# 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**

## Punto 1: Descripción del flujo de trabajo

## Ortólogos

El gen elegido fue el de la girasa Beta que es un gen específico que codifica para la subunidad B de la ADN girasa, topoisomerasa tipo II exclusiva de procariotas, crucial en procesos como replicación y transcripción del ADN mediante la introducción de superenrollamientos negativos dependientes de ATP (De la Fuente et al., 2007). Su función garantiza su presencia variedad de bacterias y una evolución conservada, lo que lo convierte en un marcador ideal para estudios de ortología. Los ortólogos de gyrB, generados por especiación, acumulan sustituciones nucleotídicas de manera lenta y predecible debido a la presión selectiva que conserva sus dominios funcionales, actuando como un reloj molecular confiable (Peeters & Willems, 2011). Este se encuentra en una zona llamada QRDR (región determinante de resistencia a quinolonas). (De la Fuente, y otros, 2007). 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. Además, se incorporó la secuencia del gen gryB de las *Pseudomonas aeruginosa PAO1* también de NCBI como outgroup tomando los CDS fasta del gen en OrthoDB, 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*, lo que permite comparar secuencias homólogas.


**Imagen1**. Página OrthoDB de donde se sacó los ortólogos de *Pseudomonas aeruginosa*

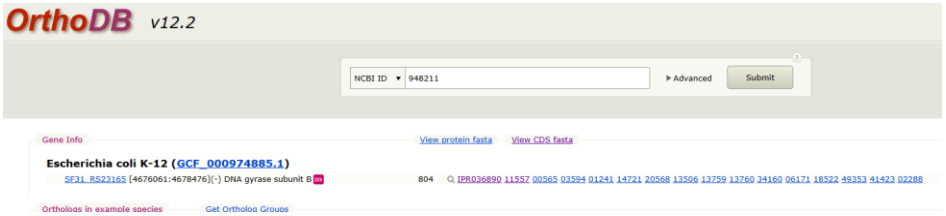


**Imagen1**. Página OrthoDB de donde se sacó los ortólogos de *Escherichia coli*

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

- Primero, se eliminaron las secuencias duplicadas, guardándolas en el archivo orto_nodups.fasta y se aplicó un filtro de longitud mínima para facilitar el procesamiento computacional, dando como resultado orto_minlen.fasta.

- 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 (Li, 2009).
- Finalmente, quedamos con un conjunto representativo de 100 secuencias, almacenado en orto_final.fasta. 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.

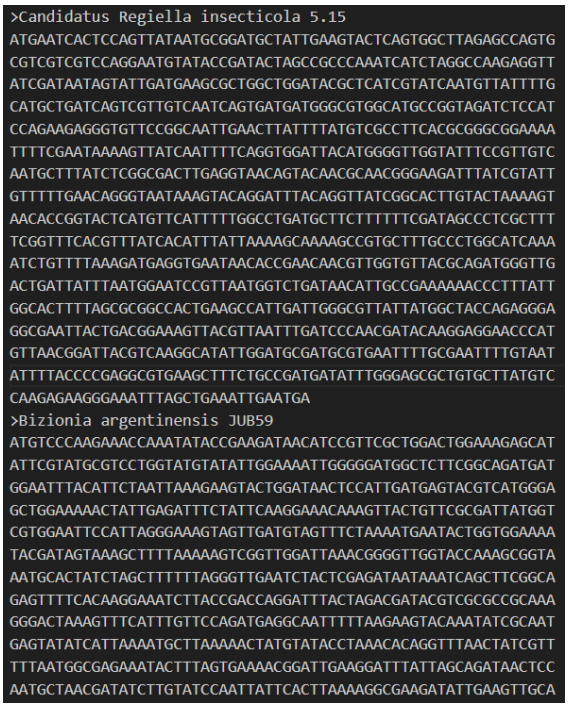

**Imagen3**. Muestra de orto_final_names.fasta.

## Alineamiento múltiple

Después se hizo un alineamiento múltiple con la herramienta MAFFT para los archivos del gen gryB de *Pseudomonas aeruginosa y Escherichia coli*, nombrado como orto_alineados_tag1.fasta. Este alineamiento es la base para todos los análisis evolutivos siguientes, ya que establece la homología de posición nucleotídica.

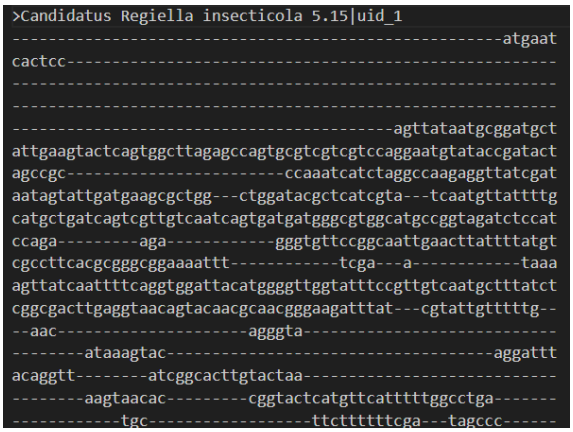


**Imagen4**. Muestra del alineamiento resultado de MAFFT

**<u>Filogenia: Tener en cuenta esta pregunta ¿Cuál es el mejor modelo de sustitución que pudieron encontrar para sus datos? Explicarlo brevemente.</u>**

La selección del modelo de sustitución nucleotídica se realizó mediante el software IQTREE que está diseñado para analizar grandes cantidades de datos evolutivos. Este programa acelera los cálculos usando varios procesadores a la vez y trabajando en paralelo (Wong et al., 2025). Este incorpora ModelFinder, un sistema de selección de modelos que evalúa varios modelos de sustitución (Wong et al., 2025).

El objetivo fue identificar el modelo que mejor describa los patrones de cambio evolutivo en los datos ya alineados. La selección se basó en criterios de información estadística, como el criterio de Información Bayesiano (BIC) (Jani, s. f.).

$$\mathrm{BIC} = k\ln(n) - 2\ln(\widehat{L})$$

**Imagen4**. Fórmula de criterio de información Bayesiano. Donde, k representa el número de parámetros del modelo (complejidad), n el tamaño de la muestra, y L la verosimilitud máxima del modelo (bondad de ajuste a los datos). El término k·ln(n) penaliza la complejidad, favoreciendo modelos más simples cuando el tamaño de muestra aumenta, mientras que -2·ln(L) premia un buen ajuste (Jani, s. f.).

El comando de ModelFinder calculó la verosimilitud logarítmica de un árbol inicial para múltiples modelos y luego calculó sus puntuaciones de BIC (Nguyen et al., n.d.). El modelo con el valor de BIC más bajo se selecciona como el mejor, pues minimizar el BIC ayuda a seleccionar el modelo de más precisión, mejor ajuste y mayor sencillez. Con el comando

*cat ortologos_alineados_tag1.fasta.iqtree | grep "Best-fit model"*
Se obtuvo que el modelo de mejor ajuste según BIC fue GTR+F+I+G4.

El componente GTR (General Time-Reversible) es el modelo reversible en el tiempo más general, que incorpora tasas de sustitución desiguales entre los seis tipos de transiciones y transversiones (Banos et al., 2025). El sufijo +F indica el uso de frecuencias de bases empíricas calculadas directamente del conjunto de datos, en lugar de frecuencias asumidas como iguales. Los modificadores +I y +G4 tienen en cuenta la heterogeneidad de las tasas de sustitución entre los distintos sitios del alineamiento; +I modela la proporción de sitios

invariables (completamente conservados), mientras que +G4 utiliza una distribución gamma discreta con 4 categorías para modelar la variación de las tasas entre los sitios variables (Banos et al., 2025). El informe completo de este análisis se guardó en el archivo orto_alineados_tag1.fasta.iqtree.

Finalmente, se reconstruyó el árbol filogenético mediante el método de Máxima Verosimilitud (ML) en IQTREE bajo el modelo de sustitución GTR+F+I+G4. El análisis arrojó una verosimilitud logarítmica óptima de -295068.028 (Best score found: -295068.028). Este valor representa el máximo ajuste alcanzado entre el modelo y los datos del alineamiento, indicando una alta congruencia del modelo con los patrones de variación observados de las secuencias.

## Árbol filogenético

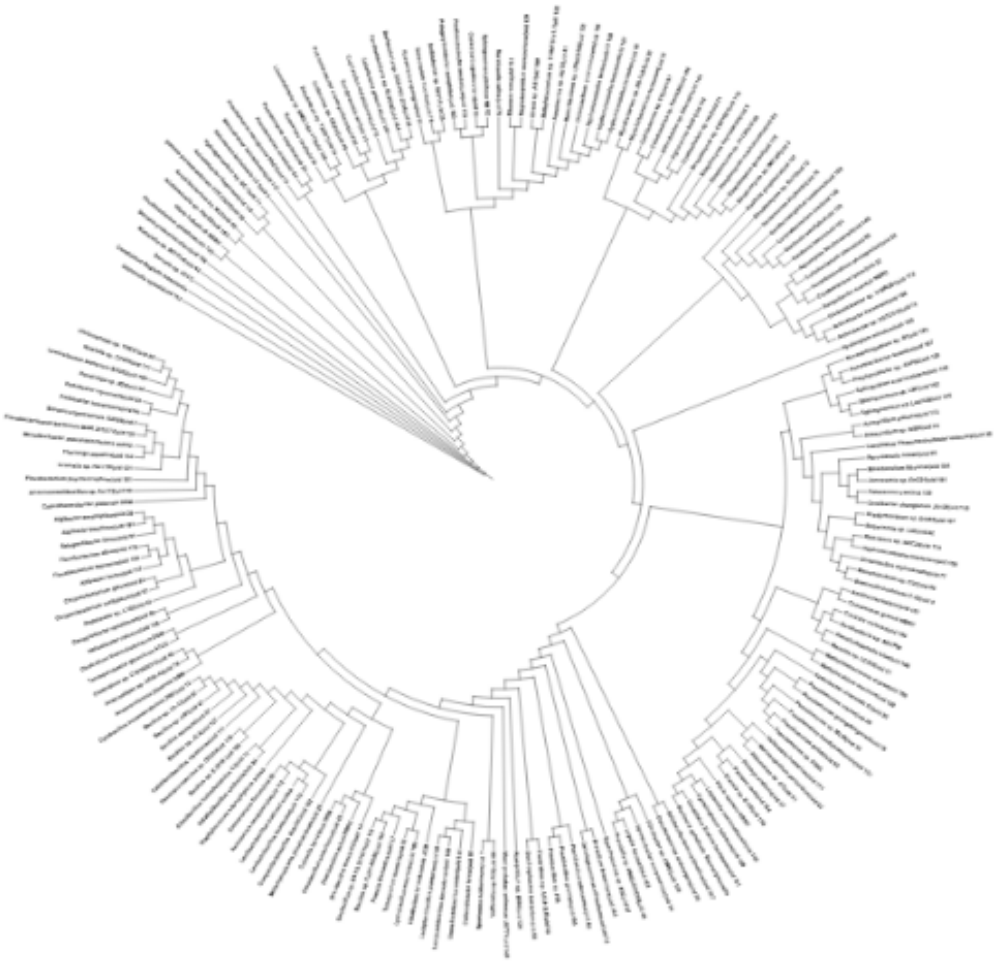

**Figura1**. Árbol filogenético del gen gyrB en *Escherichia coli* y *Pseudomonas aeruginosa* con visualización en IQTreeS, 2007)

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, según sus relaciones evolutivas y según su grado de similitud o divergencia.

Para el outgrop 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*.

Para el objetivo de examinar la filogenia de *Escherichia coli* a partir del gen gyrB, con el árbol 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.

## 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<br>- Reconstrucción de árbol filogenético | - Entender la base génica de las funciones |

**Tabla1**. Características de ortólogos vs patólogos

## 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 |

**Tabla2**. Características de ortólogos vs patólogo

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

Wikipedia. (1 de Octubre de 2025). *Chitinophaga* . Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Chitinophaga_agri?utm_source=chatgpt.com

Wikipedia. (1 de Octubre de 2025). *Niastella*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Niastella

Wikipedia. (28 de julio de 2025). *Regiella insecticola*. Obtenido de Wikipedia: https://en.wikipedia.org/wiki/Regiella_insecticola?utm_source=chatgpt.com

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

Enterobacteriaceae. (n.d.). *NCBI Taxonomy Browser*. Recuperado de https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=543

Vibrionaceae. (n.d.). *NCBI Taxonomy Browser*. Recuperado de https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=641

Pseudomonadaceae. (n.d.). *NCBI Taxonomy Browser*. Recuperado de https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=135621

Burkholderiaceae. (n.d.). *NCBI Taxonomy Browser*. Recuperado de https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=119060

Li, W. (2009). *CD-HIT: Cluster Database at High Identity with Tolerance*. Recuperado de https://www.bioinformatics.org/cd-hit/

Jani. (s. f.). *Choosing the best model: A friendly guide to AIC and BIC*. Medium. Recuperado de https://medium.com/%40jshaik2452/choosing-the-best-model-a-friendly-guide-to-aic-and-bic-af220b33255f

Wong, T. K. F., Ly-Trong, N., Ren, H., Baños, H., Roger, A. J., Susko, E., Bielow, C., De Maio, N., Goldman, N., Hahn, M. W., Huttley, G., Lanfear, R., & Minh, B. Q. (2025). *IQ-TREE 3: Phylogenomic inference software using complex evolutionary models* [Software]. GitHub. Recuperado de https://github.com/iqtree/iqtree3

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (n.d.). *Advanced tutorial* [Software documentation]. IQ-TREE. Recuperado de https://iqtree.github.io/doc/Advanced-Tutorial iqtree.github.io

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., & Minh, B. Q. (n.d.). *Tutorial* [Software documentation]. IQ-TREE. Recuperado de https://iqtree.github.io/doc/Tutorial iqtree.github.io

Last update: Jun 5, 2025, Contributors: Hector Banos, Diep Thi Hoang, Dominik Schrempf, Heiko Schmidt, Jana Trifinopoulos, Minh Bui, Thomas Wong, Nhan Ly-Trong, Hiroaki Sato. Recuperado de https://iqtree.github.io/doc/Substitution-Models