Proyecto integrador en Bioinformática

Entrega 3

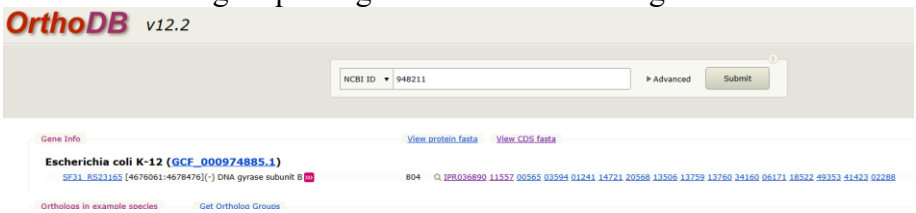
Laura del Sol González
Sara Flórez Hernández
Maria Camila Gómez Villegas
Sergio Andrés Morales Toro

Universidad EIA
Envigado – Antioquia
Noviembre 2025

Descripción del flujo de trabajo: Tener en cuenta esta pregunta ¿Cuál es el mejor modelo de sustitución que pudieron encontrar para sus datos? Explicarlo brevemente.

El flujo de trabajo para el análisis filogenético utilizando el gen de la gyrB sigue la siguiente metodología.

Inicialmente se obtuvo la secuencia de referencia de Escherichia coli K-12 substr. MG1655 (Gene ID: 948211) desde la base de datos NCBI. Después en la página de OrthoDB para este gen se descargaron los CDS fasta ya que estos son los que contienen las secuencias codificantes del gen que se guardaron como ortologs.fasta.



Para asegurar la independencia filogenética y la calidad del conjunto de datos, se hizo la siguiente depuración.

- Se eliminaron las secuencias duplicadas, guardándolas en el archivo orto_nodups.fasta y se aplicó un filtro de longitud mínima de 150 nucleótidos, dando como resultado orto_minlen.fasta. Este umbral se estableció para facilitar el procesamiento computacional. Después se usó la herramienta CD-HIT para reducir la redundancia de secuencias y solo dejar las representativas, obteniendo el archivo orto_clean.fasta (). CD-HIT utiliza un algoritmo de eliminación de listas que prioriza la secuencia más larga para descartar secuencias con una identidad superior a un umbral determinado importante para mitigar sesgos en la reconstrucción filogenética. Finalmente, se seleccionó un subconjunto representativo de 100 secuencias, almacenado en orto_final.fasta, para optimizar el balance entre diversidad taxonómica. Los identificadores de estas secuencias fueron estandarizados para mejorar la legibilidad, se conservaron solo los nombres de los organismos al inicio de cada secuencia, generando el archivo orto_final_names.fasta.

Para enraizar el árbol filogenético, se incorporó la secuencia de Pseudomonas aeruginosa PAO1 tomada de NCBI como grupo externo u outgroup. La elección de este taxón, pertenece a las Gammaproteobacteria con una divergencia temprana respecto al linaje de E. coli, lo que permite polarizar correctamente las relaciones evolutivas en el árbol. También de la página de OrthoDB se descargaron los CDS fasta y se les dió el mismo tratamiento que a las secuencias anteriores. El conjunto de datos final, se le hizo un alineamiento múltiple (MSA) con la herramienta MAFFT, resultando en el archivo orto_alineados_tag1.fasta. Este alineamiento es la base para todos los análisis evolutivos siguientes, ya que establece la homología de posición nucleotídica.

La selección del modelo de sustitución nucleotídica óptimo se realizó mediante el software IQ-TREE, que evalúa una serie de modelos bajo criterios de información estadística. El objetivo fue identificar el modelo que mejor describe los patrones de cambio evolutivo en

los datos alineados. La selección se basó en …(ampliar) cuyas fórmulas son …(ampliar). El modelo con los valores ... más bajos es considerado el de mejor ajuste. Best-fit model according to BIC: GTR+F+I+G4

La elección de un modelo …(ampliar) permite una estimación más precisa de las longitudes de rama y la topología en comparación con modelos más restrictivos. El reporte completo de este análisis, que incluye los parámetros del modelo seleccionado, se guardó en orto_alineados_tag1.fasta.iqtree.

Después, se procedió a calcular una matriz de distancias evolutivas, la cual cuantifica el número esperado de sustituciones por sitio entre cada par de secuencias, corregido por el modelo seleccionado. Esta matriz se almacenó en el archivo orto_alineados_tag1.fasta.mldist (10). Finalmente, la reconstrucción del árbol filogenético se realizó mediante el método de Máxima Verosimilitud (ML) implementado en IQ-TREE . El principio de ML busca encontrar la topología del árbol y las longitudes de rama que maximicen la probabilidad de observar el alineamiento de secuencias dado, bajo el modelo de sustitución ..... La función de verosimilitud se define como $L(T, M \mid D) = P(D \mid T, M)$, donde T representa la topología y longitudes de rama, M el modelo de sustitución con sus parámetros, y D los datos del alineamiento. El algoritmo calcula la probabilidad de los datos en cada sitio de forma independiente y combina estos valores para identificar el árbol con la verosimilitud global más alta. El árbol filogenético resultante, constituye el resultado final de este análisis, permitiendo la inferencia de relaciones evolutivas entre los taxones estudiados.

Resultados obtenidos:

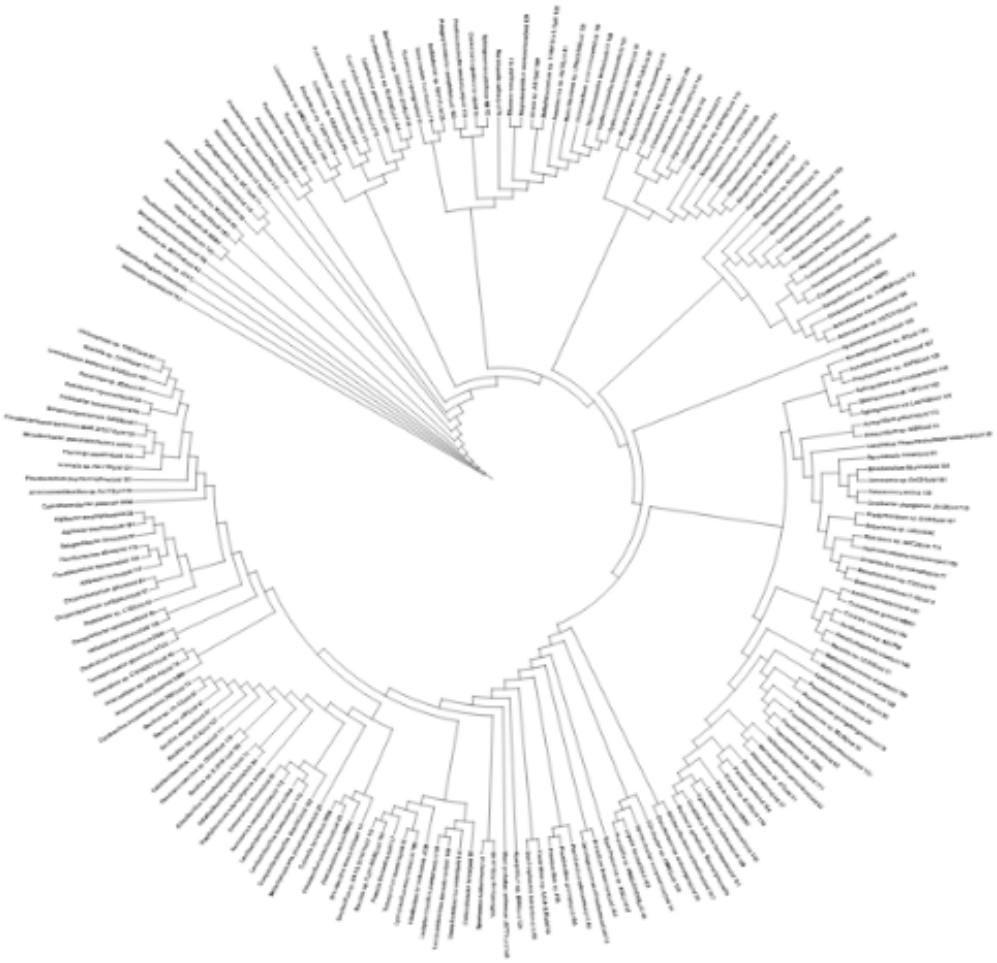


Figura1. Árbol filogenético del gen gyrB en *Escherichia coli* y *Pseudomonas aeruginosa* con visualización en IQTree

El gen escogido fue gyrB que es un gen específico que codifica para la subunidad B de la ADN girasa, se encuentra en una zona llamada QRDR (región determinante de resistencia a quinolonas) específica para la subunidad (De la Fuente, y otros, 2007). Este gen tiene una importancia significativa como marcador molecular para la identificación y filogenia de bacterias (Peeters & Willems, 2011). En el árbol filogenético obtenido, cada rama representa a una especie que tiene una secuencia del gen gyrB. El árbol muestra cómo se relacionan evolutivamente esas secuencias del gen gyrB, según su similitud o divergencia.

En el árbol filogenético obtenido, cada rama representa una especie que contiene la secuencia del gen gyrB. La estructura del árbol refleja las relaciones evolutivas entre esas secuencias, según su grado de similitud o divergencia.

Como punto de referencia se utilizó *Pseudomonas aeruginosa* como outgroup, con el fin de orientar la raíz del árbol y establecer la dirección de los cambios evolutivos. La elección de esta especie se debió a su lejanía filogenética con *E. coli* y a que también posee el gen gyrB, lo que permite comparar secuencias homólogas. En el árbol, las especies que aparecen más

cercanas entre sí tienen secuencias del gen gyrB más similares, mientras que aquellas situadas en ramas más alejadas presentan mayor divergencia.

Por ejemplo, especies como *Gilliamella apicola* y *Candidatus Regiella insecticola* se ubican cerca de la raíz, mostrando una mayor similitud con el outgroup, mientras que *Niastella* sp. y *Chitinophaga* sp. presentan secuencias del gen gyrB más divergentes con respecto a *P. aeruginosa*.

El objetivo principal del análisis fue examinar la filogenia de *Escherichia coli* a partir del gen gyrB, utilizando el árbol obtenido para inferir la dirección de los cambios evolutivos con base en la raíz establecida previamente mediante *Pseudomonas aeruginosa*. Dado que *E. coli* pertenece a la familia Enterobacteriaceae, se esperaba que se agrupara en un clado cercano a especies como *Serratia* sp. y *Klebsiella* sp., lo cual sería coherente con su relación filogenética y su pertenencia al mismo grupo taxonómico. Sin embargo, en el árbol obtenido no se observó la presencia de *E. coli* ni de sus variantes comunes en bases de datos, como *E. coli* K12, *E. coli* O157:H7 o *Escherichia* sp. *strain X*. Esta ausencia podría deberse a errores durante la fase procedimental, por ejemplo, a la exclusión accidental de la secuencia de *E. coli* en el momento del alineamiento, o a que el software seleccionara secuencias homólogas de gyrB sin incluir exactamente la correspondiente a *E. coli*.

Aun así, el árbol filogenético permitió identificar diferentes familias bacterianas pertenecientes a diversos filos, entre ellas *Enterobacteriaceae*, *Vibrionaceae*, *Pseudomonadaceae*, *Burkholderiaceae*, *Rhizobiaceae*, *Bradyrhizobiaceae*, *Actinomycetaceae*, *Streptomycetaceae*, *Flavobacteriaceae*, *Paenibacillaceae*, *Bacillaceae*, *Lactobacillaceae*, *Clostridiaceae*, *Helicobacteraceae* y *Synechococcaceae*, entre otras.
Entre las bacterias del árbol, aquellas pertenecientes a las familias Vibrionaceae (*Vibrio* spp.), Pseudomonadaceae (*Pseudomonas* spp.), Halomonadaceae (*Halomonas*) y Acinetobacteraceae (*Acinetobacter*) se ubican dentro de la clase Gammaproteobacteria, la misma que *E. coli*. Aunque comparten el mismo filo (Proteobacteria), presentan una distancia evolutiva mayor, lo que indica una separación evolutiva más antigua dentro del grupo.

También se observaron familias correspondientes a otras clases dentro del filo Proteobacteria, como las Alphaproteobacteria (*Rhizobium*, *Sphingomonas*, *Paracoccus*, *Bradyrhizobium*) y las Betaproteobacteria (*Burkholderia*, *Cupriavidus*, *Acidovorax*). Estas familias presentan una relación filogenética más lejana con *E. coli*, ya que pertenecen a linajes que se separaron con anterioridad en la evolución de los proteobacterios. Es decir, comparten un origen común más antiguo, por lo que su similitud genética es menor.

En conjunto, el análisis del gen gyrB permitió observar la diversidad bacteriana y las relaciones evolutivas existentes entre distintos grupos taxonómicos, evidenciando la utilidad de este marcador molecular para la construcción de árboles filogenéticos precisos, incluso cuando la especie de interés no se encuentra representada directamente en los resultados.

Punto 2: Preguntas

1. **¿Qué diferencia hay entre ortólogos y parálogos?**

Para entender la diferencia entre genes ortólogos y parálogos, primero debemos tener claro qué es cada uno.
Por un lado, los genes ortólogos son aquellos que se originan a partir de un único gen ancestral común y están presentes en dos especies diferentes. En este caso, ocurre un evento de especiación, en el que ambas especies heredan una copia del gen del ancestro común.
Por otro lado, los genes parálogos surgen a partir de un evento de duplicación génica dentro del mismo genoma. Es decir, un gen se duplica y, con el tiempo, cada copia puede acumular mutaciones y desarrollar funciones especializadas. Aunque ambos genes provienen del mismo gen ancestral y eran idénticos al inicio, las diferencias adquiridas en su secuencia o función hacen que se consideren parálogos. Este tipo de relación puede presentarse tanto dentro de una misma especie como entre especies diferentes.

Características	Ortólogos	Parálogos
Evento de origen	Especiación o división del linaje entre 2 especies	Duplicación de un gen
Relación principal	Entre especies	Dentro de una misma especie
Mecanismos evolutivos	Divergencia por aislamiento reproductivo	Divergencia que se da en la copia del gen redundante
Función molecular	- Anotación funcional de genomas - Reconstrucción de árbol filogenético	- Entender la base génica de las funciones

2. **¿Por qué se debe incluir un outgroup en el árbol? ¿En qué casos se debe hacer?**
Un outgroup o un grupo externo son especies o grupos de especies que no pertenecen al grupo que se está siendo analizado (Wikipedia, 2022). El grupo externo se utiliza para ayudar a enraizar el árbol filogenético e inferir la dirección del cambio evolutivo. Sin un grupo externo, un árbol filogenético carece de raíz y no puede mostrar la dirección de la evolución. El grupo externo permite a los investigadores distinguir entre rasgos ancestrales y derivados e interpretar correctamente las relaciones evolutivas.

3. **¿Cuál es la diferencia entre el enfoque frecuentista y bayesiano para la reconstrucción de árboles filogenéticos?**

Para este caso debemos entender que es cada una, por un lado tenemos el enfoque frecuentista que lo que busca es encontrar un árbol filogénico que hace la posibilidad de observar los datos del alineamiento de secuencias se la más alta posible, mientras que el enfoque bayesiano hace uso del teorema de bayes para calcular la probabilidad siguiente en un árbol filogénico, dados los datos observados y un conocimiento previo explicito, por lo tanto sabiendo esto podemos identificar las siguientes diferencias entre ellos:

Características	Ortólogos	Parálogos
Pregunta	¿Qué árbol hace que estos datos sean más probables?	¿Cuál es la probabilidad de que este árbol sea correcto?
salida	Un único árbol, el que mejor explica los datos	Distribución de probabilidad entre los arboles
Medida de soporte	Remuestreo, repetir el experimento muchas veces con datos similares	Probabilidades posteriores, con destrucción y datos la probabilidad de que una clado sea cierto
¿Usa conocimiento previo?	No usa	Si usa

Referencias

De la Fuente, M., Dauros, P., Bello, H., Domínguez , M., Sepúlveda, M., Zemelman, R., & Gonzáles, G. (2007). *Mutaciones en genes gyrA y gyrB en cepas de bacilos Gram negativos aisladas en hospitales chilenos y su relación con la resistencia a fluoroquinolonas*.

Koonin, E. V. (2005). *Orthologs, paralogs, and evolutionary genomics*.

Peeters, K., & Willems, A. (2011). *The gyrB gene is a useful phylogenetic marker for exploring the diversity of Flavobacterium strains isolated from terrestrial and aquatic habitats in Antarctica*.

Wikipedia. (3 de Julio de 2022). *Wikipedia*. Obtenido de Grupo externo (cladística): https://es.wikipedia.org/wiki/Grupo_externo_(clad%C3%ADstica)

Wikipedia. (30 de Julio de 2025). *Wikipedia*. Obtenido de Serratia: https://es.wikipedia.org/wiki/Serratia

https://www.bioinformatics.org/cd-hit/