UNIVERSIDAD AUTÓNOMA DEL ESTADO DE MORELOS

INSTITUTO DE INVESTIGACIÓN EN CIENCIAS BÁSICAS Y APLICADAS

CENTRO DE INVESTIGACIÓN EN CIENCIAS

"IDENTIFICACIÓN IN SILICO DE ELEMENTOS DE REGULACIÓN EN SECUENCIAS DE ORGANISMOS BACTERIANOS"

TESIS QUE PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS

PRESENTA
KAREL JOHAN ESTRADA GUERRA

TUTOR
DR. ENRIQUE MERINO PÉREZ

MIEMBROS DEL COMITÉ TUTOR

DRA. SONIA DÁVILA RAMOS (tutor de seguimiento)

DR. ARMANDO HERNÁNDEZ MENDOZA

DR. ALEJANDRO GARCIARRUBIO GRANADOS

CUERNAVACA, MORELOS Noviembre 2017

A mis tres emes.

AGRADECIMIENTOS

A Enrique Merino, por tu amistad, por tu alegría, por tu apoyo, por tu ayuda, por tu paciencia, por todo lo que me enseñas y por empujarme y hacerme crecer, sin ti no existiría este trabajo tío Lucas. 10000000000000000 gracias.

A mi familia.

A Mabel Rodríguez, por viajar conmigo, por tu amor.

A mis padres Carlos y Ana, por regalarme la vida y el ejemplo.

A mi hermano Carlos Estrada, por siempre estar, por tu luz.

A Alejandro Garciarrubio, por tantos años de enseñanza y amistad, por tu sabiduría, por darme la oportunidad de conocer el mundo de las 4 letras.

A Sonia Dávila y Armando Hernández por acompañarme estos dos años. Muchas gracias por el apoyo y los consejos.

A todos mis compañeros de trabajo, a Alejandro Sanchéz por darme la oportunidad de superarme y por la paciencia, a Alejandro Abdala y Tina Godoy por ayudarme a sufrir menos con bases de datos y R, a Verónica Jiménez, Leticia Vega, Alejandra Escobar, Arturo Pimentel y Jerome Verleyen por todo el apoyo.

NOMENCLATURA

SD - Shine-Dalgarno

anti-SD - Anti Shine-Dalgarno

mRNA - RNA mensajero

pb - Pares de bases

NCBI - National Center for Biotechnology Information

RBS - Sitio de unión al ribosoma

MAST - Motif Alignment and Search Tool

MEME - Multiple Expectation Maximitation for Motif Elicitation

CDS - Secuencia codificante

tRNA - RNA de transferencia

UTR - Región no traducida del RNA mensajero

ΔG - Cambio en la energía libre de Gibbs

RPS1 - Proteína ribosomal S1

CAI – índice del uso de codones

EM - Esperanza-maximización

siRNAs - RNA pequeño de interferencia

miRNA - MicroRNA

IF - Factores de iniciación

RESUMEN

Alrededor del codón de inicio de un mRNA se han identificado diferentes elementos que influyen en el proceso del inicio de la traducción, como son: el tipo de codones que existen río abajo del mismo, la estructura secundaria del mRNA, la presencia de la proteína ribosomal S1 y la ausencia de la región 5' UTR, pero sin duda alguna, la secuencia Shine-Dalgarno (SD), es el elemento más importante que define la frecuencia con la que el mRNA será traducido.

En el presente trabajo, abordamos el comportamiento de estos elementos de regulación de la traducción a través de estudios *in silico* partiendo de las secuencias de 12,112 genomas bacterianos reportadas en la base de datos Refseq (NCBI) (ftp://ftp.ncbi.nlm.nih.gov/genomes/refseq/bacteria/) a junio de 2017. Hicimos un énfasis particular en la predicción y anotación del gen ribosomal 16S y la identificación cualitativa y cuantitativa de las secuencias Shine-Dalgarno presente en cada genoma. Estudiamos la distribución filogenética de aquellos genomas que carecen o poseen un Shine-Dalgarno con secuencia atípica, y evaluamos la correlación entre el contenido de GC genómico y el cambio en la energía libre de Gibbs (ΔG) con la que la secuencia anti-SD en el extremo 3' del 16S ribosomal se unirá, en promedio, con el sitio de unión al ribosoma (RBS) de los diferentes mRNAs. Además caracterizamos la calidad de los RBS que existe entre el primer y segundo cistrón de los operones, y genes cuyos RBS se encuentran localizados en la secuencia codificante del cistrón ubicado río arriba o en una región intergénica menor a 15 bases. Los resultados generados en el presente estudio ofrecen una visión global y actualizada de los distintos elementos que intervienen en el inicio de uno de los procesos más importantes de los organismos bacterianos: la traducción de los mRNAs.

1. INTRODUCCIÓN

La traducción es el proceso biológico por medio del cual se realiza la síntesis de proteínas. En las células eucariotas la transcripción se realiza en el núcleo y la traducción en el citoplasma; mientras que en las procariotas ambos procesos se realizan en el citoplasma de manera acoplada.

La traducción se lleva a cabo en los ribosomas, los cuales son descritos en función de su velocidad de sedimentación (medida en Svedbergs). Así nos encontramos que la sedimentación de los ribosomas bacterianos es aproximadamente 70S constando de dos subunidades, la pequeña de 30S y una mayor de 50S.

Una proteína es ensamblada secuencialmente por la adición de aminoácidos en la dirección N-terminal al C-terminal y este proceso involucrará tres etapas: iniciación, elongación y terminación. La etapa de la iniciación comienza con la formación de un complejo formado por la subunidad menor del ribosoma, un tRNA con el anticodón UAC, el aminoácido metionina químicamente modificado y la ayuda de tres factores de iniciación (IF1, IF2 y IF3). Es esta etapa de inicio de la traducción la que mayormente regula la síntesis de proteínas. Comúnmente la traducción en organismos procariontes es iniciada por la interacción entre la secuencia Shine-Dalgarno en el extremo 5' UTR de un mRNA y la secuencia de anti-SD en el extremo 3' del 16S rRNA que se conserva de una manera muy significativa en el 99% de los organismos procariontes (Fig. 1 y 2).

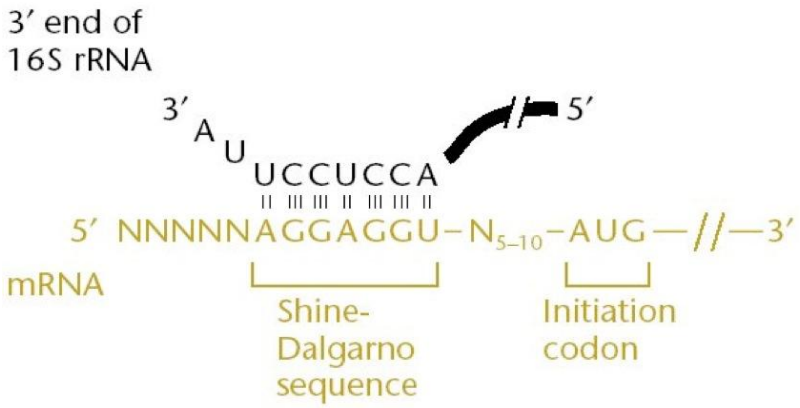

Fig. 1. Extremo 3’ de 16S rRNA hibridando con la secuencia Shine-Dalgarno del mRNA.

Fig. 2. Logo de la conservación de la secuencia del extremo 3' del 16S en organismos procariontes y arqueobacterias. El contenido informacional de la conservación está determinado en bits, siendo el valor de 2, el máximo posible de una posición 100% conservada. Nakagawa et al. 2010

En bacterias un mensajero puede codificar varios genes con múltiples codones de inicio (generalmente AUG). ¿Cómo identifica el ribosoma dónde unirse para comenzar la síntesis de proteínas? Para tratar de responder esta pregunta, varios grupos de investigación secuenciaron pequeñas secciones de mRNA (alrededor de 20 nucleótidos de largo) cerca de varios codones de inicio AUGs. John Shine, un científico que ya tenía en cuenta que la firma CCUCC en la cola de los rRNAs 16S podría estar relacionado con el proceso de iniciación [47] buscó el complemento de CCUCC, -GGAGG- en las regiones 5' UTR de los mensajeros. Su hallazgo lo conocemos hoy como la secuencia "Shine-Dalgarno", que determina el inicio de la traducción en la mayoría de los genes de Bacterias y Archaeas.

La secuencia Shine-Dalgarno se encuentra generalmente a 8 bases río arriba del codón de inicio de una secuencia codificante (CDS) y se complementa en la cola del 16S ribosomal de la subunidad ribosomal pequeña 30S. En bacterias, la interacción entre el Shine-Dalgarno y las secuencias anti-SD facilitan la formación del pre-complejo de iniciación que comprende una subunidad 30S y tres factores de iniciación (IF1, IF2 e IF3) alrededor del codón de inicio del mRNA [27]. La importancia de esta interacción para la iniciación eficiente de la traducción ha sido verificada experimentalmente, tanto para eubacterias, como para arqueobacterias, y se conoce que existe una relación significativa entre la similitud de las secuencias Shine-Dalgarno de un gen con el anti-SD y la eficiencia de inicio de la traducción del mismo. Aunque se asume que la

secuencia Shine-Dalgarno está conservada entre organismos bacterianos, diferentes estudios genómicos han demostrado que el grado en que los mRNAs de un organismo contienen una secuencia Shine-Dalgarno consenso, varía significativamente. En promedio, sólo el 77.0 % de genes presentan Shine-Dalgarno [29]. Está reportado que un Shine-Dalgarno de mayor longitud no produce necesariamente un aumento de la traducción, esto es debido a posibles estructuras secundarias adicionales o a que una complementariedad muy fuerte hará mayor el tiempo de despeje del sitio de unión y evitará la formación de los polisomas [26].

Para predecir la hibridación entre dos moléculas de RNAs, normalmente se utiliza el cálculo del cambio en la energía libre de Gibbs que ésta produce (ΔG, en kilocalorías por mol). De esta manera se pueden identificar y cuantificar, por ejemplo, blancos para secuencias siRNAs y miRNAs, los posibles sitios de unión de estructura secundarias para una sub-secuencia dentro de una mRNA o la secuencia Shine-Dalgarno favorable si evaluamos el ΔG entre la secuencia anti-SD y las 15 bases antes del codón de inicio de un gen, que es la región del mRNA en donde está comúnmente la secuencia Shine-Dalgarno. Los valores de ΔG para un Shine-Dalgarno funcional se ubican entre -3.5kcal/mol y -4.4kcal/mol [19], clasificándose entonces como débil a un Shine-Dalgarno por encima de este rango, o como fuerte si este valor está por debajo.

1.1 Elementos adicionales al Shine-Dalgarno que afectan la frecuencia del inicio de la traducción

1.1.1 Estructuras secundaria del RNA alrededor del Shine-Dalgarno

Adicional a la secuencia Shine-Dalgarno, existen otros factores que pueden afectar la eficiencia del inicio de traducción de mRNAs bacterianos. Uno de ellos es la presencia de estructuras secundarias alrededor del Shine-Dalgarno. Los mRNAs sin una secuencia Shine-Dalgarno son generalmente menos estructurados en su región de iniciación de la traducción y muestran un mínimo de plegamiento alrededor del codón de inicio. Se conoce que la estructura en las regiones UTRs sirven para modular la accesibilidad de la subunidad ribosomal 30S y con esto favorecer el inicio de la traducción. Los elementos de estructura secundaria del mRNA poseen un mayor

efecto inhibitorio en la traducción cuando en ellos se incluyen o se encuentran cercanas al codón de inicio o a la secuencia Shine-Dalgarno (Fig. 3)[25].

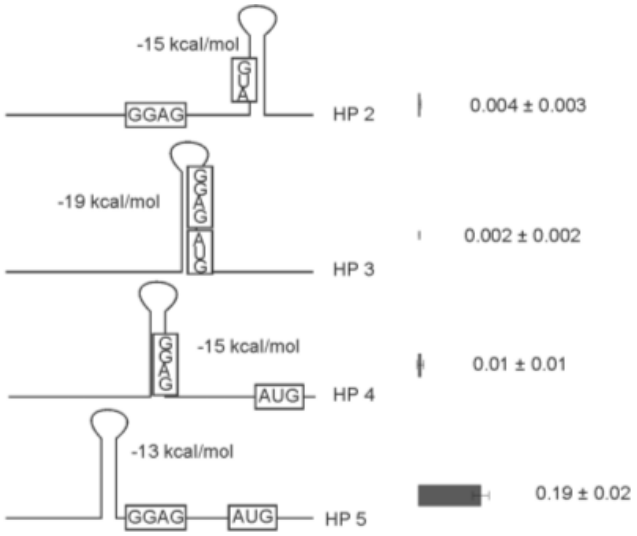

Fig. 3. Representación esquemática (lado izquierdo de la figura) y eficiencia en la traducción (lado derecho de la figura). Se puede observar una fuerte inhibición de la traducción cuando hay presencia de "hairpins" sobre las regiones con Shine-Dalgarno, mientras que si aumenta la distancia entre el "hairpin" y el Shine-Dalgarno se pierde gradualmente su influencia en la traducción (IIya et al., 2013).

1.1.2 Presencia de secuencias tipo-Shine-Dalgarno

La velocidad de la síntesis de proteínas sobre el mRNA no es uniforme. Existen pausas transitorias de los ribosomas que son originadas tanto por estructuras secundarias del mRNA, el uso preferencial de codones y secuencias similares al Shine-Dalgarno que tienden a interactuar de manera innecesaria con el gen rRNA 16S. Aunque en casos excepcionales, las secuencias del tipo Shine-Dalgarno pueden generar pausas que resultan beneficiosas para el plegado de las proteínas, en términos generales, éstas derivan en la reducción de la síntesis de proteínas por lo que se conoce existe una selección natural en contra de su presencia en regiones codificantes [31].

1.1.3 Traducción de genes con distancias intergénicas menores a 15 pb.

En operones bacterianos, es posible que el inicio de la traducción de un mRNA se encuentre sobrepuesta al término de la traducción del mRNA localizado inmediatamente río arriba. En estos casos de mRNA policistrónicos con regiones

intergénicas menores a 15 pb, la secuencia Shine-Dalgarno del gen río abajo se encuentra en la región codificante del gen inmediatamente anterior (Fig. 4). La mayoría de los arreglos intercistrónicos frecuentemente poseen una dependencia muy marcada de un Shine-Dalgarno, así como la presencia de regiones ricas en A/U. Adicionalmente a estos dos requisitos, se ha demostrado que la traducción del cistrón río arriba es requerida para que se lleve a cabo la traducción del cistrón río abajo y que la distancia intercistrónica juega un papel primordial en el tipo de traducción que se llevará a cabo (Fig. 5).

Recientemente se ha reportado otro posible modelo para el inicio de la traducción llamado "70S-scanning initiation" que sugiere que el complejo 70S ribosomal no se disocia necesariamente luego de la terminación, sino que continúa escaneando el mRNA alrededor del codón de término en busca de una nueva secuencia Shine-Dalgarno, y que no siempre existe una pérdida o disociación de los factores de iniciación IF1 e IF3, pudiéndose encontrar estos factores presentes en el complejo 70S [36].

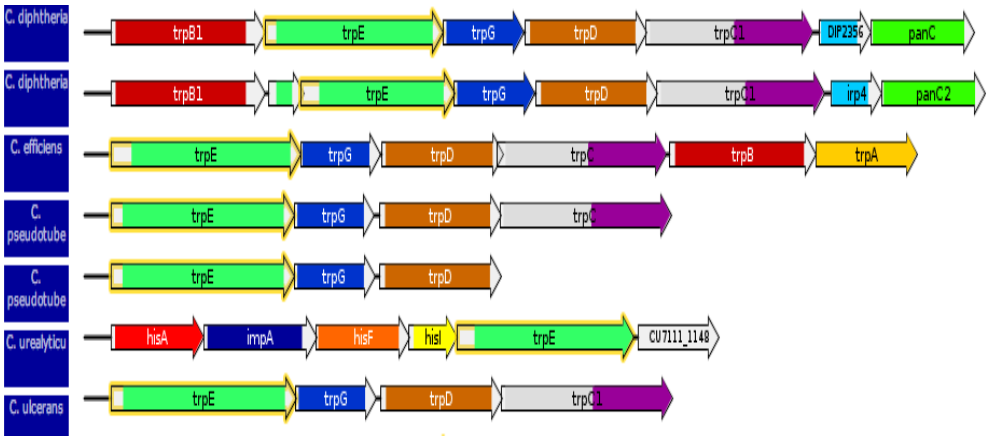

Fig. 4. Contexto genómico del operón de biosíntesis de triptófano en Actinobacteria. Algunos de los genes de dicho operón están sobrelapados, por lo que las secuencias Shine-delgarno de dichos genes se encontrarán en la región codificante del cistrón ubicado río arriba.

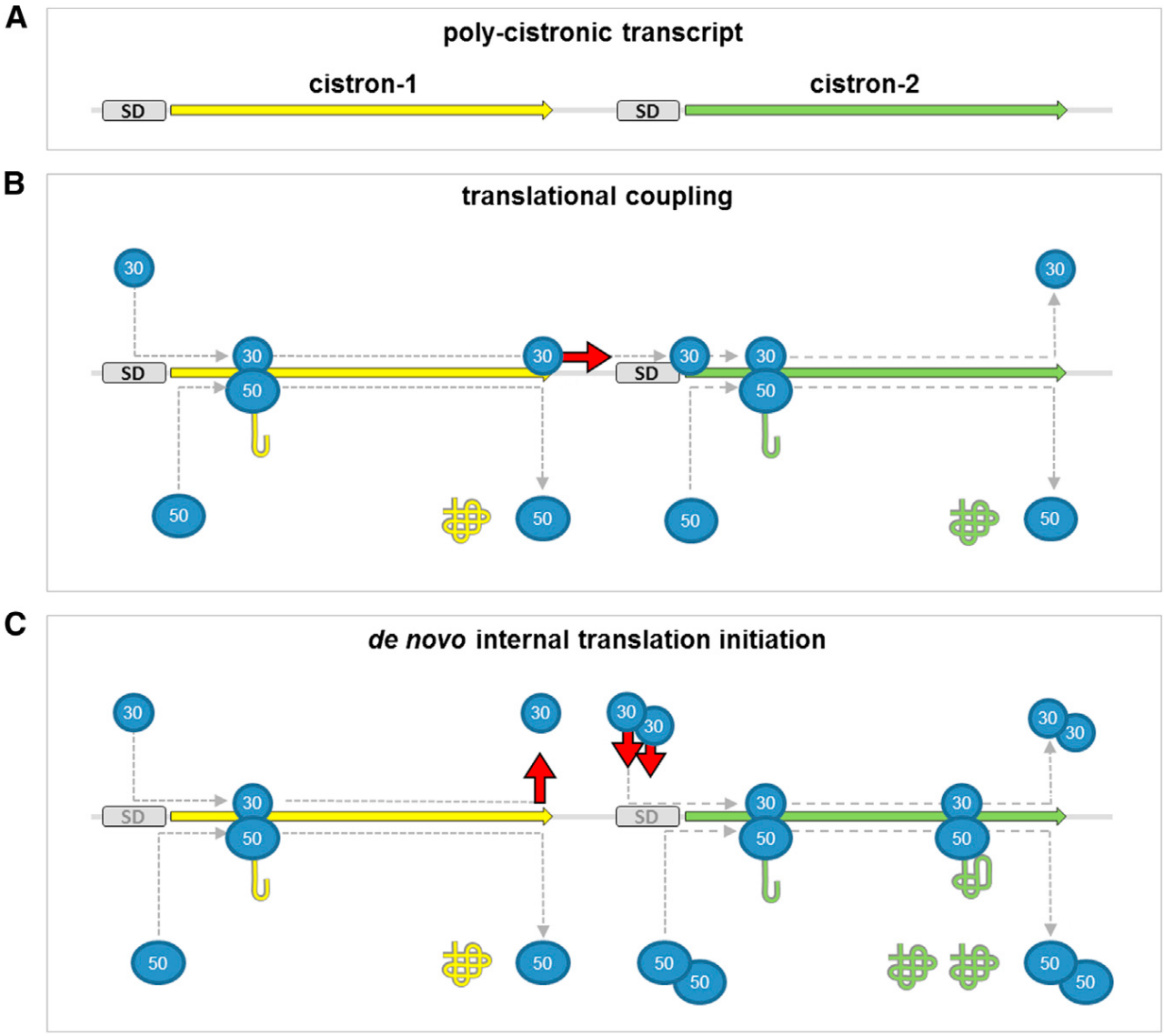

Fig. 5. Modelos de traducción para mensajeros policistrónicos en procariotes. (A) Transcrito con dos cistrones, cada uno con una secuencia Shine-Dalgarno. (B) Acoplamiento de la traducción donde la subunidad 30S ribosomal permanece asociada luego de la terminación y la 50S se une para la reiniciación. Generalmente sucede esto si la distancia intercistrónica es menor a 20 bases. (C) Reclutamiento *de novo* de ambas subunidades ribosomales. Esto puede producir mayores tasas de expresión entre las proteínas. (Quax et.al., 2013).

1.1.4 Traducción de genes adyacentes modulados por el gen que se encuentra río arriba

Existen estudios donde se expone una clara correlación entre la presencia de una secuencia Shine-Dalgarno y los niveles de expresión de genes predichos [28], pero hasta hoy, la medida en la que una secuencia Shine-Dalgarno determina la eficacia de

la traducción, es un tema incompleto. Adicionalmente a la "calidad" de las secuencias Shine-Dalgarno en mensajeros policistrónicos, se ha descrito un fenómeno llamado "acoplamiento traduccional" y se refiere a la interdependencia de la eficiencia de traducción entre genes codificados dentro de un operón [22]. El grado de acoplamiento se puede cuantificar midiendo cómo la traducción de un gen se modula por la tasa de traducción de un gen río arriba (Fig. 6).

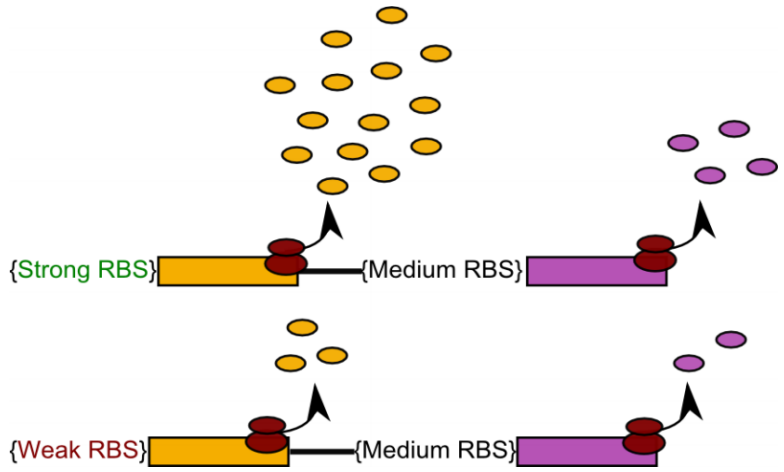

Fig. 6. Representación de la modulación de los cistrones dependiendo de la tasa de traducción del cistrón anterior (Levin -Karp et al., 2013).

Los operones, al contener genes funcionalmente relacionados, permiten a los procariotas co-regular su expresión dando una ventaja cuando iguales cantidades de los productos proteicos son requeridas, pero requerirá ajustes cuando éste no sea el caso. Se ha demostrado que la traducción es un factor determinante de la expresión modulada de genes agrupados en operones y que el uso de codones generalmente es el mejor indicador *in sílico* de una diferencia en la producción de una proteína [24]. También se ha reportado la importancia del espacio entre el RBS y el codón de inicio de la traducción [23]. Se conoce que 5 bases es el espacio óptimo para el inicio de la traducción.

1.2 Mecanismos alternativos de traducción en organismos procariontes y arqueobacterias

Pese a la gran conservación del mecanismo de traducción de organismos procariontes

y arqueobacterias basadas en la interacción entre el 16S rRNA y el Shine-Dalgarno en el extremo 5´ de mRNAs, se han descrito mecanismos alternativos y excepcionales que no requieren la secuencia Shine-Dalgarno para la iniciación de la traducción:

1. Mecanismo mediado por la proteína ribosomal S1 (RPS1)
2. Mecanismo utilizado por mRNAs sin líder

Se conoce también que algunas especies tales como cianobacterias presentan estos mecanismos alternos de traducción y que organismos de los phyla Proteobacteria, Tenericutes y Bacteroidetes no cuentan en su genoma con 16S ribosomales que contengan la secuencia anti-SD.

1.2.1 Mecanismo mediado por una proteína ribosomal S1

La proteína ribosomal S1 (RPS1) es componente de la subunidad 30S ribosomal y es conocida por su alta afinidad por RNAs de cadena-sencilla y zonas ricas en pirimidinas (A/U). Por ejemplo, en *Escherichia coli*, la proteína RPS1 interactúa con la región 5′ UTR del mRNA, iniciando eficientemente la traducción a pesar de la ausencia de secuencias Shine-Dalgarno. La proteína RPS1 le confiere a la subunidad 30S ribosomal la actividad de chaperona de RNA que es esencial para el acoplamiento (docking) y desnaturalización de mRNA estructurados, y para posicionar correctamente al codón de inicio dentro del canal de traducción [32], confiriéndole al ribosoma las condiciones necesarias para iniciar la traducción de los mRNA a pesar de sus estructura secundaria.

1.2.2 Mecanismo utilizado para la traducción de mRNA sin líder

Se ha visto que algunos mRNA no poseen secuencia líder en organismos procariontes y arqueobacterias. En estos casos el mRNA puede ser traducido mediante su unión directa con la subunidad pequeña del ribosoma 70S que incluirá un tRNA con el residuo de N-formilmetionina (fMet). Particularmente en arqueobacterias, por ejemplo, en *Halobacterium salinarum*, que pertenece a las Euryarchaeota, los mRNAs sin líderes tienen 15 veces mayor actividad en la traducción de mensajeros en relación a los mRNAs con presencia de secuencia Shine-Dalgarno.

Como hemos dicho, el inicio de la traducción en bacterias generalmente ocurre mediante la interacción del rRNA 16S y el extremo 5´ UTR del mRNA (Fig. 7A) que

junto con los factores de iniciación (IF1, IF2 y IF3) y el aminoacil-tRNA fMet-tRNA$_f$ forman el complejo de iniciación 30S de tal forma que el codón de iniciación queda ubicado dentro de la parte del sitio P de la subunidad menor del ribosoma. Se ha observado que un aumento en la concentración del factor IF2 estimula la capacidad de reclutamiento del fMet-tRNAf (tRNA-iniciador) y por tanto el pegado de este al sitio P [35]. Estos resultados sugieren que el reconocimiento 30S-mRNA (sin líder) es llevado a cabo por un complejo 30S-IF2-tRNA-iniciador (Fig. 7B)[33].

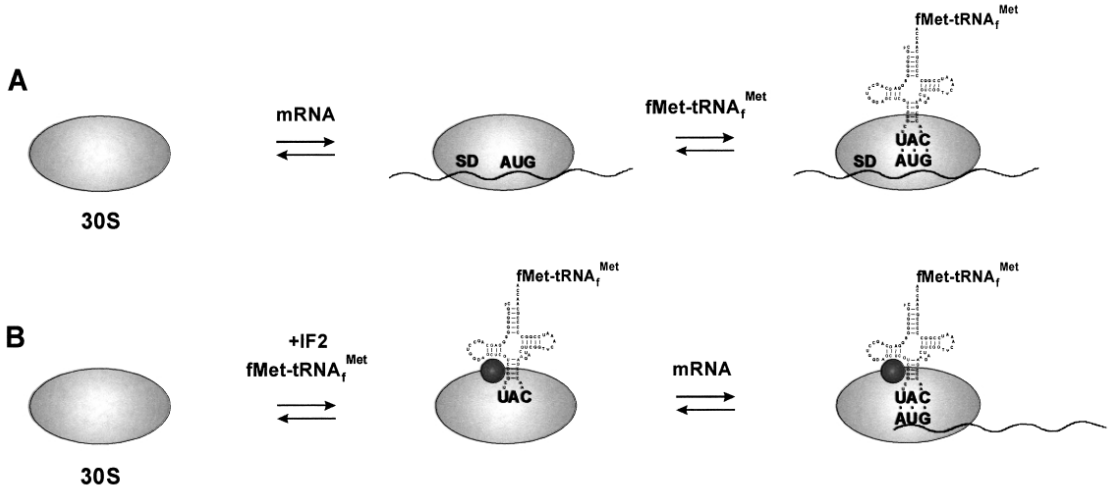

Fig. 7. Vías de inicio de la traducción en procariotes. (A) Reclutamiento de un ribosoma procariótico por un mRNA que contiene una región de inicio de traducción canónica a través la interacción Shine-Dagarno/anti-SD. (B) Reconocimiento del codón de inicio de un mRNA sin líder por un complejo del 30S y el RNAt iniciador. (Moll, Isabella, et al. 2002)

En términos generales, las observaciones iniciales de los elementos anteriormente descritos que intervienen en el inicio de la traducción en bacterias, fueron descritos a partir de observaciones en organismos modelo, como *Escherichia coli* y *Bacillus subtilis.* Gracias a los avances significativos en las nuevas tecnologías de secuenciación, actualmente existen mas de 10,000 genomas secuenciados en su totalidad, por lo que es posible y conveniente realizar un estudio global que los incluya.

2. OBJETIVOS

2.1 Objetivo General

Identificar y caracterizar las diferencias y similitudes de los elementos que intervienen en el proceso del inicio de la traducción canónica (dada por la interacción Shine-Dagarno/anti-SD) y no-canónica en organismos procariontes mediante un estudio de genómica comparativa a partir de sus correspondientes secuencias genómicas.

2.2 Objetivos particulares

Identificación y análisis de los siguientes eventos de regulación de la expresión génica:

1. Identificación de sitios Shine-Dalgarno.
2. Distribución filogenética de las características de la secuencia Shine-Dalgarno (tamaño, grado de conservación, energía libre (ΔG) de las secuencias Shine-Dalgarno de los mRNAs de un organismo con la región anti-SD de sus correspondientes RNAs ribosomales 16S, presencia de secuencias no-canónicas Shine-Dalgarno)
3. Análisis de energía libre (ΔG) de las secuencias Shine-Dalgarno de los mRNAs de un organismo con la región anti-SD de sus correspondientes RNAs ribosomales 16S en relación con el contenido de GC genómico.
4. Identificación de especies de bacterias que carecen del mecanismo de traducción "canónico".
5. Caracterización del grado de conservación de la secuencia Shine-Dalgarno en primeros *versus* segundos cistrones de operones bacterianos.

3. METODOLOGÍA

3.1 Selección de genomas

Fueron seleccionados los genomas de las bacterias depositados en la base de datos Refseq (NCBI). Se desarrollaron scripts de Perl para elegir un solo genoma por especie, teniendo en cuenta el siguiente orden: en primer lugar los genomas de referencia, clasificación que se aplica a genomas de alta calidad, curados manualmente y seleccionados por NCBI y la comunidad como estándar contra el cual se compararon otros datos; en segundo lugar los genomas representativos, referente a genomas seleccionados manualmente o computacionalmente como mejor de un clado o especie y en tercer lugar los genomas más recientes en el orden genome - chromosomes - scaffolds – contigs de completitud. Los análisis están actualizados al total de genomas (13,208 bacterias) disponibles en esta base de datos hasta junio de 2017. Los scripts desarrollados durante mi estudio, son incluidos en el Apéndice 9.1 de esta tesis.

3.2 Cálculo de energía libre de interacción entre moléculas de RNA.

El cálculo de energía libre de regiones de RNA se realizó utilizando el programa Free2bind con tamaños de análisis variable [13]. Free2bind hace el cálculo de energía libre en función de la posición de los nucleótidos. El algoritmo requiere una secuencia de ácido nucleico corta como el "decodificador" que se alinea sucesivamente con una secuencia de "mensaje" más larga en la que se codifica la información. En cada alineamiento, el algoritmo calcula una energía libre de hibridación de los nucleótidos. El cálculo de la energía libre utiliza programación dinámica extendida para identificar la conformación de energía libre mínima en bucles internos, y el modelo de enlace de hidrógeno del vecino más cercano para estimar el valor asociado de la energía libre para esa conformación.

Como parte de la validación de Free2bind, comparamos sus resultados con los obtenidos con el programa RNAhybrid [45], otra herramienta muy utilizada para encontrar la energía libre mínima de hibridación de dos RNAs. Los resultados mostraron una correlación importante (r=0.8) entre ambos programas como se puede ver en la figura 8.

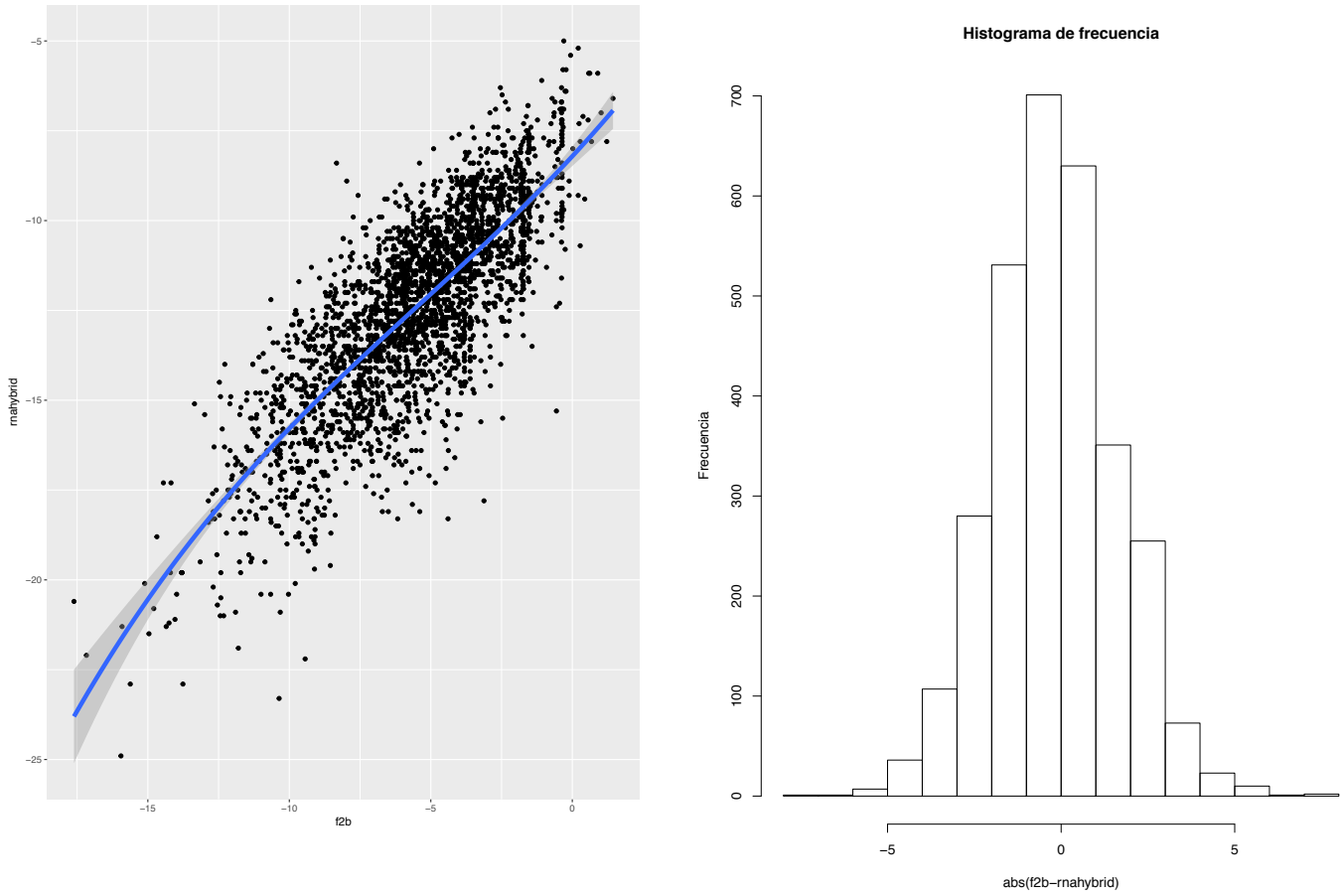

Fig. 8. Histograma de frecuencia (derecha) y curva de regresión de los valores de Free2bind y RNAhybrid, para el cálculo de las energía libre de los UTRs de *E. coli* y la región anti-SD en el extremo 3´ del gen ribosomal 16S.

3.3 Cálculo de energía libre de la interacción anti-Shine-Dalgarno / Shine - Dalgarno

Para este análisis se consideró la región 3' del gen ribosomal 16S de cada genoma con los 15 primeros nucleótidos río arriba de cada región codificante y se evaluó el cambio en la energía libre de la interacción con el programa Free2bind [13]. Fue seleccionado el 16S de cada genoma mejor anotado en cuanto a tamaño y extremo 3’

3.4 Identificación de motivos de secuencia conservada

Los posibles motivos de secuencia conservada alrededor del codón de inicio de la traducción fueron encontrados utilizando algoritmos de maximización de la expectancia (EM) con la paquetería MEME [1].

3.5 Filogenia

Se utilizó el software ssu-align [41] para realizar el alineamiento de las secuencias y se obtuvo la filogenia con Jmodeltest [43] y Phylm [42]. La visualización se realizó con iTOL [40].

3.6 Anotación de la proteína ribosomal S1

Se obtuvieron de la base de datos curada Swiss-Prot [49] las secuencias de las proteínas S1 anotadas; dichas secuencias fueron alineadas usando el programa Mafft [38]. El alineamiento múltiple así construido fue usado como dato de entrada para el programa hmmbuild del paquete de análisis HMMER [18] para construir un modelo de Cadenas de Markov Escondidas de la proteína S1. Por último se utilizó el programa hmmrsearch del mismo paquete de análisis HMMER, para buscar sobre los 6 marcos de lecturas de los genomas las proteínas S1.

4. RESULTADOS

Una parte fundamental para el desarrollo del proyecto consistió en la selección de los organismos y las secuencias tanto de los "mensajeros" como de los genes ribosomales 16S con los que trabajamos. Con el objeto de tener una primera aproximación de estudio, se seleccionaron 1,000 genomas no-redundantes de un total de 10,000 genomas considerando que en ningún caso se tuvieran más de una cepa de la misma especie. Utilizando los programas de análisis para identificación de motivos conservados MEME-MAST [1], se detectaron los posibles sitios Shine-Dalgarno de los genes y se construyeron matrices de probabilidad que representaran la frecuencia relativa del motivo sobre-representado.

Derivado de análisis de la región 3´ de los 16Ss, descubrimos que en muchos más de los casos esperados, la anotación de los genes 16S no incluye la región correspondiente a la secuencia anti-SD. Cabe mencionar que este problema de anotación, había sido reportado desde el 2008 [20], pero aún persiste. Debido a lo anterior, trabajamos en el desarrollo de un algoritmo para hacer nuestra predicción de los genes 16S y asegurar la correcta anotación del extremo 3' de dichos genes ribosomales.

Se realizó un "pipeline" para la selección de los genes ribosomales 16S y los resultados confirmaron la anotación incompleta de estos en NCBI para el 40% de los organismos analizados . Cabe mencionar que tratamos de alertar de esta situación a la comunidad científica a través de una nota corta en revistas de Bioinformáticas, pero para algunos de los revisores se trataba de un problema a resolver directamente con NCBI. Unos meses después actualizamos nuestros análisis con un mayor número de organismos y descubrimos que finalmente NCBI mejoró la anotación del extremo 3' de los 16S, encontrando incompletos, sólo en el 2% de los organismos seleccionados. ¿Habrá tenido que ver nuestro intento de publicación en esta mejoría de las bases de datos ?.

4.1 Corrección de la anotación de los extremos 3´de los genes ribosomales 16S.

Primeramente, del conjunto de 1,000 genomas, se seleccionaron aquellos genes ribosomales 16S que tuvieran una longitud esperada de genes canónicos o "normales",

definida a partir de la distribución de tamaños de los genes 16S, tal y como se muestra en las figuras 9 y 10.

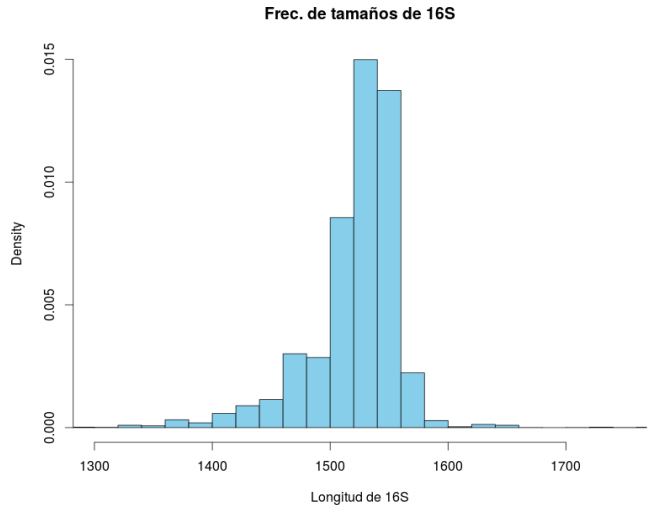

Fig. 9. Histograma de frecuencia de tamaños de los genes 16S reportados en Genbank para el año 2017.

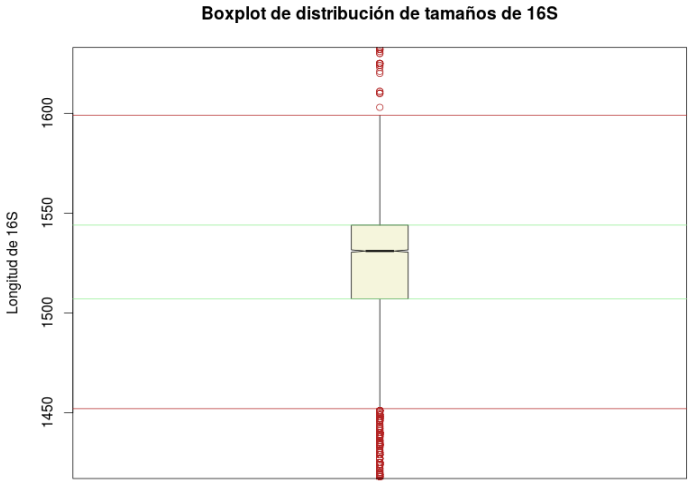

Fig. 10. Gráfico de caja de frecuencia de tamaños de los genes ribosomales 16S reportados en la base de datos Genbank para el año 2017.

Con la información del histograma y boxplot anteriores decidimos completar todos aquellos genes 16S mayores a 1,300 b, llevándolos hasta 1,600 pb para buscar el posible sitio anti-SD. Se obtuvo un archivo en formato fasta con la secuencia de los

genes 16S filtrando sólo aquellos mayores de 1,299 pb y menores de 1,601 pb y luego le añadimos las bases de la secuencia genómica necesarias para llegar a 1,600 pb. Con el archivo anteriormente construido, se hizo un análisis con el programa MEME con solo 1,000 secuencias obtenidas al azar para buscar el motivo del anti-SD. Se obtuvo la matriz para el motivo "GATCACCTCCTT[AT]" y posteriormente con esta matriz se hizo un MAST sobre todas las secuencias para saber cuáles finalmente tienen la firma del anti-SD y en qué posición. Finalmente obtuvimos que 4,387 genes (42% de 10,397 – total analizado) estaban reportados incompletos, no contenían la firma CCTCC y otro 18% estaban reportados parcialmente.

El trabajo de curar las secuencias de los genes ribosomales 16S es un objetivo colateral de nuestro proyecto dado que queremos incluir en nuestro estudio la mayor cantidad posible de organismos y no podríamos analizar las interacciones Shine-Dagarno/anti-SD de los rRNAs 16S con sus secuencias anti-SD mal anotadas. Por esto vimos la necesidad implementar un algoritmo predictivo de genes ribosomales 16S completos como herramienta a utilizar para mejorar la anotación de este importante gen en lugar del programa predictor de rRNA por excelencia, RNAmmer [50] utilizado por NCBI.

4.2 Desarrollo de herramienta para identificación de los genes ribosomales 16S:

Para el desarrollo de dicha herramienta se utilizó el programa MEME y nuestros modelos curados de genes 16S para buscar las matrices más características de éstos. Se seleccionaron aquellos parámetros (i.e. número de motivos a identificar), con los que se obtuviera correctamente los extremos 5´y 3´de este gen. Una vez con nuestras matrices pudimos realizar la anotación de los nuevos 16S buscando con el programa MAST sobre los genomas bacterianos.

Fig. 11. Motivo del extremo 3' del 16S ribosomal. Se puede apreciar la firma CCTCC del anti-SD.

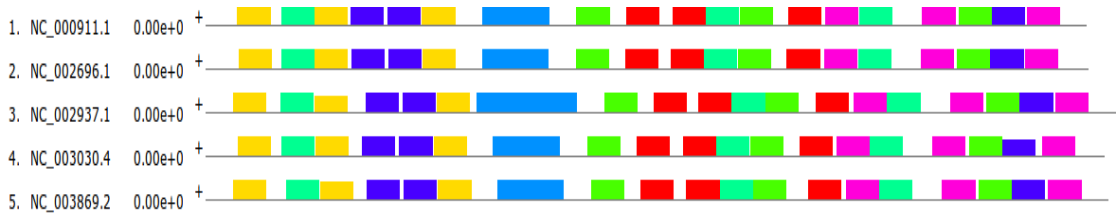

Fig. 12. Consenso de motivos a lo largo de los 16S ribosomales.

Con la matriz obtenida (Fig. 12) para los motivos presentes en los 16S, utilizamos el programa MAST (Fig. 13) para anotar estos sobre los 16S (completados) y se utilizaron scripts de Perl para evaluar y finalmente extraer la estructura correcta de los genes ribosomales 16S. Dichos scripts se incluyen en el apéndice 9.1 de esta tesis.

Trabajando con los genomas disponibles en NCBI hasta junio de 2017 (13,208 Bacterias), utilizamos nuestro "pipeline" de anotación de los genes ribosomales 16S para finalmente obtener el conjunto total de genes incluidos en el estudio.

4.3 Análisis de la variación de los valores de ΔG de la interacción Shine-Dagarno/anti-SD.

Se utilizó el software Free2bind para el cálculo del ΔG entre las 15 bases río arriba del codón de inicio de todos los mensajeros de cada una de las bacterias (12,112) y las últimas 15 bases del extremo 3' de los genes ribosomales 16S de cada bacteria, respectivamente. Esta colección de valores de ΔG por organismo, fue utilizada para obtener (de acuerdo al corte de -4.4 kcal/mol) el conjunto de genes con secuencias, así como los estadísticos de la media y la mediana que fueron utilizadas en nuestros estudios comparativos. La Fig. 13 muestra un ejemplo de la salida obtenida por el programa free2bind.

```
Delta-G for best pairing: -8.979195

Length of bound region between sequences: 6

seq1 binding begins on base index: 10
seq2 binding begins on base index: 2

| = pair is a member of a helix
0 = pair is a member of a loop
P = bulge in top strand
b = bulge in bottom strand

seq1   0: auagucgauuGGGAGGaaga
          :           ||||||
seq2   0:         uuUCCUCCacuaggucggcg
```

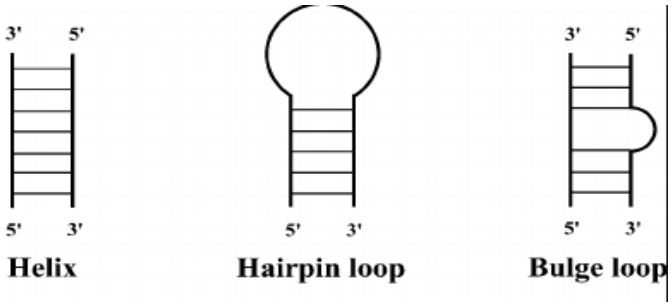

Fig. 13. Análisis del valor de ΔG usando el Software Free2bind.

4.4 Especies de secuencias UTRs sin anti-SD canónico

De acuerdo a lo reportado [18], se conocían 15 organismos que no presentaban anti-SD en sus genes ribosomales 16S, pertenecientes en su mayoría a las clases Mollicutes y Flavobacteriia. Adicional a estos 15 organismos, nuestros estudios detectaron otras 125 especies, incluídas 5 dentro de una nueva clase (Cytophagia), representadas en la lista 1, mostrada en el apéndice 9.2. Con esta lista de organismos, nos preguntamos si tendría alguno de ellos pudiera tener un nuevo motivo conservado, en lugar del motivo correspondiente a la secuencia anti-SD canónico. Por ello, se usó nuevamente el programa MEME para analizar todas las regiones 3' de los 16S obteniendo un nuevo "motivo" conservado sin anti-SD: "TGGAACATCTCAT"

Como podemos ver en el siguiente alineamiento, al comparar este motivo, con el motivo del anti-SD vemos como el núcleo de este "CCTCC" está modificado en dos bases y presenta un tercer cambio dos bases arriba:

Motivo para el anti-SD: "TGGATCACCTCCTT"

Identity: 10/13 (76.9%)

```
anti-SD                 1 TGGATCACCTCCT      13
                            ||||.||.|||.|
nuevo-anti-SD           1 TGGAACATCTCAT      13
```

A continuación, intentamos buscar la existencia de una secuencia complementaria a este nuevo motivo (TGGAACATCTCAT) en las regiones UTRs cercanas al codón de inicio de los mRNA de estas especies, con el fin de ver si pudiera existir un nueva Shine-Dalgarno correspondiente. De manera inesperada, los resultados obtenidos no mostraron ningún motivo conservado, ni siquiera el correspondiente al Shine-Dalgarno canónico.

Como parte del estudio realizado, también hemos detectado una región conservada en los genes ribosomales 16S (motivo 19 de nuestra matriz de búsqueda), encontrada alrededor de 100 pb antes del extremo 3', que invariablemente se pierde (modifica), en todos los genes 16S sin anti-SD detectados con nuestro método.

En la **tabla 1**, presentamos un resumen de los organismos con nuevo anti-SD o sin

anti-SD y notamos que a pesar de pertenecer a 3 phyla distintos y 6 clases distintas, estas especies tienen dos elementos en común: todas necesitan de un Eucariote que funja como hospedero, ya sea a través del parasitismo, como el caso de los micoplasmas, o como simbiontes primarios de insectos, como el caso de las proteobacterias (*Hodkinia cicadicola*, *Zinderia insecticola* y *Carsonella ruddi*) y por otra parte todas cuentan con genomas extremadamente pequeños, posiblemente explicado a partir de la primera característica.

Tabla1:

Resumen de especies con nuevo anti-SD o sin anti-SD:

Flavobacteriia (nuevo anti-SD)	*Mollicutes (sin anti-SD)*	*Cytophagia (nuevo anti-SD)*	*Alphaproteobacteria (sin anti-SD)*	*Betaproteobacteria (sin anti-SD)*	*Gammaproteobacteria (sin anti-SD)*
Candidatus_sulcia_muelleri_CARI	Mycoplasma suis str. Illinois	Cardinium endosymbiont of Bemisia tabaci	Candidatus Hodgkinia cicadicola Dsem	Candidatus Zinderia insecticola CARI	Candidatus Carsonella ruddii
Candidatus_sulcia_muelleri_DMIN	Mycoplasma suis KI3806	Cardinium endosymbiont of Encarsia pergandiella			
Candidatus_sulcia_muelleri_GWSS	Mycoplasma haemofelis str. Langford	Fibrella_aestuarina			
Candidatus_sulcia_muelleri_SMDSEM	Mycoplasma haemocanis	Fibrella_sp._ES10-3-2-2			
.....	Mycoplasma ovis	Fibrisoma_limi			
	Mycoplasma parvum				
	Mycoplasma wenyonii				

4.5 Relación los valores de ΔG de la interacción Shine-Dagarno/anti-SD con el contenido de GC genómico:

En la figura 14 se puede observar que sí existe una relación entre un alto contenido de GC genómico y un menor valor ΔG de la interacción Shine-Dagarno/anti-SD.

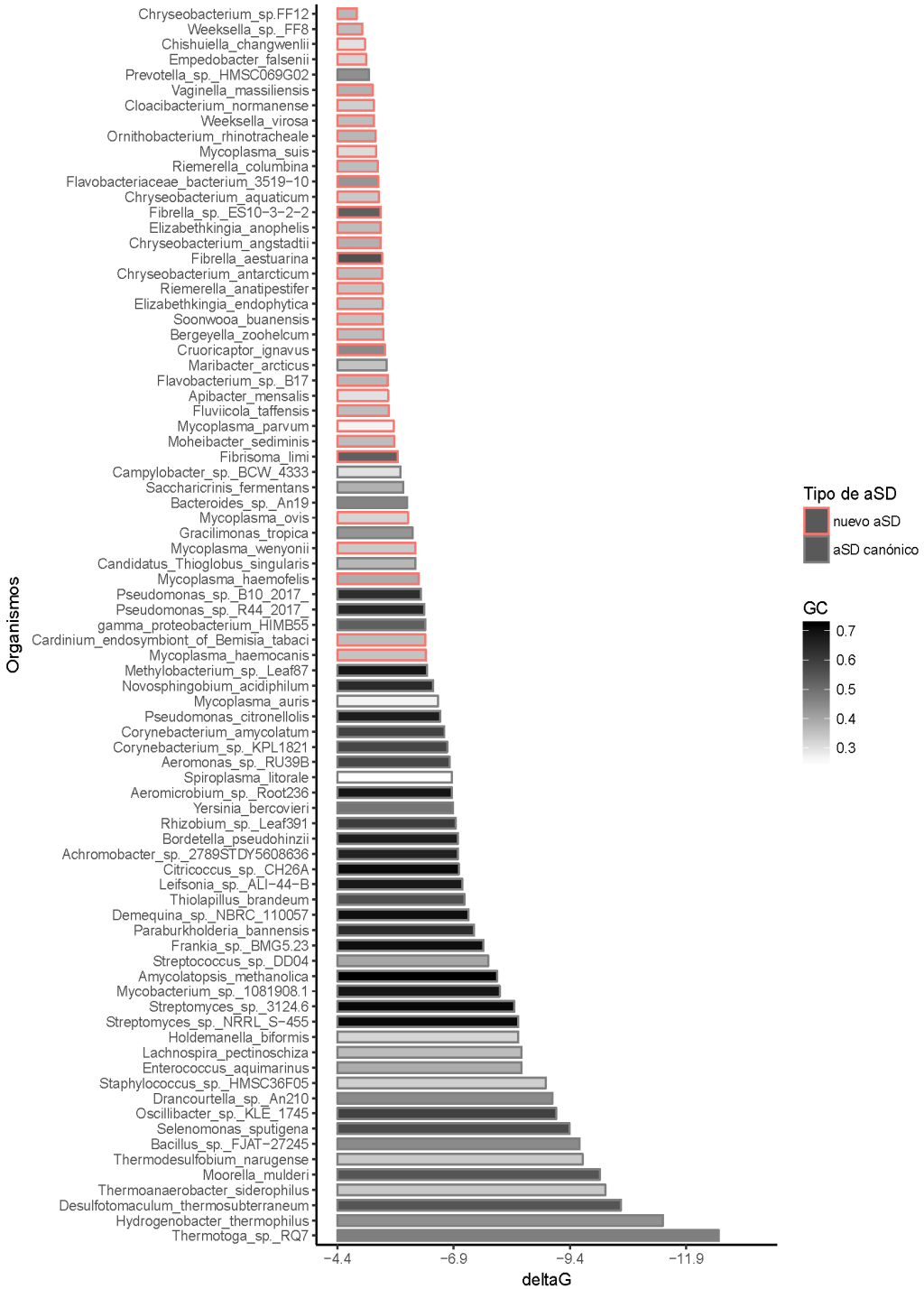

Fig. 14. Valor ΔG *versus* contenido de GC por organismo.

Para la gráfica anterior se utilizó el mismo conjunto de organismos que veremos en la filogenia, es decir, un conjunto de organismos representativo para el tipo de anti-SD canónico y el conjunto de los representativos con anti-SD no-canónico.

Hasta abajo se ven aquellos con mayor porcentaje de contenido de GC genómico, todos con anti-SD canónico y correspondiendo a los de menor ΔG. En este grupo se puede notar la relación de estas características con la presencia de organismos termófilos. Por otro lado, la parte superior de la gráfica agrupa los de menor contenido de GC, mayor ΔG y corresponden mayormente a los organismos con anti-SD nuevo o con ausencia de este.

Es importante destacar los casos de los organismos *Mycoplasma auris*, *Spiroplasma litorales* y *Prevotella sp.HMSC*. Los dos primeros pertenecen a la clase Mollicutes, de bajo contenido de GC y donde tenemos caracterizado organismos sin anti-SD, pero estos en particular sí presentan el anti-SD canónico. *Prevotella* por otro lado es una Bacteroidia con anti-SD, pero que a pesar de ésto, sus mensajeros carecen de secuencias Shine-Dalgarno.

4.6 Relación de la interacción Shine-Dalgarno/anti-SD de los primeros y segundos cistrones de los operones.

Para este estudio se tomaron 568 organismos con la anotación de sus operones de la base de datos del Gene Context Tool NG [44]. De cada operón, se tomaron el primero y segundo cistrón cuyas secuencias no estuvieran sobrepuestas en ninguna de sus bases (genes no-sobrelapantes) y contaran con al menos 30 genes con secuencias Shine-Dalgarno. Posteriormente se corrió free2bind y se obtuvo la media de los valores ΔG de las interacciones Shine-Dalgarno/anti-SD de los primeros cistrones de los operones, mismos que fueron promediados. El mismo valor promedio fue obtenido para los segundos cistrones de los operones. Los valores así obtenidos para cada organismo fueron comparados mediante la gráfica de la Fig. 15, en donde cada punto representa a un organismo cuya ordenada en X representa el valor promedio de los primeros cistrones y el valor de Y representa el valor promedio de los segundos cistrones. En la Fig. 15 se puede observar un sesgo hacia valores menores de ΔG (mejor interacción) en los 5' del segundo cistrón. Aunque este sesgo pudiera parecer

pequeño (puntos cercanos a la línea de 45 grados), hay que considerar que cada punto representa la media de cientos o miles de valores (genes de un organismo) y por tanto la desviación de la gráfica anterior es estadísticamente muy significativa. Para comprobar lo anterior se realizó la prueba no paramétrica de Kolmogórov-Smirnov utilizada para comparar dos muestras, la prueba arrojó valores de D = 0.09507 y p-value = 0.01179, por lo que podemos rechazar la hipótesis nula de que los conjuntos provengan de una misma distribución de referencia, validando que el sesgo apreciado no es producto del azar.

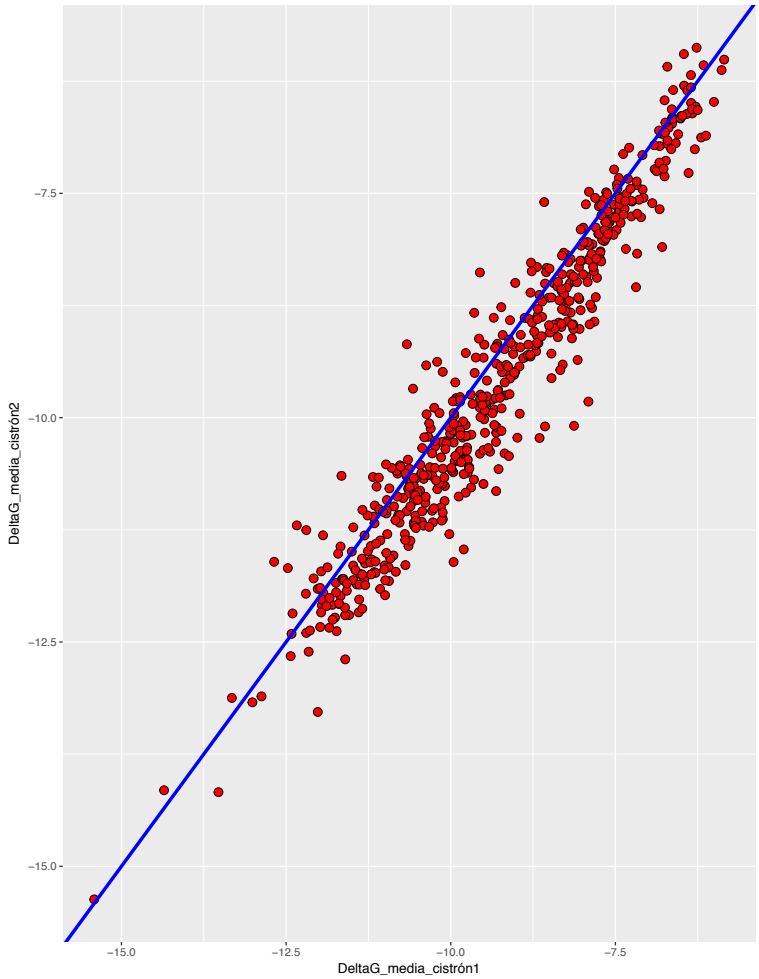

Fig.15. Comparación de los valores ΔG de las interacciones Shine-Dalgarno/anti-SD del primero y segundo cistrón de operones.

4.7 Relación de las interacciones Shine-Dalgarno/anti-SD de los genes sobrelapantes.

Para ver la relación de la interacción Shine-Dalgarno/anti-SD entre genes sobrelapantes o no-sobrelapantes, calculamos los valores de ΔG para el conjunto de genes por organismos con distancia intergénica menor a 5 pb, la promediamos y la graficamos junto al promedio de los valores del ΔG de los genes con al menos 15 pb de distancia intergénica. El resultado se muestra en la figura 16 en donde se puede observar un sesgo a un mejor valor de interacción (ΔG más negativos) con los UTRs de los genes sobrelapantes. Como en el caso anterior, para comprobar nuestra hipótesis realizamos la prueba no paramétrica de Kolmogórov-Smirnov utilizada para comparar dos muestras, la prueba arrojó valores de D = 0.20751 y p-value < 2.2e-16, por lo que validamos que el sesgo observado es altamente significativo y no es producto del azar.

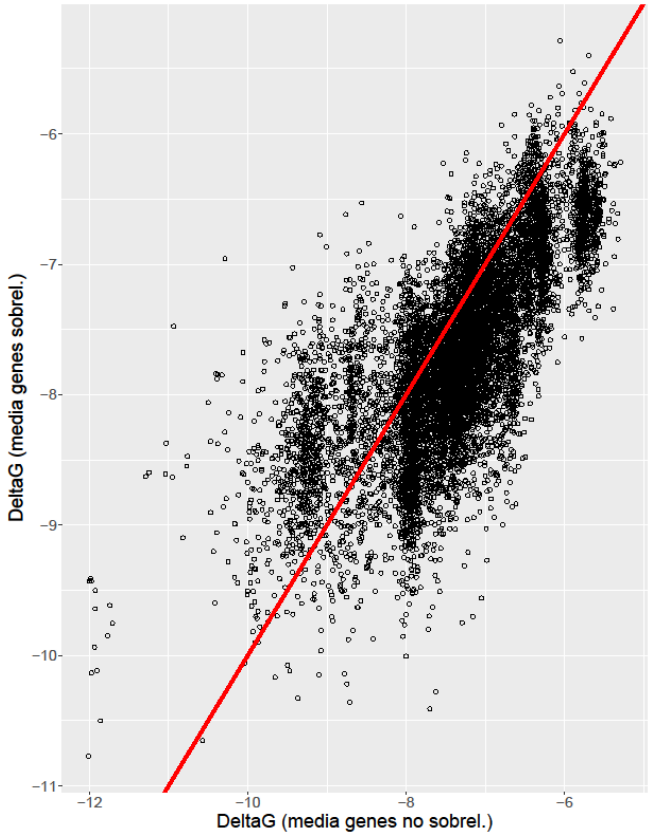

Fig.16. Comparación de los valores ΔG de las interacciones Shine-Dalgarno/anti-SD de los genes sobrelapantes.

4.8 Análisis filogenético de organismos y sus elementos característicos de inicio de traducción.

Para mostrar un árbol representativo se tomaron al azar 4 de cada 1,000 secuencias de rRNA de bacterias con el anti-SD canónico (CCTCC) y se seleccionaron los representantes de los rRNA con el "nuevo anti-SD" y "sin anti-SD" encontrados en este estudio, dejando un máximo de 3 individuos por género. Adicionalmente se agregó la información por organismo del contenido de GC, el porciento de genes con Shine-Dalgarno y el número de copias de la proteína ribosomal S1.

El alineamiento sirvió para identificar y caracterizar 3 tipos de 16S rRNA:

1) Con anti-SD canónico (ACCTCC-polyT).
2) Con nuevo-anti-SD (ATCTCa-polyT) la a minúscula no está 100% conservada.
3) Sin anti-SD.

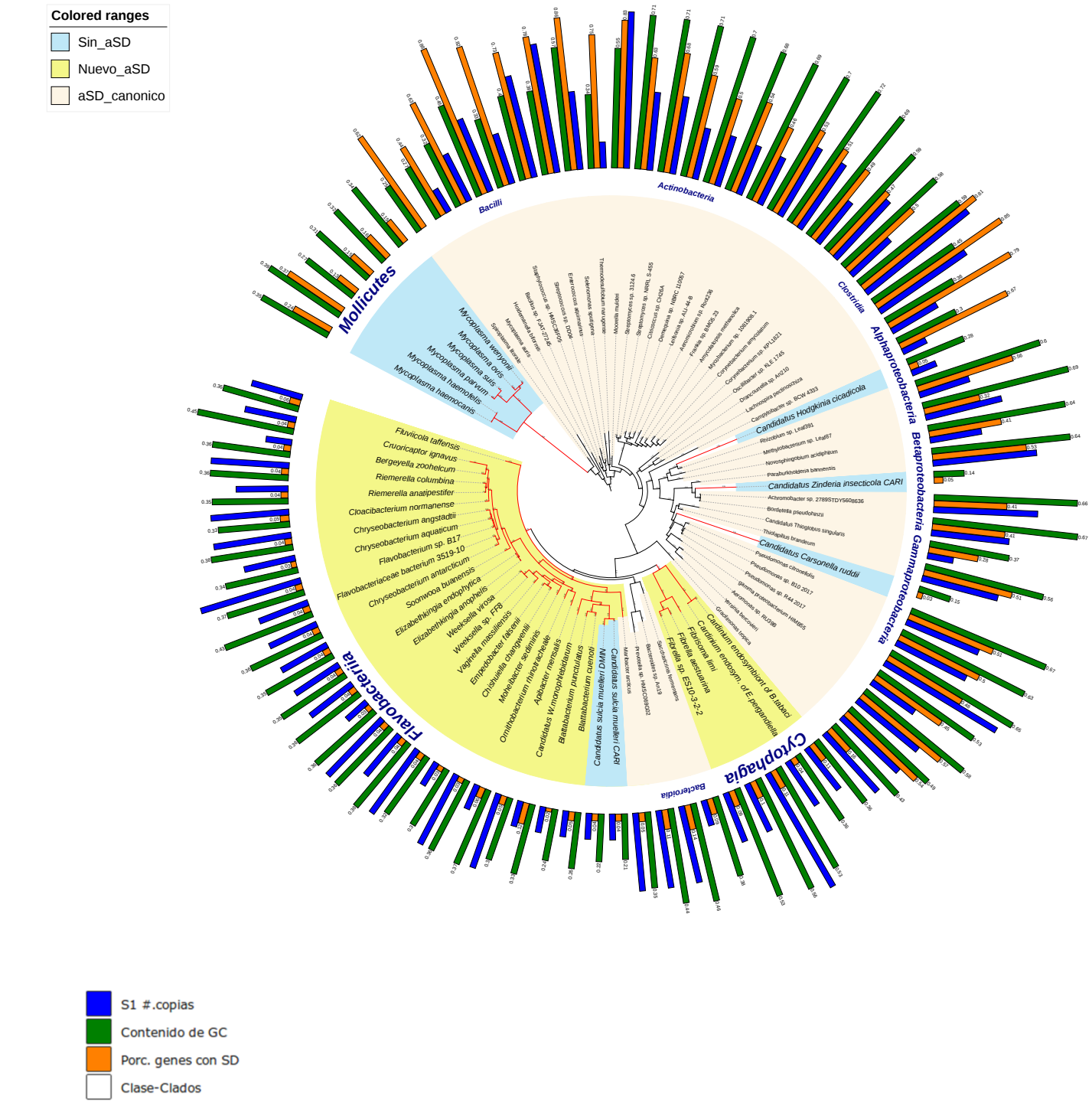

Fig.17: Árbol filogenético de organismos con diferente tipo de secuencia Shine-Dalgarno.

Del árbol anterior se puede concluir que los tres tipos de rRNA tienen un origen cercano a ser monofilético y bien separados, aunque sí encontramos, organismos con

secuencias anti-SD canónicos entre los organismos sin anti-SD, como es el caso de la flavobacteria *Maribacter arcticus*.

A pesar de la creencia casi generalizada sobre las secuencias Shine-Dalgarno en organismos bacterianos como una elemento *sine qua non* e invariable para el inicio de la traducción, los resultados del presente estudio muestran con claridad una visión diferente, en donde el inicio de la traducción no es dependiente de la interacción Shine-Dalgarno/anti-SD, y pueden resumirse en cuatro casos:

1) Organismos en los que la secuencia anti-SD canónico es sustituida por una secuencia conservada y similar al anti-SD canónico, pero con mRNAs carentes de su correspondiente secuencia complementaria con las que se podría establecer interacciones tipo Watson-Crick
2) Organismos en los que la secuencia anti-SD canónico es eliminada completamente del extremo 3´ de los rRNAs y sus mRNAs sin ninguna secuencia conservada en los UTRs de sus correspondientes mRNAs. Encontramos que la distribución de estos organismos está sesgada a un número limitado de phyla (Proteobacteria, Tenericutes y Bacteroidetes) y asociada con organismos bacterianos simbiontes, como es el caso de los Micoplasmas y de las tres Proteobacterias, en su mayoría parásitos de vertebrados, con los genomas bacterianos más pequeños que existen y un bajo contenido de GC. Por ejemplo, se ha estimado que más del 10% de las especies de insectos se ven beneficiados de endosimbiontes bacterianos que suministran nutrientes esenciales para su crecimiento, comprobando en éstos su bajo contenido de GC y que sus genomas son mucho más pequeños que los genomas de otras bacterias intracelulares o de vida libre [48].
3) Organismos del clado formado por Bacteroidia: *Bacteroides sp._An19*, *Prevotella sp._HMSC069G02* y *Saccharicrinis fermentans*, que a pesar de tener anti-SD canónico en sus genes rRNA 16S, no presentan secuencias Shine-Dalgarno en la region 5' UTR de sus correspondientes mRNAs.
4) Algunos organismos pertenecientes al género de las micoplasmas que contienen secuencias Shine-Dalgarno, pero sus genes ribosomales 16S carecen

de secuencias anti-SD, lo que podría sugerir que la pérdida de sus secuencias anti-SD es un fenómeno reciente.

Adicionalmente a lo anterior, nuestro estudio identificó a organismos cuyos genes ribosomales 16S carecen de secuencias anti-SD y que también carecen de la proteína S1. Proponemos que los mRNAs de dichos organismos, como *Candidatus Zinderia insecticola CARI* y *Candidatus Carsonella ruddii* (Beta y Ganma proteobacteria, respectivamentes), carecen de secuencia líder, y es este atributo lo que les permite que dichos mRNAs sean traducidos.

Se caracterizó la relación que tiene el cambio en la energía libre de la hibridación Shine-Dalgarno/anti-SD con el contenido de GC genómico y vimos la correlación de este valor entre el primer y el segundo cistrón, notando que la región 5' de este último presenta un sesgo a tener un mejor (más bajo) valor de ΔG así como la correlación con genes sobrelapantes, siendo también favorecidos estos en general con un menor valor de ΔG.

A pesar de corroborar que la forma de inicio de traducción canónica dependiente de la interacción entre la secuencia Shine-Dalgarno/anti-SD no es única, se confirma que la interacción sí es la más frecuente. El hecho de que nuestro estudio que consideró secuencias de organismos de todos los grupos filogenéticos secuenciados en la actualidad, no haya podido identificar ninguna otra secuencia, diferente a la Shine-Dalgarno con su contraparte funcional en los rRNAs 16S, sugiere que dicha interacción fue establecida previo a la división de bacterias y arqueas y constituye un evento que fue "fijado" en dichos grupos de organismos debido al fuerte costo funcional de su pérdida; no obstante, en algunos casos excepcionales, como el de organismos simbiontes, formas alternativas de inicio de traducción fueron posible establecerse.

6. CONCLUSIONES

En el presente trabajo se logra profundizar en el estudio y caracterización de los rRNA y los sistemas para el inicio de la traducción de los mensajeros en Bacteria y Arqueobacteria, presentando la lista con la mayor cantidad de especies descritas hasta el momento carentes de la secuencia anti-SD canónico en los rRNAs 16S.

Se obtuvo el porcentaje de genes con Shine-Dalgarno por bacteria y la presencia o no de la proteína ribosomal S1 para las 12,212 bacterias incluidas en el estudio y se caracterizó la relación que tiene el cambio en la energía libre de la hibridación Shine-Dalgarno/anti-SD con el contenido de GC genómico observando también la correlación del valor de esta hibridación entre el primer y el segundo cistrón de los operones.

7. PERSPECTIVAS

Las secuencias similares al Shine-Dalgarno producen una pausa en la traducción [11] y tal vez esa sea la causa de que sean evitadas dentro de las regiones codificantes [31]. Sin embargo, se ha visto que que las pausas en la traducción pueden favorecer el plegamiento de proteínas multidominio y aumentar la solubilidad [46]. Como parte del trabajo a seguir queremos buscar la conservación de estas secuencias tipo-Shine-Dalgarno en regiones codificantes de genes ortólogos y de esta forma ver si existe selección positiva para su conservación en proteínas multidominio. La ubicación los aminoácidos que codifican estas secuencias tipo Shine-Dalgarno en estructuras secundarias o tridimensionales, será analizada en términos de que pudieran delimitar dominios funcionales de plegamiento de proteínas.

8. BIBLIOGRAFÍA

1. Bailey TL, Williams N, Misleh C, Li WW (2006). "MEME: discovering and analyzing DNA and protein sequence motifs". Nucleic Acids Res 34 (Web Server issue): W369-373. doi:10.1093/nar/gkl198. PMC 1538909. PMID 16845028.
2. Durbin, Richard; Sean R. Eddy, Anders Krogh, Graeme Mitchison (1998). Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids. Cambridge University Press. ISBN 0-521-62971-3.
3. Benton, D. et al. (2006). GenBank. Nucleic Acids Research 34 (Database): pp. D16-D20.
4. Felsenstein, J. (1981). "Evolutionary trees from DNA sequences: A maximum likelihood approach". Journal of Molecular Evolution 17 (6): 368–376. doi:10.1007/BF01734359. PMID 7288891.
5. Illumina HiSeq. http://www.illumina.com/systems/hiseq 2500 1500.ilmn
6. PACBIO RS II. http://www.pacificbiosciences.com/
7. Starmer J, Stomp A, Vouk M, Bitzer D (2006) Predicting Shine–Dalgarno Sequence Locations Exposes Genome Annotation Errors. PLoS Comput Biol 2(5): e57. doi:10.1371/journal.pcbi.0020057
8. Palleja A, Harrington ED, Bork P (2008) Large gene overlaps in prokaryotic genomes: result of functional constraints or mispredictions? BMC Genomics 9: 335.
9. Gu W, Zhou T, Wilke CO (2010) A Universal Trend of Reduced mRNA Stability near the Translation-Initiation Site in Prokaryotes and Eukaryotes. PLoS Comput Biol 6(2): e1000664. doi:10.1371/journal.pcbi.1000664
10. Nakagawa S., Niimura Y., Miura K., Gojobori T. (2010). Dynamic evolution of translation initiation mechanisms in prokaryotes. Proc. Natl. Acad. Sci. U.S.A. 107, 6382–6387. 10.1073/pnas.1002036107
11. Li, G.W., Oh, E. & Weissman, J.S. The anti-SD sequence drives translational pausing and codon choice in bacteria. Nature 484, 538–541 (2012).
12. Osterman I. A., Evfratov S. A., Sergiev P. V., Dontsova O. A. Comparison of mRNA features affecting translation initiation and reinitiation. Nucleic Acids Res. 41:474–486. 2013
13. Starmer et al., PLoS Computat. Biol. (2006), Vol. 2, pp. e57. http://www.unc.edu/~starmer/free2bind/
14. PHYLIP is a free package of programs for inferring phylogenies. http://evolution.genetics.washington.edu/phylip.html
15. Redes Neuronales.http://www.uta.cl/charlas/volumen16/Indice/Ch-csaavedra.pdf
16. Jose Ramon Hilera Gonzales, Victor Jose Martınez Hernando, “Redes neuronales artificiales. Fundamentos, modelos y aplicaciones”, Addison-Wesley Iberoamericana S.A, ISBN 0-201-87895-X, De la Edición RA-MA, 1995.
17. RNAfold web server will predict secondary structures of single stranded RNA or DNA sequences. http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi
18. Zhou, R.J. Weber, J.W. Allwood, R. Mistrik, Z. Zhu, Z. Ji, S. Chen, W.B. Dunn, S. He, M.R. Viant. HAMMER: automated operation of mass frontier to construct in silico mass spectral fragmentation libraries
19. Lim, K., Furuta, Y. and Kobayashi, I., 2012. Large variations in bacterial ribosomal RNA genes. *Molecular biology and evolution*, p.mss101.
20. Lin, Yu-Hsiang, et al. "Questionable 16S ribosomal RNA gene annotations are frequent in completed microbial genomes." *Gene* 416.1 (2008): 44-47.

21. SHINE, John, and Lynn DALGARNO. "Terminal-Sequence Analysis of Bacterial Ribosomal RNA." *European Journal of Biochemistry* 57.1 (1975): 221-230.
22. Levin-Karp, Ayelet, et al. "Quantifying translational coupling in E. coli synthetic operons using RBS modulation and fluorescent reporters." *ACS synthetic biology* 2.6 (2013): 327-336.
23. Chen, Hongyun, et al. "Determination of the optimal aligned spacing between the Shine–Dalgarno sequence and the translation initiation codon of Escherichia coli m RNAs." *Nucleic acids research* 22.23 (1994): 4953-4957.
24. Quax, Tessa EF, et al. "Differential translation tunes uneven production of operon-encoded proteins." *Cell reports* 4.5 (2013): 938-944.
25. Osterman, Ilya A., et al. "Comparison of mRNA features affecting translation initiation and reinitiation." *Nucleic acids research* 41.1 (2012): 474-486.
26. Komarova, A. V., et al. "Extensive Complementarity of the Shine–Dalgarno Region and 3"-End of 16S rRNA Is Inefficient for Translation in vivo." *Russian Journal of Bioorganic Chemistry* 27.4 (2001): 248-256.
27. Jacob, William F., Melvin Santer, and Albert E. Dahlberg. "A single base change in the Shine-Dalgarno region of 16S rRNA of Escherichia coli affects translation of many proteins." *Proceedings of the National Academy of Sciences* 84.14 (1987): 4757-4761.
28. Ma, Jiong, Allan Campbell, and Samuel Karlin. "Correlations between Shine-Dalgarno sequences and gene features such as predicted expression levels and operon structures." *Journal of bacteriology* 184.20 (2002): 5733-5745.
29. Omotajo, Damilola, et al. "Distribution and diversity of ribosome binding sites in prokaryotic genomes." *BMC genomics* 16.1 (2015): 604.
30. Altschul, Stephen; Gish, Warren; Miller, Webb; Myers, Eugene; Lipman, David (1990). "Basic local alignment search tool". Journal of Molecular Biology 215 (3): 403–410. doi:10.1016/S0022-2836(05)80360-2. PMID 2231712.
31. Diwan, Gaurav D., and Deepa Agashe. "The Frequency of Internal Shine–Dalgarno-like Motifs in Prokaryotes." *Genome biology and evolution* 8.6 (2016): 1722-1733.
32. Duval, Mélodie, et al. "Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation." *PLoS Biol* 11.12 (2013): e1001731.
33. Moll, Isabella, et al. "Leaderless mRNAs in bacteria: surprises in ribosomal recruitment and translational control." *Molecular microbiology* 43.1 (2002): 239-246.
34. Grill, Sonja, et al. "Selective stimulation of translation of leaderless mRNA by initiation factor 2: evolutionary implications for translation." *The EMBO journal* 19.15 (2000): 4101-4110.
35. Gualerzi, Claudio O., et al. "Translation initiation in bacteria." *The Ribosome*. American Society of Microbiology, 2000. 475-494.
36. Yamamoto, Hiroshi, et al. "70S-scanning initiation is a novel and frequent initiation mode of ribosomal translation in bacteria." *Proceedings of the National Academy of Sciences* 113.9 (2016): E1180-E1189.
37. Li, Weizhong, and Adam Godzik. "Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences." *Bioinformatics* 22.13 (2006): 1658-1659.
38. Katoh, Kazutaka, et al. "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform." *Nucleic acids research* 30.14 (2002): 3059-3066.
39. McWilliam, Hamish, et al. "Analysis tool web services from the EMBL-EBI." *Nucleic*

acids research 41.W1 (2013): W597-W600.
40. Letunic, Ivica, and Peer Bork. "Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees." *Nucleic acids research* 44.W1 (2016): W242-W245.
41. Nawrocki, Eric P., and Sean R. Eddy. "ssu-align: a tool for structural alignment of SSU rRNA sequences." (2010).
42. Guindon, Stéphane, and Olivier Gascuel. "A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood." *Systematic biology* 52.5 (2003): 696-704.
43. Posada, David. "jModelTest: phylogenetic model averaging." *Molecular biology and evolution* 25.7 (2008): 1253-1256.
44. D. A. Abdala, R. Ciria and E. Merino GeConT 3: gene context analysis for orthologous proteins, conserved domains and metabolic pathways
45. Rehmsmeier, Marc, et al. "Fast and effective prediction of microRNA/target duplexes." Rna 10.10 (2004): 1507-1517.
46. Vasquez, Kevin A., et al. "Slowing translation between protein domains by increasing affinity between mRNAs and the ribosomal anti-Shine–Dalgarno sequence improves solubility." *ACS synthetic biology* 5.2 (2015): 133-145.
47. Shine, J., & Dalgarno, L. (1975). Determinant of cistron specificity in bacterial ribosomes. Nature, 254(5495), 34?38. https://doi.org/10.1038/254034a0
48. Molloy, Sheilagh. "A tiny alternative." Nature Reviews Microbiology 7.9 (2009).
49. Apweiler, Rolf, et al. “Uniprot: the universal protein knowledgebase.” Nucleic acid research 32.suppl_1 (2004): D115-D119.
50. Lagesen K, Hallin PF, Rødland E, Stærfeldt HH, Rognes T Ussery DW RNammer: consistent annotation of rRNA genes in genomic sequences

9. APÉNDICE

9.1 Algunos de los scripts de perl utilizados durante el estudio:

9.1.1 Descarga de base de datos (bacterias):

ext_allbest_genome_ncbi.pl

```
my (%lines, %HQ_genome);
my $file_to_open= shift; # assembly_summary.paths

open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
  chomp;
  %lines= ();
  %HQ_genome= ();
  my $HQ= 0;
  my $file_to_open2= $_;
  (my $org= $file_to_open2)=~s/\.\/([^\/]+).*/$1/;
  open IN2, $file_to_open2 or die "Cant read $file_to_open2\n";
  while(<IN2>){
    next if /^#/;
    chomp;
    if( /^.*\s+strain=([^\t]+)\s+latest.*\s+(\d+)\/\d{2}\/\d{2}/ ){
      my ($strain, $year)= ($1, $2);
      ($strain= $strain)=~s/ /_/g;
      if( /^.*\s+latest/ && /^.*\s+Genome/ && /^.*\s+(ftp\S+)/ ){
        $lines{$year}{$strain}{genome}= $1;
      }elsif( /^.*\s+latest/ && /^.*\s+Chromosome/ && /^.*\s+(ftp\S+)/ ){
        $lines{$year}{$strain}{chromosome}= $1;
      }elsif( /^.*\s+latest/ && /^.*\s+Scaffold/ && /^.*\s+(ftp\S+)/ ){
        $lines{$year}{$strain}{scaffold}= $1;
      }elsif( /^.*\s+latest/ && /^.*\s+Contig/ && /^.*\s+(ftp\S+)/ ){
        $lines{$year}{$strain}{contig}= $1;
      }
      if( /^.*\s+reference genome\s+.*/ && /^.*\s+(ftp\S+)/ ){
        $HQ_genome{$strain}{reference}= $1;
```

```
            $HQ_genome{$strain}{year}= $year;

        }

        if( /^.*\s+representative genome\s+.*/ && /^.*\s+(ftp\S+)/ ){

            $HQ_genome{$strain}{representative}= $1;

            $HQ_genome{$strain}{year}= $year;

        }


    }elsif(  /^.*\s+([^\t]+)\s+latest.*\s+(\d+)\/\d{2}\/\d{2}/ ){ # solo por unos pocos (287 de 13208) que no tienen la nomenclatura "strain=" solo tienen el nombre del strain. ver: lista_sin_strain.txt

        my ($strain, $year)= ($1, $2);

        if( /^.*\s+latest/ && /^.*\s+Genome/ && /^.*\s+(ftp\S+)/ ){

            $lines{$year}{$strain}{genome}= $1;

        }elsif( /^.*\s+latest/ && /^.*\s+Chromosome/ && /^.*\s+(ftp\S+)/ ){

            $lines{$year}{$strain}{chromosome}= $1;

        }elsif( /^.*\s+latest/ && /^.*\s+Scaffold/ && /^.*\s+(ftp\S+)/ ){

            $lines{$year}{$strain}{scaffold}= $1;

        }elsif( /^.*\s+latest/ && /^.*\s+Contig/ && /^.*\s+(ftp\S+)/ ){

            $lines{$year}{$strain}{contig}= $1;

        }

        if( /^.*\s+reference genome\s+.*/ && /^.*\s+(ftp\S+)/ ){

            $HQ_genome{$strain}{reference}= $1;

            $HQ_genome{$strain}{year}= $year;

        }

        if( /^.*\s+representative genome\s+.*/ && /^.*\s+(ftp\S+)/ ){

            $HQ_genome{$strain}{representative}= $1;

            $HQ_genome{$strain}{year}= $year;

        }

    }

  }

foreach my $strain ( sort keys %HQ_genome ){

    $HQ_genome{$strain}{reference} ? (print "$org\t$strain\treference\t$HQ_genome{$strain}{year}\t$HQ_genome{$strain}{reference}\n") : (print "$org\t$strain\trepresentative\t$HQ_genome{$strain}{year}\t$HQ_genome{$strain}{representative}\n");

    $HQ++;

  }

  unless ($HQ){

  OUTER:
```

```
    foreach my $year ( sort {$b<=>$a} keys %lines ){
     foreach my $strain ( sort keys $lines{$year} ){
       if( $lines{$year}{$strain}{genome} ){
          print "$org\t$strain\tGenome\t$year\t$lines{$year}{$strain}{genome}\n";
          last OUTER;
       }
     }
     foreach my $strain ( sort keys $lines{$year} ){
       if( $lines{$year}{$strain}{chromosome} ){
          print "$org\t$strain\tChrm\t$year\t$lines{$year}{$strain}{chromosome}\n";
          last OUTER;
       }
     }
     foreach my $strain ( sort keys $lines{$year} ){
       if( $lines{$year}{$strain}{scaffold} ){
          print "$org\t$strain\tScaff\t$year\t$lines{$year}{$strain}{scaffold}\n";
          last OUTER;
       }
     }
     foreach my $strain ( sort keys $lines{$year} ){
       if( $lines{$year}{$strain}{contig} ){
          print "$org\t$strain\tContig\t$year\t$lines{$year}{$strain}{contig}\n";
          last OUTER;
       }
     }
     }
    }
}
```

downloads_from_ncbi.pl

```
my $file_to_open= shift; # ext_best_genome_ftp_resume.txt";

open IN, $file_to_open or die "Cant read $file_to_open\n";

while(<IN>){

   chomp;

   my($org, $strain, $ftp)= (split)[0,1,-1];

   print "$org, $strain, $ftp\n";

   my $command= "wget -r -l1 -np \"$ftp/\" -A \"*.fna.gz,*feature_table.txt.gz,*assembly_stats.txt,*.faa.gz,*.gff.gz\"  -R
\"*cds_from_genomic.fna.gz\" -P ${org}_-_$strain";
```

```
   system("$command");
}
```

taxids_NCBI.pl

```
my $file_to_open= shift; # recibe el  org_vs_taxIds.txt ex:
# Escherichia_coli     511145    K-12_substr._MG1655
# Sphingobacterium_spiritivorum            525373    ATCC_33861
# Paenibacillus_sp._FSL_H7-0357            1536774  FSL_H7-0357
# Demequina_aurantiaca          676200    NBRC_106265

open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
   my($org, $taxid, $strain)= (split)[0,1,2];
   my $urlcore= "https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/";
   my $url_taxid= "wwwtax.cgi?mode=Undef&id=$taxid";
   my ($superkingdom, $phylum, $class, $order, $family, $genus, $species);

   system("wget \"$urlcore$url_taxid\"");

   open IN2, $url_taxid or die "Cant read $url_taxid\n";
   open OUT, ">${org}_-_$strain/taxonomy.txt" || die "can't write taxonomy.txt";
   while(<IN2>){
      if( /^.*superkingdom">([^<]+)/ ){
         $superkingdom= $1;
      }
      if( /^.*phylum">([^<]+)/ ){
         $phylum= $1;
      }
      if( /^.*class">([^<]+)/ ){
         $class= $1;
      }
      if( /^.*order">([^<]+)/ ){
         $order= $1;
      }
      if( /^.*family">([^<]+)/ ){
         $family= $1;
      }
```

```
    if( /^.*genus">([^<]+)/ ){
      $genus= $1;
    }
    if( /^.*species">([^<]+)/ ){
      $species= $1;
    }
  }
  $superkingdom ||= 'NA';
  $phylum ||= 'NA';
  $class ||= 'NA';
  $order ||= 'NA';
  $family ||= 'NA';
  $genus ||= 'NA';
  $species ||= 'NA';

  print OUT
"Organism\tTaxID\tSuperkingdom\tPhylum\tClass\tOrder\tFamily\tGenus\tSpecies\n$org\t$taxid\t$superkingdom\t$phylu
m\t$class\t$order\t$family\t$genus\t$species\n";
  unlink $url_taxid;
}
```

9.1.2 Extracción de 16S

ext_16s_from_NCBI_rna_from_genomic.pl

```
my ($dir, $fasta, $name, %seq);

my $file_to_open= shift; #  rna_from_genomic_fna.paths == ( find . -name '*_rna_from_genomic.fna' > rna_from_genomic_fna.paths
)
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
  my ($name, %seq);
  chomp;
        if( /^\.\/([^\/]+)\/(\S+)/ ){
    ($dir, $fasta)= ($1, $2);
  }
  print STDERR "$dir, $fasta\n";
  chdir $dir;
  open IN2, $fasta or die "Cant read $fasta\n";
  my $good= 0;
```

```
  while(<IN2>){
    chomp;
    if( /^>(.*product=16S ribosomal.*$)/ ){
      $name= $1;
      $good++; next
    }elsif( /^>/ ){
      $good= 0; next
    }
    $seq{$name}.= $_ if $good;
  }
  open OUT, ">${dir}_16S.fna";
  foreach my $ribo ( sort keys %seq ){
    print OUT ">$ribo\n$seq{$ribo}\n";
  }

  chdir '..';

#   exit;
}
```

ext_16s_from_NCBI_rna_from_genomic_2.pl

```
my ($dir, $fasta, $name, %seq);

my $file_to_open= shift; #  16S_from_genomic_fna.paths == ( find . -name '*_16S.fna' > 16S_from_genomic_fna.paths )
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
  my ($name, %seq);
  chomp;
        if( /^\.\/([^\/]+)\/(\S+)/ ){
    ($dir, $fasta)= ($1, $2);
  }
  print STDERR "$dir, $fasta\n";
  chdir $dir;
  open IN2, $fasta or die "Cant read $fasta\n";
  my $good= 0;
  while(<IN2>){
```

```
    chomp;
    if( /^>(.*product=16S ribosomal.*\[location=complement\(<\d+\.\..*$)/ ){
      $name= $1;
      $good= 0; next
    }elsif( /^>(.*product=16S ribosomal.*\[location=complement\(\d+\.\..*$)/ ){
      $name= $1;
      $good++; next
    }elsif( /^>(.*product=16S ribosomal.*\[location=.*\.\.\d+.*$)/ ){
      $name= $1;
      $good++; next
    }elsif( /^>/ ){
      $good= 0; next
    }
    $seq{$name}.= $_ if $good;
  }
  open OUT, ">${dir}_best16S.fna";
  foreach my $ribo ( sort {length($seq{$b}) <=> length($seq{$a})} keys %seq ){
    print OUT ">$ribo\n$seq{$ribo}\n";
    last;
  }

  chdir '..';

#   exit;
}
```

9.1.3 Cálculo de energía libre (Free2bind)

sacar_cola_15pb_16s.pl

```
my($name, %seq_15tail_rev);

my $file_to_open= shift; # fasta "bacteria_newanti-SD_16S.fna";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
  chomp;
```

```
   if( /^>(\S+)/ ){
      $name= $1;
   }
   $seq_20tail_rev{$name}= substr($_, -15);
}

foreach my $seq ( sort keys %seq_15tail_rev ){
   print ">$seq\n$seq_15tail_rev{$seq}\n";
}
```

run_free2bind.pl

```
use Parallel::ForkManager;
my $pm = new Parallel::ForkManager(900);

my (%good, $org, $end, $gene, %sobrel);

my $file_to_open= "/scratch01/karel/f2b/sobrelapantes.paths";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
   chomp;
   open IN2, $_;
   while(<IN2>){
      chomp;
      $sobrel{$_}++;
   }
}


my $file_to_open= "/scratch01/karel/f2b/best16s_tail.paths";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
   chomp;
   if( /^\.\/([^\/]+)/ ){
      $good{$1}++;
   }
}
```

```
my $file_to_open1= "/scratch01/karel/f2b/ur.paths";
open IN1, $file_to_open1 or die "Cant read $file_to_open1\n";
while(<IN1>){
   chomp;
   my $ur= $_;
   if( /^\.\/([^\/]+)/ ){
      $org= $1;
      next unless $good{$org};
      print STDERR "$org\n";
   }
   open IN2, "$org/best16S_tail.fna";
   while(<IN2>){
      chomp;
      next if /^>/;
      $end= reverse(substr $_ ,2,15);
   }
#    exit;
   open IN3, "$ur";
   while(<IN3>){
      chomp;
      if( /^>(\S+)/ ){
         $gene= $1;
         next
      }
      next unless $sobrel{$gene};
      $pm->start and next;
      system("echo 'gene $gene' >> /scratch01/karel/f2b/bacterias/${org}/${org}_$gene.f2b_ur_tmp");
      my $utr= substr $_, -15;
      my $command= "/home/karel/bin/free2bind/free_align.pl $utr $end >>
/scratch01/karel/f2b/bacterias/${org}/${org}_$gene.f2b_ur_tmp";
      system("$command");
      $pm->finish;
   }
$pm->wait_all_children;
   system("cat /scratch01/karel/f2b/bacterias/${org}/*.f2b_ur_tmp >
/scratch01/karel/f2b/bacterias/${org}/${org}.f2b_sobrelapantes");
```

```
   unlink </scratch01/karel/f2b/bacterias/${org}/*.f2b_ur_tmp>;
#    exit;
}
```

ext_deltaG.pl

```
chomp(my $file_to_open= `ls -1`);

open OUT, ">${file_to_open}.deltaG";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
   chomp;
   if( /^Delta-G for best pairing:.*\s+(\S+)$/ ){
      print OUT "$1\n";
   }

}
```

9.1.4 Extracción de resultados

porciento_genes_conSD_por_org.pl

```
my ($org);
my $file_to_open= "/free/databases/ncbi_bacteria/db_230617/f2b.paths";
my $file_to_open= "/free/databases/ncbi_bacteria/db_230617/f2b2paths";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
   chomp;
   my $f2b= $_;
   ($org= $f2b)=~ s/\.\/([^\/]+).*/$1/;
   print STDERR "$org\n";
   my ($gene, $cnt, $cntSD);

   open IN2, $f2b or die "Cant read $f2b\n";
   open SD, ">$org/genes_with_SD.txt2";
   open nonSD, ">$org/genes_without_SD.txt2";
```

```
  while(<IN2>){
    if( /^gene (\S+)/ ){
      $gene= $1;
      $cnt++;
      next;
    }
    if( /^Delta-G for best pairing:.*= (\S+)/ ){
      my $deltag= $1;
      $deltag <= -4.4 ? ($cntSD++, print Shine-Dalgarno "$gene\t$deltag\n") : (print nonSD "$gene\t$deltag\n");
    }
  }
  next if $cnt==0;
  print "$org\t",sprintf( "%2.2f", $cntSD/$cnt),"\n";
#   exit;
}
```

allf2b_stats_genes.pl

```
my %data;

my ($org, $cnt);
# my $file_to_open= "/free/databases/ncbi_bacteria/db_230617/f2b.paths";
my $file_to_open= "/free/databases/ncbi_bacteria/db_230617/f2b2paths";
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
  chomp;
  my $f2b= $_;
  ($org= $f2b)=~ s/\.\/([^\/]+).*/$1/;

  open IN2, "$org/genes_with_SD.txt2" or die "Cant read $org\n";
  open IN3, "$org/genes_without_SD.txt2" or die "Cant read $org\n";

  while(<IN2>){
    chomp;
    push @{$data{$org}{wSD}}, (split)[1];
  }
```

```
  while(<IN3>){
    chomp;
    push @{$data{$org}{woutSD}}, (split)[1];
  }

#   exit;
}

open allSDstat, ">allf2b_stats_genes_with_SD.txt2";
open allnonSDstat, ">allf2b_stats_genes_without_SD.txt2";

foreach my $org ( sort keys %data ){
  print STDERR "$org\t", $cnt++,"\n";
  open SDstat, ">$org/f2b_stats_genes_with_SD.txt2";
  open nonSDstat, ">$org/f2b_stats_genes_without_SD.txt2";
            my($mean, $std, $median, $mode);
  if( @{$data{$org}{wSD}} > 30 ){
            mean(\@{$data{$org}{wSD}}) ? ($mean = sprintf( "%2.2f", mean(\@{$data{$org}{wSD}}))) : ($mean ='NA');
            stdev(\@{$data{$org}{wSD}}) ? ($std = sprintf( "%2.2f", stdev(\@{$data{$org}{wSD}}))) : ($std ='NA');
            median(\@{$data{$org}{wSD}}) ? ($median = sprintf( "%2.2f", median(\@{$data{$org}{wSD}}))) : ($median ='NA');
    mode(\@{$data{$org}{wSD}}) ? ($mode = sprintf( "%2.2f", mode(\@{$data{$org}{wSD}}))) : ($mode ='NA');
            print SDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";
            print allSDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";

  }else{
    $mean ='NA'; $std ='NA'; $median ='NA'; $mode ='NA';
            print SDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";
  }

  if( @{$data{$org}{woutSD}} > 30 ){
            mean(\@{$data{$org}{woutSD}}) ? ($mean = sprintf( "%2.2f", mean(\@{$data{$org}{woutSD}}))) : ($mean ='NA');
            stdev(\@{$data{$org}{woutSD}}) ? ($std = sprintf( "%2.2f", stdev(\@{$data{$org}{woutSD}}))) : ($std ='NA');
            median(\@{$data{$org}{woutSD}}) ? ($median = sprintf( "%2.2f", median(\@{$data{$org}{woutSD}}))) : ($median
='NA');
    mode(\@{$data{$org}{woutSD}}) ? ($mode = sprintf( "%2.2f", mode(\@{$data{$org}{woutSD}}))) : ($mode ='NA');
            print nonSDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";
            print allnonSDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";
```

```
   }else{
     $mean ='NA'; $std ='NA'; $median ='NA'; $mode ='NA';
      print nonSDstat "$org\tmean: $mean\tstd: $std\tmedian: $median\t mode: $mode\n";
   }

#    exit;
}

sub mean{
   my($data) = @_;
   if (not @$data) {
        die("Empty array\n");
   }
   my $total = 0;
   foreach (@$data) {
        $total += $_;
   }
   my $average = $total / @$data;
   return $average;
}
sub stdev{
   my($data) = @_;
   if(@$data == 1){
        return 0;
   }
   my $average = &mean($data);
   my $sqtotal = 0;
   foreach(@$data) {
        $sqtotal += ($average-$_) ** 2;
   }
   my $std = ($sqtotal / (@$data-1)) ** 0.5;
   return $std;
}
sub median {
              my($data) = @_;
```

```
    my @a = sort {$a <=> $b} @$data;
    my $length = scalar @a;
    return undef unless $length;
    ($length % 2)
      ? $a[$length/2]
      : ($a[$length/2] + $a[$length/2-1]) / 2.0;
}
sub mode {
            my($data) = @_;
    my(%freq, %freq_per_pos, @mode, $mode);
    foreach (@$data) { $freq{sprintf( "%3.1f", $_)}++ }; # para que tome el numero con 1 valor despues de la coma
            foreach my $val ( sort keys %freq ){
            push @{$freq_per_pos{$freq{$val}}}, $val
    }
    my @sorted_array = sort {$a <=> $b}  keys %freq_per_pos;
            foreach my $elem ( @{$freq_per_pos{pop(@sorted_array)}} ){
        push @mode, $elem;
    }
    @sorted_array > 0 ? ($mode = join(',', @mode)) : ($mode = 'NA');
            return $mode;
}
```

gc_content.pl

```
my ($cnt, $all);

my $file_to_open= shift;
open IN, $file_to_open or die "Cant read $file_to_open\n";
while(<IN>){
    next if /^>/;
    chomp;
            $cnt+= $_=~ tr/GgCc//;
    $all+= length($_);
}

print "GC content: ",sprintf( "%2.2f\t", $cnt/$all), "\n";
```

9.2 Lista.1. Lista completa de organismos sin anti-SD en sus 16S:
(*) reportados

(+) encontrados nuevos

* Flavobacteriaceae bacterium 3519-10

* Riemerella anatipestifer

* Weeksella virosa

* Blattabacterium sp. Blattella germanica

* Blattabacterium sp. str. BPLAN

* Candidatus Sulcia muelleri CARI

* Candidatus Sulcia muelleri DMIN

* Candidatus Sulcia muelleri GWSS

* Candidatus Sulcia muelleri SMDSEM

* Candidatus Hodgkinia cicadicola Dsem

* Candidatus Zinderia insecticola CARI

* Candidatus Carsonella ruddii

* Mycoplasma haemofelis

* Mycoplasma suis str. Illinois

* Mycoplasma suis KI3806

+Cardinium_endosymbiont_of_Bemisia_tabaci

+Cardinium_endosymbiont_of_Encarsia_pergandiella

+ Fibrella_aestuarina

+ Fibrella_sp._ES10-3-2-2

+ Fibrisoma_limi

+ Blattabacterium_cuenoti

+ Blattabacterium_punctulatus

+ Blattabacterium_sp._Blaberus_giganteus_

+ Blattabacterium_sp._Blatta_orientalis_

+ Blattabacterium_sp._Blattella_germanica_

+Blattabacterium_sp._Mastotermes_darwiniensis_

+ Blattabacterium_sp._Nauphoeta_cinerea_

+Blattabacterium_sp._Periplaneta_americana_

+ Fluviicola_taffensis

+ Apibacter_mensalis

+ Bergeyella_zoohelcum

+ Chishuiella_changwenlii

+ Chryseobacterium_angstadtii

+ Chryseobacterium_antarcticum

+ Chryseobacterium_aquaticum

+ Chryseobacterium_arachidis

+ Chryseobacterium_arthrosphaerae

+ Chryseobacterium_artocarpi

+ Chryseobacterium_balustinum

+ Chryseobacterium_caeni

+ Chryseobacterium_carnipullorum

+ Chryseobacterium_chaponense

+ Chryseobacterium_contaminans

+ Chryseobacterium_cucumeris

+ Chryseobacterium_daeguense

+ Chryseobacterium_formosense

+ Chryseobacterium_gallinarum

+ Chryseobacterium_gambrini

+ Chryseobacterium_gleum

+ Chryseobacterium_greenlandense

+ Chryseobacterium_gregarium
+ Chryseobacterium_haifense
+ Chryseobacterium_halperniae
+ Chryseobacterium_hispalense
+ Chryseobacterium_indologenes
+ Chryseobacterium_indoltheticum
+ Chryseobacterium_jeonii
+ Chryseobacterium_joostei
+ Chryseobacterium_koreense
+ Chryseobacterium_kwangjuense
+ Chryseobacterium_limigenitum
+ Chryseobacterium_luteum
+ Chryseobacterium_molle
+ Chryseobacterium_oranimense
+ Chryseobacterium_palustre
+ Chryseobacterium_piperi
+ Chryseobacterium_piscicola
+ Chryseobacterium_polytrichastri
+ Chryseobacterium_scophthalmum
+ Chryseobacterium_shigense
+ Chryseobacterium_soli
+ Chryseobacterium_solincola
+ Chryseobacterium_sp._6021061333
+ Chryseobacterium_sp._BGARF1
+ Chryseobacterium_sp._BLS98
+ Chryseobacterium_sp._CBo1
+ Chryseobacterium_sp._CCH4-E10
+ Chryseobacterium_sp._CF284
+ Chryseobacterium_sp._CF299
+ Chryseobacterium_sp._CF314
+ Chryseobacterium_sp._CF356
+ Chryseobacterium_sp._CF365
+ Chryseobacterium_sp._FF12
+ Chryseobacterium_sp._FH1
+ Chryseobacterium_sp._FH2
+ Chryseobacterium_sp._G972
+ Chryseobacterium_sp._Hurlbut01
+ Chryseobacterium_sp._IHB_B_10212
+ Chryseobacterium_sp._IHB_B_17019
+ Chryseobacterium_sp._J200
+ Chryseobacterium_sp._JAH
+ Chryseobacterium_sp._JM1
+ Chryseobacterium_sp._Leaf180
+ Chryseobacterium_sp._Leaf201
+ Chryseobacterium_sp._Leaf394
+ Chryseobacterium_sp._Leaf404
+ Chryseobacterium_sp._Leaf405
+ Chryseobacterium_sp._MOF25P
+ Chryseobacterium_sp._OV259
+ Chryseobacterium_sp._OV705
+ Chryseobacterium_sp._RU33C
+ Chryseobacterium_sp._RU37D
+ Chryseobacterium_sp._StRB126
+ Chryseobacterium_sp._UNC8MFCol
+ Chryseobacterium_sp._VT16-26
+ Chryseobacterium_sp._YR005
+ Chryseobacterium_sp._YR203
+ Chryseobacterium_sp._YR459
+ Chryseobacterium_sp._YR460
+ Chryseobacterium_sp._YR480
+ Chryseobacterium_taiwanense
+ Chryseobacterium_takakiae

+ Chryseobacterium_tenax
+ Chryseobacterium_ureilyticum
+ Chryseobacterium_vrystaatense
+ Chryseobacterium_zeae
+ Cloacibacterium_normanense
+ Cruoricaptor_ignavus
+ Elizabethkingia_anophelis
+ Elizabethkingia_endophytica
+ Elizabethkingia_genomosp._2
+ Elizabethkingia_genomosp._3
+ Elizabethkingia_genomosp._4
+ Elizabethkingia_meningoseptica
+ Elizabethkingia_miricola
+ Elizabethkingia_sp._C1558
+ Elizabethkingia_sp._F8124
+ Elizabethkingia_sp._G4070
+ Elizabethkingia_sp._HvH-WGS333
+ Empedobacter_falsenii
+ Flavobacterium_sp._B17
+ Moheibacter_sediminis
+ Flavobacteriaceae_bacterium_3519-10
+ Ornithobacterium_rhinotracheale
+ Riemerella_anatipestifer
+ Riemerella_columbina
+ Soonwooa_buanensis
+ Vaginella_massiliensis
+ Weeksella_sp._FF8
+ Weeksella_sp._HMSC059D05
+ Weeksella_virosa
+Candidatus_Walczuchella_monophlebidarum
+ Candidatus_Hodgkinia_cicadicola_Dsem
+ Candidatus_Zinderia_insecticola_CARI
+ Candidatus_Carsonella_ruddii
+ Mycoplasma_haemocanis
+ Mycoplasma_haemofelis
+ Mycoplasma_ovis
+ Mycoplasma_parvum
+ Mycoplasma_suis
+ Mycoplasma_wenyonii
+ Candidatus_sulcia_muelleri_CARI
+ Candidatus_sulcia_muelleri_DMIN
+ Candidatus_sulcia_muelleri_GWSS
+ Candidatus_sulcia_muelleri_SMDSEM