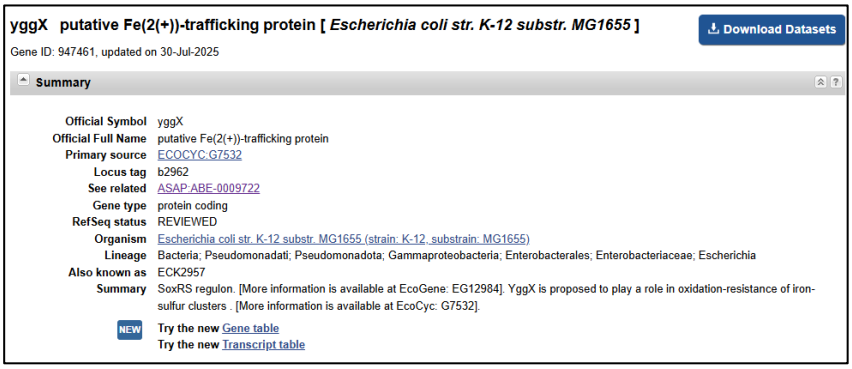
Julieta López - Sofía Upegui – Sebastián Reina – Mateo Pulgarín

Entrega 3 bioinformática

El gen yggX codifica una proteína de aproximadamente 91 aminoácidos que desempeña un papel crucial en la respuesta bacteriana al estrés oxidativo y en el mantenimiento de la homeostasis del hierro. Estudios en *Escherichia coli* han demostrado que *yggX* forma parte del regulón SoxRS, un sistema regulador que responde al estrés oxidativo generado por especies reactivas de oxígeno (ROS) (Pomposiello y Demple, 2001; Skovran et al., 2004). La deleción de *yggX* en *E. coli* reduce drásticamente la capacidad de la célula para resistir agentes generadores de superóxido como el paraquat (Duval & Lister, 2013).

Desde un punto de vista evolutivo, *yggX* es un gen altamente conservado entre las eubacterias, lo que sugiere que su función es esencial para la supervivencia bacteriana bajo condiciones ambientales adversas. Su papel en el control del metabolismo del hierro y la mitigación del daño oxidativo lo convierte en un marcador molecular de interés para estudios filogenéticos, especialmente en el contexto de la evolución de los mecanismos antioxidantes en bacterias. (Osborne et al., 2005)

Este gen fue seleccionado para el análisis filogenético tras verificar su presencia en los archivos del organismo ancestral mediante la terminal. En esta etapa se confirmó que el gen (*yggX*) estaba presente en la muestra secuenciada. Posteriormente, se realizó una búsqueda de este en la base de datos de NCBI, desde donde se descargó su secuencia genética. Es importante resaltar que este paso fue crucial para definir qué gen utilizar en el análisis filogenético, ya que otros genes inicialmente considerados generaban archivos demasiados pesados para las corridas computacionales, debido no solo a la longitud de sus secuencias, sino también a las anotaciones asociadas a ellas.



yggX putative Fe(2(+))-trafficking protein [*Escherichia coli str. K-12 substr. MG1655*]

Download Datasets

Gene ID: 947461, updated on 30-Jul-2025

Summary

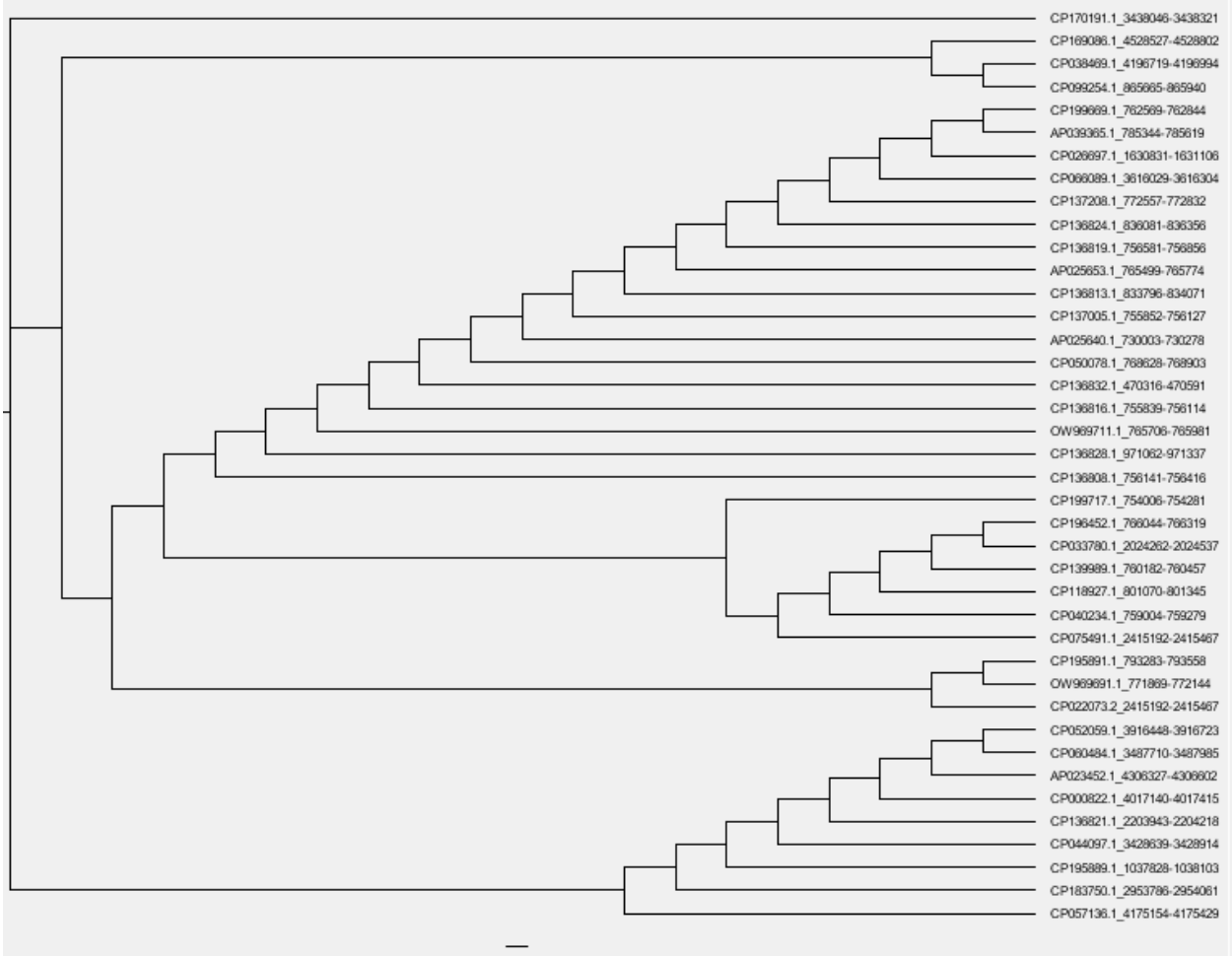
Official Symbol	yggX
Official Full Name	putative Fe(2(+))-trafficking protein
Primary source	ECOCYC:G7532
Locus tag	b2962
See related	ASAP:ABE-0009722
Gene type	protein coding
RefSeq status	REVIEWED
Organism	Escherichia coli str. K-12 substr. MG1655 (strain: K-12, substrain: MG1655)
Lineage	Bacteria; Pseudomonadati; Pseudomonadota; Gammaproteobacteria; Enterobacterales; Enterobacteriaceae; Escherichia
Also known as	ECK2957
Summary	SoxRS regulon. [More information is available at EcoGene: EG12984]. YggX is proposed to play a role in oxidation-resistance of iron-sulfur clusters . [More information is available at EcoCyc: G7532].

NEW Try the new Gene table

Try the new Transcript table

Se continúo el análisis usando la herramienta BLAST, la cual compara la secuencia del gen con otras presentes en bases de datos, con el fin de identificar secuencias que presentan similitudes con el gen de interés. A partir de los resultados obtenidos, se seleccionaron las mejores coincidencias considerando criterios como el *E-value* (valor esperado), que representa la probabilidad de que una alineación específica ocurra por azar y no debido a una verdadera relación biológica entre las secuencias (sequenceserver, 2023). Este análisis permitió identificar genes ortólogos, es decir, aquellos que presentan similitud entre sí debido a una ascendencia común. Dichos genes conservan la misma función, pero se encuentran en diferentes organismos que evolucionaron a partir de un ancestro común. Es importante diferenciar los genes ortólogos de los parálogos, los cuales se originan por duplicación y posterior divergencia, adquiriendo funciones distintas o cierto grado de especialización (*Computational Gene Identification*, n.d.)

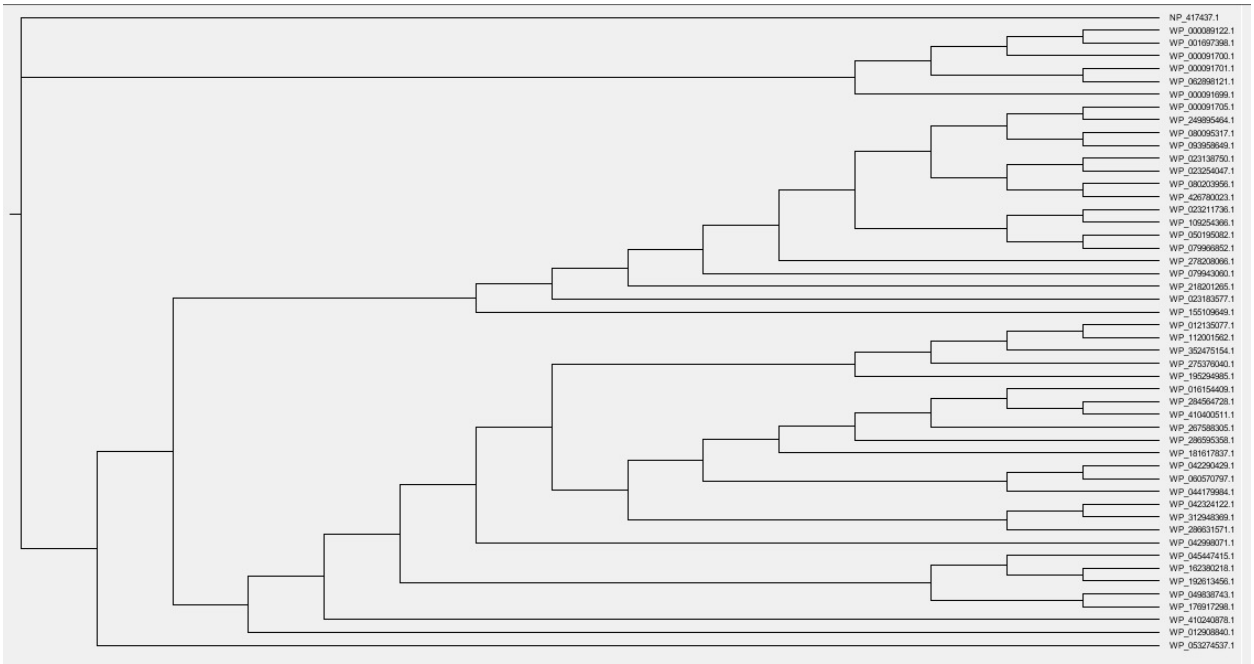
Una vez identificados los genes ortólogos, se alinearon en MAFFT. Luego, se eliminaron una las regiones mal alineadas para construir el árbol filogenético (Figura 2). Este mostró una gran cantidad cepas del organismo *Escherichia Coli,* indicando que las secuencias eran muy parecidas entre sí. Para encontrar la información diferente a cepas de esta bacteria, se aplicó un filtro que elimina los genes relacionados con *E.coli,* lo que entregó un árbol filogenético conteniendo bacterias y hongos diferentes conteniendo el gen de interés. Adicionalmente, se optó por emplear otro enfoque que permitiera no solo conservar la similitud, sino también capturar una mayor diversidad de secuencias.



Como solución a este inconveniente se realizó un BLASTp, que compara secuencias de aminoácidos contra una base de datos de proteínas, a diferencia de un BLASTn que compara secuencias de nucleótidos contra una base de datos de nucleótidos. Una de las ventajas de utilizar BLASTp se basa en que las proteínas en bacterias mutan más lento que muchos otros genes, lo que permite un análisis filogenético más amplio y específico. De este proceso salió entonces información de organismos con esta misma proteína, y se procedió hacer el árbol filogenético con el software IQtree para ambos blasts, el de genes y el de proteínas.

Desde una perspectiva molecular, el uso de BLASTp en la búsqueda de genes ortólogos se justifica porque compara directamente las secuencias de aminoácidos, que reflejan de mejor los cambios evolutivos que influyen en la estructura y función de las proteínas. Según Moreno-Hagelsieb y

Latimer (2008), las secuencias proteicas tienden a conservar mejor su función incluso cuando las secuencias de ADN discrepan por mutaciones sinónimas, permitiendo así reconocer homologías evolutivas reales y reducir los llamados “falsos positivos” que pueden presentarse al trabajar con nucleótidos. Por eso en este análisis filogenético se decidido emplear también BLASTp, aprovechando esa capacidad de detectar similitudes funcionales. Además del uso de esas herramientas de sustitución o de filtrados para esas regiones de poca información aumenta la sensibilidad del método, haciendo que aplicar el BLASTp sea realmente una herramienta más precisa para el estudio evolutivo en proteínas.



La reconstrucción filogenética puede abordarse desde dos principales enfoques estadísticos: el frecuentista y el bayesiano. El primero se basa en el principio de máxima verosimilitud, y busca identificar el árbol que presenta la mayor probabilidad de haber generado las secuencias observadas bajo un modelo evolutivo determinado. El segundo estima una distribución de probabilidad sobre el conjunto de árboles posibles, incorporando información previa y utilizando métodos de muestreo, como *Markov Chain Monte Carlo* (MCMC), para inferir las relaciones evolutivas (Schrempf et al., 2025).

En el script para generar el árbol filogenético se utilizó el flag -m MFP (ModelFinder Plus), el cual le indica al software IQ-TREE que evalúe automáticamente diferentes modelos de sustitución molecular y seleccione el que mejor se ajuste a los datos del alineamiento.

- Modelo para el BLASTn: K2P+G4
  El modelo K2P (Kimura 2-Parameter) (Kimura, 1980) asume que las tasas de sustitución entre bases no son iguales, diferenciando entre transiciones (cambios entre purinas o entre pirimidinas) y transversiones (cambios entre purinas y pirimidinas). Este modelo considera que las transiciones ocurren con mayor frecuencia, pero mantiene la suposición de que todas las bases tienen la misma frecuencia en el genoma. El sufijo +G4 no forma parte del modelo de sustitución en sí, sino que representa una corrección mediante una distribución

gamma con cuatro categorías discretas, la cual permite modelar la variabilidad en las tasas evolutivas entre los distintos sitios del gen. Esto significa que no todos los nucleótidos cambian a la misma velocidad a lo largo del tiempo: algunos son más propensos a mutar que otros.

- Modelo para el BLASTp: JTT+G4
  El modelo JTT (Jones–Taylor–Thornton) (David et al., 1992) se basa en una matriz empírica de sustitución entre aminoácidos, construida a partir del análisis de un amplio conjunto de proteínas. Esta matriz describe la probabilidad de que un aminoácido se sustituya por otro durante la evolución, permitiendo estimar relaciones evolutivas más precisas a nivel proteico. Al igual que en el caso anterior, el sufijo +G4 corresponde a una corrección mediante una distribución gamma con cuatro categorías, que modela la heterogeneidad de las tasas de sustitución entre los sitios de la proteína (Schrempf et al., 2025) (Arenas, 2015).

En el árbol generado a partir de BLASTn, se encontraron menos ramas y bifurcaciones en comparación con el árbol obtenido mediante BLASTp. El árbol de BLASTn presentó una rama principal que se divide en muchas especies muy cercanas las unas a las otras y con *bootsraps* altas, y por tanto, confiables. Esto está dado por la sensibilidad de BLASTn ante mutaciones y pequeños cambios en el código genético, los cuales pueden impedir que dos secuencias sean consideradas idénticas, mostrando únicamente variantes muy cercanas con pequeños cambios en el código.

Por otro lado, el BLASTp aprovecha la degeneración del código genético y se enfrenta a las mutaciones de manera más amplia, pues un cambio en un nucleótido no siempre significa un cambio del aminoácido codificado, lo que significa que la misma proteína continua en producción y a través del tiempo, las generaciones y la diferenciación de microorganismos hay menos variaciones. Esto explica la gran cantidad de ramas en el árbol filogenético del blastp que muestran *bootstraps* más bajas que en el BLASTn, por la degeneración del código genético, pero una cantidad de ortólogos importante para mostrar cómo se ha pasado el gen desde el mismo inicio.

Adentrándonos en el árbol filogenético del BLASTn, podemos encontrar una variedad de hongos y bacterias que comparten el gen analizado. Destacan especialmente diversas especies del género *Citrobacter*, un grupo de enterobacterias gram negativas conocidas por su capacidad infecciosa en estados de desequilibrio fisiológico. Entre estas, por ejemplo, se encuentra *citrobacter freundii,* una bacteria anaerobia oportunista con gran capacidad de infección de las vías urinarias humanas y multirresistente a diferentes tipos de antibiótic*os* (Wanger et al., 2017). La presencia predominante de estas especies resulta coherente, ya que, al igual que *E, coli,* suelen habitar en el sistema digestivo de humanos y animales, un entorno caracterizado por altos niveles de estrés oxidativo y competencia por el hierro. Adicionalmente, estas toman la mayor parte del árbol teniendo en cuenta que el ADN muta muy rápido, por lo que solo los genes muy similares se tomarán para un BLASTn y se mantendrá conservada la proteína para este género.

Por otro lado el BLASTp mostró mayor flexibilidad en los organismos encontrados, pues aunque a pesar de que la presencia de *citrobacter sp*. todavía es dominante, se encontraron códigos para proteínas protectoras ante el estrés oxidativo similares en microorganismos como *Salmonella entérica* y *Superficieibacter*, las cuales son más lejanas a *E. coli* pero muestran mayor

conservación de la proteína a través de más tipos de microorganismos y no solo *citrobacter sp,* lo cual muestra la importancia de esta proteína a través de diferentes ambientes, no solo en el intestino.

¿Porque se debe incluir un outgroup?

Un outgroup es una especie o secuencia que está relacionada evolutivamente con el grupo de interés (ingroup), pero que divergió antes que los demás miembros del grupo. Su función principal es raizar el árbol filogenético, es decir, determinar la dirección de la evolución y cuál es el antepasado común más reciente. Sin un outgroup, el árbol es no enraizado, y aunque muestra relaciones de similitud, no indica qué linajes son más antiguos o derivados (Berkeley Evolution, s. f.).

Cuando se aplica:

- Permite orientar el árbol (saber cuál es el nodo basal y cuál es más reciente).
- Ayuda a distinguir homologías verdaderas de similitudes por convergencia.
- Facilita la interpretación evolutiva de los cambios en las secuencias o rasgos.

Referencias

How BLAST E-values are calculated and what they mean. (2023, abril 7). *Sequenceserver.com*. https://sequenceserver.com/blog/blast-e-value-meaning/

Berkeley Evolution. (s. f.). *Tree-Building Basics: Outgroup*. The Tree Room. University of California, Berkeley. Recuperado de https://evolution.berkeley.edu/the-tree-room/how-to-build-a-tree/tree-building-basics/

Computational Gene Identification. (s/f). Uam.es. Recuperado el 2 de noviembre de 2025, de http://www.pdg.cnb.uam.es/cursos/Leon_2003/pages/Genomas_Anal_Anot/2_2_Anal.html

Wanger, A., Chavez, V., Huang, R. S. P., Wahed, A., Actor, J. K., & Dasgupta, A. (2017). Overview of Bacteria. *Microbiology and Molecular Diagnosis in Pathology*, 75–117. https://doi.org/10.1016/B978-0-12-805351-5.00006-5

Schrempf, D., Trifinopoulos, J., Meaningseeking, Bui, M., & Thomaskf. (2025, abril 19). *Compilation guide*. Iqtree.org. http://www.iqtree.org/doc/iqtree-doc.pdf

Kimura, M. (1980). A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. *Journal of molecular evolution*, *16*, 111–120. https://doi.org/10.1007/BF01731581

David T. Jones, William R. Taylor, Janet M. Thornton. (Junio 1992). The rapid generation of mutation data matrices from protein sequences, *Bioinformatics*, 8, 3, 275 - 282, https://doi.org/10.1093/bioinformatics/8.3.275

Arenas, M. (2015). Trends in substitution models of molecular evolution. *Frontiers in Genetics*, *6*, 319. https://doi.org/10.3389/fgene.2015.00319

Moreno-Hagelsieb, G., & Latimer, K. (2008). *Choosing BLAST options for better detection of orthologs as reciprocal best hits*. *Bioinformatics, 24*(3), 319–324. https://doi.org/10.1093/bioinformatics/btm585