UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

Maestría y Doctorado en Ciencias Bioquímicas

IDENTIFICACIÓN DE PROMOTORES σ54 BACTERIANOS CON BASE EN LA CONSERVACIÓN DE SECUENCIA NUCLEOTÍDICA.

TESIS

QUE PARA OPTAR POR EL GRADO DE:

Maestro en Ciencias

PRESENTA:

Lic. Maricela Carrera Reyna

TUTOR PRINCIPAL

Dr. Enrique Merino Pérez

IBT, UNAM.

MIEMBROS DEL COMITÉ TUTOR

Dr. José Luis Puente García

IBT, UNAM.

Dr. Víctor Manuel González Zúñiga

CCG, UNAM.

Cuernavaca, Morelos. Octubre, 2020

Agradecimiento al programa de posgrado en Ciencias Bioquímicas de la Universidad Nacional Autónoma de México (UNAM). El desarrollo de este proyecto fue realizado gracias al apoyo de beca para estudiantes, que otorga el Consejo Nacional de Ciencia y Tecnología (CONACyT), con el dinero derivado de los impuestos del pueblo mexicano.

El presente trabajo se desarrolló bajo la asesoría del Dr. Enrique Merino Pérez, en el Laboratorio número 6 en el grupo de Genómica Computacional, adscrito al Departamento de Microbiología Molecular del Instituto de Biotecnología, campus Cuernavaca, Morelos, México, de la Universidad Nacional Autónoma de México.

Tutor principal:

Dr. Enrique Merino Pérez

Instituto de Biotecnología (IBT) UNAM

Comité Tutoral:

Dr. José Luis Puente García

Instituto de Biotecnología (IBT), UNAM.

Dr. Víctor Manuel González Zúñiga

Centro de Ciencias Genómicas, UNAM.

Miembros del jurado:

Presidente - Dr. Mario Soberón Chávez
Secretario - Dra. Clarita Olvera Carranza
Vocal - Dr. David Romero Camarena
Vocal - Dra. Katy Juárez López
Vocal - Dr. Ricardo Oropeza Navarro

Dedicatoria

A mi científica favorita: Lucía.

Te amo infinitamente.

Agradecimientos

Al Dr. Enrique Merino Pérez, muchísimas gracias. Este trabajo fue posible en gran medida a su apoyo y dedicación como tutor, que con paciencia y respeto ha encaminado mi formación académica. Gracias por todo su tiempo, confianza y optimismo para el desarrollo de este trabajo. También quiero agradecer al Ing. Ricardo Ciria Mérce, que gracias a su trabajo (que en conjunto con el Dr. Merino) hace más amena la vida bioinformática. Gracias También a la Dra. Rosa María Gutiérrez Ríos por sus consejos y orientación, además del apoyo personal.

Al comité tutoral por su tiempo, comentarios y su bien atinada asesoría, el tiempo se pasaba volando en los tutorales.

Muchas gracias también a los miembros del jurado revisor, sus observaciones, tiempo y sugerencias para mejorar el escrito fueron esenciales para lograr esta tesis.

A la comunidad de Stack Overflow, que durante mi aprendizaje del lenguaje de programación Perl, fue de gran utilidad.

Al Lic. J. Antonio Bolaños Guillen de la Unidad de Docencia por siempre ser tan accesible para resolver cualquier duda sobre los trámites burocráticos, y facilitar los mismos. Es un elemento muy valioso en esta institución.

Muchas gracias a l@s chic@s del grupo de la Dr. Guadalupe Espín, su siempre buena vibra y actitud hicieron cada reunión un evento muy especial. A mis compañer@s y amig@s que fueron relevantes en el transcurso de mi maestría, tanta ayuda, cariño, y ánimos , los agradeceré siempre, soy afortunada en tenerlos en mi vida tanto académica como personal. En particular, agradezco mucho a Joselyn Chavez, por su ayuda para aprender R y generar las gráficas para la presente tesis.

Y por su puesto, esta tesis, ni mi maestría hubiera sido posible sin la ayuda y amor incondicional de ¡mi madre!, que es una mujer excepcional. Así como el apoyo y cariño de mis hermanos y mi familia Murphy-Pérez. Gracias infinitas, los amo mucho.

Gracias a mi Murpholino. Por su amor, paciencia, los chistes, café y demás... Gracias por acompañarme en esta aventura de descubrir el mundo científico y la dicha de ser y estar con nosotras.

A mi preciosa hija Lucía, que es la razón principal para seguir.

TABLA DE CONTENIDOS

Índice de figuras

índice de Tablas

Tabla Página

índice de Gráficas

ABREVIATURAS

RNA: Ácido Ribonucleico.

DNA: Ácido Desoxirribonucleico.

RNAP: RNA polimerasa.

TSS: Sitio de inicio de transcripción.

tRNA: RNA de transferencia.

rRNA: RNA ribosomal.

sRNA: RNA pequeños.

mRNA: RNA mensajero.

RBP: Proteínas de unión a RNA.

TF: Factores transcripcionales.

HMM: Modelos Ocultos de Markov.

PSSM: Matrices de puntuación de posiciones específicas.

EM: Máxima expectancia.

EBP54: Proteína potenciadora de la unión de σ 54 (Enhancer Binding Protein of σ 54).

MEME: Maximización de la Expectancia Múltiple para la obtención de Motivos.

MAST: Herramienta de alineación y búsqueda de motivos.

RBD: Base de Datos de Regulón.

COG: Clúster de Grupos Ortólogos.

KEGG: Base de Datos de genes y genomas de Kioto.

UTR: Región sin traducir.

BLAST: Herramienta utilizada para comparar y alinear secuencias de tipo local.

TFBS: Sitio de unión de un factor transcripcional.

Resumen

La transcripción es un proceso fundamental que permite la expresión de información genética. La RNA polimerasa dependiente de DNA(RNAP), utiliza una cadena del dúplex de DNA como plantilla para producir moléculas de RNA complementarias que sirven en la traducción (tRNA, rRNA), síntesis de proteínas (mRNA) y regulación (sRNA). Aunque el núcleo de la RNAP es catalíticamente competente para la síntesis de RNA, la selectividad del inicio de la transcripción requiere un factor sigma (σ) para el reconocimiento y apertura del promotor. La expresión de factores σ alternativos proporciona un mecanismo poderoso para controlar la expresión de conjuntos discretos de genes (un regulón σ, o sigmulón) en respuesta a señales específicas relacionadas con el desarrollo, la nutrición o el estrés. Los factores σ determinan la transcripción de genes particulares uniéndose a promotores a los que reconocen de manera específica. Los principales factores σ y sus respectivas secuencias de reconocimiento han sido descritos para la mayoría de los organismos modelo; sin embargo, para los organismos menos caracterizados, los consensos de secuencias promotoras distintas a los promotores *housekeeping* o σ70, son mayormente desconocidos. Debido a lo anterior y al elevado costo, tanto económico como en tiempo, para la caracterización experimental de promotores, es importante la elaboración de protocolos computacionales para la identificación *in silico* de las secuencias de promotores que reconocen el complejo RNAP en conjunto con la gama de factores σ presentes en eubacterias. En la presente tesis se desarrolló un protocolo computacional para la identificación de los promotores dependientes del factor σ54. Pese a que los genes transcritos por la interacción de RNAP y este factor σ han sido caracterizados en diferentes organismos modelo, la intensión de estudiar al factor σ54 como modelo de estudio fue la de sentar las bases para el estudio de factores σ menos caracterizados. Este trabajo partió de una primera hipótesis para definir el conjunto inicial de secuencias o secuencias semilla en un proceso cíclico de construcción de modelo búsqueda de hits (pattern-finding/pattern-matching). Este proceso consideraba como tendencia general que los factores σ, dirigen la transcripción los genes que los codifican; es decir, se autotranscriben. Para nuestra sorpresa, el estudio detallado en grupos de organismos poco caracterizados reveló que dicha hipótesis sólo se cumplía en la minoría de los casos. Un segundo supuesto que resultó ser parcialmente cierto, fue que los motivos de secuencias conservadas en las regiones de regulación de genes que pertenecen a un mismo sigmulón (genes trascritos por el mismo factor σ), deberían de estar enriquecidos en las secuencias promotoras en común. La presente investigación reveló que en

muchos casos existen regiones altamente conservadas, ya sea por motivos asociados a una regulación en común, o por cierto tipo de sesgo en secuencias repetidas con funciones aún no conocidas, o bien, por la presencia de secuencias con poco contenido informacional. En el desarrollo de nuestra investigación se intentaron varios protocolos de enriquecimiento de la señal del motivo diana, que lograron superar los problemas que resultaron por el limitado cumplimento de las dos hipótesis antes mencionadas. El protocolo modificado partió de un motivo consenso general para el factor σ54, que se fue refinando para cada grupo filogenético estudiado mediante un proceso cíclico y la elección de matrices en base a criterios de especificidad, sensibilidad y precisión de las predicciones, logrando una precisión de 94.7%. Con este análisis, se ha logrado tener una visión más íntegra del sigmulón σ54 y el posible origen evolutivo de este factor σ en bacterias. El análisis de enriquecimiento de los genes predichos cuya expresión es regida por el factor σ54 nos permitió definir los procesos metabólicos principalmente regulados en todas las bacterias por este factor σ, así como procesos celulares específicos de un clado filogenético particular. Esperamos que el conocimiento generado a partir de esta investigación pueda ser empleado, en un futuro próximo, en análisis similares de otros factores σ en bacterias, además que estas metodologías sirvan para generar información o predicciones que complementen los datos experimentales, tanto en organismos modelo como los menos estudiados.

ABSTRACT

Transcription is a fundamental process that allows the expression of genetic information. DNA-dependent RNA polymerase (RNAP) uses a strand of the DNA duplex as a template to produce complementary RNA molecules that serve in translation (tRNA, rRNA), protein synthesis (mRNA), and regulation (sRNA). Although the nucleus of RNAP is catalytically competent for RNA synthesis, the selectivity of transcription initiation requires a sigma factor (σ) for promoter recognition and openness. The expression of alternative σ factors provides a powerful mechanism to control the expression of discrete sets of genes (a σ regulon, or sigmulon) in response to specific signals related to development, nutrition, or stress. σ factors determine the transcription of particular genes by binding to promoters that they specifically recognize. The main σ factors and their respective recognition sequences have been described for most model organisms; however, for less characterized organisms, the consensus of promoter sequences other than the *housekeeping* or σ70 promoters are largely unknown. Due to the above and the high cost, both economic and time, for the experimental characterization of promoters, it is important to develop computational protocols for *the in silico* identification of promoter sequences that recognize the RNAP complex in conjunction with the range of σ factors present in eubacteria. In the present thesis, a computational protocol was developed for the identification of promoters dependent on the σ54 factor. Although the genes transcribed by the interaction of RNAP and this factor σ have been characterized in different model organisms, the intention of studying factor σ54 as a study model was to lay the foundations for the study of less characterized σ factors. Our work started from a first hypothesis to define the initial set of sequences or seed sequences in a cyclical process of construction of a pattern search of hits (pattern-finding/pattern-matching). This process considered as a general trend that the σ factors direct transcription the genes that encode them; that is, they self-transcribe. To our surprise, the detailed study in groups of poorly characterized organisms revealed that this hypothesis was only fulfilled in the minority of cases. A second assumption that turned out to be true was the reason for the conserved sequences in the regulatory regions of genes that belonged to the same

sigmulon (genes transcribed by the same σ factor), should be enriched in the common promoter sequences. Our study revealed that in many cases there are highly conserved regions, either for reasons associated with common regulation, or for a certain type of bias in repeated sequences with functions not yet acknowledged, or due to the presence of sequences with few informational contents. In the development of our research, various protocols for enriching the target motif signal were attempted, which managed to overcome the problems that resulted from the limited compliance with the two aforementioned hypotheses. Our modified protocol starts with a consensus motif general for the σ54 factor, which was refined for each phylogenetic group studied through a cyclical process and the choice of matrices based on criteria of specificity, sensitivity, and precision of the predictions, achieving a precision of 94.7%. With our study, we have acquired a complete view of the σ54 sigmulon and the possible evolutionary origin of this σ factor in bacteria. The enrichment analysis of the predicted genes whose expression is governed by the σ54 factor allowed us to define the metabolic processes mainly regulated in all bacteria by this σ factor, as well as specific cellular processes of a particular phylogenetic clade. We hope that the knowledge generated from this research can be before long employed, in similar analyzes of other σ factors in bacteria.

Identificación de promotores σ54 bacterianos con base en la conservación de secuencia nucleotídica.

INTRODUCCIÓN

En organismos bacterianos uno de los pasos más importantes en la regulación de la expresión genética es el inicio de la transcripción, en el que la RNA polimerasa dependiente de DNA (RNAP) es la enzima clave. La RNAP es la maquinaria catalítica para la síntesis de RNA a partir de una secuencia de DNA que sirve como patrón o molde. Aunque el núcleo de la RNAP es catalíticamente competente para la síntesis de RNA, la selectividad del inicio de la transcripción requiere un polipéptido adicional conocido como **factor σ**(**σ**) para el reconocimiento y apertura del promotor (Borukhov y Nudler, 2003). Juntos, el factor σ y el "core" de la RNAP, que consta de cinco subunidades (2α, β, β' y ω), forman una enzima específica de iniciación, la holoenzima RNAP como se esquematiza en la **figura 1.** Los factores σ especifican la transcripción bacteriana uniéndose a un promotor característico y reclutando así la RNA polimerasa asociada a ese promotor (R.R. Burgess, 2001). **El promotor** es la secuencia de DNA conservada que señala y dirige la transcripción de un gen o grupo de genes adyacentes que es reconocida por la holoenzima RNAP y que determina el sitio inicio y la frecuencia de la transcripción. Los factores σ son una familia de proteínas relativamente pequeñas que pueden asociarse de manera reversible con la RNAP (Campbell *et al.*, 2002). A cada subfamilia se le ha asignado un papel funcional global y cada factor σ, reconoce una secuencia promotora consenso distinta. Los promotores se consideran factores clave para la transcripción, ya que su reconocimiento, es el paso inicial en la expresión génica y parte de la regulación transcripcional (Wösten y Wo, 1998).

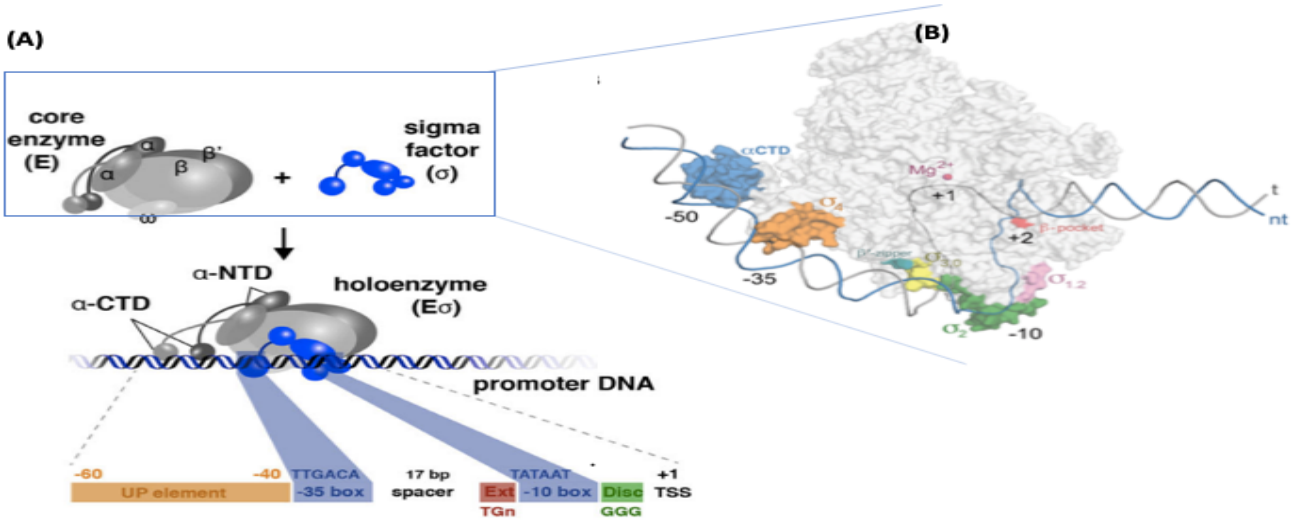


Figura 1 Modelo del complejo abierto y la interacción de la RNAP-DNA . (A) Representación esquemática de la holoenzima RNAP compuesta por el núcleo central de la enzima (α2ββ´ Ω) y un factor σ que es responsable del reconocimiento del promotor. UP es el elemento promotor río arriba que hace contacto con α-CTD, mientras que los elementos promotores –10 y –35 interaccionan con diferentes partes o regiones de σ. Ext representa el elemento promotor –10 extendido y Disc el sitio discriminador. TSS representa el inicio de la transcripción, que está compuesto principalmente por purinas. Las secuencias y las distancias corresponden al promotor de consenso para el factor de mantenimiento celular (σ70) de *E. coli*. **(B)** Se esquematizan las interacciones de la holoenzima RNAP y su región promotora, donde se aprecia que la apertura ocurre primeramente sobre la caja -10 del promotor. La RNAP se muestra como una superficie transparente gris, excepto parches involucrados en el reconocimiento específico de la secuencia promotora. La estructura principal del DNA se esquematiza en dos colores, gris y azul (la hebra molde). El modelo fue creado combinando coordenadas del PDB 4G7H, 3UGO, 1LB2 y 1L9Z. Tomado de (Feklistov, 2013).

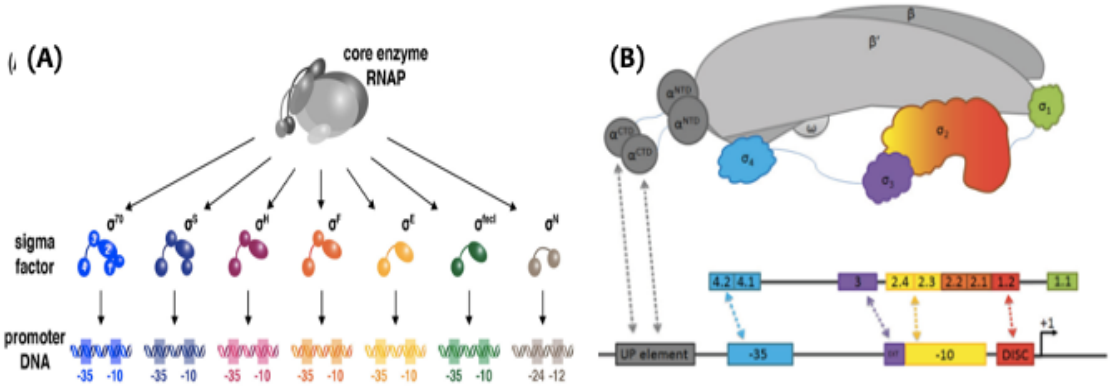
Factores σ.

El primer factor σ descubierto fue el factor σ70 de la bacteria *Escherichia coli,* organismo modelo, altamente estudiado. Su descubrimiento alrededor del año 1969, fue un resultado inesperado de los intentos de comprender la estructura de las subunidades de la RNA polimerasa.(BURGESS *et al.*, 1969). Para 1985, se hizo el hallazgo de que

los factores σ primarios de *E. coli* y *B. subtilis* son proteínas homólogas, aunque son diferentes en cuanto a tamaño. En un gel SDS-PAGE, el factor σ de *E. coli* migra a 90 kDa y su masa es de 70 kDa, mientras que el σ de *B. subtilis* su masa es de 43 kDa por lo que se les nombró σ70 y σ43 respectivamente (antes σ55). La nomenclatura de los factores σ se tornó un tema polémico, *Losick et al*, propusieron que se les nombrara de acuerdo al alfabeto, además que la asociación de las letras, fuesen de acuerdo con los nombres de los genes que los codifican, siendo así que a los factores primarios se les renombró como σA, σ37 como σB (codificado por *sigB*), σ28 a σD (codificado por *sigD*), y así con los demás (Losick *et al.*, 1986). Este sistema demostró ser bastante adaptable y solo se enfrentó a un serio desafío con la expansión de la familia σ para incluir factores σ ECF (Factores extracitoplasmáticos), que, junto con el aumento exponencial de secuencias en las bases de datos, llevaron a la comprensión de que algunas bacterias tienen más (y a veces mucho más) de 26 factores σ. Hubo algunas excepciones a este sistema de nombres, en gran medida por razones históricas. En *E. coli*, por ejemplo, los factores σ están codificados por genes *rpo* (por ejemplo, *rpoD* para σA y *rpoH, rpoS*, etc. para factores σ alternativos)(Helmann, 2019).

El factor σ70 (o σA) es el factor σ central en *E. coli* y muchas otras especies ya que dirige la transcripción de la mayoría de los genes requeridos para el crecimiento exponencial o *housekeeping* (de mantenimiento*)*. El sitio de unión para la familia de promotores σ70 se define por dos hexámeros de consenso, TTGACA y TATAAT, ubicados aproximadamente a -35 y -10, respectivamente en relación con el sitio de inicio de la transcripción (TSS, por sus siglas del inglés transcription start site). Dichos hexámeros se encuentran separados por un espacio de 15 a 21 pb. Adicionalmente a estos dos hexámeros, existe un elemento ubicado río arriba llamado "*Up-Element*" (Feklistov, 2013). El factor σ realiza dos funciones principales:1) dirigir el núcleo catalítico de RNAP hacia el promotor río arriba del sitio de inicio de la transcripción +1 y 2) ayudar en el inicio de la separación de la cadena del DNA de doble hélice, formando la "burbuja de transcripción" (Wösten y Wo, 1998; R.R. Burgess, 2001). Cada promotor genético utiliza una región promotora específica aproximadamente 40 pb río arriba del sitio de inicio de la transcripción, y, por lo tanto, diferentes factores σ juegan un papel importante en la regulación de diferentes genes (Mishra *et al.*, 2018). Este proceso, que incluye la

asociación del factor σ con RNAP para reconocer y abrir al DNA en el sitio del promotor, seguido de la disociación de σ para permitir el alargamiento (elongación), que luego puede activar enzimas RNAP adicionales, se denomina ciclo σ (Mauri y Klumpp, 2014). Hay muchos tipos de subunidades σ, y como se ha mencionado anteriormente, cada uno reconoce una secuencia promotora única. Además, cada σ único se compone de un número variable de dominios estructurados. Los factores σ más simples tienen dos dominios, pocos tienen tres, y la mayoría, llamados factores σ de mantenimiento, tienen 4 dominios, a los cuales se les ha dado los nombres de σ (4), σ (3), σ (2) y σ (1.1) (como se observa en la **figura 2**). Todos los dominios están unidos por enlaces peptídicos muy flexibles que pueden extenderse distancias muy largas. Cada uno de estos dominios utilizan sitios de unión al DNA, o dominios que reconocen secuencias y conformaciones específicas en el DNA. Más comúnmente, estas secuencias reconocidas se encuentran en los sitios -35 y -10 río arriba del sitio +1. Uno de estos motivos de unión al DNA, es el motivo de hélice-giro-hélice (HTH), que ayuda a reconocer específicamente los promotores de DNA en las posiciones -35 y -10. Este motivo HTH, utilizado por la mayoría de los factores σ, mantiene su especificidad y precisión al unirse en el surco principal del DNA, donde puede interactuar con los pares de bases en la doble hélice del DNA (Bervoets y Charlier, 2019). Dado que los factores σ están vinculados exclusivamente a la expresión génica en organismos procariotas, la variedad de factores σ en una célula dicta cómo y qué genes se transcriben (R R Burgess, 2001). La función especializada en las células está altamente moderada por su arsenal de subunidades σ. El desarrollo celular y la diferenciación son directamente afectados y llevados a cabo por "cascadas" de factores σ (R.R. Burgess, 2001). En las primeras etapas de desarrollo, los ¨genes tempranos¨ son transcritos por factores bacterianos básicos σ. Se ha observado como tendencia general en *Bacillus subtilis* que sus factores σ reportados, tienden a autotranscibirse. El ciclo de vida de este organismo modelo está altamente regulado, donde los factores σ juegan un papel importante en la regulación de la transición de estadía celular, como se observa en la **figura 3,** donde se esquematizan las cascadas que definen procesos celulares importantes de este organismo (Freyre-González *et al.*, 2013). Por lo tanto, estos genes se transcriben para generar nuevos factores σ, que a su vez activan genes adicionales, y así sucesivamente (Mauri y Klumpp, 2014).



(C)

Species	σ-factor	-35		-10	Function
E. coli					
	σ^{70}	TTGACA	- 17nt -	TATAAT	Housekeeping genes
	σ^{S}/σ^{38}	TTGACA	- 17nt -	TATACT	Stationary phase
	σ^{E}/σ^{24}	GAACT	- 17nt -	TCTGAt	Extracytoplasmic function
	σ^{F}/σ^{28}	TAAAgt	- 15nt -	GCCGATAA	Motility genes
	σ^{H}/σ^{32}	TTGAAA	- 15nt -	CCCCATNT	Heat shock response
	$\sigma^{Fecl}/\sigma^{19}$	aaAAT	- 17nt -	tTGTNt	Extracytoplasmic function
		-24		-12	
	σ^{N}/σ^{54}	TGGCAGG	- 4nt -	TTGCA	Nitrogen metabolism
B. subtilis		-35		-10	
	σ^{A}	TTGACA	- 17nt -	TATAAT	Housekeeping genes
	$\sigma^{B}(/\sigma^{C})$	GTTTaa	- 13-15nt -	GGGWAW	General stress response
	σ^{F}	NGTATWWN	- 14nt -	GKNNANNNTW	Early forespore genes
	σ^{G}	GHATA	- 17-18nt -	MAWAMTA	Late forespore genes
	σ^{H}	RNAGGAWWW	- 11-12nt -	RNNGAATWW	Sporulation and Stationary phase
	σ^{M}	tgaAAC	- 16-17nt -	CGTC	Extracytoplasmic function
	σ^{W}	tgaAAC	- 16-17nt -	CGTM	Extracytoplasmic function
	σ^{X}	tgtaACtttt	- 12-13nt -	CGWC	Extracytoplasmic function

Figura 2. Composición de los dominios de factores σ y su interacción con el promotor. (A) Los siete factores σ de *E. coli*, que se unen competitivamente a regiones similares del núcleo de RNAP, pero interactúan con secuencias promotoras específicas centradas en las posiciones –10 y –35 para los seis miembros de la familia σ70 y alrededor de –12 y –24 para σ54 (σ^{N}), el único representante de la familia σ54 **. (B)** Representación esquemática de la holoenzima RNAP, la estructura del factor σ, y la estructura representativa del promotor. En gris la estructura de las subunidades de la RNAP, en colores que coinciden con la representación de la región promotora, los 4 dominios de la subunidad σ y sus regiones de interacción. **(C)** Lista de diferentes factores σ, presentes en *Escherichia coli* y *Bacillus subtilis*, con sus respectivas secuencias promotoras consenso y función en la célula. Tomado y modificado de (Davis *et al.*, 2017; Bervoets y Charlier, 2019).

Como se ha mencionado con anterioridad, además del factor σ70 existen un conjunto muy grande de factores σ alternativos que alteran el reconocimiento del promotor, logrando así cambiar los patrones de transcripción para regular genes involucrados con funciones específicas. Estas funciones incluyen: esporulación; respuesta al choque térmico; entrada y mantenimiento de la fase estacionaria; expresión de genes de flagelos y control del metabolismo del nitrógeno. Algunos factores son específicos para las diferentes clases de bacterias, incluidos los factores involucrados en la esporulación. El número de factores σ puede variar dependiendo del tamaño de genoma del organismo y del tipo de vida que éste tiene. Como regla general, cuanto más complejos sean los nichos de ciclo vital y ambiental de una bacteria, mayor será el número de factores σ con los tipos de promotores correspondientes. Por ejemplo, en *Mycoplasma genitalium* únicamente se ha reportado un único factor σ (Torres-Puig *et al.*, 2015), en *E. coli* el número de factores σ es 7 (Cook y Ussery, 2013), en *B. subtilis* se han reportado alrededor de 19 factores σ (Haldenwang, 1995; Freyre-González *et al.*, 2013), mientras que *Streptomyces coelicolor* codifica para 65 factores σ (Bentley *et al.*, 2002). Cercanos al promotor, existen sitios en el DNA llamados *operadores* que permiten la unión de proteínas llamados factores transcripcionales (TF, por sus siglas del inglés Transcription Factors), que interaccionan con la RNAP, favoreciendo el reconocimiento del promotor (activadores), o impidiendo o constituyendo un impedimento estérico a su paso (represores) (Pérez-Rueda *et al.*, 2009; Wang *et al.*, 2012).

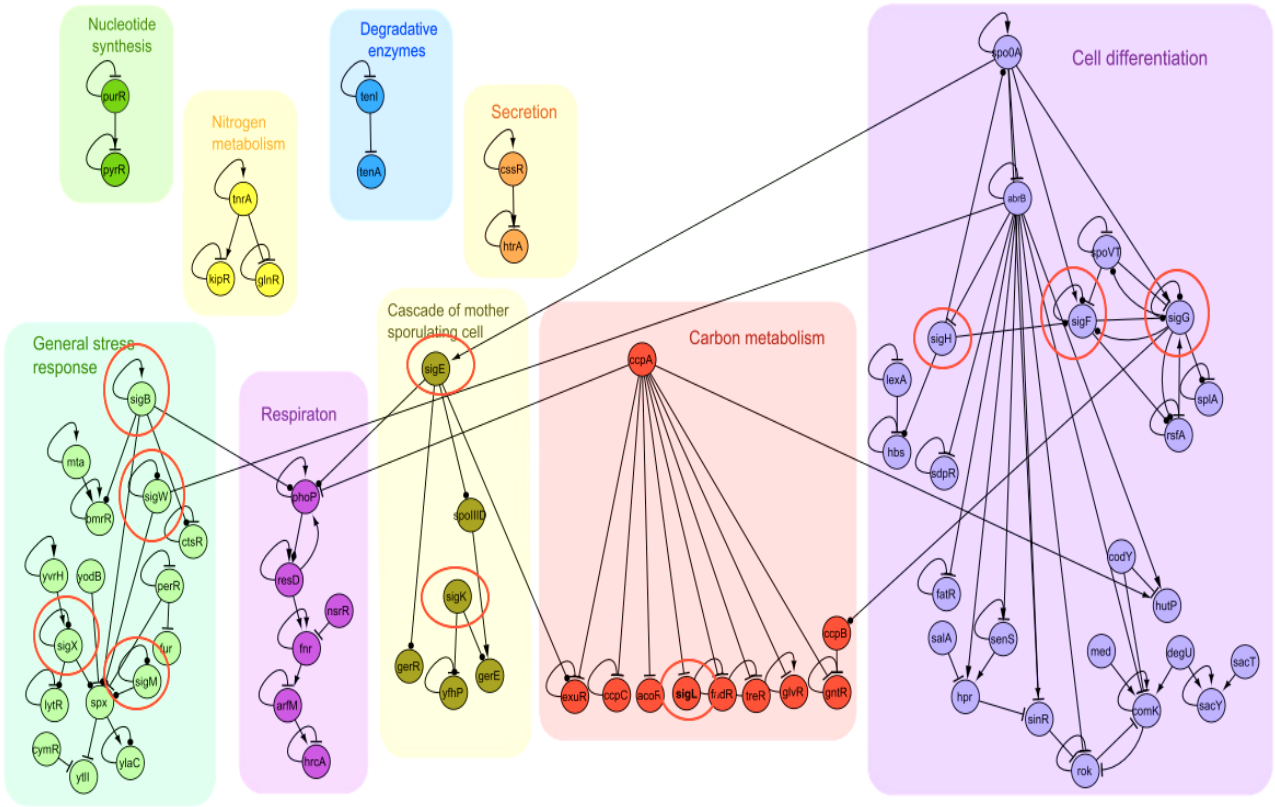


Figura 3 . Red de modulación de procesos celulares mediados por Factores Transcripcionales (TFs) en *Bacillus subtilis*. Se representa cada evento celular, separado por color, con sus respectivas cascadas de regulación mediada por TFs. En círculos color rojo se resaltan los factores σ que participan en cada proceso. Tomado de (Freyre-González *et al.*, 2013).

El factor σ54

El factor σ54 (también conocido como RpoN, σN, NtrA y codificado por el gen *rpoN* en *E. coli y por sigL en B. subtilis*) se une al núcleo de la RNA polimerasa través de una región que es similar a σ70 (Tintut *et al.*, 1995). Mediante microscopía de dispersión de rayos X se ha evaluado su estructura y se encontró que σ54 y σ70 comparten una gran similitud estructural. Sin embargo, a pesar de tener sitios conservados similares, el mecanismo de acción de σ54 es diferente al de σ70 (Svergun *et al.*, 2000). En la **figura 4**, se esquematizan dicho mecanismo, y se observa que una vez que RNAP-σ54 une al promotor característico en las regiones -24 /-12 con respecto del inicio de la transcripción, requiere la entrada de energía libre, proveniente de la hidrólisis de ATP y de un activador

asociado para iniciar la transcripción (Lin *et al.*, 2014). En la mayoría de los casos, el activador se une a un elemento potenciador ubicado río arriba del promotor y, por lo tanto, se denomina proteína potenciadora de la unión de σ 54 (EBP^{54}, por sus siglas del inglés Enhancer Binding Protein of σ 54) (Morett y Segovia, 1993). Los EBP^{54} se unen al DNA como dímeros inactivos, pero al recibir la señal apropiada, se ensamblan en anillos oligoméricos, con hexámeros que constituyen el estado activo oligomérico (Weiss *et al.*, 1991). Los promotores dependientes de RpoN no tienen los elementos conservados -35 y -10 que se encuentran típicamente en los promotores reconocidos por factores σ en la familia σ70 (Paget, 2015). La mayoría de los promotores dependientes de σ54 descritos hasta ahora contienen un motivo conservado de GG-10 pb-GC(Buck y Cannon, 1992b, 1992a). Se ha reportado que el consenso para 186 elementos promotores dependientes de RpoN de 47 especies bacterianas es YT-GGCACG (caja -24)- separados por 4 pb-TTGCWNN (sitio -12) (Barrios *et al.*, 1999)**.** σ54 constituye una familia de proteínas evolutivamente separada y se encuentra ampliamente distribuida entre el reino bacteriano, aunque algunos phyla carecen de esta proteína. El sistema ***σ***54-RNAP está ampliamente representado en las Proteobacterias Alpha y Gamma (Francke *et al.*, 2011; Tsoy *et al.*, 2016).

No se ha logrado asignar una función fisiológica común para los genes dependientes de σ54. Estudios en diversos organismos de los genes dependientes de σ54 indican que los productos de estos genes usualmente están involucrados en la asimilación de nitrógeno, pero no es una regla general. En *Escherichia coli,* se ha identificado una lista representativa de genes dependientes de σ54, mediante el análisis conjunto de experimentos con microarreglos de DNA de genes inducidos por la limitación de nitrógeno y análisis bioinformáticos. Se reporta que *E. coli* tiene aproximadamente 30 operones dependientes de σ54 y la mitad de ellos están involucrados en el metabolismo y asimilación de nitrógeno. Reitzer y Schneider en 2001 plantean que una posible relación fisiológica entre los genes dependientes de σ54 puede basarse en el hecho de que la asimilación de nitrógeno consume energía e intermediarios del metabolismo central. Debido a esto, los productos de los genes dependientes de σ54 que no están involucrados en el metabolismo del nitrógeno pueden prevenir el agotamiento de los metabolitos y los recursos energéticos en diferentes entornos (Reitzer y Barbara L

Schneider, 2001). Entre otros procesos fisiológicos controlados por σ54, además de la asimilación y fijación de nitrógeno (Hoover *et al.*, 1990), se encuentran: el transporte de ácido dicarboxílico, la oxidación de hidrógeno, la utilización de alginato, la degradación de compuestos aromáticos, la utilización de formiatos y la formación de *pilus* (Buck *et al.*, 2000; Tsang y Hoover, 2014). Basados en análisis de la expresión de transcriptoma, se reportó que existe un efecto antagónico global en la expresión génica entre σ54, σS y σ28 (motilidad mediada por flagelo). La interacción reguladora entre estos tres factores resulta ser bastante compleja. Se dice que aproximadamente el 60% de los genes en el regulón de RpoN están bajo el control recíproco de RpoS. Como ejemplo, se ha visto que RpoN controla positivamente la expresión del factor σF (σ28 ó FliA) y σS (RpoS) lo controla negativamente (Dong et al., 2011;Ahmar et al., 2018; Lobanovska et al., 2019).

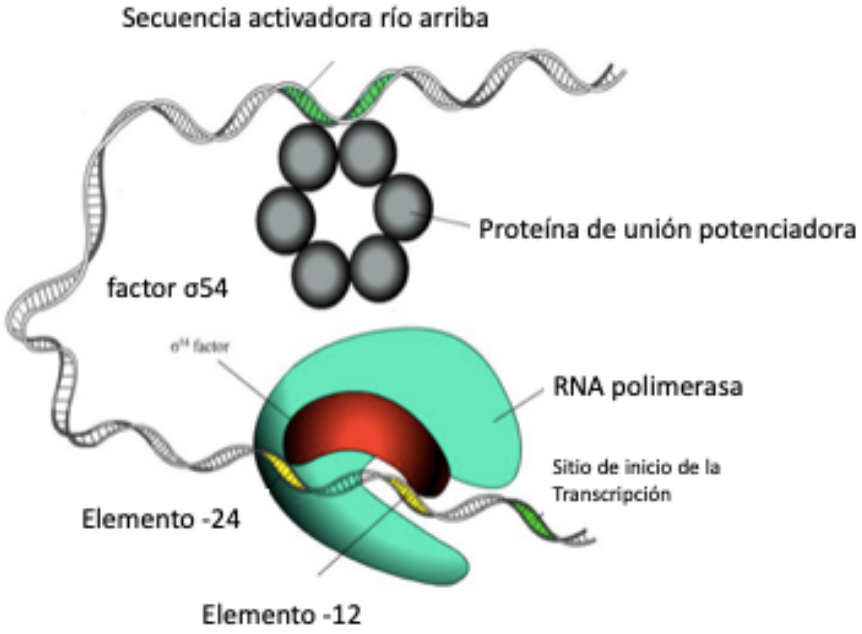


Figura 4. Esquema de la estructura básica del promotor σ54 y su proceso biológico. Se muestra las interacciones del complejo RNApolimerasa-σ54, tanto con las regiones promotoras -24 y -12 como de la forma oligomérica de la proteína de unión potenciadora, esta a su vez unida a la secuencia activadora río arriba. En rojo el factor σ54. Tomada y modificada de Hao Lin 2014 [Hao Lin 2014].

Modelos y algoritmos para identificar promotores.

En las últimas tres décadas se han desarrollado e implementado una serie de algoritmos para la identificación de promotores procarióticos, basados en características particulares de las secuencias promotoras, o motivos conservados en secuencias de DNA. A continuación, se mencionan algunos de los más importantes.

A) Matriz de peso de posición.

Matriz de peso de posición (PSSM, por sus siglas del inglés, Position-Specific Scoring Matrix). Uno de los modelos más utilizados son las PSSMs, debido a que capturan la diversidad de las secuencias que pueden ser reconocidas específicamente por un factor de transcripción; estas matrices se construyen a partir de un conjunto de secuencias que se sabe contienen algún patrón o sitio conocido, que se alinean y con base en este alineamiento se construye una matriz en donde en cada posición se anota el número de instancias de cada uno de los posibles cuatro nucleótidos y se representa una colección de sitios sobre-representados como un consenso (degenerado) o un logo de secuencia (Sequence logo) (Mrazek, 2009).

B) Algoritmos basados en Modelos ocultos de Márkov.

HMM (por sus siglas del inglés, Hidden Markov Model). Un modelo de Markov, también conocido como cadena de Markov, describe una secuencia de eventos que ocurren uno tras otro en secuencia. Cada evento determina la probabilidad del siguiente evento. Una cadena de Markov puede verse como un proceso que se mueve en una dirección de un estado al siguiente con una cierta probabilidad, la cual es conocida como la probabilidad de transición. No siempre es posible conocer u observar de forma directa la secuencia de estados por la que transita el modelo; es decir dichos estados se encuentran ocultos, a ello se debe su nombre de *Modelos ocultos de Markov*. A pesar de que en dichos modelos los estados se encuentran ocultos, es posible inferir dichos estados a partir de señales relacionadas y funciones que definen que tan probable es que las señales indirectas representen a los estados reales (Eddy, 2004). En términos

de secuencias biológicas, de DNA o proteína, los modelos ocultos de Markov sirven para representar motivos conservados, en donde cada nucleótido o aminoácido de un conjunto de secuencias alineadas, representan a los estados del sistema a los que se asignan probabilidades que definen estados posteriores (nucleótidos o aminoácidos) (Mrazek, 2009).

C) Redes neuronales artificiales.

Son un modelo matemático/computacional basado en la manera en que aprenden las redes neuronales naturales. La unidad operacional de estas redes es llamada neurona, haciendo símil al sistema biológico. Estas neuronas artificiales son "conectadas" entre sí mediante funciones matemáticas que interaccionan entre sí para generar un valor de salida o respuesta. En la red, el valor de salida de una neurona afecta a neuronas posteriores y la intensidad y la forma en la que lo hace, activación o inhibición, puede variar de neurona a neurona de acuerdo a un factor que es evaluado cada ciclo de aprendizaje. Adicionalmente a este factor de propagación, los valores que determinan el umbral limitante que define si una neurona se activará, es evaluado mediante funciones matemáticas cuyos parámetros son modificados ciclo con ciclo. Las redes neuronales artificiales "aprenden" a dar una respuesta modificando los parámetros de conectividad y umbral de activación para cada neurona de tal forma que minimicen una función de pérdida que evalúa la red conjuntamente al presentarles conjuntos de datos verdaderos con el que realizan su entrenamiento (Askary *et al.*, 2009).

D) Máquinas de vectores de soporte.

Son modelos de aprendizaje supervisados que analizan los datos utilizados para el análisis de clasificación y regresión. Dado un conjunto de ejemplos de entrenamiento, cada uno marcado como perteneciente a una categoría de dos posibles categorías. Un algoritmo de entrenamiento de máquinas de vectores de soporte construye un modelo que asigna nuevos ejemplos a una categoría u otra, convirtiéndolo en un clasificador lineal binario no probabilístico. De forma pragmática, las máquinas de vectores de soporte pueden conceptualizarse como un modelo que mapean los puntos de entrada a un espacio de características de una dimensión mayor, para luego encontrar el

hiperplano que los separe y maximice el margen entre las clases (Byvatov y Schneider, 2003).

E) Método de Maximización de la Expectancia Múltiple para la obtención de Motivos.

MEME, por sus siglas del inglés, Multiple Expectation maximization for Motif Elicitation. Este método está implementado en la famosa MEME-Suit (http://meme-suite.org/) que es una colección de herramientas para el análisis de motivos de secuencias conservadas en conjuntos de secuencias de ácidos nucleicos (DNA o RNA) o de aminoácidos. MEME se encuentra en línea a través de una interfaz web, o para descargar gratuitamente. Proporciona algoritmos computacionalmente eficientes para descubrir y analizar los motivos de secuencia característicos como, por ejemplo, de los sitios de unión de factores de transcripción (TFs), los sitios de unión de proteínas de unión a RNA (RBP) y elementos promotores o en proteínas, motivos que representan dominios funcionales o estructurales conservados. Entre sus herramientas se encuentra MEME y MAST (Motif Alignment & Search Tool) que se utilizan en este proyecto (Bailey *et al.*, 1998, 2006)De manera breve, MEME funciona mediante la búsqueda de patrones de secuencia repetidos y sin indeles (inserciones o deleciones), que existen en un conjunto de secuencias proporcionadas por el usuario que pueden provenir de resultados experimentales, por ejemplo, del conjunto de promotores de genes coexpresados, o de análisis *in silico*, como secuencias río arriba de genes ortólogos que pueden contener sitios de unión (la "señal") para el mismo factor de transcripción. Un motivo MEME es un patrón de secuencia que ocurre repetidamente en una o más secuencias en el conjunto de entrada. Los motivos MEME permiten variaciones en cualquier posición del patrón. MEME divide patrones que contienen indeles en múltiples motivos. Los motivos pueden aparecer en cualquier orden, varias veces o no aparecer en ninguna secuencia. De manera automática, MEME puede elegir el ancho y el número de ocurrencias[1] de cada motivo para optimizar el valor de la relevancia estadística de la sobre-representación del motivo. Dicha relevancia estadística es representada por el estadístico E (E-value) del

[1] Ocurrencia de ocurrir. En probabilidad, materialización de un evento aleatorio. La probabilidad de ocurrencia se refiere a la frecuencia de un evento o característica sin distinguir entre la prevalencia o incidencia.

motivo y corresponde a la probabilidad de encontrar un patrón igualmente bien conservado en un conjunto de secuencias generadas de manera aleatoria. Para identificar los motivos relevantes desde un punto de vista biológico, la elección del ancho del motivo debe de realizarse de manera cuidadosa, misma que la realiza el usuario, y debe tener en cuenta los siguientes criterios:

1. El mejor ancho de motivos.
2. El número de ocurrencias en cada secuencia.
3. La composición de cada motivo.

Dado un conjunto de secuencias de nucleótidos, el algoritmo de MEME utiliza varios tipos de funciones bien conocidas:

- Máxima Expectancia (EM).
- Heurística basada en EM para elegir el punto de partida de EM.
- Relación basada bajo el criterio de máxima verosimilitud que se basa en el uso de un valor llamado log-likelihood[1]. Heurística para determinar el mejor número de parámetros sin modelo.
- Inicio múltiple para buscar en posibles anchos de motivos (cuando no se indica éste).
- Búsqueda Greedy para encontrar múltiples motivos.

El algoritmo de descubrimiento de motivos busca un conjunto de secuencias cortas similares en un conjunto de secuencias mucho más largas. El problema es más fácil cuando las instancias de motivos son largas y muy similares entre sí (Bailey, 2003). Se vuelve mucho más difícil cuando las instancias de motivos son cortas y/o degeneradas, o las secuencias de entrada son muy largas. Descubrir motivos que correspondan al sitio de unión de un factor transcripcional (TFBS, por sus siglas del inglés transcription factor binding site), en un conjunto de secuencias de DNA (por ejemplo, regiones genómicas

[1] El log-likelihood es el índice de la verosimilitud. La verosimilitud está en función de los parámetros de un modelo estadístico. Dado un conjunto de datos, es igual a la probabilidad condicional de los datos, dada una hipótesis.

río arriba de genes), es una tarea difícil debido a la tendencia de los sitios de unión a ser cortos y degenerados, y al hecho de que las regiones promotoras a menudo son difíciles de identificar con precisión. El problema tiende a ser más crítico en la identificación de TFBS en organismos eucariotas porque dichos sitios tienden a ser más cortos y más variables (Bailey *et al.*, 2006).

MEME encuentra uno o más motivos en una colección de secuencias de DNA o de proteínas no alineadas, usando la técnica de máxima expectancia. La maximización de la expectancia (EM) es un algoritmo que encuentra las mejores estimaciones para los parámetros del modelo cuando a un conjunto de datos le falta información o tiene variables latentes ocultas[1]. Si bien esta técnica se puede utilizar para determinar la función de máxima verosimilitud o el modelo de "mejor ajuste" para un conjunto de datos, EM lleva las cosas un paso más allá y trabaja en conjuntos de datos incompletos. Esto se logra insertando valores aleatorios para los puntos de datos faltantes y luego estimando un segundo conjunto de datos. El nuevo conjunto de datos se utiliza para refinar las conjeturas agregadas al primero, y el proceso se repite hasta que se cumplan los criterios de terminación del algoritmo. Si bien los algoritmos de Estimación de máxima verosimilitud (conocida también como EMV y, en ocasiones, MLE por sus siglas en inglés del maximum likelihood estimation) es un método habitual para ajustar un modelo y estimar sus parámetros. y Maximización de expectancia pueden determinar los parámetros de "mejor ajuste", ambos adoptan enfoques significativamente diferentes. MLE necesita conocer todos los parámetros primero para construir un modelo, que generalmente es más preciso que EM, pero no puede funcionar si falta información. EM puede adivinar los valores de los parámetros que faltan y ajustar el modelo según sea necesario con unos pocos pasos adicionales: Se genera un valor inicial para los parámetros del modelo y se le asigna una distribución de probabilidad, conocida como la distribución “Esperada". Los datos recientemente observados se introducen en el modelo. Usando ecuaciones diferenciales y probabilidad condicional, la distribución de

[1] En estadística, las **variables latentes** (o variables ocultas, en contraposición a las variables observables), son las variables que no se observan directamente, sino que son inferidas (a través de un modelo matemático) a partir de otras variables que se observan (medidos directamente). Una ventaja de utilizar variables latentes es que reduce la dimensionalidad de los datos.

probabilidad de la distribución esperada se ajusta para aumentar la probabilidad de "mejor ajuste". Estos pasos se repiten hasta que la distribución esperada no cambie de la distribución observada, hasta alcanzar la estabilidad (Do y Batzoglou, 2008). La esencia del algoritmo de Maximización de la expectancia es usar los datos observados disponibles del conjunto de datos para estimar los datos faltantes y luego usar esos datos para actualizar los valores de los parámetros. Mismo que se esquematiza en la **figura 5**. Este algoritmo es la base de muchos algoritmos de agrupación llamados no supervisados, en el campo del aprendizaje automático o *machine learning* (Saeys *et al.*, 2007; Sampaio *et al.*, 2020).

Para encontrar el mejor perfil de alineamiento intenta "adivinar" la posición del motivo en las secuencias de entrada de las regiones que lo forman. Dado un perfil M, el algoritmo MEME evalúa la probabilidad (*likelihood*, máxima verosimilitud) de que cada región de secuencia de una longitud m se ajuste al perfil con respecto al fondo de las secuencias, mientras que el resto de la secuencia debe ajustarse al fondo mejor que el perfil. De acuerdo con este principio, se calcula un valor de probabilidad, para cada posición de cada secuencia de entrada. Este es el paso E (Expectancia). Luego, el algoritmo construye un nuevo perfil de alineación al juntar todas las regiones de secuencia de longitud m, pero ponderando cada una con el valor de probabilidad en la posición correspondiente. Este es el paso M (Maximización)(Bailey, 2003).

El algoritmo utiliza la búsqueda greedy (o codicioso), es decir, en cada paso solo se guardan las mejores alineaciones parciales, con la esperanza de que eventualmente conduzcan a la óptima. Obviamente, cuanto más conservado esté el motivo, más probable es que el algoritmo lo encuentre. El algoritmo comienza construyendo un perfil diferente de cada k-mer[1] en las secuencias de entrada, utilizando un valor de frecuencia de 1/2 para los nucleótidos del oligo y 1/6 para los demás. Luego, para cada perfil (cada k-mer en la secuencia de entrada) realiza un solo paso de E y un solo paso M. El perfil de puntuación (probabilidad de frecuencias) de máxima verosimilitud más alto obtenido (después de la iteración única) se optimiza aún más con pasos EM adicionales, hasta que no se obtenga un aumento adicional (un valor más grande) en la puntuación. Finalmente, se reporta el perfil y el oligo (motivo reportado) se elimina de las secuencias de entrada. Luego, el algoritmo se reinicia, hasta que se hayan generado varios perfiles que se pueden especificar como entrada. Por lo tanto, MEME puede detectar múltiples motivos dentro del mismo conjunto de secuencias en una sola ejecución y representa estos motivos como matrices de probabilidad de letras dependientes de la posición que describen la probabilidad de cada letra posible en cada posición del patrón, también conocida como matriz de peso de posición específica (PSWM) o matriz de puntuación de posición específica (PSSM), es una representación de motivos (patrones) comúnmente utilizada en secuencias biológicas (Bailey, 1994).

[1] Distribución de DNA k-mers ('palabras' de DNA de longitud k a lo largo de una secuencia).

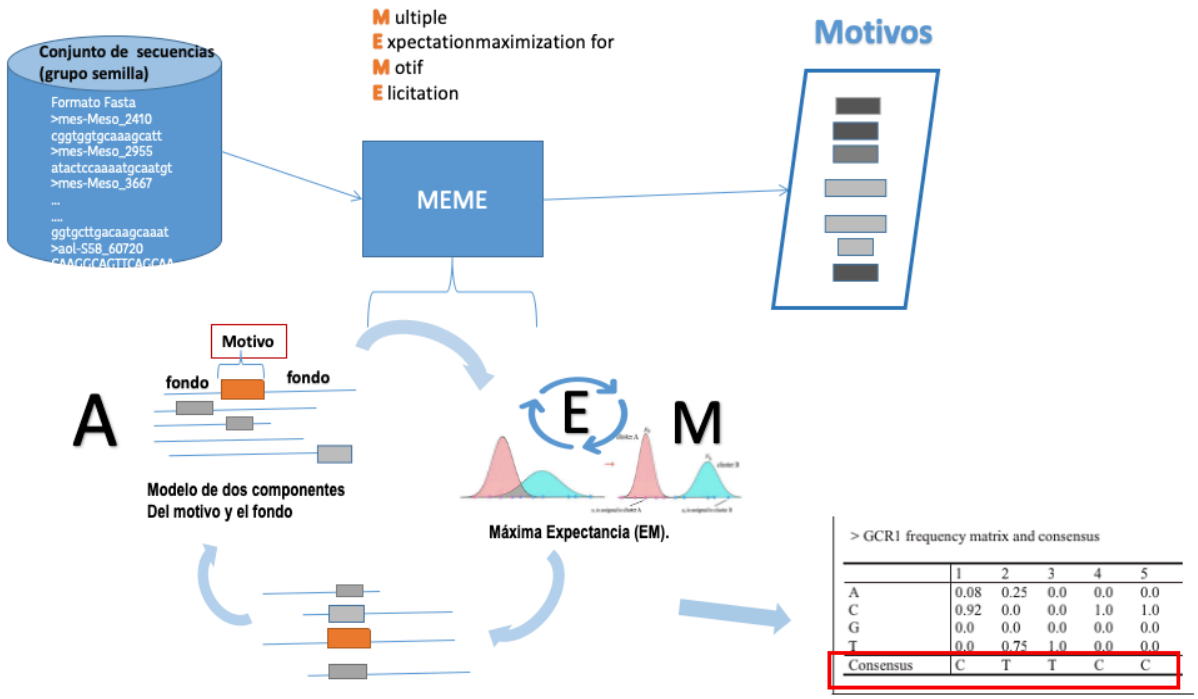


Figura 5. Diagrama de flujo para algoritmo EM y MEME. El usuario de MEME proporciona como entrada, un conjunto de secuencias de DNA, en formato Fasta, el cuál debe estar previamente "curado", es decir, que pertenezcan a datos de secuencias relacionadas (para no incluir ruido al análisis). MEME, internamente, trata de "adivinar" el motivo, el usuario debe proporcionar datos sobre el motivo a buscar (si se conocen o intuyen). MEME construye el modelo de dos componentes, que se esquematiza en el punto A, donde el motivo, es la señal detectada, del resto de la secuencia, que se denomina fondo, y que utiliza parámetros de su propio modelo. Con el algoritmo de EM, se busca encontrar la probabilidad de ocurrencia de ese patrón, maximizando la expectancia, guardando en cada iteración el mejor ajuste. Una vez que se estabilizan los parámetros, y se encontró la mejor representación del motivo, se reporta el consenso (recuadro rojo). Inicialmente, se considera un conjunto de valores iniciales de los parámetros. Se da al sistema un conjunto de datos observados incompletos con el supuesto de que los datos observados provienen de un modelo específico. El siguiente paso se conoce como "Expectancia" o paso E. En este paso, utiliza los datos observados para estimar o adivinar los valores de los datos faltantes o incompletos. Básicamente se usa para actualizar las variables. El siguiente paso se conoce como "Maximización" o paso M. En este paso, se usan los datos completos generados en el anterior "Expectancia" - paso E, para actualizar los valores de los parámetros. Se usa para actualizar la hipótesis. Ahora, en el cuarto paso, se verifica si los valores son convergentes o no, si es así, deténgase de lo contrario, repita los pasos 2 y 3, es decir, "Expectancia" y "Maximización" hasta que se produzca la convergencia. El ciclo se repite hasta que los parámetros se estabilizan, esto se hace para cada secuencia de los datos. El resultado final es una matriz de puntuación específica que representa los motivos sobre-representados del grupo de secuencia.

MEME y sus modelos de fondo.

La sensibilidad de la búsqueda de motivos TFBS se puede mejorar mediante el uso de un "modelo de secuencia de fondo de orden superior”(Bailey, 2003). MEME, hace uso de archivos de modelo de fondo tipo Markov. Utilizando cadenas de Markov, se define la probabilidad de encontrar una letra en una posición dada que depende de las letras que la preceden en la secuencia.

Estos modelos son necesarios para estimar el número de veces que se espera que una matriz “encuentre“ un sitio de unión o motivo, sólo por suerte y no uno biológicamente relevante. Un archivo de modelo de fondo especifica todas las frecuencias de k-mer hasta un máximo de k elegido por el usuario. MEME usa un modelo de fondo para normalizar la distribución sesgada de letras y grupos de letras en sus secuencias. Por ejemplo, en nuestro caso que buscamos sitios en secuencias reguladoras de genes y estas suelen tener diferente contenido rico en AT a diferencia de las regiones codificantes, se recomienda usar un modelo creado a partir de todas las secuencias reguladoras del organismo (para representar mejor la región donde hacemos la búsqueda).

Un modelo de orden 0 se ajusta para sesgos de letra única, un modelo de orden 1, nos indica la frecuencia con la que se presenta cada uno de los cuatro nucleótidos dependiendo de cuál es el nucleótido que se encuentra antes en la secuencia (frecuencia de G, cuando antes se encuentra una C) se ajusta para sesgos de dímero (por ejemplo, contenido de GC en secuencias de DNA), es decir, la frecuencia con la que se encuentra el di-nucleótido GC. MEME utiliza el modelo de fondo para cuatro propósitos: durante EM como el "modelo nulo”, para calcular la relación de probabilidad logarítmica de un motivo, para calcular la significancia (valor E) de un motivo y para crear la PSSM. Si se da la opción, MEME solo lee el modelo de Markov en un archivo hasta el orden dado,

ignorando las entradas de orden superior. En caso de no especificar uno el modelo nulo (Bailey, 1994).

El modelo de fondo aumenta su complejidad al usar cadenas de Markov de ordenes mayores a cero. Haciendo que el modelo sea más estricto y contendrá información más precisa de la composición nucleotídica del contexto donde se hará la búsqueda, pero esto resulta contraproducente ya que, al usar un orden muy alto, tiende al sobre-ajustamiento, esto provocaría que fuese difícil encontrar sitios relevantes (Bailey *et al.*, 2006).

Para generar archivo de background de Markov se utiliza el siguiente comando:

fasta-get-markov [opciones]

Donde las opciones son*: -dna -m [número]* archivo_dna archivo_salida que indica que se trata de secuencias de DNA y -m el orden de markov a utilizar.

Para MEME un comando básico sería:

meme [options] <primary sequence file>
Donde las opciones son *-dna* (indicando que se trata de DNA*), -oc [directorio de salida],* -bfile [archivo de fondo], *-mod [oops, zoops, o anr].*

Los resultados de MEME los devuelve como un documento HTML interactivo que brinda una descripción completa del contenido del motivo de las secuencias de entrada sea de DNA o RNA, en este caso, nosotros utilizamos secuencias de DNA. Todos los resultados se presentan en grupos ordenados según el significado estadístico. En este archivo HTML muestra los motivos como alineamientos locales múltiples de (subconjuntos de) las secuencias de entrada. Los "diagramas de bloques" muestran las posiciones relativas de los motivos en cada una de las secuencias de entrada. Los botones en la salida HTML MEME permiten que uno o todos los motivos se envíen para su análisis por otros programas depositados en sitios web. Al hacer clic en un botón, se pueden enviar todos los motivos al servidor web MAST (http://meme-suite.org/doc/mast.html), donde se pueden buscar en varias bases de datos de

secuencias (o secuencias cargadas), secuencias que coincidan con los motivos (Bailey *et al.*, 1998).

MEME-MAST.

Una vez que MEME ha generado el PSSM, se utiliza un segundo algoritmo, MAST, para encontrar las mejores coincidencias para ese PSSM en los datos de entrada. El valor p le indica qué tan bien un sitio en particular coincide con el PSSM encontrado por MEME. Cuanto menor es el valor p, más significativa es la coincidencia. Esto es útil, por ejemplo, para saber si el motivo de interés también está presente en otros genes o genomas. MAST es una herramienta que se puede utilizar para buscar secuencias que coincidan con uno o más motivos. Se puede usar para buscar secuencias que contienen motivos encontrados por MEME, por otras herramientas de descubrimiento de motivos o que se toman de una base de datos de motivos. El sitio web de MAST, al que se accede a través de la misma URL que el sitio web de MEME, proporciona numerosas bases de datos de nucleótidos y proteínas para la búsqueda. Las consultas MAST pueden contener cualquier número de motivos, y puntúa (da un valor o puntaje) cada secuencia en la base de datos seleccionada utilizando todos los motivos. MAST puede buscar secuencias de DNA para encontrar coincidencias con los supuestos motivos del sitio de unión del TFBS encontrados por MEME en un conjunto de secuencias promotoras (Bailey *et al.*, 1998).

En este proyecto se hizo uso de esta herramienta, para poder localizar motivos de secuencias promotoras en las regiones 5´ de genes bacterianos, usando como grupo “semilla” los genes ortólogos que codifican para el factor σ54. Ya que se ha observado que en conjunto MEME-MAST son herramientas que permiten descubrir motivos a partir de secuencias no alineadas, el análisis de diferentes tamaños de ventana (especialmente si no se conoce el tamaño del motivo), genera más de un motivo si éste tiene espacios (huecos), es óptimo para secuencias genómicas o proteicas, identifica el perfil de los motivos, analiza simultáneamente múltiples motivos y además es flexible en cuanto a la

presencia de un motivo, que puede estar ausente en algunas de las secuencias de entrada, o aparecer varias veces en una, en comparación con otros programas. Y a su vez, realizar la búsqueda de coincidencias con las matrices PSS generadas por MEME en genomas completos de diferentes organismos (Bailey *et al.*, 2006). Se han desarrollado, y siguen desarrollándose muchos métodos para la predicción de TFBS en regiones reguladoras de genes relacionados, sin embargo, el problema es extremadamente desafiante en todos los niveles: desde la construcción de soluciones candidatas, hasta la evaluación de las mejores soluciones(Kanhere y Bansal, 2005; Saeys *et al.*, 2007; Askary *et al.*, 2009; Novichkov *et al.*, 2010; de Avila e Silva *et al.*, 2011; Sallet *et al.*, 2013; Bharanikumar *et al.*, 2018; Coelho *et al.*, 2018; Amin *et al.*, 2019). A manera de introducción describimos, de forma breve, los fundamentos de algunos métodos ya implementados, dentro de estos, los métodos que utilizamos para desarrollar un protocolo computacional que permita identificar los promotores en genomas de bacterias con base en la conservación de secuencia nucleotídica, tomando como modelo de estudio promotores reconocidos por el factor σ54.

ANTECEDENTES

La identificación de promotores en genomas bacterianos mediante análisis *in silico* ha sido de interés desde hace varias décadas. Los principales métodos computacionales para identificar promotores pueden dividirse en:

Métodos basados en la conservación de secuencia representados mediante matrices de frecuencia relativa o de probabilidades.

El método más común para identificar los promotores σ70 *in silico* utiliza matrices de peso de posición (PWM) y depende de la conservación relativa del sitio de unión del factor σ. Para organismos modelo ya se han calculado las matrices PWM y en base a ellas se pueden identificar potenciales promotores (K. *et al.*, 2016). Sin embargo, debido a que los promotores regulados positivamente poseen promotores alejados del consenso y a los bajos requerimientos de transcripción de genes cuyos productos se requieren en bajas concentraciones, las predicciones de promotores basada en motivos son poco precisos. En organismos alejados de los organismos modelo, se desconocen las matrices PWM para muchos factores σ por lo que no es factible identificar promotores con este tipo de métodos (Mulligan *et al.*, 1984).

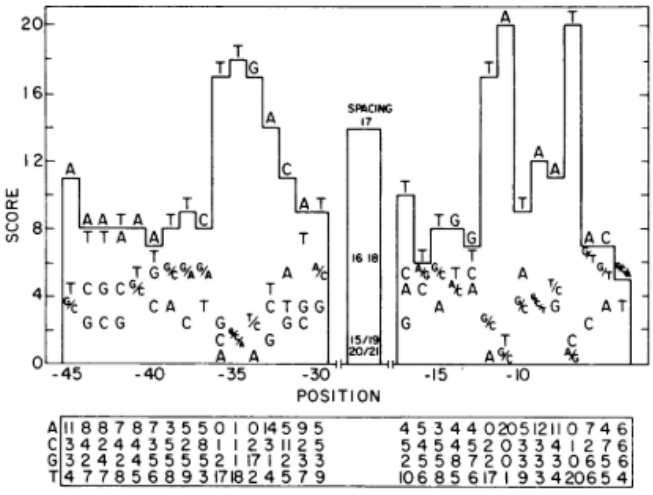


Figura 6. Posibilidad de ponderaciones estadísticas para evaluar las puntuaciones de similitud del promotor σ70 de *E. coli.* El histograma de la figura muestra las puntuaciones (tabuladas en el panel inferior) para cada base en cada una de las posiciones utilizado en la evaluación de la región -35, espaciador y región -10. Las puntuaciones individuales fueron calculadas a partir del alineamiento de secuencias de promotores σ70 reportados. figura tomada de Mulligan et al. 1984.

Métodos basados en la frecuencia relativa de lecturas obtenidas en estudios de RNA-seq.

Las tecnologías de secuenciación de nueva generación y en particular las del tipo RNA-seq, han hecho posible tener la capacidad de cuantificar la abundancia relativa de los transcritos de diferentes organismos, tanto bacterianos como organismos superiores. La frecuencia relativa del número de lecturas de cada base del genoma ha sido usada como medida para identificar sitios de inicio y término de la transcripción. Pese a lo anterior, esta manera de identificar dichos límites transcripcionales sigue siendo poco precisa debido a variaciones significativas en los valores de las lecturas por nucleótidos aún en regiones que deberían de ser similares (p. ej. Lecturas de nucleótidos de un mismo gen)(Osmundson *et al.*, 2013).

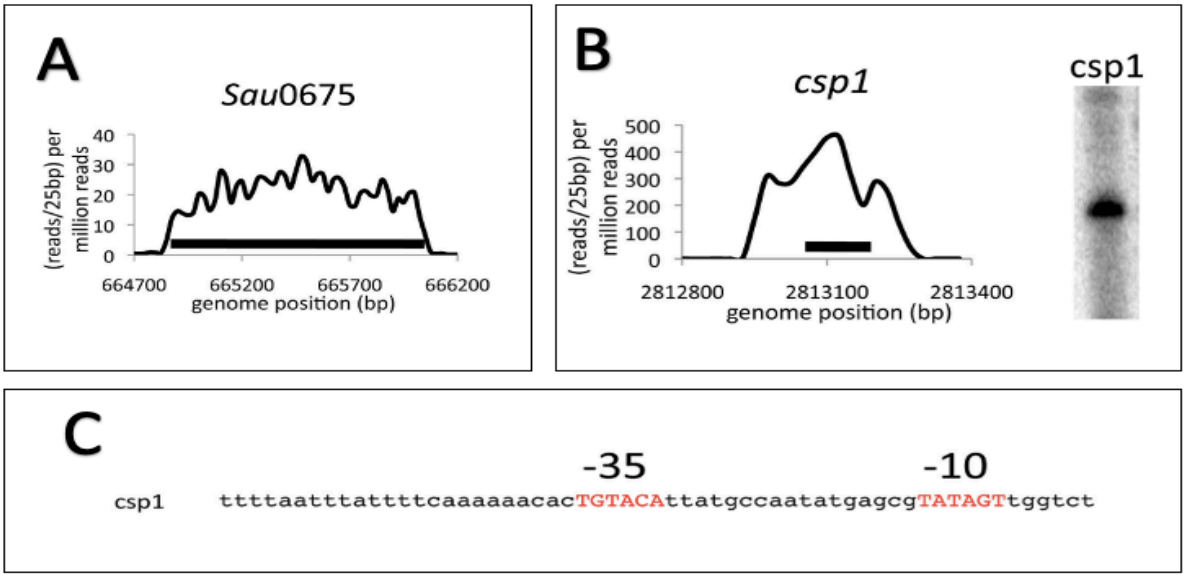


Figura 7 . Datos de RNA-seq utilizados para identificar secuencias promotoras en *S. aureus*. A) El mapeo de lecturas de RNA-seq inmediatamente río arriba del codón de inicio para un gen predicho, a partir de análisis de RNA-seq en *S. aureus*. Las lecturas se incrementan justo después del inicio de transcripción. Los datos son representados como lecturas sobre 25 pb por millón de lecturas totales y el eje x muestra la posición a lo largo del genoma de *S. aureus* 8325. La barra negra representa la secuencia de codificación del gen de *S. aureus*. B) Lecturas de RNA-seq del gen *csp1*, donde se observa evidencia de posible región promotora justo después del inicio de su transcripción. Resultado de ensayos de transcripción *in vitro* muestran la actividad de la RNA polimerasa del promotor mostrado en el panel C. C) La secuencia promotora inferida de los datos de RNA-seq se muestran en rojo.. Tomado y modificado de Osmudson et al. 2013.

Métodos basados en el cálculo de la estabilidad relativa del DNA.

El análisis de las propiedades estructurales de las regiones promotoras en genomas tanto de procariotas como de eucariotas indica que éstas tienen en común varias características, tales como: menor estabilidad, mayor curvatura, menor capacidad de flexión en comparación con sus regiones vecinas o localizada en las proximidades (Kanhere, 2005; Wang y Benham, 2007; Kumar *et al.*, 2016; Kumar y Bansal, 2017). Debido al requerimiento de la apertura de la doble cadena de DNA en el proceso del inicio de la transcripción, la región del DNA que contiene un promotor es, en términos generales, menos estable y, por lo tanto, más propensa a la fusión, en comparación con otras regiones genómicas. El cálculo de la estabilidad relativa de la molécula de DNA se puede expresar en términos de energía libre y ha sido empleado para predecir exitosamente los promotores de 913 organismos bacterianos. Cabe hacer notar que el contenido genómico de GC no fue un factor determinante para dicha predicción ya que el valor de dichos genomas variaba entre el 17% y el 75% (Rangannan y Bansal, 2010). El cambio de energía libre estándar ($\Delta G\,37^{o}$) correspondiente a la transición de fusión de un conjunto de nucleótidos de la secuencia de DNA doble hebra a hebra única, así como los valores de energía correspondientes a las 10 secuencias dinucleótidas únicas, se toman de los parámetros unificados obtenidos a partir de estudios de fusión en 108 oligonucleótidos reportados por *SantaLucía* (Allawi y SantaLucia, 1997). Basados en estas observaciones, desarrollaron un protocolo codificado en el programa: "*PromPredict*", para la identificación de promotores microbianos (Rangannan y Bansal, 2009).

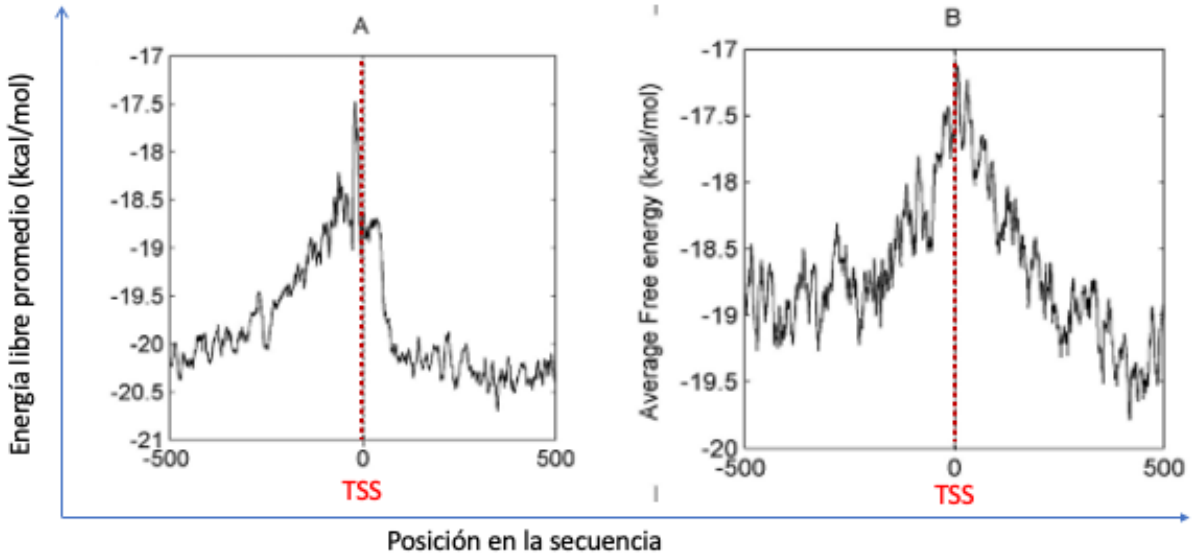


Figura 8 . Perfil de energía libre global alrededor de un TSS bacteriano. La figura muestra el promedio de perfiles de energía libre de **A**) *Escherichia. coli* (227 promotores) y **B**) *Bacillus subtilis* (89 promotores). La posición de la secuencia de nucleótidos se muestra en el eje x. Los valores más negativos de energía libre indican una mayor estabilidad. Los perfiles se extienden desde 500 nt río arriba hasta 500 nt río abajo del sitio de inicio de la transcripción (posicionado en 0, mostrado como línea discontinua en rojo). Figura modificada de Kanhere y Bansal 2005.

La mayoría de los algoritmos disponibles para la predicción de promotores bacterianos están enfocados en la identificación de secuencias promotoras dependientes de un factor σ específico, como lo es el σ70. Pese a lo anterior, el desarrollar un método de predicción óptimo es esencial para conocer más sobre la regulación de los genes bacterianos para los cuales aún no se sabe qué σ factor los regula.

JUSTIFICACIÓN

Se sabe que las secuencias promotoras desempeñan un papel central en la expresión génica por lo que la identificación de promotores transcripcionales resulta ser de gran importancia para diversos estudios teóricos o experimentales que requieran conocer las secuencias que son reconocidas por los factores σ de un organismo u organismos particulares. A pesar de que existen diferentes metodologías moleculares para identificar el inicio de transcripción, dichas metodologías suelen ser costosas y consumen mucho tiempo, por lo que los métodos *in silico* son una alternativa atractiva. No obstante, a que se ha definido la secuencia consenso de varios factores σ de organismos modelo y algunos de sus promotores correspondientes, el número de organismos para los que no existe dicha información ha crecido y continúa creciendo con gran rapidez debido a los avances en las tecnologías de secuenciación. Actualmente es un tema de investigación activa en biología molecular y un desafío en bioinformática. Debido a lo anterior, el presente proyecto pretende desarrollar un protocolo computacional que permita identificar los promotores en genomas de bacterias con base en la conservación de secuencia nucleotídica. Consideramos tomar como modelo inicial de estudio a los promotores reconocidos por el factor σ54 debido a que su alta conservación de secuencia podría facilitar definir las estrategias computacionales de estudio para la identificación global de promotores bacterianos.

OBJETIVOS

Objetivo general

Desarrollar un protocolo computacional que permita identificar los promotores en genomas de bacterias con base en la conservación de secuencia nucleotídica, tomando como modelo de estudio promotores reconocidos por el factor σ54.

Objetivos particulares

1. Desarrollar programas computacionales para identificar un conjunto de secuencias genómicas de organismos representativos de las diferentes especies bacterianas.
2. Agrupar las secuencias genómicas en término de las relaciones filogenéticas de sus correspondientes organismos.
3. Identificar los genes ortólogos que codifican al factor σ54 dentro del genoma de los organismos representativos de estudio.
4. Definir el grupo de secuencias de regulación inicial, denominado secuencias semilla, correspondientes a las unidades transcripcionales en donde el gen *rpoN* es el primero de operón dentro de la clase Gammaproteobacteria.
5. Evaluar el desempeño del programa MEME al variar sus distintos parámetros de análisis y valores de corte que permitan la obtención de motivos de secuencias, expresados en términos de matrices de frecuencias de nucléotidos, que permitan identificar promotores σ54 en el conjunto de secuencias genómicas de organismos de la clase Gammaproteobacteria.
6. Aplicación del protocolo considerando el conjunto de secuencias genómicas, en donde se obtendrá una matriz MEME para cada clase filogenética contemplada de estudio.
7. Búsqueda de promotores σ54 en el conjunto de regiones de regulación 5´ acotadas según los parámetros definidos en pasos anteriores.
8. Análisis de las implicaciones biológicas de los genes identificados por estar regulados por el factor σ54.

HIPÓTESIS

Los genes que codifican para un factor σ, tienden a dirigir la transcripción específica de los genes que los codifican; es decir, se autotranscriben. En grupos filogenéticamente cercanos pertenecientes a la misma clase, existe una alta conservación en la secuencia de los miembros de una misma familia de factores σ. Esta conservación de los factores σ, implica la conservación de las secuencias nucleotídicas que reconocen.

METODOLOGÍA

Para el desarrollo del presente trabajo se utiliza la base de datos KO (grupo de ortología KEGG) de funciones moleculares representadas en términos de ortólogos funcionales, KEGG (*Kyoto Encyclopedia of Genes and Genomes*), que contiene genes-proteínas ortólogos caracterizados experimentalmente. Un ortólogo funcional se define manualmente en el contexto de redes moleculares de KEGG (mapas de rutas KEGG, jerarquías BRITE y módulos de KEGG). En este caso, el identificador para el gen de *rpoN*; RNA polymerase σ54 factor, es el KO: K03092. También se hace uso de la base de datos pública de grupos de proteínas ortólogas COG de NCBI (por sus siglas en inglés, Clusters of Orthologous Groups of Genes)(Tatusov, 2000). Esta base de datos agrupa a las proteínas en base a la similitud de sus secuencias de aminoácidos bajo la suposición de que dicha similitud, por arriba de un cierto valor de corte, pudiera implicar que dichas proteínas están relacionadas filogenéticamente y son homólogas. Se denomina homólogo al gen o proteína que descienden de un ancestro común y que han evolucionado por eventos de especiación, y, por tanto, se encuentran en organismos distintos**.** En algunos casos las proteínas homólogas conservan su función ancestral. Existen al menos tres subtipos de homología : ortología, paralogía y xenología. Los genes ortólogos son aquellos comparten el último ancestro común y cuya divergencia se debe a un proceso de especiación, es decir el mismo gen en distintas especies. Los parálogos son genes que, debido a una duplicación dentro de un mismo genoma, ya no comparten el último ancestro. Usualmente, las copias del gen son funcionalmente distintas (Fitch, 1970). La ortología y la paralogía son conceptos clave de la genómica evolutiva. Una distinción clara entre ortólogos y parálogos es crítica para la construcción de una clasificación evolutiva robusta de genes y una anotación funcional confiable de genomas recién secuenciados. Los genes que clasifica la Xenología, son los genes derivados de la transferencia horizontal entre organismos (Koonin, 2005).

1. Identificación de organismos de estudio no-redundantes.

En este proyecto consideramos a los organismos pertenecientes a los dominios de bacterias cuyos genomas han sido secuenciados en su totalidad y se encuentran depositados en la base de datos KEGG. Hasta el momento del desarrollo de este trabajo, año 2019, existen en la base de datos KEGG la secuencia nucleotídica de 5,158 organismos, siendo 4,852 los que corresponden a organismos bacterianos. Con el propósito de evitar sesgos en los análisis, introducidos por la redundancia de cepas secuenciadas, solo se considera un organismo por especie; por ejemplo, una sola cepa de *Escherichia coli*, de las más de sesenta cepas de *E. coli* secuenciadas y depositadas en la base de datos KEGG.

2. Agrupamiento de organismos de acuerdo con su filogenia.

Se desarrollaron algoritmos Ad hoc escritos en lenguaje de programación Perl (se encuentran en https://github.com/Haly-en/Programas_maestria_metodologia) para agrupar a los organismos representativos por especie, agrupados de acuerdo a su asignación taxonómica según los niveles de clase. Se tomaron únicamente aquellas agrupaciones de clases que contenían más de 4 organismos representativos a nivel de especie, para poder trabajar con mayor información y lograr comparaciones estadísticas.

3. Identificación inicial de genes que codifican para el factor σ54.

Para poder estudiar un grupo de secuencias reguladoras ortólogas, es necesario primero obtener el grupo de genes ortólogos y después obtener la secuencia reguladora correspondiente a cada gen. La identificación de genes ortólogos de las distintas familias de genes que codifican para el factor σ54 se realizó con base a la descripción funcional de los grupos de ortología KO (*Kegg orthologous*) de la base de datos KEGG. Mismos que para tener mayor certeza de su agrupamiento por ortología y tener un agrupamiento sin secuencias parálogas, se utilizó el método de análisis Bidirectional-best-hits (BBH)

(Overbeek *et al.*, 1999). Este método de identificación de ortólogos consiste en comparar las secuencias de genes o proteínas de dos organismos de manera bidireccional, e identificar aquellos pares de secuencias que hayan sido las más parecidas de forma recíproca. En este trabajo, las secuencias de proteínas fueron consideradas para realizar las comparaciones utilizando el programa BLAST (Altschul *et al.*, 1990). Debido a que en algunos organismos existen más de una copia parálogas del gen que codifica para el factor σ54, las proteínas ortólogas de organismos pertenecientes al mismo clado, fueron agrupadas utilizando el inverso del valor de Bitscore de BLAST como medida de distancia entre un par de secuencias. El método empleado considera la frecuencia de vecinos cercanos como criterio de agrupamiento, de tal forma que el primer grupo está conformado por la proteína que tuviera mayor número de vecinos obtenidos en las comparaciones de BLAST, y todos sus correspondientes vecinos. Las proteínas restantes no incluidas en este primer grupo son nuevamente agrupadas de acuerdo con el mismo criterio. Los ciclos de agrupamiento se repiten hasta que todas las proteínas hayan sido agrupadas en algún grupo de proteínas parálogas. Luego, se debe ubicar la secuencia directamente arriba del primer gen del operón; esto es porque en bacterias para obtener la región reguladora de un gen, debe tomarse en cuenta que los genes suelen encontrarse en configuración de operón y, por lo tanto, la región anterior al TSS de un gen no siempre podría contener la región reguladora.

4. Identificación de genes en operones y su posición relativa dentro de la unidad transcripcional correspondiente.

En los genomas bacterianos, la unidad transcripcional[1] (TU, del inglés transcription unit) básica es el operón. Un operón es un grupo de genes contiguos en el genoma y se encuentran en la misma cadena del DNA que son co-transcritos y comparten una misma región reguladora (Jacob y Monod, 1961). Para la predicción de operones se utilizó una

[1] Una unidad de transcripción es un conjunto de uno o más genes transcritos a partir de un único promotor. Una TU también puede incluir sitios de unión a proteínas reguladoras que afectan a este promotor y un terminador. Un operón complejo con varios promotores contiene, por lo tanto, varias unidades de transcripción. Según la definición de operón, al menos una unidad de transcripción debe incluir todos los genes en el operón.

metodología computacional desarrollado en nuestro grupo de Investigación (Taboada et al., 2018). Dicha metodología utiliza una Red Neuronal Artificial que tiene dos variables de entrada: a) Las distancias intergénicas entre pares de genes y b) La relación funcional entre genes contiguos, tal y como se define en la base de datos STRING (Jensen et al., 2009). La precisión de las predicciones de operones alcanzadas por este método para los organismos modelo de *E. coli* y *B. subtilis* fueron las más altas, jamás antes alcanzadas (94.6 y 93.3%, respectivamente). Por esto, mediante el uso de programas escritos en Perl, verificamos la posición de los genes que codifican para los factores σ54 dentro de sus respectivos operones, con particular atención en aquellos que se encontraran en la primera posición de un operón (o cabeza de operón).

5. Obtención de las regiones intergénicas 5´ de los operones en donde el gen que codifica para el factor σ54 es cabeza.

De los genes definidos en los pasos anteriores (aquellos genes que codifican al factor σ54 y que se encuentran en la primera posición de un operón) se tomaron sus correspondientes regiones intergénicas 5´ que se obtuvieron de acuerdo con la información de posiciones de genes de archivos tipo GFF (*General Feature Format*) (Wang *et al.*, 2016) y programas *ad hoc* realizados en lenguaje de programación *Perl.*

6. Identificación de motivos conservados (promotores) y su representación mediante matrices MEME.

La identificación de motivos de secuencia conservados fue realizada mediante el programa MEME (*Multiple expectation maximization for Motif Elicitation*) (Bailey *et al.*, 2006). Se utilizaron diferentes tamaños de ventana de análisis para incluir las dos cajas de los promotores bacterianos del factor σ54 y se realizaron pruebas variando el valor de los parámetros y de las opciones que se enlistan en la **Tabla** 1. Este proceso se realizó primero en el conjunto de prueba, es decir, secuencias intergénicas de genes que codifican al factor σ54 y que se encuentran en la primera posición de un operón, de la clase Gammaproteobacteria.

Tabla 1. Variables de prueba en MEME. Se enlistan las variables que se analizaron para llegar a parámetros estándar.

Variables	Opciones para analizar
-mod	*oops* | *zoops* | *anr*
-maxw	no definida l definida (17 nt)
-nmotifs	dos (6 nt) | uno
background de Markov	k | 0 | 1 | 2 | 3

7. Búsqueda genómica mediante MAST de los motivos conservados (promotores) en regiones intergénicas usando matrices MEME.

Para identificar nuevas secuencias genómicas con regulación potencial por el factor σ54 en el conjunto de secuencias intergénicas 5´ de los genomas de organismos de estudio, se utilizaron las matrices MEME obtenidas en el paso anterior (**paso 6**) para la búsqueda de patrones o coincidencia de patrones mediante el uso del programa de búsqueda MAST (Bailey *et al.*, 1998). En el primer conjunto de prueba se realizaron ensayos con diferentes valores de corte para definir el más adecuado para el protocolo. Una vez inferidos, se procedió a realizar la búsqueda con todos los conjuntos de estudio utilizando los valores de corte en MAST: 0.25,0.5, 1 y 2, en cada ciclo descrito en el paso siguiente de nuestra metodología.

8. Realización del proceso cíclico de identificación de motivos conservados (promotores) y búsquedas genómicas de secuencias con los motivos definidos.

El proceso cíclico denominado *pattern-finding/pattern-matching o búsqueda de patrones/coincidencia de patrones*, fue efectuado durante cinco ciclos, en donde se partía de las secuencias de regulación de un ciclo previo, para definir, utilizando el

programa MEME, los motivos conservados de secuencia que representaban a los promotores σ54. Dichos motivos fueron expresados como matrices de frecuencias con los que fueron buscadas, usando el programa MAST, nuevas secuencias dentro del conjunto de regiones intergénicas de los genomas de estudio. En cada ciclo, fueron definidas matrices MEME más representativas y con valores estadísticos de más significativos.

RESULTADOS

Organismos representativos agrupados por clase e identificación de sus genes ortólogos descritos y clasificados como genes que codifican para el factor σ54.

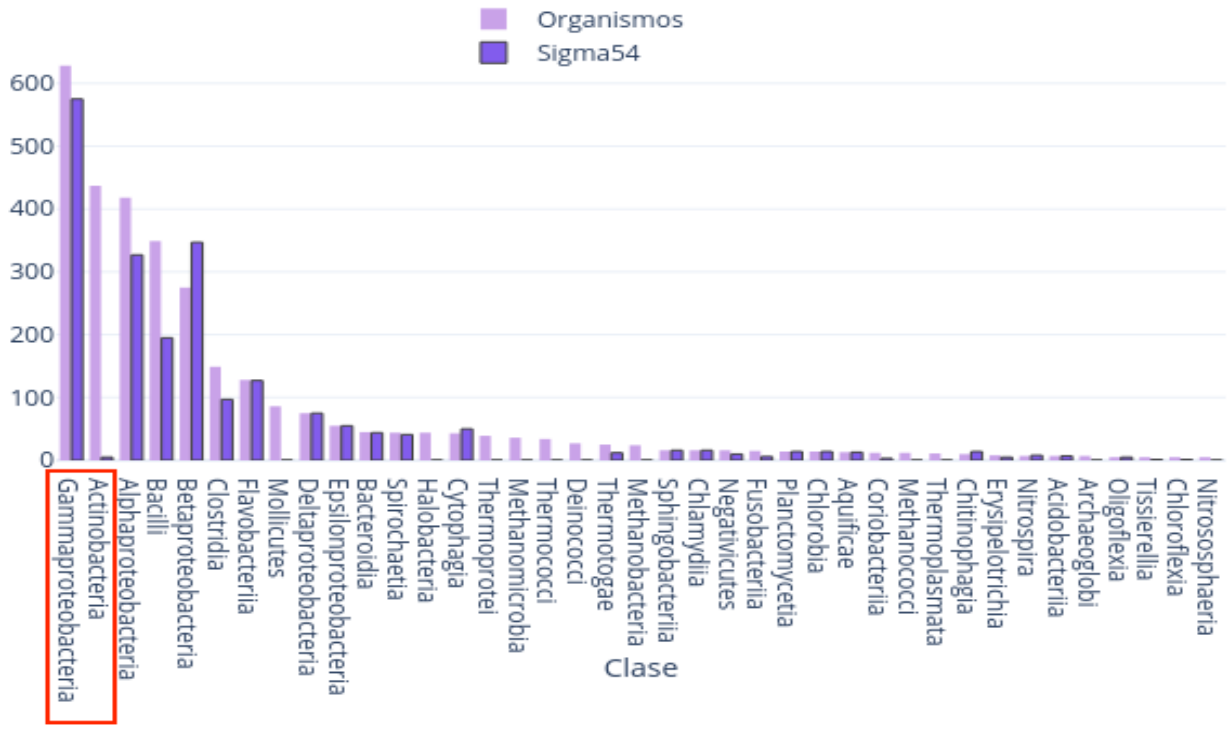
El primer paso de nuestra metodología se llevó acabo utilizando una serie de algoritmos que se encuentran descritos secuencialmente y alojados en el repositorio[1]: https://github.com/Haly-en/Programas_maestria_metodologia de GitHub[2], escritos en el lenguaje de programación Perl, para la identificación de organismos de estudio no-redundantes. Mediante este programa, se logró agrupar a organismos representativos por especie, agrupados de acuerdo a su asignación taxonómica según los niveles de clase. Sólo se consideraron aquellas clases que contaron con más de 4 organismos representativos a nivel de especie. Con esta consideración del número de organismos mínimos por grupo se terminó con un total de 39 clases de estudio, de las 76 clases existentes en la base de datos, que cumplen con estos criterios **(Tabla 2),** es decir, el 51.3 %**.**

Posteriormente, para cada conjunto de organismos por clase, se agruparon sus diferentes genes ortólogos descritos como “RNA polymerase sigma-54 factor” con base a grupo de ortología, utilizando su identificador KO (*Kegg orthologous*) que es K03092, con el programa: Identificación_de_sigma_factors.pl depositado en el repositorio mencionado con anterioridad. Los resultados se muestran en la **gráfica 1**, donde se

[1] Un repositorio es un directorio donde se almacenan los archivos de un proyecto. En GitHub, puede estar ubicado en el almacenamiento de GitHub o en un repositorio local en tu computadora.

[2] GitHub es un sistema de gestión de proyectos y control de versiones de código, así como una plataforma de red social diseñada para desarrolladores. Permite trabajar en colaboración con otras personas de todo el mundo, planificar proyectos y realizar un seguimiento del trabajo**.**

puede observar la distribución de genes que codifican para el factor σ54 por clase, así como el número de organismos. En la **gráfica 2** se muestran los porcentajes de los organismos (de todas las clases de estudio) que cuentan con 0, 1, 2, 3 o 4 genes que codifican al factor σ54. En la **gráfica 3** se muestran los porcentajes de los organismos que cuentan con el gen *rpoN (62.53 %)*, así como el porcentaje de aquellos que, dentro de una unidad transcripcional, son primer gen o se encuentran internamente (62 % y 38 %, respectivamente). Además, con este programa se tomaron las secuencias de regiones de regulación 5´, para obtener el grupo de datos "semilla" de estudio.

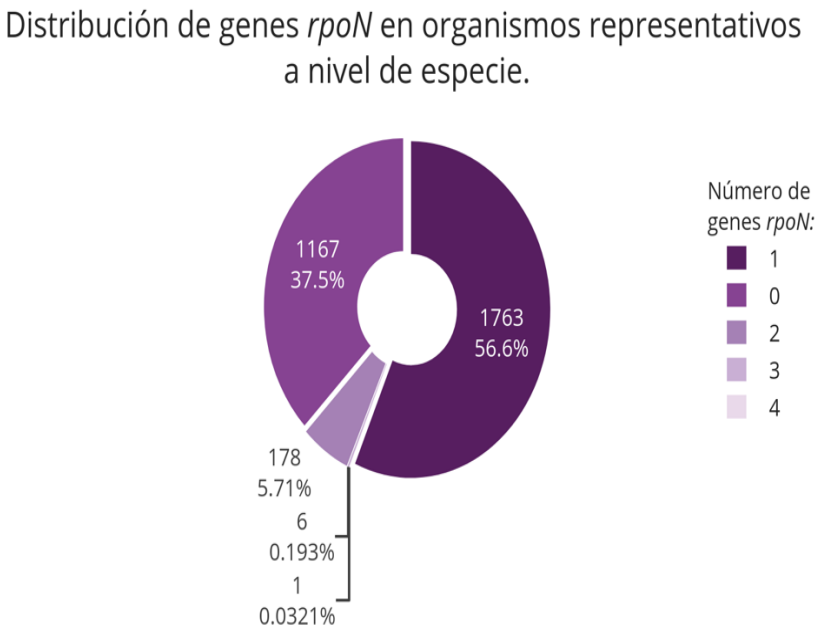


Gráfica 1. Número de organismos representativos a nivel de clase que cuentan con genes que codifican al factor σ54. Se grafican el número de organismos representativos (barra púrpura tenue) y el número de genes que corresponden al factor σ54 (barra púrpura obscura). Se resalta en el rectángulo rojo, al grupo de las Gammaproteobaterias, que es el grupo de estudio inicial y que consta de 628 organismos representativos. En términos generales, cada organismo cuenta con un sólo gen del factor σ54. También se resalta a la clase Actinobacteria, un grupo bastante singular, ya que muy pocos, 5 organismos, en más de 400 secuenciados, cuentan con el gen que codifica para el factor σ54.

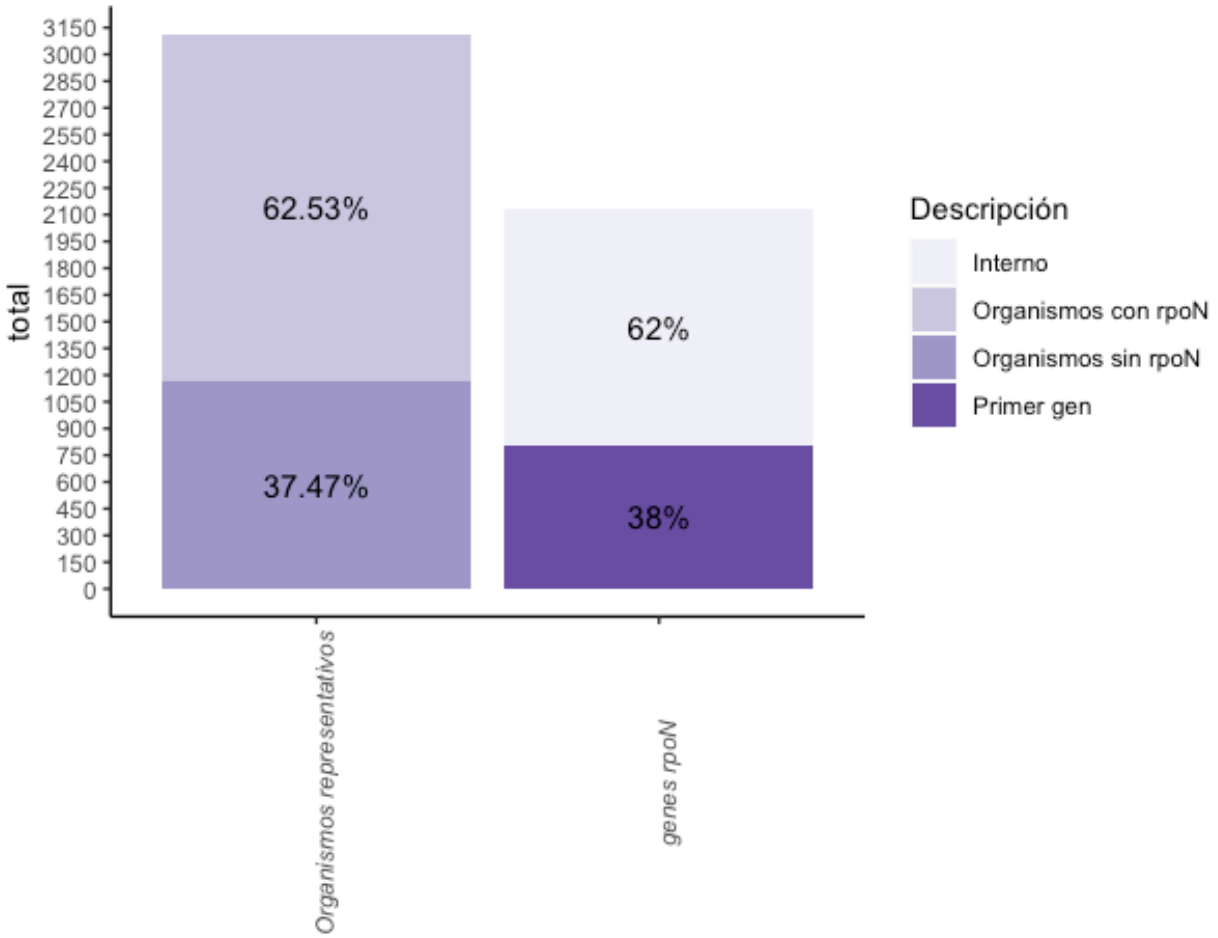
Tabla 2. Clases de estudio a nivel de especie. Se enlistan los datos para cada una de las clases que cuentan con el mínimo de organismos de estudio, así como el número de factores σ54, en rojo se resalta el grupo de prueba inicial. El porcentaje de genes que codifican para el factor σ RpoN por clase también se agrega en la quinta columna.

N	Clase	OrganismosRepresentativos	RpoN	Porcentaje
1	Gammaproteobacteria	628	575	91.6
2	Actinobacteria	437	5	1.1
3	Alphaproteobacteria	418	327	78.2
4	Bacilli	349	195	55.9
5	Betaproteobacteria	275	347	126.2
6	Clostridia	149	97	65.1
7	Flavobacteriia	128	127	99.2
8	Mollicutes	86	0	0
9	Deltaproteobacteria	75	74	98.7
10	Epsilonproteobacteria	55	55	100
11	Bacteroidia	45	44	97.8
12	Spirochaetia	44	41	93.2
13	Halobacteria	44	0	0
14	Cytophagia	43	50	116.3
15	Thermoprotei	39	0	0
16	Methanomicrobia	36	0	0
17	Thermococci	34	0	0
18	Deinococci	27	0	0
19	Thermotogae	25	12	48
20	Methanobacteria	24	0	0
21	Sphingobacteriia	16	16	100
22	Chlamydiia	16	16	100
23	Negativicutes	16	10	62.5
24	Fusobacteriia	15	6	40
25	Planctomycetia	14	14	100
26	Chlorobia	14	14	100
27	Aquificae	13	13	100
28	Coriobacteriia	12	3	25
29	Methanococci	12	0	0

30	**Thermoplasmata**	11	0	0
31	**Chitinophagia**	10	14	140
32	**Erysipelotrichia**	8	5	62.5
33	**Nitrospira**	7	8	114.3
34	**Acidobacteriia**	7	7	100
35	**Archaeoglobi**	7	0	0
36	**Oligoflexia**	5	5	100
37	**Tissierellia**	5	1	20
38	**Chloroflexia**	5	1	20
39	**Nitrososphaeria**	5	0	0



Gráfica 2. Distribución de genes *rpoN* presentes en organismos representativos a nivel de especie. La gráfica representa el porcentaje de organismos que cuentan con 0-4 genes que codifican para el factor σ54. Código de colores indicado en el lado derecho. El organismo que tiene 4 genes es *Rhodobacter sphaeroides* es una bacteria fotosintética no azufrada púrpura y se considera una posible fuente de producción de H_2.



Gráfica 3. Distribución del gen *rpoN* dentro de los organismos de estudio y su posición relativa dentro de la unidad transcripcional . La gráfica representa el porcentaje de organismos que cuentan genes que codifican para el factor σ54, alguno de ellos, como se muestra en la gráfica 2, cuentan con más de una copia; el porcentaje de organismos que no contiene el gen, es el 37.47 % (de los organismos analizados en el presente trabajo). El código de colores indicado en el lado derecho. Así mismo en la segunda barra, se muestra el porcentaje de genes *rpoN*, de los diferentes organismos de estudio, que se encuentran tanto en la primera posición dentro de un operón (38 %), como los que están internamente dentro de una unidad transcripcional (62 %).

Búsqueda de parámetros para la obtención de matrices MEME que representan potenciales secuencias promotoras.

De los datos generados en los pasos anteriores, tomamos como conjunto de prueba los datos de la clase Gammaproteobacterias, por ser el grupo con mayor número de organismos representativos a nivel de especie, 628, y de genes que codifican para el factor σ54 (RpoN) reportados en la base de datos de KEGG.

Adicionalmente, en este conjunto se agrupan varios miembros de organismos modelo, de los cuales podemos tener información reportada, como es el caso de *Escherichia coli* de la base de datos de RegulónDB (http://regulondb.ccg.unam.mx/). (Gama-Castro *et al.*, 2011). Esto resulta relevante ya que proporciona información de alta calidad para entrenar y evaluar los resultados de las predicciones derivados del método aquí desarrollado. Se obtuvieron las regiones 5´ de regulación y trabajamos con aquellas secuencias mayores a 40 pb, para tener suficiente longitud de secuencia para realizar la búsqueda de motivos, además estos genes deben encontrarse en la primera posición dentro de una unidad transcripcional, en la **gráfica 3**, se grafica el número de genes totales de estudio y su posición relativa en su unidad transcripcional correspondiente. Trabajamos con un total de 54 secuencias, que cumplían estos requerimientos, del grupo de prueba de la clase Gammaproteobacteria. Para visualizar su contexto genómico se utilizó la herramienta Gene Context tool, desarrollada en nuestro grupo de investigación (Martinez-Guerrero *et al.*, 2008), mismo que se esquematiza en la **figura 9**. Con el grupo de prueba se siguió la metodología planteada anteriormente, particularmente centrándonos en el paso número 6.6, que trata de la identificación de motivos conservados (promotores) y su representación mediante matrices MEME, para obtener los mejores parámetros para trabajar con este programa (MEME) posteriormente, con todas las clases de estudio y tener un criterio más amplio de qué esperar una vez implementado el protocolo.

Los resultados de pruebas derivadas del uso de MEME y el conjunto de datos de Gammaproteobacterias, tras modificar los tamaños de ventana, el modo (*-oops*, *-zoops*, *-anr*, descritos en la introducción), el número de motivos y los diferentes modelos de Markov , se muestran en **las figuras del 10-12**, donde se incluyen únicamente los ejemplos representativos de las pruebas. Las diferentes pruebas en donde se variaron los valores de los parámetros y valores de corte del programa MEME nos permitió llegar a la conclusión de que los motivos de secuencias estadísticamente más significativos dentro del conjunto de secuencias de las regiones de regulación de los genes que codifican para el factor σ54, no siempre corresponden al de las secuencias de los promotores σ54.

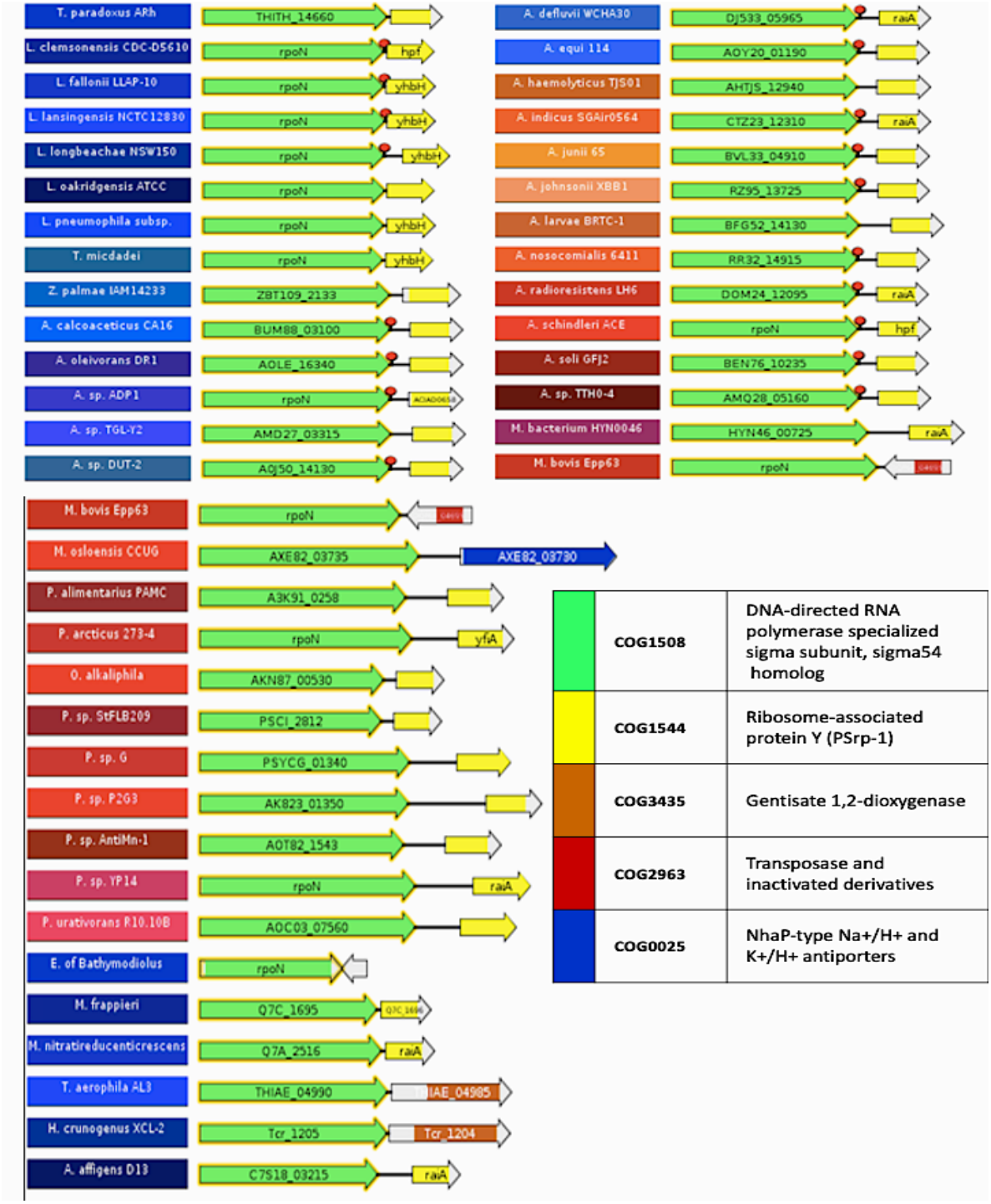


Figura 9. Contexto genómico de un conjunto semilla de Gammaproteobacteria (44 secuencias). En verde se muestra al gen que utilizamos para la búsqueda, con un gen vecino a la derecha. La caja de colores describe cada una de las funciones descritas por su grupo de ortología **COG, el COG1508** agrupa a RpoN.

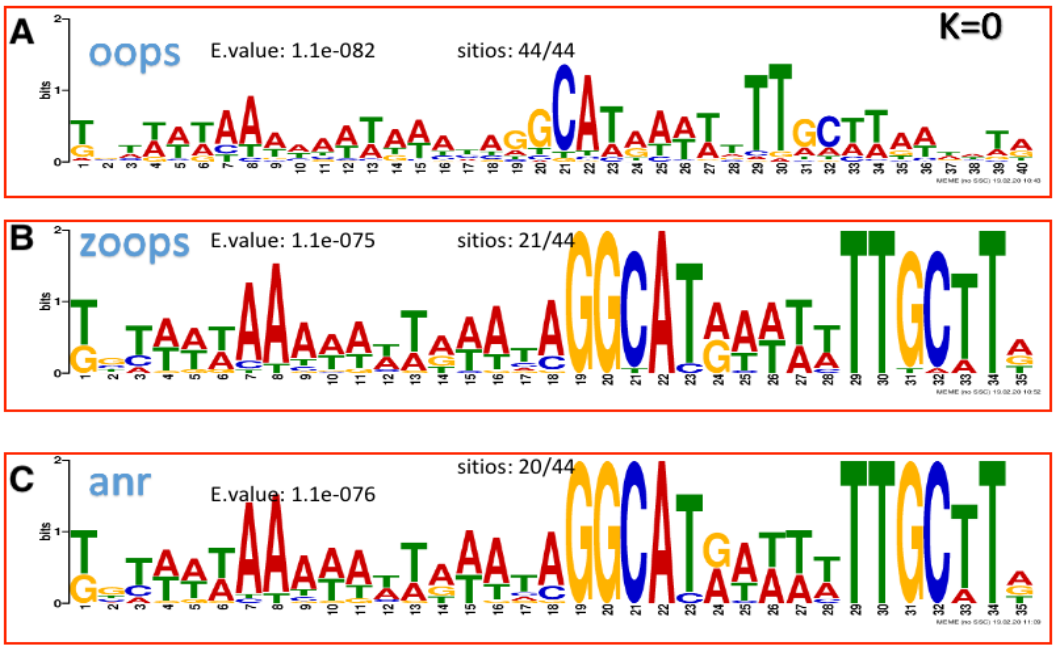


Figura 10. Logo de los resultados de MEME, con modelo de fondo 0 y amplitud de motivo no definido. Se muestran los logos de los resultados de MEME, utilizando el modo A) una ocurrencia por secuencia, oops B) cero o más ocurrencias por secuencias, zoops c) cualquier número de repeticiones en la secuencia, anr. Para todos los casos se utilizó como modelo de fondo k-mer, 0. De las 57 secuencias originales, para evitar duplicados (sobre-representación de secuencias), se hizo uso el programa CDHIT con un 0.70 % de identidad. Los sitios se indican 44/44, siendo 44 el número de secuencias después de aplicar CDHIT. Así mismo se anota el valor E, para cada uno de los logos. Un valor E más pequeño representa mayor significancia estadística.

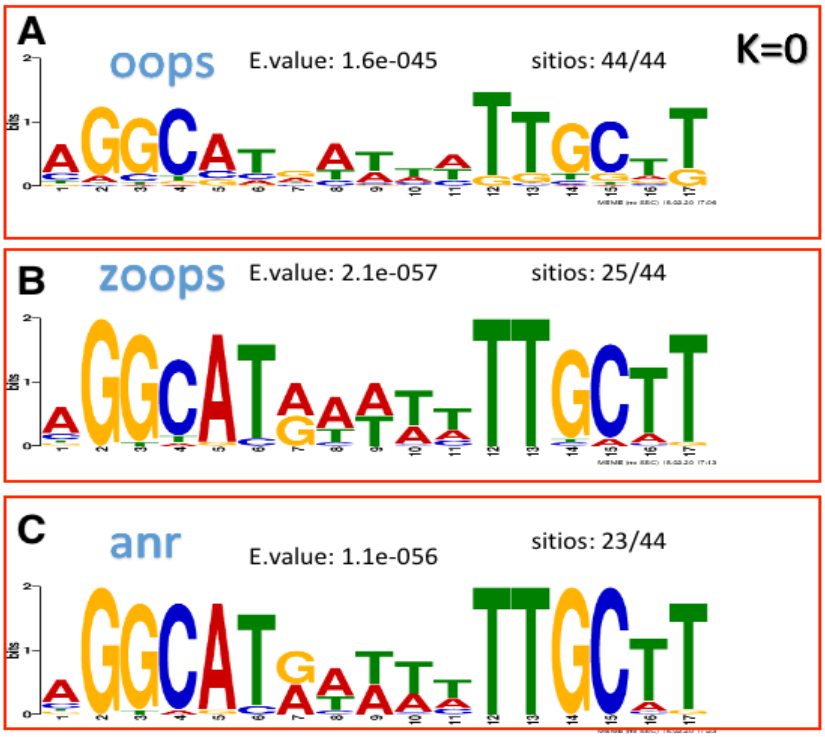


Figura 11. Logo de los resultados de MEME, con modelo de fondo 0 y amplitud de motivo definido (17 pb). Se muestran los logos de los resultados de MEME, utilizando el modo A) una ocurrencia por secuencia, oops B) cero o más ocurrencias por secuencias, zoops c) cualquier número de repeticiones en la secuencia, anr. Todos los casos utilizando como modelo de fondo k-mer, 0. Y amplitud definida a 17 pares de base.

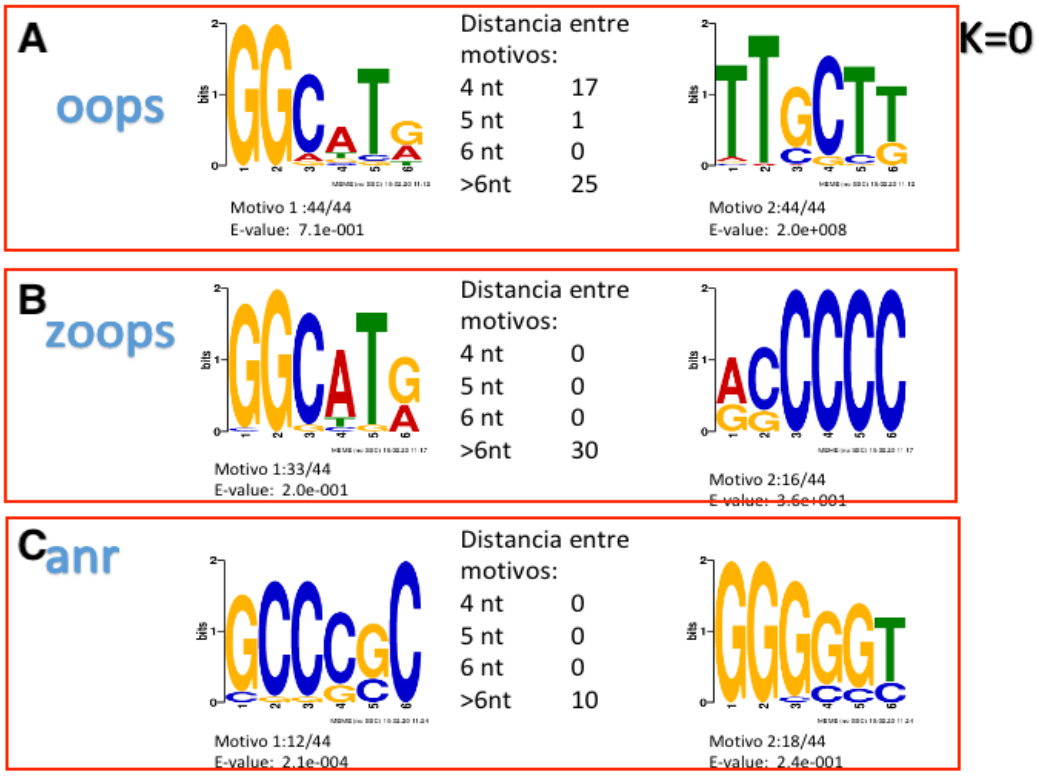


Figura 12. Logo de los resultados de MEME, con modelo de fondo 0, amplitud de motivo definido (6 pb) y búsqueda de dos motivos. Se muestran los logos de los resultados de MEME, utilizando el modo A) una ocurrencia por secuencia, oops B) cero o más ocurrencias por secuencias, zoops **c**) cualquier número de repeticiones en la secuencia, anr. Todos los casos utilizando como modelo de fondo k-mer, 0. Y amplitud definida a 17 pares de base. En el centro de los motivos, se anota la distancia entre motivos, así como el número de motivos a 4-6 nucleótidos.

Basándonos en las observaciones de los resultados obtenidos con las diferentes opciones de ejecución de MEME, notamos que en general cuando se busca un motivo de longitud definida para la señal de un promotor típico de σ54, con amplitud definida de 17 pares de bases, se obtiene motivo con un valor E altamente significativo, además de que las probabilidades de secuencia son similares a la secuencia consenso reportada para promotores de reconocimiento de σ54. También probamos realizar la búsqueda de dos motivos en cada secuencia, como se observa en la **figura 12**, los valores E buscando una ocurrencia por secuencia (*oops*) representa un consenso muy parecido al promotor σ54. Sin embargo, probando el método de cero o una ocurrencia por secuencia (*zoops*), que en este caso esperamos que las secuencias contengan ninguna o una "señal" estadísticamente significativa del motivo por secuencia, pero esto no ocurrió ni con el

uso de modelo de fondo de Markov de orden más elevado. Considerando lo anterior, decidimos tomar como referencia al motivo con el valor E estadísticamente más significativo, que suele ser el primer motivo que reporta MEME. Analizamos el uso de modelo de fondo de Markov del orden 1, en modo oops y *zoops* una vez anclados al motivo número 1. En la **figura 13,** se observa que el tomar como referencia al motivo primero resulta en una mejora sustantiva a la hora de buscar uno o dos motivos. Lo anterior es aún más notable cuando se utiliza el modo de búsqueda de *zoops*, panel B de **la figura 13.**

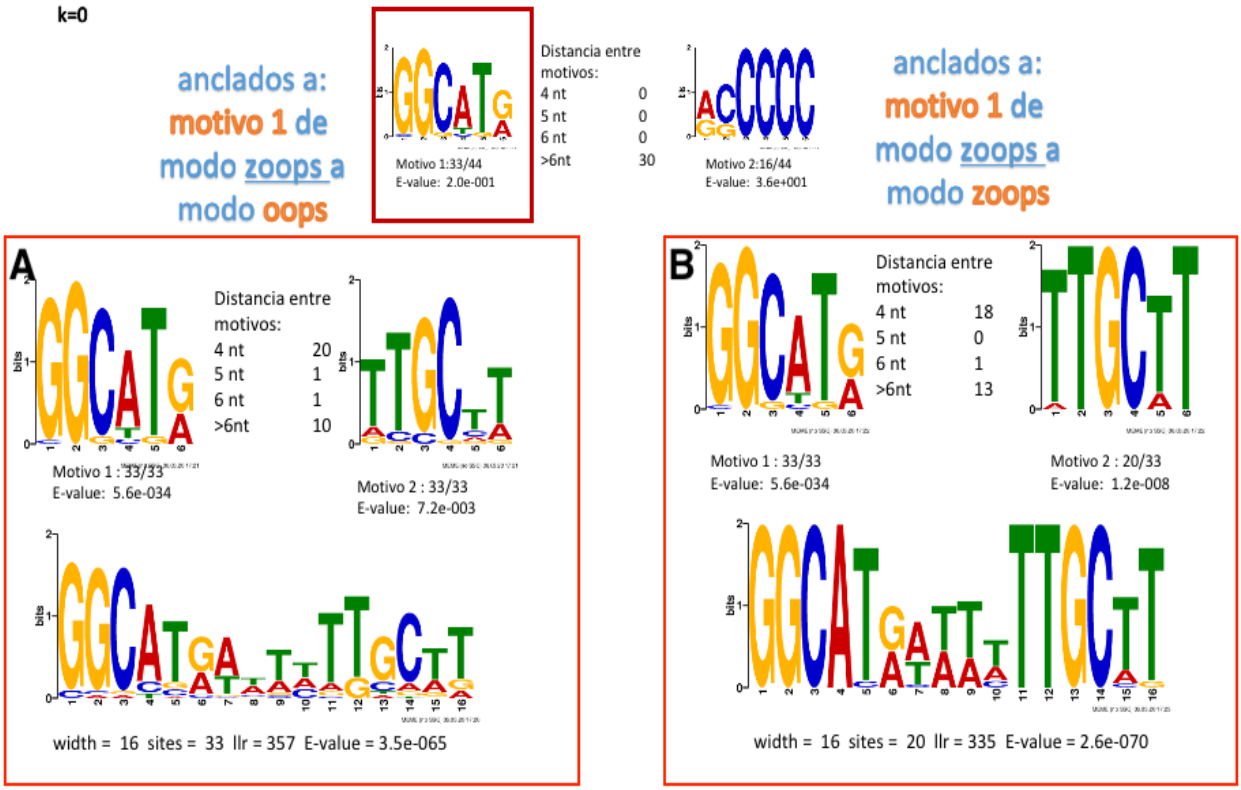


Figura 13. Logos de los resultados reportados por MEME, anclados al primer motivo, modo oops y zoops. Anclados al motivo uno de los análisis de modo zoops, de la Figura 12 (inciso B). En A, se muestran los datos generados vía MEME con motivos separados y un solo motivo de amplitud máxima de 17 pb buscando una ocurrencia por secuencia (oops). En el panel B, mismos parámetros, pero utilizando la búsqueda de cero o una ocurrencia por secuencia (modo zoops). En todos los casos utilizando un modelo de fondo de orden 1.

Resultados del análisis al emplear MAST para la búsqueda.

Considerando una de nuestras hipótesis iniciales de estudio referente a que los genes que codifican para un factor σ, tienden a dirigir la transcripción específica de los genes que los codifican; es decir, se que se autotranscriben, en nuestra primera aproximación metodológica, el ciclo de análisis *pattern-finding/pattern-matching* se inicia tomando como secuencias semilla a aquellas secuencias de regulación de los genes que codifican a los factores σ54. Los primeros resultados muestran que esta hipótesis es cierta para la minoría de los casos, ya que no se cumple para todos los factores σ estudiados. Dado lo anterior, se realizó una segunda aproximación metodológica, que *considera* como hipótesis alternativa que la similitud de los promotores σ54 de organismos de diferentes clases filogenéticas son suficientemente significativas como para identificar a algunos genes transcritos por este factor σ que pudieran utilizarse como secuencias semillas para el inicio del proceso cíclico de análisis. Considerando esta segunda hipótesis, se agruparon las regiones de regulación de los genes que codifican al factor σ54 de los 628 organismos representativos de la clase *Gammaproteobacteria*. Las secuencias así compiladas fueron analizadas con el programa MEME para identificar el motivo de secuencias correspondiente al promotor *σ54* derivado de este análisis *(WDH**TGGCAC**RVMNH**TTGC**HWW)*, misma que se representa en la ***figura 14,*** y cuya matriz correspondiente fue posteriormente empleada por el programa MAST dentro del ciclo identificación-búsqueda. En la literatura, la secuencia consenso reconocida por el complejo RNAP- *σ54 en 47 especies bacterianas*, es: Y**TGGCAC**G-NNNN-**TTGC**WNN (Barrios *et al.*, 1999), muy similar a la secuencia resuelta una vez empleado MEME en las secuencias analizadas en esta primera prueba.

Para el proceso del ciclo de identificación-búsqueda se analizaron diferentes valores de corte para evaluar la precisión de predicción. Para esto, se utilizaron 5 ciclos avaluando 4 valores de corte de MAST (0.25,0.5,1,2), dando un total de 20 matrices, de las cuales se evalúan el número de hits de la matriz generada en cada ciclo, tomando como verdaderos al hit que ocurre en más de 10 conjuntos de los resultados obtenidos al variar las matrices producidas por MEME. Para elegir a la predicción óptima, se seleccionan los resultados de MAST con la matriz con mayor precisión al evaluar los coeficientes de sensibilidad y especificidad de las 20 matrices generadas. En las **figuras 15 y 16** se esquematiza el proceso de selección de mejor predicción y la estadística para la evaluación de la predicción de hits aprobados del organismo modelo *E. coli*, utilizando como datos de referencia el sigmulón[1] de σ54 reportado en RegulonDB, el programa predice 19 hits, es decir 19 genes que son regulados por el factor σ54, de los 94 genes reportados en RegulonDB.

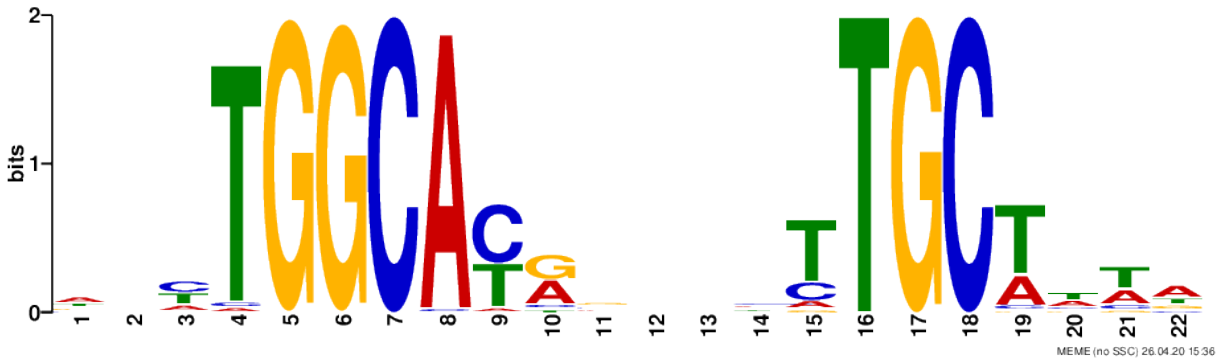


Figura 14. Logo representativo de la matriz utilizada en el ciclo seleccionado como estadísticamente significativo. Se hace una representación gráfica de la conservación de las secuencias de nucléotidos. En el eje de las abscisas se representa las posiciones de cada nucleótido, y en el de las ordenadas, su frecuencia representada en bits, mostrando las bases más conservadas alrededor de una región promotora.

[1] Sigmulón: Grupo de genes que pertenecen a regulones específicos, los cuales son transcritos por las subunidades alternativas (factores σ) de RNAP que son responsables de reconocer los promotores de los genes que pertenecen a regulones particulares (sigmulones) implicados en diversas funciones celulares.

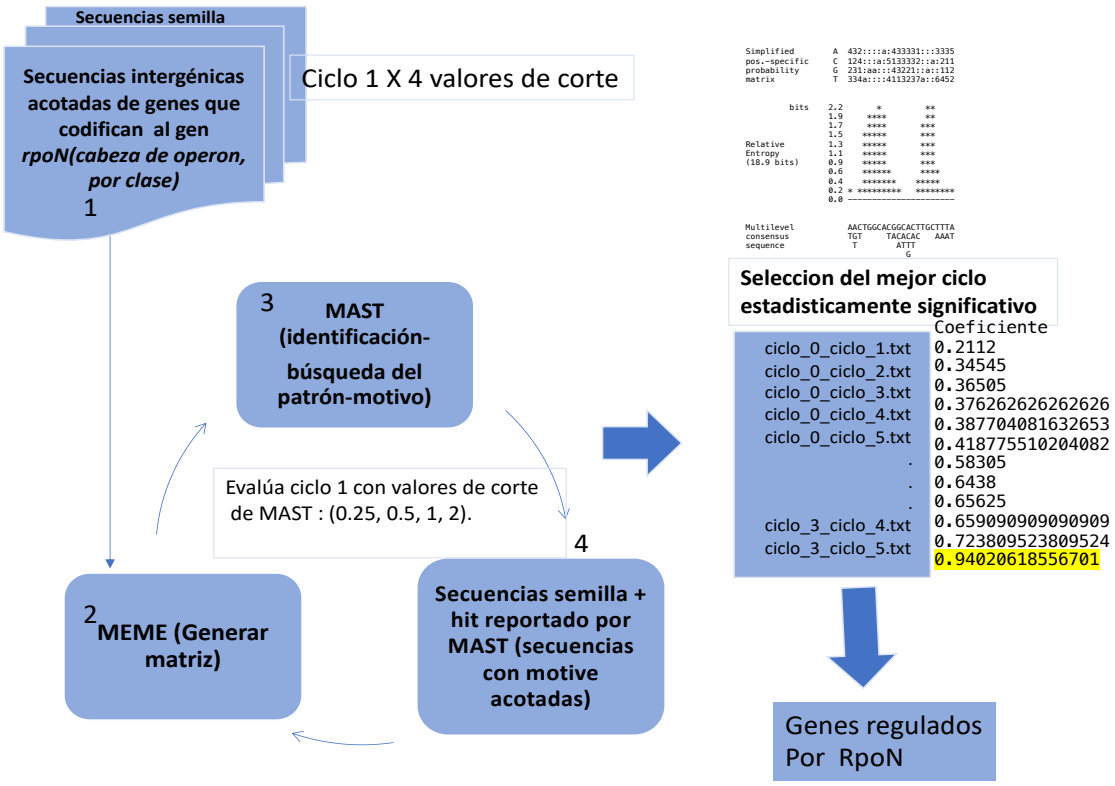
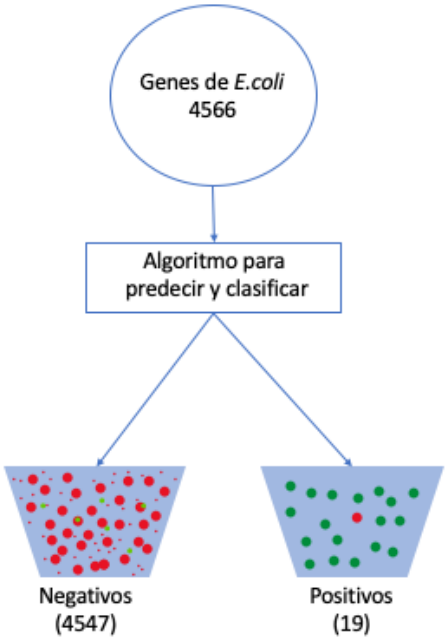


Figura 15. Proceso cíclico del protocolo. Se esquematiza el proceso cíclico metodológico del protocolo, en el paso número 1, se toma la región intergénica 5´ acotada de las secuencias cabeza de operón (secuencias semilla iniciales), seleccionadas a partir de los criterios previamente descritos, en el paso 2, con el uso del programa MEME, se genera una matriz representativa de las secuencias semilla, utilizando esta matriz, MAST busca e identifica los patrones descritos por la matriz y reporta los hits con determinado valor de corte (primer valor 0.25), luego en 4, añade las secuencias de los hits reportados (acotados a las posiciones que MAST reporta), generando un nuevo set de secuencias semilla + hits con las cuales se vuelve a generar la matriz con MEME y se repite el paso 3 y 4, luego 2 cambiando el valor de corte al siguiente valor (ejemplo: 0.5), y vuelve a iterar los pasos 3 y 4, así sucesivamente hasta completar los ciclos con los diferentes valores. Una vez que termina el proceso, se procede a hacer el análisis estadístico, donde se selecciona el ciclo que utiliza la matriz que mejor sobre-representa las secuencias, basándose en el coeficiente mayor entre sensibilidad y especificidad. Se subraya en amarillo el valor del coeficiente que en este ejemplo es el mayor. De este ciclo de análisis, se toman los resultados de hits predichos como genes transcritos derivados de la interacción del factor σ54 y el núcleo de RNAP.



Resultado de la predicción	Genes que de acuerdo con RDB, son regulados por RpoN		
	Regulado	No regulado	Total
Positivo	$18(VP)_a$	$1(FP)_b$	$19_{(a+b)}$
Negativo	$76(FN)_c$	$4471(VN)_d$	$4547_{(c+d)}$
Totales	94	4472	$n=4566_{(a+b+c+d)}$

$$\text{Precisión} = \frac{VP}{VP+FP} = \frac{18}{18+1} = 94.7\%$$

$$\text{Sensibilidad} = \frac{VP}{VP+FN} = \frac{18}{18+76} = 19.1\%$$

$$\text{Especificidad} = \frac{VN}{VN+FP} = \frac{4471}{4472+1} = 99.9\%$$

$$\text{NPV} = \frac{4471}{4547} = 98.3\% \quad \text{PPV} = \frac{18}{19} = 94.7\%$$

Figura 16. Precisión, Sensibilidad y Especificidad de la predicción vs RegulónDB. De los 4,566 genes que contiene *Escherichia coli, e*l conjunto de los 19 genes predichos por el algoritmo elaborado en este proyecto serían potenciales genes transcritos por el factor σN y representan los elementos positivos del conjunto. De este conjunto, las predicciones que fueron reportados y corroborados, en la base de datos RegulonDB (nuestro "estándar de oro") como parte del sigmulón, representan a los verdaderos positivos. Aquellos genes predichos por nuestro algoritmo pero que no se encuentran reportados en RegulonDB, constituyen a los verdaderos negativos. El conjunto de genes que no fueron incluidos en las predicciones de nuestro algoritmo, constituye los elementos negativos del conjunto. De estos elementos clasificados como negativos, los genes que de manera consistente con nuestro algoritmo y con RegulonDB, no son transcritos por el factor σ54, constituyen al conjunto de falsos positivos. Finalmente, aquellos no listados por nuestro algoritmo, pero que según RegulonDB, si son transcritos por el factor σ54, constituyen a los falsos negativos del conjunto. También se calculan el porcentaje de los valores predictivos positivos y negativos (PPV y NPV, respectivamente) que miden la precisión de una prueba diagnóstica.

Análisis de la aplicación de los parámetros definidos con el grupo de prueba en las diferentes clases de estudio.

Para observar el comportamiento de la aplicación de nuestros parámetros definidos para los organismos pertenecientes a la clase Gammaproteobacteria en otro tipo de clase, se utilizaron otros dos conjuntos de datos de microorganismos pertenecientes a las clases Alphaproteobacteria y Bacilli. Se seleccionó al conjunto de datos de la clase

Alphaproteobacterias por considerarse parafiléticas[1] a las Gammaproteobacterias, así como contener suficientes datos para el análisis, asimismo las Alphaproteobacterias abarcan entre sus integrantes a varios organismos que metabolizan componentes del carbono, intracelulares como simbiontes de plantas que son fijadores de nitrógeno y además, existen diversos trabajos donde se hace análisis genómicos amplios acerca de la transcripción de genes mediada por σ54, de algunos organismos en las Alphaproteobacterias (por ejemplo, *Caulobacter crecentus*) (revisado en (Tsang y Hoover, 2014; Tsoy *et al.*, 2016)). Y también se utilizó como tercer conjunto de prueba a la clase Bacilli ya que es una clase de las bacterias del filo Firmicutes, que en su mayoría son Gram positivas, a diferencia de las Alphaproteobacterias, además de que entre sus organismos representativos modelo se encuentra *Bacillus subtilis*. Al observar los datos de los resultados obtenidos en los primeros pasos, donde se crean las matrices de posición de peso mediante MEME, del algoritmo, se notó que la profundidad taxonómica, a nivel de especie, no era suficiente para evitar la redundancia en sus secuencias genómicas; es decir, no han divergido suficiente para generar divergencia en sus secuencias que permitieran identificar sitios conservados por selección funcional. Con el propósito de evitar sesgos en nuestros análisis introducidos por la redundancia en las secuencias de organismos a nivel de especie, solo se consideraron organismos representativos a nivel de género. En la **tabla 3** se incluye el número de organismos y genes que codifican al factor RpoN a nivel de género de las clases que lograron aprobar el primer filtro de los requisitos requeridos (en el primer filtro, más de 4 organismos), además de este primer filtro, se consideraron aquellas secuencias intergénicas de longitud >40 pb, suficientes datos de genes en primera posición en un operón (por conjunto), estas especificaciones se fueron definiendo a lo largo del desarrollo del presente proyecto. En la **gráfica 4** se grafican los datos de la tabla 3. Las clases que a pesar de contar con un número de organismos representativos mayores a nuestro mínimo requerido (5 organismos), no cuentan con la presencia del gen *rpoN* son resaltadas.

[1] En taxonomía filogenética o cladísta, un taxón parafilético se define como uno que incluye el ancestro común más reciente, pero no todos sus descendientes.

Tabla 3. Número de organismos representativos a nivel de género y la distribución de genes que codifican al factor RpoN. Se enlistan el número de organismos representativos presentes a nivel de género en cada conjunto de las diversas clases de estudio, así como el número de factores σ54 presentes en cada conjunto, que fueron analizados por el algoritmo desarrollado en este proyecto. En *negritas* se resaltan las clases que, a pesar de contar con organismos representativos considerables, el porcentaje de genes que codifican para el factor RpoN es poco significativo. El porcentaje de genes que codifican para el factor RpoN por clase se muestra en la columna 5.

N	Clase	Organismos representativos	RpoN	Porcentaje
1	Gammaproteobacteria	193	163	84
2	Alphaproteobacteria	142	115	80.9
3	**Actinobacteria**	**122**	**3**	**2.4**
4	Betaproteobacteria	92	90	97.8
5	Clostridia	63	43	68.2
6	Bacilli	55	45	81.8
7	Flavobacteriia	49	48	97.9
8	Deltaproteobacteria	40	39	97.5
9	**Halobacteria**	**27**	**0**	**0**
10	Cytophagia	22	22	100
11	**Thermoprotei**	**21**	**0**	**0**
12	Bacteroidia	18	18	100
13	**Methanomicrobia**	**15**	**0**	**0**
14	Planctomycetia	12	12	100
15	Epsilonproteobacteria	10	8	80
16	Aquificae	9	9	100
17	Thermotogae	9	5	55.5
18	Spirochaetia	9	9	100
19	**Coriobacteriia**	**9**	**1**	**11.1**
20	**Mollicutes**	**8**	**0**	**0**
21	Chitinophagia	7	7	100
22	**Thermoplasmata**	**7**	**1**	**14.2**
23	**Fusobacteriia**	**6**	**2**	**33.3**
24	Negativicutes	6	4	66.6
25	**Deinococci**	**6**	**0**	**0**

Gráfica 4. Numero de organismos representativos a nivel de género que cuentan con genes que codifican al factor σ54. Se grafican el número de organismos representativos por género (barra púrpura tenue) y el número de ellos que cuentan con al menos un gen que codifica para el factor σ54 (barra púrpura obscura). Los organismos están agrupados de acuerdo a su correspondiente clase filogenética. Se marca en el rectángulo rojo, al grupo de las Gammaproteobacterias, grupo que cuenta con el mayor número de organismos representativos a nivel de género, 193. También se resalta a la clase Actinobacteria que, a nivel de género, solo el 2% de sus 122 organismos representativos cuentan con gen que codifica para el factor σ54.

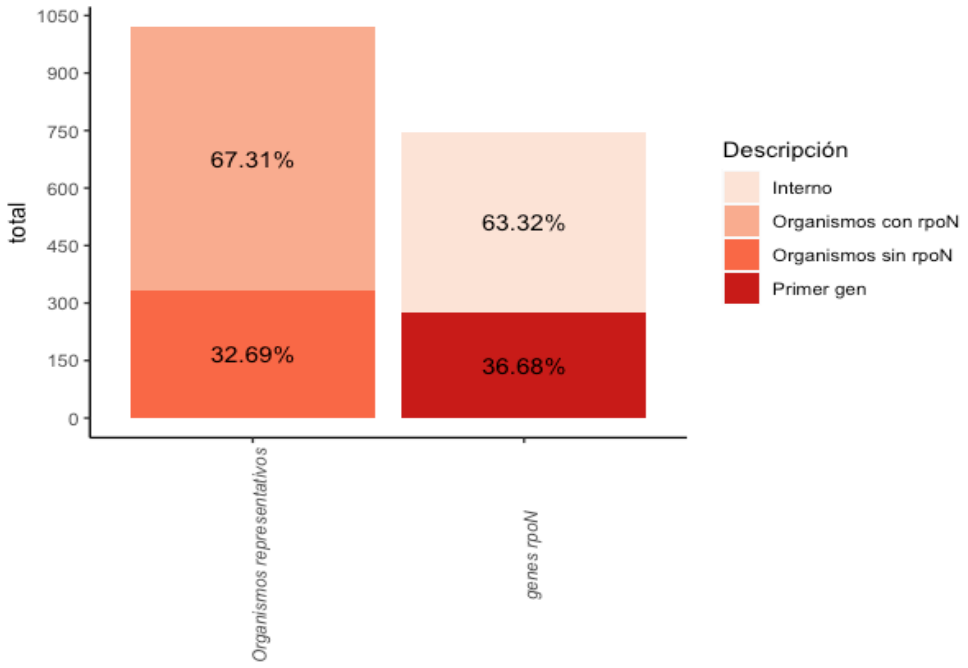
De acuerdo con los datos derivados de agrupar organismos representativos a nivel de especie y género (por clase), se puede observar en las **gráficas 1 y 4** que existen clases filogenéticas para las que la mayoría (si no es que todos) de sus organismos representativos, no cuentan con genes *rpoN*, éstas clases se clasifican dentro del clado de las terrabacterias, como es el caso las Actinobacterias, Cianobacterias y Deinococcus, que son organismos mayormente asociados con hábitats terrestres (Battistuzzi *et al.*, 2004). Y además comparten similitud en la estructura de su pared celular. Dentro de este clado terrestre (Terrabacteria), que incluye Cyanobacterias, los phyla grampositivos (Actinobacteria y Firmicutes) y dos phyla con paredes celulares que difieren estructuralmente de los phyla grampositivos y gramnegativos típicos (Chloroflexi y Deinococcus–Thermus) , poseen adaptaciones importantes como la resistencia a los peligros ambientales (por ejemplo, desecación, radiación ultravioleta y alta salinidad) y

fotosíntesis oxigenada (Pisani *et al.*, 2007). Además, las propiedades únicas de la pared celular en taxones grampositivos, que probablemente evolucionaron en respuesta a condiciones terrestres, han contribuido a la patogenicidad en muchas especies y puede ser que hayan divergido tanto que los genes *rpoN* se perdieron a lo largo de su evolución. *Studholme* ha reportado la ausencia de RpoN en diversas bacterias gram positivas con alto contenido de GC, como es el caso de extremófilo *Thermotoga maritima*, y patógenos especializados, como *Mycoplasma pneumoniae* (Studholme, 2000).

Estadística de los conjuntos de datos de estudio.

Se determinó el nivel taxonómico (género) como el adecuado, para implementar nuestro protocolo de manera global en las 14 clases de estudio finales. A este nivel se espera que las especies hayan divergido lo suficiente para generar divergencia en sus secuencias que permiten identificar sitios conservados por selección funcional. Además, se pretende evitar el sesgo propiciado por la redundancia de secuencia genómica. Para elegir la distancia filogenética de trabajo debe tenerse en cuenta lo anterior para llegar a un punto de equilibrio, ya que, si nos movemos a un nivel mucho más alto, el número de organismos representativos disminuye y la cantidad de información no sería suficiente.

A nivel de género se obtuvieron un total de 1022 organismos representativos totales, de los cuales, 688 organismos cuentan con el gen *rpoN,* es decir el 67.31%. De éstos, el 7.55% cuenta con dos genes *rpoN,* el total de genes que codifican para σ54 es de 747; se realizó la separación de estos genes en dos conjuntos, aquellos que son el primer gen dentro de un operón y aquellos que se encuentran internamente, resultando en 274 y 473 respectivamente, los porcentajes se muestran en la **gráfica 5**. El 36.6% de los genes *rpoN* se encuentran en la primera posición dentro de un operón.



Gráfica 5. Distribución del gen *rpoN* dentro de los organismos de estudio a nivel de género y su posición relativa dentro de la unidad transcripcional . La gráfica representa el porcentaje de organismos a nivel de género que cuentan genes que codifican para el factor σ54, alguno de ellos, como se muestra en la gráfica 2, cuentan con más de una copia; el porcentaje de organismos que no contiene el gen, es el 32.69 % (de los organismos analizados en el presente trabajo). El código de colores indicado en el lado derecho. Así mismo en la segunda barra, se muestra el porcentaje de genes *rpoN, de los 747*, 473 se encuentran internamente en un operón, de los diferentes organismos de estudio, y que se encuentran en la primera posición dentro de un operón son 274.

Resultados de la implementación del protocolo sobre las 14 clases de estudio.

Una vez evaluado nuestro protocolo de estudio sobre los grupos de prueba, los organismos de las clases Gammaproteobacteria, Alphaproteobacteria y Bacilli, se realizó el análisis de las demás clases filogenéticas. Dado el gran volumen de resultados generados para cada uno de los organismos representativos, dentro de cada una de las 14 clases filogenéticas, se construyó un repositorio en la nube que puede ser consultado vía internet en la liga https://biocomputo.ibt.unam.mx/webdata/Indice_de_tablas.html .

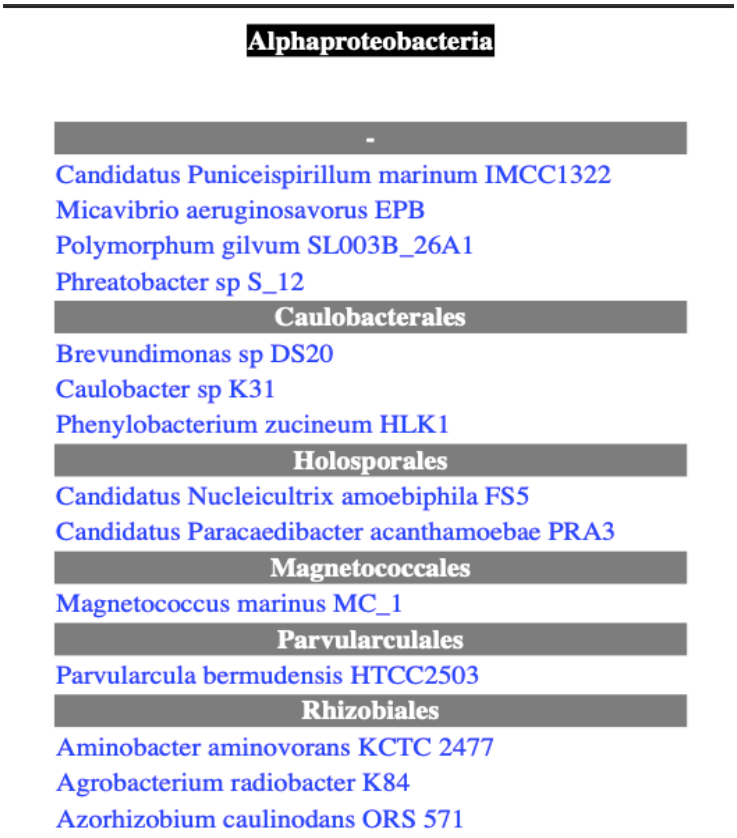
La dirección antes mencionada muestra el índice web de acceso a los resultados de acuerdo a la clase filogenética en cuestión, en la **tabla 4** se encuentran las direcciones

web para consultar los resultados de cada clase. Cada liga de las clases filogenéticas apunta a una página web con la lista de organismos analizados, mismos que están separados a nivel de orden, tal y como se muestra en la **Tabla 5** para la clase Alphaproteobacteria como ejemplo.

Tabla 4. Direcciones web de los resultados, representados en tablas. En azul se representa el hipervínculo a la dirección web escrita en la segunda columna, en esta dirección se encuentran las tablas con los resultados obtenidos en este proyecto, para cada uno de los organismos representativos de las 14 clases de estudio.

Transcripción dependiente del factor Σ 54	
Clase	Link de la dirección web
Alphaproteobacteria	https://biocomputo.ibt.unam.mx/webdata/tabla_Alphaproteobacteria html
Aquificae	https://biocomputo.ibt.unam.mx/webdata/tabla_Aquificae.html
Bacilli	https://biocomputo.ibt.unam.mx/webdata/tabla_Bacilli.html
Bacteroidia	https://biocomputo.ibt.unam.mx/webdata/tabla_Bacteroidia.html
Betaproteobacteria	https://biocomputo.ibt.unam.mx/webdata/tabla_Betaproteobacteria.h ml
Chitinophagia	https://biocomputo.ibt.unam.mx/webdata/tabla_Chitinophagia.html
Clostridia	https://biocomputo.ibt.unam.mx/webdata/tabla_Clostridia.html
Cytophagia	https://biocomputo.ibt.unam.mx/webdata/tabla_Cytophagia.html
Deltaproteobacteria	https://biocomputo.ibt.unam.mx/webdata/tabla_Deltaproteobacteria. tml
Epsilonproteobacteria	https://biocomputo.ibt.unam.mx/webdata/tabla_Epsilonproteobacteri a.html
Flavobacteriia	https://biocomputo.ibt.unam.mx/webdata/tabla_Flavobacteriia.html
Gammaproteobacteria	https://biocomputo.ibt.unam.mx/webdata/tabla_Gammaproteobacter a.html
Planctomycetia	https://biocomputo.ibt.unam.mx/webdata/tabla_Planctomycetia.html
Spirochaetia	https://biocomputo.ibt.unam.mx/webdata/tabla_Spirochaetia.html

Tabla 5. Página web con el índice de resultados de organismos que pertenecen a la clase Alphaproteobacteria, divididos por orden. El nombre de cada organismo direcciona a una página web con los resultados de la predicción de genes transcritos por el factor σ54 de dicho organismo.



Finalmente, los genes identificados en nuestro estudio, por ser regidos por la transcripción específica dirigida por el factor σ54 (potenciales miembros del sigmulón de σ54), son mostrados de manera tabular en un archivo con formato de tipo HTML que contienen la información en tablas de 4 columnas que contienen, el ID del gen, el valor E de la predicción, el número del grupo COG y la descripción funcional de dicho grupo. La **tabla 6** es un ejemplo de dichos datos, en este caso se trata del organismo *Escherichia coli* K12 MG1655, miembro de la clase de las Gammaproteobacterias. Se compilaron los conjuntos genes (por grupo de ortología COG) identificados en todas las clases, en la **tabla 7**, para poder visualizar y analizar los diferentes resultados obtenidos para cada clase de estudio.

Tabla 6 Genes predichos por ser transcritos por el factor σ54 en *Escherichia coli* K12 MG1655. Como ejemplo representativo de los archivos en formato HTML que muestran las predicciones de nuestro estudio de los genes potencialmente transcritos por el factor σ54 en un organismo, se muestran los resultados del análisis en *Escherichia coli K12.* Estos resultados pueden ser obtenidos directamente de nuestro repositorio web cuya dirección es https://biocomputo.ibt.unam.mx/webdata/Indice_de_tablas.html. Los genes desplegados de manera continua del mismo color, corresponden a genes de un mismo operón, en donde el primer gen del operón, es el que tiene la señal de σ54 identificada en el proceso y su correspondiente valor de predicción está indicado en la columna E-val.

Escherichia coli K_12 MG1655			
Ids Gene	**E-val**	**COGs**	**Descripcion**
eco-b2725	0.0079	-------	regulator of the transcriptional regulat
eco-b2710	0.015	COG1773 COG0426	anaerobic nitric oxide reductase flavoru
eco-b2711		COG0446	NADH:flavorubredoxin reductase
eco-b0450	0.0019	COG0347	nitrogen regulator GlnK
eco-b0451		COG0004	ammonia/ammonium transporter
eco-b2726	0.09	COG0375	hydrogenase 3 nickel incorporation prote
eco-b2727		COG0378	hydrogenase isoenzymes nickel incorporat
eco-b2728		COG0298	hydrogenase 3 maturation protein HypC
eco-b2729		COG0409	Fe-(CN)2CO cofactor assembly scaffold pr
eco-b2730		COG0309	hydrogenase maturation protein, carbamoy
eco-b2731		COG3604	DNA-binding transcriptional activator Fh
eco-b0331	0.1	COG2513	2-methylisocitrate lyase
eco-b0333		COG0372	2-methylcitrate synthase
eco-b0334		COG2079	2-methylcitrate dehydratase
eco-b0335		COG0365	propionyl-CoA synthetase
eco-b3421	0.056	COG1690	RNA-splicing ligase
eco-b4475		COG0430	RNA 3'-terminal phosphate cyclase
eco-b0656	0.21	COG3039	IS5 transposase and trans-activator
eco-b2221	0.11	COG1788	acetyl-CoA:acetoacetyl-CoA transferase s
eco-b2222		COG2057	acetyl-CoA:acetoacetyl-CoA transferase s
eco-b2223		COG2031	short chain fatty acid transporter
eco-b2224		COG0183	acetyl-CoA acetyltransferase
eco-b1748	0.24	COG4992	succinylornithine transaminase
eco-b1747		COG3138	arginine N-succinyltransferase
eco-b1746		COG1012	aldehyde dehydrogenase
eco-b1745		COG3724	N-succinylarginine dihydrolase
eco-b1744		COG2988	succinylglutamate desuccinylase
eco-b4002	0.17	COG3678	zinc responsive, periplasmic protein wit
eco-b3461	0.22	COG0568	RNA polymerase, sigma 32 (sigma H) facto
eco-b1783	0.45	COG2766	protein kinase YeaG
eco-b1784		COG2718	DUF444 domain-containing protein YeaH
eco-b1012	0.47	COG2141	pyrimidine oxygenase
eco-b1011		COG1335	peroxyureidoacrylate/ureidoacrylate amid
eco-b1010		COG0251	putative aminoacrylate peracid reductase
eco-b1009		COG0596	putative aminoacrylate hydrolase
eco-b1008		COG0778	putative malonic semialdehyde reductase
eco-b1007		COG1853	flavin reductase
eco-b1006		COG2233	pyrimidine:H(+) symporter

Significado estadístico de la sobre-representación de genes ortólogos presentes en el sigmulón del factor σ54.

De acuerdo al número de genes identificados por tener de un promotor reconocido por el factor σ54 en los diferentes grupos de genes ortólogos de la base de datos COGs, se identificaron los grupos mayormente regulados dentro de las clases filogenéticas de estudio**.** El análisis de enriquecimiento de genes regulados por el factor σ54 en las diferentes clases filogenéticas, mostraron que los genes transcritos por este factor σ en bacterias corresponden a aquellos relacionados con el metabolismo del nitrógeno, tales como nitrogenasas, factores transcripcionales en respuesta al nitrógeno, proteínas fijadoras de nitrógeno y transportadores de nitrato, nitrito y amonio, urea-amidohidrolasas, genes relacionados con el metabolismo de aminoácidos, transporte de fructosa, al igual que el propio gen *rpoN* que codifica al factor σ54 (en algunos casos). Adicionalmente, también se observó un enriquecimiento en los genes relacionados con la biogénesis del flagelo, sobre todo en los organismos de las Proteobacterias (Alpha, beta, Delta, Épsilon y Gamma) y Bacilli, como se puede observar en la **tabla 7**.

Los resultados derivados de implementar el protocolo desarrollado en esta investigación se encuentran agrupados mediante la técnica de clusterización K-means[1] (MacQueen, 1967) en la **gráfica 6**. Se utilizaron los datos de la **tabla 7** que muestra el significado estadístico de la sobre-representación de genes ortólogos transcritos por el factor σ54. En este análisis se puede observar que existen 4 grupos. En el grupo 1 únicamente se encuentra la clase Deltaproteobacteria, que, aún cuando comparte gran similitud de los genes predichos con otras Proteobacterias, el algoritmo la separa, se puede observar en la gráfica 6 que este clúster 1 se encuentra cercano al clúster donde se agrupan las Proteobacterias que comparten mayor parecido, Alpha, Beta y Gamma, excluyendo a las Deltaproteobacterias y las Épsilón de este grupo denominado 3. Esto quiere decir que algunos de los genes predichos están presentes en este grupo, pero no

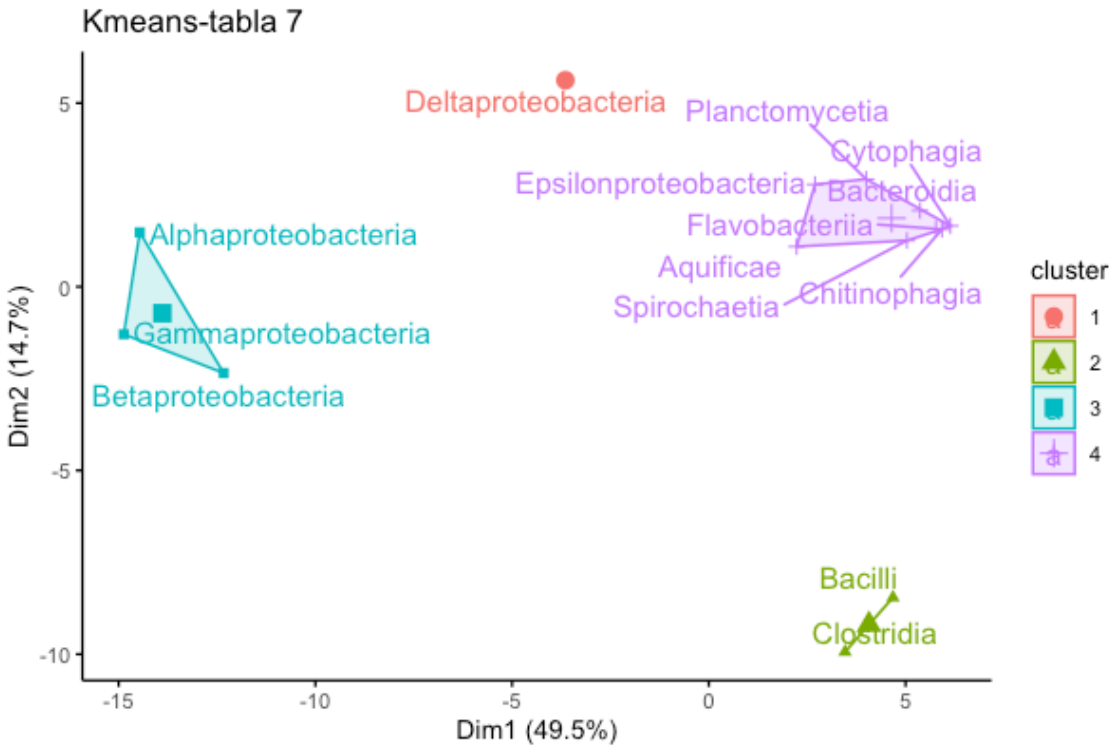
[1] El método k-means, agrupa las observaciones en K clusters distintos, donde el numero k lo determina el usuario antes de ejecutar el algoritmo. Encuentra los K mejores ,entendiendo como mejor agrupación aquel cuya varianza interna sea lo más pequeña posible. K-means trata de agrupar los datos por su distancia, juntando a aquellos más parecidos o cercanos. Es un método bastante utilizado por su sencillez y velocidad, es un método heurístico.

todos. Los principales genes presentes en los grupos 1, 3 y 4 son relacionados con la motilidad celular, metabolismo y transporte de aminoácidos y producción y conversión de energía. El grupo 2, está conformado por las clases Clostridia y Bacilli, que presentan COGs con descripción de transporte y metabolismo de carbohidratos, además de COGS relacionados con el transporte y metabolismo de lípidos, mismos que no se encuentran en las otras clases, por eso el algoritmo de K-means las conjunta en un grupo bastante alejado de todos los clústers. Es importante mencionar que existen organismos en estas dos clases en los cuales los genes *rpoN* se autotranscriben, al igual que en las clases de las Proteobacterias: Alpha,Gamma, Beta y la clase Planctomycetia. Sin embargo, comparten más similitud al clúster 4, según la posición que le confiere K-means y lo que se aprecia en la tabla 7. En el grupo 4, se observa un enriquecimiento de los genes predichos de algunas chaperonas como GroES, Chaperona molecular clase DnaJ con dominio de dedo de Zn C-terminal y GrpE, donde se encuentran reunidas las clases Planctomycetia, Aquificae, Cytophagia, Bacteroidia, Flavobacteriia, Chitinophagia Spirochaetia y Épsilonbacteria. Otras chaperonas que sólo se observan en el grupo 3, son las relacionadas con la biosíntesis del flagelo, no presentes en otras clases.

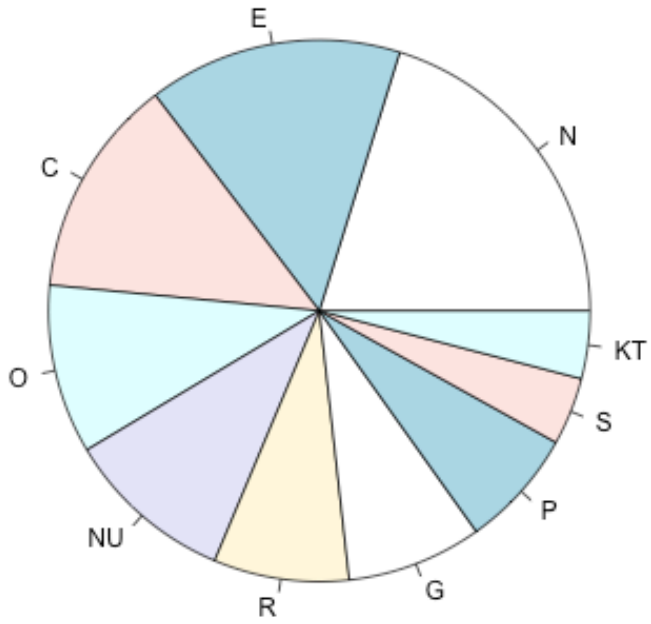
En **tabla 7**, se puede apreciar que los grupos de genes ortólogos mayormente regulados dentro de clases Spirochaetia y Deltaproteobacteria tienden compartir los grupos de COGS, por ejemplo: COG2710, COG0347, COG1348, COG0004, COG0535 y COG1842; los primeros cuatro, relacionados con el transporte de iones y metabolismo. Esto hace sentido en contexto filogenético, ya que los clados de estas clases son cercanos entre sí (ver árbol en **figura 17**).

Se han descrito, en diferentes organismos, genes no relacionados entre sí (funcionalmente) transcritos a partir de promotores dependientes de σ54; Cases et al. mediante su método de predicción lograron un 80% de confiabilidad de sus predicciones, basados en encontrar la región de unión del activador de la transcripción mediada por el factor σ54 y la región promotora, además tener en cuenta evidencia del contexto cromosomal y la similitud con bacterias cercanamente relacionadas, reportaron el sigmulón σ54 de *Pseudomonas putida KT2440*; las funciones descritas para los genes en el sigmulón de esta especie, coinciden con la de los genes predichos mediante

nuestro protocolo, esta concordancia sustenta nuestras predicciones, y pueden considerarse bona fide (Cases *et al.*, 2003; Pátek y Nešvera, 2011). Entre los genes predichos una vez implementado nuestro algoritmo, el 59% de los genes rpoN con potenciales secuencias promotoras que reconoce la RNAP-σ54, se encuentran en la primera posición de un operón. El porcentaje de genes rpoN que se autotranscribe es de 5.87%, del total de genes de los organismos analizados en el presente trabajo, a nivel de género.



Gráfica 6. Significado estadístico de la sobre-representación de genes ortólogos transcritos por el factor σ54 clusterizado mediante Kmeans. De acuerdo al número de genes identificados por tener un promotor reconocido por el factor σ54 en los diferentes grupos de genes ortólogos de la base de datos COGs, se identificaron los grupos mayormente regulados dentro de las clases filogenéticas de estudio. Se realizó una clusterización con 4 centros utlizando la técnica kmeans. Datos de la tabla 7.



Gráfica 7. Clasificación de los COG por categorías funcionales y el número de éstos presente en los resultados de las predicciones. Abreviaturas de una letra para las categorías funcionales: recombinación y reparación; K, transcripción; O, chaperonas moleculares y funciones relacionadas; M, estructura de la pared celular y biogénesis y membrana externa; N, secreción, motilidad y quimiotaxis; T, transducción de señales; P, transporte y metabolismo de iones inorgánicos; C, producción y conversión de energía; G, metabolismo y transporte de carbohidratos; E, metabolismo y transporte de aminoácidos; F, metabolismo y transporte de nucleótidos; H, metabolismo de las coenzimas; Y, metabolismo de los lípidos; U,Tráfico, secreción y transporte vesicular intracelular; R, predicción funcional general solamente; S, sin predicción funcional.

En la **gráfica 7,** es la representación gráfica de los COGs predichos por el protocolo desarrollado en esta investigación que se observan en la tabla 7, tomando en cuenta las categorías funcionales de cada COG, como se puede observar, la mayor parte de los COGs predichos se encuentran clasificados como N (secreción, motilidad y quimiotaxis), E (metabolismo y transporte de aminoácidos), C (producción y conversión de energía), y O (chaperonas moleculares y funciones relacionadas).

Tabla 7. Significado estadístico de la sobre-representación de genes ortólogos transcritos por el factor σ54. De acuerdo al número de genes identificados por tener un promotor reconocido por el factor σ54 en los diferentes grupos de genes ortólogos de la base de datos COGs, se identificaron los grupos mayormente regulados dentro de las clases filogenéticas de estudio. Cuanto más pequeño es el valor de E, mayor es el valor estadísticamente representativo, siendo cero, el valor de mayor significancia estadística. Por el contrario, la anotación "---" significa que el valor de sobre-representación no es significativo.

COG	Descripcion	Alphaproteobacteria	Betaproteobacteria	Deltaproteobacteria	Epsilonproteobacteria	Gammaproteobacteria	Bacilli	Clostridia	Aquificae	Bacteroidia	Chitinophagia	Cytophagia	Flavobacteriia	Planctomycetia	Spirochaetia
COG2710	Nitrogenase molybdenum-iron protein, alpha and beta chains	0	0	0	---	0	---	---	0.0001	0	---	---	---	---	0
COG0347	Nitrogen regulatory protein PII	0	0	0	---	0	---	---	0	0	---	---	---	---	0
COG1348	Nitrogenase subunit NifH (ATPase)	0	0	0	---	0	---	---	---	0.0001	---	---	---	---	0
COG5554	Nitrogen fixation protein	0	0	---	0	0	---	---	---	---	---	---	---	---	---
COG2223	Nitrate/nitrite transporter	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG1251	NAD(P)H-nitrite reductase	0	0	---	---	0	---	---	0.0003	---	---	---	---	---	---
COG0715	ABC-type nitrate/sulfonate/bicarbonate transport systems, pe	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG0600	ABC-type nitrate/sulfonate/bicarbonate transport system, per	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG2146	Ferredoxin subunits of nitrite reductase and ring-hydroxylat	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG1116	ABC-type nitrate/sulfonate/bicarbonate transport system, ATP	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0004	Ammonia permease	0	0	0	---	0	---	---	0	---	---	---	---	---	0
COG2986	Histidine ammonia-lyase	---	0	---	---	0	0.0004	---	---	---	---	---	---	---	---
COG0378	Ni2+-binding GTPase involved in regulation of expression and	0	0	---	---	0	---	---	---	---	---	---	0.0007	---	---
COG2371	Urease accessory protein UreE	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG2370	Hydrogenase/urease accessory protein	0	0	---	---	0.0004	---	---	---	---	---	---	---	---	---
COG0832	Urea amidohydrolase (urease) beta subunit	0	0	---	---	0.0003	---	---	---	---	---	---	---	---	---
COG0831	Urea amidohydrolase (urease) gamma subunit	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0830	Urease accessory protein UreF	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0829	Urease accessory protein UreH	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0804	Urea amidohydrolase (urease) alpha subunit	0	0	---	---	0.0002	---	---	---	---	---	---	---	---	---
COG0174	Glutamine synthetase	0	0	0	---	0	---	---	0	---	---	---	---	0.0008	---
COG3968	Uncharacterized protein related to glutamine synthetase	---	---	0	---	---	---	---	---	---	---	---	---	0.0001	0
COG2610	H+/gluconate symporter and related permeases	---	0.0002	---	---	---	0	0	---	---	---	---	---	---	---
COG1071	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase	---	0	---	---	---	0	0	---	---	---	---	---	---	---
COG0508	Pyruvate/2-oxoglutarate dehydrogenase complex, dihydrolipoam	---	0.0009	---	---	---	0	0	---	---	---	---	---	---	---
COG0334	Glutamate dehydrogenase/leucine dehydrogenase	---	---	---	---	---	0	0	---	---	---	---	0.0003	---	---
COG0154	Asp-tRNAAsn/Glu-tRNAGln amidotransferase A subunit and relat	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0022	Pyruvate/2-oxoglutarate dehydrogenase complex, dehydrogenase	---	0	---	---	---	0	0	---	---	---	---	---	---	---
COG1815	Flagellar basal body protein	0	0	0	0	0	---	---	---	---	---	---	---	0.0002	---
COG1677	Flagellar hook-basal body protein	0	0	0	0	0	---	---	---	---	---	---	---	0.0002	---
COG1558	Flagellar basal body rod protein	0	0	0	0	0	---	---	---	---	---	---	---	0.0002	---
COG4786	Flagellar basal body rod protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG3144	Flagellar hook-length control protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG2063	Flagellar basal body L-ring protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG1843	Flagellar hook capping protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG1766	Flagellar biosynthesis/type III secretory pathway lipoprotei	0	0	0	---	0	---	---	---	---	---	---	---	0.0004	---
COG1749	Flagellar hook protein FlgE	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG1706	Flagellar basal-body P-ring protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG1536	Flagellar motor switch protein	0	0	0	---	0	---	---	---	---	---	---	---	0.0006	---
COG1344	Flagellin and related hook-associated proteins	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG1317	Flagellar biosynthesis/type III secretory pathway protein	0	0	0	---	0	---	---	---	---	---	---	---	0.0002	---
COG1157	Flagellar biosynthesis/type III secretory pathway ATPase	0	0	0	---	0	---	---	---	---	---	---	---	0.0002	---
COG3190	Flagellar biogenesis protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG2882	Flagellar biosynthesis chaperone	---	0	0	---	0	---	---	---	---	---	---	---	0.0001	---
COG1987	Flagellar biosynthesis pathway, component FliQ	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG1886	Flagellar motor switch/type III secretory pathway protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG1868	Flagellar motor switch protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG1684	Flagellar biosynthesis pathway, component FliR	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG1580	Flagellar basal body-associated protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---

Continúa tabla en la siguiente página...

COG1345	Flagellar capping protein	0	---	0	0	0	---	---	---	---	---	---	---	---	---
COG1338	Flagellar biosynthesis pathway, component FliP	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG1256	Flagellar hook-associated protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG4787	Flagellar basal body rod protein	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG1516	Flagellin-specific chaperone FliS	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG1419	Flagellar GTP-binding protein	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG1377	Flagellar biosynthesis pathway, component FlhB	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG1360	Flagellar motor protein	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG1334	Uncharacterized flagellar protein FlaG	0	---	---	---	0	---	---	---	---	---	---	---	---	0
COG1291	Flagellar motor component	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG1261	Flagellar basal body P-ring biosynthesis protein	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG4969	Tfp pilus assembly protein, major pilin PilA	---	0	0.001	---	0	---	---	0.0004	---	---	---	---	---	---
COG1508	DNA-directed RNA polymerase specialized sigma subunit, sigma	0	0	---	---	0	0.0002	0.0004	---	---	---	---	---	0.0004	---
COG1842	Phage shock protein A (IM30), suppresses sigma54-dependent t	0	---	0.0003	---	0	---	---	---	---	---	---	---	---	0
COG1191	DNA-directed RNA polymerase specialized sigma subunit	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG4177	ABC-type branched-chain amino acid transport system, permeas	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG0683	ABC-type branched-chain amino acid transport systems, peripl	0	0	---	---	0	---	---	0.0002	---	---	---	---	---	---
COG0577	ABC-type antimicrobial peptide transport system, permease co	0.0004	0.0003	---	---	0	---	---	---	0	---	---	---	---	---
COG0559	Branched-chain amino acid ABC-type transport system, permeas	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG0410	ABC-type branched-chain amino acid transport systems, ATPase	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG4674	Uncharacterized ABC-type transport system, ATPase component	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG4603	ABC-type uncharacterized transport system, permease componen	0	0.0001	---	---	0	---	---	---	---	---	---	---	---	---
COG3845	ABC-type uncharacterized transport systems, ATPase component	0	0.0001	---	---	0	---	---	---	---	---	---	---	---	---
COG2031	Short chain fatty acids transporter	0.0001	---	---	---	0	---	0	---	---	---	---	---	---	---
COG1744	Uncharacterized ABC-type transport system, periplasmic compo	0	0	---	---	0.0004	---	---	---	---	---	---	---	---	---
COG1176	ABC-type spermidine/putrescine transport system, permease co	0	0.0009	---	---	0.0002	---	---	---	---	---	---	---	---	---
COG1079	Uncharacterized ABC-type transport system, permease componen	0	0.0001	---	---	0	---	---	---	---	---	---	---	---	---
COG0765	ABC-type amino acid transport system, permease component	0.0008	0	---	---	0	---	---	---	---	---	---	---	---	---
COG0444	ABC-type dipeptide/oligopeptide/nickel transport system, ATP	0	0	---	---	0.0005	---	---	---	---	---	---	---	---	---
COG2204	Response regulator containing CheY-like receiver, AAA-type A	0	0	0	---	0	---	0.0002	---	---	---	---	---	---	---
COG1145	Ferredoxin	0	0	---	---	0	---	---	0.0004	---	---	---	0.0006	0.0004	---
COG1433	Uncharacterized conserved protein	0	0	0	0	0	---	---	---	---	---	---	---	---	---
COG0576	Molecular chaperone GrpE (heat shock protein)	---	---	---	---	---	---	---	---	0	---	0	0	0	0.001
COG0443	Molecular chaperone	---	---	---	---	---	---	---	---	0	0	0.0005	0.0006	0	---
COG3951	Rod binding protein	0	0	0	---	0	---	---	---	---	---	---	---	---	---
COG3829	Transcriptional regulator containing PAS, AAA-type ATPase, a	0	---	---	---	---	0	0	0	---	---	---	---	---	---
COG3604	Transcriptional regulator containing GAF, AAA-type ATPase, a	0	---	---	0.0001	0	---	---	0	---	---	---	---	---	---
COG3383	Uncharacterized anaerobic dehydrogenase	0	0	---	---	0	---	0.0002	---	---	---	---	---	---	---
COG1513	Cyanate lyase	0	0	---	---	0	---	---	0	---	---	---	---	---	---
COG0631	Serine/threonine protein phosphatase	0.0001	0	---	---	0	---	---	---	---	---	---	---	0.0001	---
COG0624	Acetylornithine deacetylase/Succinyl-diaminopimelate desucci	---	0	---	---	0.0002	0	0	---	---	---	---	---	---	---
COG0484	DnaJ-class molecular chaperone with C-terminal Zn finger dom	---	---	---	---	---	---	---	---	0	---	0	0	0	---
COG0459	Chaperonin GroEL (HSP60 family)	---	---	---	---	---	---	---	---	---	0	0	0	0	---
COG0402	Cytosine deaminase and related metal-dependent hydrolases	0	0	---	---	0	---	0.0003	---	---	---	---	---	---	---
COG0243	Anaerobic dehydrogenases, typically selenocysteine-containin	0	0	---	---	---	---	0	0	---	---	---	---	---	---
COG0234	Co-chaperonin GroES (HSP10)	---	---	---	---	---	---	---	---	---	0.0001	0	0	0	---
COG0007	Uroporphyrinogen-III methylase	---	0	---	---	0	---	---	0	---	---	---	---	0.0004	---
COG4789	Type III secretory pathway, component EscV	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG4656	Predicted NADH:ubiquinone oxidoreductase, subunit RnfC	0	0	---	---	---	---	0	---	---	---	---	---	---	---
COG4647	Acetone carboxylase, gamma subunit	0	0	---	---	---	0	---	---	---	---	---	---	---	---
COG3933	Transcriptional antiterminator	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG3716	Phosphotransferase system, mannose/fructose/N-acetylgalactos	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG3715	Phosphotransferase system, mannose/fructose/N-acetylgalactos	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG3707	Response regulator with putative antiterminator output domai	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG3665	Uncharacterized conserved protein	0	0	---	---	0	---	---	---	---	---	---	---	---	---
COG3444	Phosphotransferase system, mannose/fructose/N-acetylgalactos	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG3199	Uncharacterized conserved protein	0	0	---	---	---	---	0.0007	---	---	---	---	---	---	---
COG2893	Phosphotransferase system, mannose/fructose-specific compone	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG2604	Uncharacterized protein conserved in bacteria	---	---	0	0	0	---	---	---	---	---	---	---	---	---
COG2513	PEP phosphonomutase and related enzymes	---	0	---	---	0	---	0.0002	---	---	---	---	---	---	---
COG2057	Acyl CoA:acetate/3-ketoacid CoA transferase, beta subunit	---	---	---	---	0.0001	0	0	---	---	---	---	---	---	---
COG1788	Acyl CoA:acetate/3-ketoacid CoA transferase, alpha subunit	---	---	---	---	0.0001	0	0	---	---	---	---	---	---	---
COG1455	Phosphotransferase system cellobiose-specific component IIC	---	---	---	---	---	0	0	---	---	---	---	---	---	0.0001
COG1454	Alcohol dehydrogenase, class IV	---	---	0	---	---	0	0	---	---	---	---	---	---	---
COG1447	Phosphotransferase system cellobiose-specific component IIA	---	---	---	---	---	0	0	---	---	---	---	---	---	0.0001
COG1440	Phosphotransferase system cellobiose-specific component IIB	---	---	---	---	---	0	0	---	---	---	---	---	---	0.0001
COG1335	Amidases related to nicotinamidase	0	---	---	---	0	---	---	---	---	---	---	0.0002	---	---
COG1301	Na+/H+-dicarboxylate symporters	0	0	---	---	---	---	0	---	---	---	---	---	---	---
COG1221	Transcriptional regulators containing an AAA-type ATPase dom	---	---	---	---	0	0	0	---	---	---	---	---	---	---
COG1142	Fe-S-cluster-containing hydrogenase components 2	---	---	0	---	0	---	0.0004	---	---	---	---	---	---	---
COG1104	Cysteine sulfinate desulfinase/cysteine desulfurase and rela	0	0	---	---	0.0003	---	---	---	---	---	---	---	---	---
COG1012	NAD-dependent aldehyde dehydrogenases	---	0.0009	---	---	---	0	---	---	---	---	---	---	0.0008	---
COG0633	Ferredoxin	0	0	---	---	---	0	---	---	---	---	---	---	---	---
COG0535	Predicted Fe-S oxidoreductases	0	---	0	---	---	---	---	---	---	---	---	---	---	0.0001
COG0455	ATPases involved in chromosome partitioning	0	---	0	---	0	---	---	---	---	---	---	---	---	---
COG0413	Ketopantoate hydroxymethyltransferase	---	---	---	---	0	---	0	---	---	---	0.0002	---	---	---
COG0183	Acetyl-CoA acetyltransferase	0	---	---	---	---	0	0	---	---	---	---	---	---	---
COG0071	Molecular chaperone (small heat shock protein)	---	---	---	---	---	---	---	---	0.0001	---	---	---	0	0

Fin de la tabla 7.

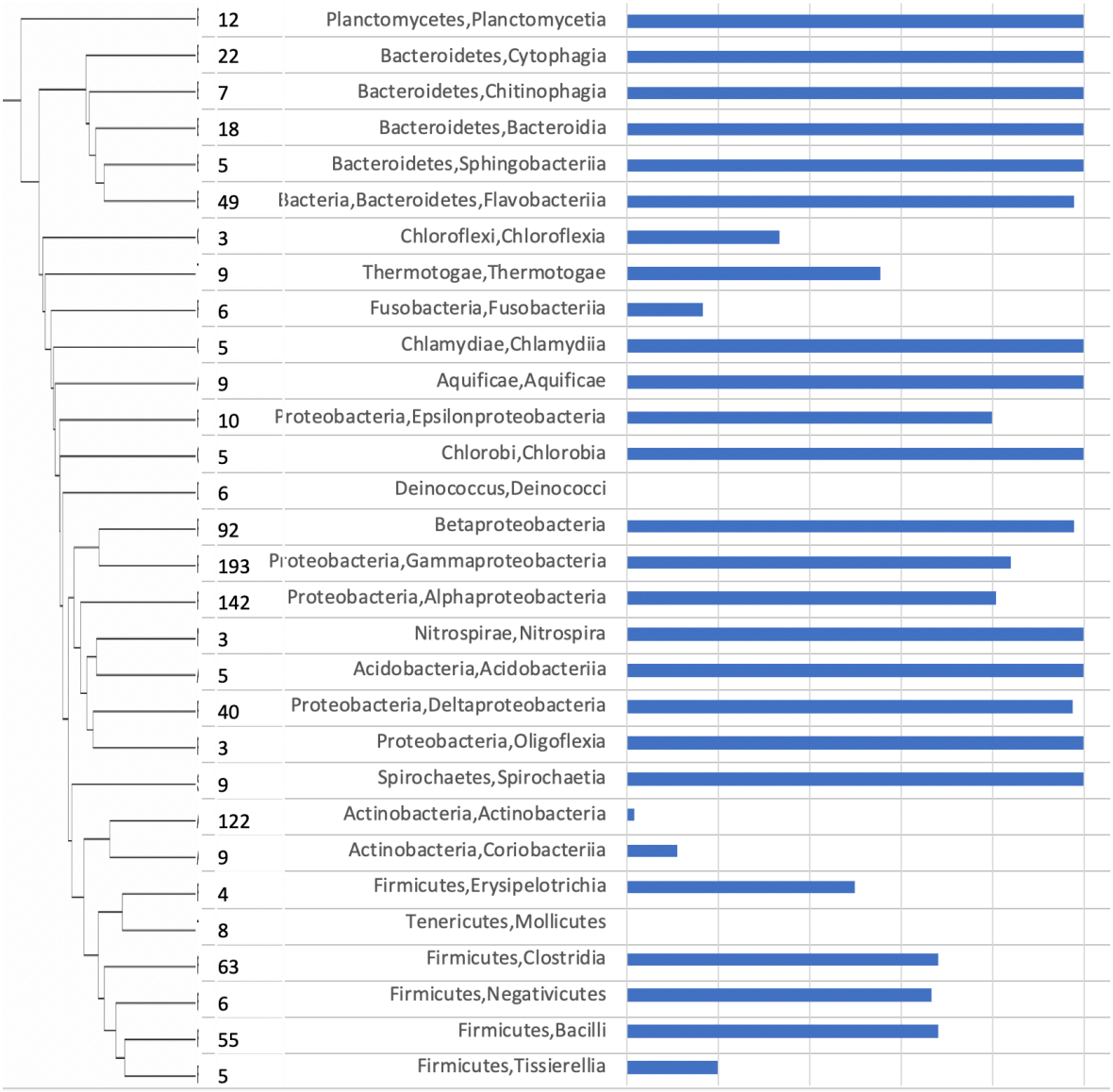
Análisis filogenético de la prevalencia del factor σ54 en bacterias.

Con el propósito de analizar la distribución filogenética del factor σ54 en bacterias se ubicaron dentro de un árbol las diferentes clases consideradas en el presente trabajo y la frecuencia de organismos representativos que contenían el gen que codifica para el factor σ de estudio.

El filo de Proteobacteria, las bacterias Gram negativas, incluyen una variedad muy grande de patógenos, muchas se mueven utilizando flagelos. El análisis del rRNA 16S indica que podría no ser monofilético, dividiéndose en dos grupos: Thiobacteria (bacterias del sulfato y relacionadas) y Rhodobacteria (bacterias púrpuras y relacionadas). Se ha propuesto que Aquificae forma parte de Thiobacteria, sin embargo, es probable que este acercamiento se deba a una gran transferencia genética horizontal entre los Aquificae y Épsilonproteobacteria (Battistuzzi *et al.*, 2004; van Passel *et al.*, 2014). Esto sucede con el gen σ54, como se puede observar en la similitud de logos reportados por nuestro análisis. Lo mismo sucede con el grupo denominado Rhodobacteria (Alpha, Gamma y Beta-proteobacterias) la similitud de logos es muy marcada.

Cabe destacar, algunos grupos filogenéticos en donde pudo haberse perdido selectivamente el factor σ54, como el caso de la clase Deinoccoci del dominio de las Bacterias y del phylum Deinococcus, que tienen una pared celular gruesa como las Gram-positivas, pero presentan una estructura que incluye dos membranas como las gram-negativas, y están más cercanas filogenéticamente hablando de las Proteobacterias. En general, las bacterias Gram-positivas, tienden a disminuir o a perder el factor σ54, como es el caso de las Actinobacterias y Firmicutes. Como se puede observar en la **figura 17**, los grupos de las bacterias Gram-positivas, la distribución de este factor **σ** es mucho menor que en las Gram-negativas.(Battistuzzi y Hedges, 2009)

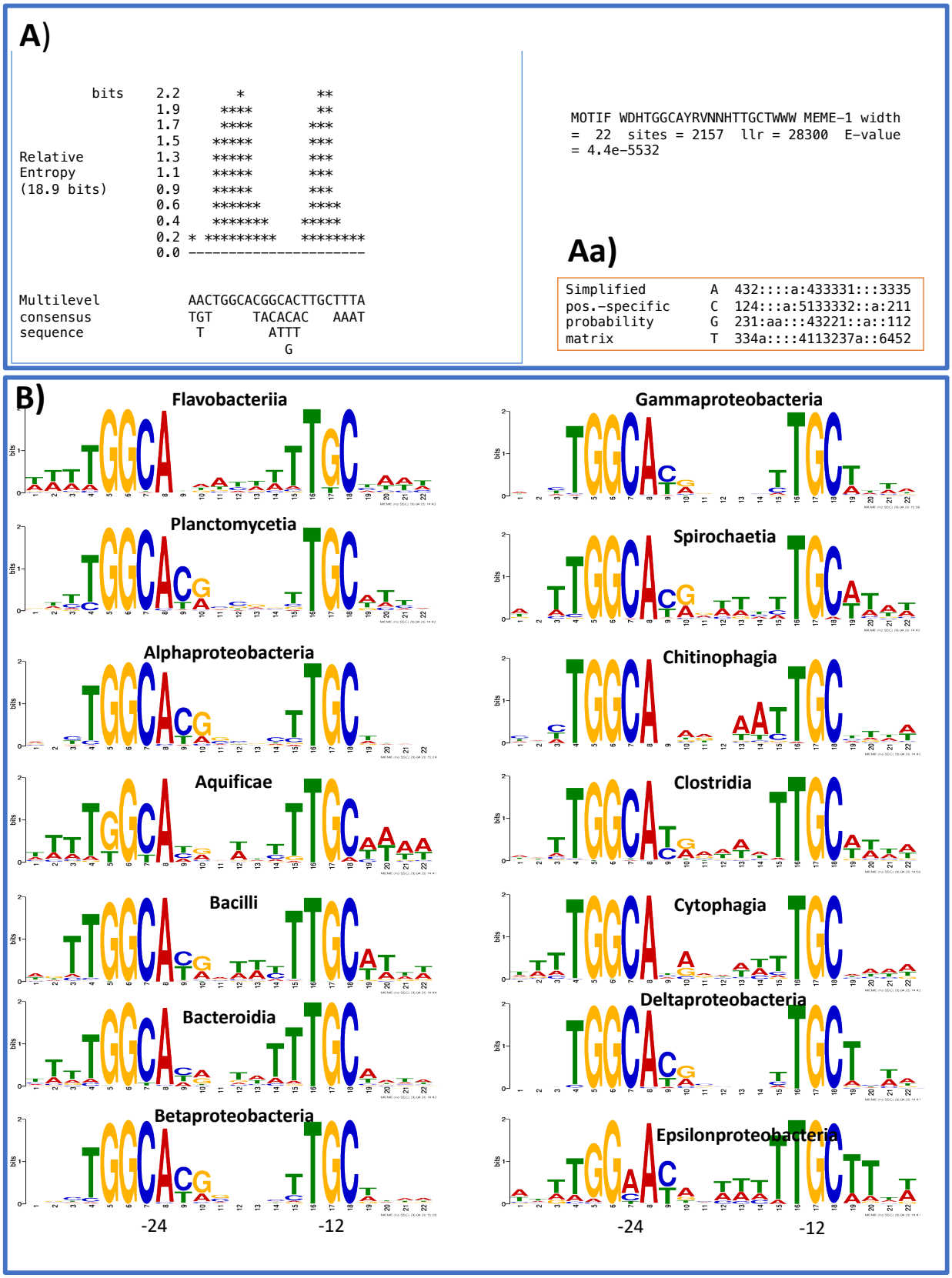
Figura 17. Distribución filogenética del factor σ54. El árbol filogenético fue realizado en base a la secuencia ribosomal 16S de los organismos de nuestro proyecto utilizando el programa ss-align para el alineamiento de las secuencias de acuerdo al modelo de covarianza de bacterias. Posteriormente se usan dichos alineamientos como datos de entrada al programa jmodeltest quien determina el tipo de modelo filogenético que mejor se adapta a las secuencias. En un paso posterior, mediante el uso del paquete PHYLIP y con el modelo filogenético identificado, se evalúa el cálculo de las distancias filogenéticas entre todos los pares de secuencias, mismas que son compiladas en una matriz de distancias. Finalmente, con base a dicha matriz de distancias, el programa agrupa y construye un árbol en base al método de distancia Fitch-Margoliash que estima filogenias bajo un modelo de árbol aditivo en el que se supone un reloj evolutivo. El número de organismos representativos a nivel de género agrupados por clase filogenética se anota delante de cada nombre y las barras de color azul presentan el porcentaje de estos que presentan el gen *rpoN*.



Análisis de las secuencias consenso predichas por clase filogenética.

Como se describió anteriormente, una matriz de peso posicional o PSSM, representa la frecuencia relativa de sus elementos, nucleótidos en el caso de secuencias de DNA, de una colección de secuencias alineadas o motivos. Y en nuestro caso, los motivos corresponden a los promotores transcripcionales reconocidos por el factor σ54 en las diferentes clases filogenéticas analizadas aquí, mismas que se observan en la **figura 18**. En dicha figura se puede apreciar que existe una gran conservación en la secuencia entre los promotores σ54, pese a que pertenecen a grupos filogenéticamente distantes. En todos los grupos, los motivos GGCA (caja -24) y TGC (caja -12) son altamente conservados. En adición a estas bases, también se puede observar que, con menor frecuencia, también se conservan otras bases, co que antecede a la caja -24 de un promotor σ54 .

Figura 18. Representación por logos de los motivos de secuencias de promotores σ54 de las diferentes clases filogenéticas de estudio. En A) se muestra un histograma de frecuencias y su secuencia consenso multinivel, un ejemplo de los resultados reportados por MEME para 2157 secuencias de Gammaproteobacterias, así cómo su respectiva PSSM en el apartado Aa). A partir del proceso cíclico de análisis basado en pattern-finding/pattern-matching, se generaron 20 diferentes modelos MEME por cada una de las 14 clases filogenéticas contempladas en nuestro estudio de las cuales se seleccionaron aquellas que fueron evaluadas por ser las más precisas. Dichas matrices se representan en forma de logos en el panel B. En el eje de las abscisas se representan la posición de los nucleótidos dentro del motivo. En el eje de las ordenadas se representa el contenido informacional de la base conservada dentro del motivo. Para secuencias de DNA con cuatro posibles bases, A, C, G y T, el máximo contenido informacional de una columna es 2 bits.


A)
bits
Relative Entropy (18.9 bits)
2.2
1.9
1.7
1.5
1.3
1.1
0.9
0.6
0.4
0.2
0.0
Multilevel consensus sequence
AACTGGCACGGCACTTGCTTTA
TGT TACACAC AAAT
T ATTT
G
MOTIF WDHTGGCAYRVNNHTTGCTWWW MEME-1 width = 22 sites = 2157 llr = 28300 E-value = 4.4e-5532
Aa)
Simplified probability matrix
A 432::::a:433331:::3335
C 124:::a:5133332::a:211
G 231:aa:::43221::a::112
T 334a::::4113237a::6452
B)
Flavobacteriia
Gammaproteobacteria
Planctomycetia
Spirochaetia
Alphaproteobacteria
Chitinophagia
Aquificae
Clostridia
Bacilli
Cytophagia
Bacteroidia
Deltaproteobacteria
Betaproteobacteria
Epsilonproteobacteria
-24
-12
-24
-12

DISCUSIÓN

Es bien conocido que el inicio de la transcripción es el principal nivel de la regulación de la expresión genética en bacterias. La decisión de qué genes serán transcritos, puede mediarse por diferentes elementos como los factores transcripcionales, los *riboswitches* y los factores σ. Son estos últimos, los factores que proporcionan la especificidad a la RNA polimerasa para reconocer el sitio en donde iniciará la transcripción (Wösten y Wo 1998; Bervoets y Charlier 2019).

A pesar de que los factores σ y sus correspondientes secuencias consenso se han descrito en organismos modelo y para los factores σ principales, la creciente secuenciación de nuevos organismos y la definición de promotores para σ menos conocidos hacen importante la elaboración de protocolos computacionales para la identificación *in silico* de las secuencias de promotores de las diferentes σ bacterianas. En la presente tesis se desarrolló un protocolo computacional para la identificación de los promotores dependientes del factor σ54. Pese a que los genes transcritos por este factor σ han sido caracterizados en diferentes organismos modelo (Reitzer y Barbara L. Schneider, 2001; Cases *et al.*, 2003; Batchelor *et al.*, 2008; Lin *et al.*, 2014; Tsang y Hoover, 2014; Hocq *et al.*, 2019), la intención de estudiar al factor σ54 como modelo de estudio fue la de sentar las bases para el estudio de factores σ menos caracterizados e intentar ampliar el sigmulón de este factor σ en organismos poco caracterizados y alejados de los organismos más estudiados.

Nuestro trabajo partió de una primera hipótesis para definir el conjunto inicial de secuencias o secuencias semilla en un proceso cíclico de construcción de modelo/ búsqueda de hits (*pattern-finding/pattern-matching)*. Este proceso consideraba como tendencia general que los factores σ dirigen la transcripción de los genes que los codifican; es decir, se autotranscriben. Para nuestra sorpresa, el estudio detallado en grupos de organismos poco caracterizados del factor σ54, evidenció que dicha hipótesis sólo se cumple en 5.87% de los genes *rpoN* que se analizaron en este proyecto. Una posible explicación de este fenómeno es que la transcripción de estos genes se encuentre regulada por otros factores σ, posiblemente algún factor *housekeeping*, y

además posiblemente activada mediante factores transcripcionales en respuesta a las concentraciones intracelulares de alguna molécula relacionada con el nitrógeno. Ya que la regulación de la transcripción es un mecanismo complejo que integra señales internas y externas para la expresión de genes en las células, según requieran las condiciones ambientales y fisiológicas, es posible que la transcripción de este factor σ54 se encuentre finamente regulado. (Mauri y Klumpp, 2014)

Un segundo supuesto que resultó no ser siempre verdadero para la mayoría de los factores σ, fue que los motivos de secuencias conservadas en las regiones de regulación de genes que pertenecen a un mismo sigmulón (genes trascritos por el mismo factor σ), deberían de estar enriquecidos en las secuencias promotoras en común. Nuestra investigación reveló que en muchos casos existen regiones altamente conservadas, ya sea por motivos asociados a una regulación en común, o por cierto tipo de sesgo en secuencias repetidas con funciones aún no conocidas, o bien, por la presencia de secuencias con poco contenido informacional. Debido a lo anterior, en estos casos, el análisis de conservación de secuencias no fue útil para definir un motivo que representara el posible consenso de los promotores reconocidos por σ54. Para esos casos, la secuencia semilla para iniciar el ciclo de pattern-finding/pattern-matching se identificó con el consenso general de reconocimiento del factor σ54 en diferentes clases filogenéticas durante nuestras primeras pruebas, además de que se han identificado experimentalmente e *in silico,* diversos promotores dependientes de RpoN, y mediante alineamientos de dichas secuencias se ha obtenido su consenso, siendo Y**TGGCAC**G-NNNN-**TTGC**WNN (Barrios *et al.*, 1999).

Para comprobar nuestra segunda hipótesis de investigación, que versa que, *en grupos filogenéticamente cercanos pertenecientes a la misma clase, existe una alta conservación en la secuencia de los miembros de una misma familia de factores σ,* se obtuvieron los motivos para cada conjunto de estudio y se muestran los logos generados en la **figura 18** , esta hipótesis fue verificada al observar la similitud de los logos de los promotores consenso del factor σ54 en diferentes clases filogenéticas. También cabe resaltar que esta prevalencia de secuencia en organismos filogenéticamente cercanos, es notorio en aquellas clases que comparten un ancestro común temprano, como se

observa para las clases Aquificae y Epsilonbacteria, la región -24 del promotor, difieren en secuencia con el resto de las clases de estudio, además las Epsilonbacteria, que presenta menor grado de conservación, se encuentran más próximas a aquellas bacterias que han perdido el gen *rpoN.*

Durante el desarrollo de nuestra investigación se intentaron varios protocolos de enriquecimiento de la señal que lograron superar los problemas que resultaron por el limitado cumplimento de las dos hipótesis antes mencionadas. Nuestro protocolo modificado partió de motivos consenso generales para el factor σ54, que se fue refinando para cada grupo filogenético estudiado mediante un proceso cíclico y la elección de matrices en base a criterios de especificidad, sensibilidad y precisión de las predicciones. El proceso cíclico efectuado en nuestro estudio de *pattern-finding/pattern-matching* resultó ser mucho mejor que el proceso estándar en donde los resultados son obtenidos con un solo ciclo ya que nuestro alcance en precisión supera a los reportados en la literatura, también se disminuyó considerablemente el número de falsos positivos en nuestro análisis, que ha sido un problema importante en los métodos implementados hasta la fecha, que se ha tratado de evitar mejorando los conjuntos de datos de entrenamiento, *Hao Lin et al*. en el año 2014, implementó un método de predicción basados en evidencia experimental de 166 genes dependientes de σ54. Las tasas alcanzadas por su predictor fueron superiores al 90%, 97%, y 93%, en sensibilidad, especificidad, precisión, respectivamente, verificaron sus predicciones con los 6 genes depositados en aquel entonces en RegulónDB. Mas tarde *Liang et al.* crearon la base de datos almacenando 210 promotores σ54 con 297 genes regulados, así como sus productos en 43 especies diferentes (Liang *et al.*, 2017), basados en la herramienta llamada iPro54-PseKNC desarrollada por *Hao Lin et al*. (antes mencionada); sin embargo, estos métodos han sido implementados solamente en algunos organismos selectos, como lo son las enterobacterias y organismos modelo. La precisión de nuestro protocolo, corroborada contra los datos de RegulonDB, genes del sigmulón de σ54, fue de 94.7% y la especificidad de 99.9 %, o sea que, se logra clasificar adecuadamente la proporción de los verdaderos negativos y verdaderos positivos, además de que el

algoritmo desarrollado aquí es aplicable en la diversidad de organismos de las diferentes clases filogenéticas.

Teniendo en cuenta lo anterior, y considerando la alta conservación de las secuencias nucleotídicas reconocidas por el factor σ54 en diferentes clases filogenéticas, se consideró como secuencias semillas del ciclo inicial de análisis, aquellas secuencias reconocidas con el programa MAST y utilizando una matriz consenso obtenida de la comparación de las regiones de regulación de genes que codifican al factor σ54 en organismos de los diferentes grupos filogenéticos de nuestro estudio.

Con nuestra investigación hemos logrado tener una visión más íntegra del sigmulón σ54 y el posible origen evolutivo de este factor σ en bacterias. El estudio de enriquecimiento de genes regulados por el factor σ54 en las diferentes clases filogenéticas de estudio, indicó que los genes mayormente transcritos por este factor σ en bacterias corresponden a aquellos relacionados con el metabolismo del nitrógeno, tales como nitrogenasas, factores transcripcionales en respuesta al nitrógeno, proteínas fijadoras de nitrógeno y transportadores de nitrato, nitrito y amonio, urea-amidohidrolasas, al igual que el gen *rpoN* que codifica al factor σ54. Adicionalmente, también se observó un enriquecimiento en los genes relacionados con la biogénesis del flagelo. El análisis de enriquecimiento de los genes predichos por ser transcritos por el factor σ54 nos permitió definir los procesos metabólicos principalmente regulados por este factor σ, en todas las bacterias. Así como procesos celulares específicos de un clado filogenético particular. En **tabla 7**, se puede apreciar los grupos de genes ortólogos mayormente regulados dentro de clases Spirochaetia y Deltaproteobacteria tienden compartir los grupos de COGS: COG2710, COG0347, COG1348, COG0004, COG0535 y COG1842; los primeros cuatro, relacionados con el transporte de iones y metabolismo. Esto hace sentido en contexto filogenético, ya que los clados de estas clases son cercanos entre sí (ver árbol en **figura 17**). En el grupo 4 derivado de clusterizar los datos, de la tabla 7, se observa un enriquecimiento de los genes predichos de algunas chaperonas como GroES, Chaperona molecular clase DnaJ con dom de dedo de Zn C-terminal y GrpE, donde se encuentran reunidas las clases Planctomycetia, Aquificae, Cytophagia, Bacteroidia, Flavobacteriia, Chitinophagia, Spirochaetia y Épsilonbacteria.

Otras chaperonas que sólo se observan en el grupo 3, de las Alpha, Gamma y Betaproteobacterias, son las relacionadas con la biosíntesis del flagelo, no presentes en otras clases.

Otro aspecto que es importante mencionar, es la falta del factor σ54 en los grupos filogenéticos de Deinococci, Mollicutes, Halobacterias, Thermoprotei y Methanomicrobia, y la perdida parcial en organismos de las clases Fusobacteria, Coriobacteria y la mayoría de las actinobacterias, que abre la pregunta del origen de este factor σ en el grupo de los procariotes. Es importante mencionar que este factor σ54 no es homólogo al factor σ70 a pesar de que estos dos grupos de factores σ pueden compartir una gran similitud estructural, pero con mecanismos de acción completamente diferentes. Como se mencionó en la sección de introducción de esta tesis, el mecanismo de acción de σ54 difiere del de σ70 ya que una vez que RNAP-σ54 une al promotor se requiere la entrada de energía libre proveniente de la hidrólisis de ATP y de la unión de una proteína activadora que se une a gran distancia y río arriba del promotor, que se asemeja a las proteínas conocidas como “enhancers” en organismos eucariontes. Por lo tanto, el identificar el origen del factor σ54 es un tema de gran relevancia cuya investigación queda abierta para ser abordada (Buck *et al.*, 2000). Los resultados del análisis filogenético en el presente trabajo, ubican el origen del factor σ54 más parsimonioso, justo después de la división de los dominios arqueobacterias y bacterias, ya que como se sabe, la RNA polimerasa de las arqueobacterias no poseen factores σ (Karr, 2014), mientras que los grupos más cercanos a la raíz del árbol de bacterias, comprendido por los Planctomycetes claramente se puede observar que el 100% de los miembros analizados poseen al factor σ54.

CONCLUSIONES

1) La primera hipótesis de nuestro trabajo que consideraba que *los factores σ tienden dirigir la transcripción de los genes que los codifican*, resultó sólo ser cierta para un pequeño grupo de genes, ya que del total de genes *rpoN* estudiados, en el 94.13%, de los casos no encontramos evidencia de la auto-transcripción del factor σ54.

2) La segunda hipótesis de nuestra investigación, *en grupos filogenéticamente cercanos pertenecientes a la misma clase, existe una alta conservación en la secuencia de los miembros de una misma familia de factores σ,* fue verificada al observar la similitud de los logos de los promotores consenso del factor σ54 en diferentes clases filogenéticas, que se muestra en la **figura 18**.

3) Los motivos de secuencias estadísticamente más significativos dentro del conjunto de secuencias de las regiones de regulación de los genes que codifican para el factor σ54, no siempre corresponden al de las secuencias de los promotores σ54, por lo que fue necesario modificar la estrategia metodológica para identificar las secuencias semillas del ciclo inicial de análisis.

4) El proceso cíclico efectuado en nuestro estudio de *pattern-finding/pattern-matching* resultó ser mucho mejor que el proceso estándar en donde los resultados son obtenidos con un solo ciclo.

5) La comparación de nuestras predicciones de los genes transcritos por el factor σ54 en *Escherichia coli*, con lo reportado en la base de datos RegulonDB, indican que son altamente específicos (alto número de verdaderos negativos), aunque poco sensibles (bajo número de verdaderos positivos), con una precisión con un valor de 94.7% y una especificidad de 99.9%. Nuestro algoritmo se comporta como un predictor aceptable ya que logra clasificar adecuadamente la proporción tanto de los verdaderos negativos como de los verdaderos positivos, considerando las medidas de precisión, sensibilidad y especificidad que se obtuvieron.

6) La profundidad taxonómica originalmente contemplada en nuestro trabajo de investigación, al incluir organismos no redundantes a nivel de especie no resultó ser suficientemente lejana para evitar la redundancia en sus secuencias genómicas; es decir, no a esa profundidad taxonómica, las secuencias no han divergido lo suficiente para generar diferencias en sus secuencias que permitan identificar sitios conservados debido a la selección funcional de ser reconocidas por el factor σ de estudio. Considerando lo anterior, nuestra metodología se repitió una segunda vez considerando exclusivamente a secuencias de regulación provenientes de organismos representativos a nivel de género.

7) En todos los grupos, los motivos altamente conservados son GGCA (caja -24) y TGC (caja -12). En adición a estas bases, también se puede observar que, con menor frecuencia, también se conservan otras bases purinas, que anteceden a ambas cajas y otras bases, particulares de cada clase filogenética.

8) El estudio de enriquecimiento de genes regulados por el factor σ54 en las diferentes clases filogenéticas de estudio, indicó que los genes mayormente transcritos por este factor σ en bacterias corresponden a aquellos relacionados con el metabolismo del nitrógeno, tales como nitrogenasas, factores transcripcionales en respuesta al nitrógeno, proteínas fijadoras de nitrógeno y transportadores de nitrato, nitrito y amonio, urea-amidohidrolasas, al igual que el gen *rpoN* que codifica al factor σ54. Adicionalmente, también se observó un enriquecimiento en los genes relacionados con la biogénesis del flagelo en las clases Alpha, Beta y Gamma de las proteobacterias.

9) Los resultados de nuestro análisis filogenético ubican el origen del factor σ54 más parsimonioso, justo después de la división de los dominios arqueobacterias y bacterias, ya que como se sabe, la RNA polimerasa de las arqueobacterias no poseen factores σ, mientras que los grupos más cercanos a la raíz del árbol de bacterias, comprendido por los Planctomycetes claramente se puede observar que el 100% de los miembros analizados poseen al factor σ54.

10) En algunos grupos filogenéticos, el factor σ54 se ha perdido, total o parcialmente, es decir, en ninguno de los organismos representativos se observa la

presencia de σ54 o sólo alguno de los integrantes aún lo conservan. Ejemplos de pérdida total del factor σ54 se observan en las clases Deinococci del phylum Deinococcus y la clase Molicutes del phylum Tenericutes. Como ejemplo de la pérdida casi total o parcial del factor σ54 se puede mencionar a las clases Actinobacteria y Coriobacteria del phylum Actinobacteria, la clase Fusobacteria del phylum Fusobacteria, la clase Cloroflexia del phylum Cloroflexi y a la clase Tissierella del phylum Firmicutes.

11) Los parámetros definidos para los programas y los protocolos de análisis que permitieron identificar las secuencias consenso del sitio de reconocimiento del factor σ54 en bacterias, son una base fundamental para estudios similares contemplando otros factores σ.

PERSPECTIVAS

Como perspectiva a futuro, el protocolo que se desarrolló en este trabajo de investigación puede complementarse con datos derivados de los análisis de RNA-seq de los más de 2613 organismos para los cuales ya se cuenta con esta información, además de que, para algunos de estos, existen diferentes condiciones de crecimiento y variantes génicas (mutantes); lo que permitiría ampliar la predicción del sigmulón no solo de σ54 si no, de los diferentes factores σ. Los análisis globales de los transcriptomas públicamente disponibles proporcionarán información valiosa para poder estudiar los genes que cambian su expresión diferencialmente cuando exista datos del transcriptoma de una mutante del factor σ a estudiar y los genes que aumentan su transcripción bajo condiciones de crecimiento particulares. Esto permitiría mejorar la precisión y sensibilidad de las predicciones. Asimismo, se pretende combinar lo que se sabe de las propiedades estructurales como la curvatura y la flexibilidad, de las regiones promotoras, para mejorar su identificación en los genomas, no sólo para los genes dependientes del factor σ54, si no para los genes regulados por otros factores σ. Haciendo uso del programa ya implementado basado en estas observaciones, que desarrollaron Rangannan y Bansal en 2009, llamado: "*PromPredict*", para la identificación de promotores microbianos (Rangannan y Bansal, 2009) y permite calcular la energía de

desestabilización al rededor de una potencial región promotora, dándole más peso a las predicciones reportadas por el algoritmo aquí descrito.

Con la experiencia derivada del presente trabajo, se espera poder sortear los retos que se vayan presentando en la mejora del algoritmo.

Los puntos que se deben considerar para mejorar las predicciones de promotores reconocidos por los distintos factores σ serían los siguientes:

- Análisis global de los transcriptomas disponibles. Para estudiar los genes que cambian su expresión diferencialmente cuando existe una mutación de algún σ particular que es no esencial. Esto proporcionaría información de qué otros genes aumentan su transcripción concomitantemente ante la mutación (sea ausencia, activación, atenuación etc.) de un factor σ particular; lo cuál indicaría que estos genes se encuentran potencialmente regulados por este σ específicamente o de forma antagónica con otro factor σ , y que se están transcribiendo bajo determinada condición de crecimiento (cuando exista el reporte). Para los casos que no exista información de las condiciones experimentales de crecimiento, se podría observar el cambio de expresión del gen que codifica para el factor σ e inferir su posible participación en la regulación. Y al contrario cuando no exista un mutante de factores σ, pero sí de las condiciones de crecimiento, se asociaría el nivel de expresión tanto de los genes regulados como el/los genes que pertenezcan a factores σ a esta.

- El análisis de las propiedades estructurales de las regiones promotoras, mediante el uso *PromPredict*", de tales como: menor estabilidad, mayor curvatura, menor capacidad de flexión en comparación con sus regiones vecinas o localizada en las proximidades (Kanhere, 2005; Wang y Benham, 2007; Kumar *et al.*, 2016; Kumar y Bansal, 2017). Debido al requerimiento de la apertura de la doble cadena de DNA en el proceso del inicio de la transcripción, la región del DNA que contiene un promotor es, en términos generales, menos estable y, por lo tanto, más propensa a la fusión, en comparación con otras regiones genómicas. El

cálculo de la estabilidad relativa de la molécula de DNA se puede expresar en términos de energía libre y ha sido empleado para predecir exitosamente los promotores de organismos bacterianos.

En cuanto al análisis de los datos obtenidos para las 14 clases de estudio presentadas en este trabajo, es preciso analizar cada conjunto de forma particular para poder examinar en detalle cada una de las redes de regulación de la transcripción regida por el factor σ54. Así como asociar la función biológica de los diversos genes predichos como probable actor de la orquesta dirigida por el factor σ54.

Como última perspectiva, aunque no por ello, menos importante, la continuación de los resultados obtenidos durante el desarrollo de esta tesis, podrían ser de gran relevancia para abordar aspectos evolutivos sobre el origen evolutivo de este importante factor σ54 y sus correspondientes redes de regulación en diferentes grupos filogenéticos.

BIBLIOGRAFÍA

Ahmar, R. Al, Kirby, B. D. y Yu, H. D. (2018) “Pyrimidine Biosynthesis Regulates the Small-Colony Variant and Mucoidy in Pseudomonas aeruginosa through Sigma Factor Competition”, *Journal of Bacteriology*, 201(1), p. JB.00575-18. doi: 10.1128/jb.00575-18 PMID - 30322853.

Allawi, H. T. y SantaLucia, J. (1997) “Thermodynamics and NMR of Internal G·T Mismatches in DNA”, *Biochemistry*, 36(34), pp. 10581–10594. doi: 10.1021/bi962590c PMID - 9265640.

Altschul, S. F. *et al.* (1990) “Basic local alignment search tool”, *Journal of Molecular Biology*, 215(3), pp. 403–410. doi: 10.1016/S0022-2836(05)80360-2.

Amin, R. *et al.* (2019) “iPromoter-BnCNN: a Novel Branched CNN Based Predictor for Identifying and Classifying Sigma Promoters. (arXiv:1912.10251v1 [q-bio.QM])”, (2018), pp. 1–7. Disponible en: http://arxiv.org/abs/1912.10251.

Askary, A. *et al.* (2009) “N4: A precise and highly sensitive promoter predictor using neural network fed by nearest neighbors”, *Genes & Genetic Systems*, 84(6), pp. 425–430. doi: 10.1266/ggs.84.425.

de Avila e Silva, S., Echeverrigaray, S. y Gerhardt, G. J. L. (2011) “BacPP: Bacterial promoter prediction—A tool for accurate sigma-factor specific assignment in enterobacteria”, *Journal of Theoretical Biology*, 287, pp. 92–99. doi: 10.1016/j.jtbi.2011.07.017.

Bailey, T. L. (1994) “FITTING A MIXTURE MODEL BY EXPECTATION MAXIMIZATION TO DISCOVER MOTIFS IN BIOPOLYMERS UCSD Technical Report CS94-351”.

Bailey, T. L. *et al.* (1998) “Combining evidence using p-values : application to sequence homology searches”, 14(1), pp. 48–54.

Bailey, T. L. (2003) “Discovering Novel Sequence Motifs with MEME”, *Current Protocols in Bioinformatics*, 00(1), pp. 2.4.1-2.4.35. doi: 10.1002/0471250953.bi0204s00.

Bailey, T. L. *et al.* (2006) “MEME: Discovering and analyzing DNA and protein sequence motifs”, *Nucleic Acids Research*, 34(WEB. SERV. ISS.), pp. 369–373. doi: 10.1093/nar/gkl198.

Barrios, H., Valderrama, B. y Morett, E. (1999) “Compilation and analysis of σ54-dependent promoter sequences”, *Nucleic Acids Research*, 27(22), pp. 4305–4313. doi: 10.1093/nar/27.22.4305 PMID - 10536136.

Batchelor, J. D. *et al.* (2008) “Structure and Regulatory Mechanism of Aquifex aeolicus NtrC4: Variability and Evolution in Bacterial Transcriptional Regulation”, *Journal of Molecular Biology*, 384(5), pp. 1058–1075. doi: 10.1016/j.jmb.2008.10.024.

Battistuzzi, F. U., Feijao, A. y Hedges, S. B. (2004) “A genomic timescale of prokaryote evolution: Insights into the origin of methanogenesis, phototrophy, and the colonization

of land", *BMC Evolutionary Biology*, 4, pp. 1–14. doi: 10.1186/1471-2148-4-44.

Battistuzzi, F. U. y Hedges, S. B. (2009) "A Major Clade of Prokaryotes with Ancient Adaptations to Life on Land", *Molecular Biology and Evolution*, 26(2), pp. 335–343. doi: 10.1093/molbev/msn247.

Bentley, S. D. *et al.* (2002) "Complete genome sequence of the model actinomycete Streptomyces coelicolor A3(2)", *Nature*, 417(6885), pp. 141–147. doi: 10.1038/417141a PMID - 12000953.

Bervoets, I. y Charlier, D. (2019) "Diversity, versatility and complexity of bacterial gene regulation mechanisms: opportunities and drawbacks for applications in synthetic biology", *FEMS Microbiology Reviews*, 43(3), pp. fuz001-. doi: 10.1093/femsre/fuz001 PMID - 30721976.

Bharanikumar, R., Premkumar, K. A. R. y Palaniappan, A. (2018) "PromoterPredict: sequence-based modelling of Escherichia coli σ 70 promoter strength yields logarithmic dependence between promoter strength and sequence", *PeerJ*, 6, p. e5862. doi: 10.7717/peerj.5862.

Borukhov, S. y Nudler, E. (2003) "RNA polymerase holoenzyme: Structure, function and biological implications", *Current Opinion in Microbiology*, 6(2), pp. 93–100. doi: 10.1016/S1369-5274(03)00036-5.

Buck, M. *et al.* (2000) "The bacterial enhancer-dependent σ54 (σ(N)) transcription factor", *Journal of Bacteriology*, 182(15), pp. 4129–4136. doi: 10.1128/JB.182.15.4129-4136.2000.

Buck, M. y Cannon, W. (1992a) "Specific binding of the transcription factor sigma-54 to promoter DNA.", *Nature*, 358(6385), pp. 422–4. doi: 10.1038/358422a0.

Buck, M. y Cannon, W. (1992b) "Specific binding of the transcription factor sigma-54 to promoter DNA", *Nature*, 358(6385), pp. 422–422. doi: 10.1038/358422a0.

Burgess, R R (2001) "Encyclopedia of Genetics", *Article Titles: S*, (Nature2211969), pp. 1831–1834. doi: 10.1006/rwgn.2001.1192.

Burgess, R.R. (2001) "Sigma Factors", en *Encyclopedia of Genetics*. Elsevier, pp. 1831–1834. doi: 10.1006/rwgn.2001.1192.

BURGESS, R. R. *et al.* (1969) "Factor Stimulating Transcription by RNA Polymerase", *Nature*, 221(5175), pp. 43–46. doi: 10.1038/221043a0 PMID - 4882047.

Byvatov, E. y Schneider, G. (2003) "Support vector machine applications in bioinformatics.", *Applied bioinformatics*, 2(2), pp. 67–77. Disponible en: http://www.ncbi.nlm.nih.gov/pubmed/15130823.

Campbell, E. A. *et al.* (2002) "Structure of the Bacterial RNA Polymerase Promoter Specificity σ Subunit", *Molecular Cell*, 9(3), pp. 527–539. doi: 10.1016/s1097-2765(02)00470-7 PMID - 11931761.

Cases, I., Ussery, D. W. y De Lorenzo, V. (2003) “The σ 54 regulon (sigmulon) of Pseudomonas putida”, *Environmental Microbiology*, 5(12), pp. 1281–1293. doi: 10.1111/j.1462-2920.2003.00528.x.

Coelho, R. V. *et al.* (2018) “Bacillus subtilis promoter sequences data set for promoter prediction in Gram-positive bacteria”, *Data in Brief*, 19, pp. 264–270. doi: 10.1016/j.dib.2018.05.025.

Cook, H. y Ussery, D. W. (2013) “Sigma factors in a thousand E. coli genomes: Genomics update”, *Environmental Microbiology*, 15(12), pp. 3121–3129. doi: 10.1111/1462-2920.12236 PMID - 23992563.

Davis, M. C. *et al.* (2017) “The essential activities of the bacterial sigma factor”, *Canadian Journal of Microbiology*, 63(2), pp. 89–99. doi: 10.1139/cjm-2016-0576 PMID - 28117604.

Do, C. B. y Batzoglou, S. (2008) “What is the expectation maximization algorithm?”, *Nature Biotechnology*, 26(8), pp. 897–899. doi: 10.1038/nbt1406.

Dong, T., Yu, R. y Schellhorn, H. (2011) “Antagonistic regulation of motility and transcriptome expression by RpoN and RpoS in Escherichia coli”, *Molecular Microbiology*, 79(2), pp. 375–386. doi: 10.1111/j.1365-2958.2010.07449.x.

Eddy, S. R. (2004) “What is a hidden Markov model?”, *Nature Biotechnology*, 22(10), pp. 1315–1316. doi: 10.1038/nbt1004-1315.

Feklistov, A. (2013) “RNA polymerase: in search of promoters”, *Annals of the New York Academy of Sciences*, 1293(1), pp. 25–32. doi: 10.1111/nyas.12197.

Fitch, W. M. (1970) “Distinguishing Homologous from Analogous Proteins”, *Systematic Zoology*, 19(2), p. 99. doi: 10.2307/2412448.

Francke, C. *et al.* (2011) “Comparative analyses imply that the enigmatic sigma factor 54 is a central controller of the bacterial exterior”, *BMC Genomics*, 12(1), p. 385. doi: 10.1186/1471-2164-12-385 PMID - 21806785.

Freyre-González, J. A. *et al.* (2013) “Lessons from the modular organization of the transcriptional regulatory network of Bacillus subtilis”, *BMC Systems Biology*, 7(1), p. 127. doi: 10.1186/1752-0509-7-127 PMID - 24237659.

Gama-Castro, S. *et al.* (2011) “RegulonDB version 7.0: transcriptional regulation of Escherichia coli K-12 integrated within genetic sensory response units (Gensor Units)”, *Nucleic Acids Research*, 39(Database), pp. D98–D105. doi: 10.1093/nar/gkq1110.

Haldenwang, W. G. (1995) “The sigma factors of Bacillus subtilis.”, *Microbiological reviews*, 59(1), pp. 1–30. doi: 10.1128/MMBR.59.1.1-30.1995.

Helmann, J. D. (2019) “Where to begin? Sigma factors and the selectivity of transcription initiation in bacteria”, *Molecular Microbiology*, 112(2), pp. 335–347. doi: 10.1111/mmi.14309.

Hocq, R. *et al.* (2019) “σ 54 (σ L) plays a central role in carbon metabolism in the

industrially relevant Clostridium beijerinckii", *Scientific Reports*, 9(1), pp. 1–13. doi: 10.1038/s41598-019-43822-2.

Hoover, T. R. *et al.* (1990) "The integration host factor stimulates interaction of RNA polymerase with NIFA, the transcriptional activator for nitrogen fixation operons", *Cell*, 63(1), pp. 11–22. doi: 10.1016/0092-8674(90)90284-L.

Jacob, F. y Monod, J. (1961) "Genetic regulatory mechanisms in the synthesis of proteins", *Journal of Molecular Biology*, 3(3), pp. 318–356. doi: 10.1016/S0022-2836(61)80072-7.

K., K. *et al.* (2016) "Effective Feature Selection for Classification of Promoter Sequences", *PLOS ONE*, 11(12), p. e0167165. doi: 10.1371/journal.pone.0167165 PMID - 27978541.

Kanhere, A. (2005) "Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes", *Nucleic Acids Research*, 33(10), pp. 3165–3175. doi: 10.1093/nar/gki627.

Kanhere, A. y Bansal, M. (2005) "A novel method for prokaryotic promoter prediction based on DNA stability", *BMC Bioinformatics*, 6(1), p. 1. doi: 10.1186/1471-2105-6-1 PMID - 15631638.

Karr, E. A. (2014) "Transcription Regulation in the Third Domain", en *Advances in Applied Microbiology*, pp. 101–133. doi: 10.1016/B978-0-12-800259-9.00003-2.

Koonin, E. V. (2005) "Orthologs, Paralogs, and Evolutionary Genomics", *Annual Review of Genetics*, 39(1), pp. 309–338. doi: 10.1146/annurev.genet.39.073003.114725.

Kumar, A. y Bansal, M. (2017) "Unveiling DNA structural features of promoters associated with various types of TSSs in prokaryotic transcriptomes and their role in gene expression", *DNA Research*, 24(1), pp. 25–35. doi: 10.1093/dnares/dsw045 PMID - 27803028.

Kumar, A., Manivelan, V. y Bansal, M. (2016) "Structural features of DNA are conserved in the promoter region of orthologous genes across different strains of Helicobacter pylori", *FEMS Microbiology Letters*. Editado por A. van Vliet, 363(18), p. fnw207. doi: 10.1093/femsle/fnw207.

Liang, Z. Y. *et al.* (2017) "Pro54DB: a database for experimentally verified sigma-54 promoters", *Bioinformatics (Oxford, England)*, 33(3), pp. 467–469. doi: 10.1093/bioinformatics/btw630.

Lin, H. *et al.* (2014) "IPro54-PseKNC: A sequence-based predictor for identifying sigma-54 promoters in prokaryote with pseudo k-tuple nucleotide composition", *Nucleic Acids Research*, 42(21), pp. 12961–12972. doi: 10.1093/nar/gku1019.

Lobanovska, M., Tang, C. M. y Exley, R. M. (2019) "crossm pilE in Neisseria meningitidis", 201(20), pp. 1–16.

Losick, R., Youngman, P. y Piggot, P. J. (1986) "Genetics of Endospore Formation in Bacillus Subtilis", *Annual Review of Genetics*, 20(1), pp. 625–669. doi:

10.1146/annurev.ge.20.120186.003205.

MacQueen, J. (1967) “Some methods for classification and analysis of multivariate observations”, en *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*. Berkeley, Calif.: University of California Press (Fifth Berkeley Symposium on Mathematical Statistics and Probability), pp. 281–297. Disponible en: https://projecteuclid.org/euclid.bsmsp/1200512992.

Martinez-Guerrero, C. E. *et al.* (2008) “GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways”, *Nucleic Acids Research*, 36(suppl_2), pp. W176–W180. doi: 10.1093/nar/gkn330.

Mauri, M. y Klumpp, S. (2014) “A Model for Sigma Factor Competition in Bacterial Cells”, *PLoS Computational Biology*, 10(10), pp. 29–34. doi: 10.1371/journal.pcbi.1003845.

Mishra *et al.* (2018) “Towards a universal structural and energetic model for prokaryotic promoters”, *Biophysical Journal*, 115(7), pp. 1180–1189. doi: 10.1016/j.bpj.2018.08.002 PMID - 30172386.

Morett, E. y Segovia, L. (1993) “The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains.”, *Journal of Bacteriology*, 175(19), pp. 6067–6074. doi: 10.1128/JB.175.19.6067-6074.1993.

Mrazek, J. (2009) “Finding sequence motifs in prokaryotic genomes--a brief practical guide for a microbiologist”, *Briefings in Bioinformatics*, 10(5), pp. 525–536. doi: 10.1093/bib/bbp032.

Mulligan, M. E. *et al.* (1984) “Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity”, *Nucleic Acids Research*, 12(1Part2), pp. 789–800. doi: 10.1093/nar/12.1part2.789 PMID - 6364042.

Novichkov, P. S. *et al.* (2010) “RegPredict: an integrated system for regulon inference in prokaryotes by comparative genomics approach”, *Nucleic Acids Research*, 38(Web Server), pp. W299–W307. doi: 10.1093/nar/gkq531.

Osmundson, J., Dewell, S. y Darst, S. A. (2013) “RNA-Seq Reveals Differential Gene Expression in Staphylococcus aureus with Single-Nucleotide Resolution”, 8(10), pp. 1–12. doi: 10.1371/journal.pone.0076572.

Overbeek, R. *et al.* (1999) “The use of gene clusters to infer functional coupling”, *Proceedings of the National Academy of Sciences*, 96(6), pp. 2896–2901. doi: 10.1073/pnas.96.6.2896.

Paget, M. (2015) “Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution”, *Biomolecules*, 5(3), pp. 1245–1265. doi: 10.3390/biom5031245.

van Passel, M. W. J., Nijveen, H. y Wahl, L. M. (2014) “Birth, Death, and Diversification of Mobile Promoters in Prokaryotes”, *Genetics*, 197(1), pp. 291–299. doi: 10.1534/genetics.114.162883.

Pátek, M. y Nešvera, J. (2011) “Sigma factors and promoters in Corynebacterium

glutamicum", *Journal of Biotechnology*, 154(2–3), pp. 101–113. doi: 10.1016/j.jbiotec.2011.01.017.

Pérez-Rueda, E., Janga, S. C. y Martínez-Antonio, A. (2009) "Scaling relationship in the gene content of transcriptional machinery in bacteria", *Molecular BioSystems*, 5(12), p. 1494. doi: 10.1039/b907384a.

Pisani, D., Cotton, J. A. y McInerney, J. O. (2007) "Supertrees Disentangle the Chimerical Origin of Eukaryotic Genomes", *Molecular Biology and Evolution*, 24(8), pp. 1752–1760. doi: 10.1093/molbev/msm095.

Rangannan, V. y Bansal, M. (2009) "Relative stability of DNA as a generic criterion for promoter prediction: whole genome annotation of microbial genomes with varying nucleotide base composition", *Molecular BioSystems*, 5(12), pp. 1758–1769. doi: 10.1039/b906535k PMID - 19593472.

Rangannan, V. y Bansal, M. (2010) "High-quality annotation of promoter regions for 913 bacterial genomes", *Bioinformatics*, 26(24), pp. 3043–3050. doi: 10.1093/bioinformatics/btq577.

Reitzer, L. y Schneider, Barbara L (2001) "Metabolic Context and Possible Physiological Themes of sigma 54 -Dependent Genes in Escherichia coli Downloaded from http://mmbr.asm.org/ on February 19 , 2018 by Danish Veterinary and Agricultural Library", *Microbiology and Molecular Biology Reviews*, 65(3), pp. 422–444. doi: 10.1128/MMBR.65.3.422.

Reitzer, L. y Schneider, Barbara L. (2001) "Metabolic Context and Possible Physiological Themes of ς54-Dependent Genes in Escherichia coli", *Microbiology and Molecular Biology Reviews*, 65(3), pp. 422–444. doi: 10.1128/MMBR.65.3.422-444.2001.

Saeys, Y., Inza, I. y Larranaga, P. (2007) "A review of feature selection techniques in bioinformatics", *Bioinformatics*, 23(19), pp. 2507–2517. doi: 10.1093/bioinformatics/btm344.

Sallet, E. *et al.* (2013) "Next-Generation Annotation of Prokaryotic Genomes with EuGene-P: Application to Sinorhizobium meliloti 2011", *DNA Research*, 20(4), pp. 339–354. doi: 10.1093/dnares/dst014.

Sampaio, M. *et al.* (2020) "Predicting Promoters in Phage Genomes Using Machine Learning Models", en *Advances in Intelligent Systems and Computing*, pp. 105–112. doi: 10.1007/978-3-030-23873-5_13.

Studholme, D. (2000) "The biology of enhancer-dependent transcriptional regulation in bacteria: insights from genome sequences", *FEMS Microbiology Letters*, 186(1), pp. 1–9. doi: 10.1016/S0378-1097(00)00082-3.

Svergun, D. I. *et al.* (2000) "Low resolution structure of the sigma54 transcription factor revealed by X-ray solution scattering.", *The Journal of biological chemistry*, 275(6), pp. 4210–4. doi: 10.1074/jbc.275.6.4210.

Tatusov, R. L. (2000) "The COG database: a tool for genome-scale analysis of protein

functions and evolution", *Nucleic Acids Research*, 28(1), pp. 33–36. doi: 10.1093/nar/28.1.33.

Tintut, Y., Wang, J. T. y Gralla, J. D. (1995) "A novel bacterial transcription cycle involving sigma 54.", *Genes & Development*, 9(18), pp. 2305–2313. doi: 10.1101/gad.9.18.2305.

Torres-Puig, S. *et al.* (2015) "A novel sigma factor reveals a unique regulon controlling cell-specific recombination in Mycoplasma genitalium", *Nucleic Acids Research*, 43(10), pp. 4923–4936. doi: 10.1093/nar/gkv422.

Tsang, J. y Hoover, T. R. (2014) "Themes and Variations: Regulation of RpoN-Dependent Flagellar Genes across Diverse Bacterial Species", *Scientifica*, 2014, pp. 1–14. doi: 10.1155/2014/681754.

Tsoy, O. V *et al.* (2016) "Nitrogen Fixation and Molecular Oxygen: Comparative Genomic Reconstruction of Transcription Regulation in Alphaproteobacteria", *Frontiers in Microbiology*, 7, p. 1343. doi: 10.3389/fmicb.2016.01343 PMID - 27617010.

Wang, H. y Benham, C. J. (2007) "Annotation of Promoter Regions in Microbial Genomes Based on DNA Structural and Sequence Properties", en *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, pp. 212–224. doi: 10.1007/978-3-540-48540-7_18.

Wang, X., Chen, J. y Quinn, P. (eds.) (2012) *Reprogramming Microbial Metabolic Pathways*. Dordrecht: Springer Netherlands (Subcellular Biochemistry). doi: 10.1007/978-94-007-5055-5.

Wang, Z., Jensen, M. A. y Zenklusen, J. C. (2016) "A Practical Guide to The Cancer Genome Atlas (TCGA)", en Mathé, E. y Davis, S. (eds.) *Methods in molecular biology (Clifton, N.J.)*. New York, NY: Springer New York (Methods in Molecular Biology), pp. 111–141. doi: 10.1007/978-1-4939-3578-9_6.

Weiss, D. S. *et al.* (1991) "The phosphorylated form of the enhancer-binding protein NTRC has an ATPase activity that is essential for activation of transcription", *Cell*, 67(1), pp. 155–167. doi: 10.1016/0092-8674(91)90579-N.

Wösten, M. M. S. M. y Wo, M. M. S. M. (1998) "Eubacterial sigma-factors", *FEMS Microbiology Reviews*, 22(3), pp. 127–150. doi: 10.1111/j.1574-6976.1998.tb00364.x PMID - 9818380.