

$$c(n) = 2(\ln(n-1) + 0.5772156649) - (2(n-1)/n) \tag{4.5}$$

Examinando la ecuación 4.4, podemos notar que cuando $E(h(x)) \to 0$, entonces $s \to 1$. En otras palabras, las instancias anómalas, que en promedio requieren menos particiones para ser aisladas, obtienen un puntaje de anomalía muy cercano a uno. Por otro lado, cuando $E(h(x)) \to n-1$, entonces $s \to 0$. En este caso, si la instancia x se aísla en los niveles más profundos de los árboles, no se considera anómala y obtiene una puntuación de anomalía muy cercana a cero.

La Figura 4.11 presenta de manera visual los diferentes puntajes de anomalía obtenidos sobre una región específica del espacio de atributos. En donde el *bosque de aislamiento* utilizado para generar estos puntajes se construyó a partir del conjunto de datos que se muestra en la figura (puntos azules). Cuanto más oscuro sea el color de la región, mayor será su puntaje de anomalía.

Es importante destacar que una de las características más relevantes de este algoritmo es la posibilidad de construir cada *árbol de aislamiento* utilizando un submuestreo de los datos en lugar del conjunto de datos completo. Este hecho permite que el algoritmo pueda manejar grandes conjuntos de datos de manera eficiente y obtenga mejores resultados en ciertas situaciones [49].

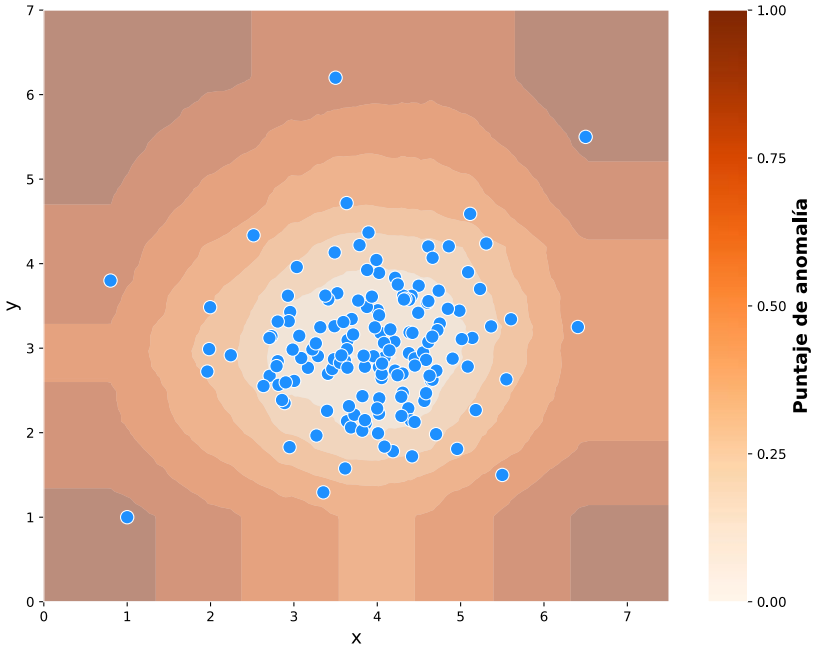

Figura 4.11: Puntuaciones de anomalía obtenidas mediante el algoritmo del *bosque de aislamiento* para una región específica en el espacio de atributos.

4.3.5. Otras estrategias

En adición a los algoritmos mencionados anteriormente, en este trabajo de tesis se implementan otras estrategias computacionales que utilizan diferentes criterios para identificar los genomas más atípicos del virus SARS-CoV-2. Estas estrategias emplean algunas de las herramientas discutidas en el Capítulo 3. En particular, el cálculo de distancias, la estimación de la *entropía* y la reducción de la dimensionalidad desempeñan un papel fundamental en estas.

La primer estrategia conlleva el cálculo de la *distancia euclidiana* y de la *distancia de Wasserstein* entre los diferentes usos de codones del virus y el uso de codones humano. Estas medidas, además de permitir analizar la evolución del uso de codones del virus en relación al del humano, también se utilizan como un criterio para identificar los genomas más atípicos del virus. Considerando anómalos a los genomas que muestran una mayor divergencia en estas medidas.

La segunda estrategia consiste en calcular la *distancia log-cociente esperada* para cada uso de codones del conjunto de datos y, posteriormente, utilizar esta medida como otro criterio para la detección de anomalías. Para computar esta distancia, en primer lugar, se aplica la transformación *log-cociente centrado* a

todos los usos de codones, descrita en el Capítulo 3. Luego, se determina la *distancia euclidiana* esperada de las instancias ya transformadas. Finalmente, se identifican como anomalías aquellos genomas que estén más alejados del resto, es decir, que presenten una mayor *distancia log-cociente esperada*.

Una tercera estrategia implica calcular la *entropía* para cada uso de codones del virus, tal como se describe en el Capítulo 3. La *entropía*, además de ser un indicador del *sesgo en el uso de codones* de cada genoma, también sirve como criterio para seleccionar los genomas más atípicos del conjunto de datos. Estos genomas se caracterizan por presentar un valor de *entropía* y, en consecuencia, un *sesgo en el uso de codones* que difiere significativamente de los demás genomas.

Por último, otra estrategia empleada consiste en reducir la dimensionalidad de los datos, representando los diferentes usos de codones del virus con solo dos atributos. Esto se logra mediante el uso de algoritmos como *PCA* o *Isomap*. Posteriormente, se calcula la *distancia euclidiana* esperada de cada instancia en ese nuevo subespacio. Aquellas instancias que se encuentren más alejadas del resto se identifican como anómalas. Esta estrategia se puede clasificar como un método de detección de anomalías *basado en técnicas subespaciales* (Fig. 4.2), en el cual, después de obtener la representación de los datos en baja dimensionalidad, es más sencillo distinguir aquellas instancias anómalas.

Para cada una de las estrategias mencionadas, se calcula una medida que, junto con un umbral específico para cada una de ellas, permite detectar los genomas más anómalos del virus SARS-CoV-2.

4.4. Resumen del capítulo

En este capítulo, se ha proporcionado una breve definición de los objetivos de la detección de anomalías. Se abordó de manera general la motivación que impulsa la identificación de estos elementos anómalos en varios campos de estudio, reiterando la utilidad de detectar genomas anómalos para un mismo organismo o virus. Además, se describieron algunas características generales de la detección de anomalías y se detallaron los algoritmos de aprendizaje no supervisado que se emplearán en la metodología de este trabajo de tesis.

La Sección 4.1 abordó la clasificación general de las técnicas de detección existentes, incluyendo *técnicas de aprendizaje supervisado*, *semi-supervisado* y *no supervisado*, las cuales dependen de la disponibilidad de etiquetas en los datos. Dado que el conjunto de datos que se analizará carece de etiquetas, en este trabajo se han empleado técnicas de aprendizaje no supervisado.

La Sección 4.2 menciona los tipos de anomalías existentes, que incluyen *anomalías puntuales*, *anomalías contextuales* y *anomalías colectivas*. En este trabajo,

el enfoque se centra en la detección de *anomalías puntuales* debido a la naturaleza del conjunto de datos que se va a analizar, el cual consiste en genomas del virus SARS-CoV-2 representados por su uso de codones.

La primera parte de la Sección 4.3 presenta la clasificación de diversos algoritmos de aprendizaje no supervisado, según la estrategia computacional que implementan. Las cuales incluyen algoritmos *basados en la agrupación*, *basados en los vecinos más cercanos*, *basados en técnicas estadísticas*, *basados en técnicas de subespacio* y *otros* (los cuales adoptan estrategias diferentes a los anteriores).

La segunda parte de la Sección 4.3 detalla el funcionamiento de los algoritmos empleados en este trabajo. En resumen:

- Algoritmo *DBSCAN*: Este algoritmo genera diferentes agrupaciones con los elementos que se encuentran en las regiones más densas del espacio. Aquellos que no forman parte de ninguna agrupación se identifican como anómalos.
- Algoritmo de *k-medias* para la detección de anomalías: Este método agrupa el conjunto de datos en k grupos utilizando el algoritmo de *k-medias*. Luego, calcula la distancia de cada elemento al centroide de su agrupación respectiva, obteniendo un puntaje de anomalía para ese elemento.
- Algoritmo del *valor atípico local* (*LOF*): Este método compara la densidad local de cada elemento con la de sus k vecinos más cercanos. Si su densidad local es inferior a la de sus vecinos, su puntaje de anomalía es más alto.
- Algoritmo del *bosque de aislamiento*: Este algoritmo aísla cada instancia del conjunto de datos utilizando árboles binarios que generan particiones recursivas en el espacio de atributos. Entre menos particiones se necesiten para aislar a una instancia, mayor será su puntaje de anomalía.

La última parte de la Sección 4.3 describe otras estrategias para la detección de genomas anómalos del virus SARS-CoV-2, algunas de las cuales son más específicas para este conjunto de datos. Estas estrategias se desarrollaron haciendo uso de herramientas previamente mencionadas en el Capítulo 3.

El próximo capítulo detalla el procedimiento propuesto en la metodología y presenta los resultados obtenidos. Finalmente, el Capítulo 6 aborda las conclusiones generales de este trabajo.

Capítulo 5

Resultados y discusión

En este capítulo se describe de manera detallada los resultados obtenidos al implementar la metodología descrita en el Capítulo 1. Toda la implementación se realizó utilizando el lenguaje de programación *Python* y las principales bibliotecas utilizadas fueron: *NumPy* [50], *SciPy* [51], *Pandas* [52], *Matplotlib* [53] y *Scikit-learn* [54]. El código se generó utilizando una serie de *Jupyter Notebooks*, los cuales siguen las mismas secciones mencionadas en este capítulo. La colección de *Jupyter Notebooks* y el archivo de datos empleado se encuentran disponibles aquí.

Siguiendo la metodología descrita en la introducción, en la Sección 5.1 se aborda la obtención y preprocesamiento del conjunto de datos a utilizar. En la Sección 5.2 se obtienen las medidas de *distancia* y *entropía* empleadas en el conjunto de datos, describiendo las diferencias por regiones geográficas y su evolución temporal. En la Sección 5.3 se aplican los métodos de reducción de dimensionalidad, *PCA* e *Isomap*, con el fin de examinar la variación del conjunto de datos bajo las nuevas representaciones. En la Sección 5.4 se implementan los algoritmos de detección de anomalías mencionados en el Capítulo 4. Finalmente, la Sección 5.5 presenta un resumen con los genomas que fueron identificados con mayor frecuencia como anomalías por todos los criterios empleados.

5.1. Conjunto de datos

El conjunto de datos crudos se obtuvo de GISAID [12], donde cada línea dentro de este archivo (Fig. 5.1, arriba) define la secuencia de nucleótidos para un gen en específico, así como la descripción del genoma del cual proviene (institución de secuenciación, fecha de secuenciación, identificador del genoma proveniente, estado de la República del cual proviene, entre otros). Este conjunto de datos incluye secuencias desde el 1 de enero de 2020 al 16 de junio de 2022.

Para generar el conjunto de datos a utilizar, en primer lugar se procesan estas secuencias para crear una tabla que represente el archivo original en su totalidad. Cada fila de esta tabla, contiene el uso de codones de cada gen, junto con su descripción (Fig. 5.1, abajo). El objetivo final del preprocesamiento es obtener los genomas completos de cada virus SARS-CoV-2 secuenciado, agregando todos los usos de codones de los genes identificados para un mismo individuo.

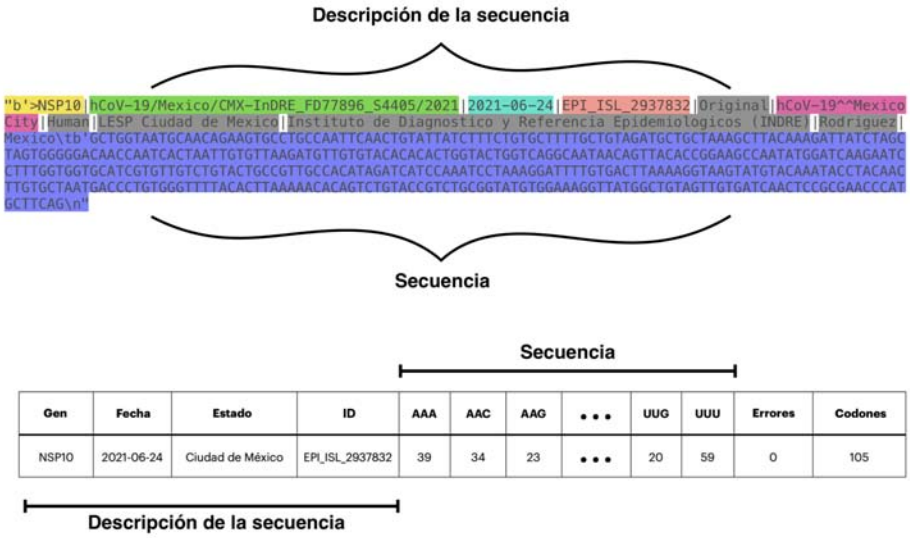

Figura 5.1: Preprocesamiento de una secuencias del archivo de GISAID [12]. Arriba: Línea de texto con la secuencia de nucleótidos de un gen y sus descripciones respectivas. Abajo: Secuencia preprocesada representada como fila de una tabla.

5.1.1. Preprocesamiento y extracción del uso de codones

El archivo original contiene 1,415,436 genes secuenciados (líneas de texto). Cada una de estas líneas se divide y filtra individualmente para formar una fila (Fig. 5.1, abajo) con su descripción y su secuencia de nucleótidos representada como uso de codones. Las filas obtenidas se unen para formar la Tabla 5.1.

En esta tabla, la columna *ID* indica el genoma del cual proviene el gen. Las columnas del 0 al 63 especifican el uso de codones. La columna *Errores* contiene los codones que no se pudieron obtener debido a nucleótidos no especificados en la secuencia de texto original, mientras que la columna *Codones* contiene el número de codones que sí se obtuvieron. Por lo tanto, el total de codones de la secuencia es la suma de *Errores* y *Codones*.

	Gen	Fecha	Estado	ID	0	1	...	62	63	Errores	Codones
0	Spike	2020-11-23	Ciudad de México	EPI_ISL_1005568	39	34	...	20	59	0	1274
1	Spike	2020-12-21	Ciudad de México	EPI_ISL_1005563	39	34	...	20	59	0	1274
2	Spike	2020-12-23	Ciudad de México	EPI_ISL_1005576	38	34	...	20	59	0	1274
...	...	...	...	...	...	...	...	...	...	...	...
1415433	N	2022-02-21	Estado de Mexico	EPI_ISL_13560138	22	6	...	9	3	0	417
1415434	NS9b	2022-02-21	Estado de Mexico	EPI_ISL_13560138	3	3	...	2	0	0	95
1415435	NS9c	2022-02-21	Estado de Mexico	EPI_ISL_13560138	0	2	...	2	0	0	74

Tabla 5.1: Conjunto de secuencias de genes preprocesados. Cada renglón contiene las descripciones y el uso de codones de diferentes genes.

5.1.2. Obtención del genoma completo y limpieza de datos

Para obtener el uso de codones de los genomas virales completos, se procedió a agrupar los genes de la Tabla 5.1 con el mismo identificador de genoma (columna *ID*). Solo se consideraron las agrupaciones que contienen todos los genes mencionados en el Capítulo 2. Aquellas agrupaciones que no cumplen con este criterio fueron descartadas.

El uso de codones del genoma es el resultado de agregar el uso de codones de todos los genes dentro de cada agrupación. Del mismo modo, el error (codones sin especificar) se calcula sumando todos los errores de estos genes. Si este error supera el 10 % del total de codones, se descarta la secuencia correspondiente. La Figura 5.2 muestra estas secuencias descartadas por estado. A continuación, se aplica la operación de cierre mencionada en el Capítulo 3 a los genomas restantes, generando así un conjunto de genomas del virus SARS-CoV-2 representados por la frecuencia relativa de su uso de codones. El conjunto de datos a utilizar se presenta en la Tabla 5.2, que contiene 51,322 instancias y 69 atributos.

	Fecha	ID	Estado	0	...	61	62	63	Errores	Codones
0	2020-01-01	EPI_ISL_913918	Oaxaca	0.035689	...	0.013952	0.019816	0.033667	9	9891
1	2020-01-01	EPI_ISL_914878	Tamaulipas	0.035963	...	0.014042	0.019699	0.033842	1	9899
2	2020-01-01	EPI_ISL_933664	Estado de Mexico	0.035869	...	0.013944	0.019905	0.033849	3	9897
...	...	...	...	...	...	...	...	...	...	...
51319	2022-06-15	EPI_ISL_13454152	Ciudad de México	0.036473	...	0.014122	0.019811	0.034034	43	9843
51320	2022-06-16	EPI_ISL_13454153	Ciudad de México	0.036329	...	0.014066	0.019632	0.033900	1	9882
51321	2022-06-16	EPI_ISL_13454154	Ciudad de México	0.036310	...	0.013857	0.019723	0.034287	1	9887

Tabla 5.2: Conjunto de datos a utilizar. Cada fila contiene la descripción y el uso de codones como frecuencia relativa de un genoma completo del virus SARS-CoV-2.

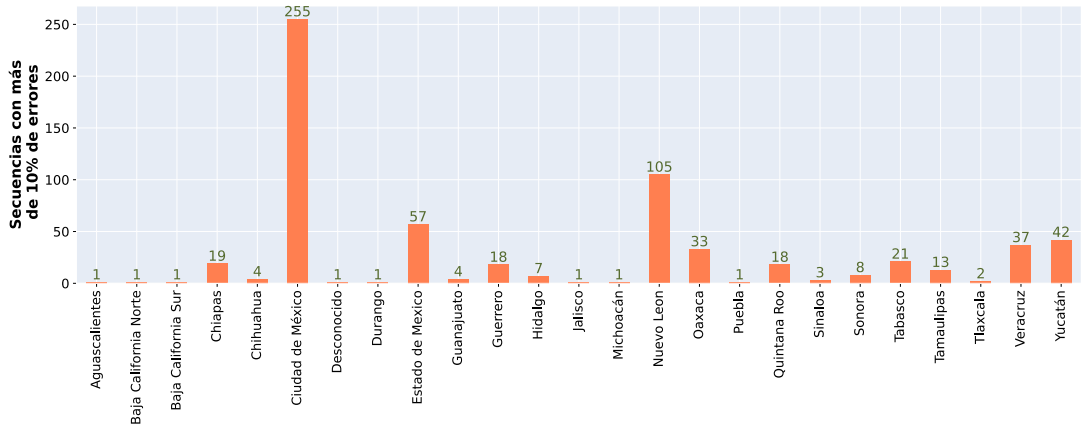

Figura 5.2: Secuencias con más del 10 % del total de codones sin especificar.

En la Figura 5.3, se presenta la distribución de los genomas del conjunto de datos a utilizar por estado. La Ciudad de México es la región con la mayor cantidad de genomas, alrededor de 25 % del conjunto de datos, seguida por el Estado de México y Yucatán. Por otro lado, se observa que Tlaxcala, Nayarit y Durango cuentan con la menor cantidad de estos. Además, se registra una pequeña cantidad de genomas cuyo estado de origen es desconocido.

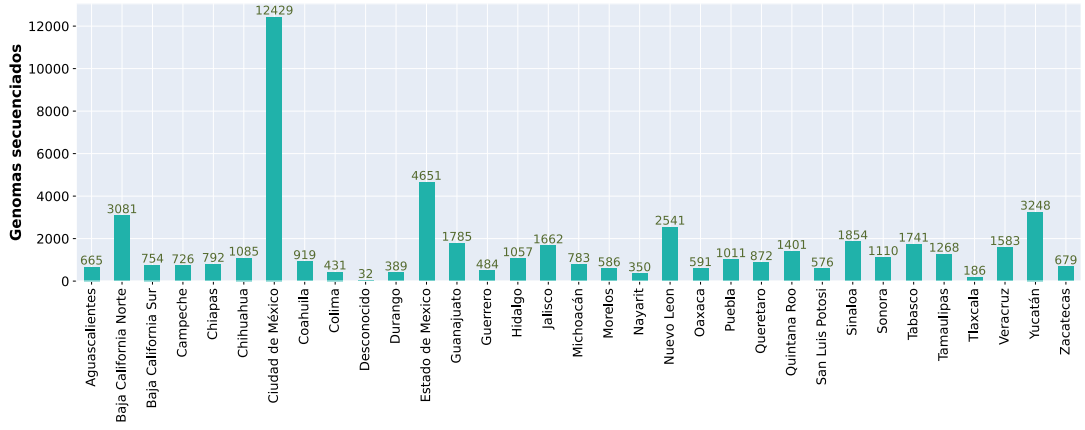

Figura 5.3: Distrbución de los 51,322 genomas del conjunto de datos por estado.

La Figura 5.4 muestra la cantidad de genomas secuenciados por día. En esta se puede apreciar cómo la cantidad de genomas secuenciados entre enero de 2020 y abril de 2021 es mucho menor que los obtenidos después de este periodo. La mayor cantidad de genomas secuenciados se registró durante las primeras semanas del 2022 y luego disminuyó considerablemente alrededor de abril de 2022.

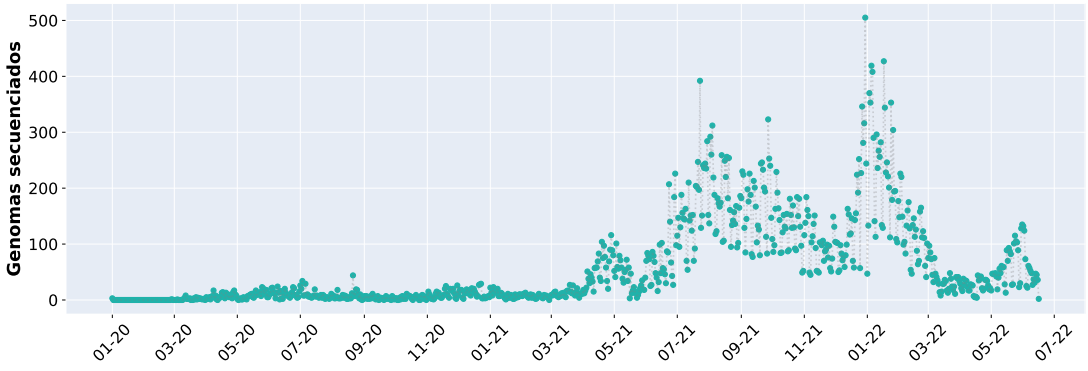

Figura 5.4: Cantidad de genomas secuenciados por fecha.

5.2. Distancias y entropía

Después de obtener el conjunto de datos, se lleva a cabo el cálculo de las distancias y la *entropía* mencionadas en el Capítulo 3. Primeramente, se obtiene la distancia entre cada genoma del virus y el genoma humano obtenido de [19]. Específicamente, se emplean tanto la *distancia euclidiana* como la *distancia de Wasserstein* para comparar los usos de codones de ambos genomas. Los resultados se presentan en las columnas *Dist_euclidiana* y *Dist_wasserstein* de la Tabla 5.3.

Posteriormente, se calcula la *entropía* de la distribución de cada uso de codones, que se presenta en la columna *Entropía* de la misma tabla. Finalmente, se aplica la transformación *log-cociente centrado* (también mencionada en el Capítulo 3) a todo el conjunto de datos y se obtiene la *distancia euclidiana* esperada de cada una de estas instancias, es decir, se calcula la *distancia log-cociente esperada*. Este resultado se muestra en la columna *Dist_logcociente* de la Tabla 5.3.

	Fecha	ID	Estado	Dist_euclidiana	Dist_wasserstein	Entropía	Dist_logcociente
0	2020-01-01	EPI_ISL_913918	Oaxaca	0.095266	2.415117	5.658083	0.209493
1	2020-01-01	EPI_ISL_914878	Tamaulipas	0.095288	2.422777	5.657209	0.200889
2	2020-01-01	EPI_ISL_933664	Estado de Mexico	0.095266	2.427153	5.657930	0.205896
...	...	...	...	...	...	...	...
51320	2022-06-16	EPI_ISL_13454153	Ciudad de México	0.095309	2.424662	5.657191	0.245656
51321	2022-06-16	EPI_ISL_13454154	Ciudad de México	0.095365	2.432371	5.657354	0.238075

Tabla 5.3: Resultados de calcular la *distancia euclidiana* y la *distancia de Wasserstein* entre el uso de codones de cada virus y el del humano. Así como la *entropía* y la *distancia log-cociente esperada* del uso de codones de cada virus.

5.2.1. Diferencias por región geográfica

Los resultados de la Tabla 5.3 se presentaron de acuerdo a la región geográfica. En la Figura 5.5 se muestran estas medidas según su estado de procedencia. Adicionalmente, empleando la visualización de estos datos, se seleccionaron los 50 genomas con los valores más divergentes de cada medida y luego se formó la Figura 5.6 con el conteo de los mismos según su estado de procedencia.

La *distancia euclidiana* al uso de codones humano (Fig. 5.5A) revela que la mayoría de los genomas virales se encuentran a una distancia de aproximadamente 0.095. Destacan el genoma EPI_ISL_12474718 del estado de Querétaro, siendo el más distante de todos, y el genoma EPI_ISL_8184724 de la Ciudad de México, como el más cercano. Además, se observa que algunos genomas del estado de Puebla están considerablemente más cerca en comparación con el resto. Al tomar los 50 genomas con la *distancia euclidiana* más pequeña (Fig. 5.6A), se encontró que más de la mitad de estos corresponden al estado de Puebla.

En la *distancia de Wasserstein* al uso de codones humano (Fig. 5.5B), se observa que la mayoría de los genomas se encuentran a una distancia menor de 2.6, alrededor del 99.7 % de los genomas del conjunto de datos. Donde el genoma EPI_ISL_2102650 del estado de Nuevo León es el más cercano de todos. Entre los pocos que tienen una distancia mayor a 2.8, destaca el genoma EPI_ISL_3055556 del estado de Puebla como el más distante. Al analizar los 50 genomas con la mayor *distancia de Wasserstein* (Fig. 5.6B), se observa que los genomas más cercanos pertenecen a los estados de Jalisco, Tabasco y Puebla.

En la Figura 5.5C se muestra la variabilidad de la *entropía* en los genomas analizados. Se destaca el genoma EPI_ISL_12474718 del estado de Querétaro, que presenta la menor *entropía*, y el genoma EPI_ISL_3055561 del estado de Puebla, con la *entropía* más alta. Además, se observa un grupo de genomas también del estado de Puebla que exhiben una *entropía* más elevada en comparación con el resto. Estos elementos presentan un valor de *entropía* similar al observado en el uso de codones humano. Al examinar los 50 genomas con la mayor *entropía* (Fig. 5.6C), se observa que la mitad de ellos provienen del estado de Puebla.

En cuanto a la *distancia log-cociente esperada* (Fig. 5.5D), la mayoría de los genomas exhiben valores inferiores a 2, salvo unas pocas decenas que sobrepasan este límite, siendo notablemente aquellos pertenecientes al estado de Puebla. Donde el genoma EPI_ISL_305556, también de Puebla, tiene el valor más alto de esta distancia. Si tomamos los 50 genomas con la mayor *distancia log-cociente esperada* (Fig. 5.6D), observamos que muchos de ellos son de Puebla.

En comparación con otros estados de la república, Puebla, Tabasco y Tlaxcala muestran una mayor cantidad de genomas divergentes bajo estos criterios, como se puede observar en la Figura 5.6.

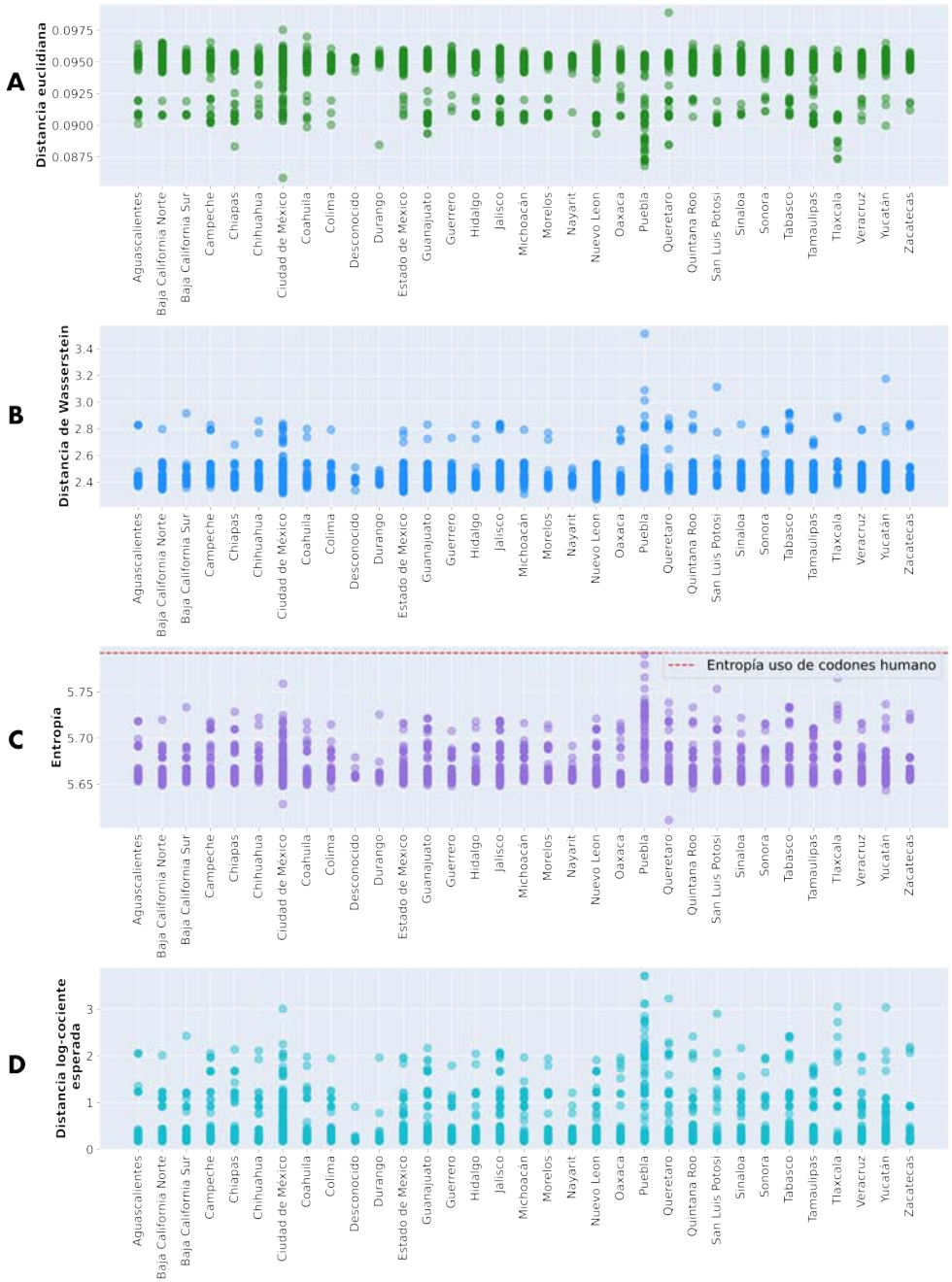

Figura 5.5: Distancias y *entropía* de cada genoma por región geográfica. *Distancia euclidiana* (**A**) y *Distancia de Wasserstein* (**B**) entre el uso de codones de cada virus y el uso de codones humano. *Entropía* (**C**) y *Distancia log-cociente esperada* (**D**) del uso de codones de cada virus.

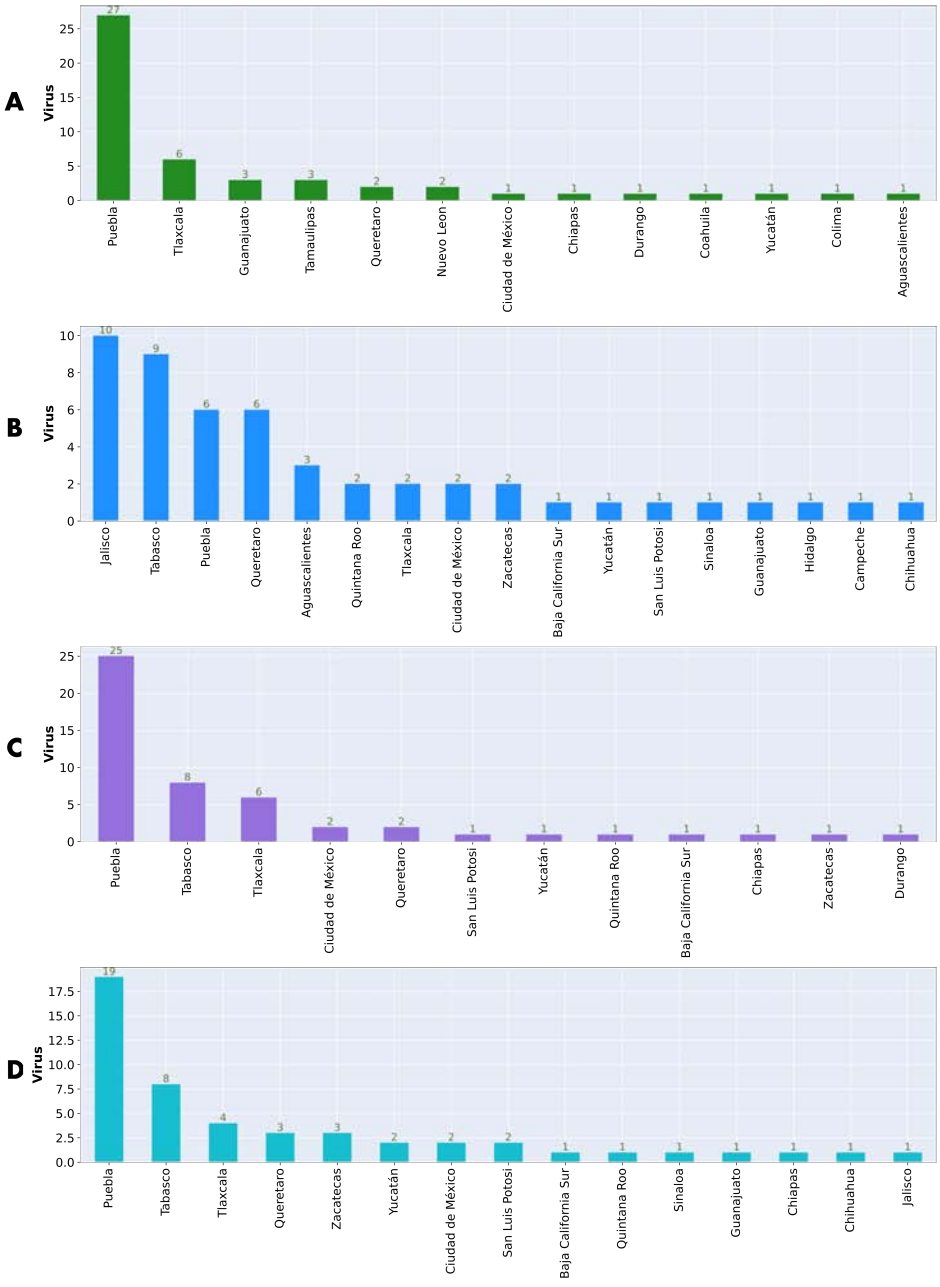

Figura 5.6: Conteo por estado de los 50 genomas más divergentes. Menor *distancia euclidiana* (**A**) y mayor *distancia de Wasserstein* (**B**) entre los usos de codones virales y el uso de codones humano. Mayor *entropía* (**C**) y mayor *distancia log-cociente esperada* (**D**) de los usos de codones virales.

A fin de obtener una perspectiva más general de las diferencias por región geográfica, se procedió a calcular el uso de codones promedio de cada región. Para ello, se agruparon los genomas de la Tabla 5.2 por estado y luego se determinó el uso de codones promedio de cada grupo. Posteriormente, se calcularon las medidas previamente mencionadas sobre este nuevo conjunto (Fig. 5.7).

Los valores de *distancia euclidiana* (Fig. 5.7A) revelan que el uso de codones promedio de los estados de Tlaxcala, Puebla y Morelos son los más similares al del humano. Sin embargo, al analizar la *distancia de Wasserstein* (Fig. 5.7B), se observa que los estados con un uso de codones promedio más similar al del humano son Oaxaca, Nuevo León y el Estado de México, mientras que Tlaxcala, Michoacán y Puebla se destacan como los más distantes.

Esto nos indica que al considerar los usos de codones como puntos en un espacio euclidiano, Tlaxcala y Puebla se encuentran cercanos al uso de codones humano. Por otro lado, si se considera la forma de las distribuciones de los usos de codones, estos estados resultan ser los más distantes al uso de codones humano.

En cuanto a la *entropía* (Fig. 5.7C) y la *distancia log-cociente esperada* (Fig. 5.7D), se observa que también destacan los estados de Puebla y Tlaxcala, presentando los valores más elevados. Esto nos sugiere que el *sesgo en el uso de codones* de estos estados es menor en comparación al resto, y que, en promedio, su uso de codones se distancia más del resto de estados en el espacio euclidiano.

5.2.2. Evolución temporal

A fin de mostrar la evolución temporal de las medidas obtenidas, los resultados de la Tabla 5.2 se presentan organizados según su fecha de secuenciación, como se puede observar en la Figura 5.8. En esta figura se puede apreciar cómo, alrededor de abril de 2021, comienzan a surgir un número significativo de genomas con valores divergentes en todas las medidas, mientras que después de abril de 2022 se observa una considerable disminución de los mismos.

La *distancia euclidiana* (Fig. 5.8A) muestra como los genomas más similares aparecen en julio de 2021 y en diciembre de 2021. Donde el genoma más cercano al genoma humano, EPI_ISL_8184724, se registra a finales de diciembre de 2021.

En la *distancia de Wasserstein* (Fig. 5.8B) se observa como los genomas más divergentes (lejanos) se registran también en julio de 2021. A principios de ese mes se registró la presencia del genoma con el uso de codones más distante al del humano, EPI_ISL_3055556 del estado de Puebla.

En la Figura 5.8C se puede apreciar la variabilidad de la *entropía* del uso de codones de cada genoma a lo largo del tiempo. Donde de manera similar, julio de 2021 y diciembre de 2021 son los meses con los genomas más notables. Específicamente, durante la segunda mitad de julio de 2021 se registró el genoma

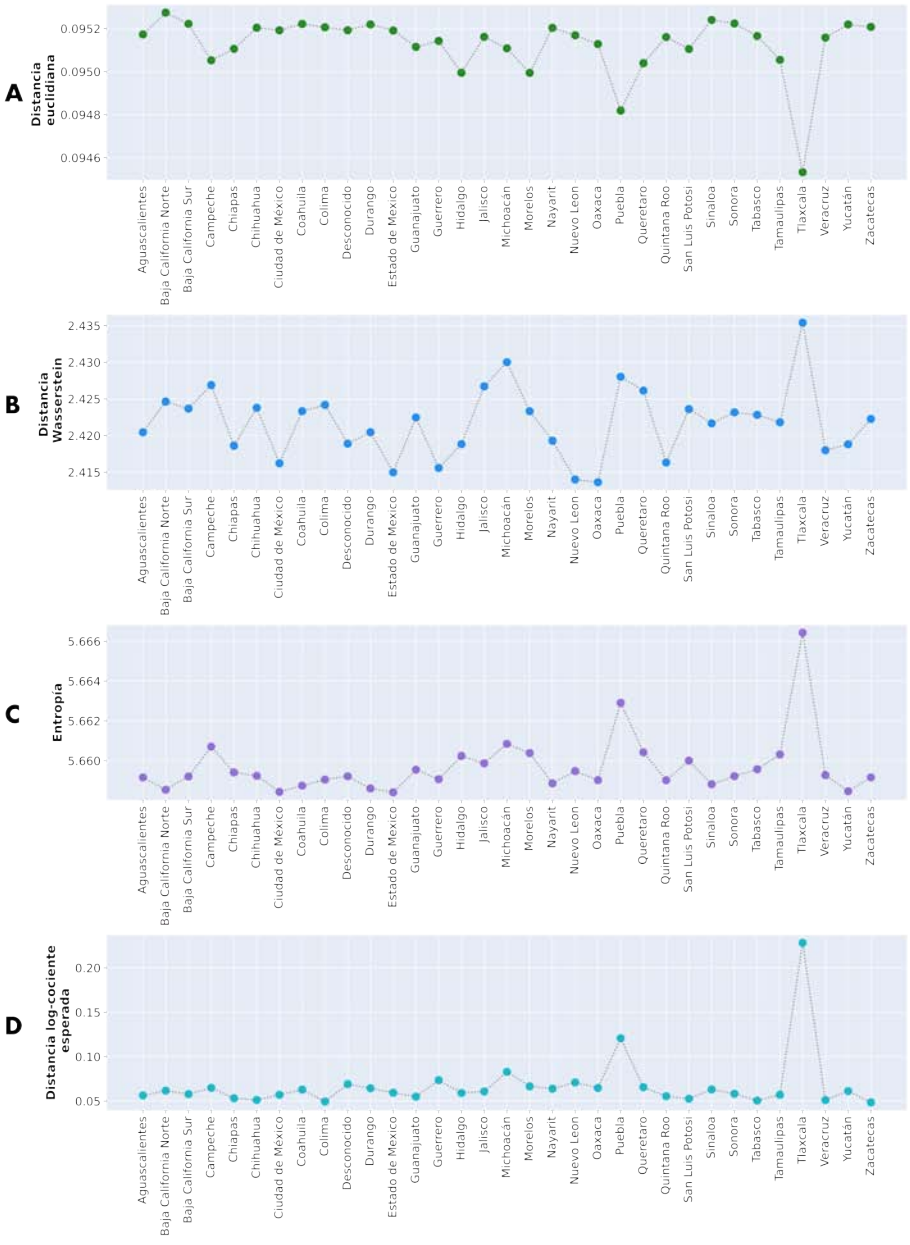

Figura 5.7: Distancias y *entropía* del uso de codones promedio por región geográfica. *Distancia euclidiana* (**A**) y *Distancia de Wasserstein* (**B**) entre los usos de codones promedio y el uso de codones humano. *Entropía* (**C**) y *Distancia log-cociente esperada* (**D**) de los usos de codones promedio.

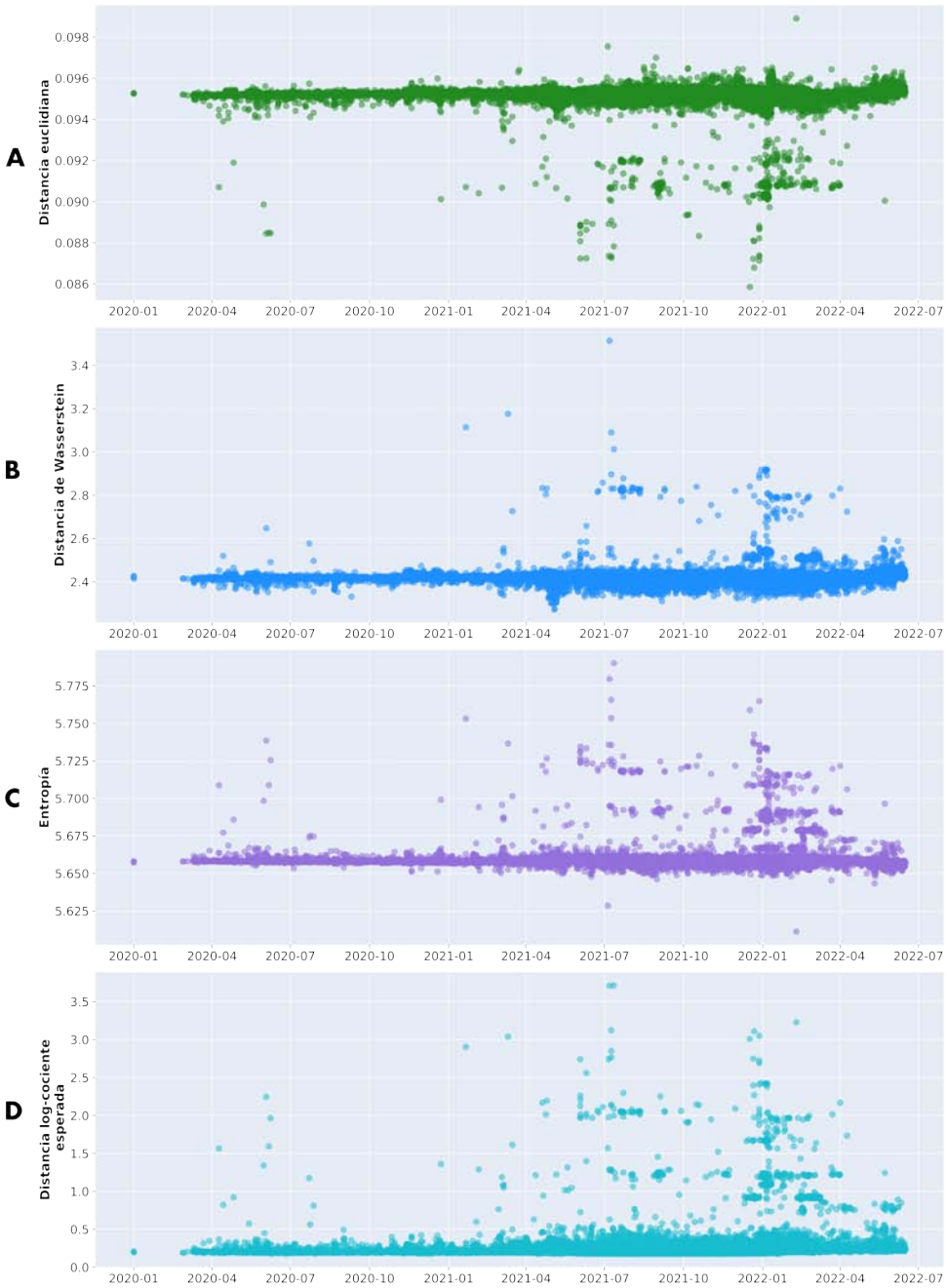

Figura 5.8: Distancias y *entropía* de cada genoma por fecha de secuenciación. *Distancia euclidiana* (**A**) y *Distancia de Wasserstein* (**B**) entre cada uso de codones viral y el uso de codones humano. *Entropía* (**C**) y *Distancia log-cociente esperada* (**D**) del uso de codones de cada virus.

EPI_ISL_3055561 con la *entropía* más alta de todos. Por otro lado, en febrero de 2022, se registró el genoma EPI_ISL_12474718 con la *entropía* más baja de todos.

Finalmente, la Figura 5.8D, que muestra la *distancia log-cociente esperada*, revela cómo también los genomas correspondientes a julio y diciembre de 2021 exhiben los mayores valores. Donde los genomas EPI_ISL_3055561 y EPI_ISL_3055556, ambos de julio de 2021, muestran la distancia esperada más alta y la segunda distancia esperada más alta, respectivamente.

Con el fin de estudiar la tendencia general que tienen los resultados a medida que pasa el tiempo, se calcularon los usos de codones representativos de cada semana. El conjunto de datos de la Tabla 5.2 consta de 122 semanas, por lo tanto, se obtuvieron 122 usos de codones representativos. Las distancias y *entropía* calculados para estos elementos se muestra en la Figura 5.9.

La *distancia euclidiana* (Fig. 5.9A) muestra como la primera semana de enero de 2022 cuenta con el uso de codones más divergente. Además, tanto para la *distancia euclidiana* como para la *distancia de Wasserstein* (Fig. 5.9B), se puede notar que a partir de abril de 2022 estos usos de codones promedio tienden a distanciarse cada vez más del uso de codones humano.

De igual forma, la primera semana de 2022 presenta el mayor valor tanto de *entropía* (Fig. 5.9C) como de *distancia log-cociente esperada* (Fig. 5.9D). Después de esta semana, se observa de manera más evidente como la *entropía* de los usos de codones representativos tiende a disminuir y su *distancia log-cociente esperada* tiende a aumentar.

5.2.3. Distancias contra entropía

Con el propósito de analizar la correlación existente entre las medidas computadas, se realizó una gráfica de dispersión entre cada una de las distancias y la *entropía*. La Figura 5.10 muestra los resultados obtenidos.

La Figura 5.10A muestra la gráfica de dispersión entre el valor de *distancia euclidiana* y la *entropía* obtenidos previamente (Tabla 5.3). En esta se puede observar que la mayoría de los genomas se agrupan en una región específica de este espacio. Estas dos variables exhiben una correlación de Pearson negativa significativa (-0.8885). Esto indica que, en general, la *distancia euclidiana* al uso de codones humano es mayor en los genomas virales que presentan un valor de *entropía* bajo (*sesgo en el uso de codones* alto).

La gráfica de dispersión entre los valores de la *distancia de Wasserstein* y la *entropía* (Tabla 5.3) se muestra en la Figura 5.10B. En esta, también se muestra una región donde la mayor parte de los genomas se encuentra agrupada. Sin embargo, la correlación de Pearson entre las dos variables no es tan significativa como en la Figura 5.10A. A pesar de esto, se puede notar que algunos genomas

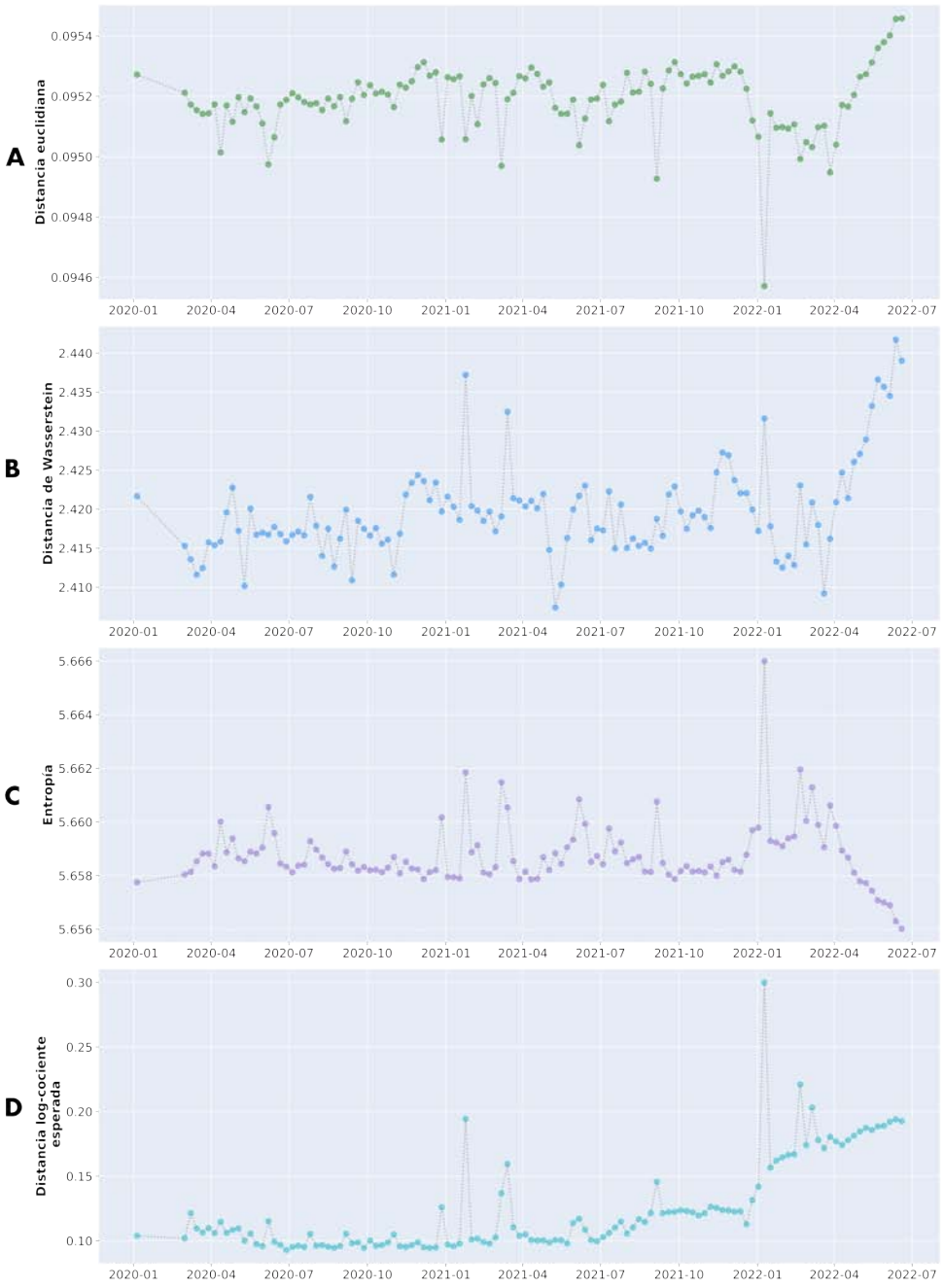

Figura 5.9: Distancias y *entropía* de los usos de codones promedio por semana. *Distancia euclidiana* (**A**) y *Distancia de Wasserstein* (**B**) entre los usos de codones promedio y el uso de codones humano. *Entropía* (**C**) y *Distancia log-cociente esperada*(**D**) de los usos de codones promedio.

con un alto valor de *distancia de Wasserstein* también presentan un alto valor de *entropía* (*sesgo en el uso de codones* bajo).

Por último, la Figura 5.10C muestra la gráfica de dispersión entre la *distancia log-cociente esperada* y la *entropía*. De igual forma, se puede observar que la mayoría de los genomas del conjunto de datos se agrupan en una región específica, mientras que solo un puñado de estos muestran la relación lineal presente en la figura. Es importante destacar que estas dos variables muestran una correlación positiva significativamente alta, con un coeficiente de correlación de Pearson de 0.9358. Lo que en general nos dice que entre más bajo sea el *sesgo en el uso de codones* del genoma, mayor sera la *distancia log-cociente esperada*.

5.3. Reducción de la dimensionalidad

En esta sección, se emplean los métodos de reducción de dimensionalidad que se detallan en el Capítulo 3. El objetivo es reducir los 64 atributos que describen a cada genoma en la Tabla 5.2 a solo 2 atributos. Con estas nuevas representaciones se facilita la visualización del conjunto de datos en un espacio bidimensional, permitiendo así observar de forma sencilla las variaciones de estas instancias en los nuevos subespacios. Posteriormente, se identifican las instancias anómalas seleccionando aquellas que presentan el mayor valor de distancia esperada (*distancia euclidiana*) en los nuevos subespacios, tal como se menciona en el Capítulo 4.

5.3.1. Análisis de componentes principales (PCA)

La nueva representación obtenida con el método de *análisis de componentes principales* se muestra en la Tabla 5.4. Cada instancia en esta tabla se representa mediante dos atributos, los primeros dos *componentes principales* (*PC1* y *PC2*). Estos dos componentes capturan aproximadamente el 67 % de la variabilidad total de los datos. La Figura 5.11 muestra el porcentaje de contribución de los primeros 15 componentes a la varianza total. Por último, se calcula la distancia esperada para cada vector en este nuevo subespacio, columna *Distancia*.

Los resultados obtenidos al aplicar *PCA* se visualizan en la Figura 5.12. La proyección de los datos sobre los dos primeros *componentes principales* se presenta en la Figura 5.12A, donde se observa como la mayoría de los genomas se agrupan entorno a una región determinada. Al examinar las instancias con la mayor distancia esperada, se encontró que los genomas EPI_ISL_3055561, EPI_ISL_3055556 y EPI_ISL_3055553 presentan los valores más altos, respectivamente.

La Figura 5.12B muestra la distancia esperada a través del tiempo. En esta

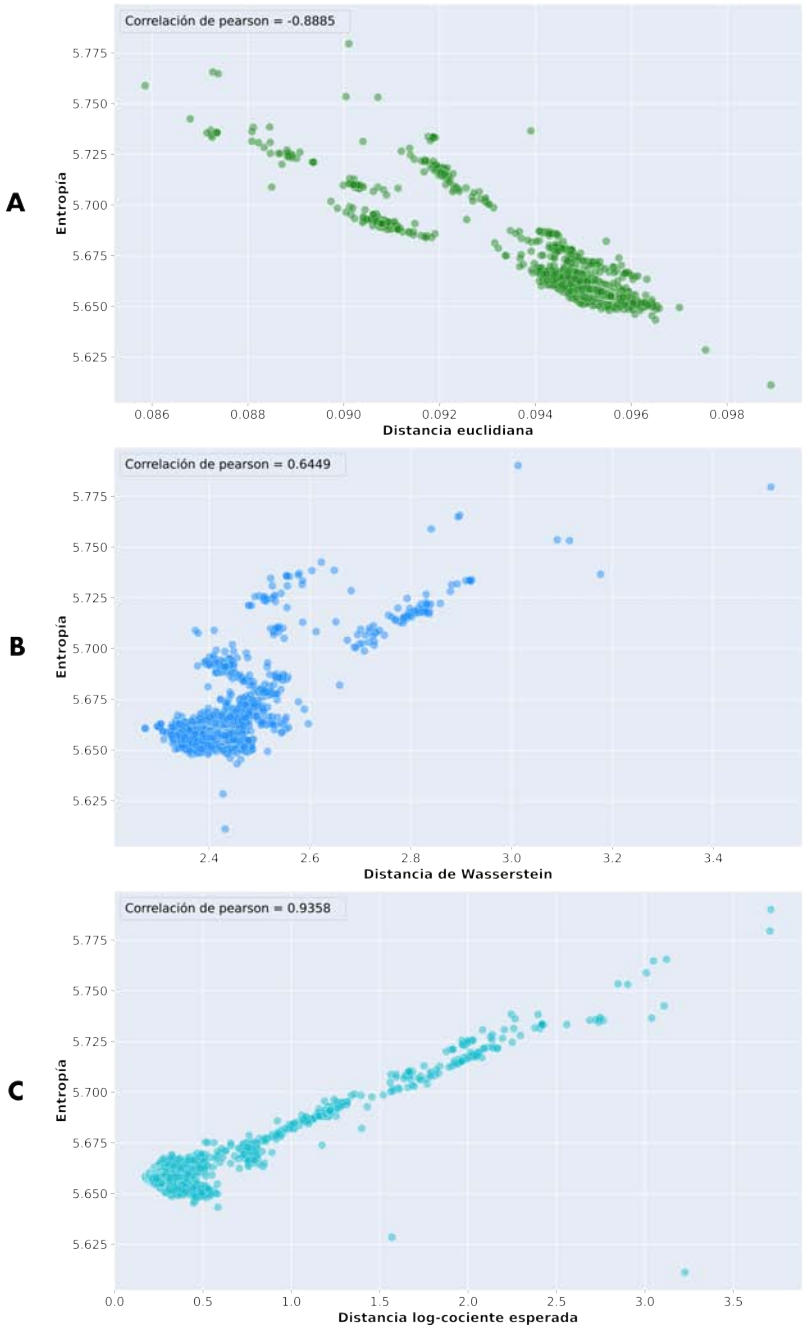

Figura 5.10: Distancias contra *entropía* obtenidas para los usos de codones virales (Tabla 5.3). *Distancia euclidiana* contra *entropía* (**A**), *Distancia de Wasserstein* contra *entropía* (**B**) y *Distancia log-cociente esperada* contra *entropía* (**C**).

	Fecha	ID	Estado	PC1	PC2	Distancia
0	2020-01-01	EPI_ISL_913918	Oaxaca	-0.040065	0.057259	0.108857
1	2020-01-01	EPI_ISL_914878	Tamaulipas	-0.042507	0.026700	0.109350
2	2020-01-01	EPI_ISL_933664	Estado de Mexico	-0.042139	0.060417	0.109732
...	...	...	...	...	...	...
51320	2022-06-16	EPI_ISL_13454153	Ciudad de México	-0.003315	-0.118541	0.155450
51321	2022-06-16	EPI_ISL_13454154	Ciudad de México	-0.013471	-0.151782	0.175567

Tabla 5.4: Usos de codones de cada genoma viral representados con los primeros dos *componentes principales* y su distancia esperada dentro del nuevo subespacio.

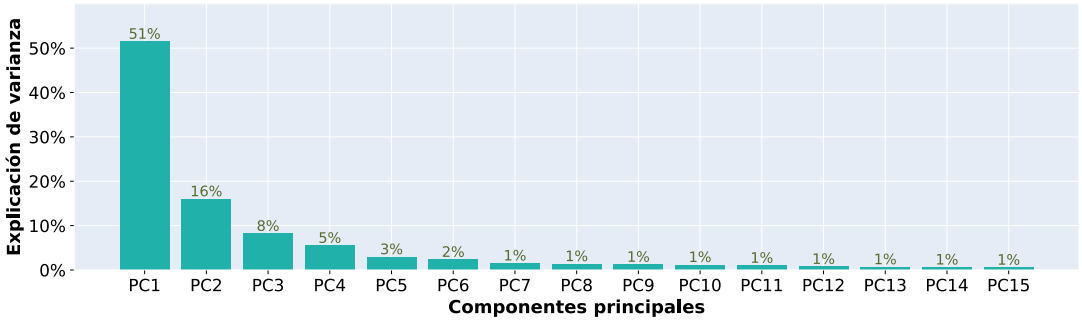

Figura 5.11: Porcentaje de contribución de los primeros 15 *componentes principales* de *PCA* a la varianza total del conjunto de genomas de la Tabla 5.2.

se observa como los genomas más distantes en la proyección se registran en julio y diciembre de 2021, similar a las medidas computadas en la Sección 5.2.

La Figura 5.12C presenta el recuento por estado de los 30 genomas con la mayor distancia esperada, destacando especialmente los estados de Puebla y Tabasco, que cuentan con la mayor cantidad de estos genomas. Adicionalmente, en la Figura 5.12A y en la Figura 5.12B se etiquetan estos 30 genomas con la mayor distancia esperada según su estado de procedencia.

5.3.2. Mapeo isométrico (Isomap)

Debido a algunas limitaciones computacionales solo fue posible aplicar este método reduciendo el tamaño del conjunto de genomas (Tabla 5.2). Para lograrlo, se seleccionó aleatoriamente solo el 78 % de las instancias de cada estado, lo que resultó en una disminución del tamaño general del conjunto de datos a 40,030

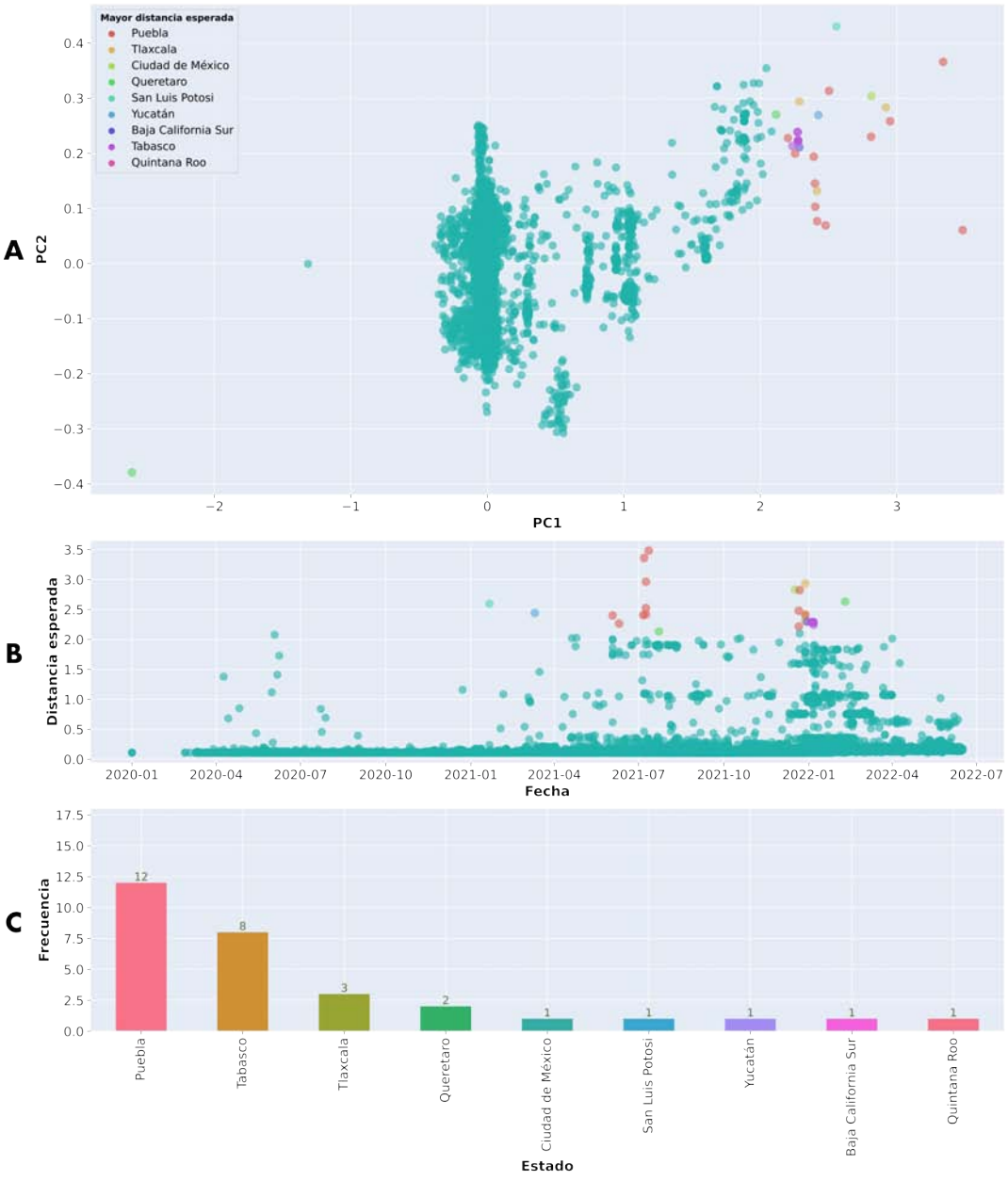

Figura 5.12: *Análisis de componentes principales* sobre el conjunto de genomas del SARS-CoV-2. (**A**) Usos de codones proyectados sobre *PC1* y *PC2*. (**B**) Distancia esperada de cada genoma viral en relación a su fecha de registro. (**C**) Conteo por estado de los 30 genomas con mayor distancia esperada.

instancias, pero manteniendo la misma proporción de genomas por estado. Los resultados de aplicar *Isomap* sobre este conjunto se muestra en la Tabla 5.5.

	Fecha	ID	Estado	1	2	Distancia
0	2021-10-21	EPI_ISL_5846244	Oaxaca	0.010186	0.006455	0.016669
1	2021-04-20	EPI_ISL_2402190	Oaxaca	-0.001729	0.014643	0.011918
2	2021-09-07	EPI_ISL_4232414	Oaxaca	-0.003389	-0.006899	0.018234
...	...	...	...	...	...	...
40029	2021-08-18	EPI_ISL_7813058	Desconocido	0.002980	-0.005230	0.011734
40030	2021-08-04	EPI_ISL_7813048	Desconocido	0.003044	-0.010827	0.014403

Tabla 5.5: Uso de codones representados en un espacio bidimensional obtenido con el método de *Isomap* y su distancia esperada en este nuevo subespacio.

La Tabla 5.5 presenta los vectores bidimensionales correspondientes a cada uso de codones. Las columnas *1* y *2* indican su posición en cada dimensión, mientras que la columna *Distancia* presenta su distancia esperada en el nuevo subespacio.

El conjunto de nuevos vectores se presentan en la Figura 5.13A, donde se observa como la mayoría de las instancias se concentran en una sola región de este subespacio y solo unos pocos se alejan bastante de esta región. Los 30 genomas con la mayor distancia esperada se etiquetaron según su estado de origen.

En la Figura 5.13B se representa la distancia esperada en relación a la fecha de registro de cada genoma. Se observa que los genomas registrados en julio y diciembre de 2021 presentan las mayores distancias esperadas en el subespacio, lo cual coincide con resultados previos. De igual forma, los genomas que presentan la mayor distancia esperada se etiquetaron según su estado de procedencia.

La Figura 5.13C presenta el recuento de los 30 genomas más distantes por estado. Se observa nuevamente que los estados de Puebla y Tlaxcala destacan.

5.4. Algoritmos de detección de anomalías

En esta sección, se procede a implementar los algoritmos de detección de anomalías, mencionados en el Capítulo 4, sobre la representación de los genomas (Tabla 5.2). Los resultados de cada algoritmo se presentan por fecha y estado, además de resaltar los estados con una mayor cantidad de genomas anómalos.

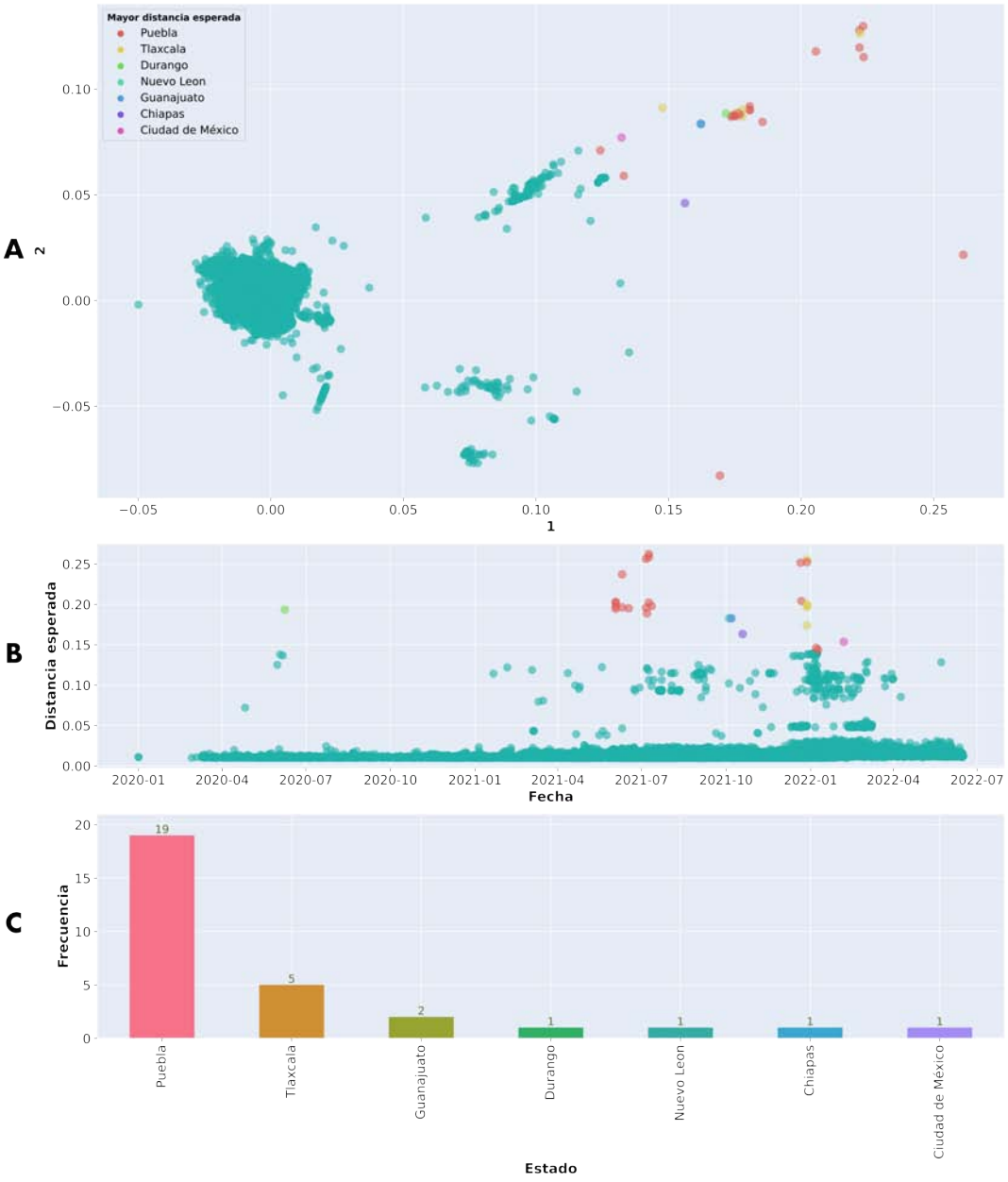

Figura 5.13: *Mapeo isométrico* sobre el conjunto de genomas del SARS-CoV-2. (**A**) Usos de codones representados por vectores bidimensionales. (**B**) Distancia esperada de cada genoma viral en relación a su fecha de registro. (**C**) Conteo por estado de los 30 genomas con mayor distancia esperada.

5.4.1. DBSCAN

El algoritmo *DBSCAN* se aplicó con un radio de vecindad *eps* igual a 0.005. Este valor se eligió después de analizar la distribución de las distancias esperadas del conjunto de genomas (Fig. 5.14). En dicha distribución, se encontró que aproximadamente el 1.84 % de las instancias se encuentran a una distancia esperada mayor a 0.005. Por esta razón, al seleccionar este valor, se espera que algunas de estas instancias sean clasificadas como anomalías.

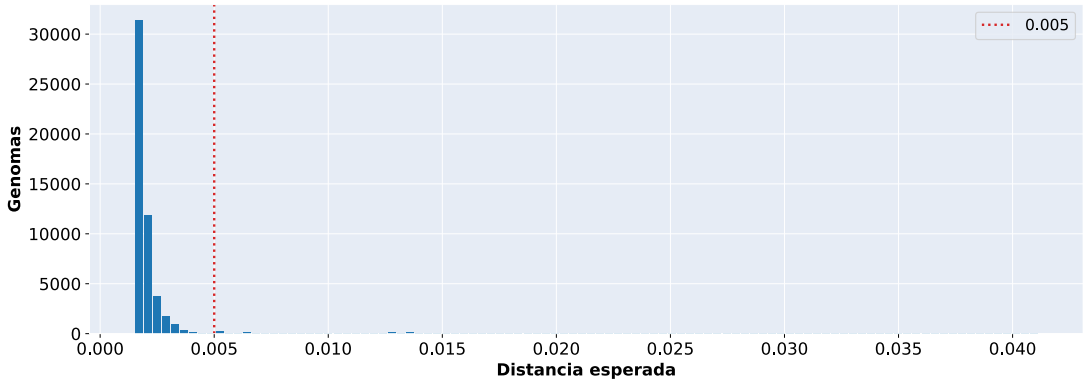

Figura 5.14: Distancias esperadas de los usos de codones de la Tabla 5.2.

En este caso se consideraron como *puntos central* aquellos que cuentan con al menos 10 puntos en su vecindad. Este parámetro se eligió de manera empírica.

En la Figura 5.15 se presentan los resultados obtenidos por este algoritmo. La Figura 5.15A muestra la asignación de etiquetas realizada por *DBSCAN* a cada genoma a lo largo del tiempo. Este identifica cuatro grupos diferentes (etiquetas 0, 1, 2 y 3), donde la mayor parte de los genomas pertenecen al grupo 0. El algoritmo encontró 30 genomas anómalos (etiqueta -1), la mayoría de los cuales se localizan en los meses de julio y diciembre de 2021. Al examinar la procedencia de cada genoma anómalo (Fig. 5.15B), se observa que la mayoría proviene del estado de Puebla, seguido por el estado de Tlaxcala.

5.4.2. Detección con el algoritmo de k-medias

En este caso, el algoritmo de *k-medias* se aplicó a subconjuntos de los datos en lugar de al conjunto de datos completo. Se utilizaron dos criterios diferentes para la selección de estos subconjuntos. El primer criterio se basó en los estados de registro, mientras que el segundo se centró en la selección de periodos de tiempo con la misma cantidad de genomas registrados.

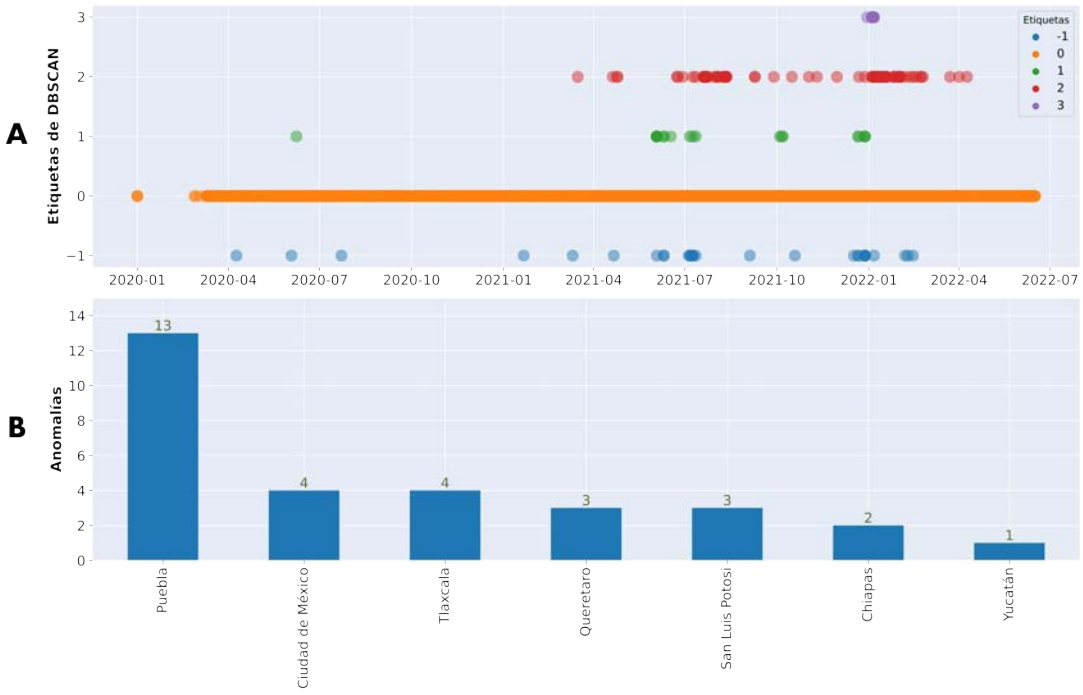

Figura 5.15: Algoritmo *DBSCAN* aplicado al conjunto de genomas del SARS-CoV-2 ($eps = 0.005$ y $MinPts = 10$). (**A**) Etiqueta de cada genoma a lo largo del tiempo. (**B**) Estados de origen de los genomas clasificados como anómalos.

5.4.2.1. Aplicación por estado

Este procedimiento consistió en dividir el conjunto de datos en 32 subconjuntos distintos, correspondientes a los estados de la república. La Figura 5.3 muestra la cantidad de genomas en cada uno de estos subconjuntos. La selección de los centroides para cada subconjunto se determinó dividiendo el número total de instancias de cada uno de estos sobre 20. La Figura 5.16A muestra la cantidad de usos de codones representativos computados para cada estado.

Posteriormente se procedió a determinar la distancia promedio de cada uso de codones con respecto a los usos de codones representativos de su estado. Por ejemplo, para un genoma del estado de Nuevo León, se calcula la distancia promedio entre su uso de codones y los 128 usos de codones representativos del estado.

La distancia promedio se utiliza como una medida o puntuación para evaluar que tan atípico es un genoma en comparación con los usos de codones representativos del estado del cual proviene. La Figura 5.16B muestra esta puntuación para cada genoma según su estado. También se presenta el umbral que indica los

50 genomas con las puntuaciones de anomalía más altas, donde se destacan numerosos genomas del estado de Puebla con valores superiores a este umbral. Sin embargo, el genoma EPI_ISL_8184724 de la Ciudad de México es el que presenta la mayor diferencia en comparación con los genomas de su estado.

En la Figura 5.16C se presenta el recuento, por estado, de los 50 genomas con las puntuaciones de anomalía más altas. Se puede observar que más del 50 % de estos genomas pertenecen al estado de Puebla, seguido por el estado de Tabasco.

5.4.2.2. Aplicación por periodos de tiempo

El segundo criterio consistió en dividir el conjunto de datos en 10 períodos de tiempo diferentes, cada uno con la misma cantidad de instancias (5,132 genomas). Luego, se eligieron los centroides para cada periodo dividiendo su número total de instancias sobre 20. Por último, se calculó la distancia promedio de cada uso de codones del virus con respecto a los usos de codones representativos de su periodo. Esta distancia promedio se interpretó como un puntaje de anomalía.

En la Figura 5.17A se presenta la distancia promedio o puntaje de anomalía de las instancias según su período correspondiente. En esta se puede observar el umbral seleccionado que identifica los 50 genomas con los puntajes de anomalía más altos. Los períodos 2 y 7 se destacan por tener la mayor cantidad de genomas con puntajes por encima de dicho umbral. Asimismo, se puede apreciar que el primer y último período contienen la mayor cantidad de días, lo cual concuerda con la información de los genomas registrados por día mostrada en la Figura 5.4.

La Figura 5.17B presenta la distancia promedio de cada genoma según su estado de registro. Los tres genomas que presentan la mayor distancia promedio son EPI_ISL_8183503, EPI_IS_3055556 y EPI_ISL_3055561, todos del estado de Puebla. Al examinar el recuento por estado de los 50 genomas con la mayor distancia promedio (Fig. 5.17C), se observa que la mitad de ellos pertenecen al estado de Puebla, seguido por los estados de Tabasco y Tlaxcala.

Se puede notar que tanto en los subconjuntos por estado como por periodos de tiempo, los genomas más sobresalientes son aquellos del estado de Puebla.

5.4.3. Valor atípico local (LOF)

El algoritmo *LOF* se computó utilizando un valor de vecinos k igual a 50. La puntuación de anomalía obtenida para cada genoma y el umbral requerido para encontrar las instancias más anómalas se muestran en la Figura 5.18.

En la Figura 5.18A se muestra este puntaje para cada genoma a lo largo del tiempo. Donde es evidente que los genomas que exhiben la puntuación más alta

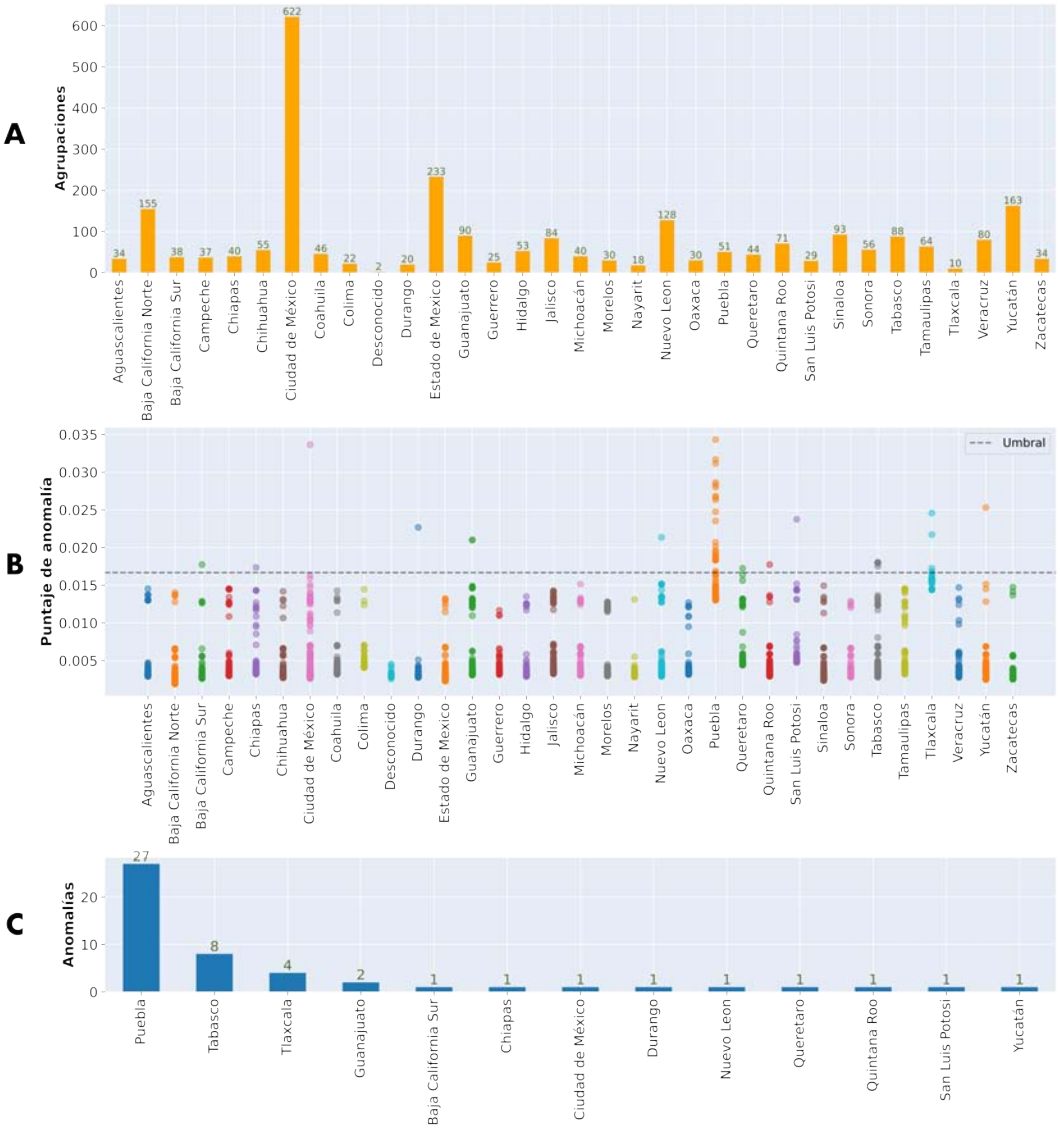

Figura 5.16: Detección de anomalías con *k-medias* (subconjuntos por estado). (**A**) Número de usos de codones representativos por estado. (**B**) Distancia promedio de cada uso de codones del virus a los usos de codones representativos de su estado. (**C**) Conteo por estado de los 50 genomas con mayor distancia promedio.

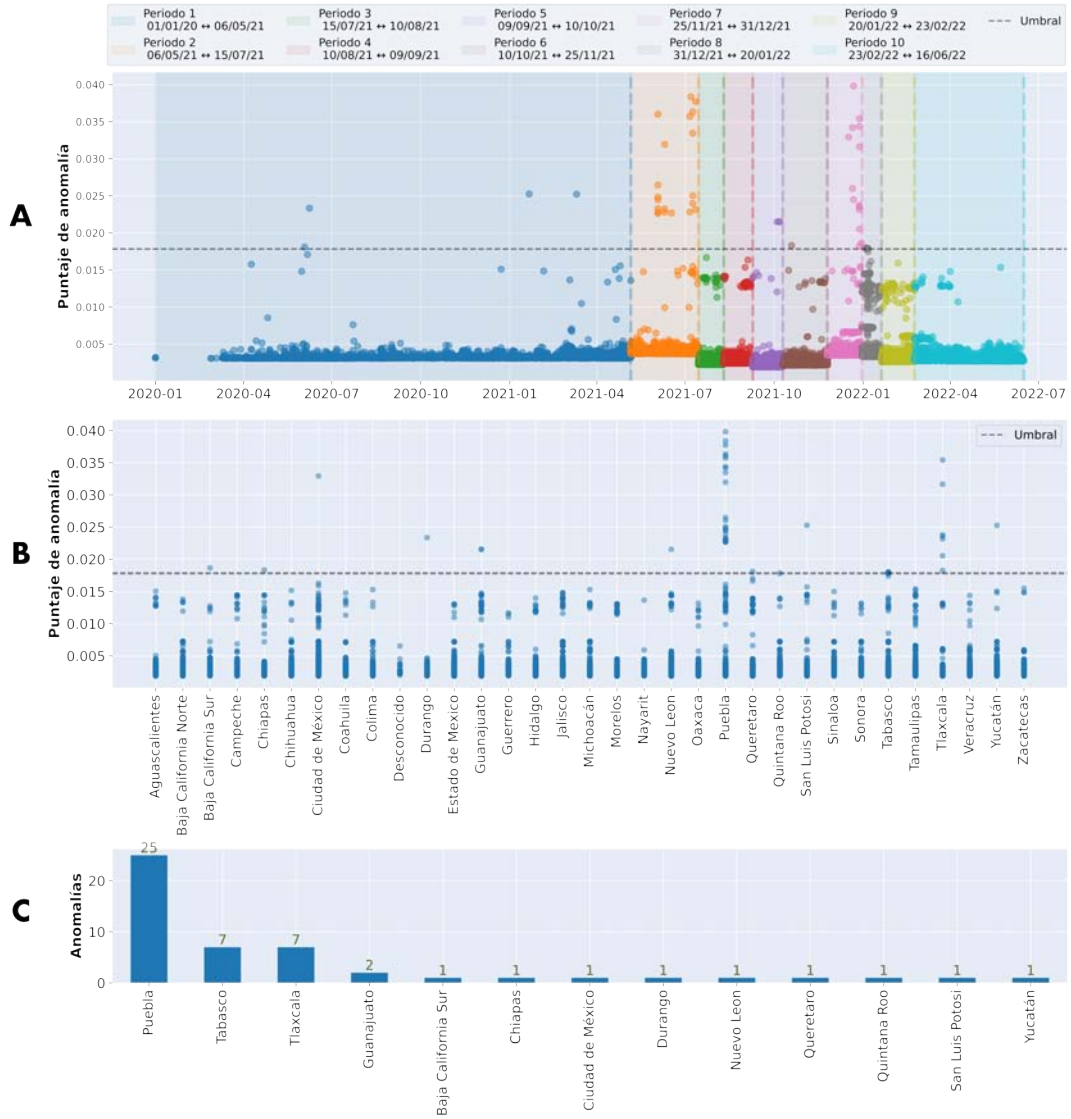

Figura 5.17: Detección de anomalías con *k-medias* (subconjuntos por periodos). (**A**) Distancia promedio de cada uso de codones del virus a los usos de codones representativos de su periodo (por fecha). (**B**) Distancia promedio de cada uso de codones del virus a los usos de codones representativos de su periodo (por estado). (**C**) Conteo por estado de los 50 genomas con mayor distancia promedio.

son los registrados entre diciembre de 2021 y abril de 2022, antes y después de este periodo el puntaje es insignificante en comparación.

La Figura 5.18B muestra la puntuación de anomalía por región geográfica. En este caso, 20 estados presentan al menos una instancia que supera el umbral seleccionado, destacando los genomas EPI_ISL_8988618 de Nayarit, EPI_ISL_12480199 del Estado de México y EPI_ISL_9092809 de Nuevo León con los puntajes de anomalía más altos. El conteo de estos genomas por estado se muestra en la Figura 5.18C. En la cual destacan la Ciudad de México y el estado de Yucatán como las regiones con la mayor cantidad de genomas que sobrepasan este umbral.

5.4.4. Bosque de aislamiento

En este caso, se utilizaron 100 *árboles de aislamiento* para computar el algoritmo. Cada árbol se construyó seleccionando aleatoriamente sólo el 25 % del total de instancias. Una vez generados todos estos árboles, se empleó el conjunto completo de éstos para calcular el puntaje de anomalía de cada instancia, tal como se detalló el algoritmo en el Capítulo 4.

La Figura 5.19A muestra el puntaje de anomalía de cada genoma según su fecha de registro. En la figura también se aprecia el umbral seleccionado, que identifica los 52 genomas con el puntaje de anomalía más alto. Por encima de este umbral, se observa que los genomas con los puntajes más altos se concentran en las fechas cercanas a julio y diciembre de 2021.

En la Figura 5.19B se presenta el puntaje de anomalías de cada genoma según su estado de registro. Se puede observar que el estado de Puebla tiene la mayor cantidad de genomas con puntajes por encima del umbral establecido, lo cual concuerda con hallazgos previos. Específicamente, los genomas EPI_ISL_3055556, EPI_ISL_8183503 y EPI_ISL_3055561 del estado de Puebla se destacan como los genomas más anómalos entre todos los genomas analizados.

Por último, la Figura 5.19C muestra el conteo por estado de los genomas que están por encima del umbral, donde la mayoría de estos pertenecen a Puebla.

5.5. Resumen del capítulo

En esta tesis se empleó una metodología que consideró 11 criterios para determinar la clasificación de un genoma como una anomalía. Como se mencionó previamente, solo se consideraron como anomalías aproximadamente el 0.01 % (entre 50 y 52 genomas) de las instancias más divergentes según la mayoría de estos criterios. Esto se hizo con el propósito de evitar una sobrerrepresentación

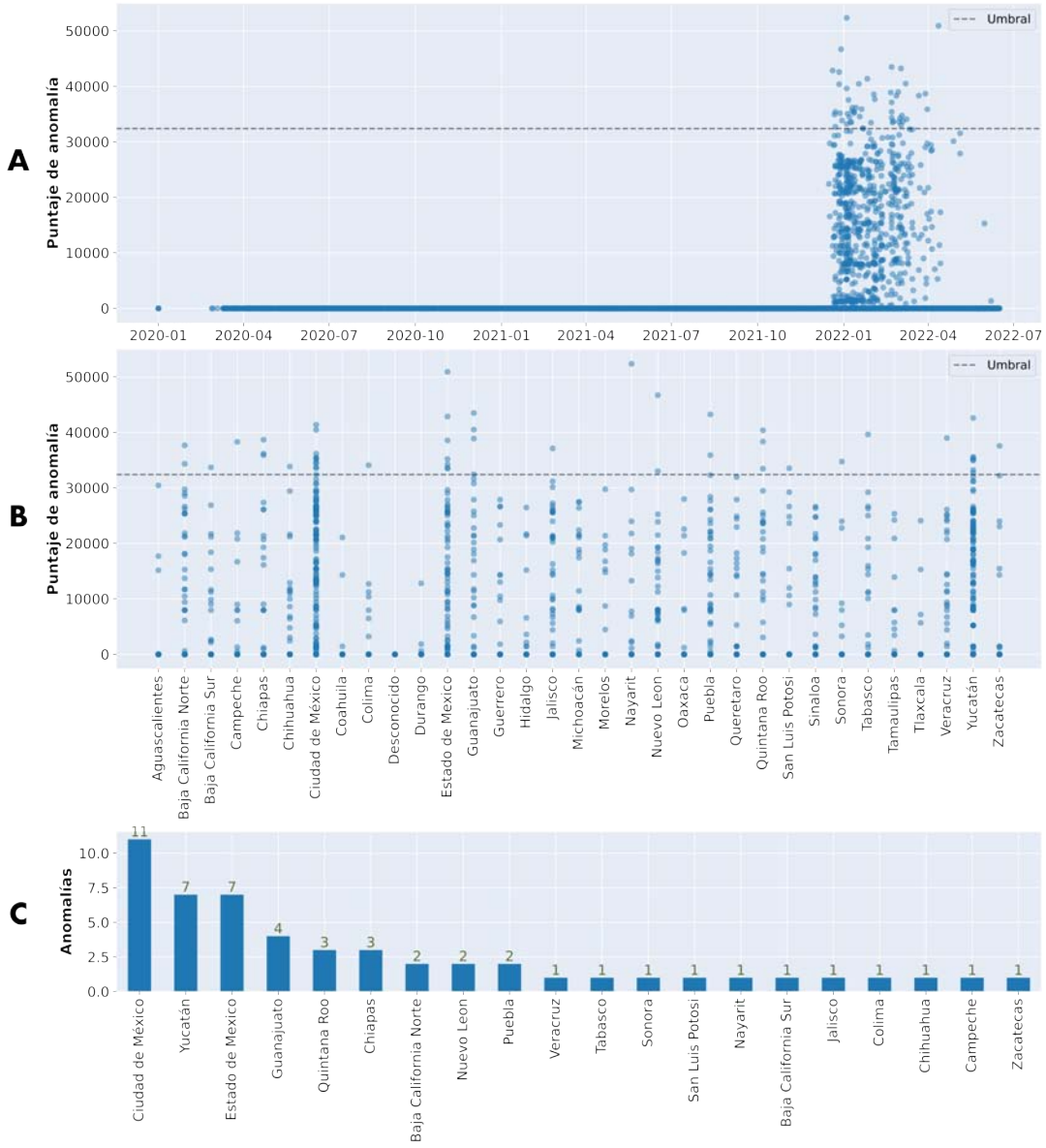

Figura 5.18: Algoritmo *Local Outlier Factor* aplicado al conjunto de genomas del SARS-CoV-2 ($k = 50$). (**A**) Puntaje de anomalía de cada genoma a lo largo del tiempo. (**B**) Puntaje de anomalía de cada genoma por estado. (**C**) Conteo por estado de los 52 genomas con mayor puntaje de anomalía.

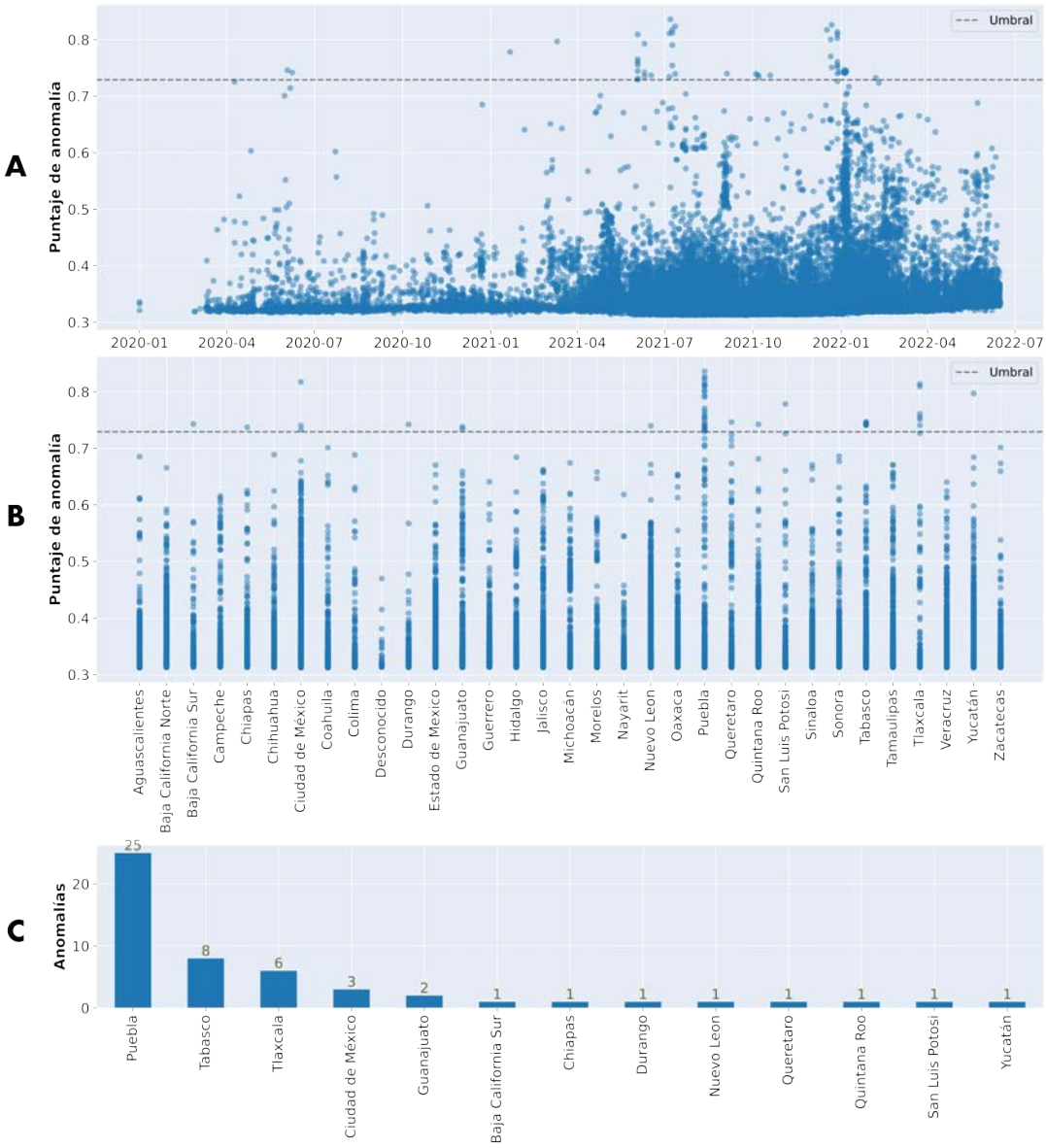

Figura 5.19: Algoritmo del *bosque de aislamiento* aplicado al conjunto de genomas del SARS-CoV-2 (100 *árboles de aislamiento*). (**A**) Puntaje de anomalía de cada genoma por fecha. (**B**) Puntaje de anomalía de cada genoma por estado. (**C**) Conteo por estado de los genomas con puntaje de anomalía encima del umbral.

de genomas de la Ciudad de México. Los criterios empleados en este capítulo abarcaron:

1. Los genomas con la menor *distancia euclidiana* entre su uso de codones y el uso de codones del genoma humano.
2. Los genomas con la mayor *distancia de Wasserstein* entre su uso de codones y el uso de codones del genoma humano.
3. Los genomas con la mayor *entropía* en su uso de codones.
4. Los genomas con la mayor *distancia log-cociente esperada* en el espacio de uso de codones.
5. Los genomas con la mayor distancia esperada en la proyección de *PCA*.
6. Los genomas con la mayor distancia esperada en el espacio generado por *Isomap*.
7. Los genomas clasificados como anomalías por el algoritmo de *DBSCAN*.
8. Los genomas con el mayor puntaje de anomalía obtenido por el algoritmo de *k-medias*, aplicado a subconjuntos por estado.
9. Los genomas con el mayor puntaje de anomalía obtenido por el algoritmo de *k-medias*, aplicado a subconjuntos por periodos de tiempo.
10. Los genomas con el mayor puntaje de anomalía obtenido por el algoritmo del *valor atípico local* (*LOF*).
11. Los genomas con el mayor puntaje de anomalía obtenido por el algoritmo del *bosque de aislamiento*.

Se identificaron 176 genomas diferentes como anomalías según al menos uno de estos criterios. La Figura 5.20A muestra la gráfica de dispersión de estos genomas anómalos, organizados por fecha y estado. En dicha figura, el color y el tamaño de cada instancia indican la cantidad de criterios por los cuales ese genoma en particular fue clasificado como una anomalía. Se puede observar que los genomas clasificados por más criterios pertenecen al estado de Puebla, específicamente durante los meses de julio y diciembre de 2021.

La Figura 5.20B muestra el recuento por estado de los 176 genomas anómalos identificados. Destaca, una vez más, el estado de Puebla en comparación con los demás, ya que cuenta con 33 de estos genomas anómalos.

La Figura 5.20C muestra la cantidad de genomas identificados como anomalías en cada mes. Los meses de junio y julio de 2021 presentan varias instancias

anómalas, pero es evidente que la mayoría de las anomalías se detectaron hacia finales de 2021 y principios de 2022.

La Tabla 5.6 muestra los primeros 30 genomas identificados por la mayoría de estos criterios como anomalías. La mayor parte de los genomas de la tabla fueron secuenciados durante los meses de junio, julio y diciembre del año 2021. Destacan los genomas EPI_ISL_3055556 y EPI_ISL_3055553, ambos provenientes del estado de Puebla, siendo los únicos en haber cumplido con 10 criterios. Además, se observa que 17 de estos 30 genomas pertenecen al estado de Puebla. Al revisar el laboratorio donde se realizó la secuenciación de cada genoma, mostrado también en la Tabla 5.6, se aprecia que el laboratorio *LABOPAT* es responsable de la mayoría de estos genomas.

Finalmente, en el siguiente capítulo se abordará la conclusión de este trabajo, así como una breve exploración del trabajo a futuro que se podría considerar.

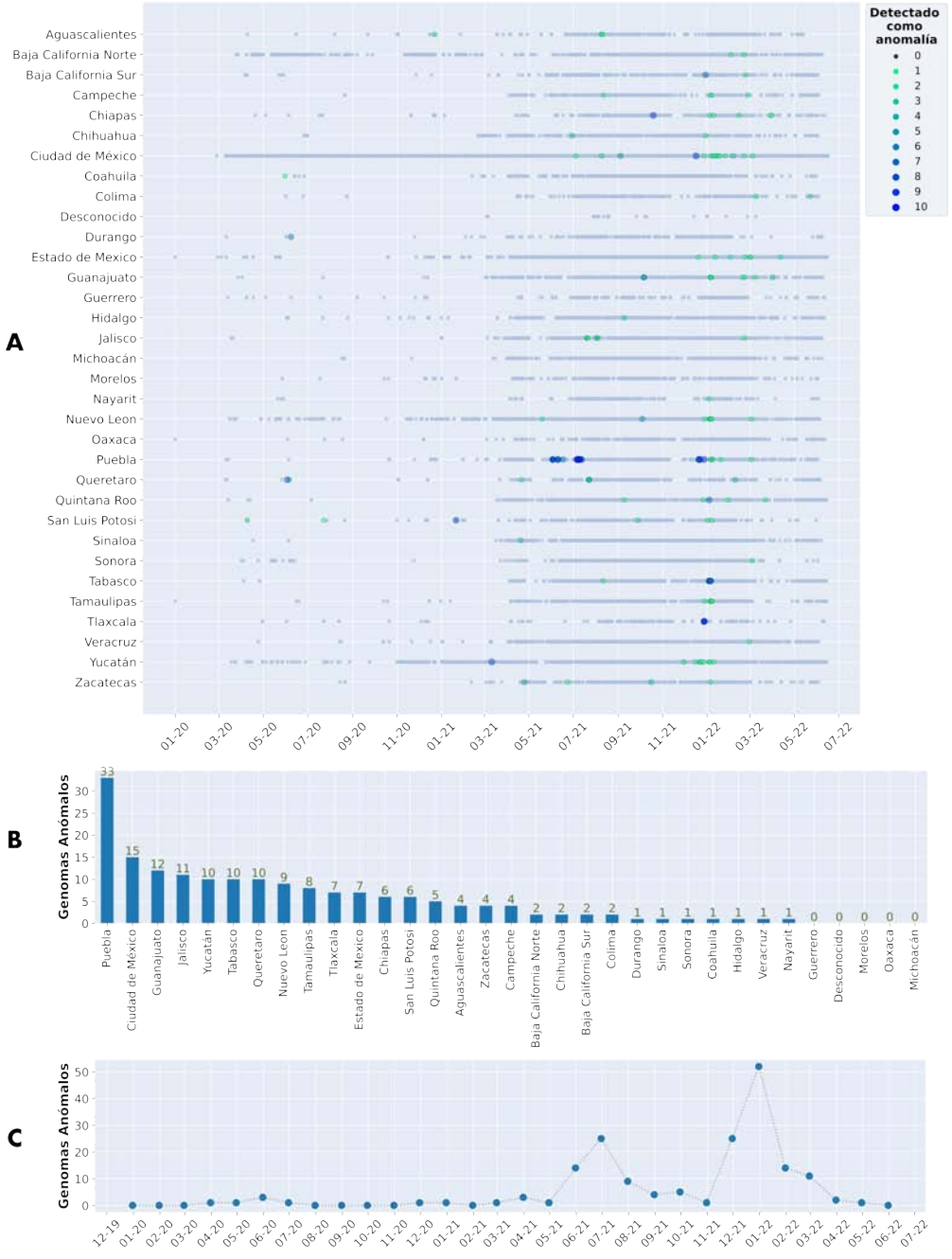

Figura 5.20: Genomas anómalos detectados por los todos los criterios. (**A**) Genomas por fecha y estado, el tamaño y color de cada instancia indican la cantidad de veces que este se identificó como anomalía por diferentes criterios. (**B**) Conteo por estado de todas las anomalías. (**C**) Conteo por mes de todas las anomalías.

	Fecha	ID	Laboratorio	Estado	Anomalía
1	2021-07-07	EPI_ISL_3055556	PUE-LABOPAT-71_26796	Puebla	10
2	2021-07-09	EPI_ISL_3055553	PUE-LABOPAT-88_27202	Puebla	10
3	2021-06-10	EPI_ISL_3055582	PUE-LABOPAT-17_20285	Puebla	9
4	2021-07-06	EPI_ISL_3055554	PUE-LABOPAT-70_26696	Puebla	9
5	2021-07-09	EPI_ISL_3055552	PUE-LABOPAT-91_27256	Puebla	9
6	2021-07-09	EPI_ISL_3055557	PUE-LABOPAT-85_27142	Puebla	9
7	2021-07-12	EPI_ISL_3055561	PUE-LABOPAT-92_21416	Puebla	9
8	2021-10-19	EPI_ISL_6570840	CHP_InDRE_FB49584_S9184	Chiapas	9
9	2021-12-17	EPI_ISL_8184724	PUE-VB21-86817-LABOPAT	Ciudad de México	9
10	2021-12-21	EPI_ISL_8184727	PUE-VE21-87386-LABOPAT	Puebla	9
11	2021-12-28	EPI_ISL_8317248	PUE-VE21-088274-LABOPAT	Puebla	9
12	2021-12-28	EPI_ISL_8317271	TLA-VTX21-088689-LABOPAT	Tlaxcala	9
13	2021-12-28	EPI_ISL_8324835	TLA-VTX21-088696-LABOPAT	Tlaxcala	9
14	2020-06-03	EPI_ISL_3463635	QUE-InDRE-IBT-21066	Queretaro	8
15	2021-01-21	EPI_ISL_1483388	SLP-UASLP_A025	San Luis Potosi	8
16	2021-03-11	EPI_ISL_3235368	YUC-NYGC-1051-SM-2	Yucatán	8
17	2021-06-03	EPI_ISL_3055575	PUE-LABOPAT-2_17717	Puebla	8
18	2021-06-03	EPI_ISL_3055578	PUE-LABOPAT-4_17900	Puebla	8
19	2021-12-22	EPI_ISL_8183503	PUE-VZ21-87471-LABOPAT	Puebla	8
20	2021-12-22	EPI_ISL_8184726	PUE-VZ21-87392-LABOPAT	Puebla	8
21	2021-06-03	EPI_ISL_3055576	PUE-LABOPAT-1_17712	Puebla	7
22	2021-06-03	EPI_ISL_3055577	PUE-LABOPAT-9_19117	Puebla	7
23	2021-07-09	EPI_ISL_3055555	PUE-LABOPAT-80_27103	Puebla	7
24	2021-12-21	EPI_ISL_8184728	PUE-VZ21-87470-LABOPAT	Puebla	7
25	2021-12-28	EPI_ISL_8324841	VTX21-088693-LABOPAT	Tlaxcala	7
26	2021-12-30	EPI_ISL_9430890	BCS_InDRE_FB817_E03313393054_S10605	Baja California Sur	7
27	2022-01-04	EPI_ISL_9749690	ROO_InDRE_FB2490_E23313497660_S11077	Quintana Roo	7
28	2022-01-04	EPI_ISL_9749691	TAB_InDRE_FB2495_E27913507088_S11078	Tabasco	7
29	2022-01-04	EPI_ISL_9749692	TAB_InDRE_FB2502_E27913480959_S11079	Tabasco	7
30	2022-01-04	EPI_ISL_9749693	TAB_InDRE_FB2504_E27913493464_S11081	Tabasco	7

Tabla 5.6: Los 30 genomas identificados como anomalías por la mayor cantidad de criterios. Se muestra la fecha de secuenciación, el identificador del genoma, el laboratorio que lo secuenció, el estado del cual proviene el genoma y el número de criterios que identificaron el genoma como una anomalía.

Capítulo 6

Conclusiones

En este trabajo de tesis, se confirma la hipótesis planteada inicialmente, ya que se proporcionó evidencia de que es posible utilizar diversos algoritmos de aprendizaje no supervisado, así como herramientas computacionales, para detectar los genomas más atípicos dentro de un conjunto de genomas representados como su uso de codones. En concreto, en un conjunto de genomas del virus SARS-CoV-2 secuenciados en México desde el 1 de enero de 2020 hasta el 16 de junio de 2022.

Se comenzó discutiendo las características generales del virus, como su composición genómica y la estructura de este. Abordando brevemente el funcionamiento de sus diferentes genes y señalando que durante el proceso de replicación viral podrían ocurrir mutaciones en ellos. Se mencionó que algunas de estas mutaciones son capaces de provocar cambios significativos en el virus, generando así nuevas variantes de éste. Además, se destacó que se han utilizado métodos de detección de anomalías para tratar de identificar a estas nuevas variantes.

A continuación, se describió el concepto de uso de codones de una secuencia genética. Este fenómeno biológico, desde un punto de vista computacional, posibilitó representar los genomas virales completos con solo 64 variables. La frecuencia de aparición de cada codón permitió tratar a cada genoma como una distribución de probabilidad. De la misma forma, debido a la naturaleza composicional del uso de codones, cada genoma bajo esta representación existe en un espacio simplicial de 63 dimensiones. El conjunto de genomas analizados en esta tesis se representaron según su uso de codones, lo que permitió la aplicación de los diferentes algoritmos y herramientas computacionales abordados en este trabajo.

Herramientas computacionales como la *distancia euclidiana* y la *distancia de Wasserstein*, posibilitaron comparar el uso de codones de cada genoma con respecto al uso de codones humano. Por otro lado, la *entropía* y la *distancia log-cociente esperada* permitieron caracterizar cada genoma teniendo en cuenta únicamente las características inherentes de este conjunto de datos. Los resultados de estas medidas no solo facilitaron el análisis de las diferencias por estado y de

la evolución del uso de codones del virus, sino que también se utilizaron como criterio para identificar los genomas más atípicos dentro del conjunto de genomas.

Analizando estas medidas, se encontró que los estados de Puebla, Tabasco y Tlaxcala presentan la mayor cantidad de genomas con las medidas más divergentes. Asimismo, al examinar el uso de codones promedio por estado, se observó que tanto Puebla como Tlaxcala son los que muestran las mayores diferencias.

El análisis temporal de las medidas mostró que gran parte de estos genomas divergentes se secuenciaron en los meses de julio y diciembre de 2021. Por otro lado, la evolución del uso de codones promedio por semana reveló como después de los primeros meses de 2022, en general, el uso de codones del virus tiende a distanciarse cada vez más del uso de codones humano. Además, se observó que el genoma del virus presenta un *sesgo en el uso de codones* cada vez menor (mostrando menos preferencia por codones sinónimos) y los genomas tienden a alejarse cada vez más del resto de los genomas en el conjunto de datos. Estos resultados coinciden con algunas de las investigaciones previamente mencionadas en los antecedentes de este trabajo, en las cuales el virus presenta una desoptimización en su uso de codones en comparación con el uso de codones humano.

En esta tesis, se emplearon dos algoritmos de reducción de dimensionalidad: el *análisis de componentes principales* y el *mapeo isométrico*. Esto se hizo con el fin de representar cada uso de codones del conjunto de datos únicamente con dos atributos, donde las representaciones generadas por estos algoritmos conservan diferentes aspectos del conjunto de datos original. Los resultados no solo facilitaron la visualización de los datos, sino que también permitieron establecer otro criterio para identificar a los genomas más atípicos del conjunto. En estos nuevos subespacios, la distancia esperada se interpretó como un puntaje de anomalía.

Tanto en la proyección obtenida con *PCA* como en la representación generada por *Isomap*, los genomas que presentaron una mayor distancia esperada, y por consiguiente, se consideraron más anómalos al resto, fueron aquellos secuenciados en los estados de Puebla, Tabasco y Tlaxcala durante mediados y finales de 2021.

Los algoritmos de detección de anomalías que se implementaron en la tesis asumieron características diferentes de lo que es un elemento anómalo. Mientras que el algoritmo *DBSCAN* y la detección de anomalías con *k-medias* se basaron en la agrupación de datos para identificarlos, el algoritmo del *valor atípico local* se basó en los vecinos más cercanos para su detección. Por otro lado, el algoritmo del *bosque de aislamiento* adoptó un enfoque único al aislar a cada instancia.

Como resultado de la aplicación de los algoritmos *DBSCAN*, *k-medias* y *bosque de aislamiento*, la mayoría de los genomas anómalos identificados tienen su origen en los estados de Puebla, Tabasco y Tlaxcala. Por otro lado, la mayor parte de las anomalías identificadas por *LOF* provienen de la Ciudad de México y de Yucatán.

Finalmente, al considerar todos los criterios abordados en esta tesis para la detección de anomalías, se logró identificar un total de 176 genomas anómalos

distintos. Al analizar detenidamente los resultados, se observó que la mayoría de los criterios detectaron repetidamente como anomalías a un mismo conjunto de genomas, la mayoría de los cuales pertenecen al estado de Puebla. Estos genomas anómalos, detectados por la mayoría de los criterios, fueron principalmente identificados a mediados y finales de 2021. Además, se aprecia que el laboratorio *LABOPAT* fue donde se secuenciaron la mayor parte de estas anomalías.

En este trabajo se ha demostrado cómo, a partir de diferentes supuestos de lo que constituye a una anomalía, se logró identificar un conjunto de genomas anómalos. Estos genomas muestran que en su mayoría se secuenciaron en torno a fechas específicas, en ciertos estados de la República y en laboratorios particulares.

Como se mencionó anteriormente, estas secuencias anómalas podrían indicar desde errores en la secuenciación hasta mutaciones en el genoma del virus. Por lo tanto, los resultados podrían estar vinculados a los procedimientos específicos implementados en dichos laboratorios, algo de gran relevancia para las entidades responsables de la vigilancia y control de estos datos. De la misma manera, futuros estudios podrían enfocarse en analizar la composición específica de las secuencias anómalas encontradas, buscando irregularidades y comparándolas con aquellas secuencias que no fueron detectadas como anomalías.

Adicionalmente, es posible comparar los resultados obtenidos a lo largo del tiempo con las primeras apariciones de las variantes del SARS-CoV-2 registradas en México. Esto se podría hacer con el propósito de determinar si existe alguna correlación entre las anomalías detectadas y el surgimiento de estas variantes.

Referencias

[1] P. Lyman *et al.*, "How much information? 2003," UC Berkeley, Tech. Rep., 2003, accessed: march 2023. [Online]. Available: https://groups.ischool.berkeley.edu/archive/how-much-info-2003/ 1

[2] L. Clifton, D. Clifton, P. Watkinson, and L. Tarassenko, "Identification of patient deterioration in vital-sign data using one-class support vector machines." *Proc. Comput. Sci. Inf. Syst*, pp. 125–131, 01 2011. 1, 35

[3] C. Phua, V. Lee, K. Smith-Miles, and R. Gayler, "A comprehensive survey of data mining-based fraud detection research," *CoRR*, vol. abs/1009.6119, 09 2010. 1, 35

[4] P. García-Teodoro, J. Díaz-Verdejo, G. Maciá-Fernández, and E. Vázquez, "Anomaly-based network intrusion detection: Techniques, systems and challenges," *Computers & Security*, vol. 28, no. 1, pp. 18–28, 2009. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0167404808000692 1, 35

[5] K. G. Mehrotra, C. K. Mohan, and H. Huang, *Anomaly Detection Principles and Algorithms*, 1st ed. Springer Publishing Company, Incorporated, 2017. 1, 35, 44

[6] Q. Ferré, J. Chèneby, D. Puthier, C. Capponi, and B. Ballester, "Anomaly detection in genomic catalogues using unsupervised multi-view autoencoders," *BMC Bioinformatics*, vol. 22, no. 1, p. 460, Sep 2021. [Online]. Available: https://doi.org/10.1186/s12859-021-04359-2 1

[7] L. Buck *et al.*, "Anomaly detection in mixed high-dimensional molecular data," *Bioinformatics*, vol. 39, no. 8, p. btad501, 08 2023. [Online]. Available: https://doi.org/10.1093/bioinformatics/btad501 1

[8] M. T. Pervez *et al.*, "A comprehensive review of performance of next-generation sequencing platforms," *BioMed Research International*, vol. 2022, Sep 2022, Article ID 3457806. [Online]. Available: https://doi.org/10.1155/2022/3457806 2

[9] T. Lencz and A. Darvasi, "Genome," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 292–293. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840006409 2, 10

[10] J. Perona, "Dna," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2001, pp. 339–340. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012374984000437X 2

[11] D. Simón, J. Cristina, and H. Musto, "Nucleotide composition and codon usage across viruses and their respective hosts," *Frontiers in Microbiology*, vol. 12, 6 2021. 2, 5, 19

[12] S. Khare *et al.*, "Gisaid's role in pandemic response," *China CDC Weekly*, vol. 3, p. 1049, 2021. [Online]. Available: https://weekly.chinacdc.cn//article/id/21792cdf-a54a-4a11-b6fe-68d50f817d91 2, 6, 11, 12, 29, 55, 56

[13] M. Ibba, "Genetic code," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 234–235. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840006094 3, 11

[14] M. Dilucca, S. Forcelloni, A. G. Georgakilas, A. Giansanti, and A. Pavlopoulou, "Codon usage and phenotypic divergences of sars-cov-2 genes," *Viruses*, vol. 12, 5 2020. 4, 5

[15] W. Hou, "Characterization of codon usage pattern in sars-cov-2," *Virology Journal*, vol. 17, 9 2020. 4, 5, 9, 11, 12

[16] L. L. Maldonado, A. M. Bertelli, and L. Kamenetzky, "Molecular features similarities between sars-cov-2, sars, mers and key human genes could favour the viral infections and trigger collateral effects," *Scientific Reports*, vol. 11, 12 2021. 5

[17] E. Posani *et al.*, "Temporal evolution and adaptation of sars-cov-2 codon usage," *Frontiers in Bioscience - Landmark*, vol. 27, 1 2022. 5, 6, 9, 17, 19

[18] G. Hahn *et al.*, "Unsupervised outlier detection applied to sars-cov-2 nucleotide sequences can identify sequences of common variants and other variants of interest," *BMC Bioinformatics*, vol. 23, 12 2022. 5, 6, 7, 17, 36

[19] Y. Nakamura, T. Gojobori, and T. Ikemura, "Codon usage tabulated from international DNA sequence databases: status for the year 2000," *Nucleic Acids Research*, vol. 28, no. 1, pp. 292–292, 01 2000. [Online]. Available: https://doi.org/10.1093/nar/28.1.292 6, 59

[20] J. Cui, F. Li, and Z. L. Shi, "Origin and evolution of pathogenic coronaviruses," *Nature Reviews Microbiology*, vol. 17, pp. 181–192, 3 2019. 9, 10

[21] P. V'kovski *et al.*, "Coronavirus biology and replication: implications for sars-cov-2," *Nature Reviews Microbiology*, vol. 19, pp. 155–170, 3 2021. 9, 12, 14, 15, 16

[22] WHO. (2020) Who director-general's opening remarks at the media briefing on covid-19 - 11 march 2020. Accessed: march 2023. [Online]. Available: https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020 9

[23] J. F.-W. Chan *et al.*, "Genomic characterization of the 2019 novel human-pathogenic coronavirus isolated from a patient with atypical pneumonia after visiting wuhan," *Emerging Microbes and Infections*, vol. 9, no. 1, pp. 221–236, 2020, pMID: 31987001. [Online]. Available: https://doi.org/10.1080/22221751.2020.1719902 9, 12

[24] H. Singh, N. Dahiya, M. Yadav, and N. Sehrawat, "Emergence of sars-cov-2 new variants and their clinical significance," *Canadian Journal of Infectious Diseases and Medical Microbiology*, vol. 2022, 2022. 9, 16, 17

[25] S. Sessions, "Genome size," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 301–305. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840006392 10

[26] B. Guttman, "Virus," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 291–294. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840016260 10, 14

[27] G. Fermin, "Chapter 2 - virion structure, genome organization, and taxonomy of viruses," in *Viruses*, P. Tennant, G. Fermin, and J. E.

Foster, Eds. Academic Press, 2018, pp. 17–54. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780128112571000024 10, 15

[28] A. R. Fehr and S. Perlman, "Coronaviruses: An overview of their replication and pathogenesis," *Coronaviruses: Methods and Protocols*, pp. 1–23, 2 2015. 10

[29] I. Schildkraut, "Codons," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2001, p. 65. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840002813 11

[30] P. Sharp, "Codon usage bias," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2001, pp. 67–69. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123749840002795 11, 16

[31] N. Zhu *et al.*, "A novel coronavirus from patients with pneumonia in china, 2019," *New England Journal of Medicine*, vol. 382, no. 8, pp. 727–733, 2020, pMID: 31978945. [Online]. Available: https://doi.org/10.1056/NEJMoa2001017 15

[32] F. Pereira and A. Amorim, "Evolution: Viruses," in *Brenner's Encyclopedia of Genetics (Second Edition)*, 2nd ed., S. Maloy and K. Hughes, Eds. San Diego: Academic Press, 2013, pp. 566–568. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B978012374984000499X 16

[33] WHO, "Tracking sars-cov-2 variants," World Health Organization, marzo 2023, accessed: abril 2023. [Online]. Available: https://www.who.int/activities/tracking-SARS-CoV-2-variants 16

[34] M. Greenacre, *Compositional Data Analysis in Practice*, 1st ed. Chapman and Hall/CRC, 07 2018. 20, 21, 22, 24, 31

[35] S. S. Skiena, *The Data Science Design Manual*, 1st ed. Springer Cham, 07 2017. 23, 28, 30

[36] S.-H. Cha, "Comprehensive survey on distance/similarity measures between probability density functions," *Int. J. Math. Model. Meth. Appl. Sci.*, vol. 1, 01 2007. 23

[37] Y. Rubner, C. Tomasi, and L. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No.98CH36271)*, 1998, pp. 59–66. 25

[38] K. Mehrotra, C. Mohan, and H. Huang, *Anomaly Detection Principles and Algorithms*. Springer Cham, 01 2017. 26

[39] K. P. Murphy, *Probabilistic Machine Learning: An introduction*. MIT Press, 2022. [Online]. Available: probml.ai 27, 29, 30, 32, 33

[40] D. J. C. MacKay, *Information theory, inference and learning algorithms*. Cambridge University Press, 2003. 28

[41] A. Géron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O'Reilly Media, Inc., 2017. 29, 30, 32

[42] J. B. Tenenbaum, V. de Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *Science*, vol. 290, no. 5500, pp. 2319–2323, 2000. [Online]. Available: https://www.science.org/doi/abs/10.1126/science.290.5500.2319 33

[43] M. A. Pimentel, D. A. Clifton *et al.*, "A review of novelty detection," *Signal Processing*, vol. 99, pp. 215–249, 2014. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S016516841300515X 35, 41

[44] M. Goldstein and S. Uchida, "A comparative evaluation of unsupervised anomaly detection algorithms for multivariate data," *PLoS ONE*, vol. 11, 4 2016. 35, 36, 37, 38, 39, 41, 46

[45] C. C. Aggarwal, *Outlier Analysis*, 2nd ed. Springer Publishing Company, Incorporated, 2016. 35, 37, 41

[46] V. Chandola, A. Banerjee, and V. Kumar, "Anomaly detection: A survey," *ACM Comput. Surv.*, vol. 41, no. 3, jul 2009. [Online]. Available: https://doi.org/10.1145/1541880.1541882 36, 37, 39, 44, 45

[47] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, ser. KDD'96. AAAI Press, 1996, p. 226–231. 41, 42

[48] M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander, "Lof: Identifying density-based local outliers," in *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, ser. SIGMOD '00. New York, NY, USA: Association for Computing Machinery, 2000, p. 93–104. [Online]. Available: https://doi.org/10.1145/342009.335388 45

[49] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation forest," in *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ser. ICDM '08. USA: IEEE Computer Society, 2008, p. 413–422. [Online]. Available: https://doi.org/10.1109/ICDM.2008.17 48, 49, 50

[50] C. R. Harris *et al.*, "Array programming with NumPy," *Nature*, vol. 585, no. 7825, pp. 357–362, Sep. 2020. [Online]. Available: https://doi.org/10.1038/s41586-020-2649-2 55

[51] P. Virtanen *et al.*, "SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python," *Nature Methods*, vol. 17, pp. 261–272, 2020. 55

[52] Wes McKinney, "Data Structures for Statistical Computing in Python," in *Proceedings of the 9th Python in Science Conference*, Stéfan van der Walt and Jarrod Millman, Eds., 2010, pp. 56 – 61. 55

[53] J. D. Hunter, "Matplotlib: A 2d graphics environment," *Computing in Science & Engineering*, vol. 9, no. 3, pp. 90–95, 2007. 55

[54] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011. 55