Universidad Nacional Autónoma de México
Maestría y Doctorado en Ciencias Bioquímicas

Análisis del papel de las chaperonas de RNA en la regulación genética mediada por sRNAs en bacterias

TESIS

QUE PARA OPTAR POR EL GRADO DE:
Maestro en Ciencias

PRESENTA:
Ing. Walter Josué Hernández Santos

TUTOR PRINCIPAL

Dr. Enrique Merino Pérez
Instituto de Biotecnología, UNAM

MIEMBROS DEL COMITE TUTOR

Dr. José Luis Puente García
Instituto de Biotecnología, UNAM

Dr Cei Abreu Goodger
LANGEBIO-CINVESTAV, IPN

Cuernavaca, Morelos. Julio de 2019

Universidad Nacional Autónoma de México
UNAM
Biblioteca Central
Dirección General de Bibliotecas de la UNAM

UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

Todo el material contenido en esta tesis esta protegido por la Ley Federal del Derecho de Autor (LFDA) de los Estados Unidos Mexicanos (México).

El uso de imágenes, fragmentos de videos, y demás material que sea objeto de protección de los derechos de autor, será exclusivamente para fines educativos e informativos y deberá citar la fuente donde la obtuvo mencionando el autor o autores. Cualquier uso distinto como el lucro, reproducción, edición o modificación, será perseguido y sancionado por el respectivo titular de los Derechos de Autor.

A mi familia

El presente trabajo se realizó bajo la dirección del Dr Enrique Merino Pérez en el grupo de Genómica Computacional adscrito al Departamento de Microbiología Molecular del Instituto de Biotecnología de la Universidad Nacional Autónoma de México.

Este proyecto se realizó con fondos de CONACyT mediante el proyecto CONACYT FC 2015-2 887 «Generando nuevos paradigmas dentro de lo biología sintética aplicado al estudio de estresomos bacterianos» otorgado al Dr. Enrique Merino Pérez

Este trabajo se realizó gracias a la Beca CONACyT para estudios de posgrado nivel maestría para estudiantes nacionales durante el periodo Enero 2017 - Dic 2018 con número de becario 619341.

Partes de este trabajo fueron presentadas en:
Hernández-Santos, W. J., Rodríguez-Escamilla, Z. and Merino-Pérez E. Analysis of Bacterial Stressomes using Synthetic Biology en *International Symposium on Fuctional Genomics and Systems Biology*, Mayo 2017.

Agradecimientos

A los integrantes del comité tutoral que participaron en las diferentes etapas del proyecto: Dr. Cei Abreu Goodger y Dr. José Luis Puente García.

A los sinodales del comité de revisión de esta tesis por su tiempo invertido: Dr. Jose Luis Reyes Taboada, Dr. José Adelfo Escalante Lozada, Dra. Blanca Itzel Taboada Ramírez, Dr Ricardo Oropeza Navarro y Dr. Victor Manuel González Zúñiga.

A la M. en C. Maria Luisa Tabche y Dr. Raúl Noguez Moreno por su atención y paciencia

A la Dra Rosa María Gutierrez por sus atinados consejos y su valiosa amistad.

Al M. en C. José Ricardo Ciria Merce por su apoyo metodológico y técnico así como por sus atinados consejos y su sincera amistad.

A mis compañeros y amigos de laboratorio, sin ningún orden particular: Edgar, Natali, Maria del Carmen, Nori, Lizeth, Joselyn.

A Ingrid, por su invaluable e incondicional paciencia, apoyo, cariño y afecto en estos años.

y Finalmente al Dr. Enrique Merino Pérez, asesor de esta tesis, por cruzar la barrera de la academia y convertirse en un inigualable amigo.

Mi más sincera gratitud.

Walter Josué Hernández Santos

Índice

Índice de figuras

Índice de tablas

Presentación

El presente proyecto se enmarca en el área de la Genómica Computacional y Bioinformática y tiene como objetivo el analizar la relación que pudiera tener la presencia de chaperonas de RNA con la regulación genética mediada por sRNAs en bacterias.

La literatura actual sugiere que los procesos de regulación genética que implican en algún punto el uso de sRNAs dependen fuertemente de proteínas con actividad de chaperona de RNA. Sin embargo, estos estudios se han realizado únicamente en organismos modelo, o bien con un conjunto reducido de proteínas con actividad de chaperona. Por lo anterior, en este trabajo analizaremos computacionalmente si la presencia de sRNAs está supeditada en todos los casos a la presencia de Hfq, la principal y más estudiada chaperona de RNA, en fondos genéticos alejados de los organismos modelo. Con ello esperamos aportar información relevante que sirva para ratificar o refutar la veracidad de esta regla implícita en un contexto generalizado válido entre los organismos procariontes.

Este proyecto es una de las piezas iniciales de una nueva línea de investigación de nuestro grupo de trabajo cuyo objetivo central es el de implementar estrategias de biología sintética para analizar tipos novedosos de estrés generados artificialmente y regulados por sRNAs en bacterias, por lo que el entendimiento de estos mecanismos resultará de vital importancia para el avance de dicho objetivo.

Resumen

Este trabajo se realizó 3 en etapas diferentes. En la primera se utilizó una estrategia combinada de una búsqueda de blast y perfiles PFam usando de cadenas de Markov escondidas para determinar el patrón global de la presencia/ausencia de la chaperona Hfq o uno de sus homólogos en una base de datos inicial de 13,575 organismos obtenidos de los repositorios del NCBI. Los resultados apuntaron a una distribución poco homogénea de la presencia de las proteínas del CLAN CL0527 que incluye a la chaperona de sRNA Hfq.

En la segunda etapa se desarrolló una metodología computacional que permite predecir nuevos sRNAs usando como valor de entrada la secuencia genómica de cualquier organismo procariota. Este método está soportado en estudios de secuenciación masiva tomados del SRA (*Sequence Raw Archive*) del NCBI. El algoritmo desarrollado en este trabajo identifica características estructurales, energéticas y de secuencia de los sRNAs reportados actualmente para calibrarse y así lograr predicciones más acertadas que otros algoritmos existentes. Este método demostró alcanzar un 81 % de precisión al aplicarse a una lista control de sRNAs de *E. coli*. El algoritmo se formalizó en implementaciones Linux bajo licencia de software libre.

Finalmente, en la tercera etapa se construyó WebSRI, un servidor web desarrollado con la finalidad de hacer el algoritmo desarrollado en esta tesis accesible a usuarios que no cuentan con amplios conocimientos del sistemas operativos basados en el Kernel Linux.

Con los datos encontrados se concluye que la distribución de la chaperona Hfq no es universal como puede suponerse a partir de la busqueda bibliográfica y se abre un nuevo panorama de organismos que poseen, al menos en un nivel teórico, sRNAs sin tener una copia de la chaperona HFq.

Summary

This work was done 3 in different stages. In the first one, a blast-Pfam profile combined lookup was performed using hidden Markov chains to determine the general pattern of the presence/absence of the Hfq chaperone or one of its homologs in an initial database of 13,575 organisms obtained from the NCBI repositories. The results pointed to an inhomogeneous distribution of the proteins in the CLAN CL0527 that includes the RNA chaperone HFq.

In the second stage, a computational methodology was developed to predict new sRNAs using the genome sequence of any prokaryotic organism as an input value. This method is supported in massive sequencing studies taken from the SRA (Sequence Raw Archive) repository of the NCBI. The algorithm developed in this work identifies the structural, energetic and sequence characteristics of the sRNAs that are currently reported, to be calibrated and therefore, achieve more accurate predictions than other existing algorithms. This method proved to achieve 81 % accuracy when applied to an E. coli sRNA control list. The algorithm was formalized in Linux implementations under a free software license.

Finally, WebSRI was built in the third stage, a web server developed with the purpose of making the algorithm developed in this thesis accessible to users who do not have extensive knowledge of operating systems based on the Linux kernel.

With the data found, it is concluded that the distribution of the chaperone Hfq is not universal, as can be assumed from the bibliographic search and we give a new panorama of organisms that have, at least at a theoretical level, sRNAs without having a copy of the chaperone HFq.

1. Antecedentes

Los organismos existen en ambientes generalmente dinámicos en los que las condiciones varían de un momento a otro, en ocasiones sin un patrón aparente. Este dinamismo del ambiente produce una gran cantidad de efectos sobre el metabolismo de los organismos, en algunos casos hace menos disponibles algunos nutrientes, aumenta la concentración de partículas tóxicas o provoca cambios en las temperaturas típicas de crecimiento, entre otros muchos efectos. En general, este fenómeno en un contexto biológico se reconoce como estrés metabólico y en adelante en este documento se discutirán algunas de las respuestas regulatorias que los organismos bacterianos poseen y su implicación en la contención de los efectos negativos del estrés, haciendo énfasis en las respuestas que implican el uso de RNAs pequeños (*small RNAs* o sRNAs).

Los RNAs pequeños son cadenas de ácidos nucleicos de tamaños menores a 300 bases con actividad reguladora que participan en el mantenimiento celular en condiciones de estrés en organismos procariotas.[1,2] El entendimiento de los mecanismos mediante los cuales estas moléculas realizan su función ha permitido el desarrollo de novedosas tecnologías. Nuestro laboratorio ha generado circuitos reguladores basados en sRNAs artificiales diseñados a semejanza de los nativos de especies modelo[3], denominados riborreguladores, los cuales alteran la expresión de un gen blanco al unirse a la secuencia Shine-Dalgarno (SD) de su respectivo mRNA, inhibiendo de este modo el proceso de traducción. A la fecha, estos riborreguladores se encuentran en fase de prueba en el organismo modelo *E. coli*, sin embargo, en nuestro laboratorio buscamos aplicar en el corto o mediano plazo esta tecnología en mas especies de organismos.

La literatura actual se limita al análisis de los sRNAs en organismos modelo por su facilidad de cultivo y por el entendimiento previo de su actividad metabólica, como *E. coli*, en cuyo contexto se sabe que los sRNAs dependen de una proteína con actividad chaperona llamada Hfq. Sin embargo, la literatura no hace mención del escenario homólogo en organismos filogenéticamente alejados, por lo que antes de extrapolar la tecnología de los riborreguladores a estos, resulta crucial entender si estos organismos alejados tienen esquemas de regulación semejantes a los reportados en organismos modelo, así como detectar las particularidades que estos mecanismos, si los hubiera, puedan tener y que permitan abundar en la generación de herramientas de manipulación y análisis.

Un análisis inicial de nuestro laboratorio[4] realizado en base al genoma y proteoma de *Planctomycetes limnophilus* (*P. limnophilus*), una bacteria filogenéticamente lejana a organismos modelo como *E. coli* o *B. subtilis*, sugiere que en este organismo no existe un análogo a la chaperona de RNA Hfq, lo que podría representar una ventana de oportunidad para determinar los mecanismos por medio de los cuales este y otros organismos diferentes a los organismos modelo realizan sus procesos de regulación de respuestas a estrés y si estos involucran sRNAs.

2. Marco teórico

2.1. Redes de regulación en bacterias

Todos los organismos tienen la capacidad de *detectar* los cambios de su entorno que inducen presión de estrés en el metabolismo, de este modo, su sobrevivencia dependerá esencialmente de la capacidad de responder y adaptarse a dichos cambios.[5] Ante estas situaciones, los organismos han desarrollado sistemas de transducción de señales con objeto de coordinar variaciones metabólicas para así recuperar el estado de homeostasis, esta modulación puede ocurrir durante las etapas de inicio, elongación o terminación en los niveles transcripcional, postranscripcional y traduccional.[2]

A nivel transcripcional existen activadores (elementos de control positivos) o represores (elementos de control negativos). Algunos de estos elementos son específicos para un sólo gen, mientras que otros regulan un número mayor de genes, lo cual constituye un regulón. En el nivel postranscripcional han evolucionado diversos sistemas de regulación que buscan intervenir el proceso de transferencia de información desde el gen hasta la proteína funcional. Estos mecanismos controlan principalmente la estabilidad del mRNA y su tasa de inicio de traducción y, adicionalmente, tienen la capacidad de modular la estabilidad y actividad de las proteínas mediante modificaciones postraduccionales.

En bacterias, existen diferentes vías que determinan las respuestas que el organismo tomará en función del estímulo percibido, entre las cuales podemos mencionar: i) El control transcripcional mediado por factores sigma alternativos (ver Sección 2.3), ii) El control transcripcional mediado por Factores Transcripcionales (TFs), que pueden tener un efecto de activación o represión de la transcripción de acuerdo a la posición relativa de su unión en el DNA respecto al promotor del gen regulado, iii) El control al inicio de la traducción mediado por RNAs pequeños o sRNAS (de los que se habla en este documento en la Sección 2.4), iv) El control postraduccional mediado por chaperonas que ayudan al plegamiento adecuado de las proteínas y v) El control postraduccional mediado por proteólisis.

Debido a que en condiciones de estrés, estas respuestas son las más elevadas, a las redes de regulación de esta naturaleza se les conoce como *Sistemas de respuesta a estrés*. A partir de su origen, son muchos los tipos de estrés que enfrentan las bacterias, algunos de los más estudiados se enlistan en la Tabla 1.

Los sistemas de respuesta a estrés muestran una amplia similitud entre organismos procariontes, aunque en algunos casos, como el estrés calórico, los mecanismos son muy símiles con los propios de eucariontes y arqueas. Sin embargo, pese a la amplia similitud sistemática, los niveles de activación de la respuesta varían significativamente entre organismos.[6]

Tabla 1: Algunos tipos de estrés en bacterias y el tipo de respuesta asociados

Estímulo	Estrés	estrategia	efectores
Calor	HeatShock	Síntesis de proteínas de respuesta (Chaperonas y Proteasas)	σE,$\sigma 32$[7]
Inanición	*starvation*	Respuesta generalizada de estrés	RpoS (σ^S)[8] ppGpp
Hierro	por carga Iónica	Factores de transcripción (síntesis de solutos)	*kdp, porU*[9]
Presión	Turgencia	sistemas de dos componetes	KdpD[10]

2.2. El estrés en bacterias

2.3. Las respuestas bacterianas al estrés

Son muchas las estrategias que toman las bacterias para contender con el estrés, el uso de estas dependerá de la naturaleza del estrés y de la cantidad de decisiones que la célula tendrá que tomar para mantener su homeostásis. Las respuestas se pueden clasificar a su vez por el nivel en el que estas toman efecto. Pudiendo ser este el transcripcional, postranscripcional o traduccional. Entre las principales respuestas a nivel transcripcional se encuentran las siguientes:

- **Uso de factores sigma alternativos**
  Esta es la estrategia primaria para contender con el estrés a nivel transcripcional, esta regulación es caracterizada por el uso de distintos factores σ, que son péptidos que interactúan directamente con el núcleo catalítico de la RNA polimerasa formando así la holoenzima funcional. La afinidad de este complejo por los promotores que transcribirá depende entonces del factor sigma con el que esté formado. Así, los genes u operones que participan en una respuesta específica contendrán un promotor que será reconocido por uno de estos factores sigma alternativos.

  En condiciones normales (sin estrés), σ^{70} dirige la transcripción de los genes *housekeeping*, mientras que en condiciones de estrés, son factores alternativos quienes forman la holoenzima y logran dirigir la transcripción de genes de respuesta específicos,[11] entre los más estudiados se encuentran: σ^S; regulador maestro del estrés.[12] σ^{54} como regulador de genes asociados a nitrógeno, σ^{32} y σ^{24} que coordina los genes de respuesta a choque térmico, entre otros.[13–15]

- **Uso de represores**
  Este mecanismo se caracteriza por el uso de represores de la transcripción que se unen a un elemento de control del DNA para de este modo evitar su transcripción. Un ejemplo de este fenómeno es el represor HrcA que hibrida con un tipo particular de elemento de control conocido como CIRCE (*controlling inverted repeat of chaperone expression*) que se encuentra río arriba de los operones que codifican

para las proteínas de respuesta al estrés por calor o *heat-shock* principalmente en bacterias Gram-positivas y algunas Gram-negativas.

- **Proteólisis**
  Cuando así lo requiere, el organismo opta por degradar algunas de sus proteínas para de este modo regular sus funciones, un ejemplo de esta respuesta es el sistema SOS que se activa en respuesta a efectos genotóxicos y que es mediado por una serie de proteasas autorreguladas. Se ha propuesto que los sistemas de proteólisis permiten también regular la disponibilidad de los factores sigma y participar de modo adyacente en otros procesos de respuesta a estrés[16].

- **sRNAs**
  Los sRNAs coordinan respuestas específicas de estrés como las de fase estacionaria, estrés oxidativo por patogénesis, oxígeno, entre otros. De estos se habla con mayor profundidad en las siguientes secciones de este documento.

Todos estos elementos crean un complejo sistema de redes de regulación que permite a las bacterias adaptarse a los cambios de su ambiente y sobrevivir.

2.4. El papel de los sRNAs en la regulación genética

En un contexto generalizado, los sRNAs son porciones de RNA de longitud heterogénea entre 50 y 300 nucleótidos que tienen funciones regulatorias que ocurren en su mayoría como una consecuencia del apareamiento con la región Shine-Dalgarno de los mRNAs de los genes blanco, aunque muchos sRNAs también titulan la actividad de ciertas proteínas.[17]

Hasta ahora la vasta mayoría de los sRNAs que se han verificado experimentalmente tienen funciones de regulación genética en múltiples procesos celulares como control de la calidad de proteínas, resistencia ácida, balance de hierro y virulencia,[18,19] el mecanismo de acción es bastante variado, aunque en todos los casos se trata de la hibridación con el mensajero del gen blanco en la región Shine-Dalgarno, afectando de forma positiva o negativa el paso de la polimerasa y con esto la secuencia de procesamiento.[17,20,21]

Los sRNAs son codificados por el propio cromosoma bacteriano, aunque no necesariamente en regiones cercanas a los genes sobre los que ejercen su regulación, por lo que en algunos reportes se suele referir a ellos como *trans-encoded* sRNAs. Casi todos los sRNAs de este grupo se sintetizan bajo condiciones específicas de estrés.[22] Se ha referido frecuentemente a los sRNAs como análogos de los miRNAs y siRNAs de eucariontes (RNAs monocatenarios que interfieren en la expresión de diversos genes) dado que todas estas clases de RNA tienen funciones reguladoras. Sin embargo, los sRNAs difieren de los RNAs reguladores eucariontes en varios aspectos como su biogénesis, mientras que en los miRNAs y siRNAs se requiere del procesamiento de los transcritos por un intrincado sistema en el que intervienen complejos como DICER (nucleasas que degradan al precursor de doble cadena) y RISC (encargado de la ruptura de la doble hebra del precursor para liberar al siRNA)[23] encargados de procesar), en bacterias los

sRNAs se sintetizan como un solo fragmento que rara vez es procesado.[24]

En un nivel teórico, los sRNAs, a diferencia de otras estructuras genéticas no pueden ser determinados por criterios de similitud de secuencia, como los alineamientos locales o globales. Los sRNAs suelen estar bien organizados estructuralmente en dos dominios estructurales funcionales. A) un dominio de interacción que brinda especificidad para el reconocimiento del mRNA blanco y B) un dominio estructurado que consiste a su vez de dos elementos, el primero consta de un terminador rho-independiente formado por una estructura secundaria estable, mientras que el segundo es una región adyacente rica en Uracilos como se muestra en la Figura 1.

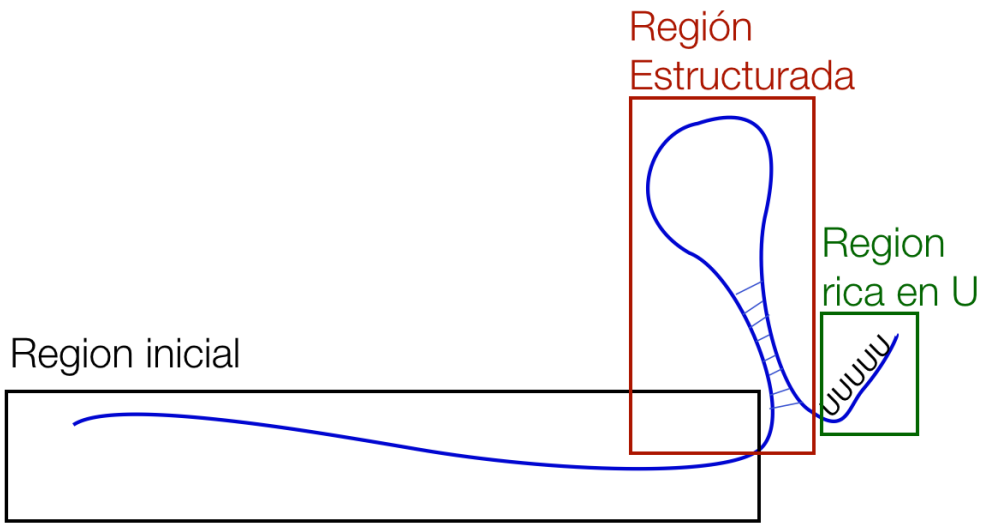


Figura 1: Modelo de un sRNA Rho-independiente

2.5. Los mecanismos de acción de los sRNAs

Como ocurre con todos los reguladores, los sRNAs son moléculas cuyos niveles de expresión cambian de un momento a otro. Cuando la concentración de estos se incrementa rapidamente, el efecto biológico se puede observar muy rápidamente, aunque contrario a lo que lo anterior pudiera sugerir.Los sRNAs poseen promotores que no difieren significativamente de los promotores de los genes que codifican para proteínas. El transcrito primario de un sRNA es, por lo general, activo para regulación sin la necesidad de procesamiento.[18] A pesar de lo anterior, existen casos reportados de transcritos de sRNAs procesados, pero a la fecha se desconoce si estos procesamientos influyen directamente en la actividad del sRNA.[1,25]

2.6. Los mecanismos de regulación de los sRNAs

La síntesis, estabilidad y concentración de los sRNAs son factores determinantes para que funcionen correctamente, por ello resulta importante entender las vías por medio de las cuales los sRNAs son regulados. La principal estrategia consiste en el uso

de proteínas reguladoras, algunas de las más importantes se muestran en la Tabla 2.

Tabla 2: Algunos ejemplos de circuitos reguladores de sRNAs

Familia de proteínas	Proteína reguladora	Estrés	sRNA	Blancos
LysR	OxyR[26]	Estrés Oxidativo	*OxyS*	*fhlA* *yobF* *wrbA* *ybaY*
	GcvA[27]		*GcvB*	*dpp* *opp*
Dos componentes	OmpR[28]	Choque osmótico	*OmrA* *OmrB*	*ampT* *cirA*
	RscB[29]	Superficie celular	*RprA*	*RpoS*
	LuxO[30]	Quorum sensing	Qrr -4 (Vc)	*hapR* *luxR*
Factor sigma	SigE[31]	Estrés periplásmico	*MicA* *RybB*	*ompA* *sigE*
Represor Fur	Fur[32,33]	Hierro	*RyhB* (Ec) (Vc) *PrrF* (Pa)	*sodB* *sdh* Proteínas de unión a Hierro
Familia Mar	Mar, SoxS,[34] Rob	Oxidativo, Antibiótico	*MicF*	*ompF*
CRP	CRP	Glucosa	Spot 42	*galK*

La región 5' no traducida (5' UTR) del gen *rpoS* es una de las más estudiadas como blancos de sRNAs, ya que es regulado positivamente por tres sRNAs: *DsrA*, *RprA* y *ArcZ*.[35] En la fase de crecimiento exponencial, estos se sintetizan en niveles bajos cuando la traducción de *rpoS* es inhibida por la formación de un tallo-asa del mRNA, lo que bloquea el acceso al sitio de unión al ribosoma (RBS). Cuando inicia la transición a la fase estacionaria, o bien ante un estímulo de estrés, estos sRNAs comienzan a incrementarse en síntesis e hibridan en la región líder del mensajero de *rpoS* inhibiendo de este modo la formación de la estructura inicial y haciéndolo accesible al ribosoma.

Algunos sRNAs tienen una complementariedad limitada con sus genes blanco, lo cual permite que en ocasiones un solo sRNA regule a más de un gen, tal es el caso de *RyhB* que participa en la regulación de hierro en bacterias.

2.7. La identificación de los sRNAs

2.7.1. Enfoque experimental

Aunque se sabe de la existencia de diferentes especies de RNA desde los primeros análisis de la estabilidad de este, el descubrimiento y comprobación del papel de estas especies ha resultado en un desafío complejo con muchos años de investigación invertidos. Los sRNAs se empezaron a describir de modo casual o con métodos experimentales poco usuales. Los primeros sRNAs de los que se tiene reporte son:

Spot 42, el primer sRNA reportado que se descubrió de modo fortuito en la bacteria *E. coli* durante la realización de experimentos de búsquedas genéticas con RNA radiomarcado que luego fue aislado de geles.[36]

OxyS, el segundo sRNA en ser publicado fue detectado mientras se buscaba mapear la transcripción de su gen regulador *oxyR*. Por su parte, el tercer sRNA, *csrB*, se detectó por co-purificación de la proteína CsrA.[37,38]

MicF, ***DicF*** y ***DsrA***, que fueron hallados usando mapeos con nucleasas e hibridacion RNA-DNA cuando se analizaba la fisiología de *E. coli* al ser afectada por los fragmentos genéticos que los contenían.[39–41]

En la actualidad la alta capacidad de secuenciación así como el desarrollo de técnicas y aplicaciones para el análisis de la expresión genética han permitido el desarrollo de númerosas estrategias para la identificación de nuevos sRNAs.[42,43] Se habla típicamente de un centenar de sRNAs en *E. coli* mientras que un número por mucho menor que comprende al resto de sRNAs pertenecen a un grupo reducido de organismos mayoritariamente representados por *Bacillus subtilis*, *Vibrio cholerae*, *Pseudomonas aeruginosa* *Staphylococcus aureus* y *Listeria monocytogenes*. La mayoría de estos sRNAs se descubrieron por medio de detección directa con microarreglos, aislamiento por *shotgun* y copurificación con proteínas. Por su parte, una minoría de sRNAs se ha descubierto por marcado directo o por búsquedas funcionales.[43,44]

Marcado directo y secuenciación
Los primeros enfoques experimentales que buscaban encontrar nuevos sRNAs utilizaban marcado con ^{32}P-ortofosfato, las bandas de geles que exhibían tamaños menores a 370 nt eran escindidas del gel para luego ser procesadas por nucleasas y secuenciadas (*fingerprinting*). Este método resultaba un poco lento e ineficiente, sin embargo sentó las bases de los futuros acercamientos. Una versión mejorada de este procedimiento buscaba extraer en un paso previo el RNA de la muestra, sin embargo no contaban con la inestabilidad del RNA al momento de la extracción por lo que haciéndolo de este modo, la calidad de la técnica se reducía considerablemente. Un enfoque diferente consideraba el uso de bromuro de etidio para marcar los sRNAs, con este método se determinaron los sRNAs *BS190* y *BS203* de *B. subtilis*.[45,46]

Copurificación
Desde su descubrimiento como un *host factor* para el bacteriófago Qβ, la chaperona Hfq se asoció como una molécula necesaria para la estabilización de sRNAs, por lo que un paso lógico consistió en buscar a los sRNAs en experimentos de asociación con esta chaperona.[1,44]

Una aproximación frecuentemente utilizada en la década pasada se basó en el método SELEX,[47] un procedimiento experimental que permite la extracción de oligomeros con afinidad por una molécula blanco a partir de un *pool* de oligonucleótidos. En este caso la base del procedimiento fue la transcripción con la T7 polimerasa de librerías de secuencias de *E. coli* de longitudes aleatorias de 50 a 500 bases que luego fueron incubadas con Hfq[48] La ventaja de esta aproximación es que genera porciones de RNA de todo el genoma, por lo que no depende del estímulo de estrés.[49] Adicional a los anteriores, una minoría de sRNAs se han detectado por coinmunprecipitación con sus blancos.[50]

Microarreglos
Los microarreglos permiten el análisis simultáneo de la expresión genética de un genoma completo, sin embargo, los microarreglos convencionales están diseñados para detectar la expresión de los marcos abiertos de lectura, por lo que esta aproximación requirió del uso de sondas diseñadas para ambas cadenas de las regiones intergénicas. Estas se diseñaron a partir de co-inmunoprecipitaciones de Hfq. En combinación con métodos de genómica comparada, estas técnicas permitieron la detección de muchos sRNAs relativos a la fase de crecimiento.[44,51] Los microarreglos permitieron la identificación de algunos sRNAs. Sin embargo, el tamaño tan pequeño de los sRNAs los hacen malos candidatos de amplificación.

Adicional a los enfoques mencionados arriba, los sRNAs también han sido buscados por clonación por shotgun, sin embargo, en esta aproximación se requiere de un gran número de análisis debido al gran número de clonas que se producen debido al uso de fragmentos de longitudes heterogeneas propias del método.[51]

2.7.2. Enfoque teórico-computacional

Los primeros métodos de búsqueda de sRNAs que utilizaron una estrategia computacional buscaban estructuras de conservación de secuencia en las regiones intergénicas de bacterias filogenéticamente cercanas. Estos métodos permitieron predecir con un aceptable grado de precisión 42 nuevos sRNAs en *E. coli*, sin embargo, cuando estos se implementaron en organismos filogenéticamente alejados entre sí, la calidad de las predicciones disminuyó considerablemente. Dos de estos algoritmos[42,44] basaban su búsqueda en la identificación de regiones con alto grado de conservación de secuencia, mientras que el tercero[52] además de esto restringía la búsqueda a regiones con tamaños menores a 300 bases que pudieran ser marcados como estables estructuralmente.

Con el abaratamiento de los sistemas de cómputo de alto rendimiento se desarrollaron nuevos métodos para predecir sRNAs que quedarían en calidad de putativos hasta su verificación experimental, las bases de análisis de todos ellos fueron variadas desde implementaciones comparativas hasta métodos que utilizaban Inteligencia Artificial (IA) con Maquinas de Soporte Vectorial (MSV) cuya base matemática busca aumentar la predicción a partir de calibraciones con un set positivo. Cada uno de ellos permitió avanzar en la búsqueda de sRNAs nuevos. Los paquetes más importantes se mencionan en la Tabla 3.

Otros algoritmos cuyo enfoque principal no es el de predecir sRNAs, pero que se pueden aplicar a la búsqueda de estos se han propuesto en los últimos años, el más importante es SPAR, que a partir de alineamientos de metagenomas busca predecir nuevos marcos de lectura correspondientes a cualquier especie de RNA no reportada.

Tabla 3: Principales algoritmos de prediccion aplicados a la búsquieda de nuevos sRNAs

Método base	Herramienta	Principales características
Genómica comparativa	QRNA[53]	Busca predecir estadísticamente nuevos ORFs
	ERPIN[54]	Usa alineamientos múltiples para inferir perfiles de estructura
	ISI[55]	Predice la conservación intergenética
	MSARI[56]	Usa métodos mixtos de distribución
Estructura del RNA	RNAz[57]	Usa IA con MSV para encontrar estructuras
Señales de transcripción	sRNAscanner[58]	Usa datos *training*
	sRNApredict[59]	Busca promotores por señales de transcripción
Independientes de secuencia	PsRNA[60]	Usa ortologos de KEGG para calibrar las regiones intergénicas
	NAPP[61]	Calcula la ocurrencia de regiones de 50 nt en 1000 genomas

2.7.3. Los retos de identificación de los sRNAs

A pesar de la creciente cantidad de métodos que buscan determinar nuevos sRNAs, el problema permanece aun sin una estrategia que lo resuelva de forma definitiva, sobre todo cuando estas búsquedas se realizan en organismos atípicos. En la mayoría de los casos, de forma experimental resulta imposible la identificación de los genes responsables de los sRNAs por métodos como las mutaciones sin sentido y por inserción debido a condiciones como el tamaño del sRNA o el tamaño de los genes que los codifican.

La bioinformática tomó como materia prima la enorme cantidad de datos que se produjo gracias al advenimiento de las técnicas de secuenciación masiva, permitiendo que se identificaran como funcionales secuencias que se encontraban contenidas entre un promotor y un terminador que no representaban, al menos teóricamente, marcos de lectura (*Open Reading Frame* u ORFs) codificados.

El problema actual es el de identificar correctamente a un sRNA, tanto computacional como experimentalmente. En principio, algunos de ellos solo se sintetizan bajo condiciones específicas, su tamaño es variable y algunos de ellos son reacios a los métodos convencionales de extracción y purificación gracias a sus estructuras tridimensionales, por lo que no hay una regla general para identificarlos.[62] Dado que no son secuencias codificantes no es posible identificarlos por medio de búsquedas de ORFs y en el caso de la regulación positiva por sRNAs la formación de la doble cadena ocurre lejos del RBS y del sitio de inicio de la traducción.

Considerando su conservación restringida a especies cercanas, resulta imposible identificar de forma generalizada a los sRNAs por análisis de homología con respecto a los ya reportados. En un principio, estas restricciones permitieron únicamente hacer búsquedas en el organismo modelo *Escherichia coli*; sin embargo, cada vez resulta mas importante la búsqueda de estas moléculas en organismos alejados.[63] A pesar de estas limitaciones, los esfuerzos computacionales han tenido aciertos y actualmente utilizan las estrategias antes mencionadas combinadas con nuevos métodos como el *machine learning* y la predicción de estructuras secundarias para proponer candidatos,[64] aunque generalmente la tasa de éxito se mantiene relativamente baja mientras que los costos computacionales de estos métodos siguen incrementándose.[63] Por todo lo anterior resulta de gran importancia el estudio de sRNAs en diferentes contextos, en la Figura 2 se muestra el número de publicaciones que involucran sRNAs que se han predicho según *Web of knowledge* al 10 de diciembre de 2018.

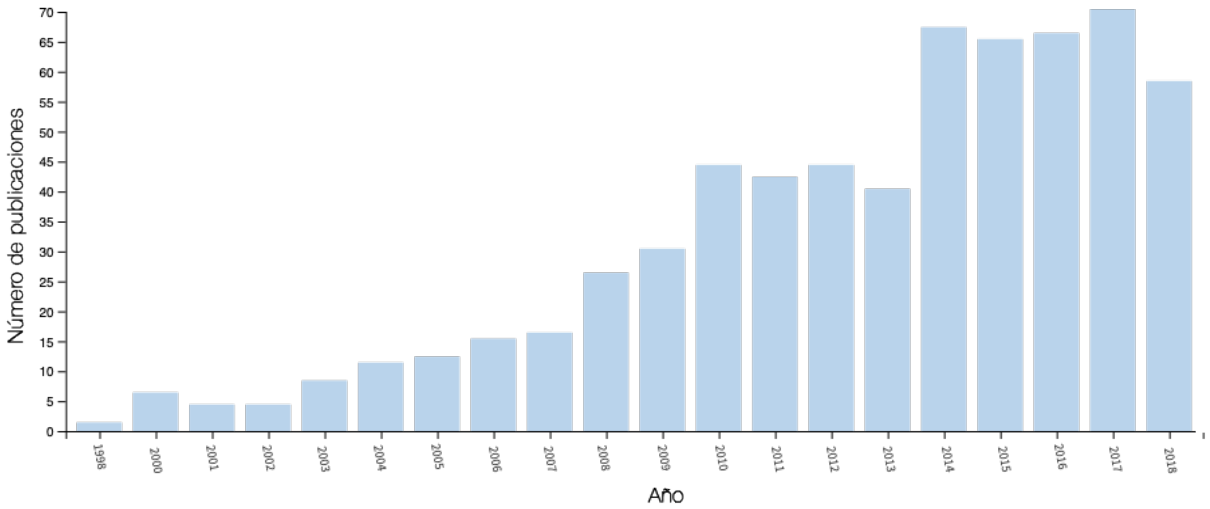


Figura 2: Progresión de estudios de sRNAs en WoK.

2.8. Las chaperonas de RNA

Debido a la naturaleza propia del RNA, este exhibe una alta capacidad de formación de estructuras de doble cadena debido principalmente a la complementaridad de las bases que lo constituyen.[65]

La capacidad biológica del RNA dependerá muchas veces de su estructura tridimensional, sin embargo, dada su alta energía , los RNAs pueden plegarse incorrectamente y propiciar la formación de estructuras secundarias o terciarias termodinámicamente estables pero inadecuadas o incorrectas que pueden fácilmente llevar a la pérdida de la función biológica.[66]Adicionalmente el proceso de plegamiento correcto del RNA puede pasar por varias estructuras intermedias con alta estabilidad, lo cual se traduce en una posibilidad de que el proceso se torne lento o ineficiente, dicha hipótesis ha sido abordada y demostrada como verdadera en estudios *in vitro*.[67]

Cuando los procesos de plegamiento de los ácidos nucleicos se analizaron en su contexto nativo, la duración de estos fenómenos fue mucho menor que su escenario homólogo *in vitro*, por lo que inicialmente se pensó que había más factores involucrados en estos procesos. Así fue como eventualmente las investigaciones señalaron a un grupo heterogéneo de proteínas que no comparten estructura, secuencia ni motivos funcionales,[68] pero que apoyan al RNA a llegar a su estructura funcional y estable. Estas proteínas son conocidas como Chaperonas de RNA y suelen ser definidas estrictamente como proteínas que se unen de manera transitoria al RNA en un modo inespecífico y que permiten resolver en términos cinéticos las estructuras incorrectamente plegadas de este.[69]

Las chaperonas de RNA tienen la particularidad de unirse de manera débil e inespecífica a las moléculas de RNA sin requerimientos energéticos adicionales (como los provenientes de la hidrólisis de ATP), y que una vez que la estructura está resuelta estas ya no participan en el fenómeno biológico.[70,71]

A la fecha, se distinguen varias clases de proteínas que forman parte del grupo de chaperonas de RNA. De acuerdo con *The RNA Chaperone Activity Website* (www.projects.mfpl.ac.at/rnachaperones)[69] estas son:

- **hnRNPs**: Se encuentran por lo general en organismos eucariotas, uno de los ejemplos más importantes de esta familia es YRA1 que forma parte del complejo TREX en *Saccharomyces cerevisiae* y que está involucrada en los procesos de exportación del mRNA[72]
- **Porteínas contenedoras de dominios cold-shock**: Son chaperonas con funciones involucradas en los procesos de resistencia al estrés por frío, las más estudiadas son CspA de *E. coli* y CspB de *B. subtilis*.[73,74]
- **Proteinas ribosomales**: Son proteínas que forman parte del ribosoma, de estas la más frecuentemente estudiada es la proteína S1 que forma parte de la

subunidad ribosomal 30S y que participa de manera importante en la traducción de muchos mRNAs, se une arriba del Shine-Dalgarno y permite la formación o estabilización de la estructura de estos mRNAs para su correcta traducción.[75]

- **Proteínas bacterianas semejantes a histonas**: Son proteínas con alta homología con las histonas bacterianas, en el caso de StpA, esta también puede unirse a RNA no estructurado para promover su correcto plegamiento[76]

- **Proteínas de nucleocápsides virales.**: Por lo general presentes en eucariontes, el ejemplo representativo de este grupo es HCV, una proteína presente en el núcleo del virus de la hepatitis, donde funge como constituyente de la cápside que resguarda el RNA viral[77]

- **Proteínas RCA huérfanas**: Son proteínas que no se han asociado a otras funciones, o que se han asociado unicamente en forma teórica. De este, la más estudiada es FinO, una proteína presente en *E. coli* que está involucrada en la facilitación de las interacciones de RNA del tipo sentido-antisentido.[78] Otros ejemplos de esta familia son DnaX, Cyp55 y Ltp28.[69]

- **Proteinas SM-like** : Son un grupo de chaperonas que se han asociado, entre otras cosas, con la interacción con sRNAs. Están presentes en organismos de los tres dominios. De este grupo la más estudiada es la chaperona Hfq de la que se habla mas adelante en este documento.

Cada una de estas subclasificaciones involucra diferentes proteínas que están presentes en diferentes organismos, por ejemplo, en eucariotas se ha reportado que las versiones de las proteínas SM juegan papeles relevantes en procesos como el splicing alternativo y otras funciones ajenas a los sRNAs.[79]

2.9. Chaperonas de sRNAs

Una clase particular de chaperonas RNA ejercen su función al favorecer la interacción de los sRNAs y sus blancos de regulación (mRNAs). Estas proteínas pertenecen a la categoría SM-like que se ha identificado en Pfam como un CLAN (CL0527) que incluye a 5 familias de proteínas: DUF903 (PF06004) de función desconocida, SM-ATX (PF14438), LSM14 (PF12701), LSM (PF01423) y Hfq (PF17209), el clan es llamado LSM (*Like-Sm*) por su similitud estructural con los motivos encontrados en muestras de antígenos de una paciente de lupus llamada Stephanie Smith.[79,80]

En muchas revisiones de la literatura se sugiere que estas proteínas son esenciales para la correcta resolución de los fenómenos de regulación por sRNAs. De estas, la más estudiada es Hfq, una chaperona relacionada íntimamente con el bacteriófago Q, de la que se habla en la Sección 2.9.1

2.9.1. La chaperona Hfq

Hfq, también mencionada en la bibliografía como HF-I,[81] fue encontrada a mediados del siglo pasado como un *Host Factor* del bacteriófago Q (de donde su nombre fue derivado[82,83]), es una proteína con actividad de chaperona de RNA que forma una estructura cuaternaria toroidal en forma de dona donde se piensa reside su capacidad de interactuar con moléculas de RNA, su función principal es la de favorecer la interacción de los sRNAs con los mRNAs. Dichos sitios de interacción se han podido mapear de forma computacional solo para un número limitado de ejemplos pero se piensa que este sitio debe ser rico en A y U.[84] El sRNA y el mRNA se unen a Hfq en diferentes caras[83] y hay reportes que sugieren que estas moléculas atraviesan un periodo de competencia por la unión inicial con la chaperona.[85]

El complejo Hfq-sRNA tiene varios escenarios y funciones posibles. Primero, Hfq puede ayudar a la interacción entre el sRNA y su mRNA blanco, comúnmente en una región que contiene el RBS y de este modo inhibe el inicio de la traducción de la proteína[2] (Figura 3a). Por otro lado y de modo controversial, Hfq también puede promover la síntesis de proteínas al facilitar la unión del sRNA con mRNAs blancos cuya traducción es inhibida por estructuras secundarias en su región líder que incluyen al RBS en tallos estables, y por tanto de manera normal no son traducidos. En este caso, la unión del sRNA con la región 5' del mRNA desestabiliza la estructura de inhibición traduccional, lo que resulta en la síntesis de la proteína[86–88] (Figura 3b).

Durante su estadía en el citoplasma, los sRNAs pueden protegerse de la degradación por ribonucleasas cuando, mediados por Hfq forman un complejo estable (Figura 3c), aunque cuando se requiere que un sRNA sea degradado es Hfq quien favorece la interacción de los sRNAs con las ribonucleasas (Figura 3d). Adicionalmente a estas formas de regulación debidas a interacción mediada por Hfq del sRNA con su blanco. Hfq también puede promover la poliadenilación del extremo 3' del mRNA para que este sea degradado por exonucleasas (Exo en la Figura 3e) del tipo 3'-5'. El desempeño de Hfq dependerá de la información codificada en la secuencia del RNA con que interactúa. Ejemplos de sRNAs que tienen funciones reguladoras positivas y negativas simultáneas se han reportado escasamente.[88]

La mayoría de publicaciones hacen referencia a la interacción de Hfq con sRNAs y mRNAs a pesar de la basta cantidad de clases diferentes de sRNA que existen. Aunque se sabe que Hfq se asocia también con otras clases de RNA[1,89,90] entre las que se encuentran *RNA decoys*, y transcritos de función dual (mRNA o sRNA).[91]

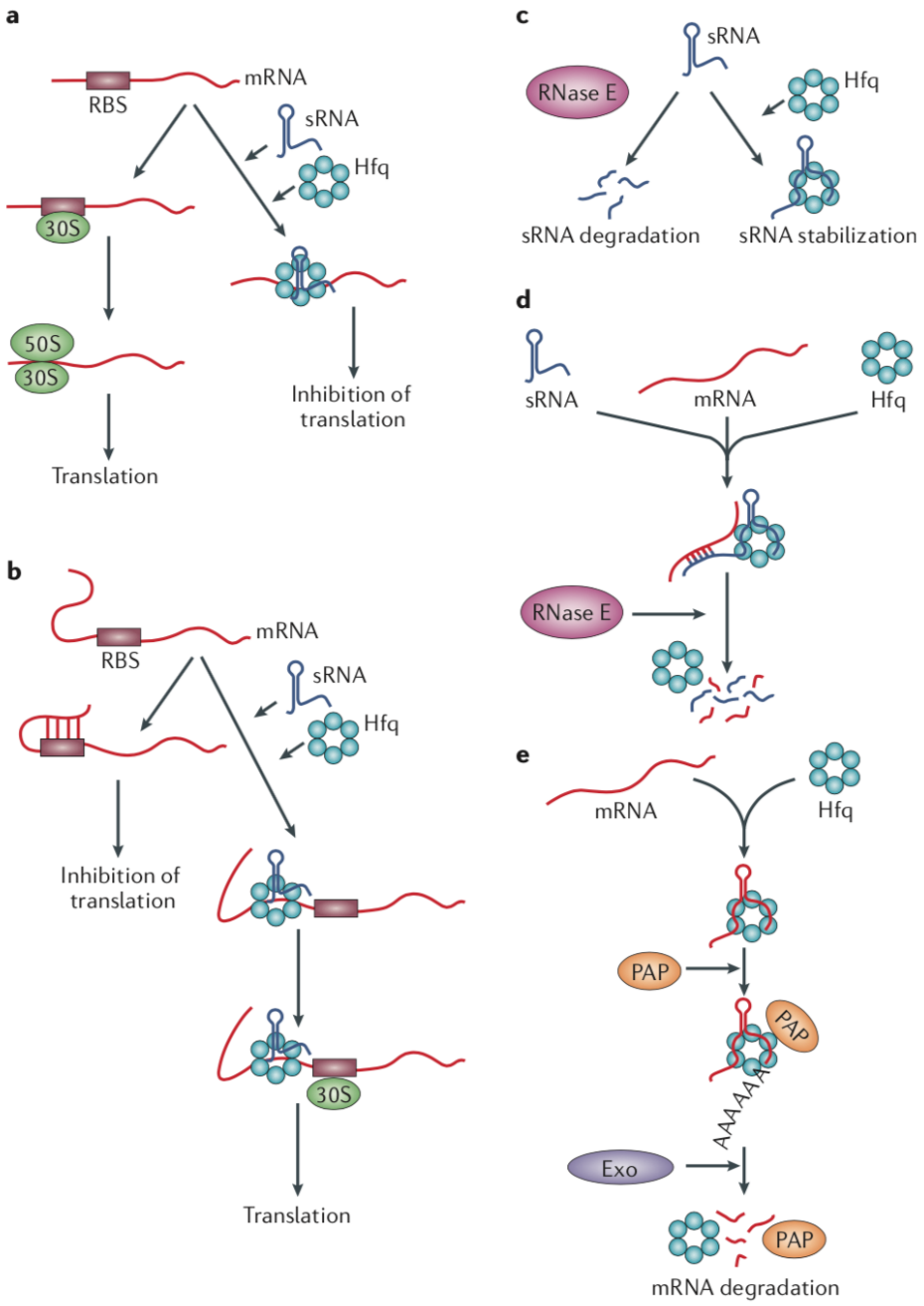
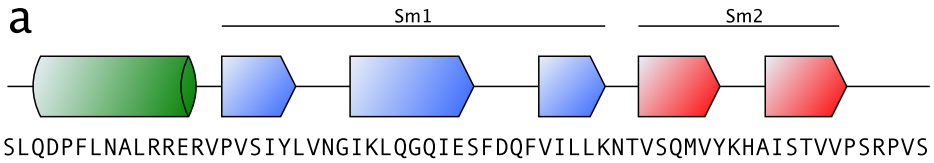


Figura 3: Diferentes mecanismos de acción de Hfq en la regulación genética, tomado de Vogel y Luisi (2011).[82]

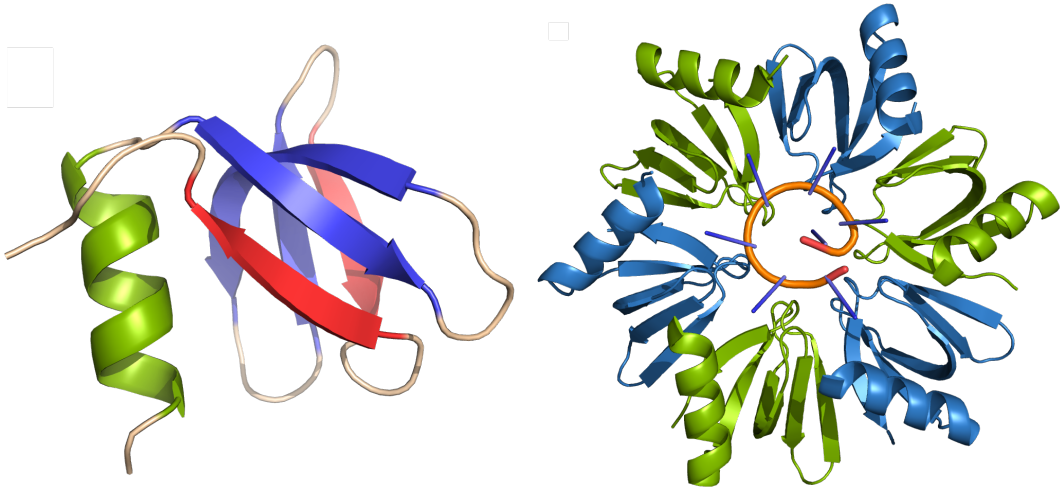
El gen *hfq* codifica para unidades proteicas que se unen para formar el complejo funcional en forma de anillo, cada una de estas unidades también llamadas protómeros[92] tiene una longitud típica de 65 residuos, (la secuencia común se muestra en la Figura 4a) y está formada por un α hélice (verde en la Figura 4a y b) y 5 hojas β antiparalelas (flechas en la Figura 4a y b). Hay dos motivos estructurales bien caracterizados sobre la secuencia de Hfq llamados Sm1 y Sm2 y son comunes en toda la familia de proteínas Sm-Hfq-LSm.[93]

Estas estructuras están altamente conservadas en la mayoría de las proteínas Hfq de arqueas y bacterias a pesar de las distancias genéticas asociadas a divergencia evolutiva con la excepción del dominio Sm2 que presenta algunas variaciones interespecie y que se ha asociado con el número de unidades formadoras de complejo.

El complejo funcional, de aproximadamente 65 angstroms, está compuesto generalmente por 6 protómeros idénticos (Figura 4c) aunque en algunas ocasiones el complejo está formado por 6 protómeros diferentes o bien por 7 subunidades no necesariamente idénticas, como ocurre con la versión de la proteína en humanos. Los hetero-oligómeros también pueden presentarse en bacterias que expresan múltiples homólogos de Hfq o por la unión de proteínas Hfq que se sintetizan con doble carga de los dominios Sm.[82,83]



(a)



(b) (c)

Figura 4: Hfq y su modelo de interacción con sRNA. a) definición de la estructura secundaria del modelo basado en la proteína de *E. coli*.[94] b) estructura tridimensional de un protómero de *E. coli*. c) interacción de Hfq (PDBID: 1KQ2) con sRNA (naranja) por la región proximal, los protómeros están distinguidos en colores diferentes.

El complejo proteico dispone estructuralmente de dos caras y en su centro se forma un poro por medio del cual interactúa con las moléculas de RNA, la cara proximal es responsable de una interacción mas fuerte pues cada uno de los protómeros puede interactuar con 3 bases del ácido nucleico, en contraste con la porción distal en la que

esta interacción se limita a un contacto por protómero. La porción distal parece preferir los motivos de secuencia AAYAAYAA, ARN o ARNN (donde R equivale a una purina y N a cualquier base).[94]

En algunas clases de bacterias Hfq tiene una extensión carboxy-terminal sin estructura aparente de hasta 100 residuos adicionales al núcleo conservado.[95,96] Diversos estudios sugieren que este dominio tiene un fuerte impacto en la regulación del comportamiento de Hfq,[97] la mayoría de las estructuras depositadas en el PDB carecen de este segmento.[98]

3. Justificación

A pesar de la importancia de los procesos de regulación por sRNAs, estos se han estudiado principalmente en organismos modelo pues por lo general en ellos se tienen bien caracterizadas sus funciones bioquímicas, lo cual facilita el análisis de los resultados obtenidos. Sin embargo, no se cuenta con estudios actualizados que permitan obtener una imagen global de la distribución filogenética de la proteína Hfq o cualquiera de sus homólogos ni un análisis crítico de las implicaciones que su ausencia tendría en la regulación de la expresión genética en procesos de respuesta a estrés a los que podría estar asociada.

Asimismo, la dificultad de detectar experimentalmente la identidad de las proteínas como chaperonas de RNA genera una brecha considerable entre la información que se dispone para la proteína Hfq de la que se tiene para otras familias de chaperonas como la familia FinO, por lo que resulta preciso considerar estas proteínas como alternativas a las chaperonas SM-like al menos en estudios teóricos.

Aunque nunca se menciona de manera explícita, la naturaleza de los análisis actualmente reportados sugieren que Hfq, es un interactor crucial en fenómenos de modulación genética por sRNAs. Sin embargo, experimentos bioinformáticos realizados en nuestro laboratorio sugieren que esta proteína esta se encuentra está ausente en al menos un grupo de organismos alejados filogenéticamente de los modelos tradicionales como *E. coli*. Por lo anteriormente expuesto, resulta primordial una búsqueda global de esta proteína y de las proteínas representativas de otras familias en un análisis exhaustivo en organismos y entender si su ausencia en ciertos organismos bacterianos conlleva también a la ausencia de la regulación mediada por sRNAs.

A la fecha existe un número pequeño de programas computacionales y algoritmos que buscan determinar nuevos sRNAs, sin embargo, el porcentaje de éxito predictivo de estos se mantiene bajo mientras que el nivel de conocimientos computacionales requeridos dificultan su uso por un publico más generalizado que carece de conocimientos avanzados de informática.

Así, con base a lo dicho anteriormente, es clara la necesidad de desarrollar un método de análisis que permita identificar sRNAs en genomas bacterianos que resulte eficiente y fácilmente implementable. De este modo se podrán correlacionar la presencia de estas chaperonas de sRNAs con los sRNAs que estos organismos pudieran tener, al mismo tiempo que se podría abundar en el análisis de las diferencias y semejanzas que existan entre los sRNAs de diferentes phyla de organismos.

Por los motivos anteriores, este trabajo busca aportar información acerca de la distribución filogenética de las proteínas con actividad de chaperona de sRNA y busca generar una metodología que permita identificar fácilmente potenciales sRNAs. De este modo, ofreceremos un panorama amplio de los fenómenos de regulación que abra la puerta a futuros análisis y métodos mediados por chaperonas de RNA y sRNAs.

4. Hipótesis

Existen sRNAs reguladores en organismos que carecen de la chaperona de RNA Hfq.

5. Objetivos

5.1. Objetivo general

Caracterizar si la presencia de sRNAs en genomas bacterianos se da de manera exclusiva en aquellos que posean el gen que codifica para la chaperona de RNA Hfq.

5.2. Objetivos particulares

- Determinar de forma computacional la distribución de la chaperona Hfq, en base a búsquedas genómicas.
- Desarrollar un método computacional eficiente y reproducible que permita predecir sRNAs en los genomas bacterianos.
- Analizar la relación de las distribuciones de Hfq y de sRNAs a diferentes niveles taxonómicos.
- Analizar la distribución de las proteínas con actividad de chaperonas de RNA que usen como sustrato sRNAs diferentes a Hfq.

6. Estrategia experimental

Para alcanzar los objetivos listados anteriormente se diseñó e implementó una estrategia metodológica con un enfoque teórico computacional que se resume en las siguientes secciones, los resultados de tal estrategia se detallan en la secció̀in 7 de este documento.

6.1. Identificación de homólogos de Hfq en los diferentes *Phyla* bacterianos

Para identificar potenciales homólogos de Hfq y su distribución bacteriana, como un paso inicial se realizaron búsquedas de secuencias pareadas (por medio de BLAST) y búsquedas usando modelos de cadenas de Markov escondidas (HMMs) usando el programa hmmsearch y tomando como punto de partida los datos que se encuentran en la base de datos de PFam sobre los genomas secuenciados de las bacterias disponibles en los repositorios del NCBI.

6.2. Identificación de sRNAs

Se identificaron sRNAs nativos en los genomas bacterianos por medio de la localización computacional de elementos característicos, como promotores y terminadores rho-independientes que se encuentren dispuestos a distancias menores a 300 pb esta búsqueda fue posteriormente complementada con el cálculo de la energía libre del RNA, lo que que permitió discriminar formas comunes a los sRNAs consistentes con las reportadas en la bibliografía.

6.2.1. Identificación de terminadores transcripcionales rho-independientes y sitios de inicio de transcripción

Para obtener valores de predicción significativos, utilizamos el algoritmo desarrollado en nuestro grupo que se basa en una modificación de un método *ad hoc* para la identificación de atenuadores transcripcionales[99] complementándolo con el análisis de datos de secuenciación masiva, de este método se habla en detalle en este documento en las secciones 7.7.2 y 7.7.8.

7. Resultados

7.1. Construcción de base de datos y búsquedas pareadas

Se obtuvieron las secuencias de los genomas y proteomas, así como los registros de secuenciación para arqueas y bacterias alojados en el repositorio `ftp` del NCBI (ncbi.nlm.nih.gov/genomes/). Se identificaron redundancias de genomas y cuando los hubo se eligieron los datos del organismo clasificado como *reference* para un total de 13,575 organismos repartidos en 367 arqueas y 13,208 bacterias. Con los datos antes mencionados se realizó una búsqueda global por medio del algoritmo BLASTP 2.2.28+ con un valor de corte (e-value) de $1x10^{-5}$. Para ello se tomó como referencia la secuencia de Hfq de *Escherichia coli* (NP_418593.1 UniProt: P0A6X3.2, Anexo 11.1.1) y la versión de LSm de *Thermovibrio ammonificans* (WP_013536983.1, Anexo 11.1.2).

Como resultado de este análisis se encontró la que distribución de la chaperona Hfq está restringida a solo 8 *phyla*, mientras que su homólogo Sm, dió positivo solo en 7. Una interpretación gráfica de este resultado se puede observar en la Figura 5.

El análisis de este resultado mostró organismos para los cuales la búsqueda arrojó un valor positivo para ambas proteínas, lo que sugiere la presencia de organismos con copias de ambas proteínas. Para corroborar esto se decidió realizar un análisis más robusto usando modelos de Pfam como se detalla en la Sección 7.2.

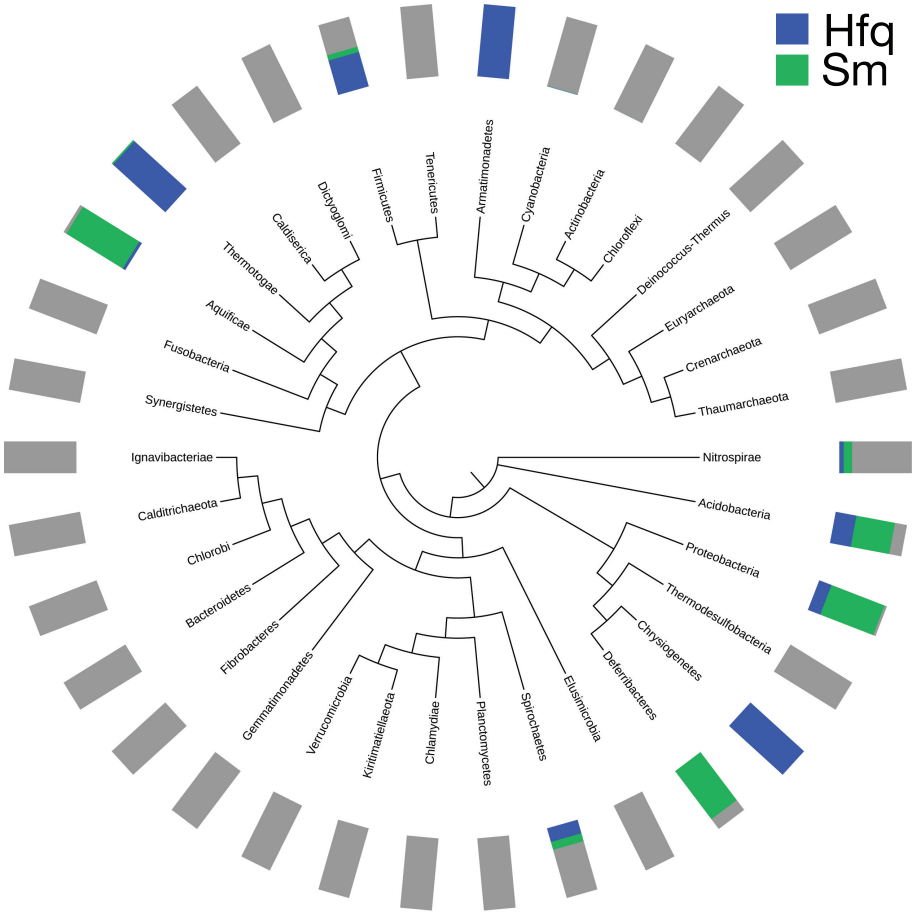


Figura 5: Distribución de las proteínas Hfq y Sm por Phylum identificadas usando búsquedas pareadas con el programa BLAST. Las barras grises indican ausencia de HFq o Sm. Las azules y verdes representan la proporción de las distribución encontrada.

7.2. Búsquedas de Hfq y sus homólogos usando modelos de Markov escondidos

En la siguiente etapa se realizaron búsquedas usando los modelos de cadenas de Markov definidos en la base de datos de Pfam (http://pfam.xfam.org)[100] para cada una de las proteínas del clan LSM (CL0527). Los detalles de dichos modelos se encuentran en el Anexo 11.2.

Los análisis de búsquedas ayudados por matrices se realizaron con tres niveles de astringencia, estos fueron: sin restricciones de e-value y con valores de corte de $1x10^{-8}$

y $1x10^{-15}$, los resultados se compararon con las distribuciones presentadas por PFam en su sitio web (`pfam.xfam.org`). La comparativa permitió descartar los dos primeros bloques de resultados para finalmente conservar los generados con valores de corte de $1x10^{-15}$. Como control negativo, no se encontró presencia significativa para los grupos LSM14 y LSM-ATX en los genomas procariotas de la base de datos construida en este trabajo, lo cual concuerda con la bibliografía, pues estos se han reportado únicamente en genomas eucariotas. La distribución de las proteínas del CLAN LSM se muestra en la Figura 6. La visualización de estos datos se realizó con ayuda de la plataforma iTOL.[101]

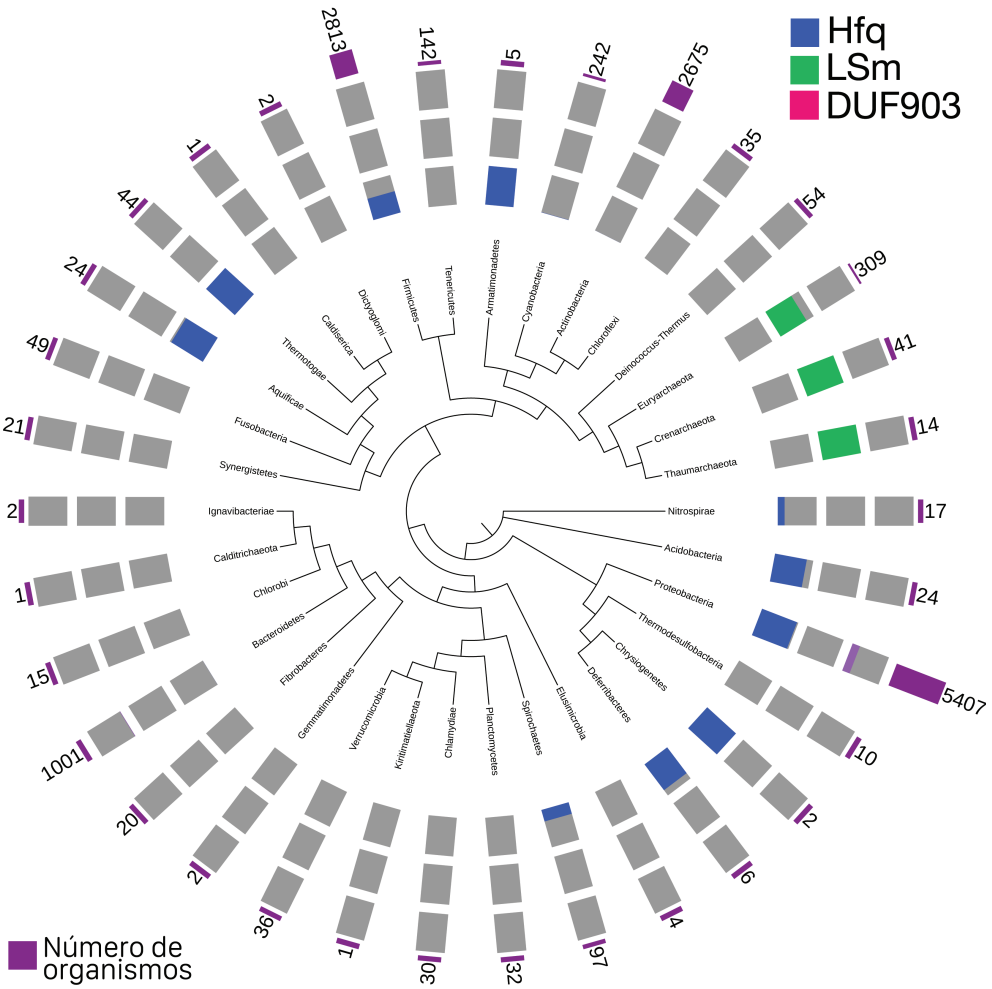


Figura 6: Distribución de las proteínas del Clan LSM en arqueas y bacterias organizados por Phylum. En cada anillo de barras se encuentra representada la proporción encontrada en el análisis de cada perfil de proteínas estudiado. El color gris representa la ausencia de *hits*.

7.3. Análisis filogenético y construcción de árboles

Con todos los datos producidos en los puntos anteriores se obtuvo una imagen global de la cobertura y distribución del CLAN LSM en arqueas y bacterias. Para representar las proporciones en un formato informativo se utilizó una estrategia de representación mediante árboles filogenéticos. Para esto, el primer paso fue el de elegir al representante filogenético de cada *phylum* y a partir de éstos construir una matriz de distancia y su árbol filogenético asociado. La aproximación inicial fue la de escoger al organismo de cada grupo taxonómico con mayor número de proteínas, quedando seleccionados los que se enlistan en la Tabla 4.

Con los organismos representantes de cada grupo se utilizó una metodología basada en el trabajo del grupo de Peer Bork[102] como se explica a continuación:

Se buscaron las proteínas de los organismos que coincidieron con los 31 núcleos de genes (Clusters of Orthogous o COGs[103]) que han sido reportados como marcadores de distancia genética por medio del algoritmo HMMRSEARCH en su versión 3.1b1. Las secuencias de proteínas identificadas para cada COG de estudio, se alinearon por medio de HMMALIGN en la misma versión. Las columnas con un contenido de gaps mayor o igual 25 % fueron descartadas por considerarse poco informativas para el alineamiento. Posteriormente, la secuencia de cada uno de los 31 bloques fueron colocadas en orden sucesivo para construir un *mega-alineamiento* con las secuencias concatenadas. Dicho mega-alinemiento fue usado como dato de entrada para el cálculo de distancias filogenéticas usando el programa Protdist de la suite Phylip.[104] Dichas distancias fueron usadas por el programa Neighbor de la misma suite Phylip que utiliza el método Neighbor Joining para la construcción de un árbol filogenético. Un esquema general del proceso se muestra en la Figura 7.

Sin embargo, cuando los resultados de esta caracterización se compararon con los reportes puntuales del NCBI taxonomy, el método demostró tener inconsistencias con las reportadas en este sitio. Por lo que esta aproximación se complementó con una estrategia basada en la secuencia de los genes 16S rDNA. Dicha estrategia es en esencia el mismo principio que la versión anterior con la diferencia de que esta parte de secuencias de RNA, para alinear dichas secuencias se utilizó el algoritmo **ssu-align**,[105] mientras que la construcción de la las relaciones filogenéticas se calcularon con los paquetes jmodeltest[106] y phyml.[107]

7.4. Extrapolación a diferentes niveles taxonómicos

Los resultados de los análisis de los puntos anteriores sugieren en su conjunto que la distribución de la proteína Hfq y sus homólogos es particularmente diferente a la distribución generalizada que suponen diferentes revisiones de la regulación por sRNAs en bacterias[18]. Por tal motivo, se repitieron los análisis realizados en los pasos anteriores a otros niveles taxonómicos. De este modo, el número de clasificaciones a analizar se resume en la Tabla 5.

Tabla 4: Número de *phyla* y organismos depositados en el NCBI

	Phylum	N.O.	Representante
1	Acidobacteria	24	*Candidatus Solibacter usitatus*
2	Actinobacteria	2675	*Nonomuraea sp.* ATCC 55076
3	Aquificae	24	*Persephonella marina*
4	Armatimonadetes	5	*Fimbriimonas ginsengisoli*
5	Bacteroidetes	1001	*Chitinophaga pinensis*
6	Caldiserica	1	*Caldisericum exile*
7	Calditrichaeota	1	*Caldithrix abyssi*
8	Cand. Cloacimonetes	1	Candidatus *Cloacimonas acidaminovorans*
9	Cand. Korarchaeota	1	Candidatus *Korarchaeum cryptofilum*
10	Cand. Saccharibacteria	1	Candidatus *Saccharibacteria oral* taxon TM7x
11	Chlamydiae	30	*Parachlamydia acanthamoebae*
12	Chlorobi	15	*Pelodictyon phaeoclathratiforme*
13	Chloroflexi	35	*Roseiflexus sp.* RS-1
14	Chrysiogenetes	2	*Desulfurispirillum indicum*
15	Crenarchaeota	41	*Sulfolobus solfataricus*
16	Cyanobacteria	242	*Acaryochloris marina*
17	Deferribacteres	6	*Denitrovibrio acetiphilus*
18	Deinococcus-Thermus	54	*Deinococcus peraridilitoris*
19	delta/epsilon sub.	373	*Archangium gephyra*
20	Dictyoglomi	2	*Dictyoglomus thermophilum*
21	Elusimicrobia	4	*Elusimicrobium minutum*
22	Euryarchaeota	309	*Haloterrigena turkmenica*
23	Fibrobacteres	20	*Fibrobacter succinogenes*
24	Firmicutes	2813	*Paenibacillus donghaensis*
25	Fusobacteria	49	*Sebaldella termitidis*
26	Gemmatimonadetes	2	*Gemmatimonas aurantiaca*
27	Ignavibacteriae	2	*Ignavibacterium album*
28	Kiritimatiellaeota	1	*Kiritimatiella glycovorans*
29	Nitrospirae	18	*Nitrospira moscoviensis*
30	Planctomycetes	32	*Singulisphaera acidiphila*
31	Proteobacteria	5407	*Paraburkholderia xenovorans*
32	Spirochaetes	97	*Sediminispirochaeta smaragdinae*
33	Synergistetes	21	*Cloacibacillus porcorum*
34	Tenericutes	142	*Spiroplasma citri*
35	Thaumarchaeota	14	*Nitrososphaera viennensis*
36	Thermodesulfobacteria	10	*Thermodesulfatator indicus*
37	Thermotogae	44	*Mesotoga prima*
38	Verrucomicrobia	36	*Opitutaceae bacterium* TAV5

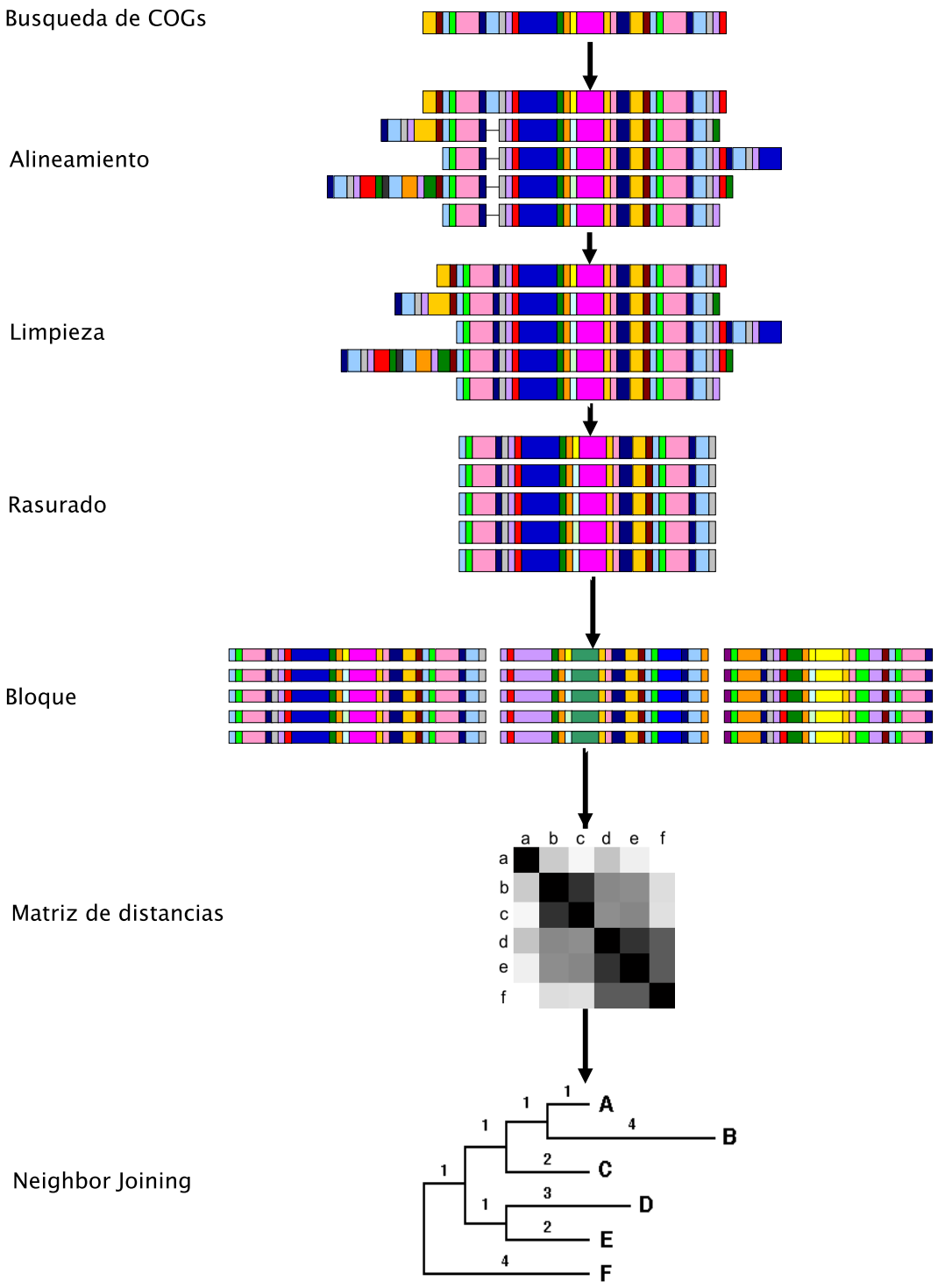


Figura 7: Estrategia de análisis filogenético utilizada para la construcción de árboles a partir de los representantes seleccionados.

Tabla 5: Número de clasificaciones de acuerdo al nivel taxonómico

Nivel	Clasificaciones
Dominio	2
Phylum	44
Clase	102
Orden	220
Familia	488
Género	2089

7.5. Ajuste de metodología

Los resultados del agrupamiento taxonómico mostrado en la Figura 6 presentaron algunas inconsistencias derivadas de la profundidad de secuenciación de los organismos con los que se construyó la base de datos. Algunos organismos se agruparon en clados incorrectos por lo que se decidió hacer un ajuste a la metodología y construir una sub base de datos que incluyera solo a los organismos de los cuales se tiene certeza que la secuenciación es completa.

Con los ajustes presentados el número de especies a analizar se redujo hasta 2685 organismos repartidos en 178 arqueas y 2507 bacterias. Con esta nueva sub base de datos se repitió la búsqueda de cadenas de markov con un e-value de $1x10^{-15}$ los resultados de este nuevo análisis se muestram en la Figura 8. En este caso la búsqueda de perfiles con modelos de Markov demostró que no existen prganismos con más de una copia de las proteínas del CLAN LSm y además muestra que la presencia las proteínas integrantes de este CLAN tienen una distribución restringida. En la sección 7.6 se procedió a buscar otras chaperonas que pudieran tener actividad sobre sRNA.

7.6. Búsqueda de chaperonas de RNA diferentes a Hfq

Se realizó una búsqueda en las base de datos de Pfam de las palabras clave *RNA chaperone* y *small RNA*, los hits de las familias de Pfam encontrados fueron posterirmente comparados con la base de datos construida obteniendo valores positivos para ProQ en 40 de los elementos. Posteriormente, se realizó una búsqueda global sobre el conjunto de la sub base de datos construida con los prganismos totalmente secuenciados. Para esto se utilizó la matriz de Pfam para esta proteína y se comparó con la base de datos mediante modelos de Markov. Los resultados de este análisis se muestran en la Tabla 6 e indican que, aunque ProQ es la segunda chaperona de RNA más abundante, la presencia de esta está limitada a una muy baja proporción del phylum de las proteobacterias. Lo que rechaza la posibilidad de que esta esté presente en los organismos donde no está Hfq o alguna de sus homólogas.

El primer paso fue entonces buscar a ProQ en todos los *phylumns* bacterianos, se realizó con el mismo nivel de astringencia que con *E. coli*, el resultado apunta a que ProQ solo está contenido en el phylum protobacteria, este resultado se muestra en la Figura 9.

.

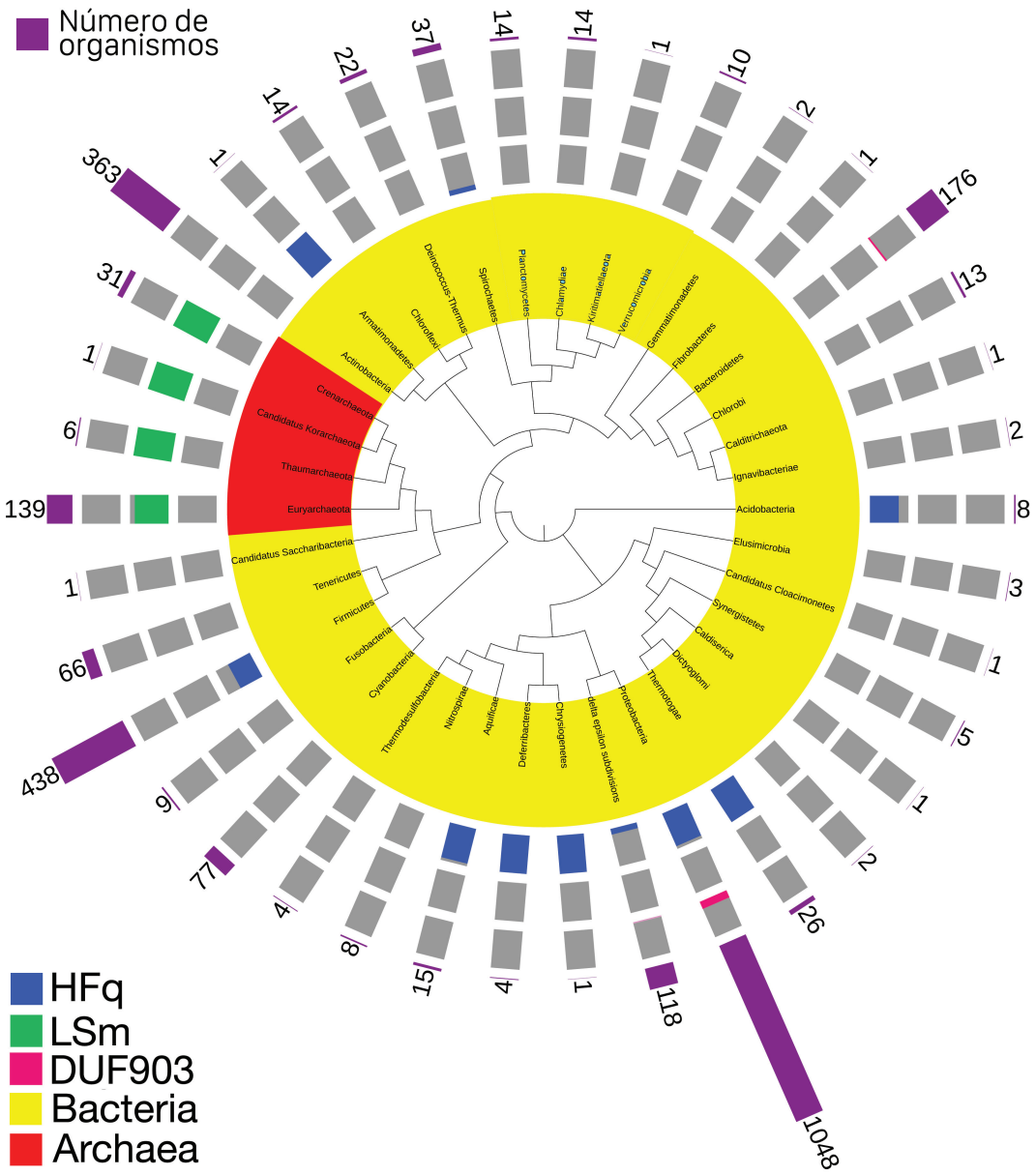


Figura 8: Distribución de Hfq, SM y DUF en la base de datos construida en este trabajo. En los anillos de barras se representa la proporción encontrada para cada uno de los perfiles de proteínas analizadas. El color gris representa ausencia de *hits*.

Tabla 6: Familias PFAM encontradas en la base de datos diferentes al CLAN LSM

Hits	PFAM ID	Family	Descripción
1	PF00011	HSP20	Hsp20/alpha crystallin family
1	PF00085	Thioredoxin	
1	PF02613	Nitrate red del	Nitrate reductase delta subunit
1	PF08264	Anticodon 1	Anticodon-binding domain of tRNA
1	PF09754	PAC2	Proteasome assembly chaperone
1	PF12895	ANAPC3	Cyclosome, subunit 3
2	PF00004	AAA	ATPase family
2	PF00118	Cpn60 TCP1	TCP-1/cpn60 chaperonin family
2	PF00166	Cpn10	Chaperonin 10 Kd subunit
2	PF01435	Peptidase M48	Peptidase family M48
2	PF02518	HATPase c	Histidine kinase- and HSP90-like ATPase
3	PF00313	CSD	Cold-shock' DNA-binding domain
3	PF00403	HMA	Heavy-metal-associated domain
3	PF01588	tRNA bind	Putative tRNA binding domain
5	PF00012	HSP70	
8	PF00226	DnaJ	
40	PF04352	ProQ	ProQ/FINO family

7.7. Predicción de sRNAS

Se desarroló un método computacional que busca predecir nuevos sRNAs en genomas procariotes, el cual se detalla en las siguientes secciones:

7.7.1. Anotación de regiones intergénicas

Como lo sugiere la literatura, y tal como lo realizan los algoritmos expuestos en la Sección 2.7.2, el primer paso fue el de suponer que los sRNAs nuevos están codificados en las regiones catalogadas como intergénicas. Dado que los sRNAs pueden estar en cualquiera de las dos cadenas de las regiones, se desarrolló un algoritmo que toma las secuencias entre dos genes y las etiqueta con los nombres de los genes adyacentes. Como indicador de la orientación, el algoritmo escribe en mayúsculas los genes que cuyo sentido de transcripción es hacia la derecha,, mientras que aquellos que lo hacen a la izquierda están anotados en minúsculas.

A modo de ejemplo, la región intergénica >eco-b3418–ECO-B3417 se encuentra entre los genes *eco-b3418* orientado hacia la izquierda y *eco-b3418* orientado a la derecha. De este modo los posibles sRNAs estarán en función de los genes aledaños y se basarán en su sentido de transcripción para brindar más información.

7.7.2. Predicción de terminadores Rho-independientes

Se buscaron referencias de la anotación y estrategia de identificación de los sRNAs de *E. coli*, con esta lista se filtraron los datos de sRNA de las tablas de anotación del genoma de este organismo de la base de datos del NCBI, para un conjunto de datos de 41 sRNAs, se revisó la literatura para identificar características similares entre ellos llegando a la conclusión de que la vasta mayoría poseen regiones terminadoras rho independientes, la lista completa se encuentra en la sección 11.3. Basados en esto, se realizó un algoritmo que parte de la búsqueda de dichas regiones terminadoras y que se detalla a continuación.

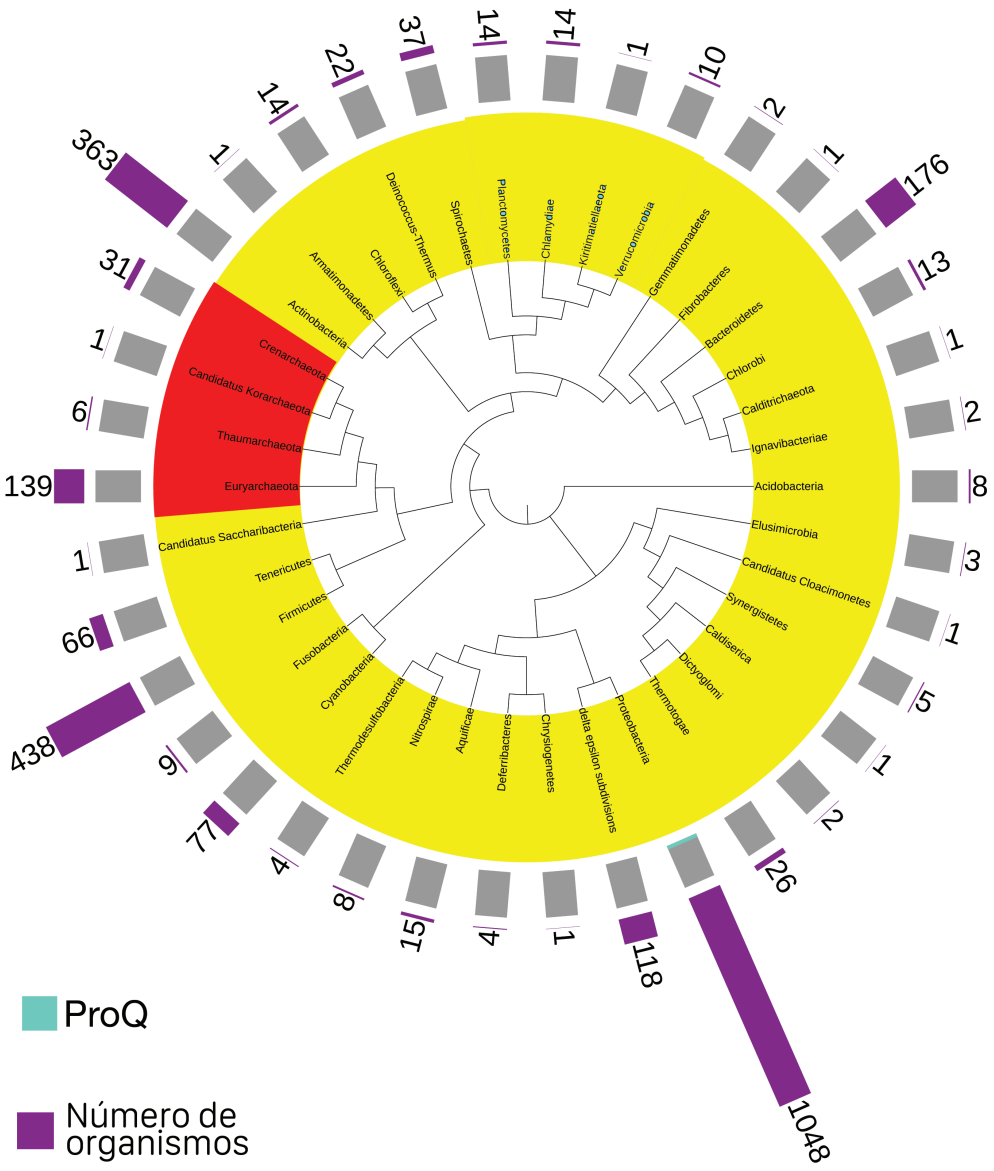


Figura 9: Distribución de ProQ en la base de datos.

7.7.3. Busqueda de regiones ricas en U

Se diseñó un algoritmo en Perl que recorre una ventana de análisis (o *window*) sobre la secuencia problema, dicha ventana de análisis se subdivide a su vez en dos regiones, la primera denominada *frame* y la segunda llamada caja de análisis o simplemente *box*. En este primer paso la caja de análisis busca encontrar una región en la que exista una de las siguientes condiciones:

- Que existan 5 uracilos consecutivos
- Que en una porción de 10 bases contiguas existan al menos 7 uracilos

Cuando encuentra alguna de estas condiciones procede al paso de la siguiente sección y desecha las secuencias que no cumplieron con ninguno de los criterios.

7.7.4. Cálculo de energía Libre y restricción de distancia

Cuando la caja de análisis encuentra una región que cumpla con los criterios establecidos, el algoritmo toma la región anterior (es decir, la secuencia *frame*) y apoyado en el programa RNAfold,[108] calcula la energía libre de esta y su posible plegamiento. Posteriormente calcula la distancia del tallo hasta el inicio de la región rica en Uracilos. El algoritmo descartará las secuencias en las que esta distancia sea mayor o igual a 2 bases.

7.7.5. Incorporación de datos de secuenciación masiva

Para verificar que las secuencias encontradas tengan identidad de terminador, se analizaron las posiciones de estas con la sumatoria de los coeficientes de *reads* de diferentes estudios de secuenciación masiva. Para ello, se siguió el siguiente proceso:

Se utilizó la suite **SRA toolkit** de NCBI para descargar diferentes estudios de secuenciación masiva para cada organismo, se procuró que cada uno de estos estudios fuera realizado bajo condiciones diferentes de cultivo. En el caso de *E. coli* se seleccionaron estudios en condiciones de medio enriquecido, medio minimo, estrés calórico y en cocultivo.

Una vez descargados los archivos fastq, cada uno de ellos se alineó al genoma de referencia con ayuda del software **bowtie2**[109] y de la suite **samtools**,[110] como paso de verificación, estos datos se visualizaron en la plataforma **tablet**.[111] Es importante señalar que este paso es el de mayor demanda de tiempo y recursos de cómputo y que los archivos fastq pueden tener un peso que supera fácilmente los 30 GB.

Se extrajeron las posiciones de alineamiento de cada archivo .sam, y considerando la longitud de los reads de cada estudio se generó una tabla de coordenadas e indice de lecturas, este archivo (al que anexamos la terminación **.ngs**) es capaz de resumir la información en un esquema mas sencillo y manejable para los propósitos de este

algoritmo que además resulta de un tamaño pequeño computacionalmente.

Los archivos .ngs de cada una de los estudios se sumaron en un único archivo combinado con la finalidad de cubrir la cantidad más alta posible de escenarios de transcripción, un esquema de este paso se muestra en la Figura 10.

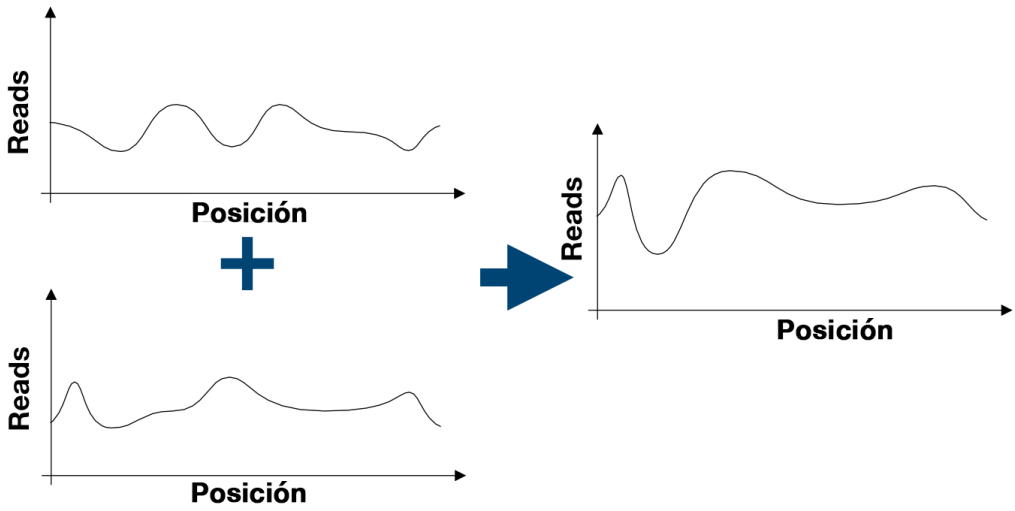


Figura 10: Esquema de suma de reads

Con los datos de la suma de *reads* se verificó que las regiones encontradas en los pasos anteriores mostrara una caída significativa en el número de *reads* alineados en esa región como un indicador del nivel de transcripción. Se obtubo un a medida indirecta de esta caida en la expresión por medio de un «coeficiente de cambio de pendiente» que toma el valor de la expresión en los extremos de la zona predicha y obtiene la relación entre ellos. De este modo, se determina matematicamente si existe dicha caida de expresión considerando el sentido del marco de lectura en cuestión. Un esquema generalizado del proceso hasta este punto se muestra en la Figura 11.

7.7.6. Validación del algoritmo de predicción de terminadores

Una de las ventajas del algoritmo expuesto hasta este punto consiste en que este puede ser modulado a partir de los valores de entrada para poder obtener un mejor resultado. En este sentido se utilizó como entrada la lista de los terminadores de los riboswitches de T-box anotados en la base de KEGG como controles positivos, mientras que una lista de 631 secuencias pertenecientes a riboswitches de gammaproteobacterias fue utilizada como control negativo. Después de un primer procesamiento, de la lista de 754 secuencias positivas el algoritmo fue capaz de identificar 601, lo que representa una precisión cercana al 80 %. Por otro lado, del set de secuencias negativas, el algoritmo prescindió de 509, lo que supone un margen de error del 19 %, número que resulta congruente con el extracto de positivos. Estos números colocan al algoritmo en indices de

precisión semejantes a los algoritmos mas complejos reportados utilizando únicamente la valoración cualitativa de las secuencias problema.

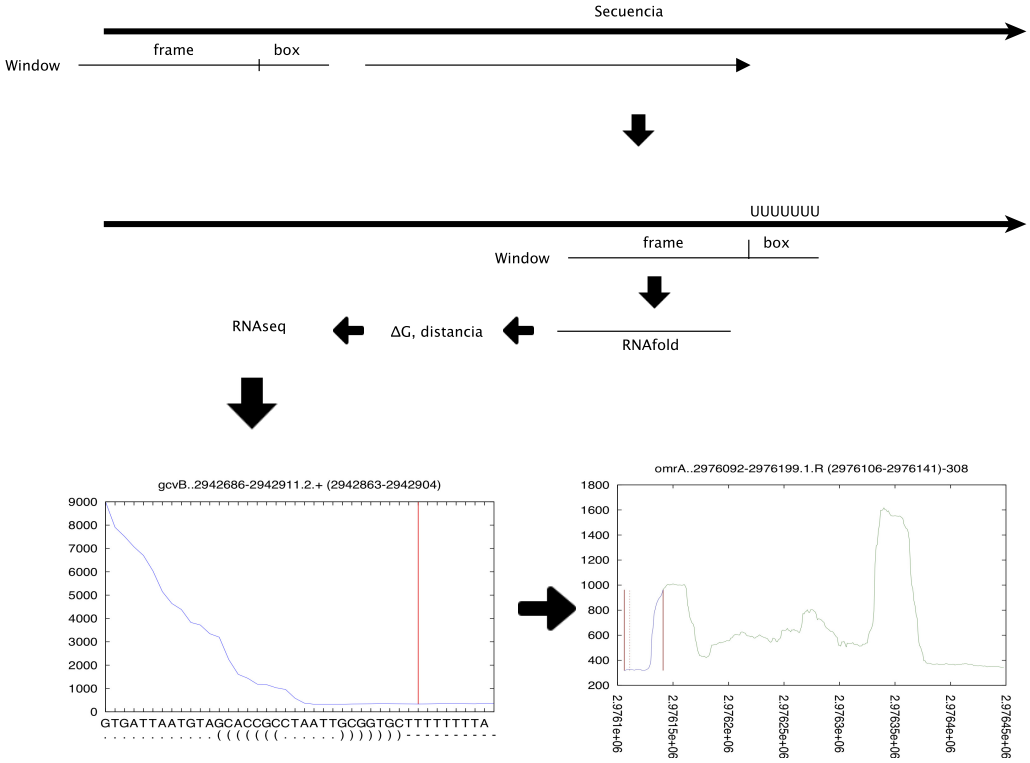


Figura 11: Proceso de búsqueda de terminadores

7.7.7. Ajuste de parámetros

Con los resultados anteriores se modificaron los parámetros iniciales del algoritmo hasta quedar como se detalla en la Tabla 7. Con estos parámetros la lista de predicciones positivas alcanzó el 88 % (665/747), mientras que la lista de negativos ignorados por el algoritmo alcanzó el 95.4 (602/631) %.

Tabla 7: Valores de los parámetros modificados

Parametro	valor	unidades
Número de U's a buscar	5	-
Energía libre máxima	11	Kcal/mol
Longitud de la región estructurada	31	bases
Distancia al terminador máxima	1	base
Región inicial (máxima)	300	bases

7.7.8. Predicción del inicio de la transcripción

Toda vez que se tiene una posición en la que existe una caída de los índices de lecturas, el paso siguiente es determinar el inicio de la transcripción de esos terminadores candidatos. Para ello, se realizó una búsqueda de mínimos de lecturas en los 300 nucleótidos rio arriba del terminador candidato, para evaluar estos puntos se buscó un cambio en la pendiente (comparando posiciones hacia atrás) y se guardaron las coordenadas del cambio máximo de esta como el inicio de la promoción de la transcripción. De este modo fue posible determinar la posición de inicio y fin de los terminadores candidatos.

7.8. sRNAsFinder

Con los scripts predictores de terminadores, de iniciadores y de análisis de datos de RNAseq se construyó un algoritmo completo en Perl que implementa la ejecución secuencial de estos bajo el nombre sRNAsFinder. Este algoritmo se empaquetó en un ejecutable que puede fácilmente ser instalado en cualquier equipo de cómputo con ambiente linux. Una vista de la pantalla de ayuda de este script se presenta en la Figura 12.

Figura 12: Pantalla principal de sRNAsFinder

sRNAsFinder usa implementaciones de gnuplot para entregar las gráficas de la expresión de los sRNAs que encuentra y presentarlos de manera visual y fácil de entender. Adicionalmente, para facilitar el análisis de los resultados sRNAsFinder ordena los sRNAs en una salida de texto plano de forma ascendente o descendente de acuerdo a cualquiera de las siguientes opciones 1) posición en el genoma, 2) energía libre del terminador o 3) longitud del sRNA.

Una de las más notorias ventajas de este algoritmo es su bajo uso de recursos de cómputo. Durante este trabajo se realizaron pruebas en un equipo portátil con procesador Intel core i5 y 8 gb de RAM. El proceso de búsqueda de sRNAs a partir de los archivos .ngs ya procesados se realizó en aproximadamente 20 minutos. Por el momento, el algoritmo se encuentra en estructura de procesamiento secuencial de un sólo hilo de procesamiento, pero implementaciones de paralelización podrían acortar considerablemente el tiempo de procesamiento en versiones futuras.

Con excepción del paquete de graficación Gnuplot, no es necesario instalar ningún paquete adicional para correr sRNAsFinder, lo que lo convierte en una herramienta fácilmente portable para ser utilizada con facilidad.

7.8.1. Indice de auto-correlación

Se añadió una rutina de validación al algoritmo mediante muestreo aleatorio. Para esto, se tomaron muestras del perfil de expresión en el archivo combinado de secuencias para conocer la posibilidad de un sesgo estadístico. Cada una de estas muestras corresponde a una ventana de 300 bases que es comparada con el perfil promedio de los sRNAs predichos por sRNAsFinder obteniendo el Indice de Auto Corelación o (AIC). Un esquema generalizado del algoritmo final se muestra en la Figura 13. El ACI es un valor porcentual que representa la probabilidad de encontrar de forma aleatoria un perfil semejante a los predichos. Un AIC bajo indica que el nivel de expresión en los perfiles de las predicciones es significativamente diferente al encontrado en el promedio aleatorio.

7.9. WebSRI

Aunque sRNAsFinder resulta en un algoritmo manejable y eficiente, su uso requiere de conocimientos intermedios de programación en linux. Para facilitar el uso del algoritmo, asi como para permitir la creación y el incremento de una base de datos con los datos de los usuarios que así lo deseen, se construyó una página web escrita en HTML5 con CSS, javascript, node y angular que aloja a este algoritmo.

WebSRI (*WEBserver for Small RNA Identification*) se aloja en un servidor de nuestro grupo y se puede acceder a su página web desde cualquier navegador conectado a internet por medio de la dirección http://biocomputo.ibt.unam.mx/websri/index.html.

En el portal, los usuarios pueden, después de registrarse, emplear el algoritmo en una interfaz sencilla y amigable que permite además que los datos de los usuarios se sumen a la base de datos que periódicamente recalcula los parámetros para ofrecer mejores predicciones.

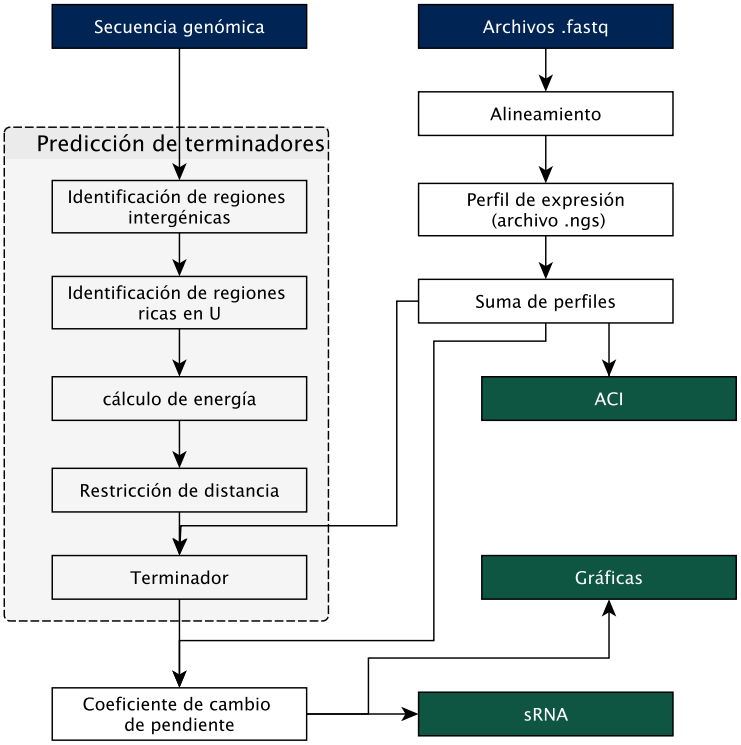


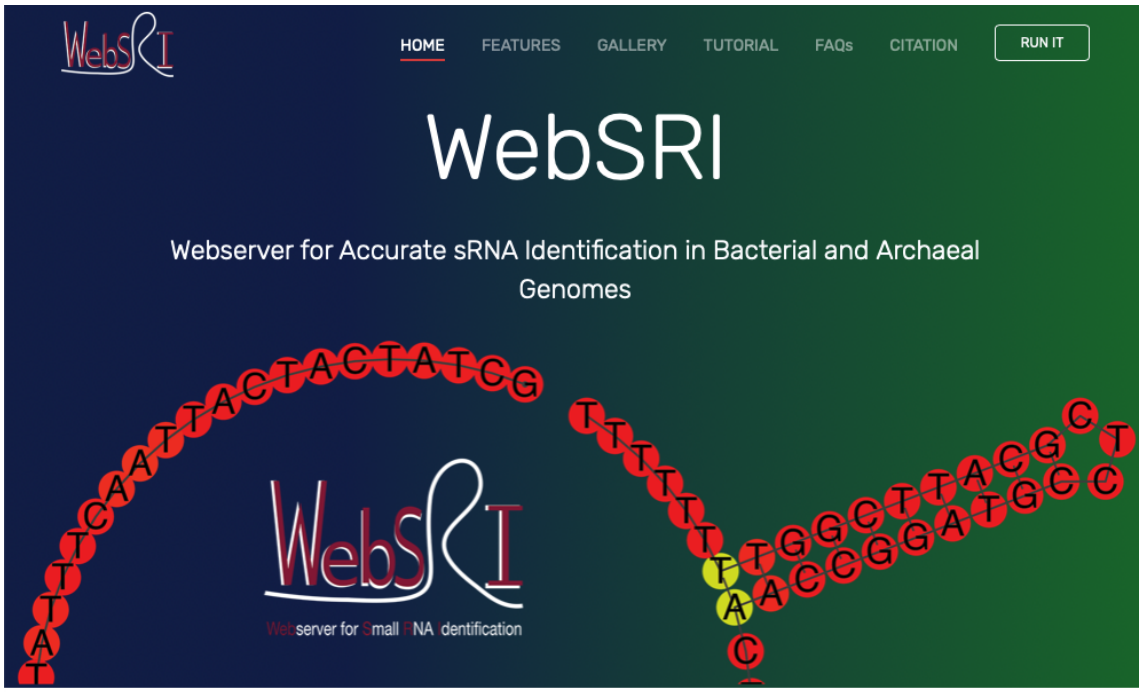
Figura 13: Esquema general de sRNAsFinder. Los cuadros coloreados en azul representan las entradas del algoritmo. Los cuadros en verde representan salidas del algoritmo. Los cuadros en blanco y gris representan rutinas de procesamiento.

Al acceder a webSRI, el usuario puede seleccionar si quiere usar sus propios datos de RNAseq, o bien usar los de la base de datos (cuando estos existen) o utilizar los del genoma mas cercano filogenéticamente. Una vista de la página principal del servidor se puede observar en la Figura 14.

7.10. Validación del método en *E. coli*

El método anterior se probó en un archivo fasta que contenía 41 secuencias correspondientes a las regiones de los sRNAs anotados en el genoma de *E. coli*, este algoritmo encontró un total de 29 de esas secuencias lo que corresponde a un 70 % de precisión.

Una de las particularidades del método aquí expuesto es que permite la corrección de los puntos de anotación de los genes de los sRNAs de *E. coli*; por ejemplo, en el caso del gen *spf*, este se encuentra anotado en las coordenadas 4 049 889 a 4 050 007, mientras que el algoritmo aquí presentado señala que su terminador termina en la posición 4 050 018.



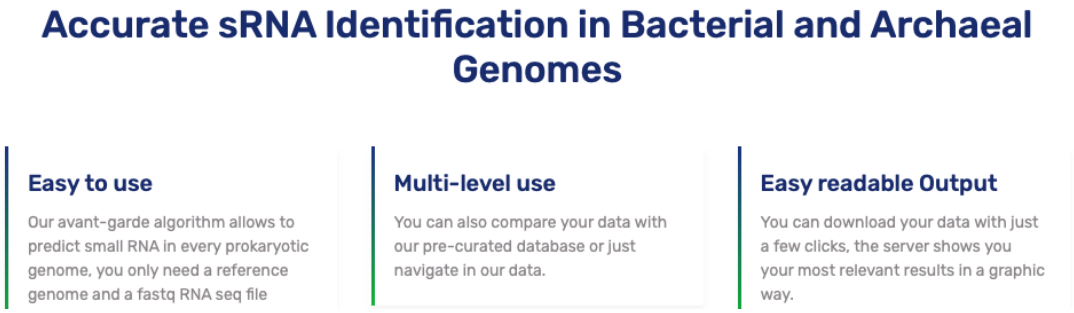


Figura 14: Pantalla principal del servidor webSRI

7.11. Búsqueda de sRNAs en *Pedobacter heparinus*

De entre todos los *phyla* que carecen de chaperonas de la familia LSm, el más representado en la base de datos es el de los bacteroidetes, de este *phylum* existen registros para 1001 organismos en la base de datos de este trabajo, lo que descarta la posibilidad de un falso negativo por falta de representantes.

Previendo la facilidad para realizar futuras pruebas *in vivo*, se seleccionó como primer candidato a *Pedobacter heparinus*, un rganismo con suficientes estudios de RNAseq publicados en el SRA del NCBI que adicionalmente se reporta como un organismo de fácil cultivo (ATCC® Medium 3) para el que existen mecanismos de transformación publicados y verificados.[112] Se utilizó la secuencia genómica de *P. heparinus* para su análisis. Se utilizaron los estudios de RNAseq SRR1796751, SRR1796752, SRR3923847

y SRR3923848 disponibles en NCBI SRA. El algoritmo encontró 327 regiones terminadoras nuevas, de las cuales 109 correspondieron a sRNAs candidatos. Las gráficas de los sRNAs nuevos PE2210 y PE 319 encontradods con sRNAsFinder se presentan en la Figura 15.

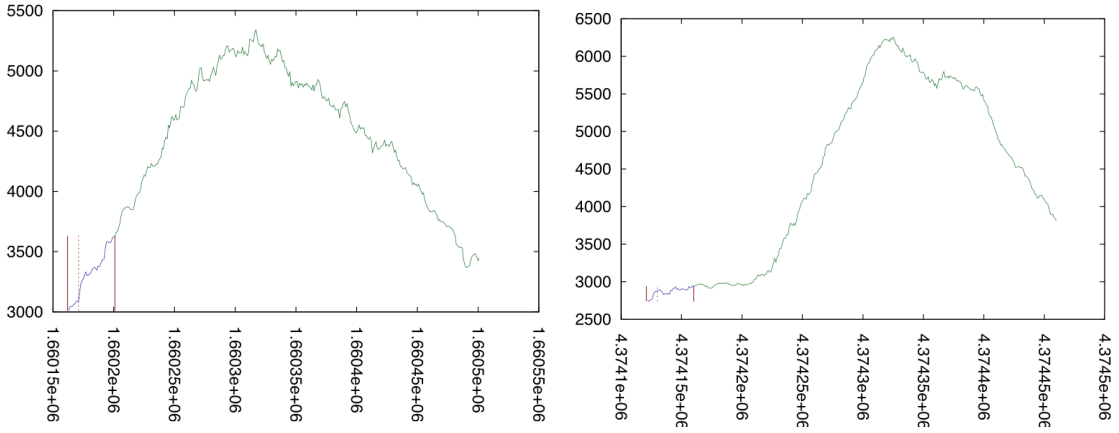


Figura 15: Salidas de sRNAsFinder en el genoma de *P. heparinus*

7.12. Fortalezas y limitaciones del método

A pesar de su capacidad de reproducir los datos de referencia con cierta fidelidad, el método aquí presentado depende directamente de la cantidad y calidad de los datos de secuenciación masiva disponibles en la bibliografía, por lo que aquellos en los que estos datos no existen no pueden ser tratados por este método. La elección de los archivos de secuenciación masiva a incorporar al análisis también resultan de gran importancia pues deben ser lo suficientemente variados para notar expresión en la mayoría de genes posibles.

A pesar de lo anterior, este método resulta en un buen acercamiento al problema de la identificación de sRNAs, pues permite flexibilizar cada uno de los parámetros involucrados en el análisis para así permitir análisis en condiciones diferentes a las de referencia. Adicionalmente, requiere muy pocos recursos de cómputo, pues consiste solo de comparaciones lógicas y cuentas de contenido de uracilos. Lo que lo hace ideal para aplicarse en casi cualquier entorno computacional.

8. Discusión

En este trabajo, una de las primeras conclusiones es que la distribución de la chaperona de RNA Hfq no es homogenea a pesar de lo que se puede inferir de la revisión bibliográfica, este trabajo sugiere que Hfq o cualquiera de sus homólogas del CLAN LSm solo están presentes en 15 *phyla* de 44 *phyla* procariotes analizados. Este resultado nos hizo cuestionarnos sobre la existencia de sRNAs en estos genomas carentes de

chaperonas de RNA.

Por otra parte, el método teórico-computacional desarrollado en esta tesis demuestra que existen organismos con presencia de sRNAs a pesar de no existir en ellos una copia detectable de Hfq. La presencia de sRNAs en estos organismos sugiere que estos poseen regulación genética mediada por sRNAs, sin embargo no ofrece una respuesta a la manera en la que esta ocurre en estos casos.

La integración de estos dos resultados permite suponer dos posibles escenarios; en el primero, los organismos carentes de Hfq poseen un análogo funcional a Hfq que realiza las labores de esta chaperona y que no pertenece al CLAN LSm, las pruebas realzadas en esta tesis demuestran que la única chaperona ajena al CLAN LSm con actividad sobre el RNA es ProQ, por lo que esta puede ser un excelente primer candidato de búsqueda.

En el segundo escenario, una chaperona no es necesaria para la correcta participación de los sRNAs en los procesos de regulación genética, de ser este el caso, el complejo sRNA-mRNA debería exhibir una energía de estabilización mayor a la que se encuentra en los pares sRNA-mRNA de los organismos que si tiene Hfq. En cualquiera de los dos escenarios, la búsqueda de alternativas resulta imperativa para determinar los mecanismos que ocurren en los organismos que no poseen chaperonas con actividad sobre sRNA.

Fue posible crear un método que permite la predicción de nuevos sRNAs en genomas alejados de los organismos modelo canónicos, sin embargo, en todos los casos las predicciones se mejoran cuando la cantidad de experimentos aumenta para cada organismo; sRNAsFinder y su servidor web WebSRI se ven afectados directamente por la cantidad de experimentos de secuenciación masiva reportados de forma pública en las bases de datos.

Por otra parte, a pesar de la dramática reducción en el tiempo de predicción en contraste con otros métodos, WebSRI requiere una plataforma de cómputo de alto rendimiento para poder hacer crecer su base de datos y requiere también de amplios conocimientos en tecnologías de programación muy dinámicas que hacen necesario un mantenimiento constante. La base de datos de WebSRI se nutre de datos tomados directamente de servidores computacionales con alta tasa de recambio, por lo que la posibilidad de verse afectado por cambios en las estructuras de archivos de estos permanece elevada.

La calibración y corroboración de los datos predichos por algoritmos como sRNAsFinder se ve beneficiada por la cantidad de sRNAs correctamente anotados en los genomas modelo, sin embargo este número permanece casi sin cambios debido a la alta dificultad que existe para aislar y caracterizar los nuevos sRNAs de forma experimental. Por otra parte, tecnologías como la Inteligencia Artificial podrían mejorar la manera en que este servidor ajuste los parámetros de calibración para ofrecer mejores predicciones.

Finalmente, resulta imperativo desarrollar métodos o estrategias que permitan determinar la veracidad de predicciones como las obtenidas por sRNAsFinder y WebSRI. Estas estrategias tienen el reto de ser aplicadas a un elevado número de organismos de diferentes necesidades metabólicas y ambientales, por lo que, de encontrarse esta estrategia resultaría en un hito en la identificación de sRNAs.

9. Conclusiones

Los resultados de este trabajo permiten emitir las siguientes conclusiones:

1. Este trabajo sugiere que la distribución de la chaperona de RNA Hfq no está presente en todos los *phyla* de organismos procariontes como se pensaba. La presencia de Hfq o cualquiera de sus homólogos está restringida a una minoría de estos *phyla*

2. Es posible construir un método teórico computacional capaz de predecir nuevos sRNAs en genomas procariontes que tome como entrada la secuencia genómica de un organismo. La inclusión de datos provenientes de estudios de secuenciación masiva aumenta la precisión de predicción y reduce la cantidad de tiempo requerido en contraste con otros métodos.

3. Existen organismos con sRNAs detectables a pesar de no poseer una proteína del CLAN LSm. Lo que sugiere que en estos organismos, la regulación por sRNAs puede depender de otra chaperona análoga en función a HFq o bien prescindir de proteínas con actividad de chaperona de RNA. Los sRNAs reguladores se encuentran distribuidos en todas (o la mayoría) de las especies bacterianas idenpendientemente de la presencia o ausencia de la chaperona Hfq.

4. Es necesario encontrar un método experimental que permita la identificación de la proteína análoga en función a Hfq, o bien permita determinar si esta es prescindible en los genomas carentes de Hfq.

10. Perspectivas

El desarrollo de este trabajo plantea dos nuevas interrogantes en referencia a los fenómenos de regulación genética mediada por sRNAs en organismos procariontes. Los organismos carentes de Hfq podrían poseer una proteína análoga en función a esta chaperona, o bien no requerir de ella. Basados en estas dos interrogantes como premisas, la perspectiva más cercana a llevarse a cabo es la de determinar de forma experimental cuál de los dos escenarios es el que existe en estos organismos. Esta estrategia deberá seguir los siguientes objetivos:

- Identificar computacionalmente nuevos sRNAs en todas las bacterias carentes de Hfq usando el servidor web webSRI y el algoritmo sRNAs-Finder.

- Identificar computacionalmente los blancos de regulación de los sRNAs encontrados en al menos un organismo, de este modo se podría determinar si las fuerzas de interacción de los pares sRNA-mRNA tienen diferentes valores en contraste con aquellos que son beneficiados por la presencia de la chaperona Hfq.
- Verificar las predicciones computacionales de sRNAs y sus correspondientes blancos usando metodologías como sistemas de expresión heterólogos o genes reporteros.
- Identificar experimentalmente las potenciales chaperonas análogas funcionales de función de Hfq en organismos os carentes de Hfq como *P. heparinus* mediante procedimientos como las bibliotecas de DNA o los tansposones.

11. anexos

11.1. Secuencias utilizadas en estre trabajo

11.1.1. Hfq de *Escherichia coli*

```
MAKGQSLQDP FLNALRRERV PVSIYLVNGI KLQGQIESFD QFVILLKNTV SQMVYKHAIS
TVVPSRPVSH HSNNAGGGTS SNYHHGSSAQ NTSAQQDSEE TE
```

11.1.2. Sm de *Thermovibrio ammonificans*

```
MKKFKTLEEA QIELIELLEQ EGEFRGTLNE LADRLNVKPE NIRPLLQLLK SSGDVLYEES
EEGLIVRPAM MVPVVPPTLT PEQEQEVQQK LKEGYKVIAC STMGGVQSRE LRSALGKRVI
VYFRNGSKVE AKLKGFDRFC LKLRNYMGNM LAYKHAISTI VYKP
```

11.2. Resumen de los Modelos de HMMR utilizados en este trabajo

Se tomaron los modelos crudos alojados en la base de datos de PFam (`pfam.xfam.org`) como e detalla a continuación.

11.2.1. Hfq (PF17209)

Seed source:	Bateman A	
Previous IDs:	none	
Type:	Domain	
Author:	Bateman A	
Number in seed:	148	
Number in full:	1994	
Average length of the domain:	63.50 aa	
Average identity of full alignment:	55 %	
Average coverage of the sequence by the domain:	73.55 %	
Model details:		
Parameter	Sequence	Domain
Gathering cut-off	21.3	21.3
Trusted cut-off	21.3	21.3
Noise cut-off	21.2	21.2
Model length:	64	
Family (HMM) version:	2	

11.2.2. LSM (PF01423)

Seed source:	Psiblast SMD1_HUMAN	
Previous IDs:	Sm;	
Type:	Domain	
Author:	Bateman A	
Number in seed:	112	
Number in full:	13259	
Average length of the domain:	69.40 aa	
Average identity of full alignment:	25 %	
Average coverage of the sequence by the domain:	50.36 %	
Model details:		
Parameter	Sequence	Domain
Gathering cut-off	23.1	23.1
Trusted cut-off	23.1	23.1
Noise cut-off	23.0	23.0
Model length:	67	
Family (HMM) version:	21	

11.3. sRNAs anotados en el genoma de *E. coli*

name	sense	left	right	name	sense	left	right
agrA	+	3648063	3648146	*rdlB*	+	1269858	1269923
agrB	+	3648294	3648377	*rdlC*	+	1270393	1270460
arcZ	+	3350577	3350697	*rdlD*	+	3700136	3700201
arrS	-	3657986	3658054	*rprA*	+	1770372	1770477
chiX	+	507204	507287	*rseX*	+	2033649	2033739
csrB	-	2924156	2924524	*rttR*	-	1287066	1287236
csrC	+	4051036	4051280	*rybB*	-	887976	888054
cyaR	+	2167114	2167200	*rydB*	-	1764713	1764780
dicF	+	1649382	1649434	*rydC*	-	1491443	1491506
dsrA	-	2025227	2025313	*ryeA*	+	1923066	1923314
esrE	+	4019978	4020229	*ryfA*	+	2653855	2654158
eyeA	+	272580	272654	*ryfD*	-	2734153	2734295
ffs	+	476448	476561	*ryhB*	-	3580927	3581016
fnrS	+	1409129	1409250	*ryjA*	-	4277927	4278066
gadY	+	3664864	3664968	*ryjB*	+	4527977	4528066
gcvB	+	2942696	2942901	*sdsR*	-	1923164	1923284
glmY	-	2691157	2691340	*sgrS*	+	77367	77593
glmZ	+	3986432	3986638	*sibA*	+	2153309	2153451
isrC	+	2071317	2071511	*sibB*	+	2153644	2153779
istR	-	3853118	3853257	*sibC*	+	3056849	3056988
mcaS	-	1405656	1405751	*sibD*	-	3194723	3194865
mgrR	-	1622817	1622914	*sibE*	-	3195099	3195240
micA	+	2814802	2814879	*sokB*	+	1492119	1492174
micC	+	1437121	1437229	*sokC*	+	16952	17006
micF	+	2313084	2313176	*sokE*	+	607734	607792
ohsC	+	2700520	2700596	*sokX*	+	2887354	2887409
omrA	-	2976102	2976189	*spf*	+	4049899	4050007
omrB	-	2976304	2976385	*sroH*	-	4190327	4190487
psrD	+	1146589	1146757	*symR*	+	4579835	4579911
psrO	+	3311225	3311398	*tff*	+	189712	189847
rdlA	+	1269323	1269389				

Referencias

[1] Zhang, A. *et al.* Global analysis of small RNA and mRNA targets of Hfq. *Molecular Microbiology* **50**, 1111–1124 (2003).

[2] Hoe, C.-H., Raabe, C. A., Rozhdestvensky, T. S. & Tang, T.-H. Bacterial sRNAs: regulation in stress. *International Journal of Medical Microbiology* **303**, 217–229 (2013).

[3] Rodriguez-Escamilla, Z., Martínez-Núñez, M. A. & Merino, E. Epigenetics knocks on synthetic biology's door. *FEMS Microbiology Letters* **363**, fnw191 (2016).

[4] Hernández-Santos, W. J., Rodríguez-Escamilla, Z. & Merino-Pérez, E. Analysis of bacterial stressomes using synthetic biology. In *International Symposium on Fuctional Genomics and Systems Biology* (2017).

[5] Boor, K. J. Bacterial stress responses: what doesn't kill them can make them stronger. *PLoS Biologyogy* **4**, e23 (2006).

[6] Caron, M.-P., Lafontaine, D. A. & Massé, E. Small RNA-mediated regulation at the level of transcript stability. *RNA Biology* **7**, 140–144 (2010).

[7] Rodrigues, J. L. & Rodrigues, L. R. Potential Applications of the Escherichia coli Heat Shock Response in Synthetic Biology. *Trends in Biotechnology* (2017).

[8] Hengge-Aronis, R. Signal transduction and regulatory mechanisms involved in control of the σs (RpoS) subunit of RNA polymerase. *Microbiology and Molecular Biology Reviews* **66**, 373–395 (2002).

[9] Csonka, L. N. Physiological and genetic responses of bacteria to osmotic stress. *Microbiological Reviews* **53**, 121–147 (1989).

[10] Laermann, V., Ćudić, E., Kipschull, K., Zimmann, P. & Altendorf, K. The sensor kinase KdpD of E scherichia coli senses external K+. *Molecular Microbiology* **88**, 1194–1204 (2013).

[11] Borukhov, S. & Nudler, E. RNA polymerase holoenzyme: structure, function and biological implications. *Current Opinion in Microbiology* **6**, 93–100 (2003).

[12] Lange, R. & Hengge-Aronis, R. The cellular concentration of the sigma S subunit of RNA polymerase in Escherichia coli is controlled at the levels of transcription, translation, and protein stability. *Genes & Development* **8**, 1600–1612 (1994).

[13] Merrick, M. & Gibbins, J. The nucleotide sequence of the nitrogen-regulation gene ntrA of Klebsiella pneumoniae and comparison with conserved features in bacterial RNA polymerase sigma factors. *Nucleic Acids Research* **13**, 7607–7620 (1985).

[14] Zhou, Y. N., Kusukawa, N., Erickson, J., Gross, C. & Yura, T. Isolation and characterization of Escherichia coli mutants that lack the heat shock sigma factor sigma 32. *Journal of Bacteriology* **170**, 3640–3649 (1988).

[15] Raina, S., Missiakas, D. & Georgopoulos, C. The rpoE gene encoding the sigma E (sigma 24) heat shock sigma factor of Escherichia coli. *The EMBO Journal* **14**, 1043–1055 (1995).

[16] Jenal, U. & Hengge-Aronis, R. Regulation by proteolysis in bacterial cells. *Current Opinion in Microbiology* **6**, 163–172 (2003).

[17] Gottesman, S. & Storz, G. Bacterial small RNA regulators: versatile roles and rapidly evolving variations. *Cold Spring Harbor Perspectives in Biology* **3**, a003798 (2011).

[18] Gottesman, S. *et al.* Small RNA regulators and the bacterial response to stress. In *Cold Spring Harbor Symposia on Quantitative Biology*, vol. 71, 1–11 (Cold Spring Harbor Laboratory Press, 2006).

[19] Fozo, E. M., Hemm, M. R. & Storz, G. Small toxic proteins and the antisense RNAs that repress them. *Microbiology and Molecular Biology Reviews* **72**, 579–589 (2008).

[20] Gerdes, K. & Wagner, E. G. H. RNA antitoxins. *Current Opinion in Microbiology* **10**, 117–124 (2007).

[21] Tramonti, A., De Canio, M. & De Biase, D. GadX/GadW-dependent regulation of the Escherichia coli acid fitness island: transcriptional control at the gadY–gadW divergent promoters and identification of four novel 42 bp GadX/GadW-specific binding sites. *Molecular Microbiology* **70**, 965–982 (2008).

[22] Waters, L. S. & Storz, G. Regulatory RNAs in bacteria. *Cell* **136**, 615–628 (2009).

[23] Ketting, R. F. *et al.* Dicer functions in RNA interference and in synthesis of small RNA involved in developmental timing in C. elegans. *Genes & Development* **15**, 2654–2659 (2001).

[24] Storz, G., Vogel, J. & Wassarman, K. M. Regulation by small RNAs in bacteria: expanding frontiers. *Molecular Cell* **43**, 880–891 (2011).

[25] Opdyke, J. A., Kang, J.-G. & Storz, G. GadY, a small-RNA regulator of acid response genes in Escherichia coli. *Journal of Bacteriology* **186**, 6698–6705 (2004).

[26] Zheng, M., Doan, B., Schneider, T. D. & Storz, G. OxyR and SoxRS regulation of *fur*. *Journal of Bacteriology* **181**, 4639–4643 (1999).

[27] Wilson, R. L., Stauffer, L. T. & Stauffer, G. V. Roles of the GcvA and PurR proteins in negative regulation of the Escherichia coli glycine cleavage enzyme system. *Journal of Bacteriology* **175**, 5129–5134 (1993).

[28] Chung, H., Bang, W. & Drake, M. Stress response of escherichia coli. *Comprehensive Reviews in Food Science and Food Safety* **5**, 52–64 (2006).

[29] Majdalani, N., Hernandez, D. & Gottesman, S. Regulation and mode of action of the second small RNA activator of RpoS translation, RprA. *Molecular Microbiology* **46**, 813–826 (2002).

[30] Freeman, J. A. & Bassler, B. L. A genetic analysis of the function of LuxO, a two-component response regulator involved in quorum sensing in Vibrio harveyi. *Molecular Microbiology* **31**, 665–677 (1999).

[31] Balbontín, R., Fiorini, F., Figueroa-Bossi, N., Casadesús, J. & Bossi, L. Recognition of heptameric seed sequence underlies multi-target regulation by RybB small RNA in Salmonella enterica. *Molecular Microbiology* **78**, 380–394 (2010).

[32] Tardat, B. & Touati, D. Two global regulators repress the anaerobic expression of MnSOD in Escherichia coli: Fur (ferric uptake regulation) and Arc (aerobic respiration control). *Molecular Microbiology* **5**, 455–465 (1991).

[33] Niederhoffer, E. C., Naranjo, C. M., Bradley, K. L. & Fee, J. A. Control of Escherichia coli superoxide dismutase (sodA and sodB) genes by the ferric uptake regulation (fur) locus. *Journal of Bacteriology* **172**, 1930–1938 (1990).

[34] Chen, S., Zhang, A., Blyn, L. B. & Storz, G. MicC, a second small-RNA regulator of Omp protein expression in Escherichia coli. *Journal of Bacteriology* **186**, 6689–6697 (2004).

[35] Majdalani, N., Cunning, C., Sledjeski, D., Elliott, T. & Gottesman, S. DsrA RNA regulates translation of RpoS message by an anti-antisense mechanism, independent of its action as an antisilencer of transcription. *Proceedings of the National Academy of Sciences* **95**, 12462–12467 (1998).

[36] Wassarman, K. M., Zhang, A. & Storz, G. Small RNAs in Escherichia coli. *Trends in Microbiology* **7**, 37–45 (1999).

[37] Altuvia, S., Weinstein-Fischer, D., Zhang, A., Postow, L. & Storz, G. A small, stable RNA induced by oxidative stress: role as a pleiotropic regulator and antimutator. *Cell* **90**, 43–53 (1997).

[38] Romeo, T. Global regulation by the small RNA-binding protein CsrA and the non-coding RNA molecule CsrB. *Molecular Microbiology* **29**, 1321–1330 (1998).

[39] Bouché, F. & Bouché, J.-P. Genetic evidence that DicF, a second division inhibitor encoded by the Escherichia coli dicB operon, is probably RNA. *Molecular Microbiology* **3**, 991–994 (1989).

[40] Sledjeski, D. & Gottesman, S. A small RNA acts as an antisilencer of the H-NS-silenced rcsA gene of Escherichia coli. *Proceedings of the National Academy of Sciences* **92**, 2003–2007 (1995).

[41] Aiba, H., Matsuyama, S.-l., Mizuno, T. & Mizushima, S. Function of micF as an antisense RNA in osmoregulatory expression of the ompF gene in Escherichia coli. *Journal of Bacteriology* **169**, 3007–3012 (1987).

[42] Argaman, L. *et al.* Novel small RNA-encoding genes in the intergenic regions of Escherichia coli. *Current Biology* **11**, 941–950 (2001).

[43] Hüttenhofer, A. & Vogel, J. Experimental approaches to identify non-coding RNAs. *Nucleic Acids Research* **34**, 635–646 (2006).

[44] Wassarman, K. M., Repoila, F., Rosenow, C., Storz, G. & Gottesman, S. Identification of novel small RNAs using comparative genomics and microarrays. *Genes & Development* **15**, 1637–1651 (2001).

[45] Ando, Y., Asari, S., Suzuma, S., Yamane, K. & Nakamura, K. Expression of a small RNA, BS203 RNA, from the yocI–yocJ intergenic region of Bacillus subtilis genome. *FEMS Microbiology letters* **207**, 29–33 (2002).

[46] Suzuma, S. *et al.* Identification and characterization of novel small RNAs in the aspS–yrvM intergenic region of the Bacillus subtilis genome. *Microbiology* **148**, 2591–2598 (2002).

[47] Djordjevic, M. SELEX experiments: new prospects, applications and data analysis in inferring regulatory pathways. *Biomolecular Engineering* **24**, 179–189 (2007).

[48] Lorenz, C., Von Pelchrzim, F. & Schroeder, R. Genomic systematic evolution of ligands by exponential enrichment (Genomic SELEX) for the identification of protein-binding RNAs independent of their expression levels. *Nature Protocols* **1**, 2204 (2006).

[49] Altuvia, S. Identification of bacterial small non-coding RNAs: experimental approaches. *Current Opinion in Microbiology* **10**, 257–261 (2007).

[50] Jedamzik, B. & Eckmann, C. R. Analysis of RNA-protein complexes by RNA coimmunoprecipitation and RT-PCR analysis from Caenorhabditis elegans. *Cold Spring Harbor Protocols* **2009**, pdb–prot5300 (2009).

[51] Vogel, J. *et al.* RNomics in Escherichia coli detects new sRNA species and indicates parallel transcriptional output in bacteria. *Nucleic Acids Research* **31**, 6435–6443 (2003).

[52] Rivas, E., Klein, R. J., Jones, T. A. & Eddy, S. R. Computational identification of noncoding RNAs in E. coli by comparative genomics. *Current Biology* **11**, 1369–1373 (2001).

[53] Rivas, E. & Eddy, S. R. Noncoding RNA gene detection using comparative sequence analysis. *BMC Bioinformatics* **2**, 8 (2001).

[54] Lambert, A. *et al.* The erpin server: an interface to profile-based rna motif identification. *Nucleic Acids Research* **32**, W160–W165 (2004).

[55] Pichon, C. & Felden, B. Intergenic sequence inspector: searching and identifying bacterial RNAs. *Bioinformatics* **19**, 1707–1709 (2003).

[56] Coventry, A., Kleitman, D. J. & Berger, B. MSARI: multiple sequence alignments for statistical detection of RNA secondary structure. *Proceedings of the National Academy of Sciences* **101**, 12102–12107 (2004).

[57] Gruber, A. R., Findeiß, S., Washietl, S., Hofacker, I. L. & Stadler, P. F. RNAz 2.0: improved noncoding RNA detection. In *Biocomputing 2010*, 69–79 (World Scientific, 2010).

[58] Sridhar, J. *et al.* sRNAscanner: a computational tool for intergenic small RNA detection in bacterial genomes. *PLoS One* **5**, e11970 (2010).

[59] Livny, J., Fogel, M. A., Davis, B. M. & Waldor, M. K. sRNAPredict: an integrative computational approach to identify sRNAs in bacterial genomes. *Nucleic Acids Research* **33**, 4096–4105 (2005).

[60] Sridhar, J., Sowmiya, G., Sekar, K. & Rafi, Z. A. PsRNA: A Computing Engine for the comparative identification of putative small RNA locations within intergenic regions. *Genomics, Proteomics & Bioinformatics* **8**, 127–134 (2010).

[61] Ott, A., Idali, A., Marchais, A. & Gautheret, D. NAPP: the nucleic acid phylogenetic profile database. *Nucleic Acids Research* **40**, D205–D209 (2011).

[62] Sharma, C. M. & Vogel, J. Experimental approaches for the discovery and characterization of regulatory small RNA. *Current Opinion in Microbiology* **12**, 536–546 (2009).

[63] Livny, J. & Waldor, M. K. Identification of small RNAs in diverse bacterial species. *Current Opinion in Microbiology* **10**, 96–101 (2007).

[64] Do, C. B., Woods, D. A. & Batzoglou, S. CONTRAfold: RNA secondary structure prediction without physics-based models. *Bioinformatics* **22**, e90–e98 (2006).

[65] Houseley, J. & Tollervey, D. The many pathways of RNA degradation. *Cell* **136**, 763–776 (2009).

[66] Treiber, D. K. & Williamson, J. R. Exposing the kinetic traps in RNA folding. *Current Opinion in Structural Biology* **9**, 339–345 (1999).

[67] Russell, R. & Herschlag, D. Probing the folding landscape of the Tetrahymena ribozyme: commitment to form the native conformation is late in the folding pathway1. *Journal of Molecular Biology* **308**, 839–851 (2001).

[68] Schroeder, R., Barta, A. & Semrad, K. Strategies for RNA folding and assembly. *Nature reviews Molecular Cell biology* **5**, 908 (2004).

[69] Rajkowitsch, L. *et al.* RNA chaperones, RNA annealers and RNA helicases. *RNA Biology* **4**, 118–130 (2007).

[70] Zhang, A., Derbyshire, V., Salvo, J. & Belfort, M. Escherichia coli protein StpA stimulates self-splicing by promoting RNA assembly in vitro. *RNA* **1**, 783–793 (1995).

[71] Russell, R. RNA misfolding and the action of chaperones. *Frontiers in Bioscience: a journal and virtual library* **13**, 1 (2008).

[72] Kashyap, A. K., Schieltz, D., Yates Iii, J. & Kellogg, D. R. Biochemical and genetic characterization of Yra1p in budding yeast. *Yeast* **22**, 43–56 (2005).

[73] Jiang, W., Hou, Y. & Inouye, M. CspA, the major cold-shock protein of Escherichia coli, is an RNA chaperone. *Journal of Biological Chemistry* **272**, 196–202 (1997).

[74] Graumann, P. & Marahiel, M. A. The major cold shock protein of Bacillus subtilis CspB binds with high affinity to the ATTGG-and CCAAT sequences in single stranded oligonucleotides. *FEBS letters* **338**, 157–160 (1994).

[75] Duval, M. *et al.* Escherichia coli ribosomal protein S1 unfolds structured mRNAs onto the ribosome for active translation initiation. *PLoS Biologyogy* **11**, e1001731 (2013).

[76] Mayer, O., Rajkowitsch, L., Lorenz, C., Konrat, R. & Schroeder, R. RNA chaperone activity and RNA-binding properties of the E. coli protein StpA. *Nucleic Acids Research* **35**, 1257–1269 (2007).

[77] Khachatoorian, R. & French, S. W. Chaperones in hepatitis C virus infection. *World Journal of Hepatology* **8**, 9 (2016).

[78] Glover, J. M. *et al.* The FinO family of bacterial RNA chaperones. *Plasmid* **78**, 79–87 (2015).

[79] Khusial, P., Plaag, R. & Zieve, G. W. LSm proteins form heptameric rings that bind to RNA via repeating motifs. *Trends in Biochemical Sciences* **30**, 522–528 (2005).

[80] Achsel, T., Stark, H. & Lührmann, R. The Sm domain is an ancient RNA-binding motif with oligo (U) specificity. *Proceedings of the National Academy of Sciences* **98**, 3685–3689 (2001).

[81] Muffler, A., Fischer, D. & Hengge-Aronis, R. The RNA-binding protein HF-I, known as a host factor for phage Qbeta RNA replication, is essential for rpoS translation in Escherichia coli. *Genes & Development* **10**, 1143–1151 (1996).

[82] Vogel, J. & Luisi, B. F. Hfq and its constellation of RNA. *Nature Reviews Microbiology* **9**, 578–589 (2011).

[83] Updegrove, T. B., Zhang, A. & Storz, G. Hfq: the flexible RNA matchmaker. *Current Opinion in Microbiology* **30**, 133–138 (2016).

[84] Brescia, C. C., Mikulecky, P. J., Feig, A. L. & Sledjeski, D. D. Identification of the Hfq-binding site on DsrA RNA: Hfq binds without altering DsrA secondary structure. *RNA* **9**, 33–43 (2003).

[85] Hussein, R. & Lim, H. N. Disruption of small RNA signaling caused by competition for Hfq. *Proceedings of the National Academy of Sciences* **108**, 1110–1115 (2011).

[86] Soper, T., Mandin, P., Majdalani, N., Gottesman, S. & Woodson, S. A. Positive regulation by small RNAs and the role of Hfq. *Proceedings of the National Academy of Sciences* **107**, 9602–9607 (2010).

[87] Mackie, G. A. RNase E: at the interface of bacterial RNA processing and decay. *Nature Reviews Microbiology* **11**, 45–57 (2013).

[88] Prévost, K. *et al.* The small RNA RyhB activates the translation of shiA mRNA encoding a permease of shikimate, a compound involved in siderophore synthesis. *Molecular Microbiology* **64**, 1260–1273 (2007).

[89] Sittka, A., Sharma, C. M., Rolle, K. & Vogel, J. Deep sequencing of salmonella rna associated with heterologous hfq proteins in vivo reveals small rnas as a major target class and identifies rna processing phenotypes. *RNA Biology* **6**, 266–275 (2009).

[90] Kavita, K., de Mets, F. & Gottesman, S. New aspects of RNA-based regulation by Hfq and its partner sRNAs. *Current Opinion in Microbiology* **42**, 53–61 (2018).

[91] Figueroa-Bossi, N., Valentini, M., Malleret, L. & Bossi, L. Caught at its own game: regulatory small RNA inactivated by an inducible transcript mimicking its target. *Genes & Development* **23**, 2004–2015 (2009).

[92] Weichenrieder, O. RNA binding by Hfq and ring-forming (L) Sm proteins: a trade-off between optimal sequence readout and RNA backbone conformation. *RNA Biology* **11**, 537–549 (2014).

[93] Kambach, C. *et al.* Crystal structures of two Sm protein complexes and their implications for the assembly of the spliceosomal snRNPs. *Cell* **96**, 375–387 (1999).

[94] Link, T. M., Valentin-Hansen, P. & Brennan, R. G. Structure of Escherichia coli Hfq bound to polyriboadenylate RNA. *Proceedings of the National Academy of Sciences* **106**, 19292–19297 (2009).

[95] Attia, A. S. *et al.* Moraxella catarrhalis expresses an unusual Hfq protein. *Infection and Immunity* **76**, 2520–2530 (2008).

[96] Schilling, D. & Gerischer, U. The Acinetobacter baylyi Hfq gene encodes a large protein with an unusual C terminus. *Journal of Bacteriology* **191**, 5553–5562 (2009).

[97] Vincent, H. A. *et al.* Characterization of Vibrio cholerae Hfq provides novel insights into the role of the Hfq C-terminal region. *Journal of Molecular Biology* **420**, 56–69 (2012).

[98] Beich-Frandsen, M., Večerek, B., Sjöblom, B., Bläsi, U. & Djinović-Carugo, K. Structural analysis of full-length Hfq from Escherichia coli. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications* **67**, 536–540 (2011).

[99] Merino, E. & Yanofsky, C. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends in Genetics* **21**, 260–264 (2005).

[100] Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research* **44**, D279–D285 (2016).

[101] Letunic, I. & Bork, P. Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research* **44**, W242–W245 (2016).

[102] Ruano-Rubio, V., Poch, O. & Thompson, J. D. Comparison of eukaryotic phylogenetic profiling approaches using species tree aware methods. *BMC Bioinformatics* **10**, 383 (2009).

[103] Tatusov, R. L., Galperin, M. Y., Natale, D. A. & Koonin, E. V. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Research* **28**, 33–36 (2000).

[104] Felsenstein, J. {PHYLIP}(Phylogeny Inference Package) version 3.6 a3. *Wiley Online Library* (2002).

[105] Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. *Bioinformatics* **25**, 1335–1337 (2009).

[106] Posada, D. jModelTest: phylogenetic model averaging. *Molecular Biology and Evolution* **25**, 1253–1256 (2008).

[107] Guindon, S. *et al.* New algorithms and Methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. *Systematic Biology* **59**, 307–321 (2010).

[108] Gruber, A. R., Lorenz, R., Bernhart, S. H., Neuböck, R. & Hofacker, I. L. The vienna RNA websuite. *Nucleic Acids Research* **36**, W70–W74 (2008).

[109] Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nature Methods* **9**, 357 (2012).

[110] Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

[111] Milne, I. *et al.* Tablet—next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2009).

[112] Su, H., Shao, Z., Tkalec, L., Blain, F. & Zimmermann, J. Development of a genetic system for the transfer of DNA into Flavobacterium heparinum. *Microbiology* **147**, 581–589 (2001).