Università degli Studi di Salerno

Dipartimento di Informatica

Corso di Laurea Magistrale in Informatica

Corso Strumenti Formali per la Bioinformatica

Costruzione di un albero filogenetico per le sequenze genomiche del virus Ebola

Candidati

Dott.ssa **Grazia Margarella**
Matricola: 0522501448

Dott. **Nicola Pio Santorsa**
Matricola: 0522501434

Anno Accademico 2022-2023

Sommario

La presente relazione ha l'obiettivo di descrivere il processo di lavoro nel costruire tramite diversi algoritmi degli alberi filogenetici, utilizzando come contesto applicativo i dati riguardanti l'epidemia del 2014 di virus Ebola.
A guidare l'analisi sono state utilizzate le domande fornite dall'assignment di Phillip Compeau [1] nel contesto del corso di *Fundamentals of Bioinformatics* del 2018 presso la *Carnegie Mellon University*.
In particolare ci si è focalizzati sulla comprensione di quale specie del virus Ebola sia precedente alle altre, quindi causa della diffusione dell'epidemia del 2014, come abbia infettato il paziente zero e come le specie del virus siano mutate nel tempo.
Nella relazione sono descritte le basi teoriche che riguardano sia l'ambito biologico che quello informatico necessarie per costruire un albero filogenetico e la sua successiva interpretazione.
Vengono descritti inoltre i dati biologici su cui si è svolta l'analisi e il loro conseguente preprocessing, in particolare focalizzandosi sui metodi di allineamento multiplo.
Il tool principale utilizzato per implementare tali algoritmi è MEGA11 di cui ne è riportata l'esecuzione e i risultati.
Per quanto riguarda gli approfondimenti biologici necessari si è utilizzato il testo di Campbell e Reece[2] e per gli approcci bioinformatici le risorse disponibili in rete del professor Compeau [1, 3, 4] e il libro *Understanding Bioinformatics* di Marketa Zvelebil e Jeremy O. Baum [5].

Indice

Glossario

Albero additivo Albero che descrive una Matrice additiva.

Albero binario non radicato Un Albero non radicato in cui ogni nodo ha grado 1 o 3.

Albero binario radicato Un Albero non radicato con una radice (di grado 2) su uno dei suoi archi.

Albero filogenetico Un albero filogenetico (o *cladogramma*, o *evolutionary tree*) è un diagramma in cui si rappresentano le relazioni evolutive tra sequenze e/o Specie. Tale definizione è approfondita nel capitolo 2.

Albero non radicato Un albero non radicato (*o unrooted-tree*) è un albero in cui non è indicato alcun nodo come radice.

Albero radicato Un albero radicato (o *rooted-tree*) è un albero in cui un nodo è candidato come radice.

Albero ultrametrico Un albero si dice ultrametrico (o *dendrogramma*) se è un albero additivo in cui le foglie e la radice sono equidistanti. Nell'analisi filogenetica viene utilizzato per determinare, in base al loro livello, un valore che indica l'età del singolo nodo, in base al meccanismo dell'Orologio molecolare. In questo caso le foglie hanno età pari a 0.

Algoritmo AdditivePhylogeny(D) Di seguito è descritto l'algoritmo:

1. Prendere un nodo foglia arbitrario j.
2. Calcolare il $LimbLength(j)$.
3. Sottrarre il $LimbLength(j)$ per ogni riga e colonna per produrre D^{bald} in cui j è un limb di lunghezza 0.
4. Rimuovere la j-esima riga e colonna della matrice per formare una matrice D^{trim} $(n-1)x(n-1)$.
5. Chiamare ricorsivamente il metodo $AdditivePhylogeny(D^{trim})$ per ottenere $Tree(D^{trim})$.
6. Identificare il punto nell'albero D^{trim} dove inserire la foglia j.
7. Attaccare j con un arco di lunghezza $LimbLength(j)$ per creare $Tree(D)$.

Per le matrici non additive si utilizza una versione approssimata di tale algoritmo con l'obiettivo di minimizzare la Discrepanza .

Allineamento Date due stringhe v e w, un allineamento di v e w è una matrice di due righe tale che la prima riga contiene i simboli di v e la seconda riga contiene i simboli di w. Ogni riga può anche contenere dei simboli di gap

−, ma una colonna non può avere due simboli di gap. Esso può essere sia un Allineamento globale che un Allineamento locale. Si utilizza inoltre una Matrice di score per definire quale sia l'allineamento migliore.

Allineamento globale Trovare l'allineamento di stringhe con lo score più alto su tutta la sequenza.

Allineamento locale Trovare le sottostringhe all'interno della sequenza che se allineate sono in grado di produrre il miglior allineamento globale.

Allineamento multiplo L'allineamento di tre o più sequenze (vedesi Capitolo 2).

Allineamento pairwise L'allineamento di una coppia di sequenze.

Discrepanza Misura della differenza tra un albero e una matrice non additiva. Definito come:
$Discrepancy(T, D) = \sum_{1 \leq i < j \leq n} (d_{i,j}(T) - D_{i,j})^2$.

Distanza di Hamming La distanza di Hamming di due strighe è il numero di mismatch tra i simboli delle stringhe.

FASTA Formato standard per rappresentare le sequenze biologiche. I file secondo questo formato hanno estensione *.fa* oppure *.fasta* e sono in plain text. Essi descrivono la sequenza primaria del genoma separata in righe di 60/80 bp e aggiungono delle informazioni addizionali come il cromosoma descritto, l'identificativo del genoma di riferimento e la sua localizzazione. Inoltre aggiunge un simbolo 1 se si tratta di una catena diretta di tipo 5'3' o -1 se di direzione 3'5'. Nato per il software FASTA che si occupa di allineamento locale.

Filogenesi character-based Metodo utilizzato per definire alberi filogenetici utilizzando proprietà anatomiche o fisiologiche dette caratteri. Possiamo però interpretare anche un allineamento di sequenze biologiche come carattere.

Filogenesi distance-based Data una matrice delle distanze, bisogna trovare un albero che rappresenti tale matrice. Definiamo $d_{i,j}(T)$ la distanza tra i nodi i e j nell'albero T calcolata sommando i pesi degli archi da i a j. Possiamo dire che l'albero T rappresenti la matrice D se per ogni coppia di i e j, $d_{i,j}(T) = D_{i,j}$.

LimbLength Si definisce limb (o *ramo*) l'arco da una foglia al suo genitore e *LimbLength(i)* è la lunghezza del limb per la foglia i.

Matrice additiva Una matrice si dice additiva quando è una matrice delle distanze per cui esiste un albero che la rappresenti. Una matrice può essere definita additiva se rispetta i criteri del *Four Point Theorem*, il quale dichiara che: una matrice delle distanze D è additiva se e solo la condizione dei quattro punti è verificata per ogni quadrupla (i, j, k, l). Tale condizione è verificata se due delle somme seguenti sono uguali e la terza è minore o uguale alla altre due somme:

- $D_{i,j} + D_{k,l}$
- $D_{i,k} + D_{j,l}$
- $D_{i,l} + D_{j,k}$

.

Matrice di distanza O *distance matrix*, è una matrice D che rappresenta le distanze tra coppie di n organismi che soddisfa tre proprietà:

- **Simmetria**: $D_{i,j} = D_{j,i}$ per tutte le coppie i, j;
- **Non negatività**: $D_{i,j} \geq 0$ per tutte le coppie i, j;
- **Disuguaglianza triangolare**: per tutti gli i, j, e k, $D_{i,j} + Dj, k \geq D_{i,k}$.

Dunque l'elemento $D_{i,j}$ rappresenta il numero di simboli differenti tra l'i-esima e la j-esima riga di un allineamento multiplo, secondo la Distanza di Hamming.

Matrice di score Una tabella di valori rappresentanti lo score da applicare quando si allineano due nucleotidi o aminoacidi. In generale lo score viene determinato per le quattro operazioni di Allineamento come match, mismatch, inserzione e delezione. Queste matrici vengono utilizzate per calcolare la qualità di un allineamento.

Motifs Sono una regione (una sottosequenza) di una proteina o di una sequenza di DNA che ha una specifica struttura e sono classificati come importanti per le funzionalità che svolgono.

Orologio molecolare Metodo per misurare gli intervalli temporali dei cambiamenti evolutivi in modo assoluto. Le modificazioni a carico delle sequenze di basi di alcune regioni del DNA si verificano a velocità abbastanza costanti; tale circostanza rende possibile datare episodi evolutivi avvenuti in passato.

Parsimony La parsimony (o *parsimonia*) è un principio per cui si indica come migliore un albero filogenetico con il numero più piccolo di mutazioni. Misurata tramite il Parsimony score.

Parsimony score Somma delle distanza di Hamming lungo ogni arco di un albero filogenetico.

Profile Dato un Allineamento multiplo di un insieme di stringhe, un profile è la frequenza per ogni carattere (anche il blank) per ogni colonna.

Sequenza genomica Sequenza di simboli sull'alfabeto a quattro lettere $\Sigma = \{A, T, G, C\}$ che rappresenta il sequenziamento di frammenti di DNA.

Sequenza proteica Sequenza di simboli sull'alfabeto a venti lettere
$\Gamma = \{A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y\}$
che indica il sequenziamento di una proteina, la quale è composta da 20 diversi aminoacidi uniti da legami peptidici.

Simple tree Si definisce simple tree un albero con nessun nodo di grado 2. Per quanto riguarda le matrici additive sussiste il seguente teorema: c'è un unico e solo simple tree che descrive una matrice additiva (ed esiste un algoritmo che lo produce).

Speciazione Formazione di una nuova Specie, quando una specie di divide in due. In un albero filogenetico si assume che ogni nodo interno corrisponda ad un evento di speciazione.

Specie Una specie biologica consiste in un gruppo di popolazioni i cui individui hanno la possibilità di incrociarsi tra loro generando prole vitale e fertile, mentre tali condizioni non possono verificarsi con membri di altre specie.

Specie vicine Due specie si dicono vicine quando condividono lo stesso genitore. Sussiste il teorema per cui ogni simple tree con almeno tre foglie ha almeno una coppia di foglie vicine.

Teorema del LimbLength Se D è una matrice additiva e j è un nodo foglia dell'albero $Tree(D)$, allora $LimbLength(j)$ è uguale al valore minimo di $(D_{i,j} + D_{j,k} - D_{i,k})/2$ tra tutte i nodi foglia j e k dell'albero $Tree(D)$.

Timetree Albero filogenetico scalato nel tempo.

Virus RNA I virus a RNA sono virus che utilizzano l'RNA come materiale genetico. Questo acido nucleico di solito è presente come filamento singolo, sebbene siano presenti gruppi di virus che utilizzano un RNA a doppio filamento..

1 Introduzione

1.1 Che cos'è l'Ebola

L'Ebola è un Virus RNA a filamento singolo negativo e prende il nome dal fiume omonimo vicino alle città centrafricane in cui è stato individuato la prima volta nel 1976.
Questo agente patogeno causa una grave forma di febbre emorragica.
Inizia con l'attaccare le cellule del sistema immunitario neutralizzandone la risposta e permettendo la proliferazione del virus.
Esso si trasmette tramite lo scambio di fluidi corporei e un individuo infetto risulta contagioso solo dopo la comparsa dei sintomi che si verificano tra i 2 e i 21 giorni dopo l'esposizione.
L'epidemia del 2014 è stata la più mortale di questo virus provocando un totale di 28.652 casi confermati con 11.325 decessi in dieci Paesi[6]. Il paziente zero è stato identificato in un bambino di 2 anni della Guinea. Nel corso dell'analisi proveremo a ricostruire come questo bambino si sia infettato.
A supporto di questo scopo vi sono gli alberi filogenetici.
Questo tipo di struttura dati permette di evidenziare le differenze evolutive tra diversi campioni biologici tramite il concetto di distanza.
In particolar modo gli alberi filogenetici permettono di descrivere delle relazioni di discendenza tra varie specie e dunque permettono di determinare in base ai nodi interni i progenitori comuni e in base alle ramificazioni i cambiamenti genetici tra una specie e l'altra.
L'analisi filogenetica dell'Ebola è dunque uno studio effettuato tramite la comparazione dei genomi delle diverse specie del virus e su come siano mutate nel tempo.
Nel corso dell'analisi utilizzeremo dunque le seguenti specie del virus Ebola:

- **Zaire** (EBOV);
- **Sudan** (SUDV);
- **Bundibugyo** (BDBV);
- **Tai Forest** (TAFV);
- **Reston** (RESTV): questa specie non provoca danni all'essere umano ed è una variante sviluppatasi nelle Filippine.

1.2 Ipotesi Iniziale

Una prima analisi iniziale può essere data dalla visualizzazione dei luoghi di sviluppo delle diverse specie.
Nella seguente figura, i pin indicano da sinistra a destra la **Guinea**, ossia l'epicentro dell'epidemia del 2014, **Tai Forest**, **Zaire**, **Bundibugyo** e **Sudan**.

Da questa prima visualizzazione, una prima ipotesi formulabile è che la specie di Ebola scatenenante l'epidemia sia stata la **Tai Forest (TAFV)**, dal momento che è la più vicina all'epicentro.

1.3 Struttura della Relazione

Nel capitolo 2 verranno descritti gli algoritmi necessari allo svolgimento dell'analisi, in particolar modo ci si concentrerà sui metodi di allineamento multiplo utilizzati, ossia Clustal W e MUSCLE, e sugli algoritmi di costruzione degli alberi filogenetici, come il Neighbor-Joining Algorithm, l'UPGMA e il Maximum Parsimony.
Nel capitolo 3 verranno riportati gli alberi risultanti dagli algoritmi citati e il processo necessario per la loro costruzione tramite il tool MEGA11.
Nel capitolo 4 verranno analizzati, ispezionati e commentati i risultati ottenuti dalla fase precedente, ed infine nel capitolo 5 verranno riportate le conclusioni tratte dall'analisi.

2 Stato dell'Arte

In questo capitolo sono descritti alcuni degli algoritmi utilizzati per la costruzione degli alberi filogenetici. In prima battuta ci si è focalizzati sul preprocessing delle sequenze genomiche da analizzare, ossia sul loro allineamento multiplo tramite gli algoritmi utilizzati dal tool di riferimento MEGA11. In seguito sono stati descritti gli algoritmi di costruzione degli alberi filogenetici.

2.1 Multiple Alignment

Multiple Sequence Alignment (MSA) è la tecnica di allineamento di tre o più sequenze biologiche (che possono essere proteiche o nucleotidiche) di lunghezza simile.
Questa tecnica rivela molte informazioni biologiche in più rispetto al Allineamento pairwise e permette di :

- Identificare la conservazione di pattern di sequenze e Motifs in una famiglia di sequenze proteiche e nucleotidiche;
- Identificare la conservazione di strutture e funzionalità importanti in residui di amminoacidi nelle proteine;
- Mostrare relazioni evolutive di proteine e geni.

Dall'esecuzione di un algoritmo di Allineamento multiplo viene restituita una struttura che evidenzia le eguaglianze e diseguaglianze tra le sequenze prese in esame con un relativo score che indica la loro somiglianza. L'obiettivo degli algoritmi di questo tipo è quello di ottimizzare questo valore.
Esistono diversi algoritmi che risolvono questa problematica, però la grande sfida ancor oggi aperta è sullo sviluppo di una soluzione con costo computazionale basso considerando un grande numero di sequenze molto lunghe.
Tra gli algoritmi presenti ne spiccano principalmente due che risultano essere i più utilizzati e con tempi di attesa ragionevoli, questi sono il **ClustalW** e il **MUSCLE**.

2.1.1 ClustalW

L'algoritmo *ClustalW* è un algoritmo che segue un approccio euristico ed effettua Allineamento globale di sequenze.
La prima implementazione dell'algoritmo avvenuta nel 1994, mostrata nel seguente paper [7] è stata realizzata in linguaggio C.
Le matrici di score che utilizza sono *PAM* e *BLOSUM*. L'algoritmo presenta diverse varianti, le quali sono esaustivamente descritte in [8].
Tra le particolarità del seguente algoritmo troviamo che :

- Applica diverse matrici di score mentre si effettua l'allineamento in base al grado di similarità delle sequenze;
- Applica uno schema pesato per incrementare l'affidabilità dell'allineamento di sequenze divergenti (il fattore di peso per ogni sequenza viene determinato dalla lunghezza del ramo dell'albero guida);

- Applica variando "gap open penalty" e "gap extension penalty" per evitare gap non necessari in regioni conservative.

L'esecuzione dell'algoritmo è caratterizzata dai seguenti passi :

1. Inizialmente effettua un *Pairwise Alignment* di tutte le sequenze tra di loro utilizzando l'algoritmo di *Needleman-Wunsch* e calcola gli score di similarità;

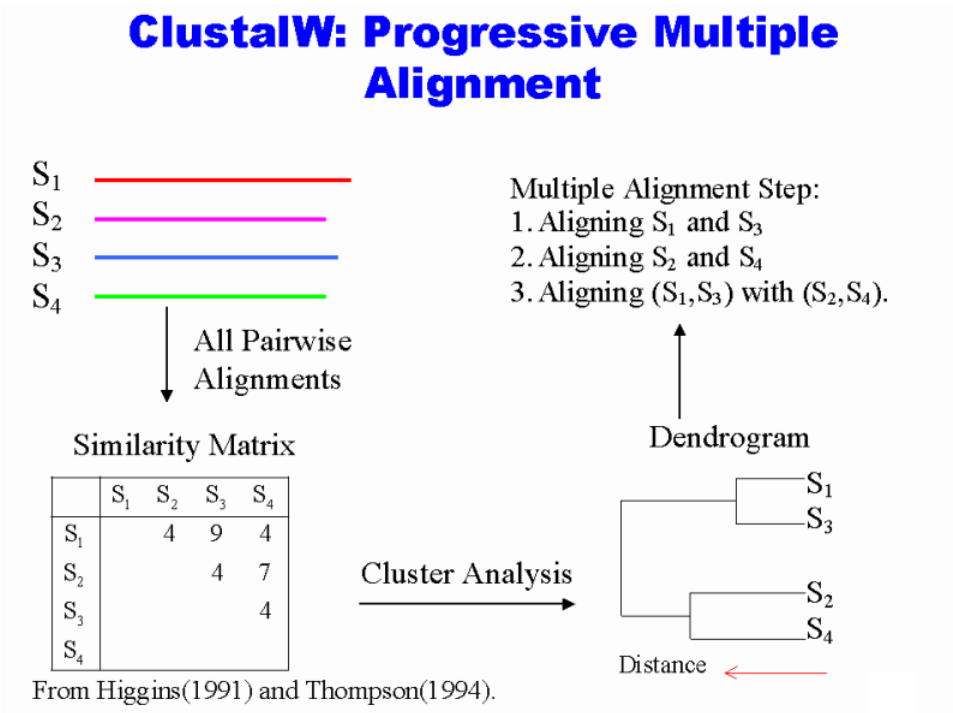
2. Dagli score di similarità vengono calcolate le distanze e da queste viene creata una matrice delle distanze triangolare;

3. Basandosi sulle distanze si crea un albero guida usando algoritmi come il *Neighbour Joining* o *UPGMA*(si vedranno più avanti).
   Nell'albero creato la lunghezza del ramo è proporzionale alla divergenza stimata lungo ogni ramo;

4. Dall'albero vengono assegnati pesi per ogni sequenza dipendentemente dalla distanza dalla radice dell'albero: Sequenze più vicine riportano un peso basso, mentre sequenze divergenti un peso alto;

5. Procedendo dalla punta dell'albero verso la radice, sulle sequenze viene effettuato l'Allineamento pairwise in modo progressivo utilizzando un algoritmo di programmazione dinamica globale;

6. Durante l'alignment lo score per ogni posizione residua è basata sulla variazione della matrice dei pesi che è scelta basata sulla similarità delle sequenze allineate.
   I pesi delle sequenze e posizioni specificano la penalità dei gap che sono basati sulla Matrice di score, similarità delle sequenze e la lunghezza delle stesse.



Uno svantaggio del seguente metodo è che non è adatto per il confronto di sequenze che si differenziano di molto per la loro lunghezza poichè il metodo è di tipo globale. Le complessità del seguente algoritmo, come vengono anche mostrate nel paper [9], sono le seguenti:

Distance Matrix	$O(N^2L^2)$
Neighbor Joining	$O(N^4)$
Progressive Alignment	$O(N^3 + NL^2)$
Total	$O(N^4 + L^2)$

Tabella 1: Tabella Complessità Temporale dell'algoritmo

Distance Matrix	$O(N^2 + L)$
Neighbor Joining	$O(N^2)$
Progressive Alignment	$O(NL + L^2)$
Total	$O(N^2 + L^2)$

Tabella 2: Tabella Complessità Spaziale dell'algoritmo

2.1.2 MUSCLE

Un altro metodo di allineamento molto usato è il **MUSCLE**, presentato nei seguenti paper [9, 10], che si differenzia dal precedente per un costo computazionale minore e per la possibilità di interrompere l'algoritmo prima del termine dell'esecuzione. L'allineamento ottenuto risulta essere meno corretto rispetto a quello restituito al termine dell'esecuzione, ma comunque accettabile.
Un'implementazione del seguente algoritmo in linguaggio *C++* è visibile al seguente Link. Il seguente algoritmo utilizza due misure di distanza per effettuare l'allineamento pairwise e sono :

- **Distanza kmer** : Questa distanza non richiede che le sequenze siano allineate, e si basa sul concetto di *kmer*, il quale indica una sequenza contigua di simboli di lunghezza k.
  Il fatto che questa distanza non richiede nessun tipo di allineamento rende l'esecuzione molto veloce;

- **Distanza di Kimura** : Questa distanza è la più utilizzata in tutti i modelli di sostituzione nucleotidica per stimare le differenze genetiche ed è uno tra i modelli più precisi esistenti.[1]

Questo algoritmo a differenza di altri, preferisce l'utilizzo del UPGMA rispetto al Neighbor Joining poichè gli autori del paper [10] hanno notato una correttezza nei risultati migliore utilizzando il primo metodo rispetto al secondo.
Qui di seguito si descrive l'esecuzione dell'algoritmo suddiviso in 3 stage :
Stage 1 : L'obiettivo di questo stage è di produrre un Allineamento multiplo preferendo sulla velocità piuttosto che sull'accuratezza dello stesso.

[1] Per approfondire : Articolo Distanza Kimura

- 1.1 : Viene calcolata la distanza kmer per ogni paia di sequenze in input restituendo poi la matrice di distanza D1;
- 1.2 : La matrice D1 viene clusterizzata tramite algoritmo UPGMA producendo l'albero binario TREE1;
- 1.3 : Viene costruito un allineamento progressivo seguendo l'ordine dei rami dell'albero. Ad ogni foglia viene costruito un profile da una sequenza in input. I nodi dell'albero sono visitati in prefix order e ad ogni nodo interno viene costruito un pairwise alignment a partire da due profile.Ne verrà creato uno nuovo che verrà assegnato al nodo.

Alla fine di questo stage il primo allineamento MSA1 è disponibile e l'utente può interrompere l'esecuzione.

Stage 2 : La causa principale degli errori nel primo alignement è basata sull'approssimazione dei kmer. Per questo motivo si applica la distanza di Kimura sull'allineamento precedente per effettuare una nuova stima dell'albero.
Questa distanza risulta essere più accurata rispetto alla precedente ma richiede che le sequenze siano allineate.

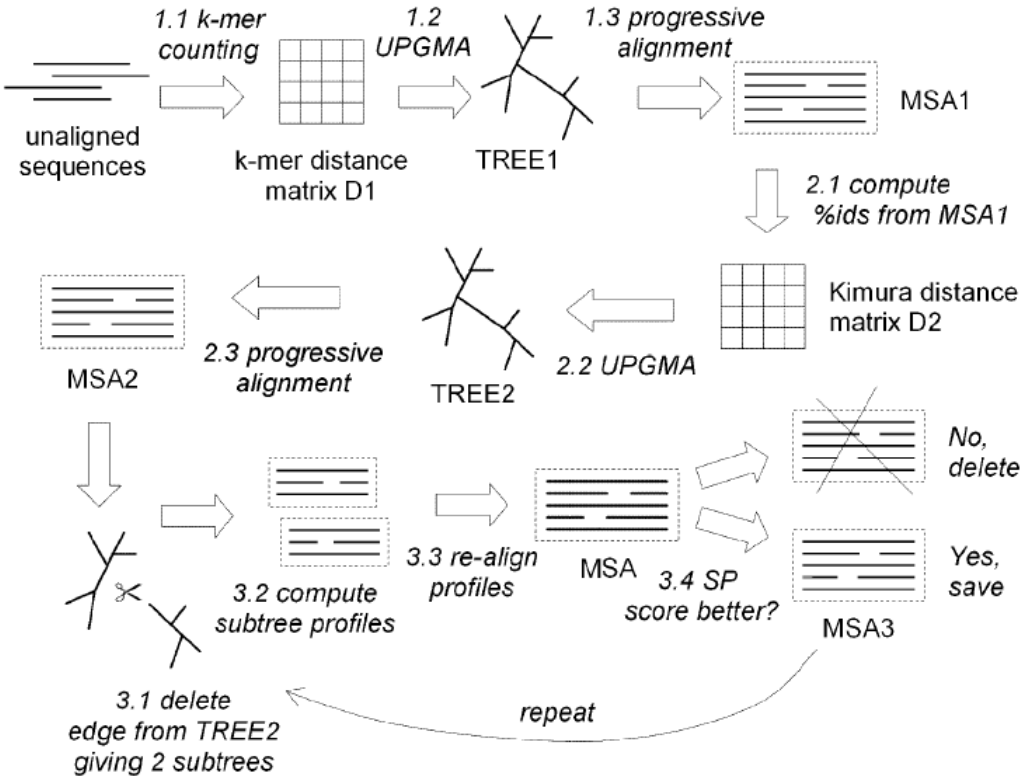
- 2.1 Viene calcolata la distanza di Kimura per ogni paia di sequenze in input prese dal MSA1.Restituirà la matrice di distanza D2;
- 2.2 La matrice D2 viene clusterizzata tramite algoritmo UPGMA producendo l'albero binario TREE2;
- 2.3 Si effettua un allineamento progressivo seguendo TREE2, producendo l'allineamento multiplo MSA2.
  Questo allineamento risulta ottimizzato poichè vengono calcolati solo gli allineamenti dei sottoalberi il cui ordine dei rami è cambiato rispetto al TREE1.

Alla fine di questo stage il secondo allineamento MSA2 è disponibile e l'utente può interrompere l'esecuzione.

Stage 3 : Raffinamento

- 3.1 Viene scelto un arco casuale del TREE2;
- 3.2 TREE2 viene suddiviso in due sottoalberi cancellando l'arco e viene calcolato il profile del multiple alignment di ogni sottoalbero;
- 3.3 Viene ricalcolato un nuovo multiple alignment riallineando i due profili;
- 3.4 Se lo score è migliorato si mantiene il nuovo allineamento, altrimenti si scarta.

Gli step da 3.1 a 3.4 vengono ripetuti finchè non si raggiunge la convergenza o un limite definito dall'utente.



Muscle-p 1 e 2 Stage	$O(N^2L + NL^2)$
Muscle	$O(N^2L + NL^2 + N^3L)$

Tabella 3: Tabella delle Complessità Temporali MUSCLE

Muscle-p 1 e 2 Stage	$O(N^2 + NL + L^2)$
Muscle	$O(N^2 + NL + L^2)$

Tabella 4: Tabella delle Complessità Spaziali MUSCLE

2.2 Phylogenetic Trees

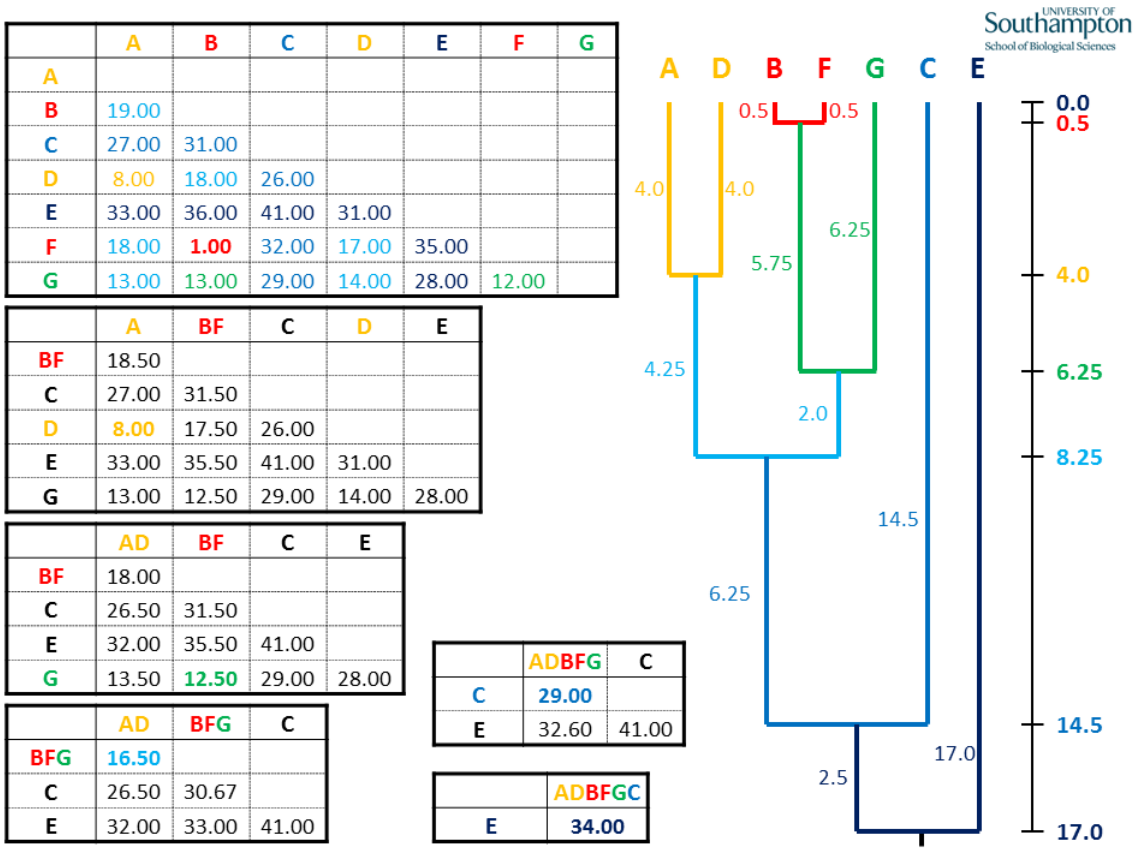
Un **albero filogenetico** è un diagramma che rappresenta le linee di discendenza evolutiva di diverse specie, organismi o geni da un antenato comune.
Questi sono utili per organizzare la conoscenza della diversità biologica, per strutturare le classificazioni e per fornire informazioni sugli eventi che si sono verificati durante l'evoluzione.
Un albero di questo tipo può essere di tipo *rooted* oppure *unrooted*. In un Albero radicato, ogni nodo con discendenti rappresenta l'antenato comune più recente di

quelli, e le lunghezze degli archi in alcuni alberi possono essere interpretate come stime temporali. Ogni nodo è chiamato unità tassonomica. I nodi interni sono generalmente chiamati unità tassonomiche ipotetiche, in quanto non possono essere osservati direttamente.
Un Albero non radicato illustra solo la parentela dei nodi foglia e non richiede che la radice sia nota o dedotta.
Un albero filogenetico può essere un Albero additivo oppure non additivo, a seconda del tipo di matrice di distanza utilizzata.
La maggior parte delle matrici di distanza generate risultano essere non additive, motivo per cui nella seguente relazione si tratteranno approfonditamente algoritmi per matrici non additive.
Se si intende approfondire ulteriormente le matrici additive, si consiglia l'Algoritmo AdditivePhylogeny(D).

2.2.1 UPGMA

L'**UPGMA**, anche detto *Unweighted Pair Group Method with Arithmetic Mean*, è un metodo agglomerativo gerarchico di clustering. La sua invenzione viene attribuita a Sokal e Michener.
Questa euristica viene utilizzata in alcuni algoritmi di multiple alignment (come abbiamo visto in Clustal e MUSCLE) e nell'ambito filogenetico per la creazione di un Albero ultrametrico.
Il seguente algoritmo prende in input una matrice di distanza(accetta sia matrici triangolari che complete) e da essa effettua il procedimento che vediamo qui di seguito:

- Crea un cluster per ogni sequenza presente nella matrice di distanza. Ogni cluster rappresenterà una foglia dell'albero;
- Cerca i due cluster più vicini C1 e C2 e calcola la distanza media
  $D_{avg}(C1, C2) = \sum_{\text{i in C1, j in C2}} \frac{D_{i,j}}{|C1| \cdot |C2|}$
  con $|C|$ che denota il numero di elementi nel cluster C.
- Si effettua il merge dei due cluster C1 e C2 in un cluster C;
- Si crea un nuovo nodo C e si connette ad esso C1 e C2, i due cluster da cui è composto e infine si imposta l'età di C come $D_{avg}(C1, C2)/2$;
- Si aggiorna la matrice di distanza ricalcolando la distanza tra ogni cluster;
- L'algoritmo viene iterato fintanto non si avrà un singolo cluster che conterrà tutti gli elementi.



	A	B	C	D	E	F	G
A							
B	19.00						
C	27.00	31.00					
D	8.00	18.00	26.00				
E	33.00	36.00	41.00	31.00			
F	18.00	**1.00**	32.00	17.00	35.00		
G	13.00	13.00	29.00	14.00	28.00	12.00	

	A	BF	C	D	E
BF	18.50				
C	27.00	31.50			
D	**8.00**	17.50	26.00		
E	33.00	35.50	41.00	31.00	
G	13.00	12.50	29.00	14.00	28.00

	AD	BF	C	E
BF	18.00			
C	26.50	31.50		
E	32.00	35.50	41.00	
G	13.50	**12.50**	29.00	28.00

	AD	BFG	C
BFG	**16.50**		
C	26.50	30.67	
E	32.00	33.00	41.00

	ADBFG	C
C	**29.00**	
E	32.60	41.00

	ADBFGC
E	**34.00**

L'immagine precedentemente mostrata è stata presa dal seguente sito[2] dove se si vuole si può approfondire l'algoritmo.
Uno svantaggio del seguente algoritmo è che non tutte le distanze della matrice sono rispettate nell'albero. La seguente problematica viene ignorata il più delle volte. Nonostante ciò questo algoritmo resta uno tra i più utilizzati nell'ambito bioinformatico oltre all'algoritmo Neighbor Joining.

UPGMA	O(N^2)

Tabella 5: Complessità Temporale UPGMA

UPGMA	O(N^2)

Tabella 6: Complessità Spaziale UPGMA

2.2.2 Neighbor Joining

Anche questo algoritmo, come il precedente viene utilizzato sia in ambito Multiple Alignment che in ambito filogenetico per la creazione di un albero a partire da una matrice di distanza.

[2] Esempio UPGMA

L'ideazione di questo algoritmo è stata talmente importante che ancora oggi il paper[11] risulta essere tra i top 20 paper più citati in tutti i campi scientifici.

Questo algoritmo utilizza una matrice denominata Neighbor-Joining Matrix, la quale viene costruita al primo passo a partire da una matrice di distanza D.
Il valore di ogni cella della seguente matrice NJM sarà calcolato come :

$$D^*_{i,j} = (n-2) \cdot D_{i,j} - TotalDistance(i) - TotalDistance(j)$$

con TotalDistance*(i)* metodo che restituisce la somma delle distanze da i a tutte le foglie.
Se la matrice delle distanze D iniziale è additiva, allora il più piccolo valore di D^* corrisponde a due elementi della tabella che risultano essere vicini nell'albero della matrice D.
Vediamo adesso i passi di esecuzione del seguente algoritmo:

1. Si crea la matrice D^* a partire dalla matrice di distanza passata in input, utilizzando la formula sopra descritta;

2. Si cerca l'elemento minore nella matrice delle distanze D^* e indicheremo questo valore con la nomenclatura $d^*_{i,j}$;

3. Calcoliamo il delta della coppia (i, j) identificata come minima utilizzando la seguente formula :
   $Delta_{i,j}$ = (TotalDistance(i) - TotalDistance(j)) / (n-2) dove n è il numero di sequenze della matrice.

4. Si impone :
   $\text{LimbLength(i)} = \dfrac{D_{i,j} + Delta_{ij}}{2}$
   $\text{LimbLength(j)} = \dfrac{D_{i,j} - Delta_{ij}}{2}$
   N.B. Questa formula funziona ottimamente su matrici non additive.

5. Si cancellano le righe e colonne i e j e le si sostituiscono con il nodo merge m. Ricostruiamo quindi una nuova matrice D' con le stesse righe e colonne di prima, meno i e j, aggiungendo però come riga e colonna m.
   I nuovi valori saranno calcolati rispetto all' elemento k-esimo come :
   $D'_{k,m} = (D_{i,k} + D_{j,k} - D_{i,j})/2$

6. Applichiamo l'algoritmo di Neighbor Joining a D' per ottenere l'albero di D';

7. Riattacchiamo i e j all'albero ottenuto nello scorso step, utilizzando i valori calcolati precedentemente come distanza, per ottenere l'albero di D.

Neighbor Joining	O(N^3)

Tabella 7: Complessità Temporale Neighbor Joining

Neighbor Joining	O(N^2)

Tabella 8: Complessità Spaziale Neighbor Joining

2.2.3 Maximum Parsimony Problem

In filogenetica, il Maximum Parsimony è un criterio di ottimalità in base al quale viene preferito l'albero filogenetico che minimizza il Parsimony score.
In altre parole, in base a questo criterio, l'albero più piccolo possibile che spiega i dati è considerato il migliore.
Alcune delle idee alla base del Maximum Parsimony sono state presentate da James S. Farris nel 1970 and Walter M. Fitch nel 1971.
Esistono diversi approcci volti alla creazione dei vari alberi filogenetici, che vengono poi analizzati dal criterio presentato, e due di questi vengono descritti brevemente qui di seguito.

Exhaustive Search

Il seguente metodo costruisce tutti i possibili alberi e successivamente calcola tutti i parsimony score.
E' facilmente intuibile che il seguente metodo risulta impraticabile per risolvere problemi con più di 10 sequenze ma comunque una soluzione accessibile se non si supera questo limite superiore.

Branch And Bound Algorithm

Questa risulta essere una strategia volta ad evitare il calcolo di tutti i possibili alberi con il metodo visto in precedenza.
Non risulta essere l'approccio ottimo per risolvere questa problematica dato che nel caso peggiore l'esecuzione di questo algoritmo non garantisce una complessità di tempo polinomiale.
L'algoritmo genera un albero iniziale T_0 randomicamente e utilizza il suo score come limite superiore. Costruisce in seguito un nuovo albero T con 2 sequenze e aggiunge successivamente le altre sequenze iterativamente su tutti i possibili rami creando nuovi alberi che vengono mantenuti solamente se il loro score è minore di quello imposto come limite superiore. Alla fine dell'esecuzione si applica il maximum parsimony e si seleziona quindi il tree migliore.
Il fatto che il Maximum Parsimony sia un problema combinatoriamente complesso implica che l'esecuzione del seguente algoritmo risulta essere impraticabile per un numero di sequenze superiore a 25 per motivi di efficienza e memoria.

Altri metodi

Sono numerosi gli algoritmi sviluppati ed introdotti per l'applicazione del criterio del Maximum Parsimony.
Se si intende approfondire ulteriormente tale argomento, si consiglia la tesi di dottorato citata [12] in cui si approfondiscono tutti i metodi avanzati per risolvere il problema del Maximum Parsimony.

3 Costruzione Tramite Tool MEGA

In questo capitolo vengono descritti i risultati ottenuti tramite il tool MEGA e la loro esecuzione.
La versione di MEGA che è stata utilizzata per lo svolgimento della seguente relazione è MEGA11 ed è scaricabile gratuitamente dal seguente sito.

3.1 Specifiche Hardware del Dispositivo Utilizzato

Il software MEGA è stato eseguito su un computer con le seguenti componenti hardware :

- CPU : Intel(R) Core(TM) i5-8600K CPU 4.00 GHz
- RAM : 16 GB DDR4
- Scheda Video : Radeon (TM) RX 480

Quindi le prestazioni che verranno indicate sugli algoritmi di alignment saranno quindi riferite al dispositivo qui sopra descritto.

3.2 Primo Passo: Multiple Alignment

3.2.1 Sequenze Prese in Studio

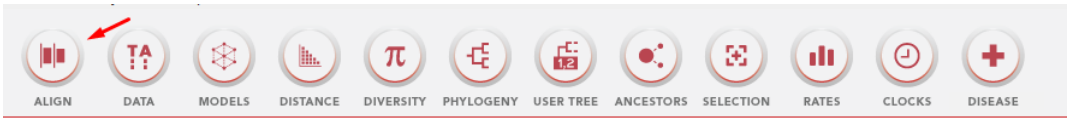
Le sequenze che andremo ad allineare sono riportate nella seguente tabella.
La loro lunghezza si assesta in un range di 18,875 – 18,959 coppie di basi (bp).

Accession Number	Virus Species	Location	Date
KJ660348	????	Gueckedou, Guinea	2014
FJ217161	Bundibugyo (BDBV)	Bundinbugyo, Uganda	2007
KC545393	Bundibugyo (BDBV)	Isiro, DRC	2012
AF272001	Zaire (EBOV)	Yambuku, DRC	1976
KC242792	Zaire (EBOV)	Mekouka, Gabon	1994
KC589025	Sudan (SUDV)	Luwero, Uganda	2012
FJ968794	Sudan (SUDV)	Sudan	1976
FJ217162	Tai Forest (TAFV)	Tai Forest, Ivory Coast	1994
AF522874	Reston (RESTV)	Philippines	1990
FJ621583	Reston (RESTV)	Philippines	2008

Tabella 9: Sequenze Ebola

3.2.2 Finestra di Allineamento MEGA

All'apertura del programma si nota immediatamente la presenza di una barra superiore mostrata qui in figura, per poter accedere alla finestra di allineamento di MEGA, dobbiamo cliccare sul primo elemento della barra evidenziato da una freccia rossa.



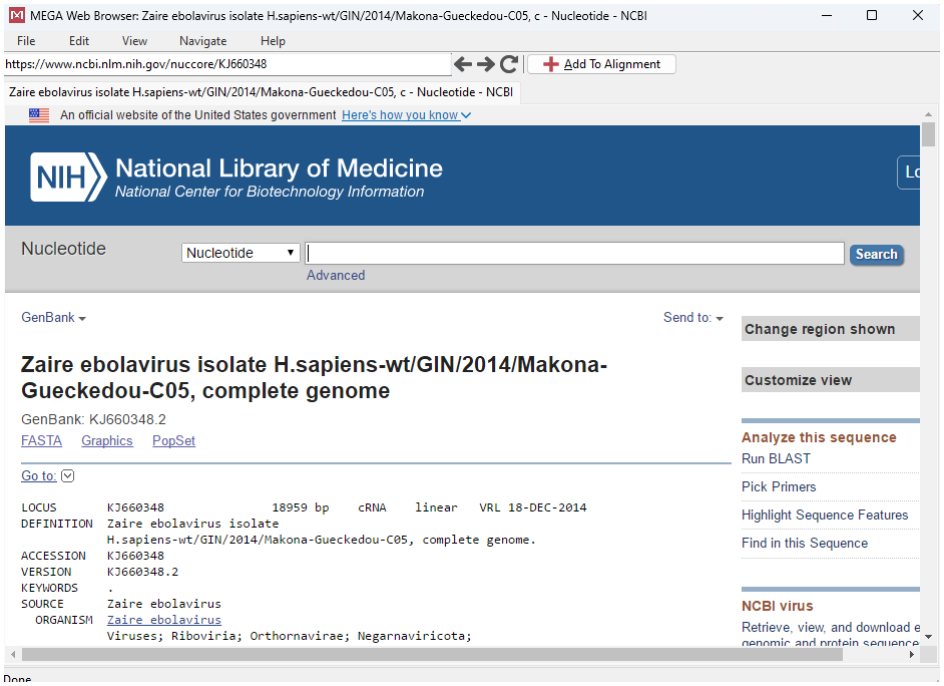
Al click si presenterà un menu a tendina, clicchiamo il primo item del menu con scritto **Edit/Build Alignment**.
Successivamente si aprirà una finestra che ci chiederà se vogliamo creare un nuovo allineamento o aprire una sessione già avviata, selezioniamo la prima opzione.
Infine specifichiamo se l'allineamento a cui siamo interessati è di sequenze di DNA o Proteine.
Al termine di queste operazioni la finestra per effettuare allineamento tramite il software MEGA sarà aperta e funzionante.
L'operazione successiva sarà quella di caricare le sequenze da allineare, abbiamo due opzioni :
Il caricamento da file o tramite GenBank.
Vedremo entrambi gli approcci nei paragrafi che seguono.

3.2.3 Caricamento Tramite File Delle Sequenze

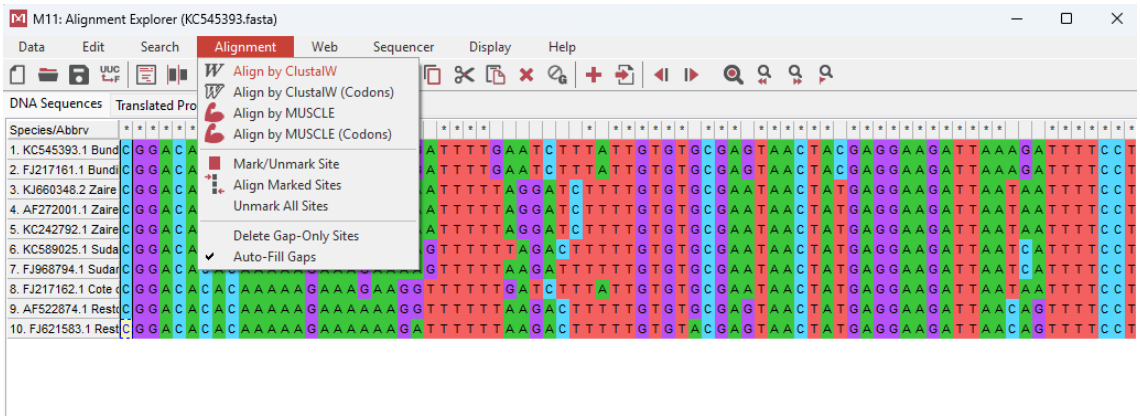
Nella pagina di allineamento di MEGA, selezioniamo nella barra superiore la sezione **Edit** e da questa selezioniamo l'item **Insert Sequence From File**.
A questo punto il sistema operativo in uso aprirà una finestra File Chooser che permetterà all'utente di selezionare le sequenze scaricate nei vari formati, in particolar modo quelli in formato FASTA.

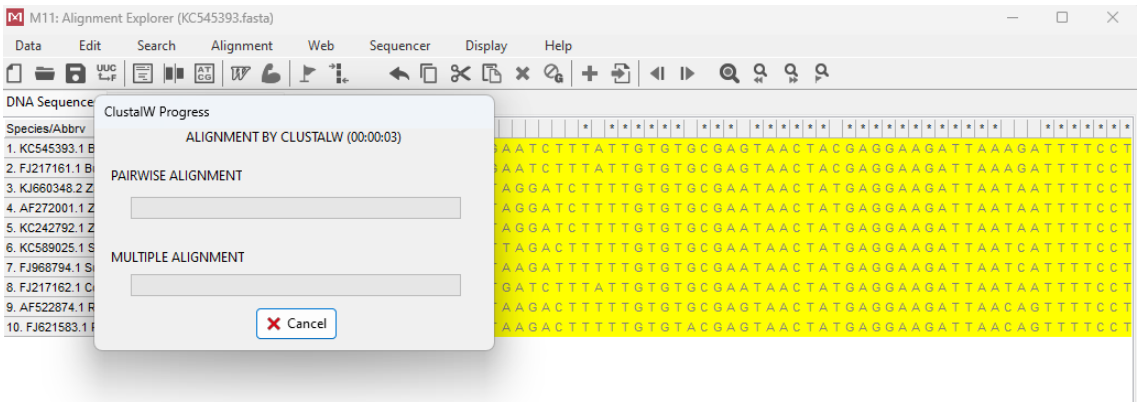
3.2.4 Caricamento Tramite GenBank Delle Sequenze

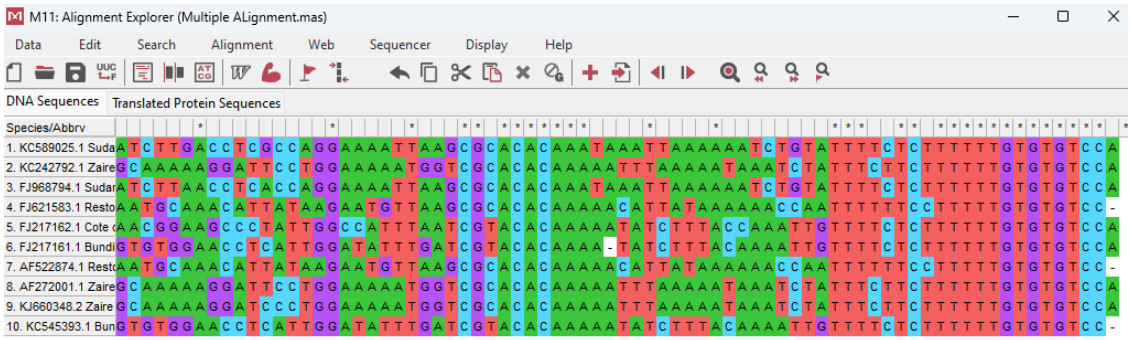
Nella pagina di allineamento di MEGA, selezioniamo nella barra superiore la sezione **bank**, e da questa selezioniamo l'item **Query GenBank**.
Dopo che la pagina si è caricata utilizziamo la barra di ricerca del sito per cercare le sequenze che ci interessano, nel nostro caso specifico utilizziamo i codici descritti in tabella sotto la colonna *Accession Number*.
Trovata la sequenza desiderata sul sito, clickiamo sulla scheda il pulsante **Add to Alignment**, selezioniamo **Import all Sequence**, scegliamo le preferenze sul labeling della sequenza e clickiamo ok.
La sequenza verrà quindi aggiunta alle sequenze da allineare.



3.2.5 Allineamento con ClustalW

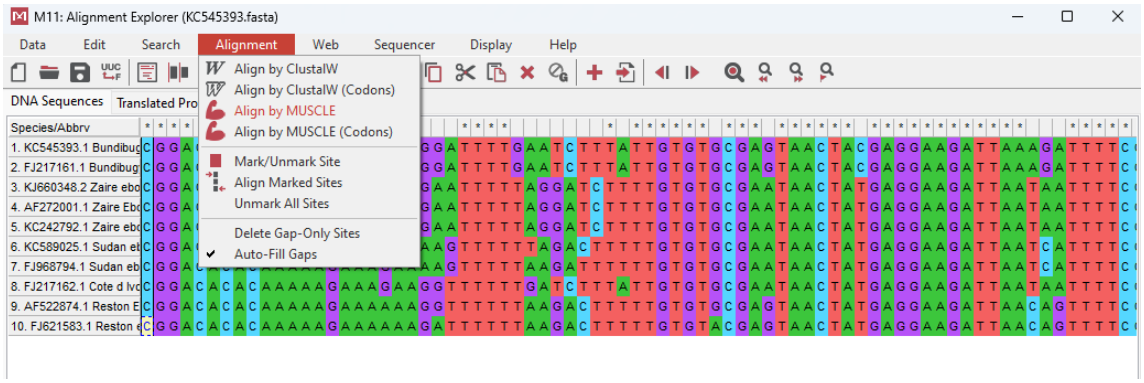


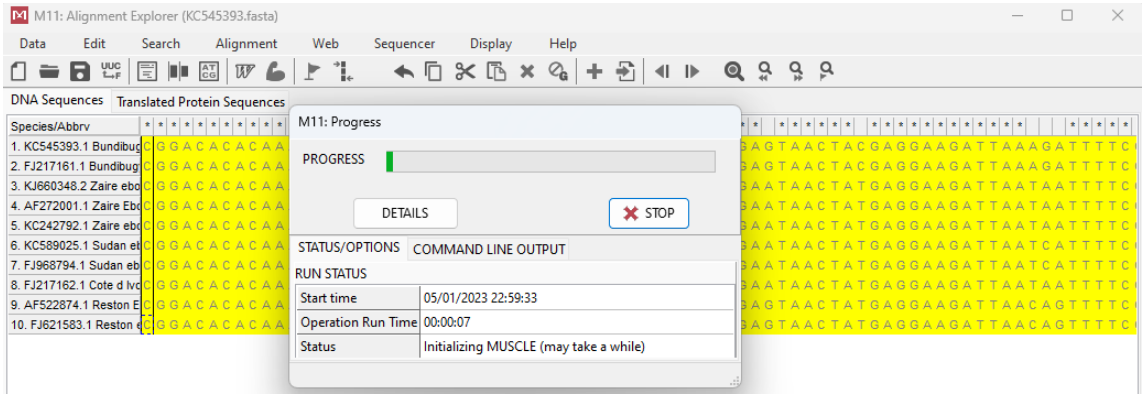


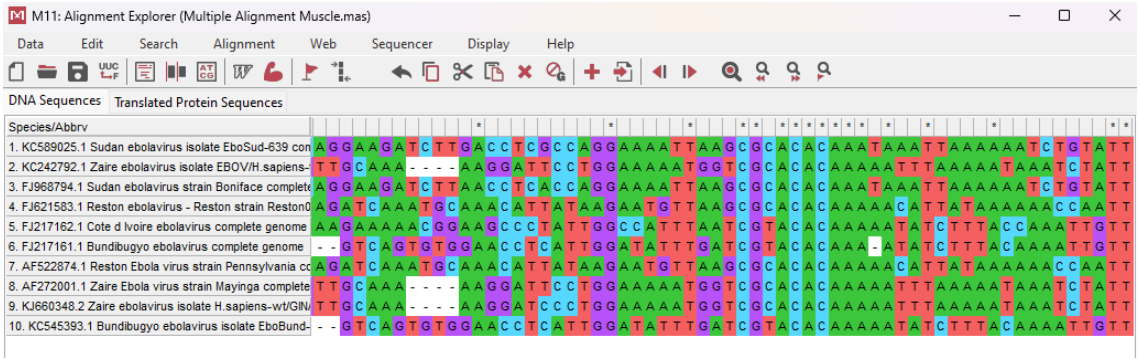


Al termine dell'allineamento salvare la sessione di allineamento tramite il menu di mega.
Questo file salvato verrà poi utilizzato per effettuare la creazione degli alberi filogenetici.
Tempo di esecuzione : 25 minuti circa.

3.2.6 Allineamento con MUSCLE







Al termine dell'allineamento salvare sessione di allinamento tramite il menu di mega. Questo file salvato verrà poi utilizzato per effettuare la creazione degli alberi filogenetici.
Tempo di esecuzione : 20 minuti circa.

3.2.7 Confronto tra i due allineamenti

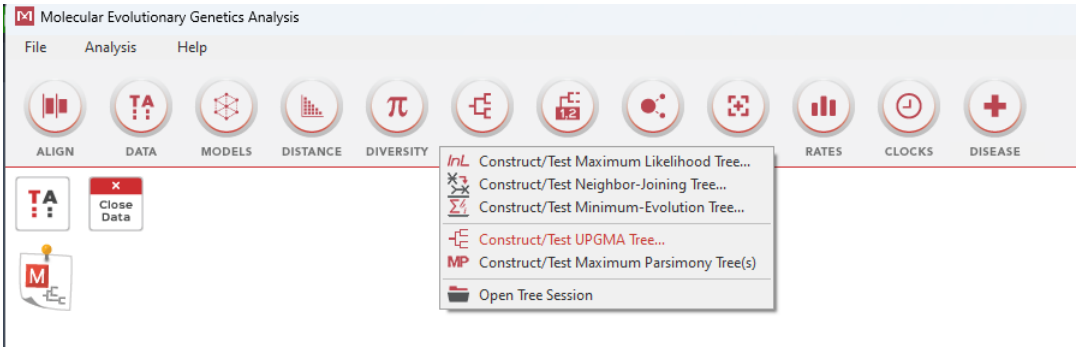
Entrambi hanno tempi di esecuzione molto alti dovuti non tanto al numero di sequenze ma più alla loro lunghezza che risulta essere alta.
Nonostante ciò i tempi sono stati minori di quanto ci aspettavamo, con MUSCLE che ha avuto un esecuzione più veloce rispetto a quella di Clustal.

L'algoritmo di Multiple Alignment che utilizzeremo come riferimento alle analisi seguenti per lo sviluppo di alberi filogenetici è il ClustalW.
Questa scelta è stata fatta poichè il seguente algoritmo permette la creazione di un albero filogenetico, rimuovendo lo step intermedio della creazione della tabella delle distanze.

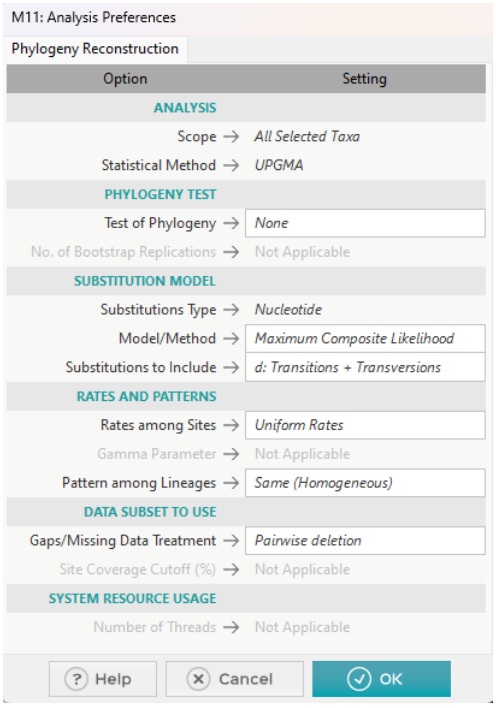
3.3 Secondo Passo: Creazione Alberi Filogenetici

Adempiuto il requisito fondamentale dell'allineamento delle sequenze, passiamo alla creazione degli alberi filogenetici.
In questa sezione creeremo 3 alberi filogenetici con 3 algoritmi diversi che abbiamo visto nel capitolo 2 :
Gli algoritmi UPGMA, Neighbor Joining e Maximum Parsimony Score.
Per poter accedere alla finestra di filogenetica di MEGA, dobbiamo cliccare sull'elemento della barra evidenziato da una freccia rossa in figura.
Al click verrà generata un menu a tendina da cui si potrà selezionare l'algoritmo da utilizzare.

3.3.1 Albero Filogenetico: UPGMA

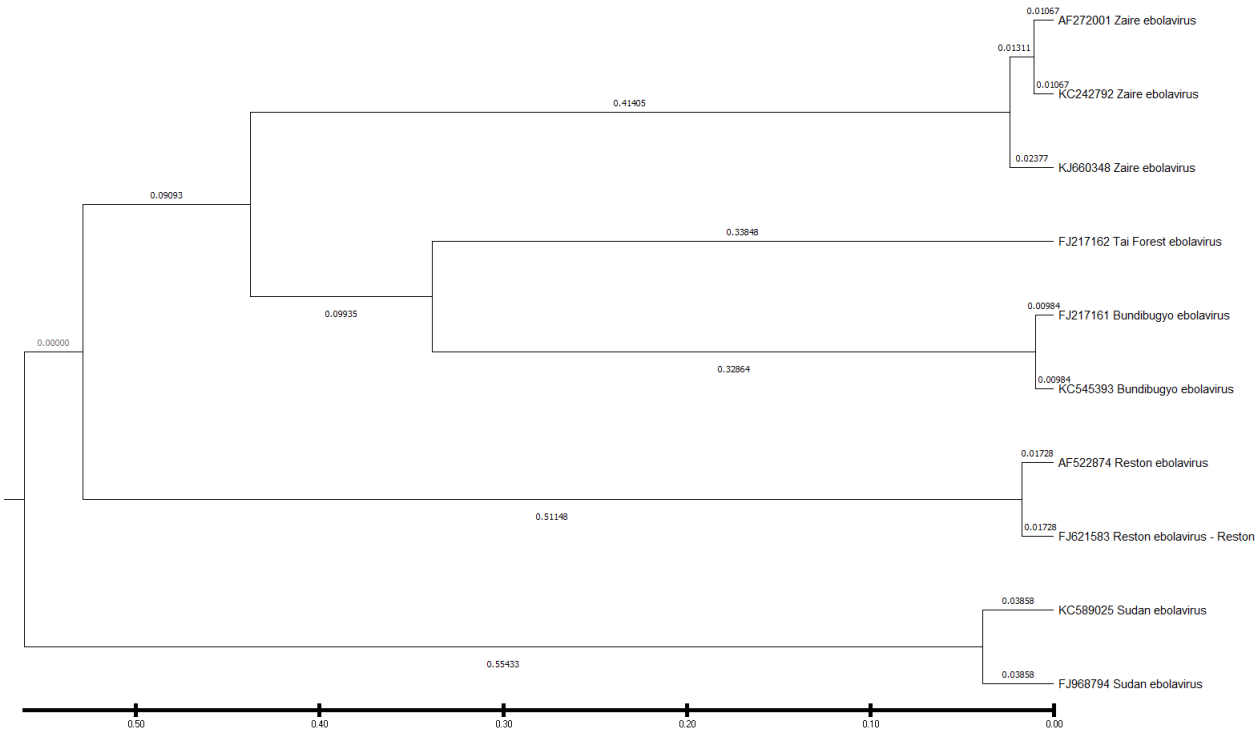


Successivamente alla selezione dell'algoritmo UPGMA, MEGA chiede i file delle sequenze su cui operare. L'utente può decidere di inserire direttamente le sequenze oppure di inserire la sessione di allineamento precedentemente creata, la nostra scelta ricadrà sulla seconda.
Successivamente a questa scelta MEGA chiederà se le sequenze su cui si sta effettuando l'elaborazione risultano essere codificanti, nel nostro caso specifico le sequenze di Ebola non risultano esserlo, per cui selezioneremo il no.
In seguito MEGA mostrerà una finestra dove l'utente a piacimento potrà modificare i parametri di esecuzione dell'algoritmo, la nostra scelta è stata quella di mantenerli con i valori di default.

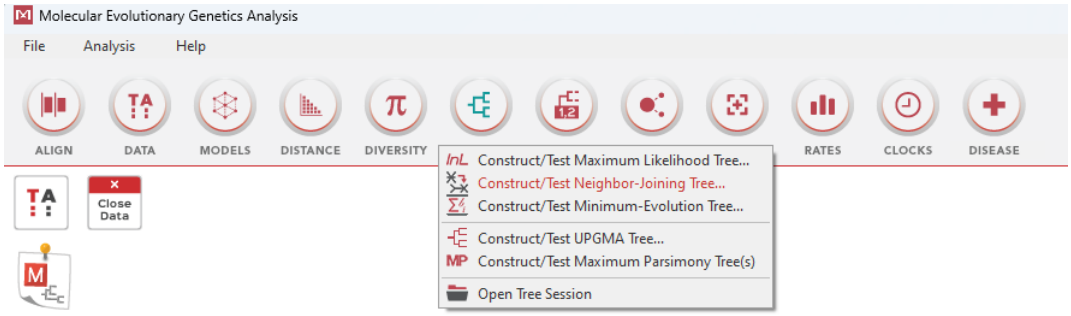


Dopo aver confermato i parametri dell'algoritmo compare la finestra di MEGA dove è stato costruito l'albero filogenetico desiderato.

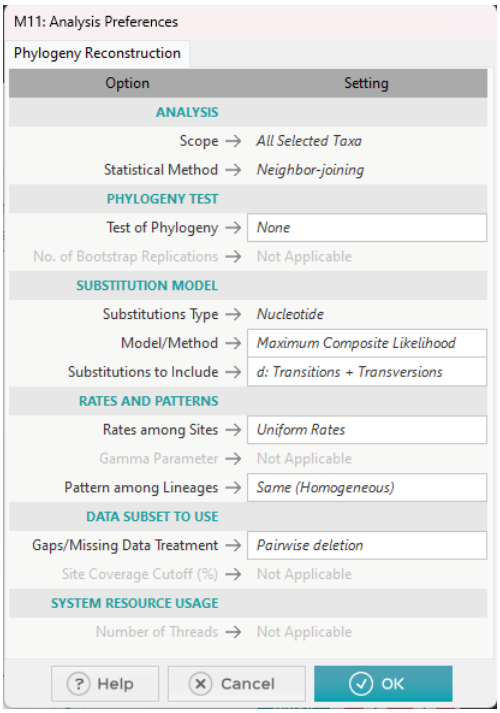
Prima di salvare l'immagine effettuiamo delle modifiche all'albero per renderlo più visibile e maggiormente descrittivo.
Inizialmente accediamo tramite la barra al lato alla sezione **Branch Length** (verificare se la sezione presenta una spunta) e modifichiamo la sezione **Precision** con le cifre decimali significative preferite. La nostra scelta è ricaduta su 5 cifre significative.
Successivamente, tramite la sezione layout, adattiamo la dimensione dell'albero alla finestra tramite il pulsante **Auto-Size Tree**.
Infine, dopo aver effettuato le seguenti modifiche, tramite barra dell'applicazione posta al di sopra della finestra, accediamo al menu a tendina **Image** e selezioniamo formato dell'immagine desiderata e percorso dove salvarla.
Il risultato sarà il seguente :



3.3.2 Albero Filogenetico: Neighbor Joining

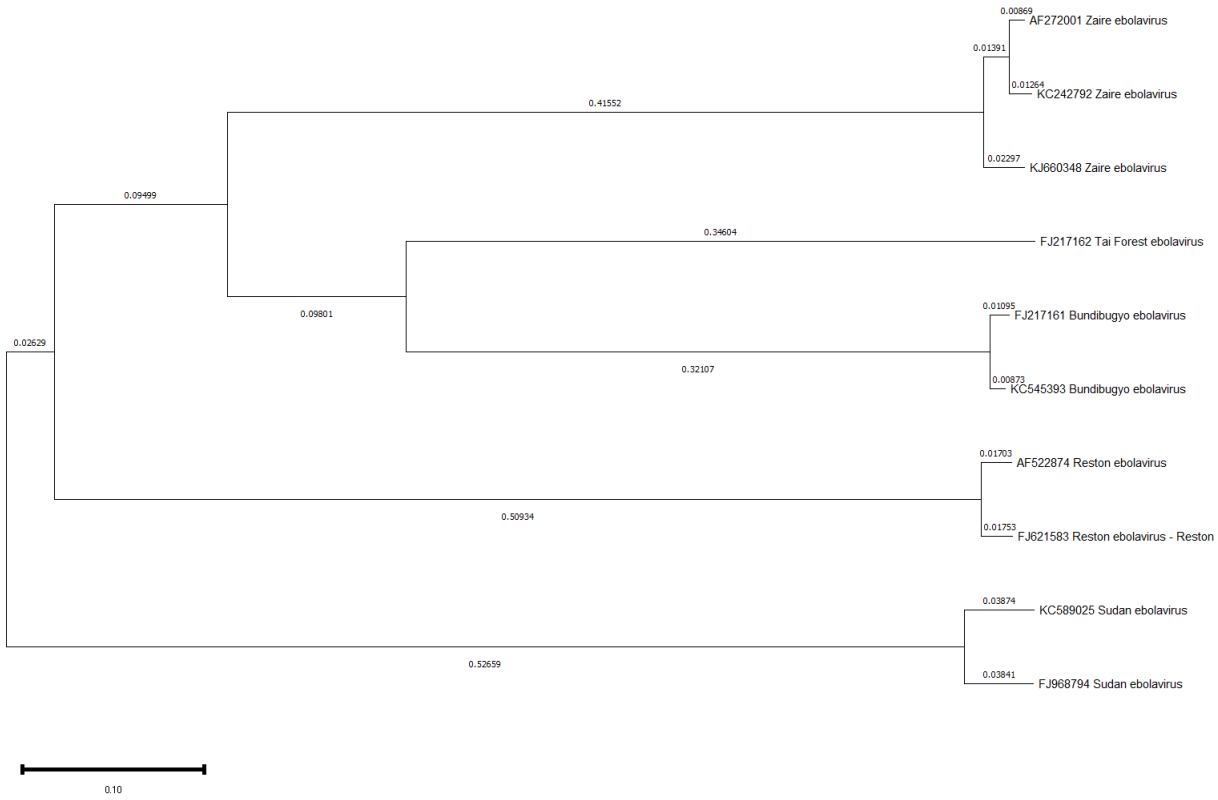


Successivamente alla selezione dell'algoritmo Neighbor, MEGA chiede i file delle sequenze su cui operare. L'utente può decidere di inserire direttamente le sequenze oppure di inserire la sessione di allineamento precedentemente creata, la nostra scelta ricadrà sulla seconda.
Successivamente a questa scelta MEGA chiederà se le sequenze su cui si sta effettuando l'elaborazione risultano essere codificanti, nel nostro caso specifico le sequenze di Ebola non risultano esserlo, per cui selezioneremo il no.
In seguito MEGA mostrerà una finestra dove l'utente a piacimento potrà modificare i parametri di esecuzione dell'algoritmo, la nostra scelta è stata quella di mantenerli con i valori di default.

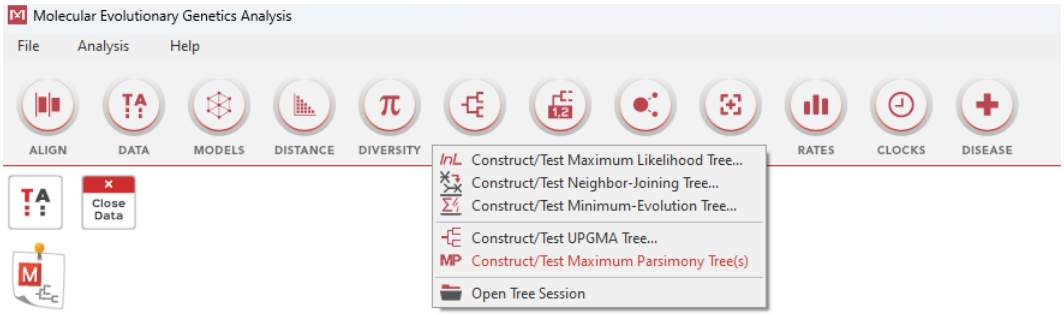


Dopo aver confermato i parametri dell'algoritmo compare la finestra di MEGA dove è stato costruito l'albero filogenetico desiderato.

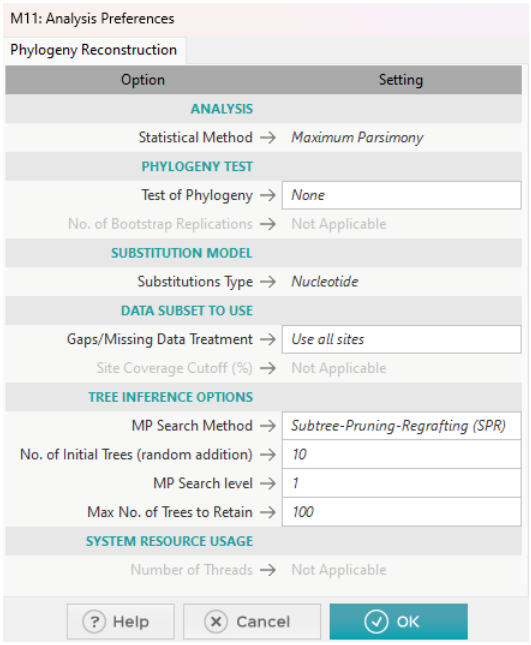
Prima di salvare l'immagine effettuiamo delle modifiche all'albero per renderlo più visibile e maggiormente descrittivo.
Inizialmente accediamo tramite la barra al lato alla sezione **Branch Length** (verificare se la sezione presenta una spunta) e modifichiamo la sezione **Precision** con le cifre decimali significative desiderate. La nostra scelta è ricaduta su 5 cifre significative.
Successivamente, tramite la sezione layout, adattiamo la dimensione dell'albero alla finestra tramite il pulsante **Auto-Size Tree**.
Infine, dopo aver effettuato le seguenti modifiche, tramite barra dell'applicazione posta al di sopra della finestra, accediamo al menu a tendina **Image** e selezioniamo formato dell'immagine desiderata e percorso dove salvarla.
Il risultato sarà il seguente :



3.3.3 Albero Filogenetico: Maximum Parsimony Score

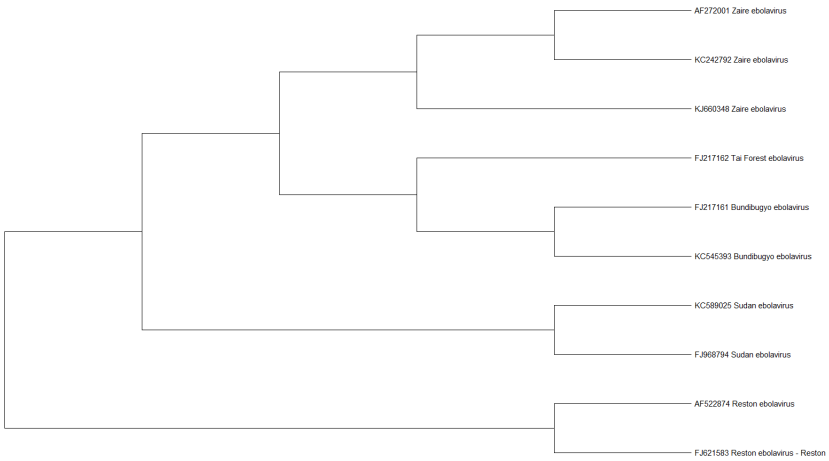


Successivamente alla selezione dell'algoritmo Maximum Parsimony, MEGA chiede i file delle sequenze su cui operare. L'utente può decidere di inserire direttamente le sequenze oppure di inserire la sessione di allineamento precedentemente creata, la nostra scelta ricadrà sulla seconda.
Successivamente a questa scelta MEGA chiederà se le sequenze su cui si sta effettuando l'elaborazione risultano essere codificanti, nel nostro caso specifico le sequenze di Ebola non risultano esserlo, per cui selezioneremo il no.
In seguito MEGA mostrerà una finestra dove l'utente a piacimento potrà modificare i parametri di esecuzione dell'algoritmo, la nostra scelta è stata quella di mantenerli con i valori di default.



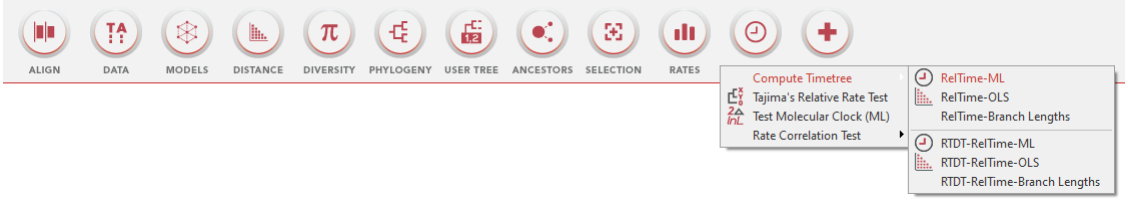
Dopo aver confermato i parametri dell'algoritmo compare la finestra di Mega dove è stato costruito l'albero filogenetico desiderato.
Prima di salvare l'immagine effettuiamo delle modifiche all'albero per renderlo più

visibile e maggiormente descrittivo.
Tramite la sezione layout, adattiamo la dimensione dell'albero alla finestra tramite il pulsante **Auto-Size Tree**.
Dopo aver effettuato la seguente modifica, tramite barra dell'applicazione posta al di sopra della finestra, accediamo al menu a tendina **Image** e selezioniamo formato dell'immagine desiderata e percorso dove salvarla.
Il risultato sarà il seguente:

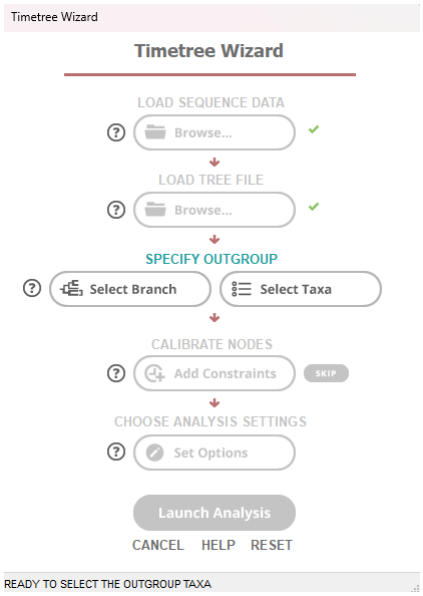


3.3.4 Time Tree

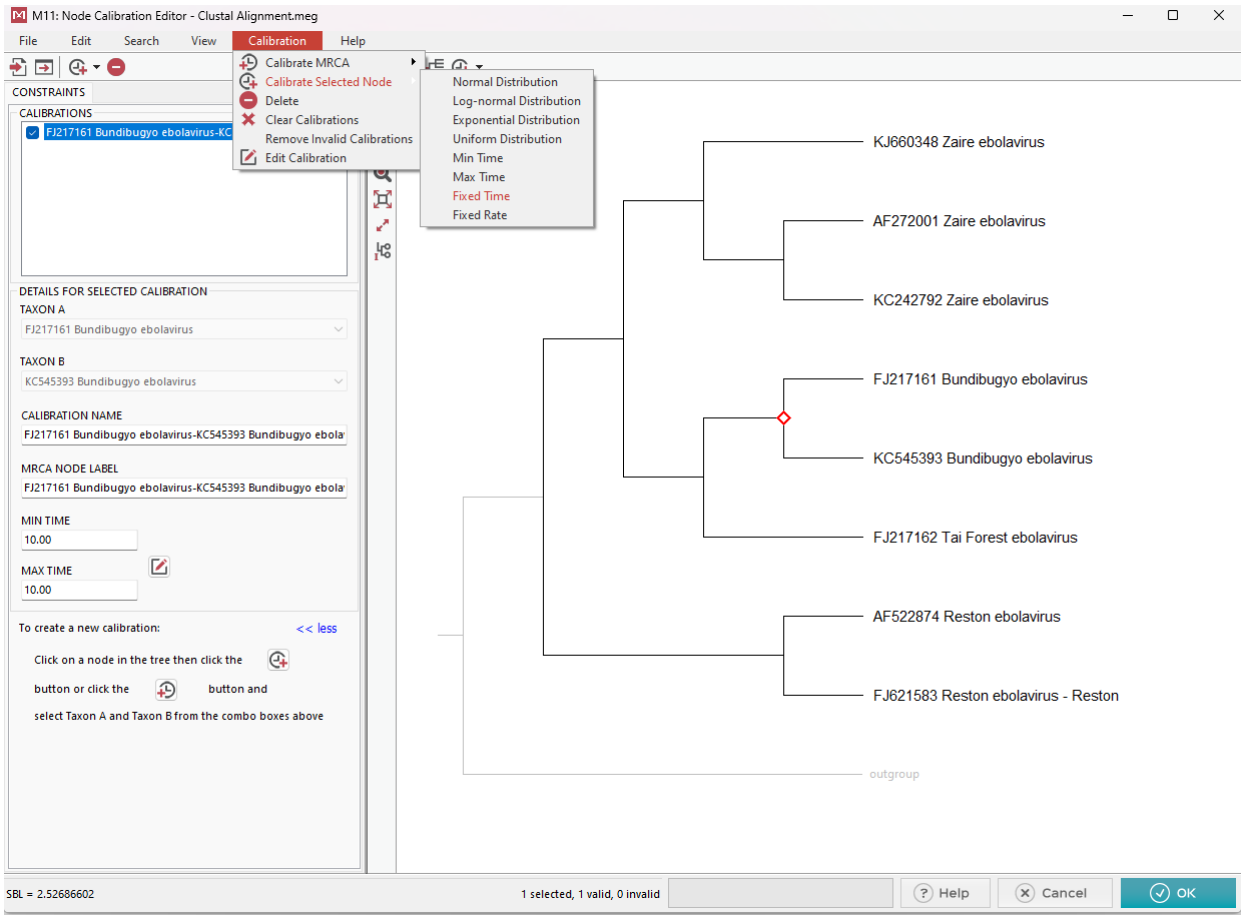
Un altro albero utile al nostro studio è il *Time Tree*.
Per calcolare un albero di questo tipo necessitiamo di un alignment e di un file che rappresenti un albero. Mentre il primo requisito è ampiamente rispettato, non possiamo dire lo stesso sul secondo.
Questo tipo di file è facilmente reperibile tramite le finestre filogeniche di MEGA, dove accedendo al menu a tendina *File*, e successivamente sull'item *Export Current Tree(Newick)*, l'applicazione ci permette di salvare l'albero in un formato specifico di tipo *.nwk*. L'albero che utilizzeremo per questo scopo sarà quello calcolato con algoritmo UPGMA.
Reperiti quindi i seguenti file, come nei casi precedenti, tramite la barra dell'applicazione MEGA clickiamo la sezione *Clocks*, e nel menu a tendina che compare selezioniamo l'item *Compute Timetree(RealTime-ML)*.



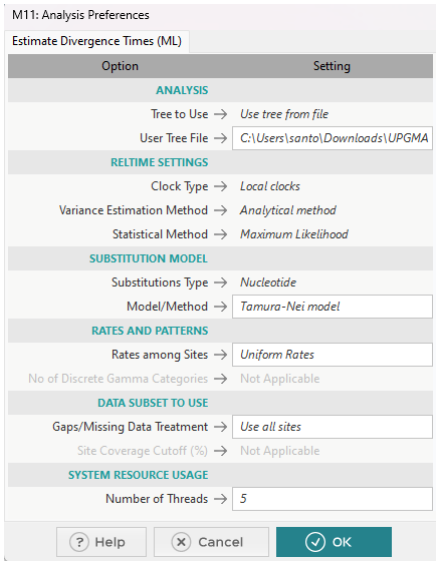
Al click, si aprirà la finestra seguente che richiederà prima di tutto l'allineamento e l'albero, come avevamo anticipato in precedenza, ed altre informazioni come la selezione dell'outgroup.



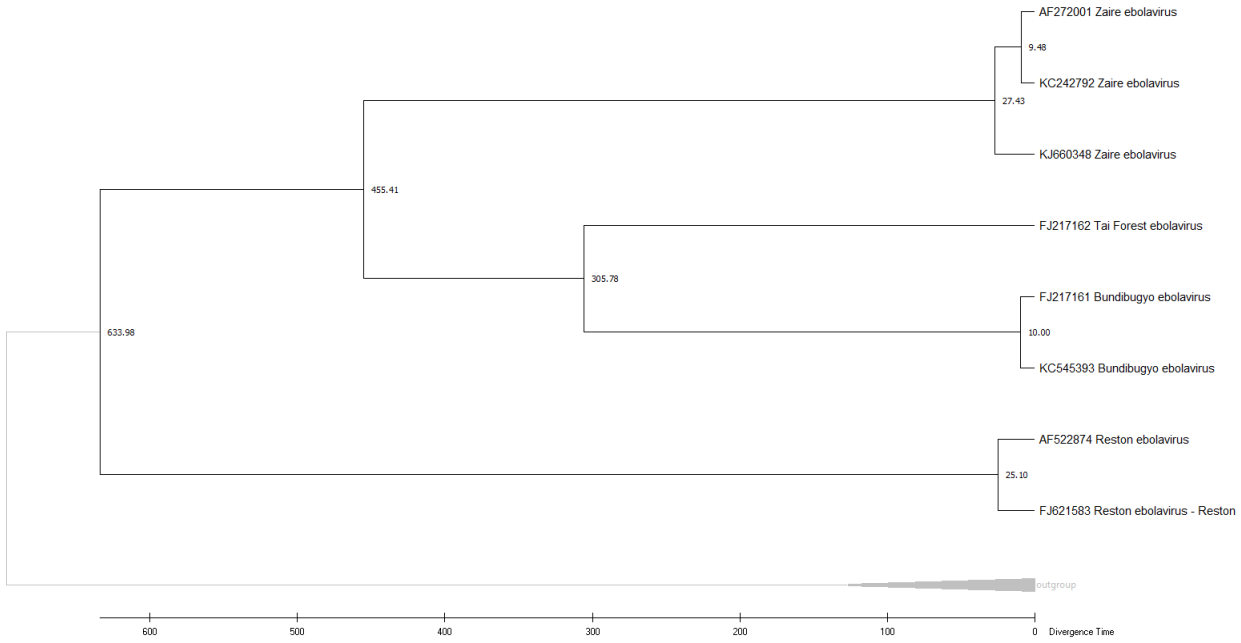
Nella selezione dell'outgroup è possibile specificare sia le tassonomie da includere che direttamente il ramo che le include. La nostra scelta è ricaduta nella 2 opzione, e abbiamo selezionato il ramo che include KC589025 e FJ968794 come outgroup.
Ultima opzione che il software richiede (facoltativa) è la selezione dei vincoli. Il vincolo che abbiamo inserito nel nostro studio è stato di natura temporale ed assume che le due sequenze BDBV (FJ217161 e KC545393) si sono separate dal loro antenato comune 10 anni fa.
Per far ciò dalla finestra dei vincoli selezioniamo il nodo antenato che unisce le due sequenze tramite cursore, clickiamo *Calibration* sulla barra superiore della finestra, successivamente l'item *Calibrate Selected Note* e infine *Fixed Time*, e impostiamo come valore massimo e minimo 10.



Dopo aver selezionato i vincoli, il software, come nei casi precedenti, ci fa modificare i parametri di esecuzione dell'algoritmo, che anche in questo caso noi lasciamo di default.



Al termine, clickiamo sul pulsante per far proseguire l'esecuzione del metodo e dopo pochi secondi compare l'albero cosi' calcolato.



4 Analisi dei risultati

Nella seguente sezione andremo ad analizzare gli alberi e i dati ottenuti dall'analisi filogenetica delle sequenza di Ebola e tramite questi risponderemo alle domande poste da *Phillip Compeau* in [1].

4.1 Riepilogo Sequenze Ebola

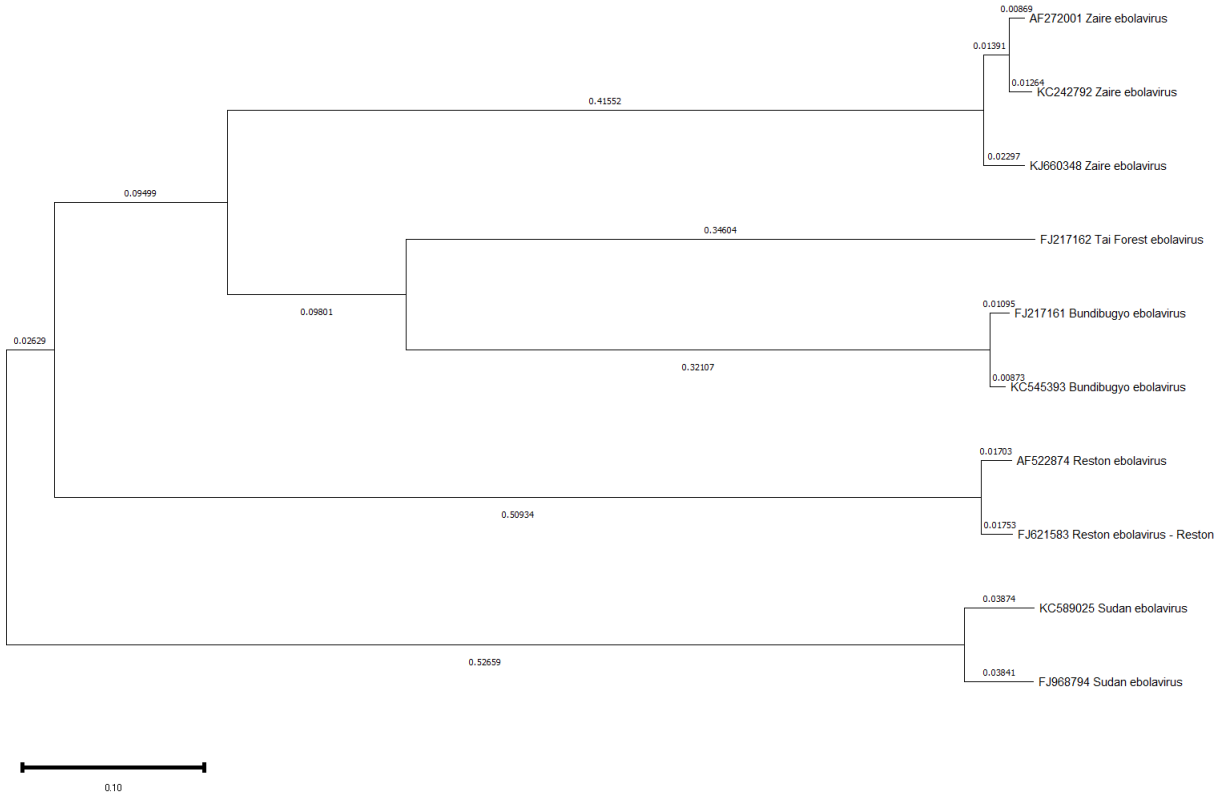
Accession Number	Virus Species	Location	Date
KJ660348	????	Gueckedou, Guinea	2014
FJ217161	Bundibugyo (BDBV)	Bundinbugyo, Uganda	2007
KC545393	Bundibugyo (BDBV)	Isiro, DRC	2012
AF272001	Zaire (EBOV)	Yambuku, DRC	1976
KC242792	Zaire (EBOV)	Mekouka, Gabon	1994
KC589025	Sudan (SUDV)	Luwero, Uganda	2012
FJ968794	Sudan (SUDV)	Sudan	1976
FJ217162	Tai Forest (TAFV)	Tai Forest, Ivory Coast	1994
AF522874	Reston (RESTV)	Philippines	1990
FJ621583	Reston (RESTV)	Philippines	2008

Tabella 10: Sequenze Ebola

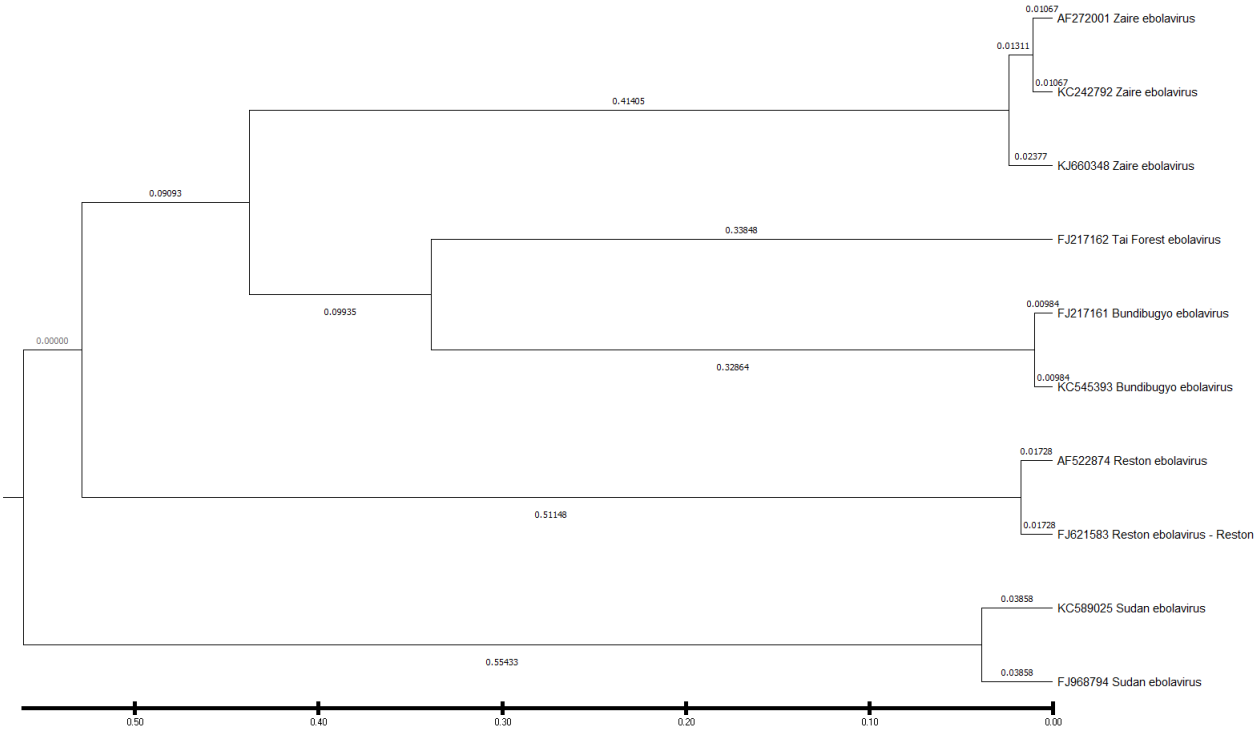
4.2 Validazione Ipotesi Iniziale

Basandosi sull'albero filogenetico ottenuto dall'esecuzione dell'algoritmo Neighbour Joining, quale specie di Ebola ha scatenato l'epidemia del 2014? Da cosa lo supponiamo? E' la stessa specie che abbiamo premesso nell'introduzione?

Secondo l'albero ottenuto tramite il Neighbor-Joining, le sequenze più vicine alla sequenza di Ebola del 2014, identificata con KJ660348, sono di tipo EBOV. In particolare sono il AF272001 e il KC242792. Questo risultato è diverso da quello ipotizzato nel capitolo 1, ossia la Tai Forest (TAFV).


0.00869
AF272001 Zaire ebolavirus
0.01391
0.01264
KC242792 Zaire ebolavirus
0.41552
0.02297
KJ660348 Zaire ebolavirus
0.09499
0.34604
FJ217162 Tai Forest ebolavirus
0.09801
0.01095
FJ217161 Bundibugyo ebolavirus
0.02629
0.32107
0.00873
KC545393 Bundibugyo ebolavirus
0.01703
AF522874 Reston ebolavirus
0.50934
0.01753
FJ621583 Reston ebolavirus - Reston
0.03874
KC589025 Sudan ebolavirus
0.52659
0.03841
FJ968794 Sudan ebolavirus
0.10

4.3 Analisi Albero Filogenetico UPGMA



4.3.1 Radice dell'albero

Dove si trova la root dell'albero prodotto dall'UPGMA?

La root dell'albero è da individuare nel ramo che divide le specie di Sudan (SUDV) (KC589025 e FJ968794) dalle altre specie.

4.3.2 Distanza Totale

Qual è la distanza totale D_{max} dell'albero prodotto dall'algoritmo UPGMA? (Distanza da ogni foglia alla radice)

$D_{max} = 0.52876$. In questo caso sono state calcolate le distanze per i singoli rami e si è scelto il valore più comune. In un solo ramo, il valore per la distanza varia ed è 0.59291. Come descritto in precedenza (capitolo 2), l'algoritmo UPGMA non rispetta la matrice delle distanze e dunque le distanze possono variare tra di loro.

4.3.3 Distanza tra epidemia 2014 e fattore scatenante

Qual è la distanza nell'albero UPGMA tra la foglia KJ660458 (Epidemia Ebola 2014) e il nodo interno che la mette in relazione con i suoi antenati?

$D_{2014} = 0.02377$. Questa distanza è da individuarsi nel ramo che collega KJ660458 e il nodo interno da cui si genera il sottoalbero che descrive AF272001 e KC242792.

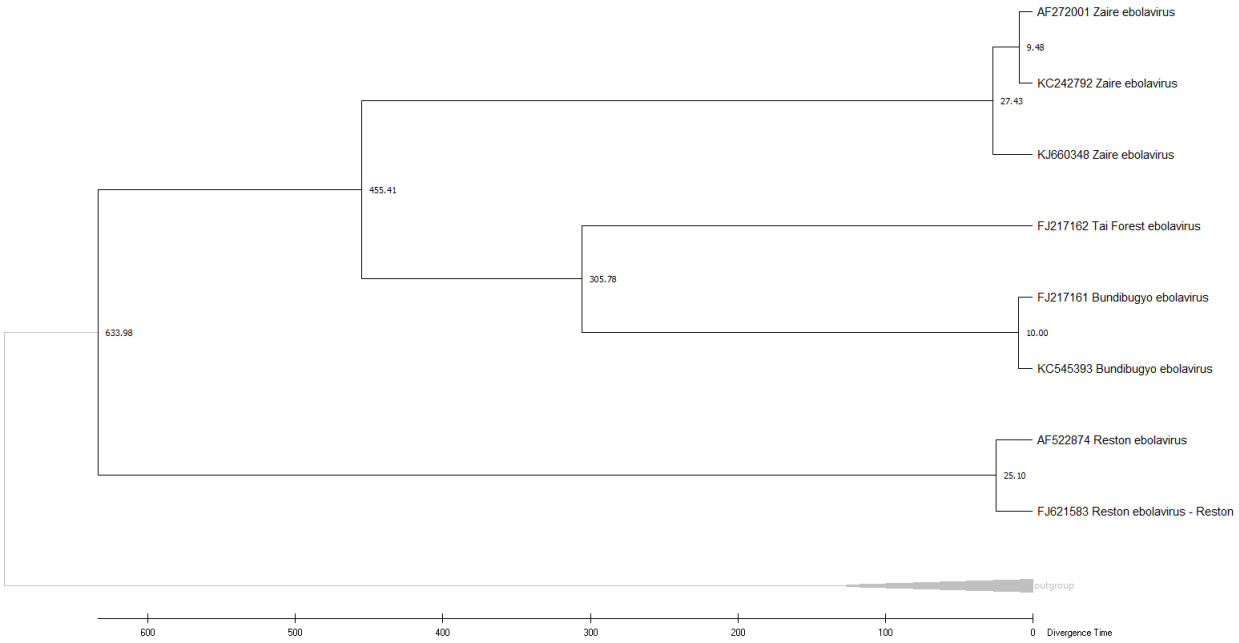
4.3.4 Distanza di Mutazione Ebola 2014

Secondo l'albero UPGMA, quanto tempo ci è voluto per il virus Ebola del 2014 per dividersi da altri virus Ebola nell'albero dopo il loro antenato comune più recente?

Tale distanza è pari alla differenza tra $D_{max} - D_{2014} = 0.50499$.

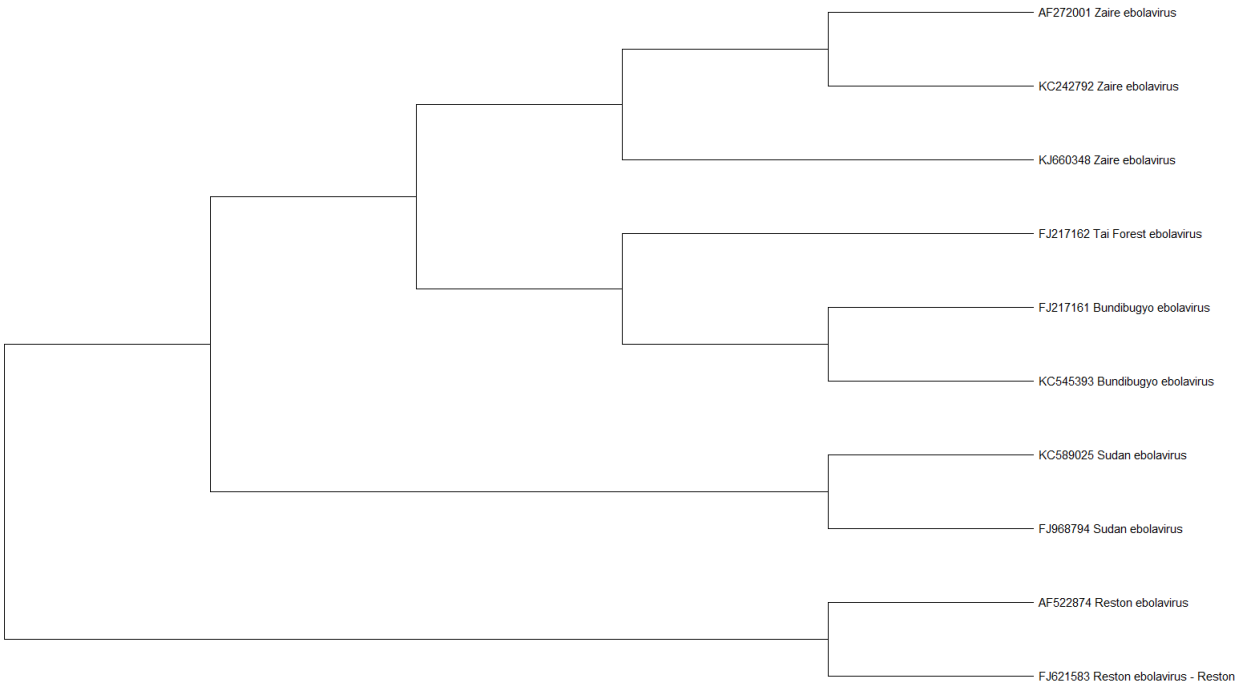
4.3.5 Distanza Temporale Mutazione Ebola 2014

Secondo il Timetree risultante, quanti anni fa l'Ebola del 2014 si è separata dal suo antenato comune?



La risposta è 27.43, come evidenziato dall'albero in figura.

4.4 Analisi Albero Filogenetico Maximum Parsimony



4.4.1 Radice dell'albero

L'albero prodotto dal maximum parsimony è radicato o non radicato?

L'albero è non radicato.

4.4.2 Cambiamento di base Nucleotidica

Quante basi nucleotidiche sono cambiate tra la sequenza di Ebola del 2014 e il suo antenato?

I cambiamenti di basi nucleotidiche che sono state identificate sono 10623.
Valore che ci è stato possibile calcolare facendoci restituire da MEGA un file Excel[3] che descrive tutti i cambiamenti di base per ogni sequenza rispetto al suo diretto antenato.
Questo punto è stato svolto, al contrario dei precedenti, con la versione del software MEGA7 poichè con la versione corrente rispetto al assignment di Compeau i risultati risultavano essere largamente differenti (circa 600).
Supponiamo, dato che il software MEGA non risulta essere molto chiaro nella sua documentazione, che questa discrepanza sia data dal cambiamento degli algoritmi o

[3]Nella finestra del Max Parsimony, dalla barra superiore della finestra clickiamo su *Ancestors*, selezionamo dalla tendina l'item *Show All*,senza cui non potremmo esportare il file, e infine, sempre dalla stessa, clickiamo sull'item *Export Change List*, dove si aprirà una finestra di dialogo che permetterà la selezione del formato del file generato e il percorso di memorizzazione.

dal loro perfezionamento negli anni, dato che a parità di allineamenti, esplorando il file restituitoci anche le basi mutate cambiano in quantità e posizione. Riteniamo inoltre che il risultato ottenuto con MEGA7 sia più attendibile, al contrario di quello ottenuto nella versione di MEGA11, poichè il virus Ebola preso in esame muta molto velocemente da epidemia ad epidemia, il che rende questo virus particolarmente difficile da analizzare.

5 Conclusioni

In questa analisi si è dunque voluto descrivere, tramite gli alberi filogenetici, quale delle specie di Ebola virus è stata responsabile del focolaio iniziale dell'epidemia del 2014. Come mostrato nel capitolo 4, l'ipotesi iniziale di considerare la specie il cui campione di rilevamento è il più vicino al focolaio è stata smentita. Dal virus di tipo Tai Forest (TAFV), si è optato per la specie Zaire (EBOV). Possiamo dunque aggiornare nel seguente modo la tabella delle specie di Ebola analizzate:

Accession Number	Virus Species	Location	Date
KJ660348	Zaire (EBOV)	Gueckedou, Guinea	2014
FJ217161	Bundibugyo (BDBV)	Bundinbugyo, Uganda	2007
KC545393	Bundibugyo (BDBV)	Isiro, DRC	2012
AF272001	Zaire (EBOV)	Yambuku, DRC	1976
KC242792	Zaire (EBOV)	Mekouka, Gabon	1994
KC589025	Sudan (SUDV)	Luwero, Uganda	2012
FJ968794	Sudan (SUDV)	Sudan	1976
FJ217162	Tai Forest (TAFV)	Tai Forest, Ivory Coast	1994
AF522874	Reston (RESTV)	Philippines	1990
FJ621583	Reston (RESTV)	Philippines	2008

Ma come può una specie che è stata rilevata nel Congo provocare un focolaio in Guinea senza infettare nessuno nel tragitto? Per quanto riguarda i virus ed altri agenti patogeni, molto spesso si trovano negli animali, e molto spesso si parla di fenomeni di spillover (o salto di specie). Tali eventi si verificano quando una popolazione animale, detta serbatoio, entra in contatto con una popolazione di specie differente, detta ospite, e trasmette il virus.
In questo caso gli animali che avrebbero potuto trasferire questo virus potrebbe essere stati i pipistrelli della frutta o i pipistrelli angolani dalla coda libera. Nella seguente figura sono riportate le rotte migratorie di questi animali, che coprono anche la rotta che ci era necessaria per descrivere il passaggio del virus dal Congo alla Guinea.

Inoltre tutte le specie di Ebola virus sono rilevate nelle rotte migratorie dei pipistrelli e molto probabilmente le varie epidemie di Ebola sono state il frutto di diversi spillover durante il passare del tempo.

Riferimenti bibliografici

[1] Phillip Compeau. *Software Challenge: Constructing Evolutionary Trees (Answer Key)*, corso 02-604: Fundamentals of Bioinformatics Spring 2018.

[2] Neil A Campbell, Jane B Reece, and Eric J Simon. *Principi di biologia*. Pearson Italia Spa, 2008.

[3] Phillip Compeau. *Great Ideas in Computational Biology Course Resources*. https://compeau.cbd.cmu.edu/teaching/great-ideas-in-computational-biology/ , 2022 (estratto gennaio 2023).

[4] Phillip Compeau & Pavel Pevzner. *Bioinformatics Algorithms YT Course*. https://www.youtube.com/playlist?list=PLQ-85lQlPqFPhJxNcuSOxLTNmbp3NzVtU, 2017 (estratto gennaio 2023).

[5] Marketa Zvelebil & Jeremy O. Baum. *Understanding Bioinformatics*. Garland Science, Taylor & Francis Group, LLC, 2008.

[6] Istituto Superiore di Sanità. *Epidemia da virus Ebola 2014-2016*. https://www.epicentro.iss.it/ebola/epidemia-africa-2014, 2023 (estratto gennaio 2023).

[7] Gibson TJ. Thompson JD, Higgins DG. Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nov 1994.

[8] K. Chaichoompu, S. Kittitornkun, and S. Tongsima. Mt-clustalw: multithreading multiple sequence alignment. In *Proceedings 20th IEEE International Parallel & Distributed Processing Symposium*, pages 8 pp.–, 2006.

[9] Robert Edgar. Muscle: A multiple sequence alignment method with reduced time and space complexity. *BMC bioinformatics*, 5:113, 09 2004.

[10] Robert C. Edgar. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research*, 32(5):1792–1797, 03 2004.

[11] Nei M. Saitou N. *The neighbor-joining method: a new method for reconstructing phylogenetic trees.* Mol Biol Evol., 1987 Jul.

[12] Karla Esmeralda Vazquez Ortiz. *Advanced methods to solve the maximum parsimony problem*. Theses, Université d'Angers, June 2016.