Desenvolupament d'un sistema de filogènia vírica en Python

Anna Folch Salvador

Resum—Els avenços científics de la metagenòmica han portat al descobriment d'un gran nombre d'organismes i les seves relacions evolutives. Uns d'ells són els bacteriofags que, a pesar de ser els organismes més abundants de la biosfera, encara no estan completament descrits. No obstant això gràcies a l'anàlisi filogenètica s'han realitzat grans descobriments relatius a l'origen evolutiu de les espècies víriques. A més, el brot de pandèmia ocasionada pel SARS-CoV-2 ha fet incrementar els esforços de recerca en aquesta àrea. Aquest projecte s'ha centrat en el desenvolupament d'un programa informàtic, anomenat ViPhy, per a l'alineament de seqüències de proteïnes i la generació d'arbres filogenètics de bacteriòfags. ViPhy també es pot aplicar a altres tàxons i, a diferència d'altres programes, no limita el nombre de seqüències a alinear i és de codi obert.

Paraules clau—Arbres filogenètics, Filogenòmica, Bacteriòfags, Aminoàcids, SARS-CoV-2

Abstract—Scientific advances in metagenomics have led to the discovery of a significant number of organisms and their evolutionary relationships. One of them are bacteriofages that, despite being the most abundant organisms in the biosphere, are not yet fully characterized. However, thanks to the phylogenetic analysis, significant discoveries have been made regarding the evolutionary origin of the viral species and outbreak of the SARS-CoV-2 pandemics has increased the research efforts in this area. This project has focused in the development of a computer program, called ViPhy, for the alignment of protein sequences and the generation of bacteriophage phylogenetic trees. ViPhy can also be applied to other taxons and, unlike other programs, does not limit the number of sequences to be aligned and is open code.

Index Terms—Phylogenetic tree, Phylogenomics, Bacteriophage, Amino acids, SARS-CoV-2

1 Introducció - Context del treball

L'anàlisi filogenètica ens permet conèixer les relacions evolutives entre espècies, el que ens facilita trobar un ancestre comú entre elles i, alhora, poder realitzar una classificació real dels organismes, en una estructura en forma d'arbre. De fet, la filogenètica, com una disciplina de la biologia evolutiva, ha demostrat ser una eina amb gran eficiència per descobrir els orígens de les malalties i l'estudi de la seva futura propagació [1].

Els bacteriòfags, o virus que infecten bacteris, són els organismes més abundants de la Terra, amb una estimació de més de 10^{31} partícules víriques, el qual equival a una quantitat deu vegades superior al total de bacteris existents a la biosfera [2]. A diferència d'altres éssers vius, l'evolució dels bacteriòfags (o "fags") no és determinada únicament per una herència vertical (i.e. de pares a fills) sinó també per la transferència horitzontal o "*Horizontal gene transfer*" (HGT), a partir d'altres bacteriòfags, plasmidis i cromosomes bacterians.

Els mètodes d'inferència filogenètica molecular clàssica es basen en l'ús de gens marcadors. Això vol dir que, per exemple en bacteris, és comú fer servir el gen de la subunitat 16S de l'RNA ribosòmic com a marcador universal. El motiu és que aquest gen és essencial en tots els bacteris i es pot utilitzar la seva seqüència, adequadament alineada, per inferir com han evolucionat els diferents organismes i quina és la seva afinitat filogenètica [3] [4]. Per contra, en els fags, tot i que es poden trobar alguns gens amb funcions més rellevants, com ara la terminasa, cap d'ells és realment essencial. A més, donat que els virus intercanvien gens constantment, la història d'un dels seus gens difícilment recapitula la seva evolució al complet. Tot això fa necessària la recerca de nous enfocaments, basats en l'anàlisi de genomes virals sencers, que permetin inferir la seva filogènia.

Les tècniques de "*DNA-DNA hybridization*" (DDH) són utilitzades per estimar la similitud general entre genomes. Malauradament, són tedioses i amb una alta propensió a errors, per tant, poden ser substituïdes o complementades per procediments "*in silico*", realitzats mitjançant una simulació computacional [5]. Amb el temps, per poder

estudiar l'evolució dels fags (i d'altres virus, com el SARS-CoV-2), s'han desenvolupat tècniques de filogènia, basades en el genoma complet dels virus [6], que alineen fragments dels genomes dels fags i computen la distància entre ells. Una de les eines més utilitzades per realitzar un alineament de seqüències és el "Basic Local Alignment Search Tool" (BLAST) [7]. Aquest *software* és capaç de trobar les millors coincidències o *hits* entre les seqüències comparades. Les seves capacitats són utilitzades tant per inferir en relacions funcionals o evolutives entre seqüències, com per ajudar a identificar la família a la qual pertany cada gen.

Aquest projecte s'ha centrat en el desenvolupament d'una eina per a la classificació i construcció d'arbres filogenètics vírics, basat en l'alineament de genomes prèviament seleccionats per l'usuari.

2 OBJECTIUS

L'objectiu principal del projecte ha estat la generació d'una *pipeline* de codi obert, anomenada ViPhy, per a l'anàlisi filogenètic de genomes virals complets a partir del llenguatge de programació Python. S'ha pres com a model la coneguda eina web VICTOR [2], focalitzant-se en els bacteriòfags i resolent les limitacions presents en el servei web.

Els objectius secundaris han estat els següents:

- Implementació dels processos de transcripció i traducció *in silico* d'una seqüència de nucleòtids a proteïnes.
- Integració del programari existent com PHYML i BLAST [7] per l'alineament de seqüències i l'estimació de filogènies.
- Desenvolupament de biblioteques de funcions, capaces de suportar el càlcul de distàncies entre genomes.
- Desenvolupament de biblioteques pel suport estadístic via *bootstrap*.
- Validació de la pipeline utilitzant grups taxonòmics establerts de bacteriòfags i altres virus com el SARS-CoV-2.
- Redacció d'un manual d'usuari de la pipeline

3 ESTAT DE L'ART

Tal com s'ha comentat en l'apartat anterior, el servei web independent conegut com a VICTOR és un dels més utilitzats per alineament de seqüències i generació d'arbres filogenètics de tota mena d'éssers vius. Tanmateix, VICTOR presenta unes limitacions importants. Dues de les més destacables és que no és "open source" i només pot comparar fins a un màxim de 100 genomes o seqüències proteiques [2]. A més, presenta problemes en el cas de treballar amb genomes no anotats, i fa necessari pujar manualment la seqüència de DNA o RNA a l'eina. Aquesta opció, però, no és de gaire utilitat en el cas de virus, a causa de la gran variabilitat que presenten les seqüències de DNA o RNA.

Tanmateix, aquest projecte ha desenvolupat una eina que supera algunes de les mancances de VICTOR anomenades. Així, no estableix cap limitació en el nombre de genomes a analitzar i, com s'ha mencionat anteriorment, és de codi lliure i fàcilment accessible des de la plataforma Github [8].

4 METODOLOGIA

4.1 Anàlisi de requeriments

4.1.1 Determinació de plataformes de desenvolupament

El programa ViPhy s'ha desenvolupat en el llenguatge Python versió 3.8.3, utilitzant les biblioteques necessàries per poder dur a terme i suportar la computació biològica o filogenètica necessària. Per tal d'evitar problemes de dependència amb la versió de Python s'ha creat, amb les biblioteques utilitzades, un entorn virtual independent a partir de Conda, que també es pot trobar al repositori Github anteriorment mencionat.

També s'ha fet servir Biopython, un software de codi obert escrit en el mateix llenguatge de programació que la resta del projecte, que disposa d'un conjunt d'eines de gran utilitat en la computació biològica i la bioinformàtica [10]. De la mateixa manera, entre les seves funcionalitats es troba una interfície per generar i interpretar una gran varietat d'arxius, entre els quals es troben els fitxers de tipus fasta i genbank, els dos formats d'entrada acceptats pel programa.

Els genomes dels bacteriòfags necessaris per elaborar l'anàlisi filogenètica s'han extret de la web del "National Center for Biotechnology Information" (NCBI) [9], divisió de la "National Library of Medicine" (NLM). El NCBI alberga un conjunt de bases de dades de gran rellevància en camps com la biotecnologia, biomedicina, bioquímica, biologia molecular o genètica. Una mostra d'això és GenBank, una base de dades pública amb milions de seqüències de nucleòtids i anotacions bibliogràfiques i biològiques de suport [11] comunament referenciada en estudis en aquests àmbits.

Respecte als arbres filogenètics, "*Neighbor Joining*" (NJ) i "FastME" són dos mètodes que permeten la seva construcció amb certa rapidesa, sent especialment efectius per estudis a gran escala o anàlisis *bootstrap* [12]. A pesar de l'eficiència de FastME i de què és utilitzat per VICTOR, el programa desenvolupat en aquest projecte utilitza NJ per minimitzar les dependències externes.

4.1.2 Fitxer de configuració

L'eina ViPhy inclou un fitxer de configuració, escrit en format "JavaScript Object Notation" (*JSON*), en el qual cada usuari pot indicar les seves especificacions. A la *Figura Suplementària 1* es pot veure una mostra del contingut del fitxer *JSON*, seguint una estructura

clau-valor.

En el fitxer de configuració *JSON* es troben els següents camps:

- Tipus d'anàlisi a realitzar segons les dades d'entrada. Les opcions disponibles són, per una banda, "protein" o "amino acid" i, per l'altre, "nucleic acid" o "nucleotide".
- Correu identificatiu per accedir a la base de dades de NCBI i descarregar els fitxers seleccionats.
- Noms dels directoris, amb els seus *path*, on es volen emmagatzemar les dades d'entrada i sortida, així com altres dades temporals creades durant la mateixa execució del programa.
- Seqüències específiques que es volen obtenir accedint directament de les bases de dades de NCBI.
- Altres paràmetres de caràcter numèric que permeten concretar el valor del paràmetre *e-value* (valor que indica el nombre d'encerts que s'esperen trobar per casualitat en alinear dues seqüències) i el nombre de rèpliques de *bootstrap.*
- Fórmula de distància seleccionada per l'usuari. Entre les opcions es troben "*d0*", "*d4*" i "*d6*", sent la darrera la funció definida per defecte.
- Conjunt de booleans que permeten decidir les sortides que es pretenen obtenir un cop finalitzada l'execució del programa.

4.2 Definició del workflow

El programa ViPhy s'executa en diferents etapes successives. La primera és el preprocessament, en el que es produeix la recollida de les seqüències i, si és necessari, la traducció a proteïna i la conversió en un format comú de totes les dades.

A continuació es construeix una base de dades de tipus BLAST i es realitza un alineament per cada parell de seqüències disponible. El resultat és un vector de cobertura anomenat *coverage vector*. Utilitzant aquest vector es pot calcular la distància genètica per cada gen i obtenir el total de rèpliques de *bootstrap* indicades al fitxer de configuració. L'últim pas consisteix en la generació d'arbres consens a partir de les dades anteriorment processades.

4.3 Preprocessament

Aquesta primera fase estandarditza les seqüències de nucleòtids o proteïnes (extretes de fitxers en format fasta i/o genbank) per evitar futurs problemes de format durant l'execució de la resta del procés. En la *Figura Suplementaria 2* es mostren els fluxos que el codi realitza internament segons la configuració establerta per l'usuari. Com es pot veure a la Figura, els processos divergeixen segons si s'analitzen nucleòtids o proteïnes. En el primer cas, la seqüència de nucleòtids es tradueix a aminoàcids, en el segon cas això no es requereix.

En el cas dels bacteriòfags és més recomanable realitzar un alineament d'aminoàcids que de nucleòtids, donat que la seqüència proteica presenta menys variabilitat evolutiva que la nucleòtida. Efectivament, moltes mutacions sobre la seqüència de DNA són silencioses i no provoquen canvis en la proteïna corresponent.

4.3.1 Transcripció i traducció

Biològicament, el procés de transcripció precedeix al de traducció. Aquest terme fa referència al procés de síntesi d'RNA, prenent com a motlle una cadena de DNA [13]. La transferència de la informació d'una seqüència a l'altra es realitza seguint les regles de complementarietat de les bases nitrogenades, amb la diferència que en aquesta nova seqüència resultant s'inclourà el nucleòtid conegut com a uracil (U), en els casos en què abans hi havia una timina (T). La transcripció no ha estat implementada en el codi com si ho ha estat la traducció.

Durant la traducció es llegeix una molècula anomenada "*messenger RNA*" (mRNA) en grups de tres bases (codons), els quals s'utilitzen per sintetitzar una proteïna fent servir una taula de traducció. Que el codi genètic es llegeixi en grups de tres parells de bases significa que, en una molècula de DNA de doble cadena, poden obrir-se fins a 6 possibles marcs de lectura (3 per cada sentit de la lectura).

Com es mostra al diagrama de flux de la F*igura Suplementaria 3*, en el programa ViPhy la traducció es realitza en dos passos: primer es calculen els tres primers marcs de lectura i, seguidament els tres restants. Més tard, com indica la *Figura Suplementaria 2*, aquests 6 marcs obtinguts es combinaran com una única seqüència i finalitzarà el preprocessament.

4.4 Alineament de seqüències

Durant aquesta fase té lloc l'alineament i creació del vector de *coverage*. El primer pas consisteix a crear una base de dades amb tots els proteomes acceptats i obtinguts durant l'etapa anterior. A continuació, s'aplica BLAST per alinear localment cada una de les seqüències d'aminoàcids sobre la base de dades acabada de crear [6]. En el projecte s'aplica específicament BLASTp, el qual compara les seqüències de proteïnes individuals amb la base de dades de proteomes creada pel programa.

BLAST busca un *hit* d'alta puntuació o "high-scoring segment pair" (HSP), que s'expandeix el màxim possible tant per la dreta com per l'esquerra. Aquest representa el nivell de similitud entre dues seqüències genètiques alineades. Un cop finalitzat el procés d'alineament, s'extreu la informació continguda en aquests HSP i es guarda en el vector de *coverage*. El vector generat té la mateixa longitud que la seqüència alineada amb la base de dades i indica les posicions on s'ha trobat aquestes coincidències. Seguidament, i fent ús d'aquest vector, es calcula la distància genoma a genoma mitjançant una fórmula de distància específica [11].

4.5 Càlcul de distàncies

Existeixen diverses maneres per calcular la distància entre

dues seqüències, per exemple "Genome BLAST Distance Phylogeny" (GBDP) utilitza un total de 10 fórmules diferents amb petites diferències entre elles [15]. D'entre aquestes s'han escollit 3 (concretament *d0, d4* i *d6)* perquè l'usuari pugui comparar els resultats.

La fórmula *d0* calcula la proporció del genoma que cobreix el HSP (1). Per fer-ho, té en compte la mida dels HSPs i la longitud total del vector de *coverage*. Per altra banda, *d4* computa el nombre total de parells de bases idèntiques o identitats en un HSP, en relació amb la cobertura total d'aquest HSP (2). Una de les característiques que distingeix *d4* de les altres dues fórmules analitzades és que no té en compte la longitud total dels genomes, fet que la fa més robusta en el cas de trobar-se amb genomes incompletament seqüenciats [15].

A pesar de tot, la principal fórmula que s'ha implementat en el programa ViPhy ha estat *d6* (*3*), que consisteix en una combinació entre d0 i d4. En efecte, per calcular la distància entre dues seqüències, *d6* té en compte el nombre total de parells de bases idèntiques, però en relació amb la longitud total dels dos proteomes, al contrari que en el cas de *d0*.

$$d_0(X,Y) = 1 - \frac{H_{XY} + H_{YX}}{\lambda(X,Y)} \qquad (1)$$

$$d_4(X,Y) = 1 - \frac{2 * I_{XY}}{H_{XY} + H_{YX}} \qquad (2)$$

$$d_6(X,Y) = 1 - \frac{2 * I_{XY}}{\lambda(X,Y)} \qquad (3)$$

- **XY:** Execució de BLAST del genoma Y (query) contra el genoma X (db).
- **YX:** Aplicació de BLAST de X (db) contra Y (query).
- $\mathbf{H_{XY}}$: Posicions del vector de *coverage* cobertes pels HSPs trobats al alinear una seqüència Y sobre X.
- $\boldsymbol{\lambda_{XY}}$**:** Suma de les longituds totals dels vectors de *coverage* obtinguda en alinear la seqüència Y amb X, i a la inversa .
- $\mathbf{I_{XY}}$**:** Nombre d'identitats en els HSPs, és a dir, nombre de bases iguals trobades en un alineament de seqüències. Per resoldre les situacions en què es produeix una superposició o *overlapping*, és calculat utilitzant la *fórmula (4),* que barreja la informació de tots dos HSPs.

$$I_{XY} = \frac{\frac{\sum(\nu x)}{|\nu x|} + \frac{\sum(\nu y)}{|\nu y|}}{2} \qquad (4)$$

El resultat d'aplicar una de les fórmules és una matriu en què apareixen totes les distàncies calculades de l'alineament de cada possible parell de seqüències.

4.5.1 Bootstrap

El *bootstrapping* és una tècnica de remostreig comunament utilitzades en estadística i en els estudis filogenètics per poder estimar la importància de les branques d'un arbre, ja que permeten realitzar nous càlculs significatius amb un nombre limitat de dades. El nombre de mostres pel *bootstrap*, definit per defecte al programa ViPhy, és de 100 rèpliques. La duració del procés s'incrementarà en relació amb la quantitat de rèpliques indicades, però alhora també augmenta la fiabilitat de les dades. Si després de realitzar el procés un nombre elevat de vegades es troba que, a vegades, es tornen a obtenir els valors inicials, això implica un major suport a aquell resultat [16].

Donat el vector de *coverage* i segons el nombre de rèpliques prèviament seleccionades en el fitxer de configuració (*Figura Suplementaria 1*), es realitzen la mateixa quantitat de mostres de *bootstrap*. Les rèpliques són mostrejos amb reemplaçament de les dades originals per crear conjunts de dades fictícies, però amb la mateixa longitud de la seqüència inicial [14].

Per crear cada una d'aquestes mostres es fa servir la funció que es pot veure a la *Figura Suplementària 4*. En aquest fragment del codi del programa ViPhy es mostra que, a partir d'un diccionari on es recullen els vectors de *coverage* per cada parell de seqüències, es pot calcular un altre vector de la mateixa longitud, però amb els valors en posicions diferents. De fet, es pot veure que el fragment de codi utilitza un generador de nombres aleatoris per calcular quin serà el valor a copiar del vector de *coverage* original a una posició concreta d'un nou vector.

L'usuari pot indicar si vol rebre com a sortida la matriu de distància de l'arbre original, així com de cada una d'aquestes matrius creades per cada replicat de *bootstrap*. Com més properes són dues seqüències genèticament, menor serà la distància entre elles, per tant, 0 simbolitza una completa similitud i 1 una diferència total entre elles.

4.6 Arbres filogenètics

Entre les possibles maneres de representar arbres filogenètics, en aquest projecte se n'han aplicat dues realitzables mitjançant el mòdul *Phylo* de Biopython. L'usuari té l'opció de seleccionar si prefereix una d'elles, cap d'elles o totes dues.

La primera opció permet crear un arbre amb valors (o "support") a les seves branques, donat un conjunt de rèpliques de *bootstrap*. Per aquest fi s'utilitza com a model un arbre inicial i es calculen un conjunt de mostres vía

bootstrap del vector de *coverage*. La comparació dels resultats, per cadascun dels replicats, resulta en uns valors per cada branca. El codi encarregat de realitzar aquest procés es recull a la *Figura Suplementària 5*.

Per altra banda, també existeix la possibilitat de crear un arbre de consens mitjançant la comparació directa de les mostres de *bootstrap* obtingudes. Com es pot veure en el codi d'aquest procés (*Figura Suplementària 6)*, es pot indicar un llindar mínim pels valors de les branques. D'aquesta manera si una branca no supera el llindar establert, no es mostrarà la unió de dues seqüències, a pesar d'haver similitud entre elles.

Un cop finalitzada l'última etapa del codi es crearà, per cada arbre, un document en format *newick*, perquè l'usuari pugui visualitzar-los de la manera que cregui convenient.

5 RESULTATS

5.1 Test amb resultats controlats

Per validar el funcionament del programa, la primera prova realitzada ha estat partir de dades controlades i fer servir la fórmula de distància d6. En concret, s'han utilitzat quatre seqüències curtes d'aminoàcids (amb els noms "Original_seq", "Test_seq_1", "Test_seq_2" i "Test_seq_3"), no existents a la natura, per poder calcular i comprovar manualment el resultat (*Figura 1*).

En primer lloc es van comparar les seqüències amb elles mateixes **(Figura 1A)**, i es va trobar un únic alineament (HSP) de la mateixa mida del vector de *coverage*, tal com s'esperava. Les coincidències eren del 100% i les distàncies 0.

Figura 1: Representació gràfica del resultat d'un alineament amb dades controlades entre parells de seqüències. Els HSPs de les seqüències estan marcats en diferents colors per cada HSP i cas analitzat. En el cas **(A)** es compara una seqüència amb ella mateixa. Només es troba un HSP que s'indica en verd. **(B)** recull l'alineament de dues seqüències amb petites regions de similitud, alhora que també presenta alguna mutació en el HSP amb un blau més fosc. A **(C),** es mostra un cas d'*overlapping* entre dos HSPs, marcant aquesta superposició amb color taronja. Per últim, en el cas **(D)** es compara la seqüència original amb una segona generada de forma aleatòria.

A continuació es varen alinear seqüències diferents. En el cas de "Original_seq" i "Test_seq_1" (Figura 1B), es poden observar tres HSPs diferents, tant en alinear la primera seqüència contra la segona, com a la inversa, a pesar que hi ha una petita mutació en el HSP. La distància entre elles calculada a partir de la fórmula *d6*, és de 0.89.

En el cas de "Original_seq" i "Test_seq_2" **(Figura 1C)**, hi ha una superposició o *overlapping*. Quan succeeix això, s'aplica l'algoritme de *coverage*, en el que es fusiona la informació dels dos HSPs. A la *Figura 2* es pot observar el vector de *coverage* generat a causa de l'*overlapping*. La fórmula de distància dona com a resultat un valor de 0.89435.

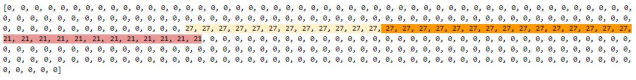

Figura 2: Vector de *coverage* resultant de comparar les seqüències "Original_seq" i "Test_seq_2". La zona marcada en color taronja implica que s'ha produït un *overlapping*, mentre que els altres dues zones ressaltades amb altres colors representen els trossos dels HSPs que no s'han superposat.

Si en comptes d'utilitzar l'algoritme de coverage es fes servir *greedy-with-trimming* el resultat seria diferent, ja que aquest algoritme elimina la regió en la qual es produeix la superposició, disminuint el nombre total d'identitats (veure apartat **4.5**). La distància obtinguda en aplicar aquest segon algoritme seria de 0.91768. Com es pot veure el valor resultant seria superior al calculat anteriorment, per tant, les seqüències es considerarien més distants. El programa ViPhy fa servir exclusivament l'algoritme de *coverage* i no el *greedy-with-trimming* per evitar aquest tipus d'efectes.

Finalment, en alinear seqüències completament dispars (*Figura 1D*), ja que entre elles no existeix cap mena de similitud, no es troba cap HSP i la distància entre elles és 1 exactament, és a dir, el màxim possible; tal com s'esperava.

5.2 Test proteína vs DNA traduït

A continuació, el programa Viphy es va provar amb seqüències reals de bacteriòfags. El test va comparar les diferències produïdes en alinear directament proteïnes o DNA traduït (*Figura 1*). Aquesta opció, tal com s'ha indicat en apartats anteriors, està a disposició de l'usuari al fitxer de configuració, així com l'elecció de la fórmula de la distància.

En aquest primer test es van fer servir un total de 15 seqüències de la família dels fags *Tectiviridae*, extretes de la base de dades del NCBI. Aquestes seqüències havien estat analitzades anteriorment [17], trobant-se l'arbre filogenètic que es presenta a la *Figura Suplementària 7*.

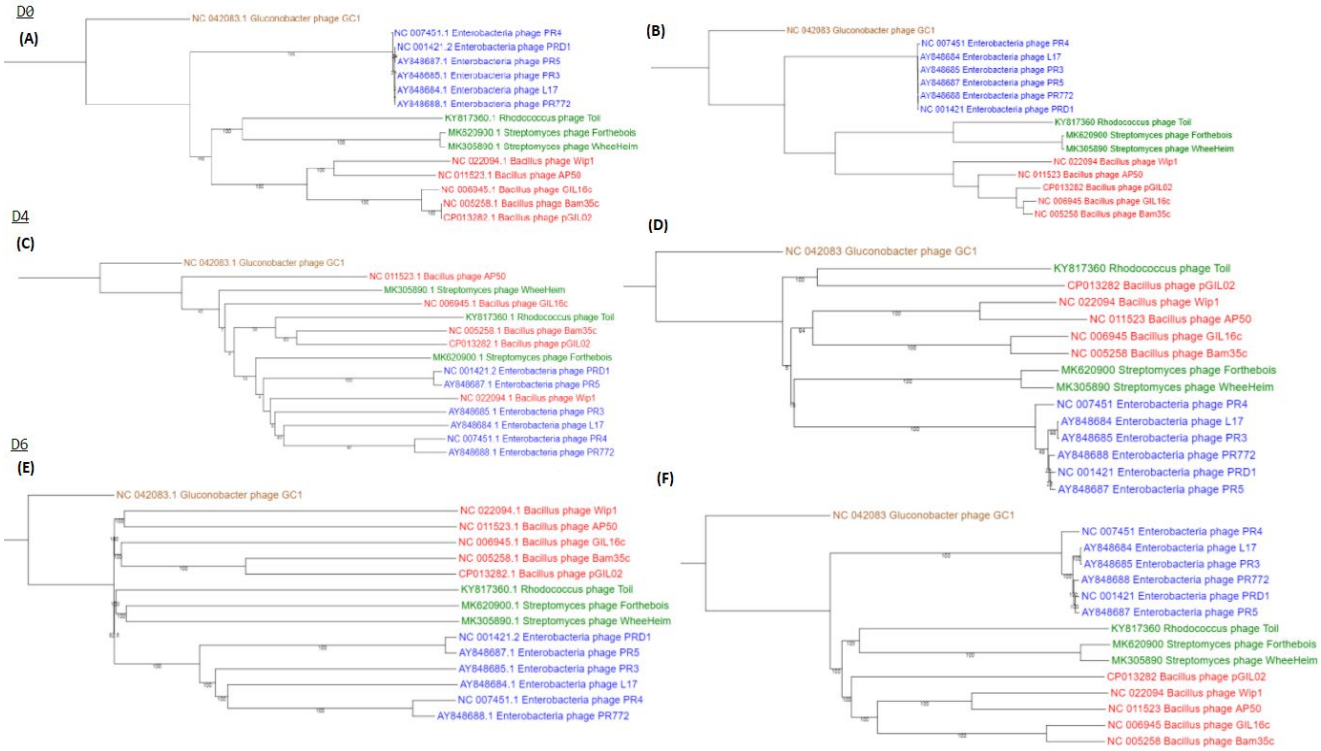

Figura 3: Arbres filogenètics generats per el programa ViPhy, partir del genoma complet de 15 bacteriòfags de la família dels *Tectiviridae*. A cada fila es mostra els arbres resultants d'aplicar una fórmula diferent per calcular la distància. Segons la columna, es poden diferenciar segons el tipus de dades d'entrada (A, C, E, alineaments basats en DNA. B, D, F, alineaments basats en proteïna. A-B, alineaments amb d0; C-D, d4; E-F, d6). Els colors, per la seva part, diferencien els gèneres de la família dels *tectiviridae* (Alphatectivirus (blau); Betatectivirus, (vermell); Gammatectivirus, (marró) i Deltatectivirus (verd)) de la mateixa manera que succeeix a la *Figura Suplementària 7*.

A la *Figura* 3 es presenten els arbres filogenètics generats pel programa ViPhy a partir d'aquestes 15 seqüències. Els arbres estan separats en dues columnes segons el tipus de test prèviament seleccionat per l'usuari, és a dir, *'nucleotide*" (esquerra) o "*protein*" (dreta). Segons les dades d'entrada les longituds de les branques varien, sent generalment molt més llargues si es parteix de seqüències de nucleòtids. És de destacar que això succeeix pel fet que ViPhy, a diferència del programa VICTOR, genera 6 seqüències d'aminoàcids a partir d'una seqüència de DNA, que després concatena. L'increment de la longitud de la seqüència d'aminoàcids afecta la distància final de la seqüència analitzada, la qual varia, mentre que els HSPs poden o no mantenint la mateixa mida ja que tenen més espai per expandir-se. Això es fa més evident en el gènere dels *Alphatectivirus* (Figura 3). En canvi, els arbres basats en proteïnes, generats pel programa ViPhy (Figura 3 B, D, F), presenten unes branques més curtes que en el cas del DNA. Aquest fet també és palès en la publicació original (Figura 9).

A la figura també es presenten els arbres obtinguts segons la fórmula aplicada. De forma inesperada, per *d0* les distàncies entre espècies calculades a partir de proteïnes i DNA són les que menys varien (Figura 3A i B). A més, els arbres generats al aplicar aquesta fórmula són els més semblants respecte al prèviament publicat (comparar Figures 3 A, B amb Figura *Suplementària 7*). En canvi, en el cas de *d4* i *d6* la longitud de les branques es dispara si els arbres es calculen amb nucleòtids, tal com s'esperava. Això es deu al fet que, quan s'incrementa la longitud total de la seqüència, els HSPs poden seguir expandint-se, juntament amb el nombre d'identitats. A pesar de tot, les identitats són de menor qualitat, és a dir, que a pesar de créixer a l'alineament es troba un percentatge menor de bases idèntiques.

En el cas de *d4,* per a seqüències de nucleòtid, les agrupacions d'espècies varien respecte a les calculades amb aminoàcids. Tanmateix, els valors a les branques són molt més baixos, generalment per sota del 70% (*Figura 3C*). Els valors que superen aquest llindar apareixen en branques que destaquen sobre la resta pel baix nivell de canvi genètic que presenten i que es compleixen també en la resta d'arbres de la *Figura 3*. En canvi, en aplicar d4 en seqüències d'aminoàcids (*Figura 3D*) es visualitza una lleugera reorganització de les espècies i alhora s'incrementen aquests valors de suport de les branques.

Finalment, la fórmula *d6* és la que dóna els arbres més dispars segons si es parteix de seqüències de nucleòtids o de proteïnes. Presenta aquestes variacions significatives entre resultats a causa de l'increment de la longitud de la seqüència de nucleòtids en ser traduïda, que fa que les branques de l'arbre filogenètic es distanciïn visiblement (*comparar Figura 3E amb 3F*), però sempre respectant l'ancestre comú més recent entre les espècies alineades.

L'eina VICTOR resol els overlappings aplicant l'algorisme

de greedy-with-trimming i construeix l'arbre filogenètic a partir del software FastME, que segons altres estudis ha demostrat tenir una major precisió topològica que NJ [12]. Això resulta en distàncies diferents a les calculades per ViPhy. Amb tot, els seus arbres filogenètics resultants (Figura 4), presenten poques diferències al comparar-los amb els arbres de la Figura 3 pels alineaments realitzats a partir de seqüències de proteïnes (comparar *Figures 3B, D i E* amb *Figures 4B, D i E*).

En ambdues figures les fórmules d0 i d6 recolzen la idea que els fags WheeHeim ("MK305890") i Forthebois ("MK620900") són gèneres dins de la família dels tectiviridae, tal com indica l'article d'on s'ha extret la Figura Suplementària 7.

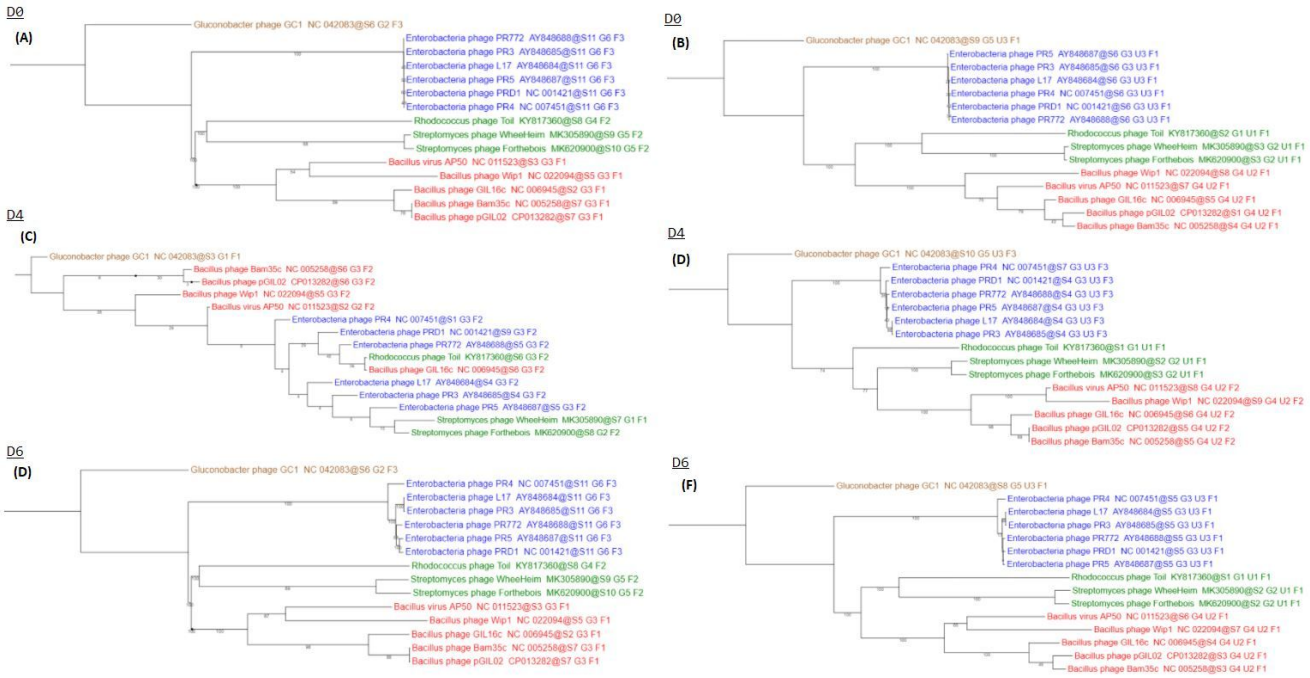

Figura 4: Arbres filogenètics, obtingut a partir del *software* VICTOR del genoma complet de 15 bacteriòfags de la família dels *Tectiviridae*. Verticalment es distingeixen tres files en les que es separen els arbres segons la fórmula de la distància aplicada (A-B, alineaments amb d0; C-D, d4; E-F, d6). Per cada columna, els arbres es diferencien segons el tipus de dades d'entrada (A, C, E, alineaments basats en DNA; esquerra. B, D, F, alineaments basats en proteïna; dreta). Els colors, diferencien els gèneres de la família dels *tectiviridae* (Alphatectivirus (blau); Betatectivirus, (vermell); Gammatectivirus, (marró) i Deltatectivirus (verd)) de la mateixa manera que succeeix a les *Figures 3 i Suplementària 7*.

5.3 Test amb més de 100 seqüències

En aquesta prova es va demostrar que el programa ViPhy supera el límit de 100 seqüències que permet analitzar VICTOR, gràcies al fet que és capaç d'escalar quan resulta necessari.

Es van utilitzar un total de 176 seqüències, extretes de l'article "Evidence for shared ancestry between Actinobacteria and Firmicutes bacteriophages" [18]. Seguint amb les pautes establertes per VICTOR, van definir-se 100 rèpliques de *bootstrap* utilitzat la fórmula *d6* per calcular la distància i es va indicar un e-value de 0,1.

El resultat es pot veure a la *figura 5*, en la que es presenta un arbre filogenètic amb un alt suport de *bootstrap* a les seves branques. A l'arbre es pot veure, per exemple, proximitat entre un dels fags que infecten *Firmicutes (FP Oengus) i els fags que infecten Gordonia, Rodococcus, Streptomyces i fins i tot els Arthrobacters.* Aquest fet suggereix que, entre ells, existeix un ancestre comú relativament proper. De fet, el fag Oengus FP comparteix més contingut genètic amb altres grups esmentats sí que amb altres espècies del seu mateix gènere, un fet que ja s'havia observat prèviament (18).

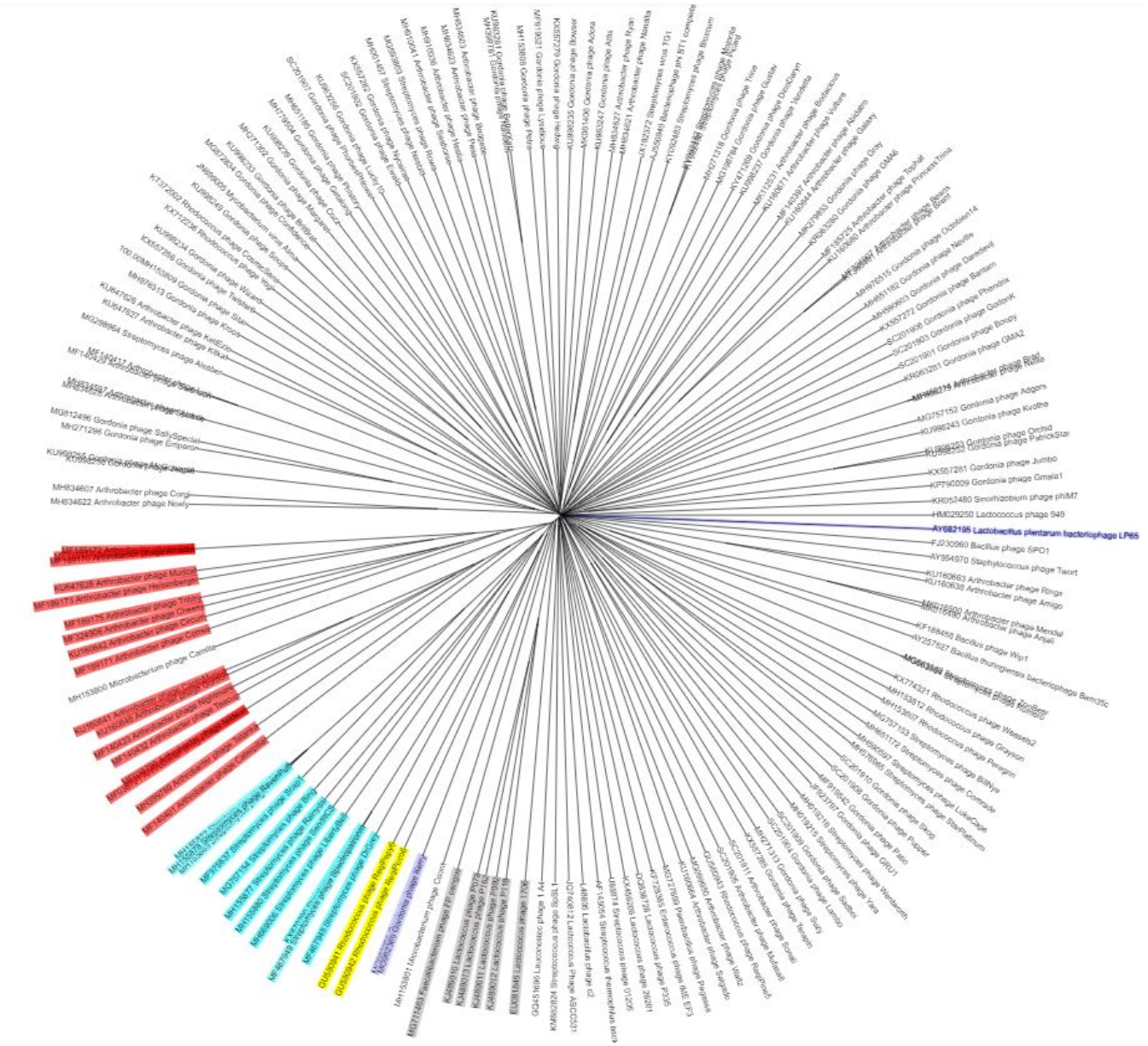

Figura 5: Arbre filogenètic arrelat format a partir de 167 seqüències genètiques i generat a partir de l'eina iTOL[20]. Es mostren els suports de les banques per a 100 replicats de *bootstrap*. Els colors denoten els gèneres *Firmicutes* (gris), *Arthrobacter* (vermell), *Rodococcus* (groc), *Streptomyces* (blau clar) i *Gordonia* (morat).

5.4 Test SARS-Cov2

A causa de la pandèmia global ocasionada per la COVID-19 s'ha cregut convenient realitzar un quart test amb l'objectiu de crear un arbre filogenètic a partir de l'alineament de les seqüències que pertanyen a la família dels *Coronaviridae*.

Per identificar l'origen dels patogens és imprescindible realitzar estudis filogenètics [19]. Amb aquesta idea i, després de realitzar un seguit d'anàlisis de recombinació i identificació de les diferents regions del genoma, es va indicar que el SARS-CoV-2 no és un recombinant de cap *sarbecovirus*, sinó que té una forta relació amb els virus dels ratpenats, amb un suport del 95%, així com dels pangolins.

L'arbre filogenètic generat amb el programa ViPhy es mostra a la *figura 6.* La branca corresponent a la soca de Wuhan es troba més propera a altres soques que també són espècies de "Severe Acute Respiratory Syndrome" (SARS), com per exemple la soca de Kenya (amb identificador "KY352407") amb un valor superior al 90%. Així i tot, l'elevat suport que es mostra en altres branques ens permet suposar que el SARS-CoV-2 pugui compartir ancestres comuns amb els virus que tenen ratpenats com amfitrions. De fet, entre les seqüències que no són de la mateixa espècie que el SARS, el seu ancestre calculat comú més recent és el "Bat coronavirus BM46" amb identificador "GU190215" (*Figura 6*).

Per validar els resultats de ViPhy respecte a la similitud del SARS de Wuhan i el seu ancestre calculat més proper de ratpenats, s'ha fet servir el software "Mauve". Donat

que aquesta eina permet comparar la similitud entre múltiples seqüències, s'ha afegit un altre SARS (específicament la soca de Kenya). Com es veu a la imatge de la *Figura Suplementària 8*, es repeteixen els resultats prèviament obtinguts amb ViPhy.

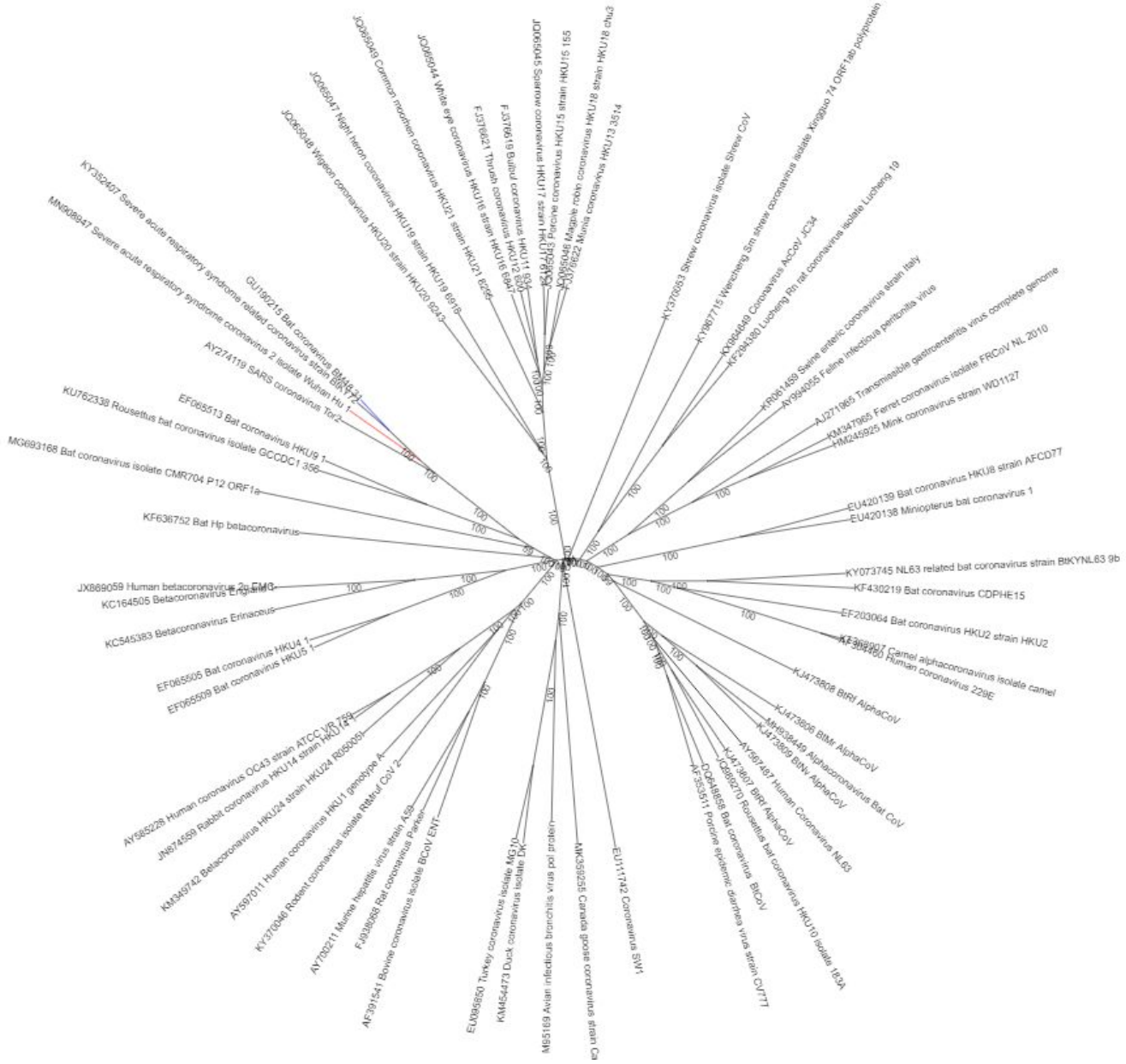

Figura 6: Arbre filogenètic del SARS-CoV-2 amb suport a les branques aplicant la fórmula de distancia d6 i amb un e-value de 0.001. En vermell es marca la branca de la soca de Wuhan ("MN908947") i en blau el virus dels ratpenats ("UG19215") més proper al SARS-CoV-2

6 CONCLUSIONS

S'ha desenvolupat un nou programa de codi lliure, anomenat ViPhy, que permet alinear seqüències de proteïna per obtenir arbres filogenètics. El programa es va inspirar en el servei web independent conegut com a VICTOR.

S'ha validat el programa ViPhy amb diversos tests, incloent-hi la generació d'un arbre genealògic de fags i seqüències víriques com és el cas del SARS-Cov-2, i s'ha trobat que els arbres resultants s'apropen considerablement a arbres prèviament publicats.

S'ha demostrat que el programa ViPhy supera algunes limitacions del software VICTOR, com poder comparar més de 100 seqüències i ser *open source*.

Per a futures actualitzacions, seria possible afegir els algoritmes de *greedy* i *greedy-with-trimming* com a opcions juntament amb l'algoritme de *coverage*, així com una opció addicional al fitxer de configuració per controlar el nombre de decimals de les distàncies computades.

AGRAÏMENT

PRIMERAMENT, M'AGRADARIA AGRAIR ALS MEUS TUTORS IVAN ERILL I JOAN SERRA-SAGRISTÀ PER TOT EL SUPORT QUE M'HAN PROPORCIONAT, PER ATENDRE ELS MEUS DUBTES I PER L'ALTA DISPONIBILITAT. HI HAVIA MOMENT EN QUÈ NO SEMBLAVA QUE EL PROJECTE POGUÉS TIRAR ENDAVANT I SEMPRE HEU ESTAT PRESENTS PER DONAR-ME SUPORT DAVANT DE QUALSEVOL PROBLEMA QUE POGUÉS SORGIR.

GRÀCIES, TAMBÉ TOTS ELS AUTORS QUE HAN DECIDIT COMPARTIR ELS GENOMES EN DIFERENTS BASES DE DADES PÚBLIQUES PERQUÈ SIGUIN ACCESSIBLES PER LA RESTA DEL MÓN, AIXÍ COM PER LA RESTA DE DESCOBRIMENTS I ARTICLES QUE AJUDEN A CRÉIXER LA NOSTRA SOCIETAT.

TAMBÉ VULL AGRAIR A JETBRAINS PER PROPORCIONAR-ME UNA LLICÈNCIA GRATUÏTA DURANT EL TEMPS QUE DURAVA EL TREBALL DE FI DE GRAU.

BIBLIOGRAFIA

[1] Mendoza-Revilla J. Aportes de la filogenética a la investigación médica. RMH [Internet]. 16 jul.2012; 23(2):119. Available from: https://revistas.upch.edu.pe/index.php/RMH/article/view/1042

[2] Meier-Kolthoff J.P., Göker M., VICTOR: genome-based phylogeny and classification of prokaryotic viruses, Bioinformatics, Volume 33, Issue 21, 01 November 2017, Pages 3396–3404.

[3] Janda J.M., Abbott S.L., 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: Pluses, perils, and pitfalls. J Clin Microbiol. 2007;45: 2761–2764. 10.1128/JCM.01228-07

[4] Clokie, M. R., Millard, A. D., Letarov, A. V., \& Heaphy, S. (2011). Phages in nature. Bacteriophage, 1(1), 31–45. https://doi.org/10.4161/bact.1.1.14942

[5] Stefan R. Henz, Daniel H. Huson, Alexander F. Auch, Kay Nieselt-Struwe, Stephan C. Schuster, Whole-genome prokaryotic phylogeny, Bioinformatics, Volume 21, Issue 10, Pages 2329–2335.

[6] Auch, A. F., Klenk, H. P., & Göker, M. (2010). Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. Standards in genomic sciences, 2(1), 142–148. https://doi.org/10.4056/sigs.541628

[7] Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. doi: 10.1016/S0022-2836(05)80360-2. PMID: 2231712.

[8] Folch, A. (2020) ErillLab/ViPhy [Source Code]. https://github.com/ErillLab/ViPhy

[9] National Center for Biotechnology Information (NCBI)[Internet]. Bethesda (MD): National Library of Medicine (US), National Center for Biotechnology Information; [1988] – [cited 2020 Oct. 01]. Available: https://www.ncbi.nlm.nih.gov/

[10] Cock, P.J.A. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics 2009 Jun 1; 25(11) 1422-3 https://doi.org/10.1093/bioinformatics/btp163 pmid:19304878

[11] Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016 Jan 4;44(D1):D67-72. doi: https://doi.org/10.1093/nar/gkv1276. Epub 2015 Nov 20. PMID: 26590407; PMCID: PMC4702903.

[12] Desper R, Gascuel O. Getting a tree fast: Neighbor Joining, FastME, and distance-based methods. Curr Protoc Bioinformatics. 2006 Oct;Chapter 6:Unit 6.3. doi: 10.1002/0471250953.bi0603s15. PMID: 18428768.

[13] Guo, Jiannan. "Transcription: the epicenter of gene expression." Journal of Zhejiang University. Science. B vol. 15,5 (2014): 409-11. doi: https://doi.org/10.1631/jzus.B1400113

[14] Ranganathan, Shoba \& Nakai, Kenta \& Schönbach, Christian. (2019). Encyclopedia of Bioinformatics and Computational Biology: ABC of Bioinformatics. Publisher: Elsevier. ISBN: 9780128114322

[15] Meier-Kolthoff, J.P., Auch, A.F., Klenk, H. et al. Genome sequence-based species delimitation with confidence intervals and improved distance functions. BMC Bioinformatics 14, 60 (2013).

[16] Richard M. Kliman, *Encyclopedia of Evolutionary Biology*. 1st Edition. Academic Press, 2016. Hardcover ISBN: 9780128000496. eBook ISBN: 9780128004265

[17] Caruso, Steven M.; deCarvalho, Tagide N.; Huynh, Anthony; Morcos, George; Kuo, Nansen; Parsa, Shabnam; Erill, Ivan. 2019. "A Novel Genus of Actinobacterial Tectiviridae" *Viruses* 11, no. 12: 1134. https://doi.org/10.3390/v11121134

[18] Koert M, López-Pérez J, Mattson C, Caruso S, Erill I (2020) Evidence for shared ancestry between Actinobacteria and Firmicutes bacteriophages. bioRxiv, 842583, ver. 3 peer-reviewed and recommended by Peer Community in Genomics. https://doi.org/10.1101/842583

[19] Boni, M.F., Lemey, P., Jiang, X. *et al.* Evolutionary origins of the SARS-CoV-2 sarbecovirus lineage responsible for the COVID-19 pandemic. *Nat Microbiol* 5, 1408–1417 (2020). https://doi.org/10.1038/s41564-020-0771-4

[20] Ivica Letunic, Peer Bork, Interactive Tree Of Life (iTOL) v4: recent updates and new developments, *Nucleic Acids Research*, Volume 47, Issue W1, 02 July 2019, Pages W256–W259, https://doi.org/10.1093/nar/gkz239

APÈNDIX

A1. ACRÒNIMS I SIGLES

BLAST	Basic Local Alignment Search Tool
DDH	DNA-DNA Hybridization
DNA	Deoxyribonucleic acid
GBDP	Genome BLAST Distance Phylogeny
HGT	Horizontal Gene Transfer
HSP	High-scoring segment pair
JSON	JavaScript Object Notation
NCBI	National Center for Biotechnology Information
NJ	Neighbor Joining
NLM	National Library of Medicine
RNA	Ribonucleic acid
SARS	Severe Acute Respiratory Syndrome
VICTOR	Virus Classification and Tree Building Online Resource

A2. GRÀFICS, DIAGRAMES, CODI I FIGURES

```
{
      "analysis_type": "protein",
      "user_email": "A.N.Other@example.com",
      "input_folder" : "Inputs",
      "working_folder" : "WorkingFolder",
      "output_folder" : "Outputs",
      "genome_accessions": [["KC545383.1"]],
      "e_value": 0.1,
      "distance_function": "d6",
      "replicates": 100,
      "cutoff": 0.7,
      "majority_or_support_tree": "both",
      "get_original_newick_tree": "True",
      "get_original_distance_matrix": "True",
      "get_bootstrap_distance_matrix": "True"
}
```

Figura Suplementària 1: Mostra del contingut del fitxer de configuració *JSON*. En aquest fitxer es recullen els directoris, els *accessions* o noms dels fitxers que es volen descarregar de NCBI, com aquest cas "KC545383.1" (Betacoronavirus Erinaceus), i altres camps a omplir segons les intencions de l'usuari. La major part de les variables que es visualitzen contenen el seu valor per defecte, en canvi, en el camp "genome_accessions" es presenta una mostra real del seu format.

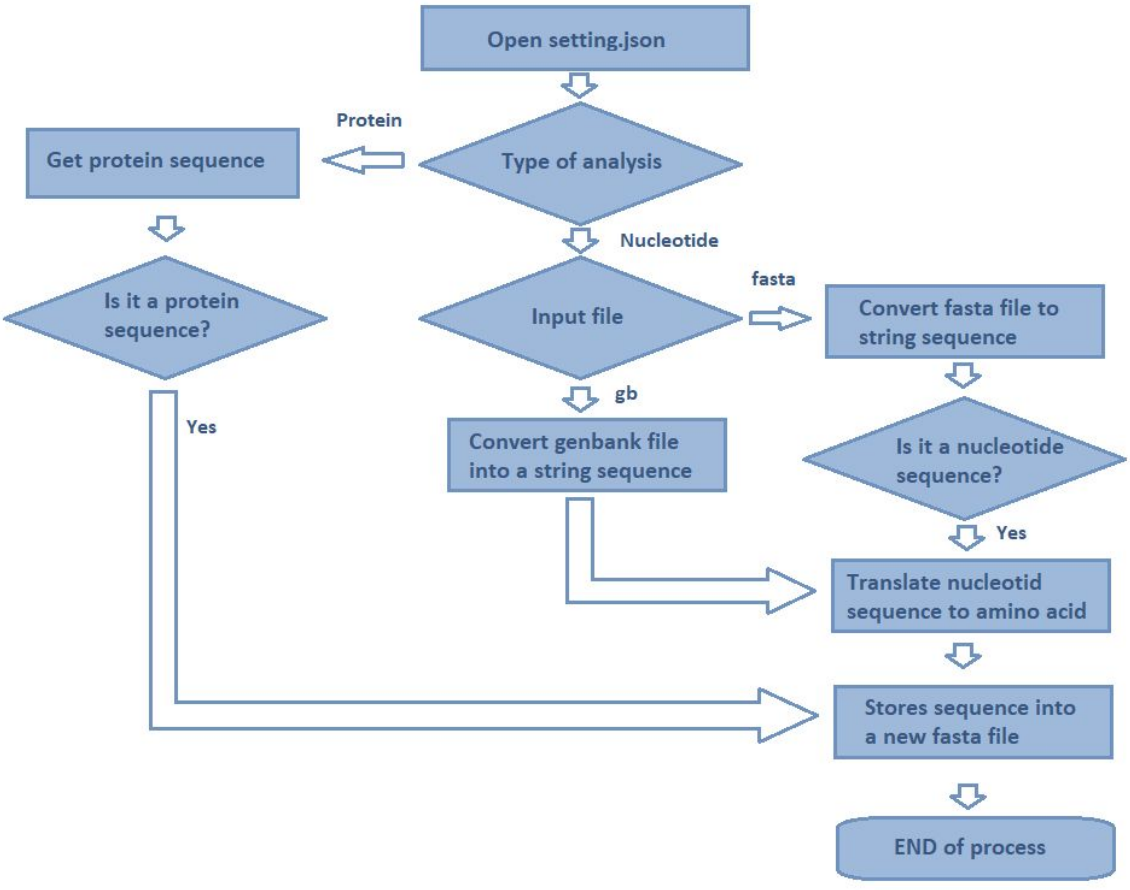

Figura Suplementària 2: Diagrama de flux de la fase de preprocessament del projecte en el que es detallen tots els possibles estats pels que pot passar el codi. Segons el tipus d'anàlisi seleccionat per l'usuari el procés es divideix en nucleòtids (esquerra) o proteïnes (dreta).

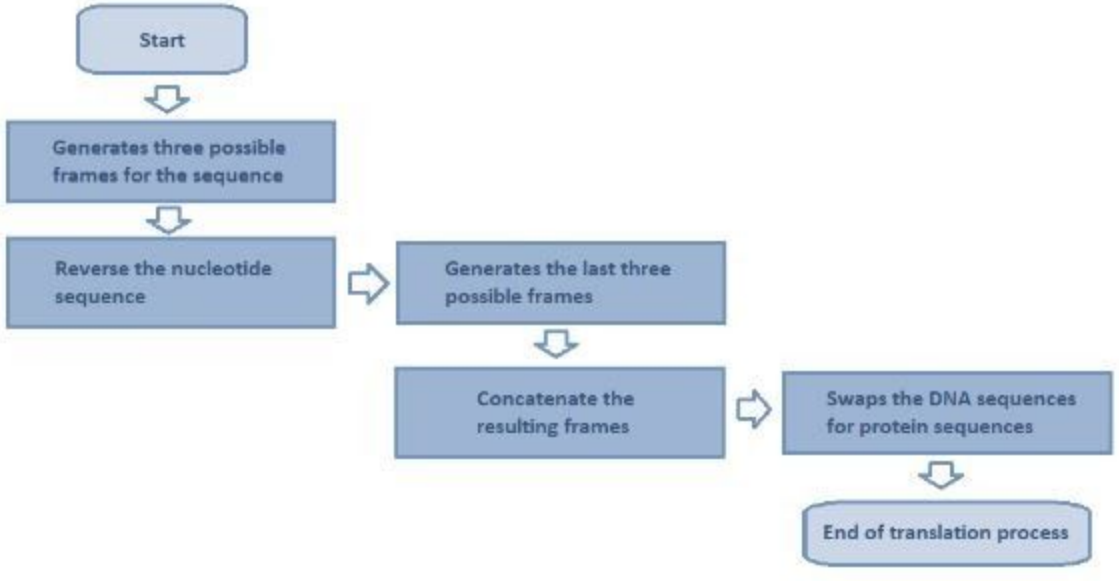

Figura Suplementària 3: Diagrama de flux que detalla els passos que segueix al realitzar el procés de traducció que prèviament s'havia indicat a la Figura 2

```
def bootstrap(dict):
    for key in dict.keys():
        length = len(dict[key])
        aux_list = []
        for i in range(0, length):
            position = randrange(length)
            aux_list.append(dict[key][position])
        dict[key] = aux_list
    return dict
```

Figura Suplementària 4: Fragment del codi en què es recull la funció encarregada de generar les rèpliques de *bootstrap.* Està format per un bucle exterior que calcula la longitud del vector de *coverage* i un bucle interior on es reorganitzen el contingut d'aquest de forma aleatòria.

```
def get_support_tree(original_tree, trees, output_folder):
    print("Support tree: ")
    tree_with_support = Consensus.get_support(original_tree, trees)
    Phylo.write(tree_with_support, output_folder + "/support_consensus_tree.nwk", "newick")
```

Figura Suplementària 5: Fragment del codi on es genera un arbre amb suport a les branques. Per fer-ho, es fa ús d'una funció proporcionada per la llibreria *Phylo* anomenada "get_support" a la que se li ha de passar com a paràmetre un arbre en format *newick*, sobre el que s'escriurà el suport calculat, i una llista d'arbres generat a partir de rèpliques de *bootstrap*. El resultat d'aquest procés es guarda a la carpeta de sortida del projecte.

```
def majority_consensus_tree(tree_list, cutoff):
    print("Majority consensus tree: ")
    majority_tree = Consensus.majority_consensus(tree_list, cutoff)
    Phylo.write(majority_tree, output_folder + "/majority_consensus_tree.nwk", "newick")
```

Figura Suplementària 6: Fragment del codi on es genera un arbre de consens segons les regles de la majoria. La llibreria *Phylo* ens proporciona una funció anomenada "majority_consensus" que, a partir d'una llista d'arbres, originats via *bootstrap,* i un llindar entre 0 i 1, permet calcular un arbre consensuat. L'arbre generat es desa, juntament amb la resta de resultats, a la carpeta de sortida del projecte.

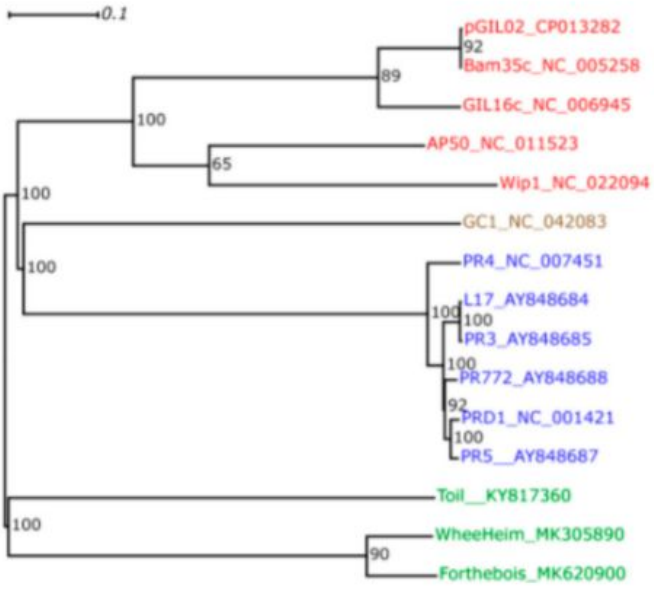

Figura Suplementària 7: Arbre filogenètic extret de l'article "A Novel Genus of Actinobacterial Tectiviridae" [17]. Els colors denoten els gèneres tectiviridae (Alphatectivirus, en blau; Betatectivirus, en vermell; Gammatectivirus, en marró) i Deltatectivirus (verd), el nou gènere que proposen a l'article.

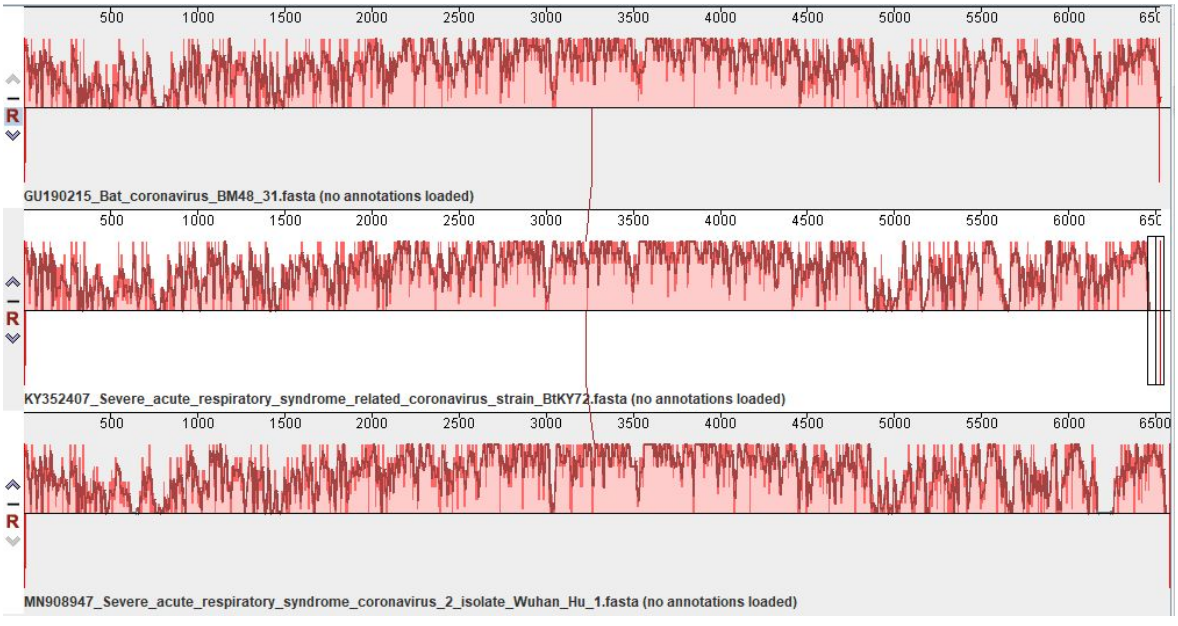

Figura Suplementària 8: Visualització de l'alineament de tres espècies de coronavirus mitjançant l'anàlisi "Prograssive Mauve" proporcionat per l'eina software amb el mateix nom. L'alineament s'ha realitzat a partir de deus soques dels SARS, concretament la de Wuhan i la de Kenya, juntament amb la seqüència "Bat coronavirus MB48".