Universidad Nacional Autónoma de
México y Universidad Michoacana
de San Nicolás de Hidalgo

Instituto de Física y Matemáticas

Posgrado Conjunto en Ciencias Matemáticas
UMSNH-UNAM

Análisis espacial de genes de expansiones de familias génicas en Cianobacteria

T E S I S

que para obtener el grado de

Maestra en Ciencias Matemáticas

presenta

Johanna Atenea Carreón Baltazar

Asesora :
Dra. Adriana Haydeé Contreras Peruyero

Morelia, Michoacán, México

noviembre de 2025

Índice general

Agradecimientos

Agradezco a todos los que supieron estar en esta etapa de mi vida. A los que me sostuvieron cuando dudé y celebraron cuando avancé. Su presencia, apoyo y silencio también fueron parte del camino...

Investigación realizada gracias al Programa de Apoyo a Proyectos de Investigación e Innovación Tecnológica (PAPIIT) de la UNAM **IN114323**. Agradezco a la DGAPA-UNAM la beca recibida.

Resumen

La expansión génica, entendida como la duplicación y conservación de genes dentro de un mismo genoma, es un proceso evolutivo clave en la diversificación funcional y la adaptación de las cianobacterias. Sin embargo, se desconoce si la distribución espacial de los genes duplicados ocurre al azar o responde a principios estructurales o evolutivos. En este trabajo se analizaron 29 genomas completos de cianobacterias para evaluar la organización espacial de genes pertenecientes a familias génicas expandidas. Se identificaron genes duplicados, se calcularon las distancias genómicas entre ellos y se compararon con modelos nulos basados en simulaciones que redistribuyen aleatoriamente los duplicados a lo largo del genoma. Los datos observados y simulados se ajustaron a diferentes distribuciones teóricas, y las diferencias se evaluaron mediante pruebas de bondad de ajuste como Kolmogorov-Smirnov. Además, se examinó el clúster biosintético de scytonemin, un pigmento protector frente a radiación UV, como caso particular de posible organización no aleatoria. Los resultados indican que la distribución de genes duplicados no es completamente aleatoria, lo que sugiere la existencia de restricciones biológicas o evolutivas en la arquitectura genómica. Este enfoque metodológico proporciona una base cuantitativa para explorar la organización funcional y la evolución del contenido génico en bacterias.

Palabras clave: expansión génica, cianobacterias, duplicación de genes, organización genómica, escitonemina.

Abstract

Gene expansion, understood as the duplication and retention of genes within a genome, is a key evolutionary process driving functional diversification and environmental adaptation in cyanobacteria. However, it remains unclear whether the spatial distribution of duplicated genes occurs randomly or follows structural or evolutionary principles. This study analyzed 29 complete cyanobacterial genomes to evaluate the spatial organization of genes belonging to expanded gene families. Duplicated genes were identified, genomic distances between them were calculated, and results were compared with null models based on simulations that randomly redistributed duplicates across the genome. Observed and simulated data were fitted to theoretical probability distributions, and differences were assessed using goodness-of-fit tests such as Kolmogorov–Smirnov. Additionally, the biosynthetic cluster of scytonemin—a UV-protective pigment—was examined as a specific case of potentially non-random organization. The results indicate that the distribution of duplicated genes is not entirely random, suggesting the presence of biological or evolutionary constraints shaping genome architecture. The methodological framework developed here provides a quantitative basis for exploring the functional organization and evolutionary dynamics of gene content in bacteria.

Introducción

La expansión génica —la duplicación y conservación de genes dentro de un mismo genoma— es un proceso evolutivo clave que permite la innovación funcional, la adaptación a diversos entornos y la aparición de nuevas rutas metabólicas [1]. En procariotas como las cianobacterias, los genes de familias génicas expandidas se han asociado con la producción de metabolitos secundarios, resistencia a antibióticos y otras funciones ecológicas relevantes [2].

A pesar de su importancia, aún se desconoce si la organización espacial de estos genes dentro del genoma ocurre de forma aleatoria o si responde a principios estructurales, funcionales o evolutivos. Estudios en procariotas —como Acaryochloris marina o Staphylococcus aureus— han reportado que las copias duplicadas tienden a agruparse o estar sobrerrepresentadas en regiones específicas del genoma, lo cual indica que su distribución puede no ser aleatoria [3, 4]. Abordar esta pregunta implica no solo identificar correctamente los genes asociados a familias de expansión génica, sino también caracterizar su distribución física y contrastarla con modelos aleatorios simulados.

Este trabajo analiza 29 genomas completos de cianobacterias para evaluar la distribución espacial de genes duplicados —a los que nos referimos como genes secundarios— en contraste con las copias primarias o ancestrales dentro de las mismas familias. Se diseñaron simulaciones de genomas artificiales con múltiples familias génicas; en cada caso, algunas familias se seleccionaron aleatoriamente y sus duplicados fueron insertados uniformemente a lo largo del genoma. También se generó una simulación alternativa, en la que ciertos genes se seleccionaron aleatoriamente y se marcaron como duplicaciones, sin inserciones, para representar una hipótesis nula sin estructura espacial.

Tras identificar los genes secundarios, se calcularon las distancias genómicas entre ellos y se evaluó su ajuste a modelos estadísticos teóricos, tanto originales como transformados logarítmicamente. Finalmente, se examinó el clúster biosintético de escitonemina —un pigmento protector frente a la radiación UV— para evaluar si también presenta organización no aleatoria. Un clúster biosintético es un conjunto de genes contiguos que codifican colectivamente un metabolito.

Los resultados sugieren que la distribución de genes de familias génicas expandidas en cianobacterias no es completamente aleatoria, lo cual apunta a posibles restricciones biológicas o evolutivas. El enfoque metodológico desarrollado ofrece una base para futuros estudios sobre la arquitectura y evolución del contenido génico en bacterias.

Capítulo 1

Marco Teórico

El objetivo de este capítulo es mostrar las bases biológicas y matemáticas necesarias para comprender este estudio de la distribución espacial de copias secundarias.

1.1. Cianobacterias: características generales

Las cianobacterias son un grupo diverso de microorganismos procariotas capaces de realizar fotosíntesis oxigénica, similar a la de las plantas, lo que las convierte en algunos de los organismos más antiguos y ecológicamente más importantes de la Tierra [5]. Presentan una amplia variedad morfológica, que va desde células unicelulares hasta formas multicelulares filamentosas con diferenciación celular, lo que les confiere una notable adaptabilidad a distintos ambientes [6]. Además, poseen la capacidad de fijar nitrógeno atmosférico en condiciones de bajo contenido de nitrógeno disponible, función que las hace esenciales en los ciclos biogeoquímicos [5]. Esta combinación de características les permite colonizar una gran diversidad de hábitats y desempeñar un papel crucial en los ecosistemas donde habitan.

Algunas especies de cianobacterias producen metabolitos secundarios, como la *escitonemina*, un pigmento protector frente a la radiación ultravioleta que contribuye a su adaptación a ambientes extremos.

Por su diversidad funcional, su relevancia evolutiva y su presencia en una amplia gama de ambientes, las cianobacterias constituyen un componente esencial de los ecosistemas tanto acuáticos como terrestres.

1.2. Genómica de cianobacterias

1.2.1. Estructura y organización del genoma de las cianobacterias

El genoma de las cianobacterias es típicamente circular, aunque también existen casos lineales [7]. Su tamaño varía desde 1.4 Mb hasta más de 9 Mb (en algunos casos se han reportado hasta 12 Mb)[8, 9]. Estructuralmente, estos genomas incluyen cromosomas y plásmidos con genes adaptativos, como los relacionados con la fijación de nitrógeno o la producción de metabolitos secundarios, compuestos no esenciales para el crecimiento pero importantes en funciones ecológicas como defensa o comunicación.

Algunos genes se organizan en operones —bloques de genes que se transcriben juntos— o en regiones funcionalmente conservadas llamadas clústeres biosintéticos (BGCs), que codifican rutas para la síntesis de estos compuestos especializados [10, 11].

1.2.2. Técnicas para el estudio genómico

El análisis genómico de cianobacterias permite comprender su diversidad genética, mecanismos de adaptación, evolución y funciones biológicas. Este tipo de estudio emplea diversas herramientas bioinformáticas para procesar y comparar grandes volúmenes de datos. Una etapa inicial consiste en la anotación génica, es decir, la identificación de genes y la predicción de sus posibles funciones. Posteriormente, se puede llevar a cabo un análisis de pangenoma, el cual considera el conjunto completo de genes presentes en un grupo de organismos relacionados. Esta estrategia permite clasificar los genes en conservados (presentes en todos los genomas) y accesorios (presentes solo en algunos), lo que ofrece información clave sobre la especialización funcional y ecológica de las cepas.

Además, se pueden aplicar técnicas complementarias que enriquecen el análisis. Por ejemplo, la búsqueda de BGCs permite identificar genes asociados a la producción de metabolitos secundarios, compuestos clave para la defensa, la competencia o la comunicación entre microorganismos. Otro enfoque relevante es el análisis de duplicaciones génicas, que revela eventos evolutivos como expansiones de familias de genes. A esto se suman los análisis filogenéticos, que permiten estudiar las relaciones evolutivas entre organismos, y las simulaciones computacionales, que facilitan modelar la dinámica de expansión o pérdida génica. La tabla 1.1 resume los métodos y destaca en negritas el software utilizado en este trabajo.

Cuadro 1.1: Tipos de análisis genómicos, objetivos y herramientas bioinformáticas

Tipo de análisis	**Objetivo**	**Herramientas bioinformáticas**
Anotación génica	Identificar genes y predecir funciones a partir de genomas	**RAST** [12], Prokka [13]
Pangenoma	Analizar variabilidad funcional entre cepas mediante genes conservados y accesorios	**GET-HOMOLOGUES** [14], **PPanGGOLiN** [15]
Duplicaciones génicas	Detectar familias génicas duplicadas como posible señal de adaptación	OrthoFinder [16]
Análisis filogenético	Inferir relaciones evolutivas a partir de genes conservados	FastTree [17]
Análisis de BGCs	Detectar genes de metabolitos secundarios y eventos de reclutamiento	antiSMASH [18], BiG-SCAPE [19], EvoMining [20]

1.2.3. Mecanismos de expansión génica

La expansión génica se refiere al proceso mediante el cual múltiples genes homólogos —es decir, que comparten un ancestro común— aparecen en un mismo genoma. Estas expansiones pueden originarse por distintos mecanismos evolutivos, como duplicaciones génicas (por ejemplo mediante errores durante la replicación del ADN), o mediante transferencia horizontal de genes [21].

En cianobacterias, aunque estos mecanismos no han sido tan ampliamente estudiados como en otros grupos de organismos, se han identificado casos en los que ciertas familias génicas están representadas por múltiples copias dentro de un mismo genoma. Estas expansiones pueden involucrar desde un solo gen hasta segmentos más grandes que incluyen varios genes [3]. La presencia de múltiples copias puede ofrecer ventajas evolutivas, ya que se ha observado que algunas mantienen su función original, mientras que otras pueden adquirir nuevas funciones o adaptarse a condiciones ambientales cambiantes [1].

Este tipo de procesos contribuye a la variabilidad genética dentro del grupo de las cianobacterias, lo cual puede influir en su capacidad para adaptarse a distintos hábitats o para desarrollar funciones especializadas [3].

1.3. Distribución espacial de genes en el genoma

En bacterias, los genes suelen organizarse en patrones influenciados por presiones evolutivas. Uno de los más comunes es la formación de operones, donde genes contiguos se transcriben juntos, lo que facilita su regulación coordinada [22].

En el caso de familias génicas expandidas, sus copias pueden aparecer agrupadas o dispersas en el genoma según el mecanismo evolutivo involucrado (como duplicación local, reordenamientos genómicos o transferencia horizontal), lo que puede influir en su función o regulación [23].

Aunque aún son limitados los estudios sobre la distribución espacial específica de familias de genes expandidas en cianobacterias, investigaciones en otras bacterias han sugerido que las expansiones recientes suelen permanecer agrupadas, mientras que aquellas más antiguas tienden a dispersarse a lo largo del genoma [24].

1.4. Fundamentos estadísticos

En esta sección se introducirán las herramientas estadísticas necesarias para modelar la distribución de las familias de genes expandidas dentro de las cianobacterias estudiadas.

1.4.1. Funciones de distribución

En la siguiente tabla se muestran algunas de las funciones de densidad de probabilidad (PDF) más utilizadas en este trabajo. Una función de densidad es una función matemática que describe la distribución de probabilidad de una variable aleatoria continua. Estas funciones se utilizan para modelar la probabilidad relativa de que la variable tome valores específicos y sirven para evaluar la adecuación de los modelos estadísticos a la distribución observada de las copias génicas pertenecientes a familias expandidas.

Cuadro 1.2: Funciones de densidad de probabilidad usadas en este trabajo [25]

Distribución	**PDF**	**Parámetros**
Uniforme	$f(x;a,b) = \frac{1}{b-a},\ a \leq x \leq b$	$a < b$ (intervalo)
Exponencial	$f(x;\lambda) = \lambda e^{-\lambda x},\ x \geq 0$	$\lambda > 0$
Gamma	$f(x;k,\theta) = \frac{1}{\Gamma(k)\theta^k} x^{k-1} e^{-x/\theta},\ x \geq 0$	$k > 0$ (forma), $\theta > 0$ (escala)
Weibull	$f(x;k,\lambda) = \frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1} e^{-(x/\lambda)^k},\ x \geq 0$	$k > 0$ (forma), $\lambda > 0$ (escala)
Log-Gamma	$f(x;\mu,\sigma,\gamma) = \frac{1}{\sigma\Gamma(\gamma)}\left(\frac{\log x-\mu}{\sigma}\right)^{\gamma-1} e^{-\left(\frac{\log x-\mu}{\sigma}\right)},\ x > 0$	μ (ubicación), $\sigma > 0$ (escala), $\gamma > 0$ (forma)
Johnson SU	$f(x) = \frac{\delta}{\lambda\sqrt{2\pi}\sqrt{1+\left(\frac{z^2}{\delta^2}\right)}} e^{-\frac{z^2}{2}},$ $z = \gamma + \delta \sinh^{-1}\left(\frac{x-\xi}{\lambda}\right)$	γ (forma 1), $\delta > 0$ (forma 2), ξ (ubicación), $\lambda > 0$ (escala)

1.4.2. Ajuste de distribuciones

El ajuste de una distribución consiste en encontrar un modelo probabilístico que describa adecuadamente el comportamiento de un conjunto de datos observados [26].

El proceso de ajuste de una distribución puede abordarse en cuatro etapas:

1. **Visualización exploratoria**: se examinan los datos mediante histogramas, curvas de densidad o gráficos Q-Q.
2. **Selección de distribuciones**: se eligen distribuciones teóricas que podrían ajustarse bien a los datos.
3. **Estimación de parámetros**: se utilizan métodos como el método de máxima verosimilitud (MLE), que busca los valores de los parámetros θ que maximizan la función de verosimilitud $L(\theta \mid x_1, \ldots, x_n) = \prod_{i=1}^{n} f(x_i \mid \theta)$.
4. **Evaluación del ajuste**: se aplican pruebas de bondad y ajuste como Kolmogorov-Smirnov y sus p valores para evaluar la adecuación del modelo.

Transformación de variables aleatorias

En algunos casos, transformar una variable aleatoria puede facilitar su modelado, estabilizar la varianza o aproximarla a una distribución teórica conocida. Para ello, se emplea la teoría de transformación de variables aleatorias. El procedimiento general es el siguiente.

Sea X una variable aleatoria de distribución desconocida y $Y = g(X)$ una transformación monótona y diferenciable. Si se obtiene la densidad de probabilidad de Y denotada como $f_Y(y)$, entonces la densidad de X puede recuperarse mediante:

$$f_X(x) = f_Y(g(x)) \cdot \left| \frac{d}{dx} g(x) \right|. \tag{1.1}$$

Además, si g es estrictamente creciente, las funciones de distribución acumulada se relacionan por:

$$F_X(x) = F_Y(g(x)), \tag{1.2}$$

y si g es estrictamente decreciente:

$$F_X(x) = 1 - F_Y(g(x)). \tag{1.3}$$

Pruebas de bondad de ajuste

Una vez propuesta una distribución teórica, es necesario evaluar su ajuste a los datos observados. Para ello se aplican pruebas de bondad de ajuste, que cuantifican la discrepancia entre la distribución empírica y la teórica. Estas pruebas contrastan la hipótesis nula H_0: *"Los datos provienen de la distribución teórica especificada"*, frente a la alternativa H_1: *"Los datos no siguen dicha distribución"*. Si se rechaza H_0, concluimos que no deberíamos usar el modelo, pero no rechazar H_0 no garantiza que el modelo sea correcto, ya que puede deberse a un bajo poder estadístico [26].

Prueba de Kolmogorov-Smirnov (K-S)

Este trabajo utiliza la prueba de Kolmogorov-Smirnov para evaluar el ajuste entre una distribución empírica y una teórica. La prueba se basa en la mayor diferencia absoluta entre la función de distribución empírica $F_n(x)$ y la función teórica $F(x)$:

$$D_n = \sup_x |F_n(x) - F(x)|$$

El valor D_n cuantifica la discrepancia entre ambas distribuciones, y su significancia se evalúa a través de un *p-valor*. Si dicho valor es pequeño (por ejemplo, menor o igual a 0.05), se considera que hay suficiente evidencia para rechazar la hipótesis nula [27].

1.4.3. Comparación de modelos

En este trabajo, el criterio principal para aceptar o rechazar un modelo se basa principalmente en el *p-valor* de la prueba K-S, y complementariamente en el valor del estadístico D.

1.5. Planteamiento del problema e hipótesis

1.5.1. Contextualización del problema

Una pregunta clave en la organización genómica es si las copias génicas de familias expandidas se distribuyen aleatoriamente o siguen un patrón. Este trabajo aborda la cuestión en cianobacterias mediante un **modelo nulo** de distribución aleatoria comparado con datos reales.

Se identificaron copias duplicadas en cada genoma y se calcularon sus distancias, ajustando distribuciones de probabilidad. También se simularon genomas con inserciones aleatorias de genes duplicados, manteniendo número total de copias y tamaño genómico.

Se estudió además el clúster biosintético de *escitonemina* para evaluar patrones distintos al azar. Todas las distribuciones, observadas y simuladas, fueron comparadas mediante pruebas de bondad de ajuste, como Kolmogorov-Smirnov, para determinar si las diferencias reflejan organización no aleatoria.

1.5.2. Justificación de la investigación

Determinar si las copias génicas expandidas muestran organización no aleatoria puede evidenciar clústeres funcionales, como los BGCs, asociados a metabolitos secundarios. Como muchos BGCs contienen genes duplicados, este análisis apoya estrategias computacionales para su identificación.

El trabajo también aborda el reto de distinguir patrones aleatorios y estructurados cuando los datos observados y simulados comparten la misma familia de distribuciones pero con distintos parámetros, lo que plantea interrogantes sobre su interpretación biológica.

Además, el enfoque de análisis circular, donde distancias entre puntos uniformemente distribuidos se aproximan mediante una exponencial [28], respalda conceptualmente el uso de distribuciones exponenciales en los modelos nulos.

1.5.3. Formulación de la hipótesis

Para evaluar la organización espacial de las familias génicas expandidas, se comparan los datos observados con un **modelo nulo de aleatoriedad**, generado mediante simulaciones que redistribuyen genes duplicados aleatoriamente. Este modelo sirve como referencia para definir las distribuciones teóricas a ajustar.

Las hipótesis estadísticas se plantean como:

H_0 : Las distancias observadas entre genes duplicados pueden modelarse mediante una distribución teórica específica (p. ej. gamma, exponencial, Weibull).

H_1 : Las distancias observadas no siguen la distribución teórica especificada.

De esta manera, si el *p-valor* es menor a $\alpha = 0{,}05$, se rechaza H_0 y se concluye que los datos no provienen de dicha distribución. Si el *p-valor* es mayor, no se rechaza H_0, lo que implica que los datos son compatibles con esa distribución.

El contraste con el modelo nulo de aleatoriedad se logra comparando los ajustes de las distribuciones teóricas en datos reales y simulados. Una discrepancia sistemática en los ajustes sugiere que la organización observada no puede explicarse solamente por azar.

Capítulo 2

Metodología

El código fuente empleado en este capítulo está disponible en el repositorio de GitHub: github.com/Johannaatenea/analisis-familias-expansion-cianobacterias. Incluye scripts en Bash y Python, cuadernos interactivos, archivos de entrada y resultados intermedios que respaldan cada etapa del flujo de trabajo.

2.1. Selección del conjunto de genomas y preprocesamiento

Este estudio se basa en 29 genomas completos de cianobacterias, seleccionados por su clasificación taxonómica y disponibilidad en NCBI con ensamblado a nivel de cromosoma. Se incluyeron géneros como *Nostoc*, *Nodularia*, *Lyngbya*, *Anabaena* y *Chlorogloeopsis*, priorizando diversidad filogenética y calidad del ensamblado. Genomas incompletos o muy fragmentados fueron descartados para evitar sesgos en el cálculo de distancias. La descarga se realizó mediante `ncbi-genome-download` (v0.3.3), ejecutando el siguiente comando por separado para cada género:

```
ncbi-genome-download --genera "GENERO" -l "complete,chromosome"
--formats fasta -o salida/ bacteria
```

La anotación funcional se realizó con el servidor web de `RAST`, cargando los archivos `.fasta` y obteniendo como salidas archivos `.faa`, `.gbk` y `.txt` con secuencias proteícas y descripciones funcionales. Para identificar familias ortólogas se utilizó `GET_HOMOLOGUES` (v3.5.1) con archivos `.gbk` como entrada en un entorno `conda`. Se define como ortólogos a genes en diferentes genomas que derivan de un ancestro común, y como familia génica al conjunto de ortólogos agrupados por similitud. La detección se realizó con:

```
get_homologues.pl -d data_gbks -M -t 0 -c -n 1
```

donde `-M` usa un método para comparar genes entre genomas, `-t 0` asegura que se usen todos los genomas, `-c` busca solo genes muy parecidos (ortología estricta), y `-n 1` indica que se use un solo núcleo del procesador.

Posteriormente, la matriz de presencia-ausencia se generó con:

```
compare_clusters.pl -o sample_intersection -m -d familias
```

2.2. Procesamiento del grupo de estudio y filtrado

El agrupamiento inicial con *GET_HOMOLOGUES* produjo 51,286 familias génicas. No obstante, debido a que muchas de estas podrían corresponder a regiones poco conservadas o no informativas, se aplicó un filtrado en tres etapas para reducir el conjunto y enfocarse en un número manejable de familias:

1. **Prevalencia genómica:** se conservaron las familias presentes en al menos el 60 % de los 29 genomas. Este paso, basado en la matriz de presencia-ausencia, redujo el conjunto a 3,711 familias.

2. **Expansión génica:** con el fin de enfocarse en familias candidatas a eventos de duplicación, se conservaron solo aquellas que contenían al menos dos genes en al menos uno de los genomas. Para implementarlo se construyó la matriz de abundancia, en la cual cada celda representa el número de copias de una familia específica en un genoma dado. Tras este filtrado, conservamos 678 familias potenciales.

3. **Longitud de alineamiento:** se ejecutó *BLASTp* [29] (v2.12.0) para comparar pares de genes dentro de cada familia, y se conservaron solo aquellas con alineamientos superiores a 500 residuos en todas las comparaciones. Esto eliminó familias con secuencias cortas, propensas a resultados inconsistentes, dejando un conjunto final de 130 familias.

Estas 130 familias constituyen un subconjunto de alto interés para el análisis de familias de expansión génica, distribución espacial y posible expansión funcional (véase la lista completa de familias en la carpeta Resultados/130_Familias_de_interes).

2.3. Clasificación de los genes

Los análisis descritos se encuentran en el cuaderno de Python (v3.10.10) `Genes_primarios_y_secundarios.ipynb` en el cual se usan bibliotecas como `pandas` (2.2.2), `seaborn` (0.13.2), `matplotlib` (3.9.0) y `scikit-learn` (1.6.1).

A partir de las familias filtradas y los resultados de *BLASTp*, se calcularon dos métricas por gen: el promedio de porcentaje de identidad (*pident*, proporción de residuos coincidentes entre secuencias alineadas) y el *e-value* (probabilidad de obtener una coincidencia por azar), considerando todas sus comparaciones dentro de la familia para caracterizar su similitud interna. Se eliminaron casos donde un genoma estuviera representado por un único gen en una familia, conservando solo familias con al menos dos genes del mismo genoma, condición necesaria para evaluar expansión génica.

Se propusieron dos enfoques para clasificar a los genes como **primarios** (originales de la familia) o **secundarios**, entendiendo por estos últimos a los genes adicionales en la misma familia que se originan por duplicación de los primarios o por adquisición mediante transferencia horizontal. Una familia puede contener ambos tipos. En este trabajo se emplearon dos métodos para la clasificación de genes en primarios y secundarios:

- **Umbrales** (ver Anexo A.0.1): un gen fue clasificado como primario si su *e-value* promedio fue igual o menor al mejor valor de la familia (tolerancia de 10^{-1}) o si su *pident* promedio estaba en el percentil 90. Si ningún gen cumplía estos criterios, se eligió como primario aquel con mejor *e-value* y mayor *pident*. El resto que no cumplía lo anterior se clasificó como secundario.

- **K-means ($k = 2$)** (ver Anexo A.0.2): se aplicó sobre los valores estandarizados de *e-value* y *pident*. El clúster que contenía al gen con métricas óptimas se etiquetó como primario; el otro, como secundario.

2.4. Análisis estadístico mediante el calculo de distancias entre genes pertenecientes en familias génicas expandidas

Esta sección fue desarrollada utilizando los cuadernos `Graficos_y_ajustes.ipynb`, `Diferencias_entre_duplicados_en_pb.ipynb` y `Diferencias_entre_duplicados_en_núm_de_genes.ipynb`.

Antes de realizar los análisis estadísticos, se calcularon las distancias entre genes clasificados como *secundarios* con el objetivo de caracterizar su distribución espacial dentro de cada genoma. Este análisis se llevó a cabo empleando dos medidas complementarias: (1) la distancia en número de genes que separan a los genes pertenecientes a dichas familias y (2) la distancia en kilobases (Kb) (donde una kilobase equivale a mil pares de bases de ADN) calculada a partir de sus posiciones genómicas. Para cada genoma, se partió de diccionarios que contenían listas ordenadas de genes secundarios según su posición en el genoma, obtenidas mediante los métodos de clasificación por umbrales y *K-means*.

Sea $\{G_1, G_2, \ldots, G_n\}$ la lista de genes duplicados secundarios ordenada por su aparición en el genoma y sea

$$r_i = \text{pos}(G_i) \in \{1, \ldots, n\}$$

la posición *ordinal* de G_i en la anotación genómica (es decir, el número de orden del gen dentro del genoma). Entonces, la distancia en número de genes entre duplicados consecutivos en dicha lista se define como

$$d_i = r_{i+1} - r_i, \qquad i = 1, \ldots, n-1,$$

y, por la circularidad del genoma, se consideró también la distancia que une el último con el primero:

$$d_n = r_1 + N - r_n,$$

donde N es el número total de genes en el genoma (ver Anexo B.0.1 para un ejemplo).

Para la métrica basada en kilobases, se utilizaron las posiciones de inicio y fin de cada gen, extraídas de los archivos de anotación generados por `RAST` (ver carpeta `Cianobacterias`).

Sea g_i el i-ésimo gen duplicado en el orden en que aparecen a lo largo del genoma, con $p_{g_i,\text{start}}$ y $p_{g_i,\text{stop}}$ sus posiciones inicial y final, respectivamente. La distancia entre duplicados consecutivos se calculó como:

$$d_i = \frac{\text{mín}\left(p_{g_{i+1},\text{start}},\ p_{g_{i+1},\text{stop}}\right) - \text{máx}\left(p_{g_i,\text{start}},\ p_{g_i,\text{stop}}\right)}{1000}, \quad \text{para } i = 1, 2, \ldots, n-1$$

y la distancia circular entre el último y el primero como:

$$d_{\text{circular}} = \frac{\text{mín}\left(p_{g_1,\text{start}},\ p_{g_1,\text{stop}}\right) + L - \text{máx}\left(p_{g_n,\text{start}},\ p_{g_n,\text{stop}}\right)}{1000}$$

donde L es la longitud total del genoma en pares de bases (ver Anexo B.0.2 para un ejemplo).

Nota: aquí el índice i no representa el identificador original del gen, sino su posición ordinal en la lista ordenada de duplicados según la coordenada genómica.

A partir de estas distancias, se construyeron histogramas que representan la frecuencia de las diferentes separaciones entre genes identificados como secundarios, permitiendo visualizar sus patrones de agrupamiento o dispersión a lo largo del genoma. Para complementar el análisis descriptivo, se ajustaron modelos estadísticos teóricos — incluyendo distribuciones exponencial, gamma, log-normal y Weibull — a las distribuciones observadas. La selección del modelo más adecuado se basó en el p-valor para el estadístico de Kolmogorov-Smirnov. Este enfoque metodológico es fundamental para identificar y caracterizar patrones espaciales relevantes en la organización de genes pertenecientes de familias génicas expandidas.

2.5. Simulación

Con el objetivo de contrastar los patrones observados en los datos reales, se implementaron dos simulaciones independientes que generaron genomas artificiales con características similares a los genomas reales del estudio, particularmente en cuanto al tamaño y al número de genes pertenecientes a familias génicas expandidas. Ambas simulaciones tuvieron como propósito evaluar si la distribución observada de estos genes en los genomas reales podría explicarse mediante un modelo de distribución aleatoria.

Se realizaron simulaciones utilizando los cuadernos `Simulación_por_inserción` y `Simulación_por_selección`. En ambas, se generaron 29 genomas artificiales con un tamaño promedio similar al de los genomas reales y una cantidad comparable de genes etiquetados como pertenecientes a familias génicas expandidas. En la primera, estos genes fueron insertados aleatoriamente siguiendo una distribución uniforme; en la segunda, se seleccionó aleatoriamente un subconjunto de genes ya existentes en cada genoma y se les asignó dicha etiqueta.

Para ambas simulaciones, se calcularon las distancias entre genes etiquetados como pertenecientes a familias génicas expandidas, en términos del número de genes que los separan, siguiendo el mismo procedimiento descrito en secciones anteriores. Cabe destacar que, debido a la naturaleza artificial de los genomas simulados, no fue posible estimar posiciones genómicas reales (en kilobases), por lo que las distancias solo

pudieron calcularse en número de genes. Posteriormente, se construyeron histogramas y se realizaron ajustes a distribuciones teóricas para comparar los patrones generados artificialmente con los observados en los genomas reales. Este enfoque permitió obtener una línea base sobre cómo se comportarían los genes de familias expandidas si su distribución fuera efectivamente aleatoria.

2.6. El caso de la Escitonemina

Esta sección se basa en los cuadernos `Diferencias_entre_duplicados_en_num_de_genes_scytonemin.ipynb`, `Graficos_y_ajustes_para_scytonemin.ipynb`, y el script de terminal `detectar_scytonemin.sh`.

Para evaluar el patrón de distribución del clúster de *escitonemina* en los genomas del grupo de estudio, se utilizó como secuencia de consulta un archivo en formato `.faa` que contiene las proteínas que conforman dicho clúster. Esta secuencia fue comparada contra todos los genomas mediante `BLASTp`. Se consideró que un gen pertenecía al clúster si cumplía dos condiciones: un valor de *e-value* menor a 1×10^{-10} y un porcentaje de identidad (*pident*) mayor a 80 %. Un genoma fue clasificado como portador del clúster únicamente si presentaba al menos el 90 % de los genes del clúster original. A partir de estos resultados, se extrajeron las posiciones genómicas de los genes del clúster en cada genoma identificado para calcular las distancias (en número de genes) entre ellos, de forma análoga al análisis realizado para las familias génicas expandidas.

Estas distancias se almacenaron en un diccionario estructurado por genoma y se complementaron con estadísticas descriptivas como la media, mediana, desviación estándar y rango intercuartílico. Posteriormente, se construyeron histogramas para visualizar la distribución observada de las distancias entre genes del clúster. Aunque se intentaron ajustes con distintos modelos estadísticos, no se observaron buenos ajustes en las distribuciones, por lo que no se propuso ningún modelo específico. Finalmente, se realizó una visualización de la ubicación de los genes del clúster a lo largo de cada genoma, lo que permitió explorar gráficamente su posible agrupamiento o dispersión.

2.7. Transformación de los datos

En esta sección se aborda el proceso de transformación logarítmica aplicado a las distancias entre genes etiquetados como pertenecientes a familias génicas expandidas en cada uno de los genomas. Este procedimiento se realizó para los distintos escenarios abordados en este trabajo: distancias en número de genes y en kilobases, tanto para los datos reales (clasificados por *K-means* y umbrales), como para las simulaciones.

Dado que en todos los casos las distribuciones de las distancias presentaban un sesgo positivo, se aplicó una transformación logarítmica con el objetivo de reducir dicho sesgo y aproximar las distribuciones a formas más simétricas. Esta transformación mejora la calidad de los ajustes estadísticos.

El procedimiento consistió en transformar los datos X (las distancias) mediante $Y = \log(X)$, ajustar distribuciones teóricas sobre los datos transformados Y, y posteriormente recuperar la forma de la distribución original X. En particular, si $f_Y(y)$ es la densidad ajustada sobre los datos transformados, la densidad correspondiente a la variable original se puede recuperar mediante:

$$f_X(x) = f_Y(\log(x)) \cdot \frac{1}{x}$$

Este enfoque permite trabajar en el espacio transformado, donde los modelos estadísticos se ajustan mejor, y luego interpretar los resultados en la escala original de las distancias. Para cada conjunto de datos transformados se probaron múltiples modelos de distribución teórica.

Una vez seleccionadas las distribuciones con mejor ajuste sobre los datos transformados, se procedió a recuperar la distribución para los datos originales usando la relación anterior. Esto permitió generar una propuesta de distribución para los datos originales con base en los parámetros obtenidos en el espacio transformado.

Finalmente, se generaron visualizaciones que permitieron comparar la calidad de los ajustes antes y después de la transformación logarítmica. Estas incluyen histogramas, curvas ajustadas, y gráficos QQ-plot. Los análisis descritos en esta sección se llevaron a cabo en los notebooks `Graficos_y_ajustes.ipynb`, `Simulación_por_inserción.ipynb` y `Simulación_por_selección.ipynb`.

Capítulo 3

Resultados

3.1. Clasificación de genes primarios y genes pertenecientes a familias génicas expandidas

En esta sección se presentan los resultados obtenidos al clasificar los genes de los genomas analizados en dos categorías: **primarios**, que corresponden a genes únicos o no duplicados, y **secundarios**, aquellos que forman parte de familias génicas expandidas. Por conveniencia, en lo que sigue nos referiremos a estos últimos simplemente como *genes secundarios*.

La clasificación como planteamos en la sección pasada se hizo basadose en dos métodos: **umbrales** y **agrupamiento con K-means**. A continuación, se describen los resultados generales y las estadísticas por genoma, así como las coincidencias y diferencias entre ambos métodos.

Cuadro 3.1: Resumen comparativo de la clasificación de genes por método. Las filas "Promedio genes por genoma" indican el número promedio de genes primarios o secundarios en un genoma individual, mientras que "% genes del total clasificados" corresponde al porcentaje global de genes primarios o secundarios respecto al total clasificado.

Métrica	**Umbrales**		**K-means**	
	Primarios	**Secundarios**	**Primarios**	**Secundarios**
Total genomas analizados	26	26	26	26
Número total de genes	1295	1232	1686	871
Promedio genes por genoma	49.8 %	47.4 %	64.9 %	33.5 %
% genes del total clasificados	51.3 %	48.8 %	65.9 %	34.1 %
Total genes clasificados	2527		2557	
Total genes en 26 genomas	167,686		167,686	

La coincidencia en la clasificación de genes secundarios entre los métodos de umbrales y *K-means* fue variable entre los genomas analizados. La Figura 3.1 ilustra, mediante un gráfico de barras apiladas, las coincidencias y discrepancias en la clasificación de

genes secundarios para cada genoma.

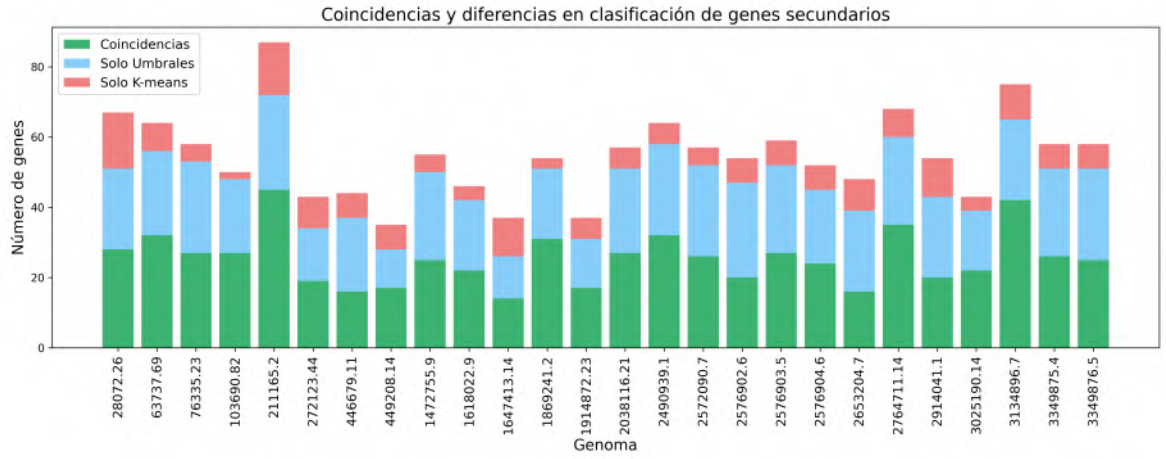

Figura 3.1: Distribución de coincidencias y discrepancias entre ambos métodos de clasificación en genes secundarios

3.2. Distribución de genes duplicados en los genomas reales

3.2.1. Distancias observadas

En esta sección se analizan las distancias entre genes clasificados como secundarios utilizando los métodos de umbrales y *K-means*, tanto en términos del número de genes como de la longitud genómica en kilobases (KB).

Con umbrales, las distancias en número de genes mostraron una media de 145.30, mediana de 93.00 y alta dispersión (desviación estándar DE = 159.60), con un máximo de 1309. En *K-means*, los valores fueron mayores: media de 209.61, mediana de 134.50, DE = 250.29 y máximo de 2117.

En kilobases, los patrones se mantienen: umbrales arrojó una media de 147.42 KB y mediana de 94.65 KB, mientras que *K-means* alcanzó 213.57 KB de media y 141.39 KB de mediana. En ambos casos, las distancias mínimas fueron muy pequeñas (0.02 KB), lo que indica cercanía física entre algunos genes duplicados.

3.2.2. Ajuste de distribuciones teóricas

En esta sección se analizan las diferencias entre genes secundarios utilizando dos métodos de agrupamiento: *K-means* y umbrales. Las diferencias se evaluaron tanto en kilobases como en número de genes, explorando sus distribuciones mediante histogramas y ajustes teóricos.

La Figura 3.2 muestra que todas las distribuciones presentan alta frecuencia de distancias pequeñas y sesgo a la derecha, siendo más marcada en el método por umbrales.

Para caracterizar estas distribuciones se ajustaron modelos teóricos, normalizando previamente las diferencias como densidades de probabilidad para facilitar la comparación visual entre curvas.

La Figura 3.3 muestra sólo los casos con al menos un ajuste aceptable. Para *K-means*, sólo la distribución gamma fue significativa ($p = 0.1025$). En umbrales, tanto gamma ($p = 0.4330$) como Weibull mínima ($p = 0.2220$) ofrecieron buenos ajustes.

En cambio, al aplicar los ajustes sobre diferencias en kilobases, ninguna distribución superó el umbral de significancia estadística.

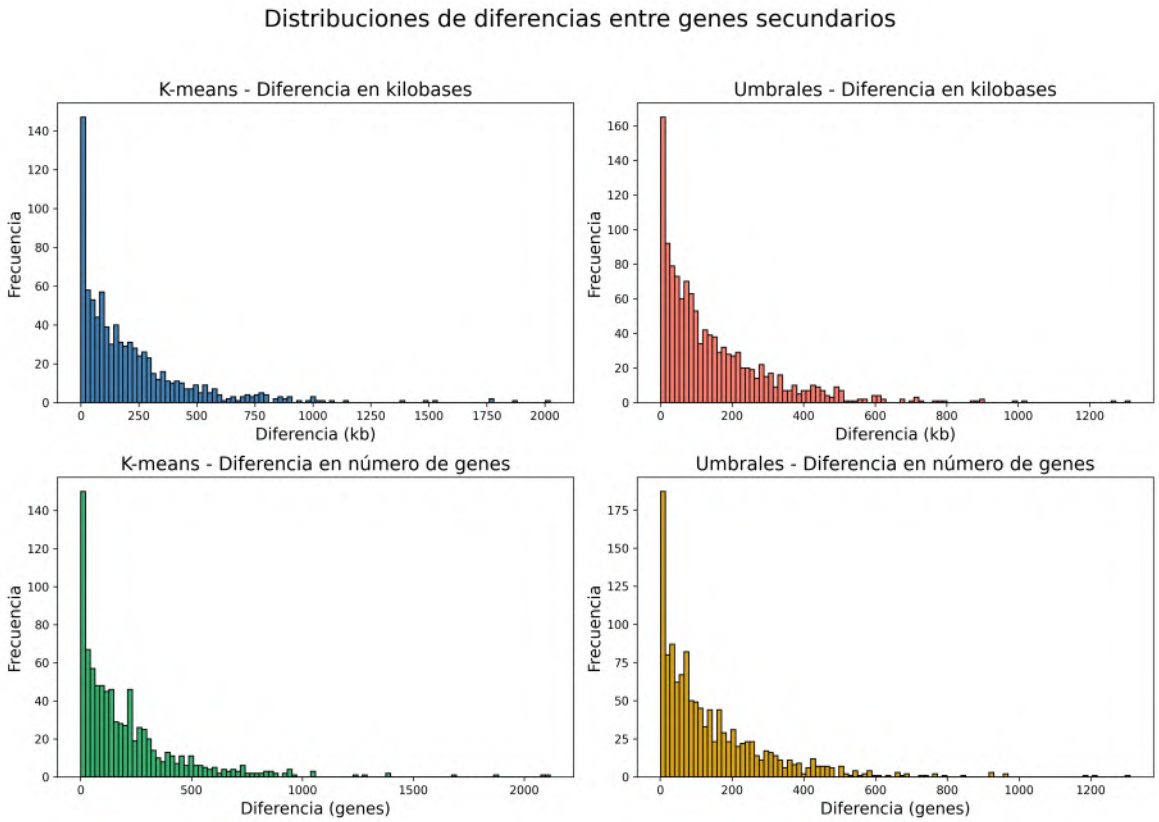

Figura 3.2: Distribuciones de diferencias entre genes secundarios según el método de agrupamiento.

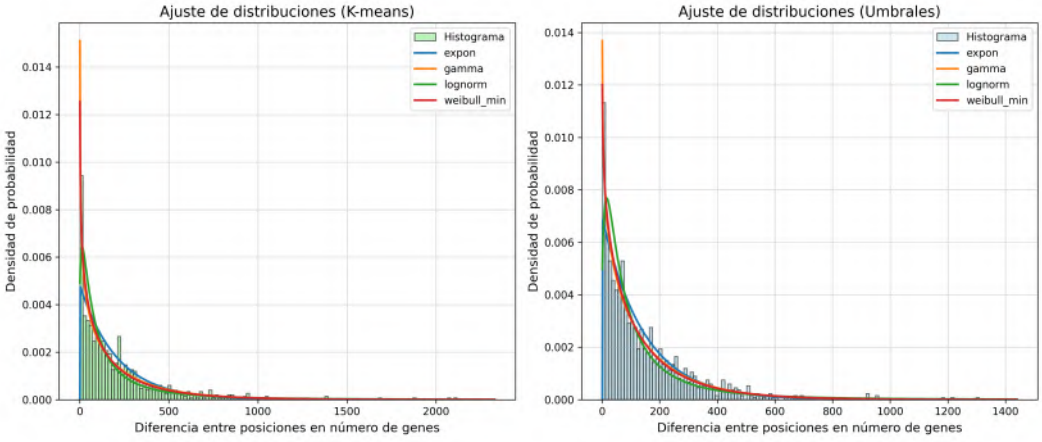

Figura 3.3: Ajuste de distribuciones a diferencias entre genes secundarios según K-means y umbrales (en número de genes).

3.2.3. Transformación logarítmica y nuevos ajustes

Dado el sesgo derecho en las distribuciones originales, se aplicó una transformación logarítmica a las diferencias entre genes secundarios, tanto en kilobases como en número de genes. Esta transformación duplicó las combinaciones evaluadas, al incluir datos originales y transformados para ambos métodos (K-means y umbrales) y ambas métricas (número de genes y kilobases). Para mantener la claridad visual, la Figura 3.4 muestra sólo los ajustes correspondientes al método **umbrales en número de genes**, donde se observó una mejora notable en el ajuste, con varias distribuciones que superaron el umbral de significancia.

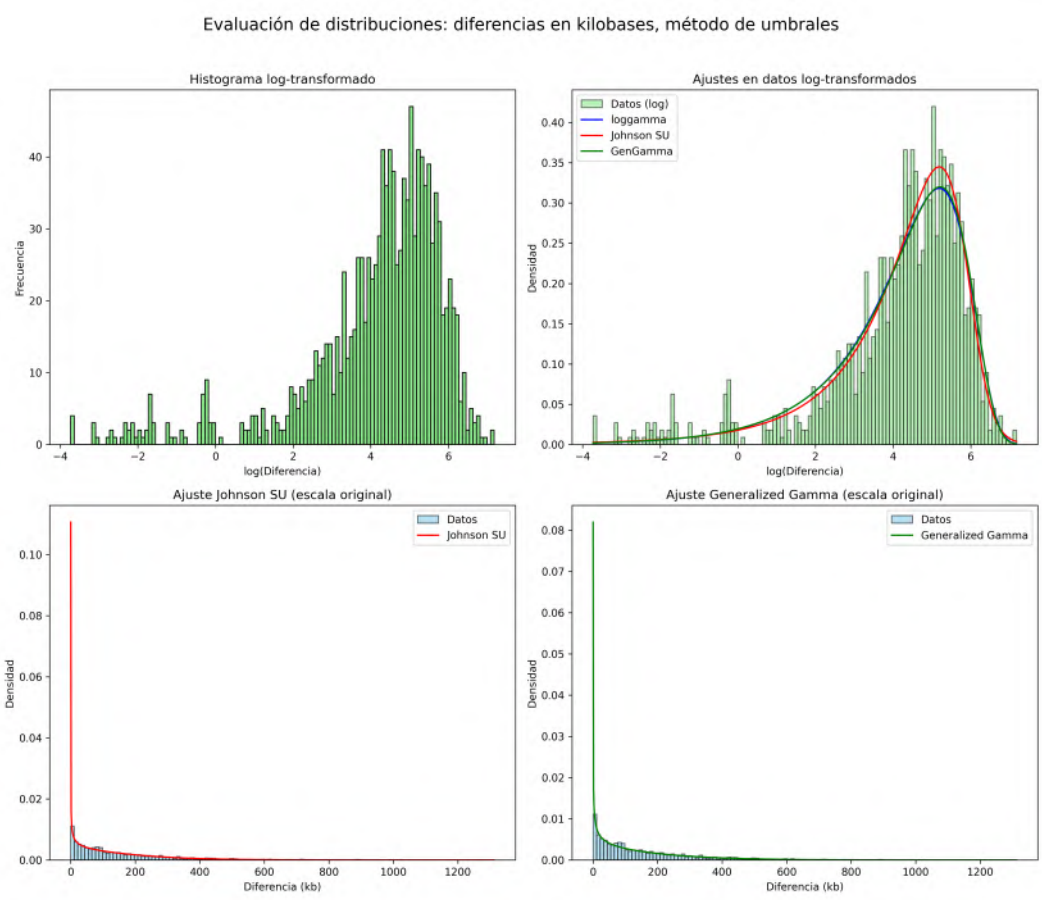

Figura 3.4: Evaluación de distribuciones log-transformadas (KB, umbrales).

Un resumen completo se presenta en los cuadros 3.2 y 3.3, con las mejores distribuciones por caso.

Cuadro 3.2: Mejores ajustes de distribuciones por método y unidad (I).

Método y unidad	**Distribución**	**KS-Statistic**	**KS-pvalue**
K-means (kb)	Johnson SU	0.0340	0.2722
Umbrales (kb)	Johnson SU	0.0219	0.5889
	Generalized Gamma	0.0312	0.1767

Cuadro 3.3: Mejores ajustes de distribuciones por método y unidad (II).

Método y unidad	Distribución	KS-Statistic	KS-pvalue
K-means (número genes)	LogGamma	0.0389	0.1460
	Johnson SU	0.0370	0.1878
Umbrales (número genes)	LogGamma	0.0224	0.5575
	Johnson SU	0.0226	0.5450

3.3. Distribución de genes en los genomas simulados

Para evaluar la naturaleza estadística de las diferencias entre genes duplicados en los escenarios simulados, se ajustaron diferentes distribuciones tanto a los datos en escala original como a los datos transformados mediante la función logarítmica. Solo se presentan aquellos ajustes que superaron la prueba de Kolmogorov-Smirnov.

3.3.1. Simulación por inserción

En la simulación por **inserción**, varias distribuciones como la exponencial, gamma, log-normal y Weibull lograron buenos ajustes en la *escala original*. Tras la transformación logarítmica, los ajustes mejoraron aún más, destacando las distribuciones *GenGamma*, *Johnson SU* y *Weibull*, con p-valores superiores a 0.99 (ver Tabla 3.4).

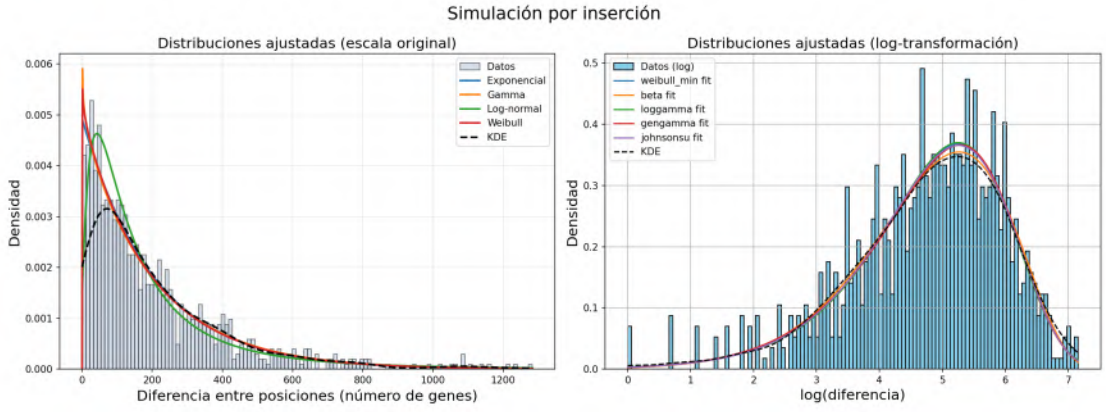

Figura 3.5: Ajustes en datos originales y log-transformados (inserción).

3.3.2. Simulación por selección

En la simulación por **selección**, las distribuciones exponencial, gamma y Weibull lograron buenos ajustes en la escala original. Con la transformación logarítmica, mejoraron significativamente, destacando *GenGamma*, *Johnson SU* y *LogGamma*, con p-valores superiores a 0.88 (ver Tabla 3.4).

Los resultados respaldan la utilidad de la transformación logarítmica y demuestran que es posible obtener buenos ajustes estadísticos incluso con distintos mecanismos simulados.

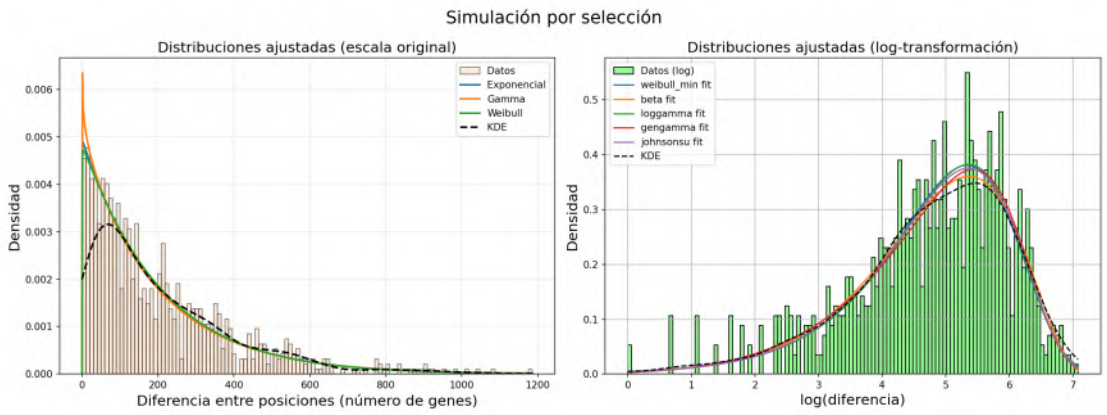

Figura 3.6: Ajustes en datos originales y log-transformados (selección).

Cuadro 3.4: Distribuciones con buen ajuste en ambas simulaciones.

Simulación	**Transformación**	**Distribución**	**KS-Statistic**	**p-valor**
Inserción	Original	Exponencial	0.0153	0.9760
Inserción	Original	Gamma	0.0198	0.8405
Inserción	Original	Weibull	0.0169	0.9422
Inserción	Log-transformada	Johnson SU	0.0135	0.9941
Inserción	Log-transformada	GenGamma	0.0116	0.9994
Inserción	Log-transformada	Weibull	0.0140	0.9906
Selección	Original	Exponencial	0.0269	0.4832
Selección	Original	Gamma	0.0319	0.2799
Selección	Original	Weibull	0.0256	0.5487
Selección	Log-transformada	Johnson SU	0.0185	0.8928
Selección	Log-transformada	GenGamma	0.0153	0.9766
Selección	Log-transformada	LogGamma	0.0188	0.8804

3.4. Distribución del clúster de Escitonemina

Para el análisis de las distancias entre los genes que conforman el clúster *escitonemina*, se evaluaron 18 genomas donde fue detectada su presencia. Las distancias fueron calculadas en kilobases.

A nivel global, la media fue de 360.12 KB y la mediana de apenas 1 KB, lo que revela una fuerte asimetría hacia valores bajos. Esto también se refleja en la desviación estándar (994.47 KB) y en el rango observado (0 a 6136 KB). El rango intercuartílico (IQR) fue de solo 1 KB (Q1 = 1 KB, Q3 = 2 KB), lo que indica una alta concentración de datos en un intervalo muy reducido.

Dado que la mayoría de los valores son muy pequeños y se agrupan en un solo intervalo, no se ajustaron modelos estadísticos a los histogramas. En su lugar, se proporciona una visualización (véase Anexo C) complementaria que muestra la ubicación relativa de los genes del clúster de escitonemina en cada genoma.

Capítulo 4

Discusión de resultados

4.1. Resumen general de hallazgos

En este estudio se analizaron 29 genomas completos de cianobacterias, identificando genes primarios y secundarios (pertenecientes a familias génicas expandidas) mediante dos métodos complementarios: umbrales y K-means. Se detectó que los genes secundarios, pertenecientes a familias génicas expandidas, tienden a agruparse espacialmente dentro de los genomas, mostrando distancias menores a las esperadas bajo un modelo de distribución aleatoria. La aplicación de una transformación logarítmica sobre las distancias mejoró significativamente el ajuste de modelos estadísticos teóricos, permitiendo una caracterización más precisa de los patrones observados. Además, el análisis del clúster *escitonemina* reveló una fuerte conservación espacial, consistente con su probable función biológica conservada.

4.2. Comparación entre métodos de clasificación de genes

Los resultados muestran diferencias sustanciales en la asignación de genes secundarios según el método empleado. Mientras que el enfoque por *umbrales* identificó 1232 genes secundarios, *K-means* clasificó 871, lo que representa una discrepancia considerable. Esta divergencia puede atribuirse a la naturaleza de cada método: los umbrales aplican criterios fijos sobre *e-value* y *pident*, mientras que *K-means* realiza agrupamientos basados en similitud general.

La coincidencia media de entre 25 y 30 genes secundarios por genoma indica que ambos métodos capturan subconjuntos parcialmente solapados. Esta complementariedad sugiere que combinar o comparar ambos enfoques puede ofrecer una visión más completa del fenómeno de expansión génica.

4.3. Distribución espacial de genes duplicados

4.3.1. Patrones observados en genomas reales

En los genomas reales, las distancias entre genes duplicados presentaron alta variabilidad, con valores mínimos cercanos a cero y máximos que superan los 2000 genes

o kilobases. Esto indica la coexistencia de duplicaciones muy cercanas y otras muy distantes. Las distancias fueron sistemáticamente menores cuando se utilizó el método de **umbrales**, reflejando la mayor sensibilidad de este enfoque.

En la escala original, los ajustes a distribuciones teóricas mostraron resultados limitados. Para *K-means* únicamente la distribución *gamma* alcanzó un ajuste aceptable ($p = 0{,}1025$), mientras que para umbrales tanto *gamma* ($p = 0{,}4330$) como *Weibull mínima* ($p = 0{,}2220$) mostraron mejores ajustes, indicando que H_0 (que los datos siguen la distribución) *no se rechaza* solo en estos casos.

Tras aplicar la transformación logarítmica, los ajustes mejoraron notablemente para ambos métodos. Distribuciones como *LogGamma* y *Johnson SU* alcanzaron p-valores altos (ver Tablas 3.2 y 3.3), lo que sugiere que los datos reales son compatibles con estas distribuciones teóricas tras la corrección de sesgo

4.3.2. Ajustes en simulaciones aleatorias

Para las simulaciones (mecanismos de inserción y selección), se ajustaron las mismas distribuciones teóricas tanto en escala original como transformada. En la escala original, varias distribuciones lograron buen ajuste, destacando exponencial, gamma y Weibull (ver Tabla 3.4). Tras la transformación logarítmica, los ajustes fueron aún más significativos, con p-valores superiores a 0.88 en todos los casos, indicando que los datos simulados siguen de manera consistente la distribución esperada según cada modelo.

4.3.3. Comparación de ajustes entre datos reales y simulados

La comparación se puede resumir en la Tabla 4.1, mostrando los ajustes más representativos y si H_0 se rechaza o no (solo en datos reales). Es importante notar que en el caso del *p-valor* de la simulación, solo se considera el mejor valor obtenido, ya sea de la simulación por inserción o por selección. Además, únicamente se comparan los ajustes de distancias en número de genes, ya que los resultados en kilobases no son directamente comparables.

Cuadro 4.1: Comparación de ajustes entre datos reales y simulados.

Método	**Escala**	**Distribución**	**p-valor (simulación)**	**p-valor (real)**	**H_0**
K-means	Genes (original)	Gamma	0.8405	0.1025	No se rechaza
Umbrales	Genes (original)	Gamma	0.8405	0.4330	No se rechaza
Umbrales	Genes (original)	Weibull min	0.9422	0.2220	No se rechaza
K-means	Genes (log)	LogGamma	0.9994	0.1460	No se rechaza
K-means	Genes (log)	Johnson SU	0.9941	0.1878	No se rechaza
Umbrales	Genes (log)	LogGamma	0.9994	0.5575	No se rechaza
Umbrales	Genes (log)	Johnson SU	0.9941	0.5450	No se rechaza
Ambos	–	GenGamma	0.9994	0.05	Se rechaza
Ambos	–	Exponencial	0.9760	0.05	Se rechaza
Ambos	–	Weibull	0.9906	0.05	Se rechaza

En los datos reales, las distribuciones ajustadas a los datos transformados logarítmicamente (LogGamma y Johnson SU) no rechazaron H_0, mientras que en la escala original solo algunas, como Gamma y Weibull mínima, mostraron ajustes suficientes, sugiriendo que la organización de genes duplicados no es completamente aleatoria.

Los datos simulados presentan p-valores altos, como se esperaba, y permiten comparar los ajustes con los datos reales. Tras la transformación logarítmica, los datos reales se aproximan a los patrones observados en simulaciones, aunque algunas distribuciones presentes solo en simulaciones (GenGamma, Exponencial, Weibull) muestran que ciertos patrones reales no se explican únicamente por azar.

En conclusión, la transformación logarítmica facilita identificar distribuciones que reflejan la organización espacial de genes duplicados y permite comparar coherentemente datos reales y simulaciones, apoyando la hipótesis de que la distribución de estos genes *no es totalmente aleatoria*.

4.4. Análisis específico del clúster de Escitonemina

En los 18 genomas que contienen el clúster de *Escitonemina*, las distancias entre genes fueron extremadamente pequeñas (mediana = 1 KB; IQR = 1–2 KB), lo que indica un fuerte agrupamiento físico. Debido a esta concentración, no fue posible ajustar distribuciones teóricas, pero el patrón observado respalda la hipótesis de organización no aleatoria. La cercanía entre genes podría reflejar necesidades funcionales sugiriendo que, al menos en este caso, la expansión génica responde a presiones biológicas que favorecen su agrupamiento.

Capítulo 5

Conclusiones y perspectivas

Este estudio sugiere que la distribución espacial de genes pertenecientes a familias génicas expandidas en cianobacterias no es completamente aleatoria. La combinación de análisis empíricos y simulaciones indica la existencia de patrones de agrupamiento que pueden modelarse estadísticamente, especialmente tras aplicar una transformación logarítmica. No obstante, los resultados también muestran que la metodología utilizada para clasificar genes (umbrales vs. *K-means*) influye de manera significativa en la detección de dichos patrones.

A futuro, sería relevante:

- Evaluar otros métodos de clasificación de genes pertenecientes a familias génicas expandidas, incluyendo enfoques más sofisticados como redes de similitud o herramientas bioinformáticas especializadas.
- Ampliar la muestra de genomas analizados, tanto en número como en diversidad taxonómica, para verificar si los patrones observados se mantienen en distintos grupos de cianobacterias o incluso en otros linajes bacterianos.
- Analizar clústeres biosintéticos distintos a la *Escitonemina*, con el fin de explorar si el patrón de agrupamiento es generalizable o particular de ciertos compuestos.
- Considerar factores genómicos adicionales —como la orientación de los genes, las regiones intergénicas o la presencia de operones— que podrían influir en la organización espacial de genes duplicados.

Estas líneas de trabajo permitirán refinar la hipótesis sobre la no aleatoriedad en la expansión génica y profundizar en sus posibles implicaciones funcionales y evolutivas.

Apéndice A

Clasificación de genes

A.0.1. Clasificación de genes por umbrales

Se clasificaron los genes de cada familia en dos categorías: **primarios**, aquellos genes considerados originales de la familia, y **secundarios**, provenientes de duplicaciones o transferencias horizontales, con base en su desempeño en el genoma a partir de los valores medios de e-value y porcentaje de identidad.

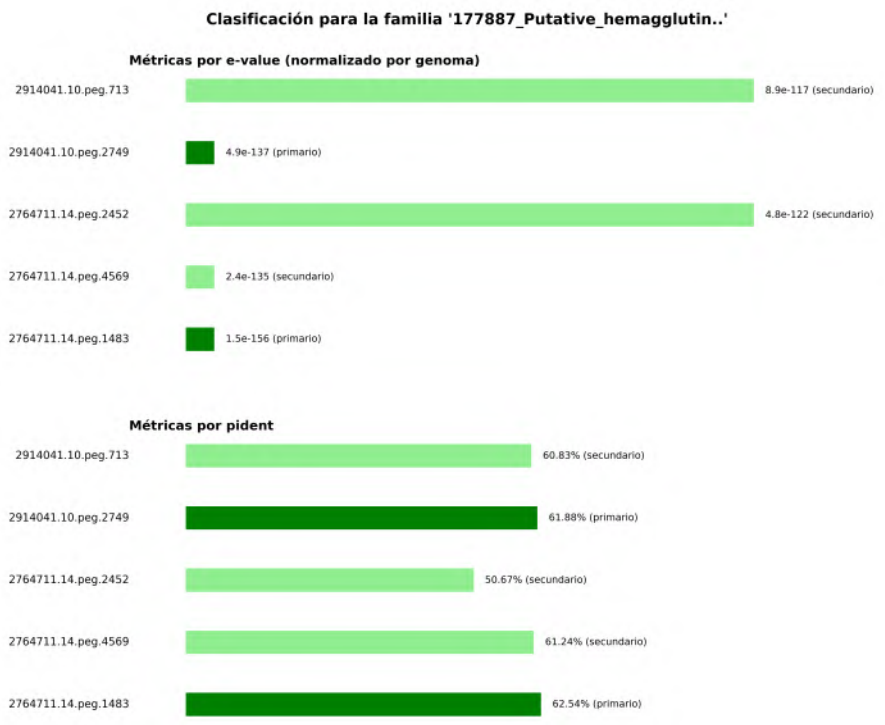

Figura A.1: Clasificación de algunos genes de la familia 177887_Putative_hemagglutin. En la parte superior se muestran las barras del porcentaje de identidad (pident) para cada gen. En la parte inferior se representan los e-values, normalizados de forma independiente dentro de cada genoma. Los genes se agruparon por genoma y se usaron dos tonalidades: verde oscuro para los genes primarios y verde claro para los secundarios.

A.0.2. Clasificación de genes mediante K-means (con $k = 2$)

Este enfoque se basó en los valores estandarizados de *mean e-value* y *mean pident* para cada gen dentro de la familia. Se formaron dos clústeres: el que contenía al gen con las métricas óptimas se etiquetó como **primario**, y el otro, como **secundario**.

Este método permite una clasificación automática y basada en la agrupación natural de los datos.

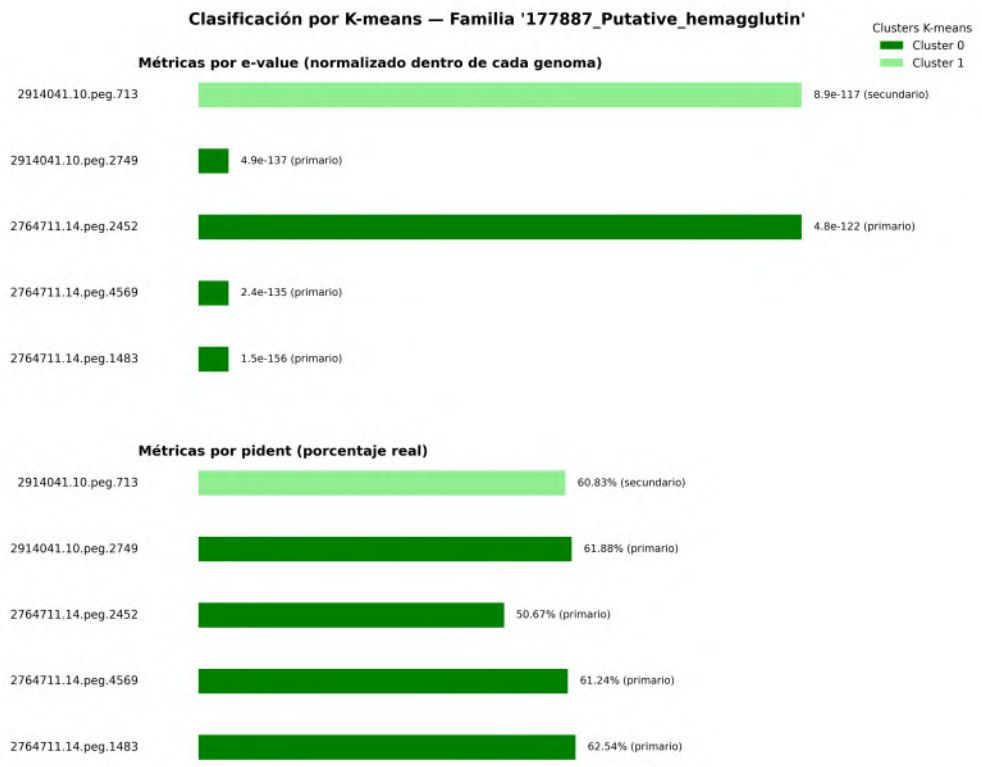

Figura A.2: 0.9 Clasificación para algunos de los genes de la familia *177887_Putative_hemagglutin* según la clasificación con K-means. La parte superior muestra los porcentajes medios de identidad (pident) y la parte inferior los valores de e-value normalizados por genoma para resaltar diferencias relativas. Los genes se codificaron con colores según su etiqueta: verde oscuro para "primario" y verde claro para "secundario".

Este ejemplo muestra cómo K-means divide los genes en dos grupos según sus características, facilitando la identificación de genes primarios y secundarios en familias con expansión génica.

Apéndice B

Cálculo de distancias

B.0.1. Ejemplo del cálculo de distancias en número de genes

Ejemplo para el genoma `103690.82`, el cual tiene $N = 5854$ genes:
Posiciones ordinales (lista ordenada):

$$(r_1, \ldots, r_{18}) = (446,\ 1364,\ \ldots,\ 5250).$$

Cálculo primera distancia

$$d_1 = r_2 - r_1 = 1364 - 446 = 918 \quad \text{(genes)}.$$

Cálculo distancia circular (último → primero)

$$d_{18} = r_1 + N - r_{18} = 446 + 5854 - 5250 = 1050 \quad \text{(genes)}.$$

Observación: los valores d_i representan la diferencia en la numeración ordinal de genes (p. ej. separación $446 \to 1364 = 918$ genes), no coordenadas en pares de bases.

B.0.2. Ejemplo del cálculo de distancias en kilobases

Genoma `1914872.23`, longitud total $L = 5294{,}286$ Kb

Sea la lista ordenada de genes duplicados secundarios con sus posiciones iniciales $\{p_{\text{start}}\} = [914806, 1257459, \ldots, 5215133]$ y finales $\{p_{\text{stop}}\} = [913391, 1258517, \ldots, 5216221]$.

Aplicando

$$d_i = \frac{\text{mín}\left(p_{g_{i+1},\text{start}}, p_{g_{i+1},\text{stop}}\right) - \text{máx}\left(p_{g_i,\text{start}}, p_{g_i,\text{stop}}\right)}{1000}$$

obtenemos, por ejemplo, para G_1 y G_2:

$$d_1 = \frac{\text{mín}(1257459, 1258517) - \text{máx}(914806, 913391)}{1000} = \frac{1257459 - 914806}{1000} = 342{,}653\,\text{Kb}.$$

Para la distancia circular entre el último y el primero:

$$d_{\text{circular}} = \frac{\text{mín}(914806, 913391) + L - \text{máx}(5215133, 5216221)}{1000}$$

$$= \frac{913391 + 5294286 - 5216221}{1000} = 991{,}456\ \text{Kb}.$$

Apéndice C

Distribución del clúster de Escitonemina

Figura C.1: Visualización circular de los genes del clúster de escitonemina en cada genoma analizado. Cada subfigura representa un genoma, y cada punto marca la posición relativa de un gen identificado, se considera la estructura circular de los genomas bacterianos.

Apéndice D

Procedimiento para la Revisión de Originalidad de las Tesis de Posgrado

Johanna Atenea Carreón Baltazar

Análisis espacial de genes de expansiones de familias génicas en Cianobacteria.pdf

Universidad Michoacana de San Nicolás de Hidalgo

Detalles del documento

Identificador de la entrega
trn:oid:::3117:529461741

Fecha de entrega
18 nov 2025, 8:30 a.m. GMT-6

Fecha de descarga
18 nov 2025, 8:32 a.m. GMT-6

Nombre del archivo
Análisis espacial de genes de expansiones de familias génicas en Cianobacteria.pdf

Tamaño del archivo
3.1 MB

36 páginas

9618 palabras

54.768 caracteres

Bibliografía

[1] G. C. Conant and K. H. Wolfe, "Turning a hobby into a job: how duplicated genes find new functions," *Nature Reviews Genetics*, vol. 9, no. 12, pp. 938–950, 2008, doi: 10.1038/nrg2482.

[2] P. M. Shih, D. Wu, A. Latifi, S. D. Axen, D. P. Fewer, E. Talla, A. Calteau, F. Cai, N. T. de Marsac, R. Rippka, M. Herdman, K. Sivonen, T. Coursin, T. Laurent, L. Goodwin, M. Nolan, K. W. Davenport, C. S. Han, E. M. Rubin, *et al.*, "Improving the coverage of the cyanobacterial phylum using diversity-driven genome sequencing," *Proceedings of the National Academy of Sciences*, vol. 110, no. 3, pp. 1053–1058, 2013, doi: 10.1073/pnas.1217107110.

[3] S. R. Miller, A. M. Wood, R. E. Blankenship, M. Kim, and S. Ferriera, "Dynamics of gene duplication in the genomes of chlorophyll d-producing cyanobacteria: implications for the ecological niche," *Genome Biology and Evolution*, vol. 3, pp. 601–613, 2011, doi: 10.1093/gbe/evr060.

[4] J. F. Sánchez-Herrero, *et al.*, "Distribution patterns of gene duplications in Staphylococcus aureus and related species," *Frontiers in Molecular Biosciences*, vol. 7, p. 160, 2020, doi: 10.3389/fmolb.2020.00160.

[5] A. Herrero, A. M. Muro-Pastor, and E. Flores, "Nitrogen control in cyanobacteria," *Journal of Bacteriology*, vol. 183, no. 2, pp. 411–425, 2001, doi: 10.1128/JB.183.2.411-425.2001.

[6] E. Flores and A. Herrero, "Compartmentalized function through cell differentiation in filamentous cyanobacteria," *Nature Reviews Microbiology*, vol. 8, no. 1, pp. 39–50, 2010, doi: 10.1038/nrmicro2242.

[7] E. S. Donkor, "Sequencing of bacterial genomes: principles and insights into pathogenesis and development of antibiotics," *Genes (Basel)*, vol. 4, no. 4, pp. 556–572, 2013, doi: 10.3390/genes4040556.

[8] C. Cassier-Chauvat, T. Veaudor, and F. Chauvat, "Comparative Genomics of DNA Recombination and Repair in Cyanobacteria: Biotechnological Implications," *Frontiers in Microbiology*, vol. 7, Nov. 2016, doi: 10.3389/fmicb.2016.01809.

[9] R. Prabha, D. P. Singh, P. Somvanshi, and A. Rai, "Functional profiling of cyanobacterial genomes and its role in ecological adaptations," *Genomics Data*, vol. 9, pp. 152–159, 2016, doi: 10.1016/j.gdata.2016.06.005.

[10] R. V. Popin, D. O. Alvarenga, R. Castelo-Branco, D. P. Fewer, and K. Sivonen, "Mining of cyanobacterial genomes indicates natural product biosynthetic gene clusters located in conjugative plasmids," *Frontiers in Microbiology*, vol. 12, p. 748, 2021, doi: 10.3389/fmicb.2021.684535.

[11] A. Calteau, D. P. Fewer, A. Latifi, T. Coursin, T. Laurent, J. Jokela, *et al.*, "Phylum-wide comparative genomics unravel the diversity of secondary metabolism in cyanobacteria," *BMC Genomics*, vol. 15, p. 977, 2014, doi: 10.1186/1471-2164-15-977.

[12] R. K. Aziz, D. Bartels, A. A. Best, M. DeJongh, T. Disz, R. A. Edwards, *et al.*, "The RAST Server: rapid annotations using subsystems technology," *BMC Genomics*, vol. 9, p. 75, 2008, doi: 10.1186/1471-2164-9-75.

[13] T. Seemann, "Prokka: rapid prokaryotic genome annotation," *Bioinformatics*, vol. 30, no. 14, pp. 2068–2069, 2014, doi: 10.1093/bioinformatics/btu153.

[14] B. Contreras-Moreira and P. Vinuesa, "GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis," *Applied and Environmental Microbiology*, vol. 79, no. 24, pp. 7696–7701, 2013, doi: 10.1128/AEM.02411-13.

[15] G. Gautreau, A. Bazin, M. Gachet, R. Planel, L. Burlot, M. Dubois, *et al.*, "PPanGGOLiN: Depicting microbial diversity via a partitioned pangenome graph," *PLoS Computational Biology*, vol. 16, no. 3, 2020, doi: 10.1371/journal.pcbi.1009687.

[16] D. M. Emms and S. Kelly, "OrthoFinder: phylogenetic orthology inference for comparative genomics," *Genome Biology*, vol. 20, p. 238, 2019, doi: 10.1186/s13059-019-1832-y.

[17] M. N. Price, P. S. Dehal, and A. P. Arkin, "FastTree 2 – Approximately maximum-likelihood trees for large alignments," *PLoS ONE*, vol. 5, no. 3, p. e9490, 2010, doi: 10.1371/journal.pone.0009490.

[18] K. Blin, S. Shaw, A. M. Kloosterman, Z. Charlop-Powers, G. P. van Wezel, M. H. Medema, and T. Weber, "antiSMASH 6.0: improving cluster detection and comparison capabilities," *Nucleic Acids Research*, vol. 49, no. W1, pp. W29–W35, 2021, doi: 10.1093/nar/gkab335.

[19] J. C. Navarro-Muñoz, N. Selem-Mojica, M. W. Mullowney, *et al.*, "A computational framework to explore large-scale biosynthetic diversity," *Nature Chemical Biology*, vol. 16, pp. 60–68, 2020, doi: 10.1038/s41589-019-0400-9.

[20] P. Cruz-Morales, H. E. Ramos-Aboites, C. Licona-Cassani, N. Selem-Mojica, P. M. Mejía-Ponce, V. Souza-Saldívar, and F. Barona-Gómez, "EvoMining reveals the origin and fate of natural product biosynthetic enzymes," *Microbial Genomics*, vol. 2, no. 11, p. e000088, 2016, doi: 10.1099/mgen.0.000260.

[21] H. Kaessmann, "Origins, evolution, and phenotypic impact of new genes," *Genome Res.*, vol. 20, no. 10, pp. 1313–1326, Oct. 2010, doi: 10.1101/gr.101386.109.

[22] F. Jacob, *et al.*, "The operon: a group of genes with expression coordinated by an operator," *Comptes Rendus de l'Académie des Sciences*, vol. 250, pp. 1727–1729, 1960, doi: 10.1016/j.crvi.2005.04.005.

[23] T. J. Treangen and E. P. C. Rocha, "Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes," *PLoS Genetics*, vol. 5, no. 1, p. e1000326, 2009, doi: 10.1371/journal.pgen.1001284.

[24] G. Achaz, E. P. C. Rocha, P. Netter, and E. Coissac, "Origin and fate of repeats in bacteria," *Nucleic Acids Research*, vol. 30, no. 13, pp. 2987–2994, 2002, doi: 10.1093/nar/gkf391.

[25] N. L. Johnson, S. Kotz, and N. Balakrishnan, *Continuous Univariate Distributions, Volume 1*, 2nd ed. New York, NY, USA: Wiley, 1994.

[26] L. Wasserman, *All of Statistics: A Concise Course in Statistical Inference*. New York, NY, USA: Springer, 2004, doi: 10.1007/978-0-387-21736-9.

[27] F. J. Massey, "The Kolmogorov-Smirnov test for goodness of fit," *Journal of the American Statistical Association*, vol. 46, no. 253, pp. 68–78, 1951, doi: 10.1080/01621459.1951.10500769.

[28] N. I. Fisher, *Statistical Analysis of Circular Data*. Cambridge, UK: Cambridge University Press, 1993.

[29] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990, doi: 10.1016/S0022-2836

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo

A quien corresponda,

Por este medio, quien abajo firma, bajo protesta de decir verdad, declara lo siguiente:

- Que presenta para revisión de originalidad el manuscrito cuyos detalles se especifican abajo.
- Que todas las fuentes consultadas para la elaboración del manuscrito están debidamente identificadas dentro del cuerpo del texto, e incluidas en la lista de referencias.
- Que, en caso de haber usado un sistema de inteligencia artificial, en cualquier etapa del desarrollo de su trabajo, lo ha especificado en la tabla que se encuentra en este documento.
- Que conoce la normativa de la Universidad Michoacana de San Nicolás de Hidalgo, en particular los Incisos IX y XII del artículo 85, y los artículos 88 y 101 del Estatuto Universitario de la UMSNH, además del transitorio tercero del Reglamento General para los Estudios de Posgrado de la UMSNH.

Datos del manuscrito que se presenta a revisión		
Programa educativo	Posgrado Conjunto en Ciencias Matemáticas UNAM-UMSNH	
Título del trabajo	Análisis espacial de genes de expansiones de familias génicas en Cianobacterias	
	Nombre	**Correo electrónico**
Autor/es	Johanna Atenea Carreón Baltazar	
Director	Adriana Haydeé Contreras Peruyero	haydeeperuyero@matmor.unam.mx
Codirector		
Coordinador del programa	Elmar Wagner	elmar.wagner@umich.mx

Uso de Inteligencia Artificial		
Rubro	**Uso (sí/no)**	**Descripción**
Asistencia en la redacción	Sí	Se utilizó inteligencia artificial para complementar ideas, mejorar la coherencia del texto y asegurar una redacción fluida.

Formato de Declaración de Originalidad y Uso de Inteligencia Artificial

Coordinación General de Estudios de Posgrado
Universidad Michoacana de San Nicolás de Hidalgo

Uso de Inteligencia Artificial		
Rubro	**Uso (sí/no)**	**Descripción**
Traducción al español	No	
Traducción a otra lengua	No	
Revisión y corrección de estilo	Sí	Se empleó inteligencia artificial para mejorar la claridad, formalidad y ortografía del texto, así como para optimizar su estilo general.
Análisis de datos	Sí	Se utilizó inteligencia artificial como apoyo en la generación y depuración de códigos para el análisis de los datos.
Búsqueda y organización de información	Sí	Se empleó inteligencia artificial para localizar y organizar referencias bibliográficas que respaldaran el contenido del marco teórico.
Formateo de las referencias bibliográficas	Sí	Se utilizó inteligencia artificial para verificar la consistencia y el formato de las referencias bibliográficas.
Generación de contenido multimedia	No	
Otro	No	

Datos del solicitante	
Nombre y firma	Johanna Atenea Carreón Baltazar
Lugar y fecha	Morelia, Michoacán a 13 de Octubre de 2025.