Entrega 3: Ortología, Filogenia y Análisis Matemático del gen *lacZ* (β-galactosidasa, GH2)

Proyecto de Bioinformática
Tomás Gonzalez
Sofía Cabezas
Santiago Berty
3 de noviembre de 2025

Resumen

Se analizaron ortólogos de *lacZ* (β-galactosidasa; familia GH2) partiendo de la secuencia de *Escherichia coli*. El conjunto de secuencias se filtró por longitud (600 aa a 1300 aa) y calidad ($\leq$20 % residuos ambiguos), reteniéndose 113 secuencias. Se construyó un alineamiento múltiple con MAFFT y se recortó automáticamente con trimAl para eliminar regiones poco informativas. La filogenia se infirió con IQ-TREE≈2, seleccionando el mejor modelo de sustitución (MFP), con soportes SH-aLRT e *ultrafast bootstrap* de 1000 réplicas. Se utilizó como outgroup una β-galactosidasa GH2 de Bacillales para enraizar el árbol. Los clados principales recuperados son coherentes con la taxonomía bacteriana, destacando la conservación funcional de *lacZ* en bacterias heterótrofas.

1. Introducción

El gen *lacZ* codifica una β-galactosidasa (GH2) ampliamente distribuida en bacterias, responsable de la hidrólisis de β-D-galactósidos. Estructuralmente, las GH2 suelen presentar un dominio central tipo barril TIM $(\beta/\alpha)_8$ y motivos catalíticos conservados. En este trabajo se evalúa la ortología y la historia evolutiva de *lacZ* mediante la construcción de un alineamiento múltiple de ortólogos y la inferencia filogenética por máxima verosimilitud (ML) (Emms y Kelly, 2019; Page y Holmes, 1998).

Alineamiento y recorte. Se generó un alineamiento con MAFFT (opción `-auto`) y se aplicó trimAl para remover columnas con señal filogenética pobre.

Filogenia. El árbol ML se reconstruyó con IQ-TREE 2, activando *ModelFinder* (MFP) para seleccionar el mejor modelo según BIC, y calculando soportes SH-aLRT = 1000 y UFBoot = 1000. El outgroup se fijó con `-o Bacillales_betagalc_outgroup`. El modelo óptimo reportado fue **BLOSUM62+F+R4**. Los archivos de salida principales incluyen `lacZ_tree.plusOG.treefile` (Newick) y el PDF con el árbol (Nguyen, Schmidt, von Haeseler, y Minh, 2015; Kalyaanamoorthy, Minh, Wong, von Haeseler, y Jermiin, 2017).

2. Análisis matemático y modelo estadístico (IQ-TREE)

IQ-TREE aplica el principio de **máxima verosimilitud (ML)** para estimar la topología T y los parámetros θ que maximizan la probabilidad de observar el alineamiento D bajo un modelo evolutivo M:

$$\ln L(T, M, \theta|D) = \sum_{i=1}^{n} \ln P(D_i|T, M, \theta)$$

donde D_i representa la columna i del alineamiento. El algoritmo de Felsenstein calcula $P(D_i|T, M, \theta)$ integrando sobre los posibles estados internos de los nodos del árbol (Felsenstein, 1981).

Selección del modelo. IQ-TREE prueba decenas de modelos de sustitución y selecciona el que optimiza la verosimilitud penalizada según el **criterio de información bayesiano (BIC)**:

$$BIC = -2 \ln L + k \ln n$$

donde k es el número de parámetros y n el número de sitios. El modelo elegido (**BLOSUM62+F+R4**) incluye:

- una matriz empírica BLOSUM62 que representa las tasas relativas de sustitución entre aminoácidos;
- corrección por frecuencias observadas (+F);
- heterogeneidad de tasas entre sitios con distribución gamma discreta de cuatro categorías (+R4).

Esta selección automática de modelo se realiza mediante **ModelFinder** (Kalyaanamoorthy y cols., 2017).

Soportes estadísticos. IQ-TREE calcula:

- **UFBoot2 (Ultrafast Bootstrap):** re-muestrea el alineamiento 1000 veces y calcula la frecuencia con que cada clado aparece en las réplicas, estimando su confiabilidad.
- **SH-aLRT (Approximate Likelihood Ratio Test):** compara la verosimilitud del nodo frente a topologías alternativas, proporcionando un soporte complementario al bootstrap.

El uso conjunto de SH-aLRT y UFBoot2 proporciona un soporte robusto y eficiente, interpretado como significativo cuando ambos valores superan 80–90 % (Nguyen y cols., 2015).

3. Resultados e Interpretación

La topología ML de *lacZ* muestra:

- Un clado principal que agrupa *Enterobacterales* (*E. coli*, *Salmonella*, *Klebsiella*), con altos valores de soporte, coherente con las relaciones taxonómicas conocidas.
- Subclados diferenciados para *Gammaproteobacteria* y *Betaproteobacteria*, lo que sugiere divergencia evolutiva acorde con la filogenia de especies.
- El outgroup de *Bacillales* se posiciona externamente, enraizando el árbol y confirmando su utilidad para orientar la polaridad de los caracteres.

En conjunto, el árbol respalda que *lacZ* es un gen ubicuo y conservado en bacterias heterótrofas, con divergencia estructural moderada dentro de la familia GH2.

4. Preguntas Teóricas

(i) Diferencia entre ortólogos y parálogos. **Ortólogos** son genes homólogos en distintas especies que se originan por *especiación* y tienden a conservar función. **Parálogos**

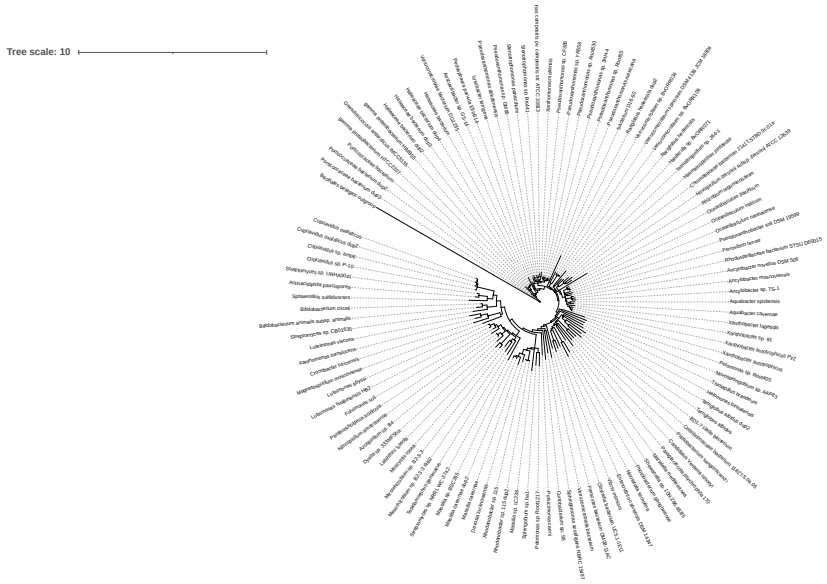

Figura 1: **Árbol ML de ortólogos de *lacZ*** (β-galactosidasa, GH2). Modelo: BLOSUM62+F+R4; soportes: SH-aLRT=1000 y UFBoot=1000. Outgroup: `Bacillales_betagalc_outgroup`.

son homólogos dentro de un mismo genoma que surgen por *duplicación*, pudiendo divergir funcionalmente (Page y Holmes, 1998; Emms y Kelly, 2019).

(ii) ¿Por qué incluir un outgroup? ¿Cuándo se debe usar? El **outgroup** permite *enraizar* el árbol, definiendo la dirección evolutiva. Se usa cuando se dispone de un linaje externo cercano pero no perteneciente al ingroup, lo que permite distinguir estados ancestrales de derivados.

(iii) Enfoque frecuentista vs bayesiano en filogenia. En el enfoque **frecuentista** (ML), se estiman los parámetros y topología que maximizan la verosimilitud, y la confianza se evalúa mediante técnicas de remuestreo (*bootstrap*). El enfoque **bayesiano** infiere una *distribución posterior* sobre árboles y parámetros combinando la verosimilitud con distribuciones previas (*priors*). Los soportes se interpretan como probabilidades a posteriori. Ambos enfoques pueden coincidir, pero el bayesiano tiende a integrar mejor la incertidumbre y la heterogeneidad de modelos.

5. Conclusiones

El análisis de ortólogos de *lacZ* (β-galactosidasa, GH2) evidencia coherencia filogenética con los principales linajes bacterianos. El modelo BLOSUM62+F+R4 explica adecuadamente la heterogeneidad de sustituciones y refuerza la conservación funcional del gen. El outgroup de *Bacillales* permite enraizar correctamente el árbol, confirmando que la evolución de *lacZ* sigue la divergencia esperada entre bacterias Gram negativas y positivas.

Referencias

Emms, D. M., y Kelly, S. (2019). Orthofinder: Phylogenetic orthology inference for comparative genomics. *Genome Biology*, *20*(1), 238. doi: 10.1186/s13059-019-1832-y

Felsenstein, J. (1981). Evolutionary trees from dna sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, *17*(6), 368–376. doi: 10.1007/BF01734359

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., y Jermiin, L. S. (2017). Modelfinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. doi: 10.1038/nmeth.4285

Nguyen, L.-T., Schmidt, H. A., von Haeseler, A., y Minh, B. Q. (2015). Iq-tree: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Molecular Biology and Evolution*, *32*(1), 268–274. doi: 10.1093/molbev/msu300

Page, R. D. M., y Holmes, E. C. (1998). *Molecular evolution: A phylogenetic approach*. Oxford, UK: Blackwell Science.