UNIVERSIDAD NACIONAL AGRARIA LA MOLINA

DEPARTAMENTO DE ESTADÍSTICA INFORMÁTICA

CURSO: Lenguaje de Programación II - Principios de Bioinformática

TEMA: Árbol filogenético de 5 especies del género Passiflora

Integrantes	Código
Quinte Saenz Ximena (Biología)	20191138
Olortegui Perez Pacheco Ariana (Biología)	20191418
Garces Quispe, Adryana Luisa (Estadística)	20220764
Jesus Mamani, Angelo Miguel (Estadística)	20220767

Profesora: Ana Cecilia Vargas Paredes

2024

I. INTRODUCCIÓN

El siguiente análisis presenta varios árboles filogenéticos de cinco especies del género *Passiflora*: *P. caerulea, P. actinia, P. elegans, P. organensis y P. misera.* Estos árboles se han construido utilizando RStudio , aprovechando bibliotecas especializadas en bioinformática y visualización de datos.

El género Passiflora, conocido comúnmente como "flores de la pasión", es notable por su diversidad morfológica y ecológica. Este estudio filogenético nos permite explorar las relaciones evolutivas entre estas cinco especies seleccionadas, proporcionando insights sobre su historia evolutiva y posibles adaptaciones.

Para construir este árbol, se han utilizado secuencias de ADN de las especies mencionadas, específicamente el ARN Ribosomal 5S. Estas secuencias se alinearon utilizando un programa que es muy utilizado en bioinformática llamado "MAFFT", y se construyeron diferentes arboles filogenéticos empleando los siguientes métodos: Máxima Verosimilitud, Neighbor Joining y UPGMA.

El código en RStudio utiliza bibliotecas como "ape" para el manejo de secuencias y la construcción del árbol, "adegenet" para el análisis de datos genéticos, y "phangorn" que proporciona métodos para inferir árboles filogenéticos utilizando diversos algoritmos.

Este análisis nos permitirá observar la cercanía evolutiva entre estas especies de *Passiflora*, lo cual puede proporcionar información valiosa sobre la diversificación del género y potencialmente guiar futuros estudios taxonómicos y de conservación. Además sabremos si los árboles tienen diferencias notables entre ellos.

II. MARCO TEÓRICO

Filogenia

La filogenia es la ciencia que estudia las relaciones evolutivas entre los organismos. A través del análisis filogenético, los científicos pueden reconstruir los patrones de ascendencia y descendencia que han llevado a la diversidad actual de formas de vida.

Métodos de Construcción de Árboles Filogenéticos

- **Neighbor Joining (NJ):** Es un método de inferencia filogenética utilizado para crear árboles filogenéticos. Es especialmente útil para grandes conjuntos de datos debido a su eficiencia computacional.
- **Máximo de Verosimilitud (ML):** Este método estima la topología del árbol filogenético que maximiza la probabilidad de observar los datos dados un modelo de evolución molecular.
- **UPGMA:** Es un método jerárquico de agrupamiento que asume una tasa constante de evolución (reloj molecular). Aunque es simple y rápido, su suposición de tasas de evolución iguales puede no ser realista en muchos casos.

III. MATERIALES Y MÉTODOS

A. Descarga de Datos:

Los datos se descargaron en formato fasta usando el código de la Figura 1 en JupyterNotebook, se utilizó la librería “Biopython”.

```
from Bio import Entrez, SeqIO

# Establecer el correo electrónico (NCBI requiere que proporciones un correo electrónico)
# Si desea usar su propio correo, debe crearse una cuenta en el NCBI
Entrez.email = "20220764@lamolina.edu.pe" # En este caso, se está usando un correo que ya tiene una cuenta en NCBI

# Lista de especies y secuencias específicas
species_sequences = {
    "KX378102.1": "Passiflora caerulea isolate SBV08 5S ribosomal RNA gene, complete sequence",
    "KX378057.1": "Passiflora organensis isolate PM31 5S ribosomal RNA gene, complete sequence",
    "KX378119.1": "Passiflora misera isolate SF20 5S ribosomal RNA gene, complete sequence",
    "KX378094.1": "Passiflora actinia isolate PM69 5S ribosomal RNA gene, complete sequence",
    "KX378121.1": "Passiflora elegans isolate SF24 5S ribosomal RNA gene, complete sequence"
}
# definimos un diccionario que contiene los identificadores de las secuencias (por ejemplo, "KX378102.1") como claves,
# y las descripciones de las secuencias como valores. Esto nos permite buscar secuencias específicas.


# Obtener los identificadores de las secuencias
id_list = list(species_sequences.keys()) # Convertimos las claves del diccionario (que son los identificadores de las secuencias)
                                         # en una lista que se utilizará para buscar las secuencias en la base de datos NCBI.


# Descargar las secuencias en formato FASTA
ids = ",".join(id_list) # Junta todos los identificadores en una sola cadena separada por comas.
handle = Entrez.efetch(db="nucleotide", id=ids, rettype="fasta", retmode="text")
sequences = handle.read()
handle.close()

# Guardar las secuencias en un archivo FASTA
file_name = "sequence.fasta"
with open(file_name, "w") as out_file:
    out_file.write(sequences)
print(f"Secuencias descargadas y guardadas en {file_name}")

# Leer y mostrar un resumen de las secuencias en formato FASTA
for record in SeqIO.parse(file_name, "fasta"): # Itera sobre cada secuencia en el archivo FASTA.
    print(f">{record.id} {species_sequences[record.id]}")
    print(record.seq) # Imprime la secuencia de nucleótidos.
    print(f"Longitud de la secuencia: {len(record.seq)} nucleótidos")
    print(f"Secuencia guardada en {file_name}")
```

Figura 1. Código utilizado para descargar datos en JupyterNotebook.

B. Alineamiento de datos

Subimos el archivo “secuencia.fasta” a la página MAFFT para continuar con el alineamiento de secuencias, no configuramos ninguna opción específica en la página. Esperamos un rato hasta que la página procese la secuencia y la descargamos en formato fasta con el nombre de “secuencia_alineada.fasta”.

C. Construcción de árboles filogenéticos

1. Árbol Neighbor-Joint

Utilizamos la librería `ape` para construir el árbol filogenético utilizando el método Neighbor-Joining con ayuda de la función `nj()`.

2. Árbol UPGMA

Construimos el árbol UPGMA utilizando la librería `ape`, con la implementación de la función `upgma()`.

3. Árbol de Máxima Verosimilitud

Finalmente, construimos el árbol de máxima verosimilitud utilizando las librerías `ape` y `phangorn` en RStudio, empleando las funciones `pml()` y `optim.pml()`.

IV. RESULTADOS

Los árboles filogenéticos construidos con los tres métodos muestran las relaciones evolutivas entre las secuencias analizadas. Cada método puede generar resultados ligeramente diferentes debido a sus suposiciones y algoritmos subyacentes.

1. ÁRBOL FILOGENÉTICO CON MÉTODO DE NEIGHBOR JOINING

```r
library(adegenet)
file_path <- "C:\\Users\\ARIANA\\Desktop\\UNALM IX\\Principios de Bioinformatica\\Trabajo encargado\\secuencia_alineada.fasta"
dna <- fasta2DNAbin(file=file_path)
dna
csv_path<- "C:\\Users\\ARIANA\\Desktop\\UNALM IX\\Principios de Bioinformatica\\Trabajo encargado\\secuencia_alineada.csv"
annot <- read.csv(csv_path, header=TRUE,row.names = 1)
annot

library(ape)
D <- dist.dna(dna, model = "TN93")

length(D)
temp <- as.data.frame(as.matrix(D))
print(D)

#Visualización de la matriz de distancia
print(temp)

# Cargar el paquete ade4
library(ade4)
table.paint(temp, cleg=0, clabel.row=0.5, clabel.col=0.5)

#Construyendo árbol utilizando NJ:
library(ape)
tree <- nj(D)
class(tree)
tree <- ladderize(tree)
tree

plot(tree, cex = 0.6)
title("Un simple Arbol NJ")
```

A

```r
#Arbol enraizado con una de las anotaciones
library(ape)
library(adegenet)
windows(width = 8, height = 5)
tree2 <- root(tree, out = 1)
tree2 <- ladderize(tree2)
windows(width = 8, height = 5)
plot(tree2, show.tip=FALSE, edge.width=1)
title("Arbol enraizado NJ")
unique_ids <- unique(annot$ID)
myPal <- colorRampPalette(c("red", "yellow", "green", "blue", "gray", "orange"))(length(unique_ids))
id_colors <- setNames(myPal, unique_ids)
tip_colors <- id_colors[as.character(annot$ID)]
tiplabels(annot$ID, bg = tip_colors, cex = 0.5)
legend("bottomright", fill = myPal, legend = unique_ids, ncol = 2)
axisPhylo()
```

B

Figura 2. A) Importación de datos y construcción de matriz de distancias.
B) Construcción de árbol enraizado.

El código presentado en la Figura 2 nos da como resultado el siguiente árbol:

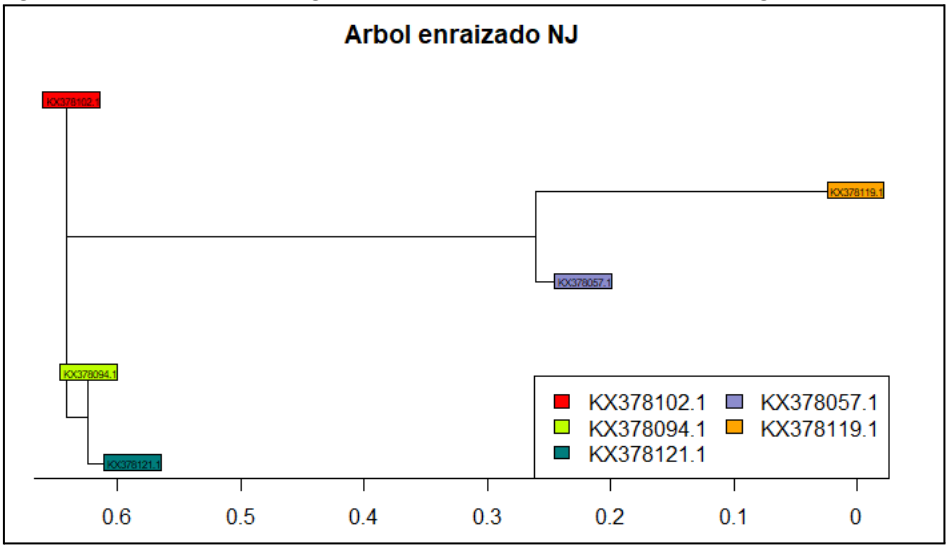

Figura 3. Árbol filogenético enraizado construido con el método Neighbor-Joint

2. ÁRBOL FILOGENÉTICO CON MÉTODO UPGMA

```
#ÁRBOL FILOGENÉTICO CON METODO UPGMA-----------------------------------------

# Instalar los paquetes si no están ya instalados
install.packages("ape")
install.packages("phangorn")

# Cargar los paquetes
library(ape)
library(phangorn)

# Leer el archivo de alineamiento en formato FASTA
alineamiento <- read.dna("secuencia_alineada.fasta", format = "fasta")

# Calcular la matriz de distancias usando el modelo K80
distancias <- dist.dna(alineamiento, model = "K80")
#Se calcula una matriz de distancias genéticas entre las secuencias de ADN
#Especifica el uso del modelo evolutivo Kimura 80 para calcular las distancias

# Construir el árbol filogenético usando el método UPGMA
arbol_upgma <- upgma(distancias)

# Dibujar el árbol filogenético
plot(arbol_upgma, main = "Árbol Filogenético UPGMA")
```

El código presentado nos da como resultado el siguiente árbol:

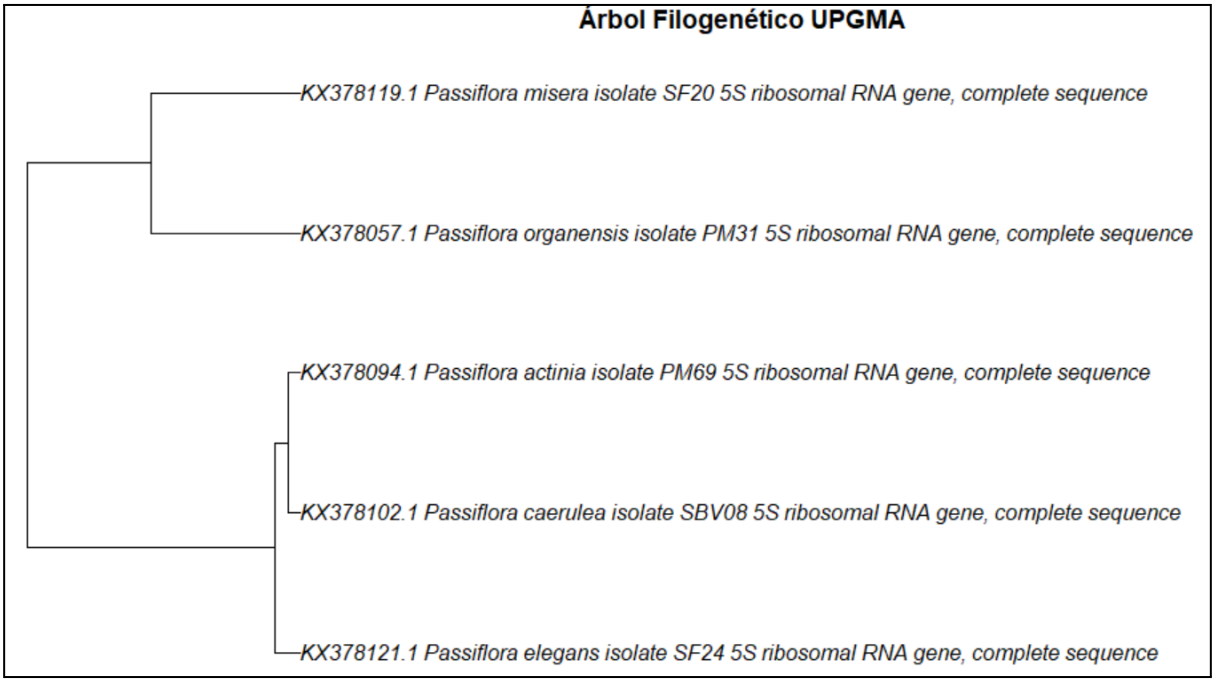

Figura 4: Árbol filogenético construido con el método UPGMA

3. ÁRBOL FILOGENÉTICO CON MÉTODO DE MÁXIMA VEROSIMILITUD

```
# ÁRBOL CON MÁXIMA VEROSIMILITUD-----------------------------------------------

# Cargar las librerías necesarias
library(ape)
library(phangorn)

# Cargar el archivo FASTA y el archivo CSV con anotaciones
file_path <- "secuencia_alineada.fasta"  # Ruta al archivo FASTA con secuencias de pasiflora
csv_path <- "secuencia_alineada.csv"     # Ruta al archivo CSV con anotaciones

dna <- read.dna(file_path, format = "fasta")
annot <- read.csv(csv_path, header = TRUE, row.names = 1)

# Convertir las secuencias a formato phyDat para análisis filogenético
dna_phy <- as.phyDat(dna)

# Crear un árbol inicial usando Neighbor Joining
tre_ini <- nj(dist.dna(dna, model = "TN93"))
# Se crea un árbol filogenético inicial utilizando el método de Neighbor Joining
# basado en las distancias de ADN calculadas con el modelo TN93.

# Estimar la verosimilitud del árbol inicial
fit_ini <- pml(tre_ini, dna_phy, k = 4)
# Se estima la verosimilitud utilizando el método de Máxima Verosimilitud
# con 4 categorías de gamma.
```

```
# Optimizar el árbol inicial
fit <- optim.pml(fit_ini, optNni = TRUE, optBf = TRUE, optQ = TRUE, optGamma = TRUE)
# tasas de sustitución de nucleótidos (optBf)
# las frecuencias de bases (optQ)
# la forma de la distribución gamma (optGamma)


# Mostrar los resultados de la optimización
print(fit)
```

```
> # Mostrar los resultados de la optimización
> print(fit)
model: F81+G(4)
loglikelihood: -1457.812
unconstrained loglikelihood: -1535.235
Discrete gamma model
Number of rate categories: 4
Shape parameter: 999.9958

Rate matrix:
          a         c         g         t
a 0.0000000 0.6006270 1.0236998 0.6846836
c 0.6006270 0.0000000 0.5259954 1.3818226
g 1.0236998 0.5259954 0.0000000 1.0000000
t 0.6846836 1.3818226 1.0000000 0.0000000

Base frequencies:
        a         c         g         t
0.2314471 0.2986006 0.2631213 0.2068311
```

```
# Comparar los modelos inicial y optimizado usando AIC (criterio de información de Akaike)
aic_ini <- AIC(fit_ini)
aic_opt <- AIC(fit)
cat("AIC inicial:", aic_ini, "\nAIC optimizado:", aic_opt, "\n")
```

```
> # Comparar los modelos inicial y optimizado usando AIC
> aic_ini <- AIC(fit_ini)
> aic_opt <- AIC(fit)
> cat("AIC inicial:", aic_ini, "\nAIC optimizado:", aic_opt,
"\n")
AIC inicial: 2977.221
AIC optimizado: 2947.624
```

```
# Enraizar y organizar el árbol optimizado
tre_opt <- root(fit$tree, 1)
tre_opt <- ladderize(tre_opt)

# Definir una paleta de colores pasteles
colores <- c("#E41A1C", "#377EB8", "#4DAF4A", "#984EA3", "#FF7F00")

# Crear un gráfico del árbol optimizado
windows()
par(mar = c(5, 4, 4, 2) + 0.1) # Ajustar márgenes
plot(tre_opt, show.tip = FALSE, edge.width = 2, cex = 0.8) # Mostrar etiquetas de las especies
#show.tip = TRUE -> Para poder observar a que tipo de especie de passiflora nos referimos

# Obtener nombres únicos de las especies
unique_species <- unique(annot$ID)

# Asignar colores a cada especie
species_colors <- colores[match(annot$ID, unique_species)]

# Añadir etiquetas de las especies con colores diferentes
tiplabels(annot$ID, bg = species_colors, col = "black", cex = 0.7)

# Añadir el eje temporal
axisPhylo()

# Ajustar título del gráfico
title("Árbol de Máxima Verosimilitud de Pasiflora", cex.main = 1.2)
```

El código presentado nos da como resultado el siguiente árbol:

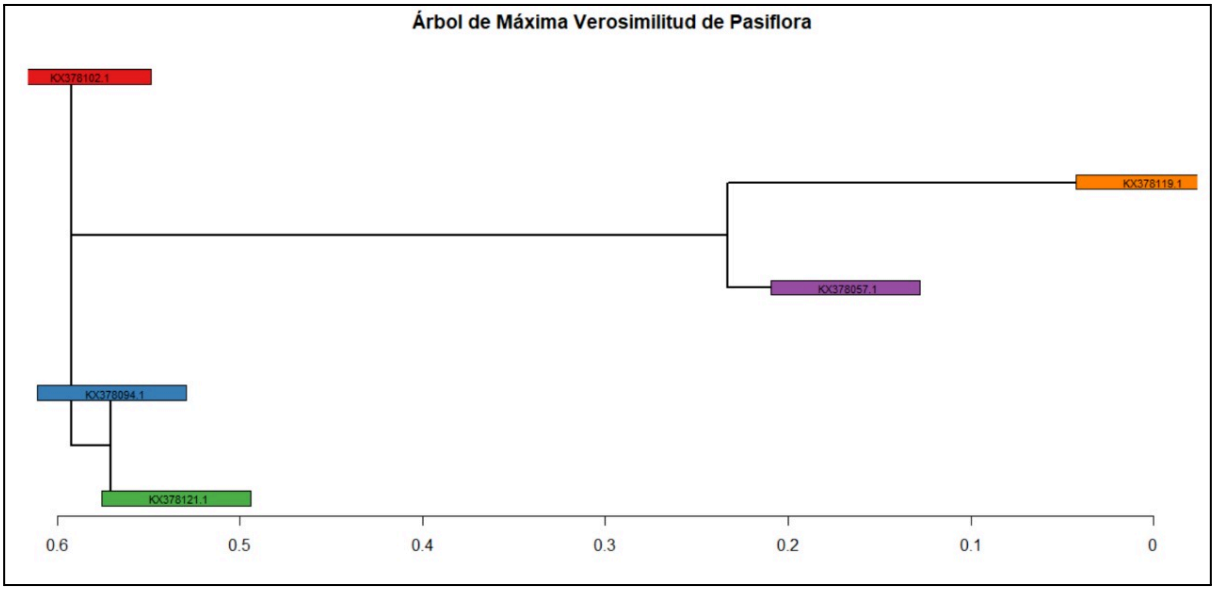

Figura 5: Árbol filogenético construido con el método de máxima verosimilitud

- Optimización: La optimización del árbol mejora el ajuste, lo que se refleja en un menor AIC comparado con el modelo inicial.
- Visualización: El árbol muestra las relaciones filogenéticas entre las secuencias de Passiflora, con etiquetas y colores para identificar fácilmente cada especie.

V. DISCUSIONES

- Con respecto al Árbol enraizado construido con el método Neighbor-Joining (NJ)

  En la **Figura 3** obtenemos como resultado un árbol Neighbor-Joining que muestra las relaciones evolutivas entre diferentes especies especificando un ancestro en común de todas las secuencias de *Pasiflora*, en la cual las ramas más cortas están estrechamente relacionadas. Este tipo de árbol es útil para visualizar la relación entre grandes conjuntos de datos.

  1. **Estructura:**
     - La escala en la parte inferior del gráfico indica la distancia genética.
     - Las ramas indican la distancia genética entre las secuencias.
  2. **Relaciones evolutivas:**
     - La secuencia raíz se ubica en el extremo izquierdo.
     - Las ramas convergen en el lado izquierdo antes de divergir hacia las hojas ubicadas en el lado derecho.
  3. **Interpretación de la distancia:**
     - Hay una menor diferencia genética entre *P. caerulea, P. actinia y P. Elegans*
     - A diferencia de *P. organensis y P. misera* que presentan una diferencia genética mayor al 10%.

- Con respecto al Árbol Filogenético UPGMA

  En la **Figura 4** tenemos como resultado:

  1. **Estructura:**
     - El árbol se lee de izquierda a derecha.
     - Las líneas representan ramas evolutivas.
     - Los puntos donde las ramas se dividen se llaman nodos y representan ancestros comunes.
  2. **Relaciones Evolutivas:**
     - Las especies que comparten ramas más cercanas están más estrechamente relacionadas.
     - P. actinia, P. caerulea y P. elegans forman un grupo más cercano entre sí.
     - P. misera y P. organensis están más distantes evolutivamente del resto.
  3. **Interpretación de la distancia:**
     - Las ramas más largas indican mayor distancia evolutiva.
     - P. misera parece ser la especie más divergente del grupo.
  4. **Ancestros comunes:**
     - El nodo más a la izquierda representa el ancestro común de todas estas especies.
     - Cada bifurcación representa un punto de divergencia evolutiva.

- Con respecto al Árbol Filogenético de Máxima Verosimilitud

```
> # Mostrar los resultados de la optimización
> print(fit)
model: F81+G(4)
loglikelihood: -1457.812
unconstrained loglikelihood: -1535.235
Discrete gamma model
Number of rate categories: 4
Shape parameter: 999.9958

Rate matrix:
          a         c         g         t
a 0.0000000 0.6006270 1.0236998 0.6846836
c 0.6006270 0.0000000 0.5259954 1.3818226
g 1.0236998 0.5259954 0.0000000 1.0000000
t 0.6846836 1.3818226 1.0000000 0.0000000

Base frequencies:
        a         c         g         t
0.2314471 0.2986006 0.2631213 0.2068311
```

La siguiente salida muestra los detalles de la optimización del modelo filogenético ajustado.

- **model: F81+G(4)**: Indica que se utilizó el modelo Felsenstein 1981 (F81) con un modelo de gamma discretizado en 4 categorías para modelar la variación en las tasas de sustitución entre sitios.
- **loglikelihood: -1457.812**: Esta es la log-verosimilitud del árbol optimizado. Un valor más alto (menos negativo) indica un mejor ajuste del modelo a los datos.
- **unconstrained loglikelihood: -1535.235**: La log-verosimilitud del modelo sin restricciones. Este valor sirve como comparación para evaluar la mejora del ajuste del modelo con las restricciones impuestas.
- **Number of rate categories: 4**: Indica que el modelo gamma utiliza 4 categorías para discretizar la distribución gamma, que modela la variación en las tasas de sustitución entre sitios.
- **Shape parameter: 999.9958**: El parámetro de forma (alfa) de la distribución gamma. Un valor muy alto sugiere que las tasas de sustitución son bastante homogéneas entre los sitios (poca variación).

Si observamos la matriz:

```
Rate matrix:
          a         c         g         t
a 0.0000000 0.6006270 1.0236998 0.6846836
c 0.6006270 0.0000000 0.5259954 1.3818226
g 1.0236998 0.5259954 0.0000000 1.0000000
t 0.6846836 1.3818226 1.0000000 0.0000000
```

Las tasas de sustitución no son simétricas, lo que indica que la tasa de cambio de un nucleótido a otro puede diferir según la dirección del cambio.

Respecto a la frecuencia de las bases:

```
Base frequencies:
        a         c         g         t
0.2314471 0.2986006 0.2631213 0.2068311
```

Estas frecuencias indican que la base C (citosina) es la más común en el conjunto de datos, mientras que la base T (timina) es la menos común.

Entonces de acuerdo a estos resultados, podemos concluir que el modelo ha sido ajustado a los datos de secuencia y ha producido un árbol filogenético con una log-verosimilitud de -1457.812. La matriz de tasas muestra las tasas de sustitución entre diferentes pares de nucleótidos, y las frecuencias base proporcionan una idea de la composición de nucleótidos en el conjunto de datos. La discretización de la distribución gamma en 4 categorías permite modelar la variación en las tasas de sustitución entre los sitios de la secuencia. El parámetro de forma gamma alto sugiere que la variación entre los sitios es baja.

Ahora centrándonos más en el gráfico del árbol:

En la **Figura 5** tenemos lo siguiente:

1. **Estructura:**
   - El árbol está organizado jerárquicamente, donde cada nodo representa un ancestro común y las ramas representan la divergencia evolutiva entre diferentes especies.
   - Las longitudes de las ramas indican la cantidad de cambio evolutivo. Las ramas más largas representan mayor cantidad de cambios en las secuencias de ADN.
2. **Nodos:**
   - Los nodos internos (donde las ramas se dividen) representan los ancestros comunes de las secuencias que se bifurcan de ese punto.
3. **Especies Representadas:**
   - Rojo (KX378102.1): Esta secuencia se agrupa sola, lo que sugiere que es más distante evolutivamente de las otras secuencias.
   - Azul (KX378094.1) y Verde (KX378121.1): Estas dos secuencias están más cercanamente relacionadas entre sí que con las otras secuencias, indicando que comparten un ancestro común más reciente.
   - Morado (KX378057.1) y Naranja (KX378119.1): Estas secuencias también forman un grupo distinto, pero están separadas por un nodo que sugiere divergencia evolutiva.

Este árbol proporciona una representación visual de las relaciones evolutivas entre las diferentes especies de Pasiflora analizadas, indicando que algunas especies están más estrechamente relacionadas que otras y sugiriendo posibles caminos evolutivos a través de sus divergencias genéticas.

VI. CONCLUSIONES

- Todos los árboles filogenéticos coinciden en que las especies con código de accesión KX378057 (*P. organensis)* y KX378119 (*P. misera*) son cercanas evolutivamente.

- El árbol filogenético obtenido por el método Neighbor-Joint (NJ) y por Máxima Verosimilitud concluyen que las especies con código de accesión KX378094 (*P. actinia*) y KX378121 (*P. elegans*) son cercanas evolutivamente, además KX378102 (*P. caerulea*) sería el ancestro de las especies anteriormente mencionadas.

- El árbol filogenético obtenido por el método UPGMA concluye que las especies con código de accesión KX378094 (*P. actinia*) y KX378102 (*P. caerulea*) son cercanas evolutivamente, además KX378121 (*P. elegans*) sería cercana evolutivamente a estas dos especies simultáneamente.

- Gracias a los árboles filogenéticos generados se llegó a la conclusión de que las distintas especies de *Passiflora* son cercanas evolutivamente, lo que proporciona información valiosa para posteriores estudios de taxonomía y conservación de especies.

VII. BIBLIOGRAFÍA

Guantes, R., Aguirre, J., & Bajic, D. (2014). Biología de sistemas. En A. Sebastián & A. Pascual-García (Eds.), *Bioinformática con Ñ* (Cap. 9). Arquitectura Viva.

Zou, Y., Zhang, Z., Zeng, Y., Hu, H., Hao, Y., Huang, S., & Li, B. (2024). Common methods for phylogenetic tree construction and their implementation in R. Bioengineering, 11(480).