



Genómica Computacional

Reporte de Proyecto : *Similitudes de algunos elefantes modernos con los mamuts*

Sebastián Alejandro Gutiérrez Medina
sebasguti1511@ciencias.unam.mx

Alejandro Terrazas Rivera
alexterrazas@ciencias.unam.mx

Aquino Chapa Armando Abraham
armandoaac@ciencias.unam.mx

13 de junio de 2024

1. Introducción

La historia evolutiva de los elefantes y sus parientes extintos, como el mamut lanudo (*Mammuthus primigenius*), que vagó por las frías y secas regiones de la estepa-tundra hasta hace unos 4.000 años, ha sido una de las especies extinguidas más estudiadas, ha fascinado a los científicos durante décadas, pues las adaptaciones del mamut a las duras condiciones árticas, como un pelaje espeso, orejas pequeñas y una importante capa de grasa, difieren de gran manera con los elefantes africanos (*Loxodonta africana*) y asiáticos (*Elephas maximus*) modernos, que habitan en climas más cálidos.

Gracias a avances en genómica junto con la obtención del DNA de varios especímenes de mamuts se ha logrado secuenciar su DNA[2], lo que ha permitido explorar con más detalle las relaciones evolutivas entre la familia *Elephantidae*, gracias a esto, los estudios genómicos comparativos, incluidos los análisis del DNA mitocondrial (mtDNA), han sugerido que los mamuts comparten una relación evolutiva más estrecha con los elefantes asiáticos que con los africanos, lo cual se ve respaldado por hallazgos que indican una divergencia genética entre los mamuts y los elefantes asiáticos hace aproximadamente entre 5,8 y 7,8 millones de años, mientras que los elefantes africanos divergieron antes, hace entre 6,6 y 8,8 millones de años.

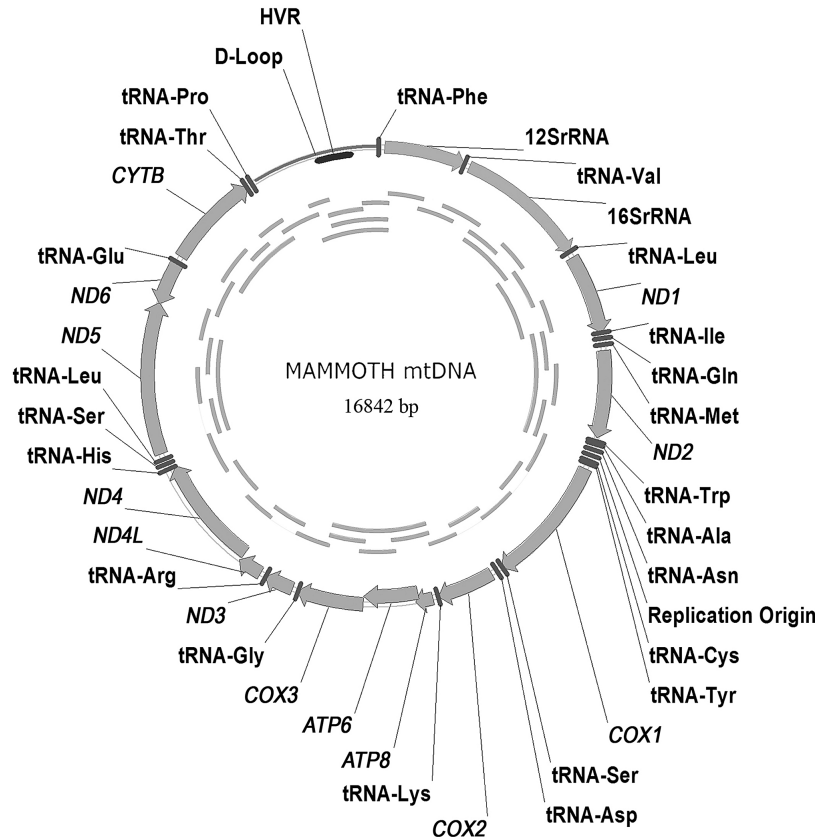
Las particulares estructuras sociales matrilineales de los elefantes, en las que las hembras rara vez se dispersan de sus manadas, dan lugar a genealogías de DNA mitocondrial distintas que se diferencian de gran manera de los patrones de DNA nuclear.

Pero, ¿qué es el DNA mitocondrial?, para responder esto debemos mencionar que son las mitocondrias y sus funciones en una célula. Las mitocondrias son orgánulos de células eucariontes implicadas en la producción de energía de las células y se constituyen de su propio cromosoma circular (a diferencia del cromosoma en forma de "X" del DNA nuclear) de 16569 de bases para los humanos, que contiene 37 genes codificantes de proteínas y genes de ARN.

A diferencia del ADN nuclear, sólo podemos heredar DNA mitocondrial de nuestras madres, por lo que solo tenemos una versión a comparación del genoma nuclear, donde tenemos la versión de nuestro padre y madre. Cada célula contiene múltiples copias de este mtDNA.

En el caso del mamut lanudo, el genoma mitocondrial del mamut contiene[3]:

- 13 genes codificadores de proteínas
- 22 genes de ARNt
- Dos genes de ARNr
- La región de control D-loop



A lo largo de este proyecto se tiene como objetivo aprovechar las secuencias del genoma mitocondrial para determinar entre diferentes especies actuales de la familia *Elephantidae*, mas específicamente el elefante africano (*Loxodonta africana*), elefante asiático (*Elephas maximus*) y la Musaraña Elefante (*Elpehantulus sp. VB001*), cual está más estrechamente emparentada con el mamut lanudo (*Mammuthus primigenius*), lo que permitirá comprender mejor su historia evolutiva y sus adaptaciones.

Para ello realizaremos el siguiente procedimiento:

1. Eliminar las HVR de las secuencias de mtDNA.
2. Alineamiento global de las secuencias de mtDNA.
3. Estimar la distancia genética utilizando el modelo de Jukes-Cantor.
4. Creación de gráficos para una mejor apreciación de los resultados.

2. Métodos

2.1. Recolección de datos:

Se realizó una colecta de secuencias de mtDNA correspondientes al mamut lanudo (*Mammuthus primigenius*), elefante africano (*Loxodonta africana*), elefante asiático (*Elephas maximus*) y musaraña elefante (*Elephantulus sp. VB001*) en el GenBank del NCBI. De igual forma, se realizó una colecta de fragmentos de regiones hipervariables (*HVR*) para las especies mencionadas, aunque no se encontraron fragmentos de *HVR* para la musaraña elefante.





Nombre	Imagen	mtDNA	Fragmento <i>HVR</i>
<i>Mammuthus primigenius</i>		mtDNA	HVR
<i>Loxodonta africana</i>		mtDNA	HVR
<i>Elephas maximus</i>		mtDNA	HVR
<i>Elephantulus sp. VB001</i>		mtDNA	No se encontró

Tabla 1: Información de la recolección de datos (clic sobre las secuencias para ir al NCBI)

2.2. Eliminación de las regiones hipervariables (*HVR*) de las secuencias:

El DNA mitocondrial cuenta con una región específica, que no contiene ningún gen, conocida como D-loop o región de control de aproximadamente 1,1 kb que es una secuencia no codificante en el genoma mitocondrial, pero es fundamental para la replicación y la transcripción del DNA mitocondrial. El D-loop contiene también dos regiones hipervariables sin función conocida: denominadas como regiones I y II (*HVR-I* y *HVR-II*). Estas regiones constituyen la mayor parte del D-loop (cada uno tiene una longitud de entre 400-500 pb)[1].

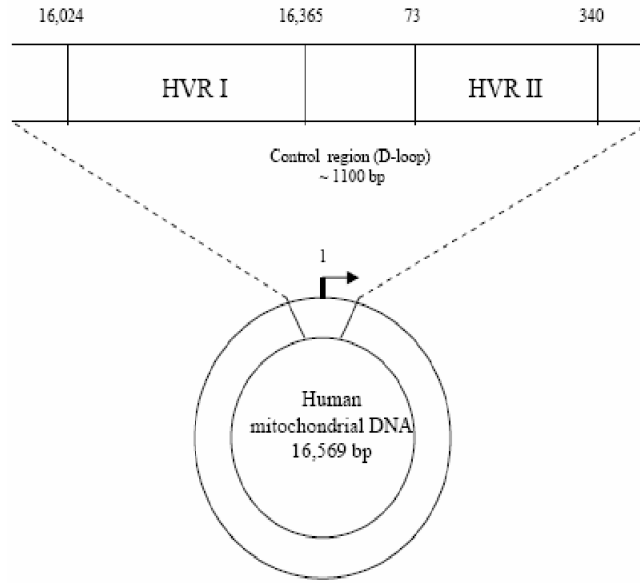


Figura 1: D-loop del mtDNA del ser humano

Cómo se aprecia en la imagen, en el caso de los humanos las dos HVR tienen posiciones aproximadamente 16024-16400 (para HVR-I) y posiciones 1-500 (para HVR-II).

Las regiones hipervariables al no tener funciones conocidas las eliminaremos de nuestras secuencias. Desafortunadamente no encontramos las regiones completas, sino fragmentos, por lo que sólo eliminaremos estos fragmentos encontrados.

2.3. Alineamiento global de las secuencias:

El propósito del alineamiento de secuencias es medir su similitud para ver que tan parecidas o que tan diferentes pueden ser las secuencias entre sí. En nuestro caso, esto nos ayudará para indicarnos que tan estrechamente emparentados pueden estar los elefantes que vamos a analizar con el mamut lanudo.

Nosotros utilizamos el algoritmo de Needleman-Wunsch que fue tratado durante el curso, el cual consiste en:

- Crear una tabla de tamaño $(m + 1) \times (n + 1)$ para las secuencias digamos s y t de longitud m y n . El espacio extra es para la inserción del gap.
- Completar las entradas de la tabla $(m : 1)$ y $(1 : n)$ con la fórmula:

$$M_{i,1} = \sum_{k=1}^i \sigma(s_k, -)$$

$$M_{1,j} = \sum_{k=1}^j \sigma(-, t_k)$$

- Comenzando desde la parte superior izquierda, calcular cada entrada usando la relación recursiva:

$$M_{i,j} = \max \begin{cases} M_{i-1,j} + \sigma(s_i, -) & \text{Fila de arriba} \\ M_{i,j-1} + \sigma(-, t_j) & \text{Columna izquierda} \\ M_{i-1,j-1} + \sigma(s_i, t_j) & \text{Diagonal} \end{cases}$$

- Realizar el procedimiento de trace-back para obtener el alineamiento óptimo recordando cada elección de la relación recursiva. Para cada paso diagonal hay una correspondencia o discrepancia (match/mismatch) en el

alineamiento: Para cada paso vertical se inserta un gap en la secuencia superior: Para cada paso horizontal se inserta un gap en la secuencia del costado[4].

Por ejemplo, supongamos que queremos realizar el alineamiento de las secuencias $s = ATACAC$ y $t = AAGCGC$. Siguiendo el algoritmo de Needleman-Wunsch con $\sigma(s_i, -) = -1$, $\sigma(-, t_j) = -1$ y $\sigma(s_i, t_j) = \pm 1$:

s \ t	-	A	A	G	C	G	C
-	0	-1	-2	-3	-4	-5	-6
A	-1	1	0	-1	-2	-3	-4
T	-2	0	0	-1	-2	-3	-4
A	-3	-1	1	0	-1	-2	-3
C	-4	-2	0	0	1	0	-1
A	-5	-3	-1	-1	0	0	-1
A	-6	-4	-2	-2	0	-1	1

s \ t	-	A	A	G	C	G	C
-	.	←	←	←	←	←	←
A	↑	↖	↖	←	←	←	←
T	↑	↑	↖	↖	↖	↖	↖
A	↑	↖	↖	←	←	←	←
C	↑	↑	↑	↖	↖	←	↖
A	↑	↖	↖	↖	↑	↖	↖
A	↑	↑	↑	↖	↖	↖	↖

Segunda Fila: $\max(-2, -2, 1) = 1$ $\max(0, 0, -3) = 0$ $\max(-1, -3, -4) = -1$
 $\max(-2, -4, -5) = -2$ $\max(-3, -5, -6) = -3$ $\max(-4, -6, -7) = -4$

Tercer Fila: $\max(-3, -2, 0) = 0$ $\max(-1, 0, -1) = 0$ $\max(-1, -1, -2) = -1$
 $\max(-2, -2, -3) = -2$ $\max(-3, -3, -4) = -3$ $\max(-4, -4, -5) = -4$

Cuarta Fila: $\max(-4, -1, -1) = -1$ $\max(-2, 1, 1) = 1$ $\max(0, -1, -2) = 0$
 $\max(-1, -2, -3) = -1$ $\max(-2, -3, -4) = -2$ $\max(-3, -4, -5) = -3$

Quinta Fila: $\max(-5, -4, -2) = -2$ $\max(-3, -2, 0) = 0$ $\max(-2, 0, -1) = 0$
 $\max(-1, 1, -2) = 1$ $\max(0, -2, -3) = 0$ $\max(-1, -1, -4) = -1$

Sexta Fila: $\max(-6, -3, -3) = -3$ $\max(-4, -1, -1) = -1$ $\max(-2, -1, -1) = -1$
 $\max(-2, -1, 0) = 0$ $\max(-1, 0, -1) = 0$ $\max(-1, -1, -2) = -1$

Figura 2: Algoritmo Needleman-Wusch para $s = ATACAC$ y $t = AAGCGC$

La matriz de la izquierda representa el score del alineamiento óptimo (máximo), mientras que la matriz de la derecha contiene las flechas que realizan el seguimiento de la mejor elección realizada para obtener el score óptimo. Las marcas amarillas corresponden a la alineación final óptima.

Por tanto, la alineación óptima con un score igual a 1 es:

$$\begin{aligned} s' &= ATA - CAC \\ t' &= A - AGCGC \end{aligned}$$

2.4. Estimación de la distancia genética utilizando el modelo de Jukes-Cantor

El modelo de Jukes-Cantor es uno de los modelos más utilizados para estimar distancias genéticas entre secuencias de ADN. Fue propuesto por Jukes y Cantor en 1969 y se basa en la suposición de que todos los nucleótidos cambian a otros nucleótidos con la misma probabilidad. Un simple conteo de diferencias entre dos secuencias no refleja completamente su distancia genética debido a la posibilidad de múltiples sustituciones en la misma posición a lo largo del tiempo. El modelo de Jukes-Cantor ajusta este conteo para proporcionar una estimación más precisa de las sustituciones reales.

2.4.1. Fórmula del Modelo de Jukes-Cantor

El modelo de Jukes-Cantor es un modelo de sustitución de un solo parámetro que asume que:

- Todas las posiciones de la secuencia evolucionan de manera independiente.



- La probabilidad de sustitución de un nucleótido por otro es la misma para todas las posiciones y todos los pares de nucleótidos.
- No se considera la posibilidad de inserciones o deleciones.

La fórmula utilizada para calcular la distancia genética d entre dos secuencias es:

$$d = -\frac{3}{4} \ln \left(1 - \frac{4}{3}p \right)$$

Donde p es la proporción de sitios diferentes entre las dos secuencias.

2.4.2. Derivación del Modelo de Jukes-Cantor

Para derivar la fórmula de Jukes-Cantor, consideramos que todas las posiciones en una secuencia evolucionan independientemente y nos centramos en una posición en particular. Queremos determinar la probabilidad de una sustitución en esta posición en un tiempo dado, denominada α . Asumiendo que todas las posibles sustituciones desde una base a cualquiera de las otras tres son igualmente probables, podemos describir la cadena de Markov de un paso con la siguiente matriz de transición:

$$M_{JC} = \begin{pmatrix} 1-\alpha & \alpha/3 & \alpha/3 & \alpha/3 \\ \alpha/3 & 1-\alpha & \alpha/3 & \alpha/3 \\ \alpha/3 & \alpha/3 & 1-\alpha & \alpha/3 \\ \alpha/3 & \alpha/3 & \alpha/3 & 1-\alpha \end{pmatrix}$$

La teoría de procesos de Markov indica que la probabilidad de una sustitución después de t pasos de tiempo puede calcularse como $M(t) = M^t$. Esta matriz se puede expresar en términos de sus valores propios λ_i y vectores propios v_i :

$$M_{JC}(t) = \sum_{i=1}^4 \lambda_i^t v_i v_i^T$$

Los valores propios son $\lambda_1 = 1$ y $\lambda_2, \lambda_3, \lambda_4 = 1 - \frac{4}{3}\alpha$, y los vectores propios son:

$$\begin{aligned} v_1 &= \frac{1}{4}(1, 1, 1, 1) \\ v_2 &= \frac{1}{4}(-1, -1, 1, 1) \\ v_3 &= \frac{1}{4}(1, -1, -1, 1) \\ v_4 &= \frac{1}{4}(1, -1, 1, -1) \end{aligned}$$

Sustituyendo estos valores en la ecuación, obtenemos:

$$M_{JC}(t) = \begin{pmatrix} r(t) & s(t) & s(t) & s(t) \\ s(t) & r(t) & s(t) & s(t) \\ s(t) & s(t) & r(t) & s(t) \\ s(t) & s(t) & s(t) & r(t) \end{pmatrix}$$

donde $s(t) = \frac{1}{4} - \frac{1}{4}(1 - \frac{4}{3}\alpha)^t$.



2.4.3. Estimación de la Distancia Genética

Esta matriz se puede traducir directamente en la proporción de diferencias observadas entre dos secuencias, d . Queremos expresar t como una función de d . Usando la aproximación $\ln(1+x) \approx x$ para x pequeño, obtenemos:

$$t \approx -\frac{3}{4\alpha} \ln\left(1 - \frac{4}{3}d\right)$$

Estimamos que si la probabilidad de sustitución por paso de tiempo es α , entonces en t pasos de tiempo, la distancia genética esperada es:

$$K = t\alpha$$

Finalmente, eliminando t , obtenemos la fórmula de Jukes-Cantor:

$$K \approx -\frac{3}{4} \ln\left(1 - \frac{4}{3}d\right)$$

Esta relación implica que K tiende a infinito cuando d tiende a $\frac{3}{4}$, la distancia esperada entre secuencias no relacionadas, y da $K \approx d$ para secuencias muy similares.

2.4.4. Varianza de la Estimación

La varianza de K se puede estimar usando el método delta:

$$\text{Var}(K) \approx \left(\frac{\partial K}{\partial d}\right)^2 \text{Var}(d)$$

donde

$$\text{Var}(d) = \frac{d(1-d)}{n}$$

y

$$\frac{\partial K}{\partial d} = \frac{1}{1 - \frac{4}{3}d}$$

Por lo tanto,

$$\text{Var}(K) \approx \frac{d(1-d)}{n(1 - \frac{4}{3}d)^2}$$

donde n es la longitud de la secuencia.

2.4.5. Aplicación del Modelo a Secuencias de Ejemplo

Para ilustrar el uso del modelo, se calcularon las distancias genéticas entre secuencias de diferentes especies de elefantes y una musaraña utilizando archivos de secuencias en formato de texto. Los resultados se almacenan y visualizan en gráficos.

3. Resultados

A continuación, se presentan los resultados obtenidos del análisis genómico comparativo entre el mamut lanudo (*Mammuthus primigenius*), el elefante africano (*Loxodonta africana*), el elefante asiático (*Elephas maximus*) y la musaraña elefante (*Elephantulus sp. VB001*).

3.1. Eliminación de las regiones hipervariables (HVR)

Cómo se mencionó anteriormente, realizamos una colecta de las HVR del mtDNA de los elefantes descritos, pero desafortunadamente sólo encontramos fragmentos para algunos de ellos:





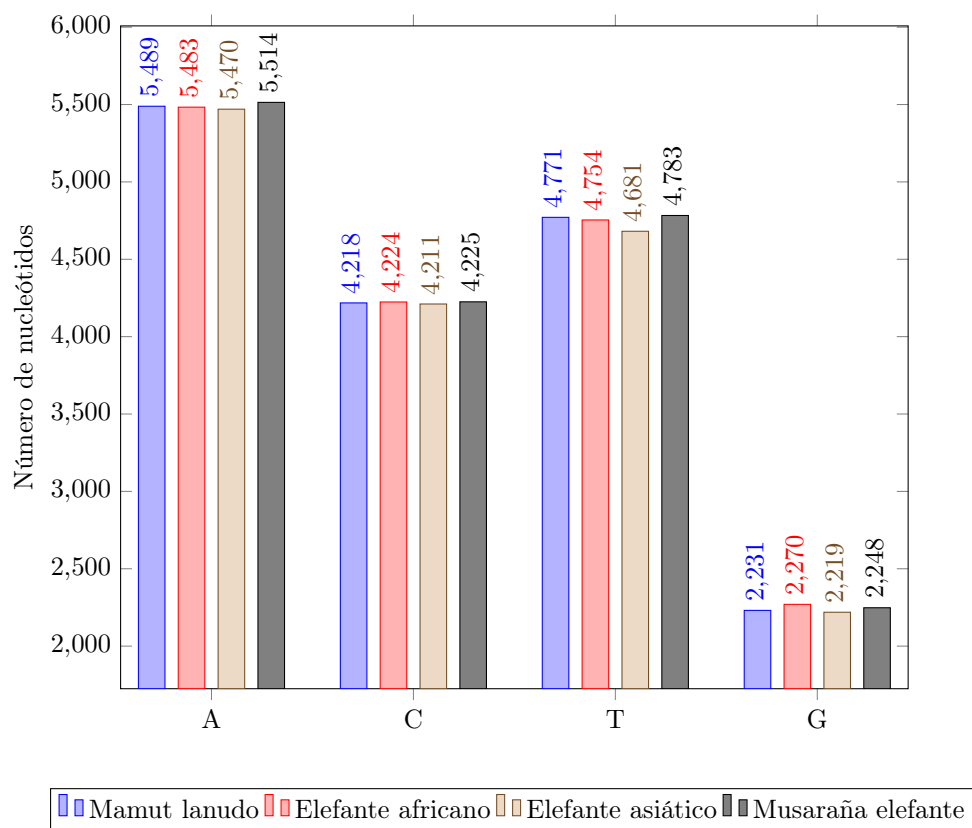
Nombre	Imagen	mtDNA	Frag. HVR	Pos. HVR	mtDNA sin HVR
<i>Mammuthus primigenius</i>		16770 pb	61 pb	16585-16645	16709
<i>Loxodonta africana</i>		16866 pb	135 pb	15422-15556	16731
<i>Elephas maximus</i>		16902 pb	321 pb	15418-15738	16581
<i>Elephantulus sp. VB001</i>		16677 pb	-	-	-

Tabla 2: Información sobre el mtDNA de los elefantes y mamut después de la eliminación de fragmentos HVR

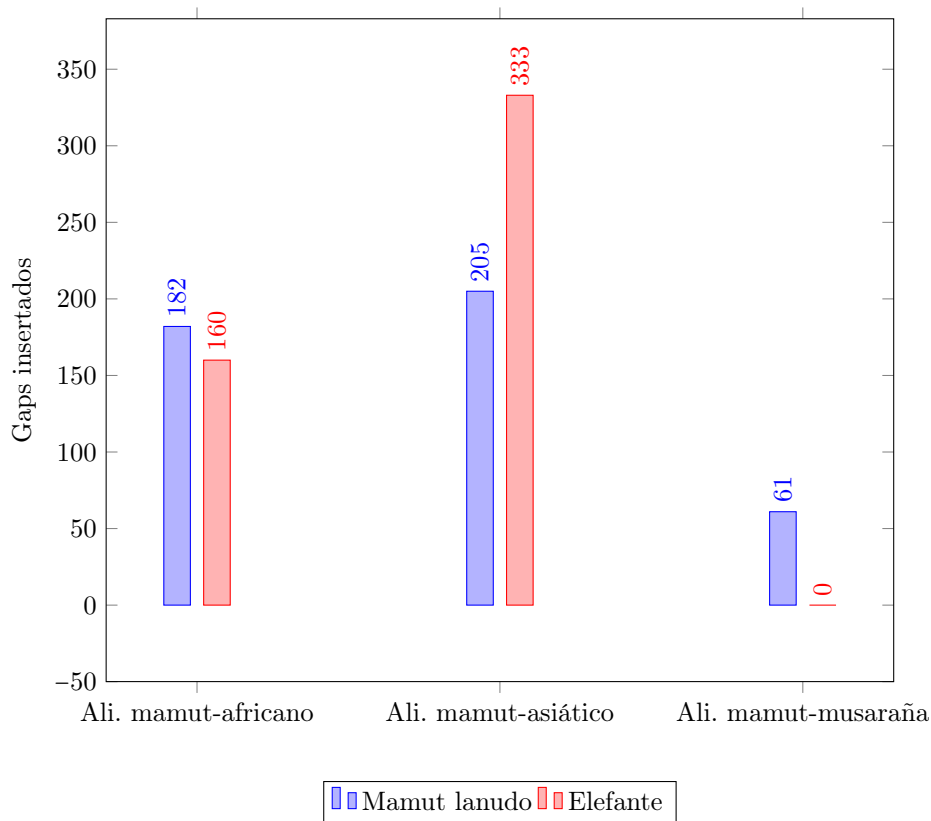
3.2. Alineamiento global de las secuencias

Después de la eliminación de las regiones hipervariables (HVR) de las secuencias de mtDNA recolectadas, procedimos con el alineamiento global de las secuencias utilizando el algoritmo de Needleman-Wunsch. Esta etapa es crucial para evitar sesgos en las estimaciones de las distancias genéticas debido a las regiones altamente variables y no codificantes.

Primero, veamos una tabla que compara el número de nucleótidos que tiene cada secuencia de mtDNA de los elefantes y el mamut:



El alineamiento de las secuencias de mtDNA ajustadas (sin HVR) nos permitió medir la similitud entre las secuencias y obtener una idea clara de las relaciones evolutivas entre las especies analizadas. A continuación se muestran los resultados del alineamiento donde se muestra cuantos gaps fueron insertados en cada secuencia por cada alineamiento, ya que recordemos que nosotros estamos utilizando el algoritmo de Needleman-Wunsch que solo permite dos secuencias, no estamos realizando un alineamiento múltiple:



3.3. Estimación de distancias genéticas

Utilizando el modelo de Jukes-Cantor, estimamos las distancias genéticas entre cada par de secuencias. Las distancias calculadas son las siguientes:

Par de Especies	Distancia Genética
<i>Mammuthus primigenius</i> y <i>Loxodonta africana</i>	0.0669
<i>Mammuthus primigenius</i> y <i>Elephas maximus</i>	0.0760
<i>Mammuthus primigenius</i> y <i>Elephantulus sp. VB001</i>	0.0036

Tabla 3: Distancias genéticas estimadas utilizando el modelo de Jukes-Cantor.

3.4. Visualización de resultados

Para una mejor apreciación de los resultados, se crearon gráficos que ilustran las distancias genéticas entre las especies analizadas. A continuación se presentan estos gráficos:

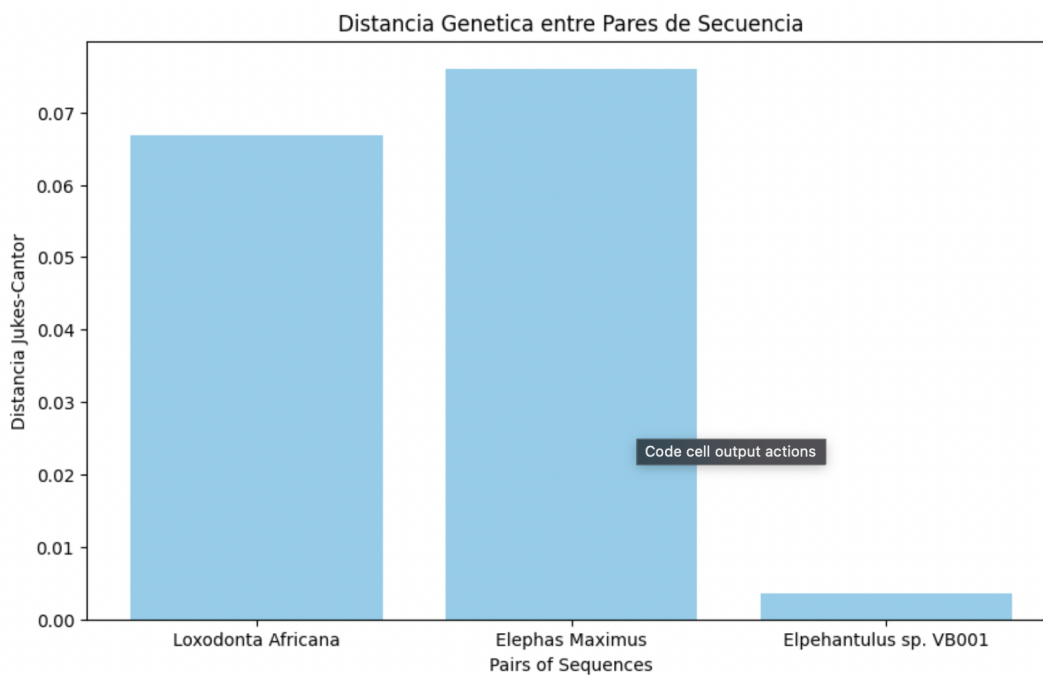


Figura 3: Distancia Genética entre Pares de Secuencia

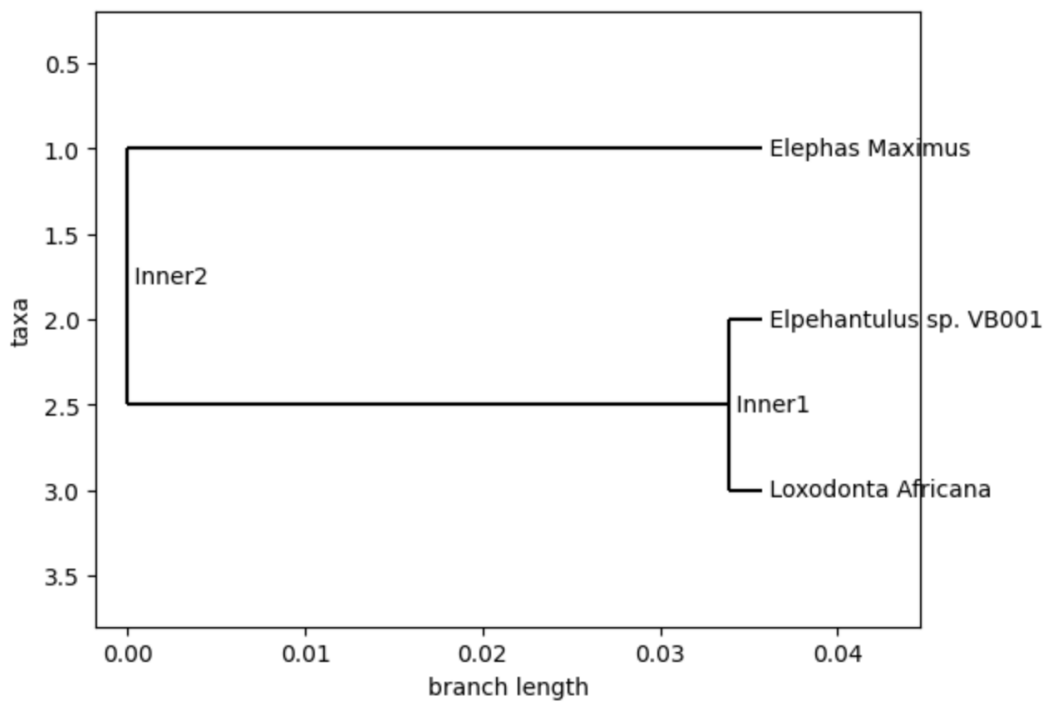


Figura 4: Árbol Filogenético Basado en Distancias Jukes-Cantor

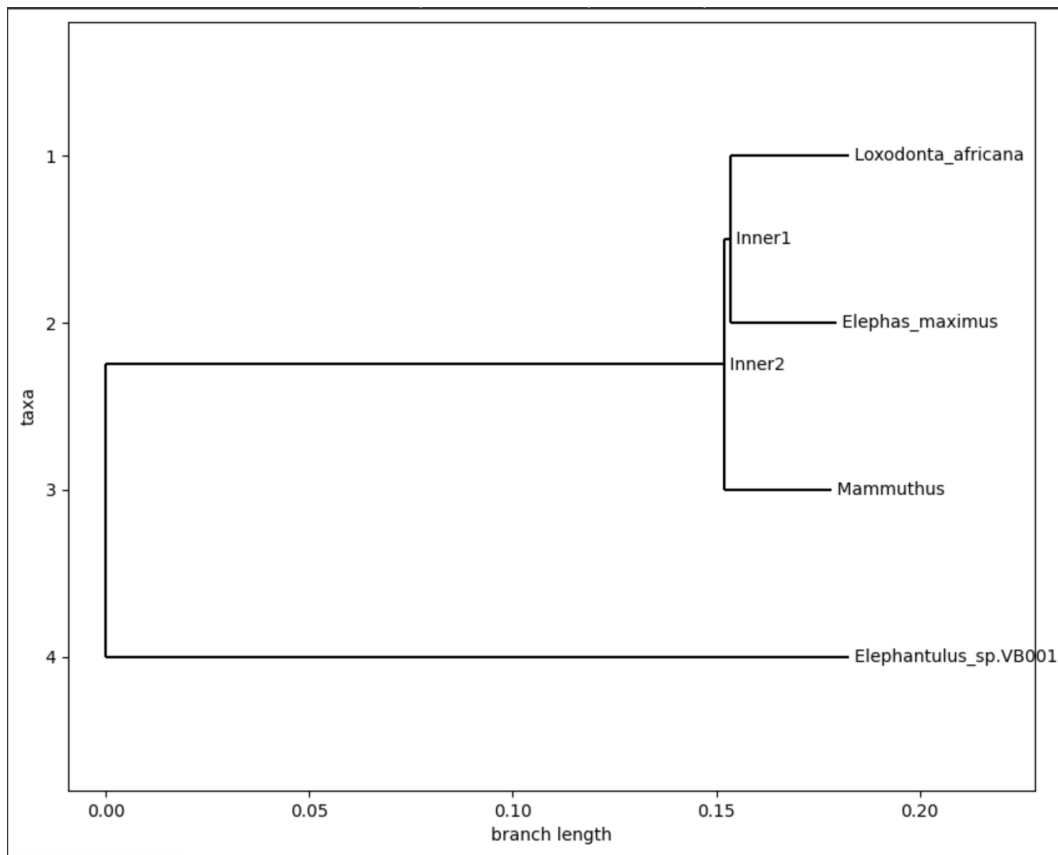


Figura 5: Árbol Filogenético Basado en Distancias Jukes-Cantor

Las distancias genéticas más pequeñas indican una mayor similitud entre las secuencias y, por ende, una relación evolutiva más estrecha. Observamos que la distancia genética entre el mamut lanudo y el elefante asiático es menor que la distancia entre el mamut lanudo y el elefante africano. Esto sugiere que el mamut lanudo está más estrechamente emparentado con el elefante asiático.

4. Discusión

En el artículo *Complete Mitochondrial Genome and Phylogeny of Pleistocene Mammoth* *Mammuthus primigenius* se presenta la secuencia del genoma mitocondrial de un mamut lanudo encontrado en Siberia donde fueron extraídos restos de la espinilla y articulación del tobillo conservados de la época del Pleistoceno, siendo la secuencia de genoma mitocondrial más antigua determinada hasta esa fecha. En este estudio se proporciona una reconstrucción filogenética de la familia *Elephantinae* que sugiere que *Mammuthus primigenius* y *Elephas maximus* son especies hermanas que divergieron poco después de que su ancestro común se separara de la línea de *Loxodonta africana*[3].

Esto nos indica que el mamut lanudo (*Mammuthus primigenius*) y el elefante asiático (*Elephas maximus*) son especies hermanas que divergieron después de separarse del elefante africano (*Loxodonta africana*), por lo cual el artículo nos proporciona una gran comprensión de cómo los mamuts lanudos están relacionados con los elefantes modernos y sustenta nuestros resultados de que el elefante asiático (*Elephas maximus*) está más estrechamente emparentado con el mamut lanudo ya que se separó después del elefante africano.



Es muy interesante cómo para llegar a resultados similares realizamos procedimientos mucho más sencillos tomando en cuenta que nuestro objetivo es mucho más acotado que el del artículo, ya que para llegar a los resultados se tuvieron que realizar procedimientos más complicados y complejos cómo PCR para determinar la secuencia completa de mtDNA del mamut, o también que realizó una alineación múltiple (a diferencia de nosotros que utilizamos el algoritmo de Needleman-Wunsch que sólo permite la alineación de dos secuencias) de los genomas mitocondriales completos utilizando el programa MUSCLE, para posteriormente utilizar este resultado para realizar el análisis filogenético.

5. Conclusión

A partir del árbol filogenético y las distancias genéticas, podemos concluir que **el elefante asiático (*Elephas maximus*) es el que se encuentra más estrechamente emparentado con el mamut lanudo (*Mammuthus primigenius*)**. Gracias a las distancias genéticas obtenidas, podemos determinar que de todos los elefantes analizados, el elefante asiático es el más parecido al mamut. Esto a su vez nos puede indicar de que este elefante y mamut compartieron un ancestro común más reciente antes de ramificarse (lo cuál se muestra en el artículo mencionado en la sección de Discusión) del elefante africano (*Loxodonta africana*).

Referencias

- [1] Nello Cristianini y Matthew W.Hahn. *Introduction to Computational Genomics*. Cambridge University Press, 2006.
- [2] Webb Miller et al. “Sequencing the nuclear genome of the extinct woolly mammoth”. En: *Nature* 456.7220 (2008), págs. 387-390.
- [3] Yuri K. Moliaka et al. “Complete Mitochondrial Genome and Phylogeny of Pleistocene Mammoth *Mammuthus primigenius*”. En: *PLoS Biology* (2006).
- [4] Shoba Ranganathan et al. *Encyclopedia of Bioinformatics and Computational Biology*. Elsevier, 2019.