00591

UNIVERSIDAD NACIONAL AUTÓNOMA
DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

GRUPO DE BIOLOGÍA COMPUTACIONAL

Señales conservadas en regiones intergénicas bacterianas: riboswitches y más allá

T E S I S
QUE PARA OBTENER EL GRADO DE
D O C T O R E N C I E N C I A S
P R E S E N T A :
C E I A B R E U G O O D G E R

DIRECTOR DE TESIS:
DR. ENRIQUE MERINO PÉREZ

CUERNAVACA, MORELOS SEPTIEMBRE DE 2005

m347791

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.

Director de Tesis:

Dr. Enrique Merino

Miembros del Comité Tutoral:

Dr. Alejandro Garcíarrubio

Dr. Enrique Merino

Dr. Miguel Ángel Cevallos

Miembros del Comité Tutoral Ampliado:

Dr. Alejandro Garcíarrubio

Dr. Enrique Morett

Dr. Julio Collado

Dr. Lorenzo Segovia

Dr. Miguel Ángel Cevallos

Miembros del Jurado:

Dr. Enrique Merino

Dr. Humberto Flores

Dr. José Luis Puente

Dr. Juan Miranda

Dr. Julio Collado

Dr. Lorenzo Segovia

Dr. Pablo Vinuesa

Agradezco al CONACyT y la DGEP-UNAM por becas otorgadas durante la realización de este proyecto.

ÍNDICE

ABSTRACT

The increasing number of publicly available whole-genome sequences opens opportunities to address new and important biological questions with comparative genomic approaches. In this case, the motivation was to identify conserved regulatory motifs, without using any knowledge of regulon structure or metabolic pathways, so that general and unbiased searches could be performed.

A method was developed, to search for over-represented motifs in the regulatory regions of genes clustered by their orthologous relationships, as defined in the COG database. Initially, "seed motifs" were located for each group of orthologous genes, which were then used to identify other members of the putative regulon by cycling between searching all upstream regions from fully sequenced bacterial genomes and redefining each motif from the new set of matching regulatory regions. The resulting "refined motifs" represent candidate regulatory elements. In order to evaluate the likelihood that these "refined motifs" represent biologically important elements, those that match against known proteins or RNA genes were first eliminated. The remaining motifs were then assigned a possible function; by calculating a *p*-value of how likely they are to regulate the genes of a given COG or KEGG pathway. The genome context congruence of the genes putatively regulated by each "refined motif" was also verified using our web server *GeConT*. The comparison of the final set of motifs against the Rfam database revealed that all currently reported riboswitches could be located by our method. In addition, many other highly conserved RNA cis-regulating elements were identified, such as the Gram positive T-box. Finally, *RibEx*, a web server to search any sequence for known riboswitches as well as predicted regulatory elements, was developed. This server also allows the visual inspection of the identified motifs, in relation to predicted attenuators and open reading frames.

Our results show that for a great many regulatory elements, their conservation is strong enough to be detected in a single group of orthologous genes, without using information of known regulons or metabolic pathways. The method detects previously unknown regulatory sites and can even group whole sets of uncharacterized genes into

possible regulons. The generated data should be of great use in directing new experiments since it defines not only the set of regulated genes, but also the location of the regulatory sequence, as well as possible functional relationships according to the KEGG or COG assignations.

RESUMEN

La cantidad de genomas completamente secuenciados sigue aumentando, lo cual permite afrontar nuevas preguntas biológicamente relevantes, mediante enfoques de genómica comparativa. En este caso la motivación fue identificar elementos conservados de regulación sin utilizar conocimiento previo de regulones o caminos metabólicos, permitiendo búsquedas globales con los menores sesgos posibles.

Se desarrolló un método para encontrar motivos sobre-representados en las regiones de regulación de genes agrupados por sus relaciones de ortología, definidos en la base de datos COG. Inicialmente, se localizaron "motivos semilla" para cada grupo, los cuales fueron utilizados para encontrar nuevos miembros del probable regulón ciclando entre localizar cada motivo en las regiones de regulación de todas las bacterias y redefinir el motivo con las nuevas secuencias encontradas. Los "motivos refinados" resultantes son los candidatos a ser elementos de regulación. Para evaluar la factibilidad de que estos candidatos fueran realmente elementos regulatorios de importancia biológica, primeramente se eliminaron todos aquellos que estuvieran contenidos en algún gene. A los motivos restantes se les asignó una función tentativa, al calcular un valor de probabilidad de que regularían a los genes de un COG o camino metabólico del KEGG. La congruencia del contexto genómico de cada "motivo refinado" se verificó usando nuestro servidor web *GeConT*. El conjunto final de motivos se comparó posteriormente a la base de datos Rfam, donde se encontró que el método fue capaz de encontrar todos los riboswitches reportados, además de varios elementos de regulación basados en

estructuras en el RNA, como la T-box de organismos Gram positivos. Finalmente, se desarrolló *RibEx*, un servidor web que permite buscar riboswitches y elementos regulatorios predichos en cualquier secuencia. El servidor permite la inspección visual de los motivos encontrados, en relación a marcos abiertos de lectura y atenuadores predichos.

Los resultados muestran que para una gran cantidad de elementos regulatorios, su conservación permite que sean detectados en un solo grupo de genes ortólogos, sin utilizar información de regulones o caminos metabólicos conocidos. El método permite la detección de sitios de regulación previamente desconocidos y hasta puede agrupar genes de función desconocida en posibles regulones. Los datos generados serán de gran utilidad para dirigir nuevos experimentos, puesto que definen tanto los genes regulados como la posición y secuencia de los elementos regulatorios además de posibles funciones, de acuerdo a las asignaciones del KEGG o COG.

1. INTRODUCCIÓN

Los seres vivos se caracterizan, entre otras cosas, por su capacidad de responder a estímulos de su exterior. Desde el organismo unicelular más primitivo, hasta el más complejo de los animales es capaz de percibir cambios en su medio y reaccionar de una manera que cree apropiada. Muchas de estas respuestas, en un momento dado, requerirán de cambios en el estado de expresión del genoma: genes que estaban prendidos necesitarán ser apagados y genes latentes necesitarán ser activados. Uno de los primeros trabajos en donde se estudió este fenómeno, fue la caracterización de cómo *Escherichia coli* mantiene control sobre los genes necesarios para utilizar lactosa [38]. En ausencia de lactosa, la proteína represora LacI se encuentra unida al DNA manteniendo reprimido al operón *lacZYA*. Cuando la bacteria empieza a internalizar lactosa, ésta se une a LacI desactivándolo, causando que se desprenda del DNA y que así pueda iniciarse la transcripción. En este caso la señal del medio es la concentración de lactosa y LacI es el elemento específico capaz de percibir la señal y transformarla en una respuesta.

En general, los reguladores transcripcionales suelen ser proteínas que se unen al DNA en sitios específicos (regiones operadoras o de regulación) y que son capaces de dirigir la expresión de los genes contiguos. Las regiones de regulación contienen por ende la información de qué genes se deben expresar en qué condiciones. De nada sirve tener un gene para la asimilación de lactosa, si no se puede regular apropiadamente: si el gene está activo en ausencia del azúcar, representa un desperdicio de recursos para la célula y si está apagado cuando es requerido, su utilidad es prácticamente nula. El descifrar entonces la información contenida en las regiones de regulación de los organismos, llevará a entender mucho mejor la manera en que éstos interactúan con su medio, sus capacidades y sus limitaciones. El presente trabajo pretende realizar un análisis exhaustivo de las unidades de regulación compartidas entre organismos distantes, es decir: las más conservadas, las más antiguas.

1.1. Búsquedas de sitios de regulación

Para poder activar coordinadamente los genes necesarios para una respuesta, los seres vivos han adoptado diversas estrategias. Las bacterias suelen agrupar genes funcionalmente relacionados en operones, genes contiguos que se transcriben como una sola unidad y que por lo tanto responden a una misma señal vía su región de regulación [39]. Adicionalmente, cuando distintas regiones de regulación comparten un mismo elemento (el sitio de pegado de un regulador transcripcional), los genes aledaños a estos sitios pueden responder coordinadamente a un mismo estímulo, a pesar de encontrarse en regiones distantes del genoma, conformándose así un regulón. Desde organismos multicelulares hasta procariontes utilizan esta última estrategia de corregulación.

La existencia de regulones ha sido aprovechada para encontrar elementos o sitios de regulación. Si se conoce gran parte de los genes que forman parte de un regulón, una estrategia factible es averiguar las secuencias o motivos que están enriquecidos en las regiones controladoras del regulón, pero ausentes del resto del genoma [18, 57, 85]. Los motivos así obtenidos son muy buenos candidatos a ser los sitios donde se une un regulador transcripcional. Este tipo de estrategias ha sido sobre todo muy útil a partir de la reciente explosión de resultados de experimentos con microarreglos, de los cuales se pueden descubrir nuevos regulones, al inferir corregulación a partir de coexpresión [23, 42]. La limitante de predecir sitios de regulación se presenta cuando la señal es muy débil, ya sea porque el sitio de pegado no está muy conservado, o porque existen muchos falsos positivos dentro del conjunto de posibles genes corregulados, diluyéndose así la señal. Para fortalecerla, se pueden agregar regiones de regulación adicionales provenientes de genes ortólogos de organismos cercanos. En general, se observa que los sitios de pegado de proteínas al DNA divergen rápidamente y solamente dentro de un mismo organismo, o en organismos muy cercanos, se conservan lo suficiente como para ser detectados.

1.2. RNA y regulación

El momento durante la expresión genética donde la regulación es más importante es sin duda el inicio de la transcripción, viéndose involucrados los promotores y otros sitios reguladores del DNA. Sin embargo, existen procesos críticos para la regulación posterior al inicio de la transcripción que dependen de RNA.

1.2.1. Atenuación

Uno de los primeros mecanismos de regulación en el que se mostró que el RNA juega un papel importante es la atenuación [34]. Descrito inicialmente para el operón biosintético de triptofano en *E. coli* [35], hoy se sabe que es una estrategia ampliamente utilizada por las bacterias para la regulación de sus genes [58]. La atenuación transcripcional se basa en la posibilidad que tiene el líder no traducido de un mRNA para adoptar diversas estructuras mutuamente excluyentes, una de los cuales es un terminador intrínseco. El terminador es una estructura tipo tallo y asa el cual ocasiona que la RNA polimerasa que lo acaba de transcribir haga una pausa, inmediatamente río-abajo de esta estructura deberá haber una secuencia rica en uracilos, que ocasiona que la polimerasa se desprenda, dándose por terminada la transcripción. Para que pueda modularse esta terminación, existe una estructura sobrelapada al terminador llamado anti-terminador, que si se forma impide que se estructure el terminador, permitiendo que continúe la transcripción. A veces también existe una tercera estructura, un anti-antiterminador, el cual análogamente evita la formación del anti-terminador, por lo cual el terminador sí se forma y la transcripción se ve terminada. Los mecanismos que las bacterias utilizan para dirigir la estabilización de la estructura apropiada en la atenuación son varios, pero uno recurrente es que el mismo ribosoma traduzca un pequeño péptido líder, que se localiza río arriba del atenuador. El péptido líder es rico en el aminoácido cuya concentración regula la expresión de la unidad transcripcional y por lo tanto, en ausencia del aminoácido (específicamente del tRNA cargado), el ribosoma se detiene. Al ser el anti-terminador la estructura más estable, continúa la transcripción de los genes necesarios para sintetizar el aminoácido en cuestión. Cuando la concentración del aminoácido

aumenta a cierto nivel, el ribosoma ya no se detiene y al avanzar impide la formación de la estructura del anti-terminador, viéndose favorecido el terminador y ocasionando así la terminación de la transcripción (Figura 1).

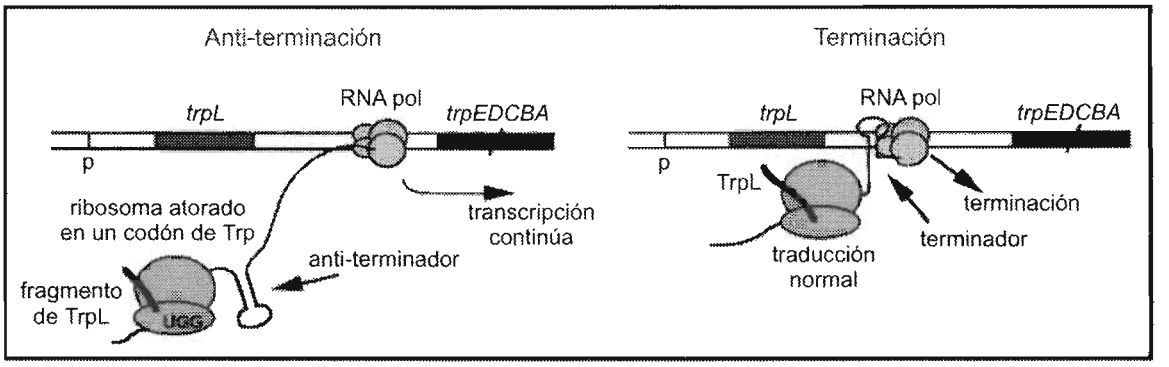

Figura 1. Esquema de atenuación mediado por el ribosoma. En ausencia de triptofano, el ribosoma se detiene en los codones Trp, se forma un anti-terminador y la transcripción continúa. En presencia de triptofano, el ribosoma traduce correctamente el péptido TrpL, planchando la estructura del anti-terminador, ocasionando así la terminación del transcrito. Modificado de [34].

1.2.2. RNA no codificante (*trans*-reguladores)

Recientemente se ha destacado la relevancia de pequeños RNA no codificantes (ncRNA), que no necesitan ser traducidos para ejercer su función [37]. En eucariontes se ha descrito una familia de pequeños RNA nucleolares (snoRNA) involucrados en la modificación de nucleótidos en rRNA y tRNA principalmente [6]. En cambio los micro-RNA (miRNA) y los pequeños RNA interferentes (siRNA) regulan genes codificantes, al actuar directamente sobre diversos mRNA [63]. Estas dos últimas familias de ncRNA son fragmentos muy pequeños, de alrededor de 21 nucleótidos, que se unen por complementariedad de bases a su mRNA blanco, dirigiendo así su degradación [83]. El mecanismo por el cual actúan los siRNA, nombrado interferencia por RNA o RNAi, es de sumo interés ya que pretende ser explotado para fines terapéuticos; en principio se puede dirigir la degradación específica de cualquier mensajero dentro de una célula, como el de genes virales u oncogenes [79]. En bacterias se han encontrado también pequeños ncRNA, algunos de los cuales se unen a sus mRNA blanco para interferir con la traducción o estabilidad, otros interactúan directamente con proteínas para modular su

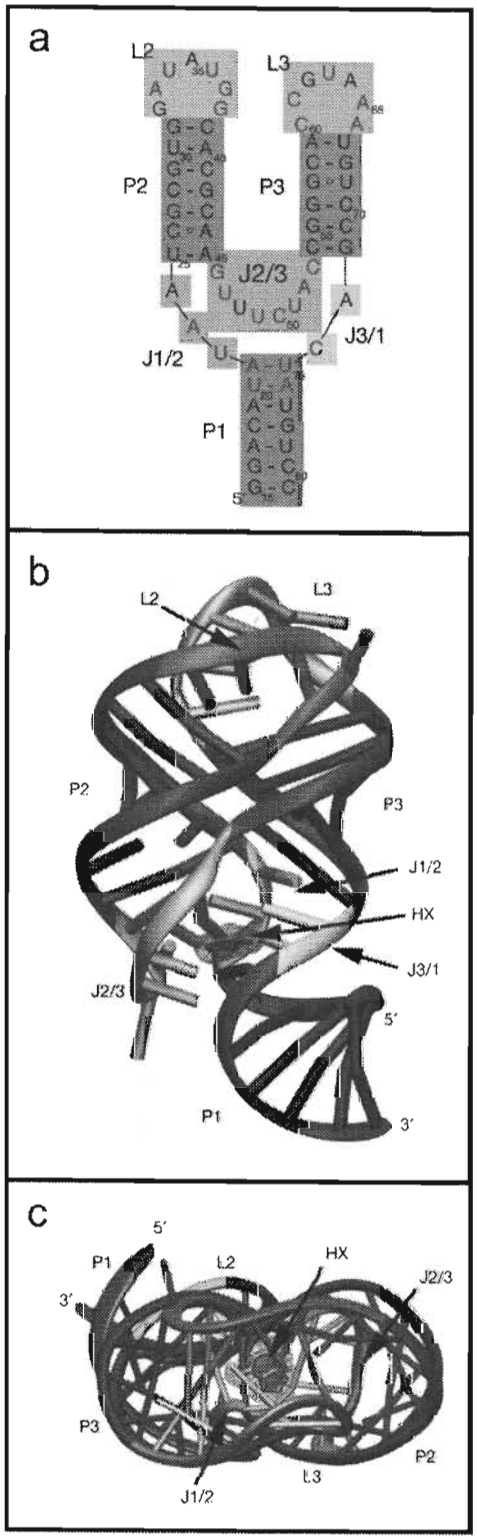

Figura 2. Riboswitch de guanina. a) Estructura secundaria, b) estructura terciaria lateral y c) vista superior. Se puede ver que el RNA envuelve completamente a la HX, un análogo de guanina. Modificado de [12]

actividad [50]. En *E. coli*, el número de pequeños ncRNA conocidos se acerca al 2% del número de los genes codificantes [27].

1.2.3. Riboswitches y otros *cis*-reguladores

Otra clase novedosa de ncRNA son los llamados riboswitches (revisado en: [54, 90]). A diferencia de los antes mencionados, los riboswitches no son moléculas libres, sino son parte íntegra de un mRNA y presentan su efecto, de represión o activación, exclusivamente sobre éste. Se caracterizan por unir pequeños metabolitos, como vitaminas o nucleótidos, con alta afinidad y especificidad, en total ausencia de proteínas, mediante una compleja estructura formada por el RNA (Figura 2) [12]. A pesar de la notoriedad reciente de los riboswitches, se conocen desde hace algún tiempo elementos regulatorios integrados al RNA. Muchas tRNA sintetasas y otros genes involucrados en el metabolismo de aminoácidos en organismos Gram positivos, se encuentran regulados por una compleja estructura llamada T-box, capaz de reconocer una molécula de tRNA no cargada [29]. En este mismo grupo de organismos, varios genes de la biosíntesis de pirimidinas dependen para su regulación del acoplamiento entre la proteína PyrR y una estructura de tipo tallo y asa en el RNA [48]. Y en *Bacillus subtilis* la regulación del metabolismo de triptofano se da en parte por la proteína TRAP que,

activada por triptofano, se une a la región líder del mRNA [5]. Lo que distingue a los riboswitches es que reconocen directamente a la señal o metabolito, sin necesidad de ningún intermediario, ya sea tRNA o proteína. Sin embargo, dado que todos estos sistemas funcionan como "interruptores" cuando una molécula (aunque se trate de una proteína) se une al RNA, el término riboswitch pudiera empezar a ser extendido para incluirlos [32].

1.3. Descubrimiento de los riboswitches

Después de los trabajos sobre regulación de Jacob y Monod [38], parecía muy claro que las proteínas eran las encargadas de modular la expresión genética. Sin embargo, existían algunos sistemas regulatorios que no se podían explicar adecuadamente. Se había visto que los genes de biosíntesis de la vitamina B_{12} (cobalamina) en *E. coli* y *Salmonella typhimurium* se reprimían en presencia de este cofactor, pero el regulador proteico que se suponía debía existir permanecía elusivo [49, 70]. Un caso similar ocurría para la biosíntesis de la vitamina B_1 (tiamina) [59]. En estos casos se observó que la región líder no-traducida del transcrito era particularmente larga y contenía regiones críticas para el funcionamiento de la represión. Un análisis cuidadoso llevó a la conclusión de que estas regiones podían formar estructuras secundarias estables las cuales se encontraban conservadas en distintos organismos, surgiendo la posibilidad de que la estructura del RNA detectara directamente a las vitaminas [26, 60]. Esta propuesta se justificaba ya que se conocía desde hace algún tiempo la capacidad que tenía el RNA de adoptar estructuras que podían reconocer con alta afinidad y especificidad a una variedad de pequeñas moléculas [24]. Aunque estos aptámeros habían sido generados mediante procesos de selección *in vitro*, parecía plausible que existieran moléculas con esta capacidad en la naturaleza. El término riboswitch se acuñó cuando se probó experimentalmente que el RNA unía directamente al metabolito, en ausencia de proteínas, ocasionando la modulación de la expresión genética [94]. Desde entonces, mediante diversas estrategias tanto computacionales como experimentales, se han ido describiendo laboriosamente varios riboswitches, capaces de reconocer directamente a las

vitaminas tiamin-pirofosfato (TPP) [73], adenosil-cobalamina (Ado-CBL) [62, 74, 89], S-adenosil-metionina (SAM) [25, 97] y flavin-mononucleótido (FMN) [88, 95], a los nucleótidos adenina [53] y guanina [52], y a los aminoácidos lisina [33, 75, 81] y glicina [55] (Figura 3).

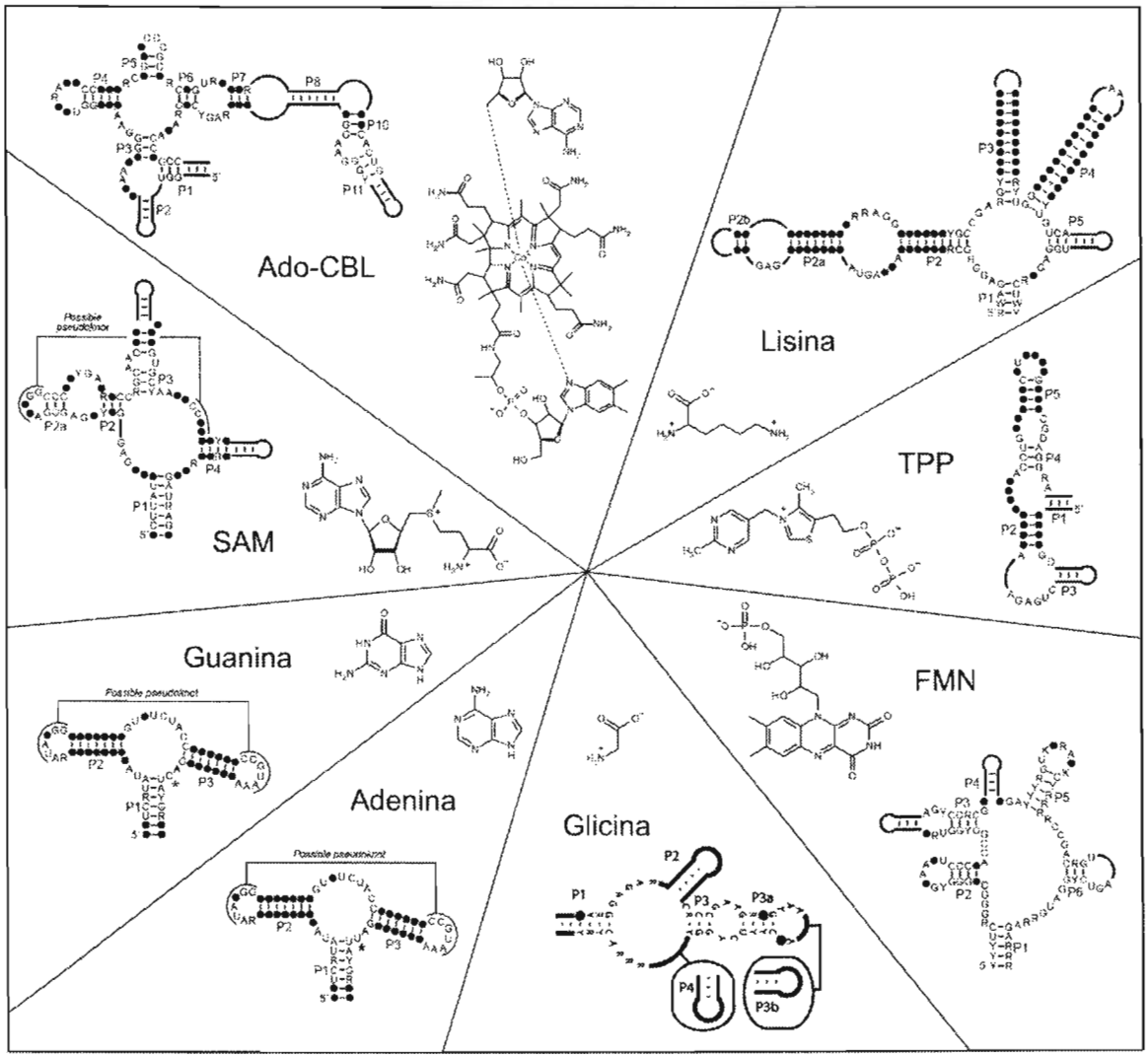

Figura 3. Modelos de las estructuras secundarias y de los metabolitos unidos por los distintos riboswitches conocidos. Modificado de [78].

1.4. Origen de los riboswitches

Una de las teorías más aceptadas del origen de la vida considera que hubo un periodo donde la molécula más importante era el RNA. Apoyo para esta idea viene de que actualmente el RNA está involucrado en muchos procesos cruciales para los seres vivos. Los genes deben pasar por un RNA mensajero intermediario, este mensajero es traducido por un complejo aparato que depende principalmente de RNA ribosomal para su funcionamiento y de RNA de transferencia como conector entre el mensaje y los distintos aminoácidos. Aún los procesos que generalmente se consideran exclusivos de otras moléculas, son desempeñados en algunas ocasiones por RNA. Existen virus cuyos genomas son de RNA en lugar de DNA y las ribozimas son moléculas de RNA capaces de catalizar reacciones bioquímicas en ausencia de proteínas [22].

Muchos de los riboswitches actuales pudieran ser remanentes del mundo del RNA, donde cofactores y otras moléculas podrían haber provocado cierto efecto al ser reconocidos por una molécula de RNA. Las ribozimas primitivas probablemente dependían de estos cofactores como reguladores alostéricos, en semejanza a como hoy ocurre con las enzimas. De esta manera, el aptámero capaz de unir a un metabolito, podía ser aprovechado de múltiples maneras, acoplado a un RNA catalítico o a un RNA génico. El DNA pasaría a sustituir al RNA como repositorio de la información genética y poco a poco las proteínas irían sustituyendo a las moléculas de RNA, sobre todo las enzimas que debido a su mayor complejidad pueden presentar ventajas considerables en velocidades de catálisis [16]. El RNA empezaría a quedar relegado a ser principalmente un intermediario en el procesamiento de la información genética. Algunos RNA reguladores se mantendrían, sobre todo aquellos cuya función requería una interacción con otro segmento de RNA, como los miRNA y siRNA, puesto que la eficiencia de la interacción por apareamiento de bases es mucho mayor que la lograda por una proteína. Los riboswitches originales, tenderían a verse sustituidos por proteínas más eficientes. Sin embargo, en algunos casos, los riboswitches podrían tener ventajas insuperables por las proteínas, como el hecho de poder reaccionar inmediatamente a una señal al empezar a ser transcrito el gene, sin necesidad de que la célula haya invertido previamente en producir una cantidad de proteína regulatoria. Si una proteína, entonces, no podía

presentar una ventaja selectiva, el riboswitch probablemente se mantendría. Los riboswitches que reconocen TPP, Ado-CBL y FMN pudieran caer en esta categoría, por su amplia distribución entre los genomas secuenciados hasta el momento. Cabe notar que se ha observado la presencia de aptámeros funcionales específicos para la TPP acoplados a genes relacionados con el metabolismo de tiamina en los tres reinos [80]. En eucariontes, los mecanismos de acción tendrían que ser diferentes, por no tener acoplados los procesos de transcripción y traducción. Además, los aptámeros se han encontrado en los intrones y regiones 3' no traducidas, en lugar de en las regiones 5' lideres como en las bacterias. Así, en eucariontes podrían estar relacionados con el procesamiento o la estabilidad del mensajero. De hecho se ha visto *in vivo* que la regulación y el splicing del gene *thiA* del hongo filamentoso *Apsergillus oryzae* depende de un aptámero de TPP, presente en uno de sus intrones [45]. Sin embargo, no todos los riboswitches se encuentran tan ampliamente distribuidos. El que reconoce a guanina, por ejemplo, se encuentra principalmente en firmicutes, haciendo más parsimonioso pensar que se originó justo previo a la separación de este grupo y que se transfirió horizontalmente a los demás organismos aislados que lo contienen. Recientemente se descubrieron cuatro nuevos riboswitches, cuya distribución es mucho más reducida y su función aún no muy clara [10]. Estos también pudieran ser casos de riboswitches surgidos más recientemente, pues de lo contrario habría que invocar muchísimos eventos independientes de pérdida, lo cual no resulta muy convincente.

1.5. Mecanismo de acción de los riboswitches

Los riboswitches conocidos se encuentran sumamente conservados, probablemente debido a que deben interaccionar con una molécula que no ha cambiado en el tiempo, como lo es una vitamina, un aminoácido o un nucleótido. La parte más conservada es por consiguiente el aptámero o elemento sensor, que es la región que reconoce al metabolito. Adicionalmente, para llevar a cabo su función biológica regulatoria, un riboswitch debe poder prender o apagar la expresión de los genes con los cuales está asociado. El segmento del riboswitch que de esto se encarga, el elemento

efector, se basa generalmente en la formación en el RNA de estructuras mutuamente excluyentes. De tal modo, la unión del metabolito estabiliza una de las estructuras, impidiendo la formación de otra. Las estructuras que hasta ahora se han descrito constan de atenuadores transcripcionales, secuestradores del sitio de entrada del ribosoma comúnmente conocido como Shine-Dalgarno (SD) e inclusive un caso de una ribozima integrada al mensajero que se activa con la unión del ligando [65, 96]. Es común que distintos organismos utilicen un mismo elemento sensor acoplado a distintos elementos efectores. Aunque no es una regla, muchos organismos Gram negativos aprovechan secuestradores del SD (también conocidos como atenuadores traduccionales), mientras que los Gram positivos suelen acoplar sus riboswitches a atenuadores transcripcionales.

1.6. Detección de nuevos riboswitches

Al inicio de este proyecto, en el laboratorio se estaba planteando realizar un método de búsqueda de nuevos riboswitches, el cual llevó a un trabajo de tesis de licenciatura [67]. La idea fue tomar todos los genes para los que se tuviera información funcional relacionada con el proceso biosintético de algún metabolito y buscar en sus regiones de regulación un patrón conservado. Se empezó a hacer una lista de vitaminas, nucleótidos y cofactores para realizar las búsquedas, cuando surgió la idea de paralelamente intentar desarrollar un método más completo y robusto, que no dependiera de funciones conocidas y con el cual se pudiera explorar la mayor parte del universo bacteriano conocido. Así, se propuso analizar todos los genes de todos los genomas completos, agrupados por ortología, con la hipótesis de que de existir un nuevo riboswitch, éste probablemente estaría regulando al menos uno de estos grupos de genes y que a partir de este grupo se pudiera empezar la caracterización.

Los trabajos que hasta ahora se habían propuesto la identificación de riboswitches constaron en análisis dirigidos a un sistema de regulación en particular. En varios casos, como se menciona en la sección "Descubrimiento de los riboswitches", existía el conocimiento de un regulón pero no se encontraba el regulador. Muchos grupos realizaron análisis de genómica comparativa y en caso de haber evidencia de una

estructura secundaria conservada, se prosiguió a probar experimentalmente si esta estructura podía unir al metabolito regulador. Durante el desarrollo de esta tesis, un grupo realizó un trabajo con un enfoque bastante similar, de buscar riboswitches en un gran número de genes, sin utilizar conocimiento de regulones [10]. Sin embargo, se enfocaron exclusivamente a genes de los que *Bacillus subtilis* tuviera una copia, mientras que el presente trabajo considera a todas las bacterias completamente secuenciadas.

2. HIPÓTESIS

- Los riboswitches y otros elementos regulatorios muy conservados y ampliamente distribuidos tenderán a regular al menos un grupo de genes ortólogos, se conozca o no la función de estos genes.

- Algunos genes ortólogos, a pesar de encontrarse en genomas filogenéticamente distantes, presentarán un mismo sistema de regulación conservado.

3. OBJETIVOS

- Elaborar un método general para la detección de elementos regulatorios muy conservados (específicamente de tipo riboswitch), que no requiera utilizar información de anotación para tener la cobertura más amplia posible.
    - Redescubrir riboswitches conocidos para probar la efectividad del método.
    - Descubrir nuevos candidatos de riboswitch, para dirigir posteriores análisis computacionales y experimentales.

- Realizar un análisis sobre la naturaleza de los elementos más conservados en las regiones regulatorias bacterianas.

estructura secundaria conservada, se prosiguió a probar experimentalmente si esta estructura podía unir al metabolito regulador. Durante el desarrollo de esta tesis, un grupo realizó un trabajo con un enfoque bastante similar, de buscar riboswitches en un gran número de genes, sin utilizar conocimiento de regulones [10]. Sin embargo, se enfocaron exclusivamente a genes de los que *Bacillus subtilis* tuviera una copia, mientras que el presente trabajo considera a todas las bacterias completamente secuenciadas.

2. HIPÓTESIS

- Los riboswitches y otros elementos regulatorios muy conservados y ampliamente distribuidos tenderán a regular al menos un grupo de genes ortólogos, se conozca o no la función de estos genes.
- Algunos genes ortólogos, a pesar de encontrarse en genomas filogenéticamente distantes, presentarán un mismo sistema de regulación conservado.

3. OBJETIVOS

- Elaborar un método general para la detección de elementos regulatorios muy conservados (específicamente de tipo riboswitch), que no requiera utilizar información de anotación para tener la cobertura más amplia posible.
    - Redescubrir riboswitches conocidos para probar la efectividad del método.
    - Descubrir nuevos candidatos de riboswitch, para dirigir posteriores análisis computacionales y experimentales.
- Realizar un análisis sobre la naturaleza de los elementos más conservados en las regiones regulatorias bacterianas.

- Identificar qué otros elementos se encuentran tan conservados en adición a los riboswitches.

- Presentar los resultados de tal forma que puedan ser aprovechados por la mayoría de la comunidad científica.

4. DESARROLLO Y MÉTODOS

A grandes rasgos, el procedimiento seguido consistió en obtener grupos de regiones de regulación ortólogas y aplicar un algoritmo de descubrimiento de patrones o motivos. Los motivos encontrados posteriormente son localizados dentro de todas las regiones de regulación de los genomas completamente secuenciados, dejando atrás la limitante de los grupos de ortólogos iniciales. Como producto final del análisis, se obtiene una serie de motivos, que corresponden a los posibles elementos conservados de regulación, junto con una lista de los genes que son candidatos a ser regulados por estos elementos (Figura 4).

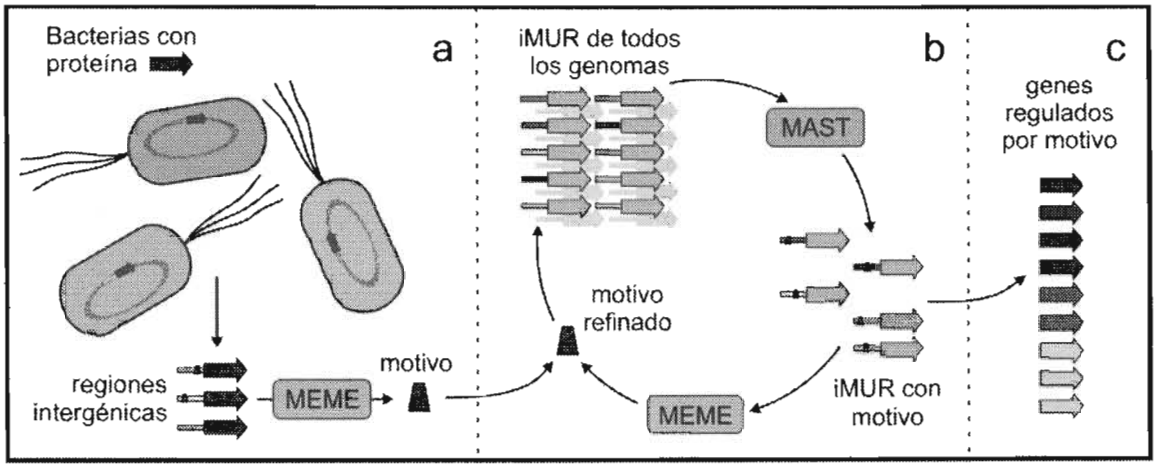

Figura 4. Resumen del método para detectar elementos regulatorios conservados. a) Se toman todas las regiones intergénicas de las proteínas de un mismo COG (flecha roja) y con MEME se encuentran motivos sobre-representados. b) Cada motivo es refinado al ciclar entre buscarlo con MAST en todas las regiones intergénicas (iMUR) de genomas completamente secuenciados y redefinir el motivo con MEME a partir de las nuevas secuencias encontradas. c) Al final, se obtiene una lista de genes regulados por el

- Identificar qué otros elementos se encuentran tan conservados en adición a los riboswitches.

- Presentar los resultados de tal forma que puedan ser aprovechados por la mayoría de la comunidad científica.

4. DESARROLLO Y MÉTODOS

A grandes rasgos, el procedimiento seguido consistió en obtener grupos de regiones de regulación ortólogas y aplicar un algoritmo de descubrimiento de patrones o motivos. Los motivos encontrados posteriormente son localizados dentro de todas las regiones de regulación de los genomas completamente secuenciados, dejando atrás la limitante de los grupos de ortólogos iniciales. Como producto final del análisis, se obtiene una serie de motivos, que corresponden a los posibles elementos conservados de regulación, junto con una lista de los genes que son candidatos a ser regulados por estos elementos (Figura 4).

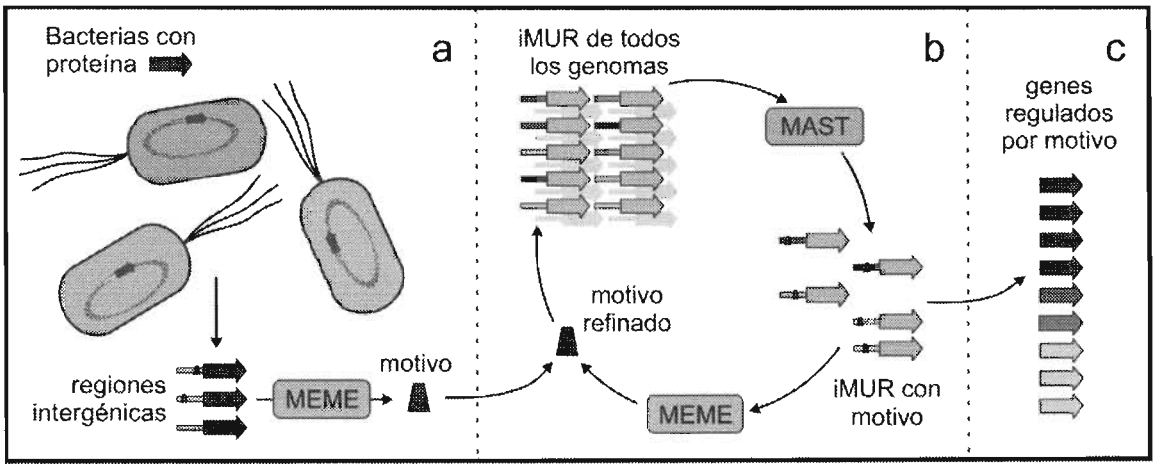

Figura 4. Resumen del método para detectar elementos regulatorios conservados. a) Se toman todas las regiones intergénicas de las proteínas de un mismo COG (flecha roja) y con MEME se encuentran motivos sobre-representados. b) Cada motivo es refinado al ciclar entre buscarlo con MAST en todas las regiones intergénicas (iMUR) de genomas completamente secuenciados y redefinir el motivo con MEME a partir de las nuevas secuencias encontradas. c) Al final, se obtiene una lista de genes regulados por el

motivo refinado, incluyendo miembros de otros COGs además del inicial. Se presume que todos éstos pertenecen a un mismo regulón.

4.1. Genomas no-redundantes

Existe un gran sesgo en cuanto a los genomas bacterianos que han sido elegidos para ser secuenciados. Por un lado, principalmente se trabaja con aquellos de interés económico: causantes de enfermedades (*Helicobacter pylori, Mycoplasma genitalium, Streptococcus pneumoniae*), productores de enzimas (*Bacillus halodurans/subtilis*), antibióticos (*Streptomyces avermitilis/coelicolor*), etanol (*Zymomonas mobilis*), de importancia para la industria agrícola (*Agrobacterium tumefaciens, Mesorhizobium loti, Pseudomonas syringae*), etc. Por otro lado, se han secuenciado varios genomas sumamente parecidos, en el caso extremo cuatro cepas de *E. coli* y cinco de *Salmonella*. Esta redundancia, aunque de sumo interés, por ejemplo para estudios de genómica comparativa donde se desea dilucidar las bases de las diferencias de patogenicidad de las distintas cepas, puede ser perjudicial en un análisis como el presente si no es debidamente manejada.

Cuando se utiliza un algoritmo de descubrimiento de patrones se le proporciona un conjunto de secuencias y el algoritmo regresa el patrón más sobre-representado. Se asume que las secuencias son independientes, es decir, que no hay redundancia en ellas más allá del motivo que se desea descubrir. Debido a la gran redundancia que existe en los genomas secuenciados, un paso importante consiste en generar un conjunto de genomas no-redundantes. Se empleó CVTree (Árbol de Vectores de Composición, por sus siglas en inglés), un programa que toma los proteomas de genomas completos y mediante su contenido de oligopéptidos calcula una distancia "evolutiva" entre cada par de genomas [69]. Los autores de CVTree comentan que estas distancias se pueden utilizar para construir árboles filogenéticos que son equiparables al árbol estándar construido a partir del RNA ribosomal de la subunidad pequeña. Con los datos arrojados por CVTree, se eligió una distancia de corte, tal que las cepas representativas de un mismo género nunca fueran consideradas como no-redundantes. Para dar una idea del nivel de

agrupación que se empleó, con el corte de distancia de 0.40 elegido, quedan como un solo grupo las 4 cepas de *E. coli*, las 5 de *Salmonella* y las 2 cepas de *Shigella*. El total de 230 bacterias cuyo genoma completo se encontraba disponible durante la escritura de esta tesis se redujeron a 145 organismos no-redundantes.

4.2. Regiones de regulación exclusivamente intergénicas

Para analizar regiones de regulación bacterianas, se suele tomar un fragmento de secuencia equivalente a 400 nucleótidos río arriba de cada gene más 50 nucleótidos río adentro, al cual se le ha denominado MUR (Región Mínima Río-arriba, por sus siglas en inglés) [Gabriel Moreno-Hagelsieb, comunicación personal]. La MUR contiene en general la mayoría de los elementos regulatorios del gene en cuestión y hasta los riboswitches, que suelen ser elementos muy grandes, caben dentro de ella. Sin embargo, al emplear una técnica de descubrimiento de patrones, se debe tener sumo cuidado de no tener "contaminación" alguna de secuencias más conservadas que los motivos que se desea encontrar. En este caso, se debe eliminar toda región codificante de las MUR, ya que estas regiones presentan por mucho una mayor conservación que los elementos de regulación. Para ello, se construyó una base de datos de regiones llamadas iMUR (MUR intergénica) de la siguiente manera. Primeramente se extrajeron todas las posiciones de genes (tanto de proteínas como de RNA) de los archivos GenBank de los genomas completamente secuenciados [14], mediante el programa GBKencoder [Ricardo Ciria, comunicación personal]. Adicionalmente, se produjo para cada replicón un archivo donde la secuencia nucleotídica de los genes estuviera marcada (colocada en mayúsculas, por ejemplo), pero las secuencias intergénicas se mantuvieron intactas. Posteriormente, se procesaron estos archivos y cada región intergénica fue asignada al gene que le correspondía. En caso de ser muy grandes las regiones intergénicas, se recortaron a 400 nucleótidos en semejanza con las MUR.

4.3. Grupos de genes ortólogos

Una hipótesis sobre la cual descansa el presente trabajo, es que algunos genes ortólogos en genomas distantes son regulados por un mismo elemento. Esto en general no ocurre cuando la regulación se da por proteínas que se unen al DNA, o si ocurre, el sitio de unión de dicha proteína generalmente varía tanto que no es distinguible como un patrón conservado. Sin embargo, para los riboswitches conocidos, esta hipótesis parece cumplirse, por lo cual se espera que en el presente proyecto se recuperen mayoritariamente señales tipo riboswitch, en lugar de elementos de regulación que dependan de interacciones DNA-proteína.

Aunque la definición de ortólogo implica una relación de descendencia vertical, donde solamente un evento de especiación separó a los genes, computacionalmente se suele definir ortología mediante un concepto conocido como *mejor hit bi-direccional*. Dos genes de organismos diferentes son mejores hits bi-direccionales, y por lo tanto ortólogos, si a) el primer gene encuentra al segundo con la mejor calificación al compararse contra todos los genes del segundo organismo y b) el segundo gene encuentra al primero con la mejor calificación al compararse contra todos los genes del primer organismo. Como grupos iniciales de genes ortólogos se tomó la base de datos COG (Clusters de Grupos de genes Ortólogos, por sus siglas en inglés) [82]. Ésta se construye inicialmente tomando los mejores hits bi-direccionales entre tres genomas de linajes diferentes. Posteriormente, se van agregando a este triángulo más proteínas siempre y cuando sean mejores hits bi-direccionales de al menos dos proteínas que ya se encuentren en el grupo. La versión más reciente agrupa las proteínas de 66 genomas en 4,873 grupos de ortólogos o COGs. La asignación de COG se realiza por dominio proteico, por lo que una proteína puede tener más de un COG si presenta más de un dominio. Debido a que al momento de la escritura de esta tesis se contaba ya con 230 genomas bacterianos completamente secuenciados, fue necesario asignar COGs a las proteínas del resto de los genomas. Primeramente se tomó las asignaciones extendidas a 179 organismos disponibles en los datos del servidor STRING versión 6 [91]. Para este servicio, asignan COGs a las proteínas de genomas nuevos con un proceso muy similar al que se empleó para la construcción de la base de datos original. Además, realizan agrupaciones

automatizadas en genes que no han podido ser asignados, resultando en una serie de grupos nuevos llamados NOGs (Grupos de Ortólogos No-supervisados, por sus siglas en inglés). Posteriormente, se empleó una rutina para procesar el resto de los genomas, en la cual un COG es asignado a una proteína siempre y cuando al menos 3 de sus ortólogos tuvieran asignados dicho COG y que el número de sus ortólogos pertenecientes a este COG representara al menos el 30% del total de las proteínas con el COG. Esto último con el fin de evitar expandir los COGs con muchas proteínas parálogas. Al final de todo el proceso se logró contar con 477,320 proteínas con al menos un COG o NOG asignado, de un total de 680,174 proteínas bacterianas (70%).

4.4. Regiones de regulación ortólogas

Al tener definidos un conjunto de regiones de regulación intergénicas (iMUR) y asignaciones de ortología de acuerdo a los COGs, se realizó una combinación de ambas para producir grupos de regiones de regulación ortólogas. Para hacer esto de la mejor manera posible, se tomó en cuenta la agrupación en operones que presentan los genes bacterianos, donde la región de regulación de cada gene de un operón es realmente la misma: la correspondiente al primer gen del operón. Debido a que no se cuenta con una información detallada de la estructura de los operones de todos los genomas, se eligió trabajar con predicciones computacionales basados principalmente en distancias intergénicas [61]. Así, para cada COG o NOG, se tomaron todas las proteínas así asignadas dentro del conjunto de genomas no-redundantes y para cada una de estas proteínas se localizó el operón al cual pertenecía y se tomó el iMUR correspondiente al primer gene de este operón. Cada conjunto de iMURs así obtenidos representa un grupo de regiones de regulación ortólogas.

4.4.1. Depuración de los grupos de regiones de regulación ortólogas

A pesar de haber eliminado la mayor parte de la redundancia a nivel de los genomas, aún puede existir debido principalmente a fenómenos de transferencia

horizontal. Dado que es muy importante no tener secuencias redundantes, se tiene que revisar cada grupo de regiones de regulación ortólogas y eliminarlas de ser encontradas. Para esto se empleó una rutina que utiliza BLAST [3] para comparar cada par de secuencias dentro de un conjunto. Si la región alineada correspondía al menos al 90% de la secuencia más chica con un valor de expectancia significativo (10^{-12}) las secuencias fueron consideradas redundantes. Después de todas las comparaciones, se realizó un proceso de agrupación o *clustering* y para cada grupo redundante se eligió una sola secuencia, la más larga en caso de existir diferencia de tamaño.

Debido a la eliminación de secuencias redundantes a varios niveles, primero por genoma y posteriormente por comparación de secuencias, muchos grupos ortólogos ya no contenían suficientes secuencias como para realizar correctamente un proceso de descubrimiento de patrones. Se eligió arbitrariamente un corte de mínimo 5 secuencias no-redundantes para que el grupo fuera considerado en el estudio. Así, se comenzó el análisis con 4,781 grupos de regiones de regulación ortólogas que pasaron este criterio. 4,079 de éstas pertenecientes a un COG y las 702 restantes a un NOG.

4.5. Descubrimiento de patrones con MEME

Para descubrir patrones o motivos conservados en los grupos de regiones de regulación ortólogas se utilizó un programa del dominio público: MEME (EM Múltiple para Elicitación de Motivos, por sus siglas en inglés) [8]. Este programa regresa de un conjunto de secuencias, uno o más motivos o patrones sobre-representados. Se basa en ajustar una mezcla de modelos a la serie de secuencias que se le proporciona. Un modelo es para el motivo que se busca y recibe tantas variables independientes como el tamaño del motivo, mientras que el modelo de fondo es más sencillo, recibiendo solamente las frecuencias de cada carácter (nucleótidos en este caso). Se prueba con un ciclo del algoritmo EM [46] todos los puntos de partida para elegir el más prometedor y para afinar la mezcla de modelos se prosigue a realizar ciclos de EM hasta convergir. MEME además compara los motivos que encuentra de distintos tamaños, para elegir el que más probablemente sea de relevancia biológica, con la ventaja de que el usuario no tiene que

conocer exactamente el tamaño. En general, se elige un E-value de corte, el rango de tamaño del motivo y el número máximo de motivos a encontrar y MEME los reporta, siempre y cuando no rebasen el valor de corte. De acuerdo al conocimiento que se tiene de los motivos que se espera encontrar, se le debe proporcionar su tipo de distribución a MEME, que lo usa para ajustar un mejor modelo al motivo. Las opciones son que el motivo: (a) se encuentre presente en una copia en todas las secuencias, (b) en sólo algunas, o (c) en más de una copia por secuencia. De acuerdo a lo que se esperaba de los elementos de regulación, se empleó el parámetro (b) que optimiza a MEME para encontrar motivos presentes en solo algunas de las secuencias de entrada. Además, se le pidió que buscara hasta 8 motivos, de cualquier tamaño entre 8 y 30 nucleótidos y con un valor de corte de 10^{-3}. Estos parámetros fueron elegidos después de hacer múltiples pruebas. En general el valor de corte empleado permite encontrar elementos de regulación presentes en solamente unas pocas de las secuencias de un grupo ortólogo. El cortar los motivos en un máximo de 30 nucleótidos es para evitar que motivos diferentes (pero contiguos) queden mezclados en uno sólo.

Existen grupos de ortólogos que pueden presentar más de un motivo de regulación, en algunos casos relacionados (motivos que son parte de un mismo elemento de regulación) y en otros independientes (un subconjunto de secuencias contiene un motivo y otro subconjunto un motivo totalmente diferente), por lo cual resultó conveniente permitir a MEME buscar hasta 8 motivos, los cuales fueron separados antes de continuar con el análisis. De no aplicarse este paso, uno de los motivos podía "opacar" a los demás, los cuales se perderían durante el resto del proceso. Además, motivos que constaran de más de 95% de G/C o A/T fueron eliminados, debido a su bajo contenido informacional. Al final se obtuvieron 3,778 motivos, aunque a este nivel muchos de estos motivos representan variantes de un mismo elemento de regulación, por lo que no son del todo independientes. Cabe mencionar que el procedimiento es sumamente robusto, por lo que si se varían un poco los parámetros de tamaño de motivo o valor de corte, los resultados son muy similares.

4.6. Localización de los motivos en los genomas con MAST

Para cada motivo encontrado por MEME se realizó una búsqueda usando MAST (Herramienta para Buscar y Alinear Motivos, por sus siglas en inglés) [9]). Esta herramienta es la contraparte de MEME; toma precisamente un motivo encontrado por éste (en forma de una matriz de peso posición-dependiente), una base de datos de secuencias y regresa las secuencias que contienen el motivo, la posición del motivo y una calificación de probabilidad. La búsqueda con MAST fue efectuada contra la base de datos completa de iMURs, es decir todas las regiones de regulación intergénicas de todos los genomas completos. De esta manera, para cada motivo, originalmente encontrado en un subconjunto de las regiones de regulación de un grupo de genes ortólogos, MAST localizó todos los genes que podrían estar regulados por el mismo elemento.

4.7. Agrupamiento inicial de motivos redundantes

Al tener la lista de genes que contienen cada motivo, se realizó otro proceso de agrupación. En este caso, se tomó cada uno de los grupos de genes y se compararon sus elementos. Si el 70% del grupo más pequeño estaba contenido en el grande y el 60% del grande estaba contenido en el pequeño, se consideró que esencialmente se trataba del mismo conjunto de genes. De este modo se redujeron a 1,042 los grupos de genes con los cuales se siguió el trabajo. Cada uno de estos grupos representa ya no un conjunto de genes ortólogos, si no un conjunto de genes aparentemente regulados por un mismo elemento.

4.8. Refinamiento de los posibles elementos de regulación

Una vez contando con una serie de grupos para los cuales suponemos existe un mecanismo de regulación en común (por tener al menos un motivo conservado en sus regiones río-arriba), es posible aplicar un proceso de refinamiento para definir estos

elementos de regulación de la mejor manera. El método que se siguió se parece bastante al empleado por PSI-BLAST [4], en donde se construye una matriz inicial con las secuencias más parecidas y con esta matriz se realizan ciclos de búsqueda seguido por una regeneración de la matriz tomando en cuenta los nuevos elementos identificados. De esta manera, utilizando PSI-BLAST se pueden recuperar homólogos más distantes que los que se pueden encontrar con BLAST. En este trabajo, los ciclos se realizaron construyendo matrices con MEME (en este caso pidiendo sólo un motivo y utilizando el modelo inicial de que el motivo estuviera presente una vez en todas las secuencias), seguido por una búsqueda con MAST de todas las iMUR de los genomas no-redundantes. Estos dos pasos se realizaron partiendo de los 1,042 grupos y los ciclos se continuaron para cada grupo hasta que el número de elementos encontrado convergiera. Al final de los ciclos, se realizó nuevamente un proceso de agrupamiento de los redundantes, obteniendo así 663 grupos de genes con motivos refinados.

4.9. Definición final de los motivos de regulación

Después de los ciclos de refinamiento, se realizó un último ciclo de MEME-MAST, para definir al elemento completo de regulación. Para esto, se permitió a MEME encontrar hasta 6 motivos para cada grupo, siempre y cuando fueran significativos y estuvieran presentes en todas las secuencias. Con el conjunto de motivos así definidos, se hizo una última búsqueda con MAST, pero esta vez utilizando tanto el conjunto de iMURs como una base de datos de regiones codificantes. Esta comparación permitió eliminar aquellos motivos que no fueran realmente intergénicos. A pesar de las precauciones que se tomaron para evitar incluir elementos génicos en el análisis, se encuentran aún por varias razones. Existen genes muy pequeños (y algunos muy conservados, como la proteína ribosomal L36 de ~35 aminoácidos), que durante el proceso automatizado de anotación de algunos genomas, no son asignados y por ende fueron considerados inicialmente como regiones intergénicas. También por procesos automatizados, los principios de algunos genes no son correctamente asignados, pudiendo quedar un primer fragmento anotado como intergénico. Inclusive, pudieran existir

pseudo-genes que, a pesar de ya no ser funcionales, mantengan suficiente conservación de secuencia como para ser detectados. Todos estos casos pueden ser descartados si se encuentran motivos que están presentes en la base de datos de regiones codificantes. Con esto se eliminan 51 casos, reduciendo el número de grupos de motivos de regulación a 612.

4.10. Procesamiento y anotación de los motivos de regulación

Una de las primeras cosas que es interesante saber de los motivos de regulación predichos, es su distribución filogenética. Se contó para cada conjunto de motivos a cuántos operones diferentes estaba regulando, el número de genomas (no-redundantes), ordenes y finalmente phyla. Ordenando por estos valores, es posible saber cuales de los elementos de regulación encontrados son los más ampliamente distribuidos.

Para empezar a caracterizar funcionalmente a los elementos, se consideró las funciones de los genes que regularían de acuerdo a los COGs, así como a los caminos metabólicos, de acuerdo al KEGG (Enciclopedia de Genes y Genomas de Kyoto, por sus siglas en inglés) [40]. Esta relación se hizo mediante el cálculo de un valor de probabilidad o *p*-value, asumiendo una distribución hipergeométrica. Esta distribución es similar a la binomial pero para eventos no independientes. Sirve para calcular, para n eventos, la probabilidad de obtener p aciertos (y n-p fallos) dentro de un conjunto total de F elementos, de los cuales M son los elementos que se buscan, donde después de cada evento no se regresa el elemento tomado. En el caso de esta evaluación, se cuenta con un número total de F genes (los de todas las bacterias), de los cuales M están asignados a un COG o KEGG determinado. Un motivo regula a n genes, de los cuales p pertenecen al COG o KEGG en cuestión y se desea saber la probabilidad de que esto ocurra por azar. La manera de calcularlo es resolviendo lo siguiente:

$$\frac{\begin{array}{c}\text{combinaciones para obtener} \\ \text{los } p \text{ genes regulados de los} \\ M \text{ pertenecientes al COG}\end{array} \quad \text{x} \quad \begin{array}{c}\text{combinaciones para obtener} \\ \text{los } n\text{-}p \text{ genes regulados de} \\ \text{los } F\text{-}M \text{ fuera del COG}\end{array}}{\begin{array}{c}\text{combinaciones con las que se pueden obtener} \\ \text{grupos de } n \text{ genes de un total de } F\end{array}}$$

Matemáticamente se representa con la siguiente fórmula:

$$p-value = \frac{\binom{M}{p}\binom{F-M}{n-p}}{\binom{F}{n}}$$

Con estos *p*-values por un lado se define la posible función del elemento de regulación (dada la función del COG o KEGG con mejor *p*-value) y por otro lado se pueden ordenar todos estos elementos para considerar primero a los más significativos.

Existe también una base de datos de familias de RNA (tanto codificantes como regulatorios) llamado Rfam [28]. Esta incluye todos los riboswitches ya conocidos y algunos otros elementos de regulación basados en RNA. Se tomaron los modelos para cada elemento de Rfam para comparar con las predicciones de este trabajo y así anotar los elementos conocidos. Con los resultados iniciales, se encontraron todos excepto uno de los riboswitches conocidos, el elemento *ykoK*, cuya distribución es tan restringida que no fue significativa su presencia dentro de un COG.

Al asignar función de estas maneras, resultó evidente que aún existía cierta redundancia entre los resultados. Esto se daba por tener matrices diferentes, que eran más específicas a cierto grupo taxonómico, como Gram positivas o negativas y que a pesar de definir realmente a un mismo elemento de regulación, encontraban suficientes genes diferentes como para mantenerse como grupos independientes. Así, de todos los grupos de genes enriquecidos con un mismo COG, se tomó solamente el que más significativamente estuviera asociado con éste (el de menor *p*-value). Se terminó entonces

con una predicción de 358 elementos de regulación independientes. De éstos, 17 resultaron ser conocidos (de acuerdo al Rfam) y los 341 restantes fueron predicciones de elementos nuevos (ver Figura 5 para el diagrama del número de elementos encontrados en cada paso del método).

4.11. Predicción de estructura secundaria conservada

Se emplearon dos estrategias con el propósito de evaluar si cada elemento predicho presentaba evidencia de conservación a nivel de estructura secundaria. En la primera se tomaron lo tres motivos más significativos de cada grupo y se generó para ellos un alineamiento incluyendo hasta las 20 mejores secuencias del conjunto no-redundante. Sobre estas se corrió el programa RNAalifold que predice una estructura secundaria consenso para una serie de secuencias alineadas [36]. La segunda estrategia consistió en tomar todas las secuencias que presentaban los motivos de un grupo (incluyendo secuencias de genomas redundantes) y se corrió un BLAST entre todas estas, generando una serie de regiones alineadas. Se prosiguió a eliminar las regiones idénticas y se empleó el programa QRNA para detectar los casos que presentaban evidencia de variabilidad consistente con un modelo de estructura de RNA, vs. estructura codificante vs. mutaciones aleatorias [71].

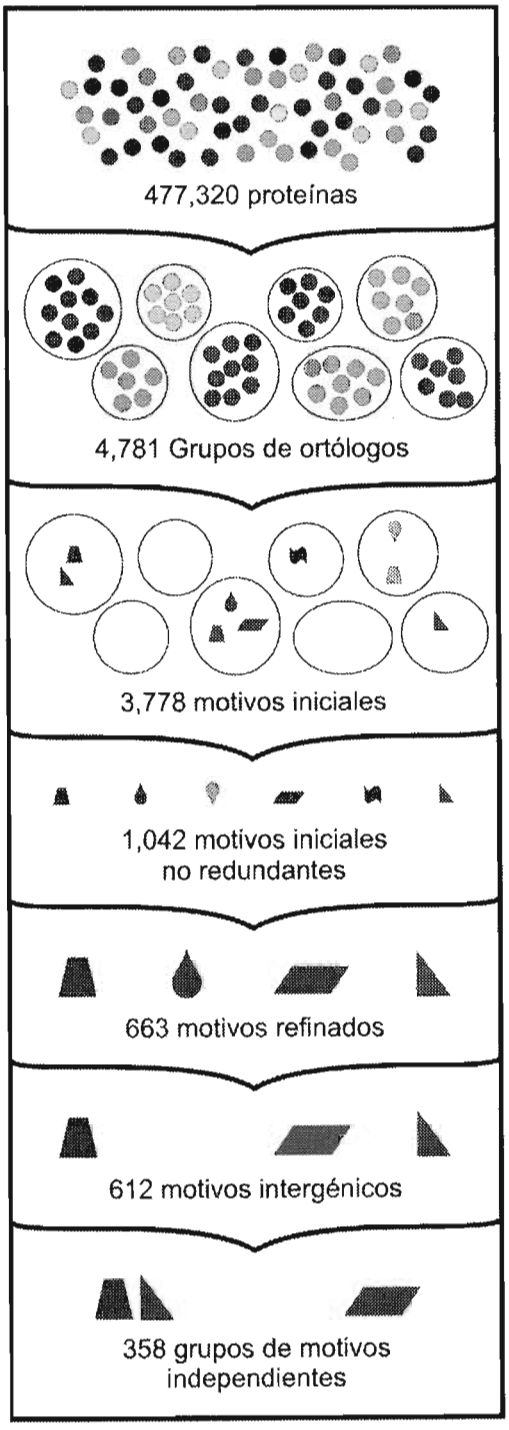

Figura 5. Esquema del número de elementos que se obtienen en cada paso del método. Ver texto para más detalles.

4.12. Variantes del método

La información que mayormente determina los motivos que se encuentran son las secuencias iniciales de un grupo de ortólogos. La razón por la que un riboswitch no pudo ser encontrado es que no está presente en suficientes secuencias de un mismo grupo. Esto ocurrirá con cualquier motivo que no se encuentre tan ampliamente distribuido. Sin embargo, existen variantes muy sencillas con las cuales se pueden recuperar más elementos de regulación. Dado que existirán aquellos que estén principalmente presentes en firmicutes o en proteobacterias, si se parte inicialmente de los genomas que pertenecen a sólo uno de estos grupos, la representatividad de los motivos será mucho mayor y se facilitará su detección. El elemento *ykoK* que no había sido recuperado inicialmente es encontrado dentro de las firmicutes de esta manera.

4.13. Servidores web

Para el análisis y la presentación de los resultados, se elaboraron algunos servidores web donde el usuario puede consultar los resultados del proyecto de una manera muy detallada así como buscar los motivos predichos en cualquier secuencia. Los servidores se programaron en Perl, utilizando el módulo de CGI para generar código de HTML al momento de ejecución. Se empleó la librería de gráficos GD para cuando se necesitaba producir imágenes y Javascript para las tareas más avanzadas, como intercomunicación de ventanas y algunos efectos gráficos. Para incluir predicciones de atenuadores transcripcionales y traduccionales, se empleó un algoritmo ya publicado desarrollado por Enrique Merino [58].

5. DISCUSIÓN DE RESULTADOS

Como producto final del trabajo de esta tesis se produjeron tres artículos que fueron aceptados en revistas internacionales de alto impacto. En los artículos se describe el método, algunos de los resultados obtenidos y los servidores web que se crearon para que la comunidad pudiera disponer fácilmente de la información generada. A continuación se presenta un recuento de cómo fueron producidos estos artículos.

5.1. Fase inicial del proyecto

Después de aplicar el método descrito en esta tesis, el resultado fue una serie de grupos de genes, cada grupo predicho como corregulado por uno o más motivos conservados. No se contó con la información de la base de datos Rfam desde el principio, por lo que inicialmente el proceso de evaluación fue un poco diferente. En primera instancia, se buscaron manualmente los riboswitches hasta entonces reportados, para ver si se habían encontrado. Esto resultó tedioso, ya que requería localizar el COG al que pertenecía cada gene reportado como regulado por un riboswitch, seguido de la enumeración de los grupos con buenos *p*-values para ese COG y posteriormente la comparación manual de los motivos con la secuencia reportada para buscar una coincidencia. No obstante, el proceso fue exitoso, ya que los siete riboswitches conocidos hasta ese momento habían sido encontrados por el método y los motivos coincidían perfectamente con la secuencia que formaba parte de las estructuras secundarias propuestas (ver manuscrito anexo para el ejemplo del riboswitch de TPP). En segundo lugar, para el análisis tanto de los riboswitches conocidos, como del resto de las predicciones, resultó sumamente útil analizar el contexto genómico en el que se encontraban. Es decir, en lugar de tomar en cuenta solamente al gene directamente regulado por cada elemento, ver también qué otros genes se encontraban en la vecindad, tanto río abajo como río arriba del gene en cuestión. Se desarrolló una herramienta web para este fin (*GeConT* [19]) y con el despliegue gráfico del contexto fue mucho más fácil descubrir las funciones en común de un grupo de genes.

Durante esta primera etapa del trabajo, se intentó anotar automáticamente los resultados de varias maneras. Se compararon tanto las secuencias intergénicas de los genes de cada grupo, como los motivos mismos, contra todas las proteínas y los RNA no traducidos de los genomas disponibles y después contra una base de datos generada con todos los sitios de regulación presentes en RegulonDB [76]. De este modo, grupos corregulados que presentaran, en sus secuencias anotadas como intergénicas, coincidencias con proteínas o RNA conocidos, fueron catalogados como posibles falsos positivos. Todos estos resultados fueron vaciados a una enorme tabla disponible por Internet (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) que incluía para cada grupo:

- Un identificador, de la forma GROUP####.
- Número de operones, especies y phyla que contienen los motivos.
- Los cinco KEGG y cinco COG con *p*-values más significativos.
- La lista desglosada de las phyla que presentan los motivos.
- Resultados de las comparaciones contra proteínas y RNAs.
- Resultados de las comparaciones contra sitios de RegulonDB.
- El consenso de cada motivo y el orden típico de estos.

En la tabla, cada renglón que corresponde a un grupo particular de motivos, se coloreó con un código para indicar si se había catalogado como un riboswitch conocido (manualmente), como una probable proteína, como un RNA o por exclusión se trataba de un elemento regulatorio putativo. Además de la información visible, se agregaron ligas para acceder a la lista de todos los genes de cada grupo, la estructura secundaria para el iMUR con la mejor calificación y las matrices de los motivos mismos. También se incluyó una liga directa para mandar a *GeConT* la lista de los genes corregulados. Todos estos datos pretendían ayudar al momento de examinar los resultados y permitir que se analizaran con mayor profundidad.

Fue durante esta fase del trabajo que se publicaron los dos primeros artículos. El primero presenta *GeConT*, el servidor web que se desarrolló inicialmente para analizar la congruencia por contexto genómico de los grupos de genes corregulados que el método

arrojó, pero que es de una gran utilidad para analizar la conservación de la sintenia de genes ortólogos y en general para visualizar el contexto genómico de cualquier grupo de genes. El segundo discute el método desarrollado así como algunos de los resultados más importantes [2], inclusive propuso un nuevo candidato de riboswitch, relacionado al metabolismo de glicina, cuya existencia fue paralelamente comprobada experimentalmente [10]. A continuación se anexan los dos manuscritos.

BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 14 2004, pages 2307–2308
doi:10.1093/bioinformatics/bth216

GeConT: gene context analysis

R. Ciria, C. Abreu-Goodger, E. Morett* and E. Merino

Departamento de Biología Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México

Received on January 16, 2004; revised on February 11, 2004; accepted on February 12, 2004
Advance Access publication April 8, 2004

ABSTRACT

Summary: The fact that adjacent genes in bacteria are often functionally related is widely known. *GeConT* (Gene Context Tool) is a web interface designed to visualize genome context of a gene or a group of genes and their orthologs in all the completely sequenced genomes. The graphical information of *GeConT* can be used to analyze genome annotation, functional ortholog identification or to verify the genomic context congruence of any set of genes that share a common property.
Availability: http://www.ibt.unam.mx/biocomputo/gecont.html
Contact: emorett@ibt.unam.mx

Graphical representations have been widely used to complement theoretical and experimental analyses of many biological processes, such as networks of gene regulation and protein–protein interactions, metabolic pathways or whole genome comparisons, to mention a few. Here we present a web application, named GeConT, for visualizing the genome context of genes. Some of its applications are described.

Genome annotation and identification of functional orthologs. Significant progress has been made in the assignment of protein function based on sequence similarity using programs such as BLAST (Altschul *et al.*, 1997), or FASTA (Pearson, 1990). Nevertheless, in many cases the identification of functional orthologs is not straightforward owing to protein sequence divergence. Also, the presence of paralogs with high similarity scores can hinder the identification of strict orthologs. To overcome these problems, different algorithms have been developed based on the analysis of gene neighborhood (Overbeek *et al.*, 1999; Wolf *et al.*, 2001; Kolesov *et al.*, 2002). These algorithms are based on the fact that adjacent genes in bacteria, organized in operons, are often functionally related. Furthermore, in some cases, conserved occurrence of neighboring genes, phylogenetic profiling and gene fusions can be observed and identified, as in the case of the web servers STRING (von Mering *et al.*, 2003) and Predictome (Mellor *et al.*, 2002). GeConT is a web application that can be used to explore the general neighborhood of any gene or set of genes given by the user, and their orthologs, as defined in the COG database (Tatusov *et al.*, 2003). For genes of the newly sequenced genomes that are not yet in the COG database, GeConT assigns their most likely orthologous group based on the COG number of their closest homologs. In contrast to the other available applications, GeConT displays the genomic context of the query genes regardless of its conservation. This is critical in cases where a gene is associated with different genes of a common process. An example of GeConT in the correct re-annotation of some genes involved in thiamine biosynthesis is given in the figure. We have previously shown that some *thi* genes, as *thiF*, *thiE* and *thiO*, have been mis-annotated on several occasions since they have paralogs with high sequence similarity (Morett *et al.*, 2003a). As shown in the Figure 1, this problem can be overcome by gene context analysis.

Genome context of any sequence. GeConT can also perform BLAST searches with an input protein sequence against all publicly available completely sequenced genomes. With this option the search is not biased by any previous annotation. Consequently, homologs with no functional annotation can be analyzed based on their context.

Verifying the genome context congruence of a set of genes. As a result of a sequence analysis, a set of genes that possess a common feature might be identified, such as being regulated by a certain protein or RNA structure (Abreu-Goodger, C., Ontiveros-Palacios, N., Ciria, R. and Merino, E. manuscript in preparation). In these cases, the neighborhood for these genes can help to discern true from false positives. This approach has been successfully used in the identification of new genes regulated by a *THI* box (Morett *et al.*, 2003b).

THE PROGRAM

GeConT was written in Perl, generating HTML code 'on the fly' and it is available at http://www.ibt.unam.mx/biocomputo/gecont.html

The classification scheme that we followed to search and color the genes is based on the clusters of orthologous groups (COGs) of proteins (Tatusov *et al.*, 2003). GeConT can highlight genes based either on their COG number or on a broader functional classification. Searches can be performed using any of the following criteria.

*To whom correspondence should be addressed.

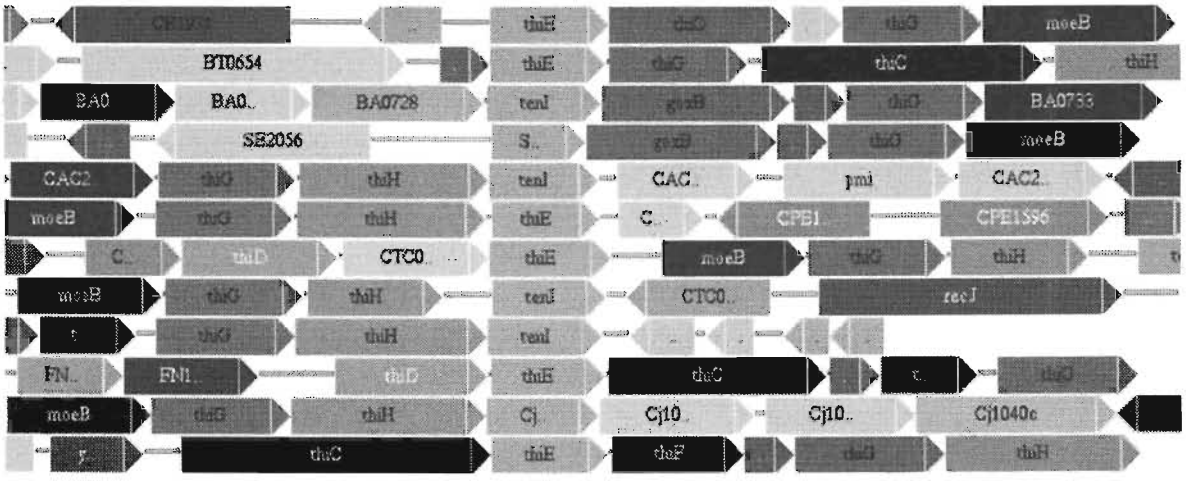

Fig. 1. Use of GeConT in genome analysis and correct gene annotation. The figure was generated using COG0352 (*thiE*) as input (only a fragment of the output is shown). Genes and intergenic distances are drawn to proportion. It is clear that the genes annotated as *moeB*, a paralog of *thiF* with high sequence similarity, *tenI*, a paralog of *thiE*, and *goxB*, a paralog of *thiO*, have been mis-annotated in several genomes. Gene context analysis is a very useful tool for correct gene annotation.

Searches based on COG. There are two different ways to select a COG: (i) by COG number or (ii) by a keyword that matches the COG functional annotation. With either of these options, the genomic context for all the members of the selected COGs is displayed.

Searches by GeneBank identification (GI) numbers or gene names. When one or more GI numbers or common names are given, GeConT displays the gene context for every entry regardless of its COG assignment.

Searches with a protein sequence. In this case, a protein sequence is used to perform a BLAST search against the database of fully sequenced genomes. The user can then select the matches to be displayed.

Other options include determining the number of genes to be displayed on either side of the input, a full description of each gene with a 'mouse-over' function and clicking on any gene to display, in a new page, the neighborhood of its corresponding COG.

ACKNOWLEDGEMENTS

The authors wish to thank Shirley Ainsworth for bibliographical assistance and critical reading of the manuscript, as well as Abel Linares, Arturo Ocadiz, Juan Manuel Hurtado and Alma Martinez for computer support.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Kolesov,G., Mewes,H.W. and Frishman,D. (2002) SNAPper: gene order predicts gene function. *Bioinformatics*, **18**, 1017–1019.

Mellor,J.C., Yanai,I., Clodfelter,K.H., Mintseris,J. and DeLisi,C. (2002) Predictome: a database of putative functional links between proteins. *Nucleic Acids Res.*, **30**, 306–309.

Morett,E., Korbel,J.O., Rajan,E., Saab-Rincon,G., Olvera,L., Olvera,M., Schmidt,S., Snel,B. and Bork,P. (2003a) Systematic discovery of analogous enzymes in thiamin biosynthesis. *Nat. Biotechnol.*, **21**, 790–795.

Morett,E., Saab-Rincon,G., Merino,E. and Bork,P. (2003b) High rate of gene displacement in vitamin biosynthesis pathways. In Andrade,M.A. (ed.), *Bioinformatics and Genomes*. Horizon Scientific Press, Norfolk, pp. 69–79.

Overbeek,R., Fonstein,M., D'Souza,M., Pusch,G.D. and Maltsev,N. (1999) The use of gene clusters to infer functional coupling. *Proc. Natl Acad. Sci., USA*, **96**, 2896–2901.

Pearson,W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Meth. Enzymol.*, **183**, 63–98.

Tatusov,R.L., Fedorova,N.D., Jackson,J.D., Jacobs,A.R., Kiryutin,B., Koonin,E.V., Krylov,D.M., Mazumder,R., Mekhedov,S.L., Nikolskaya,A.N. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.

von Mering,C., Huynen,M., Jaeggi,D., Schmidt,S., Bork,P. and Snel,B. (2003) STRING: a database of predicted functional associations between proteins. *Nucleic Acids. Res.*, **31**, 258–261.

Wolf,Y.I., Rogozin,I.B., Kondrashov,A.S. and Koonin,E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res.*, **11**, 356–372.

2 Kan, Z. *et al.* (2002) Selecting for functional alternative splices in ESTs. *Genome Res.* 12, 1837–1845
3 Modrek, B. and Lee, C.J. (2003) Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. *Nat. Genet.* 34, 177–180
4 Nurtdinov, R.N. *et al.* (2003) Low conservation of alternative splicing patterns in the human and mouse genomes. *Hum. Mol. Genet.* 12, 1313–1320
5 Li, W.H. *et al.* (1985) A new method for estimating synonymous and nonsynonymous rates of nucleotide substitution considering the relative likelihood of nucleotide and codon changes. *Mol. Biol. Evol.* 2, 150–174
6 Maquat, L.E. (2004) Nonsense-mediated mRNA decay: splicing, translation and mRNP dynamics. *Nat. Rev. Mol. Cell Biol.* 5, 89–99
7 Nagy, E. and Maquat, L.E. (1998) A rule for termination-codon position within intron-containing genes: when nonsense affects RNA abundance. *Trends Biochem. Sci.* 23, 198–199
8 Maquat, L.E. (2002) Nonsense-mediated mRNA decay. *Curr. Biol.* 12, R196–R197
9 Lewis, B.P. *et al.* (2003) Evidence for the widespread coupling of alternative splicing and nonsense-mediated mRNA decay in humans. *Proc. Natl. Acad. Sci. U. S. A.* 100, 189–192
10 Xing, Y. *et al.* (2004) The multiassembly problem: reconstructing multiple transcript isoforms from EST fragment mixtures. *Genome Res.* 14, 426–441
11 Heard, E. *et al.* (1997) X-chromosome inactivation in mammals. *Annu. Rev. Genet.* 31, 571–610
12 Modrek, B. *et al.* (2001) Genome-wide detection of alternative splicing in expressed sequences of human genes. *Nucleic Acids Res.* 29, 2850–2859
13 Boguski, M.S. *et al.* (1993) dbEST–database for expressed sequence tags. *Nat. Genet.* 4, 332–333
14 Sorek, R. *et al.* (2002) Alu-containing exons are alternatively spliced. *Genome Res.* 12, 1060–1067
15 Lynch, M. (2002) Intron evolution as a population-genetic process. *Proc. Natl. Acad. Sci. U. S. A.* 99, 6118–6123
16 Lynch, M. and Kewalramani, A. (2003) Messenger RNA surveillance and the evolutionary proliferation of introns. *Mol. Biol. Evol.* 20, 563–571

0168-9525/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.07.009

Conserved regulatory motifs in bacteria: riboswitches and beyond

Cei Abreu-Goodger, Nancy Ontiveros-Palacios, Ricardo Ciria and Enrique Merino

Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México

We present a computational approach that identifies regulatory elements conserved across phylogenetically distant organisms. Intergenic regulatory regions were clustered by orthology of the adjacent genes, and an iterative process was applied to search for significant motifs, enabling new elements of the putative regulon to be added in each cycle. With this approach, we identified highly conserved riboswitches and the Gram positive T-box. Interestingly, we identified many other regulatory systems that appear to depend on conserved RNA structures.

Comparative genomic approaches are central to analyzing the increasing number of whole-genome sequences. Although using this kind of analysis to find regulatory elements is not new, the focus has usually been on one genome or group of closely related genomes [1–3] because sequence conservation of functional intergenic regions (promoters, protein binding sites) is usually low, and quickly diverges.

It came as a surprise to many scientists when specific RNA 'riboswitches' were shown to be capable of regulating gene expression by directly sensing a metabolite without the intervention of a protein [4]. RNA riboswitches have since been shown to be involved in various metabolic processes including thiamine, riboflavin, cobalamine, adenine, guanine and lysine biosynthesis [5–11]. We assumed that this type of regulatory sequence would be easily identified given their broad phylogenetic distribution and highly conserved nature.

Searching for interesting motifs

The starting point for our work is a set of orthologous regulatory regions. To obtain these we used the Cluster of Orthologous Groups (COG) of proteins database (http://www.ncbi.nlm.nih.gov/COG/) [12] together with operon predictions based on intergenic distances [13]. In this manner, every protein from 164 fully sequenced bacterial genomes that was associated with a COG was assigned to the intergenic minimal upstream region (iMUR) of the first gene of the predicted operon to which it belongs. To avoid over-representation of similar sequences from related genomes, redundant sequences were eliminated. We obtained ~4000 clusters of orthologous regulatory regions, each belonging to a different COG.

We used the public domain motif discovery tool Multiple EM for Motif Elicitation (MEME) [14] to find a set of over-represented 'seed motifs' for each COG (Figure 1a). These motifs were used to identify other members of the putative regulon by searching in all upstream regions using the MEME counterpart Motif Alignment and Search Tool (MAST) [15]. As a result of this

Corresponding author: Enrique Merino (merino@ibt.unam.mx).
Available online 19 August 2004

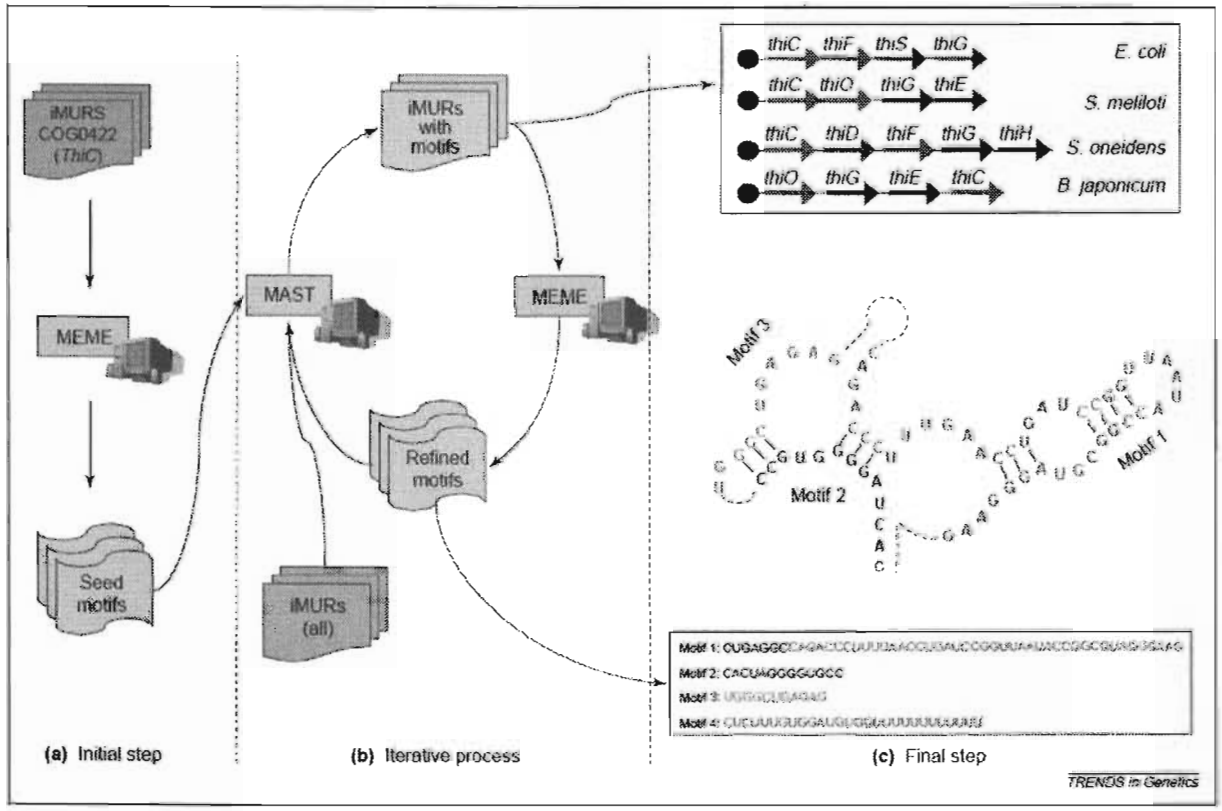

Figure 1. General procedure to identify conserved regulatory motifs. (a) The intergenic Minimal Upstream Regions (iMUR) of the operons that code for each COG (in this example *thiC*, COG0422) from fully sequenced genomes are grouped and over-represented motifs identified using MEME [14]. These motifs constitute the initial, or 'seed motifs' of the group. (b) These motifs are used to identify new members of the putative regulon in the entire set of iMURs using MAST [15]. To obtain 'refined motifs' that better represent the expanded group, we once again used MEME. This cycle is iteratively performed until no new elements are added. (c) At the final step of the process all the genes located by the 'refined motifs' are collected. In the example, the genes are found to belong to the thiamine biosynthesis metabolic pathway, and the conserved motifs correspond to important structural elements of the THI element. Abbreviations: MEME, Multiple EM for Motif Elicitation; MAST, Motif Alignment and Search Tool.

search, new members were added to each group (and some original ones were removed), and a new and more specific set of motifs was obtained, again using MEME. This cycle of locating over-represented motifs (with MEME) followed by searching for new genes containing the motifs (with MAST) was repeated until no new members were found (Figure 1b). The resulting 'refined motifs' are our candidate regulatory elements. Because many of these results are redundant, a clustering process was applied based on the number of genes in common for each motif, resulting in 672 different groups of motifs (for more details, see http://www.ibt.unam.mx/biocomputo/conserved_motifs.html). It is worth mentioning that each group can have several motifs (up to four), and that the motifs are larger than usual protein binding sites (the average length of our motif is 43 nucleotides). For example, the 'thiamine riboswitch' Group 0012 (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) was obtained from 20 different COGs that converge to form a common set of genes. Following analysis of the four different motifs that define this group, we found that the most significant one contains the *thi* box and the next two overlap with part of the structure of the THI-element; the fourth motif appears to contain part of the transcriptional terminator with its poly-U tract (Figure 1c).

Evaluating the motifs

To evaluate the likelihood that the 'refined motifs' in our groups represent biologically important regulatory elements, we first eliminated those that matched known proteins (from the non redundant GenBank database) or RNA genes (rRNA, tRNA, scRNA, snRNA and so on) from fully sequenced genomes. This step was necessary given the possible existence of unannotated small genes or erroneously assigned translation start sites resulting from automatic genome annotation. For example, ribosomal protein L36 (~40 amino acids long) is not annotated in >15 genomes but is picked up by our method, when analyzing the upstream region of ribosomal proteins S13 and S11 (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html, Group 0011 marked as a false positive). The groups of 'refined motifs' were then assigned to operons and their statistical significance was evaluated as follows. A *P*-value (assuming a hypergeometrical distribution) was calculated for each motif to be over-represented in a given Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway (http://www.genome.ad.jp/kegg/pathway.html) [16]. The lower the *P*-value, the more likely it is that the motifs can be considered to be biologically relevant for the regulation of that pathway. A similar evaluation was performed using COG assignations. We were able to

Table 1. Representative examples of conserved regulatory motifs[a]

Group and description[b]	Operons, organisms and phyla[c]	Representative KEGG pathways[d] (*P*-value)[f]	Representative COGs[e] (*P*-value)[f]	Refs
Group 0118; T-box	406; 37; 3 (firmicutes)	00970 Aa-tRNA biosynthesis (2×10^{-312}) and other amino-acid pathways	tRNA synthetases: COG0060 isoleucyl, COG0013 alanyl, COG0172 seryl, COG0441 threonyl, COG0016 phenylalanyl and others (all $<6\times10^{-29}$)	[18]
Group 0012; thiamine riboswitch	218; 106; 12	00730 thiamine metabolism (1×10^{-188})	COG0422 ThiC (8×10^{-136}), COG0351 ThiD (5×10^{-102}), COG0352 ThiE (2×10^{-80}) and COG4143 TbpA (7×10^{-67})	[4,5]
Group 0033; methionine riboswitch	166; 27; 7	00271 methionine metabolism (4×10^{-61})	COG1135 MetN ATPase (3×10^{-49}), COG2011 MetI permease (2×10^{-46}), COG0192 MetK SAM-synthetase (1×10^{-37}) and COG1464 MetQ periplasm (2×10^{-31})	[8]
Group 0013; cobalamin riboswitch	189; 77; 12	00860 porphyrin and chlorophyll metabolism (2×10^{-82})	COG4206 cobalamin receptor (1×10^{-68}), COG2087 CobP (1×10^{-33}), COG1120 cobalamin transposase ATPase (6×10^{-32}), COG0620 MetE cob-independent (3×10^{-31}) and COG1492 cobyric acid synthase (6×10^{-30})	[6]
Group 0219; IS911	68; 18; 2		COG2963 transposases (5×10^{-111}) and COG2801 transposases (1×10^{-63})	[24]
Group 0034; riboflavin riboswitch	100; 79; 6	00740 riboflavin metabolism (2×10^{-124})	COG0108 RibB (7×10^{-109}), COG3601 membrane (3×10^{-92}) COG0807 GTP cyclohydrolase II (1×10^{-68}), COG0307 RibE (6×10^{-57}) and COG0054 RibH (9×10^{-48})	[7]
Group 0027; glycine cleavage	86; 68; 7	00260 glycine metabolism (1×10^{-46})	COG0404 glycine cleavage T (1×10^{-68}), COG1003 glycine cleavage P (8×10^{-62}) and COG0403 glycine cleavage P (1×10^{-61})	
Group 0045; PyrR site	92; 41; 6 (firmicutes and/or actinobacteria)	00240 pyrimidine metabolism (2×10^{-70})	COG2065 pyrimidine attenuation (8×10^{-99}), COG0540 aspartate carbamoyl (2×10^{-55}), COG0044 cyclic amidohydrolase (1×10^{-42}) and COG2233 xanthine or uracil permease (2×10^{-41})	[21]
Group 0017; heat shock CIRCE hairpin	194; 84; 11		COG0234 GroES HSP10 (3×10^{-155}), COG1420 regulator of HS (3×10^{-86}), COG0459 GroEL HSP60 (2×10^{-69}) and COG0576 GrpE (6×10^{-49})	[22]
Group 0079; copper transport	99; 46; 4 (proteobacteria)		COG2217 cation transposases ATPase (3×10^{-98}), COG2132 multicopper oxidases (3×10^{-31}) and COG2608 copper chaperone (9×10^{-20})	
Group 0023; K^+-transport ATPase	65; 56; 8		COG2060 KdpA (8×10^{-169}), COG2156 KdpC (9×10^{-124}), COG2216 KdpB (1×10^{-120}) and COG2205 KdpD (7×10^{-49})	[25]
Group 0388; Thr-tRNA synthetase	26; 24; 1 (proteobacteria)	00260 threonyl metabolism (5×10^{-34}) 00970 Aa-tRNA biosynthesis (5×10^{-34})	COG0441 threonyl tRNA synthetase (2×10^{-71})	[19]
Group 0391; Glt, gln tRNA synthetases	24; 23; 1 (proteobacteria)	00251 glutamate metabolism (4×10^{-26}) 00970 Aa-tRNA biosynthesis (2×10^{-24})	COG0008 glutamyl- and glutaminyl-tRNA synthetases (5×10^{-59})	
Group 0204; ribosomal operon with self-regulation	30; 30; 2 (firmicutes)	03010 ribosome (1×10^{-52})	Ribosomal proteins: COG0051 S10, COG0089 L23, COG0088 L4, COG0087 L3 and COG0090 L2 (all $<6\times10^{-35}$)	[20]
Group 0262; ribosomal operon with self-regulation	9; 9; 2 (proteobacteria)	03010 ribosome (4×10^{-17})	Ribosomal proteins: COG0089 L23, COG0088 L4, COG0051 S10, COG0087 L3 and COG0185 S19 (all $<1\times10^{-15}$)	[20]

[a]A table containing the entire list of motifs can be found at http://www.ibt.unam.mx/biocomputo/conserved_motifs.html.
[b]Unique group related to the table on our web page, followed by the known or probable regulatory system.
[c]The number of operons, organisms and phyla in which the motifs are found. In case of a marked predominance of one or two phyla, it appears in parenthesis.
[d]Metabolic pathway (as defined by the Kyoto Encyclopedia of Genes and Genomes [KEGG] [16]) of the genes containing the motifs.
[e]Representative groups of orthologous genes (as defined by Cluster of Orthologous Groups [COG] [12]) in which our motifs are found.
[f]The statistical significance of over-representation of a given pathway or COG is expressed as a *P*-value (indicated in parenthesis), and was calculated assuming a hypergeometrical distribution of the signals.

assign a COG or KEGG with a *P*-value $<1\times10^{-6}$ in all the cases with the exception of only ten examples that are most probably false positives (motifs found by MEME that do not have any biological meaning). However, because many of the genes in these groups are not functionally annotated, they might still represent biologically relevant elements of unknown function. Finally, the genome context congruence of our putative regulons was verified using gene context analysis (GeConT) [17], a web tool that enables the user to view the neighbours for any set of genes together with their functions and orthologous relationships. The data on our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) is hyperlinked

to this application so that the co-regulated groups can be visualized easily.

Analyzing the nature of the conserved motifs

When sorting our results by *P*-values or by conservation, we realized that many of the first groups contain motifs that correspond to previously described riboswitches, known to regulate genes involved in the biosynthesis of different metabolites such as thiamine, riboflavin, cobalamine, adenine and guanine, in addition to the T-box regulator [18] of aminoacyl-tRNA synthetases from Gram positive bacteria (Table 1). Interestingly, we also found important sequence conservation in different families of aminoacyl-tRNA synthetases in Gram negative bacteria. For example, in *Escherichia coli* threonyl-tRNA synthetase, it is known that the mRNA leader region can adopt a tRNA-like structure that is specifically recognized by the corresponding threonyl-tRNA synthetase, establishing an auto-regulatory cycle [19]. The conserved motifs that we identified for this regulatory system correspond to parts of the stem-loop structure that resembles the threonine-tRNA anticodon CGU and to a stable structure that is similar to the acceptor arm of $tRNA^{Thr}$. Although it has not been reported previously, the motifs that we found for glutamyl and glutaminyl-tRNA synthetases could participate in a similar mechanism. We also found that many of our groups correspond to ribosomal protein operons – we detected 43 different groups of such operons, most of which correspond to specific phyla (Table 1). Self-regulation has been described for these cases because ribosomal protein L4 is known to bind to its operator, where a complex secondary structure appears to mimic the natural binding site of L4 in the ribosome [20]. We expect that most, if not all, of these operons to be auto-regulated by one of their proteins. Other well-known examples in Table 1 include a pyrimidine biosynthesis group, where our identified motifs correspond to the conserved RNA secondary structure that comprises the binding site for PyrR [21] and the 'Controlling Inverted Repeat for Chaperon Expression' (CIRCE), which constitutes the binding site for the HcrA repressor [22]. In all these examples (as occurs with riboswtiches), sequence conservation in the regulatory region is a consequence of the constraints imposed by the required RNA structure.

Another case that we found interesting was glycine cleavage. Although regulatory mechanisms have been described for the gcv operon in *E. coli* [23], we found a completely different regulatory system. The organisms that present our glycine cleavage motifs do not include *E. coli* but are mostly actinobacteria, firmicutes and α and β proteobacteria, and most of them do not encode orthologs of GcvA or GcvR, the two reported specific regulators of the gcv operon [23]. Furthermore, the reported binding sites for GcvA, do not match our motifs. Interestingly, the same glycine cleavage motifs are present in several proteins that are assigned to a Na+and alanine symporter COG. This could be part of a glycine transport system, and would make this putative regulon more similar to several riboswitches, where metabolic and transporting proteins are regulated by the same element. Many other examples of proposed regulatory systems with their conserved motifs can be found on our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html).

Concluding remarks

We have developed a computer method that detects previously reported riboswitches, other already known conserved elements and >600 statistically significant groups of conserved motifs that would appear to be biologically relevant. We have shown that for a great many regulatory elements their conservation is sufficiently strong to be detected in a single orthologous cluster of genes, without the need to consider additional elements of the regulon or metabolic pathway. In many cases, our motifs coincide with regulatory elements that are reported for specific model organisms such as *E. coli* or *Bacillus subtilis*. We are now able to propose the extent to which these systems have been conserved among fully sequenced bacteria. Several of the signals detected by our methods were related to an RNA secondary structure rather than classic DNA-binding regulators. Although we did not expect to find many conventional protein binding sites (the conserved portion being too small to be detected by our method), we were surprised by the number of structure-dependant regulatory elements that were also highly conserved at the sequence level.

Our study highlights potential new motifs to be further experimentally characterized in terms of their ability to form RNA secondary structures, such as attenuators, and their ability to bind small RNAs, cellular metabolites or regulatory proteins. All this will further help us to understand and define the regulatory mechanisms of these systems.

Acknowledgements

We wish to thank Shirley Ainsworth for bibliographical assistance, as well as Jérôme Verleyen and Roberto Bahena for computer support. We are indebted to Jose Antonio Loza for helping us with the *P*-value calculations. C.A.G was supported by fellowships from CONACyT and DGEP-UNAM. This work as partially supported by CONACyT grant 44213-Q.

References

1 Mironov, A.A. *et al.* (1999) Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes. *Nucleic Acids Res.* 27, 2981–2989
2 McGuire, A.M. *et al.* (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10, 744–757
3 Mwangi, M.M. and Siggia, E.D. (2003) Genome wide identification of regulatory motifs in Bacillus subtilis. *BMC Bioinformatics* 4, 18
4 Winkler, W. *et al.* (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature* 419, 952–956
5 Miranda-Rios, J. *et al.* (2001) A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 98, 9736–9741
6 Vitreschak, A.G. *et al.* (2003) Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA* 9, 1084–1097
7 Winkler, W.C. *et al.* (2002) An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. U. S. A.* 99, 15908–15913
8 Murphy, B.A. *et al.* (2003) Transcription termination control of the S box system: direct measurement of S-adenosylmethionine by the leader RNA. *Proc. Natl. Acad. Sci. U. S. A.* 100, 3083–3088

9 Grundy, F.J. *et al.* (2003) The L box regulon: Lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci. U. S. A.* 100, 12057–12062
10 Mandal, M. *et al.* (2003) Riboswitches control fundamental biochemical pathways in bacillus subtilis and other bacteria. *Cell* 113, 577–586
11 Vitreschack, A.G. *et al.* (2004) Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.* 20, 44–50
12 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
13 Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18, 329–336
14 Bailey, T.L. *et al.* (1994) *Fitting a mixture model by expectation maximization to discover motifs in biopolymers Proceedings of the 2nd International Conference on ISMB*, AAAI Presspp. 28–36
15 Bailey, T.L. *et al.* (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14, 48–54
16 Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.* 28, 27–30
17 Ciria, R. *et al.* GeConT: gene context analysis. *Bioinformatics* (in press)
18 Henkin, T.M. (2000) Transcription termination control in bacteria. *Curr. Opin. Microbiol.* 3, 149–153
19 Grunberg-Manago, M. (1996) Regulation of the expression of aminoacyl-tRNA synthetases and translation factors. In *Escherichia coli and Salmonella. Cellular and Molecular Biology* (Neidhardt, F.C. ed.), pp. 1432–1457, ASM Press
20 Stelzl, U. *et al.* (2003) RNA-structural mimicry in *Escherichia coli* ribosomal protein L4-dependent regulation of the S10 Operon. *J. Biol. Chem.* 278, 28237–28245
21 Bonner, E.R. *et al.* (2001) Molecular recognition of pyr mRNA by the *Bacillus subtilis* attenuation regulatory protein PyrR. *Nucleic Acids Res.* 29, 4851–4865
22 Wilson, A.C. and Tan, M. (2004) Stress response gene regulation in *Chlamydia* is dependent on HrcA-CIRCE interactions. *J Bacteriol.* 186, 3384–3391
23 Wilson, R.L. *et al.* (1995) Dna binding sites of the LysR-type regulator GcvA in the gcv and gcvA control regions of *Escherichia coli*. *J. Bacteriol.* 177, 4940–4946
24 Prére, M.F. *et al.* (1990) Transposition in *Shigella dysenteriae*: isolation and analysis of IS911, a new member of the IS3 Group of Insertion Sequences. *J. Bacteriol.* 172, 4090–4099
25 Asha, H. and Gowrishankar, J. (1993) Regulation of *kdp* operon expresión in *Escherichia coli*: evidence against turgor as signal for transcriptional control. *J. Bacteriol.* 175, 4528–4537

0168-9525/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.08.003

5.2. Evaluación los casos conocidos

Uno de los principales avances se logró al encontrar la base de datos Rfam. La información ahí contenida permitió anotar de una manera automática a todos los riboswitches ya conocidos que se fueron encontrando con el método. Además de anotar correctamente los distintos elementos regulatorios que se conocían de acuerdo al Rfam, también se realizó una evaluación formal al comparar las matrices generadas por MEME contra los modelos del Rfam. Estos modelos son curados manualmente por expertos y utilizan una combinación de estructura secundaria y elementos de secuencia primaria para determinar si una secuencia tiene o no un riboswitch. Computacionalmente, es mucho más tardado realizar una búsqueda empleando uno de estos modelos de covarianza que utilizar MAST. Sin embargo, al incluir la información de la estructura conservada, se espera que los modelos de Rfam sean mucho más sensibles y sean capaces de detectar elementos regulatorios que las matrices de MEME no podrían encontrar. La versión 7.0 del Rfam anota una gran parte de la base de datos EMBL versión 80 [41], incluyendo secuencias de genomas de bacterias, hongos, plantas, invertebrados, vertebrados y hasta de humano. De esta manera, Rfam contiene la posición de cada uno de los riboswitches presentes en todas estas secuencias. Para realizar una comparación válida de estos datos contra las matrices obtenidas con MEME fue necesario elaborar una base de datos *ad hoc* para realizar búsquedas con MAST, debido a que éste presenta una funcionalidad limitada cuando se usan secuencias muy largas. Para ello, se tomó el mismo conjunto de secuencias que empleó Rfam, pero se partieron en fragmentos de 500 nucleótidos, permitiendo un sobrelape de 150 nucleótidos. Con el conjunto de 68 x 10^6 secuencias resultantes y las matrices generadas por MEME coincidentes con un elemento regulatorio de Rfam, se realizó una búsqueda exhaustiva. Los resultados se presentan en la Tabla 1.

Nombre [a]	# en Rfam [b]	Recuperados [c]	%VP [d]	extra [e]	*p*-value [d]
T-box	475	467	**98 %**	358	10^{-07}
TPP	382	319	**83 %**	16	10^{-07}
Ado-CBL	249	225	**90 %**	34	10^{-07}
SAM	219	219	**100 %**	11	10^{-08}
Glicina	163	161	**99 %**	11	10^{-09}

FMN	136	136	**100 %**	4	10^{-08}
yybP-ykoY	127	14	11 %	0	10^{-07}
yybP-ykoY	127	31	24 %	6	10^{-07}
Purinas	100	81	**81 %**	2	10^{-07}
Lisina	98	55	56 %	1	10^{-07}
ydaO/yuaA	74	73	**99 %**	1	10^{-08}
glmS	37	30	**81 %**	7	10^{-07}
ykkC-yxkD	21	20	**95 %**	0	10^{-08}

Tabla 1. Comparación de las matrices obtenidas con MEME *vs.* Rfam. Las columnas contienen: el nombre del regulador o la molécula reconocida [a], el número de elementos independientes reportados en Rfam [b], la cantidad de estos recuperados [c] (mostrado también como % de verdaderos positivos [d]), los elementos encontrados adicionales no reportados en Rfam [c] y el *p*-value de corte utilizado en MAST [d].

Se puede observar que para 10 casos, la cobertura alcanzada es mayor al 80% (mostrado en negritas), por lo que se espera que la mayoría de los resultados de este trabajo sean tan buenos como si se hubiera definido un modelo de secuencia/estructura manualmente. Para todos los casos, los elementos adicionales encontrados en general son escasos, e insignificantes tomando en cuenta el tamaño de la base de datos. Estos pudieran tratarse de falsos positivos, sin embargo, al analizar el caso de la T-box (donde fue detectado un mayor número de adicionales), prácticamente todos los elementos detectados con MAST están en la región 5' de genes de tRNA sintetasas o relacionados de alguna manera con la biosíntesis o transporte de aminoácidos. Además, Rfam advierte que para el caso particular de la T-box, su modelo de la estructura secundaria no está completa, sugiriendo que la mayoría de los elementos aquí encontrados realmente sean verdaderos positivos. Debido a que el número de motivos definidos con MEME varía, el *p*-value óptimo de corte para MAST también variará, como puede verse en la Tabla 1. No obstante, se puede ver que usando el valor de corte de 10^{-07} en general se obtienen buenos resultados, por lo que puede ser empleado como un valor de corte universal.

En dos casos, la cobertura alcanzada no fue tan buena como se hubiera querido. Se recuperó el 56% de los riboswitches conocidos de lisina y para el elemento *yybP-ykoY* se encontró este partido en dos grupos (específicos para firmicutes y proteobacterias) logrando tan solo recuperar el 35%, aún tomando en cuenta ambos conjuntos de motivos. La estructura secundaria propuesta para ambos elementos es particularmente dependiente

de largos tallos, los cuales pueden ser representados muy bien usando modelos de covarianza, pero de no presentar suficiente conservación a nivel de secuencia, MEME no podrá definirlos correctamente. Esta es una limitante del método planteado en esta tesis.

5.3. Servidor web para encontrar elementos regulatorios

La exitosa comparación con Rfam, permitió tener mayor confianza sobre la especificidad de las matrices generadas con MEME, por lo que se decidió crear un servidor web donde los usuarios pudieran aprovechar todas las predicciones aquí realizadas. Así surgió *RibEx* (Explorador de Riboswitches) [1] el cual resume la información obtenida para cada elemento regulatorio encontrado, incluyendo el COG y KEGG más significativo, su distribución filogenética, ligas a *GeConT* para visualizar el contexto genómico y ligas a Rfam para aquellos elementos ya reportados. Pero la contribución más importante de *RibEx* es que permite hacer búsquedas sobre cualquier secuencia, representando de una manera gráfica los marcos abiertos de lectura, elementos regulatorios y atenuadores predichos ahí encontrados. La imagen resultante es interactiva, pudiéndose escoger una probable proteína o un motivo de regulación para ver su secuencia y enviarlo a BLAST o elegir un atenuador para ver su estructura secundaria. Además, para fines de comparación se incluye la opción de ver una imagen con la distribución de motivos y la calificación de todos los genes de los genomas completos no-redundantes en los cuales se encontró anteriormente el elemento regulatorio en cuestión.

Se publicó un artículo describiendo a *RibEx* en el número especial de servidores web de *Nucleic Acids Research* el cual se incluye a continuación.

W690–W692 *Nucleic Acids Research, 2005, Vol. 33, Web Server issue*
doi:10.1093/nar/gki445

RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements

Cei Abreu-Goodger and Enrique Merino*

Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, 62210 Morelos, México

Received February 14, 2005; Revised and Accepted March 30, 2005

ABSTRACT

We present RibEx (riboswitch explorer), a web server capable of searching any sequence for known riboswitches as well as other predicted, but highly conserved, bacterial regulatory elements. It allows the visual inspection of the identified motifs in relation to attenuators and open reading frames (ORFs). Any of the ORF's or regulatory elements' sequence can be obtained with a click and submitted to NCBI's BLAST. Alternatively, the genome context of all other genes regulated by the same element can be explored with our genome context tool (GeConT). RibEx is available at http://www.ibt.unam.mx/biocomputo/ribex.html.

INTRODUCTION

Ribonucleic acids have become fashionable lately. Apart from their fundamental participation in transcription and translation, RNAs are clearly some of the most functionally diverse molecules in the cell. Recently, non-translated regions of several mRNAs have been found to be capable of regulating their own expression by binding specific metabolites with high affinity in complete absence of proteins [(1), reviewed in (2)]. These regulatory elements, termed riboswitches, appear to be highly conserved, the extreme case being that of the thiamine pyrophosphate (TPP) riboswitch, which has been found in all three kingdoms of life (3). Riboswitches comprise two parts, a sensing element or aptamer, which forms a complex structure capable of binding the metabolite, and an effector element, or expression platform capable of transforming the signal into a biological response. The aptamer is the most conserved, having been selected to bind an unchanging molecule such as a vitamin or an amino acid. Upon binding, a shift between two mutually exclusive RNA secondary structures in the effector element occurs. These pairs of structures of the expression platform can represent a transcriptional terminator/anti-terminator, a Shine-Dalgarno sequester/anti-sequester or even an active/inactive ribozyme (2,4). It is not uncommon for different organisms to use the same sensing element, yet different effector elements.

FINDING RIBOSWITCHES

Although the usual method to define a riboswitch involves locating a conserved secondary structure in the RNA molecule, the highly restricted nature of the sensing element argues that sequence alone should be enough to locate riboswitches correctly. We have previously developed a computer algorithm capable of finding bacterial regulatory motifs, based exclusively on sequence conservation in the regulatory regions of orthologous groups of genes (5). The main restrictions of our method are that a regulatory element must be closely associated with at least one COG (cluster of orthologous groups of proteins) (6) and it must be present in at least five non-redundant genomes. On the other hand, the advantage is that it is an automatic process, requiring no previous regulatory information to produce relevant results, and as such, can be easily run every time that new genomes or annotations are available.

We updated our previous results (5), taking into account 223 complete genomes. From these, a reduced set of 145 non-redundant organisms was obtained using CVtree (7). We were able to recover 10 out of the 11 currently reported riboswitches. Additionally, our results included many regulatory elements that are also known to depend on structured RNA for recognition, such as the Gram-positive T-box and the PyrR protein binding site. We thus call our set of regulatory elements: riboswitch-like elements (RLEs), given the fact that almost all the identified conserved signals were RNA-dependant regulatory elements.

RibEx is a web server that allows any user to easily find any RLE in the sequence of his/her interest. Since most known riboswitches are associated with attenuators, we have included the option of searching for transcriptional and translational attenuators, which can help in selecting the most likely candidates, as has been shown by Barrick *et al.* (4). Additionally, our web server displays representative drawings of the open

*To whom correspondence should be addressed. Tel: +52 777 329 16 29; Fax: +52 777 317 23 88; Email: merino@ibt.unam.mx

© The Author 2005. Published by Oxford University Press. All rights reserved.

The online version of this article has been published under an open access model. Users are entitled to use, reproduce, disseminate, or display the open access version of this article for non-commercial purposes provided that: the original authorship is properly and fully attributed; the Journal and Oxford University Press are attributed as the original place of publication with the correct citation details given; if an article is subsequently reproduced or disseminated not in its entirety but only in part or as a derivative work this must be clearly indicated. For commercial re-use, please contact journals.permissions@oupjournals.org

reading frames (ORFs) and their corresponding regulatory elements, any of which can be selected, in order to acquire its sequence for submission to NCBI's BLAST server (8). Every RLE is linked to a list of genes that are predicted to be subject to its regulation. The genome context of these genes, analyzed with our local GeConT web server (9), in addition to the scores of the pre-computed RLEs, can be of great assistance when evaluating the likelihood of a new prediction.

A great resource when working with RNA families is the Rfam database (10). We have used their models to annotate our RLEs. As of version 7.0, Rfam contains a total of 503 families, 125 of them are non-coding, and 11 of these are annotated as riboswitches. We were able to recover automatically all but one of these riboswitches, missing the *ykoK* element. Our matrices for the most abundant riboswitches perform very well when compared with the co-variance models used by Rfam (~90% coverage when analyzing bacterial sequences). Less common riboswitches (e.g. lysine and purine) are more difficult to model with sequence-based weight-matrices. Our method thus tends to recover between 70 and 80% of these Rfam members. Our data set also contains six more RLEs that coincide with an Rfam *cis*-regulating member and 341 RLEs that do not have a match and thus remain as predicted elements. We have calculated a *P*-value, assuming a hyper-geometrical distribution, for each RLE to be over-represented in a given COG or KEGG pathway (11). Thus, we provide every RLE with a tentative functional assignation.

As far as we know there are only two servers, beside ours, that can be used to locate riboswitches in a given sequence: riboswitch finder (12) which, in its current implementation, only searches for the purine-sensing riboswitch, and Rfam, that has an option to locate riboswitches in any sequence, but as co-variance searches have high computational requirements, the sequence length is limited to 2 kb. RibEx, in addition to performing searches on larger sequences, allows the user a greater view of the regulatory potential of his sequence, by showing the ORFs and predicted attenuators. The 341 predicted RLEs also make RibEx a great complement to the curated families contained in Rfam.

THE WEB SERVER

The server is divided into modules, which are written in, and tied together with Perl. A brief description of each module follows:

Riboswitch-like elements. The program takes the sequence provided and splits it into overlapping windows of 500 nt. Each of these smaller sequences are searched for the selected RLEs with MAST (13), using matrices obtained as detailed in our previous work (5). Our method defines each RLE as several non-overlapping motifs, so we restrict the search to 500 nt to avoid false positives where the individual motifs are too far apart. When an RLE passes the selected *E*-value cutoff, the positions, size of each motif and final score of the regulatory element are recorded.

Open reading frames. ORFs are predicted, as is commonly done for bacterial genomes. The default options are for a resulting protein of at least 80 amino acids beginning with a start codon (ATG, GTG or TTG) and ending with a stop codon (TAA, TAG or TGA). By default, fully overlapped ORFs are not shown.

Attenuators. These are predicted according to an algorithm developed in our group and described elsewhere (14). The predicted secondary structure of each attenuator and its free energy is recorded. Upon clicking on the image of the attenuator, an additional window will be opened showing this information. To avoid false positives, attenuators are only searched for in the region preceding each predicted ORF.

Web output. The web page is generated 'on the fly' by a Perl script that controls all the other modules. The images are generated using the GD graphics library, and the interactivity between windows and frames is provided with Javascript.

AN EXAMPLE

Figure 1 shows a typical RibEx output. The input sequence was a region of 4000 nt from around the *thiC* gene of *Bacillus cereus ATCC14579*. Immediately upstream from one of the ORFs (drawn as blue arrows) the three motifs that comprise the TPP riboswitch (red boxes) can be seen, as well as

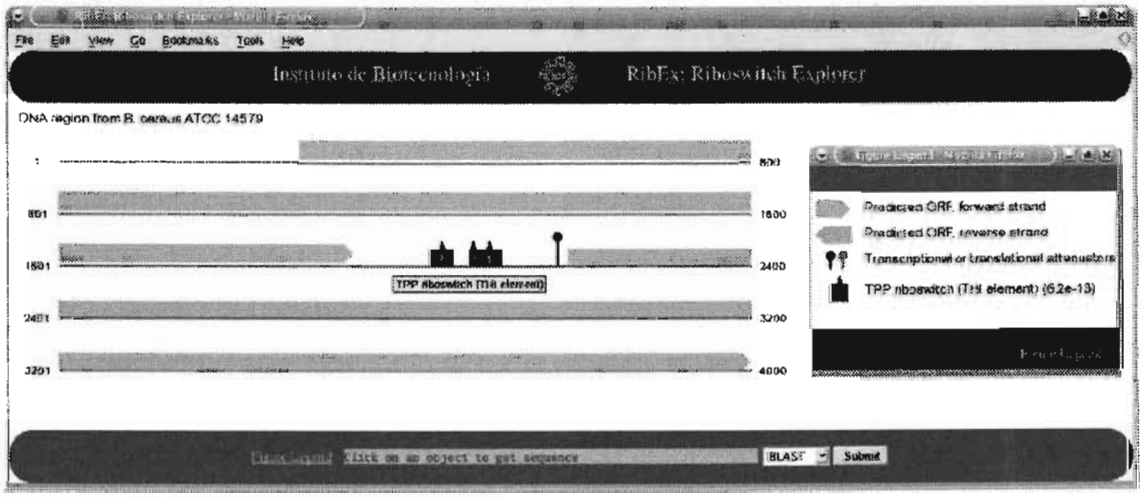

Figure 1. RibEx locates a thiamine riboswitch.

a transcriptional attenuator (black lollipop). A separate window acts as a figure legend indicating the score for each regulatory element found (in this case, only the TPP riboswitch). A typical scenario might include clicking on the second ORF, and sending the sequence to the BLAST web server, showing it to be identical to several ThiC proteins. Clicking on the TPP riboswitch motif in the figure legend box opens a window with the genes that are predicted to be regulated by this riboswitch, where the user can see how the motifs are distributed in different genomes. Taken together, and strengthened by the presence of a transcriptional attenuator, the user would have no trouble at all concluding that his sequence contains a bona fide riboswitch.

ACKNOWLEDGEMENTS

We wish to thank Ricardo Ciria for support in setting up the web server. This work was supported by CONACyT grant 44213-Q to E.M. and C.A.G. was supported by fellowships from CONACyT and DGEP-UNAM. The Open Access publication charges for this article were waived by Oxford University Press.

Conflict of interest statement. None declared.

REFERENCES

1. Winkler,W., Nahvi,A. and Breaker,R.R. (2002) Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature*, **419**, 952–956.
2. Nudler,E. and Mironov,A.S. (2004) The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.*, **29**, 11–17.
3. Sudarsan,N., Barrick,J.E. and Breaker,R.R. (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA*, **9**, 644–647.
4. Barrick,J.E., Corbino,K.A., Winkler,W.C., Nahvi,A., Mandal,M., Collins,J., Lee,M., Roth,A., Sudarsan,N., Jona,I. *et al.* (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl Acad. Sci. USA*, **101**, 6421–6426.
5. Abreu-Goodger,C., Ontiveros-Palacios,N., Ciria,R. and Merino,E. (2004) Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet.*, **20**, 475–479.
6. Tatusov,R.L., Koonin,E.V. and Lipman,D.J. (1997) A genomic perspective on protein families. *Science*, **24**, 631–637.
7. Qi,J., Wang,B. and Hao,B.L. (2004) Whole genome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.*, **58**, 1–11.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Ciria,R., Abreu-Goodger,C., Morett,E. and Merino,E. (2004) GeConT: gene context analysis. *Bioinformatics*, **20**, 2307–2308.
10. Griffiths-Jones,S., Bateman,A., Marshall,M., Khanna,A. and Eddy,S.R. (2003) Rfam: an RNA family database. *Nucleic Acids Res.*, **31**, 439–441.
11. Kanehisa,M. and Goto,S. (2000) KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, **28**, 27–30.
12. Bengert,P. and Dandekar,T. (2004) Riboswitch finder—a tool for identification of riboswitch RNAs. *Nucleic Acids Res.*, **32**, W154–W159.
13. Bailey,T.L. and Gribskov,M. (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics*, **14**, 48–54.
14. Merino,E. and Yanofsky,C. (2005) Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.*, **21**, 249–305.

5.4. Descripción de algunos resultados

Todos los resultados, en forma de tabla, están disponibles desde la página principal de *RibEx* (http://www.ibt.unam.mx/biocomputo/ribex.html). En esta sección se describirán algunos casos interesantes que se recuperaron, algunos previamente reportados, otros candidatos a ser nuevos elementos de regulación. Todos los identificadores del tipo RLE#### se refieren a los utilizados en el servidor web.

5.4.1. Recuperación de elementos regulatorios conocidos

En la Tabla 2 se incluye información sobre los distintos elementos detectados por el método, pero que ya se conocían. La mayoría de éstos fueron anotados mediante la comparación automatizada con Rfam, el resto fueron encontrados manualmente al analizar los resultados. Para la revisión y comparación manual fueron de suma utilidad los siguientes recursos disponibles por Internet: RegulonDB [76], BioCyc [44] y DBTBS (Base de datos de regulación transcripcional en *Bacillus subtilis*, por sus siglas en inglés) [51].

ID [a]	Nombre [b]	COG [c]	Mot. [d]	Estruct. [e]	Phyla [f]	Ref. [g]
RLE0001	Ado-CBL	COG0609 ABC-type Fe3+-siderophore transport system (27/220) $1e^{-40}$	4	★★	ACFP	[62, 74]
RLE0002	TPP	COG0422 Thiamine biosynthesis protein ThiC (49/89) $3e^{-112}$	3	★★	ACFPr	[60, 73, 80, 94]
RLE0003	*Sitios HrcA (CIRCE)*	*COG0234 Co-chaperonin GroES (HSP10) (64/128) $3e^{-150}$*	*1*	★★	*ACFP*	[68]
RLE0005	SAM	COG1135 ABC-type metal ion transport system, ATPase component (17/85) $3e^{-37}$	5	★★	AcFp	[30, 72, 97]
RLE0008	Sitios PyrR	COG2065 Pyrimidine operon attenuation protein (23/54) $1e^{-61}$	2	★★	AF	[48]
RLE0011	FMN	COG0108 3,4-dihydroxy-2-butanone 4-phosphate synthase (31/123) $6e^{-71}$	4	★★	AFP	[26, 88, 95]

ID [a]	Nombre [b]	COG [c]	Mot. [d]	Estruct. [e]	Phyla [f]	Ref. [g]
RLE0012	Glicina	COG0404 Glycine cleavage system T protein (22/123) $6e^{-50}$	6	✭★	AFP	[55]
RLE0013	*Sitios Fur-Fe^{+}*	*COG1629 Outer membrane receptor proteins, mostly Fe transport (20/518) $6e^{-24}$*	*1*	★	*ACFP*	[7, 86]
RLE0015	*Sitios DnaA*	*COG0592 DNA polymerase sliding clamp subunit (PCNA homolog) (33/139) $4e^{-92}$*	*3*	★	*AF*	[66]
RLE0019	glmS	COG0449 Glucosamine 6-phosphate synthetase (16/129) $6e^{-50}$	4	★★	F	[96]
RLE0022	T-box	COG0060 Isoleucyl-tRNA synthetase (24/135) $7e^{-38}$	4	★★	AF	[29, 31]
RLE0044	ydaO/yuaA	COG3773 Cell wall hydrolyses involved in spore germination (4/41) $9e^{-11}$	5	★★	ACFp	[10]
RLE0045	ykkC-yxkD	COG3665 Uncharacterized conserved protein (7/16) $4e^{-23}$	4	✭★	cFP	[10]
RLE0072	*IS200*	*COG1943 Transposase and inactivated derivatives (103/231) $9e^{-318}$*	*4*	★★	*FP*	[15]
RLE0109	Líder de treonina	COG0083 Homoserine kinase (10/72) $3e^{-27}$	4	★	P	[43]
RLE0112	Lisina	COG1757 Na+/H+ antiporter (10/88) $4e^{-28}$	4	★★	PF	[33, 75, 81]
RLE0118	Purinas	COG2252 Permeases (8/113) $6e^{-18}$	4	★★	pF	[52, 53]
RLE0129	*Sitios LexA*	*COG1974 SOS-response transcriptional repressors (11/104) $7e^{-33}$*	*1*	★✭	*P*	[17, 47]
RLE0139 RLE0223	yybP-ykoY	COG2119 Predicted membrane protein (12/39) $5e^{-42}$	4	★★	AcFP	[10]
RLE0253	Sitio S15	COG0184 Ribosomal protein S15P/S13E (10/128) $1e^{-33}$	1	★	P	[13]
RLE0277	Sitio S4	COG0100 Ribosomal protein S11 (9/133) $8e^{-22}$	4	★★	P	[77]
RLE0343	Represor Heat-Shock (ROSE)	COG0071 Molecular chaperone (8/146) $1e^{-26}$	3	✭★	P	[64]

Tabla 2. Elementos regulatorios conocidos que se recuperaron. Las columnas contienen: identificador tomado de *RibEx* [a], el nombre del regulador o la molécula reconocida (en caso de ser riboswitch) [b], COG

más significativo [c], número de motivos encontrados con MEME [d], si presenta estructuración de RNA predicho por dos métodos (★ sólo por un método, ★★ por ambos métodos, ✯ implica que una prueba fue poco convincente) [e], principales phyla que lo contienen (A = Actinobacteria, C = Cyanobacteria, F = Firmicutes, P = Proteobacteria, R = Archaea) donde una letra minúscula implica que sólo aislados organismos de ese phylum lo contienen [f] y referencias bibliográfica [g]. Reguladores que actúan a nivel de DNA se encuentran en itálicas.

5.4.1.1. Reguladores que actúan a nivel de RNA

En general se puede observar en la Tabla 2 que los riboswitches y otros reguladores que actúan a nivel del RNA presentan varios motivos. Esto es reflejo de su gran tamaño, y de que existe conservación de secuencia a lo largo del elemento. Puesto que el tamaño máximo de motivo permitido en la búsqueda es de 30 nucleótidos se espera que un riboswitch, que por lo general cubre más de 100 nucleótidos, sea representado por varios motivos. No todos los elementos regulatorios de RNA son tan grandes, sobre todo aquellas regiones a las que se les une otro RNA, ya que una corta región complementaria es suficiente para garantizar afinidad y especificidad. Además, pudieran existir otros elementos que a pesar de tener un gran tamaño (debido por ejemplo a que conforman un atenuador), tuvieran tan sólo una corta región conservada encargada de unir directamente a su ligando.

Una amplia distribución filogenética, al contrario de lo que se pensó al empezar el trabajo, no es garantía de un elemento tipo riboswitch, ya que se conocen al menos dos casos (los sitios de HrcA y de Fur) que se encuentran en prácticamente todas las phyla analizadas (ver Tabla 2). Parecería que los elementos más conservados ya han sido todos descritos. Durante el trabajo se encontró un elemento con amplia distribución filogenética asociada a genes de transporte y degradación de glicina, el cual se propuso como un nuevo riboswitch [2] y fue comprobado paralelamente [10]. Sin embargo, ningún otro elemento tan promisorio ha sido encontrado, y parecería que los riboswitches aún no encontrados estarán asociados a grupos filogenéticamente más cercanos o simplemente no presentan una buena asociación con algún COG haciéndolos indetectables por nuestro método. Dado que la idea de un riboswitch antiguo, presente en casi todas las bacterias, pero que en cada una regule genes completamente diferentes, no parece coherente; es más

factible que los riboswitches que se irán descubriendo de ahora en adelante serán de clados más específicos, lo cual hablaría posiblemente de un origen evolutivo más reciente. En todo caso, la distribución filogenética no puede ser utilizada como único criterio para elegir posibles riboswitches. En la Figura 6 se muestra la distribución filogenética de los riboswitches más importantes, los sitios HrcA y la T-box, sobre el árbol genómico de los 145 organismos no-redundantes generado utilizando CVTree [69]. A pesar de que este árbol no es necesariamente indicativo de la evolución de las especies (por ejemplo, las mycoplasma, chlamydiae y spirochaetes se separan del resto por ser parásitos obligados con genomas sumamente reducidos), se pueden observar ciertos patrones indicativos de cómo han sido heredados y perdidos los distintos elementos regulatorios. El riboswitch que une TPP parecería haber estado presente en el ancestro común del reino Bacteria, aunque se haya perdido independientemente en algunos organismos. Los elementos de este tipo presentes en archaea (y en eukaryota, aunque no se muestra) se pueden explicar por transferencias horizontales. En contraste, el riboswitch de purinas parecería haber tenido un origen mucho más reciente, previo a la separación de las firmicutes. Fuera de ese phylum, sólo existe un riboswitch de esta clase presente en el género *Vibrio*, asociado a una adenosin-deaminasa. Parecería claro que esto fue debido a una transferencia horizontal y observando el contexto con *GeConT* se puede suponer que el organismo donador fue uno cercano a *Clostridium perfringens*.

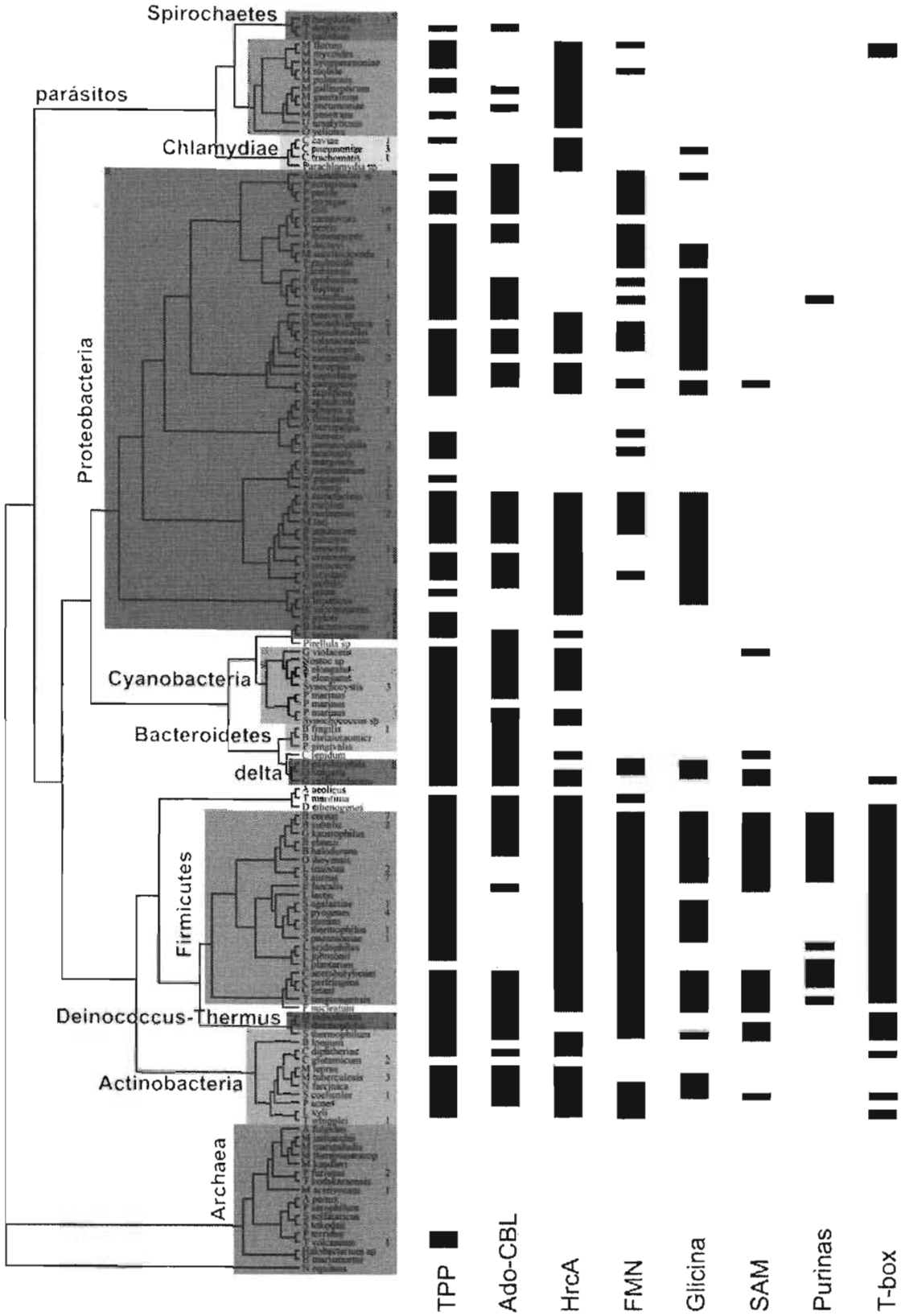

Figura 6. Distribución de algunos elementos regulatorios conservados. Árbol genómico de los 145 organismos no-redundantes, donde se puede observar la presencia o ausencia de cada regulador.

Las medidas de estructura que se emplearon son capaces de detectar la señal de los riboswitches, a pesar de que no todos los elementos son detectados por ambos métodos. No obstante, el criterio de estructurabilidad no puede ser utilizado aisladamente ya que existen casos, sobre todo los sitios palindrómicos como el de HrcA que son estructuras perfectas. Secuencias de inserción, como IS200 presentan regiones terminales conservadas tipo tallo-asa, pero pueden ser eliminados con relativa facilidad por estar siempre asociados a genes de transposasas o presentar excesivo número de copias por genoma. En general se ve que la estructurabilidad junto con la existencia de varios motivos conservados correlaciona bien con los elementos tipo riboswitch.

Vale la pena mencionar que en la Tabla 2 se incluyen ejemplos de un atenuador que utiliza un péptido líder y el ribosoma para actuar (RLE0109), sitios de unión de proteínas ribosomales al líder de su propio transcrito (RLE0253 y RLE0277) y un termo-sensor (RLE0343). Todos estos elementos actúan a nivel del RNA y se espera que el método pueda recobrar más elementos de cada una de estas categorías.

5.4.1.2. Sitios de interacción DNA-proteína

Como se ha mencionado antes, el protocolo desarrollado no pretendía encontrar sitios de pegado de proteínas al DNA, los cuales por su corto tamaño y poca conservación son mucho más difíciles de localizar que elementos tipo riboswitch. Después de una cuidadosa revisión de los resultados, fue aparente que algunos motivos sí se trataban de regiones de unión de proteína-DNA. Aún así, el sitio individual de pegado no parece contener suficiente información para ser detectado, más bien fueron localizados casos donde existen dos o más sitios contiguos, generando así una región conservada mucho mayor.

El resultado RLE0003 fue uno de los casos más significativos en todas las medidas empleadas, ya que el motivo conservado se encuentra prácticamente en todas las bacterias (con la excepción de la mayoría de las gamma-proteobacterias). Se trata de un motivo palindrómico que coincide perfectamente con CIRCE, una estructura invertida repetida que alguna vez se creyó funcionaba como un termo-sensor, pero ahora se sabe

contiene dos sitios invertidos de unión de HrcA, una proteína represora de genes de choque térmico [93]. Esta es una de las proteínas más conservadas en el mundo bacteriano y por lo tanto su sitio de unión en el DNA se encuentra bastante conservado. El hecho de que siempre sean dos sitios a una misma distancia ocasiona que sea muy fácil de encontrar con métodos como el descrito en este trabajo.

RLE0013 es un caso muy similar al de HrcA, por constar de dos sitios de pegado de la proteína Fur formando una estructura invertida repetida o palíndrome. Fur (Regulador de la Asimilación Férrica, por sus siglas en inglés) es un represor de genes de asimilación de hierro [7]. Sistemas que alteran la concentración de iones Fe^{+} deben estar altamente regulados, ya que por un lado este metal es vital para los microorganismos, pero por otro lado concentraciones elevadas son mortales [21]. Además de apagar genes que aumentan la concentración de Fe^{+}, Fur prende indirectamente a proteínas que almacenan hierro. Esto lo logra reprimiendo un RNA pequeño llamado RyhB, que normalmente mantiene apagados a los genes que codifican para estas proteínas [56]. Dada la vital importancia de mantener un control preciso sobre la concentración de hierro, no resulta sorprendente que dos sitios contiguos de Fur representen una de las señales más conservadas y extendidas en el mundo bacteriano.

Otro sitio de unión de proteína-DNA particularmente conservado es el de LexA, represor en la respuesta SOS [17]. En algunas gamma-proteobacterias el gene *lexA* se encuentra autorregulado por dos cajas contiguas para el pegado de LexA. Esta ampliada región fue la que se localizó como el motivo de RLE0129.

Curiosamente, RLE0015 que parecía interesante por tener 3 motivos conservados, estar presente en la mayoría de las firmicutes y estar siempre relacionado con genes de un mismo proceso (inicio de la replicación), resultó ser causado por sitios de unión de la proteína DnaA [66]. El inicio de replicación está controlado por varios sitios de pegado de DnaA que se encuentran río-arriba y río-abajo del gene *dnaA*, donde el cúmulo de sitios ocasionó que fuera reconocido por el método como un motivo conservado. En este caso, se detectó un primer motivo compuesto por 2 sitios consecutivos de DnaA, seguido por una región conservada rica en A/T, un segundo motivo compuesto por 2 sitios separados por tres nucleótidos y un último compuesto por 1 sitio en la cadena contraria.

Este tipo de casos son difíciles de distinguir automáticamente, ya que la correlación que existe entre los motivos, por incluir un mismo sitio, podría también interpretarse como la señal de covarianza presente en una molécula estructurada de RNA.

5.4.2. Análisis de algunas predicciones

Finalmente, en esta sección se describen algunos de los resultados encontrados que no parecen haber sido descritos anteriormente. En la Tabla 3 se reúne la información pertinente para estos elementos putativos de regulación, agregando principalmente aquellos que presentan evidencia de estructura secundaria. Por lo que se observó en la Tabla 2, se espera que los elementos más promisorios sean aquellos con al menos 3 motivos y una buena predicción de estructura secundaria. Aquellos elementos que además de cumplir con estas dos condiciones presentan un atenuador predicho en más del 30% de los genes, han sido marcados en negritas. Vale la pena mencionar que usando estas 3 condiciones bastante estrictas se recuperan 15 de los 18 elementos regulatorios dependientes de RNA incluidos en la Tabla 2. Como se ha mencionado anteriormente, la lista completa de las predicciones se encuentra en la página principal de *RibEx* (http://www.ibt.unam.mx/biocomputo/ribex.html).

ID [a]	Descripción [b]	COG [c]	Mot. [d]	Estruct. [e]	Phylum [f]
RLE0004	Terminador	COG0536 Predicted GTPase (5/117) $1e^{-08}$	1	★★	acFP
RLE0006	División celular	COG2001 Uncharacterized protein conserved in bacteria (45/65) $7e^{-120}$	2		AFP
RLE0014	ORF huérfano divergente	COG1186 Protein chain release factor B (40/136) $4e^{-112}$	2	ORF	acFP
RLE0017	Regulador divergente de enzimas NADPH	COG1733 Predicted transcriptional regulators (30/197) $6e^{-92}$	2	★	acFP
RLE0018	**ORF huérfano y promotor sigmaA**	**COG2740 Predicted nucleic-acid-binding protein implicated in transcription termination (11/47) $2e^{-28}$**	**3**	**★★ ORF**	**F**
RLE0020	División, partición de cromosomas	COG1192 ATPases involved in chromosome partitioning (18/259) $2e^{-30}$	1	★★	AFP

ID [a]	Descripción [b]	COG [c]	Mot. [d]	Estruct. [e]	Phylum [f]
RLE0023	Palíndrome, stress oxidativo	COG1017 Hemoglobin-like flavoprotein (13/39) $3e^{-35}$	1	★★	AFP
RLE0035	**Autorregulación ribosomal**	**COG0244 Ribosomal protein L10 (26/129) $6e^{-80}$**	**3**	**★★**	**cFP**
RLE0036	Autorregulación ribosomal	COG0360 Ribosomal protein S6 (29/107) $7e^{-76}$	3	★★	acFP
RLE0040	**Autorregulación ribosomal**	**COG0290 Translation initiation factor 3 (IF-3) (19/116) $4e^{-49}$**	**3**	**★★**	**aF**
RLE0042	Dos palíndromes, maltosa	COG0366 Glycosidases (9/211) $7e^{-19}$	2	★	aFp
RLE0048	**Atenuador, citosina**	**COG0504 CTP synthase (UTP-ammonia lyase) (8/131) $1e^{-22}$**	**3**	**★**	**F**
RLE0062	Sitio FruR extendido	COG4668 Mannitol/fructose-specific phosphotransferase system, IIA domain (7/50) $2e^{-18}$	3	★★	fP
RLE0064	Palíndrome, coenzima-A	COG1960 Acyl-CoA dehydrogenases (11/502) $6e^{-20}$	1	★	P
RLE0073	Operón molibdeno, independiente a sitio ModE	COG2896 Molybdenum cofactor biosynthesis enzyme (10/90) $5e^{-28}$	3	★	P
RLE0077	Atenuador, alfa-proteobacteria	COG0541 Signal recognition particle GTPase (6/131) $1e^{-15}$	2	★★	fP
RLE0081	Posible redundancia	COG3157 Hemolysin-coregulated protein (uncharacterized) (6/55) $6e^{-22}$	5	★★	P
RLE0089	**Divergente, ribosomal**	**COG0254 Ribosomal protein L31 (6/130) $9e^{-18}$**	**3**	**★★**	**aP**
RLE0097	ISH8 transposasa	COG3385 FOG: Transposase and inactivated derivatives (37/211) $8e^{-113}$	4	★	pR
RLE0110	**Autorregulación ribosomal**	**COG0051 Ribosomal protein S10 (22/130) $3e^{-46}$**	**5**	**★★**	**FP**
RLE0116	**En medio de operón ribosomal con promotor interno**	**COG0085 DNA-directed RNA polymerase, beta subunit/140 kD subunit (13/133) $4e^{-40}$**	**4**	**★★**	**P**
RLE0130	Motivo 2 trpL, motivo 1 posible elemento móvil	COG0134 Indole-3-glycerol phosphate synthase (9/106) $9e^{-20}$	2	★★	fP
RLE0131	2 palíndromes, motivo 1 es sitio de GntR	COG3265 Gluconate kinase (7/34) $2e^{-23}$	2	★	P

ID [a]	Descripción [b]	COG [c]	Mot. [d]	Estruct. [e]	Phylum [f]
RLE0133	Transporte	COG4708 Predicted membrane protein (6/17) $1e^{-24}$	2	★	Fp
RLE0134	División celular	COG0445 NAD/FAD-utilizing enzyme apparently involved in cell division (12/111) $5e^{-33}$	3	★	P
RLE0135	Transporte de sucrosa, divergente	COG1621 Beta-fructosidases (levanase/invertase) (8/45) $3e^{-24}$	3	★★	F
RLE0136	**Divergente, ribosomal**	**COG0211 Ribosomal protein L27 (10/117) $2e^{-28}$**	**4**	★	**P**
RLE0137	Atenuador histidina, con hisL	COG0040 ATP phosphoribosyltransferase (11/94) $1e^{-26}$	2	★	P
RLE0140	Atenuador de serina, alfa-proteobacteria	COG1932 Phosphoserine aminotransferase (7/64) $2e^{-23}$	2	★★	P
RLE0149	Degradación de nucleótidos, sólo en pocas gamma-proteobacteria	COG0274 Deoxyribose-phosphate aldolase (6/91) $1e^{-20}$	2	★★	Pr
RLE0153	Divergente, biosíntesis de nucleótidos, no sitio PurR	COG0035 Uracil phosphoribosyltransferase (10/101) $5e^{-33}$	2	★★	fP
RLE0154	**sraG RNA**	**COG1185 Polyribo-nucleotide nucleotidyl-transferase (9/107) $8e^{-29}$**	**3**	★★	**P**
RLE0163	Aspartato	COG1448 Aspartate/tyrosine/aromatic aminotransferase (5/45) $3e^{-18}$	2	★★	fP
RLE0168	Función desconocida	COG3513 Uncharacterized protein conserved in bacteria (5/11) $1e^{-19}$	4	★★	Fp
RLE0172	Posible redundancia	COG1189 Predicted rRNA methylase (5/70) $1e^{-13}$	5	★	aF
RLE0174	**Proteína-RNA**	**COG0724 RNA-binding proteins (RRM domain) (8/60) $3e^{-28}$**	**4**	★	**Cp**
RLE0178	**Atenuador**	**COG0219 Predicted rRNA methylase (SpoU class) (6/96) $5e^{-18}$**	**3**	★	**F**
RLE0183	Adenina	COG0742 N6-adenine-specific methylase (7/111) $2e^{-20}$	3	★	aF
RLE0189	Divergente, posible redundancia	COG1539 Dihydroneopterin aldolase (5/87) $3e^{-16}$	2	★★	P

ID [a]	Descripción [b]	COG [c]	Mot. [d]	Estruct. [e]	Phylum [f]
RLE0192	Divergente, posible redundancia	COG3115 Cell division protein (5/20) $1e^{-19}$	3	★	P
RLE0195	**Atenuador, glicerol**	**COG0578 Glycerol-3-phosphate dehydrogenase (6/83) $1e^{-21}$**	**3**	★★	**aF**
RLE0201	ileS de alfa-proteobacterias	COG0060 Isoleucyl-tRNA synthetase (6/135) $2e^{-19}$	2	★	P
RLE0206	**Autorregulación ribosomal**	**COG0048 Ribosomal protein S12 (5/131) $1e^{-14}$**	**5**	★★	**F**
RLE0209	Operón ATP sintasa	COG0711 F0F1-type ATP synthase, subunit b (6/99) $4e^{-17}$	3	★★	Fp
RLE0212	Divergente, posible redundancia	COG1217 Predicted membrane GTPase involved in stress response (5/97) $5e^{-17}$	5	★★	aP
RLE0224	**Divergente, con Ribosomal**	**COG0024 Methionine aminopeptidase (12/162) $3e^{-33}$**	**3**	★	**P**
RLE0229	Divergente con proteína que une RNA	COG1738 Uncharacterized conserved protein (10/58) $2e^{-37}$	2	★★	P
RLE0236	**Biosíntesis purinas, motivo 1 coincide con el promotor**	**COG0516 IMP dehydrogenase/GMP reductase (10/139) $1e^{-28}$**	**3**	★	**FP**
RLE0247	Palíndrome, regulador divergente, sitio BetI	COG1292 Choline-glycine betaine transporter (6/70) $1e^{-16}$	1	★	P
RLE0258	**Autorregulación ribosomal**	**COG1841 Ribosomal protein L30/L7E (9/84) $1e^{-19}$**	**4**	★★	**P**
RLE0269	Divergente, glucosalina, varios sitios CRP sobrelapados en ambas cadenas	COG1820 N-acetylglucosamine-6-phosphate deacetylase (7/76) $2e^{-20}$	2	★	P
RLE0278	Divergente, 2 sitios MetJ en cadena opuesta	COG0241 Histidinol phosphatase and related phosphatases (7/83) $3e^{-24}$	2	★	P
RLE0280	Posible sitio GadE	COG0104 Adenylosuccinate synthase (8/120) $2e^{-26}$	2	★	P
RLE0290	**Operón con posible proteína de unión a DNA**	**COG0820 Predicted Fe-S-cluster redox enzyme (7/100) $5e^{-23}$**	**3**	★	**P**

ID [a]	Descripción [b]	COG [c]	Mot. [d]	Estruct. [e]	Phylum [f]
RLE0295	Divergente, ambos sitios sobrelapan con un promotor	COG1194 A/G-specific DNA glycosylase (6/94) $1e^{-18}$	2	★	P
RLE0303	**Riboswitch SAM-II**	**COG2021 Homoserine acetyltransferase (5/50) $1e^{-15}$**	**4**	★	**P**
RLE0309	Función desconocida, sólo en alfa-proteobacteria	COG3831 Uncharacterized conserved protein (7/15) $5e^{-30}$	2	★★	P
RLE0310	Transporte de metales, posible redundancia	COG0598 Mg2+ and Co2+ transporters (7/135) $4e^{-21}$	5	✲★	P
RLE0319	Sobrelapa con promotor	COG1734 DnaK suppressor protein (6/96) $6e^{-19}$	3	★	P
RLE0327	Divergente, posible redundancia	COG3317 Uncharacterized lipoprotein (6/18) $1e^{-21}$	2	★	P
RLE0329	Posible sitio FruR	COG0205 6-phosphofructokinase (5/102) $7e^{-17}$	2	✲	P
RLE0333	**Sobrelapa con promotor sigma32**	**COG0361 Translation initiation factor 1 (IF-1) (5/135) $5e^{-17}$**	**3**	★	**P**
RLE0336	Posible redundancia	COG1674 DNA segregation ATPase FtsK/SpoIIIE and related (5/158) $5e^{-14}$	4	★★	P
RLE0350	Divergente, sólo en alfa-proteobacteria	COG1052 Lactate dehydrogenase and related dehydrogenases (6/147) $4e^{-20}$	2	★	P

Tabla 3. Ejemplos de otros elementos predichos. Las columnas contienen: identificador tomado de *RibEx* [a], una descripción posible dado un análisis del elemento [b], COG más significativo [c], número de motivos encontrados con MEME [d], si presenta estructuración de RNA predicho por dos métodos (★ sólo por un método, ★★ por ambos métodos, ✲ implica que una prueba fue poco convincente, ORF en caso de que además exista evidencia de marco de lectura abierto) [e], principales phyla que lo contienen (A = Actinobacteria, C = Cyanobacteria, F = Firmicutes, P = Proteobacteria, R = Archaea) donde una letra minúscula implica que sólo aislados organismos de ese phylum lo contienen [f].

Dentro de la Tabla 3 se pueden observar una gran variedad de funciones, incluyendo división celular, reguladores transcripcionales, enzimas asociadas a procesos redox, proteínas ribosomales, metabolismo de azúcares, aminoácidos, nucleótidos, etc. A

continuación se describen algunos resultados puntualmente, empezando por algunos ejemplos que se eliminarían y siguiendo con casos más interesantes:

- Un tipo de elemento que no se esperaba encontrar eran los terminadores transcripcionales, como RLE0004. Estos terminadores están realmente asociados al extremo 3' de los genes y son levantados erróneamente por el método. Pueden ser descartados porque generalmente incluyen sólo un motivo y la asociación con cualquier COG es más débil que la mayoría del resto de los elementos, sin embargo dentro de los resultados existen varios como RLE0004.

- Los resultados incluyen muchos elementos móviles, que por su gran repetitividad son fáciles de localizar y se incluye RLE0097 ejemplificando esto. En particular este transposón es casi exclusivo de Archaeas. En total, se encontraron 21 elementos conservados asociados a transposasas.

- En algunos casos (RLE0081, RLE0172, RLE0212 y RLE0310) parecería que la conservación se debe a la falta de divergencia entre las secuencias. Al constar de 5 motivos, prácticamente toda la región intergénica está conservada. Esto habla posiblemente de que la eliminación de redundancia no fue lo suficientemente estricta y que debieran eliminarse ya sea más número de genomas, o tener un filtro adicional al final.

- Se encuentran otros casos presentes solamente en organismos muy cercanos, y que pudieran también tratarse de redundancia (RLE0189, RLE0192 y RLE0327), sin embargo al encontrarse entre genes divergentes, la conservación también pudiera deberse a esta estructura de regulación.

- Existen muchísimos pares de genes divergentes con motivos conservados. Algunos de estos motivos coinciden con sitios proteína-DNA ya reportados (RLE0247, RLE0269, RLE0278, RLE0280) y muchísimos otros no, sobre todo por estar presente solamente en organismos poco caracterizados. Posiblemente existan constricciones funcionales en promotores divergentes que ocasionen una mayor conservación en estas regiones intergénicas. En el laboratorio se está

realizando un análisis más extenso de la conservación de motivos entre genes divergentes.

- Un ejemplo interesante de regulación divergente se presenta en RLE0135, específico de Lactobacillales, que incluye los operones divergentes *scrBR* y *scrAK*, ambos relacionados al metabolismo de sacarosa. Existe ya evidencia de que ScrR es el encargado de regular ambos operones [92].

- Existen también algunos grupos que coinciden al menos parcialmente con sitios de pegado de proteínas al DNA (RLE0062 y RLE0131) cuya conservación es inusual por alguna razón, como por incluir más de un sitio de pegado.

- Un resultado interesante fue encontrar elementos que presentan marcos de lectura abiertos (ORF) sin parecido alguno en las bases de datos (RLE0014 y RLE0018). La ausencia de codones de paro en estas regiones pudiera ser una casualidad, o se podría tratar de pequeños péptidos que no han sido caracterizados hasta ahora.

- Puesto que se sabe que muchos operones de proteínas ribosomales son auto-regulados por una de sus productos, se esperaba recobrar muchos de estos elementos, algunos ejemplos siendo RLE0035, RLE0036 y RLE0040. El elemento RLE0116 es curioso sin embargo, por estar presente en medio del operón rplJL-rpoBC que incluye proteínas ribosomales y las subunidades β y β' de la RNA polimerasa. Existe un reporte de que esta región interna incluye un sitio de procesamiento que pudiera ser lo que se esta recobrando [11]. Esta fue la categoría en la que se encontraron más elementos conservados, llegando a un total de 26. Se puede observar en la tabla que muchos de los mejores elementos, marcados en negritas, están asociados a proteínas ribosomales.

- Como ejemplos de atenuadores que emplean péptidos líder se encuentra uno de triptofano (RLE0130) y uno de histidina (RLE0137).

- No siempre será posible asignar función tentativamente. RLE0168 es un candidato interesante pero que regula genes de función desconocida.

Algunos de los ejemplos que sería interesante estudiar experimentalmente:

- Existe al menos un caso que coincide perfectamente con un pequeño RNA sraG, el cual ya había sido identificado, pero no caracterizado (RLE0154). Se encuentra divergente de *pnp*, una de las subunidades del degradosoma de RNA.

- Cuando se asocia la señal significativamente con un regulador o con una proteína que une DNA o RNA, se puede tomar el motivo conservado como candidato del sitio de unión de la proteína. RLE0017 es un buen ejemplo de esto, el motivo conservado esta siempre asociado a un miembro del COG1733 de reguladores transcripcionales. El gene río-arriba en cada caso es divergente y es un buen candidato a ser regulado, codificando en general para permeasas y varias reductasas. RLE0229 es otro caso con divergencia entre una proteína no caracterizada y otra que en un caso fue catalogado como que putativamente une RNA.

- Otro elemento entre genes divergentes es RLE0042, donde la función en común gira alrededor de maltosa, encontrándose reguladores, transportadores y enzimas involucradas en su metabolismo. A pesar de estar sumamente distribuido entre los Firmicutes, no está presente en *B. subtilis*, lo cual es probablemente la razón de que no ha sido caracterizado previamente.

- Existen algunos ejemplos que podrían contener mecanismos de atenuación (posiblemente riboswitches), debido al número de motivos y la presencia de atenuadores predichos. RLE0048 parecería depender de alguna manera de citosina, RLE0140 de serina y RLE0195 de glicerol.

- RLE0064 es un palíndrome ampliamente distribuido en alpha y beta-proteobacterias, regulando a varias enzimas asociadas a la coenzima-A.

- A pesar de que se conoce un regulador (ModE) específico para el operón de biosíntesis del cofactor para molibdeno, RLE0073 parece incluir sitios diferentes, conservados en la región líder del transcrito.

- RLE0149 podría tratarse de un regulador en la degradación de nucleótidos, pero al sólo estar presente en *Yersinia*, *Erwinia* y *Vibrio*, no parece haber sido estudiado. El Archaea *Haloarcula. marismortui* contiene el mismo gene, *deoC*, con el mismo motivo conservado, sugiriendo una posible transferencia horizontal, a pesar de que los genes aledaños no sean los mismos.

- Otro par de operones divergentes, involucrados en la biosíntesis de nucleótidos, parece ser regulado por el elemento RLE0153. A pesar de que para *E. coli* K12 se conoce un sitio de PurR para uno de estos operones, los motivos conservados no coinciden con ese sitio, sugiriendo la existencia de otro elemento regulatorio.

- Por último vale la pena mencionar a RLE0303, que se asocia a genes divergentes involucrados en el metabolismo de metionina de alpha-proteobacterias, que acaba de ser descrito independientemente como un riboswitch específico para SAM, haciéndolo el primer caso donde se conocen dos aptámeros naturales diferentes que han convergido en una misma función [20].

6. CONCLUSIONES

El método desarrollado parece ser sumamente robusto siendo capaz de encontrar todos los riboswitches reportados hasta el momento, aún aquellos que fueron encontrados posteriormente al desarrollo de esta tesis. Definitivamente conviene aplicar el método en grupos de organismos más relacionados, como firmicutes o proteobacterias (o inclusive subdivisiones de éstas), para que mecanismos de regulación específicos se encuentren conservados y así pueda detectarse un mayor número de señales. Por ahora, las únicas phyla con suficientes miembros son estos dos, sin embargo, al aumentar el número de organismos completamente secuenciados será interesante aplicar estas estrategias para encontrar novedosos sistemas regulatorios exclusivos de actinobacterias, cyanobacterias, archaeas y otros. Inclusive se puede pensar en una estrategia relacionada para estudiar elementos regulatorios en eucariontes.

Efectivamente, se pudo encontrar muchos candidatos a elementos de regulación, entre los cuales seguramente habrá muchos elementos tipo riboswitch. Sin embargo, no todos los motivos resultantes se encuentran tan conservados ni tan ampliamente distribuidos y es difícil asegurar que se traten de elementos que actúan a nivel del RNA. Además, varios de los riboswitches descubiertos recientemente están poco conservados y distribuidos, lo cual dificulta aún más la clasificación de los resultados. Falta todavía mejorar los criterios automatizados de evaluación para que se pueda confiablemente plantear experimentos sobre los mejores candidatos sin necesitar demasiados análisis manuales previos. Sin embargo, el conjunto de resultados presentado es de por sí una selección bastante enriquecida en elementos regulatorios *bona fide* y sería muy interesante que se analicen a fondo para completar su descripción.

Por ahora, el presentar los resultados en el Internet, permite a toda la comunidad científica tener acceso al fruto de este trabajo. La descripción detallada de todos los elementos regulatorios, tanto conocidos como predichos, incluyendo su posible función, distribución filogenética y el contexto de los genes que son regulados, es de una gran utilidad para cualquiera que desea profundizar en estos sistemas. El ofrecer un servicio que realiza búsquedas sobre cualquier secuencia para visualmente localizar elementos regulatorios conocidos y predichos, además de atenuadores y marcos de lectura abiertos, permite que hasta los usuarios más casuales averigüen nuevos sistemas de regulación en las secuencias de su interés.

El presente ha sido un trabajo que demuestra el gran poder de las herramientas computacionales para explotar la información disponible gracias a las técnicas masivas de secuenciación. Se vislumbra que en un futuro la cantidad de información siga creciendo exponencialmente, sobre todo con los proyectos de secuenciación de metagenomas [84, 87], por lo cual enfoques como el de esta tesis serán cada vez más importantes.

7. PERSPECTIVAS

- Implementar una mejor automatización del proceso, para que los resultados puedan mantenerse actualizados con la menor intervención humana.

- Aprovechar todos los resultados, aún los que se descartan por no ser elementos regulatorios. Si se encuentran genes no anotados, o mal anotados, proveer esta información a las bases públicas.
    - Buscar correlaciones con genes, para encontrar candidatos a unirse en el sitio.

- Realizar búsquedas dirigidas para encontrar nuevos sistemas regulatorios en aquellos clados que se sabe que carecen del elemento regulatorio clásico.

- Optimizar la determinación automatizada de estructuras secundarias conservadas, para la selección de mejores candidatos a ser riboswitches.

- Realizar un análisis más cuidadoso de la distribución de los distintos elementos regulatorios encontrados. Tratar de dilucidar el origen, así como el patrón de herencia vertical y horizontal de cada uno.
    - Profundizar en la presencia de riboswitches bacterianos en genes de eucariontes. ¿Existe evidencia para soportar un origen previo a la separación de los reinos, o llegaron por transferencia horizontal?

- Localizar elementos de regulación en bases de datos menos caracterizadas, como los metagenomas, para extender el conocimiento de su distribución y encontrar posibles genes análogos (sustituciones no-ortólogas).

- Implementar modificaciones del método para realizar búsquedas en eucariontes, donde se buscaría conservación en intrones y regiones 3' no traducidas.

- Utilizar otro método de agrupación inicial, como por KEGG o grupos más específicas de ortólogos, cuidando de no incluir parálogo alguno.

A1. GLOSARIO

Ado-CBL.....Adenosil-cobalamina, forma de la vitamina B_{12} reconocida por un riboswitch.

Aptámero.....Región de RNA capaz de reconocer específicamente y con alta afinidad a un metabolito.

BioCyc .....Conjunto de bases de datos con información acerca de regulación y caminos metabólicos.

BLAST.....Basic Local Alignment Search Tool. Herramienta utilizada para comparar y alinear secuencias.

CIRCE.....Controlling Inverted Repeat of Chaperone Expression. Operador de genes de choque térmico, sitio de unión de dos proteínas HrcA.

COG .....Cluster of Orthologous Groups of proteins. Grupos de proteínas organizados por ortología.

DBTBS.....DataBase of Transcriptional regulation in *Bacillus subtilis*.

DNA.....Ácido Deoxirribonucleico, material genético.

FMN.....Flavin-mononucleótido.

GenBank .....Base de datos de nucleótidos, públicamente disponible por el NCBI.

Gene.....Unidad básica de información genética, que codifica para un RNA o una proteína.

Hipergeométrica.....Distribución binomial sin reemplazo.

iMUR .....Intergenic Minimal Upstream Region. Una MUR recortada para incluir solamente nucleótidos intergénicos.

Mejor Hit Bi-Direccional.....Criterio para definir computacionalmente que dos genes son ortólogos. Se cumple cuando el primer gene encuentra al segundo con la mejor calificación, al ser comparado contra todos los genes del segundo organismo y que el segundo encuentre al primero con la mejor calificación, al

ser comparado contra todos los genes del primer organismo.

MEME.....Multiple EM for Motif Elicitation. Herramienta utilizada para descubrir patrones o motivos en un conjunto de secuencias.

Microarreglo .....Experimento para analizar simultáneamente la expresión de un gran número de genes de un organismo.

miRNA.....Micro RNA, involucrado en la represión de mensajeros.

Motivo.....Patrón o segmento continuo de letras (nucleótidos o aminoácidos).

MUR .....Minimal Upstream Region. Región de 400 nucleótidos río arriba de un gene y 50 río adentro. Contiene la mayoría de los elementos regulatorios de ese gene.

NCBI.....National Center for Biotechnology Information. Organización que entre otras cosas provee al público de BLAST y GenBank.

ncRNA .....RNA no codificante. Cualquier molécula funcional de RNA que no es traducido a una proteína.

NOG.....Non-suppervised Orthologous Group. Grupos de proteínas que no pertenecen a ningún COG, organizadas automáticamente por criterios de ortología.

Ortólogos.....Genes homólogos cuya separación fue debido a un evento de especiación.

Operón.....Conjunto de genes transcritos en una sola unidad y que comparten por lo tanto una región de regulación.

p-value.....Valor que representa la probabilidad de que algo haya ocurrido al azar.

Parálogos.....Genes homólogos cuya separación fue por un evento de duplicación.

Phylum .....La siguiente división taxonómica después de reino.

Procariontes.....Organismos sin núcleo, incluye a los reinos eubacteria y archaeabacteria.

PSI-BLAST.....Position-Specific Iterated BLAST. Una herramienta basada el BLAST que mediante iteraciones logra encontrar homólogos remotos.

Regulón.....Conjunto de genes regulados por un mismo elemento.

RegulonDB.....Base de datos que reúne información sobre regulación transcripcional de *E. coli* K12.

Rfam.....Base de datos de familias de RNA, tanto codificantes como no codificantes.

Riboswitch.....Elemento regulatorio presente en el RNA que reconoce directamente a un metabolito, sin necesidad de un intermediario.

Ribozima.....Molécula de RNA capaz de catalizar una reacción bioquímica.

RNA.....Ácido ribonucleico, anteriormente considerada como una molécula intermediaria, ahora se sabe que tiene muchísimas más funciones.

RNAi.....RNA interference, proceso en que un siRNA ocasiona la represión de un mensajero.

SAM.....S-adenosil-metionina.

Shine-Dalgarno.....Sitio de unión del ribosoma.

siRNA.....Small interfering RNA, involucrado en la represión de mensajeros.

snoRNA.....Small nucleolar RNA, relacionado con la modificación de nucleótidos en RNA ribosomal y de transferencia.

Splicing.....Proceso post-transcripcional en que se eliminan los intrones y se empalman los exones.

TPP.....Tiamin-pirofosfato, forma activa de la vitamina B_1 reconocida por un riboswitch.

A2. BIBLIOGRAFÍA

1. Abreu-Goodger, C. y Merino, E. **2005**. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res.* 33: W690-W692.
2. Abreu-Goodger, C., Ontiveros-Palacios, N., Ciria, R. y Merino, E. **2004**. Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet.* 20: 475-479.
3. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. y Lipman, D.J. **1990**. Basic local alignment search tool. *J. Mol. Biol.* 215: 403-410.
4. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. y Lipman, D.J. **1997**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389-3402.
5. Babitzke, P. y Yanofsky, C. **1993**. Reconstitution of Bacillus subtilis trp attenuation in vitro with TRAP, the trp RNA-binding attenuation protein. *Proc. Natl. Acad. Sci. U. S. A.* 90: 133-137.
6. Bachellerie, J.P., Cavaille, J. y Huttenhofer, A. **2002**. The expanding snoRNA world. *Biochimie.* 84: 775-790.
7. Bagg, A. y Neilands, J.B. **1987**. Ferric uptake regulation protein acts as a repressor, employing iron (II) as a cofactor to bind the operator of an iron transport operon in Escherichia coli. *Biochemistry.* 26: 5471-5477.
8. Bailey, T.L. y Elkan, C. **1994**. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 2: 28-36.
9. Bailey, T.L. y Gribskov, M. **1998**. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics.* 14: 48-54.
10. Barrick, J.E., Corbino, K.A., Winkler, W.C., Nahvi, A., Mandal, M., Collins, J., Lee, M., Roth, A., Sudarsan, N., Jona, I., Wickiser, J.K. y Breaker, R.R. **2004**. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. U. S. A.* 101: 6421-6426.
11. Barry, G., Squires, C. y Squires, C.L. **1980**. Attenuation and processing of RNA from the rplJL--rpoBC transcription unit of Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* 77: 3331-3335.
12. Batey, R.T., Gilbert, S.D. y Montange, R.K. **2004**. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature.* 432: 411-415.
13. Benard, L., Philippe, C., Ehresmann, B., Ehresmann, C. y Portier, C. **1996**. Pseudoknot and translational control in the expression of the S15 ribosomal protein. *Biochimie.* 78: 568-576.
14. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. y Wheeler, D.L. **2005**. GenBank. *Nucleic Acids Res.* 33: D34-D38.
15. Beuzon, C.R. y Casadesus, J. **1997**. Conserved structure of IS200 elements in Salmonella. *Nucleic Acids Res.* 25: 1355-1361.
16. Breaker, R.R., Emilsson, G.M., Lazarev, D., Nakamura, S., Puskarz, I.J., Roth, A. y Sudarsan, N. **2003**. A common speed limit for RNA-cleaving ribozymes and deoxyribozymes. *RNA.* 9: 949-957.
17. Brent, R. y Ptashne, M. **1981**. Mechanism of action of the lexA gene product. *Proc. Natl. Acad. Sci. U. S. A.* 78: 4204-4208.
18. Bussemaker, H.J., Li, H. y Siggia, E.D. **2000**. Regulatory element detection using a probabilistic segmentation model. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* 8: 67-74.
19. Ciria, R., Abreu-Goodger, C., Morett, E. y Merino, E. **2004**. GeConT: gene context analysis. *Bioinformatics.* 20: 2307-2308.
20. Corbino, K.A., Barrick, J.E., Lim, J., Welz, R., Tucker, B.J., Puskarz, I., Mandal, M., Rudnick, N.D. y Breaker, R.R. **2005**. Evidence for a second class of S-adenosylmethionine riboswitches and other regulatory RNA motifs in alpha-proteobacteria. *Genome Biol.* 6: R70-
21. Crosa, J.H. **1997**. Signal transduction and transcriptional and posttranscriptional control of iron-regulated genes in bacteria. *Microbiol. Mol. Biol. Rev.* 61: 319-336.
22. Doudna, J.A. y Cech, T.R. **2002**. The chemical repertoire of natural ribozymes. *Nature.* 418: 222-228.
23. Eisen, M.B., Spellman, P.T., Brown, P.O. y Botstein, D. **1998**. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. U. S. A.* 95: 14863-14868.
24. Ellington, A.D. y Szostak, J.W. **1990**. In vitro selection of RNA molecules that bind specific ligands. *Nature.* 346: 818-822.

25. Epshtein, V., Mironov, A.S. y Nudler, E. **2003**. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 100: 5052-5056.
26. Gelfand, M.S., Mironov, A.A., Jomantas, J., Kozlov, Y.I. y Perumov, D.A. **1999**. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends Genet.* 15: 439-442.
27. Gottesman, S. **2004**. The small RNA regulators of Escherichia coli: roles and mechanisms*. *Annu. Rev. Microbiol.* 58: 303-328.
28. Griffiths-Jones, S., Moxon, S., Marshall, M., Khanna, A., Eddy, S.R. y Bateman, A. **2005**. Rfam: annotating non-coding RNAs in complete genomes. *Nucleic Acids Res.* 33: D121-D124.
29. Grundy, F.J. y Henkin, T.M. **1993**. tRNA as a positive regulator of transcription antitermination in B. subtilis. *Cell.* 74: 475-482.
30. Grundy, F.J. y Henkin, T.M. **1998**. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol.* 30: 737-749.
31. Grundy, F.J. y Henkin, T.M. **2003**. The T box and S box transcription termination control systems. *Front Biosci.* 8: d20-d31.
32. Grundy, F.J. y Henkin, T.M. **2004**. Regulation of gene expression by effectors that bind to RNA. *Curr. Opin. Microbiol.* 7: 126-131.
33. Grundy, F.J., Lehman, S.C. y Henkin, T.M. **2003**. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci. U. S. A.* 100: 12057-12062.
34. Henkin, T.M. y Yanofsky, C. **2002**. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *Bioessays.* 24: 700-707.
35. Hiraga, S. y Yanofsky, C. **1973**. Inhibition of the progress of transcription on the tryptophan operon of Escherichia coli. *J. Mol. Biol.* 79: 339-349.
36. Hofacker, I.L., Fekete, M. y Stadler, P.F. **2002**. Secondary structure prediction for aligned RNA sequences. *J. Mol. Biol.* 319: 1059-1066.
37. Huttenhofer, A., Schattner, P. y Polacek, N. **2005**. Non-coding RNAs: hope or hype? *Trends Genet.* 21: 289-297.
38. Jacob, F. y Monod, J. **1961**. Genetic regulatory mechanisms in the synthesis of proteins. *J. Mol. Biol.* 3: 318-356.
39. Jacob, F., Perrin, D., Sanchez, C. y Monod, J. **1960**. [Operon: a group of genes with the expression coordinated by an operator.]. *C. R. Hebd. Seances Acad. Sci.* 250: 1727-1729.
40. Kanehisa, M., Goto, S., Kawashima, S. y Nakaya, A. **2002**. The KEGG databases at GenomeNet. *Nucleic Acids Res.* 30: 42-46.
41. Kanz, C., Aldebert, P., Althorpe, N., Baker, W., Baldwin, A., Bates, K., Browne, P., van den, B.A., Castro, M., Cochrane, G., Duggan, K., Eberhardt, R., Faruque, N., Gamble, J., Diez, F.G., Harte, N., Kulikova, T., Lin, Q., Lombard, V., Lopez, R., Mancuso, R., McHale, M., Nardone, F., Silventoinen, V., Sobhany, S., Stoehr, P., Tuli, M.A., Tzouvara, K., Vaughan, R., Wu, D., Zhu, W. y Apweiler, R. **2005**. The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 33: D29-D33.
42. Keles, S., van der, L.M. y Eisen, M.B. **2002**. Identification of regulatory elements using a feature selection method. *Bioinformatics.* 18: 1167-1175.
43. Kolter, R. y Yanofsky, C. **1982**. Attenuation in amino acid biosynthetic operons. *Annu. Rev. Genet.* 16: 113-134.
44. Krummenacker, M., Paley, S., Mueller, L., Yan, T. y Karp, P.D. **2005**. Querying and computing with BioCyc databases. *Bioinformatics.* 21: 3454-3455.
45. Kubodera, T., Watanabe, M., Yoshiuchi, K., Yamashita, N., Nishimura, A., Nakai, S., Gomi, K. y Hanamoto, H. **2003**. Thiamine-regulated gene expression of Aspergillus oryzae thiA requires splicing of the intron containing a riboswitch-like domain in the 5'-UTR. *FEBS Lett.* 555: 516-520.
46. Lawrence, C.E. y Reilly, A.A. **1990**. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins.* 7: 41-51.
47. Little, J.W., Mount, D.W. y Yanisch-Perron, C.R. **1981**. Purified lexA protein is a repressor of the recA and lexA genes. *Proc. Natl. Acad. Sci. U. S. A.* 78: 4199-4203.
48. Lu, Y. y Switzer, R.L. **1996**. Evidence that the Bacillus subtilis pyrimidine regulatory protein PyrR acts by binding to pyr mRNA at three sites in vivo. *J. Bacteriol.* 178: 5806-5809.

49. Lundrigan, M.D., Koster, W. y Kadner, R.J. **1991**. Transcribed sequences of the Escherichia coli btuB gene control its expression and regulation by vitamin B12. *Proc. Natl. Acad. Sci. U. S. A.* 88: 1479-1483.
50. Majdalani, N., Vanderpool, C.K. y Gottesman, S. **2005**. Bacterial small RNA regulators. *Crit Rev. Biochem. Mol. Biol.* 40: 93-113.
51. Makita, Y., Nakao, M., Ogasawara, N. y Nakai, K. **2004**. DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res.* 32: D75-D77.
52. Mandal, M., Boese, B., Barrick, J.E., Winkler, W.C. y Breaker, R.R. **2003**. Riboswitches control fundamental biochemical pathways in Bacillus subtilis and other bacteria. *Cell.* 113: 577-586.
53. Mandal, M. y Breaker, R.R. **2004**. Adenine riboswitches and gene activation by disruption of a transcription terminator. *Nat. Struct. Mol. Biol.* 11: 29-35.
54. Mandal, M. y Breaker, R.R. **2004**. Gene regulation by riboswitches. *Nat. Rev. Mol. Cell Biol.* 5: 451-463.
55. Mandal, M., Lee, M., Barrick, J.E., Weinberg, Z., Emilsson, G.M., Ruzzo, W.L. y Breaker, R.R. **2004**. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science.* 306: 275-279.
56. Masse, E. y Gottesman, S. **2002**. A small RNA regulates the expression of genes involved in iron metabolism in Escherichia coli. *Proc. Natl. Acad. Sci. U. S. A.* 99: 4620-4625.
57. McGuire, A.M., Hughes, J.D. y Church, G.M. **2000**. Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res.* 10: 744-757.
58. Merino, E. y Yanofsky, C. **2005**. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet.* 21: 260-264.
59. Miranda-Rios, J., Morera, C., Taboada, H., Davalos, A., Encarnacion, S., Mora, J. y Soberon, M. **1997**. Expression of thiamin biosynthetic genes (thiCOGE) and production of symbiotic terminal oxidase cbb3 in Rhizobium etli. *J. Bacteriol.* 179: 6887-6893.
60. Miranda-Rios, J., Navarro, M. y Soberon, M. **2001**. A conserved RNA structure (thi box) is involved in regulation of thiamin biosynthetic gene expression in bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 98: 9736-9741.
61. Moreno-Hagelsieb, G. y Collado-Vides, J. **2002**. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics.* 18 Suppl 1: S329-S336.
62. Nahvi, A., Barrick, J.E. y Breaker, R.R. **2004**. Coenzyme B12 riboswitches are widespread genetic control elements in prokaryotes. *Nucleic Acids Res.* 32: 143-150.
63. Nelson, P., Kiriakidou, M., Sharma, A., Maniataki, E. y Mourelatos, Z. **2003**. The microRNA world: small is mighty. *Trends Biochem. Sci.* 28: 534-540.
64. Nocker, A., Hausherr, T., Balsiger, S., Krstulovic, N.P., Hennecke, H. y Narberhaus, F. **2001**. A mRNA-based thermosensor controls expression of rhizobial heat shock genes. *Nucleic Acids Res.* 29: 4800-4807.
65. Nudler, E. y Mironov, A.S. **2004**. The riboswitch control of bacterial metabolism. *Trends Biochem. Sci.* 29: 11-17.
66. Ogura, Y., Imai, Y., Ogasawara, N. y Moriya, S. **2001**. Autoregulation of the dnaA-dnaN operon and effects of DnaA protein levels on replication initiation in Bacillus subtilis. *J. Bacteriol.* 183: 3833-3841.
67. Ontiveros-Palacios, N. **2004**. Identificación de secuencias de regulación "riboswitches" en operones de la biosíntesis de vitaminas y cofactores. *Tesis para obtener el título de Bióloga.*
68. Permina, E.A. y Gelfand, M.S. **2003**. Heat shock (sigma32 and HrcA/CIRCE) regulons in beta-, gamma- and epsilon-proteobacteria. *J. Mol. Microbiol. Biotechnol.* 6: 174-181.
69. Qi, J., Wang, B. y Hao, B.I. **2004**. Whole proteome prokaryote phylogeny without sequence alignment: a K-string composition approach. *J. Mol. Evol.* 58: 1-11.
70. Richter-Dahlfors, A.A., Ravnum, S. y Andersson, D.I. **1994**. Vitamin B12 repression of the cob operon in Salmonella typhimurium: translational control of the cbiA gene. *Mol. Microbiol.* 13: 541-553.
71. Rivas, E. y Eddy, S.R. **2001**. Noncoding RNA gene detection using comparative sequence analysis. *BMC. Bioinformatics.* 2: 8-

72. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. y Gelfand, M.S. **2004**. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. *Nucleic Acids Res.* 32: 3340-3353.
73. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. y Gelfand, M.S. **2002**. Comparative genomics of thiamin biosynthesis in procaryotes. New genes and regulatory mechanisms. *J. Biol. Chem.* 277: 48949-48959.
74. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. y Gelfand, M.S. **2003**. Comparative genomics of the vitamin B12 metabolism and regulation in prokaryotes. *J. Biol. Chem.* 278: 41148-41159.
75. Rodionov, D.A., Vitreschak, A.G., Mironov, A.A. y Gelfand, M.S. **2003**. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res.* 31: 6748-6757.
76. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., az-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C. y Collado-Vides, J. **2004**. RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res.* 32: D303-D306.
77. Schlax, P.J., Xavier, K.A., Gluick, T.C. y Draper, D.E. **2001**. Translational repression of the Escherichia coli alpha operon mRNA: importance of an mRNA conformational switch and a ternary entrapment complex. *J. Biol. Chem.* 276: 38494-38501.
78. Soukup, J.K. y Soukup, G.A. **2004**. Riboswitches exert genetic control through metabolite-induced conformational change. *Curr. Opin. Struct. Biol.* 14: 344-349.
79. Soutschek, J., Akinc, A., Bramlage, B., Charisse, K., Constien, R., Donoghue, M., Elbashir, S., Geick, A., Hadwiger, P., Harborth, J., John, M., Kesavan, V., Lavine, G., Pandey, R.K., Racie, T., Rajeev, K.G., Rohl, I., Toudjarska, I., Wang, G., Wuschko, S., Bumcrot, D., Koteliansky, V., Limmer, S., Manoharan, M. y Vornlocher, H.P. **2004**. Therapeutic silencing of an endogenous gene by systemic administration of modified siRNAs. *Nature.* 432: 173-178.
80. Sudarsan, N., Barrick, J.E. y Breaker, R.R. **2003**. Metabolite-binding RNA domains are present in the genes of eukaryotes. *RNA.* 9: 644-647.
81. Sudarsan, N., Wickiser, J.K., Nakamura, S., Ebert, M.S. y Breaker, R.R. **2003**. An mRNA structure in bacteria that controls gene expression by binding lysine. *Genes Dev.* 17: 2688-2697.
82. Tatusov, R.L., Koonin, E.V. y Lipman, D.J. **1997**. A genomic perspective on protein families. *Science.* 278: 631-637.
83. Tijsterman, M. y Plasterk, R.H. **2004**. Dicers at RISC; the mechanism of RNAi. *Cell.* 117: 1-3.
84. Tyson, G.W., Chapman, J., Hugenholtz, P., Allen, E.E., Ram, R.J., Richardson, P.M., Solovyev, V.V., Rubin, E.M., Rokhsar, D.S. y Banfield, J.F. **2004**. Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature.* 428: 37-43.
85. van Helden, J., Andre, B. y Collado-Vides, J. **1998**. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281: 827-842.
86. Vassinova, N. y Kozyrev, D. **2000**. A method for direct cloning of fur-regulated genes: identification of seven new fur-regulated loci in Escherichia coli. *Microbiology.* 146 Pt 12: 3171-3182.
87. Venter, J.C., Remington, K., Heidelberg, J.F., Halpern, A.L., Rusch, D., Eisen, J.A., Wu, D., Paulsen, I., Nelson, K.E., Nelson, W., Fouts, D.E., Levy, S., Knap, A.H., Lomas, M.W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.H. y Smith, H.O. **2004**. Environmental genome shotgun sequencing of the Sargasso Sea. *Science.* 304: 66-74.
88. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. y Gelfand, M.S. **2002**. Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and translational attenuation. *Nucleic Acids Res.* 30: 3141-3151.
89. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. y Gelfand, M.S. **2003**. Regulation of the vitamin B12 metabolism and transport in bacteria by a conserved RNA structural element. *RNA.* 9: 1084-1097.
90. Vitreschak, A.G., Rodionov, D.A., Mironov, A.A. y Gelfand, M.S. **2004**. Riboswitches: the oldest mechanism for the regulation of gene expression? *Trends Genet.* 20: 44-50.
91. von Mering C., Jensen, L.J., Snel, B., Hooper, S.D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M.A. y Bork, P. **2005**. STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic Acids Res.* 33: D433-D437.

92. Wang, B. y Kuramitsu, H.K. **2003**. Control of enzyme IIscr and sucrose-6-phosphate hydrolase activities in Streptococcus mutans by transcriptional repressor ScrR binding to the cis-active determinants of the scr regulon. *J. Bacteriol.* 185: 5791-5799.
93. Wilson, A.C. y Tan, M. **2004**. Stress response gene regulation in Chlamydia is dependent on HrcA-CIRCE interactions. *J. Bacteriol.* 186: 3384-3391.
94. Winkler, W., Nahvi, A. y Breaker, R.R. **2002**. Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature.* 419: 952-956.
95. Winkler, W.C., Cohen-Chalamish, S. y Breaker, R.R. **2002**. An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. U. S. A.* 99: 15908-15913.
96. Winkler, W.C., Nahvi, A., Roth, A., Collins, J.A. y Breaker, R.R. **2004**. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature.* 428: 281-286.
97. Winkler, W.C., Nahvi, A., Sudarsan, N., Barrick, J.E. y Breaker, R.R. **2003**. An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nat. Struct. Biol.* 10: 701-707.