UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

IDENTIFICACION DE SECUENCIAS DE REGULACION "RIBOSWITCHES" EN OPERONES DE LA BIOSINTESIS DE VITAMINAS Y COFACTORES.

T E S I S

QUE PARA OBTENER EL TITULO DE:

B I O L O G A

P R E S E N T A :

NANCY GUADALUPE ONTIVEROS PALACIOS

DIRECTOR DE TESIS: DR. ENRIQUE MERINO PEREZ

Nancy Guadalupe Ontiveros
Palacios
7 Junio 2009

ACT. MAURICIO AGUILAR GONZÁLEZ
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

"Identificación de secuencias de regulación Riboswitches¨en operones de la biosintesis de vitaminas y cofactores"

realizado por Nancy Guadalupe Ontiveros Palacios

con número de cuenta 400072162 , quien cubrió los créditos de la carrera de: Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario Dr. Enrique Merino Pérez

Propietario Dr. Mario Soberón Chavez

Propietario Dr. Lorenzo Segovia Forcella

Suplente Biol. Luis José Delaye Arredondo

Suplente Dr. Arturo Carlos II Becerra Bracho

FACULTAD DE CIENCIAS
UNIDAD DE ENSEÑANZA DE BIOLOGÍA

Consejo Departamental de Biología

M. en C. Juan Manuel Rodriguez Chavez

Dedicada...

A la vida

y a los responsables
de la mía

AGRADECIMIENTOS

Que conste que yo no quería agradecer tanto, pero esta es la única forma que tengo de agradecer a todas las personas que han dejado una huella en mi

Quiero agradecer a mis padres, "gracias a ustedes soy", a mi padre que me a motivado en todo momento para superarme y que es mi mejor ejemplo de que siempre se puede dar más y a mi madre que es una luchadora incansable de la que estoy muy orgullosa. Gracias papas por quererme tanto.

A la personita que me cambio el mundo. Moncho eres lo mejor que tengo, me enorgullece tenerte como mi hermano.

A mi abuelo que siempre me alienta. Abuelo, aun sigo intentando ser una mujer juiciosa.

A la UNAM por ser todo lo que soñé y más.

A Patricia y Jose Luis por que por su perseverancia a seguido el Taller.

A mi querido Doctor Merino, gracias Enrique por confiar en mi, por toda tu paciencia, por guiarme en todo momento, por darme la oportunidad, por todo lo que e aprendido de la tesis y todo lo que e aprendido de ti y por tu amistad. Lo logre!, y creo que hicimos un buen trabajo

Quiero agradecer a mis sinodales, por el interés que pusieron en este trabajo y por todos sus comentarios y de forma particular a Mario, por que su trabajo en Riboswitches es consecuencia directa de que se aya realizado esta tesis, a Arturo y a Lorenzo por que sus clases y sus platicas siempre me han despertado el interés biológico, evolutivo y científico y a Luis por ser tan accesible y por todo su apoyo.

Quiero agradecer a los miembros del laboratorio, a Ricardo, Paty, Ana banana, Ruy, Rosi y Norma. Gracias por sus discusiones y por su apoyo en todo momento. Somos poquitos, pero somos bien "Chingones".

Y de forma muy especial quiero agradecer al Cei, primero por ser mi amigo y cuidarme como cualquier irresponsable hermano mayor y segundo por toda la ayuda y todos los buenos consejos. Yo se que el merito de la tesis es propio, pero que bueno es tener un Cei a la mano cuando se le necesita.

Hay les va el agradecimiento más largo, pero quien nos manda ser una familia tan grande, quiero agradecer a toda mi familia, por todos los buenos ratos que e compartido con ustedes y por que como buenos tíos, primos y sobrinos, siempre han estado al tanto de mi, a mi tía Lulú y a mi tío Bernardo por todo su apoyo y por que siempre me han considerado parte importante de su familia, gracias!!!. Gracias tío Nacho por compartirme tu tiempo y por todo tu cariño, gracias también a mis tíos Gilberto, José Luis y Mundo, por que me han cuidado todo este tiempo y con su ayuda e logrado llegar hasta aquí. Y gracias muy especiales a mi tío Enrique y a mi tía Mary, por que se convirtieron en mi segunda familia y por que tomaron el riesgo de aceptar en su casa a una chavala de 17 años, gracias por su apoyo incondicional, por compartir su vida conmigo y por ser mis amigos.

Gracias mi querido Alejandro has sido un hermano para mi.

A mis compañeros y cómplices. Amigos, por ustedes no solo e vivido, e vivido intensamente, gracias por que han sido una familia para mi. Gracias Idalia por siempre estar para mi; Ana, gracias por tantas historias compartidas; Laura, tu y yo siempre estaremos soñando; Mariana, contigo las cosas siempre son más sencillas, Emanuel, por biólogos como tu el mundo se va a salvar, te adoro; Ruth, gracias por tu amistad sincera y por tu eterna alegría; Andrés, Miriam, Mariel, Itchel, Edith, Iván David, todos parte importante de lo que e vivido. Gracias a todos por tantos momentos y tantos sueños compartidos, han sido lo mejor de la carrera.

A dos pequeñas, que son un buen ejemplo de que cuando las cosas se quieren, siempre se logran, Gracias Silvia y Tere, me han llenado de buenos momentos. Y Silvia, eres la mejor compañera de casa que haya tenido.

A la Laura, eres la mejor contribución que tengo del Politécnico

A los buenos amigos, que siempre me hacen cuestionarme y revalorar lo que hago, Álvaro, Alejandro y Keko, gracias por todos los buenos momentos que e pasado en su compañía.

Quiero agradecer también a la raza de Chihuahua, a los que siempre han estado pendientes de mi, a la Pamys, al Borre (Luis C), a la Nidia, a Mario, a Betita y a los que no han estado tan al tanto, pero que son parte importante de una etapa de mi vida y que compartimos el largo y emocionante proceso de acabar una carrera, a Ivan, Moc, Pepe lupe, Mauro, mi dulce Clarita, Ivalú. Me da gusto que la mayoría ya sean Lic o Ing.

Índice

Índice de Figuras

Índice de Tablas

Índice de Cuadros

Identificación de secuencias de regulación "Riboswitches" en operones de la biosíntesis de vitaminas y cofactores

Nancy Guadalupe Ontiveros Palacios

Es nuestro interés señalar esta perspectiva para los estudiantes que desean realizar investigación en evolución; no tienen que decidir entre ser paleontólogos, taxónomos, genetistas o ecólogos, sino simplemente evolucionistas; solo deben de tener muy claros los problemas que quieren estudiar, los principios de la biología evolutiva y la mente abierta para aprender técnicas y métodos nuevos

Juan Ñuñez-Farfán y Luis E. Eguiarte (1999)

1. Prefacio

Fig 1. Modelo de la estructura del DNA

Empezare citando dos acontecimientos muy importantes en el desarrollo de las Ciencias Biológicas. El 25 de abril de 1953, el genetista estadounidense James Dewey Watson y el británico Francis Harry Compton Crick proponen el **modelo de la estructura del DNA** [1], una molécula formada por dos cadenas complementarias que se enrollan formando una doble hélice, algo parecido a una larga escalera de caracol (Fig 1). Los lados de la escalera, están constituidos por moléculas de fosfato e hidratos de carbono y los escalones están representados por bases nitrogenadas, dispuestas en parejas. Cada base está unida por un enlace de hidrógeno a una base complementaria localizada en la cadena opuesta. La adenina (A) siempre se vincula con la timina (T), y la guanina (G) siempre con la citosina (C), (A-T, G-C). Este hecho se catalogó como el descubrimiento científico de mayor relevancia en el siglo XX y con justa razón, ya que favoreció de manera decisiva el desarrollo de la Biología Molecular. El otro acontecimiento científico de gran relevancia, que hasta ahora caracteriza al siglo XXI, es la **secuenciación del genoma humano** [2, 3]. El 26 de junio del 2000 los lideres del proyecto genoma humano y de la compañía Celera Genomics anunciaron tener la secuencia completa del genoma humano. Así hace 50 años el descubrimiento de la doble hélice del DNA marco el inicio de la era de la Biología Molecular y hoy, a solo tres años, la secuenciación del genoma humano marca la consolidación de la era genómica. La Ciencia Genómica a tenido un gran avance y se ve reflejado en el creciente número de genomas secuenciados y en la gran cantidad de estudios que se realizan con los mismos. La enorme cantidad de información que se está generando con los genomas secuenciados exige el desarrollo de nuevas herramientas para su análisis y es por ésto que este trabajo es importante, ya que analiza procesos de regulación mediante genómica comparativa.

2. Marco Teórico

2.1 Procesos Genéticos

Ácidos nucleicos (DNA, RNA)

Los ácidos nucleicos: ácido desoxirribonucleico (DNA) y el ácido ribonucleico (RNA), se diferencian estructuralmente en la presencia de desoxirribosa en el primer caso, y ribosa en el segundo, así como la presencia de timina en el DNA y uracilo en el RNA.

Transmisión de la información genética de DNA a proteínas

La Replicación, Transcripción y Traducción son los procesos esenciales que dirigen la conservación y transmisión de la información genética (Fig 2).

Replicación "Es el proceso mediante el cuál se sintetizan dos moléculas hijas de DNA de doble hélice a partir de un DNA progenitor, que actúa como molde" [4]

Transcripción "Es el proceso a traves del cual se copia una hebra de DNA a una secuencia complementaria de RNA de cadena sencilla, mediante una enzima llamada RNA polimerasa" [4]

Traducción "Es el proceso por el cual la secuencia de nucleótidos de una molécula de RNA mensajero dirige la incorporación de aminoácidos a una proteína. Utiliza RNAs de transferencia, RNAs ribosomales y ribosomas" [4]

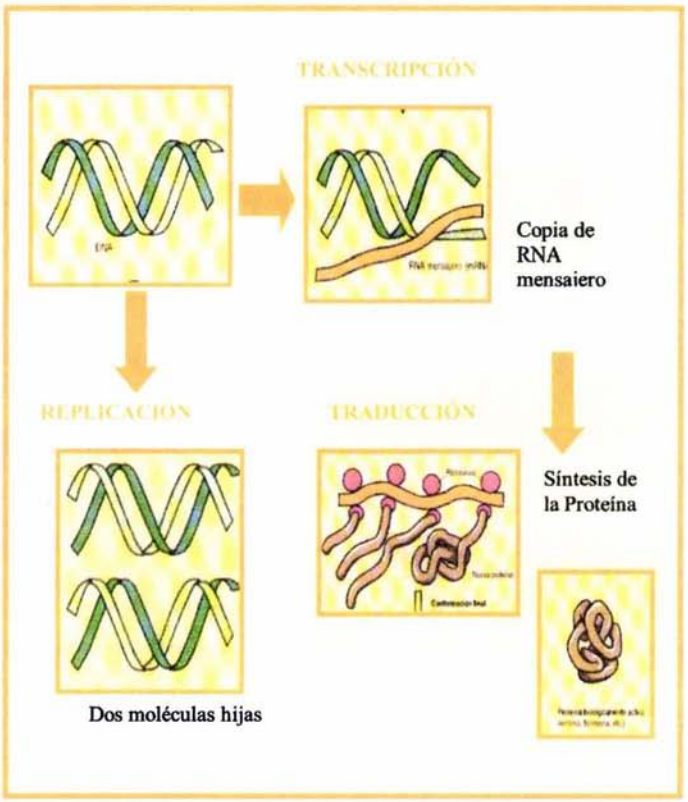


Fig 2. Procesos que dirigen la conservación y transmisión genética

Existen tres clases principales de RNA. El RNA mensajero (mRNA) que es la copia de un gen o de un conjunto de genes. El RNA de transferencia (tRNA) que es un RNA con una estructura secundaria específica, involucrado en la transferencia de aminoácidos a la cadena polipeptídica en crecimiento durante la síntesis proteica. Y el RNA ribosómico (rRNA) que forma parte de los ribosomas [5].

2.2 La regulación Genética en bacterias

Las bacterias son capaces de coordinar la expresión de varios genes agrupándolos de forma contigua en un segmento de DNA. Cada segmento de DNA que se transcribe, tiene una secuencia que es reconocida por la RNA polimerasa (RNApol), llamada promotor, que es donde se inicia la transcripción. Al grupo de genes, el promotor y las secuencias adicionales que funcionan conjuntamente en la regulación se le llama operón.

La regulación genética puede darse al nivel de la transcripción o al nivel de la traducción, ya sea activando al gen (control positivo) o apagándolo (control negativo). A si mismo la degradación del mRNA, la estabilidad de la proteína, la localización de la proteína, las interacciones proteína-proteína y la función de la proteína, están regulando a nivel postraduccional (Tabla 1).

Tabla 1
Mecanismos de Regulación genética

1) Nivel transcripcional
   - a) A nivel del inicio de la transcripción
     - i) Sustitución del factor σ de la RNApol
     - ii) Por interacción de proteínas reguladoras sobre secuencias de DNA cercanas al promotor
   - b) Terminación prematura de la transcripción: Por atenuación de la transcripción.
     - i) Terminación intrínseca
     - ii) Terminación por un factor dependiente

2) Nivel traduccional.
   - a) Regulación por la eficiencia de unión del ribosoma al mRNA
   - b) Regulación por interferencia con el sitio S-D.
     - i) Por intervención de proteínas
     - ii) Por intervención de RNAs interferentes
     - iii) Por la estructura secundaria del transcrito

3) Nivel postraduccional
   Degradación del mRNA, la estabilidad de la proteína, la localización de la proteína, las interacciones proteína-proteína y la función de la proteína

Regulación transcripcional

La **regulación por sustitución del factor σ de la RNApol** se da a nivel del reconocimiento de la secuencia promotora. La RNApol es una enzima compuesta por seis subunidades (α_2, β, β', ω y σ), de las cuales la subunidad σ es variable, por lo que el tipo de subunidad σ, reconoce una secuencia especifica de promotor y coordina la expresión de un conjunto de genes [6].

La mayoría de los procesos de regulación están dados por proteínas. En el caso de la **regulación transcripcional por proteínas,** las proteínas reconocen secuencias especificas de DNA y determinan cuales de los miles de genes de una célula se han de transcribir y cuales no. Es necesario que la proteína este en su estado activo para que sea capas de interaccionar con el DNA y para que la proteína este activa se necesita la unión de un cofactor a la misma (generalmente metabolitos relacionados a la función del gen que regula). Algunos ejemplos de proteínas de regulación génica son: proteínas con motivos hélice-giro-hélice, proteínas con motivos de dedos de Zinc, proteínas con motivos de laminas β de 2 hebras, proteínas con motivos de cremallera de leucina y proteínas con motivo de hélice-bucle-hélice.

El "operón de lactosa (*lac*)" de *Escherichia coli* es uno de los operones más estudiados en cuestión de regulación, este operón esta sujeto tanto a regulación negativa por la acción del represor Lac en presencia de Lactosa, como a regulación positiva por la CRP en presencia de Glucosa. El represor Lac es una proteína tetramérica de monómeros idénticos que esta codificada por el primer gen del operón *lac*. En ausencia de Lactosa el operón Lac se une al operador principal y a un pseudo-operador formando un lazo en el DNA que envuelve al represor y reprime la transcripción de los genes de lactosa. En presencia de lactosa se induce al operón *lac*, mediante la unión de una molécula (alolactosa) que se une en un sitio especifico del represor Lac, que provoca un cambio conformacional en este y la subsecuente disociación del represor Lac. Otro mecanismo de regulación por proteínas es la "represión por catabolito", que en el caso del operón *lac*, impide la expresión de los genes para el catabolismo de Lactosa, Arábidosa y otros azucares en presencia de Glucosa. En esta regulación interviene una proteína receptora de cAMP, o CRP, que es un homodimero con sitios de unión al DNA (motivo hélice-giro-helice) y a cAMP. A bajas concentraciones de glucosa hay un aumento de cAMP que favorece la unión de este a CRP, favoreciendo que CRP se una a un sitio cerca del promotor *lac* e induce la transcripción ya que CRP interacciona directamente con la subunidad α de la RNA polimerasa [5].

Otro ejemplo de regulación mediada por proteínas es el caso del "operón de la Arabinosa (*ara*)" de *Escherichia coli*, donde la proteína reguladora AraC ejerce un control tanto positivo como negativo. En este caso la unión de una molécula señal altera la conformación de AraC desde su forma represora, que se une a una secuencia reguladora de DNA a su forma activadora que se une a varias secuencias de DNA. Cuando hay mucha Glucosa y poca Arabinosa la proteína dimérica AraC se une tanto al operador $araO_2$ y al inductor *araI*, formando un lazo de DNA que reprime la transcripción de los genes de araBAD. Cuando hay poca Glucosa y Arabinosa a niveles

altos, el complejo CRP-cAMP actúa sobre un sitio de unión adyacente al Inductor *araI* de forma similar a la regulación en el operón *lac,* por otro lado la Arabinosa se une a AraC y altera su conformación provocando que los lazos de los homodímeros se abran y ahora la proteína AraC se une a ambos medios sitios de *araI* y se convierte en un activador, actuando conjuntamente con el complejo CRP-cAMP [5].

El complejo CRP-cAMP esta implicado en regulación coordinada de muchos operones, principalmente para aquellos que codifican enzimas del metabolismo de azucares secundarios.

En la **regulación por atenuación de la transcripción** el líder del transcrito (región entre el inicio de la transcripción y el primer gen estructural del DNA) puede adoptar una estructura secundaria correspondiente a un terminador transcripcional y otra estructura mutuamente excluyente al terminador, antiterminador. Esta regulación puede darse por dos mecanismos, terminación intrínseca y terminación por un factor dependiente [6].

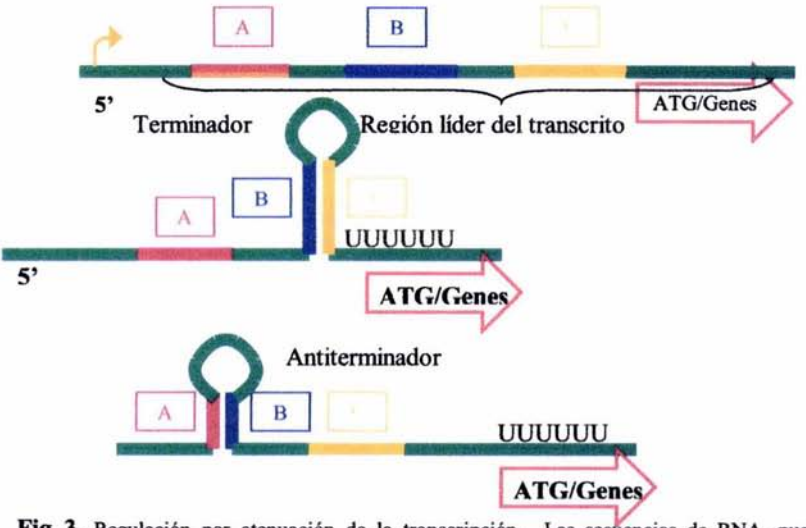


Fig 3. Regulación por atenuación de la transcripción. Las secuencias de RNA, pueden formar estructuras alternativas, de acuerdo a la coplementariedad de sus secuencias. Por ejemplo, B es complementaria tanto a A como a C, cuando B-C se complementan, se forma un terminador transcripcional, y cuando A-B complementan se evita la formación del terminador, por lo que se forma un antiterminador

Terminación intrínseca En este caso al sintetizarse un segmento del transcrito por la RNApol, se forma una orquilla seguida por una serie de residuos de U. Esta horquilla inicialmente sirve como una pausa transcripcional, que provoca que la RNApol termine la transcripción y libere tanto el transcrito como el templado del DNA [6]. (Fig 3). Algunos ejemplos de atenuación transcripcional por terminación intrínseca son:

- "Atenuación transcripcional dirigida por la traducción", en este caso el líder del transcrito tiene una región que codifica para un pequeño péptido, el cual contiene varios

codones iguales, por lo que la escases del tRNA correspondiente al codón provoca un paro del ribosoma, permitiendo que se forme el antiterminador y que se lleve a cabo la transcripción. Un ejemplo de este tipo es el operón His de *Salmonella Typhimurium.*

- "Atenuación de la transcripción dirigida por un tRNA", en este caso el líder del mRNA es capaz de censar un tRNA especifico, mediante interacciones entre el tRNA no cargado y el líder del transcrito, que estabiliza una estructura de antiterminador favoreciendo la transcripción del gen. Un ejemplo de este tipo es el gen *tyr S* de *Basillus subtilis.*
- "Atenuación de la transcripción dirigida por una proteína" algunas proteínas se unen al RNA alterando la estructura del líder del RNA, ya sea promoviendo o previniendo la formación de un terminador transcripcional. Un ejemplo de este caso es la proteína TRAP (trp RNA-binding Attenuation Protein) de *Basillus subtilis* que cuando esta activada por el triptofano interviene en la formación del terminador o como la proteína BglG de *Escherichia coli* que se une como dimero al líder del RNA cuando esta fosforilada y estabiliza una estructura de antiterminador en el operon *bgl* .

Terminación por un factor dependiente En este caso la regulación se da por una modificación de la maquinaria de transcripción que provoca la prematura terminación del transcrito. Este sistema utiliza proteínas reguladoras que se unen a sitios cercanos al inicio de la transcripción, interacciona con la RNApol y altera la respuesta a señales de terminación transcripcional [6]. Algunos ejemplos de atenuación transcripcional por un factor dependiente son:

- "Terminación dependiente de proteínas Rho", esta regulación requiere de la unión del hexamero Rho a sitios en la región rio arriba del transcrito lo cual interacciona con una RNApol pausada y causa la terminación transcripcional.
- "Antiterminación mediada por proteínas N", en este caso las proteínas N forman un complejo que se une a sitios en el inicio del transcrito e interactúa con la RNApol y la modifica en una forma resistente a los terminadores transcripcionales, un ejemplo de esta regulación es el bacteriofago λ, donde la unión de una proteína N al transcrito, en las primeras etapas de infección inicia la formación de un complejo (NusA, NusB, NusG, proteina ribosomal) resistente a la terminación.
- "Antiterminación dirigida por la traducción", en este caso el mecanismo de antiterminación se da al bloquear el acceso de la proteína Rho al transcrito debido al inicio de la traducción de un segmento del transcrito. Un ejemplo de esto es el operon tna de *E. coli*, en donde el primer gen del operon de tna, *tnaC* codifica un peptido líder que contiene un residuo de triptofano, cuando no hay triptofano disponible, se frena la traducción y se libera el ribosoma de *tnaC,* lo cual permite que el factor Rho se una al transcrito y favorezca la terminación transcripcional. En presencia de triptofano la continuidad de la traducción bloquea la unión de Rho, permitiendo que se finalice la transcripción

Asi mismo la elongación de la transcripción es un aspecto muy relacionado al control de la expresión genética. Tambien hay varios ejemplos donde la deción regulatoria depende de la traducción de un fragmento de lectura abierto (uORF) y por ultimo la regulación de la expresión genética por interacciones RNA-RNA, tanto en el caso de pequeños RNAs como en el caso de RNAs doble sentido (dsRNA) [6].

Regulación traduccional

En el caso de la regulación traduccional, el primer nivel de regulación se da por la "eficiencia de unión del ribosoma" a la secuencia Shine-Dalgarno (SD) del mRNA. Por ejemplo en *Bacillus subtilis* el rango en el que el ribosoma se asocia a la secuencia SD, va de –12 a –22 kcal/mol, estos valores influencian directamente los niveles de proteínas individuales en la célula [7].

Por otro lado la unión del ribosoma al transcrito tambien puede ser afectada por modificaciones de otras proteínas al ribosoma un ejemplo es la proteína RelA o factor de respuesta severa en *E. coli*, el cual se une al ribosoma, en respuesta a una carencia de aminoácidos que provoca la unión de un tRNA descargado al sitio aminoacil del ribosoma, la unión de RelA cataliza la síntesis de pppGpp. Por acción de una fosfohidrolasa que elimina un fosfato, nos da ppGpp. La señal ppGpp reduce la transcripción de algunos genes e incrementa la de otros, por la unión a la subunidad β de la polimerasa y la alteración de su especificidad [5].

Por otro lado la **regulación por interferencia con la secuencia SD** se da por proteínas reguladoras, RNA antiparalelo o por la estructura secundaria del transcrito (caso de los "Riboswitches") que pueden bloquear la secuencia SD impide la unión del ribosoma al mRNA [6].

La regulación de la síntesis de proteínas ribosomales es un ejemplo de "regulaciòn por intervenciòn de proteínas reguladoras". Las proteínas ribosomales, se agrupan en operones (que van de 1 a 11 genes por operon); estos operones se regulan principalmente mediante un mecanismo de retroalimentación traduccional, donde una de las proteínas de cada operon funciona como represor de la traducción, la cual se une al transcrito en un sitio cercano al inicio de la traducción y bloquea la traducción de todos los genes del operón [5].

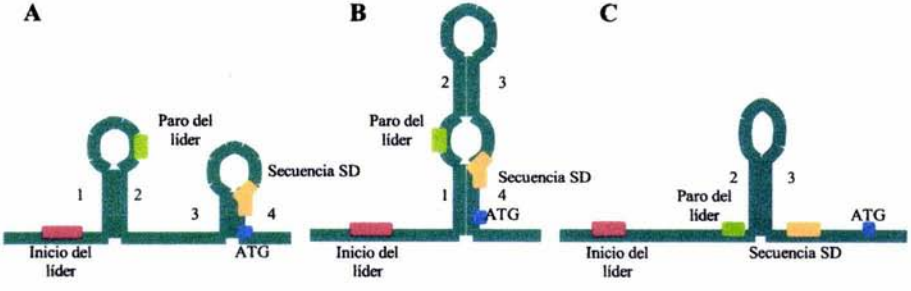


Fig 4. Regulación por atenuación traduccional. Estructuras secundarias alternativas involucradas en la atenuación traducional del gen *ermC*

Un ejemplo de regulación por la "estructura secundaria del transcrito" es el gen *ermC*, de *Bacillus subtilis* el cual codifica para 29,000-DA metilasa que confiere resistencia a la eritromicina. *ermC* se transcribe constitutivamente, en ausencia de eritromicina se transcribe poca metilasa y en presencia de eritromicina aumenta su

producción. El gen de metilasa esta precedido por un péptido líder de 19 aminoácidos. En esta región líder se forman tres estructuras secundarias alternativas. Dos de estas estructuras (Fig 4A y B) son, inhibidoras de la síntesis de metilasa debido a la inaccesibilidad de la secuencia SD del gen y del codon de inicio del gen. Una tercera estructura secundaria es activa en la síntesis de metilas (Fig 4C). Se a propuesto que la traducción del líder en presencia de eritromicina puede provocar un paro en el ribosoma, resultando en la desetabilisación de la estructura formada por las secuencias complementareas 1-4, 3-4 liberando la secuencia SD y permitiendo que se lleve a cabo la traducción [7].

2.3 La comparación genómica como una herramienta para estudiar la regulación genómica

Con la secuenciación de múltiples genomas, la Biología Computacional está entrando a una nueva era, donde la disposición de secuencias genómicas, nos permite estudiar las relaciones que hay entre los distintos organismos.

Una de las principales razones por la que se usa a la genómica comparativa para buscar secuencias de regulación es que tanto los genes homólogos como las secuencias de regulación son elementos relativamente conservados. Por otro lado las regiones no codificantes tienden a mostrar distintos grados de variación por lo que las secuencias conservadas en regiones no codificantes, pueden estar relacionadas a la regulación y expresión genética, también pueden estar relacionadas a mantener la organización estructural del genoma e incluso a funciones que aun no se han descubierto.

La comparación genómica se basa mucho en el alineamiento de secuencias (tanto secuencias nucleotidicas, como secuencias de aminoácidos). En la alineación de la secuencia se asigna un valor a cada posible alineamiento (típicamente, la suma de la similitud / los valores de identidad para cada residuo del alineamiento, menos una cantidad dependiente de la introducción de "gaps") [8].

Algunos de los programas desarrollados para el alineamiento múltiple de secuencias son: CLUSTALX [9], DIALIGN, ASSIRC, MUMmer, PipMaker/BlastZ, GLASS, WABA, LSH-ALL-PAIRS, Vmatch y MGA, entre otros [8]. Muchos de estos programas se basan en el algoritmo de Dumas and Ninio (1982) [10].

En el presente trabajo se utilizaron los programas MEME y MAST [11, 12]. MEME es una herramienta para identificar motivos conservados en grupos de secuencias, donde un motivo es un patrón de nucleótidos o de aminoácidos que se repite en un grupo de secuencias de DNA o de proteínas relacionadas. El programa MEME, obtiene motivos mediante matrices de posición. MAST es el complemento de MEME, es un programa que busca motivos obtenidos por MEME en bases de datos.

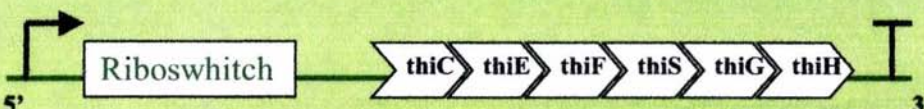
2.4 Los Riboswitches

Desde hace mas de 4 años, se han reportado secuencias de RNA, capaces de reconocer un metabolito especifico y regular la expresión de los genes relacionados a su biosíntesis y transporte. Estas secuencias que ahora se conocen como "Riboswitches", fueron inicialmente propuestas por Miranda-Rios, J. *et al.* (2001) en "A conserved RNA structure (*thi* box) is involved in regulation of thiamin biosynthetic gene expression in bacteria" donde especulan sobre la posible función del líder del operón *thi*COGE y proponen una regulación a nivel traduccional que involucra el secuestro de la secuencia SD por la estructura secundaria del líder.

Los "Riboswitches", tienen características muy particulares, algunas de estas están resumidas en la Tabla 2.

Tabla 2
"Riboswitches"

Los "Riboswitches" son elementos de regulación postranscripcional que se encuentran en la región 5' no traducida de algunos mRNA. Las características principales de estos son:



Operón de *E. coli*, que codifica para genes involucrados en la biosíntesis de tiamina

1. El mRNA forma un sitio de unión altamente selectivo a un metabolito especifico (moléculas efectoras como: vitaminas y cofactores), sin intervención de proteínas.
2. La unión del metabolito provoca un reconocimiento allostérico de la estructura del mRNA que lo estabiliza y provoca a su vez alteraciones en la expresión genética
3. El mRNA toma estructuras alternativas dependiendo de la unión del metabolito
   a. En presencia del metabolito se forma una estructura de represión genética
   b. En ausencia del metabolito se forma una estructura de antirepresión genética
4. La regulación se da en genes de la biosíntesis del metabolito

En los últimos años se han encontrado varios casos de regulación por riboswitches, basados tanto en análisis bioquímicos y moleculares, como en análisis genómicos. El riboswitch de Vitamina B_{12} [17], el riboswitch de Tiamina Pirofosfato (TPP) [20], el riboswitch del Mononucleotido de Flavina (FMN) [23], el riboswitch de S-adenosil-Metionina (SAM) [26], el riboswitch de Lisina [32], el riboswitch de Guanina [33] (Fig 5).

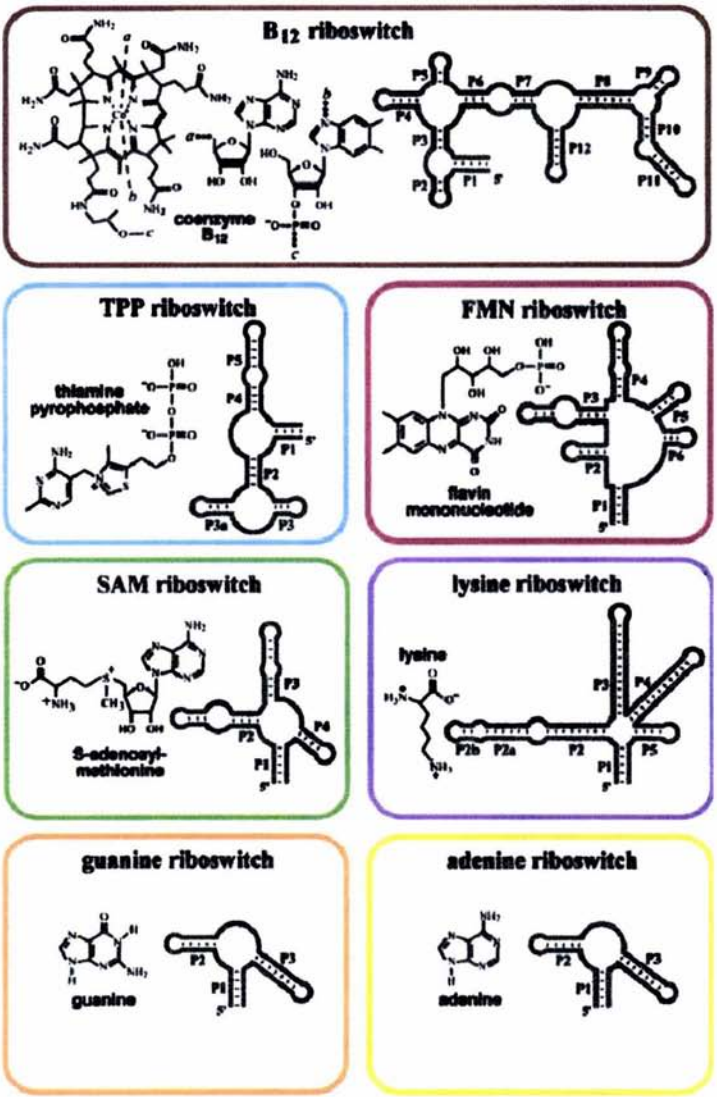


Fig 5 Riboswitches. Esquemas de 7 riboswitches reportados y el metabolito que censan [33]

2.5 Mecanismos propuestos en la regulación por Riboswitches

Las dos estrategias de regulación por riboswitches son la regulación por la formación de un terminador transcripcional y la regulación por la formación de un inhibidor traduccional y el mecanismo general de acción consiste en la unión de un metabolito especifico a la secuencia líder del RNA, que provoca un cambio conformacional en su estructura y la formación subsecuente de un anti-antiterminador o de un anti-antisecuestrador.

Regulación por un terminador transcripcional. En ausencia del metabolito parte de la secuencia que forma al termiador, se asocia a una secuencia complementaria (antiterminador), permitiendo que se termine la transcripción y en presencia del metabolito se favorece la unión del antiterminador con una secuencia complementaria (anti-antiterminador), la unión de antiterminador con el anti-antiteminador permite la formación del terminador y provoca la disociación del complejo de transcripción [29] (Fig 6A).

Regulación por un inhibidor traduccional. En ausencia del metabolito el secuestrador de la secuencia de unión del ribosoma (SD) se encuentra asociado a una secuencia complementarea (anti-antisecuestrador), dejando libre la secuencia SD para que se lleve a cabo la traducción y en presencia del metabolito, la secuencia SD se mantiene unida a la secuencia del secuestrador (anti-SD) impidiendo que el ribosoma se ancle al RNA e impidiendo la traducción [29] (Fig 6B).

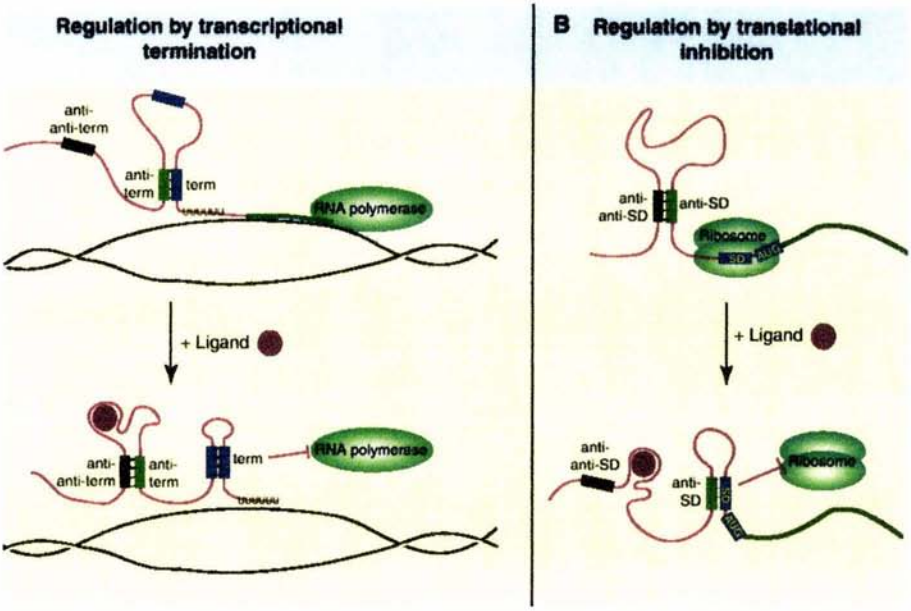
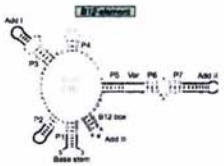
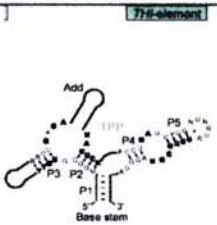
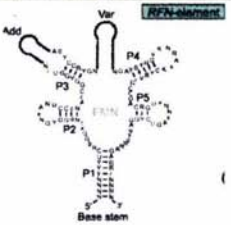


Fig 6 Mecanismos de regulación por riboswitches [29]

2.6 Características de los distintos Riboswitches

Las bacterias en su mayoría, tienen la capacidad de sintetizar todas las moléculas que requieren para su crecimiento a partir de compuestos simples y son capaces también de utilizar las fuentes ambientales (concentración de un metabolito) para reprimir la síntesis de los genes que requiere para sintetizarlas. Los Riboswitches, son una estrategia muy bien representada en Eubacterias, Arqueobacterias e incluso en algunos Eucariotes que permite regular la síntesis de algunos metabolitos mediante la regulación de sus genes. Algunas de las características particulares de 6 "Riboswitches"están resumidas en la Tabla 3.

Tabla 3
Características de los Riboswitches

Riboswitch	B_{12}	TPP	RFN
Estructura propuesta			
Precursor	Cobalamina (Vitamina B_{12})	Tiamina (vitamina B_1)	Riboflavina (vitamina B_2)
Molécula efectora	5´-deoxi-5´ adenosil Cobalamina (Ado-Cbl)[1]	Tiamina pirofosfato (TPP)[2]	Mononucleotido de Flavina (FMN)[3] Flavin Adenin Dinucleotido (FAD)[3]
Genes regulados	*btuB* (transportador de cobalamina) en *Escherichia. coli y Salmonella typhimurium* operón *cob* (síntesis de cobalamina) en *Salmonella typhimurium*	operón *thiCOGE* (síntesis de tiamina) en *Rizobium etli* operón *tenA* (síntesis de tiamina) en *Bacillus subtilis* *thiM* (kinasa de tiazol) en *Escherichia. coli*	operón *ribGBAH* (síntesis de riboflavina) en *Bacillus subtilis* *ypaA* (transportador de riboflavina) en *Bacillus subtilis*
Distribución taxonómica y regulación	**Eubacterias** **Gram-negativas, Cianobacteria, actinobacteria y VFB group** Regulación traduccional (*btuB*, operón *cob*) Regulación a nivel traduccional y	**Eubacterias** **Gram-negativas** Regulación al nivel de la traducción (*thiC*, *thiM*) **Gram-positivas** Regulación al nivel de la transcripción (*thenA*) **Arqueobacterias**	**Eubacterias** **Gram-negativas** Regulación al nivel de la traducción (*ribB, ribH2, ribD, ribE*) **Gram-positivas** Regulación al nivel de la trascripción (operón *rib* de *Bacillus/Clostridium group*)

	transcripcional (*btuB E. coli* y *Salmonella*) **Gram-positivas** Regulación a nivel de la transcripción (*Bacillus/Clostridium group*)	**Thermotogales** (transportadores de tiamina) **Eucariotes** **Plantas y Hongos** Regulación al nivel de procesamiento y estabilidad del RNA (*thiC*)[I]	Regulación a nivel transcripcional y traduccional (*ypaA*)
K_D[II] Constante de disociación	~300nM en *btuB* de *Escherichia coli*	100nM en *thiC* de *Escherichia coli* 600nM en *thiM* de *Escherichia coli* ~50nM en *thiC* de *A. thailiana*	~10nM en *ribD* y *ypaA* de *Bacillus subtilis*
Referencias	[13-17]	[18-21]	[22-25]

B_{12} element

La B_{12} box es una secuencia de 17nt que se encuentra conservada en la región líder de genes relacionados a la biosíntesis de la vitamina B_{12}, la cual reconoce a la Adenosyl Cobalamina (Ado-Cbl)[I] y reprime la expresión del gen correspondiente [13]. En un estudio posterior, Vitreschak reporto una firma ampliada de la B_{12} box, a la cual designó como: "B_{12} element", en este trabajo se reportarón 200 elementos en 66 genomas bacterianos, ampliamente distribuidos en los distintos taxa de Eubacterias (Tabla 3), sin embargo no se han encontrado en Arqueobacterias ni en Eucariontes [17].

Se a observado que la regulación del gen *btuB* de *Escherichia coli - Salmonella typhimurium* (transportador membranal CN-Cbl de la vitamina B_{12}) y el operón *cob* de *Salmonella typhimurium* (biosíntesis de Cobalamina), está dada por el elemento B_{12} que en presencia de Ado-cbl reprimen su síntesis su secuencia líder [14].

La regulación a nivel del gen *btuB* se da a nivel de la traducción por la formación de una estructura de RNA que secuestra el sitio de unión al ribosoma [13]. De forma general el elemento B_{12} es estabilizado por Ado-Cbl, el cual interacciona con la región de antisecuestrador/antiterminador, promoviendo la formación del

[I] Plantas: *Arabidopsis thaliana* (*thiC*), *Oriza sativa* (posible *thiC*), *Poa secunda* (posible *thiC*), Hongos: *Aspergillus orizae* (*thiA*/síntesis de tiazol), *Neurospora crassa* (*nmt-1*/proteína de la biosíntesis de tiamina), *Fusarium oxysporum* (*sti35*), *Fusarium solani* (*sti35*) [28]

[II] La KD refleja la cantidad de molécula efectora (metabolito), necesaria para convertir la mitad de RNA en su estructura alternativa

[I] La vitamina B_{12} o Cobalamina es un cofactor esencial en enzimas que catalizan varias transmetilaciones y rearreglos en reacciones. La Adenosyl Cobalamia es un derivado de la vitamina B_{12}

secuestrador/terminador. En ausencia de Ado-Cbl, la conformación del elemento B12 es inestable, por lo que se favorece la formación de la estructura de antisecuestrador/antiterminador impidiendo que la secuencia SD sea secuestrada o que se forme el terminador transcripcional y permitiendo el inicio de la traducción o que continué la transcripción, respectivamente. También se ha demostrado experimentalmente la existencia de un elemento regulador adicional ("enhancer"), localizado entre el elemento B_{12} y el secuestrador en los genes *cbiA* de *Salmonella typhimurium*. En este caso en ausencia de Cobalamina el "enhancer" interactúa con el elemento B_{12}, liberando el inicio de la traducción y en presencia de Cobalamina, el "enhancer" interacciona con el elemento B_{12} favoreciendo la formación del secuestrador y reprimiendo la traducción [15].

THI element

Algunos genes de biosíntesis y transporte de tiamina, contiene una secuencia altamente conservada en su región líder. El primer elemento de regulación que se encontró en estas secuencias se designo como "*thi* box" [19] y fue identificado en un grupo de genes (*thi*COGE) involucrados en la biosíntesis de la tiamina de *Rizobium etli* [18]. Un estudio posterior reporto una firma más grande y se designo a esta secuencia como "THI-element" [20].

La particularidad de este líder es que es reprimido por tiamina en un evento post transcripcional en el que la estructura secundaria de la *thi* box secuestra al sitio de unión al ribosoma, permitiendo la formación de un terminador e impidiendo la traducción del operón *thi*COGE [19].

Las tres regiones importantes para la regulación son: a) la secuencia de *thi* box, que reconoce la molécula efectora (TPP[2]), b) la región complementaria a la secuencia SD, c) y la secuencia que forma el terminador [19]. El modelo propuesto para la regulación de Tiamina esta basado en competencia entre estructuras alternas de RNA. En bacterias Gram-positivas el líder del transcrito esta organizado como un sistema de terminador-antiterminador-anti-anti-terminador donde la unión de TPP a la *thi* box, estabiliza la estructura del terminador y evita la transcripción. Sin TPP, la estructura de anti-anti-terminador es remplazada por la estructura más estable de anti-terminación y permite que la transcripción se lleve a cabo. En bacterias Gram-negativas, la unión del TPP a la *thi* box, estabiliza la estructura de secuestrador SD y reprime el inicio de la traducción, en condiciones de des represión se favorece la formación de un anti-antiSD que libera la secuencia SD y permite el inicio de la traducción [20].

El "riboswhitch" de TPP esta muy distribuido entre Eubacterias y Arqueobacterias e incluso se a encontrado también en Eucariontes: en plantas esta localizado antes de un poli A, lo cual sugiere que la unión del metabolito regula el procesamiento y estabilidad

[2] La Tiamina (Vitamina B1), es un componente esencial en el metabolismo ya que es el precursor del Tiamin Pirofosfato (TPP), que es un cofactor de enzimas esenciales en el metabolismo del Carbón. El Tiamin Pirofosfato, esta formado un motivo de pirimidina, Hydroximetilpirimidina (HMP) y un motivo de tiazol, Hydroxietiltiazole (HET)

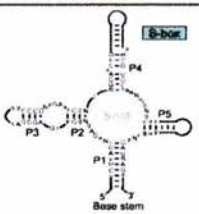
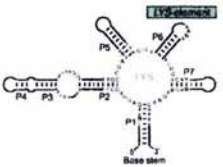
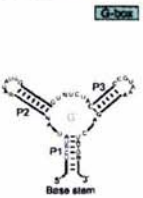
del mRNA y en el caso de hongos se localiza en un intrón, lo cual sugiere que el "splicing" del RNA esta guiado por la unión del metabolito al pre-mRNA, por lo que interviene en el procesamiento del mismo [28].

RFN element

El elemento RFN esta altamente conservado en la región rió arriba de genes de la biosíntesis de Riboflavina [22]. La Riboflavina (vitamina B2) es el precursor de las coenzimas: Flavin Mononucleotido (FMN)[3] y Flavin Dinucleotido (FAD)[3]. Los nucleotidos FMN y FAD son capaces de unirse al elemento RFN y promover la formación de un terminador en bacterias Gram-positivas y de un secuestrador SD, en bacterias Gram-negativas [23]. El líder del operón *rib* y *ypaA* (transportador de riboflavina) de *Bacillus subtilis* representa un buen ejemplo de la regulación por el elemento RFN [24].

En el caso de las bacterias Gram-positivas el elemento RFN, toma la conformación de anti-antiterminador en condiciones de represión (exceso de FMN) y previene la formación del antiterminador, permitiendo que se forme el terminador y no continúe la transcripción. Sin FMN, la estructura del elemento RFN es remplazada por la conformación de antiterminador, que evita que se forme el terminador y permite que se llevé a cabo la transcripción. En el caso de bacterias Gram-negativas el elemento RFN, previene la formación del antisecuestrador y favorece la formación del secuestrador de la secuencia SD en condiciones de represión, evitando que se el transcrito se traduzca y en condiciones de des represión el elemento RFN es remplazado por el antisecuestrador que libera la SD y permite el inicio de la traducción [23].

Tabla 3
-continuación-
Características de los Riboswitches

Riboswitch	SAM	LYS	G
Estructura propuesta			
Precursor	Metionina (Met)	Lisina (Lys)	Purinas (Adenina, Guanina)
Molécula efectora	S-adenosil-L-Metionina (SAM)[4]	Lisina (Lys)[5]	Hypoxantina y Guanina o Adenina[6]
Genes regulados	*yitJ* (síntesis de metionina)	*lysC* (monofuncional	*xpt-pbuX* (xanthin

[3] La Riboflavina (Vitamina B2), es también un componente esencial del metabolismo básico ya que es el precursor de la coenzima Flavin Adenin Dinucleotido (FAD) y el Flavin Mononucleotido (FMD), que sirven como grupos prostéticos para muchas oxidorreductasas.

	en *Bacillus subtilis* operón *metIC*	aspartocinasa) *lysA* (deamino descarboxilasa)	fosforibosiltransferasa y transportador de Xanthin) en *Bacillus subtilis* *purEKBCSQLFMNHD* (síntesis de Inosine Monofosfato IMP)
Distribución taxonómica y regulación	**Eubacterias** **Gram-negative** Regulación a nivel traduccional **Gram-positivas** Regulación a nivel transcripcional (*yitJ*)	**Eubacteria** **Gram-negativas** Regulación a nivel traduccional **Gram-positivas** Regulación a nivel transcripcional (*lysC*)	**Eubacterias** **Gram-positivas** Regulación a nivel transcripcional
K_D Constante de disociación	~4 nM	-	~5 nM en *xpt* de *Bacillus sutilis*
Referencias	[26- 27]	[30-32]	[33]

SAM element

Se ha observado que muchos de los genes involucrados en el metabolismo del sulfuro, la biosíntesis de Cisteína (Cys) y Metionina (Met) y la biosíntesis de S-adenosyl-metionin (SAM) [4] están regulados por el elemento SAM.

Las secuencias líder de los genes regulados por el elemento SAM contienen un terminador transcripcional intrínseco, un antiterminador competitivo y la S-box, que es una secuencia muy conservada a través de todos los genes y que interviene en la formación de un anti-antiterminador. En el mecanismo de regulación, en condiciones de represión la unión de SAM a la S-box, estabiliza la formación de un anti-antiterminador, el cual a su vez permite la formación del terminador y la temprana terminación de la transcripción [36, ref 40].

Lys element

La mayoría de los genes de la síntesis de Lisina en *Escherichia coli*, son reprimidos por Lisina[5] donde se identifico un elemento de regulación en genes de biosíntesis y transporte de Lisina (elemento LYS).

El mecanismo de regulación para la Lisina esta basado en la competencia entre estructuras alternativas de RNA. En las condiciones de represión el elemento LYS estabiliza la región reguladora (terminador transcripcional/secuestrador SD), por la unión de la molécula efectora Lisina. En bacterias Gram-positivas promueve la formación del

[4] SAM es una coenzima esencial en todos los organismos. Es sintetizada directamente de Metionina por la sintasa SAM y sirve como una fuente de grupos metilo en la modificación de proteínas y ácidos nucleicos
[5] La Lisina es un aminoácido esencial y es producida de Asparto a Diaminopimelato (DAP) en la mayoría de las bacterias y plantas superiores

terminador y la prematura terminación del transcrito. En bacterias Gram-negativas se favorece la formación del secuestrador-SD que reprime el inicio de la traducción. En ausencia de Lisina, la estructura secundaria del líder toma una conformación energéticamente más favorable anti-terminador/anti-secuestrador-DS, permitiendo que se lleve a cabo la transcripción o la traducción [31].

G element

El "Riboswitch" de guanina o elemento G regula a un grupo de operones involucrados en la biosíntesis, ínter conversión y transporte de purinas en *Basillus Subtilis.*

El análisis de las secuencias líder de los genes regulados por el elemento G, revelaron una confirmación típica de terminador-antiterminador-anti-antiterminador, el segmento más conservado (G-box) es capas de reconocer a la guanina y en mucho menor afinidad a la Xanthina, Hipoxantina y Adenina[6] favoreciendo la formación de un anti-antiterminador y promoviendo la formación del terminador intrínseco [33].

2.5 Métodos experimentales en el estudio de los Riboswitches

La importancia de la conservación de secuencia y de estructura secundaria de los "Riboswitches" se probo experimentalmente mediante mutaciones en las cajas de regulación (B_{12}-box, *rfn*-box y *thi*-box) y se observo que las mutaciones evitan que se lleve a cabo la represión y se provoca una perdida de la represión [13-15, 19, 25-27, 33] a si mismo se estableció que las secuencias en los distintos genomas, tienen cambios complementarios, que conservan la estructura de regulación [15, 17, 19-23, 31-33].

Además de los estudios de mutagénesis que nos permiten modificar la secuencia de RNA y observar cambios directos en la regulación, se han realizado distintos estudios experimentales para estudiar la regulación por "Riboswitches", como las pruebas de "RNasa H"[A] y [25, 27, 32] pruebas "in-line RNA"[B] que permiten estudiar los cambios conformacionales en presencia y ausencia de la molécula efectora e incluso calcular la constante de disociación [16, 21, 24, 33] o como las pruebas de Inmunofluorecencia[C], que permiten colocalizar a la molécula efectora con el sitio de unión al RNA [25] entre otras.

---

[6] Las purinas: guanina, adenina son esenciales en la síntesis de ácidos nucleicos. Xanthina e Hypoxantina, son análogos de la guanina y precursores en la formación de la misma

[A] La prueba de RNasa H se basa en el uso de oligonucleotidos complementarios al RNA y ribonucleasas que rompen específicamente, híbridos de RNA-DNA.

[B] La prueba de in-line es una estrategia molecular que tiene como principio el rompimiento de una secuencia por ataque nucleofílico a través de la acción de ribonucleasas.

[C] La Inmunofluorecencia utiliza anticuerpos fluorescentes (Fluoroforos) dirigidos.

La alta selectividad de las secuencias de "Riboswitches" por sus moléculas efectoras se ha observado en distintos experimentos. Un ejemplo es el cambio conformacional, en el líder del gen *rib,* que se da por el reconocimiento de FMN y no de riboflavina, esto se demostró mediante pruebas de RNasa H [25]. El elemento B_{12}, Ado-cbl inhibe la unión de la subunidad 30s ribosomal a *btuB*, a diferencia de la Cobalamina que no tiene efecto regulador directo en el mismo [14]. También en el elemento SAM, se observó que la unión de SAM a la S-box induce la temprana terminación del transcrito, sin embargo análogos cercanos de éste, como la Metionina o el S-adenosil-homocistein no se unen al transcrito y no son capaces de afectar la transcripción [27].

En todos los sistemas de regulación la respuesta regulatoria requiere como mínimo, una media de sensibilidad a la concentración del producto. En el caso de los Riboswitches esta media es muy pequeña (del rango de nanomoles nM), lo cual refleja su eficiencia para censar los cambios en la concentración de la molécula efectora. Una forma de medir la sensibilidad del "Riboswhitch" por la molécula efectora es mediante el uso de pruebas "in-line" ya que mediante el análisis de los rompimientos espontáneos de uniones fosfodiester del RNA se puede establecer un patrón de rompimiento que puede utilizarse para definir la conformación estructural y las características funcionales de la unión del ligando. En el caso de RFN y su interacción con FMN, la presencia de FMN en concentraciones micromolares es suficiente para potenciar la formación del terminador in vitro [24]. Incluso se observó que el elemento THI encontrado en Eucariontes, correspondiente a un gen homologo de *thiC* en *A. thaliana,* tiene un KD de ~50nM, similar a la reportada en elementos THI de *E. coli* [28].

Es importante considerar que la regulación por "Riboswitches", involucra dos conformaciones alternativas de RNA.

A. Cuando el líder del transcrito esta asociado la molécula efectora, los metabolitos, Ado-Cbl, TPP, FMN o FAD, estabilizan la estructura de regulación: elemento B_{12}, THI, RFN, respectivamente, permitiendo la formación de un terminador o de un secuestrador SD
B. Cuando el líder del transcrito no esta asociado a la molécula efectora, se forma una estructura energéticamente más favorable y evita que el elemento de regulación se forme

Estos cambios conformacionales pueden estudiarse mediante pruebas "in-line" en presencia y ausencia del metabolito

3. Justificación, Objetivos e Hipótesis

3.1 Justificación

La regulación por riboswitches representa un tipo de regulación muy conservada en organismos filogenéticamente distantes, y dado el tamaño y la complejidad de las secuencias de riboswitches, es poco probable que hayan ocurrido numerosas reinvenciones de estos, por el contrario los blancos metabólicos (vitaminas, aminoácidos y cofactores) de estos switches genéticos han estado presentes desde hace mucho tiempo y suponen por tanto que la regulación por Riboswitches es un mecanismo con muchas posibilidades de presentarse en otras biosíntesis de vitaminas y cofactores.

A pesar de los muchos reportes de Riboswitches, la búsqueda de secuencias de regulación de este tipo, siguen siendo muy limitada, ya que en su mayoría los estudios se enfocan al análisis de riboswitches, de manera individual. A diferencia de los anteriores estudios mi trabajo propone una alternativa en la que la búsqueda de secuencias de regulación, es a partir de buscar los genes relacionados a una palabra clave lo cual permite abordar varias vías metabólicas (se aborda la búsqueda en principio de 12 vías metabólicas).

3.2 Objetivos

General

El objetivo de este estudio es desarrollar un método general que permita encontrar riboswitches

Particulares

- Comparar las secuencias intergénicas de operones involucrados en la biosíntesis de vitaminas, mediante el uso de las bases de datos genómicas, herramientas de programación y aplicaciones de genómica comparativa para encontrar secuencias conservadas con posible función regulatoria (riboswitches)
- Evaluar el método con los resultados reportados para los riboswitches conocidos
- Aplicar el método a otras vitaminas y cofactores

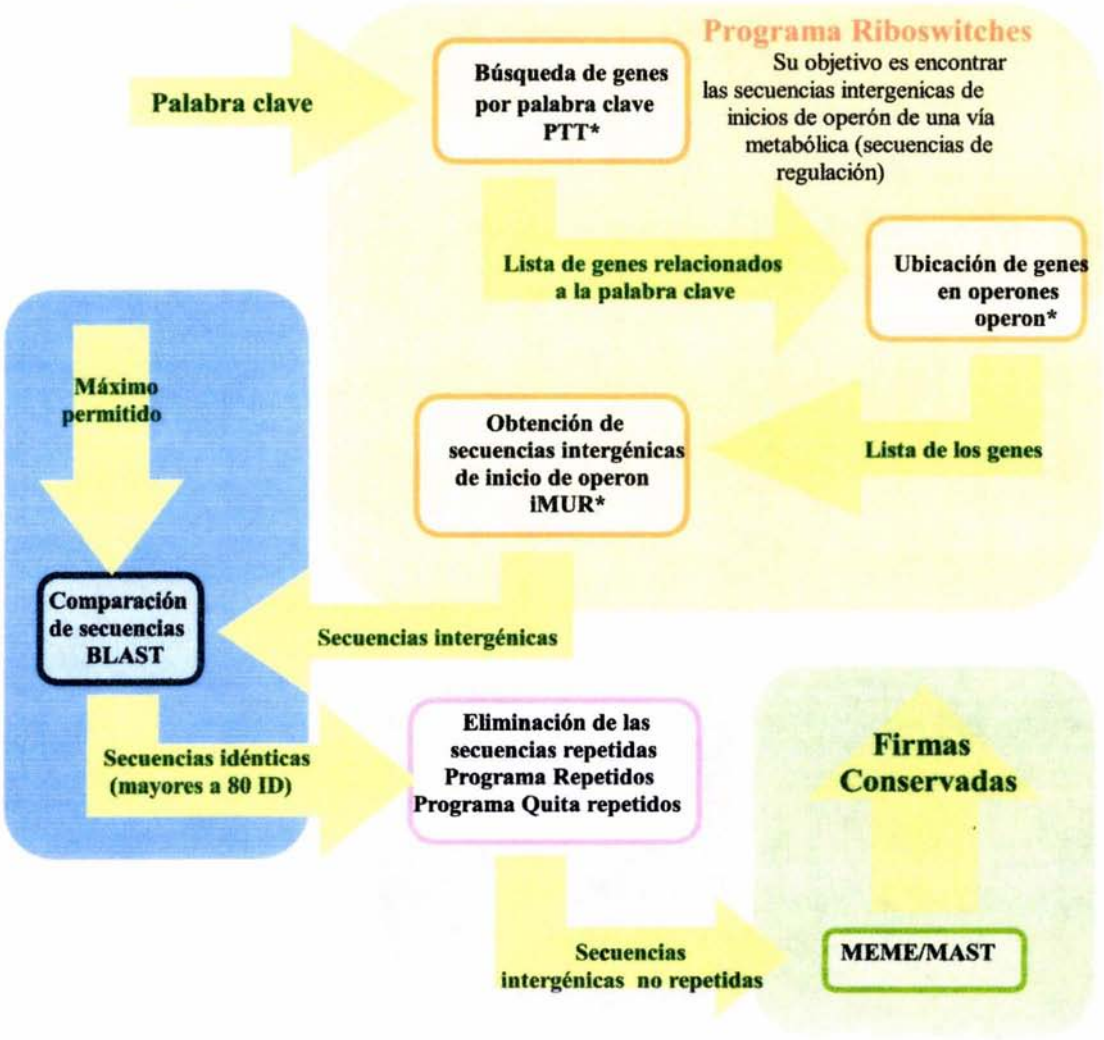
3.3 Hipótesis

Dado que los "riboswitches" son elementos muy conservados en la regulación postrascripcional de muchos genes y que sus blancos metabólicos son pequeños metabolitos (vitaminas), entonces es probable que haya "Riboswitches" en la biosíntesis de otras vitaminas y/o cofactores a los descritos a la fecha y dada su conservación, sea posible identificarlos

4. Metodología

De forma general para la búsqueda de las secuencias a comparar se generaron tres programas en lenguaje de programación PERL, "programa Riboswitches", "Programa Repetidos" y "Programa Quita Repetidos" y se compararon las secuencias con los programas MEME y MAST

Cuadro 1 Esquema general del método. La primera parte del método se enfoca a la obtención de las secuencias que se van a analizar, posteriormente se enfoca a eliminar secuencias repetidas de mi primer grupo y por ultimo a utilizar MEME/MAST, en la búsqueda de motivos conservados



Se tomó a la Tiamina como modelo y se compararon los resultados con los obtenidos en estudios de análisis comparativo para genes relacionados a la biosíntesis de la Tiamina [20]. Se utilizaron 139 genomas de bacterias completamente secuenciados (febrero 2004).

4.1 Búsqueda de genes de biosíntesis de vitaminas y cofactores

Considerando que las secuencias de regulación tipo "Riboswhitch", reconocen metabolitos pequeños, de tipo vitaminas. Y que los genes que son regulados por este tipo de secuencias, comúnmente están involucrados en la biosíntesis de la vitamina que reconocen. Se planteo utilizar palabras clave para buscar los genes de una vía metabólica común (por ejemplo la biosíntesis de Tiamina), en la búsqueda se utilizó la información del GenBank (base de datos de secuencias genéticas, NCBI) [45], organizada en una base de datos local, llamada PTT (Table of Translated Protein) (Fig 7).

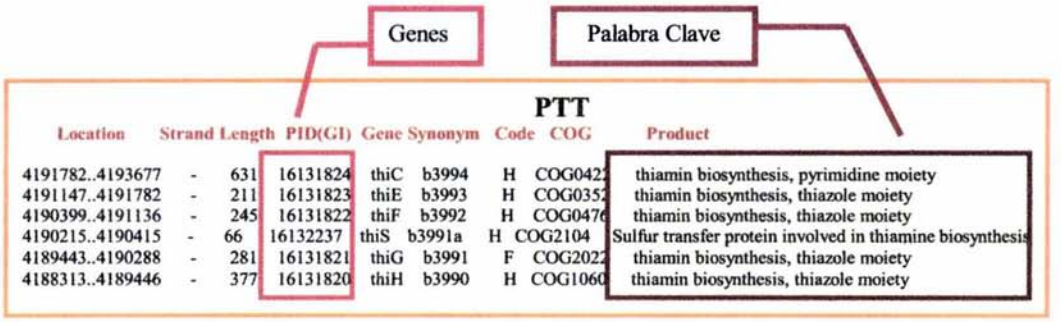


Fig 7. Base de datos PTT. La base de datos PTT contiene: la posición en la que está localizado el gen, la dirección en la que es leído, su tamaño, su número de identificación, el nombre del gen, número de identificación b, clasificación funcional de acuerdo a los COG (Cluster of Orthologous Group of proteins) [41] y para que codifica.

En base a esta información se desarrollo el "programa Riboswitches", que busca todos los genes relacionados a una palabra clave en la información reportada para el producto de cada gen. La búsqueda de palabras clave sobre la base de datos PTT, nos permite obtener los genes potencialmente relacionados a una función de acuerdo a la descripción reportada para cada uno de estos.

4.2 Ubicación de genes en operones y obtención del gen de inicio de operón

Después de obtener un grupo de genes relacionado a una misma vía metabólica, el "programa Riboswitches" ubica a los genes en operones, de acuerdo a la base de datos operón [43] (Fig 8).

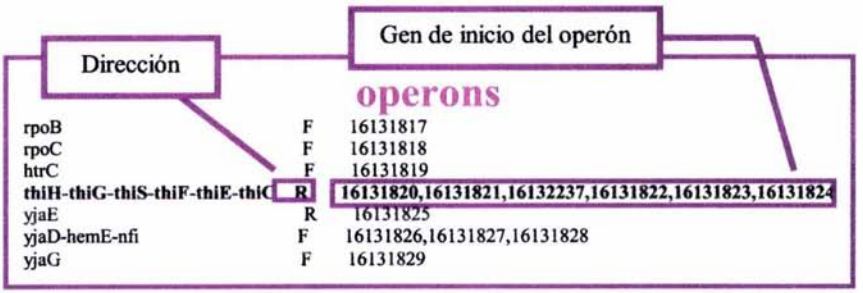


Fig 8. Base de datos de operones. La base de datos operón contiene: el grupo de genes que forman el operón, la dirección del operón y los Gi's (número de identificación).

La base de datos operón se generó en base a criterios de direccionalidad, es decir que solo aquellos genes que tienen una misma dirección, pertenecen a un operón; de distancia (los genes separados por más de un parámetro de distancia especifico, pertenecerían a un operón distinto) y de clase funcional, donde los genes relacionados funcionalmente, pertenecen a una misma clase funcional (operón) [39, 40].

Después de ubicar los genes en operones, el "programa Riboswitches" toma el gen del inicio de cada operón y finalmente se obtiene una lista de genes de inicio de operon relacionados a una misma via metabólica.

4.3 Obtención de secuencias intergénicas

Por último el "programa Riboswitches" busca la secuencia intergénica de cada uno de los genes de inicio de operón en la base de datos iMUR.

La base de datos iMUR contiene las secuencias intergénicas de los distintos genomas secuenciados. Donde una secuencia intergénica corresponde a la región entre el gen del inicio de operón y el gen más próximo. El tamaño de las regiones intergénicas es variable, pueden ir de 1 pb a más de 1000 pb (Fig 9).

Fig 9. Base de datos iMUR. Intergenic, Minimal Upstring Region

4.4 Comparación de secuencias, BLAST

Para evitar la sobrerrepresentación de secuencias redundantes de organismos filogenéticamente cercanos, se hizo una comparación de todas las secuencias intergénicas entre si mediante el programa BLAST

BLAST (Basic Local Alignment Search Tool) es un programa que encuentra regiones con un grado de similitud de secuencia. De forma muy general, el mecanismo de BLAST calcula el valor de un alineamiento en base al número y la posición de "matches" entre letras que corresponden a una posición en las secuencias comparadas, para localizar las posiciones que tienen más posibilidades de presentar un buen valor de similitud. BLAST utiliza puntos clave ("hot spots") que son regiones cortas que presentan un "match" exacto o casi exacto a partir de los cuales evalua el grado de similitud de la secuencia. BLAST presenta cinco variantes: blastn (secuencias nucleotidicas vs secuencias nucleotidicas), blastp (secuencias peptidicas vs secuencias proteicas), blastx (secuencias nucleotidicas vs secuencias proteicas), tblastn (secuencias peptidicas vs secuencias nucleotidicas) y tblastx (secuencia nucleotidica vs secuencia nucleotidica, considerando las 6 posibilidades en que un peptido puede estar sintetizado de acuerdo al los tripletes de nucleotidos).

En el estudio se utilizó el programa blastn, y un e-value de e^{-29} como valor de corte, capaz de eliminar secuencias redundantes de organismos tan cercanos como las cepas de *Escherichia coli* y las cepas de *Sallmonella*, que se sabe son genomas muy similares. Por ejemplo en el caso de la Tiamina se realizó un BLAST con 380 secuencias (iMUR) y se econtraron 144 secuencias redundantes que corresponden al 20% de la lista original de iMURs [44].

4.5 Eliminación de secuencias repetidas

Para eliminar las secuencias repetidas se utilizaron dos programas, el "programa Repetidos", que organiza las secuencias de salida del BLAST en grupos de repetidos como se observa en la Tabla 4 y el "programa Quita repetidos" que elimina las

secuencias repetidas, de la lista de secuencias intergénicas original, excepto la más grande.

En el ejemplo de Tiamina, las 144 secuencias redundantes, se agruparon en 55 grupos de repetidos, de estos grupos únicamente se tomo una secuencia (siempre la más grande), por lo que me quede con 55 secuencias representativas y 89 repetidas, que después fueron eliminadas de la lista original. De esta forma se obtiene una lista de secuencias intergénicas relacionadas metabolicamente, sin secuencias repetidas, que pueden ser comparadas para buscar pequeñas firmas conservadas.

Tabla 4
Grupos de secuencias repetidas

Cluster No. 1	33592342/B_pertussis 33602255/B_bronchiseptica
Cluster No. 2	15800146/E_coli_O157H7_EDL933 15829724/E_coli_O157H7 161284(
Cluster No. 3	18977902/P_furiosus 18977903/P_furiosus
Cluster No. 4	15674857/S_pyogenes 19745896/S_pyogenes_MGAS8232 28896217/S_
Cluster No. 5	15927293/S_aureus_N315 21283388/S_aureus_MW2
Cluster No. 6	15901320/S_pneumoniae_TIGR4 15903367/S_pneumoniae_R6
Cluster No. 7	15675116/S_pyogenes 19746086/S_pyogenes_MGAS8232 21910333/S_
Cluster No. 8	15925084/S_aureus_Mu50 15927669/S_aureus_N315 21283746/S_au
Cluster No. 9	33591548/B_pertussis 33598441/B_parapertussis 33603381/B_br
Cluster No. 10	15640093/V_cholerae 15804585/E_coli_O157H7_EDL933 15834171/
Cluster No. 11	13474823/M_loti 16265275/S_melliloti
Cluster No. 12	16761194/S_typhi 16765594/S_typhimurium_LT2 29141112/S_typh
Cluster No. 13	33595678/B_parapertussis 33600200/B_bronchiseptica
Cluster No. 14	21230174/X_campestris 21241523/X_citri
Cluster No. 15	16123342/Y_pestis_CO92 22124910/Y_pestis_KIM
Cluster No. 16	15676468/N_meningitidis_MC58 15793721/N_meningitidis_Z2491
Cluster No. 17	15800155/E_coli_O157H7_EDL933 15829733/E_coli_O157H7 161284
Cluster No. 18	15799752/E_coli_O157H7_EDL933 15829326/E_coli_O157H7 161280(
Cluster No. 19	15900614/S_pneumoniae_TIGR4 15902673/S_pneumoniae_R6
Cluster No. 20	20093831/M_kandleri 20093832/M_kandleri
Cluster No. 21	16801218/L_innocua 16804085/L_monocytogenes
Cluster No. 22	21232578/X_campestris 21244026/X_citri
Cluster No. 23	26991600/P_putida_KT2440 28872090/P_syringae
Cluster No. 24	33592626/B_pertussis 33600410/B_bronchiseptica
Cluster No. 25	15837384/X_fastidiosa 28199747/X_fastidiosa_Temecula1
Cluster No. 26	16759402/S_typhi 29142826/S_typhi_Ty2
Cluster No. 27	21232698/X_campestris 21244140/X_citri
Cluster No. 28	15676298/N_meningitidis_MC58 15677889/N_meningitidis_MC58 1
Cluster No. 29	17986612/B_melitensis 23502565/B_suis_1330
Cluster No. 30	21232748/X_campestris 21244172/X_citri
Cluster No. 31	15900616/S_pneumoniae_TIGR4 15902675/S_pneumoniae_R6
Cluster No. 32	21223918/S_coelicolor 29829218/S_avermitilis
Cluster No. 33	17988015/B_melitensis 23501125/B_suis_1330
Cluster No. 34	16800703/L_innocua 16803633/L_monocytogenes
Cluster No. 35	19552525/C_glutamicum 25027963/C_efficiens_YS-314
Cluster No. 36	16759396/S_typhi 16763799/S_typhimurium_LT2 29142832/S_typh
Cluster No. 37	15838486/X_fastidiosa 28198805/X_fastidiosa_Temecula1
Cluster No. 38	16120853/Y_pestis_CO92 22127525/Y_pestis_KIM
Cluster No. 39	30023259/B_cereus_ATCC14579 30265254/B_anthracis_Ames
Cluster No. 40	33591450/B_pertussis 33595015/B_parapertussis 33599293/B_br
Cluster No. 41	15839800/M_tuberculosis_CDC1551 15839801/M_tuberculosis_CDC
Cluster No. 42	15827068/M_leprae 15839810/M_tuberculosis_CDC1551
Cluster No. 43	15837196/X_fastidiosa 28199436/X_fastidiosa_Temecula1
Cluster No. 44	21222334/S_coelicolor 29830808/S_avermitilis
Cluster No. 45	16799417/L_innocua 16802360/L_monocytogenes
Cluster No. 46	15827072/M_leprae 15827073/M_leprae
Cluster No. 47	16272304/H_influenzae 16272364/H_influenzae
Cluster No. 48	30022723/B_cereus_ATCC14579 30264722/B_anthracis_Ames
Cluster No. 49	15834168/E_coli_O157H7 30064727/S_flexneri_2a_2457T
Cluster No. 50	33595238/B_parapertussis 33599526/B_bronchiseptica
Cluster No. 51	33594359/B_pertussis 33599940/B_bronchiseptica
Cluster No. 52	17986568/B_melitensis 23502614/B_suis_1330
Cluster No. 53	20093117/M_acetivorans 21227127/M_mazei
Cluster No. 54	15677863/N_meningitidis_MC58 15793405/N_meningitidis_Z2491
Cluster No. 55	15837558/X_fastidiosa 28199614/X_fastidiosa_Temecula1

4.6 MEME/MAST

MEME (Múltiple Em for Motif Elicidation) es un programa que obtiene motivos (secuencias altamente conservadas), de acuerdo a matrices especificas de posición. La matriz calcula la probabilidad que tiene cada posible letra de una ventana de secuencia, en cada posible posición (match), en este caso el valor de un "match" en un motivo comparado con una secuencia al azar es una variable discreta de la cual se puede calcular su distribución y a partir de esta distribución se puede calcular el valor de probabilidad (p-value) del "match" [11].

MEME utiliza un valor de "proporción de probabilidad logarítmica" (log likelihood ratio LLR) para la ocurrencia de cada motivo y a partir del LLR se estima el valor de expectancia (e-value) del motivo en un estimado de motivos (con el mismo largo y con el mismo número de secuencias encontradas por motivo), esto nos da el mismo o mayor LLR en los motivos evaluados.

MAST (Motif Alignment & Search Tool) es una herramienta de búsqueda de secuencias que contienen uno o más motivos en una base de datos dada [12]. MAST está diseñado para buscar "matches" a partir de matrices especificas de posición, de esta forma MEME y MAST son programas complementarios

El programa MEME tiene distintas opciones para dirigir la búsqueda:

- El tipo de alfabeto que va a comparar: secuencias de aminoácidos [-protein] o secuencias nucleotidicas [-dna]
- La distribución de motivos: que el motivo este por lo menos una vez en cada secuencia [-mod oops], que el motivo este cero o una vez por secuencia [-mod zoops] o que el motivo este cualquier número de veces en una secuencia de forma no sobrelapada [-mod tcm]
- El máximo número de motivos a encontrar [-nmotifs]
- Para ajustar los resultados a un máximo valor de espectancia (e-value) [-evt]
- Para ajustar la cantidad de secuencias que presenten el motivo: [nsites], mínimo número de secuencias por motivo [-minsites], máximo número de secuencias por motivo [-maxsites]
- Para ajustar el largo del motivo (ventana):[-w], mínimo largo [-minw], máximo largo [-maxw], entre otros.

Las secuencias intergénicas sin repetidos se compararon con el programa MEME y se obtuvieron motivos conservados para cada grupo de secuencias comparadas. Los motivos definidos en el programa MEME fueron posteriormente utilizados para realizar una búsqueda de secuencias con MAST. Para la comparación con MAST se utilizaron todas las secuencias intergénicas de la base de datos iMUR, lo cual nos permite encontrar más secuencias relacionadas a los motivos que inicialmente no fueron encontradas por la palabra clave. Finalmente se volvió a realizar un MEME a las secuencias encontradas por MAST para obtener motivos más representativos.

En el método se utilizó la opción "dna" para secuencias nucleotidicas y se utilizó el rango de largo de motivo definido por MEME (–minw 8, -maxw 50). Además para el

primer MEME se utilizo la opción zoops, considerando que no todas las secuencias que se están comparando presentan motivos que corresponden a secuencias de "Riboswitches", esta opción le da más flexibilidad a la búsqueda ya que el motivo puede estar cero o una vez por secuencia, a diferencia del primer MEME en el segundo MEME se utilizo la opción "oops", ya que las secuencias que se compararon fueron la salida de MAST que es consecuencia de las matrices obtenidas por el primer MEME, esto supone que todas las secuencias presentan por lo menos uno o más motivos y esta segunda búsqueda nos ayuda a eliminar motivos que no están en todas las secuencias y a reconocer un motivo más fuerte.

Por otra parte para limitar la primera búsqueda a un número limitado de motivos se utilizo –nmotifs 6 que nos da los 6 primeros motivos con menor e-value, este parámetro se determinó en la primera prueba del método, que fue hecha con Tiamina, y se vio que 6 motivos era suficiente para reconocer motivos que corresponden a la firma de Riboswitches.

El MAST se corrió después de una evaluación de los motivos de MEME, donde se eliminaron todos aquellos motivos con un 95% de contenido G-C o A-T por considerarse con un bajo contenido informacional.

Para probar la eficiencia del método se probó todo el proceso utilizando palabras clave para la biosíntesis de Tiamina y se compararon los resultados con los ya reportados [19, 20] y después se reprodujo la metodología, tomando palabras clave relacionadas a la biosíntesis de otras vitaminas y cofactores. Las palabras clave que se utilizaron están ordenadas en la (Tabla 5)

Tabla 5
Palabras elegidas para la búsqueda de regulación

Palabras clave	**Metabolismo**
biotin	Metabolismo de Biotina
cobalamin	Metabolismo de Cobalamina
folate	Metabolismo de folato
glycine, imidazole	Metabolismo de Glicina
nicotina, quinolina	Metabolismo de Nicotina
pantothe, pantoate	Metabolismo de Pantotenato
pyridox	Metabolismo de Piridoxina
falvin	Metabolismo de Rivoflavina
thiami, thiazo, thylpyri	Metabolismo de Tiamina
benzoate, benzoquinone, prenylphenol, quinone, ubiquino	Metabolismo de Ubiquinona
cytosine	Metabolismo de Pirimidinas
thymine	
uracil, uridine	
adenine, adenosine	Metabolismo de Purinas
guanine, guanosine	
xanthine, xanthosine, inosine	

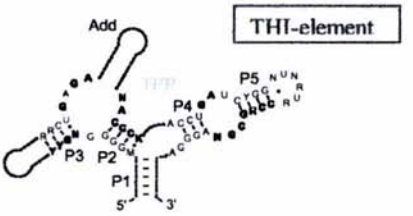
5. Resultados y Discusión

Se probaron 12 casos de distintas vías metabólicas (Resumido en la Tabla 4), de las cuales únicamente se encontraron 5 casos con motivos fuertemente conservados: Cobalamina, Riboflavina, Uracilo (Metabolismo de Pirimidinas) y Xantina (Metabolismo de Purinas). Únicamente se consideran estos casos por que son aquellos que presentan motivos conservados, por arriba de un E-value de $1e^{-10}$, que no sean secuencias ricas en G-C o en A-T y solo aquellos casos con motivos que están presentes en todas las secuencias (opción oops de MEME, que asume que cada secuencia dada, contiene por lo menos una ocurrencia de cada motivo)

A continuación se detalla el análisis de cada uno de los 5 casos encontrados en mi estudio, empezando con el caso de Tiamina que es la secuencia de regulación de tipo Riboswitches más altamente conservada, por lo que se utilizó para evaluar la eficiencia de mi método

5.1 Comparación de los resultados para genes de biosintesis de Tiamina

El elemento de regulación THI-element, fue reportado en 2002, por Rodionov et. al. A partir de un análisis comparativo con 103 secuencias genómicas, en las cuales se reportaron 170 THI-elements en 78 genomas (Fig 9) [19]



Firma de THI-element reportada (Rodionov, 2002)

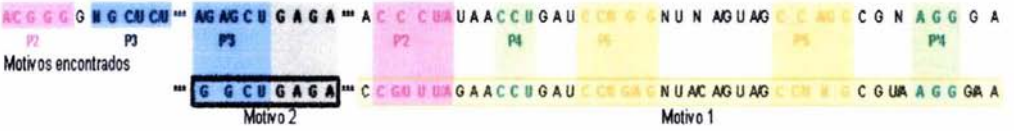
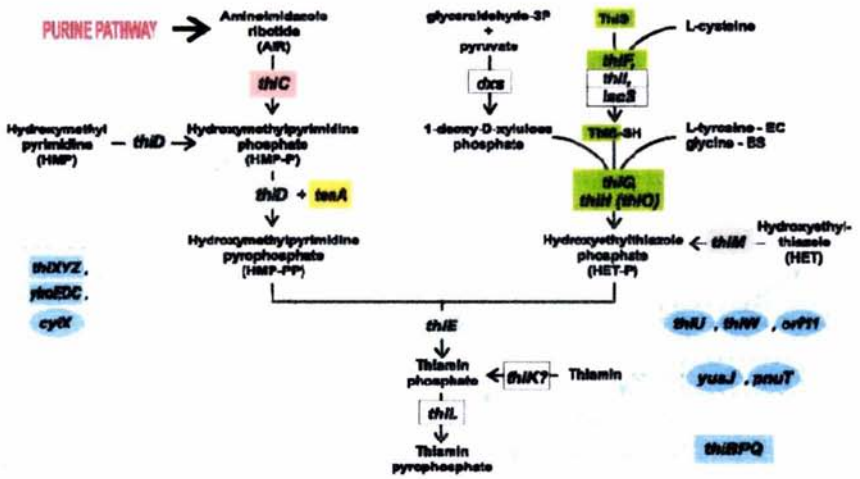


Fig 10. Estructura secundaria propuesta para el THI-element y secuencia conservada. En la secuencia conservada los colores corresponden a segmentos complementarios: rosa (P2-P'2), azul (P3-P'3), verde (P4-P'4), anaranjado (P5-P'5) de acuerdo al modelo reportado por Rodionov 2002, los (***) corresponden a segmentos variables o adicionales (Add) y el segmento gris (GAGA), corresponde a un loop conservado. Motivo 1 y Motivo 2 corresponden a los motivos encontrados en mi estudio

Los motivos 1 y 2 que se presentan en la Figura 10, se obtuvieron de la búsqueda con palabras clave para tiamina (detalle en la Tabla 6). Estos motivos se encuentran presentes en 219 secuencias geonómicas y son congruentes con la firma reportada, lo cual valida positivamente los resultados obtenidos a partir del método.

Tabla 6 Búsqueda de genes regulados por THI-element



Biosíntesis de Tiamina en bacterias: la tiamina (vitamina B1), es sintetizada por un motivo de pirimidina (hidroxietilpirimidine, HMP) y uno de thiazol (hydroxietilthiazole, HET) [19]

Palabra clave	Total de genes encontrados		Cantidad de THI-element encontrados	Principales funciones reguladas por THI-element
	Genes por palabra clave	iMURs		
thiami	673	380	98	• ***thiC***, Proteína de la biosíntesis de tiamina • ***tbpA***, Proteína del periplasma, sistema de transporte de tiamina, tipo-ABC • ***thiM***, Hydroxietilthiazole cinasa • ***thiD***, Hidroxietilpirimidin/Fosfoetilpirimidin cinasa • ***tenA***, Regulador transcripcional, posible activador • ***thiO***, Oxidoreductasa de la biosíntesis de tiamina • ***thiE***, Tiamin fosfato sintasa • ***yuaJ, ykoE***, Proteínas de Membrana
thiazo	136	71	180	
thilpyri	128	76	186	

El 70% de las firmas encontradas corresponden a los genes *thiC, tbpA, thiM, yuaJ, thiD, tenA, ykoE, thiO y thiE.* El gen *thi*C, se encuentra en muchos casos agrupado en operon con genes de biosíntesis de tiamina y está fuertemente regulado por THI-element. En el caso de *Rizobium etli*, THI-element regula al operon *thiCOGE* [17] y en el caso de *Escherichia coli* regula a *thiH-thiG-thiS-thiF-thiE-thiC.*

La tiamina en su forma activa (TPP) se sintetizada a partir de dos principales precursores, la hidroxietilpirimidina (HMP) y el hydroxietiltiazole (HET). Como se observa en la ruta metabólica de la Tabla 4, *thiC* produce HMP, que es después fosforilada por una cinasa, (*thiD*). El HET se forma a partir de tirosina o glicina (tirosina

en *E.coli* y glicina en *B. subtilis),* de cisteina y de 1-Deoxi-d-xilulosa fosfato, a través de una serie de reacciones que involucran los genes *thiF, thiS, thiG, thiI, thiH/E. coli, thiO/B. subtilis.* A si mismo *thiM* codifica para una sinasa HET, de la vía de obtención de tiazol. La formación de tiaminfosfato, precursor de la tiaminpirofosfato se da por la sintasa *thiE* a partir de HMP y HET.

De acuerdo a la biosíntesis de tiamina (Tabla 6). Mi búsqueda de secuencias conservadas, partió de los términos, "thiami", "thiazo" y "thilpyri", de los cuales se obtuvo un grupo de 918 genes relacionados a la biosíntesis de tiamina: 664 genes a partir de "thiami", 117 genes a partir de "thiazo", 19 genes compartidos por "thiami/thiazo" y 128 genes a partir de "thilpyri" (Tabla 6).

La regulación por THI-element, se extiende tanto a proteínas relacionadas a la biosíntesis de tiamina, como a proteínas de transporte de tiamina. *tbpA* y *ykoE* son proteínas del sistema de transporte ABC, donde *ykoE* está asociado a *ykoD* y *ykoC* (*ykoE-ykoD-ykoC*) y está involucrado en el transporte de HMP.

5.2 Comparación de los resultados para genes de biosíntesis de Riboflavina

El elemento de regulación RFN (Fig 11) fue reportado originalmente por Gelfand en 1999 [22]. En un estudio con una búsqueda más amplia (Vitreschak 2002) en la cual se reportaron 61 elementos en 49 genomas [23].

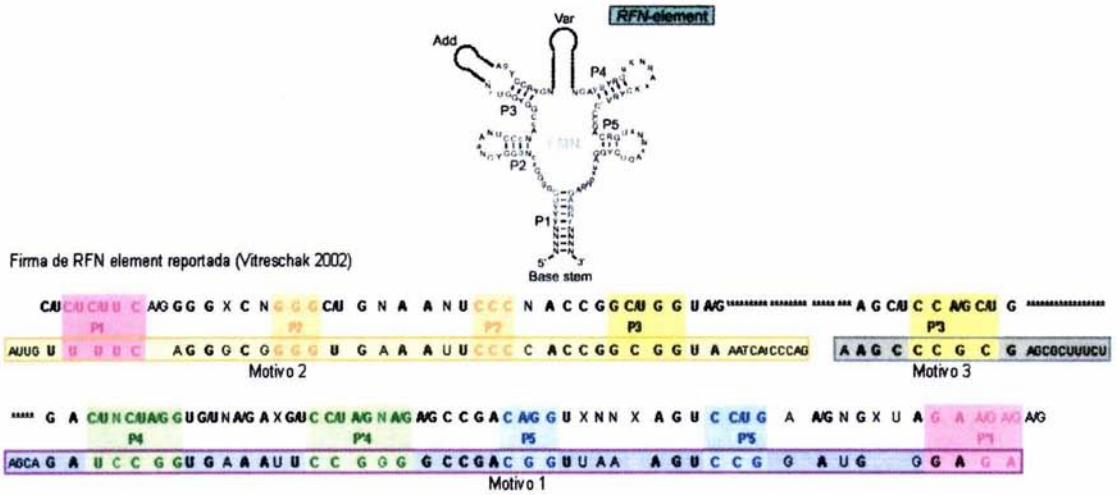
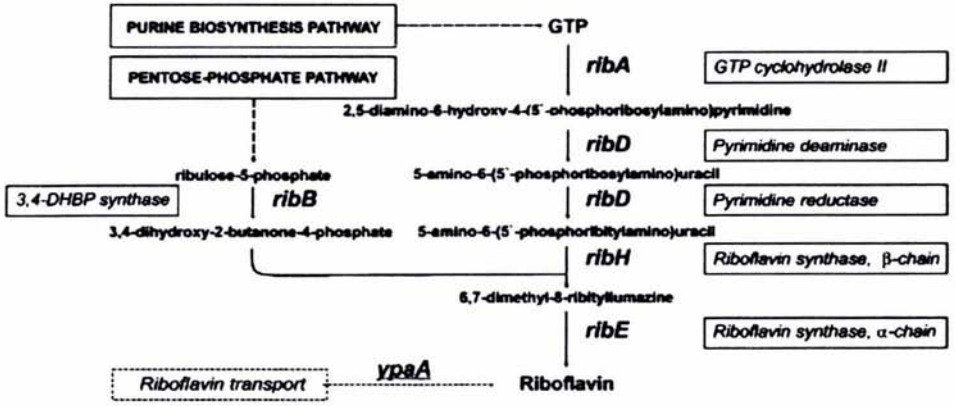


Fig 11. Estructura secundaria propuesta para el RFN-element y secuencia conservada. Tomada de la estructura propuesta por Vitreschak, 2002 En la secuencia conservada, los colores corresponden a segmentos complementarios: rosa (P1-P'1), anaranjado (P2-P'2), amarillo (P3-P'3), verde (P4-P'4), azul (P5-P'5). La primer secuencia corresponde al consenso. Motivo1, Motivo2 y Motivo3 son los motivos obtenidos con mi método

En el caso de la Riboflavina, los resultados obtenidos a partir de la palabra clave "flavi" me arrojaron tres motivos claramente relacionados al consenso: Motivo1, Motivo2 y Motivo3 (Fig 11). Los 3 motivos encontrados, coinciden con la secuencia consenso del elemento RFN reportada [22, 23] y mantienen el mismo orden (Motivo 2, 3, 1) en todas las secuencias analizadas.

Además de los 3 motivos encontrados, se encontró un cuarto motivo altamente conservado que no forma parte del elemento RFN reportado, este motivo se localiza después de los motivos 2-3-1 en la mayoría de los casos (~70%), sin embargo también está presente antes de los motivos 2-3-1.

Tabla 7 Búsqueda de genes regulados por el RFN-element



Biosíntesis de Riboflavina: la Riboflavina (vitamina B2), es sintetizada a partir de una molécula de GTP y dos moléculas de ribulosa 5-fosfato [23].

Palabra clave	Elementos de regulación encontrados	Principales funciones que presentan la regulación
flavi	De 465 secuencias intergénicas probadas se obtuvieron 42 secuencias altamente conservadas que coinciden con el elemento RFN	• ***ribB,*** 3, 4 dihidroxi-2-butano 4-fosfate sintasa • ***ribD,*** Reductasa/Deaminasa de Pirimidina • ***ribH,*** Sintasa de Riboflavina, cadena beta • ***ribE,*** Sintasa de Riboflavina, cadena alfa • ***ypaA,*** Transportador de Riboflavina

La síntesis de Riboflavina empieza al nivel de una guanosina o de un nucleótido respectivamente. Como se observa en la ruta metabólica de la Tabla 7, el motivo de ribosa del precursor púrico es convertido directamente a un motivo ribitil. El primer paso de la biosíntesis es catalizado por una ciclohidrolasa de GTP II. Después son catalizados varios intermediarios de pirimidina y por último la conversión de 5-amino-2,6-dihidroxi-4-(D-ribitylamino) pirimidina a 6, 7-dimetil-8-(D-ribitil)lumacina por la adición de cuatro motivos de carbono

Los principales genes regulados por el elemento RFN que se encontraron por mi método, son *ribB, ribD, ribH, ribE,* entre otros. Todos estos genes forman parte de la vía de síntesis de la Riboflavina, donde *ribE* codifica para una sintasa de riboflavina responsable de convertir la lumacina a Riboflavina, el gen *ribB* codifica para una sintasa DHBP responsable de la síntesis de 3,4-dihydroxyl-2-butanone 4-fosfato, el gen *ribH* codifica para una sintasa responsable en la síntesis de lumacina, el gen *ribA* que codifica para una ciclohidrolasa de GTP II responsable de la síntesis de lumacina y el gen *ribD* que codifica para una enzima bifuncional deaminasa/reductasa de pirimidina. Además de los genes relacionados a la biosíntesis de Riboflavina encontrados se observo que el gen *ypaA* que es un transportador de Flavinas presenta el elemento RFN en su región río arriba (Tabla 7)

5.3 Comparación de los resultados para genes de biosíntesis de Cobalamina

El elemento de regulación B_{12} (Fig 12) fue reportado originalmente por Vitreschak en el 2003 [17]. Donde se reportan 200 elementos B_{12} en 66 genomas.

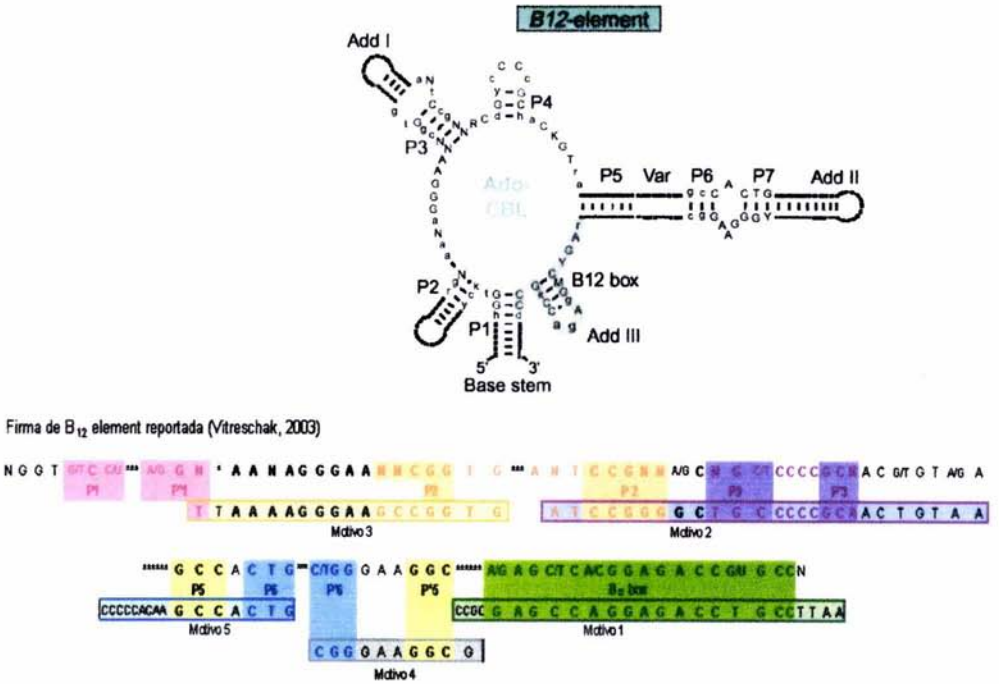
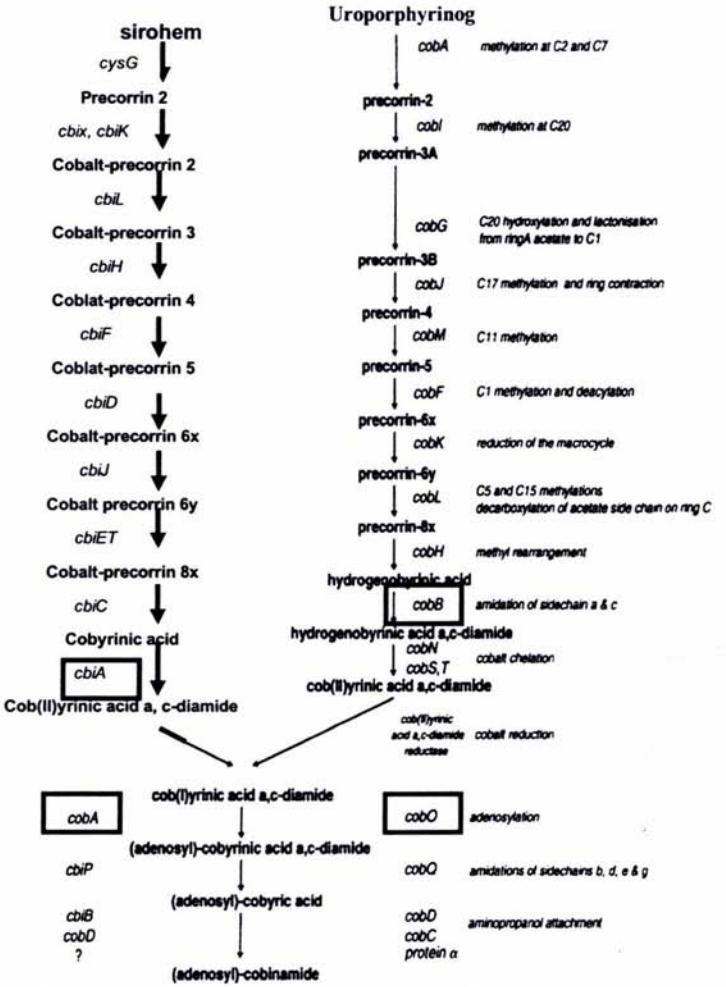


Fig 12 Estructura secundaria propuesta para B_{12}-element y secuencia conservada. La estructura propuesta para el elemento B_{12}, consiste en 7 hélices (P0-P6). En la secuencia conservada los colores corresponden a segmentos complementarios: rosa (P1-P'1), naranja (P2-P'2), morado (P3-P'3), amarillo (P5-P'5), azul (P6-P'6), verde (B_{12} box). El elemento B_{12} tiene un numero de secuencias facultativas no conservadas designadas add I y add II y una estructura interna variable (Var). Motivo 1, Motivo 2, Motivo 3, Motivo 4 y Motivo 5 corresponden a los motivos encontrados con mi método

Tabla 8 Búsqueda de genes regulados por B_{12}-element



Biosíntesis de Cobalamina: La biosíntesis de Cobalamina puede dividirse en dos vías alternas, La vía de síntesis de cobalamina (genes cob), y la síntesis de cobinamida (genes cbi).

Palabra Clave	Elementos de regulación encontrados	Principales funciones que presentan la regulación
cobalamin	De 263 secuencias intergénicas probadas se obtuvieron 58 secuencias altamente conservadas que coinciden con el elemento RFN	• ***btuB,*** Receptor de membrana externa para cobalamina • ***phuR, hasR, fhuA, cirA, fecA,*** Receptores de membrana externa, la mayoría transportadores de Fe^{+} • ***fecB, fhuD,*** sistema de transporte de hidroxamato $Fe3^{+}$, de tipo ABC, periplasmico • ***metE,*** sintasa de Metionina II, (independiente de cobalamina) • ***hoxN,*** Permeasa de alta afinidad al Níquel • ***btuR, cobA, cobO,*** ATP: adenosiltransferasa corrinoide • ***cbiA, cobB,*** ácido cobirinoica a, c/ diamino sintasa • ***cobW,*** CTPasa putativa (familia de G3E) • ***yciA,*** hidrolasa de Acyl-CoA

En el caso de la Cobalamina, se utilizó el término "cobalamin" como palabra clave, para levantar los genes relacionados al metabolismo de Cobalamina y en base a esto se levanto un grupo de 58 secuencias altamente conservadas, dentro de los genes regulados por estas secuencias, se encuentra el gen *btuB* que codifica para una proteína de membrana externa necesaria para el transporte extracelular de Cobalamina (Ver detalle en Tabla 8).

A partir de la búsqueda se encontraron 5 motivos distribuidos siempre en el mismo orden en las secuencias intergénicas analizadas: motivo 3-2-5-4-1. Estos motivos son consistentes con la firma de Cobalamina propuesta lo cual refleja la alta conservación de la firma propuesta. El motivo 1, contiene la B_{12} *box*, que es la secuencia más altamente conservada. El hecho de que la secuencia más altamente conservada obtenida (motivo 1), sea la que contiene la B_{12} *box* es congruente con la función de reconocimiento de adenosyl cobalamin de la misma.

Varios genes regulados por B_{12}-element corresponden a receptores de membrana externa, como: *phuR, hasR, fhuA, cirA, fecA,* dentro de los cuales hay muchos transportadores de Fe^{+}, como en el caso de *fecB, fhuD.*

Otra función regulada por B_{12}-element es la síntesis de 5'-deoxiadenosil-Cbl (Ado-Cbl) que se da a partir de CN-Cbl por la acción del producto del gen *btuR*. El equivalente en *Salmonella typhimurium* del gen *btuR* de *Escherichia coli* es *cobA* y en Pseudomonas denitrificands se llama *cobO.*

Así mismo también se encontraron otros genes relacionados a genes y operones de biosíntesis de Cobalamina, como son *hoxN*, que codifica para un transportador de Niquel/Coblato y *cobW* que codifica para un quelador de Cobalto.

Además de los genes de la vía metabólica de Cobalamina, se encontraron genes de vías metabólicas alternas, dependientes de Cobalamina, como *metE* corresponde a una sintasa de Metionina

5.4 Comparación de los resultados para genes de biosíntesis de Purinas

En este caso se utilizaron 7 terminos relacionados a la biosíntesis de Purinas [45], 2 relacionados a la biosíntesis de adenina (adenine-adenosine), 2 relacionados a la biosíntesis de guanina (guanine-guanosine) y dos relacionados a la biosíntesis de xantina (xanthine- xanthosine-inosine), de las cuales solo "xanthine" y "adenine", encontraron motivos altamente conservados. A partir de estas palabras se encontraron 34 genes con los motivos que se presentan en la Fig 12. Todos los genes presentan, tanto el motivo1 como el motivo 2 y siempre en el orden motivo2-motivo1, que coinciden de forma congruente con la firma reportada para la G-*box*.

Mandal (2003) reporto que el operón *xpt-pbuX* está controlado por un Riboswhitch, "G-*box*" y que es regulado por guanina, xantina e hipoxantina. Según el estudio realizado por Mandal, este regulador está presente en por lo menos 34 secuencias, distribuidas tanto en organismos Gram + como Gram -. [33].

La estructura secundaria de la G-*box* está compuesta por tres segmentos complementarios (P1-P3), las secuencias más conservadas de este caso son: el segmento complementario 1 y las regiones no complementarias (Fig 13).

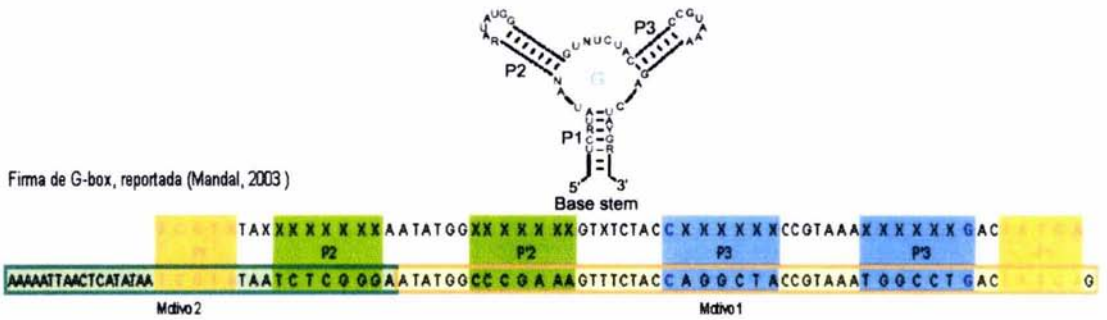


Fig 13. Estructura secundaria propuesta para la G-box y secuencia conservada. La G-box es una secuencia con tres segmentos complementarios, que corresponden a cada color: anaranjado (P1-P'1), verde (P2-P'2) y azul (P3-P'3). La primer secuencia corresponde a lo reportado por Mandal y los motivo1, motivo2, corresponden a lo encontrados en el estudio

Algunos de los genes que presentan los motivos encontrados son: *apt*, que codifica para fosforiltransferasas de adenina/guanina, *yieG* que codifica para permeasa xantina/uracilo / permeasa gunina/hipoxantina, *purE* que codifica para carboxilasa fosforibocilaminoimidazole, *guaA*, que codifica para una sintasa GMP glutamin amino transferasa y *hutH* que es un regulador transcripcional.

5.5 Comparación de los resultados para genes de biosíntesis de Pirimidinas

En este caso se utilizarón 4 palabras clave (cytosine, thymine, uracil, uridine) relacionadas a la regulación de Pirimidinas de las cuales solo "uracil" levanto motivos altamente consebados. A partir de este término se encontraron dos motivos altamente conservados en 60 secuencias.
Motivo 1: CCTTTAAATTAGTCCAGTGAGGCTGACACAAGGAG
Mortivo 2: CTTCTTTGGCAGGGAGGTTTTTTTTT

En todos los casos los motivos se encuentran en el mismo orden, motivo 1–2.
Las principales funciones encontradas para este caso fueron genes relacionados a la biosíntesis de pirimidinas. Algunos de ellos son: *pyrF* que codifica para descarboxilasa orotidin-5'-fosfato, *pyrB* que codifica para carbamoiltransferasa de aspartame, de cadena

catalítica, *pyrR* del operon de pirimidina, proteína de atenuación / fosforibosiltransferasa de uracilo y *uraA* que codifica para permeasa de xanthina/uracilo, entre otros.

Se sabe que el operón biosintetico (*pyr*) de *Bacillus subtilis* esta regulado por un mecanismo autógeno de atenuación transcripcional en el cual el primer gen del operón *pyrR*, codifica para una proteína reguladora (*PyrR*) que causa la terminación transcripcional, mediante la unión uridina núcleotido dependiente, a tres sitios específicos en *pyr*. Estos sitios están localizados en la región líder del operón entre el primer (*pyrR*) y segundo (*pyrP*) sistron. Y entre el segundo y tercero (*pyrB*) sistron (Fig 14). Se ha visto que el líder del RNA es capaz de formar estructuras secundarias de terminador transcripcional, antiterminador y anti antiterminador y se a observado tambien que PyrR se une a la secuencia del anti-anti-terminador de forma altamente especifica permitiendo que se forme el terminador transcripcional. El ácido uridilico o monofosfato de uridina (UMP) y el 5-fosforibosil-1-pirofosfato (PRPP) son correguladores con PyrR [34].

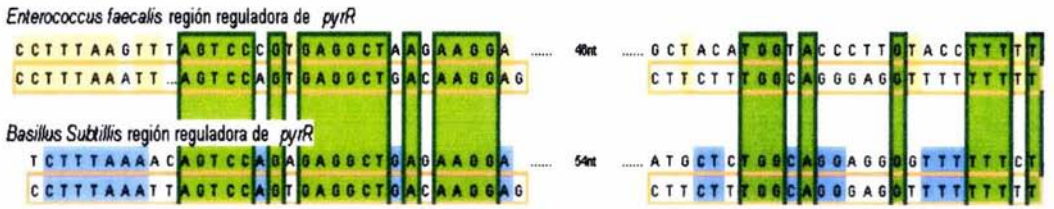


Fig 14. Región reguladora de *pyrR*. La reguión río arriba de *pyrR*, presenta sitios de unión a la proteína PyrR

5.6 Trabajo conjunto de regiones conservadas de regulación

A partir de mi proyecto de tesis se planteó una idea alterna que propone la búsqueda de secuencias de regulación a nivel global. En este trabajo se utilizan los mismos principios de búsqueda pero se analizan grupos de genes por familia de ortólogos definidas en la base de datos COG [41]. De esta forma generamos un proceso eficiente para seleccionar secuencias de regulación.

Este trabajo esta descrito en "Conserved regulatory motifs in bacteria: riboswitches and beyond" de Cei Abreu-Goodger, Nancy Ontiveros-Palacios, Ricardo Ciria and Enrique Merino (Sometido a Trends in Genetics No. A2445), Anexo a la Tesis.

Además de reconocer las firmas de los Riboswitches reportados, el método general por COGs, reconoce motivos altamente conservados en la biosíntesis de aminoácidos, ver Tabla 9.

Tabla 9
Motivos encontrados en biosíntesis de aminoácidos

Metabolismo	Gpo.	Probabilidad (p-value)	Motiv	COGs
Alanina	0116	5.87e-85	2	COG2065, COG0540, COG0284, COG2233, COG0461,
Aspartato				COG0044
	0101	1.61e-56	1	COG2502
	0286	1.49e-54	1	COG0104
Glutamato	0113	4.9e-82	3	COG0008
	0034	8.22e-67	3	COG0505, COG0458, COG2065
	0068	5.37e-59	4	COG0449
	0336	1.76e-54	3	COG0505, COG0458
Glicina	0031	3.08e-102	4	COG0290, COG3601, COG0307, COG0054, COG1985
Serina	0033	3.84e-58	3	COG0404, COG1115, COG1003, COG0403
Treonina	0275	8.74e-55	1	COG0027
Glicina	0289	5.27e-61	4	COG0112
Licina				
Glicina	0340	5.02e-86	4	COG0527, COG0083, COG0498
Licina	0122	1.41e-81	3	
Piridoxina	0157	9.25e-70	2	
Licina	0008	6.25e-193	4	COG0216, COG2890, COG1186, COG1190, COG0373, COG1435, COG3094, COG0009, COG2912, COG2884, COG2177
	0502	3.40e-109	3	COG3335
	0199	9.48e-52	1	COG1279
Metionina	0009	8.95e-86	4	COG1135, COG1464, COG2011, COG0192
	0274	6.96e-63	3	COG0024
Cisteina	0179	2.37e-56	2	COG1959, COG1104
Cisteina	0009	8.95e-86	4	COG1135, COG1464, COG2011, COG0192
Metionina				
Valina	0284	8.32e-77	4	COG0059
Leucina	0276	8.94e-66	4	COG3978, COG0028, COG0115
Isoleucina	0384	1.37e-49	4	
Arginina	0052	2.08e-42	2	COG0078
Prolina				
Histidina	0279	4.20e-46	2	COG0040, COG0141, COG0079
Fenilanina				
Fenilalanina	0320	2.00e-52	4	COG0703, COG0337
Tirosina				
Triptofano				
Triptofano	0325	1.81e-61	1	COG0376
Glutamina	0295	1.13e-21	3	COG0174

Tambien se localizaron motivos, en algunos de los casos probados en mi estudio, como son el caso de la biosíntesis de Biotina, Folato, Pantotenato y Ubiquinona (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html).

Los motivos encontrados, representan posibles secuencias de regulación a nivel de secuencia en RNA. Se conocen tres casos de síntesis de aminoácidos regulados por Riboswitches, el caso de la Metionina [26, 27], el de la Licina [31, 32] y hace poco el caso de la Glicina [38], por lo que los motivos localizados tienen muchas posibilidades de ser verdaderas secuencias de regulación

Lo reportado en la tesis y en este trabajo representa un análisis complementario y ampliado en el estudio de secuencias reguladoras que realizamos en el laboratorio. Anexo a la tesis el manuscrito senviado a "Trends in Genetics" y de esta forma redondeo el proceso realizado en la búsqueda de secuencias de regulación conservadas.

6. Discusión General

El análisis comparativo de secuencias ha demostrado que las regiones no codificantes tienden a mostrar un mayor grado de variación, por lo que las regiones altamente conservadas en estas regiones resultan ser secuencias muy interesantes para el análisis genómico. En este caso, encontrar secuencias conservadas en regiones rió arriba de operones, nos sugiere una posible función reguladora. Como se observo en este estudio, de acuerdo a nuestros parámetros, no todas las palabras clave encuentran motivos altamente conservados, sin embargo en las palabras donde si se reconocen secuencias conservadas, los motivos están relacionados con varios Riboswitches. Los Riboswitches son en tanto secuencias muy conservadas que además están distribuidas en varios grupos taxonomicos y en muchos genes. El hecho de que estas secuencias estén altamente conservadas y de que estén muy distribuidas a través de los distintos genomas, nos hace pensar que estos elementos están presentes desde hace mucho tiempo y que su conservación obedece obedece a procesos de reconocimiento de una señal especifica como pudiera ser un metabolito dentro de la via metabólica.

Por otro lado el hecho de que la regulación no involucre la intervención de proteínas y que en vez de esto haya un reconocimiento entre el RNA y pequeños metabolitos, nos sugiere que la regulación es muy especifica y que es una regulación que pudo estar presente aun antes de que las proteínas intervinieran en procesos de regulación.

Sin embargo, la pregunta interesante es ¿Por qué, si las proteínas han adquirido funciones cada vez más especificas y actualmente están involucradas en la mayoría de los procesos de regulación genética, no han sustituido a la regulación de tipo Riboswitch?. Para respondernos a esto es importante considerar que los Riboswitches están involucrados en la regulación de la biosíntesis de metabolitos secundarios como: vitaminas, cofactores y aminoácidos (al menos en lo que se sabe hasta este momento), a diferencia de éstos, las vías metabólicas centrales, están reguladas por proteínas. Se ha propuesto que las proteínas han evolucionado como un sistema de regulación más eficiente y especifico sin embargo la alta sensibilidad que presentan los Riboswitches a los cambios de concentración de la molécula efectora (nM), los hace un muy buen sistema de regulación.

Es interesante recalcar que los Riboswitches presentan diferencias de regulación entre grupos taxonómicos, como se observo, los organismos Gram +, presentan una regulación preferentemente a nivel transcripcional a diferencia de los organismos Gram -, que presentan una regulación preferentemente a nivel traduccional. Aunque también hay casos en que la regulación de un gen puede darse tanto a nivel traduccional como a nivel

transcripcional, este es el caso del gen *btuB* de *Escherichi coli* y *Salmonella thiphymurium* que es regulado por el elemento B_{12}. En éste caso la regulación traduccional del gen *btuB* requiere de una región líder de 241nt y en el caso de la regulación transcripcional de 60-100nt. Como en todos los casos reportados la regulación traduccional involucra el secuestro de la secuencia SD y la regulación transcripcional involucra un evento de atenuación y se sabe que en este caso la regulación transcripcional es secundaria a la regulación traduccional [14]. Aunque es claro que hay una diferencia en el tipo de regulación que es utilizada por organismos Gram + y Gram -, nada concreto se ha planteado respecto a las posibles razones de estas diferencias.

La conservación de la estructura secundaria de los Riboswitches también es importante en la regulación. Se ha observado en distintos Riboswitches que las mutaciones que modifican la estructura secundaria, causan una perdida de la regulación y que cambios compensatorios en las regiones complementarias pueden reestablecer la regulación [19, 33]. Considerando que los Riboswitches interaccionan con moléculas especificas, es indispensable que la región de reconocimiento de la molécula este expuesta, así mismo es indisable la formación tanto del atenuador/secuestrador como la formación de las otras estructuras competitivas (anti-terminador/anti-antiSD y anti-anti-terminador/anti-anti-antiSD). Por lo tanto, aunque la secuencia de un Riboswhitch, puede variar, siempre presenta cambios compensatorios que mantiene el sistema de regulación, no así la secuencia que es reconocida por la molécula efectora.

Particularmente en el caso de THI-element es muy contrastante la cantidad de secuencias de regulación que son levantadas en comparación de otros casos analizados. En todos los casos se utilizaron los mismos parámetros para discriminar entre secuencias conservadas, y es interesante observar que las secuencias que presentan THI-element son más del doble (170 sec.) de las que son reconocidas en cualquier otro caso, esto refleja que la secuencia de regulación THI-element, es una secuencia muy conservada.

Discusión del método

La ventaja de utilizar en la búsqueda una palabra clave es que en ella se incluye a todos los genes relacionados a la biosíntesis y transporte de un metabolito y al tener un grupo inicial de secuencias relacionadas metabolicamente es más probable encontrar motivos conservados (MEME). Por otro lado la limitante de la búsqueda por palabra clave sobre la base de datos de PTT, es que muchos genes no son localizados, por varias razones: a) en algunos casos la función reportada es inespecífica, b) en muchos otros casos la función no esta relacionada con la palabra clave, c) además de que en varios de los casos las funciones están mal asignadas y d) que muchos genes no tienen una función asignada. Sin embargo, como se observo en este estudio la cantidad de genes reconocidos por la palabra clave es suficiente para encontrar motivos conservados que después sirven de ancla para localizar otras secuencias relacionadas a la misma regulación a partir de una búsqueda de motivos en la base de datos iMUR (todas las secuencias genómicas intergénicas) con MAST.

Esta búsqueda en MAST, puede compensar las deficiencias antes mencionadas, no así el hecho de que la base de datos PTT, sigue siendo una base de datos limitada, que

esta restringida a los genomas totalmente secuenciados y a lo que se sabe de los genes hasta el momento. En la medida en que sean secuenciados más genomas, la búsqueda se vera enriquecida y en análisis puede ser más extenso y preciso.

Utilizar 6 motivos como parámetro funcionó exitosamente para el resto de los casos esperados, sin embargo se puede reevaluar la búsqueda y limitar por otra cantidad de motivos o por un e-value, de acuerdo a lo que se este buscando, considerando que los Riboswitches son firmas con pequeños motivos altamente conservados, fue congruente pensar que los primeros motivos encontraran secuencias con verdadera función biológica (en este caso regulación tipo "Riboswhitch").

Por otro lado se utilizo el rango de 8 a 50 n para el largo del motivo definido por MEME, por que las secuencias conservadas en las firmas de "Ribositches", no llegan a ser secuencias mayores a 50nc, esto se ve en que tanto las secuencias de complementariedad que forman estructuras de regulación (anti antiterminador-antiterminador-terminador o anti secuestrador-secuestrador) como la secuencia que reconoce al metabolito, son secuencias cortas

Al hacer una búsqueda en una biosíntesis específica podemos suponer que el producto metabólico de la biosíntesis o alguna molécula relacionada actúen como molécula efectora en la regulación. Como sé a observado en todos los casos de Riboswitches, los genes regulados están siempre relacionados a la síntesis y transporte de un metabolito, por lo que la búsqueda especifica nos permite asociar una función relacionada a la síntesis o transporte de un metabolito particular a genes que no tienen una función reportada. Este tipo de asociación de función puede dar la pauta para llevar a cabo estudios más dirigidos de función genética.

Un ejemplo asociación de función es el caso del gen *ypaA* que presenta asociado el elemento RFN, *ypaA* estaba reportado para *B. subtilis* como un gen que codifica para una proteína sin función asociada, que presenta cinco segmentos transmembranales [22]. La estructura transmembranal y la asociación del elemento RFN a su regulación permitió asociarle una función como transportador de Riboflavina o de compuestos relacionados, de forma paralela esta predicción fue probada experimentalmente por los mismos autores que asociaron la función al gen y se probo que *ypaA* es un transportador de Flavinas.

El análisis de secuencias de regulación tipo Riboswhitch, en el contexto de una biosíntesis nos puede ayudar a definir mejor los procesos metabólicos de los organismos.

7. Conclusiones

El método aplicado es eficiente para la búsqueda de secuencias de regulación en genes relacionados a una misma vía metabólica

Los cinco casos reportados son un ejemplo de que las secuencias de regulación tipo "Riboswhitch", están altamente conservadas en los genes de la biosíntesis de vitaminas, ácidos nucleicos, cofactores y aminoácidos.

8. Perspectivas

1. **Ampliar la búsqueda a aminoácidos.** Considerando que la atenuación es un mecanismo de control que se da en varios operones de rutas biosintéticas (sobre todo de aminoácidos), es muy probable que las biosíntesis de aminoácidos, tengan regulación de tipo "Riboswhitch", como es el caso de la Lisina y la S-methionina.
2. **Hacer una búsqueda dependiente de cada grupo taxonómico**. Cada grupo tiene una composición genómica particular, por lo que es más congruente comparar contra bases de datos formadas en base a las características de cada Taxa.
3. **Calcular la estructura secundaria de las secuencias de regulación** La secuencia secundaria pueden ser comparadas directamente con las ya reportadas, ya sea para corroborar la estructura propuesta o para proponer una nueva, además de permitirnos establecer las relaciones de complementariedad en la secuencia y su versatilidad para tener cambios conformacionales, lo cual es una característica crucial en la regulación por Riboswitches
4. **Comprobar la Regulación.** Las secuencias regulatorias predichas que se obtuvieron en base a los datos geonómicos y el análisis en silico, deben de ser evaluados en un contexto de datos experimentales.
5. **Comparar en función de los genes agrupados en KEGG.** Se conoce que la base de datos KEEG [45], contiene grupos de genes de acuerdo a la vía metabólica a la que pertenecen, en este caso se podría sustituir el grupo de genes localizados por palabra clave, por el grupo de genes de las rutas metabólicas que se quieran estudiar, sin embargo hay que considerar que la retroalimentación que se hace al comparar los motivos conservados contra todos los genes en el proceso de MAST compensa que en un principio no se seleccionen todos los genes de una vía
6. **Análisis de genes no caracterizados**. Los genes que no tiene una función definida pero que son identificados en el proceso de búsqueda, pueden tener una función asociada a la biosíntesis estudiada. Por lo que se pueden aplicar análisis de homología, de estructura y de función tanto a nivel genómico, como a nivel experimental para definir la función.

"The incredible diversity of life on this planet, most of which is microbial, is best understood in an evolutionary framework" -- Carl Woese, 2000

9. Referencias

1. Watson, J.D. and Crick, F.H. (1953) Genetical implications of the structure of deoxyribonucleic acid. *Nature*. 171:964.
2. Venter, J. C. *et al.* (2001) The secuence of the Human Genome. *Science*. 291:1304-1351
3. The International Human Genome Mapping Consortium. (2001) A physical map of the human genome. *Nature*. 409:934-941
4. Alberts B., Bray D., Lewis J., Raff M., Roberts K., Watson J. D. (1996). Biología Molecular de la célula. 3a. edición. Editorial Omega.
5. Nelson D. L. and Cox M. M. (2001). Lehninger Principios de Bioquimica. 3a. edición, Editorial Omega.
6. Henkin T. M. and Yanofski Ch. (2002) Regulation by transcription attenuation in bacteria: howRNA provides instructions for transcription termination/antitermination decisions. *BioEssays* 24:700-707
7. Vellanoweth, R.L. (1993) Translation and its regulation. Bacilllus Subtilis and Other Gram-Positive Bacteria. (Sonenshein A. L., ed. In chief), pp. 699-711.
8. Chain, P. et al. (2003) An applications-focused review of comparative genomics tools: Capabilities, Limitations and future changes. *Briefings in bioinf.*. 4(2):105-123
9. Jeanmougin and Thompson. (1998) Multiple sequence aligment with Clustal X. *Comput. Corner.* Tib 23, 403-405.
10. Dumas, J. and Ninio, J. (1982) Efficient algorithms for folding and comparing nucleic acid sequences. *Nucleic Acids Res.* 10:197-206
11. Bailey, T.L. and Elkan, Ch. "Fitting a mixture model by expectation maximization to discover motifs in biopolymers", Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology, pp. 28-36, AAAI Press, Menlo Park, California, 1994.
12. Bailey, T.L. and Gribskov, M. (1998) "Combining evidence using p-values: application to sequence homology searches". *Bioinformatics,* 14:48-54.
13. Nou, X. and Kadner, R. J. (1998) Couple Changes in Translation and Transcription during Cobalamin-Dependent Regulation of *btuB* Expresion in *Esch. coli. J Bacteriol.* 180:6719-6728.
14. Nou, X. and Kadner, R. J. (2002) Adenosylcobalamin inhibits ribosome binding to *btuB* RNA. *Proc. Natl. Acad. Sci. U.S.A.* 97:7190-7195
15. Ravnum, S. and Andersson, D.I. (2001). An adenosyl-cobalamin (coenzyme-B12)-repressed translational enhancer in the cob mRNA of *Salmonella typhimurium*. *Mor Microbiol.* 39(6):1585-1594.
16. Nahvi, A. *et al.* (2002). Genetic Control by a Metabolite Binding mRNA. *Chemm Biol.* 9:1043-1049.
17. Vitreschak, A.G *et al.* (2003). Regulation of the vitamin B_{12} metabolism and transport in bacteria by a conserved RNA structural element. *RNA*. 9:1084-1097.
18. Miranda-Rios, J. *et al.* (1997). Expression of thiamin biosynthetic genes (thiCOGE) and production of symbiotic terminal oxidase cbb (3), in Rhizobium etli. *J. Bacteriol.* 179: 6887-6893.
19. 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. USA.* 98: 9736-9741.
20. Rodionov, D.A. *et al.* (2002). Comparative Genomics of Thiamin Biosintesis in Procaryotes. *J. Biol. Chem.* 277:48949-48959.
21. Winkler, W. *et al.* (2002). Thiamine derivatives bind messenger RNAs directly to regulate bacterial gene expression. *Nature.* 419:952-956.
22. Gelfand, M.S. *et al.* (1999). A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Genome Analysis.* 15:439-442.
23. Vitreschak, A.G. *et al.* (2002). Regulation of riboflavin biosynthesis and transport genes in bacteria by transcriptional and traslational attenuation. *Nucleic Acids Res.* 30:3141-3151
24. Winkler, W.C. *et al.* (2002). An mRNA structure that controls gene expression by binding FMN. *Proc. Natl. Acad. Sci. U.S.A.* 99(25):15908-15913.
25. Mironov, A.S. *et al.* (2002). Sensing small molecules by nascent RNA: A mechanism to control transcription in bacteria. *Cell* 111:747-756.

26. Murphy, M.B. *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(6):3083-3088
27. Epshtein, V. *et al.* (2003) The riboswitch-mediated control of sulfur metabolism in bacteria. *Proc. Natl. Acad. Sci. U.S.A.* 100(9):5052-5056
28. Sudarsan, N. *et al.* (2003) Metabolite-binding RNA domains are present in the genes of eukaryotes. RNA. *RNA.* 9:644-647.
29. Lai, E. C. (2003) RNA Sensors and Riboswitches: Self-Regulating Messages. *Current Biology.* 13:R285-R291.
30. Patte, J.C. *et al.* (1998) The leader of the *Escherichia coli lycC* gene is involved in the regulation of LysC synthesis. *FEMS Microb. Lett.* 169:165-170
31. Rodionov, D.A. *et al.* (2003) Regulation of lysine byosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res.* 31(23):6748-6757
32. Grundy, F.J. *et al.* (2003) The L box regulon: Lysine sensing by leqader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci. U.S.A.* 100(21):12057-12062
33. Mandal, M. *et al.* (2003) Riboswitches control fundamental biochemical pathways in *Bacillus subtilis* and other bacteria. *Cell.* 113:577-586
34. Tomchick D. R., et al. (1998) Adaptation of an enzyme to regulatory function: structure of Bacillus subtilis PyrR, a *pyr* RNA-binding attenuation protein and uracil phosphoribosyltransferase. *Structure.* 6(3):337-350
35. Grundy, F.J. and Henkin, T. (2004) Regulation of gene expresión by effectors that bind to RNA. *Curr. Opin. Microbiol.* 7:1-6
36. Vitrschak, A.G. *et al.* (2004) Riboswitches: the oldest mechanism for the regulation of gene expresión? *Trends Gent.* 20(1):44-50
37. Nudler, E. and Mironov A.S. (2004) The Riboswhitch control of bacterial metabolism. *Trends Biochem Sci.* 29(1):11-16
38. Barrick, J.E. *et al.* (2004) New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control *Proc. Natl. Acad. Sci. U.S.A.* 101(17):6421-6426
39. Winkler et al. (2004) Control of gene expresión by a natural metabolite-responsive ribozyme Nature. 428:281-286
40. Gen Bank, the NIH genetic sequence database, an annotated collection of all publicly avaible DNA sequences (http://www.ncbi.nlm.nih.gov/)
41. Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science.* 24, 631-637
42. Salgado, H. *et al.* (2000) Operons in Escherichia coli: genomic analyses and predictions. *Proc. Natl. Acad. Sci. U.S.A.* 97:6652-6657
43. Moreno-Hagelsieb, G. and Collado-Vides, J. (2002) A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics.* 18, 329-336
44. Altschul, S.F. *et al.* (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acid Res.* 25, 3389-3402
45. Kanehisa, M. and Goto, S. (2000) KEGG: Kyoto Enciclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27-30

10. Anexo I. Artículo

Mi colaboración en este trabajo, fue principalmente aportar un principio de búsqueda, ya que se proobo que el proceso llevado a cabo es eficiente para encontrar secuencias de regulación. Durante el proceso participe en el análisis de los resultados, tanto en los casos de secuencias "Riboswhitch", como el la búsqueda de regulación asociada a los motivos encontrados, de igual forma, participe de forma conjunta con los autores en el análisis y preparación de este trabajo.

Conserved regulatory motifs in bacteria: riboswitches and beyond

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

Keywords: Riboswitches. Conserved motifs. Comparative genomics

Abstract

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

Comparative genomic approaches are central to coping with the increasing number of whole-genome sequences. Although the application of 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]. This stands to reason since sequence conservation of functional intergenic regions (promoters, protein binding sites) is usually low, and quickly diverges. It came as a surprise to many when specific RNA "riboswitches" were shown to be capable of regulating gene expression by directly sensing a metabolite without the intervention of any protein [4]. They have since been discovered to be involved in various metabolic processes including thiamine, riboflavin, cobalamine, adenine, guanine and lysine biosynthesis [5-10, reviewed in 11]. We assumed that this type of regulatory sequence would be especially prone to be 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 of proteins database [12] together with operon predictions based on intergenic distances [13]. In this manner, every protein from 141 fully-sequenced bacterial genomes that was associated to a COG was assigned the upstream intergenic region of the first gene of the operon to which it belongs. To avoid over-representation of very similar sequences from related genomes, redundant sequences were eliminated. We thus end up with just over 4,000 clusters of orthologous regulatory regions, each assigned to a different COG.

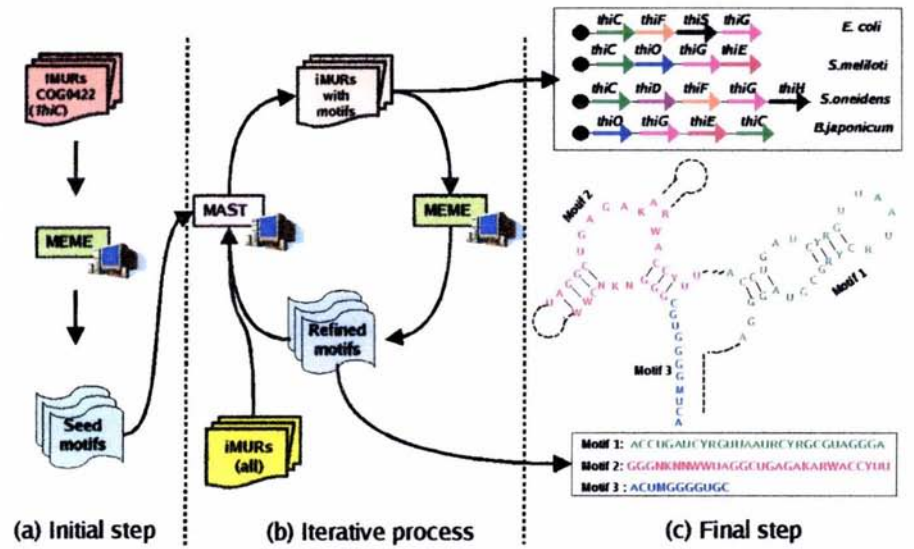


Fig 1. General procedure to identify conserved regulatory motifs. (a) The intergenic Minimal Upstream Regions (iMUR) of the operons that code for each COG (in this example ThiC, COG0422) from fully sequenced genomes are grouped and over-represented motifs identified using MEME [14]. These motifs constitute 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]. In order 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.

We used the public domain motif discovery tool MEME [14] to find a set of over-represented "seed motifs" for each COG (Fig 1a). These motifs were used to identify other members of the putative regulon by searching in all upstream regions using MEME's counterpart MAST [15]. As a result of this search, new members were added to each group (and some original ones were lost), 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 performed recursively until no new members were found (Fig 1b). The resulting "refined motifs" are our candidate regulatory elements. The putative regulons thus obtained were then clustered, resulting in 563 different groups (see: http://www.ibt.unam.mx/biocomputo/conserved_motifs.html). It is worth mentioning that each group can have several "refined motifs" (up to 4), and that the motifs are larger than usual protein binding sites (our motif's average length is 43 nucleotides). To give an example, the "thiamine riboswitch" group was obtained from 20 different COGs that actually converge to a common set of genes. Upon analysis of the three different motifs that define this group, we found that the most significant one contains the actual *thi* box, and the next two overlap with part of the structure of the THI-element (Fig 1c).

Evaluating the motifs

In order to evaluate the likelihood that the "refined motifs" in our groups represent biologically important regulatory elements, we first eliminated those that matched against known proteins (from the *nr* GenBank database) or RNA genes (rRNA, tRNA, scRNA, snRNA, etc.) from fully sequenced genomes. This step was necessary given the existence of missing small genes or erroneously assigned translation starts that occur during automatic genome annotation. For example, ribosomal protein L36 (~40 aa long) is not annotated in over 15 genomes and is strong enough as a signal to be picked up by our method, when analyzing ribosomal proteins S13 and S11 that are downstream from L36. The "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 KEGG pathway [16]. Motifs with *p*-values smaller than $1x10^{-15}$ were considered biologically relevant for the regulation of that pathway. A similar evaluation was performed using COG assignations. About 94% of our groups had *p*-values below $1x10^{-6}$ for either KEGG or COG enrichment. Some of the other 6% are cases of scarce annotation and might still represent biologically relevant elements that control mainly functionally unknown genes. Lastly, the genome context congruence of our putative regulons was verified using *GeConT* [17]. The data in our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html) is hyperlinked to this application so that the co-regulated groups can easily be visualized.

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

Description[b]	Operons / Organisms / Phyla[c]	Representative KEGG pathways[d] (*p*-value)[f]	Representative COGs[e] (*p*-value)[f]	Ref
tRNA synthetase T-box	286 / 27 / 1 (Firmicutes)	00970 Aa-tRNA biosynthesis ($1x10^{-236}$) and other aminoacid pathways	tRNA synthetases: COG0172 Seryl, COG0441 Threonyl, COG0013 Alanyl, COG0060 Isoleucyl, COG0162 Tyrosyl, COG0180 Tryptophanyl, COG0525 Valyl (all below $4x10^{-21}$)	18
Thiamine riboswitch	189 / 91 / 10	00730 Thiamine metabolism ($2x10^{-145}$)	COG042 ThiC ($4x10^{-122}$) COG0351 ThiD ($4x10^{-82}$) COG4143 TbpA ($2x10^{-65}$) COG0352 ThiE ($3x10^{-62}$) COG2145 ThiM ($1x10^{-60}$)	4,5
Methionine riboswitch	145 / 27 / 7	00271 Methion. metabolism ($8x10^{-86}$)	COG1135 MetN ATPase ($7x10^{-47}$) COG1464 MetQ periplasm ($6x10^{-38}$) COG2011 MetI permease ($4x10^{-37}$) COG0192 MetK SAM-synthetase ($9x10^{-30}$)	8

Cobalamin riboswitch	133 / 57 / 10	00860 Porphyrin, and chlorophyll metab. ($1x10^{-45}$)	COG4206 Cobalamin receptor ($2x10^{-62}$) COG0614 Cob. trans. periplasm ($1x10^{-30}$) COG0609 Cob. trans. permease ($5x10^{-26}$) COG1120 Cob. trans. ATPase ($1x10^{-25}$) COG2087 CobP ($1x10^{-24}$) COG0620 MetE cob-independent ($1x10^{-20}$)	6
IS200 hairpin	97 / 14 / 4		COG1943 Transposases ($3x10^{-171}$)	24
Riboflavin riboswitch	91 / 73 / 6	00740 Riboflavin metabolism ($2x10^{-111}$)	COG0108 RibB ($1x10^{-103}$) COG3601 Membrane ($1x10^{-85}$) COG0307 RibE ($1x10^{-57}$) COG0054 RibH ($2x10^{-53}$) COG1985 RibD ($1x10^{-32}$)	7
Glycine cleavage	59 / 51 / 4	00260 Glycine metabolism ($1x10^{-57}$)	COG0404 Glycine cleavage T ($3x10^{-58}$) COG1003 Glycine cleavage P ($2x10^{-50}$) COG0403 Glycine cleavage P ($5x10^{-33}$)	
Pyrimidine metabolism (PyrR site)	54 / 25 / 2 (Firmicutes)	00240 Pyrimidine (5x10-85) 00252 Alanine and aspartate ($2x10^{-17}$)	COG2065 Pyrimidine attenuation ($6x10^{-54}$) COG0540 Aspartate carbamoyl ($4x10^{-35}$) COG0284 Orotidine-5-P decarb ($6x10^{-30}$) COG2233 Xanthine/uracil perm ($4x10^{-28}$) COG0461 Orotate p-ribosyltrans ($1x10^{-24}$)	21
Heat shock CIRCE hairpin	53 / 33 / 5 (Firmicutes)		COG1420 Regulator of HS ($1x10^{-44}$) COG0234 GroES HSP10 ($3x10^{-43}$) COG0459 GroEL HSP60 ($3x10^{-40}$) COG0576 GrpE ($5x10^{-35}$)	22
Copper transport	50 / 31 / 3 (Proteobact.)		COG2217 Cation transp. ATPase ($9x10^{-59}$) COG2608 Copper chaperone ($1x10^{-28}$)	
K^+ -transporting ATPase operon	42 / 40 / 5		COG2060 KdpA ($4x10^{-130}$) COG2216 KdpB ($2x10^{-90}$) COG2156 KdpC ($2x10^{-79}$) COG2205 KdpD ($1x10^{-37}$)	25
Threonyl tRNA synthetase	33 / 30 / 1 (Proteobact.)	00970 Aa-tRNA biosynt. ($5x10^{-42}$) 00260 Threon. metab. ($1x10^{-40}$)	COG0441 Threonyl tRNA synthetase ($9x10^{-91}$)	19
Glt,gln tRNA synthetases	28 / 17 / 2 (Proteobact.)	00251 Glutamate metab. ($1x10^{-35}$) 00970 Aa-tRNA biosynt. ($9x10^{-34}$)	COG0008 Glutamyl- and glutaminyl-tRNA synthetases ($4x10^{-82}$)	
Ribosomal operon with L4 auto-regulation	29 / 29 / 3 (Firmicutes)	03010 Ribosome ($5x10^{-54}$)	Ribosomal proteins: COG0051 S10, COG0087 L3, COG0088 L4, COG0089 L23, COG0090 L2 (all below $1x10^{-40}$)	20
Ribosomal operon with L4 auto-regulation	19 / 19 / 3 (Proteobact. Chlamydiae)	03010 Ribosome ($2x10^{-24}$)	Ribosomal proteins: COG0087 L3, COG0088 L4, COG0089 L23, COG0090 L2, COG0185 S19, COG0091 L22, COG0092 S3, COG0197 L16 (all below $5x10^{-28}$)	20

[a]A table containing the entire list of motifs can be found at http://www.ibt.unam.mx/biocomputo/conserved_motifs.html

[b]Known or probable regulatory system.

[c]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 KEGG [16]) of the genes containing the motifs.

[e]Representative groups of orthologous genes (as defined by 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.

Analyzing the nature of the conserved motifs

The most relevant groups of conserved motifs in our study correspond to previously described riboswitches, known to regulate genes involved in the biosynthesis of different metabolites such as thiamine, riboflavin, cobalamine, adenine, guanine and lysine, as well as 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. In the case of *E. 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 very stable structure which is similar to the acceptor arm of $tRNA^{Thr}$. Although it hasn't been reported, the motifs that we find for glutamyl and glutaminyl-tRNA synthetases could participate in a similar mechanism. An even larger set of significant groups correspond to ribosomal protein operons. In fact, we detect 29 different groups of such operons, most of which correspond to specific phyla (such as the last two groups in Table 1). Autogenous regulation has been described for these cases, as ribosomal protein L4 is known to bind to its operator, where a complex secondary structure appears to mimic L4's natural binding site in the ribosome [20]. We expect that most, if not all of these operons, are auto-regulated by one or more of their highly conserved proteins. Other well described cases in Table 1 include a pyrimidine biosynthesis group, were 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 a thermo-sensor hairpin [22]. In all these cases (as occurs with riboswtiches), sequence conservation in the regulatory region is a consequence of the constraints imposed by the required RNA structure. Glycine cleavage is the last case we shall mention. 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 alpha and beta proteobacteria, and most of them do not even 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, our signal picks up several proteins assigned to a Na^+/alanine symporter COG. This could very well 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 in our web page (http://www.ibt.unam.mx/biocomputo/conserved_motifs.html).

Concluding remarks

We have developed a computer method able to identify previously reported riboswitches, other known conserved elements, as well as more than 500 groups of conserved motifs that would appear to be biologically relevant, as far as our statistical analysis could tell. We thus show that for a great many regulatory elements, their conservation is strong enough to be detected in a single orthologous cluster of genes, without the need to initially increase the signal to noise ratio by adding elements from a known regulon or metabolic pathway, since the probable regulon can be reconstructed afterwards. In many cases, our motifs coincide with regulatory elements reported for specific model organisms such as *E. coli* or *B. subtilis*. We are now able to propose the extent to which these systems have been conserved among fully sequenced bacteria. Most of the signals analyzed were found to be related to a RNA secondary structure, which is consistent with our initial focus of searching for riboswitches. Our method identified as statistically relevant more regulatory systems that depend on RNA sensors than classical

DNA-binding regulators. This is probably due to the fact that the former could be older, more structure-dependent regulatory elements, and therefore would be more conserved.

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

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 Bioinf.* 4, 1471-2105
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, 5908-15913
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 Gent.* 20, 44-50
12. Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science.* 24, 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. In *Proceedings of the 2nd International Conference on ISMB.* pp. 28-36, AAAI Press
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 Enciclopedia of Genes and Genomes. *Nucleic Acids Res.* 28, 27-30
17. Ciria, R. *et al.* (2004) 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* (2nd edn) (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. Zuber, U. and Schumann, W. (1994) CIRCE, a novel heat shock element involved in regulation of heat shock operon dnaK of Bacillus subtilis. *J. Bacteriol.* 176, 1359-1363
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. Baceriol* 177, 4940-4946

24. Beuzon, C.R. and Casadesus, J. (1997) Conserved structure of IS200 elements in Salmonella. *Nucleic Acids Res.* 25, 1355-1361
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