



Trabajo Fin de Grado

Reconstrucción computacional rápida de árboles
filogenéticos de SARS-CoV-2

Autor

Óscar Gómez Ortego

Directores

Elvira Mayordomo Cámara
Mónica Hernández Giménez

ESCUELA DE INGENIERÍA Y ARQUITECTURA
2022



DECLARACIÓN DE AUTORÍA Y ORIGINALIDAD

(Este documento debe acompañar al Trabajo Fin de Grado (TFG)/Trabajo Fin de Máster (TFM) cuando sea depositado para su evaluación).

D./Dª. _____,

con nº de DNI _____ en aplicación de lo dispuesto en el art. 14 (Derechos de autor) del Acuerdo de 11 de septiembre de 2014, del Consejo de Gobierno, por el que se aprueba el Reglamento de los TFG y TFM de la Universidad de Zaragoza,

Declaro que el presente Trabajo de Fin de (Grado/Máster) _____, (Título del Trabajo)

es de mi autoría y es original, no habiéndose utilizado fuente sin ser citada debidamente.

Zaragoza, _____

Fdo: _____

AGRADECIMIENTOS

We gratefully acknowledge all data contributors, i.e., the Authors and their Originating laboratories responsible for obtaining the specimens, and their Submitting laboratories for generating the genetic sequence and metadata and sharing via the GISAID Initiative, on which this research is based.

y especialmente a los alumnos que hacen plantillas de LaTeX.

Título del resumen

RESUMEN

Una mañana, tras un sueño intranquilo, Gregorio Samsa se despertó convertido en un monstruoso insecto. Estaba echado de espaldas sobre un duro caparazón y, al alzar la cabeza, vio su vientre convexo y oscuro, surcado por curvadas callosidades, sobre el que casi no se aguantaba la colcha, que estaba a punto de escurrirse hasta el suelo. Numerosas patas, penosamente delgadas en comparación con el grosor normal de sus piernas, se agitaban sin concierto. - ¿Qué me ha ocurrido? No estaba soñando. Su habitación, una habitación normal, aunque muy pequeña, tenía el aspecto habitual. Sobre la mesa había desparramado un muestrario de paños - Samsa era viajante de comercio-, y de la pared colgaba una estampa recientemente recortada de una revista ilustrada y puesta en un marco dorado. La estampa mostraba a una mujer tocada con un gorro de pieles, envuelta en una estola también de pieles, y que, muy erguida, esgrimía un amplio manguito, asimismo de piel, que ocultaba todo su antebrazo. Gregorio miró hacia la ventana; estaba nublado, y sobre el cinc del alféizar repiqueteaban las gotas de lluvia, lo que le hizo sentir una gran melancolía. «Bueno -pensó-; ¿y si siguiese durmiendo un rato y me olvidase de

Índice

1. Introducción y objetivos	IX
1.1. Filogenética	IX
1.2. Compresión para cálculo de filogenias	X
1.3. Sobre el COVID-19 y sus fuentes de datos	X
2. Despliegue y replicación del sistema	XIII
2.1. Despliegue	XIII
2.2. Replicación	XIV
3. Comparación de los resultados obtenidos con el estado del arte	XVII
3.0.1. Criterios de comparación	XVII
3.1. Datos obtenidos de nextstrain	XX
3.1.1. Método 2	XXI
3.2. Datos obtenidos de GISAID	XXV
3.3. Sobre el software Nextstrain y su comparación con NCD	XXVIII
4. Escalabilidad y estudio del sistema	XXXI
4.1. Optimización del sistema	XXXII
4.1.1. Uso de las distancias de compresión en árboles filogenéticos . . .	XXXII
4.1.2. Pruebas de carga	XXXIII
5. Clasificación de variantes de interés	XXXIX
6. Compresión para otro tipo de secuencias	XLI
6.1. ADN mitocondrial	XLI
6.2. Monkeypox o viruela de mono	XLII
7. Conclusiones	XLV
8. Bibliografía	XLVII
Lista de Figuras	XLIX

Capítulo 1

Introducción y objetivos

1.1. Filogenética

La filogenética se encarga de establecer una relación entre organismos mediante un árbol o cladograma dicotómico que representa una hipótesis evolutiva. Se divide en varias ramas: la filogenética morfológica y la filogenética molecular. La filogenética morfológica establece relaciones entre seres vivos en base a similitudes morfológicas o anatómicas. La filogenética molecular, base de la bioinformática, investiga las relaciones mediante el análisis de secuencias de ADN, ARN o proteínas y gracias a los algoritmos computacionales logra obtener estos árboles filogenéticos.

El método adoptado normalmente para la creación de filogenias está basado en el alineamiento múltiple de las secuencias de ADN o ARN, compuestas por los aminoacidos A,C,G,T en el caso de ADN y por A,C,G,U en el caso del ARN. Este método se basa en la detección de diferencias entre dos secuencias. También aparece la adición/borrado de gaps(denotado por " - ") en las secuencias para que coincida con más términos y así lograr la optimización global del alineamiento. No obstante, dada la intratabilidad del problema MSA(multiple sequence alignment) y su coste exponencial hace que normalmente se haga uso de heurísticas que encuentren una solución subóptima en un tiempo mucho menor.

El siguiente paso para la formación de filogenias aparecen dos métodos: los basados en secuencias y los basados en distancias. Los basados en secuencias funcionan generando todos los árboles posibles y eligiendo luego los más adecuados según los datos y otros parámetros previamente establecidos, entre estos están maxima verosimilitud y maxima parsimonia.

Los basados en distancias, concepto implícito al alineamiento, se trata de establecer distancias entre las diferentes secuencias del alineamiento, crear una matriz con ellas y realizar un agrupamiento. Los métodos de clustering más conocidos son UPGMA (Unweighted Pair-Group Method with Arithmetic) y Neighbor-joining.

1.2. Compresión para cálculo de filogenias

La idea que subyace en los métodos de compresión es librarse de los alineamientos y los inconvenientes que estos acarrean y calcular las distancias mediante algún algoritmo de compresión. Este trabajo se basará en el artículo [1], que desarrolló el método NCD. En el caso de NCD (Normalized Compressed Distance) se trata de una aproximación de la NID (Normalized Information Distance) dada su intratabilidad, que se basa en la complejidad de Kolmogorov y sigue la siguiente fórmula:

$$NCD(x, y) = \frac{Z(xy) - \min(Z(x), Z(y))}{\max(Z(x), Z(y))}$$

Siendo $NCD(x, y)$ la distancia entre las cadenas x e y, $Z(x)$ el tamaño de la compresión de x, y $Z(xy)$ el tamaño de la compresión de la concatenación de x e y. NCD establece una distancia entre 0 y 1 entre dos secuencias, siendo 0 totalmente iguales y 1 totalmente diferentes.

De esta forma el coste de la creación de la matriz depende de la eficiencia del compresor.

Es importante destacar aquí que en el artículo original de Vitanyi [2] como en la tesis de Vacca(ref) se exploran datasets de muy pocas secuencias. Vitanyi muestra árboles de como mucho 60 secuencias y Vacca habla de varios datasets de alrededor de 20 secuencias.

Aquí aparece la motivación de este trabajo, ampliar estos datasets de prueba, probar los límites de este algoritmo y compararlo con otros métodos de hoy en día como augur.

1.3. Sobre el COVID-19 y sus fuentes de datos

El coronavirus SARS-CoV-2, descubierto en enero de 2020 tras aislarse de muestras de pacientes afectados por una nueva enfermedad ahora conocida como COVID-19, evoluciona y sufre cambios genéticos, como todos los virus. Conocer estos cambios, que explican su comportamiento, es fundamental para mejorar el manejo del virus y el abordaje de la enfermedad. Desde el hallazgo del SARS-CoV-2 hasta ahora, la secuenciación de su genoma y el conocimiento de las diferentes variantes que circulan por el mundo está permitiendo conocer más sobre su origen, influencia y distribución.

Gracias a un esfuerzo sin precedentes han ido apareciendo numerosas herramientas para realizar el seguimiento de este virus. Nextstrain [1] muestra prácticamente a tiempo real la epidemiología genómica del virus y como ha ido transmitiéndose a lo largo del tiempo por los diferentes países afectados.

Actualmente, Nextstrain [1] define numerosos grandes clados filogenéticos para clasificar los genomas que se van secuenciando y que se nombran en función del

año estimado en el que emergieron seguido de una letra, estos agrupan las diferentes mutaciones que va sufriendo el virus y se detallarán posteriormente.

Por su parte, GISAID [3], actúa como base de datos y pone a disposición secuencias genéticas y datos clínicos y epidemiológicos para ayudar a los investigadores a comprender cómo evolucionan y se propagan los virus durante epidemias y pandemias. Actualmente, cuentan con 12,963,236 genomas de COVID-19 que comprenden desde 2019-12-24 hasta 2022-08-31.

Capítulo 2

Despliegue y replicación del sistema

2.1. Despliegue

Para replicar los resultados del artículo original de Cilibrasi [2] se usó el proyecto open-source del mismo autor disponible en Github: github.com/rudi-cilibrasi/ncd-covid.

El proyecto se realizó en un ordenador personal con sistema operativo ubuntu 22.04 LTS, un procesador Intel Core i5-5200U, con 4 núcleos a 2,2GHz y 4 GB de ram.

Una vez se disponía del código se descargaron una serie de herramientas necesarias:

- **Fastahack**, una utilidad para la extracción e indexado de secuencias en ficheros FASTA.
- **RocksDB**, la base de datos clave-valor de alto rendimiento, para guardar los tamaños de las compresiones de las secuencias.
- **Zpaq** como el compresor elegido para la realización del método.

La elección del compresor es de gran importancia porque sobre él recaerá la eficiencia y calidad del método. Vacca [4] publicó una tesis de licenciatura anterior a la publicación del artículo en el que se basa este trabajo [2] tratando temas como la elección de compresor y el cálculo de similitud entre árboles. Vacca explica que para elegir un compresor tiene que cumplir 4 propiedades:

- idempotencia, $Z(xx) \approx Z(x)$
- monotonicidad, $Z(xy) \geq Z(x)$
- simetría, $Z(xy) \approx Z(yx)$
- distributividad, $Z(xy) + Z(z) \leq Z(xz) + Z(yz)$

($Z(x)$ es el tamaño de la compresión de x, $Z(xy)$ es el tamaño de la compresión de la concatenación de x e y).

Realizó un estudio exhaustivo para comprobar estas características entre los compresores: Bzip2, GenCompress, Lrzip, Gzip.

Por otra parte, Cilibrasi también comparaba 3 compresores: gzip, bzip2 y PPMZ. Se establecía que el compresor ideal debería cumplir que $NCD(x, x) = 0$, con esto, gzip era el peor compresor, normalmente entre 0.5 y 1, y PPMZ el mejor(aunque el más lento) con valores entre 0.002 y 0.006.

En nuestro caso, usamos el compresor **Zpaq** propuesto por Cilibrasi. Un compresor de la familia paq que aplica conceptos relacionados con los compresores PPM (Predicción por Coincidencia Parcial) y que fue lanzado en 2009.

- **Cmake**, para la compilación del proyecto (escrito completamente en C++).

Sobre los datos que hacía uso Cilibrasi en [2] estos están a disposición pública mediante el repositorio: github.com/rudi-cilibrasi/ncd-covid-data. Además, se descargaron 29 secuencias requeridas por los autores a través de GISAID [3], 28 secuencias de COVID-19 y la secuencia de Betacoronavirus RaTG13.

2.2. Replicación

El flujo de trabajo para la obtención de la matriz de distancias no se parece en nada a los demás métodos basados en alineamiento. Este método se basa en interacción con una base de datos clave-valor que almacena los tamaños de compresión de las secuencias. Esta organizado en una serie de programas escritos en C++ como se puede observar en la figura 2.1.

- Importer/Cleaner: se encargan de procesar los ficheros FASTA de entrada y establecer las claves en la base de datos. Realiza un filtrado para quedarse

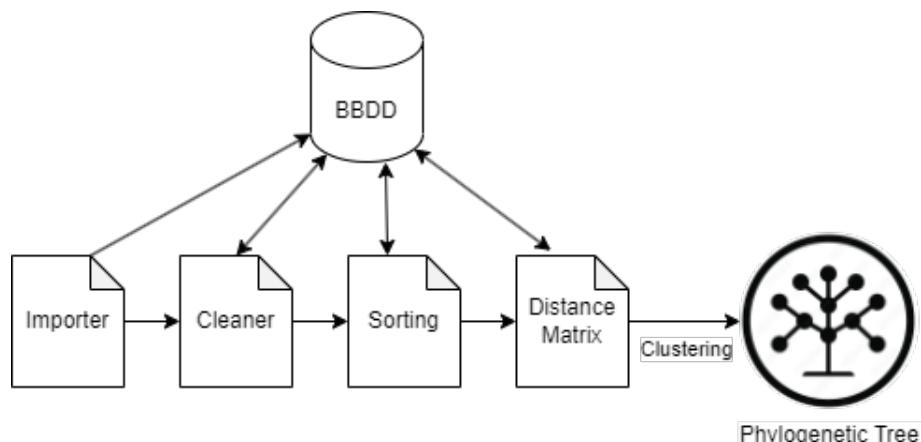


Figura 2.1: Esquema del flujo de trabajo de NCD

únicamente con secuencias que contienen los nucleótidos A, C, G y T, se presenta una de las pequeñas desventajas del método, numerosos genomas de hoy en día no cuentan con una secuenciación completa, aparecen regiones desconocidas representadas por el carácter N. Este fenómeno en métodos basados en alineamiento se puede controlar pero en NCD se descarta directamente la secuencia.

- Sorting/Matrix Creation: son 2 programas que se encargan de calcular la distancia a la secuencia de referencia y calcular la distancias de las n secuencias que constituiran la matriz entre ellas. Se explican posteriormente en 4.
- Clustering: se realiza un clustering mediante el comando MakeTree, establece el orden parcial inducido por la matriz de distancias y no añade distancias de las ramas al árbol(posteriormente se explorará este y otro método de clustering en 4.1.1).

Los resultados obtenidos de la ejecución del flujo de trabajo fueron idénticos a los del trabajo original. Se pueden encontrar en [enlace al repositorio]

Capítulo 3

Comparación de los resultados obtenidos con el estado del arte

3.0.1. Criterios de comparación

Para la comparación de los árboles filogenéticos que se irán obteniendo a lo largo del proyecto se van a realizar numerosos análisis:

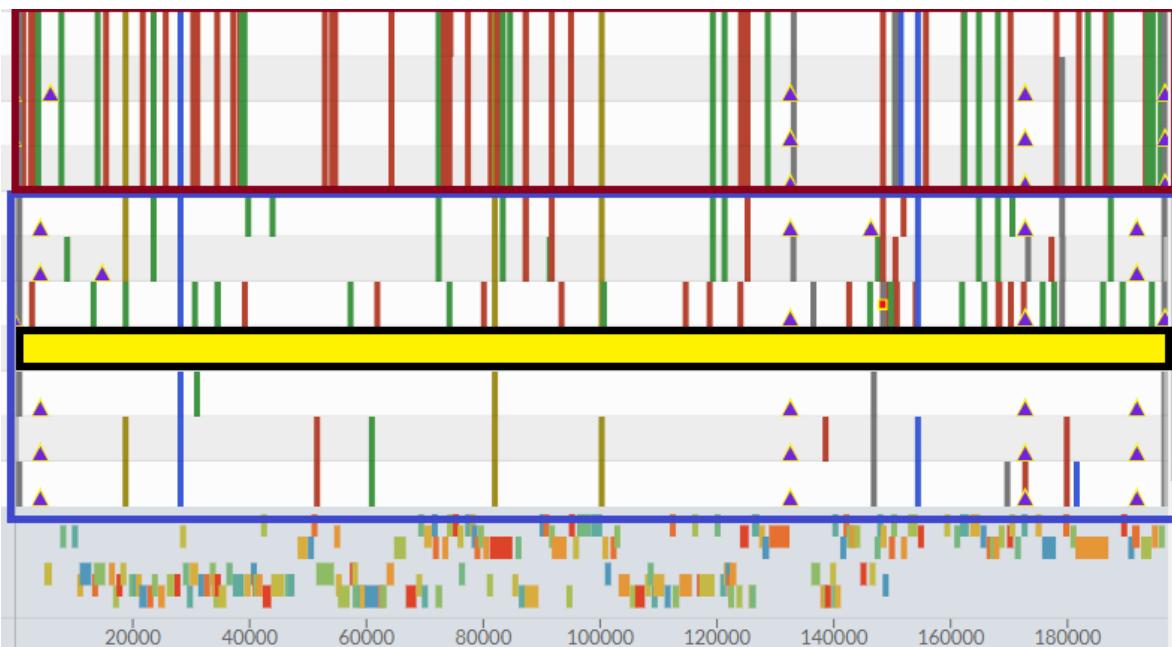
- Análisis visual por clados: Un clado es una agrupación que contiene un antepasado común y todos los descendientes de ese antepasado, en filogenética se resume a cada rama del árbol que agrupa las distintas secuencias. Dependiendo de las mutaciones que vayan sucediendo en las secuencias, se le asigna un nombre a ramas/clados que comparten las mismas mutaciones con la idea de crear clusters visuales.

Una de las maneras más sencillas de formar clados actualmente es mediante nextstrain y su herramienta nextclade.org [5] que *a grosso modo* realiza asignamiento de clados, marcado de mutaciones and comprobaciones de calidad de secuencias. Esta cuenta con diferentes datasets de virus de la actualidad y al seleccionar un conjunto de secuencias como entrada las etiqueta y clasifica. Finalmente muestra en una tabla diferentes datos sobre las secuencias y las mutaciones que han sufrido. En ciertos casos en los que el dataset es conocido como en el caso del COVID-19 o de Monkeypox se podrá ver una vista contextual de las secuencias clasificadas en el actual árbol que maneja nextstrain (imagen disponible en el github).

Para el análisis de árboles que ya hemos construido usaremos la herramienta auspice.us [1] (también forma parte de nextstrain) que permite la exploración de datasets filogenéticos.

El funcionamiento, en primera instancia, es como un visualizador de árboles común, pero permite añadir metadatos. Mediante los filtros que se asocian a

estos metadatos se aprecia una visualización mucho más rica y que, en especial, para árboles grandes es de gran ayuda.



(a) Fragmento del análisis realizado por nextclade, en este caso sobre el virus monkeypox, que muestra las mutaciones que se van realizando a lo largo de las secuencias con respecto a la de referencia (la marcada en amarillo). En el rectángulo azul se puede observar que esas secuencias comparten la mayoría de mutaciones y pertenecen a un cluster similar, mientras que las secuencias en el rectángulo rojo comparten muchas más y pertenecen a otro cluster diferente



(b) Esquema de colores para el resultado de las diferentes mutaciones en las secuencias

Figura 3.1: Captura del análisis realizado por nextclade.org

- Herramientas de comparación y visualización de árboles: a pesar de que hay métricas asociadas a la comparación de árboles como la distancia robinson-foulds o la distancia de cuartetos que se puede calcular mediante el software visualtreeccmp, estas presentan ciertas desventajas e irregularidades. Es difícil comparar árboles solamente con métricas, solo si la métrica es muy diferente se puede llegar a conclusiones. Por ello se va a hacer uso de 2 herramientas de visualización.

Estas son phylo.io [6] y iphyloC [7]. Son dos herramientas muy similares, están orientadas a la comparación de 2 árboles uno al lado del otro. Implementan algunas características que son bastante útiles para el análisis, entre ellas está el

resaltar las similitudes y diferencias entre dos árboles, identificación automática de la mejor coincidencia de orden de raíces y hojas y escalabilidad a árboles grandes.

3.1. Datos obtenidos de nextstrain

Nextstrain es un proyecto para aprovechar el potencial científico y de salud pública de los datos del genoma de patógenos. Su objetivo es ayudar a la comprensión epidemiológica de la propagación y evolución de patógenos y mejorar la respuesta a los brotes. Nextstrain se basa en tres partes: la web, en la que se puede realizar el seguimiento en tiempo real de gran cantidad de patógenos, el software, que ya hemos ido comentando, de código abierto y disponible para el público y una plataforma para compartir resultados de la comunidad.

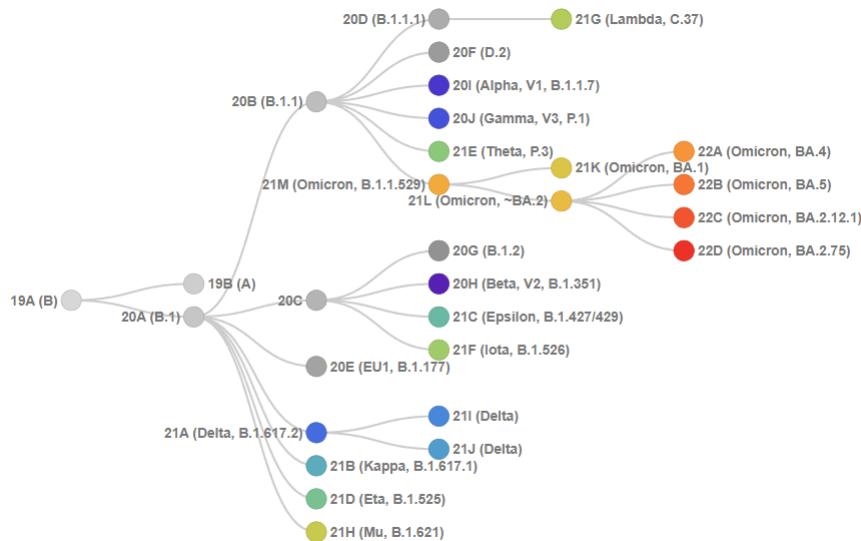


Figura 3.2: Esquema de la nomenclatura de clados en Nextstrain

El dataset de esta prueba fue obtenido de nextstrain.org con el filtro de ncov/open/global/6m. Un total de 2491 secuencias de COVID-19 que comprenden el periodo de 12/2019 hasta 7/2022.

La prueba a realizar será obtener las 100 secuencias más proximas a la de referencia. Para comparar el resultado obtenido de los 100 virus SARS-CoV-2 obtenidos con los sistemas actuales utilizamos 2 métodos. En primer lugar se comparó con un subconjunto de 479 secuencias que contienen las 100 elegidas por el programa y que contienen secuencias entre 12/2019 y 08/2021. Cabe destacar que a pesar de que el subperiodo son casi 2/3 partes del periodo original, comprenden 1/5 del total de secuencias. La mayor parte de las secuencias pertenecen a los clados 21.x y 22.x y a variantes como la Delta o la Omicron respectivamente. Esto se deba al dataset elegido, que se enfoca en un submuestreo global en los últimos 6 meses.

En segundo lugar se comparó a nivel filogenético con el mismo conjunto de 100 secuencias seleccionadas. Sin embargo, se obtuvo la filogenia con un método de la

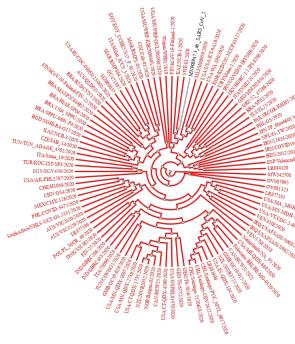
actualidad, en este caso, con la herramienta MEGA mediante un alineamiento previo (una secuencia no se tuvo en cuenta), en este caso nextstrain.org ya proporcionaba la colección de secuencias alineadas por lo que la comparación en cuanto a tiempo no se tiene en cuenta. Se realizó el método de máxima verosimilitud para la creación de la filogenia.

3.1.1. Método 2

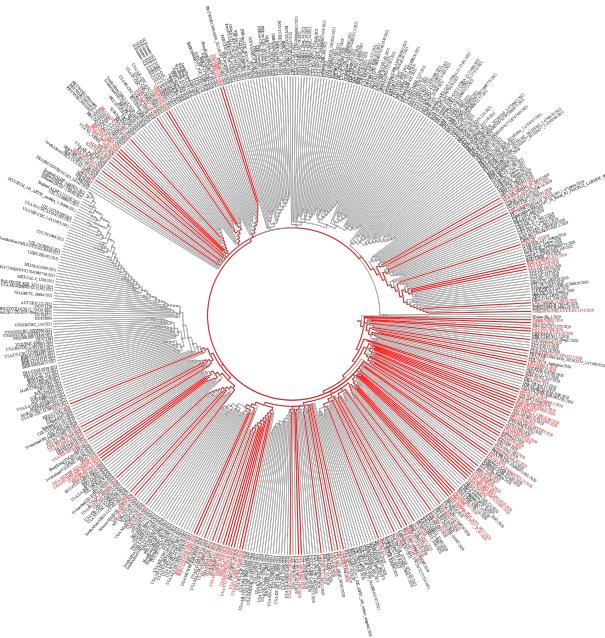
En este apartado compararemos la filogenia creada por NCD compuesta por las 100 secuencias más cercanas a la de referencia que en este caso es Wuhan-Hu-1/2019 y la filogenia generada por el software MEGA[8]. El dataset ya contaba con secuencias alineadas. Se uso el método de máxima verosimilitud para la creación del árbol.

Añadiendo ambos árboles con los metadatos a auspice.us se puede ver que la clasificación en clados es bastante buena, cabe destacar que al ser solamente 100 secuencias de un conjunto de bastantes solo aparecen los clados 19.x y 20.x correspondientes a una etapa temprana del virus que se ubica en 2020. Se puede detectar en las secuencias del clado 20.C que hay un error en un par de hojas(COL/Cali-01/2020 y MAR/RMPS-05/2020) que se ubican entre variantes del clado 20.A.

Finalmente, se cuenta con el resultado del flujo de trabajo de Nextstrain sobre las secuencias comentadas. Se encuentra disponible en nextstrain.org/community/nextstrainOGO/nextstrainSamples. Sobre él se puede hacer uso de gran cantidad de filtros, inspecciónar las diversas mutaciones y realizar la animación sobre el mapa de la expansión de las diferentes muestras en el tiempo.

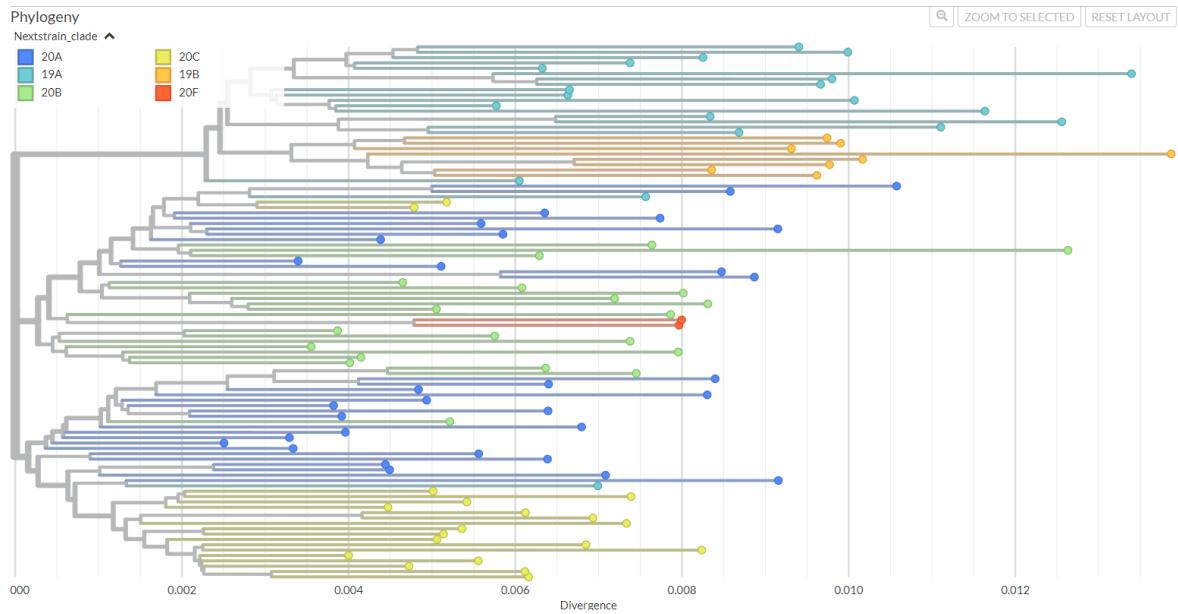


(a) Árbol salida

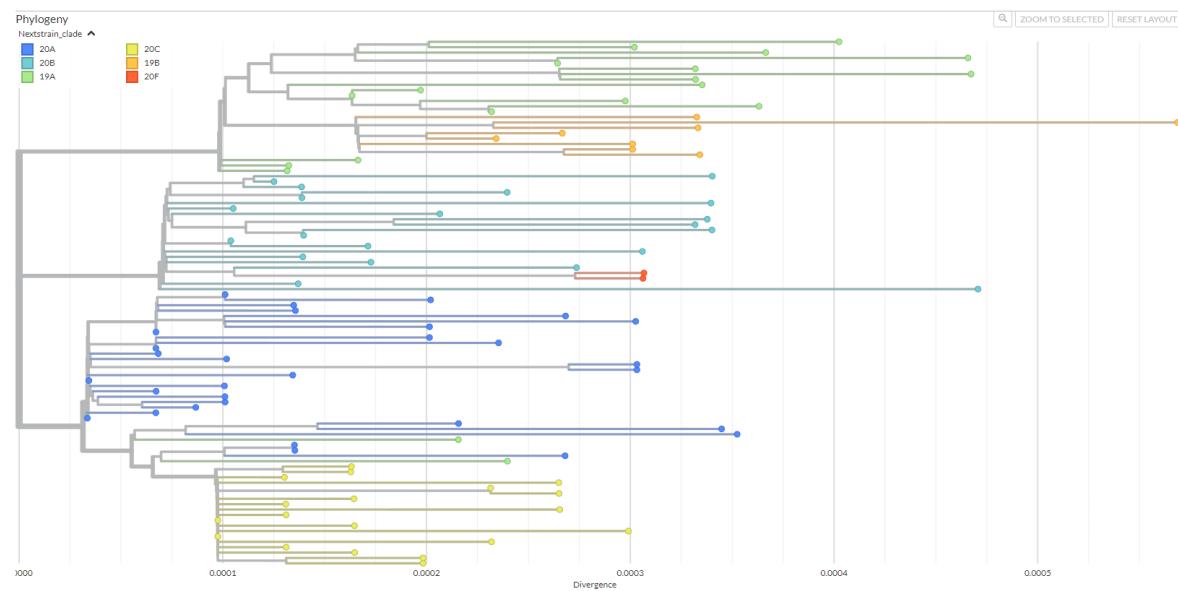


(b) Árbol subselección

Figura 3.3: Árbol salida del programa(a) y árbol con una subselección(b) de alrededor de 500 secuencias del conjunto inicial, en rojo las secuencias correspondientes al árbol a en ambos árboles



(a) árbol filogenético resultado del flujo de trabajo de NCD



(b) árbol filogenético resultado de MEGA(alineamiento y método de máxima verosimilitud)

Figura 3.4: Redactar

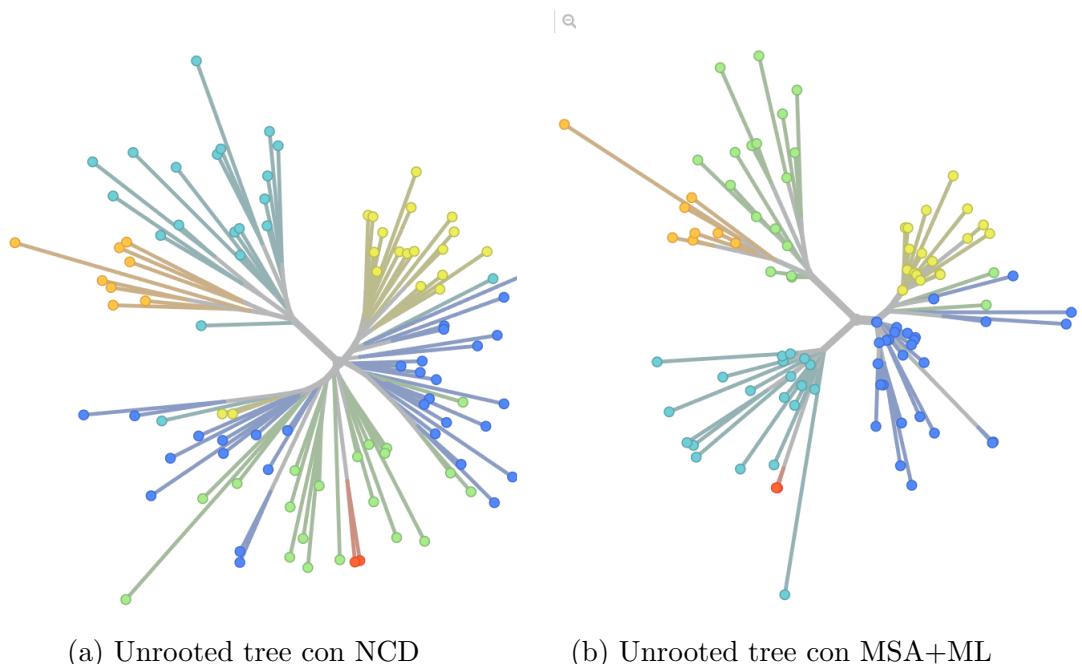


Figura 3.5: Mediante la opción unrooted tree de auspice.us se puede observar la calidad de los clusters en los árboles. Ambos casos se diferencian los clados 19.x, en la rama superior izquierda de estos. Sobre los clados 20.x, en NCD se observa una diferenciación más difusa en los clados 20.B(verde) y 20.A(azul). En el árbol de MEGA se distinguen perfectamente los clados 20.x(cyan,rojo,azul y amarillo).

3.2. Datos obtenidos de GISAID

Gisaid es una iniciativa que surgió en 2008 para el intercambio de datos del virus influenza. A día de hoy incluye datos de patógenos como influenza, COVID-19 y Monkeypox. Esto incluye secuencias genéticas y datos clínicos y epidemiológicos relacionados con virus humanos. Gisaid, al contrario que Nextstrain, requiere de un registro con verificación. Un poco menos accesible que Nextstrain pero proporciona una cantidad de datos mucho mayor. Cabe destacar que GISAID permite integración con Nextstrain. A la hora de descargar los datos se puede elegir el formato especificado por el software de Nextstrain. Una característica de GISAID es que usa otra nomenclatura para los clados:



Figura 3.6: Árbol con la nomenclatura de clados de GISAID, correspondencia con nextstrain y variante que marca el clado.

Mediante nextclade.org y el análisis que realiza siempre se puede establecer una correspondencia para secuencias nuevas, por tanto, la nomenclatura que se usará será la de nextstrain.

Para investigar la fuente de datos GISAID aprovechamos el apartado que tiene Nextstrain sobre datos de esta. Se hará uso del dataset ncov/gisaid/europe/6m de nextstrain. Al no ser un dataset abierto el procedimiento para descargar secuencias y metadatos es el siguiente: en la pestaña “download data” de nextstrain descargar el fichero ACKNOWLEDGMENTS(TSV), logearse en GISAID, introducir este fichero en GISAID en la opción análisis, seleccionar la opción de descarga *Input for the Augur pipeline*.

Este dataset cuenta con 3350 secuencias de COVID-19, fundamentalmente de Europa y centrada en el muestreo de los últimos 6 meses.

Para esta prueba se elegirán 50 secuencias linealmente separadas, es decir, no se considerarán las n secuencias más próximas a la de referencia sino que se elegirán uniformemente entre las secuencias, con el objetivo de poner en contexto las magnitudes de NCD que se están comparando y que puedan aparecer otras secuencias de variantes más recientes en comparación con la de referencia.

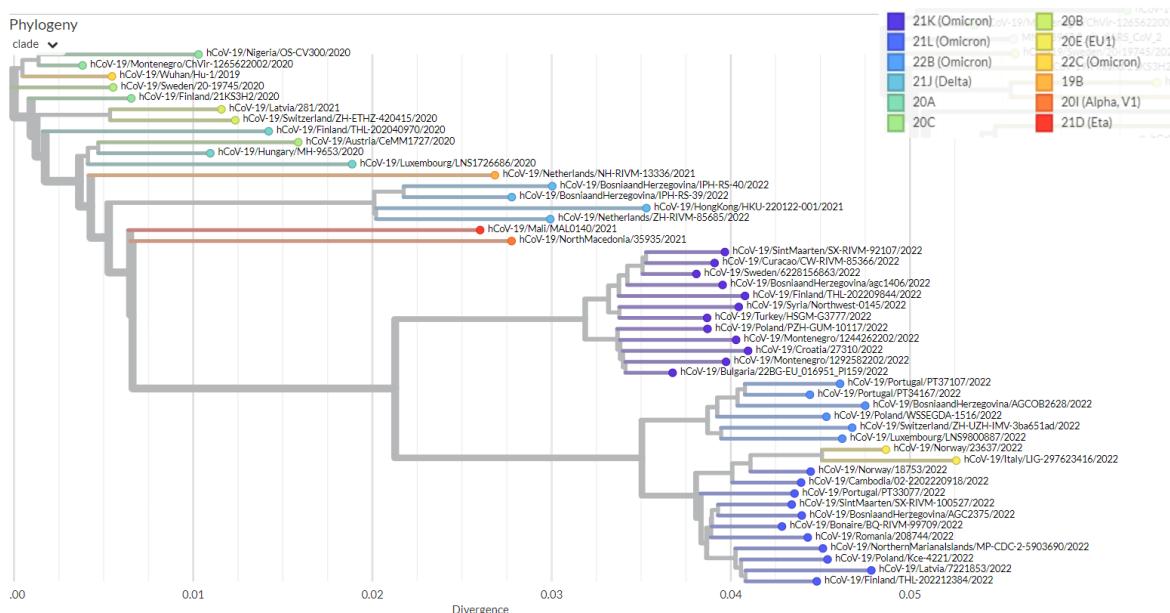


Figura 3.7: Árbol resultado del método NCD

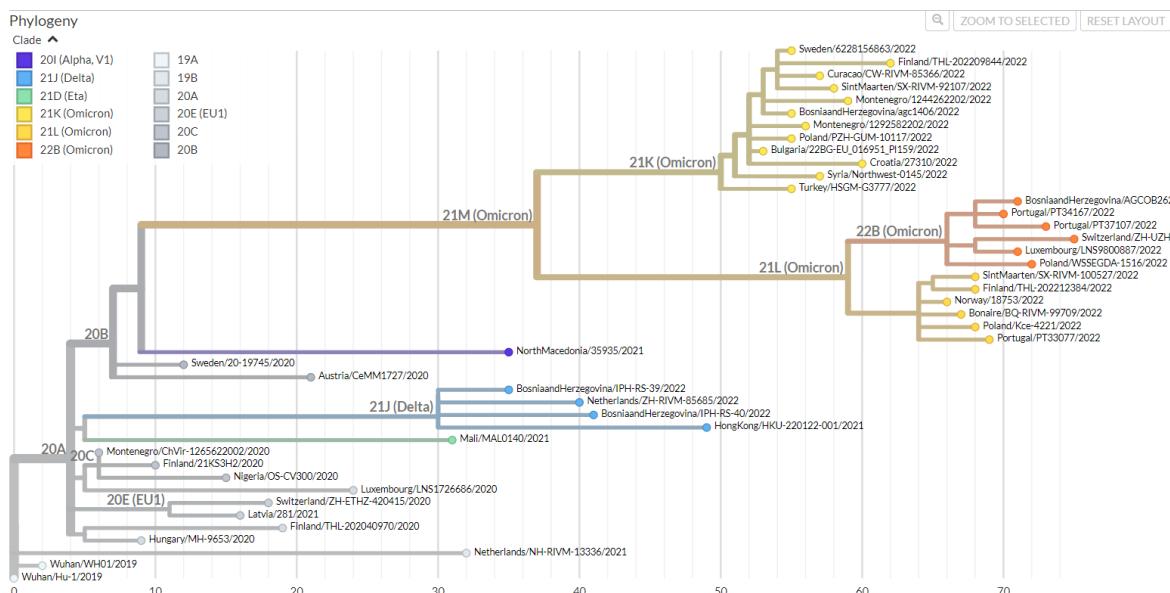


Figura 3.8: Árbol resultado del flujo de Nextstrain. *El software Nextstrain eliminó 6 secuencias entre las seleccionadas por cuestiones de gestión de calidad en el análisis.

En estos resultados, a primera vista, se tornan bastante similares y de calidad en comparación con Nextstrain. Se observa claramente una rama mayor que ubica en ambos árboles las secuencias de la variante Omicron (clados 21K, 21L y 22B), entre estos: el clado 21K es idéntico, el 22B es idéntico, las secuencias *Italy/LIG-297623416/2022* y *Norway/23637/2022* que componían el clado 22C en NCD fueron excluidas pero su ubicación original es coherente y en el clado 21L, sin tener en cuenta, las exclusiones también son idénticos.

En el resto del árbol, aparece una pequeña imprecisión en la secuencia *Mali/MAL0140/2021* en NCD tiene un parentesco muy cercano con la secuencia *NorthMacedonia/35935/2021*, sin embargo, en Nextstrain aparece unas cuantas generaciones antes.

3.3. Sobre el software Nextstrain y su comparación con NCD

Es interesante plantearse cómo han ido avanzando los proyectos por la vertiente del alineamiento y las ventajas/desventajas que tiene el método que se plantea frente a los proyectos que están a la orden del día. El software Nextstrain se compone por las siguientes herramientas:

- Nextstrain: es la de más alto nivel de todas. Actúa como padre de las siguientes herramientas, de forma que dada una entrada, genera un conjunto de tareas/comandos llamado DAG que coordina a las demás herramientas para realizar los consecutivos subtrabajos.

Un ejemplo de DAG(grafo acíclico dirigido) se encuentra disponible en el repositorio.

- Augur: esta herramienta está formada, a su vez, por un conjunto de herramientas o subcomandos entre los que destacan *augur align*, *augur tree* y *augur export* que conforman la parte del flujo de trabajo sobre análisis de múltiples patógenos bacteriales y virales.

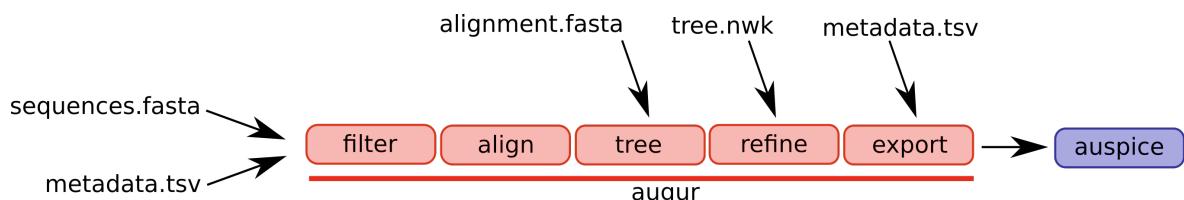


Figura 3.9: Flujo de alguno de los comandos que conforman augur.

Gracias a esta herramienta se pudo lograr la creación de un árbol filogenético de 500 secuencias de covid mediante alineamiento en solamente cuestión de minutos mientras que con NCD obtuvimos un tiempo mayor a 10 horas. Este espectacular tiempo no es debido a augur en sí sino que es causado por las herramientas que utiliza en sus comandos, mafft en augur align y raXmL en augur tree. Mafft es un programa destinado al alineamiento múltiple que inicialmente se basaba en la transformada de fourier pero que ya hoy en día ha implementado opciones para el alineamiento rápido de gran número de secuencias, alineamientos de gran precisión, y la adición de nuevas secuencias a alineamientos existentes. raxml ... Estos resultados que a la vista son mucho mejores que los que obtenemos con NCD se pueden mirar con [completar]

- Nextclade y Auspice: estas herramientas, que ya fueron comentadas en las sección 3.0.1, también cuentan con su versión software para la integración en el flujo nextstrain. Cabe destacar que con auspice en su versión software se puede lograr visualización más rica, con más información, mapa y animaciones que en la versión web.

Capítulo 4

Escalabilidad y estudio del sistema

Tradicionalmente la creación de árboles filogenéticos de gran tamaño (10.000, 30.000, 100.000 secuencias...) ha supuesto un reto, principalmente causado por uno de sus necesarios precursores, el alineamiento múltiple de secuencias.

Las dos cuestiones que preocupan al reconstruir árboles filogenéticos a partir de grandes matrices de datos son el tiempo de cálculo y la fiabilidad. La alineación múltiple de secuencias, el precursor necesario para la construcción del árbol, y la inferencia filogenética son problemas computacionales que consumen mucho tiempo. El tiempo de ejecución, t , de un algoritmo de construcción de árboles para una matriz de datos arbitraria depende principalmente del número de especies, N , porque $t = f(N)$. La relación funcional, f , oscila entre exponencial para la mayoría de los métodos de construcción de árboles que se basan en la optimización, como la parsimonia y la máxima verosimilitud hasta polinómica para soluciones aproximadas (“heurísticas”) a estos mismos o para ciertos métodos basados en la distancia, como el neighbor-joining.

Sobre el funcionamiento del sistema destacan 2 fases: el cálculo de la NCD de las distintas secuencias a la de referencia y el cálculo de la matriz de distancias entre las secuencias seleccionadas.

En este primer paso se calcula la distancia NCD de todas las respuestas frente a la secuencia de referencia. En todas las pruebas la secuencia de referencia que se ha usado es la de wuhan (Wuhan-Hu-1). El coste de este paso es $t(\text{paso1}) = n * t(\text{NCD})$ lo que en términos de O grande es $O(n)$ siendo n el número de secuencias a analizar.

Sobre el cálculo de la matriz de distancias con la que se obtendrá posteriormente la filogenia, constituye la parte más costosa del sistema. El tiempo de cálculo depende de 2 factores:

- El cálculo de la NCD entre 2 secuencias, que depende directamente de la longitud de las secuencias, en nuestro caso este cálculo con las secuencias de covid que tienen alrededor de 29500 caracteres tarda 0,3 segundos.

- El tamaño de la matriz de una forma exponencial, es decir, se realizarán $n \cdot n$ operaciones de cálculo de la NCD (en caso de que se aplique la optimización que se explica posteriormente se realizarían $\frac{n \cdot n}{2}$)

La fórmula con la que se realizan las estimaciones del tiempo es: $t(\text{matriz}) = t(\text{NCD}) \cdot n$ y en términos de O grande establece un coste $O(n^2)$.

4.1. Optimización del sistema

El sistema está basado en una base de datos clave-valor, donde la clave es una tag que identifica de manera única a las secuencias/pares de secuencias, y el valor que le corresponde es el tamaño comprimido. A nivel del cálculo del tamaño de compresión no se encontró ninguna optimización posible, sin embargo, a la hora de crear la filogenia se debe establecer la NCD de todas con todas las secuencias como se observa en (foto) formando una matriz simétrica. Por lo tanto, no se deberían calcular los $n \times n$ valores de la matriz. Se descubrió que no se tenía en cuenta la simetría de la matriz, sino que siendo 2 secuencias s_1 y s_2 se calculaba la distancia s_1-s_2 y s_2-s_1 siendo esta prácticamente igual. Este caso se producía porque se guardaba en la base de datos como una tag “ $\text{id1}+\text{id2}$ ” y a la hora de consultar si el tag se encontraba en la base no se tenía en cuenta el orden de las secuencias. Para comprobar que esta optimización realmente no alteraba los resultados originales se realizaron los árboles del artículo original obteniendo exactamente los mismos resultados. Con esta sencilla mejora se logra reducir a la **mitad** el número de NCDs a calcular y por consiguiente el tiempo de ejecución total.

4.1.1. Uso de las distancias de compresión en árboles filogenéticos

El método de clustering del artículo original era el comando MakeTree, con la matriz de distancias como entrada realizaba el método de reconstrucción de árboles de los cuartetos de mínimo coste. El programa comienza con un árbol aleatorio, y continua haciendo pequeñas modificaciones para mejorar la puntuación.

El problema con el método de clústering era que se llegaba a una salida en la que el árbol sólo representaba el orden parcial de las secuencias, las ramas no tenían longitud. Por tanto, en la parte relacional el clustering era correcto pero en la representación no se tenían en cuenta las distancias entre los diferentes clusters y a la hora de comparar con árboles actuales no se identificaban fácilmente.

Para solucionar este problema se pensó en usar el método Neighbor joining ya que se adapta perfectamente a nuestro caso de uso, requiere una matriz de distancias como

entrada y proporciona un árbol con las longitudes de las ramas.

Para adaptar la matriz de distancias del método NCD como entrada a MEGA se necesita un poco de postprocesado, esto es, transformar la matriz en triangular inferior y adaptarla al fichero de que requiere mega (formato .meg). Este postprocesado se realiza mediante un script propio.

A continuación se muestran algunos ejemplos de las diferencias en los mismos árboles agrupados con los dos métodos: y con NCD: y el árbol original en <https://nextstrain.org/monkeypox/hmpxv1>

4.1.2. Pruebas de carga

Se plantea la idea de comprobar el límite de este algoritmo y explorar la creación de árboles de gran tamaño. En los 2 artículos mencionados se hacía uso de datasets muy pequeños (menos de 100 secuencias), en este apartado trataremos de aumentar el tamaño de estos datasets, crear árboles mucho mayores y comprobar la viabilidad de este método para estos árboles grandes.

tamaño(n)	tiempoNCD(h,min)	tiempoAugur(h,min)
50	6 min	
100	32 min	
150	1 h	
250	2 h	
500	12 h 20 min	

Tabla 4.1: Tiempos del método NCD* y mediante Augur en función del tamaño del árbol. *El tiempo en NCD puede variar en base a la población de la BBDD.

Queda presente que la eficiencia en tiempo es el principal problema de este método. El coste exponencial del algoritmo y el tiempo de compresión constituyen el principal cuello de botella de este método. Se convierte en un método bastante prohibitivo para árboles de más de 500 secuencias. Por ejemplo, el tiempo estimado para un árbol de 1000 secuencias con el hardware comentado sería de alrededor de 42 horas de computo.

En la comparación con Nextstrain/Augur observamos 2 factores: el factor temporal y el factor cualitativo. Sobre lo temporal NCD no puede hacer frente al flujo de trabajo de Nextstrain, se obtienen tiempos de media x veces menores.

A nivel cualitativo, cabe destacar el uso del método de clustering Neighbor-joining. El método que llevaba a cabo el comando MakeTree(método de clustering) presentaba errores a medida que aumentaba el tamaño de árboles. El método se basaba en la generación de un árbol aleatorio y continuaba mejorandolo, para más de 200 secuencias se convertía en insostenible. El cambio en árboles de gran tamaño marcó la diferencia, en el caso de no aplicarlo, no se podría sacar ninguna conclusión o inducir errores en

ellas. En una primera instancia se cuestionó la calidad del algoritmo pensando que podía producir árboles degenerados y sin coherencia(Figura 4.3).

La conclusión es que para árboles grandes también se logran resultados bastante buenos, siguen la línea de las pruebas con árboles menores, un buena clusterización, similitud en cuanto a forma y pequeños fallos en secuencias cercanas a la de referencia(en el caso del árbol de 500 secuencias se observa el agrupamiento más difuso de los clados 20.A(azul claro), 20.B(azul oscuro) y 20.C(verde claro)). Estos errores se achacan a las bajas magnitudes de las distancias(alta similitud en las secuencias, entre dos de las secuencias más cercanas un ejemplo de distancia es 0,009575) y al margen de error del compresor. Ya en el artículo original mostraban que la distancia entre una secuencias y ella misma debería ser 0 pero era 0.003621.

También vale la pena mencionar las conclusiones que alcanzaba Vacca en su estudio: el 50 % de los árboles filogenéticos generados con la técnica de Vitányi cuentan con una similitud mayor al 80 % con los árboles obtenidos con los métodos de distancia, y que el 75 % cuentan con una similitud mayor a 75 %.

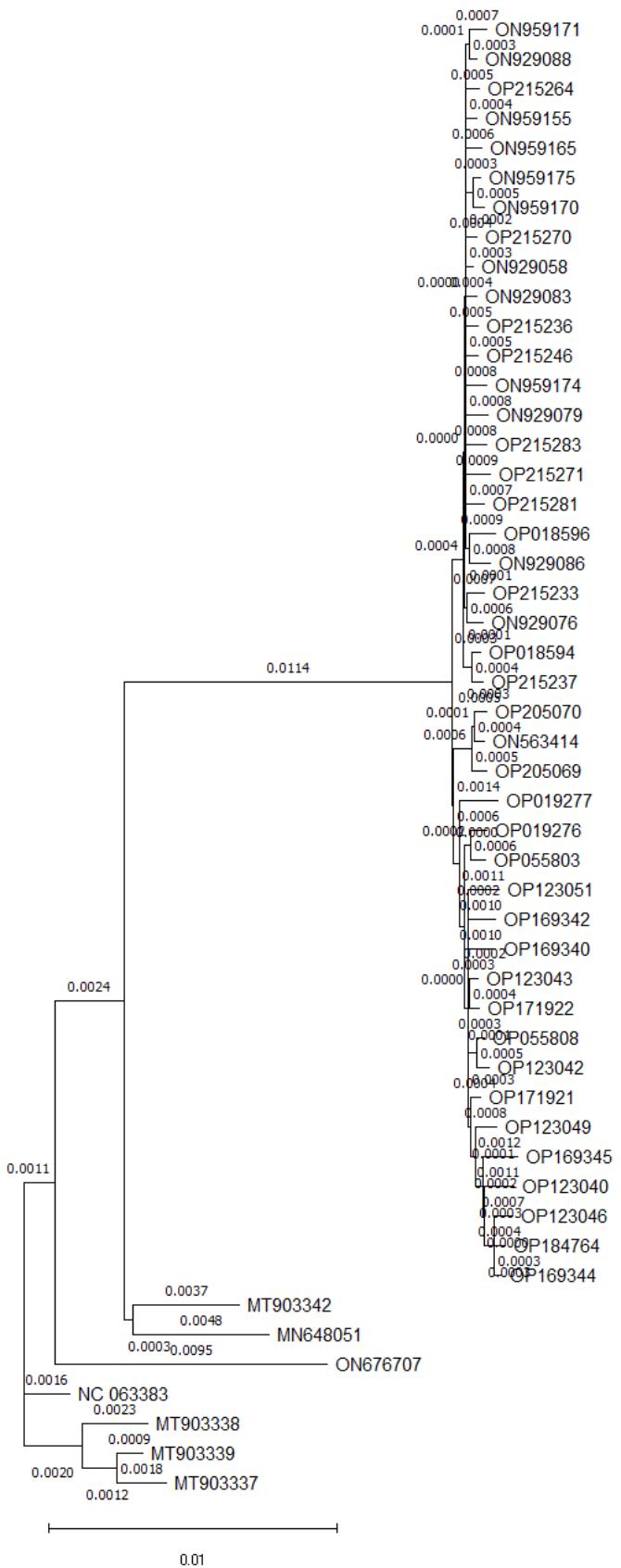


Figura 4.1: árbol con 50 secuencias del virus MPXV con branch lengths tras el postprocesamiento de los datos de la matriz de distancias NCD

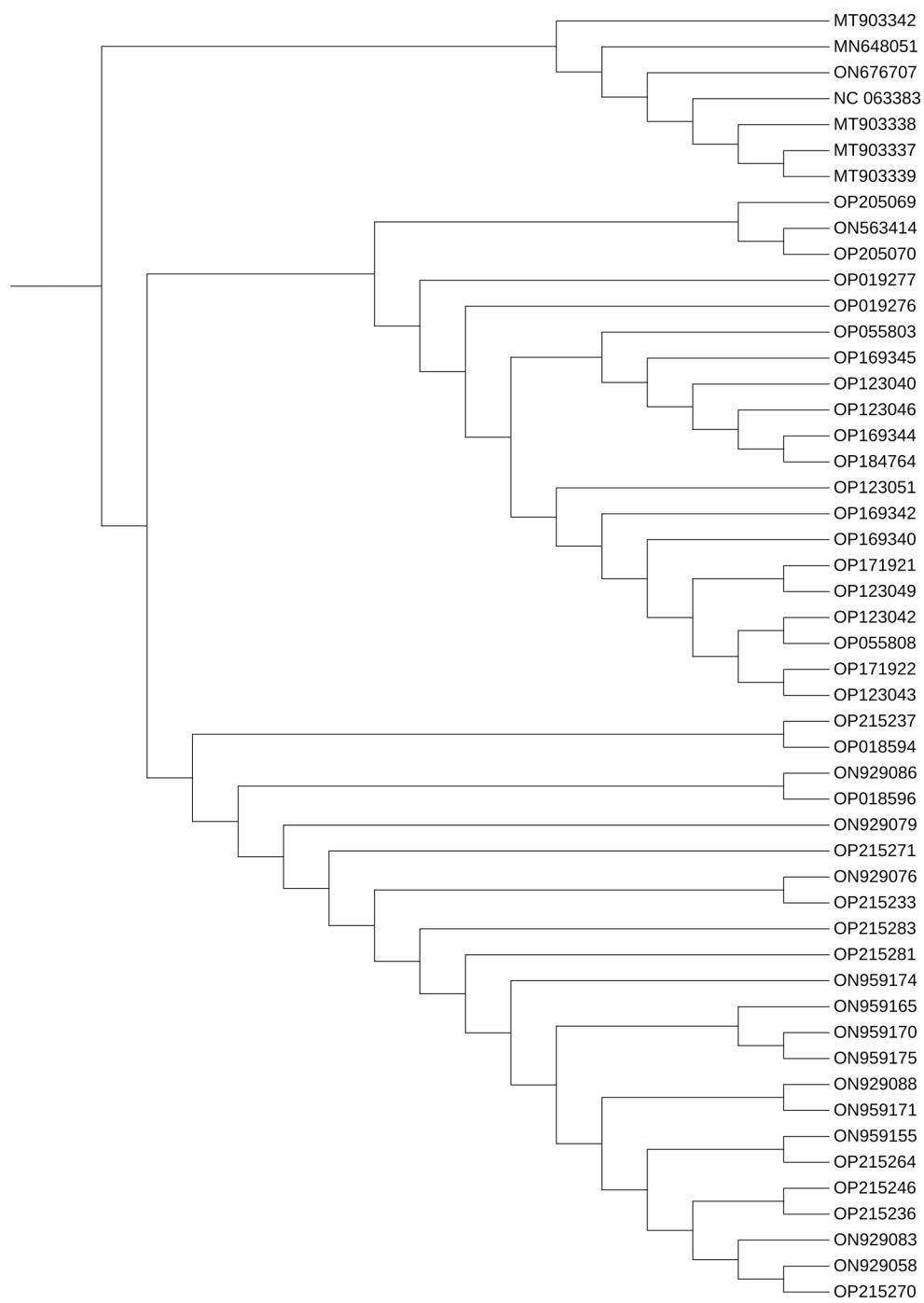
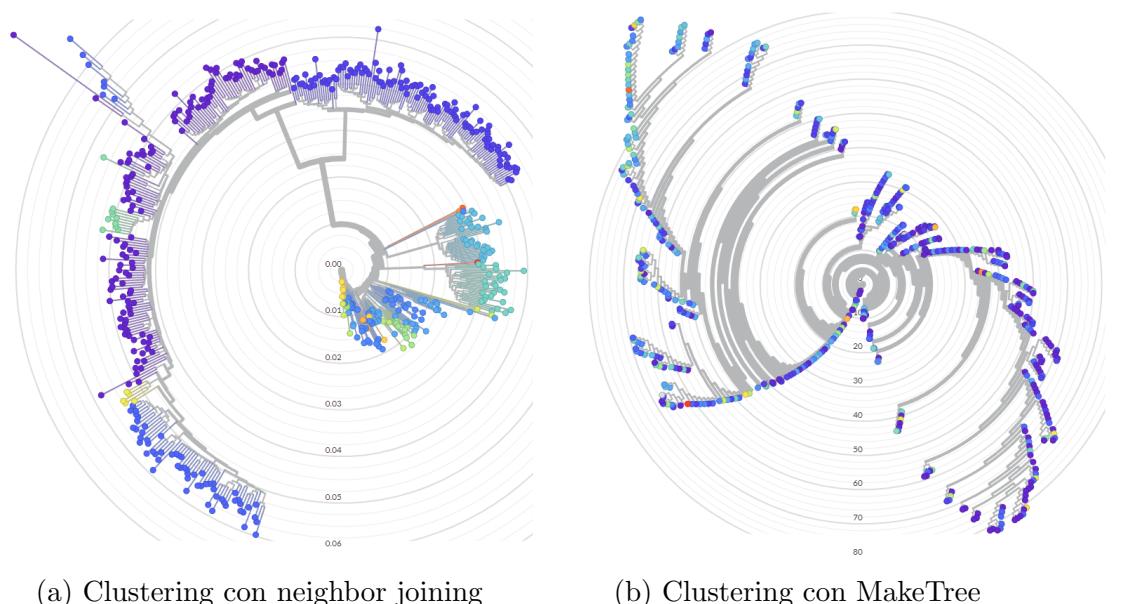


Figura 4.2: árbol con 50 secuencias del virus MPXV sin postprocesamiento salida de NCD



(a) Clustering con neighbor joining

(b) Clustering con MakeTree

Figura 4.3: Árboles formados con los dos métodos de clustering mostrados.

Capítulo 5

Clasificación de variantes de interés

XL

Capítulo 6

Compresión para otro tipo de secuencias

Estamos ante un método que se puede aplicar a multitud de ámbitos, el principal requerimiento es que la entrada sea una cadena de caracteres. En el artículo *clustering by compression* [9] de los mismo autores ponían a prueba el método con todo tipo de pruebas que no trataban secuencias de genomas, por ejemplo, clasificación de literatura, de música, reconocimiento de carácteres, árboles de lenguajes e incluso astronomía.

Aprovechando la generalidad del método NCD se plantea la búsqueda de dos datasets nuevos de secuencias que no sean de covid. En este caso los datasets que se van a investigar son secuencias de ADN mitocondrial y de la viruela de mono.

6.1. ADN mitocondrial

El ADN mitocondrial es un genoma que se encuentra en las mitocondrias, fuera del núcleo celular. Se compone de alrededor de 16.500 pares de bases y es de especial interés porque se hereda de la madre(el ADN nuclear se hereda de ambos progenitores), tiene una alta tasa de mutación y falta de recombinación.

Se obtuvieron 1524 secuencias de adn mitocondrial de GenBank con la búsqueda de ("Homo sapiens"[Organism] OR homo sapiens[All Fields]) AND (mitochondrion[filter] AND ("16400"[SLEN] : "16600"[SLEN]) AND ("2022/07/23"[PDAT] : "2022/08/31"[PDAT])), son secuencias de alrededor de 16.500 bp. Dadas las limitaciones del software NCD se realizó una prueba con un subconjunto de 197 secuencias. Ya que el software de NCD eliminina secuencias que no son completas se realizó el análisis sobre 120 secuencias. Los resultados son bastante buenos, como en este caso no se contaban con metadatos ni división en clados se realizó una comparación de búsqueda de subárboles comunes con la herramienta iphyloC [7].

Mediante esta herramienta se realizo una comparación lado a lado con el árbol con

las mismas secuencias creado por el software Nextstrain. Se observo que se formaban 2 clusters que a simple vista no se podían ver. El clúster formado por las secuencias MZ47529X.X junto a la secuencia CM045179.1, se puede observar en la figura 6.1. Por otro lado, el clúster formado por las 117 secuencias ON597XXX.X. Una de las principales diferencias que se aprecian en este clúster es la compactación, en el árbol de augur todas las secuencias están mucho más alineadas verticalmente. En NCD se encuentran mucho más escalonadas, probablemente debido al error que crea el compresor. Demás resultados y pruebas se encuentran en el repositorio(enlace al repositorio).

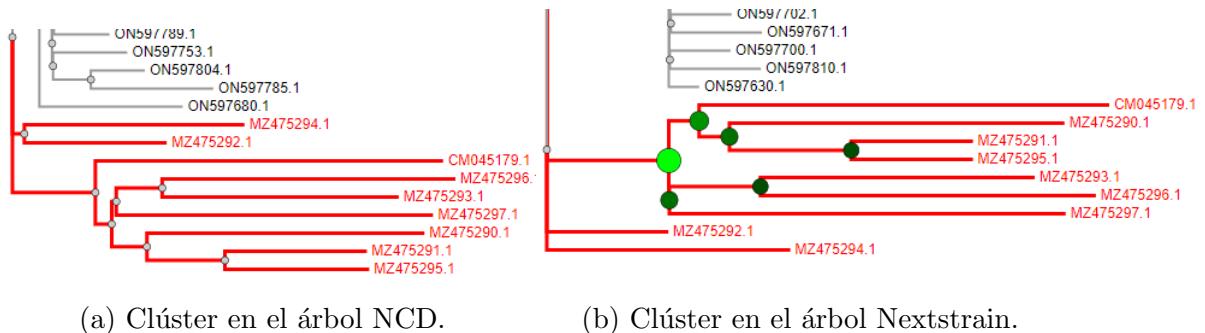


Figura 6.1: Clúster formado por las secuencias MZ47529X.X junto a la secuencia CM045179.1 en ambos métodos.

6.2. Monkeypox o viruela de mono

La viruela del mono(MPXV) es una zoonosis viral (un virus transmitido a los humanos por los animales) cuyo primer reporte de infección en humanos apareció en los 1970s [3], y que ya en 2022 aparecieron múltiples reportes en países no endémicos. A dia de 24/08/2022 ya hay registrados más de 44.000 casos en el mundo y más de 6000 en España.

Nexstrain separa el virus en 2 datasets, el mpxv y el hmpxv1, este último es el que se va a analizar. Comprende únicamente casos de transmisión humano-humano y corresponde un clado del mpxv como se puede ver en la imagen 6.4 de Happi et al [10]. Se recogieron 1262 secuencias de monkeypox de longitud 190.000-200.000 pares de bases proporcionadas por la misma página nextstrain. Se eligió como secuencia de referencia la NC-063383 y se ejecutó el algoritmo NCD para obtener un árbol de las 50 secuencias más próximas a la secuencia de referencia. En las figuras 6.2 y 6.3 se observa los resultados con NCD y Nextstrain respectivamente. Los resultados son prácticamente idénticos. Tomando como referencia la división en clados que ya hemos comentado, se pueden observar claramente la división en los 2 clústers, A y B. El

clúster A incluye muestras de los clados A(entre la que se encuentra la secuencia de referencia NC-063383), A.1 y A.2. El cluster que se forma a la derecha en la filogenia conforma el clado B, comprende muestras de los clados B.1, B.1.1, B.1.2, B.1.3, B.1.5 y B.1.7.

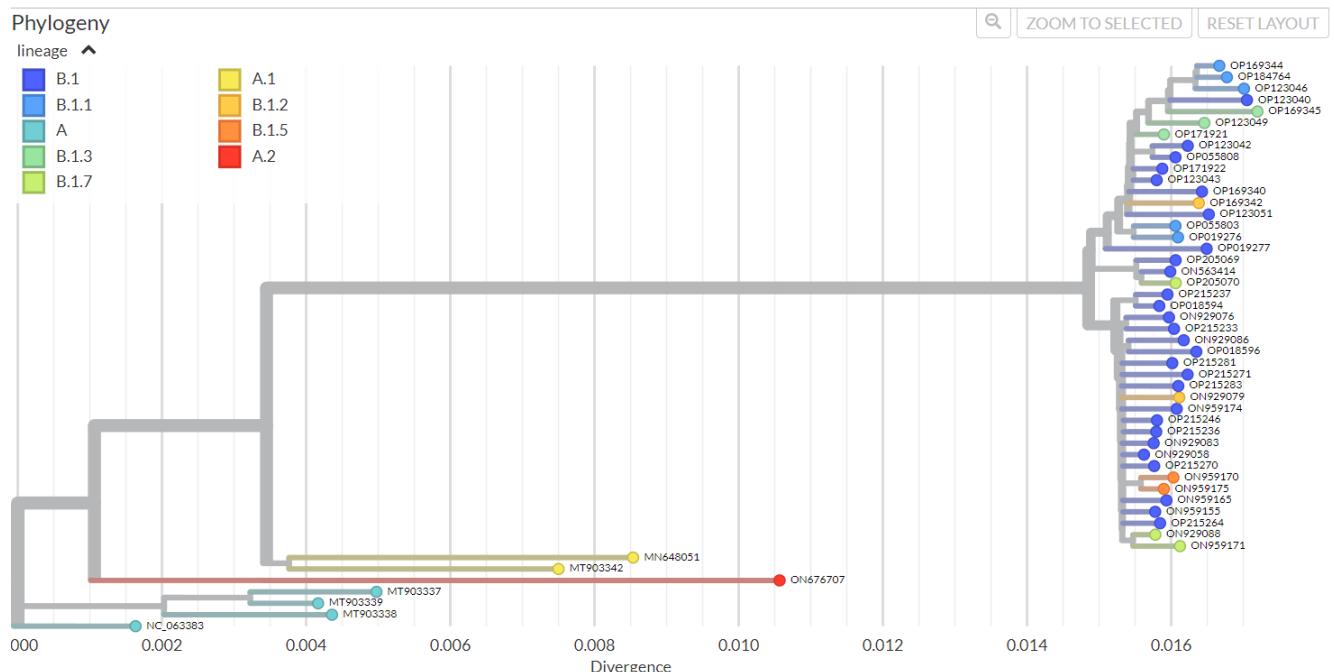


Figura 6.2: árbol salida del algoritmo NCD

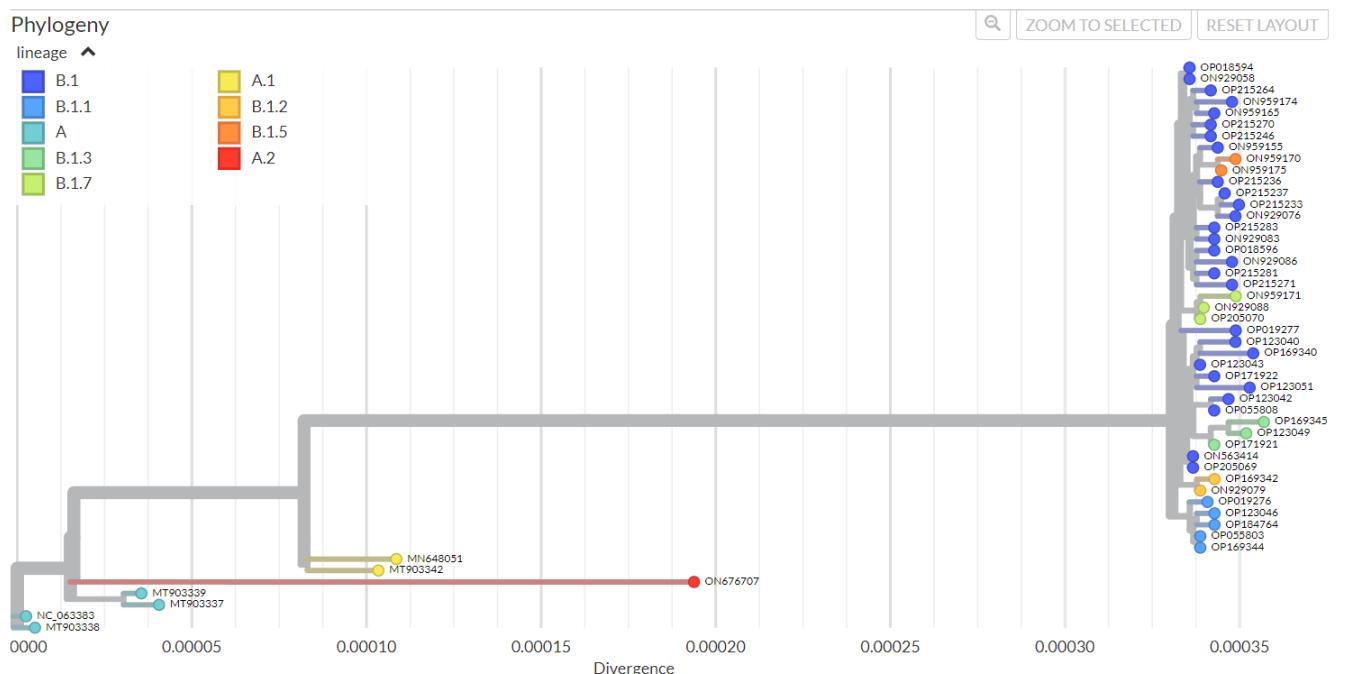


Figura 6.3: Árbol resultado del software Nextstrain

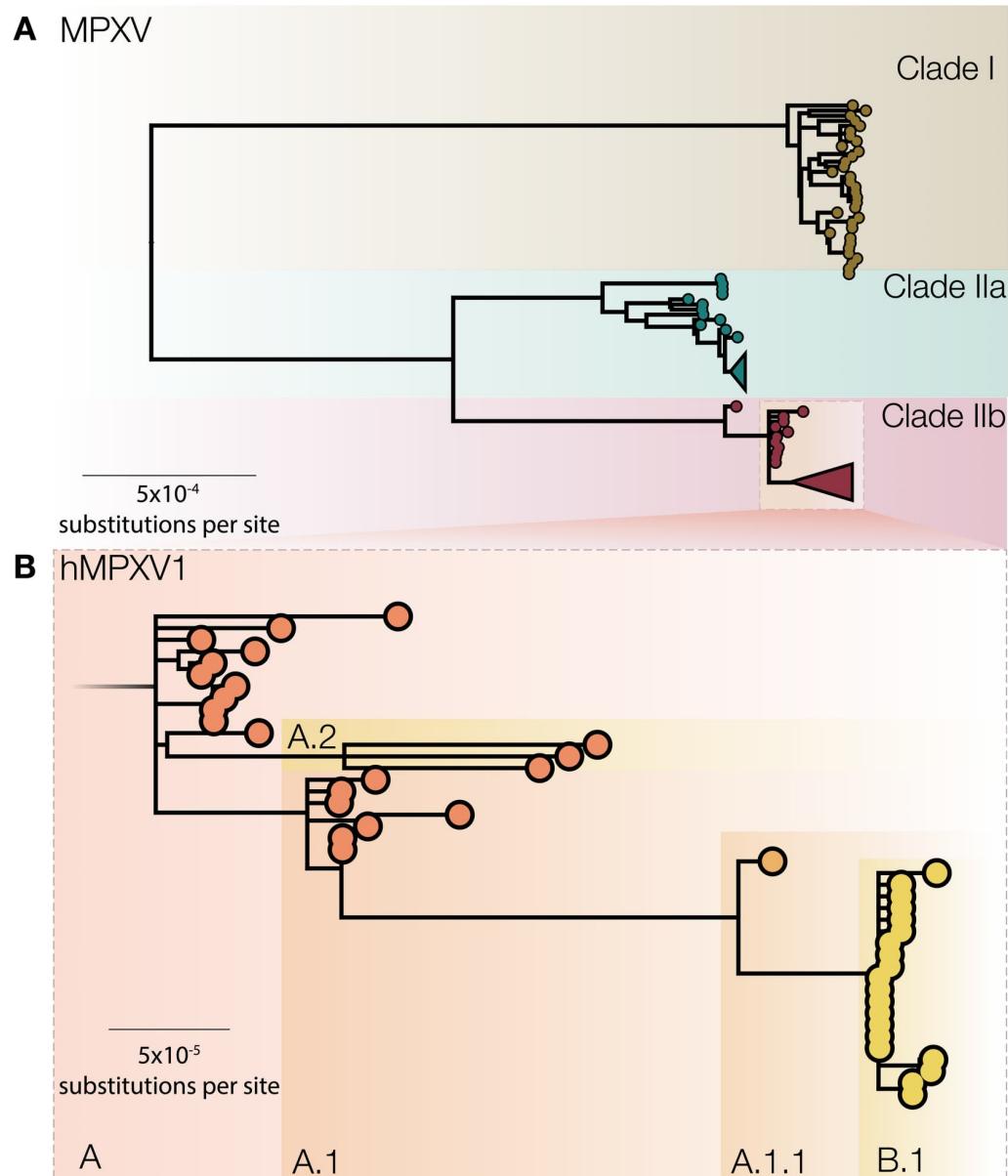


Figura 6.4: Esquema de división en clados del virus MPXV

Capítulo 7

Conclusiones

Algoritmo de carácter general que se uso para x,x,x,x y que probó su capacidad para crear filogenias suficientemente precisas como para sacar conclusiones con secuencias de covid pero que no puede hacer frente a nivel computacional, ya que la propia cualidad del método, la compresión, constituye su cuello de botella. mientras que los métodos de hoy en día, con el la reciente aparición del covid avanzaron muy rápidamente y mediante alineamiento múltiple logran unos bajos tiempos. También se logró ampliar el alcance del estudio original que probaba árboles con solamente 60 secuencias como máximo logrando árboles de hasta 500 secuencias sobre los que se realizó un análisis visual mediante la herramienta auspice.us.

Capítulo 8

Bibliografía

- [1] James Hadfield, Colin Megill, Sidney M Bell, John Huddleston, Barney Potter, Charlton Callender, Pavel Sagulenko, Trevor Bedford, and Richard A Neher. Nextstrain: real-time tracking of pathogen evolution. *Bioinformatics*, 34(23):4121–4123, 05 2018.
- [2] Rudi L. Cilibrasi and Paul M.B. Vitányi. Fast whole-genome phylogeny of the covid-19 virus sars-cov-2 by compression. *bioRxiv*, 2020.
- [3] Stefan Elbe and Gemma Buckland-Merrett. Data, disease and diplomacy: Gisaid’s innovative contribution to global health. *Global Challenges*, 1(1):33–46, 2017.
- [4] David Vacca. Estudio del uso de la ncd para la inferencia de árboles filogenéticos. Trabajo fin de grado, Facultad de Ciencias Exactas y Naturales Universidad de Buenos Aires.
- [5] Ivan Aksamentov, Cornelius Roemer, Emma B. Hodcroft, and Richard A. Neher. Nextclade: clade assignment, mutation calling and quality control for viral genomes. *Journal of Open Source Software*, 6(67):3773, 2021.
- [6] Oscar Robinson, David Dylus, and Christophe Dessimoz. Phylo.io : Interactive Viewing and Comparison of Large Phylogenetic Trees on the Web . *Molecular Biology and Evolution*, 33(8):2163–2166, 04 2016.
- [7] Muhsen Hammoud, Charles Morphy D. Santos, and João Paulo Gois. Visual comparison of phylogenetic trees through iphyloc, a new interactive web-based framework. *bioRxiv*, 2021.
- [8] Koichiro Tamura, Glen Stecher, and Sudhir Kumar. MEGA11: Molecular Evolutionary Genetics Analysis Version 11. *Molecular Biology and Evolution*, 38(7):3022–3027, 04 2021.

- [9] R. Cilibrasi and P.M.B. Vitanyi. Clustering by compression. *IEEE Transactions on Information Theory*, 51(4):1523–1545, 2005.
- [10] Christian Happi, Ifedayo Adetifa, Placide Mbala, Richard Njouom, Emmanuel Nakoune, Anise Happi, Nnaemeka Ndodo, Oyeronke Ayansola, Gerald Mboowa, Trevor Bedford, Richard A. Neher, Cornelius Roemer, Emma Hodcroft, Houriiyah Tegally, Áine O’Toole, Andrew Rambaut, Oliver Pybus, Moritz U. G. Kraemer, Eduan Wilkinson, Joana Isidro, Vítor Borges, Miguel Pinto, João Paulo Gomes, Lucas Freitas, Paola C. Resende, Raphael T. C. Lee, Sebastian Maurer-Stroh, Cheryl Baxter, Richard Lessells, Ahmed E. Ogwell, Yenew Kebede, Sofonias K. Tessema, and Tulio de Oliveira. Urgent need for a non-discriminatory and non-stigmatizing nomenclature for monkeypox virus. *PLOS Biology*, 20(8):1–6, 08 2022.
- [11] H. Farnsworth. What-if machine analysis and design. *IEEE Transactions on quantum neuroscience electronics*, 3031.
- [12] N. Sonntag. *Mis mejores recetas con repollo*. Anaconda, 2016.
- [13] S. Z. Ramírez, K. Pérez. Self conscious robots in induction heating home appliances. *IEEE transactions on anthropomorphic robots*, 2018.
- [14] Alumno Apellidos. Citar un tfm. Trabajo fin de máster, Universidad de Zaragoza, 2014.

Lista de Figuras

2.1.	Esquema del flujo de trabajo de NCD	XIV
3.1.	Captura del análisis realizado por nextclade.org	XVIII
3.2.	Esquema de la nomenclatura de clados en Nextstrain	XX
3.3.	Árbol salida del programa(a) y árbol con una subselección(b) de alrededor de 500 secuencias del conjunto inicial, en rojo las secuencias correspondientes al árbol a en ambos árboles	XXII
3.4.	Redactar	XXIII
3.5.	Mediante la opción unrooted tree de auspice.us se puede observar la calidad de los clusters en los árboles. Ambos casos se diferencian los clados 19.x, en la rama superior izquierda de estos. Sobre los clados 20.x, en NCD se observa una diferenciaión más difusa en los clados 20.B(verde) y 20.A(azul). En el árbol de MEGA se distinguen perfectamente los clados 20.x(cyan,rojo,azul y amarillo).	XXIV
3.6.	Árbol con la nomenclatura de clados de GISAID, correspondencia con nextstrain y variante que marca el clado.	XXV
3.7.	Árbol resultado del método NCD	XXVI
3.8.	Árbol resultado del flujo de Nextstrain. <i>*El software Nextstrain eliminó 6 secuencias entre las seleccionadas por cuestiones de gestión de calidad en el análisis.</i>	XXVI
3.9.	Flujo de alguno de los comandos que conforman augur.	XXVIII
4.1.	árbol con 50 secuencias del virus MPXV con branch lengths tras el postprocesamiento de los datos de la matriz de distancias NCD	XXXV
4.2.	árbol con 50 secuencias del virus MPXV sin postprocesamiento salida de NCD	XXXVI
4.3.	Árboles formados con los dos métodos de clustering mostrados.	XXXVII
6.1.	Clúster formado por las secuencias MZ47529X.X junto a la secuencia CM045179.1 en ambos métodos.	XLII

6.2. árbol salida del algoritmo NCD	XLIII
6.3. Árbol resultado del software Nextstrain	XLIII
6.4. Esquema de división en clados del virus MPXV	XLIV

Lista de Tablas

4.1. Tiempos del método NCD* y mediante Augur en función del tamaño del árbol.

**El tiempo en NCD puede variar en base a la población de la BBDD.* . XXXIII