



Universidad
Zaragoza

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./D^a. _____,

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

Agradezco a Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc,

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	1
2. Despliegue y replicación del sistema	3
2.1. Una sección	3
3. Comparación de los resultados obtenidos con el estado del arte	7
3.0.1. Criterios de comparación	7
3.1. Datos obtenidos de nextstrain	9
3.2. Datos obtenidos de GISAID	9
3.3. Escalabilidad y estudio del sistema	9
3.4. Optimización del sistema	12
3.4.1. Uso de las distancias de compresión en árboles filogenéticos . . .	12
3.5. Sobre Augur y su comparación con NCD	16
3.6. Clasificación de variantes de interés	16
4. Compresión para otro tipo de secuencias	17
4.0.1. ADN mitocondrial	17
4.0.2. Monkeypox o viruela de mono	17
5. Conclusiones	21
6. Bibliografía	23
Lista de Figuras	25
Lista de Tablas	27
Anexos	28
A. Un anexo	31

Capítulo 1

Introducción y objetivos

Y, viéndole don Quijote de aquella manera, con muestras de tanta tristeza, le dijo: Sábetete, Sancho, que no es un hombre más que otro si no hace más que otro. Todas estas borrascas que nos suceden son señales de que presto ha de serenar el tiempo y han de sucedernos bien las cosas; porque no es posible que el mal ni el bien sean durables, y de aquí se sigue que, habiendo durado mucho el mal, el bien está ya cerca. Así que, no debes congojarte por las desgracias que a mí me suceden, pues a ti no te cabe parte dellas. Y, viéndole don Quijote de aquella manera, con muestras de tanta tristeza, le dijo: Sábetete, Sancho, que no es un hombre más que otro si no hace más que otro. Todas estas borrascas que nos suceden son señales de que presto ha de serenar el tiempo y han de sucedernos bien las cosas; porque no es posible que el mal ni el bien sean durables, y de aquí se sigue que, habiendo durado mucho el mal, el bien está ya cerca. Así que, no debes congojarte por las desgracias que a mí me suceden, pues a ti no

Capítulo 2

Despliegue y replicación del sistema

Podría dedicar este capítulo a explicar como repliqué el sistema y obtuve los mismos resultados que en el artículo original.

2.1. Una sección

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Aenean commodo ligula eget dolor. Aenean massa. Cum sociis natoque penatibus et magnis dis parturient montes, nascetur ridiculus mus. Donec quam felis, ultricies nec, pellentesque eu, pretium quis, sem. Nulla consequat massa quis enim. Donec pede justo, fringilla vel, aliquet nec, vulputate eget, arcu. In enim justo, rhoncus ut, imperdiet a, venenatis vitae, justo. Nullam dictum felis eu pede mollis pretium. Integer tincidunt. Cras dapibus. Vivamus elementum semper nisi. Aenean vulputate eleifend tellus. Aenean leo ligula, porttitor eu, consequat vitae, eleifend ac, enim. Aliquam lorem ante, dapibus in, viverra quis, feugiat a, tellus. Phasellus viverra nulla ut metus varius laoreet. Quisque rutrum. Aenean imperdiet. Etiam ultricies nisi vel augue. Curabitur ullamcorper ultricies nisi. Nam eget dui. Etiam rhoncus. Maecenas tempus, tellus eget condimentum rhoncus, sem quam semper libero, sit amet adipiscing sem neque sed ipsum. Nam quam nunc, blandit vel, luctus pulvinar, hendrerit id, lorem. Maecenas nec odio et ante tincidunt tempus. Donec vitae sapien ut libero venenatis faucibus. Nullam quis ante. Etiam sit amet orci eget eros faucibus tincidunt. Duis leo. Sed fringilla mauris sit amet nibh. Donec sodales sagittis magna. Sed consequat, leo eget bibendum sodales, augue velit cursus nunc,

NCD	virus name
0.00362117	selected_SARS.CoV.2.EPI.ISL.471246
0.018043	MN908947.3.alt..SARS.CoV.2
0.448319	BetaCoV/bat/Yunnan/RaTG13/2013—EPI.ISL.402131.EPI.ISL.402131
0.788278	MG772933.1.bat.SL.CoVZC45
0.79136	MG772934.1.bat.SL.CoVZXC21
0.917906	KF569996_Coronaviridae.785
0.918252	KC881006_Coronaviridae.783
0.918699	KC881005_Coronaviridae.782
0.919072	AY278554_Riboviria.2953
0.919083	AY278741_Riboviria.2954
0.919087	FJ882963_Coronaviridae.726
0.91914	EU371561_Riboviria.3205
0.919188	AY278488_Riboviria.2951
0.919244	AY278491_Riboviria.2952
0.919297	FJ882935_Coronaviridae.722
0.919359	EU371559_Riboviria.3203
0.919393	NC_004718_Coronaviridae.806
0.919429	EU371563_Riboviria.3207
0.919486	AY357075_Riboviria.2979
0.919486	AY350750_Riboviria.2977
0.91952	AY864805_Riboviria.3030
0.91952	DQ640652_Riboviria.3098
0.919556	EU371562_Riboviria.3206
0.919635	FJ882945_Coronaviridae.724
0.919762	EU371560_Riboviria.3204
0.919774	AY394850_Riboviria.2981
0.919807	AY864806_Riboviria.3031
0.919878	EU371564_Riboviria.3208
0.919934	FJ882942_Coronaviridae.723
0.920233	FJ882954_Coronaviridae.725
0.920315	AY357076_Riboviria.2980
0.920607	KF367457_Coronaviridae.784
0.921176	AY515512_Riboviria.2987
0.921888	JX993988_Coronaviridae.779
0.923151	GQ153542_Coronaviridae.750
0.923183	GQ153543_Coronaviridae.751
0.925293	GQ153547_Coronaviridae.755
0.925569	GQ153544_Coronaviridae.752
0.925569	GQ153545_Coronaviridae.753
0.925676	GQ153548_Coronaviridae.756
0.925686	DQ648857_Riboviria.3101
0.925737	GQ153539_Coronaviridae.747
0.925737	GQ153540_Coronaviridae.748
0.925854	GQ153546_Coronaviridae.754
0.925875	GQ153541_Coronaviridae.749
0.926775	JX993987_Coronaviridae.778
0.926957	DQ412043_Riboviria.3074
0.931991	DQ412042_Riboviria.3073
0.932947	DQ648856_Riboviria.3100
0.952368	NC_014470_Coronaviridae.823
0.994546	NC_025217_Coronaviridae.835
0.994986	JF705860_Coronaviridae.768
0.994986	AY646283_Riboviria.3003
0.995034	NC_034440_Coronaviridae.847
0.995078	EF065512_Riboviria.3126
0.995078	EF065511_Riboviria.3125
0.995078	EF065510_Riboviria.3124
0.995086	EF065505_Riboviria.3119
0.995086	EF065506_Riboviria.3120
0.995086	EF065507_Riboviria.3121

NCD	virus name
0.00388619	MN908947.3.alt..SARS.CoV.2
0.442193	BetaCoV/bat/Yunnan/RaTG13/2013—EPI_ISL_402131.EPI_ISL_402131
0.786061	MG772933.1_bat_SL.CoVZC45
0.789591	MG772934.1_bat_SL.CoVZXC21
0.915357	KC881006_Coronaviridae.783
0.915702	KF569996_Coronaviridae.785
0.915805	KC881005_Coronaviridae.782
0.91631	AY278554_Riboviria.2953
0.916483	AY278491_Riboviria.2952
0.916563	AY278488_Riboviria.2951
0.916621	DQ640652_Riboviria.3098
0.916632	NC.004718_Coronaviridae.806
0.916644	EU371561_Riboviria.3205
0.916724	AY350750_Riboviria.2977
0.916724	AY357075_Riboviria.2979
0.916736	AY278741_Riboviria.2954
0.916759	AY864805_Riboviria.3030
0.916794	EU371563_Riboviria.3207
0.916874	EU371559_Riboviria.3203
0.916921	EU371562_Riboviria.3206
0.917139	EU371560_Riboviria.3204
0.91715	FJ882945_Coronaviridae.724
0.917184	AY864806_Riboviria.3031
0.917288	AY394850_Riboviria.2981
0.917415	AY357076_Riboviria.2980
0.917418	EU371564_Riboviria.3208
0.917449	FJ882942_Coronaviridae.723
0.917704	FJ882963_Coronaviridae.726
0.918051	FJ882935_Coronaviridae.722
0.918304	KF367457_Coronaviridae.784
0.918829	AY515512_Riboviria.2987
0.919125	FJ882954_Coronaviridae.725
0.921528	GQ153543_Coronaviridae.751
0.921636	GQ153542_Coronaviridae.750
0.921721	JX993988_Coronaviridae.779
0.923894	DQ648857_Riboviria.3101
0.924052	GQ153547_Coronaviridae.755
0.924297	GQ153548_Coronaviridae.756
0.924328	GQ153545_Coronaviridae.753
0.924359	GQ153540_Coronaviridae.748
0.924466	GQ153544_Coronaviridae.752
0.924497	GQ153539_Coronaviridae.747
0.924614	GQ153546_Coronaviridae.754
0.924635	GQ153541_Coronaviridae.749
0.925028	DQ412043_Riboviria.3074
0.925885	JX993987_Coronaviridae.778
0.930059	DQ412042_Riboviria.3073
0.93074	DQ648856_Riboviria.3100
0.951006	NC.014470_Coronaviridae.823
0.994806	NC.025217_Coronaviridae.835
0.995086	EF065505_Riboviria.3119
0.995086	EF065506_Riboviria.3120
0.995086	EF065507_Riboviria.3121
0.995092	EF065508_Riboviria.3122
0.995172	NC.034440_Coronaviridae.847
0.995211	EF065512_Riboviria.3126
0.995211	EF065511_Riboviria.3125
0.995211	EF065510_Riboviria.3124
0.995227	DQ648794_Riboviria.3099
0.995259	NC.038294_Coronaviridae.850

NCD	virus name
0.00362117	selected_SARS.CoV_2.EPI_ISL_471246
0.0111034	MN908947.3.alt..SARS.CoV_2
0.444846	BetaCoV/bat/Yunnan/RaTG13/2013—EPI_ISL_402131.EPI_ISL_402131
0.788416	MG772933.1.bat_SL.CoVZC45
0.791082	MG772934.1.bat_SL.CoVZXC21
0.917493	KF569996_Coronaviridae.785
0.917563	KC881006_Coronaviridae.783
0.91801	KC881005_Coronaviridae.782
0.918257	FJ882963_Coronaviridae.726
0.918381	AY278554_Riboviria.2953
0.918447	EU371561_Riboviria.3205
0.918497	AY278488_Riboviria.2951
0.918531	AY278741_Riboviria.2954
0.918553	AY278491_Riboviria.2952
0.918565	NC.004718_Coronaviridae.806
0.918597	EU371563_Riboviria.3207
0.918605	FJ882935_Coronaviridae.722
0.918658	AY357075_Riboviria.2979
0.918669	EU371559_Riboviria.3203
0.918691	DQ640652_Riboviria.3098
0.918724	EU371562_Riboviria.3206
0.918796	AY350750_Riboviria.2977
0.918829	AY864805_Riboviria.3030
0.918945	FJ882945_Coronaviridae.724
0.919072	EU371560_Riboviria.3204
0.919117	AY864806_Riboviria.3031
0.919182	EU371564_Riboviria.3208
0.919221	AY394850_Riboviria.2981
0.919244	FJ882942_Coronaviridae.723
0.919486	AY357076_Riboviria.2980
0.91954	FJ882954_Coronaviridae.725
0.91993	KF367457_Coronaviridae.784
0.920486	AY515512_Riboviria.2987
0.921053	JX993988_Coronaviridae.779
0.923045	GQ153543_Coronaviridae.751
0.923151	GQ153542_Coronaviridae.750
0.92541	DQ648857_Riboviria.3101
0.925706	GQ153547_Coronaviridae.755
0.925802	JX993987_Coronaviridae.778
0.925844	GQ153544_Coronaviridae.752
0.925951	GQ153548_Coronaviridae.756
0.925982	GQ153545_Coronaviridae.753
0.926013	GQ153540_Coronaviridae.748
0.92613	GQ153546_Coronaviridae.754
0.92615	GQ153541_Coronaviridae.749
0.92615	GQ153539_Coronaviridae.747
0.926681	DQ412043_Riboviria.3074
0.931577	DQ412042_Riboviria.3073
0.932533	DQ648856_Riboviria.3100
0.952228	NC.014470_Coronaviridae.823
0.994546	NC.025217_Coronaviridae.835
0.994897	NC.034440_Coronaviridae.847
0.994986	FJ938057_Coronaviridae.734
0.994986	AY646283_Riboviria.3003
0.995078	EF065510_Riboviria.3124
0.995078	EF065511_Riboviria.3125
0.995078	EF065512_Riboviria.3126
0.995086	EF065506_Riboviria.3120
0.995086	EF065505_Riboviria.3119
0.995086	EF065507_Riboviria.3121

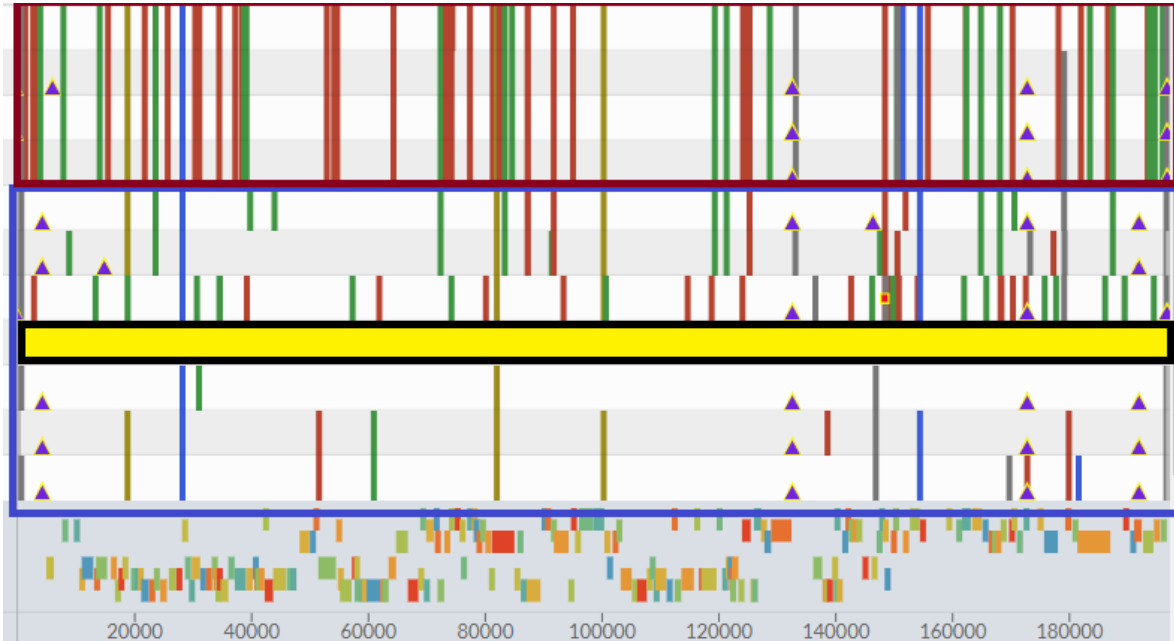
Capítulo 3

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 los 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. Por ello y 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 que a grosso modo realiza Clade assignment, mutation calling, and sequence quality checks. Esta cuenta con diferentes datasets de algunos virus de la actualidad y al seleccionar un conjunto de secuencias como entrada las etiqueta y clasifica. Finalmente muestra en una tabla difentes 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 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) ¿Igual debería hacer el ejemplo sobre el covid?
- 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 visualtreecmp, dado que estas presentan ciertas desventajas e irregularidades



(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

Nucleotide Sequence mode

In "Nucleotide Sequence" mode, the whole nucleotide sequence is shown. Line markers represent nucleotide mutations. They are colored by the resulting (query) nucleotide:



(b) Árbol subselección

Figura 3.1: 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

3.1. Datos obtenidos de nextstrain

Para comparar el resultado obtenido de los 100 virus sars cov 2 más cercanos al de referencia Wuhan 2019 con los sistemas actuales hicimos 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. En segundo lugar se comparó a nivel filogenético con el mismo conjunto de 100 secuencias seleccionadas pero se obtuvo la filogenia con uno de los métodos actuales, 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.

3.1

Alumnos	Media
0,105658	<i>EF065508</i> _{Riboviria} 3122

Tabla 3.1: Pie de tabla

3.2. Datos obtenidos de GISAID

En este caso obtenemos los datos de nexstrain.org pero estos pertenecen a Gisaid. Son datos pertenecientes a europa. Para esta prueba se eligieron 50 secuencias linealmente separadas, es decir, no se consideraran las n secuencias más próximas a la de referencia sino que se eligieran uniformemente entre las secuencias, con el objetivo de poner en contexto la magnitud de ncd que se están comparando y que puedan aparecer otras secuencias de variantes más recientes en comparación con la de referencia.

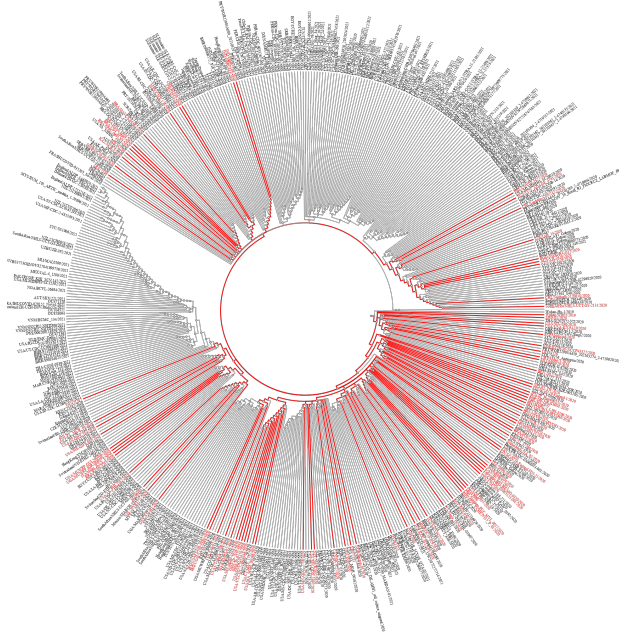
A continuación se comparará

3.3. 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. The two issues of concern in reconstructing phylogenetic trees from large data matrices are computation time and reliability. Multiple sequence alignment, the necessary precursor to tree building, and phylogenetic inference are notoriously time-consuming computational problems. The running time, t , of a tree-building algorithm for some arbitrary data matrix is dependent mainly on the number of species, N , because $t \propto N^4$: The functional



(a) Árbol salida



(b) Árbol subselección

Figura 3.2: 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

relationship, f , ranges from exponential for most treebuilding methods that rely on optimization, such as parsimony and maximum likelihood, to polynomial

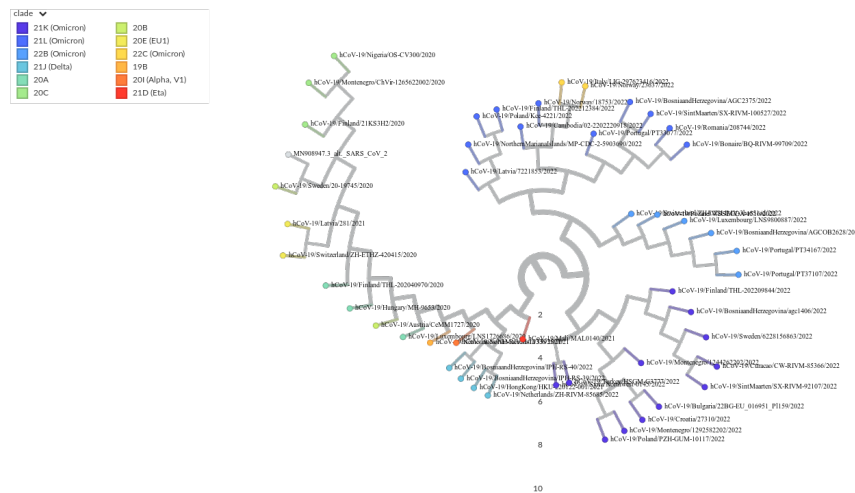


Figura 3.3: Árbol salida

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. El coste de este paso es $t(\text{paso1}) = n_{\text{seq}} * t(\text{NCD})$ 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 mi 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 $n \times n$ operaciones de cálculo de la NCD(en caso de que se aplique la optimización que se explica posteriormente se realizarían $n \times n / 2$)

la fórmula con la que se realizan las estimaciones del tiempo es: $t(\text{matriz}) = t(\text{NCD}) * n \times n$

$n * n$ y en términos de n grande establece un coste $O(n^2)$

3.4. 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/par de secuencias, y el valor que le corresponde es el tamaño comprimido. A nivel del calculo 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 $s1$ y $s2$ se calculaba la distancia $s1-s2$ y $s2-s1$ siendo esta prácticamente igual. Este caso se producía porque se guardaba en la base de datos como una tag $id1+id2z$ 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.

3.4.1. Uso de las distancias de compresión en árboles filogenéticos

El método de clustering del artículo original era el comando `maketree(ref)` que con la matriz de distancias como entrada realiza el Minimum Cost Quartet Tree Reconstruction method. The program starts with a random tree, and continues to try small modifications to improve the tree score. Eventually, the tree will stop easily improving.

El problema con el método de clústering era que se llegaba a una salida en la que el árbol solo representaba el orden parcial de las secuencias ya que las ramas no tienen 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 facilmente.

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>

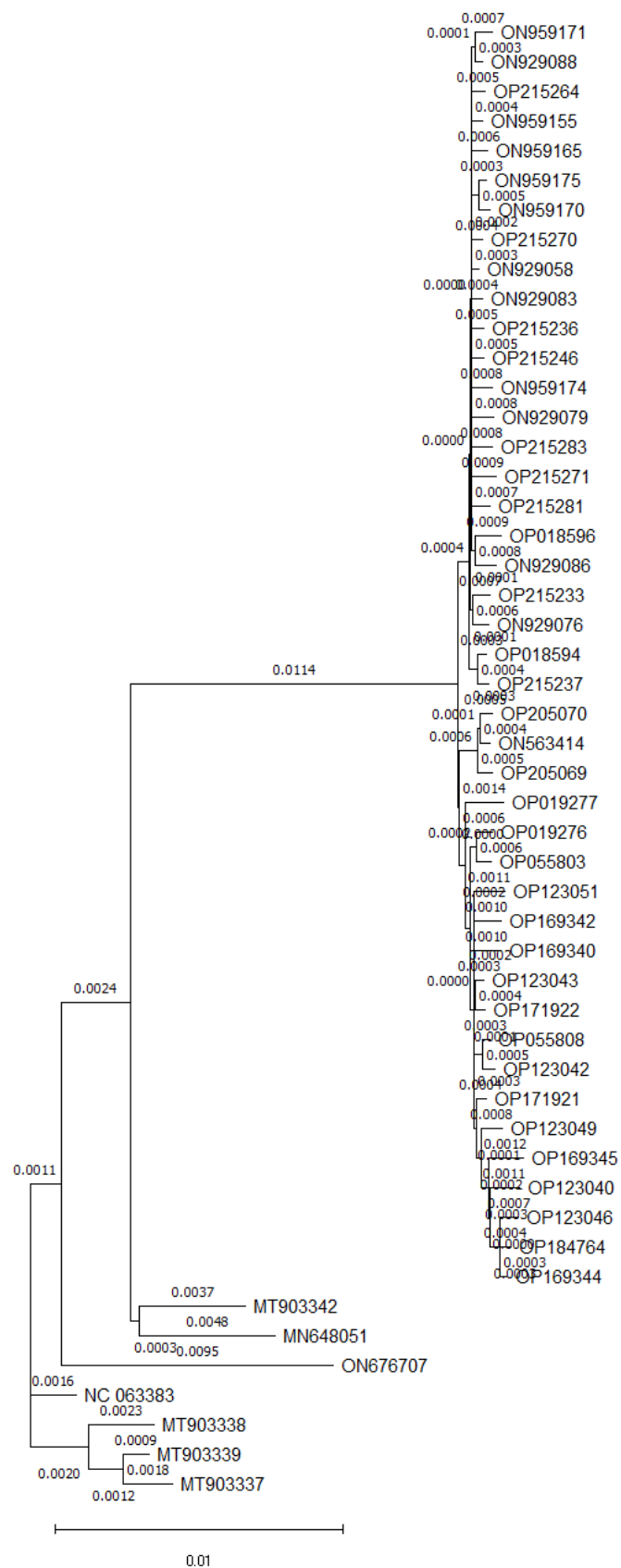


Figura 3.4: árbol con 50 secuencias del virus MPXV con branch lengths tras el postprocesamiento de los datos de la matriz de distancias NCD

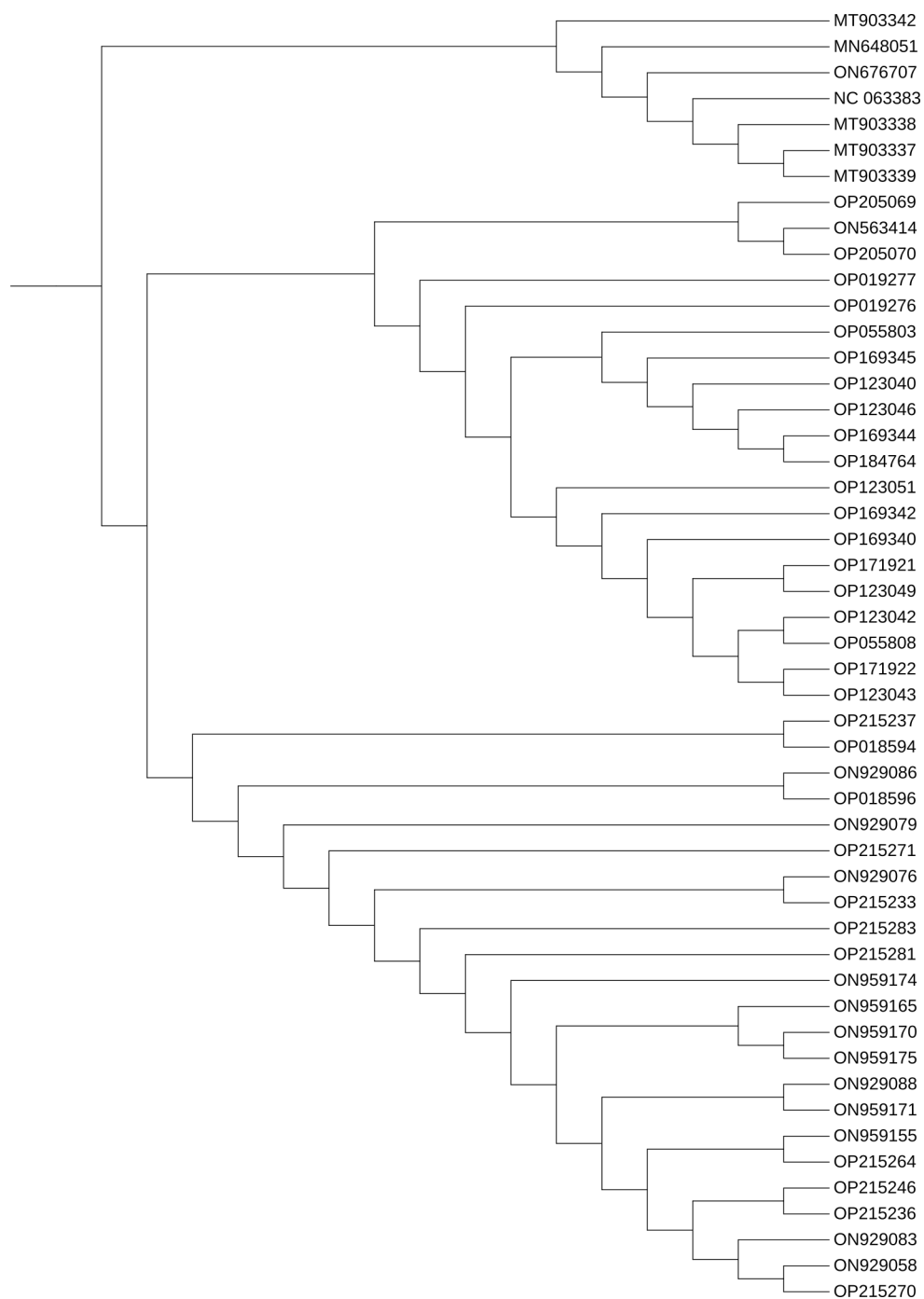


Figura 3.5: árbol con 50 secuencias del virus MPXV sin postprocesamiento salida de NCD

3.5. Sobre Augur y su comparación con NCD

Es interesante plantearse como 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. Augur es un proyecto open source para el análisis filogenético que surgió a finales de 2019. Este proyecto está formado por un conjunto de herramientas o subcomandos por ejemplo `augur align`, `augur tree`, `augur export` que proporcionan un multiple viral and bacterial pathogens analysis. Con este método 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 options for faster alignment of large numbers of sequences, higher accuracy alignments, alignment of non-coding RNA sequences, and the addition of new sequences to existing alignments. `raxml` ... Estos resultados que a la vista son mucho mejores que los que obtenemos con NCD se pueden mirar con

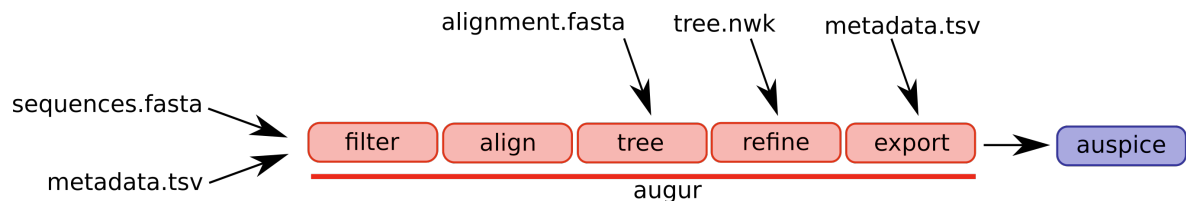


Figura 3.6: pipeline

3.6. Clasificación de variantes de interés

Capítulo 4

Compresión para otro tipo de secuencias

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.

4.0.1. ADN mitocondrial

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 elimina 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.

Tanto en estructura general como varios subárboles eran muy similares entre el árbol formado por NCD y el formado por augur. Los resultados de estas comparaciones se pueden observar en el repositorio...

4.0.2. 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 día de 24 de agosto ya hay registrados más de 44.000 casos en el mundo y más de 6000 en España.<https://www.cdc.gov/poxvirus/monkeypox/response/2022/world-map.html>

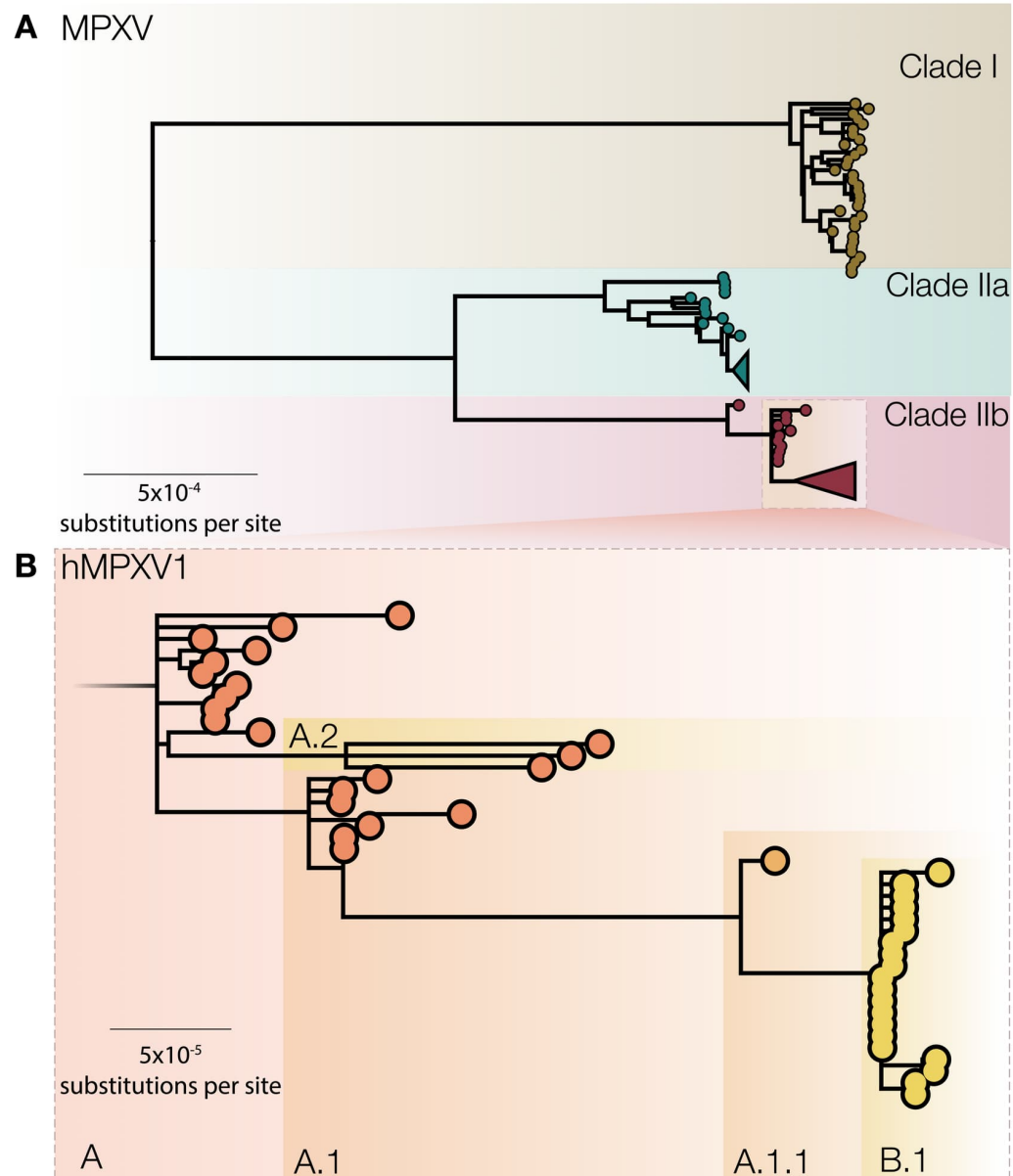


Figura 4.2: Esquema de división en clados del virus MPXV

Capítulo 5

Conclusiones

Algoritmo de caracter 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 rapidamente 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 6

Bibliografía

- [1] H. Farnsworth. What-if machine analysis and design. *IEEE Transactions on quantum neuroscience electronics*, 3031.
- [2] N. Sonntag. *Mis mejores recetas con repollo*. Anaconda, 2016.
- [3] S. Z. Ramírez, K. Pérez. Self conscious robots in induction heating home appliances. *IEEE transactions on anthropomorphic robots*, 2018.
- [4] Alumno Apellidos. Citar un tfg. Trabajo fin de grado, Universidad de Zaragoza, 2014.
- [5] Alumno Apellidos. Citar un tfm. Trabajo fin de máster, Universidad de Zaragoza, 2014.

Lista de Figuras

3.1. 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	8
3.2. 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	10
3.3. Árbol salida	11
3.4. árbol con 50 secuencias del virus MPXV con branch lengths tras el postprocesamiento de los datos de la matriz de distancias NCD	14
3.5. árbol con 50 secuencias del virus MPXV sin postprocesamiento salida de NCD	15
3.6. pipeline	16
4.1. árbol salida del algoritmo NCD	18
4.2. Esquema de división en clados del virus MPXV	19

Lista de Tablas

3.1. Pie de tabla 9

Anexos

Anexos A

Un anexo

Sed ut perspiciatis unde omnis iste natus error sit voluptatem accusantium doloremque laudantium, totam rem aperiam, eaque ipsa quae ab illo inventore veritatis et quasi architecto beatae vitae dicta sunt explicabo. Nemo enim ipsam voluptatem quia voluptas sit aspernatur aut odit aut fugit, sed quia consequuntur magni dolores eos qui ratione voluptatem sequi nesciunt. Neque porro quisquam est, qui dolorem ipsum quia dolor sit amet, consectetur, adipisci velit, sed quia non numquam eius modi tempora incidunt ut labore et dolore magnam aliquam quaerat voluptatem. Ut enim ad minima veniam, quis nostrum exercitationem ullam corporis suscipit laboriosam, nisi ut aliquid ex ea commodi consequatur? Quis autem vel eum iure reprehenderit qui in ea voluptate velit esse quam nihil molestiae consequatur, vel illum qui dolorem eum fugiat quo voluptas nulla pariatur? At vero eos et accusamus et iusto odio dignissimos ducimus qui blanditiis praesentium voluptatum deleniti atque corrupti quos dolores et quas molestias excepturi sint occaecati cupiditate non provident, similique sunt in culpa qui officia deserunt mollitia animi, id est laborum et dolorum fuga. Et harum quidem rerum facilis est et expedita distinctio. Nam libero tempore, cum soluta nobis est eligendi optio cumque nihil impedit quo minus id quod maxime placeat facere