



---

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS

INSTITUTO DE BIOTECNOLOGÍA

# Identificación de la Función de los Genes a partir de la Especificidad del Riboswitch T box

---

TESIS QUE PARA OPTAR POR EL GRADO DE DOCTOR EN CIENCIAS

PRESENTA:

Ana Gutiérrez Preciado

DIRECTOR DE TESIS:

Enrique Merino

**INSTITUTO DE BIOTECNOLOGÍA**

COMITÉ TUTOR:

Dr. Miguel Lara Flores

CENTRO DE CIENCIAS GENÓMICAS

Dr. Juan Miranda Ríos

INSTITUTO DE INVESTIGACIONES BIOMÉDICAS

Cuernavaca, Morelos, México. Junio 2013



Universidad Nacional  
Autónoma de México



**UNAM – Dirección General de Bibliotecas**  
**Tesis Digitales**  
**Restricciones de uso**

**DERECHOS RESERVADOS ©**  
**PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL**

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

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

## Índice

<b>DEDICATORIA</b>	<b>7</b>
<b>AGRADECIMIENTOS</b>	<b>8</b>
<b>RESÚMEN</b>	<b>9</b>
<b>ABSTRACT</b>	<b>12</b>
<b>INTRODUCCIÓN</b>	<b>14</b>
<b>La Regulación Génica en sus inicios</b>	<b>14</b>
<i>El descubrimiento de la regulación: el ambiente ejerce influencia sobre la respuesta metabólica</i>	14
La regulación de la expresión génica puede ser mediada por proteínas.	16
Atenuación transcripcional: regulando la expresión génica en base a elementos de RNA en <i>cis</i>	17
<b>Regulación más allá de las proteínas regulatorias</b>	<b>20</b>
<i>El descubrimiento del primer riboswitch: La T box</i>	20
<i>Otros riboswitches: el reconocimiento de metabolitos mediante RNA en lugar de proteínas</i>	24
<b>ANTECEDENTES</b>	<b>27</b>
<b>Identificación de elementos Regulatorios</b>	<b>27</b>
<b>Regulación por RNA</b>	<b>30</b>
Perspectiva Histórica: El mundo de RNA	30
RNAs no codificantes (reguladores en <i>trans</i> ): una contribución importante a la red de regulación bacteriana	32
Riboswitches: Regulando el metabolismo bacteriano	33
El Riboswitch T box	35
Tipos de riboswitches.	39
Distribución filogenética de los riboswitches	40
Riboswitches que reconocen aminoácidos	40
<b>OBJETIVO</b>	<b>44</b>
<b>Objetivos Particulares:</b>	<b>44</b>
<b>METODOLOGÍA</b>	<b>46</b>
<b>Estado del Arte de herramientas en bioinformática utilizadas en esta tesis: Una pequeña revisión</b>	<b>46</b>
RFAM	46

INFERNAL	47
MEME	48
MAST	50
<b>The Riboswitch Approach</b>	<b>51</b>
<b>Programa que construye una base de datos <i>ad hoc</i> que contiene todas las regiones intergénicas de los genomas completamente secuenciados</b>	<b>56</b>
<b>RESULTADOS Y DISCUSIÓN</b>	<b>58</b>
<b>Características bioquímicas y evolutivas del regulón T box</b>	<b>58</b>
Contribuciones puntuales de este artículo	58
Origen evolutivo del elemento T box	59
Características del codón de especificidad	61
Aminoacil tRNA sintetasas como parte del regulón T box.	62
Regulación por T box en genes de biosíntesis de aminoácidos	67
Regulación por T box en genes de transporte de aminoácidos	69
Los genes de biosíntesis y de transporte de aminoácidos son regulados por el mismo mecanismo	70
Regulación por T box de proteínas regulatorias	70
<b>Interpretación del metabolismo microbiano en base a su regulación</b>	<b>72</b>
La asignación de funciones de los genes recientemente secuenciados	72
La aproximación riboswitch / The Riboswitch Approach	73
Resultados: The Riboswitch Approach	74
Dos ejemplos en Clostridias anaeróbicas: el operón <i>por</i> y el operón <i>etf</i> . Ambos dentro del metabolismo de los aminoácidos de cadena ramificada	76
Casos involucrados en la biosíntesis de aminoácidos aromáticos	80
Conclusiones: The Riboswitch Approach	81
<b>Riboswitches en microorganismos de Cuatrociénegas</b>	<b>82</b>
¿Por qué Cuatrociénegas?	82
¿Por qué Bacillaceas?	83
<i>Bacillus coahuilensis</i> y <i>Bacillus</i> M3-13	83
Riboswitches	85
<i>Uso total de los riboswitches</i>	97
<b>Riboswitches en <i>Epilopiscium</i></b>	<b>102</b>
<b>CONCLUSIONES</b>	<b>104</b>
<b>PERSPECTIVAS</b>	<b>107</b>
1.	
<b>r la expansión del regulón T box mediante evolución experimental</b>	<b>Entende 107</b>



<i>Características estructurales del mecanismo de regulación T box: un codón provee las bases para regular específicamente en respuesta a un aminoácido.</i>	107
<i>Requerimientos en el RNA líder</i>	108
<i>Requerimientos del tRNA</i>	109
<i>Duplicación y expansión del regulón T box</i>	110
<i>¿Por qué abordar este proyecto con evolución experimental?</i>	112
<i>Metodología propuesta</i>	113
<b>2.</b>	<b>leuA en</b>
<b><i>Geobacter sulfurreducens</i></b>	<b>117</b>
Metodología propuesta	117
Resultados a esperar:	118
<b>3.</b>	<b>Rellenar</b>
<b>los huecos Regulatorios del riboswitch T box en el gen <i>trpS</i></b>	<b>120</b>
Introducción	120
Metodología empleada en una primera aproximación	125
Resultados parciales	128
Conclusiones puntuales	131
<b>4.</b>	<b>Web</b>
<b>server con el regulon T box</b>	<b>139</b>
<b>5.</b>	<b>Imbalan</b>
<b>ce: <i>hisS-aspS</i> regulation</b>	<b>141</b>
Introducción	141
Metodología propuesta	143
<b>6.</b>	<b>Regulaci</b>
<b>ón de operones de Triptofano</b>	<b>144</b>
Diferencias en la organización de los operones de triptofano de <i>E. coli</i> y <i>B. subtilis</i> y las diferentes estrategias que usan para su regulación.	145
Comparación de las estrategias regulatorias.	154
<b><i>trpS</i> termosensible en <i>Bacillus halodurans</i></b>	<b>157</b>
<b><i>TRAP</i> en <i>Oceanobacillus iheyensis</i></b>	<b>158</b>
<b><i>Clostridias</i> con represores de triptófano.</b>	<b>158</b>
<b>APÉNDICES</b>	<b>159</b>
<b>Apéndice I: Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism.</b>	<b>159</b>

<b>Apéndice II: Genome Sequence Databases: Types of Data and Bioinformatic Tools.</b>	<b>160</b>
<b>Apéndice III: An Evolutionary Perspective on Amino Acids</b>	<b>161</b>
<b>AN EVOLUTIONARY PERSPECTIVE ON AMINO ACIDS</b>	<b>161</b>
<b>The Origins of Nutrient Biosynthesis</b>	<b>162</b>
<b>What Is an Amino Acid Made Of?</b>	<b>163</b>
<b>Amino Acid Precursors and Biosynthesis Pathways</b>	<b>165</b>
<b>What Makes an Amino Acid Essential?</b>	<b>167</b>
<b>Tryptophan Synthesis: Only Created Once</b>	<b>168</b>
<b>Lysine Synthesis: Created Multiple Times</b>	<b>168</b>
<b>Synthesis on the tRNA molecule</b>	<b>169</b>
<b>How Do Metabolic Pathways Evolve? Two Different Models</b>	<b>170</b>
<b>Open Questions about Amino Acid Evolution</b>	<b>172</b>
<b>Summary</b>	<b>172</b>
<b>References and Recommended Reading</b>	<b>173</b>
<b>Apéndice IV: Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach</b>	<b>174</b>
<b>Apéndice V: Verificación de la hipótesis de la reina negra sobre la reducción genómica de las metanoarqueas</b>	<b>175</b>
Introducción	175
Firmicutes y Archaeas; Sintrofia y Metanogénesis.	176
Antecedentes	177
Hipótesis	179
Objetivo	179
Metodología Propuesta	179
Resultados	182
<b>BIBLIOGRAFÍA</b>	<b>188</b>

Este trabajo se desarrolló en el Departamento de Microbiología Molecular del Instituto de Biotecnología de la Universidad Autónoma de México, bajo la tutoría del Dr. Enrique Mérimo Pérez.

El proyecto fue financiado por los siguientes donativos: CONACyT beca de doctorado, número de becario: 220734. CONACyT 60127 "Identificación in silico y caracterización molecular, estructural y cinética de elementos de regulación de la expresión genética basada en riboswitches". CONACyT 58840 "Generación y comparación de modelos de regulación transcripcional para la red de *Escherichia coli* y *Bacillus subtilis* y su consistencia con análisis de datos de expresión global y experimentos de PCR en tiempo real". DGAPA IN215808 "Generación de modelos de regulación de la transcripción en *Bacillus subtilis* y *Escherichia coli*, análisis comparativo y su corroboración experimental". DGAPA IN212708 "Análisis de curvatura estática del DNA genomas y metagenomas.

## Dedicatoria

A mi mamá, que me enseñó a seguir mis sueños y a hacer las cosas que me hacen feliz.

A mi hermano por ser también amigo, y por siempre hacerme reír.

A Andrés por ser el mejor apoyo que podría pedir y hacerme feliz.

A mis amigos: Pedro, Pablo, Tania, Diana, Ale, la Peimbert, la Bruja, Marel, Charly, Agus, Cata, Aubin y Adolfo por ser toda la otra parte de mi vida que no tenía nada que ver con el lab. Por distraerme, apoyarme y apapacharme.

A Enrique, a David y a Valeria por ser tutores y por ser amigos al mismo tiempo.

## Agradecimientos

Al Programa de Doctorado en Ciencias Biomédicas

A Enrique Merino por formarme, por darme libertades dentro de mi proyecto y guiarme, orientándome siempre a hacer mejor la ciencia, por su apoyo constante e incondicional y por su confianza.

A Charles Yanofsky por su entusiasmo durante mi proyecto, y por sus largos emails donde siempre seguí aprendiendo de él.

A Tina Henkin por ser amiga y apoyo constante. Por orientarme y acotar los proyectos.

A Ricardo Ciria por su amistad, y por todo su apoyo bioinformático y consejos al programar. A Ale Abdala y a Ricardo por facilitarme la vida con la base de datos (y por enseñarme MySQL).

A Rosa María Gutiérrez, Katy Juárez y Ceil Abreu por la amistad y las discusiones en este proyecto (y otros).

A Tim Bailey por sentarse conmigo a modificar el programa MAST.

A Juan Miranda y a Miguel Lara, por hacer de mis tutorales unas reuniones amenas, amigables donde pudimos discutir mi proyecto cada semestre. Ellos hicieron importantes observaciones y contribuciones a esta tesis, siempre en un ambiente agradable.

A mi hermano Pedro y a Pau Juárez, por ayudarme a presentar mejor las figuras de esta tesis, de los artículos y los posters presentados en congresos.

A Jalil Saab por resolverme problemas, por su apoyo y amistad.

A Rosalva González y Gladys Aviles por todo su apoyo y paciencia.

## Resumen

La T box es un elemento de regulación comúnmente usado para modular la expresión de los genes relacionados con el metabolismo de aminoácidos en las bacterias Gram-positivas, especialmente en los Firmicutes. La regulación por T box normalmente se basa en atenuación transcripcional, donde un tRNA descargado interactúa con la región 5' del transcrito, estabilizando una estructura de antiterminación. Este antiterminador previene la formación de un terminador rho-independiente, por lo que la transcripción puede continuar. Aunque los elementos regulatorios T box, mayoritariamente están presentes en monocopia, se han observado estos elementos repetidos en tándem dos y hasta tres veces, expandiendo así el rango regulatorio de estos elementos. Considerando la distribución filogenética de las T boxes, pensamos que este elemento regulatorio se originó en un ancestro común a los Firmicutes, Chloroflexi, Deinococcus-Thermus y Actinobacteria, y que fue transferido horizontalmente a las  $\delta$ -Proteobacteria. La T box controla la expresión de genes relacionados con la biogénesis de los aminoácidos tales como los genes de aminoacil tRNA sintetasas (aaRS), de genes de biosíntesis de aminoácidos, de genes de transporte de aminoácidos y en algunos casos de genes que codifican para proteínas regulatorias de vías biosintéticas de aminoácidos. En algunos casos, encontramos también genes regulados por T boxes cuya función es parcial o totalmente desconocida. En base a la especificidad de los diferentes tipos de T box por sus correspondientes tRNAs, fuimos capaces de proponer relaciones funcionales de estos genes desconocidos con las vías metabólicas en las que pudieran participar.

En la actual era post-genómica, sólo el 3% de todos los genes se han anotado sobre la base de la evidencia experimental. A pesar de que las funciones

fácilmente se pueden predecir para muchos genes, el 25% de las predicciones llegan a estar mal. Las estrategias típicas para la asignación de funciones de los genes, en base al análisis comparativo de genomas, son particularmente útiles cuando dichos genomas han sido completamente secuenciados y para los genes homólogos cuya función ha sido caracterizada. Nuestro enfoque de asignación de funciones de los genes, que se basa en la predicción de riboswitches, es una importante alternativa a los métodos clásicos, y es especialmente adecuada cuando se trabaja con secuencias de metagenómica o genomas parcialmente secuenciados. También proporciona información cuya interpretación puede ser mejor interpretada, desde un punto de vista fisiológico, ya que se basa en la relación de la expresión génica, las necesidades metabólicas y la función del gen. Con nuestro método, el papel de genes específicos en las vías metabólicas se puede predecir, aún cuando sus homólogos no tienen ninguna función asignada, o cuando simplemente no tienen homólogos discernibles. Por lo tanto, la predicción de riboswitches, contribuye a poder mapear los genes en un contexto metabólico celular o en el de una comunidad microbiana, que nos lleva de una gran cantidad de secuencias en bruto a una visión más completa de la función génica en un contexto celular mejor comprendido. En este estudio, hemos utilizado la T box como un ejemplo representativo de la asignación de funciones basado en la aproximación riboswitch. Sin embargo, la gran especificidad de todos los riboswitches por sus metabolitos correspondientes (ya sean nucleótidos, vitaminas, aminoácidos, co-factores, etc) puede ser explotado para la anotación de la función de los genes.

Esta tesis representa un avance importante no solamente en identificar elementos de regulación en genomas, si no que la riqueza de este análisis surge al poder ubicar cada elemento identificado en un contexto metabólico. Esto amplía nuestro conocimiento, no sólo del elemento de regulación *per se*, si no también de

cómo modula la célula bacteriana su metabolismo correspondiente a aminoácidos, y cómo emplea estas estrategias para incrementar su adecuación, así como cómo toma ventaja de la T box para coordinar la biosíntesis de sus aminoácidos así como de otros procesos metabólicos. Hasta donde sabemos, es la primera vez que se logra un mapeo metabólico a partir de la identificación de los elementos de regulación.



## Abstract

The T box riboswitch is the regulatory strategy of choice by Firmicutes to regulate their amino acid related metabolism. T box regulation mainly operates at the level of transcriptional attenuation, where an uncharged tRNA binds to the 5' leader region of the mRNA, promoting an antitermination structure. This antiterminator prevents the stabilization of a transcriptional terminator, and hence, transcription occurs. T box elements are usually present in monocopy, but it has been observed to appear in tandem copies, expanding the regulatory range of this mechanism. Based on the distribution of T box regulatory elements, we propose that this regulatory mechanism originated in a common ancestor of members of the Firmicutes, Chloroflexi, Deinococcus-Thermus group, and Actinobacteria and was transferred into the  $\delta$ -Proteobacteria by horizontal gene transfer. We predicted the functional implications of T box regulation in genes encoding aminoacyl-tRNA synthetases, proteins of amino acid biosynthetic pathways, transporters, and regulatory proteins. We also considered the global impact of the use of this regulatory mechanism on cell physiology. Novel biochemical relationships between regulated genes and their corresponding metabolic pathways were revealed. Some of the genes identified, such as the quorum-sensing gene *luxS*, in members of the Lactobacillaceae were not previously predicted to be regulated by the T box mechanism. Our analyses also predicted an imbalance in tRNA sensing during the regulation of operons containing multiple aminoacyl-tRNA synthetase genes or biosynthetic genes involved in pathways common to more than one amino acid. Based on the specificity of each T box to their cognate tRNA, we are able to suggest functional relationships of the regulated genes which are poorly annotated, predicting in which metabolic pathway they might be participating.

In the current post-genomic era, only 3% of all genes have been annotated based on experimental evidence. Even though functions can readily be predicted for many genes, 25% of these are likely to be wrong. The most widely used methods for function prediction rely on sequence similarity, which might be misleading in many cases. Other methods such as genomic context or phylogenetic profiles have been developed to increase gene annotation accuracy; nevertheless these are only efficient when complete genome sequences are available. Here we propose a new approach based on T box identification, and these work for metagenomic sequences or draft genomes. Moreover, with our approach, we can obtain physiological information, based on gene expression, metabolic needs and in some cases, gene function. With this methodology, the role of specific genes can be mapped in the metabolic network. This approach can be extrapolated to all riboswitches. Riboswitches are highly conserved regulators of gene expression located in the 5' untranslated region of certain genes. When transcribed they adopt three-dimensional structures that recognize their ligands with great affinity and specificity. This specificity is a key issue for our method, allowing functional assignment with great accuracy.

This thesis represents an important step not only on identifying regulatory elements in genomes, but also the wealth of this analysis is provided by our ability to map each identified element in a metabolic context. With this, our knowledge of regulation *per se* is expanded as well as our understanding on how bacteria can modulate their amino acid metabolism and employ these strategies to increase its fitness through the T box. To our knowledge, this is the first study where a metabolic map can be achieved via the identification of regulatory elements.

## Introducción

Comienzo la introducción con los primeros estudios sobre regulación metabólica, los cuales fueron el primer intento por explicar cómo una célula responde a su ambiente. Estas observaciones pioneras llevaron al descubrimiento de proteínas regulatorias y su capacidad para percibir los nutrientes disponibles y en base a eso, generar una respuesta regulatoria. Dados estos precedentes, se establece como un dogma, que la regulación génica se basa en proteínas regulatorias. Este dogma, se ve alterado con el descubrimiento de la atenuación transcripcional, donde el ribosoma que traduce el RNA mensajero (mRNA) afecta la expresión de la síntesis de aminoácidos. Recientemente, con la llegada de la secuenciación masiva, estudios de genómica comparativa permiten identificar elementos conservados de RNAs reguladores, los riboswitches, los cuales son capaces de realizar las mismas tareas que las proteínas regulatorias, i.e., percibir señales fisiológicas y en base a éstas, modular la expresión génica.

## La Regulación Génica en sus inicios

### *El descubrimiento de la regulación: el ambiente ejerce influencia sobre la respuesta metabólica*

El panorama se sitúa en 1940, donde se conocía muy poco sobre la estructura del DNA o sobre cómo el metabolismo bacteriano estaba regulado. Jacques Monod era un estudiante de doctorado cuyo proyecto versaba sobre el crecimiento bacteriano. Él observa tasas y patrones de crecimiento y cómo varían dependiendo de las distintas combinaciones de carbohidratos administrados a dichos cultivos. El efecto que él observaba se conoce hoy en día como "represión

catabólica”, pero en ese entonces le llamaban “adaptación enzimática”, un fenómeno definido por la síntesis de ciertas enzimas únicamente en la presencia de sustratos específicos<sup>1,2</sup>. Monod observó que las enzimas para el metabolismo de lactosa únicamente estaban presentes cuando la lactosa era añadida al medio. Esta observación sugería que la bacteria era capaz de alterar los componentes intracelulares para responder a la presencia de carbohidratos en el medio<sup>3</sup>. Contemporáneamente, en el laboratorio de Andrew Lwoff se descubrieron mutantes de *Escherichia coli* cuyas enzimas involucradas en el metabolismo de lactosa eran incapaces de responder a la presencia de lactosa en el medio<sup>2</sup>. Al comparar estas cepas mutantes con las cepas silvestres, Monod se dio cuenta de que diferían en los elementos que controlaban la sincronización y la cantidad de enzimas específicas sintetizadas para responder a los cambios de concentración de lactosa en el medio.

A principios de los 60s, el término “gen” fue definido como “una molécula de DNA con una estructura que le permite autoreplicarse puede, por mecanismos desconocidos, ser traducida en una estructura de cadena polipeptídica”<sup>2</sup>. Este concepto de “gen estructural” le confería estabilidad genética a las proteínas e implicaba que éstas no estaban sujetas a un control mediado por condiciones ambientales. Esto contrastaba significativamente con las observaciones hechas por Jacob y Monod sobre cómo el crecimiento bacteriano difería dependiendo de qué fuente de carbohidratos era añadida al cultivo celular, lo cual sugería que las enzimas sí cambiaban en respuesta a las condiciones ambientales. Dado que las proteínas parecían ser traducidas en el citoplasma, y no directamente del cromosoma, Jacob y Monod pensaron que esta transferencia estructural de información debía involucrar un intermediario químico que fuera sintetizado a partir de los genes. A este intermediario hipotético, lo llamaron “mensajero estructural”<sup>2</sup>. También propusieron que la tasa de síntesis protéica (transferencia de

información) debía depender ya fuera de la actividad del gen en sintetizar el mensajero, o bien, de la actividad del mensajero en sintetizar la proteína. Jacob y Monod fueron más allá, sugiriendo que estas actividades podrían ser moduladas por características de los genes, o de sus productos correspondientes (e.g., elementos que actúan en *cis*), o por factores desconocidos (e.g., elementos que actúan en *trans*) en respuesta a cambios en las condiciones ambientales<sup>2</sup>.

### **La regulación de la expresión génica puede ser mediada por proteínas.**

Jacob y Monod observaron que la tasa de síntesis de las enzimas que participan en el metabolismo de lactosa variaba coordinadamente, y que era controlada por un elemento adicional, y desconocido, que no estaba presente en los genes estructurales de estas enzimas<sup>2</sup>. La identificación y caracterización de las mutantes para la  $\beta$ -galactosidasa y para la permeasa les permitió construir un mapa de estos *loci*, y la llamaron la región *lac*<sup>2</sup>. Con esto también dedujeron que un solo gen era el responsable de gobernar la síntesis de una enzima que podía inducir o destruir las enzimas del metabolismo de lactosa (i.e., un represor de lactosa)<sup>2</sup>. Por esas mismas épocas, se estaban realizando ya investigaciones en paralelo sobre unidades transcripcionales y proteínas regulatorias. Un ejemplo de estas investigaciones fue la caracterización del operón *trp* en el laboratorio de Charles Yanofsky <sup>4</sup>. Una observación crucial fue la observación de que se sintetizaban altos niveles de enzimas biosintéticas como resultado de una baja concentración de triptófano disponible, así como una inhibición de la síntesis de enzimas como resultado de altos niveles de triptófano (el producto final de la vía biosintética) <sup>4</sup>. Para poder entender cómo funcionaba el mecanismo que detectaba las concentraciones de triptófano, se requirió caracterización posterior de la estructura del operon *trp* en *E. coli*<sup>5,6</sup>. Las investigaciones de Matsushiro sobre transducción con el bacteriofago  $\phi$ 80, lograron mapear las características del

operón *trp*, demostrando que éste era un sólo locus que codificaba para cinco polipéptidos cuya transcripción y traducción parecía estar acoplada<sup>7,8</sup>. Este conjunto de evidencias apuntaban a que debía de existir un mecanismo que pudiera transferir la información de los genes estructurales hacia las proteínas, y que esta transferencia debía de estar controlada por moléculas regulatorias específicas codificadas en genes especializados. Sin embargo, el mecanismo de acción de éstas, o su naturaleza bioquímica no era claro. Ahora sabemos que estas moléculas regulatorias son proteínas represoras, LacI para el operón *lac*, y TrpR para el operón *trp*. Estos dos mecanismos regulatorios, uno inducido por lactosa y otro reprimido por triptófano, han resultado ser paradigmas cruciales para el descubrimiento de un gran número de otros mecanismos regulatorios. Trabajos posteriores demostraron que las proteínas regulatorias no sólo pueden actuar como represores, sino también como activadores<sup>9-13</sup>.

Estos extraordinarios proyectos establecieron un nuevo dogma: las proteínas regulatorias pueden detectar la presencia intracelular de ciertos metabolitos, y ejercer una respuesta metabólica mediante la regulación de los genes para su síntesis o para su catabolismo. Sin embargo, los mecanismos precisos que gobernaban la regulación transcripcional en esos días seguían siendo desconocidos.

#### **Atenuación transcripcional: regulando la expresión génica en base a elementos de RNA en *cis***

A principios de los 60s, Bruce Ames y sus colegas estudiaban el operón *his* de *Salmonella* demostrando que su transcripción no era regulada por una proteína represora, y que la histidina no era detectada de forma directa. Por el contrario, el tRNA de histidina parecía ser la molécula clave involucrada en obtener un desenlace regulatorio<sup>14</sup>. Este resultado inesperado planteó la pregunta de si el

tRNA de triptófano tendría un rol en la regulación del operón *trp*. Para contestar esta pregunta, Ron Baker y Charles Yanofsky construyeron una cepa de *E. coli* que carecía del represor *trp* y observaron que esta cepa aún era capaz de responder a los niveles de triptófano, sugiriendo que un segundo mecanismo regulatorio, distinto al descrito previamente donde TrpR mediaba la represión transcripcional, debía existir<sup>15</sup>. Consistentemente, Fumio Imamoto observó que al agregar triptófano a un cultivo que por generaciones había sido limitado en triptófano, se transcribía sólo la parte inicial del operón *trp*<sup>16</sup>. Esto sugería que la transcripción era terminada de forma prematura, y que este evento de término, designado como atenuación transcripcional, era regulado en respuesta a cambios en la disponibilidad del triptófano<sup>17,18</sup>.

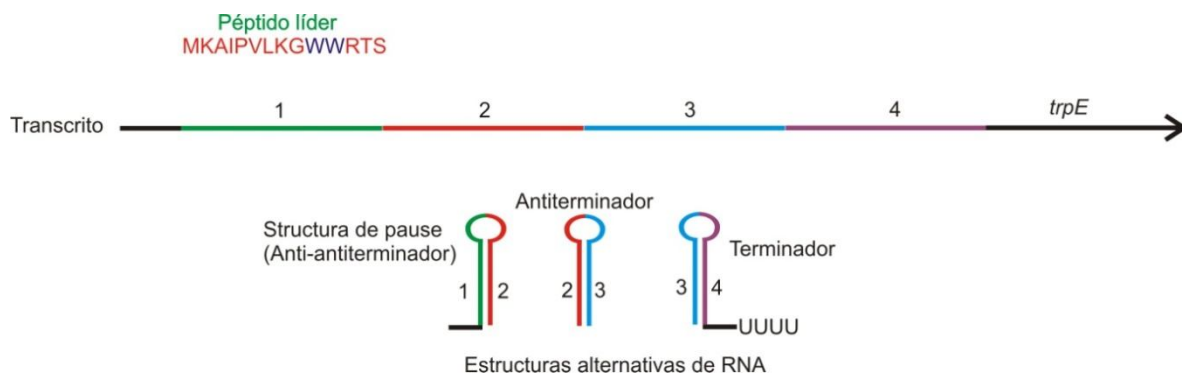


Figura 1. **Organización y funciones regulatorias de la región líder del operón *trp* de *E. coli*.** Propiedades del transcrito líder regulatorio. Figura modificada de <sup>19</sup>

A principios y a mediados de los 70s, se obtuvo la secuencia de la región 5' del operón de *trp*<sup>20</sup>. La secuencia reveló una región entre el promotor y el primer gen estructural del operón, que incluía características de un terminado transcripcional. Terry Platt y Charles Yanofsky identificaron dentro de esta región líder un posible sitio de unión a ribosoma río arriba de una pequeña región codificante de 14 codones, dos de los cuales codificaban adjacientemente para triptófano<sup>20,21</sup>. Mutaciones puntuales que no permitían la traducción de esta región codificante, reducían la regulación del operon *trp*, sugiriendo que este péptido era

un producto de esta pequeña región codificante y que podía influenciar el término de la transcripción. Este péptido era mucho más pequeño que la mayoría de las proteínas conocidas en ese entonces. ¿Cómo era posible que un péptido tan pequeño pudiera tener un rol tan importante en la regulación del operón *trp*? Se propusieron distintas alternativas que incluían la opción de que el péptido *per se* tuviera un efecto regulatorio directo, ya fuera interactuando con alguna proteína regulatoria, o uniéndose al DNA. La presencia de codones de triptófano en tándem, que eran más bien raros en regiones codificantes, fue una pieza clave, en conjunto con el análisis de la secuencia líder del operón *trp*. Se descubrió que el RNA líder podía tomar dos estructuras de hélice mutuamente excluyentes<sup>22</sup>. Una de estas estructuras, incluye la hélice de RNA conocida como terminador transcripcional que es responsable de la atenuación transcripcional. La segunda hélice de RNA, incluye a la secuencia 5' del terminador, funcionando como un antiterminador, dado que la formación del antiterminador previene la formación del terminador, permitiendo, por lo tanto, a la RNA polimerasa que continúe con la síntesis del transcrito completo. Cuál de estas hélices se formará, depende de la disponibilidad de los tRNAs de triptófano cargados con el aminoácido (aminoacetilados). Altos niveles de tRNAs de triptófano aminoacetilados (Trp-tRNA<sup>Trp</sup>) dan como resultado una traducción eficiente de los dos codones en tándem de triptófano en el péptido líder de 14 aminoácidos. Si el ribosoma transita de manera rápida de esta región, se desensambla cuando la síntesis del péptido termina, favoreciendo la formación del terminador transcripcional<sup>18,23</sup> (ver Figura 1). Por el contrario, si los Trp-tRNA<sup>Trp</sup> son escasos, el ribosoma que está traduciendo el péptido líder se detiene en alguno de estos dos codones de triptófano, lo que favorece la formación de la estructura de antiterminación, la cual previene la formación del terminador<sup>18,23</sup>.

La caracterización tan detallada del operón *trp* de *E. coli*, permitió a Yanofsky y a sus colegas expandir nuestro conocimiento sobre las complejidades de la



regulación génica e introdujo el concepto de atenuación transcripcional. Con la disponibilidad reciente de datos genómicos, se ha revelado que la atenuación transcripcional es una estrategia de regulación bastante común empleada por muchas especies bacterianas<sup>24</sup>.

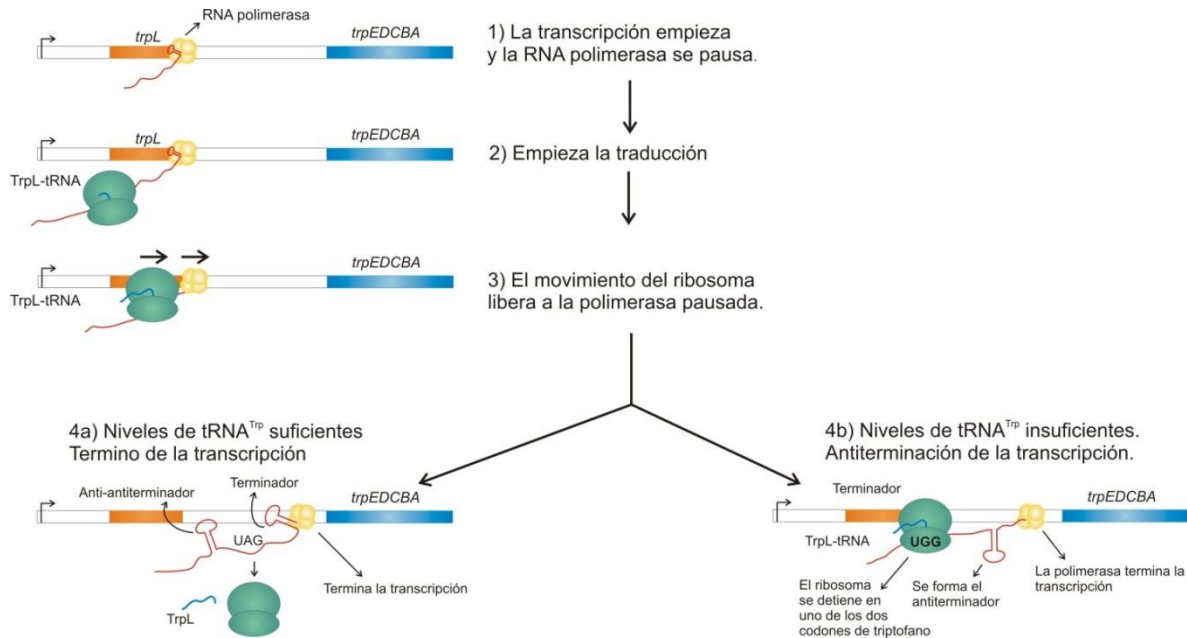


Figura 2. Organización y funciones regulatorias de la región líder del operón *trp* de *E. coli*. Regulación del operón *trp* de *E. coli* por atenuación de la transcripción. Figura modificada de<sup>19</sup>

## Regulación más allá de las proteínas regulatorias

### El descubrimiento del primer riboswitch: La T box

El mecanismo de atenuación de *trp* reveló que la disponibilidad de un aminoácido podía ser detectada de formas indirectas, mediante los efectos de la aminoacilación de los tRNAs, que podían ser monitoreados por la procesividad del ribosoma. En 1992 Frank Grundy y Tina Henkin observaron que el gen *tyrS* de *Bacillus subtilis*, codificado una tirosil tRNA sintetasa, contenía una región líder más

larga de lo usual y ésta contenía un terminador transcripcional pero carecía de una región codificante obvia para un péptido líder<sup>25</sup>. Así mismo, observaron que las características que esta secuencia poseía, estaban presentes en los genes de otras aminoacil tRNA sintetasas (aaRS) en *Bacillus*. Estas enzimas, cargan el aminoácido correspondiente a su tRNA afín y por lo tanto son esenciales para la síntesis de proteínas. Sin tener en ese entonces herramientas genómicas disponibles, la comparación de 10 secuencias era una ardua labor. Sin embargo, ellos demostraron que compartían características importantes, como una secuencia de 14 nucleótidos llamada T box que precedía a un terminador<sup>26</sup>. ¿Había algún mecanismo de regulación en común en estos genes que codificaban para aaRS? ¿Cómo podía existir la regulación en la ausencia de una proteína regulatoria? ¿Cómo podían ser controladas por el mismo mecanismo de regulación si cada gen para aaRS respondía de forma distinta a la disponibilidad de su aminoácido correspondiente?

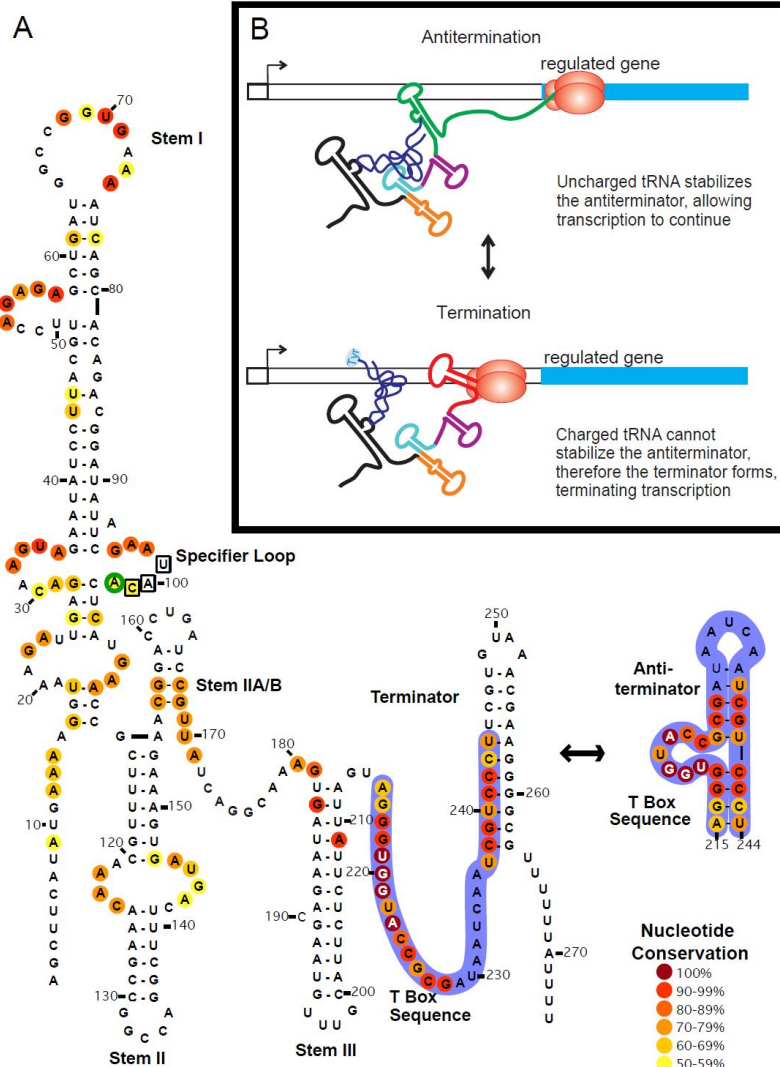
En ese entonces, la regulación de los genes para aaRS sólo había sido descrita en *E. coli* y era muy compleja. Los mecanismos conocidos para regular su expresión incluían: control del inicio de la transcripción (*alaS*)<sup>27</sup>, atenuación transcripcional (*pheS*)<sup>28</sup>, y autorregulación traduccional (*thrS*)<sup>29,30</sup>. Además, se sabía que su expresión estaba acoplada a la tasa de crecimiento celular y a los niveles de tRNAs descargados<sup>31</sup>.

Grundy y Henkin demostraron que las regiones líder para todos estos genes, que contenían una secuencia T box, se podían estructurar bajo el mismo patrón a pesar de que la secuencia variara considerablemente<sup>32</sup>. Además, demostraron que la señal responsable para una respuesta específica a la limitación de un aminoácido eran tres nucleótidos que representaban el codón que correspondía al aminoácido que cargaría la aaRS del gen río abajo. Este codón, designado como la *secuencia de*

*especificidad*, se encontraba en cada estructura en la misma posición. Alteraciones en la *secuencia de especificidad* del gen *tyrS* de un codón de tirosina hacia uno de fenilalanina era suficiente para cambiar la respuesta regulatoria congruentemente con la disponibilidad del aminoácido fenilalanina. Inserciones de una base exactamente río arriba del codo no alteraban la regulación, indicando que el mecanismo no funcionaba de modo traduccional. Reemplazar este codon, por un codón sin sentido, resultaba en una expresión muy baja que podía ser restablecida al introducir un tRNA con ese codón en particular, indicando que el tRNA es la molécula protagonista<sup>33</sup>. Gracias a estas investigaciones, ahora conocemos que el riboswitch T box utiliza el tRNA descargado como molécula señal (ver Figura 3).

Caracterización adicional de este elemento reveló, de forma similar al operon *trp* de *E. coli*, que la región líder de los genes que pertenecen a la familia T box, puede doblarse para formar una de las dos estructuras secundarias de RNA que son mutuamente excluyentes, un terminador transcripcional intrínseco que promueve la atenuación de la transcripción, o bien, un antiterminador que permite que la RNA polimersa sintetice el transcrito completo. La unión del correspondiente tRNA descargado con el RNA líder favorece la estabilización de la estructura de antiterminación; el reconocimiento del tRNA correspondiente está determinado principalmente por la unión del anticodon del tRNA con la *secuencia de especificidad* y la discriminación de un tRNA descargado ante uno cargado está mediada por la unión de las bases entre el brazo aceptor del tRNA descargado con los residuis de un asa presente en el elemento de antiterminación<sup>26,34</sup>. El reconocimiento directo del tRNA por el transcrito del RNA fue demostrado con experimentos bioquímicos<sup>35</sup>. De forma similar a como ocurre en el operón *trp* de *E. coli*, el riboswitch T box monitorea el estado de aminoacetilación de un tRNA en particular. La diferencia principal con el sistema *trp* de *E. coli* es que el el tRNA no es monitoreado por un ribosoma que traduce, si no que la T box emplea una

interacción directa del RNA líder con la molécula de tRNA en cuestión. En ambos sistemas, muchos genes del mismo organismo pueden ser regulados de forma distinta usando el mismo mecanismo, simplemente introduciendo cambios sutiles en la estructura del RNA líder, i.e., el cambio de codones del péptido líder o los cambios en la *secuencia de especificidad*.



**FIGURA 3. El mecanismo de regulación por T-box.**

**(A) Estructura modelo del RNA líder del gen *tyrS* de *Bacillus subtilis*.** Se muestra el elemento T-box, presente en la región líder del gen *tyrS* de *B. subtilis*, fue descrito originalmente por Grundy y Henkin (ver<sup>26</sup>). El rearrreglo estándar de la T-box en el RNA líder se compone de tres grandes elementos: tallo I (stem I), tallo II (stem II) y tallo III (stem III), además del *pseudoknot* formado por los tallos IIA y IIB (stem IIA/B pseudoknot), y las estructuras mutuamente excluyentes: terminador y antiterminador. El asa de especificidad (Specifier Loop) forma una burbuja o *loop* en el tallo I, el cual contiene la secuencia de especificidad (residuos UAC encuadrados que son complementarios a la secuencia del anticodón del tRNA<sup>Tyr</sup>); encerrada en un círculo verde, se resalta la purina (adenina, en este caso) conservada que siempre se encuentra adyacente a la *secuencia de especificidad*. La secuencia T-box no se encuentra apareada en el terminador, pero sí en el antiterminador (el antiterminador se muestra a la derecha del terminador). La secuencia resaltada en azul muestra los nucleótidos involucrados en la estructura del antiterminador. La estructura del antiterminador contiene un *loop* que puede interactuar con los residuos no-pareados del brazo aceptor del tRNA descargado. Se evaluó la conservación nucleotídica de las 722 T-boxes descritas en la base de datos de Rfam, y se han coloreado los nucleótidos de la T-box de acuerdo a su conservación.

**(B) Modelo de las alternativas de regulación del mecanismo T-box.** En la Figura 1(B) se muestra el modelo de las alternativas de regulación del mecanismo T-box. Mientras la RNA polimerasa (óvalos rojos) transcribe la región líder, el RNA naciente se pliega en una estructura que compete, en dos sitios, por la unión del tRNA afin. Arriba se muestra el tRNA descargado y su unión en los dos sitios: el asa de especificidad, y el bulbo del antiterminador. Esta unión estabiliza el antiterminador (segmento verde de RNA), previniendo así, la formación del terminador. Esta conformación permite que la transcripción continúe hasta la secuencia codificante que se encuentra río abajo (caja azul). El tRNA cargado (representado con una Tirosina -Tyr- unida a su región 3') puede interactuar con la secuencia específica, pero no con el antiterminador, lo cual resulta en una falla para estabilizar el antiterminador, permitiendo así que se forme la estructura más estable: el terminador (segmento rojo de RNA). Por lo tanto, la transcripción llega a su fin, antes de que la región codificante pueda ser transcrita. Los elementos conservados de la T-box son los RNAs: tallo I (en negro), tallo II (en naranja), tallo IIA/B (en azul claro) y el tallo III (en morado).

### *Otros riboswitches: el reconocimiento de metabolitos mediante RNA en lugar de proteínas*

El mecanismo T box constituye uno de los primeros ejemplos donde elementos de RNA pueden reconocer con gran afinidad y especificidad una molécula en particular, dando como resultado un cambio estructural en el RNA que afecta la expresión génica. La elucidación de las bases moleculares de cómo funciona el riboswitch T box expande nuestro horizonte sobre roles potenciales que puede tener el RNA en la regulación génica. Es importante considerar que para la época en que la T box fue reportada, la mayoría de la regulación conocida en bacterias estaba basada en proteínas regulatorias. Por ejemplo, el número de proteínas regulatorias identificadas en organismos modelo, *E. coli* y *B. subtilis*, era más de 100 para cada uno. Es bien sabido que las proteínas regulatorias son capaces de reconocer a sus substratos gracias a que son moléculas grandes compuestas de 20 aminoácidos distintos, los cuales les otorgan grandes posibilidades de adoptar una gran variedad de estructuras tridimensionalmente distintas en el espacio. Por el contrario, los elementos de RNA están compuestos sólo de 4 componentes (o ribonucleótidos) y por lo tanto tienen menos posibilidades de adoptar distintas conformaciones que sean capaces de reconocer un metabolito en específico y desencadenar una respuesta regulatoria. Con el descubrimiento del riboswitch T box, el dogma ya establecido sobre cómo las proteínas eran elementos regulatorios únicos, sufre una alteración donde se incluyen las moléculas de RNA como elementos capaces de tomar estructuras tridimensionales que son capaces de reconocer metabolitos particulares y ejercer una acción regulatoria.

Con la acumulación de secuencias de genomas completos bacterianos, no sólo se permitió identificar un número creciente de genes de aaRS que estarían potencialmente regulados por el mecanismo T box, sino también se incrementó el

regulón incluyendo a genes de biosíntesis y transporte de aminoácidos. Así mismo, reveló otros grupos de genes donde filogenéticamente se esperaba encontrar una T box, pero no fue así. Por ejemplo, genes relacionados al metabolismo de metionina en especies de *Bacillus*, *Enterococcus* o *Clostridium* parecían no tener las características comunes a los otros miembros de la familia T box. Sin embargo, compartían características similares entre sí<sup>36</sup>. La caracterización de las regiones líder de estos genes, denominada S box, reveló que la regulación ocurre a nivel de atenuación de la transcripción y que la terminación era estimulada durante el crecimiento en presencia de niveles altos de metionina<sup>36</sup>. Sin embargo, en estos casos, el tRNA parecía no tener un rol importante o directo, más bien la región líder del RNA reconoce S-adenosil metionina (SAM) como la molécula señal<sup>37-39</sup>. Como en otros sistemas de atenuación transcripcional, el RNA líder puede doblarse en estructuras de terminación y antiterminación que son mutuamente excluyentes mientras que SAM estabiliza un tercer elemento estructural que actúa como un anti-antiterminador secuestrando la secuencia que formaría parte del antiterminador. Si SAM estabiliza la formación de este anti-antiterminador, entonces la formación del antiterminador es imposible, favoreciendo así la formación del terminador. El RNA de la S box puede unir directamente a SAM con una gran afinidad y especificidad sin la participación de una proteína reguladora<sup>37-39</sup>. Esta propiedad lo define como un riboswitch, aunque difiere con la T box en el tipo de ligando que es reconocido, pero el patrón regulatorio básico es el mismo. Análisis de estructura mediante la cristalización de este RNA, demuestran que la S box forma un bolsillo donde SAM es envuelto completamente por la estructura de RNA permitiendo una unión muy específica con el ligando<sup>40</sup>.

Con la caracterización de la T box y de la S box, resaltan a la vista características muy similares. La alta especificidad de ambos riboswitches por su ligando impone restricciones tridimensionales que resultan en una alta

conservación a nivel de secuencia y estructura a pesar de que los organismos estén filogenéticamente distantes. Estos grados de conservación son una pieza clave para su predicción a nivel bioinformático, permitiendo la identificación de nuevos tipos de riboswitches<sup>41</sup>. El tamaño promedio de los riboswitches versa entre cien y doscientos nucleótidos, lo cual provee grandes ventajas para su identificación, si los comparamos con los sitios de unión de proteínas regulatorias que están compuestos de unas docenas de nucleótidos en el DNA. Actualmente, un gran número de riboswitches que unen distintos metabolitos han sido indentificados en bacterias. Estos incluyen nucleótidos (adenina<sup>42,43</sup>, guanina<sup>42,44,45</sup>, diGMP cíclico<sup>46</sup>, prequeuosina<sup>47-50</sup> y ATP<sup>51</sup>), aminoácidos (lisina<sup>52</sup>, glicina<sup>53-55</sup>, SAM<sup>36,56</sup> y *S*-adenosil homocisteina<sup>57,58</sup>), carbohidratos (glucosamina-6-fosfato(Glc6P)<sup>59</sup>), coenzimas (flavin mononucleótido<sup>60,61</sup>, tiamina pirofosfato<sup>62,63</sup>, cobalamina<sup>64,65</sup> y tetrahidrofolato<sup>66</sup>), y iones (molibdeno<sup>67</sup>, magnesio<sup>68</sup>) así como parámetros fisicoquímicos como temperatura<sup>69</sup> y pH<sup>70</sup>. Estos RNAs regulan una gran diversidad de grupos de genes, y por lo tanto tienen un gran impacto en el metabolismo celular. La regulación usualmente ocurre al nivel de atenuación transcripcional (como se describió para la *S* box) pero puede ocurrir también a nivel de inicio de la traducción (si alguna de la estructura del RNA líder secuestra el Shine Dalgarno). En la mayoría de los riboswitches que unen metabolitos, el metabolito en cuestión es el producto final de la vía biosintética que regula, apagando la expresión de los genes<sup>71</sup>.

## Antecedentes

### Identificación de elementos Regulatorios

Para poder activar coordinadamente los genes necesarios para una respuesta, las bacterias usan, por lo general, la estrategia de agrupar genes funcionalmente relacionados en operones, para que genes que necesitan ser expresados simultáneamente dentro de la célula, respondan a la misma señal vía su región de regulación<sup>72</sup>. Adicionalmente, distintos genes u operones que estén espacialmente separados dentro de un genoma, pueden estar corregulados también si comparten la región regulatoria (e.g., si tienen en su región regulatoria el mismo sitio de pegado para un regulador transcripcional) de modo que pueden estar controlados por un factor común (e.g., proteína regulatoria) o por el mismo estímulo ambiental. A este conjunto de genes u operones que están regulados por el mismo elemento de regulación y que su expresión responde coordinadamente, se le llama regulón.

La existencia de regulones ha sido aprovechada para identificar elementos o sitios de regulación (e.g., mediante un análisis de expresión de genes de tipo microarreglo). Si se conoce gran parte de los genes que forman parte de un regulón, una estrategia factible es averiguar las secuencias o motivos que están enriquecidos en las regiones controladas del regulón, pero ausentes en el resto del genoma<sup>73-75</sup>. Los motivos así obtenidos son muy buenos candidatos a ser los sitios donde se une un regulador transcripcional. Este tipo de estrategias ha sido muy útil a partir de la explosión de resultados de experimentos con microarreglos, de los cuales se descubrieron nuevos regulones en organismos modelo, al inferir la corregulación a partir de la coexpresión<sup>76,77</sup>. La limitante en predecir sitios de



regulación se presenta cuando la señal es muy débil, ya sea porque el sitio de pegado no está muy conservado, o porque existen muchos falsos positivos dentro del conjunto de posibles genes corregulados, diluyéndose así la señal. Para fortalecerla, se pueden agregar regiones de regulación adicionales provenientes de genes ortólogos de organismos cercanos. En general, se observa que los sitios de pegado de proteínas al DNA divergen rápidamente y solamente dentro de un mismo organismo, o en organismos muy cercanos, se conservan lo suficiente como para ser detectados.

El saber cómo está regulado un gen en particular provee mucha información sobre su naturaleza en general, la vía metabólica en la que puede estar participando, e inclusive, en algunos casos, bajo qué condiciones va a expresarse. Dado que la regulación génica en bacterias ocurre principalmente a nivel transcripcional, la identificación de los elementos regulatorios río arriba de las unidades transcripcionales es de crucial importancia. Como ya se ha dicho en la introducción, los primeros modelos de regulación en bacterias describen proteínas regulatorias y sus sitios de pegado, los cuales tienden a ser chicos y con un grado de conservación bajo. Recientemente, una nueva familia de elementos regulatorios ha ganado mucha importancia: los riboswitches (descritos brevemente en la introducción; en la siguiente sección se ahondará más sobre éstos). Los riboswitches, en contraste a las proteínas, no son moléculas que se encuentren libremente por la célula, sino que son parte de la unidad de transcripción a la cual regulan, ya sea activando o reprimiendo su expresión. De 1992 a la fecha, un total de 16 familias de riboswitches han sido descritas y caracterizadas experimentalmente, la mayoría une metabolitos con gran afinidad y especificidad, como aminoácidos, vitaminas, nucleótidos o cofactores. Por definición, los riboswitches unen a su molécula blanco en una completa ausencia de proteínas, gracias a que son estructuras complejas de RNA esto les impone un alto grado de

conservación tanto a nivel estructural como a nivel de secuencia. Estas características contribuyen a ubicar a los riboswitches como excelentes candidatos a ser identificados con gran precisión en búsquedas de genómica comparativa.

Si comparamos un riboswitch (T box, en este caso) con el sitio de unión de una proteína reguladora (CRP, en este caso), podemos observar lo distintos que son, sobre todo en conservación y longitud. Con este claro ejemplo se puede notar que es mucho más sencillo identificar a nivel bioinformático un riboswitch que un sitio de unión a una proteína:

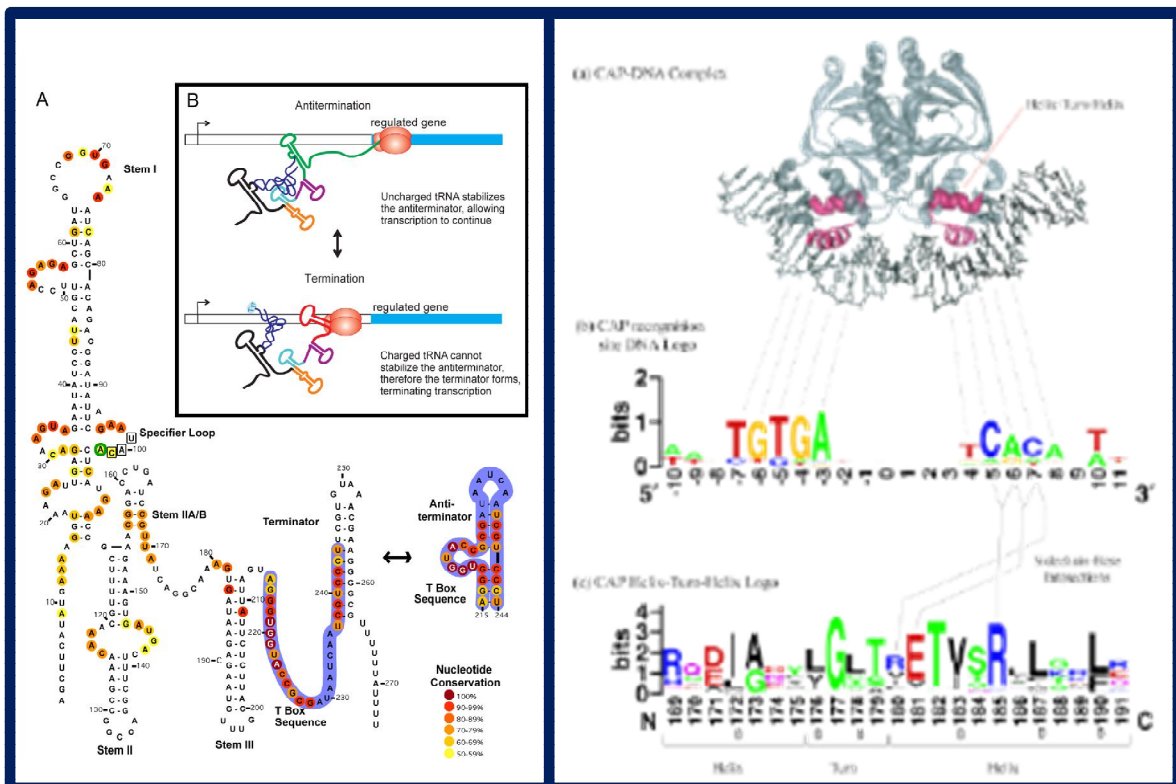


Figura 4. De lado derecho se muestra el riboswitch T box, presente en la región líder del gen *tyrS* de *B. subtilis*, fue descrito originalmente por Grundy y Henkin (ver<sup>26</sup>). Se evaluó la conservación nucleotídica de las 722 T boxes descritas en la base de datos de Rfam, y se han coloreado los nucleótidos de la T-box de acuerdo a su conservación. De lado izquierdo se muestra arriba la estructura tridimensional de la proteína CRP. Abajo muestran dos paneles de conservación de secuencia (nucleotídica arriba, proteica abajo) tomada de<sup>201</sup>.

Otra ventaja de identificar riboswitches sobre sitios de unión a proteínas es que los riboswitches se conservan más a pesar de la distancia filogenética, dado

que tienen que interactuar con un metabolito, el cual no cambia a pesar de que el ambiente y el organismo sí.

## **Regulación por RNA**

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

## **Perspectiva Histórica: El mundo de RNA**

Cuando uno se sitúa en el contexto histórico de cuando se empieza a trabajar con RNA, uno debe de considerar que en ese tiempo no era sencillo (no es que lo sea ahora tampoco) trabajarlo. El RNA es una molécula químicamente inestable en comparación con el DNA y es precisamente esta inestabilidad lo que llevó al paradigma de una historia evolutiva que planteaba al RNA como un derivado del DNA, y que en ese entonces, parecía sólo un nivel más de complejidad en la célula, que por algunas razones que no se entendían, le confería algunas ventajas. En otras palabras, se pensaba que de alguna manera el DNA había inventado el RNA.

Pero como con muchos conceptos en la historia de la ciencia, este duró poco tiempo. El rol del RNA en biología comienza a principios de los 50s cuando Francis Crick, Alexander Rich, Leslie E. Orgel y James Watson proponen que varias moléculas de RNA podían ayudar en los procesos de la célula que llevarán las instrucciones codificadas en los genes y a mediados de los 50s discuten las

implicaciones de la coexistencia de DNA y RNA<sup>78</sup>. No llegan a muchas conclusiones sino hasta finales de los 60s que Orgel, Crick y Woese sugieren que el RNA pudiera haber estado presente antes que el DNA durante la evolución y que pudo haber tenido un rol mucho más activo en funciones celulares primigenias<sup>79-81</sup>.

No fue, sino hasta 15 años después, en 1982, que hubo la primera evidencia experimental de este tipo de ideas, cuando Thomas Cech y Sydney Altman descubren, de forma independiente, que el RNA puede funcionar como enzima catalizando reacciones celulares<sup>82,83</sup>. A este nuevo tipo de RNAs catalíticos se les llamó ribozimas (por enzimas de RNA). Este descubrimiento tuvo dos consecuencias muy fuertes. La primera, inició una gran cantidad de líneas de investigación sobre qué otros roles podía tener el RNA dentro de la célula. La segunda, una reevaluación sobre el origen evolutivo de la vida<sup>78</sup>. Con esta revelación, de que el RNA puede funcionar como una enzima, podemos apreciar al RNA como la única biomolécula que tiene la habilidad de guardar material genético y transmitir esta información y además realizar operaciones que antes eran atribuidas únicamente a proteínas, para poder mantener a una célula viva. En otras palabras, el RNA es la única molécula con genotipo y fenotipo. Por tanto, que una molécula tenga las implicaciones de ambas, del DNA (sólo genotipo) y de las proteínas (sólo fenotipo) sitúa al RNA como una molécula clave en la historia de la vida<sup>78</sup>. Esto nos da el nuevo paradigma de cómo imaginamos a los primeros sistemas vivos en la tierra, aquellos que emergieron hace 4 mil millones de años, en un mundo de RNA<sup>78,80,84,85</sup>. Nuestro mejor modelo para las primeras formas de vida se sitúa con una colección de moléculas de RNA (o algo muy similar) que tenía la capacidad de replicarse a sí mismo, siendo ambos, gene y enzima. Ahora entendemos el RNA de forma muy distinta a como lo entendieron hace 50 años.

## RNAs no codificantes (reguladores en *trans*): una contribución importante a la red de regulación bacteriana

Como se describió en la introducción, uno de los primeros mecanismos de regulación en el que se mostró que el RNA juega un papel importante es en la atenuación<sup>18</sup>. Esta estrategia, descrita por Yanofsky inicialmente para el operón de triptófano, y ahora reconocida como una estrategia ampliamente usada por bacterias, es un ejemplo de una regulación en *cis* pues el RNA que juega un papel importante en la regulación es parte del mismo transcrito que contiene los genes a ser (o no) expresados. Otro ejemplo de reguladores de RNA en *cis* son los riboswitches.

Hasta ahora, he descrito cómo la regulación en bacterias fue descubierta poco a poco a través de los años. Estos hallazgos fueron hechos gracias a la genética clásica y de forma más reciente a la genómica comparativa; todos ellos respondiendo a preguntas específicas y planteadas de forma brillante. Sin embargo, los RNAs no codificantes (ncRNAs) permanecieron en el anonimato por años ya que no pudieron ser descubiertos por las metodologías clásicas. De hecho, se descubrieron por serendipia con marcaje metabólico cuando se catalogaban los RNAs de *E. coli*<sup>86-88</sup>. Los ncRNAs tienen papeles clave en la respuesta de una bacteria a situaciones como estrés, o regulan genes y factores importantes para la virulencia, pero sobretodo le ayudan a la célula a responder a ambientes que sufren cambios constantemente<sup>88</sup>. A la fecha, varios ncRNAs han sido descritos ejerciendo distintas actividades a nivel regulatorio, como control de calidad en la traducción (tmRNA), inhibidores de proteínas (CsrB), procesamiento de ribozimas (RNAseP), regulación por antisentido, ya sea en *cis* (SokA) o mediante la unión a Hfq (DsrA)<sup>88</sup>.

En el 2009, Pascale Cossart y colaboradores encontraron que dos riboswitches SAM pueden actuar en *trans* después de ser transcritos. Estos riboswitches en *trans* interactúan con la región líder de genes de virulencia (como PrfA) en *Listeria monocytogenes*, inhibiendo su expresión. De esta forma, la virulencia queda coordinada con la disponibilidad de nutrientes<sup>89</sup>.

### **Riboswitches: Regulando el metabolismo bacteriano**

La palabra *riboswitch* es acuñada por Ronald Breaker en 2002 para describir aquellos motivos de RNA que pueden actuar como un switch en la expresión génica sin proteínas como intermediarios<sup>65</sup>. Este nombre fue inspirado en las *ribozimas*, que designan aquellos RNAs que tienen una actividad parecida a la de una enzima, en el sentido que pueden promover una transformación química sin la necesidad de proteínas, a pesar de que muchas ribozimas estén asociadas con proteínas. Tanto los riboswitches, como las ribozimas, deben lidiar con baja diversidad química de ser formados sólo por cuatro componentes (bases ribonucleicas) y con éstos ser capaces de formar un “bolsillo” donde poder reconocer con alta especificidad a su ligando.

Desde 1992 que se descubrió el primer riboswitch, la T box a la fecha, se han descrito más de 25 riboswitches y es probable que muchos más estén aún en la espera de ser descubiertos. Es importante recalcar que en algunos organismos, como en *B subtilis*, se conocen más riboswitches, que el número de factores proteicos validados experimentalmente que unen metabolitos y controlan la expresión génica<sup>90</sup>.

La regulación de la expresión génica en bacterias ocurre en varios niveles, principalmente en el inicio de la transcripción, donde comúnmente, varias proteínas ayudan a la RNA polimerasa a iniciar la transcripción. Sin embargo,

existen muchos procesos moleculares que son esenciales y que ocurren una vez que la transcripción ha iniciado. Ahora sabemos que estos procesos son blanco idóneo para decisiones regulatorias. Como ya vimos, la atenuación de la transcripción involucra, ya sea la activación o la inhibición de un terminador transcripcional que se encuentra entre el promotor y los genes del operon, y es una de las estrategias preferidas por la mayoría de las bacterias. Eventos moleculares que son relevantes para la función de cada operon, son tomados en cuenta de una u otra manera para determinar si el término de la transcripción ocurrirá<sup>24</sup>. Estudios de genómica comparativa junto con predicciones de estructura secundaria de RNAs, sugieren que al menos 80 distintas familias de proteínas en bacterias, están regulados por terminación transcripcional<sup>91</sup>.

Una de las principales ventajas de regular la expresión génica con terminadores/antiterminadores transcripcionales, es que una única estructura y secuencia, de tamaño pequeño puede mediar decisiones regulatorias que son cruciales para la célula. Por tanto, es probable que un RNA a ser transcrito evolucionara para poder unir un metabolito específico (sin contradecirse con la hipótesis descrita en la sección Perspectiva Histórica: El mundo de RNA) o bien que evolucionara para unir una región de alguna proteína. Esta característica pudo después, bajo presiones selectivas, ser explotada para permitir o inhibir el término de la transcripción en respuesta a una señal fisiológica<sup>24</sup>.

El control de la transcripción a nivel de una terminación prematura, es una estrategia común de las bacterias para regular sus genes, y comúnmente se basa en el hecho de que un RNA reconozca la señal adecuada, es decir, se basa en el uso de riboswitches. Algunas de las características de los riboswitches son bastante generales, mientras que otras son más bien raras. Sobre las características generales, una de las más destacables es su dominio de "aptámero" o bien, su

capacidad de unir un metabolito. En general, este dominio es el que se encuentra mejor conservado. Esto tiene sentido, ya que los aptámeros están compuestos por sólo 4 nucleótidos, y éstos deben doblarse de una manera muy precisa para poder reconocer a su ligando, el cual suele ser constante en la historia evolutiva de los organismos. Por el contrario, la secuencia y la estructura del siguiente dominio, “la plataforma de expresión” de cada riboswitch, presenta un menor grado de conservación, pues existen muchas maneras en que estructuras de RNA pueden influenciar procesos de transcripción, traducción y procesamiento de RNA.

Los riboswitches se localizan principalmente en la región 5' de los mRNAs microbianos cuya expresión van a regular. Este arreglo permite que el riboswitch sea lo primero en ser transcrito dándole tiempo para responder a las concentraciones del metabolito en cuestión, antes de que el mRNA completo sea transcrito. En la mayoría de los Firmicutes, los riboswitch operan a nivel de término de la transcripción.

### **El Riboswitch T box**

El riboswitch T box, comúnmente modula la expresión de muchos genes relacionados con el metabolismo de aminoácidos, principalmente en los Firmicutes. La T box utiliza una molécula de tRNA descargado como señal regulatoria. Los genes que están regulados por este riboswitch presentan, en la región intergénica río arriba, una conservación de ciertas características a nivel de secuencia y estructura<sup>26,92</sup>. Para la mayoría de este conjunto de operones, segmentos de la región líder del mRNA pueden doblarse para formar una de dos estructuras alternativas, ya sea un terminador transcripcional, o bien, un antiterminador mutuamente excluyente que competirá con el terminador para estructurarse. La



formación del terminador da como resultado un término prematuro de la transcripción, reduciendo la transcripción de los genes que se encuentran río abajo de ésta. Para cada unidad transcripcional que está regulada por una T box, se requiere una unión del correspondiente tRNA descargado con el RNA líder que favorecerá la estabilización de la estructura de antiterminación. Esta unión previene la formación de un terminador y permite que la transcripción continúe a los genes que se encuentran río abajo<sup>26,34</sup>. El reconocimiento específico del tRNA descargado que corresponda a cada T box, así como la formación del antiterminador gracias a la unión del tRNA, ocurre en una ausencia absoluta de proteínas u otros factores celulares<sup>93</sup>. Como ocurre con muchos otros riboswitches donde un RNA líder reconoce una señal para en base a eso, regular la expresión de los genes que se encuentran río abajo, la T box también puede controlar el inicio de la traducción. En aquellos transcritos que tienen esta capacidad, la estructura de terminación se encuentra reemplazada por otra hélice que secuestra el Shine-Dalgarno (SD) de los genes que se encuentran río abajo, inhibiendo así el inicio de la traducción, en lugar de terminar prematuramente la transcripción. Las T boxes que regulan el término de la transcripción se encuentran principalmente en las Gram positivas con un bajo contenido en G+C, e.g., en los Firmicutes; mientras que aquellas T boxes que controlan el inicio de la traducción predominan en las Gram positivas con un alto contenido en G+C, e.g., en las Actinobacterias. La baja frecuencia con la que se encuentran genes regulados por la T box en bacterias Gram negativas hace pensar que fueron recientemente adquiridas por transferencia horizontal y nunca se ha probado experimentalmente si éstas llevan su efecto de regulación a nivel transcripcional o traduccional<sup>94,95</sup>.

Las similitudes de la T box con otros riboswitches, es la capacidad que tiene el RNA líder de reconocer directamente una molécula señal (ya sea un tRNA descargado, en caso de la T box o algún metabolito, en caso del resto de los

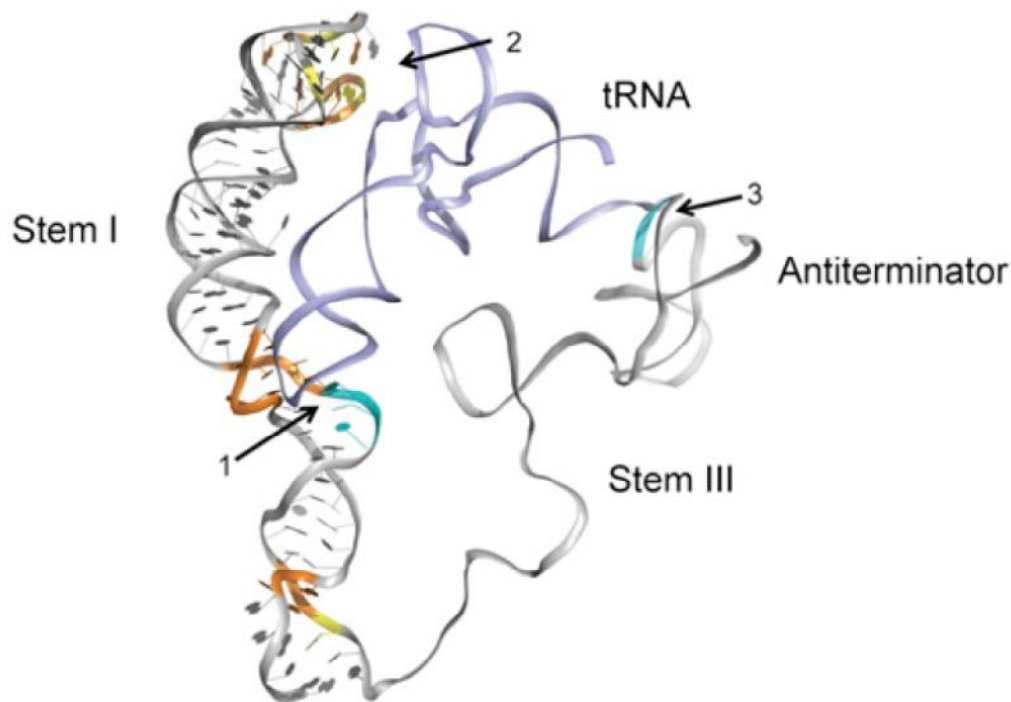
riboswitches), lo cual desencadena rearrreglos en la estructura del RNA líder y éstos determinan si los genes que se encuentran río abajo serán o no, expresados. El mecanismo T box parece estar particularmente bien adaptado para regular la expresión de aquellos genes que codifican para aminoacetilar los tRNAs, para sintetizar aminoácidos o transportarlos dentro de la célula<sup>96</sup>.

En contraste con otros riboswitches que unen metabolitos mediante el reconocimiento de características específicas de su ligando, la T box permite responder de forma específica de cada unidad transcripcional a la deficiencia de un tRNA en particular. Como ya se ha descrito anteriormente, con pequeños cambios en la secuencia de especificidad de la T box, se puede alcanzar un cambio en la respuesta regulatoria, permitiendo el reconocimiento de distintos tRNAs. Este alto grado de flexibilidad sin pérdida de especificidad, sea probablemente el responsable de que la T box sea el mecanismo de regulación de elección por los Firmicutes para sus genes de metabolismo de aminoácidos<sup>97</sup>.

Otra diferencia entre la T box y la mayoría de los riboswitches es que la T box por defecto está reprimiendo el gen. Es decir, la estructura secundaria energéticamente más estable es un terminador transcripcional y se requiere de un tRNA descargado que interactúe con la T box para que el antiterminador sea estabilizado y la transcripción pueda continuar. En los demás riboswitches, la estructura más estable es un antiterminador (o un anti-SD) y se requiere el metabolito (en general el último producto de la vía de biosíntesis para unirse y poder apagar la transcripción del mRNA en cuestión.

En abril de este año, se determinó la estructura cristalográfica de la T box con una resolución de 2.65 Å<sup>98</sup>. Esta estructura muestra la importancia del Tallo I, el cual es altamente conservado, en el reconocimiento del tRNA. Dos asas de este tallo interactúan para servir de plataforma que reconocerá las asas T y D del tRNA.

Un resultado asombroso de este trabajo es que la estructura tridimensional de la T box es altamente similar a los sitios de unión de las asas T y D de la RNasa P y del sitio de salida del ribosoma. Esto sugiere que esta plataforma podría ser un “motivo” ampliamente distribuido para unir tRNAs. En este trabajo se propone que el Tallo I es el primero en interactuar con el tRNA vía la unión de la *secuencia de especificidad* y el anticodón (independientemente de si este está cargado o no), y esto actúa como un primer “checkpoint”. Posteriormente la plataforma distal asasa mediría la longitud del brazo aceptor del tRNA. Cuando ambas condiciones se cumplen, el tRNA quedaría “asegurado” y entonces podría establecerse el estado de aminoacetilación (ver Figura 5).



**Figura 5. Estructura cristalográfica de la T box interactuando con el tRNA.** Tomada de <sup>98</sup>. En esta figura Grigg et al. muestran a la T box con su antiterminador, i.e., en el estado de encendido. El RNA de la T box se muestra en forma de listones con su tRNA (azul claro) envuelto por el Tallo I y por el antiterminador. Una conservación nucleotídica de más del 70% está coloreada en naranja/amarillo como fue descrita en nuestro trabajo<sup>97</sup> y las regiones que se sabe que interactúan con el tRNA están coloreadas en cian. El 1 representa el primer “checkpoint”, i.e., la *secuencia de especificidad* para el reconocimiento del anticodón del tRNA. El 2 representa un “checkpoint” geométrico, el cual reconoce a las asas T y D del tRNA. El 3 representa el contacto del tRNA con el antiterminador.

### Tipos de riboswitches.

Los riboswitches regulan la expresión de genes cuyas proteínas codifican para un gran rango de funciones en distintas especies bacterianas, pero se puede generalizar que todos ellos dependen de que la célula detecte cierto estado metabólico a través de un metabolito para efectuar distintas respuestas según sean necesarias. El tipo de riboswitch más simple, son los ya mencionados termosensores de RNA. Estos elementos estructurales afectan la expresión de aquellos genes río abajo, en general, mediante el secuestro de la secuencia Shine-Dalgarno (SD)<sup>99</sup>. La decisión regulatoria ocurre en respuesta a un cambio en la temperatura, en lugar de a la unión de una molécula efectora. Estos elementos regulatorios son usados comúnmente para controlar la expresión de aquellos genes que le permitirán a la bacteria responder a cambios repentinos de temperatura, incluyendo genes de heat-schock<sup>69</sup>.

Como ya se vio, los riboswitches, en contraste a los termosensores de RNA, están compuestos de dos módulos; el aptámero, un dominio que une al ligando, y la plataforma de expresión, otro dominio que permite la decisión regulatoria. El dominio que une un ligando, o metabolito, es responsable de un reconocimiento específico del ligando cuya unión afectará la estructura que se forme en el dominio subsecuente, y por tanto la expresión del gen regulado. El rearrreglo más común es aquel donde la unión del metabolito ocasiona que el RNA se estructure favoreciendo la inhibición de la expresión del gen. Por tanto, y en general, la unión de la molécula efectora apagará la expresión del gen, principalmente porque el ligando suele ser el producto final de la vía biosintética regulada por dicho riboswitch<sup>71</sup>.

## Distribución filogenética de los riboswitches

En un número importante de casos, los mecanismos regulatorios no se encuentran distribuidos de forma azarosa en el mundo bacteriano, si no que más bien tienden a agruparse. Por ejemplo, la atenuación transcripcional mediada por un péptido líder (como fue descrito en la sección Atenuación transcripcional: regulando la expresión génica en base a elementos de RNA en *cis*) es común en las proteobacterias y en organismos Gram negativos, mientras que la T box y otros riboswitches suelen ser comunes en organismos Gram positivos, sobre todo en las Firmicutes<sup>94</sup>. Cerca de 16 familias distintas de riboswitches han sido identificados a la fecha, y algunos de ellos están distribuidos en varios taxa filogenéticos, como es el caso de la Thi-box (un riboswitch que detecta los niveles de tiamina pirofosfato (TPP) en la célula), el cual está presente en bacterias, archaeas y eucarias, mientras que otros riboswitches se encuentran confinados a taxas específicos, tal es el caso de la T box, que se encuentra principalmente en Firmicutes<sup>94,96</sup>. Otro ejemplo interesante es la S box, que presenta ligeras variaciones dependiendo de en qué clado se encuentre.

## Riboswitches que reconocen aminoácidos

Estas familias representan mecanismos de regulación, presentes en las regiones líderes de ciertos mRNAs, cuya síntesis del transcrito completo está determinada por el hecho de qué juego de tallos y asas se formarán en respuesta al reconocimiento de un aminoácido

### *S-adenosil metionina*

El mecanismo S box, reconoce a la molécula S-adenosil metionina (SAM) como señal regulatoria. Los genes de esta familia, como en la T box, exhiben en su región líder una serie de características conservadas, tanto a nivel secuencia, como a nivel estructura. Ya se ha descrito cómo se identificó este mecanismo en la sección

*Otros riboswitches: el reconocimiento de metabolitos mediante RNA en lugar de proteínas*, pero puede también leerse en las referencias<sup>36,94</sup>. Los elementos conservados de la S box incluyen un terminador transcripcional, con su correspondiente, y mutuamente excluyente, antiterminador. La unión de SAM estabiliza un elemento anti-antiterminador, lo cual favorece que se forme la estructura de terminación<sup>37,39</sup>. Una interacción terciaria dentro de la S box es crucial para que SAM pueda unirse y favorecer el término de la transcripción<sup>100</sup>. Esto sugiere que el riboswitch contiene una especie de bolsillo que envuelve a SAM y que requiere un arreglo tridimensional complejo. El riboswitch S box reconoce a SAM con una especificidad muy alta, sin embargo, las bases moleculares de cómo se da esta unión, aún no se entienden del todo.

#### El regulón S box

La identificación de T boxes con una especificidad para metionina en genes en las especies de *Enterococcus* and *Streptococcus*, contrastaba con aquellos genes correspondientes en *B. subtilis* donde no parecía haber miembros del regulón T box. Los ortólogos de estos genes en *B. subtilis* contenían una longitud similar, y tenían elementos conservados suficientes para la identificación de un nuevo riboswitch, la S box. El regulón de la S box comprende a genes involucrados en el metabolismo de metionina, cuya función era en ese entonces desconocida. El conocer su mecanismo de regulación ayudó a entenderlos como parte del metabolismo del azufre. Su descripción inicial consistió en once unidades transcripcionales en el genoma de *B. subtilis*<sup>36</sup>. Ahora se sabe que el regulón de S box comprende genes para la asimilación de azufre y genes para el transporte y la síntesis de metionina y cisteína.

## Diversidad en la S box

La S box presenta cierta diversidad, dependiendo de en qué clado filogenético se encuentren, sin embargo todos reconocen SAM. Por ejemplo, el riboswitch SAM-IV se encuentra en los Actinomycetales (e.g., *Mycobacterium tuberculosis*). Tienen estructuras similares, aunque presentan variaciones en algunas hélices, reconocen el metabolito de forma muy similar y con los mismos nucleótidos<sup>101</sup>. El riboswitch SAM-II se encuentra principalmente en las  $\alpha$ -proteobacteria que no tienen al elemento SAM-I (no se han encontrado ambos en el mismo genoma)<sup>102</sup>. SAM-II difiere mucho más al resto de los riboswitches SAM, es aproximadamente de la mitad de tamaño que SAM-II, pero unen SAM con la misma afinidad<sup>101</sup>. Por último, el riboswitch SAM-III o SAM<sub>MK</sub> se encuentra en bacterias del ácido láctico (*Enterococcus*, *Streptococcus* y *Lactococcus*)<sup>56</sup>. Reconoce SAM de forma distinta a los riboswitches SAM-I o SAM-II<sup>101</sup>.

## S-adenosil homocisteína

Como se puede observar de los riboswitches SAM, existen muchas formas para reconocer un mismo metabolito, SAM, con suficiente afinidad para obtener una respuesta regulatoria y poder discriminar de su producto, la S-adenosil homocisteína (SAH). Lo opuesto es biológicamente igual de importante, pues los niveles intracelulares de SAH deben estar perfectamente regulados ya que actúa como un potente competidor de SAM inhibiendo a las enzimas que lo utilizan, por lo que a muy altos niveles es tóxico<sup>103</sup>. Las estructuras cristalográficas no esclarecen cómo es que los riboswitches pueden diferenciar estos dos compuestos. Para SAM-I, ciertas mutaciones logran disminuir la selectividad, pero no del todo<sup>101</sup>.

En genes de reciclaje de SAM e hidrolasas para SAH de proteobacteria se encontró un nuevo riboswitch que une SAH con una afinidad muy superior aquella con la que une a SAM<sup>101</sup>. Recientemente, una búsqueda masiva de riboswitches

encontró un nuevo motivo que aparentemente no distingue entre SAM y SAH<sup>57</sup>, el cual es exclusivo de las Rhodobacterales y se encuentra río arriba del gene que codifica para la SAM sintetasa (*metK*).



## Objetivo

Esta tesis tiene varios objetivos particulares que se van definiendo conforme se van planteando los distintos avances del proyecto. Sin embargo, un objetivo general, que engloba todos los análisis realizados en la presente, es entender a fondo el regulón del riboswitch T box.

### Objetivos Particulares:

- 1) Entender cómo es que el riboswitch T box se expande para regular nuevos genes relacionados al metabolismo de aminoácidos.
  - a. Predecir dónde hay T boxes en todos los genomas, con el fin de conocer mejor el regulón T box.
- 2) Establecer cuáles son las generalidades de este riboswitch, cuáles son sus particularidades y si éstas pueden asociarse al aminoácido involucrado en la decisión regulatoria.
- 3) Determinar cuáles son las restricciones evolutivas que rigen a este elemento de regulación.
  - a. Identificar, no sólo el regulón, sino también la distribución filogenética de la T box.
  - b. Identificar el origen de este elemento evolutivo.
- 4) Y por último, el objetivo que le da nombre a esta tesis, usar la información que este riboswitch nos proporciona para entender mejor el metabolismo bacteriano y cada uno de sus genes, que están bajo la regulación de la T box.

- a. Predecir la participación de genes con función desconocida, que pertenecen al regulón T box, dentro de las vías de biosíntesis de aminoácidos.
- b. En aquellos casos donde no se logre predecir completamente la función, el objetivo será identificar nuevas relaciones metabólicas de la participación de estos genes poco caracterizados en las distintas vías biosintéticas

## Metodología

### Estado del Arte de herramientas en bioinformática utilizadas en esta tesis: Una pequeña revisión

En esta sección cubro las estrategias, herramientas y bases de datos de última generación, que son a la fecha la mejor manera de encontrar motivos estructurales de RNA, con la mejor eficiencia para identificar cada uno de ellos. Me detengo poco sobre cada uno, porque es tecnología que evoluciona rápidamente.

#### RFAM

Rfam es una colección de alineamientos múltiples y de modelos de covarianza que cubre la mayoría de las familias de ncRNAs y otros elementos estructurales de RNA. Rfam es una base de datos anotada, curada y de libre acceso, pensada de una forma similar que la base de datos Pfam, que trabaja en conjunto con el software Infernal (descrito posteriormente), modelos de covarianza (CMs) para poder seguir anotando familias de RNAs en secuencias genómicas y metagenómicas. A diferencia de las proteínas, los ncRNAs presentan mucha conservación en la estructura secundaria, a pesar de no ser tan conservados a nivel de secuencia. Rfam busca dividir a los ncRNAs en familias, basándose en su historia evolutiva. Mantienen alineamientos múltiples de cada familia, los cuales son muy útiles para inferir su estructura y su función, como en el caso de las proteínas.

Rfam combina la información que proporciona la secuencia con la estructura secundaria de los ncRNAs, lo cual es representado en alineamientos múltiples, y construye modelos estadísticos llamados "Gramática Libre de Contexto para Perfiles Estocásticos" (SCFG, por sus siglas en inglés de Stochastic Context-Free Grammars) o bien, modelos de covarianza. Estos modelos, son análogos a los

modelos de Markov Escondidos que se utilizan en las proteínas para predecir las familias que componen Pfam. Cada familia en la base de datos de Rfam está representada por dos alineamientos múltiples, uno en el formato Stockholm y otro como SCFG. El primer alineamiento es un alineamiento “semilla” que es curado a mano y contiene a los miembros más representativos de una familia de ncRNAs; es anotado en base a información de estructura secundaria. Este alineamiento semilla es el que se usa para crear al SCFG, el cual puede usarse con el software Infernal (ver abajo) para identificar miembros nuevos de cada familia y agregarlos al alineamiento. Cada familia tiene sus distintos valores de corte para evitar falsos positivos, los cuales pueden ser consultados en la base de datos Rfam<sup>104</sup>.

Usar los SCFG para hacer búsquedas de familias de RNAs es computacionalmente muy costoso, de modo que para reducir tiempos de búsqueda, se utiliza inicialmente una búsqueda mediante BLAST para reducir el espacio de búsqueda. Esto representa un gran defecto para la base de datos Rfam, ya que, como se mencionó anteriormente, los ncRNAs conservan su estructura, pero difícilmente su secuencia. El segundo alineamiento es uno que contiene todas las secuencias que, con cierto valor de probabilidad, pertenecerán a la familia en cuestión, y es un resultado de dicha búsqueda usando el modelo de covarianza analizando numerosas bases de datos de secuencias. Todos los homólogos detectados por el programa, son alineados de forma automática al modelo, produciendo un alineamiento de todo lo existente, de forma automática<sup>104</sup>.

## **INFERNAL**

Infernal es un software que construye modelos de covarianza (CMs) a modo de perfiles de un consenso de estructura secundaria de RNA, y usa estos CMs para buscar RNAs homólogos en bases de datos de secuencias nucleotídicas. También

permite hacer alineamientos basados en estructura. En conjunto con Rfam, uno puede utilizar Infernal para anotar los RNAs de un genoma.

Infernal es un paquete que contiene varios programas que pueden ser utilizados en pasos:

1. *cmbuild*. Este programa sirve para construir un CM a partir de un alineamiento basado en estructura.
2. *cmcalibrate*. Calibra un CM para una búsqueda por homología.
3. *cmsearch*. Busca en la base de datos de secuencias nucleotídicas posibles homólogos para el CM en cuestión.
4. *cmalign*. Alinea los posibles homólogos a un CM.

*cmcalibrate* es necesario para obtener los e-values del CM construido y además determinará que modelo de Markov escondido es apropiado para el modelo. Cada modelo debe calibrarse sólo una vez<sup>105</sup>.

Dos trabajos independientes demostraron que Infernal es el paquete de herramientas con más sensibilidad y especificidad para detectar homólogos estructurales de RNA<sup>106,107</sup>.

## **MEME**

MEME (EM Múltiple para Elicitación de Motivos, por sus siglas en inglés, donde EM se refiere al algoritmo para maximizar la expectancia)<sup>108</sup> es un programa de dominio público que regresa de un conjunto de secuencias, uno o más motivos o patrones sobre-representados. Se basa en ajustar una mezcla de modelos a la serie de secuencias que se le proporciona. Un modelo es para el motivo que se busca y recibe tantas variables independientes como el tamaño del motivo, mientras que el

modelo de fondo es más sencillo, recibiendo solamente las frecuencias de cada carácter (nucleótidos en este caso). Se prueba con un ciclo del algoritmo EM todos los puntos de partida para elegir el más prometedor y para afinar la mezcla de modelos se prosigue a realizar ciclos de EM hasta convergir<sup>109</sup>. MEME además compara los motivos que encuentra de distintos tamaños, para elegir el que más probablemente sea de relevancia biológica, con la ventaja de que el usuario no tiene que conocer exactamente el tamaño. En general, se elige un E-value de corte, el rango de tamaño del motivo y el número máximo de motivos a encontrar y MEME los reporta, siempre y cuando no rebasen el valor de corte<sup>110</sup>. De acuerdo al conocimiento que se tiene de los motivos que se espera encontrar, se le debe proporcionar su tipo de distribución a MEME, que lo usa para ajustar un mejor modelo al motivo. Las opciones son que el motivo: (a) se encuentre presente en una copia en todas las secuencias, (b) en sólo algunas, o (c) en más de una copia por secuencia. De acuerdo a lo que uno esperarí en la búsqueda de riboswitches, lo ideal es emplear el parámetro (b) que optimiza a MEME para encontrar motivos presentes en sólo algunas de las secuencias de entrada. En general el valor de corte empleado permite encontrar elementos de regulación presentes en solamente unas pocas de las secuencias de un grupo ortólogo. El cortar los motivos en 30 nucleótidos es ideal en este caso, para evitar que motivos diferentes (pero contiguos) queden mezclados en uno sólo. Sobre todo esto ayuda a mantener como motivos independientes a las regiones conservadas de un riboswitch que están separadas por un asa de tamaño variable. Si se permite un motivo demasiado grande, puede incluir el asa y por lo tanto sólo representará correctamente a aquellos riboswitches con un asa de tamaño semejante, lo cual, sucede muy poco y no es recomendable.

Existen grupos de ortólogos que pueden presentar más de un motivo de regulación, en algunos casos relacionados (motivos que son parte de un mismo

elemento de regulación) y en otros independientes (un subconjunto de secuencias contiene un motivo y otro subconjunto un motivo totalmente diferente), por lo cual conviene acotar el número de motivos que se le permitirá a MEME encontrar. De no aplicarse este paso, uno de los motivos podría “opacar” a los demás, los cuales se perderían durante el resto del proceso. Además, conviene eliminar motivos que consten de más de 95% de G/C o A/T, debido a su bajo contenido informacional. Cabe mencionar que el procedimiento es sumamente robusto, por lo que si se varían un poco los parámetros de tamaño de motivo o valor de corte, los resultados son muy similares.

Para una revisión de metodologías similares a MEME, ver referencia<sup>111</sup>.

## MAST

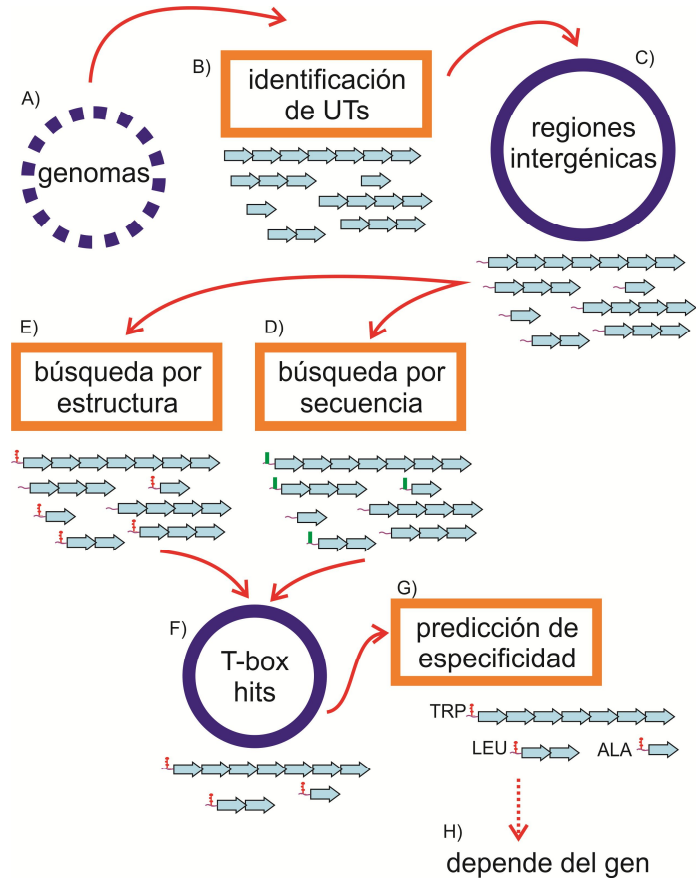
Para cada motivo encontrado por MEME conviene realizar una búsqueda usando MAST (Herramienta para Buscar y Alinear Motivos, por sus siglas en inglés)<sup>112</sup>. Esta herramienta es la contraparte de MEME; toma precisamente un motivo encontrado por éste (en forma de una matriz de peso posición-dependiente), una base de datos de secuencias y regresa las secuencias que contienen el motivo, la posición del motivo y una calificación de probabilidad. Para la búsqueda de riboswitches con MAST usamos una base de datos hecha *ad hoc* que incluye todas las regiones de regulación intergénicas de todos los genomas completos. De esta manera, para cada motivo, originalmente encontrado en un subconjunto de las regiones de regulación de un grupo de genes ortólogos, MAST localizaría todos los genes que podrían estar regulados por el mismo elemento.

## The Riboswitch Approach

El primer paso fue obtener las regiones intergénicas de todos los genomas disponibles completamente secuenciados (Figura 6A y 6C), que fueron condensados en una base de datos *ad hoc* (ver programa 1, sobre la metodología de extracción de regiones intergénicas).

Predicción de operones (Figura 6B). Fueron utilizados genomas completamente secuenciados, así como secuencias provenientes de metagenómica. Las unidades de transcripción (operones) en cada una de las secuencias genómicas y metagenómicas

se predice en base a las distancias intergénicas y sobre las relaciones funcionales de los productos protéicos de los genes contiguos, obtenido de la base de datos de STRING (como se describe en <sup>113</sup>)



**FIGURA 6. Metodología.** La idea básica de la metodología fue generar una base de datos de regiones intergénicas asociadas a un operón. Paralelamente se generaron matrices consenso que poseen la información de conservación de secuencia de distintas partes de la T-box. A su vez, una matriz de covarianza que contiene la información de la conservación estructural de la T-box. Posteriormente se buscan T-boxes utilizando estas matrices en las bases de datos primeramente mencionadas. Los hits se evalúan y si pasan cierto corte consideramos que el operón en cuestión está potencialmente regulado por una T-box



Después se construyeron matrices de posición-específica capaces de reconocer la conservación nucleotídica de las T boxes utilizando el programa MEME<sup>110</sup> (Fig 6D)<sup>41,114</sup>. Las secuencias con T boxes fueron recuperadas utilizando el programa MAST (Fig 6D)<sup>112</sup>. Cabe señalar que el programa MAST no recuperaba eficientemente las T boxes. Sabemos que en la secuencia T box existen ciertas posiciones que no deben variar para que ésta pueda reconocer el brazo aceptor del tRNA. Para poderle pedir a MAST que mantuviera fijas ciertas posiciones de la matriz fue necesario modificar el código de dicho programa, lo cual se hizo en marzo del 2010 con la colaboración de Tim Bailey (del Institute for Molecular Bioscience de la Universidad de Queensland, Australia). MAST toma las matrices de posición específica y la de logaritmos que generan MEME y utiliza éstas para buscar en la base de datos que se le proporcione. Las posiciones donde no varían los nucleótidos en la matriz generada por MEME resultan en un valor de UNO en las posiciones que deben conservarse y un CERO en el resto de los nucleótidos en esta posición. Al convertir estos valores a logartimos, resulta en un problema, el logaritmo de uno es cero, pero el logaritmo de cero no esta definido (sería menos infinito). De modo que en estas posiciones, cuando MEME convierte las matrices a logartimo, asigna un valor arbitrario de un número negativo muy grande para el caso del logartimo de cero. Este número es asignado dependiendo del tamaño de la base de datos con que fue generada dicha matriz, lo que resultaba en un valor siempre distinto. Para permitir que MAST mantuviera fijas dichas posiciones, deben hacerse dos cosas. La primera es redefinir el rango de dígitos que la variable de MAST reconocerá para leer una matriz con un valor lo más pequeño posible. Este valor tan pequeño, lo definimos empíricamente en -1,400,000, de forma que cuando MAST lee un valor tan chico, ningún otro valor biológico puede competir contra él, fijando así la presencia de ciertos nucleótidos en la posición dada.

El programa de MAST se puede acceder usualmente en:

```
/usr/local/meme_4.3.0/src/mast.c
```

MAST está escrito en C, por lo que una vez modificado el source code, debe generarse el ejecutable. El cambio fundamental fue redefinir el rango de valores que MAST acepta en las matrices logarítmicas:

```
/* number of different integral score values */  
/*#define MAST_RANGE 100*/  
#define MAST_RANGE 100000
```

Una vez realizado este cambio, deben cambiarse las matrices generadas por MEME donde se fijarán las posiciones. La matriz que MEME genera para la T box se ve así:

```
ALPHABET= ACGT
```

```
-----  
Motif 1 position-specific scoring matrix  
-----
```

```
log-odds matrix: alength= 4 w= 29 n= 66124 bayes= 8.34311 E= 3.7e-1315
```

```
 155  -482  -1431  -340  
 154  -382  -1431  -340  
-111   70   -247   90  
 -54  -73   -147   97  
 119 -1431  -447   -40  
  -2  -482   161 -1431  
-206 -1431   207 -1431  
-374 -1431   216 -1431  
-1431 -1431 -1431   168  
-1431 -1431   219 -1431  
-1431 -1431   219 -1431  
-102  -12  -1431   118  
 160 -1431 -1431 -1431  
-131  258  -1431  -318  
-1431 284  -1431 -1431  
  -42 -1431   178 -1431  
-1431 284  -1431 -1431  
-1431 -1431   219 -1431  
  29 -250   117  -199  
  29 -182    50   -34  
  33  -24   -77    12  
  47  -12  -108    -1  
  15   35  -130    24  
  16   43  -157    24  
   0  133  -177   -29  
-126  141    -8   -13  
-167  129     2    10  
-248  167   -30     4  
-448  196   -95     8  
-----
```

La matriz modificada con las posiciones fijas del asa del antiterminator, necesarias biológicamente para el reconocimiento del brazo aceptor del tRNA, deberán verse así (en negritas señalo aquellos números que fueron modificados a mano, y son los nucleótidos que interactúan con el tRNA):

ALPHABET= ACGT

```

Motif 1 position-specific scoring matrix AATTAGGGTGGTACCGCGGAAATCCCCC
-----
log-odds matrix: alength= 4 w= 30 n= 49814 bayes= 8.32364 E= 6.3e-1044
 167  -465  -1392  -450
 169  -465  -1392  -1392
-106   52   -174   91
 -63  -74   -139  107
 121  -265  -239   -46
  15  -465   146  -1392
-358 -1392   204  -1392
-241 -1392   199  -1392
-1400000  -1400000  -1400000  178
-1400000  -1400000  208  -1400000
-1400000  -1400000  208  -1400000
-112  -74  -1392  138
 169  -1400000  -1400000  -1400000
-226  253  -1392  -1392
-1392  263  -1392  -1392
 -58  -1392  174  -1392
-1392  260  -1392  -391
-1392 -1392  208  -1392
  48  -184   82  -150
 -12  -148   56   8
  30   5  -103  20
  46  30  -129  -7
  28  -6  -103  28
  15  30  -150  36
 -30  133  -103  -37
-188  137   9  -25
-400  133   53  -33
-326  144  -29  20
-458  186  -203  23
-458  198  -188  -4
-----

```

Con estas modificaciones en las matrices y en el código de MAST, se procedió a identificar por secuencia las T boxes. Paralelamente se usaron los modelos de covarianza descritos de Rfam con el programa CMsearch (Fig 6E)<sup>104,115</sup>. De esta manera, se encontraron las T boxes no sólo por conservación nucleotídica, sino también por conservación estructural. Cuando se iniciaron estos análisis, se contaba con 559 genomas bacterianos completamente secuenciados y se

caracterizaron bioinformáticamente (con la metodología previamente descrita) 1,111 T-boxes en 87 organismos. El criterio utilizado para considerar genes adyacentes como parte de un operón, fue el descrito por Taboada *et al.*<sup>113</sup>. El contexto genómico de cada T box fue analizado con nuestro servidor web GeCont<sup>116,117</sup>.

**Asignación de función de los genes regulados.** Este análisis se realizó utilizando primeramente BLAST para buscar una coincidencia contra las bases de datos COG<sup>118</sup> y KEGG<sup>119</sup>.

**La identificación de casos relevantes para la formulación de modelos metabólicos.** Como se indicó anteriormente, la T box fue descrita originalmente en la regulación de las aminoacil tRNA sintetasas, pero también regula ampliamente la expresión de genes implicados en la biosíntesis de aminoácidos. Esto es lo que llamaríamos un caso "canónico", donde una T box para cierto aminoácido está claramente regulando un proceso metabólico directamente involucrado en que la célula incremente los niveles intracelulares de ese aminoácido. Estos casos no fueron considerados para el análisis (si los genes estaban bien anotados), pues hay congruencia entre el tipo de tRNA que se está detectado y la función del metabolismo de los genes regulados. El resto de los casos, donde no había una congruencia clara, son los que fueron objetos de este estudio.

## Programa que construye una base de datos *ad hoc* que contiene todas las regiones intergénicas de los genomas completamente secuenciados

```
#!/usr/bin/perl

use strict ;
use warnings;
no warnings ('uninitialized', 'substr');
use DBI;

my $bd = "actualiza" ;
my $user = "ana" ;
my $pass = "ibt_gem_6599" ;
my $host = "localhost" ;
my $dbh = "" ;

unless ($dbh = DBI->connect("DBI:mysql:$bd;host=$host",$user,$pass)) {
    print "\nDB-ERROR connect para $bd";
    print "Error numero MySQL = $DBI::err\nError: $DBI::errstr\n" ;
    exit 1 ;
}

my $bugs = qq|"%firmicute%"| ;
#my $bugs = "bsu" ;
my ($sth_bugs, $bugi, $bug_id, $bug_def) ;
my %bichos ;

$sth_bugs = prepara ($dbh,qq|SELECT org_id, org_definition FROM orgs WHERE
org_phylo LIKE $bugs ORDER BY org_definition|) ;
#$sth_bugs = prepara ($dbh,qq|SELECT org_id, org_definition FROM orgs WHERE
org_id = "bsu" ORDER BY org_definition|) ;
$sth_bugs->execute ;
while (my @row_bugs = $sth_bugs->fetchrow_array) {
    ($bug_id, $bug_def) = @row_bugs ;
    #print "$bug_id\t$bug_def\n" ;
    $bichos{$bug_id} = $bug_def ;
}

$sth_bugs->finish() ;

my $sth_genes = prepara ($dbh,qq|SELECT gen_id, gen_name, gen_ncbi_gi FROM genes
WHERE gen_org = ?|) ; ##gen_ncbi_gi puede volver vacio
my $sth_cog = prepara ($dbh, qq|SELECT geg_name FROM gen_groups WHERE geg_type =
"cog" AND geg_gene = ?|) ;
my $sth_seqs = prepara ($dbh,qq|SELECT seq_seq FROM gen_seq WHERE seq_gen_id = ?
AND seq_type = "ur"|) ;

my (%geninfo) ;
my ($gen_id, $gen_name, $gi, $cog, $secu) ;

foreach $bugi (sort(keys(%bichos))) {
    print "$bugi\t$bichos{$bugi}...\n" ; #<STDIN> ;
    my $file = "$bugi"._ur.fna" ;
    open (FNA, ">$file") || die "Cant create $file\n" ;
    ## $bugi contiene el codigo del nombre del bicho. i.e., bsu para B.
    subtilis

    $sth_genes->execute($bugi) ;
```

```

        #$sth_seqs->execute($bugi) ;
        while (my @row_genes = $sth_genes->fetchrow_array) {
            ($gen_id, $gen_name, $gi) = @row_genes ;
            $geninfo{$gen_id} =
">".$gen_id"."|".$bichos{$bugi}".$|".$gen_name".$|".$gi" ;
#            print $geninfo{$gen_id} ;# <STDIN> ;
            $sth_cog->execute($gen_id) ;
            while ($cog = $sth_cog->fetchrow_array) {
#                $geninfo{$gen_id} .= "|"."$cog" ;
                print $geninfo{$gen_id} ; #<STDIN> ;
            }
            $sth_seqs->execute($gen_id) ;
            while ($secu = $sth_seqs->fetchrow_array) {
                #$geninfo{$gen_id} .= "\n".$secu ;
                print "$geninfo{$gen_id}\n$secu\n" ; #<STDIN> ;
                print FNA "$geninfo{$gen_id}\n$secu\n" ;
            }
        }
        close FNA ;
#        print "$file" ; <STDIN> ;
    }

sub prepara{
    my $db_handler = shift ;
    my $query      = shift ;
    my $st_handler = "" ;

    #print "en prepara, me llamaron con $db_handler, $query y $st_handler\n" ;

    unless($st_handler = $db_handler->prepare($query)) {
        print "\nDB-ERROR PREPARE tabla: org_completed";
        print "Error numero MySQL = $DBI::err\nError: $DBI::errstr\n" ;
        print "prepare = $query\n" ;
        exit ;
    }#unless
    return $st_handler ;
}#sub

```

## Resultados y discusión

### Características bioquímicas y evolutivas del regulon T box

Describo los resultados generales y de manera breve. Más detalle puede leerse en el artículo: “Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism” escrito en coautoría con Tina M. Henkin, Frank J. Grundy, Charles Yanofsky y Enrique Merino; publicado en la revista *Microbiology and Molecular Biology Reviews* en marzo del 2009. Se encuentra como Apéndice I en esta tesis.

### Contribuciones puntuales de este artículo

1. En un inicio sólo se conocía la regulación de las T boxes para genes codificantes de aminoacil tRNA sintetasas. Gracias a estudios de tipo bioinformáticos, se ha adquirido mayor conocimiento en cuánto a qué genes son regulados por la T box. De esta forma, logramos cumplir el objetivo 1a de esta tesis; predecir dónde hay T boxes en todos los genomas, con el fin de conocer mejor el regulón T box.
2. Se identificó, no sólo el regulón, sino también la distribución filogenética de la T box (objetivo 3a). Además, se logró identificar el origen de este elemento evolutivo (objetivo 3b).
3. Se pudo predecir, para algunos casos, la participación de genes con función desconocida, que pertenecen al regulón T box, dentro de las vías de biosíntesis de aminoácidos (objetivo 4a). En los casos donde no se logró predecir completamente la función, el objetivo fue identificar nuevas relaciones metabólicas de la participación de estos

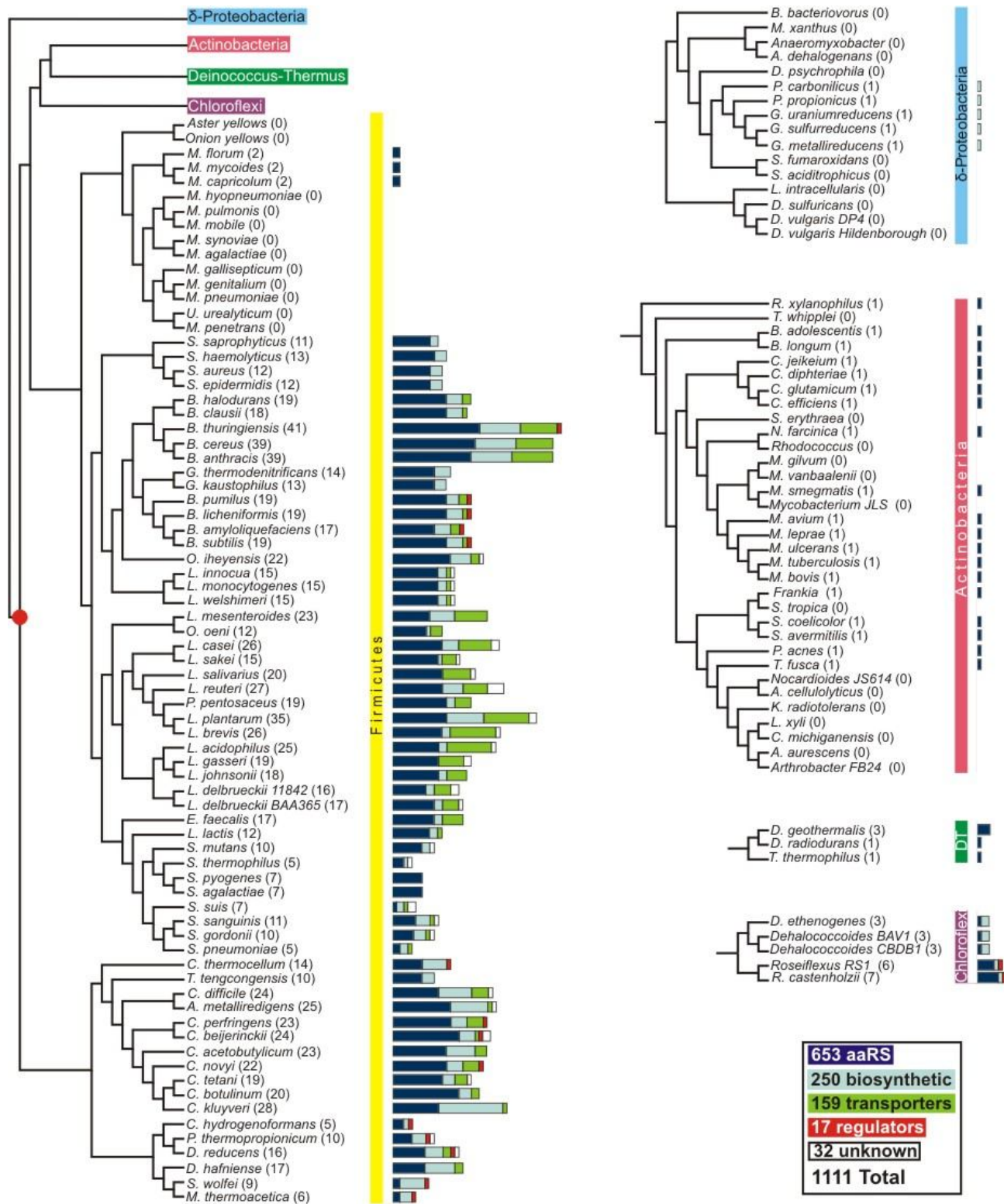
genes poco caracterizados en las distintas vías biosintéticas (objetivo 4b).

### Origen evolutivo del elemento T box

El análisis filogenético de la distribución de T boxes que se presenta en la Figura 7, sugiere que este mecanismo de regulación ya estaba presente en el ancestro común de los Firmicutes, Chloroflexi, Deinococcus-Thermus y Actinobacterias. El hecho de que un grupo tan lejano filogenéticamente, como los  $\delta$ -Proteobacterias presente este mecanismo regulatorio, creemos que se debe a un evento de transferencia horizontal, ya que son pocos los miembros de las  $\delta$ -Proteobacterias que presentan T boxes y que la mayoría de estas bacterias no lo tienen.

En particular, las especies *Geobacter* y *Pelobacter* presentan una T box en un gen de biosíntesis de leucina, *leuA*. Creemos que tanto este gen, como su región regulatoria, han sido adquiridos horizontalmente de *Clostridium acetobutylicum*. Esto se debe a que presentan homología y a que se sabe que *Clostridium* y *Geobacter* son organismos sintróficos, es decir comparten el mismo nicho, y se complementan con metabolitos (e.g. *Geobacter* necesita acetato para crecer, el cual lo adquiere de *C. acetobutylicum*).





**FIGURA 7. Distribución de las T-boxes en distintos taxa filogenéticos.** El árbol filogenético fue construido en base a una matriz de distancia filogenética de un alineamiento de secuencias concatenadas de 31 proteínas presentes en 191 especies, como fue descrito en<sup>120</sup>. El alineamiento fue generado usando el programa MUSCLE<sup>121</sup> y la reconstrucción filogenética usando el programa PROTDIST

del paquete de inferencia filogenética PHYLIP. La longitud de las barras horizontales está dibujada a escala, reflejando el número de operones regulados por una T-box, los cuales se clasifican de la siguiente manera: aminoacil-tRNA sintetasas (azul marino), genes de biosíntesis de aminoácidos (celeste), genes que codifican para proteínas regulatorias (rojo), genes de transporte de aminoácidos (verde) y genes de función desconocida (blanco).

## Características del codón de especificidad

*Cada T box es el resultado de una selección evolutiva, que está adaptada para responder de forma específica a un tRNA descargado y que dará el nivel de expresión necesario para cada transcrito que regula.*

Esta es una hipótesis que mantuve a lo largo del desempeño de esta tesis y que llevó a la generación de varias preguntas. Una de ellas surge al observar los codones empleados en la *secuencia de especificidad*. La intuición más obvia, es que si es un gen que requiere una expresión alta, es probable que el codón que utilice su T box sea aquel que el organismo prefiera, i.e., aquel cuyos tRNAs sean más abundantes. De ser cierto, la idea contraria sería válida también, si se requiere la poca expresión de un gen, el codón de su T box será un codón raro. Al analizar todas las T boxes, su uso de codones por organismo contra el uso de codones preferencial de cada organismo, resulta evidente que la intuición obvia es incorrecta. La preferencia del uso de codones de la T box, nada tiene que ver con el uso de codones de cada organismo. Por ejemplo, las T boxes de glicina, todas tienen una fuerte tendencia a utilizar el codón GGC. Este codón es el menos usado por *Bacillus halodurans* y por el contrario es el más usado por *Bacillus clausii*. Estudios de mutaciones puntuales en la T box de *tyrS* en *B. subtilis*, demostraron que reemplazar el codón de tirosina UAC por el codón de tirosina UAU afectaba dramáticamente la expresión *in vivo* de este gen. Además de nuestros análisis bioinformáticos, estudios experimentales realizados (y no publicados) en el laboratorio de Tina M. Henkin con la región líder del operon *glyQS* de *B. subtilis* sugieren que existe un fuerte sesgo a mantener una C en la tercera posición del

codón presente en la *secuencia de especificidad*, lo cual se debe, probablemente, a restricciones de la unión codón-anticodón con el tRNA, los cuales son, sin duda, muy distintos a aquellos empleados por la maquinaria de traducción. Otras restricciones estructurales de los líderes T box y su conservación se encuentran descritas en el artículo de revisión *Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism*<sup>96</sup> que forma parte de esta tesis (ver Apéndice I).

### **Aminoacil tRNA sintetasas como parte del regulón T box.**

Históricamente, el primer grupo de genes que se se encontró que era regulado por una T box fue el de aquéllos codificantes para las aminoacil tRNA sintetasas<sup>25,26</sup>. Las aminoacil tRNA sintetasas (aaRS) son las enzimas encargadas de cargar el aminoácido en su respectivo tRNA y se encuentran reguladas por T boxes en los Firmicutes<sup>96</sup>. Las aaRS se clasifican en dos grupos no homologos de enzimas; las clase I (LeuRS, IleRS, MetRS, TyrRS, GluRS, ValRS, ArgRS, LysRS1 y TrpRS) y las clase II (PheRS, SerRS, ThrRS, ProRS, AlaRS, HisRS, AspRS, AsnRs, LysRS2 y GlyRS)<sup>122</sup>. Las aaRS de clase I reconocen distintas propiedades estructurales del tRNA a las que reconocen las de clase II. Si estas características estructurales del tRNA son también reconocidas por la T box, aún no se sabe. De cualquier modo, tanto las aaRS de clase I como las de clase II se encuentran reguladas por la T box, y es, en los Firmicutes, el mecanismo de regulación más comúnmente usado para regular a estas enzimas. Esta clasificación de aaRS se basa en el tipo de dominio que contiene para unir ATP. Las de clase I contienen un Rossman fold, mientras las de clase II poseen un rearrreglo de  $\beta$ -sheet<sup>122,123</sup>. Las aaRS de clase I y de clase II están, obviamente, presentes en todos los organismos, y cada tipo de tRNA es aminoacetilado de forma exclusiva por una de las dos clases de aaRS, con la excepción del tRNA de lisina, que puede ser aminoacetilado por ambas clases de

enzimas<sup>122</sup>. Todos los Firmicutes tienen el gen que codifica para la Lisil tRNA sintetasa de clase II, el cual se encuentra, de forma conservada, presente en un supraoperon junto con genes que participan en la biosíntesis de folato, el cual no está regulado por una T box. Se conoce la regulación de este operon en *B. subtilis* y se encuentra regulado dependiendo de la tasa de crecimiento de la bacteria<sup>124</sup>. Además de esta LisRS de clase II, el gen para la LisRS de clase I está presente únicamente en *Bacillus cereus*, *Bacillus thuringiensis* y *Clostridium beijerinckii* y se encuentra regulado por una T box de Lisina.

### ***Un desbalance regulatorio para algunos operones del regulón T box***

Los genes que codifican para las aaRS se encuentran, por lo general, en una configuración monocistrónica con tres grandes excepciones: i) cuando codifican para polipéptidos en enzimas heterodiméricas, como glyQS; ii) cuando están asociados a genes biosintéticos para el mismo aminoácido, como cysS, que está localizado dentro del operón de biosíntesis para cisteína en algunos Firmicutes, y iii) cuando dos aaRS para aminoácidos distintos están codificadas en el mismo operón, tal es el caso de los operones hisS-aspS y cysS-proS. La coexpresión de genes de biosíntesis y de su correspondiente aaRS tiene mucho sentido, y parece ser un uso eficiente del mecanismo de regulación por T box, donde ambos sets de genes se expresarán en el mismo espacio y en el mismo momento, con el mismo tRNA descargado, para responder a un requerimiento celular específico, como sería la falta del aminoácido en cuestión. Sin embargo, expresar dos genes de distintas aaRS al mismo tiempo representa un problema regulatorio, ya que se espera que la expresión de cada gen responda de forma individual a los niveles del tRNA descargado que le corresponden. Ni la histidina y el aspartato ni la prolina y la cisteína tienen vías metabólicas o puntos de síntesis en común. No se sabe cuál sea la estrategia metabólica, o las ventajas evolutivas que expliquen el por qué de

regular las concentraciones de estos aminoácidos en base a la disponibilidad de sólo uno de ellos. A esto le llamamos un desbalance (o imbalance en inglés) metabólico-regulatorio. En el caso del operón *cysS-proS* presente en algunas Clostridia (*Clostridium acetobutylicum*, *C. beijerinckii*, *C. difficile*, *C. perfringens* y *C. tetani*) la regulación parece responder al tRNA de prolina. Para el particular caso de *C. tetani*, estos dos genes están cotranscritos con los genes de biosíntesis de cisteína, haciendo que este desbalance metabólico-regulatorio sea aún más fuerte. Se desconoce si existen elementos de regulación (ya sea a nivel transcripcional o traduccional) que le ayuden a la célula a superar este desbalance. La única pista que se tiene para este caso es que se sabe que tanto la prolina como la cisteína están asociados con la transferencia de grupos metilo, de modo que quizá sea necesario mantener niveles intracelulares similares de tRNAPRO y tRNACYS con este fin. Para el caso del operón *hisS-aspS*, el tRNA que se reconoce por la T box es el de aspartato, sin embargo, todas las secuencias de especificidad contienen los nucleótidos GACAC, que es el codón de aspartato (GAC) y el de histidina (CAC) sobrelapados en un nucleótido. Al predecir la estructura secundaria del Stem I el codon GAC es el único que se acomoda adecuadamente para poder interactuar con el anticodon de un tRNA, sugiriendo que la transcripción de ambas aaRS depende de la disponibilidad del tRNA de aspartato, lo cual resultará en una falla de la síntesis de HisRS cuando se necesite cargar tRNAs de histidina. No obstante, en los organismos *Bacillus halodurans*, *Bacillus licheniformis* y *Clostridium thermocellum*, el Stem I podría adoptar una estructura secundaria alternativa en la cual el codón CAC de histidina puede ser expuesto para la interacción con un tRNA de histidina, pudiendo favorecer la transcripción del operón para responder a ambas especies del tRNA. Esta hipótesis requiere validación experimental (ver la sección de Perspectivas: Imbalance: *hisS-aspS* regulation).

Existe un tercer caso donde un tRNA no esperado se encuentra controlando la expresión de un operón. En la bacteria *Pelotomaculum propionicum* se encuentra el operón *yurG-serAS* que contiene a la seril tRNA sintetasa siendo cotranscrita con genes de biosíntesis de serina, a pesar de esto, en su región intergénica se encuentra una T box que reconoce al tRNA de glicina con el codón GGC. Esto podría explicarse dado que la serina y la glicina pueden ser interconvertidas la una en la otra.

### *T boxes en tándem*

En general, existe sólo una copia de cada aaRS por genoma. Aún así en algunos organismos, se encuentran múltiples copias de algunas aaRS. Existen varios ejemplos que son cubiertos en el artículo de revisión *Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism*<sup>96</sup>, pero aquí sólo describiré el caso del organismo modelo *B. subtilis*, donde existen los parálogos *tyrS/tyrZ* y *thrS/thrZ* cada una de las cuales es regulada por el mecanismo T box. Contender con dos parálogos de una misma aaRS debe de estar sujeto a una regulación muy fina para poder mantener los niveles apropiados del tRNA cargado. Usando como ejemplo a la treonil tRNA sintetasa (*thrRS*), podemos observar que el primer parálogo, *thrS*, contiene una T box de treonina que le permite expresar a ThrRS cuando la célula así lo requiere. Pero, ¿qué pasa si la expresión de esta enzima no cumple con las concentraciones intracelulares suficientes para cargar todos los tRNAs de treonina eficientemente? *B. subtilis* resuelve esto regulando a su isoenzima, *thrZ*, con tres T boxes en tándem. Este arreglo da como resultado una regulación muy estricta, donde tres moléculas de tRNAs descargados de treonina tienen que unirse para estabilizar el RNA líder de *thrZ* para que éste pueda transcribirse sin que se forme ninguno de los tres terminadores<sup>125,126</sup>. Estos arreglos de T boxes en tándem se encuentran en copias





De este resultado en particular surgen varias preguntas obvias ¿es funcional esta T box de leucina? ¿si regula en respuesta a una deficiencia de leucina? ¿es funcional el tRNA de alanina? Para abordar estas preguntas, lo conveniente sería, como primer intento, trabajar con el DNA genómico de estas clostridias, o mandar sintetizar la region regulatoria. La razón es que son anaerobias estrictas y se requiere de instalaciones precisas para poder trabajar con ellas. Es importante considerar varios controles positivos, como la misma T box sin el tRNA, la T box de leucina de *B. subtilis* o de alguna Clostridia cercana que no tenga el tRNA.

Regulación por T box en genes de biosíntesis de aminoácidos

### Regulación por T box en genes de biosíntesis de aminoácidos

La biosíntesis de novo de aminoácidos es, en general, de costo alto para la bacteria. Esto coincide con la regulación de la expresión de los genes de biosíntesis de aminoácidos, donde en general, parecen estar regulados estrictamente. El mecanismo que se usa para regular la expresión de dichos genes, ha evolucionado en los distintos linajes y responde a cambios en los niveles intracelulares del aminoácido en cuestión<sup>19</sup>. Como se ha dicho en la introducción, las bacterias gram negativas muestran una preferencia por utilizar proteínas regulatorias como la estrategia de elección para regular el inicio de la transcripción en respuesta a la disponibilidad de cierto aminoácido, así como regiones líder con péptidos que pueden monitorear los niveles de tRNAs cargados con el aminoácido<sup>23,129</sup>. Si bien, en las bacterias gram positivas también existe el uso de proteínas regulatorias para monitorear los niveles de cierto aminoácido<sup>130,131</sup> se prefiere, en general, utilizar la estrategia de regulación vía riboswitch, especialmente la T box. La regulación de serina y glicina suele estar regulada por T boxes de glicina y por el riboswitch *gcvT* descrito previamente. Los aminoácidos con azufre, es decir, cisteína y metionina, están regulados por T boxes y por S boxes, el uso de la S box vs. la T box depende



mas bien de una historia evolutiva. Los aminoácidos de cadena ramificada, isoleucina, leucina y valina suelen estar reguladas por T boxes, y sorprendentemente en algunas Lactobacillaceas la via del pantotenato suele estar regulada por T boxes que responden a deficiencias en alguno de estos tres aminoácidos. La razón se desconoce, sin embargo hipotetizamos que la bacteria reconoce un estado metabólico en la célula que le permite decidir si conviene o no sintetizar pantotenato. La via de histidina es sumamente conservada y su regulación es poco conocida en los Firmicutes. Suelen no estar reguladas por un T box, si bien existen T boxes de histidina, estas son muy escasas. El tRNA de histidina es peculiar en el sentido que es un nucleótido mas chico que el resto de los tRNAs, i.e., el brazo aceptor del tRNA de histidina tiene solo tres nucleótidos, en lugar de cuatro. Suena razonable considerar que al tener un nucleótido menos para interactuar con el antiterminatr bulge, la regulación por T box puede ser menos eficiente que en el resto de las T boxes. Los aminoácidos aromaticos suelen ser los mas caros para la celula de sintetizar y por esta razón esos aminoácidos suelen estar regulados por T boxes en tándem (ver sección T boxes en tándem). Las vías de aspartato, asparagina, alanina, treonina y prolina no presentan particularidades de gran interés, pero si se encuentran reguladas por t boxes, aunque es un poco menos común, o menos conservado. La regulación de la biosíntesis de arginina se ve mediada por un represor transcripcional en todos los organismos excepto en *C. difficile* que parece haber perdido la proteína regulatoria y para compensar esto duplico su regulon de T boxes de arginina para poder obtener una fácil y eficiente regulación de su metabolismo de arginina. El resto de los aminoácidos que no se mencionaron (lisina, glutamina y gultamato) es porque para sus vías metabolicas no se han encontrado, a la fecha, T boxes como elementos de regulación.

En el artículo de revisión *Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism*<sup>96</sup>, incluido en el apéndice de esta tesis

se describe a profundidad cada vía metabólica y su relación con el riboswitch T box.

### Regulación por T box en genes de transporte de aminoácidos

Determinar la función de genes y de proteínas se ha convertido en un problema biológico central. Los avances en la secuenciación de genomas han mostrado la existencia de una gran cantidad de nuevos genes cuya función biológica es aún desconocida. En promedio, una tercera parte de los genes de un genoma no tienen en la actualidad función conocida, o su función es pobremente entendida.

Un caso particular es el de los genes transportadores de aminoácidos, los cuáles están vagamente anotados como “transporters”. La identificación de una T box río arriba de este gen nos indica que éste está transportando aminoácidos. Más importante aún, dado que podemos predecir qué tipo de tRNA está siendo *sensado* por la T box, podría ser predicho qué aminoácido en particular es el que se está potencialmente siendo transportado.

Un ejemplo es el gen *yvbW* de *Bacillus subtilis*, el cual está anotado como “hypothetical protein” y relacionado al grupo de genes “ $\gamma$ -aminobutyrate permease”. Este gen tiene una T box en su región intergénica la cual contiene un codón CUC (Leucina) en su *specifier sequence*. Se ha reportado experimentalmente que este gen codifica para un transportador de Leucina o algún compuesto similar y que su expresión es inducida por condiciones limitantes de Leucina<sup>132</sup>.

Siguiendo este ejemplo, puede identificarse la especificidad de todos, o de casi todos los transportadores de aminoácidos que estén regulados por una T box (se predijo qué aminoácido transportarán en un 90%). Al predecir a qué

aminoácido responde la expresión del transportador, podemos predecir qué aminoácido será el que dicho transportador introduzca a la célula.

### **Los genes de biosíntesis y de transporte de aminoácidos son regulados por el mismo mecanismo**

Un estudio comparativo de las regiones intergénicas de los operones de biosíntesis y de transporte de aminoácidos reveló que en la mayoría de los casos el elemento regulatorio coincide. Por ejemplo, *B. subtilis* y sus parientes más cercanos utilizan la proteína reguladora TRAP para regular sus genes de biosíntesis y de transporte de triptofano. Por el contrario, en otros Firmicutes se encuentra una T box de triptofano regulando el gen ortólogo de transporte, así como el operón de biosíntesis de triptófano<sup>96</sup>.

Creemos que esto se debe a la necesidad que presenta el organismo de obtener el aminoácido faltante por medio de la coordinación de la biosíntesis y de su transporte. También suponemos que es posible que las T boxes que regulan distintos genes en el mismo organismo respondan diferencialmente al mismo tipo de tRNA, lo cual permitiría una regulación ligeramente diferente.

### **Regulación por T box de proteínas regulatorias**

Se identificaron varias proteínas regulatorias que son reguladas por el mecanismo T box. En estos casos, creemos que la T box tiene un espectro de regulación mucho más amplio, ya que influye en la regulación de muchos otros genes que estarían por debajo en la cascada regulatoria.

Es curioso notar que la mayoría de los genes regulatorios identificados en este estudio se transcriben conjuntamente con aaRS. También me gustaría recalcar que muchos de estos genes son hipotéticos en el sentido de que su función no se ha probado experimentalmente, sino que ha sido asignada por métodos bioinformáticos. Para más detalle, ver el artículo de revisión *Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism*<sup>96</sup>.

## Interpretación del metabolismo microbiano en base a su regulación

En la actual era post-genómica, sólo el 3% de todos los genes se han anotado sobre la base de la evidencia experimental. A pesar de que las funciones fácilmente se pueden predecir para muchos genes, para el 25% de ellas son existe un margen de error importante, que lleva a una mala asignación<sup>133</sup>. Los métodos más utilizados para la función de predicción se basan en la similitud de secuencias, lo que podría inducir a error en muchos casos. Otros métodos tales como el contexto genómico o de perfiles filogenéticos se han desarrollado para aumentar la precisión en la anotación de genes, sin embargo, estos sólo son eficientes cuando las secuencias del genoma completo están disponibles. Aquí se propone un nuevo enfoque basado en la identificación de riboswitches. Los riboswitches se conservan muy bien como reguladores de la expresión de genes situados en la región 5' sin traducir de determinados genes. Cuando se transcriben adoptan una estructura tridimensional que reconoce a sus ligandos con gran afinidad y especificidad. Esta especificidad es una cuestión clave para nuestro método, ya que permite la asignación funcional con gran precisión

### La asignación de funciones de los genes recientemente secuenciados

Asignar bioinformáticamente con precisión una función biológica a un gen recientemente secuenciado no es trivial. Grandes proyectos de secuenciación genómica y metagenómica han dado a conocer una gran colección de genes cuya función está aún por determinar. El método más común para la asignación de una función biológica a un nuevo gen es a través de la inferencia de homología. En el mejor de los casos, una búsqueda de similitud (por ejemplo, utilizando BLAST) se encuentra un homólogo claro con una función conocida. En estos casos, es probable que ambos genes tengan la misma o una función similar. Sin embargo,

muchas proteínas tienen alta similitud de secuencia a pesar de realizar funciones diferentes (por ejemplo, los parálogos *trpE* y *pabA*<sup>134</sup>). Lo contrario también es cierto, puesto que proteínas con baja similitud pueden presentar la misma estructura y función (por ejemplo *aroE* y *ydiB*<sup>135</sup>).

### La aproximación riboswitch / The Riboswitch Approach

Saber cómo un gen en particular se regula puede dar una idea de su naturaleza en general, la vía metabólica en los que participa e incluso, en algunos casos, las condiciones en que se expresa. Puesto que la regulación génica en bacterias se desarrolla principalmente a nivel de la transcripción, la identificación de elementos reguladores en la región río arriba de las unidades de transcripción es de suma importancia. Los primeros modelos de la regulación bacteriana, como ya se ha dicho, fueron descritos para proteínas regulatorias y sus sitios de unión tienden a ser cortos y poco conservados. Por lo tanto, la asignación de funciones de los genes sobre la base de la predicción de elementos reguladores ha sido poco explorada. En contraste, los riboswitches presentan un alto grado de conservación en su estructura secundaria y en menor grado, en su secuencia nucleotídica. Por esta razón, elegimos los riboswitches como candidatos para este análisis. Además, en *B. subtilis*, 110 de los 4,105 genes están regulados por un riboswitch.

En este proyecto, parte de la tesis de doctorado, se utilizó a la T box como un ejemplo representativo de "la aproximación riboswitch". Los detalles, pueden consultarse en el artículo "*Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach*" escrito en coautoría con Enrique Merino; publicado en la revista *Clinical Microbiology and Infection* en julio del 2012. Se encuentra como Apéndice IV en esta tesis.

## Resultados: The Riboswitch Approach

Este análisis se realizó en un total de 771,528 genes. De este conjunto de genes, se identificaron 7,034 genes (0.91%) que están regulados por una T box y se puede dividir en cuatro categorías diferentes de acuerdo a sus funciones anotadas, de la siguiente manera:

- a) Los genes que tienen una clara anotación funcional y que es congruente con la T box que se encuentra río arriba, de acuerdo con la literatura. Estos incluyen aminoacil tRNA sintetasas y genes implicados en la biosíntesis de ácidos amino (para una revisión, véase la referencia<sup>96</sup>). En esta categoría, se encontraron 4,805 genes.
- b) Los genes que sólo tienen una anotación general. Hemos encontrado 857 genes en esta categoría, la mayoría de ellos corresponden a "transporte". La identificación de un elemento T box río arriba de estos genes sugiere que participan en la incorporación de aminoácidos a la célula. Por otra parte, podemos predecir el aminoácido particular para cada transportador, al saber a qué tRNA responderá, dada la secuencia de especificidad de cada T box.
- c) Los genes sin anotación funcional. Estos genes son comúnmente anotados como "proteínas hipotéticas". Un ejemplo interesante se encuentra en la ruta de biosíntesis de la metionina y la cisteína de algunas Firmicutes. Los componentes de esta vía se han determinado experimentalmente en *B. subtilis*, donde un paso, la interconversión entre metionina y cisteína, se lleva a cabo por un par de genes parálogos: *yjcl* y *yrhB*. El producto de *yjcl* realiza la síntesis de metionina a partir de cistationina, que se sintetiza a partir de cisteína. Por otro lado, el producto de *yrhB* sintetiza cisteína a partir de

cistationina. Yjcl y YrhB son 50% idénticos lo que hace que sea difícil distinguirlos a través de la comparación de secuencias. Sin embargo, sus correspondientes genes están regulados por riboswitches diferentes en muchas Firmicutes, y esto hace que sea posible predecir qué aminoácido se sintetiza (Figura 9).

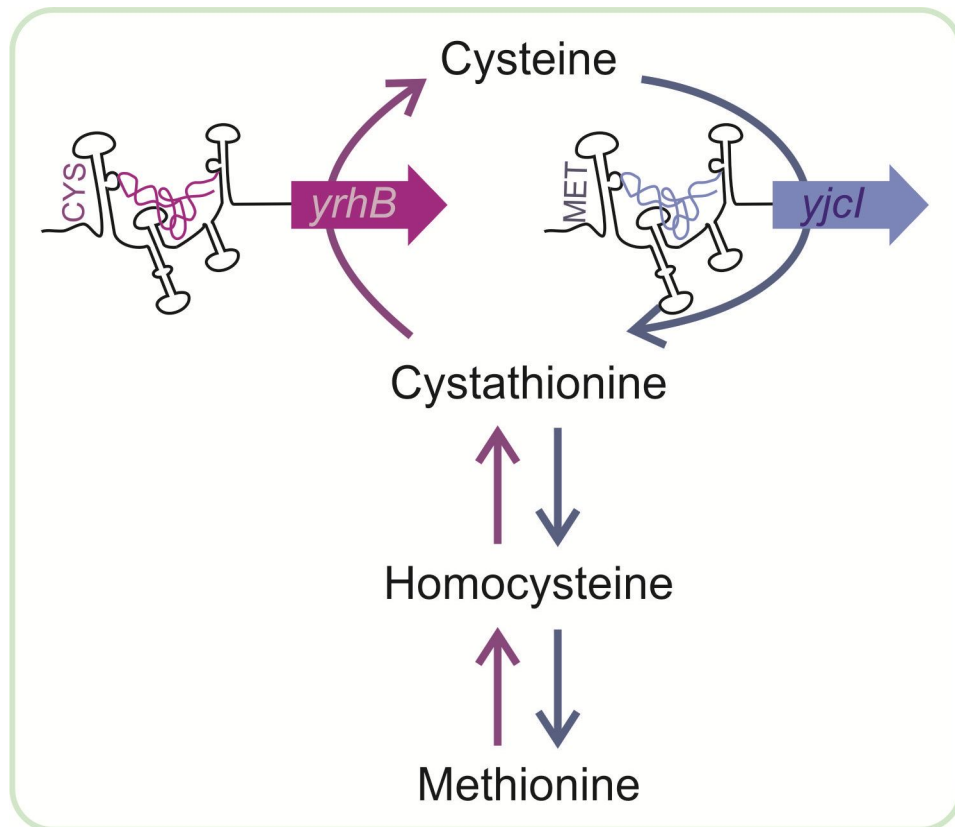


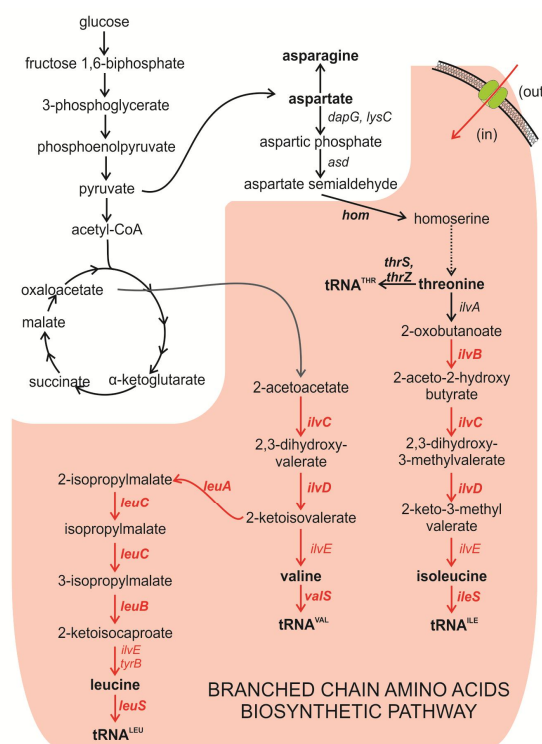
Figura 9. **Interconversión de cisteína y metionina.** YrhB y Yjcl son 50% idénticos a nivel de la secuencia de aminoácidos, lo que hace difícil determinar a través de la comparación de secuencias de genes, cuál tiene como producto la cisteína y cuál la cistationina. Sin embargo, estos genes están regulados por riboswitches en los Firmicutes, lo que hace posible la diferenciación de aminoácidos que sintetizan. Por ejemplo, en *B. subtilis*, *yjcl* está regulada por una S box, un riboswitch que pueden detectar los niveles intracelulares de S-adenosil metionina (SAM). Teniendo en cuenta esta información, es posible distinguir la dirección de la reacción enzimática para la metionina, que está de acuerdo con los datos experimentales. En otros Firmicutes, este gen no está regulado por una S box, sino por una T box. Teniendo en cuenta que somos capaces de predecir las interacciones Watson-Crick de la T box y el tRNA que se reconocería, hemos sido capaces de discriminar entre tRNA<sup>Cys</sup> o tRNA<sup>Met</sup> y por lo tanto predecir la dirección de la reacción enzimática que se llevará a cabo por la regulación gen<sup>136</sup>.



d) En esta sección se cubrirán aquellos operones cuya función no parece estar relacionada directamente con la T box por la que están siendo regulados. Las bacterias no tienen elementos de regulación azarosos río arriba de un gen. Existe una explicación metabólica a por qué se ha seleccionado esa T box para regular un operón en particular. No siempre podremos entender esta relación, y muchas veces la razón es que el gen está mal anotado. A continuación abordo los casos que me han parecido más interesantes.

**Dos ejemplos en Clostridias anaeróbicas: el operón *por* y el operón *eff*. Ambos dentro del metabolismo de los aminoácidos de cadena ramificada**

**Contexto:** El piruvato es el precursor común para los aminoácidos de cadena ramificada (o BCAAs por sus siglas en inglés: *Branched Chain Amino Acids*): isoleucina (Ile), leucina (Leu) y valina (Val). La vía biosintética para estos aminoácidos comparte los primeros cuatro pasos enzimáticos, codificados por *ilvC*, *ilvD*, *ilvE* e *ilvB-ilvN*, cuyos productos forman una enzima heterodimérica (Figura 10).



**FIGURA 10. Ruta biosintética de aminoácidos de cadena ramificada.** Los pasos dibujados en rojo representan estar regulados por una T box

En los *Firmicutes*, estos genes suelen estar agrupados de distintas formas en unidades transcripcionales: *ilvEBNCDA* o *ilvBNC-leuABCD*. Éstas se encuentran, en

su mayoría, reguladas por una T box que responde al tRNA<sup>Ile</sup> descargado, aunque algunas responden al tRNA<sup>Leu</sup>.

Identifiqué genes, en algunas *Lactobacillaceas*, que participan en la vía biosintética del pantotenato y se encuentran reguladas por BCAA T boxes. Esto resulta particularmente interesante, puesto la biosíntesis de Val, Leu y en pantotenato comparten el precursor común:  $\alpha$ -cetoisovalerato<sup>137</sup>. Una posible explicación para el uso de este tipo de regulación en los genes de biosíntesis del pantotenato, sería, que la célula tiene que “tomar en cuenta” el estado metabólico en el que se encuentra, para poder sintetizar BCAAs los metabolitos que requieren a los BCAAs como precursores.

Finalmente, quiero enfocarme en un organismo en particular, *Clostridium kluyveri*, el cual tiene características metabólicas excepcionales, como un metabolismo de azufre extremadamente activo, y que puede crecer anaeróbicamente en etanol o acetato como única fuente de carbono y de energía<sup>138</sup>. En este organismo, se identificaron 12 unidades transcripcionales reguladas por T boxes que responden a BCAAs, dos de las cuales corresponden a aminoacil tRNA sintetisas (aaRS): *leuS* e *ileS*, un transportador de Leu. De las nueve restantes, seis contienen genes de biosíntesis de BCAAs. Uno de los casos interesantes en el operón *porCDAB* el cual se encuentra regulado por una T box de Ile. Este operón codifica para proteínas similares a las subunidades de la piruvato: ferredoxin oxidoreductasa (POR). Esta enzima cataliza la descarboxilación oxidativa dependiente de tiamina pirofosfato del piruvato para formar acetil-CoA y CO<sub>2</sub><sup>139</sup>, una reacción que aparentemente no está relacionada con la biosíntesis de los BCAAs. Val e Ile son sintetizadas por los mismos cinco pasos de la vía metabólica, sólo difieren en el primer paso, el cual, para la biosíntesis de Ile, es una reacción dependiente de tiamina pirofosfato (la cual se lleva normalmente a cabo por IlvA).

Esta reacción es sumamente parecida a aquella que es realizada por las proteínas codificadas en el operón *por*<sup>140</sup>. Dado que identificamos una T box de Ile río arriba del operon *porCDAB* de *C. kluyveri*, proponemos que estas enzimas pueden estar relacionadas con la biosíntesis de Ile. Cabe recalcar que *Syntrophomonas wolfei* tiene dos copias del operon *porCDAB*, una de ellas regulada por una T box de Ile, pero la otra por una T box de Leu. Esta última relación no es clara, pues la reacción propuesta es, aparentemente, específica de la vía biosintética de Ile.

*C. kluyveri* tiene dos copias del operón *etfBA*, el cual codifica para las subunidades de una flavoproteína de transferencia de electrones que participa en el ciclo de reducción de 5-crotonil-CoA a 5-butilil-CoA<sup>138</sup>. En este ciclo, siete moléculas de NADH son oxidadas a NAD<sup>+</sup>. Una de las copias del operon *etf* está regulada por una T box de Ile. No he logrado comprender la relevancia fisiológica de que un módulo de transferencia de electrones esté siendo regulado por deficiencia de Ile, cuyo complejo proteico debe, dada su naturaleza, estar asociado a la membrana de la bacteria. Cabe recordar que la Ile es uno de los mayores reguladores de procesos metabólicos básicos en *B. subtilis* vía su interacción con la proteína regulatoria CodY<sup>141</sup>. Una posibilidad, es que niveles bajos de Ile señalicen una activación en el flujo de electrones hacia la biosíntesis de aminoácidos. Otra posibilidad es que, dado que el complejo POR mencionado anteriormente, requiera de muchos donadores de electrones, especialmente en organismos como éste que tienen un potencial redox limitado, y que de esta manera las proteínas codificadas por el operon *etf* sirvan como donadores de electrones. En la figura 10 se muestra una vía propuesta donde pudieran participar estas enzimas.

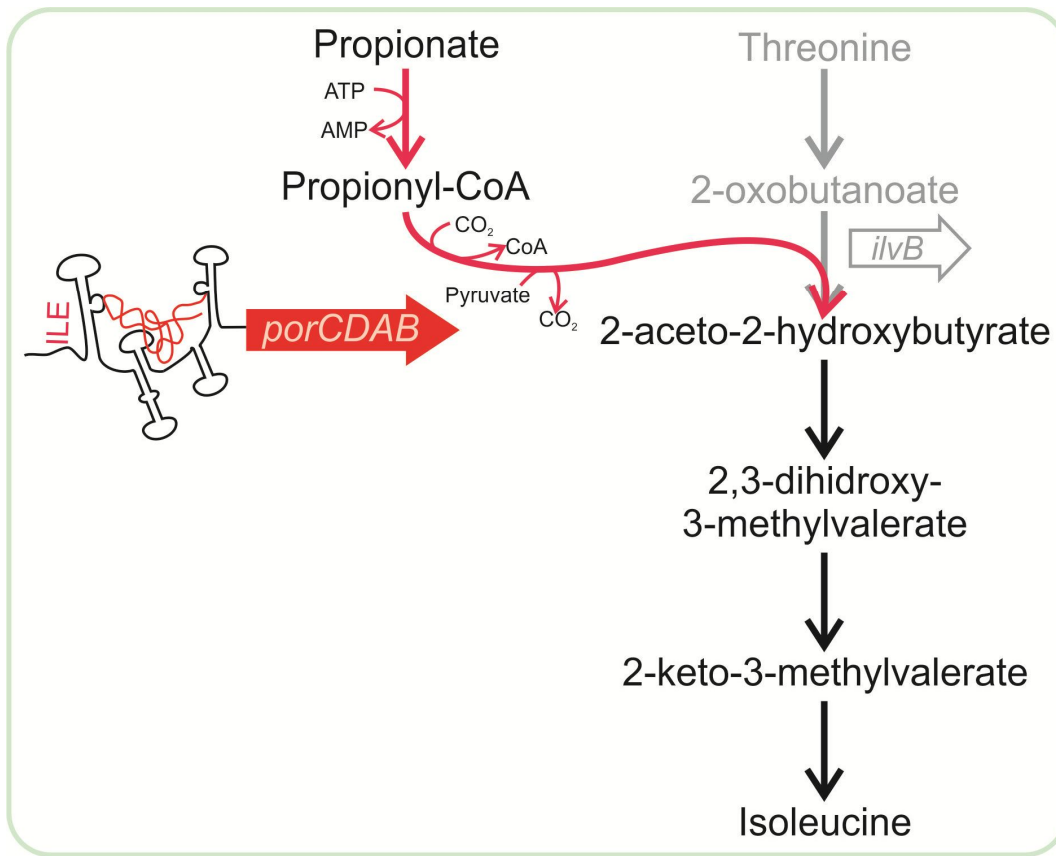


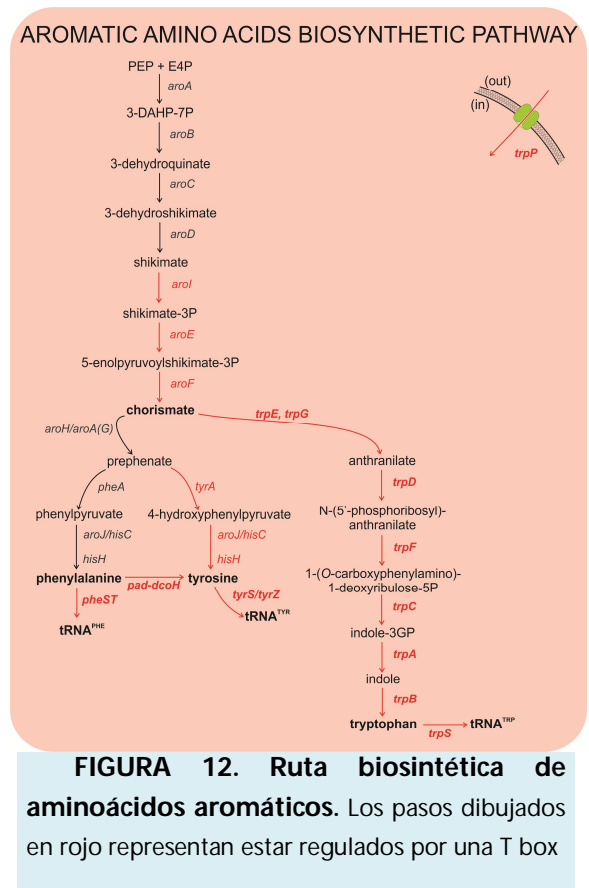
Figura 11. **El papel del operón *por* en la biosíntesis de isoleucina.** El operón *por* está regulada por una T box en respuesta a los niveles intracelulares de Ile. Esto sugiere que el complejo de la proteína podría participar en los dos primeros pasos de la biosíntesis de Ile utilizando diferentes sustratos en lugar de catalizar la reacción canónica que une la glucólisis con el ciclo de Krebs de la descarboxilación del piruvato para formar acetil-CoA y CO<sub>2</sub>

Adicionalmente, *C. kluyveri* tiene otro operón regulado por una T box de Val, el cual codifica para *fhsA* (formate-tetrahidrofolato [THF] ligasa), *fchA* (metenil-THF ciclohidrolasa) y *folD* (THF deshidrogenasa), genes que participan en la biosíntesis del THF. El hecho de que este operón esté regulado por una T box de Val, es consistente con la observación de que *valS* está mayoritariamente cotranscrita con *folC* en los Firmicutes. El THF es un cofactor importante que dona unidades de un carbono en la biosíntesis de metionina, purinas, timina y ácido pantoténico. La biosíntesis del ácido pantoténico se ramifica de la biosíntesis de Val a partir de la ceto-valina. Este primer paso se lleva a cabo por una hidroximetil transferasa, la cual usa metileno THF como cofactor. El THF puede ser sintetizado

cuando los niveles de Val son deficientes, para asegurar la biosíntesis de THF para poder sintetizar ácido pantoténico y Val.

### Casos involucrados en la biosíntesis de aminoácidos aromáticos

**Contexto:** El corismato es el precursor común de los tres aminoácidos aromáticos (Figura 12). Puede formar antranilato para generar triptófano (Trp), o pefenato que es el precursor común de fenilalanina (Phe) y tirosina (Tyr). Muchas de las proteínas que participan en la vía biosintética del corismato están codificadas en genes que se encuentran en el supraoperón *aro-trp* en *B. subtilis* y sus parientes cercanos<sup>142,143</sup>. Por el contrario, en el resto de los Firmicutes, los genes *aro* no están agrupados con los genes *trp*, exhibiendo una organización cromosomal distinta<sup>127,128</sup>. Los genes *aro* en las *Bacillaceas* distintas al grupo de *B. subtilis*, suelen estar regulados por T boxes que responden a tRNAs descargados de cualquiera de los tres aminoácidos aromáticos.



**FIGURA 12. Ruta biosintética de aminoácidos aromáticos.** Los pasos dibujados en rojo representan estar regulados por una T box

Los genes *phe* y *tyr*, llevan a cabo la conversión del corismato a Phe o Tyr, respectivamente y suelen estar regulados por T boxes de Phe y Tyr, de la misma manera que los genes *trp* suelen estar regulados por T boxes de Trp.

En *B. anthracis*, *Bacillus cereus* y *B. thuringiensis*, existe una ruta adicional para sintetizar Tyr a partir de Phe y biopterina. Esta reacción la llevan a cabo enzimas codificadas por el operon *pah-dcoH*, el cual está regulado por T boxes en tándem de Tyr.

### **Conclusiones: The Riboswitch Approach**

Las estrategias típicas para la asignación de funciones de los genes, en comparación en la Referencia<sup>144</sup>, son particularmente útiles para los genomas completamente secuenciados y de los genes homólogos con función anotada. Nuestro enfoque de asignación de funciones de los genes, que se basa en la predicción de riboswitches, es una importante alternativa a los métodos clásicos, y es especialmente adecuada cuando se trabaja con secuencias de metagenómica o genomas parcialmente secuenciados. También proporciona información más fisiológica, ya que se basa en la relación de la expresión génica, las necesidades metabólicas y la función del gen. Con nuestro método, el papel de genes específicos en las vías metabólicas se puede predecir, aún cuando sus homólogos no tienen ninguna función asignada, o cuando simplemente no tienen homólogos discernibles. Por lo tanto, la predicción de riboswitches, contribuye a poder mapear los genes en un contexto metabólico celular o en el de una comunidad microbiana, que nos lleva de una gran de secuencias en bruto a una visión más completa de la función génica en un contexto celular más amigable. En este estudio, hemos utilizado la T box como un ejemplo representativo de la asignación de funciones basado en la aproximación riboswitch. Sin embargo, la gran especificidad de todos los riboswitches por sus metabolitos correspondientes (ya sean nucleótidos, vitaminas, aminoácidos, co-factores, etc) puede ser explotado para la anotación de la función de los genes.

## Riboswitches en microorganismos de Cuatrociénegas

En la sección anterior, termina lo que fue formalmente mi proyecto de doctorado. En esta sección describo resultados de búsquedas de riboswitches en los genomas de Cuatrociénegas.

### ¿Por qué Cuatrociénegas?

Se eligió Cuatrociénegas porque sus pozas presentan el contenido más bajo de fósforo reportado en aguas continentales, lo cual ejerce una presión selectiva local en las comunidades biológicas. El valle de Cuatrociénegas tiene el mayor nivel de biodiversidad endémica en toda Norteamérica, esto se debe probablemente a que se encuentra aislado geográficamente. Estudios recientes han documentado la enorme abundancia y variedad de bacterias, *Archaea* y virus en las pozas de Cuatrociénegas<sup>145</sup>.

Estas condiciones con alta presión selectiva por falta de fósforo han permitido caracterizar nuevos tipos de organismos sub-representados en la colección de organismos comúnmente estudiados en el laboratorio, donde las condiciones de crecimiento difícilmente reproducen las del entorno natural. Los genomas de *Bacillus* que se tienen actualmente, pertenecen a cepas industriales (e.g. *Bacillus licheniformis*), cepas patógenas (como las pertenecientes al grupo *anthracis-cereus-thuringiensis*) o ambientales como *B. subtilis*. Sin embargo, en todos estos casos, existe una sobre-representación hacia cepas en ambientes con mayor disponibilidad de nutrientes. El tener organismos adaptados a distintos hábitats podría ayudarnos a entender mejor cuál es la relevancia del riboswitch T box en la regulación del metabolismo de los aminoácidos en las Bacillaceas.

## ¿Por qué Bacillaceas?

La presencia de *Bacillus* en distintos ambientes refleja una gran versatilidad en las capacidades metabólicas del género. Pueden usar una gran variedad de sustratos orgánicos e inorgánicos como fuente de nutrientes. De cualquier modo, las esporas se dispersan por aire y agua, lo cual, hace que no sea claro si una presencia tan robusta de estas bacterias en distintos ambientes es por la naturaleza resistente de las esporas, o si tienen una capacidad adaptativa que les permite adaptarse a distintos ambientes. El género *Bacillus* es el más representado en las bases de datos genómicas.

### *Bacillus coahuilensis* y *Bacillus M3-13*

Se eligieron *Bacillus coahuilensis* y *Bacillus M3-13* para ser estudiados porque estas bacterias tienen una historia evolutiva interesante, dada su adaptación a un ambiente oligotrófico. Son bacterias gram-positivas, Firmicutes, que producen esporas y fueron aisladas de una laguna en desecación en Churince, dentro del valle de Cuatrociénegas, Coahuila, México. Este ambiente rico en sulfatos pero extremadamente pobre en fósforo probablemente ha seleccionado adaptaciones genómicas únicas.

Por ejemplo, *B. coahuilensis*, ha sustituido parte de los fosfolípidos de la membrana por sulfolípidos. Esto como resultado de un posible evento de transferencia horizontal que le permitió adquirir un operón de cianobacterias que codifica enzimas pertenecientes a la biosíntesis de sulfolípidos. Otro ejemplo interesante de transferencia horizontal en *B. coahuilensis*, es la adquisición de una rodopsina sensorial, que probablemente le confiera cierta ventaja en un ambiente con tanta luz<sup>146</sup>. A pesar de que *Bacillus M3-13* y *B. coahuilensis* comparten el mismo ambiente y son organismos cercanos, han usado distintas estrategias para

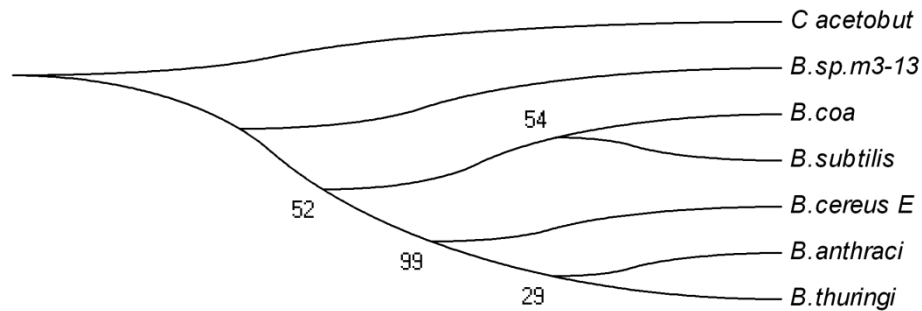


adaptarse a este ambiente oligotrófico. Por ejemplo, es probable que *Bacillus M3-13* asimile fosfonatos, pues tiene un gran número de genes que codifican para transportar y procesar fosfonatos.

El genoma de *B. coahuilensis* es hasta ahora el genoma más pequeño secuenciado de las *Bacillaceas*, y es por tanto un modelo excelente para evaluar la composición pangenómica y mínima de un grupo tan diverso como son las *Bacillaceas*<sup>146</sup>. *Bacillus M3-13* tiene un genoma de tamaño promedio de las *Bacillaceas*, i.e. alrededor de 4 millones de pares de bases. La genómica comparativa nos permitirá entender las características comunes del grupo de las *Bacillaceas*, así como las particulares sobre el uso de los riboswitches para regular la expresión de los genes.

Estudios recientes sobre la distribución filogenética de los riboswitches abordan el probable origen de estos elementos regulatorios<sup>95,96,102,147</sup>. Con el objetivo de extender estos análisis, hice un estudio comparativo de las T boxes presentes en *B. coahuilensis*, *Bacillus M3-13* y en sus parientes cercanos, *B. anthracis*, *B. cereus*, *B. thuringiensis*, *B. subtilis* y *C. acetobutylicum*.

Se eligió la comparación con *B. anthracis*, *B. cereus* y *B. thuringiensis* porque son parientes cercanos, y porque el número de T boxes presentes en estos tres organismos es mucho mayor que las presentes en los bacilos de Cuatrociénegas.



**Figura 13.** Filogenia inferida con máxima verosimilitud (PhyML) usando el 16S.

## Riboswitches

### T boxes en las Bacillaceas de Cuatrociénegas

Genoma de referencia: *Bacillus subtilis*.

*B. subtilis* cuenta con 19 T boxes: 13 regulando aminoacil tRNA sintetasas (*serS*, *thrS*, *thrZ*, *tyrS*, *tyrZ*, *trpS*, *ileS*, *glyQS*, *alas*, *hisS-aspS*, *pheST* y *leuS*); 4 regulando genes biosintéticos (de las vías de cisteína, prolina y aminoácidos de cadena ramificada); 1 que regula un gen de transporte (de leucina); y una que regula un gen regulador (AntiTRAP).

*Bacillus M3-13* presenta una regulación por T box similar a la de *B. subtilis*, pues también cuenta con 19 T boxes: 16 regulando aminoacil tRNA sintetasas (*serS*, *proS*, *lysS*, *tyrS*, *ileS*, *ileS2*, *trpS*, *argS*, *tyrS*, *valS*, *alaS*, *pheST*, *thrS*, *glyQS*, *hisS-aspS* y *leuS*); 2 regulando genes biosintéticos (de las vías de cisteína y prolina); 1 que regula un gen de transporte (de tirosina).

*Bacillus coahuilensis*, en general, tiene un número menor de riboswitches que los esperados en una familia de *Bacillaceas*. Cuenta con 16 T boxes: 12 regulando aminoacil tRNA sintetetasas (*serS*, *trpS*, *ileS*, *proS*, *glyS*, *hisS-aspS*, *alaS-foIC*, *pheST*, *tyrS*, *thrS* y *leuS*); 2 regulando genes biosintéticos (de las vías de cisteína y triptófano, este último con T boxes en tándem); 1 que regula una nitrorreductasa en ausencia de tirosina y 1 para la expresión de un probable transportador de histidina.

Estos organismos tienen un número parecido de T boxes. Un caso extremo, son las *Bacillaceas* patógenas del grupo de *Bacillus cereus* que cuentan, en promedio, con 40 T boxes por genoma, y en el otro extremo están las *Mycobacterias* que no cuentan con ninguna T box.

Es interesante notar, que *B. coahuilensis* y M3-13, fueron aislados del mismo nicho y son filogenéticamente cercanas. Uno pensaría que tienden a usar las mismas estrategias de regulación, sin embargo no es así. En el caso del operón de triptófano, es importante recalcar que las *Bacillaceas* y en general los Firmicutes, tienen dos maneras de regular sus genes de biosíntesis de triptófano. La mayoría de los Firmicutes prefieren regularlo por T boxes de Trp en tándem, mientras que *B. subtilis* y sus hermanos cercanos prefieren regularlo por la proteína TRAP<sup>127,128</sup>; ver Figura 14.

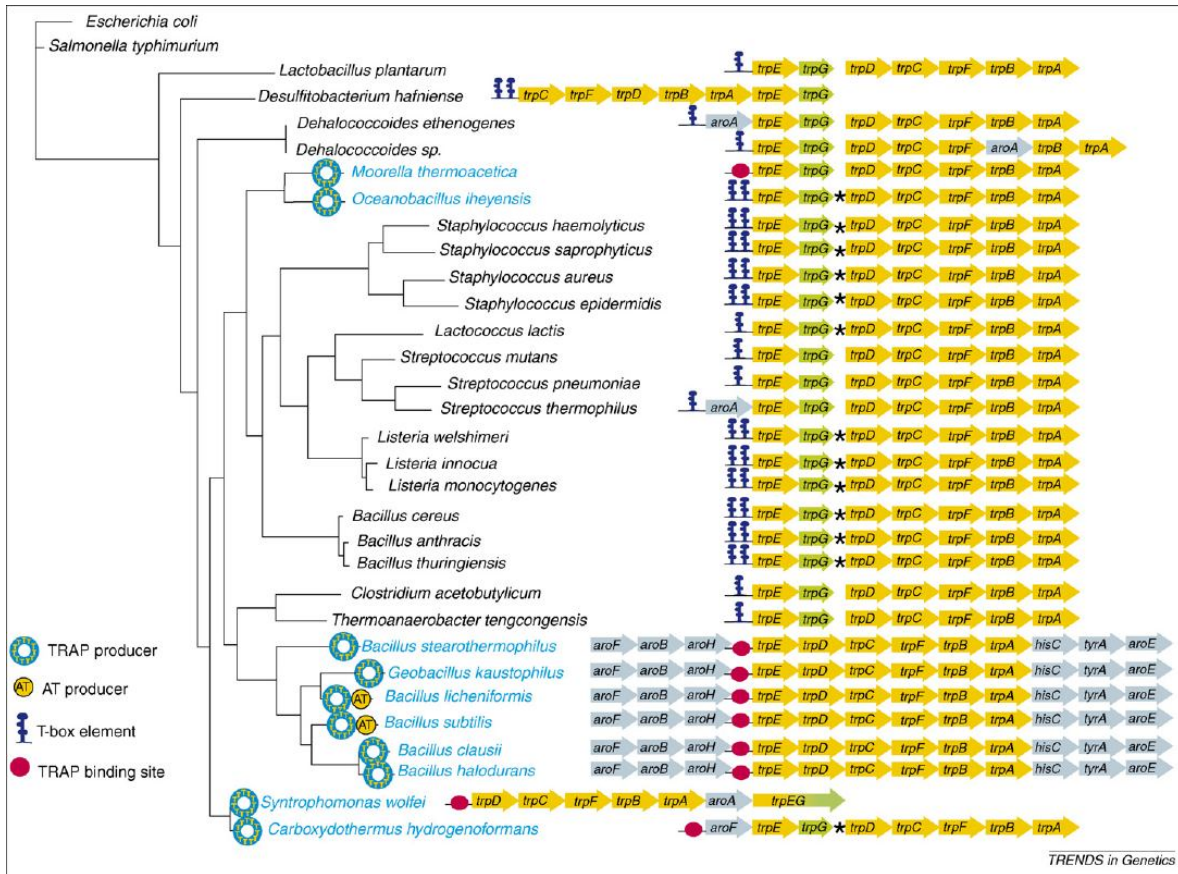


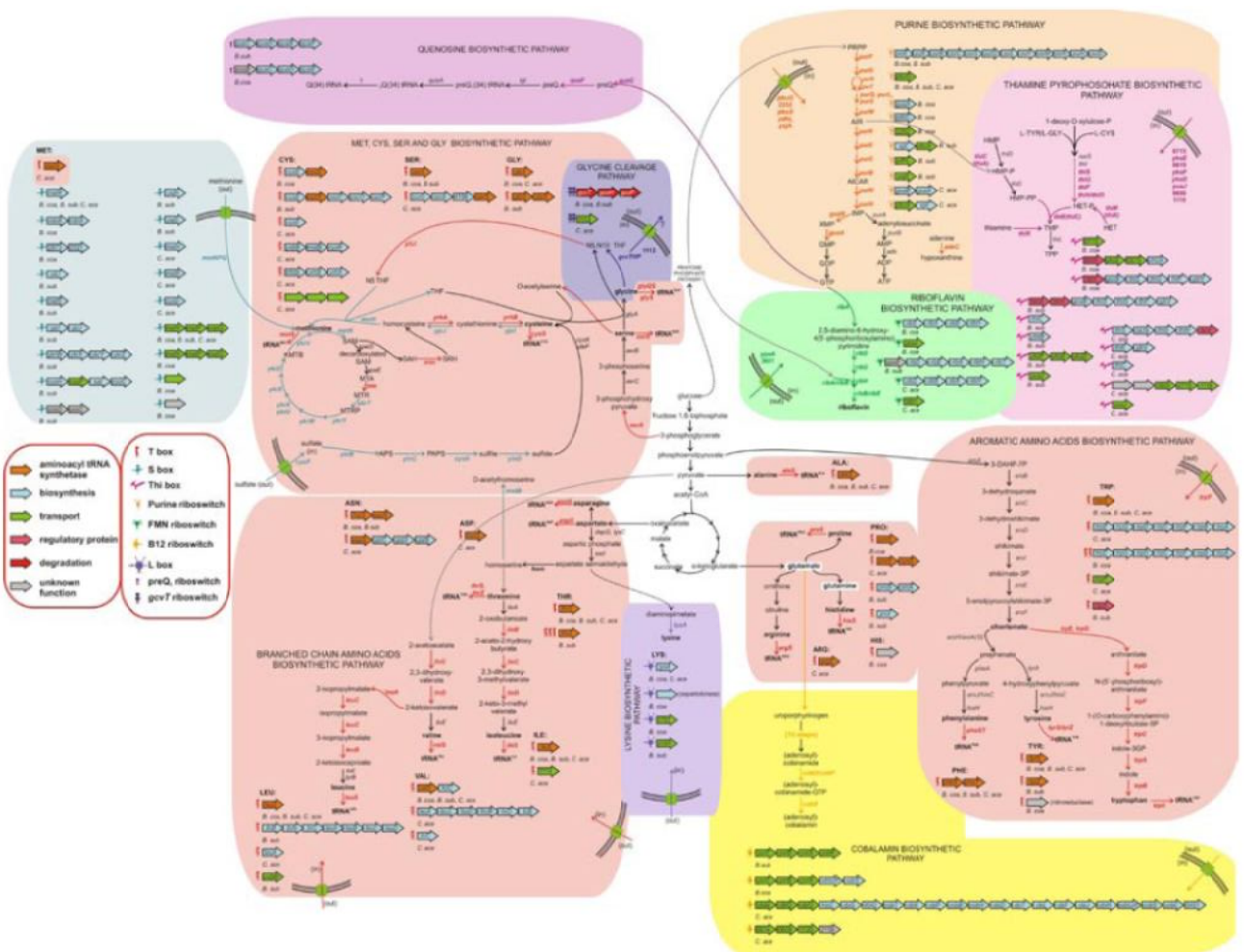
Figura 14. Contexto génico e historia evolutiva de los genes que participan en la vía biosintética de triptófano en Firmicutes. Los nombres de las especies que están en azul, tienen el gen que codifica para un ortólogo de *mtrB* (TRAP) en su genoma. Dos especies contienen el gen que codifica para AT. Los organismos que tienen TRAP pertenecen a las Clostridias o a las Bacillaceas cercanas a *B. subtilis*. Los organismos del clado de Clostridia tienen a *trpG* en su operon *trp*, mientras que *B. subtilis* y sus hermanos cercanos, no. Además, el operón *trp* de *B. subtilis* y sus parientes, existe como un suboperón *trp*, dentro de un supraoperon *aro*<sup>127,128</sup>.

*B. coahuilensis* tiene un contexto génico y regulatorio como el de las demás *Bacillaceas*, es decir, un operón *trp* con todos los genes necesarios para la vía completa, siendo estos regulados por T boxes TRP en tándem. Por el contrario, *Bacillus* M3-13 es más parecido a *B. subtilis*, pues tiene un operón *trp* sin *trpG* y dentro del supraoperón *aro*. Además, presenta TRAP binding sites en la región intergénica del operón *trp*. Este sitio de pegado es canónico (11 repeticiones

GAG/UAG espaciadas por dos o tres nucleótidos entre sí) y se encuentra, como es de esperarse seguido por un terminador transcripcional. Esto nos ayudó a definir mejor el inicio de traducción real, pues el gen *trpE*, está mal anotado en M3-13. *Bacillus* M3-13 tiene ortólogo de TRAP pero no de AntiTRAP.

Esto contrasta con el árbol expuesto en la Figura 13, donde por el gen 16S ribosomal parece que *B. coahuilensis* es más cercano a *B. subtilis* que *Bacillus* M3-13. Esto deja abiertas las siguientes preguntas: *i)* ¿*Bacillus* M3-13 adquirió la regulación por TRAP así como el contexto génico del operon *trp* por transferencia horizontal? (poco probable), *ii)* ¿es correcto el árbol filogenético de la figura 13?, *iii)* ¿Permite el gene 16S obtener la resolución suficiente para establecer relaciones filogenéticas entre estas tres bacterias? Nuestro punto de vista al respecto es que el 16S no presenta la suficiente resolución para entender la historia evolutiva dentro de las Bacilaceas, a diferencia de lo que se ha visto ya para la mayoría de las bacterias.

Figura 15. Metabolismo en las Baciláceas de Cuatrociénegas y su regulación por riboswitches.



### S boxes en las *Bacillaceas* de Cuatrociénegas

*B. subtilis* cuenta con 11 S boxes que regulan genes de biosíntesis de metionina y cisteína, vía de reciclaje de metionina (methionine salvage pathway), asimilación de azufre, de transporte de metionina y de cisteína así como genes de función desconocida.

*Bacillus M3-13* tiene un regulón más pequeño dirigido por 6 S boxes, las cuales regulan genes de la vía de reciclaje de metionina, genes de transporte (probablemente de metionina) y genes hipotéticos.

*Bacillus coahuilensis*, tiene, al igual que *Bacillus M3-13* un regulón pequeño encabezado por 7 S boxes, aunque los genes regulados difieren: 2 pertenecen a la vía de reciclaje de metionina, uno a la vía de biosíntesis de metionina, 2 transportadores de metionina, y dos genes hipotéticos.

### Elemento SECIS en las *Bacillaceas* de Cuatrociénegas

El elemento SECIS no es un riboswitch en sentido estricto, pero sí un elemento en *cis* de RNA y no reconoce exactamente al aminoácido, pero pienso que es importante considerarlo. La selenocisteína (Sec) se inserta co-traduccionalmente en las proteínas en respuesta al codón de paro UGA. Sec se encuentra en las proteínas participando en los sitios activos de oxidorreducción y es superior catalíticamente a la cisteína. Sec se usa de manera muy selectiva en proteínas y organismos. Estos codones UGA son reconocidos por una maquinaria celular compleja conocida como "selenosoma", la cual se acopla con la maquinaria de traducción. La maquinaria de inserción de Sec difiere en los tres dominios de la



vida, sin embargo parece ser que tiene un origen común<sup>148</sup>. El mRNA de la selenoproteína contiene un elemento de tallo y asa para la inserción de selenocisteína (SECIS) inmediatamente después del codón UGA que codifica para Sec. Este elemento SECIS une el factor de elongación (SelB) y forma un complejo con el tRNA<sup>Sec</sup> (codificado por *selC*), cuyo anticodon puede aparearse con el codón UGA. El tRNA<sup>Sec</sup> es inicialmente acetilado con serina por una seril-tRNA sintasa canónica, y después Ser-tRNA<sup>Sec</sup> se convierte a Sec-tRNA<sup>Sec</sup> por la Sec sintasa (SelA). SelA utiliza selenofosfato como donador de selenio, el cual es sintetizado por la selenofostato sintetasa (SelD)<sup>148</sup>.

SelD está presente en casi todos los phyla de bacterias, excepto en *Chlamydiae*, *Chlorobi*, y *Firmicutes/Mollicutes*. Dentro de los *Firmicutes*, sólo las *Clostridias* están reportadas por usar Sec en sus proteínas<sup>148</sup>.

Sorprendentemente, *Bacillus* M3-13 tiene elemento SECIS, y además cuenta con toda la vía de incorporación y síntesis de Sec: SelABD y con el tRNA<sup>Sec</sup> (*selC*). La única parte que no hace sentido es que el SECIS no está en una región codificante (bueno, sí pero la fase abierta de lectura está en la otra hebra).

Dado que es un caso extraño, se me ocurren tres hipótesis:

1. Hay en verdad un gen, en la otra hebra, que codifica para una selenoproteína y que no ha sido identificado.
2. Este organismo no incorpora Sec, sin embargo lo usa, quizá, para extraer fosfato del selenofosfato. Recordemos que el ambiente en donde vive, es uno muy pobre en fósforo, y quizá una ventaja que tenga sobre los demás organismos con los que comparte el nicho, es que puede aprovechar el selenofosfato que los demás no.



3. Este *Bacillus* usaba Sec, pero está en proceso de eliminarlo de su genoma pues el quitarlo le proveería cierta adecuación al no tener que sintetizarla o usar recursos extra en ella.

De realmente incorporar Sec, sería el primer *Bacillus* en ser reportado que haga esto.

Para probar la primera hipótesis, busqué ORFs y sí los hay y sí quedan en fase con el codón, UGA, sin embargo Blastx no arroja secuencias homólogas a esta proteína. Por lo que no puedo concluir mucho.

### Otros riboswitches o motivos de RNA presentes en las Bacillaceas de Cuatrociénegas

Por cuestiones de espacio, describo muy brevemente a cada uno, y lo presento los hits de manera muy resumida.

Nota: códigos de color. Naranja, riboswitch. Azul, genes de biosíntesis. Verde, genes de transporte. Rosa, genes regulatorios.

RF00023: *Bacterial tmRNA*, o 10Sa RNA o SsrA, se llama tmRNA por sus propiedades duales de tRNA y de mRNA. Su role es liberar el mRNA del ribosoma pausado (un hit significativo para *B. coahuilensis*; un hit significativo para *Bacillus M3-13*; genoma de referencia: *B. subtilis*: no hubo hits significativos).

RF00050: *FMN riboswitch* Flavin Mono Nucleotide. Riboswitch de riboflavina. (dos hits significativos para *B. coahuilensis*: río arriba del operón *ribEAH*: biosíntesis de riboflavina y de un gen que codifica para una proteína transportadora de riboflavina; dos hits significativos para *Bacillus M3-13*: río arriba del operón *ribEAH*:

biosíntesis de riboflavina y de un gen que codifica para una proteína transportadora de riboflavina; genoma de referencia: *B. subtilis*: un riboswitch, el cual regula al operón *ypuE-ribDEAHT*: biosíntesis de riboflavina).

RF00059: *TPP riboswitch (Thi box)*. (dos hits significativos para *B. coahuilensis*: río arriba del operón *tenA-ykoE-COG0619-ymcA*: activador transcripcional-dos proteínas de transporte- thiotransferasa y del operón *tenAI-goxB-thiSGF-yjbV*: TenI es un activador transcripcional, los demás genes pertenecen a la vía biosintética de tiamina; tres hits significativos para *Bacillus M3-13*: para el primero, se acaba el Contig, y sólo veo *thiT*, un transportador, (2) *tenA-ykoEDC*, (3) *thiEOSGF-yjbV-thiD*; genoma de referencia: *B. subtilis*: presenta cinco Thi boxes: (1) la primera regula al gen *thiC*: biosíntesis de tiamina; la (2) segunda al operón *tenAI-goxB-thiSGF-yjbV*: TenA y TenI son activadores transcripcionales, los demás genes pertenecen a la vía biosintética de tiamina; (3) se encuentra río arriba del operón *ykoFED* para el transporte de la tiamina; (4) *ylmB*: biosíntesis; (5) *yuaJ*: transporte).

RF00080: *yybP and ykoK probably a riboswitch. (M box)* (dos hits significativos para *B. coahuilensis*: uno río arriba de un gen hipotético y el otro río arriba de un gen que tiene similitud con una proteína para el control de la esporulación; cuatro hits significativos para *Bacillus M3-13*: tres están río arriba de genes hipotéticos, y el otro se encuentra río arriba de una posible proteína de membrana, pertenece a la familia de TerC; genoma de referencia: *B. subtilis*: presenta una M box río arriba del gen hipotético *yybP*).

RF00167: Purine *riboswitch*. (cinco hits significativos para *B. coahuilensis*: (1) río arriba del transportador *pubG*: xanthine/uracyl/vitamin C permease; (2) río arriba del gen de biosíntesis *guiA*; (3) se encuentra río arriba del operón de biosíntesis: *purEKBCSQLFMNHD*; (4) río arriba de un parálogo de *pubG*; (5) *adeC*: adenine deaminase; diez hits significativos para *Bacillus M3-13*: (1) río arriba de una phosphatidic acid phosphatase, (2) río arriba de una GMP sintasa, algunos nucleótidos después se encuentra el tercer (3) riboswitch de purina, río arriba del transportador *pubG*, (4) río arriba de una GMP reductasa, (5) *adeC*, (6) operón *xpt-pubG*: el gen *xpt* codifica para xanthine phosphoribosyltransferase, (7) *purEKBCSQLFMNHD*, (8) río arriba de un gen que codifica para "5'-methylthioadenosine/S-adenosylhomocysteine nucleosidase related protein VF1653" genera guanina, (9) río arriba de un parálogo de *adeC*; genoma de referencia: *B. subtilis*: presenta cinco riboswitches de purina: (1) río arriba del transportador *ydhL*: arabinose efflux permease; (2) río arriba del transportador *pbuG*; (3) se encuentra río arriba del operón de biosíntesis: *purEKBCSQLFMNHD*; (4) río arriba del operón *xpt-pbuX*; (5) *yxjA*: permeasa de nucleósidos).

RF00168: L box (*Lysine riboswitch*) (tres hits significativos para *B. coahuilensis*: (1) *lysA*: diaminopimelate (DAP) decarboxylase; (2) río arriba de una aspartate kinase podría ser *dapG* o *lysC*; (3) se encuentra río arriba de un transportador; dos hits significativos para *Bacillus M3-13*: (1) río arriba de un transportador, (2) río arriba de una aspartate kinase; genoma de referencia: *B. subtilis*: presenta una L box río arriba del transportador *yvsH*)

RF00169: *The signal recognition particle (SRP)* es una ribonucleoproteína universalmente conservada. Su función está relacionada con direccionar a las proteínas hacia las membranas. (un hit significativo para *Bacillus M3-13*; genoma de referencia: *B. subtilis*: no hubo hits significativos).

RF00174: *Cobalamin riboswitch* (un hit significativo para *B. coahuilensis*: río arriba del operón: *yvrCBA-cobUS*; tres hits significativos para *Bacillus M3-13*: (1) río arriba de un operón de tres genes que codifican para las distintas subunidades de la Ribonucleotide reductase (aerobic), (2) río arriba del operón *eutT-cobBQT-btuFC-feuC-cobUDPSC* donde *eutT*: cobalamin adenosyltransferase in ethanolamine utilization, *cobB*: Cobyric acid A,C-diamide synthase, *cobQ*: Cobyric acid synthase, *cobT*: Nicotinate-nucleotide--dimethylbenzimidazole phosphoribosyltransferas, *btuF* y *btuC*: transportadores de cobalamina, *feuC*: iron(III) dicitrate transport system (permease), *cobU*: Adenosylcobinamide-phosphate synthase, *cobD*: L-threonine 3-O-phosphate decarboxylase, *cobP*: Adenosylcobinamide-phosphate guanylyltransferase, *cobS*: Cobalamin synthase and *cobC*: Alpha-ribazole-5'-phosphate phosphatase; (3) río arriba de una proteína hipotética y un *transportador* predicho de cobalamina; genoma de referencia: *B. subtilis*: presenta un riboswitch de cobalamina río arriba de un operón de transportadores: *yvrCBA-yvqK*).

La regulación de los genes *yvrCBA* de transporte se encuentra conservada en los Firmicutes, sin embargo, tanto éstos como los genes de biosíntesis *cobUS* se encuentran normalmente regulados por este riboswitch en los Firmicutes, aquí *B. subtilis* es una excepción.

*RF00234: glmS Glucosamine-6-phosphate (GlcN6P) activated ribozyme.* GlmS es una enzima que usa la Fructosa6P y la glutamina para generar GlcN6P. La ribozima cataliza un corte en sí mismo, es decir, en su propio mRNA, autorregulándose negativamente. (un hit significativo para *B. coahuilensis*; un hit significativo para *Bacillus M3-13*; genoma de referencia: *B. subtilis*: un hit: *glmS* en los tres casos)

*RF00379:ydaO/yuaA element* Esta familia representa una estructura de RNA conservada, la cual se encuentra río arriba de los genes *ydaO* y *yuaA* de *B. subtilis* y de otras bacterias cercanas. El elemento *ydaO/yuaA* podría actuar como un switch que enciende los genes *ydaO* y *yuaA*. Se cree que este element funciona por un shock osmótico, el cual lleva a la activación de *ydaO*, predicho como transportador de aminoácidos, y al operón *yuaA-yubG* el cual codifica para las proteínas KtrA y KtrB: transportadores de K<sup>+</sup>. (un hit significativo para *Bacillus M3-13*: en el operon: **Trk system potassium uptake protein *trkA***-hypothetical protein. Esto apoya la teoría expuesta arriba, aunque no se sabe si lo que el RNA reconoce es el postasio)

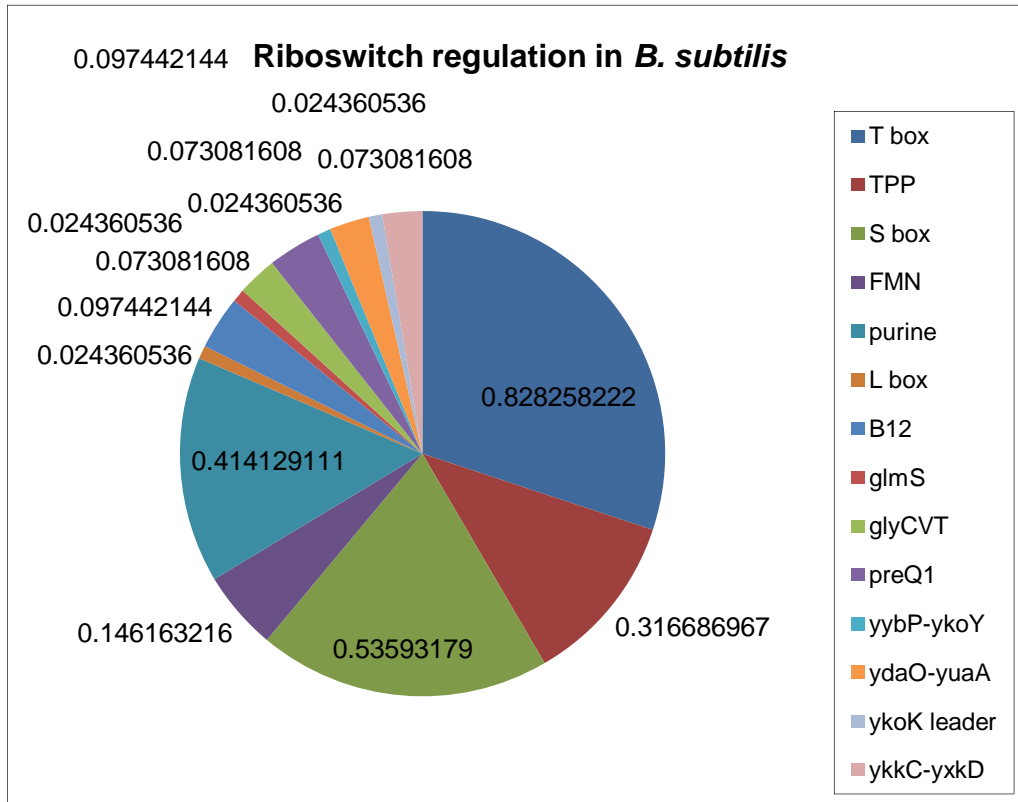
*RF00380: ykoK element (M box).* Esta familia representa una estructura de RNA conservada, la cual se encuentra río arriba del gen *ykoK* de *B. subtilis* y de genes con funciones relacionadas en otras. El roll regulatorio de este element es desconocido, sin embargo se sugiere que podría apagar genes en respuesta a un ligando metálico, como es el caso del riboswitch de cobalamina. (un hit significativo para *Bacillus M3-13*: en un operon con dos genes que son probablemente **transportadores de magnesio**. Esto apoya la teoría expuesta arriba y sugiere que el magnesio es el posible ligando)

RF00504: *Glycine riboswitch* (un hit significativo para *B. coahuilensis*; un hit significativo para *Bacillus M3-13*; genoma de referencia: *B. subtilis*: un hit: *gcvT-PA-PB* en los tres casos, y el riboswitch siempre se encuentra en tándem)

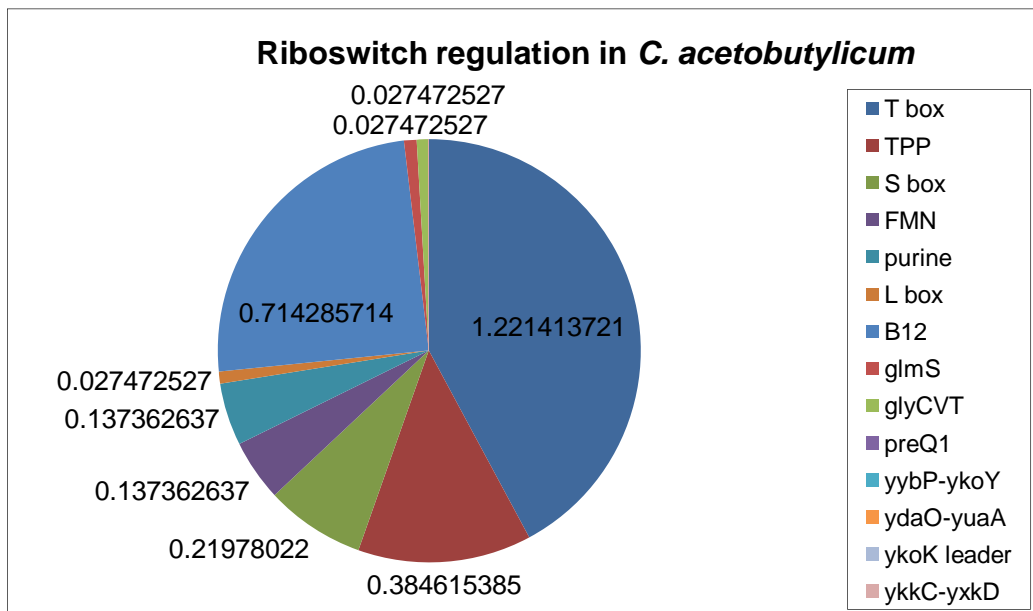
RF00522: *preQ1 riboswitch*. Esta familia de riboswitches bacterianos, unen el metabolito preQ1, un intermediario en la biosíntesis de queuosina, un nucleótido modificado de GTP, importante en el wobbling de algunos tRNAs. En *B. subtilis*, el riboswitch se encuentra en la región líder del operon *ykvJKLM (queCDEF)*, el cual codifica para los cuatro genes necesarios para la biosíntesis de queuosina. La unión de preQ1 a la región líder, podría inducir el término de la transcripción prematura (un hit significativo para *B. coahuilensis*: en el operon *exsB-queCDE* y dos hits significativos para *Bacillus M3-13*, en el operón *queCDEQ* y en su operón parálogo: *queCDEQ*).

### *Uso total de los riboswitches*

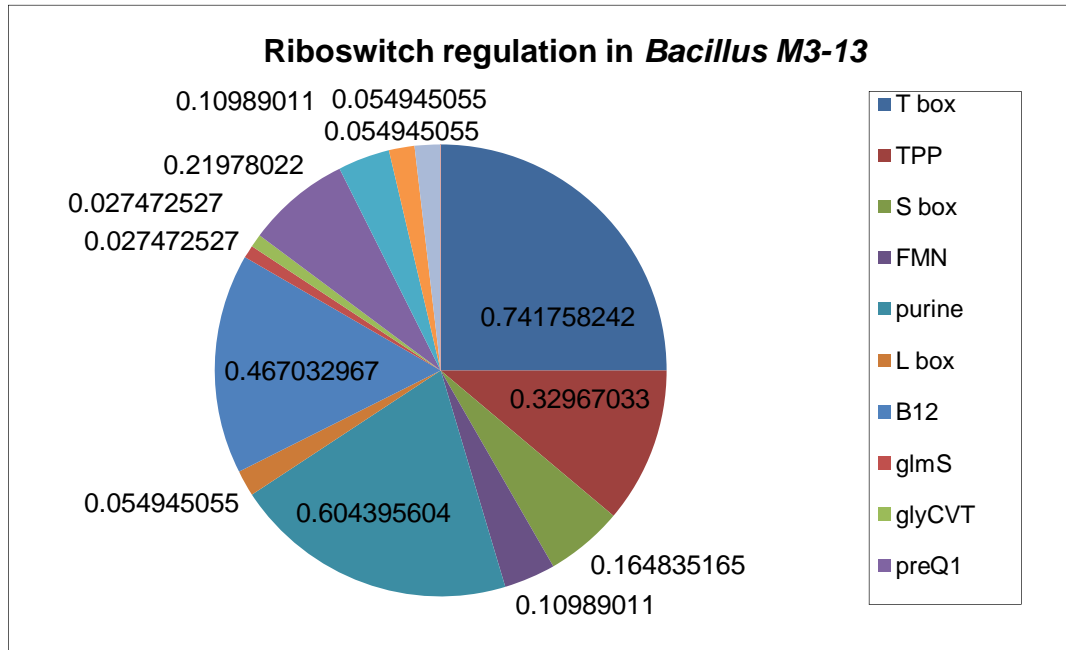
Los genes que se encuentran regulados por riboswitches se distribuyen, en porcentaje con respecto al número total de genes en los genomas correspondientes, de la siguiente manera:



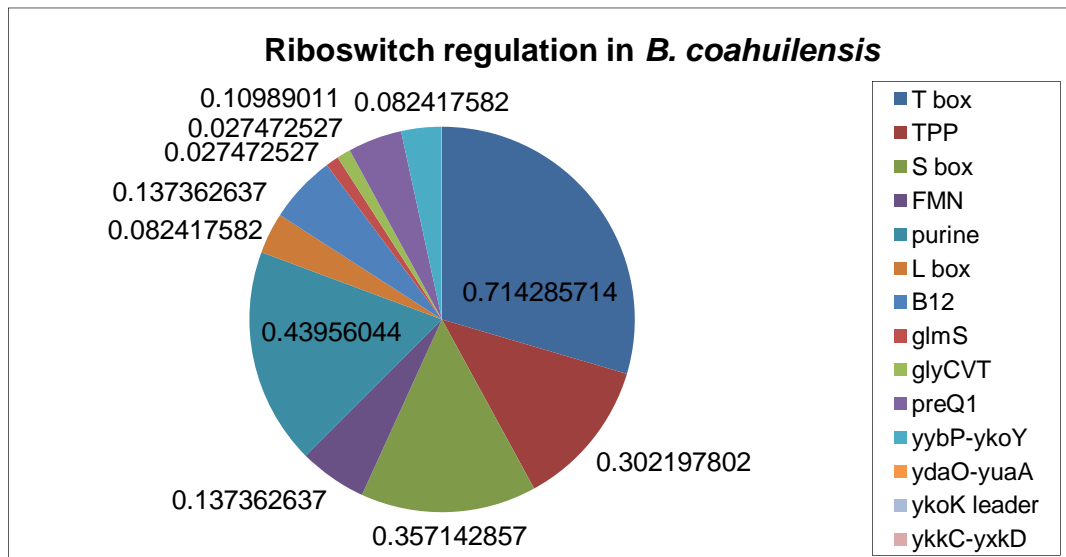
*B. subtilis*: 113 (de 4105) genes regulados por riboswitches (2.75%)



*C. acetobutylicum*: 114 (de 3848) genes regulados por riboswitches (2.96%)



*Bacillus M313*: 108 (de 4294) genes regulados por riboswitches (2.97%)



*B. coahuilensis*: 88 (de 3642) genes regulados por riboswitches (2.42%)

Figura 16. Representación por gráficas de *pay* del uso total de los riboswitches en *Bacillus M3-13*, *B. coahuilensis*, *B. subtilis* y *C. acetobutylicum*. La regulación por riboswitches en cada organismo es proporcional con el tamaño de la vía metabólica. La vía biosintética de las purinas y la vía de degradación de la glicina son ejemplos de grandes y pequeños regulones, respectivamente, y el número de genes regulados varía en consecuencia.



Como se puede apreciar en la figura 16, el uso total de los riboswitches es similar en los tres organismos, aunque diferencias importantes pueden observarse entre las *Bacillaceas* y la *Clostridia*.

*B. coahuilensis* es el organismo que menos riboswitches tiene, sin embargo no pienso que esto se deba al nicho ecológico que habita, si no al hecho de que tiene un genoma reducido.

Como ya se ha descrito<sup>96</sup>, la T box es por mucho el mecanismo de regulación más ampliamente usado en Firmicutes y está generosamente distribuido en las vías metabólicas. Alrededor del 1% de los genes de cada organismo está regulado por éste riboswitch que reconoce la proporción entre tRNAs cargados y no cargados.

#### *Comparación de la distribución de la T box de B. coahuilensis con el grupo B. anthracis/cereus/thuringiensis.*

Así mismo, de la Figura 7 se puede observar que los organismos filogenéticamente cercanos tienen una distribución parecida de T boxes. En base a esto, resulta extraño que *B. coahuilensis* tenga menos T boxes que sus “hermanos” (16 T boxes: 12 aaRS; 2 biosintéticos; 1 hipotético (transportador probablemente) y una nitrorreductasa). Al parecer *B. coahuilensis* es único, ya que la distribución de las T boxes de *Bacillus M3-13* es más parecida a aquella de la de *B. subtilis* que a la de su muy cercano *B. coahuilensis*. Cabe recordar que en organismos que tienden a una reducción génica masiva, como los patógenos o los endosimbiontes, pierden mucha regulación local, basándose más en reguladores globales y otras estrategias<sup>149</sup>.

## Conclusiones

Los riboswitches son elementos de alta importancia en la regulación metabólica. Casi el 3% de los genes en los organismos de estudio están regulados por este mecanismo basado en RNA. A pesar de las diferencias observadas en cómo se encuentran organizados los genes de distintas vías metabólicas, su expresión se basa en un mecanismo común.

Los genes involucrados en el metabolismo central son regulados similarmente en ambas especies de *Bacillus*. Aún así, existen algunos genes en *B. coahuilensis* cuya función es desconocida y son regulados por T box ó por S box. Pensamos que es posible que estos genes codifiquen para elementos que han ayudado a la bacteria a enfrentar las condiciones tan extremas como en las que habita durante el transcurso de su evolución, y son excelentes candidatos para entender mejor la adaptación de estos organismos a los ambientes oligotróficos que habitan. Esta hipótesis necesita validación experimental.

Además, se recalcó la importancia de estudiar la regulación de *B. coahuilensis* en otros genes donde no hay riboswitches, ya que cuenta con un menor número de genes regulados por este tipo de mecanismos. Las preguntas directas que surgen son: *i)* ¿*B. coahuilensis* presenta un mecanismo de regulación distinto en aquellos genes que son comúnmente regulados por riboswitches? De ser así, la hipótesis es que estos mecanismos de regulación distintos le confieren una ventaja adaptativa dado el nicho tan especial en donde vive; *ii)* o ¿quizá sólo carece de estos genes ya que tiene un genoma reducido?; *iii)* o bien, ¿presenta nuevas estrategias de regulación transcripcional, como por ejemplo, basadas en reguladores globales y/o en topología del DNA para “ahorrar” y mejorar su adecuación dado su ambiente?

## Riboswitches en *Epulopiscium*

En este caso, el modelo de estudio es una bacteria gigante, *Epulopiscium* sp. tipo B, que vive en el pez unicornio *Naso tonganus*<sup>150</sup>. Pertenece al phylum de las *Clostridias*, dentro de los *Firmicutes*. Estas células miden aproximadamente 200–300 µm de largo y 50–60 µm de ancho y se reproduce solamente por la formación de progenie interna. Esta estrategia reproductiva probablemente ha evolucionado de la formación de esporas<sup>151</sup>. Una célula de *Epulopiscium* contiene una gran cantidad de DNA (i.e., es poliploide) que se acomoda en la periferia del citoplasma. Esta peculiaridad puede ser la clave de cómo es que estas células mantienen un metabolismo activo a pesar de la baja razón superficie/volumen<sup>152</sup>.

Actualmente se ha secuenciado aproximadamente la mitad del genoma de *Epulopiscium*. En una colaboración con Esther Angert, hemos buscado riboswitches en este genoma parcial, así como otros elementos estructurales de RNA con la metodología descrita anteriormente para *B. coahuilensis*.

Los resultados son sorprendentes, puesto que dada la filogenia de este organismo, uno pensaría que tiene muchos riboswitches, en particular muchas T boxes (ver distribución de riboswitches para *C. acetobutylicum*).

En este genoma parcial, no se han encontrado T boxes, S boxes, L boxes, riboswitch de FMN, de purina, ni de glicina. Esto puede deberse a alguna peculiaridad del organismo, a que tenga un genoma reducido por vivir en simbiosis con el pez unicornio, o a que tenemos muy poca secuencia todavía.

Los riboswitches que sí se han encontrado son:

- Un **riboswitch de tiamina** regulando al operón *thiK-tenI-tenA*-COG0600, donde el primer gen es de biosíntesis, los dos siguientes son reguladores de la vía de tiamina y el último es un transportador, probablemente de tiamina.

- Un **riboswitch de cobalamina** que regula un operón de 20 genes similar al encontrado en *C. acetobutylicum*, donde la mayoría de los genes son de biosíntesis y de transporte de cobalamina con un gen de función desconocida.

- Se encuentra también un **riboswitch de quenosina**, pero no logré identificar los genes río abajo.

- Un **riboswitch de “yybP-ykoY leader”** que es un riboswitch hipotético. Este riboswitch se encuentra normalmente regulando a los genes *yybP* y *ykoY* (de ahí su nombre), que son genes de función desconocida. En el caso de *Epulopiscium* parece estar río arriba del gen que codifica para la valil tRNA sintetasa. Esto podría ser erróneo, pero de ser correcto aportaría información sobre la posible función de este riboswitch.

- Además de riboswitches, encontré un elemento de SECIS, que funciona para la incorporación del aminoácido modificado, selenocisteína, a las proteínas de este organismo. En colaboración con Héctor Romero (del Laboratorio de Organización y Evolución del Genoma, Dpto de Biología Celular y Molecular, Instituto de Biología, Facultad de Ciencias, Montevideo, Uruguay), hemos determinado que definitivamente este organismo tiene la maquinaria necesaria para incorporar selenocisteína.

Se requiere que el genoma esté totalmente secuenciado para obtener conclusiones del estudio.

## Conclusiones

Las conclusiones puntuales ya fueron descritas en cada sección de Resultados y discusión. Sin embargo, una conclusión general de esta tesis es que muestra un avance importante no solamente en identificar elementos de regulación en genomas, si no que la riqueza de este análisis surge al poder ubicar cada elemento identificado en un contexto metabólico. Esto amplía nuestro conocimiento, no sólo del elemento de regulación *per se*, si no también de cómo modula la célula bacteriana su metabolismo correspondiente a aminoácidos, y cómo emplea estas estrategias para incrementar su adecuación, así como cómo utiliza las ventajas de utilizar la T box para coordinar la biosíntesis de sus aminoácidos así como de otros procesos metabólicos. Hasta donde sabemos, es la primera vez que se logra un mapeo metabólico a partir de la identificación de los elementos de regulación.

Al principio, se plantea un Objetivo general, que plantea poder entender a fondo el regulón del riboswitch T box. Si bien no logramos entenderlo aún en su totalidad, si logramos un avance importante en esto. Podemos entender sus generalidades así como sus particularidades. Pudimos entender por qué la T box regula ampliamente el metabolismo de algunos aminoácidos, mientras que otros están escasamente controlados por este mecanismo. Por ejemplo, existen T boxes de histidina, pero son muy pocas, lo cual parece tener congruencia con el hecho de que el tRNA de histidina es ligeramente distinto al de los demás aminoácidos (tiene un nucleótido menos en el brazo aceptor), siendo ésta probablemente la causa de que no sea una interacción tan estable entre éste y su T box, lo cual hace que las bacterias prefieran utilizar otras estrategias regulatorias. Entendimos también que la T box puede presentar grandes cambios a nivel de secuencia, sin embargo son

pocos los cambios permitidos a nivel de estructura. Esto se demostró claramente con el organismo *T. marianensis*, donde las secuencias son muy distintas, sin embargo la estructura es respetada, así como los motivos estructurales del mRNA.

Gracias a este trabajo sabemos que las restricciones evolutivas que presenta la T box están impuestas por cómo ha de reconocer ésta a su tRNA, así como el mantener un estado metabólico acorde con el medio ambiente en el que la bacteria se encuentra. La T box es de gran ayuda para la mayoría de los Firmicutes en poder coexpresar los genes para poder mantener disponibles los niveles de cierto aminoácido.

Las lecciones que este proyecto nos deja son muchas, pero quiero dedicar este último párrafo a la importancia de entender la regulación para poder aprender más sobre el metabolismo bacteriano. Estamos tan acostumbrados a una anotación génica vía homología de secuencias que muchas veces perdemos de vista que las bacterias son seres que viven en constante presión selectiva y que evolucionan rápidamente para poder adaptarse a su cambiante entorno. Como muestra de esto es que un gen que sirve normalmente para una reacción en una vía particular, puede de pronto encontrar que también es útil para otro proceso celular distinto, que le conferirá ciertas ventajas al organismo. El entender la regulación génica nos permite pasar de una anotación por homología a un entendimiento mayor sobre el gen en cuestión, podemos saber cuándo será expresado y en qué momento la bacteria requerirá la proteína. Para aterrizar esto, me gustaría tomar el ejemplo del operon *por* en algunas Clostridias, el cual fue descrito en la sección: Dos ejemplos en Clostridias anaeróbicas: el operón *por* y el operón *etf*. Ambos dentro del metabolismo de los aminoácidos de cadena ramificada. Los genes del operon *por* codifican para un proceso que se encuentra normalmente en el metabolismo central bacteriano, permitiéndole al organismo utilizar el propionato como un

donador de electrones, cuando las condiciones celulares así lo requieren. Las clostridias en cuestión viven en ambientes anaeróbicos cambiantes donde no saben qué donador de electrones tendrán disponible en el ambiente. Cuando es el caso de que existe propionato la célula lo usará para sus procesos celulares centrales gracias a las proteínas codificadas por el operon *por*. Al encontrar una copia de este operon siendo regulado por deficiencia de isoleucina, podemos entender que la célula no sólo está usando el propionato para generar poder reductor, pero que también lo está utilizando para sintetizar aminoácidos, en este caso, aminoácidos de cadena ramificada. Esto le asigna a este operon una función distinta a la canónica ya descrita, y esto sólo podemos saberlo al conocer su región regulatoria y al saber que estará siendo expresado cuando necesite sintetizar isoleucina.

Por último, nos queda una pregunta que no pudimos responder del todo: ¿Cómo es que se expande a regular nuevos genes relacionados al metabolismo de aminoácidos? Para poder responderla por completo, se plantea un proyecto de investigación en la siguiente sección (Perspectivas).

## Perspectivas

### 1. Entender la expansión del regulón T box mediante evolución experimental

*En colaboración con Mike Travisano*

*(Department of Ecology, Evolution and Behavior, University of Minnesota.)*

Como ya se ha visto, la T box modula principalmente la expresión de muchos genes involucrados en el metabolismo de aminoácidos, mayoritariamente en los Firmicutes.

*Características estructurales del mecanismo de regulación T box: un codón provee las bases para regular específicamente en respuesta a un aminoácido.*

Ya se ha descrito en la Introducción de la tesis cómo las similitudes estructurales de las regiones líder de varios genes regulados por la T box llevaron a la propuesta que eran regulados por el mismo mecanismo de regulación y cómo un asa que contiene *la secuencia de especificidad* en el Stem I, permite que cada líder reconozca un tRNA específico acorde con el codón presente en dicha secuencia (Figura 3)<sup>26,94</sup>. Sólo por contexto, recordemos que si se sustituye el codón UAC de tirosina en la región líder de *tyrS* por un codón UUC de fenilalanina, el resultado es que se pierde la inducción del gen en respuesta a una limitación de tirosina, y se gana en respuesta a una limitación por fenilalanina. Este fue uno de los experimentos que demostraron que este codón, presente en el Stem I y conocido como *secuencia de especificidad* es el determinante principal de la especificidad para responder a la limitación de un aminoácido<sup>33</sup>. Este tipo de experimentos se realizó para muchas sustituciones en *tyrS* y otras regiones líder bien estudiadas, dando como resultado general que la alteración de la secuencia de especificidad puede hacer un cambio directo en la especificidad del aminoácido



reconocido en algunos, pero no en todos los casos; y que la regulación es en general poco eficiente, lo que indica que existen elementos adicionales que son determinantes en la respuesta a un aminoácido<sup>94,153-157</sup>.

### *Requerimientos en el RNA líder*

Los RNAs líder que pertenecen a la familia T box, están compuestos de una serie de arreglos que contienen características muy conservadas (Figura 3). El patrón de estructura básico que puede predecirse por análisis de genómica comparativa se ha demostrado química y enzimáticamente con la región líder de *glyQS* en *B. subtilis*, *in vivo* e *in vitro*<sup>94,155</sup>. Mutaciones que afectan los elementos conservados en el líder de *tyrS* tienen como resultado general una transcripción continua. i.e., el antiterminador no puede formarse de forma dependiente del pegado del tRNA. Esto quiere decir que los elementos de estructura conservados son importantes funcionalmente. Resultados similares se observan en otros líderes con T box. A pesar de esto, algunas características de la región líder no se encuentran, o se encuentran distintas en algunas secuencias o en algunos grupos de secuencias, pero la mayoría de las T boxes identificadas se ajustan al patrón que se ve representado en la región líder del operón *tyrS* de *B. subtilis*<sup>96</sup>. El papel que juegan la mayoría de estos elementos conservados, no es del todo claro aún. Las posibilidades obvias incluyen roles en el reconocimiento del tRNA, en la estructuración apropiada del RNA líder, en la unión de factores adicionales que pudieran ser necesarios para el reconocimiento RNA-tRNA, o bien, interacciones con la maquinaria transcripcional. De toda la región líder, la secuencia más altamente conservada es la propia T box, que forma el lado 5' del antiterminador e incluye los cuatro nucleótidos (UGGN) que se aparean con el brazo aceptor del tRNA (NCCA). Los siguientes tres nucleótidos que también forman parte del asa del antiterminador (ACC) están altamente conservados; la A está conservada en un

100% de las secuencias, mientras que las C's varían ocasionalmente<sup>93,94</sup>. Mutaciones en cualquiera de estas posiciones, afecta dramáticamente la antiterminación, lo que nos indica que estos residuos son de vital importancia para su funcionamiento.

### **Requerimientos del tRNA**

La capacidad de otros tRNAs, fuera del tRNA<sup>Tyr</sup>, para interactuar con la región líder de *tyrS*, fueron evaluados mediante mutaciones en los determinantes conocidos del RNA (i.e., la secuencia de especificidad y la posición variable en el antiterminador) para la interacción con el tRNA, buscando que pudiesen unir nuevos tipos de tRNA, y que pudieran inducir una respuesta para esta nueva limitación de algún aminoácido<sup>93,156</sup>. Esta metodología se aplicó a una gran cantidad de genes regulados por T box, y el resultado general fue que en algunos casos, era posible cambiar la especificidad del reconocimiento a un tRNA, otros tRNAs interactuaban ineficientemente con algunas regiones líder, mientras que otros no lo hacían en absoluto<sup>94,153-157</sup>. Este conjunto de resultados, sugiere que existen elementos adicionales involucrados en la interacción del líder con el tRNA. Se evaluaron los requerimientos estructurales del tRNA<sup>Tyr</sup> para la antiterminación de *tyrS*, mutando varias posiciones y usando un sistema en que la función del tRNA en la antiterminación podía ser examinada independientemente de los requisitos para la síntesis del tRNA o de otras actividades celulares<sup>94,158</sup>. Mutaciones en las hélices del tRNA<sup>Tyr</sup> son permitidas siempre y cuando se mantenga la estructura. Además, el brazo largo y variable del tRNA<sup>Tyr</sup> puede ser reemplazado por un brazo corto variable o por una inserción de una hélice larga. Sin embargo, alteraciones que afectan la estructura secundaria o terciaria del tRNA interrumpen la actividad de antiterminación, lo que sugiere que la interacción del tRNA con el líder es dependiente de la estructura terciaria completa del tRNA. Para la expresión de los genes de triptófano en *L. lactis*, se requiere una interacción entre los brazos D y T

del tRNA con el líder<sup>94,98,157</sup>, sin embargo, los apareamientos propuestos no se conservan en análisis de covarianza con otros líderes o en otros estudios con mutaciones.

### *Duplicación y expansión del regulón T box*

Es común encontrar duplicaciones de las T boxes. En los casos más simples, duplicados (o secuencias filogenéticamente muy cercanas) de las T boxes presentes en las aminacil tRNA sintetasas se encuentran regulando genes de transporte de aminoácidos. Por ejemplo, en *C. perfringens* y *C. beijerinckii*, el operón *alaRT* (el cual codifica para una probable transaminasa de alanina y un regulador transcripcional) se encuentra regulado por una T box de alanina, la cual es, probablemente, el resultado de una duplicación de la T box de alanina que regula el gen *alaS* en el ancestro común de estas bacterias. Las T boxes de alanina que se encuentran regulando el operón *alaRT* en otras especies forman, claramente, una rama filogenética distante<sup>95</sup>.

Duplicaciones múltiples de T boxes pueden llevar a una expansión rápida del regulón T box. La duplicación de la T box de treonina que se encuentra originalmente en el gen *thrS* llevó a la existencia de nuevas T boxes de treonina en los genes de transporte *brnQ* y *ykbA* en las especies de *Bacillus*: *B. anthracis*, *B. cereus* and *B. thuringiensis* y en sus genes de biosíntesis *hom* y *thrCB* en *C. difficile*. La duplicación también puede llevar a la existencia de elementos T box en tándem, o dobles parciales. En la mayoría de los casos, las T boxes en tándem están muy relacionadas unas entre otras (e.g., las T boxes en tándem de treonina del gen *hom* en *C. difficile*, o las T boxes en tándem de *thrZ* en *B. cereus*)<sup>95</sup>. En otros casos, nosotros<sup>96</sup> y otros grupos<sup>95</sup> detectaron una expansión reciente del regulón de T

box en *C. difficile*, la cual ocurrió probablemente ya que este organismo se había separado de las Clostridia para formar su propio linaje. Duplicaciones múltiples de una T box de arginina junto con la pérdida del regulador transcripcional de arginina, *AhrC*, en *C. difficile* llevó a la existencia de cinco T boxes nuevas de arginina, regulando tres unidades transcripcionales de biosíntesis de arginina, y dos de transportadores de arginina. El elemento original, la T box de arginina en el gen *argS* parece haberse perdido. Así mismo, la pérdida del riboswitch S box en las Lactobacillales (que regula la biosíntesis de metionina en algunas bacterias que pertenecen al grupo Bacillus/Clostridium) llevó a la expansión del regulón T box de metionina en este linaje. Por otro lado, los genes de biosíntesis de metionina en las especies de *Streptococcus* sp., están regulados por un nuevo factor de transcripción *MtaR/MetR*<sup>95,147,159</sup>. En algunos casos, no es posible determinar cuál es la T box ancestral. Este parece ser el caso de las T boxes de prolina que se encuentra río arriba de los operons de biosíntesis *proBA* y *proI* en *B. stearothermophilus*, *B. subtilis* and *B. licheniformis*. Estos operones aparecen asociados filogenéticamente como correspondería a su linaje vertical. En otros genomas, se encuentra el operón *proBA* regulado por una T box de Prolina, siendo éste, probablemente, el estado ancestral. Una duplicación de este operón en conjunto con la T box de prolina presente río arriba, y una pérdida subsecuente del gen *proI* en una copia pero de los genes *proBA* en otra, puede haber llevado a la configuración actual en estos tres genomas. Un escenario alternativo presenta una interrupción en el operón que lleva a una separación subsecuente de *proBA* y *proI* y una pérdida de regulación de *proI*, como parece haber sido el caso para *B. cereus*<sup>95</sup>.

Todo esto, nos lleva a querer entender cuáles son los procesos evolutivos que sustentan la expansión del riboswitch T box, tan obvia en los Firmicutes.

Un primer beneficio al entender estos procesos evolutivos, sería que lograríamos entender mejor cómo ha sido la evolución de la T box, así como la evolución de otros elementos de regulación. Otro beneficio adicional, sería que podríamos reforzar nuestro conocimiento sobre la importancia de las características estructurales de la T box. La evolución experimental puede ayudarnos a definir cuáles son los refinamientos que una T box necesita para poder reconocer óptimamente distintas especies de tRNAs. En un escenario más general, un proyecto así, podría proporcionarnos información nueva para poder entender la reproducibilidad de la evolución adaptativa.

#### *¿Por qué abordar este proyecto con evolución experimental?*

“Los cambios evolutivos ocurren en un contexto ecológico, pero trabajar en ese contexto puede ser infinitamente complejo”<sup>160</sup>. Por tanto, lo primero que debe considerarse es cómo reducir esa complejidad y como ser capaces de observar y medir los procesos mecánicos de la evolución. Para esto, las poblaciones bacterianas son ideales. Se propagan fácilmente, tienen tiempos generacionales cortos y son amigables para análisis a nivel genético. Los factores ambientales que afectan el crecimiento de la población se pueden controlar. La reproducción por fisión binaria garantiza población clonales que pueden establecerse con un fenotipo único. Desde la perspectiva de la ecología evolutiva, las poblaciones grandes y los tiempos generacionales cortos, aseguran que se pueda escalar los tiempos ecológicos y evolutivos. En otras palabras, se puede observar en tiempo real los cambios evolutivos en un contexto de ecología dinámica<sup>161</sup>.

La evolución experimental mimetiza procesos evolutivos que ocurren en la naturaleza, y por tanto, pienso que el uso de esta herramienta para poder entender la expansión del regulón T box, es fundamental.



### *b) Condiciones de crecimiento*

Una vez que se tenga esta cepa de *B. subtilis*, el siguiente paso sería evolucionarla experimentalmente (i.e., crecerla en un quimiostato por 10,000 generaciones y muestrearla a intervalos regulares para monitorear cambios graduales<sup>163-166</sup>). Una presión selectiva importante del sistema es la regulación fina de las concentraciones de glicil tRNA sintetasa, la cual, es esencial para el organismo y para su óptimo crecimiento. Además, la célula deberá ser crecida en medio mínimo para poder introducir fluctuaciones controladas de nutrientes, como agregar pulsos de glicina. Estos cambios extracelulares ejercen una presión selectiva positiva para aumentar la tasa evolutiva. Se corre aquí, el riesgo de no seleccionar por cambios que permitan una regulación fina, si no viabilidad, sobre todo en las etapas iniciales.

### *c) Control para la pureza de la cepa*

Para evitar posibles contaminaciones, se podrían agregar a la cepa marcadores de resistencia a canamicina y cloramfenicol. El cassette de resistencia a canamicina puede ser utilizado como una presión selectiva en el cultivo, mientras que el segundo marcador puede ser utilizado para identificar constantemente la integridad de nuestro cultivo en inspecciones periódicas.

### *d) Análisis de un proceso adaptativo y continuo en la regulación por T box*

Para poder seguir el cambio constante de la T box de tirosina en el proceso adaptativo que se requiere para regular el operón de *glyQS*, podría introducirse en el operón la proteína verde fluorescente (GFP) en su extremo 3', como se indica en la figura 18.



Figure 18. Construction of the Tyr regulated *glyQS-gfp* operon

Con esta construcción, podría caracterizarse la especificidad y el alcance de la T box conforme ésta evoluciona en el tiempo, y cómo su regulación correlacionará con las alteraciones en su estructura secundaria (lo cual podría verse con métodos de secuenciación, o bien, con análisis *in vitro* de corte de RNA). También es importante determinar si los cambios en la T box evolucionada correlacionan con una mejoría en la adecuación del organismo.

*e) Introducción de fluctuaciones en el medio*

La actividad de fluorescencia en las células, que será producida por la expresión del operon *glyQS-gfp* servirá para determinar y correlacionar el impacto regulatorio de la alterada y cambiante T box en la tasa de crecimiento celular y cómo ésta responde transcripcionalmente. Para poder magnificar los requerimientos de una regulación precisa de la T box, se necesitaría tener concentraciones controladas y fluctuantes de glicina y tirosina, por lo que se pueden introducir periódicamente pulsos de tirosina y glicina. Con esto esperaríamos ver cambios en la expresión del operón mientras la T box evoluciona (Figura 19).



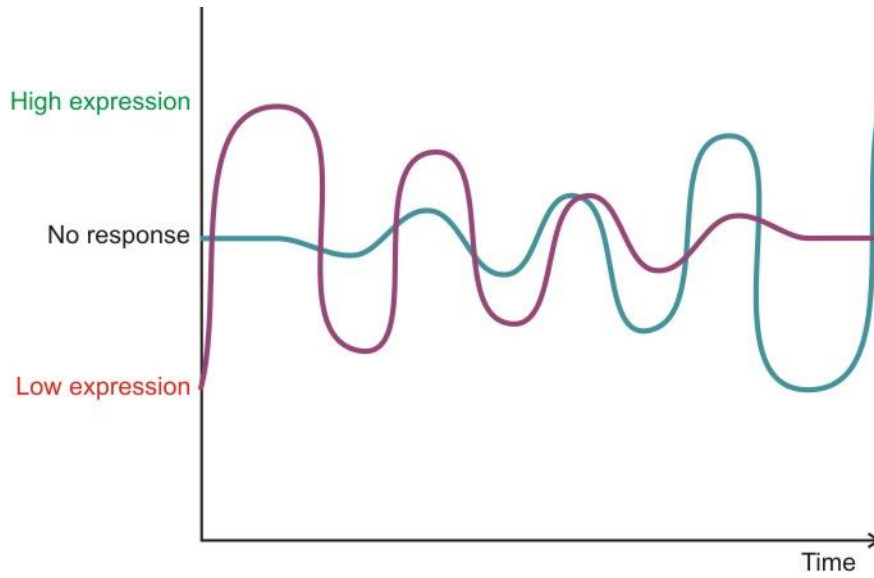


Figura 19. **Comportamiento esperado de la transcripción del operón glyQS-gfp en respuesta a la introducción de pulsos de glicina y tirosina a lo largo del tiempo evolutivo.** Las fluctuaciones en las concentraciones de los aminoácidos en el medio serán inducidas por la introducción de pulsos de glicina y tirosina a intervalos regulares. El plan es registrar cómo la expresión del operon varía a través de las 10,000 generaciones. La línea morada representa la expresión del operón construido con las variaciones de la concentración de tirosina, mientras que la línea azul representa la expresión del operón construido a través del tiempo con las variaciones de la concentración de glicina.

## 2. *leuA* en *Geobacter sulfurreducens*

En colaboración con Katy Juárez

(Departamento de Ingeniería Celular y Biocatálisis. Instituto de Biotecnología, UNAM)

*Geobacter sulfurreducens* es una  $\delta$ -Proteobacteria, la cual tiene dos copias (parálogas) del gen *leuA*. Este gen participa en la biosíntesis de Leucina, y una de estas copias esta regulada por una T box que responde a Leucina (la cual, como se explicó anteriormente, creemos que proviene de una transferencia horizontal de *C. acetobutylicum*).

**Hipótesis:** La T box que se encuentra río arriba de *leuA* es funcional y está regulando a nivel transcripcional.

### Metodología propuesta

- i) Construir una cepa auxótrofa de leucina.
- ii) Para probar que la T box no está actuando a nivel transcripcional, planeamos mapear el inicio de transcripción por RT-PCR usando el set de oligos 1,2 y 1,3 (Figura 20).
- iii) Crecer la cepa auxótrofa con y sin Leucina
- iv) Extraer RNA de ambos cultivos; cDNA
- v) RT-PCR (con una transcriptasa reversa que funcione a altas temperaturas, dado que estamos trabajando con RNA que se estructura).

### Resultados a esperar:

Si la cantidad de productos “largos” de PCR es significativamente menor que la de productos “cortos” de PCR, entonces el gen *leuA* está siendo regulado a **nivel transcripcional**.

De otra forma, asumiremos que está siendo regulado a **nivel traduccional** (Figura 20).

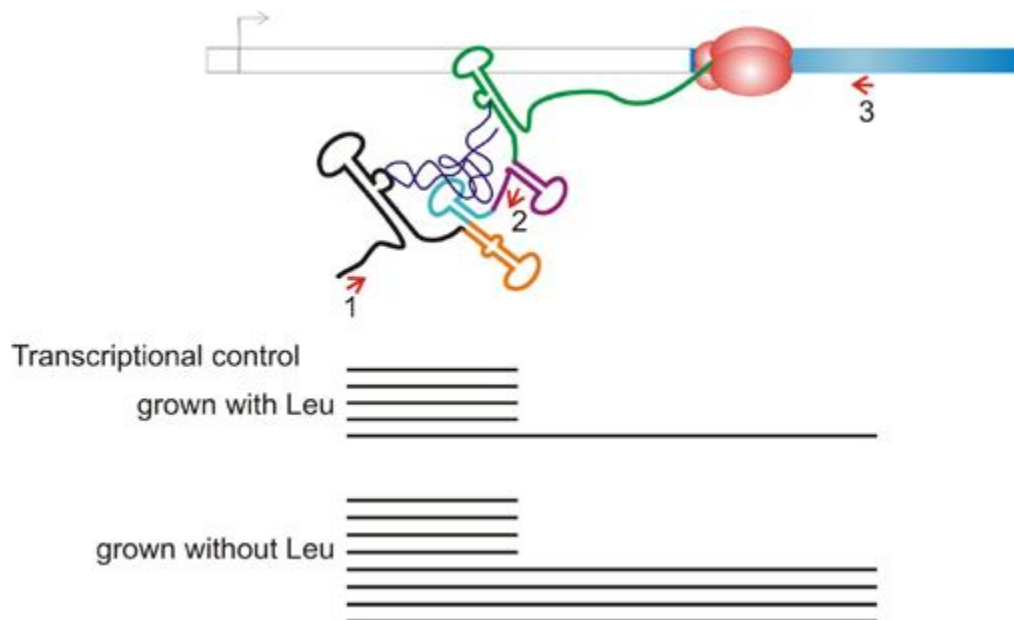


Figura 20. Resultados esperados de una T box transcripcional en *G. sulfurreducens*



### 3. Rellenar los huecos Regulatorios del riboswitch T box en el gen *trpS*

#### Introducción

##### ¿Qué es un hueco regulatorio?

Si tomamos un árbol filogenético y dibujamos una columna de presencia/ausencia de un mecanismo de regulación para cierto tipo de genes, encontraremos que en ciertos clados filogenéticos este mecanismo de regulación está ausente, por lo que en nuestro dibujo quedaría un “hueco” (Figura 21). Este hueco, donde no se presenta el mecanismo regulatorio en cuestión, debería de estar “llenado” por otro mecanismo de regulación distinto, o bien por una expresión génica constitutiva. Asumiendo que no es constitutiva (al menos en los casos de estudio de este proyecto), debe de haber un mecanismo de regulación el cual se desconoce. El propósito de este proyecto es “llenar” esos “huecos regulatorios”.

##### Ejemplos de huecos regulatorios

Existen dos ejemplos muy representativos de “huecos regulatorios” que resultaron en el hallazgo de nuevos riboswitches. El primero es el riboswitch de S box, que lo encontró Tina Henkin porque en los genomas de las *Bacillaceae* las T boxes de metionina son muy escasas, constituyendo un ejemplo de hueco regulatorio. En un principio se pensó que no existían las T boxes de metionina, hasta que nuevos genomas fueron secuenciados (*Clostridias*) donde sí había T boxes de metionina en genes de biosíntesis de metionina (*yxjG* en particular, que es un homólogo a *metE*). Con este antecedente, se supo que las T boxes de metionina sí son viables y que probablemente sí hubiera T boxes de metionina. Las secuencias (y más específicamente el parálogo de *yxjG*) fueron analizadas nuevamente, sin encontrar una T box, pero sí encontrando atenuación. Aquellas regiones intergénicas de genes relacionados a la metionina, que no tenían T box

pero sí atenuación, presentaban un 90% de identidad, mientras que las regiones codificantes sólo un 70%. En otras palabras, la región líder estaba mucho más conservada que los genes, por lo que debía de ser importante. Tina buscó regiones intergénicas parecidas a éstas y encontró cerca de 11 en el genoma de *B. subtilis*. La mitad de estos genes tenían una función desconocida, pero la otra mitad eran genes de biosíntesis de metionina. De esta manera, empezó a buscar patrones hasta que reconoció a la S box, el riboswitch de metionina. (Tina Henkin, comunicación personal)

El otro ejemplo caso de la L box, o el riboswitch de lisina. Los genes de biosíntesis de lisina tampoco presentan T boxes de lisina. Usando la misma metodología y el mismo razonamiento, se encontró simultáneamente por dos grupos distintos, el riboswitch de lisina<sup>52,167</sup>.

### ¿Por qué la T box?

El mecanismo que se ha seleccionado como una primera aproximación para este proyecto es el riboswitch T box. La regulación por T box normalmente se basa en atenuación transcripcional, donde un tRNA descargado interactúa con la región 5' del transcrito, estabilizando una estructura de antiterminación<sup>26</sup>. Como ya sabemos, el regulón de este riboswitch, es amplio, controlando la expresión de genes relacionados a aminoácidos, aaRS, de biosíntesis, de transporte y de proteínas regulatorias. También sabemos que se encuentra ampliamente distribuido en los Firmicutes y es un mecanismo versátil donde con sólo cambiar tres nucleótidos en cierta posición (el Specifier Loop) podemos cambiar la especificidad de la regulación. En otras palabras, aunque es un mecanismo muy concurrido para regular una gran variedad de genes relacionados con el metabolismo de aminoácidos, con conocer su estructura secundaria podemos

predecir a qué aminoácido está respondiendo el gen en cuestión. Otra de las ventajas es que, al ser un mecanismo muy conocido y bien estudiado, podemos predecir su presencia con un alto grado de confiabilidad, y por lo tanto estar seguros donde realmente hay una ausencia (un hueco regulatorio).

### ¿Por qué triptófano?

Dado que la T box es un mecanismo de regulación que puede responder diferencialmente a 18 de los 20 aminoácidos proteínogénicos, es conveniente acotar el proyecto a un solo aminoácido, al menos para una primera aproximación.

Se eligió el triptófano porque para este aminoácido se conoce enormemente su regulación en distintos clados filogenéticos de los Firmicutes. Su regulación empezó a ser estudiada en *B. subtilis*, donde se encontró que la expresión del operón estaba regulada por una proteína llamada TRAP (por sus siglas en inglés: Tryptophan RNA-binding Attenuation Protein). TRAP es una proteína de once subunidades, cada una de las cuales es capaz de unir una molécula de triptófano. Cuando esta proteína se encuentra en unión con once moléculas de triptófano, se encuentra en su estado activo, el cual es capaz de reconocer una secuencia de once trinucleótidos espaciados por dos nucleótidos que se encuentra en la región líder del operón de triptófano. Al unirse con esta región líder promueve la formación de un terminador impidiendo la transcripción del operón *trp* en *B. subtilis*. Si no hay suficiente triptófano en la célula para unirse a TRAP, TRAP está inactiva y por esta razón no puede unirse a la región líder del operón *trp*, la cual, en ausencia de TRAP formará una estructura de antiterminación (que es más estable *per se* y mutuamente excluyente del terminador) permitiendo que la RNA polimerasa continúe con la transcripción de los genes *trp*<sup>124,168</sup>. De esta manera, la bacteria puede reconocer los niveles intracelulares de triptófano y en base a permitir (o no) que se transcriban los genes para sintetizar la molécula de

triptófano. *B. subtilis* tiene, además, un mecanismo para reconocer los niveles intracelulares de tRNA<sup>Trp</sup> descargado. Lo hace mediante el mecanismo de regulación ya descrito: una T box que regula una segunda proteína reguladora, Anti-TRAP. Si los niveles de tRNA<sup>Trp</sup> descargado son altos, éstos interactuarán con la T box de la región líder de Anti-TRAP (codificada por el gen *yczA*) permitiendo su expresión. Anti-TRAP se une a TRAP activa inactivándola. De esta manera, aunque haya mucho triptófano intracelularmente, si los niveles de tRNA<sup>Trp</sup> descargado son altos la biosíntesis del triptófano puede continuar<sup>169,170</sup>.

Al ser *B. subtilis* un organismo modelo, por muchos años se asumió que la regulación en todas las bacterias Gram positivas sería similar a la ya descrita. Cuando uno busca homólogos de TRAP y Anti-TRAP sólo es posible encontrarlos en las *Bacillaceas* muy cercanas filogenéticamente a *B. subtilis*. Homólogos de TRAP pueden encontrarse también en un pequeño clado de *Clostridias*. Esto dio lugar a un hueco regulatorio gigante en cuanto a la regulación de la biosíntesis de triptofano en las bacterias Gram positivas. Este hueco pudo ser rellenado para la mayor parte de las Firmicutes encontrando que estaban mayoritariamente regulando sus genes de biosíntesis de *trp* por TRP-T boxes<sup>127,128</sup>. Sin embargo, tanto en biosíntesis como en tryptophanyl-tRNA sintetasas aún existen huecos regulatorios de los que no se conoce qué mecanismo de regulación ha sido seleccionado.

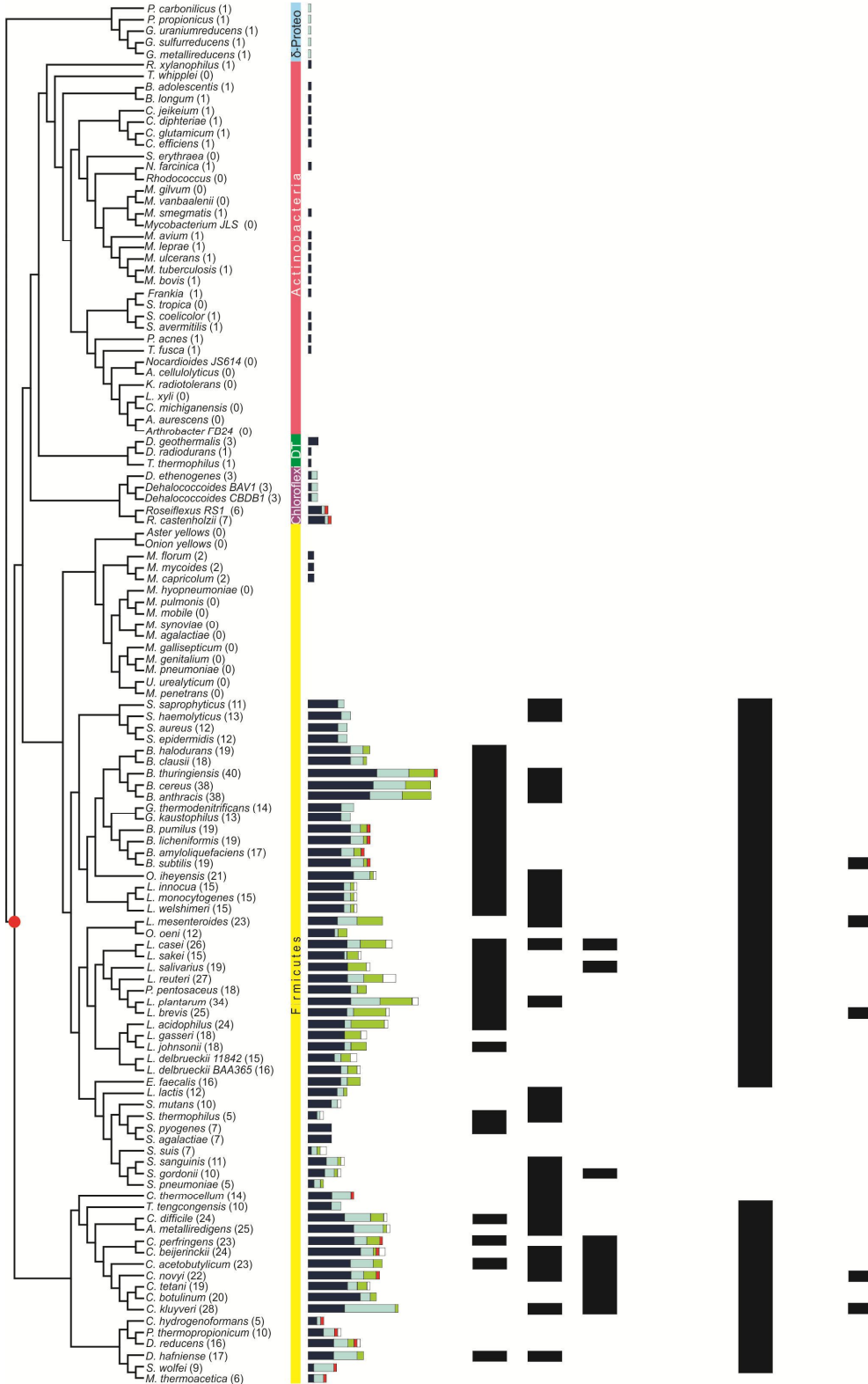
Figura 22 (siguiente página). **Árbol Filogenético de clados que tienen T boxes y huecos regulatorios para genes relacionados a Triptofano y Leucina.** El árbol filogenético fue construido para clados organismos que presentan regulación por T box en al menos un gen de al menos un organismo por clado. Se realizó en base a las distancias filogenéticas de las secuencias alineadas de 31 proteínas concatenadas en 191 especies<sup>120</sup>. Los alineamientos se generaron usando el programa MUSCLE<sup>121</sup>, y la reconstrucción filogenética se hizo mediante el programa PROTDIST del paquete PHYLIP (versión 3.57c; J. Felsenstein, University of Washington, Seattle). Los operones fueron predichos usando la metodología descrita en<sup>113</sup>. Las barras horizontales están dibujadas a escala y representan el número de operones regulados por una T box; éstos están clasificados en: aminoacil-tRNA sintetasas (azul oscuro), genes de biosíntesis de aminoácidos (azul celeste), genes que codifican para proteínas regulatorias (rojo), genes de transporte de aminoácidos (verde) y genes con función desconocida (blanco). Las columnas negras indican presencia de una T box que responde al aminoácido correspondiente río arriba de genes que caen en la categoría de aminoacil-tRNA sintetasas, de biosíntesis o de transporte. Por tanto, los espacios en blanco representan huecos regulatorios de la T box.



653 aaRS  
 250 biosynthetic  
 147 transporters  
 17 regulators  
 32 unknown  
 1099 Total

Tryptophan

Leucine



## Metodología empleada en una primera aproximación

En una primera aproximación, se realizaron ya varios scripts en el lenguaje de programación Perl que hacen lo siguiente:

1. Identifican en qué operón se encuentra el gen *trpS* para cada organismo dentro del clado de los Firmicutes cuya secuencia genómica haya sido caracterizada en su totalidad.
2. Toman la región intergénica para cada operón. Se obtuvo un total de 218 secuencias intergénicas.
3. Utilizando un modelo de covarianza implementado en el programa CMsearch del paquete Infernal, se identifican posibles T boxes en el conjunto de secuencias intergénicas.
4. Se toman aquellas secuencias intergénicas que NO tienen una T box (hueco regulatorio). Se encontraron 87 T boxes significativas mientras que 132 secuencias no presentan este elemento regulatorio.
5. Se utiliza el programa MEME<sup>110</sup> para generar matrices de secuencias sobrerrepresentadas en el conjunto de las secuencias SIN T box.

Al hacer esto, nos dimos cuenta que la primer matriz generada correspondía a la secuencia de T box. Es decir no se estaba generando un hueco regulatorio real. Una posible explicación a este hecho era la presencia parcial del riboswitch T box dentro del gen ubicado río arriba de la región intergénica en estudio. Por esta razón se cambió el paso 2 (tomar la región intergénica de cada operón) por otro programa que tomara 700 nucleótidos río arriba del operón, sin tomar en cuenta si se tratara de una región intergénica o de una región codificante. Al hacer esto se identificaron 103 T boxes (es decir, 16 más que con la metodología anterior).

Para este conjunto de 16 T boxes cuya secuencia es compartida por regiones intergénicas y regiones codificantes, se verificó la función de los genes río arriba de la T box así como en qué organismos está presente. Esto sólo ocurre en tres clados: las *Bacillaceas* (en particular para *B. cereus* y *B. subtilis*), para *Desulfitobacterium hafniense* y para los *Streptococci* (*S. dysgalactiae*, *S. equi*, *S. pyogenes* y *S. suis*). Estos clados no representan algún sesgo en particular (i.e., no todos son patógenos, no todos son de vida libre, etc...) Tampoco se encontró sesgo alguno en el tipo de genes cuya secuencia presenta una T box (ver Tabla 1).

Se realizó este análisis ya que se ha visto que en *Listeria monocytogenes* ciertos genes de virulencia están seguidos por un riboswitch, el cual actúa como terminador transcripcional.

Tabla 1. Función de los genes que sobrelapan una TRP-T box que regula al gen *trpS*.

ORGANISMO	FUNCIÓN DEL GEN
<i>Bacillus cereus</i> AH187	30S ribosomal protein S5
<i>Bacillus cereus</i> ATCC 10987	phosphoesterase PA-phosphatase related
<i>Bacillus subtilis</i> 168	phosphoribosylaminoimidazole synthetase (EC:6.3.3.1)
<i>Desulfitobacterium hafniense</i> DCB-2	hypothetical protein
<i>Streptococcus dysgalactiae</i> subsp. <i>equisimilis</i> GGS_124	acriflavin resistance protein
<i>Streptococcus equi</i> subsp. <i>equi</i> 4047	4-diphosphocytidyl-2-C-methyl-D-erythritol kinase (EC:2.7.1.148)
<i>Streptococcus equi</i> subsp. <i>zooepidemicus</i> H70	hypothetical protein
<i>Streptococcus pyogenes</i> <i>Manfredo</i> (serotype M5)	response regulator receiver protein
<i>Streptococcus pyogenes</i> MGAS10270 (serotype M2)	chromosomal replication initiator protein DnaA
<i>Streptococcus pyogenes</i> MGAS10394 (serotype M6)	putative deoxyguanosinetriphosphate triphosphohydrolase
<i>Streptococcus pyogenes</i> MGAS10750 (serotype M4)	chromosomal replication initiator protein DnaA
<i>Streptococcus pyogenes</i> MGAS2096 (serotype M12)	hypothetical protein
<i>Streptococcus pyogenes</i> MGAS5005 (serotype M1)	superoxide dismutase (EC:1.15.1.1)
<i>Streptococcus pyogenes</i> MGAS6180 (serotype M28)	hypothetical protein
<i>Streptococcus pyogenes</i> MGAS9429 (serotype M12)	seryl-tRNA synthetase (EC:6.1.1.11)
<i>Streptococcus suis</i> 05ZYH33	DNA methylase N-4/N-6 domain-containing protein
<i>Streptococcus suis</i> 98HAH33	DNA methylase N-4/N-6 domain-containing protein
<i>Streptococcus suis</i> BM407	periplasmic sensor hybrid histidine kinase
<i>Streptococcus suis</i> SC84	rod shape-determining protein RodA

Dado lo anterior, se utilizó una nueva aproximación metodológica:

1. Identificar en qué operón se encuentra el gen *trpS* para cada organismo.
2. Tomar 700 nucleótidos río arriba de cada operón. Se obtuvo un total de 218 secuencias.
3. Se utiliza el programa CMsearch del paquete Infernal<sup>105</sup> para predecir T boxes.
4. Se toman aquellas secuencias que NO tienen una T box (hueco regulatorio). Se encontraron 103 T boxes significativas.
5. Se utiliza el programa MEME<sup>110</sup> para generar matrices de secuencias sobrerrepresentadas en el conjunto de las secuencias SIN T box.
6. Se utiliza el programa MAST<sup>112</sup> para buscar estas secuencias sobrerrepresentadas en cada genoma de los Firmicutes. Se utiliza *Geobacter sulfurreducens* y *Escherichia coli* K12 como controles negativos de manera exitosa (i.e., no se encuentra ninguna secuencia en todo el genoma que presente un *hit* significativo con las matrices generadas en el paso 5).

La estadística de los resultados del MAST se puede observar en las Tablas 2 y 3 en la sección de Resultados parciales.

## Resultados parciales

### MEME

El MEME predice 5 motivos, que se describen a continuación:

---

Motif 1 Description	
bits	3.0
	2.7
	2.4
	2.1
Relative	1.8 *
Entropy	1.5 **
(10.4 bits)	1.2 * ***
	0.9 *****
	0.6 ***** *
	0.3 *****
	0.0 -----
Multilevel	TTCTCCTATATG
consensus	CCGC TAA
sequence	G

Este motivo parece estar siempre asociado a una región estructural

---

Motif 2 Description	
bits	3.0
	2.7
	2.4
	2.1
Relative	1.8 *
Entropy	1.5 *
(10.3 bits)	1.2 * *
	0.9 *** *****
	0.6 *****
	0.3 *****
	0.0 -----
Multilevel	AAAGTAGGGAT
consensus	G G AACG
sequence	C

Parece ser el Shine Dalgarno.

---

Motif 3 Description	
bits	3.0
	2.7
	2.4
	2.1
Relative	1.8 *
Entropy	1.5 *
(8.8 bits)	1.2 * *
	0.9 ** * *
	0.6 ** *****
	0.3 *****
	0.0 -----
Multilevel	CTGGTATAATG
consensus	A T CCGCT
sequence	G

Este motivo, sorprendentemente parece ser una reliquia de secuencia T box. El consenso de la T box es AGGGTGGNACCGCG. Al correr un MAST con estas matrices y una base de datos que contiene exclusivamente secuencias con T boxes, sólo se recuperan 12 (de 680) secuencias, las cuales presentan este motivo seguido por el motivo 1. Estos operones son *valS-folC*, *trpS* e *ileS*.

El motivo 4 y 5 parecen ser estadísticamente poco significativos, ya que la presencia de es clado-específicos de algunas *Bacillaceas* y de *Streptococci*.

## MAST

Una vez que un motivo conservado es representado mediante una matriz de probabilidad, el programa MAST permite identificar aquellas secuencias en una base de datos, cuya similitud con el motivo conservado sea estadísticamente significativo. Resulta sorprendente que al construir matrices con regiones intergénicas de *trpS* que carecen de una T box, MAST recupere al propio gen *trpS* y a genes relacionados con triptófano. Por ejemplo, genes de la familia isochorismatasa, genes de biosíntesis de triptófano, de folato y más sorprendentemente a Anti-TRAP (ver Tablas 2 y 3).

El isocorismato puede, mediante la enzima aquí encontrada, convertirse en corismato para posteriormente transformarse en triptófano. Este gen, con esta regulación sólo está presente en clado *Bacillus anthracis/cereus/thuringiensis*.

Los genes de biosíntesis de triptófano presentan esta regulación en *B. cereus*, *B. thuringiensis*, *Clostridium kluyveri*, el clado de las *Listeria*, *Oceanobacillus iheyensis*, *Streptococcus gordonii*, *S. pneumoniae*, *S. sanguinis*, y en clado de las *Thermoanaerobacter*. Todos estos operones tienen simultáneamente una regulación por T box.

Adicionalmente una copia paróloga de *trpB* presenta esta regulación en *Carboxydotherrnus hydrogenoformans* y *Natranaerobius thermophilus*. La región intergénica de *trpB* de *N. thermophilus* presenta, además, una TRP-T box.

Los genes de biosíntesis de folato están regulados por este elemento en el clado *Bacillus anthracis/cereus/thuringiensis*. Dos de estos genes (*pabA* y *pabB*) son parálogos de *trpE* y *trpG*, los cuales han divergido en función. La explicación de la presencia de este elemento regulatorio en los genes de folato puede ser que hayan sido duplicados, en un inicio, junto con su región regulatoria y que este organismo encontró favorable esa regulación. Sin embargo, esta no es una explicación parsimoniosa ya que todos los demás organismos que tienen ambas copias parálogas carecen de este elemento regulatorio en la región río arriba del operon de folato. Otra posibilidad es que a estos organismos en particular les resulte conveniente co-expresar estas dos vías biosintéticas, o que le resulte conveniente co-expresar a *pabA* y *pabB* porque puedan ser inespecíficas en el reconocimiento del sustrato y de esa manera incrementar el flujo hacia la biosíntesis de triptófano.

## Conclusiones puntuales

En el clado *Bacillus anthracis/cereus/thuringiensis* existen dos genes parálogos de *trpS* y ambos están regulados por una T box pero sólo una copia de estas presenta estas matrices. Esto nos hace pensar que este nuevo mecanismo de regulación en conjunto con la T box proporciona a la bacteria una regulación más estricta (o quizá más laxa) de la copia paróloga, permitiéndole entonces expresar ambos genes de manera diferencial, y expresando sólo ambos en condiciones extremas de poco tRNA<sup>Trp</sup> cargado. Un caso similar se ha descrito para los genes parálogos de *tyrS* en *B.subtilis* siendo uno regulado por una T box y el otro, de manera más estricta, por tres T boxes en tándem.

Sin duda estamos ante un nuevo mecanismo regulatorio que puede ser específico para triptófano, que está presente en algunos Firmicutes y que puede coexistir con la T box.

Tabla 2. Resultados de los MASTs corridos en todos los genomas de Firmicutes. Se grafica el número de veces que las matrices generadas con MEME (hechas sólo con regiones intergénicas de *trpS*) encuentran dicha función **con un valor de probabilidad (e value) menor igual a 0.1**

Observado	Función del gen
72	hypothetical protein
20	tryptophanyl-tRNA synthetase
15	isochorismatase family protein
10	catabolite control protein A
8	threonine synthase (EC:4.2.3.1)
7	dipeptide-binding protein
6	valyl-tRNA synthetase (EC:6.1.1.9)
6	phenylalanyl-tRNA synthetase subunit alpha (EC:6.1.1.20)
5	negative regulator of genetic competence ClpC/MecB
5	isoleucyl-tRNA synthetase (EC:6.1.1.5)
5	asparaginyl-tRNA synthetase (EC:6.1.1.22)
4	ATP-dependent Clp protease, ATP-binding subunit ClpE
3	iron chelate uptake ABC transporter, FeCT family, solute-bindingprotein
3	asparagine synthetase AsnA (EC:6.3.1.1)
2	transcriptional regulator
2	threonine synthase



2	surface lipoprotein
2	seryl-tRNA synthetase (EC:6.1.1.11)
2	PTS system, cellobiose-specific IIC component (EC:2.7.1.69)
2	protein of unknown function UPF0236
2	prolyl-tRNA synthetase (EC:6.1.1.15)
2	oxidoreductase
2	N-acetylmuramic acid 6-phosphate etherase (EC:4.2.-.-)
2	leucyl-tRNA synthetase (EC:6.1.1.4)
2	iron chelate ABC transporter solute-binding protein
2	glutaminase (EC:3.5.1.2)
2	DNA polymerase III gamma and tau subunits
2	cytidylate kinase
2	cobalt-zinc-cadmium resistance protein CzcD
2	cobalt-zinc-cadmium resistance protein
2	aspartyl-tRNA synthetase (EC:6.1.1.12)
2	alanyl-tRNA synthetase (EC:6.1.1.7)
2	anthranilate synthase component I (EC:4.1.3.27)
1	zwitermicin A resistance protein ZmaR
1	YdbJ
1	YczA
1	valyl-tRNA synthetase
1	UDP-N-acetylmuramoylalanyl-D-glutamyl-2,6-diaminopimelate ligase
1	tyrosyl-tRNA synthetase (EC:6.1.1.1)
1	two component transcriptional regulator
1	trypsin-like serine protease
1	tRNA-Val
1	tRNA (uracil-5-)-methyltransferase related enzyme
1	transposase
1	transcriptional antiterminator, putative
1	thiamin pyrophosphokinase
1	surface lipoprotein DppA
1	superfamily I DNA/RNA helicase-like protein
1	sugar transport family protein
1	S-ribosylhomocysteinase
1	spore germination protein
1	sodium/hydrogen exchanger family protein
1	RNA polymerase sigma-D factor
1	ribose-5-phosphate isomerase A (EC:5.3.1.6)
1	replication terminator protein
1	pyrroline-5-carboxylate reductase
1	pyrimidine operon attenuation protein/uracilphosphoribosyltransferase
1	putative tryptophan transport protein
1	putative threonine synthase

---

1	putative surface lipoprotein
1	putative sugar uptake protein
1	putative permease
1	putative oxidoreductase
1	putative lipoprotein
1	putative GTP-binding elongation factor
1	putative cytochrome aa3 controlling protein
1	putative CoA binding protein
1	putative chloride channel protein
1	putative cation efflux system protein
1	putative ATP-dependent Clp protease ATP-binding subunit
1	putative ABC transporter, permease protein
1	PTS system, diacetylchitobiose-specific IIC component
1	protein-export membrane protein SecD
1	prolyl-tRNA synthetase
1	proline dipeptidase
1	ProB2 (EC:2.7.2.11)
1	phosphofructokinase
1	phosphatidylglycerophosphate synthase (EC:2.7.8.5)
1	phenylacetic acid degradation protein PaaD
1	peptide deformylase (EC:3.5.1.88)
1	penicillin tolerance protein
1	PduQ protein, putative
1	oxidoreductase, Gfo/Idh/MocA family protein
1	NUDIX family hydrolase
1	NADH dehydrogenase subunit A (EC:1.6.5.3)
1	N-acetylmuramic acid 6-phosphate etherase
1	multidrug transport protein
1	molybdenum ABC transporter ATP-binding protein
1	methyltransferase type 11
1	methyltransferase
1	methyl-accepting chemotaxis sensory transducer
1	methionine sulfoxide reductase B (EC:1.8.4.12)
1	methionine sulfoxide reductase B (EC:1.8.4.11)
1	methionine sulfoxide reductase A
1	M3B family peptidase (EC:3.4.24.-)
1	M24 family metallopeptidase
1	LmbE family protein
1	ketol-acid reductoisomerase (EC:1.1.1.86)
1	inhibitor of TRAP, regulated by T box (trp) sequence RtpA
1	hypothetical membrane spanning protein
1	hydroxylamine reductase
1	hydrolase

---

1	histone acetyltransferase HPA2 and related acetyltransferases
1	histidinol-phosphate aminotransferase
1	histidine kinase
1	HD superfamily phosphohydrolase
1	GTPase
1	glutamyl-tRNA reductase (EC:1.2.1.70)
1	glucose/mannose:H <sup>+</sup> symporter
1	glucokinase regulatory protein
1	Gfo/Idh/MocA family oxidoreductase
1	gamma-glutamyl kinase (EC:2.7.2.11)
1	ferrichrome ABC transporter permease
1	elongation factor P
1	DNA polymerase III, tau subunit (EC:2.7.7.7)
1	DNA polymerase III gamma subunit
1	DNA polymerase III, gamma and tau subunits (EC:2.7.7.7)
1	DNA polymerase III gamma and tau subunits (EC:2.7.7.7)
1	DNA mismatch repair (recognition)
1	DNA mismatch repair protein MutS
1	DNA helicase
1	dipeptide-binding extracellular protein
1	diadenosine tetraphosphate (Ap4A) hydrolase and other HIT family hydrolases
1	DeoR-type transcriptional regulator, putative
1	DegV family protein
1	collagen-binding surface protein, putative
1	cell envelope-related transcriptional attenuator
1	cell division protein FtsW
1	cell division membrane protein
1	CBS domain-containing protein
1	catabolite control protein CcpA, RegM
1	branched-chain amino acid transporter
1	Bcr/CfIA subfamily drug resistance transporter
1	ATP-dependent protease ATP-binding subunit ClpX
1	ATP dependent protease
1	ATP-dependent metalloprotease FtsH (EC:3.6.4.6)
1	ATP dependent Clp protease, ATP-binding subunit, ClpE
1	aspartyl-tRNA synthetase
1	APC family amino acid-polyamine-organocation transporter
1	amino acid/peptide transporter
1	alpha/beta hydrolase superfamily protein
1	aldo/keto reductase family oxidoreductase
1	aldo/keto reductase
1	acetyltransferase
1	acetamidase/formamidase (EC:3.5.1.49)

1	ABC-type uncharacterized transport system, periplasmic component, putative
1	ABC-type molybdenum transport system, ATPase component, putative (EC:3.6.3.34)
1	ABC transporter, ATP-binding protein
1	ABC transporter ATP-binding protein
1	4-oxalocrotonate tautomerase
1	3-carboxymuconate cyclase, putative
1	2-nitropropane dioxygenase, NPD

Tabla 3. Resultados de los MASTs corridos en todos los genomas de Firmicutes. Se grafica el número de veces que las matrices generadas con MEME (hechas sólo con regiones intergénicas de *trpS*) encuentran dicha función **con un e value mayor a 0.1**

Observado	Función del gen
867	hypothetical protein
54	isoleucyl-tRNA synthetase (EC:6.1.1.5)
41	tryptophanyl-tRNA synthetase (EC:6.1.1.2)
38	valyl-tRNA synthetase (EC:6.1.1.9)
38	pseudogene
33	alanyl-tRNA synthetase (EC:6.1.1.7)
24	leucyl-tRNA synthetase (EC:6.1.1.4)
20	anthranilate synthase component I (EC:4.1.3.27)
19	membrane protein
16	transcriptional regulator
16	NAD(P)H-dependent glycerol-3-phosphate dehydrogenase (EC:1.1.1.94)
14	asparaginyl-tRNA synthetase (EC:6.1.1.22)
14	asparagine synthetase AsnA (EC:6.3.1.1)
13	excinuclease ABC subunit A
13	elongation factor G
13	ATP-dependent RNA helicase
12	xanthine/uracil permease family protein
12	tyrosyl-tRNA synthetase (EC:6.1.1.1)
12	50S ribosomal protein L28
12	GTP cyclohydrolase I (EC:3.5.4.16)
11	valyl-tRNA synthetase
11	tRNA-Arg
11	isoleucyl-tRNA synthetase
11	histidyl-tRNA synthetase (EC:6.1.1.21)
11	cell-division initiation protein DivIVA
11	aspartyl-tRNA synthetase (EC:6.1.1.12)
11	acetyltransferase
11	4-hydroxy-3-methylbut-2-enyl diphosphate reductase (EC:1.17.1.2)
10	putative permease

---

10	peptide chain release factor 2
10	DNA mismatch repair protein MutS
10	cobalamin synthesis protein/P47K family protein
10	alanyl-tRNA synthetase
...	...
5	chorismate synthase (EC:4.2.3.5)
2	tryptophan synthase subunit beta (EC:4.2.1.20)
2	bifunctional 3-deoxy-7-phosphoheptulonate synthase/chorismate mutase

---

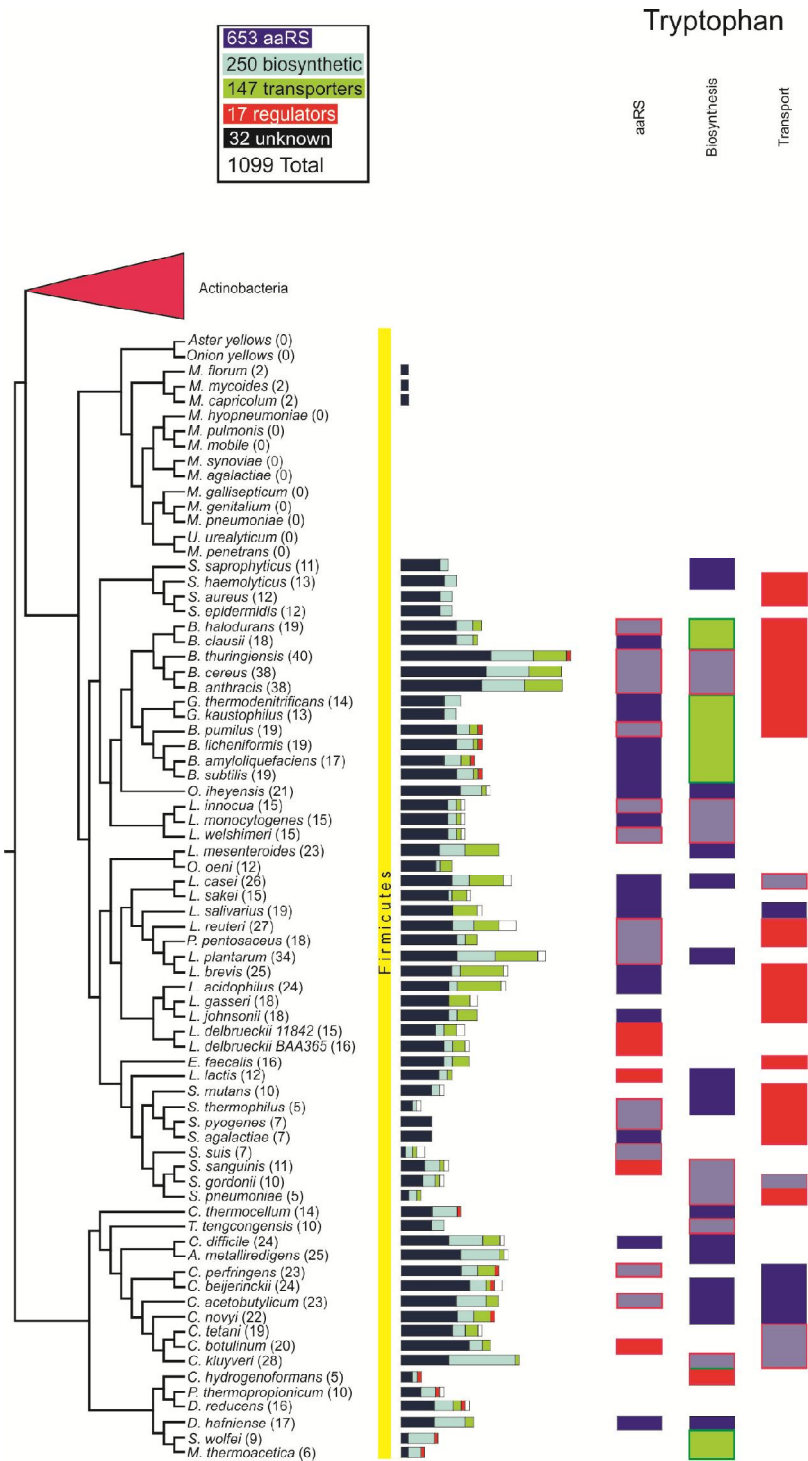


Figura 23. **Árbol Filogenético de clados que tienen T boxes de triptófano, huecos regulatorios y hits del MAST.** El árbol filogenético fue construido para clados organismos que presentan regulación por T box en al menos un gen de al menos un organismo por clado. Se realizó con la misma metodología descrita en el pie de la Figura 22. Las columnas azules indican presencia de una T box que responde al aminoácido correspondiente río arriba de genes que caen en la categoría de aminoacil-tRNA sintetasas, de biosíntesis o de transporte. Por tanto, los espacios en blanco

representan huecos regulatorios de la T box. Las columnas en rojo son aquellos hits de alguna matriz de MEME encontrada por MAST. Donde la columna azul sobrelapa con la roja, se dibuja morada con borde rojo.

Podemos rellenar muchos de estos huecos de forma bioinformática, pero se requiere una caracterización experimental. Aún así, podemos observar de la Figura 23 que aún quedan huecos regulatorios sobretodo para los *Streptococcus* y los *Staphylococcus*.

#### 4. Web server con el regulon T box

Los programas descritos en la Metodología, son lo último para identificar modelos estructurales de RNA, y sobre todo, la T box. Sin embargo, no permiten una detección optimizada del Specifier Codon para cada T box, pues solamente califican el riboswitch de manera global tomando en cuenta los elementos más conservados de éste. Se realizará un programa *ad hoc* para utilizar la información generada por estos programas y poder así identificar el Specifier Loop, sus secuencias conservadas y por ende el Specifier Codon. La idea es que sea un proceso automático que pueda actualizarse constantemente con la llegada de nuevos genomas. El pipeline de este programa se muestra a continuación (Figura 24):

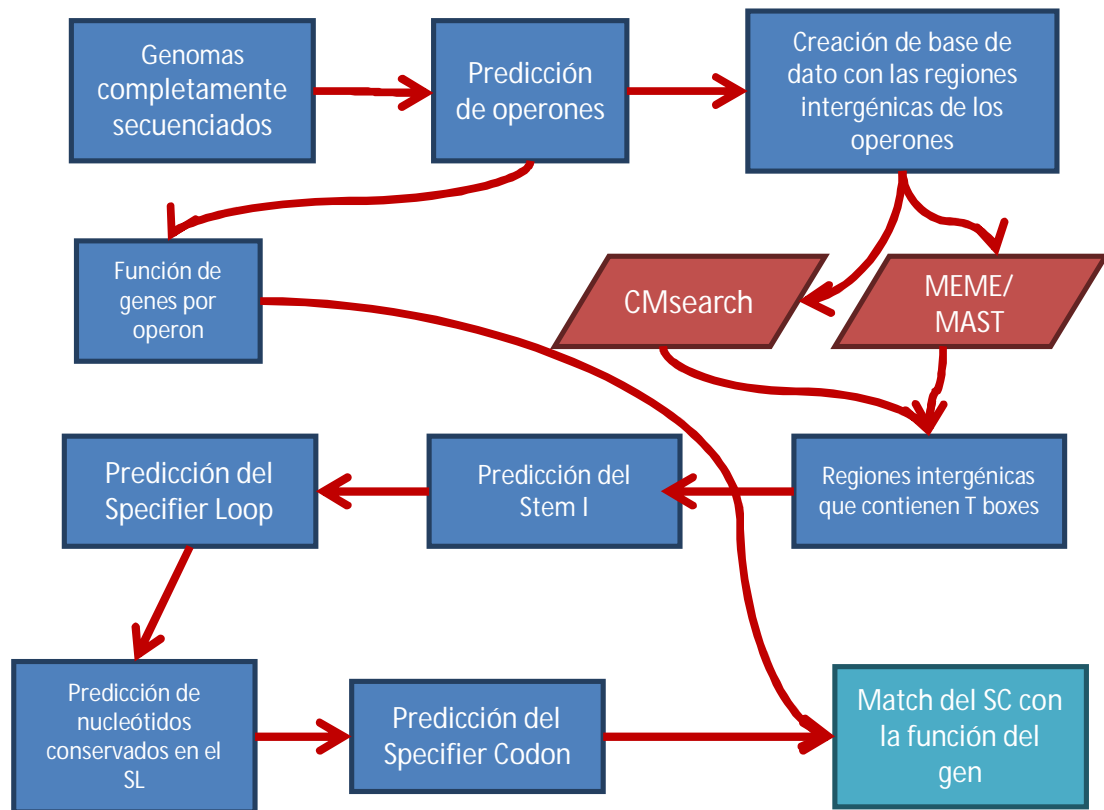


Figura 24. Metodología (pipeline) propuesta para generar un web server que identifique una T box con su secuencia de especificidad, i.e. qué tRNA reconocerá.



Este programa generará automáticamente mucha información que deberá ser representada de un modo amigable y volverla disponible para la comunidad científica. La idea, muy sencilla, es una aplicación web que permita al usuario seleccionar un subconjunto del regulon T box para poder trabajar con él. Esto podría ser, todo el regulon en un organismo; todo el regulon de la T box de un aminoácido en particular; todos los genes ortólogos de cierto grupo que estén regulados por la T box, etc.... Nos gustaría ir mas allá y pedirle, por ejemplo, todos los operones que tengan genes que estén involucrados en dos aminoácidos distintos y saber a cuál de estos está relacionada la T box que regula, esto nos podría ayudar a detectar todos los imbalances que estén ocurriendo en los Firmicutes via T box.

## 5. Imbalance: *hisS-aspS* regulation

### Introducción

Como se ha descrito ya en la sección Un desbalance regulatorio para algunos operones del regulón T box, cierto desbalance, o imbalance, existe cuando dos genes que participan en vías de aminoácidos distintos, están codificados en un mismo operon, respondiendo a un solo aminoácido. Una opción para abordar este posible problema es pensar los aminoácidos como nucleótidos. Sabemos que una cantidad similar de Guaninas y Citosinas tienen que sintetizarse para el DNA genómico, así como una cantidad similar de Adeninas y Timinas deberán sintetizarse. Desconocemos si esto pueda extrapolarse a aminoácidos, donde para el operon *hisS-aspS* ciertas cantidades de histidina debieran estar cargadas en su tRNA de forma similar que las cantidades de aspartato. Asumiendo que lo anterior es falso, y no habiendo encontrado relación metabólica alguna entre histidina y aspartato, surge la pregunta de ¿cómo puede la célula contender con este imbalance? ¿Por qué las cantidades de dichos aminoácidos responderían a la deficiencia de uno de ellos?

Se analizaron todas las estructuras de las T boxes que controlan esta estructura operónica, y se vio que en una gran mayoría el único codon posible a ser expuesto para interactuar con el tRNA fue GAC, el de aspartato. Sin embargo, en algunos organismos, cabía la posibilidad de que se doblase de una forma alterna donde ambos codones pudiesen ser expuestos (ver Un desbalance regulatorio para algunos operones del regulón T box). El caso particular que creo de más interés es la T box de este operon en *B. halodurans*. Como primer detalle, no tiene el codon para aspartato; sino para asparagina. Un segundo detalle es que parece tener dos secuencias de T box en tándem y parece ser que ambas pueden tomar la estructura de un antiterminador. Sin embargo, el detalle más importante

es que es posible que el Stem I pueda tomar una de dos estructuras mutuamente excluyentes, donde en cada caso expondría un codon distinto a ser reconocido por el tRNA (Figura 24). Cabe recordar, como se vio en la sección Regulación por T box en genes de biosíntesis de aminoácidos, que el tRNA de Histidina es un tanto distinto, siendo un nucleótido más corto, lo cuál afecta su brazo aceptor, pues en lugar de tener 4 nucleótidos para interactuar con el antiterminator bulge, sólo tiene 3, siendo ésta la posible razón por la cuál existan pocas T boxes de histidina en comparación a otros aminoácidos. Pero esta misma razón, puede ser la responsable de que esta secuencia posea dos antiterminadores distintos, uno seleccionado para reconocer el brazo aceptor del tRNA de asparagina, mientras que el otro sería necesario para reconocer el tRNA de histidina. Las perspectivas de estos hallazgos son probar que todas estas hipótesis son ciertas.

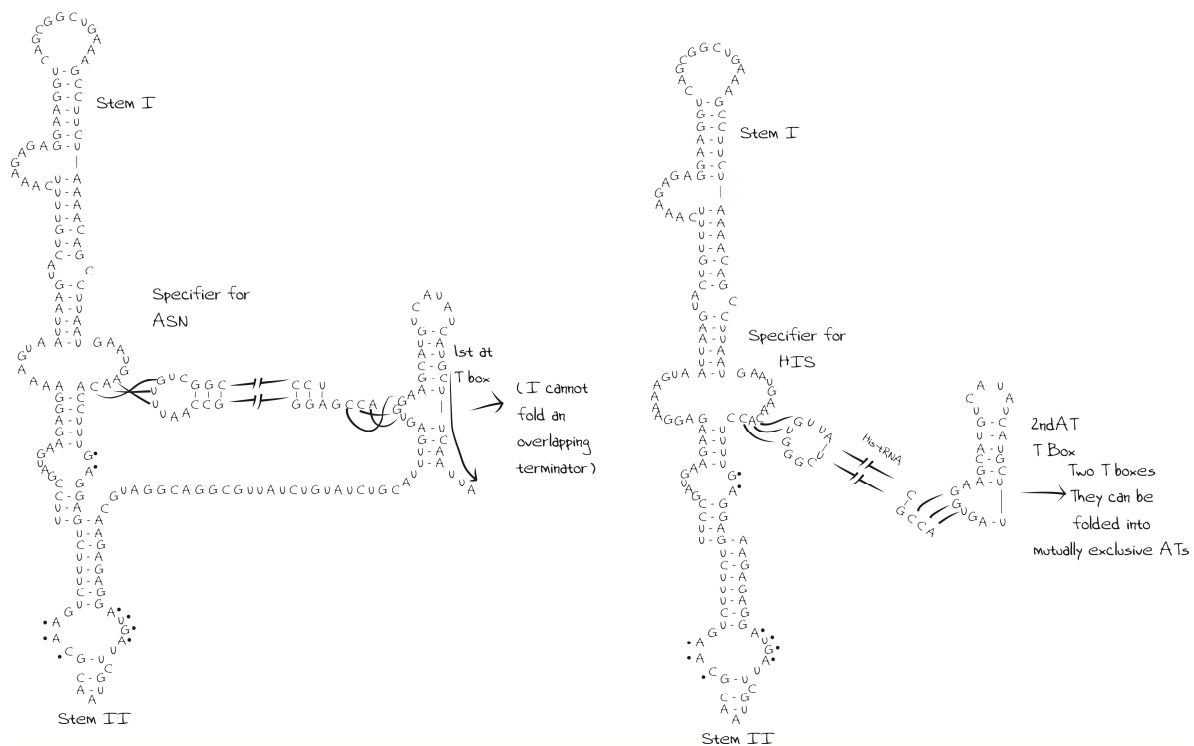


Figura 25. Dos posibles estructuras mutuamente excluyentes para el Stem I de la T box del operon *hisS-aspS* en *B. halodurans*. El de la derecha sería la conformación para reconocer a tRNA<sup>Asn</sup>, y el de la izquierda a tRNA<sup>His</sup>

## Metodología propuesta

Lo primero a realizar sería amplificar esta región regulatoria y secuenciarla para ver que, en efecto, se esté lidiando con esta T box tan peculiar.

Asumiendo que todo lo anterior es cierto, lo primero que debe de hacerse cuando uno va a medir la respuesta regulatoria de una T box, es construir una cepa que sea auxótrofa al aminoácido a detectar sus niveles, ya que esto magnifica el efecto regulatorio de la T box y es más fácil de medir y de cuantificar.

En cuanto a construir una cepa auxótrofa de histidina, es relativamente sencillo. Sólo existe una vía biosintética de histidina. Se debe de tener cuidado porque esta vía comparte pasos con la vía de biosíntesis de las purinas, por lo que el gen a mutar para construir una cepa auxótrofa de histidina es *hisB*, de este modo no se afecta la síntesis de las purinas. Para el caso de aspartato o asparagina es un poco más complicado. *B. halodurans* tiene 5 maneras distintas de sintetizar aspartato. Sin embargo, si el aminoácido que se está detectando en realidad es asparagina, y no aspartato, la elección sería construir una cepa auxótrofa a asparagina, lo cual se mucho más sencillo que a aspartato, dado que en *B. halodurans*, la asparagina se sintetiza a partir de aspartato por la enzima asparagine synthase (glutamine-hydrolysing), de modo que habría que mutagenizar un solo gen. Después creo que podría medirse los niveles de expresión vía RT-PCR para monitorear a qué tRNA está respondiendo.

## 6. Regulación de operones de Triptofano

Como se vio en la Introducción y en particular en la sección de: Atenuación transcripcional: regulando la expresión génica en base a elementos de RNA en *cis*, fue históricamente muy importante el descubrimiento de la regulación del operon *trp* en *E. coli*, la forma en que se descubrió, el momento en el que se hizo y los paradigmas que estableció en el ámbito de la genética bacteriana. Años más tarde, Yanofsky descubrió también la forma en que *B. subtilis* regulaba su biosíntesis de triptófano. La regulación de la expresión de los genes de la biosíntesis de L-triptofano (L-*trp*) se ha estudiado ampliamente en muchos organismos y por lo tanto constituye un excelente sistema para la comparación de los mecanismos de regulación. Los modelos de regulación mejor comprendidos para la vía biosintética de L-*trp* se han propuesto en *Escherichia coli* y *Bacillus subtilis* que usan los mismos siete pasos enzimáticos<sup>19,171</sup>. Dado que organismos filogenéticamente distantes, usan la misma vía para sintetizar L-*trp*, probablemente esta vía solamente ha evolucionado una vez, a partir de un ancestro común<sup>134,172</sup>.

A pesar de la conservación de enzimas y reacciones, el contexto genómico de los genes involucrados en la biosíntesis de triptofano de *E. coli* difiere apreciablemente de los de *B. subtilis*<sup>134</sup>. Además de las diferencias de cómo están organizados los genes, estos organismos también emplean distintas estrategias para cuantificar y responder a las concentraciones intracelulares de L-*trp* y al tRNA cargado o no-cargado como señales metabólicas<sup>171,173,174</sup>.

**Diferencias en la organización de los operones de triptofano de *E. coli* y *B. subtilis* y las diferentes estrategias que usan para su regulación.**

En *E. coli*, son cinco los genes que codifican los siete pasos enzimáticos que son responsables de la biosíntesis de triptofano desde el corismato, el precursor común de los tres aminoácidos aromáticos<sup>174</sup>. Estos cinco genes están organizados en el operón *trp*. Dos de estos genes, *trpG-D* y *trpC-F*, consisten en genes fusionados con actividades bifuncionales (Figura 26)<sup>134</sup>. Los cinco genes están precedidos por una región regulatoria bastante compleja, que puede cuantificar los niveles intracelulares tanto de L-trp como de tRNA<sup>Trp</sup>, ya sea cargado o no<sup>19,174</sup>. Este operón, cuenta con un sólo promotor estrictamente regulado por el represor TrpR en su forma activada por L-trp<sup>174-177</sup>. Una vez que la transcripción se encuentra más allá de la región líder, los genes estructurales del operón se regulan por atenuación<sup>174</sup>. Dependiendo de la disponibilidad de los tRNA<sup>Trp</sup> cargados, la transcripción puede terminarse en la región líder o bien transcribir todo el operón<sup>174</sup>.

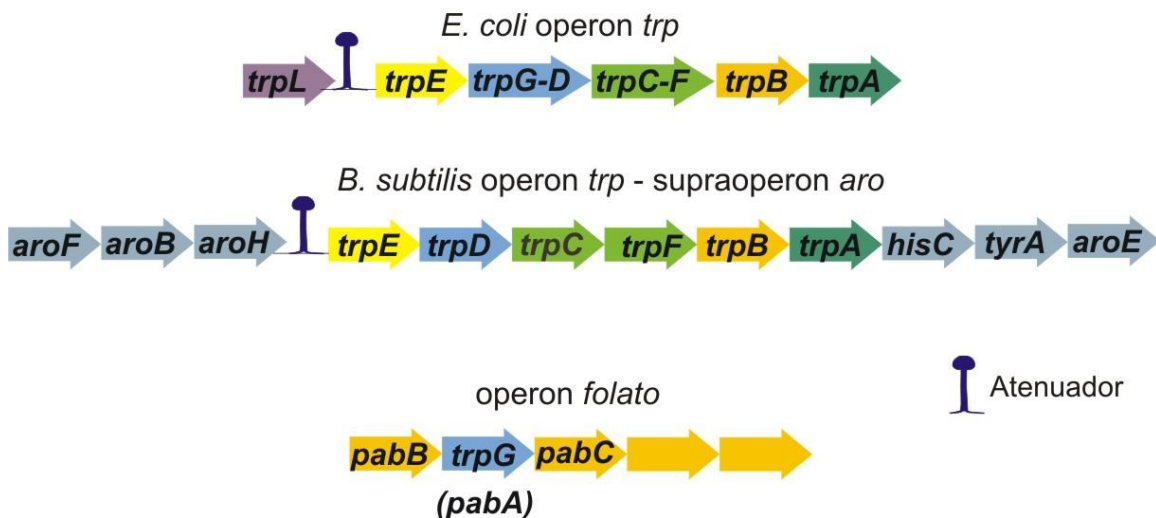


Figura 26. Organización de los operones de triptofano de *E. coli* y *B. subtilis*. Figura modificada de <sup>19</sup>

El operón de biosíntesis de triptofano de *E. coli* es una sola unidad transcripcional, mientras que en *B. subtilis* es un segmento de un supraoperón. En *B. subtilis* uno de los genes, *trpG(pabA)*, está en una unidad transcripcional diferente: el operón folato. En la región regulatoria de *E. coli* se encuentra un promotor-operador y un atenuador. Terminadores en tándem están localizados al final del operón.

En *B. subtilis*, la transcripción del operón de triptofano puede iniciar desde dos distintos promotores, el primero de ellos, inmediatamente arriba de *aroF* transcribe todo el supraoperón *aroFBH-trpEDCFBA-hisC-tyrA-aroE* y el segundo de ellos, colocado río arriba del gen *trpE*, transcribe el resto del supraoperón. No se conoce como funciona la regulación del inicio de la transcripción en el promotor río arriba de *aroF*, es decir de todo el supraoperón, pero la transcripción río arriba de *trpE* está regulado por atenuación. Existe un promotor interno traslapado con el gen *trpA* que funciona para transcribir los últimos tres genes del supraoperón, los cuales participan en la biosíntesis de los aminoácidos aromáticos. No se ha estudiado el término de la transcripción de este supraoperón<sup>19</sup>.

El contexto genómico del operón de biosíntesis de triptofano de *B. subtilis* está organizado de manera diferente (Figura 26), pues consiste en seis genes localizados dentro de un supraoperón aromático (*aro*) de doce genes, de los cuales tres, *aroFBH*, están localizados río arriba, mientras que *hisC-tyrA-aroE* río abajo de *trpEDCFBA*. Estos seis genes localizados en los extremos del operón *trp*, están involucrados en la biosíntesis de corismato y otros aminoácidos aromáticos<sup>142,173,178</sup>. El séptimo gen *trp* que se requiere para la biosíntesis de triptofano es *trpG(pabA)* y se encuentra en el operón de folato<sup>179</sup>. Este polipéptido TrpG-PabA es una glutamina amidotransferasa y cataliza la primer reacción en la vía de *trp*, y una reacción relacionada en la vía del folato<sup>179</sup>.

En *B. subtilis*, para transcribir la región del operón *trp* del supraoperón *aro* se utilizan dos promotores, uno antecediendo al gen *aroF* y otro precediendo al gen *trpE*. La RNA polimerasa (RNAPol) puede iniciar en cualquiera de estos dos promotores y es entonces sujeta a un evento regulatorio en la región líder del operón *trp* que origina el término prematuro de la transcripción, o bien, la transcripción completa del RNA mensajero (mRNA). Este evento regulatorio está basado en la disponibilidad tanto de L-trp y tRNA<sup>Trp</sup> cargado. El triptofano activa a la proteína reguladora TRAP. TRAP (por sus siglas en inglés, Tryptophan RNA-binding Attenuation Protein) es una proteína de 11 subunidades idénticas. Cada subunidad de TRAP une una molécula de triptofano, y es entonces cuando está en su forma activa que une a la región líder de mRNA promoviendo la formación de un terminador, que causa la terminación de la transcripción. Como será descrito más adelante, cuando el tRNA<sup>Trp</sup> no-cargado se acumula, se lleva a cabo la inactivación de TRAP<sup>168,178,180</sup>.

Cuando el triptofano se empieza a acumular en *E. coli*, se activa un represor, TrpR, que regula negativamente el inicio de la transcripción al unirse a uno o más de los tres operadores localizados en la región promotora del operón *trp*. Este represor también regula negativamente el inicio de la transcripción de los operones de transporte de triptofano y de biosíntesis de corismato<sup>174</sup>. En el momento en el que la concentración de L-trp es inferior a la condición de exceso de este aminoácido, la represión se libera<sup>181</sup>. La región regulatoria del operón *trp* de *E. coli* tiene un segundo "checkpoint" río abajo del promotor, en la región líder del operón. En este punto la célula puede monitorear la disponibilidad del tRNA<sup>Trp</sup> cargado mediante la síntesis de un péptido líder<sup>174</sup>.

En *B. subtilis* la respuesta al triptofano es más compleja debido a que el operón *trp* es parte de un supraoperón junto con otros genes involucrados en la



síntesis de aminoácidos aromáticos<sup>142,143,182</sup>. Como se describió antes, dos promotores sirven como sitio de inicio para la transcripción. El primero de ellos, el promotor de *aroF*, es sujeto a regulación por aminoácidos aromáticos (C. Yanofsky *et al.*, comunicación personal), sin embargo los mecanismos utilizados para ésta regulación son aún desconocidos. En cambio, la regulación del operón *trp* en *B. subtilis* es mejor conocida<sup>142,143,180,183,184</sup>. Cada molécula de RNAPol que inicia la transcripción, ya sea en el promotor de *aroF* o en el promotor del operón de *trp*, cuando empieza a transcribir la región líder del operón *trp*, es sujeta a una decisión del término de la transcripción. Si el triptofano es abundante, se activa TRAP, uniéndose entonces a trinucleótidos NAG (cuyas afinidades son en orden decreciente: GAG, UAG, AAG y CAG) repetidos en el RNA líder del operón *trp*. La unión de TRAP previene la formación de un antiterminador, favoreciendo así la formación de un terminador intrínseco y excluyendo al antiterminador para terminar la transcripción<sup>142,143,183-187</sup>. Una estructura de tallo y asa que se forma en el segmento 5' del RNA líder del operón *trp* parece contribuir a la unión de TRAP<sup>188</sup>, Figura 27.

La unión de TRAP puede regular negativamente la transcripción de la región de genes estructurales del operón *trp* reduciendo su expresión ~200 veces<sup>189</sup>. El pegado de TRAP también regula la síntesis de triptofano de una segunda forma, pues promueve la formación de una estructura de tipo tallo/asa adicional, en el pequeño fragmento de transcritos líder que no han terminado en el atenuador. Esta estructura secundaria inhibe la traducción de *trpE* y le impide al organismo sintetizar triptofano<sup>189,190</sup>. TRAP, en su estado activado por triptofano también inhibe la traducción de *trpG*, el único gen *trp* que no está en el operón *trp*<sup>191,192</sup>, ver figura 1 (página 18). La acumulación del tRNA<sup>Trp</sup> no-cargado incrementa la expresión del operón *trp*.

La figura 1 (ver página 18) muestra el segmento del transcrito líder de *E. coli*, el cual se extiende 162 nucleótidos en el 5' del +1 de *trpE*. Este transcrito puede formar tres estructuras secundarias de RNA mutuamente excluyentes: 1:2, el sitio de pausa, o el anti-antiterminador; 2:3, la estructura del antiterminador; 3:4, la estructura del terminador. Los números se refieren a los segmentos lineales en orden secuencial del transcrito líder del operón *trp*. El terminador, la estructura 3:4, funciona como un terminador intrínseco, cuando se forma dirige a la RNAPol, que está transcribiendo, en un lugar específico que precede a *trpE*. El antiterminador, la estructura 2:3, se forma siempre que el ribosoma esté traduciendo el péptido líder y se detenga en uno de los dos codones de triptofano. La formación del antiterminador previene la formación del terminador. La tercera estructura alternativa de RNA, 1:2, tiene tres funciones: funciona como una estructura de pausa a la transcripción que es esencial para acoplar la transcripción y la traducción en la región líder, funciona como un anti-antiterminador y previene la formación del antiterminador. Además, el segmento 1 codifica para un péptido líder de 14 residuos de aminoácidos, que tiene dos residuos de triptofano en tándem. La capacidad de traducir estos codones a residuos de triptofano se usa para cuantificar la presencia o ausencia de tRNA<sup>Trp</sup> cargados. Siempre que el ribosoma que esté traduciendo se detenga en cualquiera de estos dos codones de triptofano, la estructura del antiterminador se forma, previniendo así, la formación del terminador y permitiendo que la RNA polimerasa continúe con la transcripción hasta los cinco genes estructurales. Cuando el tRNA<sup>Trp</sup> es abundante, se sintetiza el péptido líder, se libera el ribosoma y el terminador se forma, dando fin a la transcripción<sup>174</sup>.

La figura 2 (ver página 19) muestra los eventos secuenciales que pueden ocurrir durante la transcripción de la región líder del operón *trp*, con sus alternativas. La formación de cada estructura de RNA depende de la posición del

ribosoma en el RNAm, al igual que la liberación del ribosoma depende de la región codificante para el péptido líder. La decisión de la terminación de la transcripción de basa en la disponibilidad del tRNA<sup>Trp</sup> cargado. Cuando el operón es transcrito, la RNA polimerasa se detiene después de la estructura de pausa (paso 1). En el paso 2 comienza la traducción y el ribosoma libera la RNA polimerasa pausada en el paso 3. Cuando la célula tiene niveles suficientes de tRNA<sup>Trp</sup> cargado, el ribosoma llega al péptido líder y es entonces liberado. La formación del anti-antiterminador permite la formación del terminador y se termina la transcripción (paso 4a). Cuando la célula no tiene niveles suficientes de tRNA<sup>Trp</sup> cargado el ribosoma sintetiza el péptido líder y se detiene en algún codón de triptofano. Esta pausa hace que se pueda formar el antiterminador, el cual previene la formación del terminador y la transcripción continúa hasta los genes estructurales del operón (paso 4b)<sup>19</sup>.

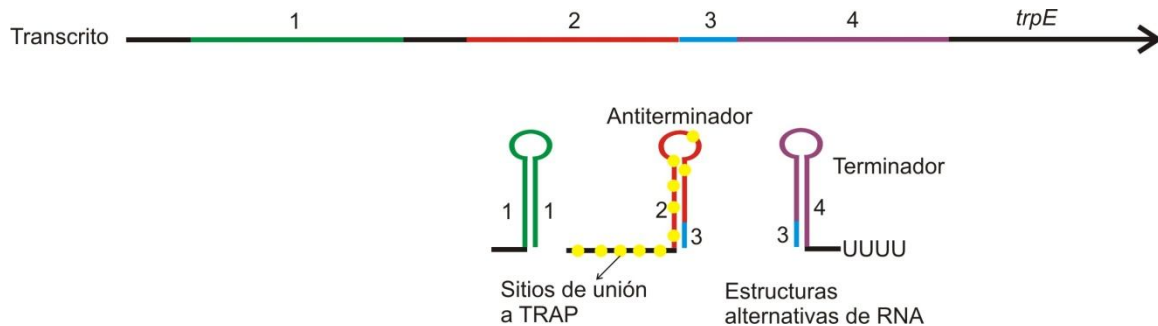


Figura 27. Organización y funciones regulatorias de la región líder del operón *trp* de *Bacillus subtilis*. Propiedades del transcrito líder del operón *trp* de *Bacillus subtilis*.

En *B. subtilis*, la transcripción puede iniciar tanto arriba del promotor de *aroF* como en la región líder del operón *trp*, donde se decide si se termina o no la transcripción. El transcrito líder del operón *trp* consta de 203 nucleótidos de largo y los segmentos de la región involucrados en la regulación se pueden observar en la figura 27. Estos segmentos forman estructuras alternativas de RNA. La estructura

2:3 es una estructura de antiterminador, mientras que la estructura 3:4, que es mutuamente excluyente de la anterior, es un terminador intrínseco. El segmento marcado con el número tres consta de tres nucleótidos y es esencial para la formación del terminador. Por lo tanto, aquí, como en el caso de *E. coli*, la formación del antiterminador imposibilita la formación del terminador. Los círculos en amarillo representan los sitios de unión a la proteína reguladora TRAP, que constan de trinucleótidos NAG repetidos y separados uno del otro por dos nucleótidos. Por tanto, la disponibilidad del triptofano determina si TRAP favorece la terminación de la transcripción, promoviendo la formación del terminador. El triptofano activa a TRAP y es por eso que la disponibilidad del triptofano puede terminar con la transcripción. La estructura 1:1, que se muestra en la figura 4, podría incrementar la función de TRAP.

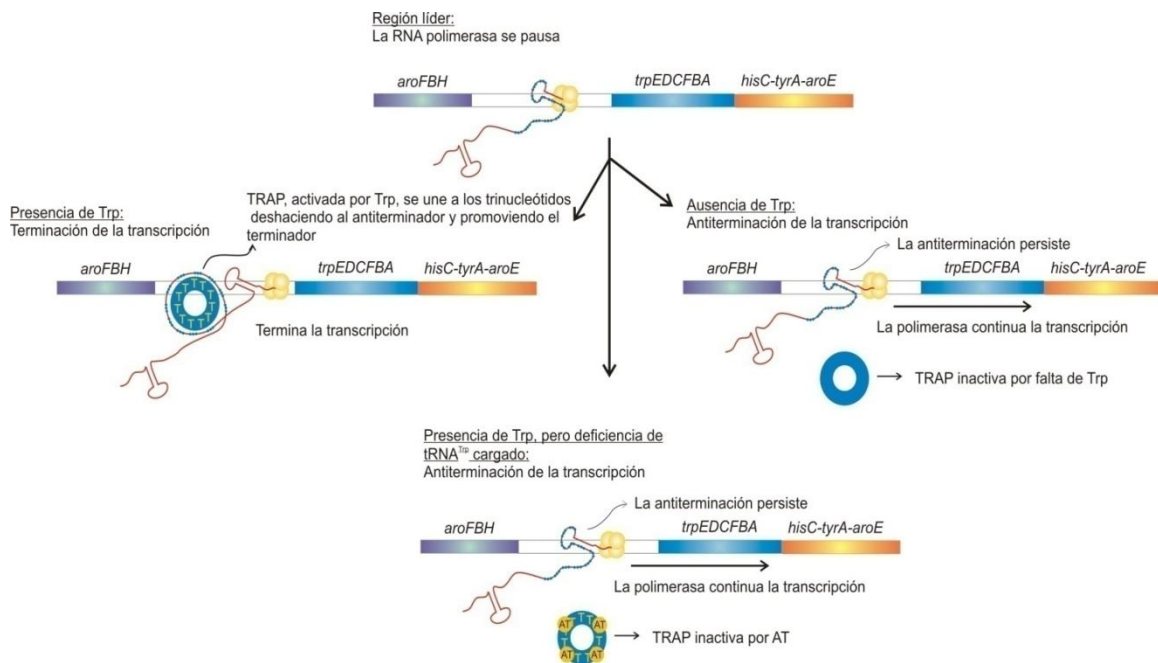


Figura 28. **Organización y funciones regulatorias de la región líder del operón *trp* de *B. subtilis*.** Regulación por atenuación de la transcripción del operón *trp* de *B. subtilis* mediante el triptofano y el tRNA<sup>Trp</sup> no-cargado. Figura tomada de <sup>19</sup>

Los eventos que se muestran en la figura 28 son alternativos y ocurren en la región líder del operón *trp*, estos eventos dependen de si TRAP está activada por triptofano o si está siendo inhibida por Anti-TRAP (AT). Cada molécula de RNAPol que entra a la región líder del operón *trp* se detiene debido a la formación de una estructura de pausa. Si hay suficiente triptofano como para activar a TRAP, ésta se une al RNA líder y previene o interrumpe la formación del antiterminador. Esto permite que se forme el terminador dando fin a la transcripción. Cuando la célula tiene niveles deficientes de triptofano TRAP se inactiva y por lo tanto el antiterminador persiste y continua la transcripción. Cuando las células tienen niveles deficientes de tRNA<sup>Trp</sup> cargado se sintetiza AT. AT se une a TRAP en su forma activa por triptofano, inactivándola, así la transcripción puede continuar en el operón<sup>169</sup>.

*B. subtilis* responde al tRNA<sup>Trp</sup> no-cargado, de la misma manera que a triptofano, como una señal regulatoria. Diversos análisis realizados por el grupo de C. Yanofsky demostraron que la acumulación del tRNA<sup>Trp</sup> no-cargado lleva a la inhibición de la habilidad de TRAP para promover la terminación de la transcripción en la región líder del operón *trp*<sup>169,170,193,194</sup>. Un análisis exhaustivo de estos resultados llevó en el 2001 al grupo de Yanofsky a identificar el operón responsable de cuantificar el tRNA<sup>Trp</sup>, la función de este operón era desconocida y nombrándosele operón *at* porque uno de sus productos es la proteína Anti-TRAP, AT<sup>169,195,196</sup>. AT puede unirse a la forma activa de TRAP por triptofano inhibiendo su capacidad de unirse al RNA líder<sup>169</sup>. La estructura del operón *at* se pueden ver en la figura 29.

## El operon *at*

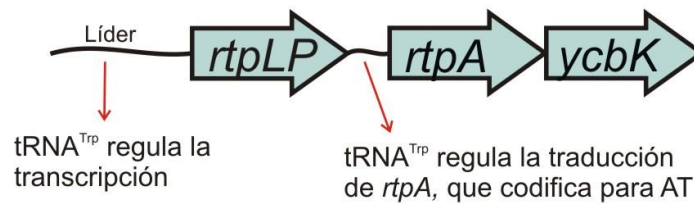


Figura 29. Organización del operón *at* de *B. subtilis* y la regulación de su expresión mediada por tRNA<sup>Trp</sup>.

La región regulatoria del operón *at* puede cuantificar el tRNA<sup>Trp</sup> no-cargado tanto traduccional como transcripcionalmente<sup>180</sup>. A nivel de la transcripción, la regulación está controlada por el mecanismo de atenuación T box. Inmediatamente después de la región líder donde está la T box hay una región codificante, *rtpLP*, de un péptido líder de diez residuos de aminoácidos, y contiene tres codones de triptofano organizados en tándem<sup>180</sup>. El codón de paro de *rtpLP* está localizado seis nucleótidos arriba de la región Shine-Dalgarno del gen *rtpA*, el gen para AT. Esta proximidad sugiere que una vez que el ribosoma termine la traducción del gen *rtpLP* podría inhibir el inicio de la traducción de *rtpA* limitando la síntesis de AT<sup>178,180</sup>. Por lo que si hay suficiente tRNA<sup>Trp</sup> cargado para la traducción completa del péptido líder *rtpLP* el ribosoma alcanza el codón de paro de *rtpLP* en donde inhibe el inicio de la traducción de *rtpA* y la proteína AT no se produce. Cuando la célula realmente tiene deficiencia de tRNA<sup>Trp</sup> cargado el ribosoma se detiene un uno de los tres codones de triptofano. Esto libera la región Shine-Dalgarno de *rtpA*, permitiendo síntesis suficiente de AT. Altos niveles de AT inhiben a TRAP en su estado activo por triptofano. De esta manera la transcripción del operón *trp* procede. Es de esta manera como las propiedades estructurales del operón *at* antes mencionadas, ayudan a determinar los niveles de tRNA<sup>Trp</sup> cargado y no-cargado, y a proporcionar una respuesta regulatoria apropiada<sup>19</sup>.

### Comparación de las estrategias regulatorias.

Las estrategias de regulación para *E. coli* y *B. subtilis* descritas anteriormente, cuantifican y responden al triptofano y al tRNA<sup>Trp</sup> no-cargado (Figura 30). Lo más probable es que en un ancestro de *E. coli* ocurrió la duplicación de *trpG* y una copia se fusionó a *trpD* dedicándose a la síntesis de triptofano, mientras que la otra copia se especializó en la síntesis de folato, convirtiéndose en *pabA*<sup>134</sup>. La diferencia, quizá mas notable, entre los dos organismos es que el operón *trp* de *B. subtilis* reside dentro de un supraoperón de 12 genes, por lo que existe la necesidad de varios promotores para su transcripción. Aunque un represor activado por triptofano podría regular la región del operón *trp*, probablemente no regularía la transcripción de *aroF*<sup>19</sup>. Otra de las diferencias, es el mecanismo que usan para regular a sus correspondientes genes *trpS*, genes codificantes de la triptofanil-tRNA sintetasa. En *B. subtilis* se utiliza el mecanismo descrito anteriormente llamado T box, mientras que *E. coli* la expresión de *trpS* depende de la tasa de crecimiento y no de la molécula de tRNA<sup>Trp</sup><sup>197</sup>. En el caso de *E. coli*, esta regulación también se lleva a cabo a nivel transcripcional<sup>197</sup>, aunque los mecanismos moleculares no se han descrito todavía. Una diferencia ancestral adicional es que la atenuación de la transcripción mediada por el ribosoma es un mecanismo común para la regulación de operones biosintéticos de aminoácidos en enterobacterias, pero no en *B. subtilis*<sup>19</sup>.

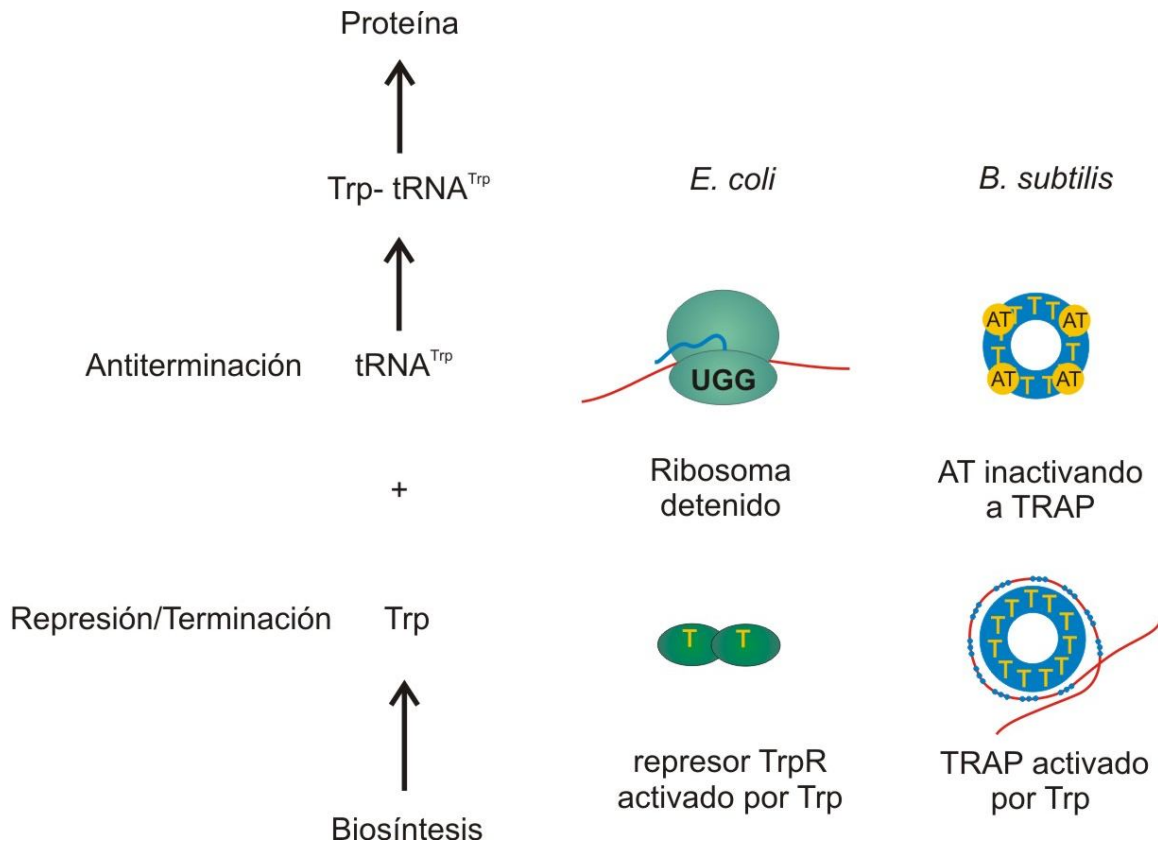


Figura 30. Comparación de las estrategias regulatorias utilizadas por *E. coli* y por *B. subtilis* para sensor y responder a las moléculas de triptofano y de tRNA<sup>Trp</sup>. Figura modificada de<sup>169</sup>.

En *E. coli* el triptofano activa el represor TrpR. Este represor en su forma activa se une a la región del operador *trp* e inhibe el inicio de la transcripción en el promotor de *trp*. En *B. subtilis* el triptofano activa a la proteína TRAP. Este regulador en su forma activa se une al RNA líder del operón *trp* ocasionando que la transcripción llegue a su fin. Cuando el tRNA<sup>Trp</sup> no-cargado se acumula en *E. coli* el ribosoma encargado de la traducción de la región del péptido líder se detiene en alguno de los dos codones de triptofano. Esta pausa permite la formación del antiterminador, previniendo la terminación de la transcripción. En *B. subtilis* la acumulación de tRNA<sup>Trp</sup> no-cargados, activa la transcripción en la región líder del operón *at*, sintetizando AT. El AT producido inactiva a TRAP para que no se una a



los sitios de unión a TRAP, aumentando la transcripción del operón *trp* y la traducción de *trpG*.

Durante mi proyecto de Licenciatura, estudié qué ocurría en el resto de los Firmicutes. ¿se comportaban igual que *B. subtilis*? Existe una gran diversidad en cuanto a los diferentes mecanismos moleculares utilizados para regular la expresión de los genes de biosíntesis de triptofano. Estos incluyen tanto a represores transcripcionales (TrpR), proteínas de unión a RNA (TRAP), o bien, sistemas de traducción acoplados a atenuación, así como elementos de regulación en *cis*, i.e., riboswitches (Figura 31).

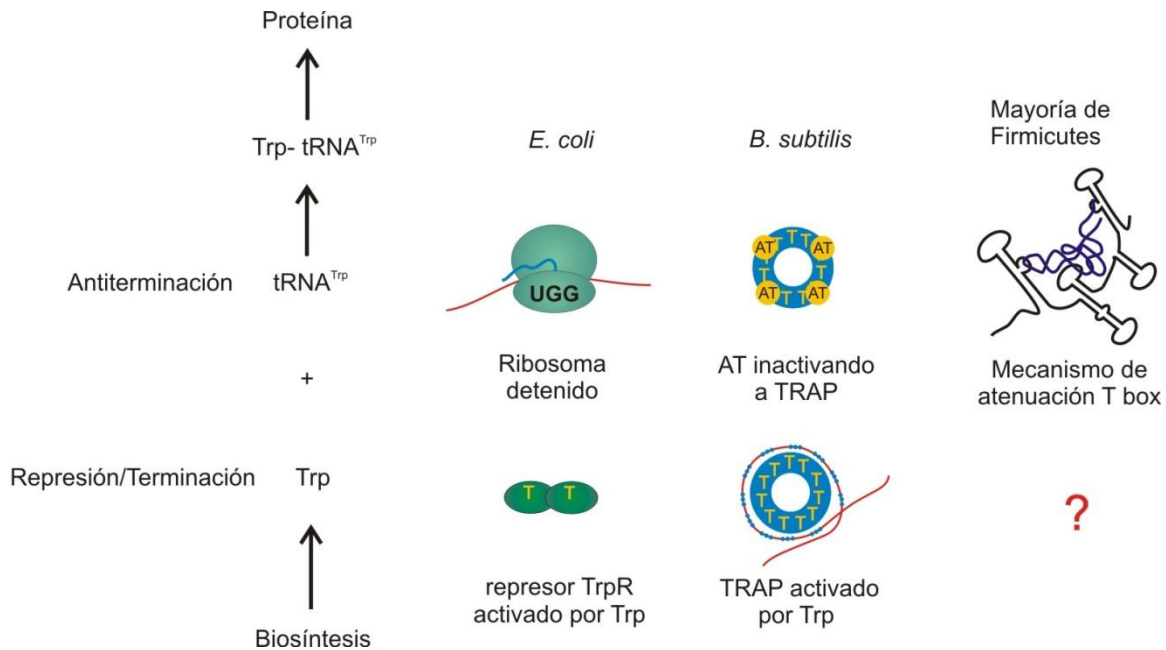


Figura 31. Esquema comparativo de los mecanismos moleculares involucrados en la regulación de los genes de biosíntesis de triptofano en organismos modelo y Firmicutes.

En cuanto a las señales reconocidas para dicha regulación, también se observan ciertas diferencias entre organismos que integran las concentraciones

intracelulares de L-Trp y el tRNA<sup>Trp</sup> cargado/descargado, o bien, exclusivamente ésta última condición. A la fecha no existe ningún estudio comparativo en cuanto a los límites de sensibilidad y la velocidad de respuesta de los diferentes mecanismos en relación a la disponibilidad del aminoácido, no obstante desde el punto de vista evolutivo dicha divergencia pudiera ser originada por los diferentes nichos que estos organismos habitan.

Es claro que los dominios catalíticos de las enzimas de biosíntesis de triptofano son homólogos para todos los organismos existiendo así una gran conservación. Por el contrario, los mecanismos de regulación han variado incluso en organismos filogenéticamente cercanos mostrando una gran plasticidad evolutiva.

### ***trpS termosensible en Bacillus halodurans***

Cabe resaltar que *B. halodurans* se comporta muy parecido a *B. subtilis*, tienen el mismo contexto génico para el operon de triptófano, así como la región regulatoria y la proteína regulatoria TRAP. Sin embargo en *B. halodurans*, no se encuentra la proteína Anti-TRAP. La pregunta que aquí resalta es ¿cómo hace *B. halodurans* para detectar los niveles de tRNA<sup>TRP</sup> descargado?

Se realizó, sin éxito, un estudio preliminar de posibles T boxes de triptófano regulando algún pequeño ORF. Estas no fueron encontradas, por lo que si el tRNA<sup>TRP</sup> descargado se sensa de alguna manera, no es mediante la T box. Recordemos que históricamente, Charles Yanosfky descubrió la proteína Anti-TRAP construyendo cepas de *B. subtilis* que contuvieran una TrpRS termosensible<sup>169,170,194</sup>. De modo que controlando con temperatura, cuándo había y cuándo no TrpRS dentro de la célula, se dio cuenta que había regulación mediante los niveles de tRNA<sup>TRP</sup> cargado/descargado y esto llevó a su grupo a descubrir la

proteína Anti-TRAP. Pienso que siguiendo una aproximación muy similar, pero en *B. halodurans*, podría encontrarse la manera en que este organismo detecta sus niveles de tRNA<sup>TRP</sup> descargado. Un control que valdría la pena hacer, es quitar el gen *yczA* de *B. subtilis* y medir si la expresión se parece a aquella de *B. halodurans*, cuyo caso sería que no cumple con el paradigma de sensar niveles de L-Trp y tRNA<sup>TRP</sup>.

### ***TRAP en Oceanobacillus iheyensis***

Otro organismo de interés en este aspecto es *O. iheyensis*. Esta bacilacea tiene un contexto génico del estilo de aquellos Firmicutes que regulan su biosíntesis de triptófano mediante T box, y así mismo, también utiliza este riboswitch para llevar a cabo la regulación de su operon *trp*. Sin embargo, esta bacteria tiene un homólogo de la proteína TRAP, y además no presenta ningún TRAP binding site estadísticamente significativo a lo largo de su genoma. ¿Cuál es el rol de esta proteína? ¿Fue funcional y dejó de serlo al adquirir T box en su operón de triptófano? O, por el contrario, ¿presenta este organismo un estado ancestral que pudiera darnos indicios de cuál era el rol de TRAP antes de ser el regulador de la biosíntesis de triptófano en *B. subtilis*? Para responder estas preguntas será necesario deletar el gen *mtrB* de *O. iheyensis* y complementarla con aquel gen de *B. subtilis*, y para tener una mejor idea, realizar lo inverso en *B. subtilis*.

### ***Clostridias con represores de triptófano.***

Durante la búsqueda del regulón T box se encontró que *C. beijerinckii* tiene un represor de triptófano. Este sería el primero reportado para este grupo de organismos. Una perspectiva interesante, es caracterizar estos represores en Clostridia

## **Apéndices**

### **Apéndice I: Biochemical Features and Functional Implications of the RNA-Based T box Regulatory Mechanism.**

## **Apéndice II: Genome Sequence Databases: Types of Data and Bioinformatic Tools.**

## Apéndice III: An Evolutionary Perspective on Amino Acids

Nature Scitable: <http://www.nature.com/scitable> es una colección de artículos gratuitos y de herramientas de aprendizaje sobre genética y biología celular. Incluye tópicos de evolución, expresión génica y muchos otros. Todo esto con el objetivo de proporcionar a estudiantes recursos gratuitos para mejorar sus conocimientos.

Para la sección de Biología Celular y la en el tópicos Cells Origins and Metabolism, escribí junto con Mariana Peimbert y Héctor Romero el siguiente artículo:

### An Evolutionary Perspective on Amino Acids

By: Ana Gutiérrez-Preciado, B.Sc. (*Departamento de Microbiología Molecular, Universidad Nacional Autónoma de México*), Hector Romero, B.Sc. (*Laboratorio de Organización y Evolución del Genoma, Facultad de Ciencias Igua*) & Mariana Peimbert, Ph.D. (*Departamento de Ciencias Naturales, Universidad Autónoma Metropolitana*) © 2010 Nature Education

Citation: Gutiérrez-Preciado, A., Romero, H. & Peimbert, M. (2010) An Evolutionary Perspective on Amino Acids. *Nature Education* 3(9):29

*Amino acids are one of the first organic molecules to appear on Earth. What are they made of and how have they evolved?*

Amino acids play a central role in cellular metabolism, and organisms need to synthesize most of them. Many of us become familiar with amino acids when we first learn about translation, the synthesis of protein from the nucleic acid code in mRNA. To date, scientists have discovered more than five hundred amino acids in nature, but only twenty-two participate in translation. In 1943, Gordon, Martin, and Synge used partition chromatography to separate and study constituents of proteins (Gordon, Martin, & Synge 1943), a major breakthrough that contributed to the rapid identification of the twenty amino acids used in proteins by all living organisms. After this initial burst of discovery, two additional amino acids, which

are not used by all organisms, were added to the list: selenocysteine (Bock 2000) and pyrrolysine (Srinivasan et al. 2002).

Aside from their role in composing proteins, amino acids have many biologically important functions. They are also energy metabolites, and many of them are essential nutrients. Amino acids can often function as chemical messengers in communication between cells. For example, Arvid Carlsson discovered in 1957 that the amine 3-hydroxytyramine (dopamine) was not only a precursor for the synthesis of adrenaline from tyrosine, but is also a key neurotransmitter. Certain amino acids — such as citrulline and ornithine, which are intermediates in urea biosynthesis — are important intermediaries in various pathways involving nitrogenous metabolism. Although other amino acids are important in several pathways, S-adenosylmethionine acts as a universal methylating agent. What follows is a discussion of amino acids, their biosynthesis, and the evolution of their synthesis pathways, with a focus on tryptophan and lysine.

## **The Origins of Nutrient Biosynthesis**

In 1953, Miller and Urey attempted to re-create the conditions of primordial Earth. In a flask, they combined ammonia, hydrogen, methane, and water vapor plus electrical sparks (Miller 1953). They found that new molecules were formed, and they identified these molecules as eleven standard amino acids. From this observation, they posited that the first organisms likely arose in an environment similar to the one they constructed in their flask, one rich in organic compounds, now widely described as the primordial soup. This hypothesis is further extended to the claim that, within this soup, single-celled organisms evolved, and as the number of organisms increased, the organic compounds were depleted.

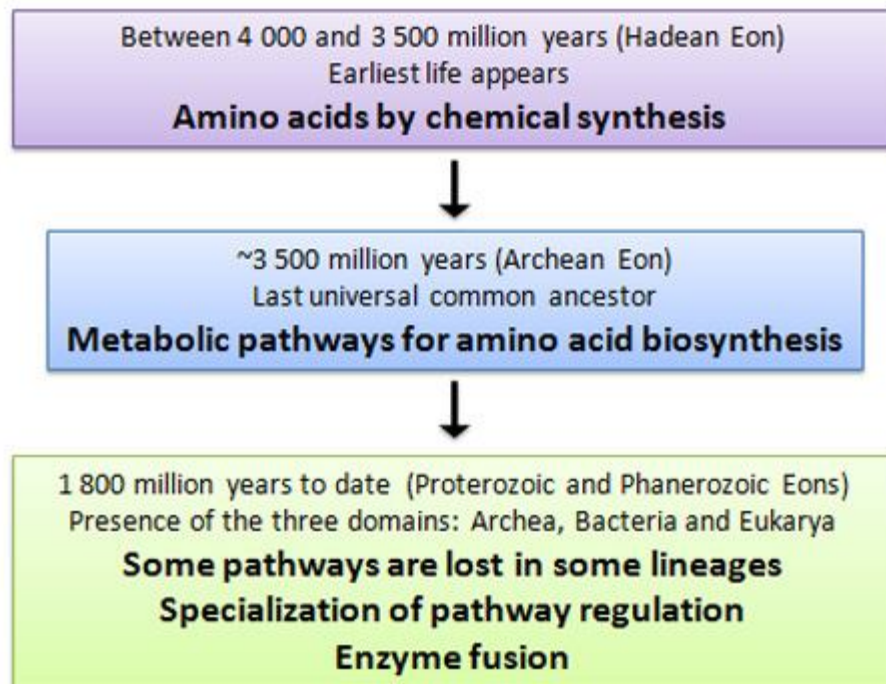
Necessarily, in this competitive environment, those organisms that were able to biosynthesize their own nutrients from elements had a great advantage over those that could not (Figure 1). Today, the vast majority of organic compounds derive from biological organisms that break down and replenish the resources for sustaining other organisms. And, rather than emerging from an electrified primordial soup, amino acids emerge from biosynthetic enzymatic reactions.

### **What Is an Amino Acid Made Of?**

As implied by the root of the word (amine), the key atom in amino acid composition is nitrogen. The ultimate source of nitrogen for the biosynthesis of amino acids is atmospheric nitrogen ( $N_2$ ), a nearly inert gas. However, to be metabolically useful, atmospheric nitrogen must be reduced. This process, known as nitrogen fixation, occurs only in certain types of bacteria. Even though nitrogen is one of the most prominent chemical elements in living systems,  $N_2$  is almost unreactive (and very stable) because of its triple bond ( $N\equiv N$ ). This bond is extremely difficult to break because the three chemical bonds need to be separated and bonded to different compounds. Nitrogenase is the only family of enzymes capable of breaking this bond (i.e., it carries out nitrogen fixation). These proteins use a collection of metal ions as the electron carriers that are responsible for the reduction of  $N_2$  to  $NH_3$ . All organisms can then use this reduced nitrogen ( $NH_3$ ) to make amino acids. In humans, reduced nitrogen enters the physiological system in dietary sources containing amino acids. All organisms contain the enzymes glutamate dehydrogenase and glutamine synthetase, which convert ammonia to glutamate and glutamine, respectively. Amino and amide groups from these two compounds can then be transferred to other carbon backbones by transamination and transamidation reactions to make amino acids. Interestingly, glutamine is the universal donor of amine groups for the formation of many other



amino acids as well as many biosynthetic products. Glutamine is also a key metabolite for ammonia storage. All amino acids, with the exception of proline, have a primary amino group (NH<sub>2</sub>) and a carboxylic acid (COOH) group. They are distinguished from one another primarily by , appendages to the central carbon atom.

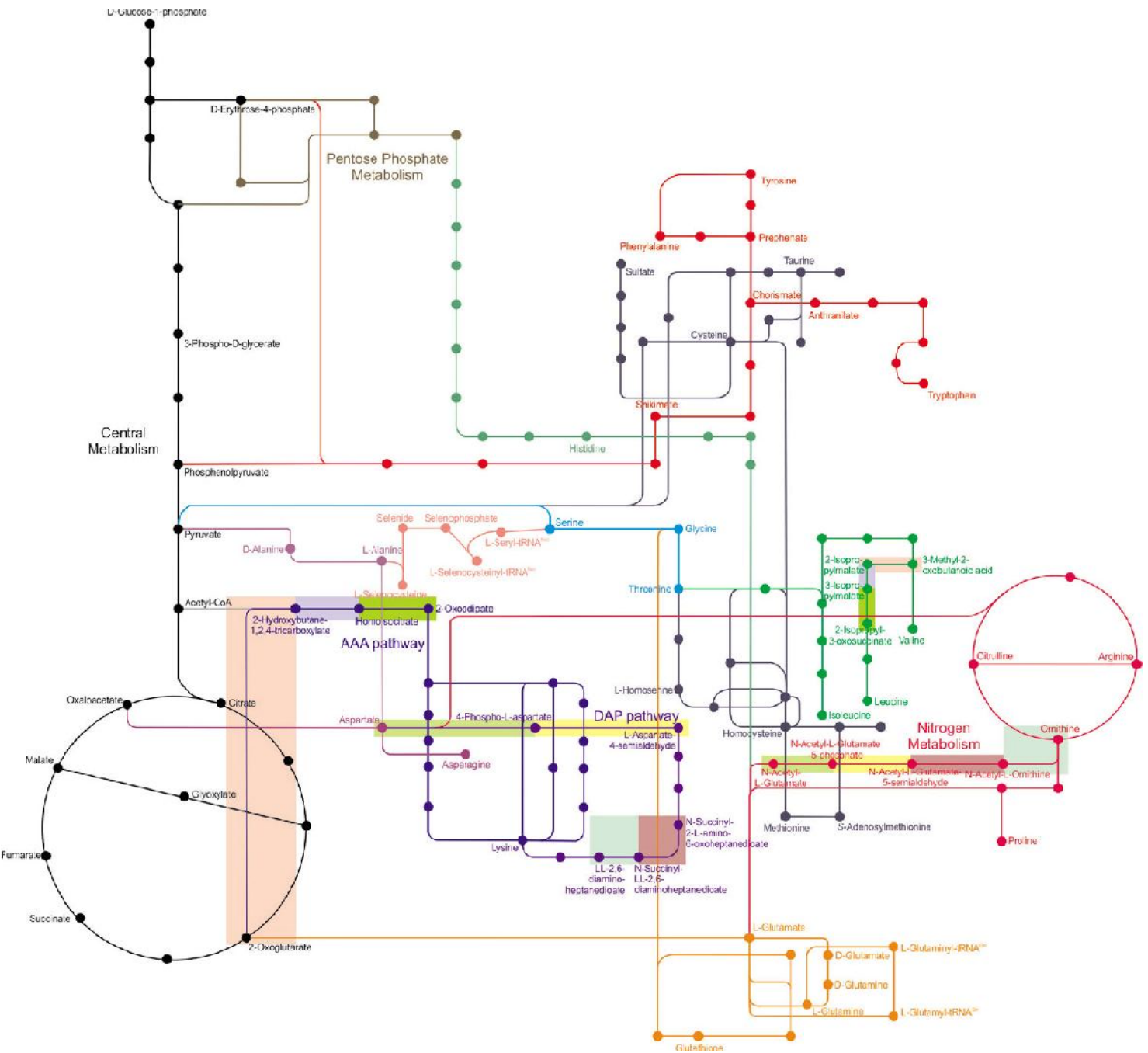


**Figure 1: Major events in the evolution of amino acid synthesis**

The way amino acids are synthesized has changed during the history of Earth. The Hadean eon represents the time from which Earth first formed. The subsequent Archean eon (approximately 3,500 million years ago) is known as the age of bacteria and archaea. The Proterozoic eon was the gathering up of oxygen in Earth's atmosphere, and the Phanerozoic eon coincides with

## Amino Acid Precursors and Biosynthesis Pathways

In the study of metabolism, a series of biochemical reactions for compound synthesis or degradation is called a pathway. Amino acid synthesis can occur in a variety of ways. For example, amino acids can be synthesized from precursor molecules by simple steps. Alanine, aspartate, and glutamate are synthesized from keto acids called pyruvate, oxaloacetate, and alpha-ketoglutarate, respectively, after a transamination reaction step. Similarly, asparagine and glutamine are synthesized from aspartate and glutamate, respectively, by an amidation reaction step. The synthesis of other amino acids requires more steps; between one and thirteen biochemical reactions are necessary to produce the different amino acids from their precursors of the central metabolism (Figure 2). The relative uses of amino acid biosynthetic pathways vary widely among species because different synthesis pathways have evolved to fulfill unique metabolic needs in different organisms. Although some pathways are present in certain organisms, they are absent in others. Therefore, experimental results about amino acid metabolism that are achieved with model organisms may not always have relevance for the majority of other organisms.



**Figure 2: Amino acid metabolism in context**

Four metabolism pathways are depicted: central metabolism (in black), pentose phosphate metabolism (in brown), nitrogen metabolism (in magenta), and various amino acid metabolism pathways (all other colors). Nodes (dots) represent metabolites, and lines represent enzymes and intermediates. The nitrogen metabolism pathway overlaps with the biosynthesis of arginine and proline, with glutamate as the shared precursor. Histidine biosynthesis branches off the pentose phosphate metabolism. Lysine (AAA) biosynthesis can be synthesized through different pathways, the aminoacidate (AAA) pathway or the diaminoipitate (DAP) pathway (shown in dark blue). There are gene homologs between different biosynthetic pathways. In the dark blue pathways,

shaded rectangles represent homologies between enzymes. Similarly, the AAA pathway contains enzymes that share homologies with the branched chain amino acid (BCAA) pathways, whereas the DAP pathway contains homologies with the arginine biosynthetic pathway. In various pathways, homologous enzymes are denoted by shaded rectangles. Different shaded colors indicate different pairs of homologous enzymes.

## What Makes an Amino Acid Essential?

Not all the organisms are capable of synthesizing all the amino acids, and many are synthesized by pathways that are present only in certain plants and bacteria. Mammals, for example, must obtain eight of twenty amino acids from their diets. This requirement leads to a convention that divides amino acids into two categories: essential and nonessential (given a certain metabolism). Because of particular structural features, essential amino acids cannot be synthesized by mammalian enzymes (Reeds 2000). Nonessential amino acids, therefore, can be synthesized by nearly all organisms. The loss of the ability to synthesize essential amino acids likely emerged very early in evolution, because this dependence on other organisms for the source of amino acids is common among all eukaryotes, not just those of mammals.

How do certain amino acids become essential for a given organism? Studies in ecology and evolution give some clues. Organisms evolve under environmental constraints, which are dynamic over time. If an amino acid is available for uptake, the selective pressure to keep intact the genes responsible for that pathway might be lowered, because they would not be constantly expressing these biosynthetic genes. Without the selective pressure, the biosynthetic routes might be lost or the gene could allow mutations that would lead to a diversification of the enzyme's function. Following this logic, amino acids that are essential for certain organisms might not be essential for other organisms subjected to different selection pressures. For example, in 2000, Ishikawa and colleagues completed the genome sequence of the endosymbiont bacteria *Buchnera*, and in it they found the genes

for the biosynthetic pathways necessary for the synthesizing essential amino acids for its symbiotic host, the aphid. Interestingly, those genes for the synthesis of its "nonessential" amino acids are almost completely missing (Shigenobu et al. 2000). In this way, Buchnera provides the host with some amino acids and obtains the other amino acids from the host (Baumann 2005; Pal et al. 2006).

### **Tryptophan Synthesis: Only Created Once**

Free-living bacteria synthesize tryptophan (Trp), which is an essential amino acid for mammals, some plants, and lower eukaryotes. The Trp synthesis pathway appears to be highly conserved, and the enzymes needed to synthesize tryptophan are widely distributed across the three domains of life. This pathway is one of three that compose aromatic amino acids from chorismate (Figure 2, red pathway). (The other amino acids are phenylalanine and tyrosine.) Trp biosynthetic enzymes are widely distributed across the three domains of life (Xie et al. 2003). The genes that code for the enzymes in this pathway likely evolved once, and they did so more recently than those for other amino acid synthesis pathways. Researchers made this contention because all organisms containing this Trp synthesis pathway use homologous enzymes (Merino, Jensen, & Yanofsky 2008). As another point of distinction, the Trp pathway is the most biochemically expensive of the amino acid pathways, and for this reason it is expected to be tightly regulated.

### **Lysine Synthesis: Created Multiple Times**

To date, scientists have discovered six different biosynthetic pathways in different organisms that synthesize lysine. These pathways can be grouped into the diaminopimelic acid (DAP) and aminoadipic acid (AAA) pathways (Figure 2, dark blue). The DAP pathway synthesizes lysine (Lys) from aspartate and pyruvate. Most bacteria, some archaea, fungi, algae, and plants use the DAP pathways. On the

other hand, the AAA pathways synthesize Lys from alpha-ketoglutarate and acetyl coenzyme A. Most fungi, some algae, and some archaea use this route. Why do we observe this diversity, and why does it occur particularly for Lys synthesis? Interestingly, the DAP pathways retain duplicated genes from the biosynthesis of arginine, whereas the AAA pathways retain duplicated genes from leucine biosynthesis (Figure 2), indicating that each of the pathways experienced at least one duplication event during evolution (Hernandez-Montes et al. 2008; Velasco et al. 2002). Fani and coworkers performed a comparative analysis of the synthesis enzyme sequences and their phylogenetic distribution that suggested that the synthesis of leucine, lysine, and arginine were initially carried out with the same set of versatile enzymes. Over the course of time came a series of gene duplication events and enzyme specializations that gave rise to the unambiguous pathways we know today. Which of the pathways appeared earlier is still a source of query and debate.

To support this hypothesis, there is evidence from a fascinating archaea, *Pyrococcus horikoshii*. This organism can synthesize leucine, lysine, and arginine, yet its genome contains only genes for one pathway. Such a gap indicates that *P. horikoshii* has a mechanism similar to the ancestral one: versatile enzymes. Biochemical experiments are needed to further support the idea that these enzymes can use multiple substrates and to rule out the possibility that amino acid synthesis in this organism does not arise from enzymes yet unidentified.

### **Synthesis on the tRNA molecule**

Selenocysteine (SeC) (Bock 2000) is a genetically encoded amino acid not present in all organisms. Scientists have identified SeC in several archaeal, bacterial, and eukaryotic species (even mammals). When present, SeC is usually confined to

active sites of proteins involved in reduction-oxidation (redox) reactions. It is highly reactive and has catalytic advantages over cysteine, but this high reactivity is undermined by its potential to cause cell damage if free in the cytoplasm. Hence, it is too dangerous, and no pool of free SeC is available. How, then, is this amino acid synthesized for use in protein synthesis? The answer demonstrates the versatility of synthesis strategies deployed by organisms forced to cope with singularities. The synthesis of SeC is carried out directly on the tRNA substrate before being used in protein synthesis. First, SeC-specific tRNA (tRNA<sup>Sec</sup>) is charged with serine via seril-tRNA synthetase, which acts in a somehow promiscuous fashion, serilating either tRNA<sup>Ser</sup> or tRNA<sup>Sec</sup>. Then, another enzyme modifies Ser to SeC by substituting the OH radical with SeH, using selenophosphate as the selenium donor (Figure 2, pink pathway). This synthesis is a form of a trick to avoid the existence of a free pool of SeC while still maintaining a source of SeC-tRNA<sup>Sec</sup> needed for protein synthesis. Strictly speaking, this mechanism is not an actual synthesis of amino acids, but rather a synthesis of aminoacylated-tRNAs. However, this technique involving tRNA directly is not exclusive to SeC, and similar mechanisms dependent on tRNA have been described for asparagine, glutamine, and cysteine. Owing to its appearance of SeC across all three domains of life, scientists wonder if it is an ancestral mechanism for amino acid biosynthesis or simply a coincidence of selection pressures.

### **How Do Metabolic Pathways Evolve? Two Different Models**

In 1945, Horowitz proposed the first accepted model for metabolic pathway evolution (Horowitz 1945). Called the retrograde model, it states that after an enzyme consumes all its substrate available, another enzyme capable of producing the aforementioned substrate is required, so the last enzyme evolved to the preceding one by a gene duplication and selection mechanism. In other words,

enzymes evolve from others with similar substrate specificity, and the substrate of the last enzyme is the product of the preceding one. Also, the active site must bind both the substrate and the product. This model became very popular, but as more genes have been sequenced and more phylogenetic analyses performed, this mechanism has become less seemingly plausible and therefore unpopular. An alternative model, the patchwork assembly model, proposes that ancestral enzymes were generalists, so they could bind a number of substrates to carry out the same type of reaction. Gene duplication events followed by evolutionary divergence would result in enzymes with high affinity and specificity for a substrate. In other words, enzymes are recruited from others with the same type of chemical reaction. Whole genome analysis of *Escherichia coli* supports the patchwork evolution model (Teichmann et al. 2001). Duplication of whole pathways does not occur very often; nevertheless, examples include tryptophan (to synthesize paraminobenzoate) and histidine (to synthesize nucleotides) biosynthesis, as well as lysine, arginine, and leucine biosynthesis (see aforementioned example).

Other mechanisms, such as gene fusion, might occur in the process of pathway evolution. When gene fusions occur between the genes for different proteins of the same pathway, a mechanism that facilitates ligand binding is provided because the substrate of one domain is the product of the other; thus, passive diffusion becomes unnecessary. Fusions can also result in the tight regulation of fused domains. Histidine biosynthesis is a good example of gene fusion; at least seven genes of this pathway underwent fusion events in different phylogenetic lineages. This assertion means that fusions must be relatively recent because they occurred after the lineages arose (Fani et al. 2007). Another important pathway evolution mechanism is horizontal gene transfer, which allows the rapid acquisition of fully functional enzymes and pathways.



## Open Questions about Amino Acid Evolution

Amino acids are one of the first organic molecules to appear on Earth. As the building blocks of proteins, amino acids are linked to almost every life process, but they also have key roles as precursor compounds in many physiological processes. These processes include intermediary metabolism (connections between carbohydrates and lipids), signal transduction, and neurotransmission. Recent years have seen great advances in understanding amino acid evolution, yet many questions on the subject of amino acid synthesis remain. What was the order of appearance of amino acids over evolutionary history? How many amino acids are used in protein synthesis today? How many were present when life began? Were there initially more than twenty used for building blocks, but intense selective process streamlined them down to twenty? Conversely, was the initial set much less than twenty, and did new amino acids successively emerge over time to fit into the protein synthesis repertoire? What are the tempo and mode of amino acid pathway evolution? These questions are waiting to be tackled — with old or new hypotheses, conceptual tools, and methodological tools — and are ripe for a new generation of scientists.

## Summary

Scientists now recognize twenty-two amino acids as the building blocks of proteins: the twenty common ones and two more, selenocysteine and pyrrolysine. Amino acids have several functions. Their primary function is to act as the monomer unit in protein synthesis. They can also be used as substrates for biosynthetic reactions; the nucleotide bases and a number of hormones and neurotransmitters are derived from amino acids. Amino acids can be synthesized from glycolytic or Krebs cycle intermediates. The essential amino acids, those that are needed in the diet, require more steps to be synthesized. Some amino acids

need to be synthesized when charged onto their corresponding tRNAs. We have discussed only two biosynthetic routes: the Trp pathway, which appears to have evolved only once, and the Lys pathway, which seems to have evolved independently in different lineages. Prevailing evidence suggests that metabolic pathways themselves seem to be evolving following the patchwork assembly model, which proposes that pathways originated through the recruitment of generalist enzymes that could react with a wide range of substrates. The study of the evolution of amino acid metabolism has helped us understand the evolution of metabolism in general.

## References and Recommended Reading

---

- Baumann, P. Biology bacteriocyte-associated endosymbionts of plant sap-sucking insects. *Annual Review of Microbiology* **59**, 155–189 (2005) doi:10.1146/annurev.micro.59.030804.121041.
- Bock, A. Biosynthesis of selenoproteins — an overview. *Biofactors* **11**, 77–78 (2000).
- Fani, R. *et al.* The role of gene fusions in the evolution of metabolic pathways: The histidine biosynthesis case. *BMC Evolutionary Biology* **7 Suppl 2**, S4 (2007) doi:10.1186/1471-2148-7-S2-S4.
- Gordon, A. H., Martin, A. J. & Synge, R. L. Partition chromatography in the study of protein constituents. *Biochemical Journal* **37**, 79–86 (1943).
- Hernandez-Montes, G. *et al.* The hidden universal distribution of amino acid biosynthetic networks: A genomic perspective on their origins and evolution. *Genome Biology* **9**, R95 (2008) doi:10.1186/gb-2008-9-6-r95.
- Horowitz, N. H. On the evolution of biochemical syntheses. *Proceedings of the National Academy of Sciences* **31**, 153–157 (1945).
- Merino, E., Jensen, R. A. & Yanofsky, C. Evolution of bacterial *trp* operons and their regulation. *Current Opinion in Microbiology* **11**, 78–86 (2008) doi:10.1016/j.mib.2008.02.005.
- Miller, S. L. A production of amino acids under possible primitive earth conditions. *Science* **117**, 528–529 (1953).
- Pal, C. *et al.* Chance and necessity in the evolution of minimal metabolic networks. *Nature* **440**, 667–670 (2006) doi:10.1038/nature04568.
- Reeds, P. J. Dispensable and indispensable amino acids for humans. *Journal of Nutrition* **130**, 1835S–1840S (2000).
- Shigenobu, S. *et al.* Genome sequence of the endocellular bacterial symbiont of aphids *Buchnera* sp. APS. *Nature* **407**, 81–86 (2000) doi:10.1038/ng986.
- Srinivasan, G., James, C. M. & Krzycki, J. A. Pyrrolysine encoded by UAG in archaea: Charging of a UAG-decoding specialized tRNA. *Science* **296**, 1459–1462 (2002) doi:10.1126/science.1069588.
- Teichmann, S. A. *et al.* The evolution and structural anatomy of the small molecule metabolic pathways in *Escherichia coli*. *Journal of Molecular Biology* **311**, 693–708 (2001) doi:10.1006/jmbi.2001.4912.
- Velasco, A. M., Leguina, J. I. & Lazcano, A. Molecular evolution of the lysine biosynthetic pathways. *Journal of Molecular Evolution* **55**, 445–459 (2002) doi:10.1007/s00239-002-2340-2.
- Xie, G. *et al.* Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiology and Molecular Biology Reviews* **67**, 303–342 (2003) doi:10.1128/MMBR.67.3.303-342.2003.

## **Apéndice IV: Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach**

## Apéndice V: Verificación de la hipótesis de la reina negra sobre la reducción genómica de las metanoarqueas

El tema de esta sección diverge mucho del tema principal de esta tesis. Sin embargo lo incluyo, pues lo realicé durante el proyecto de doctorado, en una estancia corta en el laboratorio del Dr. Andrés Moya, en el Insitut Cavanilles de Biodiversitat i Biologia Evolutiva, en el 2012. Este proyecto se realizó en colaboración con Enrique Merino y Bruno Contreras.

### Introducción

#### Bacterias sintróficas

En microbiología, existen muchos ejemplos de sintrofia, un proceso metabólico donde dos organismos distintos cooperan para degradar algún metabolito (y obtener energía de ello) y que por su cuenta no podrían hacerlo. La mayoría de las reacciones sintróficas son procesos de fermentaciones secundarias, donde los microorganismos fermentan los productos que a su vez, otros microorganismos anaeróbicos han fermentado<sup>198</sup>.

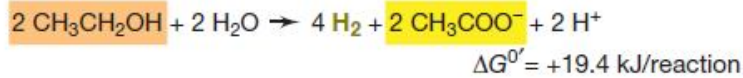
Ecológicamente hablando, las bacterias sintróficas actúan como uniones clave en partes anóxicas del ciclo del carbono. Sin ellas, existiría un cuello de botella en los ambientes anóxicos donde otros aceptores de electrones (aparte del CO<sub>2</sub>) serian limitantes. Por tanto, cuando los ambientes son óxicos, o bien, abundan otros aceptores de electrones, las relaciones que las bacterias sintróficas establecen no son necesarias<sup>198</sup>.

### Firmicutes y Archaeas; Sintrofia y Metanogénesis.

Todas las bacterias sintróficas son anaerobias estrictas, y están clasificadas metabólicamente, ya que pueden pertenecer tanto al phylum de los Firmicutes como al de las  $\delta$ -proteobacterias. En este proyecto, sólo se incluirán a los Firmicutes. El corazón de las reacciones sintróficas es la transferencia de  $H_2$  entre especies distintas cuya producción está ligada al consumo de  $H_2$  por su *socio*. El que consuma  $H_2$  puede ser una bacteria denitrificadora, una ferroreductora, sulfatorreductora, o bien acetógenos o metanógenas.

Metanogénesis es la producción biológica de  $CH_4$ , esto es un proceso prioritario en ambientes anóxicos y es llevado a cabo por un grupo importante de Archaeas, las metanógenas, que también son anaerobias estrictas. Las metanógenas en general, utilizan  $CO_2$  como el aceptor de electrones terminal para su respiración anaeróbica, reduciéndolo completamente hasta  $CH_4$ , con  $H_2$  como donador de electrones. Son muy pocos los sustratos que pueden convertirse a  $CH_4$  por las metanógenas, principalmente acetato. Para poder producir  $CH_4$  las metanógenas necesitan agruparse y coexistir con bacterias sintróficas, quienes pueden proveerlas de los precursores para la metanogénesis<sup>198</sup>, ver Figura 1.

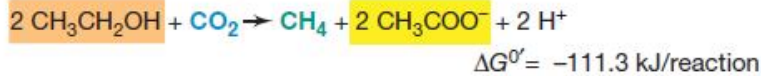
**Ethanol fermentation:**



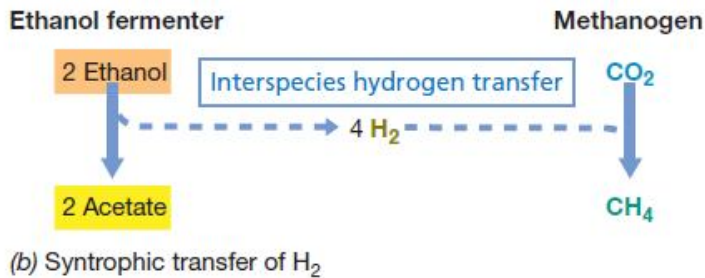
**Methanogenesis:**



**Coupled reaction:**



(a) Reactions



**Figure 14.9** Syntrophy: Interspecies H<sub>2</sub> transfer. Shown is the fermentation of ethanol to methane and acetate by syntrophic association of an ethanol-oxidizing syntroph and a H<sub>2</sub>-consuming partner (in this case, a methanogen). (a) Reactions involved. The two organisms share the energy released in the coupled reaction. (b) Nature of the syntrophic transfer of H<sub>2</sub>.

Figura 1. Tomada de 198

## Antecedentes

### La Hipótesis de la Reina Negra

Existen varias hipótesis para estudiar coevolución, una de ellas es la recientemente propuesta por Richard Lenski, *The Black Queen Hypothesis*<sup>199</sup>. Esta hipótesis explica cómo coevolucionan las comunidades microbianas, cómo interactúan entre ellas, cómo se estructuran para beneficiarse unas de las otras; pero más importantemente, explica la evolución reductiva. Es decir, qué beneficios puede traerle a un organismo la pérdida de ciertos genes.

La *Hipótesis de la Reina Negra* versa sobre cómo un microorganismo obtiene una mayor adecuación al perder ciertos genes. En pocas palabras, al coexistir distintos linajes bacterianos, éstos forman relaciones metabólicas entre sí. Aquellos capaces de evolucionar con una tasa más rápida, perderán genes que no son necesarios al coexistir con otros microorganismos en un nicho ecológico dado. A este grupo se le conoce como *beneficiarios*, los cuales, suelen perder genes o funciones biológicas que presentan para la célula un costo energético o nutricional alto. Por el contrario, aquellos compañeros que conservan esta función, conocidos como *ayudantes*, pueden permear esos metabolitos al ambiente, donde los beneficiarios pueden tomarlos<sup>199</sup>.

En una tierra primitiva y anóxica antes de que las cianobacterias llenaran de oxígeno el planeta, la simbiosis metabólica entre las metanógenas y las sintróficas debe de haber sido prevalente. Existe la hipótesis (*Syntrophy Hypothesis*) de que los organismos eucariotes tienen su origen en una relación sintrófica donde finalmente un microorganismo terminó por adentrar al otro, dado que las membranas de las sintróficas y las metanogénicas deben de estar en contacto para favorecer los intercambios de gas y no perder al volátil hidrógeno. Esto formaría consorcios estables, y en algunos casos, obligatorios<sup>200</sup>.

Independientemente de si la hipótesis sobre sintrofia sea o no cierta, la sintrofia (comiendo en compañía de) constituye un muy buen modelo para estudiar la evolución reductiva y probar la *Hipótesis de la Reina Negra*, donde los Firmicutes sintróficos serían los *ayudantes* y las Archaeas metanógenas serían los *beneficiarios*.

## Hipótesis

Las bacterias sintróficas (en particular los Firmicutes) siguen junto con las metanoarchaeas una estrecha relación coevolutiva que puede explicarse bajo la *Hipótesis de la Reina Negra*, lo cual lleva a la pérdida de genes de estas últimas.

## Objetivo

Generar conocimiento genómico global de los Firmicutes sintróficas y Archaeas metanógenas que nos permita explicar los papeles que han jugado mutuamente en su coevolución.

### Objetivo Particular

Desarrollar una metodología computacional que permita analizar los genomas de Firmicutes y de Archaeas, y hacer comparaciones de los genomas de interés contra sus parientes más cercanos, de modo que podamos discernir que funciones biológicas se han perdido en las Archaeas que sean aquellas que le pudieran estar permeando los Firmicutes.

## Metodología Propuesta

1. Se consideraran todos los genomas secuenciados de *Clostridias* y *Methanoarchaeas* para el análisis (Figura 2). A modo de control, se tomará un grupo de *Bacillus* aeróbicos y se comparará con las *Methanoarchaeas*. Resulta obvio que éstos *Bacillus* no establecen relaciones con estas Archaeas, pues ni siquiera pueden coexistir en un mismo nicho ecológico.
  - a. Los genomas considerados son los que están completamente secuenciados, en NCBI, anotados y ensamblados en su



totalidad. Estos son: 195 *Clostridias*, 105 *Bacillus* y 130 Metanoarchaeas.

2. Calcular tamaño de cada genoma, longitud de genes y de regiones intergénicas (¿existe una correlación de longitudes en genomas chicos y genomas grandes?). Esta información nos ayudaría a discernir entre evolución reductiva y otras posibles restricciones ambientales.
3. Para resolver los planteamientos anteriores, se utilizarán scripts en lenguaje Perl hechos *ad hoc*.
4. Realizar una reconstrucción metabólica de los genomas en cuestión y unos cuantos fuera de los grupos de interés para tener un punto de comparación. ¿Qué funciones han perdido las Archaeas y cuales han ganado/conservado los Firmicutes que les permita mantener estas relaciones sintróficas? ¿Qué otras relaciones metabólicas, además de la síntesis de metano, han conseguido realizar en conjunto? ¿En todos los aspectos metabólicos las Archaeas están jugando un papel de *beneficiarios* o en algunas reacciones los roles se invierten?
  - a. Para la reconstrucción metabólica se utilizará los mapas metabólicos de KEGG y la asignación de grupos de genes ortólogos COG por métodos hechos *ad hoc* usando Cadenas de Markov Escondidas (*Hidden Markov Models*).

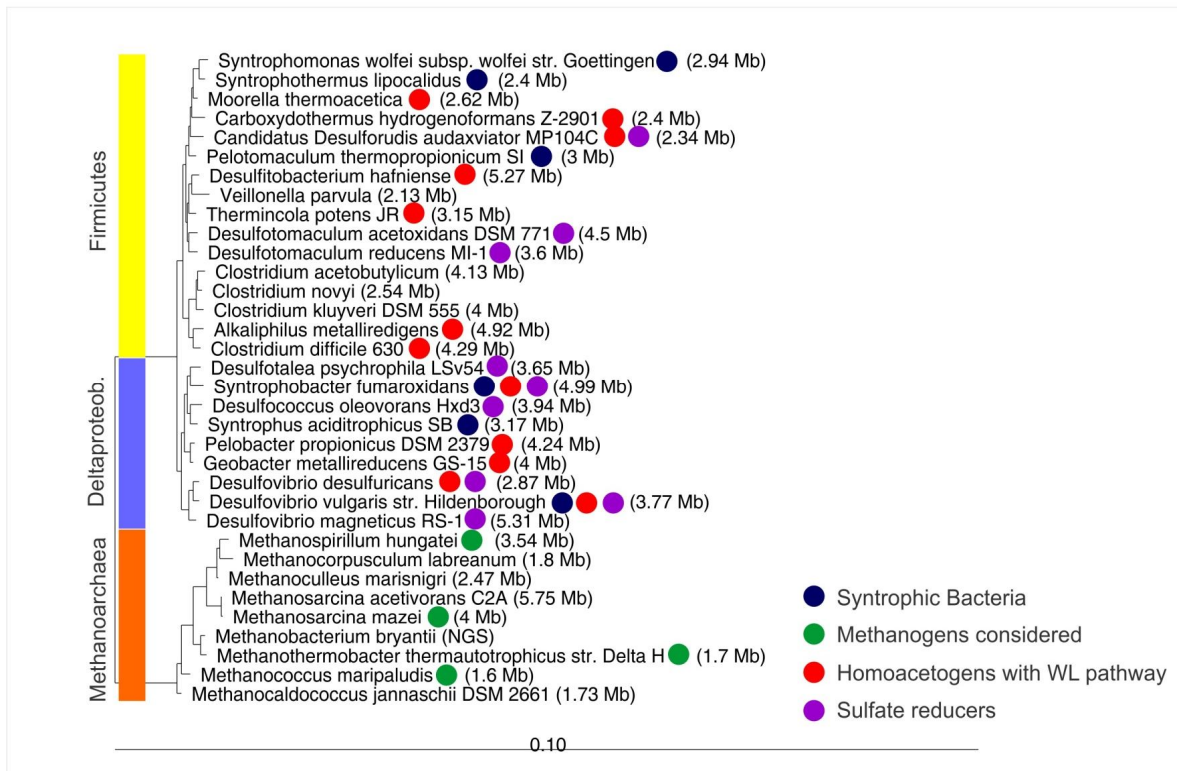


Figura 2. El árbol filogenético fue construido para los organismos que se considerarán en este proyecto. Se realizó en base a las distancias filogenéticas de las secuencias alineadas de 31 proteínas concatenadas en 191 especies<sup>120</sup>. Los alineamientos se generaron usando el programa MUSCLE<sup>121</sup>, y la reconstrucción filogenética se hizo mediante el programa PROTDIST del paquete PHYLIP (versión 3.57c; J. Felsenstein, University of Washington, Seattle).

## Resultados

No observo reducción en los genomas de los metanoarqueas (en algunos sí, pero en otros sucede todo lo contrario, parece que buscan acumular genes; Figura 3). Por esta razón, decidí ver si podía clasificarlas, ya sea como "helpers" o "beneficiaries" juzgándolas por su genoma. Un paréntesis es que algo que no me gusta de este artículo de Lenski<sup>199</sup> es que busca categorizar entre helpers y beneficiaries. Yo pienso que Lenski está viendo un proceso metabólico particular y no está viendo el panorama completo, donde algunos ayudantes podrían estarse beneficiando de los beneficiarios. Decidí empezar por las metanoarqueas (tengo 33 genomas) y construí dos pangenomas; uno de las metanoarqueas y otro de todas aquellas arqueas que no son metanógenas (éste último con el objetivo de más adelante ver cuáles son los genes que comparten todas que las metanoarqueas pudieran haber perdido... aún no llego a este paso de mi metodología).



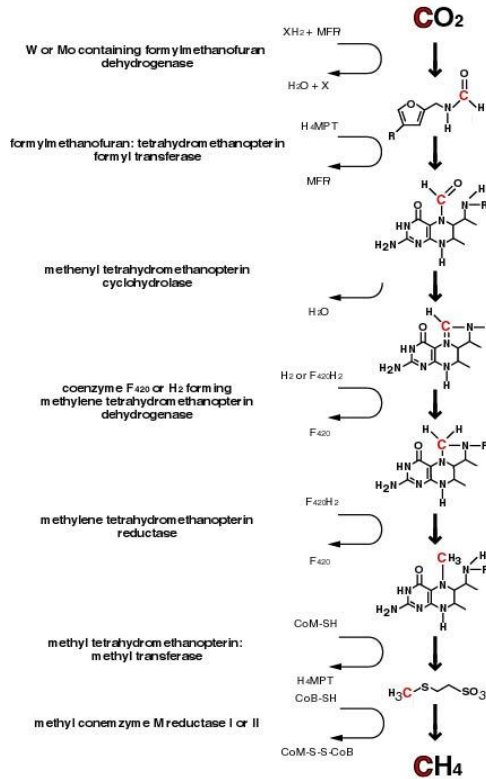


Figura 5. Vía metabólica para la metanogénesis.

5 son subunidades ( $\alpha$ ,  $\gamma$ , C, D y  $\beta$ ) de la Methyl CoM reductase (MCR), que cataliza la reducción de la metil-CoM ( $\text{CH}_3\text{-SCoM}$ ) y la coenzima B ( $\text{HS-CoB}$ ) para formar metano y consecuentemente  $\text{CoM-S-S-CoB}$  (2.8.4.1 from EC), que es el último paso en la metanogénesis.

Me quedan pues, 7 genes que son proteínas hipotéticas. La primera de estas proteínas únicas corresponde al COG1625, el cual está anotado como una proteína hipotética. En la mayoría de los genomas (23 de 33), este gen se encuentra monocistrónico y de forma divergente al operon que contiene los genes que codifican para la Methyl CoM reductase. Este contexto se presenta en 21 de 33 genomas. En aquellos genomas donde no se encuentra transcrito en forma monocistrónica, está asociado, en 3 genomas, con un dominio para el radical SAM

mientras que en el resto de los genomas está cotranscrito con genes distintos, excepto por otro genoma que se cotranscribe con un factor de maduración para una hidrogenasa.

Se realizaron análisis de predicción de estructura y modelado de proteínas. Las predicciones de compartimentos celulares, predicen que es una proteína citoplásmica (psipred y psort; los resultados pueden verse en: <http://zhanglab.ccmb.med.umich.edu/I-TASSER/output/S118083/> y en: <http://bioinf.cs.ucl.ac.uk/psipred/result/483166>). Esta proteína se puede modelar (ver los resultados de hhpred; aquí para predicciones locales: [http://toolkit.tuebingen.mpg.de/hhpred/results/COG1625\\_2](http://toolkit.tuebingen.mpg.de/hhpred/results/COG1625_2) y aquí para predicciones globales: [http://toolkit.tuebingen.mpg.de/hhpred/waiting/COG1625\\_3](http://toolkit.tuebingen.mpg.de/hhpred/waiting/COG1625_3)) obteniendo los mejores resultados con proteínas correspondientes a la biosíntesis del cofactor de Molibdeno (MoCo) y al parecer usa SAM, que es un sustrato para la transferencia de grupos metilo. La presencia de 7 motivos con residuos de cisteína son consistentes con que tiene centros de oxidorreducción con clusters de hierro y azufre.

Dado que la metanogénesis comienza con CO<sub>2</sub> y metanofurano por una enzima codificada por varias subunidades (Figura 5), la cual utiliza MoCo, y que la vía a partir de ahí sintetiza compuestos de tetrahydromethanopterin (THMPT) tiene sentido pensar que esta proteína (COG1625) une ya sea a MoCo o a THMPT. Es importante recalcar la similitud estructural entre MoCo y THMPT, las cuales comparten la "pterina" en su estructura. De esta forma existen varias posibilidades para pensar en una función a COG1625, una de ellas es que esta enzima ayude en la síntesis de MoCo para a su vez sintetizar metanofurano. Esta opción parece poco probable, ya que este gen sólo está presente en las metanógenas. Otra opción, es

que la enzima una THMPT (en lugar de MoCo) u otro compuesto con la estructura anillada "pterina" para otro fin, probablemente el de transferirle grupos metilo, dada su interacción con SAM.

Por esta razón pienso que es posible que la enzima funcione como una vía de "escape" de grupos metilos. Dada su transcripción divergente con el operón que codifica para la Methyl CoM reductase, es probable que estén anticorregulados. Por tanto, estos grupos metilo originados de la reducción de  $\text{CO}_2$  y unidos a THMPT, no terminarían en metano, si no que se transferirían de THMPT a SAM para ser posteriormente usados en otras vías de biosíntesis para el metabolismo celular. No estoy muy segura de que esto tenga sentido a un nivel metabólico global, pues la reducción de  $\text{CO}_2$  a  $\text{CH}_4$  es altamente exergónica (-130kJ), pero dada la orientación de las unidades transcripcionales, y su conservación en los genomas, parece razonable especular que ésta pudiera ser su función.

Otro de los genes que son únicos en las metanoarqueas y que no se conoce su función, es aquel que pertenece al COG4065. Esta familia de genes, se encuentra de forma monocistrónica en la mayoría de los genomas, mientras que en otros está asociado con pequeños ORFs. Sólo en dos genomas se encuentra cotranscrito con subunidades de la Tetrahydromethanopterin S-methyltransferase, lo cual sugiere fuertemente que está jugando un papel en metanogénesis. Sin embargo su estructura no puede ser modelada por lo que no sabemos qué forma pueda adoptar ni qué metabolitos, cofactores o sustratos pueda unir.

Los 5 genes restantes, que son únicos en metanoarqueas y que no se conoce su función, se encuentran cotranscritos en el mismo operón para 12 de los

33 genomas. Estos genes pertenecen a los COGs: COG4070, COG4029, COG4050, COG4051 y COG4052. Cuando estos genes están en el mismo operón se encuentran cotranscritos, ya sea con un ABC transporter ATP-binding protein (COG4555) o con un Activator of 2-hydroxyglutaryl-CoA dehydratase (HSP70-class ATPase domain) (COG1924).

De estos 5 genes, sólo el COG4050 puede ser modelado y su estructura se asemeja a aquella de una Alpha subunit 2-hydroxyisocaproyl-COA dehydratase; atypical dehydratase, lyase de *Clostridium difficile*. Es posible pensar que el COG4050 esté formando un polipéptido con aquel del COG1924, ya que ambas parecen ser subunidades de una dehidratasa.

COG4051: En base a similitud de secuencia distante, esta familia de genes, podría tentativamente, pertenecer a una familia de proteínas que unan ácidos nucleicos. Esta proteína no puede ser modelada.

COG4052: Tampoco puede ser modelada. Tiene similitud con la proteína C de Methyl-coenzyme M reductase



## Bibliografía

1. Karstrom, H. Enzymatische Adaptation bei Mikroorganism. *Ergeb. Enzymforsch.* **7**, 350–376 (1938).
2. Jacob, F. & Monod, J. Genetic Regulatory Mechanisms in the Synthesis of Proteins. *Journal of molecular biology* **3**, 318–356 (1961).
3. Monod, J. The growth of bacterial cultures. *Ann Rev Microbiol* **3**, 371–394 (1949).
4. Yanofsky, C. The Tryptophan Synthetase System. *Bacteriology Reviews* **24**, 221–245 (1960).
5. Yanofsky, C. & Ito, J. Nonsense codons and polarity in the tryptophan operon. *Journal of molecular biology* **21**, 313–34 (1966).
6. Yanofsky, C., Horn, V., Bonner, M. & Stasiowski, S. Polarity and Enzyme Functions in Mutants of the First Three Genes of the Tryptophan Operon of Escherichia coli. *Genetics* **69**, 49–433 (1971).
7. Matsushiro, A. Specialized transduction of tryptophan markers in Escherichia coli K12 by bacteriophage phi-80. *Virology* **19**, 475–482 (1963).
8. Morse, D., Mosteller, R. & Yanofsky, C. Dynamics of synthesis, translation, and degradation of trp operon messenger RNA in E. coli. *Cold Spring Harb Symp Quant Biol* **34**, 725–740 (1969).
9. Emmer, M., deCrombrughe, B., Pastan, I. & Perlman, R. Cyclic AMP receptor protein of E. coli: its role in the synthesis of inducible enzymes. *Proceedings of the National Academy of Sciences of the United States of America* **66**, 480–7 (1970).
10. Zubay, G., Schwartz, D. & Beckwith, J. Mechanism of activation of catabolite-sensitive genes: a positive control system. *Proceedings of the National Academy of Sciences of the United States of America* **66**, 104–10 (1970).
11. Englesberg, E., Irr, J., Power, J. & Lee, N. Positive Control of Enzyme Synthesis by Gene C in the Positive Control of Enzyme Synthesis by Gene C in the L-Arabinose System. *Journal of Bacteriology* **90**, 946–957 (1965).
12. Gross, J. & Englesberg, E. Determination of the order of mutational sites governing L-arabinose utilization in Escherichia coli B/r by transduction with phage P1bt. *Virology* **9**, 314–331 (1959).
13. Sheppard, D. E. & Englesberg, E. Further evidence for positive control of the L-arabinose system by gene araC. *Journal of molecular biology* **25**, 443–54 (1967).
14. Roth, J. R. *et al.* Transfer RNA and the Control of the Histidine Operon. *Cold Spring Harbor Symposia on Quantitative Biology* **31**, 383–392 (1966).

15. Baker, R. & Yanofsky, C. Transcription initiation frequency and translational yield for the tryptophan operon of *Escherichia coli*. *Journal of molecular biology* **69**, 89–102 (1972).
16. Imamoto, F. Translation and Transcription of the Tryptophan Operon. *Progress in Nucleic Acid Research and Molecular Biology* **13**, 339–407 (1973).
17. Kasai, T. Regulation of the expression of the histidine operon in *Salmonella typhimurium*. *Nature* **249**, 523–527 (1974).
18. Yanofsky, C. Attenuation in the control of expression of bacterial operons. *Nature* **289**, 751–758 (1981).
19. Yanofsky, C. The different roles of tryptophan transfer RNA in regulating trp operon expression in *E. coli* versus *B. subtilis*. *Trends in genetics: TIG* **20**, 367–74 (2004).
20. Platt, T. & Yanofsky, C. An Intercistronic Region and Ribosome-Binding Site in Bacterial Messenger RNA. *Proceedings of the National Academy of Sciences* **72**, 2399–2403 (1975).
21. Platt, T., Squires, C. & Yanofsky, C. Ribosome-protected regions in the leader-trpE sequence of *Escherichia coli* tryptophan operon messenger RNA. *Journal of molecular biology* **103**, 411–20 (1976).
22. Lee, F., Squires, C. L., Squires, C. & Yanofsky, C. Termination of transcription in vitro in the *Escherichia coli* tryptophan operon leader region. *Journal of molecular biology* **103**, 383–93 (1976).
23. Yanofsky, C. Transcription Attenuation: Once Viewed as a Novel Regulatory Strategy. *Journal of bacteriology* **182**, 1–8 (2000).
24. Henkin, T. M. & Yanofsky, C. Regulation by transcription attenuation in bacteria: how RNA provides instructions for transcription termination/antitermination decisions. *BioEssays: news and reviews in molecular, cellular and developmental biology* **24**, 700–7 (2002).
25. Henkin, T. M., Glass, B. L. & Grundy, F. J. Analysis of the *Bacillus subtilis* tyrS gene: conservation of a regulatory sequence in multiple tRNA synthetase genes. *Journal of bacteriology* **174**, 1299–306 (1992).
26. Grundy, F. J. & Henkin, T. M. tRNA as a positive regulator of transcription antitermination in *B. subtilis*. *Cell* **74**, 475–82 (1993).
27. Putney, S. D. & Schimmel, P. An aminoacyl tRNA synthetase binds to a specific DNA sequence and regulates its gene transcription. *Nature* **291**, 632–635 (1981).
28. Stringer, M. *et al.* *Escherichia coli* phenylalanyl-tRNA synthetase operon is controlled by attenuation in vivo. *Journal of molecular biology* **171**, 263–279 (1983).
29. Springer, M. *et al.* Translational Control in *E. coli*: the Case of Threonyl-tRNA Synthetase. *Bioscience Reports* **8**, 619–632 (1988).

30. Romby, P. *et al.* Molecular mimicry in translational control of *E. coli* threonyl-tRNA synthetase gene. Competitive inhibition in tRNA aminoacylation and operator-repressor recognition switch using tRNA identity rules. *Nucleic acids research* **20**, 5633–5640 (1992).
31. Grunberg-Manago, M. No Title. *Escherichia coli and Salmonella typhimurium: Cellular and Molecular Biology* 1386–1409 (1987).
32. Grundy, F. J. & Henkin, T. M. Conservation of a Transcription Antitermination Mechanism in Aminoacyl-tRNA Synthetase and Amino Acid Biosynthesis Genes in Gram-positive Bacteria. *Journal of molecular biology* **235**, 798–804 (1994).
33. Grundy, F. J., Rollins, S. M. & Henkin, T. M. Interaction between the acceptor end of tRNA and the T box stimulates antitermination in the *Bacillus subtilis* tyrS gene: a new role for the discriminator base. *Journal of bacteriology* **176**, 4518–26 (1994).
34. Henkin, T. M. tRNA-directed transcription antitermination. *Molecular microbiology* **13**, 381–387 (1994).
35. Yousef, M. R., Grundy, F. J. & Henkin, T. M. Structural transitions induced by the interaction between tRNA(Gly) and the *Bacillus subtilis* glyQS T box leader RNA. *Journal of molecular biology* **349**, 273–87 (2005).
36. Grundy, F. J. & Henkin, T. M. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Molecular microbiology* **30**, 737–49 (1998).
37. Mcdaniel, B. A. M., Grundy, F. J., Artsimovitch, I. & Henkin, T. M. Transcription termination control of the S box system: Direct measurement of S-adenosylmethionine by the leader RNA. *Proceedings of the National Academy of Sciences* **100**, 3083–3088 (2003).
38. Epshtein, V., Mironov, A. S. & Nudler, E. The riboswitch-mediated control of sulfur metabolism in bacteria. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5052–6 (2003).
39. Winkler, W. C., Nahvi, A., Sudarsan, N., Barrick, J. E. & Breaker, R. R. An mRNA structure that controls gene expression by binding S-adenosylmethionine. *Nature Structural Biology* **10**, 701–707 (2003).
40. Montange, R. K. & Batey, R. T. Structure of the S-adenosylmethionine riboswitch regulatory mRNA element. **441**, 10–13 (2006).
41. Abreu-goodger, C., Ontiveros-palacios, N., Ciria, R. & Merino, E. Conserved regulatory motifs in bacteria: riboswitches and beyond. **20**, 475–479 (2004).
42. Mandal, M. *et al.* Riboswitches Control Fundamental Biochemical Pathways in *Bacillus subtilis* and Other Bacteria. **113**, 577–586 (2003).

43. Mandal, M. & Breaker, R. R. Adenine riboswitches and gene activation by disruption of a transcription terminator. **11**, 29–35 (2004).
44. Batey, R. T., Gilbert, S. D. & Montange, R. K. Structure of a natural guanine- responsive riboswitch complexed with the metabolite hypoxanthine. **1562**, 27–31 (2004).
45. Kim, J. N., Roth, A. & Breaker, R. R. Guanine riboswitch variants from *Mesoplasma florum* selectively recognize 2'-deoxyguanosine. *Proceedings of the National Academy of Sciences of the United States of America* **104**, 16092–7 (2007).
46. Sudarsan, N. *et al.* Riboswitches in eubacteria sense the second messenger cyclic di-GMP. *Science (New York, N.Y.)* **321**, 411–3 (2008).
47. Roth, A. *et al.* A riboswitch selective for the queuosine precursor preQ1 contains an unusually small aptamer domain. *Nature structural & molecular biology* **14**, 308–17 (2007).
48. Kang, M., Peterson, R. & Feigon, J. Structural Insights into riboswitch control of the biosynthesis of queuosine, a modified nucleotide found in the anticodon of tRNA. *Molecular cell* **33**, 784–90 (2009).
49. Klein, D. J., Edwards, T. E. & Ferré-D'Amaré, A. R. Cocrystal structure of a class-I preQ1 riboswitch reveals a pseudoknot recognizing an essential hypermodified nucleobase. *Nature structural & molecular biology* **16**, 343–344 (2009).
50. Meyer, M. M., Roth, A., Chervin, S. M., Garcia, G. a & Breaker, R. R. Confirmation of a second natural preQ1 aptamer class in Streptococcaceae bacteria. *RNA (New York, N.Y.)* **14**, 685–95 (2008).
51. Watson, P. Y. & Fedor, M. J. The ydaO motif is an ATP-sensing riboswitch in *Bacillus subtilis*. *Nature chemical biology* **8**, 963–5 (2012).
52. Grundy, F. J., Lehman, S. C. & Henkin, T. M. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 12057–62 (2003).
53. Mandal, M. *et al.* A Glycine-Dependent Riboswitch That Uses Cooperative Binding to Control Gene Expression. **306**, 275–279 (2004).
54. Kwon, M. & Strobel, S. a Chemical basis of glycine riboswitch cooperativity. *RNA (New York, N.Y.)* **14**, 25–34 (2008).
55. Sherman, E. M., Esquiaqui, J., Elsayed, G. & Ye, J.-D. An energetically beneficial leader-linker interaction abolishes ligand-binding cooperativity in glycine riboswitches. *RNA (New York, N.Y.)* **18**, 496–507 (2012).
56. Fuchs, R. T., Grundy, F. J. & Henkin, T. M. The S MK box is a new SAM-binding RNA for translational regulation of SAM synthetase. **13**, 226–233 (2006).

57. Wang, J. X., Lee, E. R., Morales, D. R., Lim, J. & Breaker, R. R. Riboswitches that sense S-adenosylhomocysteine and activate genes involved in coenzyme recycling. *Molecular cell* **29**, 691–702 (2008).
58. Edwards, A. L., Reyes, F. E., Héroux, A. & Batey, R. T. Structural basis for recognition of S-adenosylhomocysteine by riboswitches. *RNA (New York, N.Y.)* **16**, 2144–55 (2010).
59. Winkler, W. C., Nahvi, A., Roth, A., Collins, J. A. & Breaker, R. R. Control of gene expression by a natural metabolite-responsive ribozyme. *Nature* **428**, 281–6 (2004).
60. Gelfand, M. S., Mironov, A. S., Jomantas, J., Kozlov, Y. I. & Perumov, D. A conserved RNA structure element involved in the regulation of bacterial riboflavin synthesis genes. *Trends in genetics: TIG* **15**, 439–42 (1999).
61. Mironov, A. S. *et al.* Sensing small molecules by nascent RNA: a mechanism to control transcription in bacteria. *Cell* **111**, 747–56 (2002).
62. Miranda-Rios, J., Navarro, M. & Sobero, M. A conserved RNA structure ( thi box ) is involved in regulation of thiamin biosynthetic gene. *Proceedings of the National Academy of Sciences* **98**, 9736–9741 (2001).
63. Miranda-Rios, J. The THI-box riboswitch, or how RNA binds thiamin pyrophosphate. *Structure (London, England: 1993)* **15**, 259–65 (2007).
64. Franklund, C. V & Kadner, R. J. Multiple transcribed elements control expression of the Escherichia coli btuB gene . Multiple Transcribed Elements Control Expression of the Escherichia coli btuB Gene. *Journal of bacteriology* **179**, 4039–4042 (1997).
65. Nahvi, A. *et al.* Genetic control by a metabolite binding mRNA. *Chemistry & biology* **9**, 1043 (2002).
66. Ames, T. D., Rodionov, D. A., Weinberg, Z. & Breaker, R. R. A eubacterial riboswitch class that senses the coenzyme tetrahydrofolate. *Chemistry & biology* **17**, 681–5 (2010).
67. Regulski, E. E. *et al.* A widespread riboswitch candidate that controls bacterial genes involved in molybdenum cofactor and tungsten cofactor metabolism. *Molecular microbiology* **68**, 918–32 (2008).
68. Groisman, E., Cromie, M., Shi, Y. & Latifi, T. A Mg<sup>2+</sup>-responding RNA that controls the expression of a Mg<sup>2+</sup> transporter. *Cold Spring Harb Symp Quant Biol.* **71**, 251–8 (2006).
69. Morita, M. T. *et al.* evidence for a built-in RNA thermosensor Translational induction of heat shock transcription factor 32: evidence for a built-in RNA thermosensor. 655–665 (1999).
70. Nechooshtan, G., Elgrably-Weiss, M., Sheaffer, A., Westhof, E. & Altuvia, S. A pH-responsive riboregulator. *Genes & development* **23**, 2650–62 (2009).

71. Henkin, T. M. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes & development* **22**, 3383–90 (2008).
72. Jacob, F., Perrin, D., Sanchez, C. & Monod, J. The Operon: A Group of Genes Whose Expression is Coordinated by an Operator. *C. R. Hebd. Seances Acad. Sci* **1729**, 1727–1729 (1960).
73. Bussemaker, H., Li, H. & Siggia, E. Regulatory element detection using a probabilistic segmentation model. *Proc Int Conf Intell Syst Mol Biol* **8**, 67–74 (2000).
74. McGuire, A. M., Hughes, J. D. & Church, G. M. Conservation of DNA Regulatory Motifs and Discovery of New Motifs in Microbial Genomes. *Genome Research* **10**, 744–757 (2000).
75. Van Helden, J., André, B. & Collado-Vides, J. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of molecular biology* **281**, 827–42 (1998).
76. Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* **95**, 14863–14868 (1998).
77. Keleş, S., Van der Laan, M. & Eisen, M. B. Identification of regulatory elements using a feature selection method. *Bioinformatics (Oxford, England)* **18**, 1167–75 (2002).
78. Lehman, N. RNA in evolution. *Wiley interdisciplinary reviews. RNA* **1**, 202–13 (2010).
79. Crick, F. H. C. The origin of the genetic code. *Journal of Molecular Biology* **38**, 367–379 (1968).
80. Orgel, L. E. Evolution of the genetic apparatus. *Journal of molecular biology* **38**, 381–93 (1968).
81. Woese, C. R., Dugre, D. H., Saxinger, W. C. & Dugre, S. A. The Molecular Basis for the Genetic Code. *Proceedings of the National Academy of Sciences* **55**, 966–974 (1966).
82. Kruger, K. *et al.* Self-splicing RNA: autoexcision and autocyclization of the ribosomal RNA intervening sequence of Tetrahymena. *Cell* **31**, 147–57 (1982).
83. Guerrier-Takada, C., Gardiner, K., Marsh, T., Pace, N. & Altman, S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. *Cell* **35**, 849–57 (1983).
84. Gesteland, R., Cech, T. & Atkins, J. *The RNA World*. (Cold Spring Harbor Laboratory Press: Cold Spring Harbor, NY, 2006).
85. Woese, C. R. *The Genetic Code: The Molecular Basis for Genetic Expression*. (Harper & Row: New York, NY, 1967).

86. Mizuno, T., Chou, M. Y. & Inouye, M. A unique mechanism regulating gene expression: translational inhibition by a complementary RNA transcript (micRNA). *Proceedings of the National Academy of Sciences of the United States of America* **81**, 1966–70 (1984).
87. Andersen, J. *et al.* The Isolation and characterization of RNA coded by the micF gene in *Escherichia coli*. *Nu* **15**, 2089–2101 (1987).
88. Gottesman, S. The small RNA regulators of *Escherichia coli*: roles and mechanisms\*. *Annual review of microbiology* **58**, 303–28 (2004).
89. Loh, E. *et al.* A trans-Acting Riboswitch Controls Expression of the Virulence Regulator PrfA in *Listeria monocytogenes*. *Cell* **139**, 770–779 (2009).
90. Breaker, R. R. Prospects for riboswitch discovery and analysis. *Molecular cell* **43**, 867–79 (2011).
91. Merino, E. & Yanofsky, C. Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends in genetics: TIG* **21**, 260–4 (2005).
92. Henkin, T. M. & Grundy, F. J. Sensing metabolic signals with nascent RNA transcripts: the T box and S box riboswitches as paradigms. *Cold Spring Harbor symposia on quantitative biology* **71**, 231–7 (2006).
93. Grundy, F. J., Winkler, W. C. & Henkin, T. M. tRNA-mediated transcription antitermination in vitro: codon-anticodon pairing independent of the ribosome. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 11121–6 (2002).
94. Grundy, F. J. & Henkin, T. M. THE T BOX AND S BOX TRANSCRIPTION TERMINATION CONTROL SYSTEMS. *Frontiers in Bioscience* **8**, d20–31 (2003).
95. Vitreschak, A. G., Mironov, A. A., Lyubetsky, V. A. & Gelfand, M. S. Comparative genomic analysis of T-box regulatory systems in bacteria. 717–735 (2008).doi:10.1261/rna.819308.negative
96. Gutiérrez-Preciado, A., Henkin, T. M., Grundy, F. J., Yanofsky, C. & Merino, E. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiology and molecular biology reviews MMBR* **73**, 36–61 (2009).
97. Gutiérrez-Preciado, A., Henkin, T. M., Grundy, F. J., Yanofsky, C. & Merino, E. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiology and molecular biology reviews: MMBR* **73**, 36–61 (2009).
98. Grigg, J. C. *et al.* T box RNA decodes both the information content and geometry of tRNA to affect gene expression. *Proceedings of the National Academy of Sciences* **110**, 7240–7245 (2013).
99. Narberhaus, F., Waldminghaus, T. & Chowdhury, S. RNA thermometers. *FEMS microbiology reviews* **30**, 3–16 (2006).



100. McDaniel, B. a, Grundy, F. J. & Henkin, T. M. A tertiary structural element in S box leader RNAs is required for S-adenosylmethionine-directed transcription termination. *Molecular microbiology* **57**, 1008–21 (2005).
101. Batey, R. T. Recognition of S-adenosylmethionine by riboswitches. *Wiley interdisciplinary reviews. RNA* **2**, 299–311 (2013).
102. Barrick, J. E. & Breaker, R. R. The distributions, mechanisms, and structures of metabolite-binding riboswitches. *Genome biology* **8**, R239 (2007).
103. Ueland, P. M. Pharmacological and Biochemical Aspects of S-adenosylhomocysteine and S-adenosylhomocysteine Hydrolase. *Pharmacological Reviews* **34**, 223–253 (1982).
104. Burge, S. W. *et al.* Rfam 11.0: 10 years of RNA families. *Nucleic acids research* **41**, D226–32 (2013).
105. Nawrocki, E. P., Kolbe, D. L. & Eddy, S. R. Infernal 1.0: inference of RNA alignments. **25**, 1335–1337 (2009).
106. Freyhult, E. K., Bollback, J. P. & Gardner, P. P. Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. *Genome Research* **17**, 117–125 (2007).
107. Nawrocki, E. P. & Eddy, S. R. Query-dependent banding (QDB) for faster RNA similarity searches. *PLoS computational biology* **3**, e56 (2007).
108. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proceedings of the Second International Conference on Intelligent Systems for Molecular Biology* 28–36 (1994).
109. Lawrence, C. & Reilly, A. An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences. *Proteins* **7**, 41–51 (1990).
110. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic acids research* **37**, W202–8 (2009).
111. Tompa, M. *et al.* Assessing computational tools for the discovery of transcription factor binding sites. *Nature biotechnology* **23**, 137–44 (2005).
112. Bailey, T. L. & Gribskov, M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics (Oxford, England)* **14**, 48–54 (1998).
113. Taboada, B., Verde, C. & Merino, E. High accuracy operon prediction method based on STRING database scores. *Nucleic acids research* **38**, e130 (2010).
114. Abreu-goodger, C. & Merino, E. RibEx: a web server for locating riboswitches and other conserved bacterial regulatory elements. **33**, 690–692 (2005).



115. Eddy, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC bioinformatics* **3**, 18 (2002).
116. Ciria, R., Abreu-Goodger, C., Morett, E. & Merino, E. GeConT: gene context analysis. *Bioinformatics (Oxford, England)* **20**, 2307–8 (2004).
117. Martinez-Guerrero, C. E., Ciria, R., Abreu-Goodger, C., Moreno-Hagelsieb, G. & Merino, E. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic acids research* **36**, W176–80 (2008).
118. Tatusov, R. L. A Genomic Perspective on Protein Families. *Science* **278**, 631–637 (1997).
119. Ogata, H. *et al.* KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic acids research* **27**, 29–34 (1999).
120. Ciccarelli, F. D. *et al.* Toward automatic reconstruction of a highly resolved tree of life. *Science (New York, N.Y.)* **311**, 1283–7 (2006).
121. Edgar, R. C. MUSCLE User Guide. **32**, 1–15 (2004).
122. Woese, C. R., Olsen, G. J., Ibba, M. & Söll, D. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiology and molecular biology reviews: MMBR* **64**, 202–36 (2000).
123. Wolf, Y. I., Aravind, L., Grishin, N. V & Koonin, E. V Evolution of aminoacyl-tRNA synthetases--analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome research* **9**, 689–710 (1999).
124. Saizieu, A. De, Vankan, P., Vockler, C. & Loon, A. P. G. M. Van RNA-binding attenuation protein (TRAP) regulates the steady-state levels of transcripts of the *Bacillus subtilis* folate operon. *Microbiology* **143**, 979–989 (1997).
125. Putzer, H., Gendron, N. & Grunberg-Manago, M. Co-ordinate expression of the two threonyl-tRNA synthetase genes in *Bacillus subtilis*: control by transcriptional antitermination involving a conserved regulatory sequence. *The EMBO journal* **11**, 3117–27 (1992).
126. Gendron, N., Putzer, H. & Grunberg-Manago, M. Expression of both *Bacillus subtilis* threonyl-tRNA synthetase genes is autogenously regulated. *Journal of bacteriology* **176**, 486–94 (1994).
127. Gutierrez-Preciado, a, Jensen, R. a, Yanofsky, C. & Merino, E. New insights into regulation of the tryptophan biosynthetic operon in Gram-positive bacteria. *Trends in genetics: TIG* **21**, 432–6 (2005).
128. Gutiérrez-Preciado, A., Yanofsky, C. & Merino, E. Comparison of tryptophan biosynthetic operon regulation in different Gram-positive bacterial species. *Trends in Genetics* **23**, 422–426 (2007).

129. Yanofsky, C. Transcription Attenuation. *J Biol Chem* **263**, 609–612 (1988).
130. Even, S. *et al.* Global Control of Cysteine Metabolism by CymR in *Bacillus subtilis*. *Journal of bacteriology* **188**, 2184–2197 (2006).
131. Shivers, R. P. & Sonenshein, A. L. *Bacillus subtilis* *ilvB* operon: an intersection of global regulons. **56**, 1549–1559 (2005).
132. Rollins, S. M. The mRNA/tRNA interaction promoting T box transcriptional antitermination. (2002).
133. Brown, D. P., Krishnamurthy, N. & Sjo, K. Automated Protein Subfamily Identification and Classification. *PLoS computational biology* **3**, 1526–1538 (2007).
134. Xie, G., Keyhani, N. O., Bonner, C. A. & Jensen, R. A. Ancient Origin of the Tryptophan Operon and the Dynamics of Evolutionary Change †. **67**, 303–342 (2003).
135. Michel, G. *et al.* Structures of shikimate dehydrogenase AroE and its Paralog YdiB. A common structural framework for different activities. *The Journal of biological chemistry* **278**, 19463–72 (2003).
136. Gutiérrez-Preciado, A. & Merino, E. Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach. *Clinical microbiology and infection the official publication of the European Society of Clinical Microbiology and Infectious Diseases* **18 Suppl 4**, 35–9 (2012).
137. Webb, M. E., Smith, A. G. & Abell, C. Biosynthesis of pantothenate. *Natural product reports* **21**, 695–721 (2004).
138. Liesegang, H. *et al.* The genome of *Clostridium kluyveri*, a strict anaerobe. (2008).
139. Ikeda, T. *et al.* Anabolic five subunit-type pyruvate:ferredoxin oxidoreductase from *Hydrogenobacter thermophilus* TK-6. *Biochemical and biophysical research communications* **340**, 76–82 (2006).
140. Voet, D. & Voet, J. G. Biochemistry. *Biochemistry* 764–776 (1995).
141. Shivers, R. P. & Sonenshein, A. L. Activation of the *Bacillus subtilis* global regulator CodY by direct interaction with branched-chain amino acids. **53**, 599–611 (2004).
142. Henner, D. J. & Yanofsky, C. No Title. *Bacillus subtilis and other Gram positive bacteria: biochemistry, physiology and molecular genetics* 269–280 (1993).
143. Gollnick, P., Babitzke, P., Merino, E. & Yanofsky, C. No Title. *Bacillus subtilis and its closest relatives: from genes to cells*, 233–244 (2002).

144. Korbelt, J. O., Jensen, L. J., Von Mering, C. & Bork, P. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature biotechnology* **22**, 911–7 (2004).
145. Souza, V. *et al.* From the Cover: An endangered oasis of aquatic microbial biodiversity in the Chihuahuan desert. *Proceedings of the National Academy of Sciences* **103**, 6565–6570 (2006).
146. Alcaraz, L. D. *et al.* The genome of *Bacillus coahuilensis* reveals adaptations essential for survival in the relic of an ancient marine environment. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 5803–5808 (2008).
147. Gelfand, M. S. Evolution of transcriptional regulatory networks in microbial genomes. 420–429 (2006).doi:10.1016/j.sbi.2006.04.001
148. Zhang, Y., Romero, H., Salinas, G. & Gladyshev, V. N. Dynamic evolution of selenocysteine utilization in bacteria: a balance between selenoprotein loss and evolution of selenocysteine from redox active cysteine residues. *Genome biology* **7**, R94 (2006).
149. Brinza, L., Calevro, F. & Charles, H. Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of *Buchnera*. *BMC genomics* **14**, 73 (2013).
150. Angert, E. R., Clements, K. D. & Pace, N. R. The largest bacterium. *Nature* **362**, 239–41 (1993).
151. Miller, D. a, Suen, G., Clements, K. D. & Angert, E. R. The genomic basis for the evolution of a novel form of cellular reproduction in the bacterium *Epulopiscium*. *BMC genomics* **13**, 265 (2012).
152. Mendell, J. E., Clements, K. D., Choat, J. H. & Angert, E. R. Extreme polyploidy in a large bacterium. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 6730–4 (2008).
153. Putzer, H., Laalami, S., Brakhage, a a, Condon, C. & Grunberg-Manago, M. Aminoacyl-tRNA synthetase gene regulation in *Bacillus subtilis*: induction, repression and growth-rate regulation. *Molecular microbiology* **16**, 709–18 (1995).
154. Marta, P. T., Ladner, R. D. & Grandoni, J. A. A CUC Triplet Confers Leucine-Dependent Regulation of the *Bacillus subtilis* *ilv-leu* Operon. *Journal of bacteriology* **178**, 2150–2153 (1996).
155. Luo, D. *et al.* Structure and regulation of expression of the *Bacillus subtilis* *valyl*-tRNA synthetase gene. *Journal of bacteriology* **179**, 2472–2478 (1997).
156. Grundy, F. J., Hodil, S. E., Rollins, S. M. & Henkin, T. M. Specificity of tRNA-mRNA interactions in *Bacillus subtilis* *tyrS* antitermination. *Journal of bacteriology* **179**, 2587–2594 (1997).

157. Guchte, M. Van De, Ehrlich, S. D. & Chopin, A. Identity elements in tRNA-mediated transcription antitermination: implication of tRNA D- and T-arms in mRNA recognition. *Microbiology* **147**, 1223–1233 (2001).
158. Grundy, F. J. *et al.* tRNA determinants for transcription antitermination of the *Bacillus subtilis* tyrS gene. *RNA (New York, N.Y.)* **6**, 1131–1141 (2000).
159. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Comparative genomics of the methionine metabolism in Gram-positive bacteria: a variety of regulatory systems. **32**, 3340–3353 (2004).
160. Darwin, C. *The Origin of Species*. (Gramercy: 1995).at  
<<http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/0517123207>>
161. Rainey, P., Buckling, a, Kassen, R. & Travisano, M. The emergence and maintenance of diversity: insights from experimental bacterial populations. *Trends in ecology & evolution* **15**, 243–247 (2000).
162. Grundy, F. J. & Henkin, T. M. Kinetic Analysis of tRNA-Directed Transcription Antitermination of the *Bacillus subtilis* glyQS Gene In Vitro. **186**, 5392–5399 (2004).
163. Lenski, R. E. & Travisano, M. Dynamics of adaptation and diversification: a 10,000-generation experiment with bacterial populations. *Proceedings of the National Academy of Sciences of the United States of America* **91**, 6808–14 (1994).
164. Lenski, R. E. *et al.* Evolution of competitive fitness in experimental populations of *E. coli*: what makes one genotype a better competitor than another? *Antonie van Leeuwenhoek* **73**, 35–47 (1998).
165. Quance, M. a & Travisano, M. Effects of temperature on the fitness cost of resistance to bacteriophage T4 in *Escherichia coli*. *Evolution; international journal of organic evolution* **63**, 1406–16 (2009).
166. Souza, V., Travisano, M., Turner, P. E. & Eguiarte, L. E. Does experimental evolution reflect patterns in natural populations? *E. coli* strains from long-term studies compared with wild isolates. *Antonie van Leeuwenhoek* **81**, 143–53 (2002).
167. Rodionov, D. A., Vitreschak, A. G., Mironov, A. A. & Gelfand, M. S. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch ? **31**, 6748–6757 (2003).
168. Gollnick, P., Babitzke, P., Antson, A. & Yanofsky, C. Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annual review of genetics* **39**, 47–68 (2005).
169. Valbuzzi, a & Yanofsky, C. Inhibition of the *B. subtilis* regulatory protein TRAP by the TRAP-inhibitory protein, AT. *Science (New York, N.Y.)* **293**, 2057–9 (2001).

170. Sarsero, J. P., Merino, E. & Yanofsky, C. A *Bacillus subtilis* operon containing genes of unknown function senses tRNA Trp charging and regulates expression of the genes of tryptophan biosynthesis. **97**, (2000).
171. Yanofsky, C. Using studies on tryptophan metabolism to answer basic biological questions. *The Journal of biological chemistry* **278**, 10859–78 (2003).
172. Gutiérrez-Preciado, A., Romero, H. & Peimbert, M. An Evolutionary Perspective on Amino Acids. *Nature Education* **3**, 29 (2010).
173. Gollnick, P. & Babitzke, P. Transcription attenuation. *Biochimica et biophysica acta* **1577**, 240–50 (2002).
174. Landick, R., Turnbough, C. L. J. & Yanofsky, C. *Escherichia coli and Salmonella: cellular and molecular biology*. 1263–1286 (Washington,DC, 1996).
175. Kuroda, M. I. & Yanofsky, C. Evidence for the transcript secondary structures predicted to regulate transcription attenuation in the trp operon. *The Journal of biological chemistry* **259**, 12838–43 (1984).
176. Otwinowski, Z. *et al.* Crystal structure of trp repressor/operator complex at atomic resolution. *Nature* **335**, 321–329 (1988).
177. Zhang, H. *et al.* The Solution Structure of the trp Repressor-Operator DNA Complex. *Journal of molecular biology* **238**, 592–614 (1994).
178. Chen, G. & Yanofsky, C. Features of a leader peptide coding region that regulate translation initiation for the anti-TRAP protein of *B. subtilis*. *Molecular cell* **13**, 703–11 (2004).
179. Slock, J., Stahly, D. P., Han, C. Y., Six, E. W. & Crawford, I. P. An apparent *Bacillus subtilis* folic acid biosynthetic operon containing *pab*, an amphibolic *trpG* gene, a third gene required for synthesis of para-aminobenzoic acid, and the dihydropteroate synthase gene. *Journal of bacteriology* **172**, 7211–26 (1990).
180. Chen, G. & Yanofsky, C. Tandem transcription and translation regulatory sensing of uncharged tryptophan tRNA. *Science (New York, N.Y.)* **301**, 211–3 (2003).
181. Yanofsky, C., Kelley, R. L. & Horn, V. Repression is relieved before attenuation in the trp operon of *Escherichia coli* as tryptophan starvation becomes increasingly severe. *Journal of bacteriology* **158**, 1018–1024 (1984).
182. Berka, R. M., Cui, X. & Yanofsky, C. Genomewide transcriptional changes associated with genetic alterations and nutritional supplementation affecting tryptophan metabolism in *Bacillus subtilis*. *Proceedings of the National Academy of Sciences of the United States of America* **100**, 5682–7 (2003).

183. Babitzke, P. & Gollnick, P. Posttranscription Initiation Control of Tryptophan Metabolism in *Bacillus subtilis* by the trp RNA-Binding Attenuation Protein (TRAP), anti-TRAP, and RNA Structure. *Journal of bacteriology* **183**, 5795–5802 (2001).
184. Babitzke, P. Regulation of transcription attenuation and translation initiation by allosteric control of an RNA-binding protein: the *Bacillus subtilis* TRAP protein. *Current opinion in microbiology* **7**, 132–9 (2004).
185. Antson, A. A. *et al.* The structure of trp RNA-binding attenuation protein. *Nature* **374**, 693–700 (1995).
186. Antson, a a *et al.* Structure of the trp RNA-binding attenuation protein, TRAP, bound to RNA. *Nature* **401**, 235–42 (1999).
187. McElroy, C., Manfredo, A., Wendt, A., Gollnick, P. & Foster, M. TROSY-NMR Studies of the 91kDa TRAP Protein Reveal Allosteric Control of a Gene Regulatory Protein by Ligand-altered Flexibility. *Journal of Molecular Biology* **323**, 463–473 (2002).
188. Du, H., Yakhnin, A. V, Dharmaraj, S. & Babitzke, P. trp RNA-Binding Attenuation Protein-5' Stem-Loop RNA Interaction Is Required for Proper Transcription Attenuation Control of the *Bacillus subtilis* trpEDCFBAOperon trp RNA-Binding Attenuation Protein-5J Stem-Loop RNA Interaction Is Required for Proper Tra. (2000).doi:10.1128/JB.182.7.1819-1827.2000.Updated
189. Merino, E., Babitzke, P. & Yanofsky, C. trp RNA-binding attenuation protein (TRAP)-trp leader RNA interactions mediate translational as well as transcriptional regulation of the *Bacillus subtilis* trp operon. *Journal of Applied Entomology* **177**, 6362–6370 (1995).
190. Du, H. & Babitzke, P. trp RNA-binding attenuation protein-mediated long distance RNA refolding regulates translation of trpE in *Bacillus subtilis*. *The Journal of biological chemistry* **273**, 20494–503 (1998).
191. Yang, M., Saizieu, A. De, Loon, A. P. Van & Gollnick, P. Translation of trpG in *Bacillus subtilis* is regulated by the trp RNA-binding attenuation protein (TRAP). *Journal of bacteriology* **177**, 4272–4278 (1995).
192. Du, H., Tarpey, R., Babitzke, P. & Babitzke, P. The trp RNA-binding attenuation protein regulates TrpG synthesis by binding to the trpG ribosome binding site of *Bacillus subtilis* . The trp RNA-Binding Attenuation Protein Regulates TrpG Synthesis by Binding to the trpG Ribosome Binding Site of *Bacillus* . **179**, (1997).
193. Steinberg, W. Temperature-Induced Derepression of Tryptophan Biosynthesis in a Tryptophanyl-Transfer Ribonucleic Acid Synthetase Mutant of *Bacillus subtilis*. *Journal of Applied Entomology* **117**, 1023–1034 (1974).
194. Lee, A. I., Sarsero, J. P. & Yanofsky, C. A temperature-sensitive trpS mutation interferes with trp RNA-binding attenuation protein (TRAP) regulation of trp gene expression in *Bacillus subtilis*. *Journal of bacteriology* **178**, 6518–6524 (1996).

195. Valbuzzi, A. & Yanofsky, C. Zinc is required for assembly and function of the anti-trp RNA-binding attenuation protein, AT. *The Journal of biological chemistry* **277**, 48574–8 (2002).
196. Valbuzzi, A., Gollnick, P., Babitzke, P. & Yanofsky, C. The anti-trp RNA-binding attenuation protein (Anti-TRAP), AT, recognizes the tryptophan-activated RNA binding domain of the TRAP regulatory protein. *The Journal of biological chemistry* **277**, 10608–13 (2002).
197. Hall, C. V & Yanofsky, C. Regulation of tryptophanyl-tRNA synthetase formation. *Journal of bacteriology* **151**, 918–23 (1982).
198. Madigan, M. T., Martinko, J. M., Stahl, D. & Clark, D. P. *Brock Biology of Microorganisms (13th Edition)*. (Benjamin Cummings: 2010).at  
<<http://www.amazon.com/exec/obidos/redirect?tag=citeulike07-20&path=ASIN/032164963X>>
199. Morris, J. J., Lenski, R. E. & Zinser, E. R. The Black Queen Hypothesis: Evolution of Dependencies through Adaptive Gene Loss. *mBio* **3**, 36–12 (2012).
200. Moreira, D. & Lopez-Garcia, P. Symbiosis between methanogenic archaea and delta-proteobacteria as the origin of eukaryotes: the syntrophic hypothesis. *Journal of molecular evolution* **47**, 517–30 (1998).
201. Crooks, G. E., Hon, G., Chandonia, J.-M. & Brenner, S. E. WebLogo: a sequence logo generator. *Genome research* **14**, 1188–90 (2004).



# Biochemical Features and Functional Implications of the RNA-Based T-Box Regulatory Mechanism

Ana Gutiérrez-Preciado,<sup>1</sup> Tina M. Henkin,<sup>2</sup> Frank J. Grundy,<sup>2</sup>  
Charles Yanofsky,<sup>3</sup> and Enrique Merino<sup>1\*</sup>

Department of Molecular Microbiology, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos 62210, México<sup>1</sup>; Department of Microbiology and Center for RNA Biology, The Ohio State University, Columbus, Ohio 43210-1292<sup>2</sup>; and Department of Biological Sciences, Stanford University, Stanford, California 94305-5020<sup>3</sup>

INTRODUCTION .....	36
GENERAL FEATURES OF T-BOX RIBOSWITCH RNA.....	37
AMINOACYL-tRNA SYNTHETASE GENES .....	39
T-BOX REGULATION OF AMINO ACID BIOSYNTHETIC GENES.....	43
Regulation of Serine and Glycine Biosynthetic Genes by the T-Box and <i>gcvT</i> Riboswitches .....	44
Pathways for Synthesis of the Sulfur-Containing Amino Acids Methionine and Cysteine Are Regulated by S-Box and T-Box Riboswitches .....	44
The Branched-Chain Amino Acids Isoleucine, Leucine, and Valine and Their Relationship to the Pantothenate Pathway .....	46
Histidine Biosynthesis: Possible Consequences of a Weak tRNA-T-Box Interaction .....	49
Aromatic Amino Acid Biosynthesis: Prediction of Tight Regulation by Tandem T Boxes.....	49
Biosynthetic Genes for Aspartate and Asparagine, Key Precursors of Many Other Amino Acids .....	51
Alanine Biosynthesis Involves T-Box Regulation of Operons Containing Biosynthetic and Regulatory Genes .....	52
Threonine Biosynthesis .....	52
Proline Biosynthesis .....	52
Regulation of Arginine Biosynthesis in the <i>Firmicutes</i> Is Mediated Predominantly by a DNA Binding Transcriptional Repressor Protein.....	53
Amino Acid Biosynthetic Pathway Genes That Are Not Regulated by the T-Box Mechanism.....	53
REGULATION OF AMINO ACID TRANSPORTER GENES .....	53
Shared Regulatory Mechanisms for Biosynthetic and Transporter Genes .....	53
REGULATION OF SYNTHESIS OF REGULATORY PROTEINS .....	55
OTHER IMPORTANT FEATURES OF THE T-BOX MECHANISM.....	56
<i>ileS</i> Is the Gene Most Widely Regulated by the T-Box Mechanism .....	56
Over- and Underrepresentation of T-Box Regions in Genomes .....	56
Single versus Tandem T-Box Elements .....	56
T-Box Sequences Containing a tRNA Gene .....	56
EVOLUTIONARY ORIGIN OF T-BOX ELEMENTS.....	57
EXPECTED INSIGHTS ON T-BOX REGULATION FROM ANALYSES OF NEW GENOME SEQUENCES.....	57
CONCLUSIONS .....	57
ACKNOWLEDGMENTS .....	58
REFERENCES .....	58

## INTRODUCTION

The regulation of bacterial gene expression is often based on RNA recognition of an appropriate signal. The term “riboswitch” has been used to describe *cis*-acting RNA regulatory elements that undergo significant structural shifts in response to a specific regulatory signal. This recognition occurs in the absence of the action of an RNA binding protein or a translating ribosome. The shift in riboswitch structure regulates the expression of RNA sequences located downstream on that RNA. Riboswitch RNAs regulate the expression of genes en-

coding proteins with a broad range of functions in a variety of bacterial species. These regulated genes include those specifying enzymes concerned with the charging of amino acids onto tRNAs, the synthesis or transport of amino acids, and the synthesis of cofactors, nucleotides, and metal ions (55).

The T-box family of riboswitches commonly modulates the expression of many genes concerned with amino acid metabolism in gram-positive bacteria, especially members of the family *Firmicutes*. The T-box mechanism utilizes uncharged tRNA as the signal molecule. Genes in this family exhibit a conservation of a set of sequence and structural features in the untranslated “leader region” of the RNA upstream of the regulated coding sequences (43, 58). For most operons in the T-box family, segments of these upstream leader RNAs can fold to form either of two alternative hairpin structures, an intrinsic transcription terminator or a competing transcription antiter-

\* Corresponding author. Mailing address: Instituto de Biotecnología, UNAM, Av. Universidad #2001, Col. Chamilpa, CP 62210 Cuernavaca, Morelos, México. Phone: 52 777 329 1634. Fax: 52 777 313 8673. E-mail: merino@ibt.unam.mx.



minator. The formation of the terminator hairpin results in a premature termination of transcription, reducing the transcription of the downstream coding region(s). In each T-box-regulated operon, the proper pairing of an appropriate uncharged tRNA with the leader RNA promotes the stabilization of the alternate antiterminator structure. This pairing prevents the formation of the terminator, which allows continued transcription into the downstream gene or genes of the operon (43, 56). The specific recognition of the cognate uncharged tRNA and the tRNA-directed formation of a transcription antiterminator can occur in the absence of any other cellular factor(s) (50). Like other riboswitches in which a leader RNA transcript senses a signal in regulating downstream gene expression, T-box RNAs can also control translation initiation. In transcripts with this capability, the terminator helix is replaced by a helix that can sequester the Shine-Dalgarno (SD) sequence of a downstream coding region, thereby inhibiting the initiation of translation rather than prematurely terminating transcription. T-box RNAs that regulate transcription termination are most commonly observed in low-G+C gram-positive bacteria, while T-box RNAs that control translation initiation predominate in high-G+C gram-positive bacteria and in gram-negative bacteria (organisms in which T-box-mediated regulation is less common) (47, 107).

The common feature of the T-box mechanism and other riboswitch mechanisms is the ability of a leader RNA to directly sense a signal molecule (an uncharged tRNA or some other specific molecule), resulting in a rearrangement of the leader RNA that determines whether the downstream coding sequences will be expressed. The T-box mechanism is particularly well suited for regulating the expression of genes encoding proteins involved in the aminoacylation of tRNA and amino acid biosynthesis and in amino acid transport. Aminoacylated tRNA is also sensed in some bacterial species using different RNA regulatory strategies. For example, in *Escherichia coli*, the *trp* transcription attenuation mechanism is used; in this mechanism, the translation of a leader peptide coding region modulates the formation of alternative leader RNA structures, determining whether or not transcription termination will occur in the leader region of the *trp* operon (117, 118). The major difference between the *E. coli trp* transcription attenuation mechanism and the T-box riboswitch mechanism is that tRNA charging is sensed indirectly by a translating ribosome in the former mechanism, compared to the direct binding of the tRNA to the leader RNA in the T-box mechanism.

The high level of conservation of the sequences and structural features of T-box leader RNAs permits the identification of many genes that are likely to be regulated by this mechanism. This information, in conjunction with the identification of the products of the downstream coding sequences, permits prediction of the specificity of their regulatory responses. Recently, Vitreschak et al. (107) reported the identification of 805 T-box leader sequences in 96 partial and completely sequenced bacterial genomes. In their paper, they provided an overview of the use of this regulatory strategy and discussed the evolutionary relationships of T-box leader sequences with regard to the origin of this highly conserved regulatory mechanism. Here, we report genomic analyses with a larger set of genomes using a different set of parameters. We identified 1,111 T-box leader sequences in 87 completely sequenced bacterial genomes; 472

of the genomes examined did not contain an identifiable T-box-controlled gene. This analysis of T-box elements allowed us to identify arrangements of this regulatory element that differ from its standard mode. We will discuss the predicted physiological roles of the identified T-box-regulated genes and the functional implications of T-box-mediated regulation for the functioning of the corresponding metabolic pathways. We will also discuss the distribution of T-box versus other regulatory mechanisms for genes concerned with different classes of amino acids. In addition to citing examples supporting the previously described role of T-box elements in regulating the expression of individual genes or operons, we predict that T-box-mediated regulation controls the synthesis of regulatory proteins that in turn regulate additional sets of genes. These features increase the overall impact of the T-box mechanism in modulating various cellular activities.

### GENERAL FEATURES OF T-BOX RIBOSWITCH RNA

A T-box RNA consists of a segment of leader RNA with conserved features that allow recognition of, and pairing with, a specific uncharged tRNA (Fig. 1). These features include the capacity of the leader RNA to form alternative secondary structures, one of which can serve as an intrinsic transcription terminator or as an anti-SD (ASD) helix that pairs with an SD sequence, blocking translation initiation. The major structures formed within the T-box RNA, in addition to the segments that can form the terminator/antiterminator (or ASD/anti-ASD) elements, are designated stem I, stem II, the stem IIA/stem IIB pseudoknot, and stem III (47, 88) (Fig. 1A). The sequestration of sequences that form the 5' strand of the terminator (or the ASD helix) into a competing antiterminator structure (or anti-ASD helix) allows the transcription or translation of the downstream coding sequence. In each T-box RNA, the terminator (or ASD helix) is predicted to be more stable than the competing antiterminator structure; therefore, binding of uncharged tRNA is required to stabilize the competing antiterminator (or anti-ASD helix) structure.

Each T-box RNA is presumably the result of evolutionary selection, preparing it to respond to a specific uncharged tRNA. Specific tRNA binding requires the pairing of the tRNA anticodon with a single codon sequence designated the "specifier sequence" in the T-box RNA (Fig. 1A). The specifier sequence is positioned at a discrete location in the specifier loop within stem I (Fig. 1A). Due to the redundancy of the genetic code, each of the amino acids, with the exception of Met and Trp, is charged onto more than one tRNA species. In theory, any of these tRNA species could be sensed by the T-box mechanism. However, initial studies of T-box leader sequences (121), in agreement with results described previously by Vitreschak et al. (107) and our analyses of T-box elements in fully sequenced genomes, exhibited a strong bias toward the presence of a C in the third position of the specifier sequence of most T-box RNAs. This preference is not influenced by the codon usage or the tRNA abundance in a specific organism. For example, glycine (Gly)-specific T-box RNAs exhibit a strong preference for the GGC codon. This is the least used Gly codon in *Bacillus halodurans* but the most often used Gly codon in *Bacillus clausii*. Mutational studies of the *Bacillus subtilis tyrS* gene demonstrated that the replacement of the

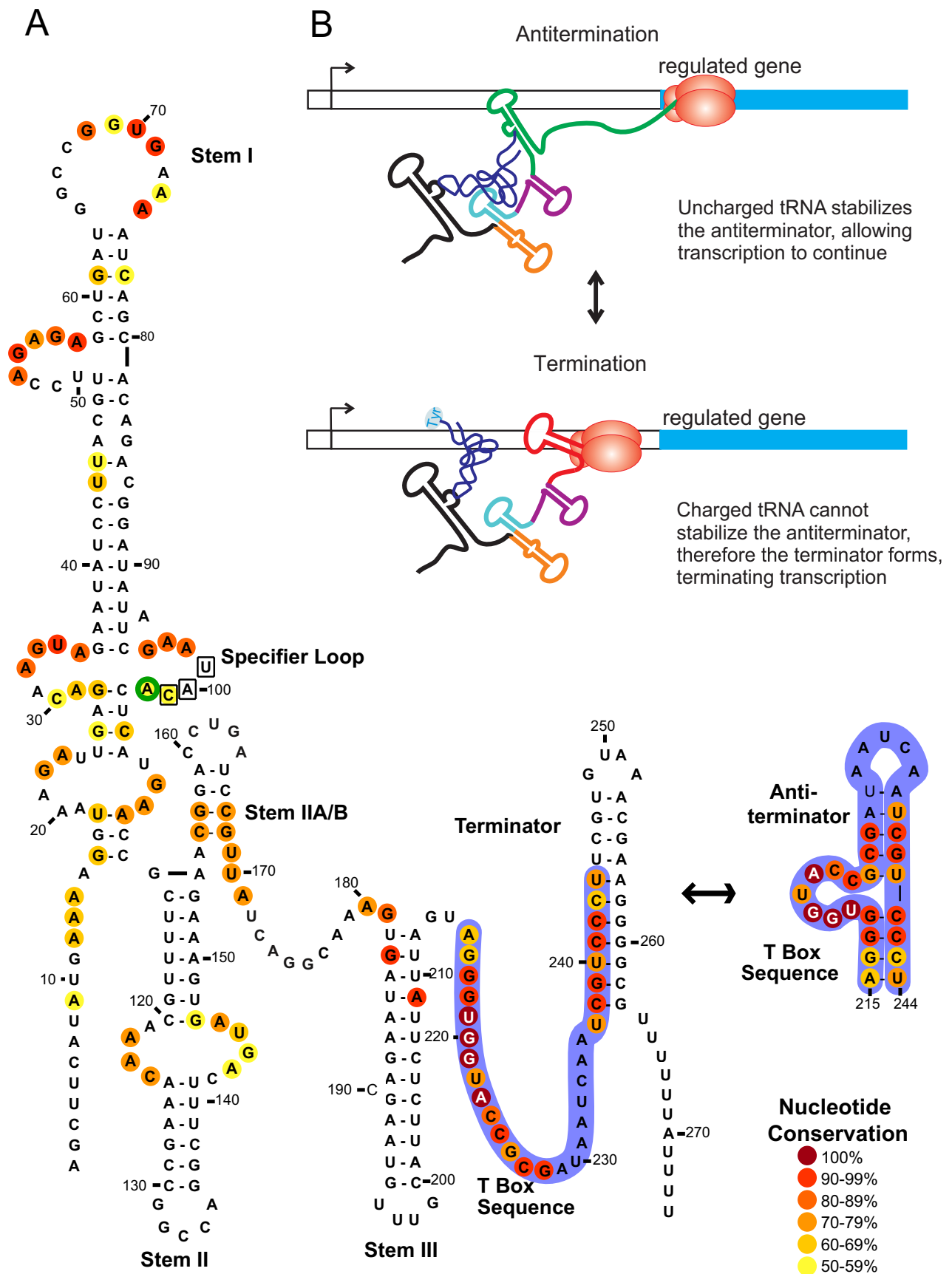


FIG. 1. The T-box RNA regulatory system. (A) Structural model of the *B. subtilis tyrS* T-box leader RNA. The T-box element present in the *B. subtilis tyrS* leader region was originally described by Grundy and Henkin (see reference 43). The standard T-box leader RNA arrangement consists of three major elements, stem I, stem II, and stem III plus the stem IIA/stem IIB pseudoknot, and the competing terminator and

UAC Tyr codon with a UAU Tyr codon resulted in a dramatic decrease in levels of expression *in vivo*. Moreover, initial studies with the *B. subtilis glyQS* leader region suggested that the bias for C in the third position of the specifier sequence is due to structural constraints for codon and/or codon-anticodon pairing that differ from those involved in translation (E. Caserta, F. Grundy, and T. Henkin, unpublished results). The universally conserved U34 residue 5' to the anticodon in tRNA was also predicted to pair with a conserved purine residue 3' to the specifier sequence (83; F. Grundy and T. Henkin, unpublished data). This pairing is supported by the structural analysis of the *B. subtilis glyQS* leader RNA/tRNA<sup>Gly</sup> complex (121).

The 14 most highly conserved residues of the entire T-box RNA represent the "T-box sequence" (AGGGUGGNACC GCG) (Fig. 1A); the recognition of this sequence in the leader regions of several aminoacyl-tRNA synthetase (aaRS) genes led to the initial prediction of a conserved regulatory mechanism (44). This sequence was shown to form the 5' side of the antiterminator structure, including the 7-nucleotide bulge (43). Discrimination between uncharged and charged tRNAs is mediated by the pairing of the four unpaired residues at the 3' end of the tRNA (5'-NCCA-3') with the first four residues of the antiterminator bulge (5'-UGGN-3'); the N residue in the antiterminator bulge covaries with the corresponding position of the tRNA, a position that often plays an important role in tRNA identity for recognition by the cognate aaRS. The presence of the amino acid at the 3' end of a charged tRNA prevents the interaction of its 3' end with the antiterminator RNA; hence, the antiterminator structure does not form. Both charged and uncharged tRNAs can interact with the leader RNA at the specifier sequence, but only uncharged tRNA has been shown to stabilize the antiterminator sequence. Thus, importantly, each T-box sequence monitors the ratio between the charged and uncharged forms of a specific tRNA rather than the absolute amount of the uncharged tRNA (51). T-box leader RNAs are further characterized by the existence of a set of conserved primary sequence elements at specific locations relative to the structural features (43). Mutational studies have demonstrated that many of the sequence and structural features conserved in T-box leader RNAs are important for function (49, 88, 112), but overall primary sequence conservation is low. Interestingly, the two main features of tRNAs (the anticodon and the base immediately preceding the CCA at its 3' end) that are recognized by the T-box RNAs are also usually central elements for recognition by the cognate aaRS (14). aaRSs have been

grouped into two nonhomologous sets of enzymes, class I (LeuRS, IleRS, MetRS, TyrRS, GluRS, ValRS, ArgRS, LysRS1, and TrpRS) and class II (PheRS, SerRS, ThrRS, ProRS, AlaRS, HisRS, AspRS, AsnRS, LysRS2, and GlyRS) (113). Class I and class II aaRSs recognize different tRNA structural features. Whether these or other tRNA structural features are also recognized by the T-box system remains to be determined.

In addition to the classical T-box RNA arrangement described by Grundy and Henkin (47), a reduced version of the T-box RNA, which is predicted to regulate the initiation of translation in the *Actinobacteria*, has also been described (107). This type of T-box RNA contains a smaller variant of the stem I structure, where the specifier sequence is placed in the terminal loop. Unusual examples of partially duplicated T-box RNAs in which a single stem I is followed by double or triple copies of the antiterminator/terminator were also reported (107), but the function(s) of these extra copies is not yet understood.

Using previously described position-specific matrices associated with leader RNAs that have T-box features (1, 2), we performed genome-scale searches using the MAST (6) and covariance models of Rfam with the program cmsearch (27, 42). Although other genomic studies used only primary sequence conservation (111), the use of covariance models for RNA secondary-structure prediction improves the accuracy of riboswitch identification. In the specific case of the T boxes, covariance analyses facilitated the identification of the specifier sequence. In the present study, 559 fully sequenced bacterial genomes were examined and 1,111 T-box leader sequences were identified in 87 organisms. The criteria for concluding that adjacent genes are within the same operon were described previously by Janga et al. and Salgado et al. (65, 89). The genomic contexts of the significant matches identified were further analyzed using our Web GeConT server (19, 72). Our findings are organized according to the type of gene or genes that were identified downstream of a T-box element, as described in the following sections. The complete list of T-box genes with a known or predicted function is shown in Fig. 3 to 8 and 10. Specific examples are discussed below to emphasize particular points of interest.

### AMINOACYL-tRNA SYNTHETASE GENES

The first group of genes to be identified that are regulated by the T-box mechanism encode aaRSs (43, 57). The accumulation of uncharged tRNA<sup>Tyr</sup> was shown to be the intracellular

---

antiterminator structures. The specifier loop, an internal bulge in stem I, contains the specifier sequence (boxed UAC residues complementary to the anticodon sequence of tRNA<sup>Tyr</sup>); the conserved purine (an adenine) following the specifier sequence is inside a green circle. The T-box sequence is unpaired in the terminator form and is paired in the antiterminator form (the antiterminator is shown to the right of the terminator). The sequence highlighted in blue shows the nucleotides involved in the antiterminator structure. The antiterminator structure has a bulge that interacts with the unpaired residues at the acceptor end of an uncharged tRNA. Nucleotide conservation in all 722 T-box sequences analyzed was evaluated using a multiple sequence alignment obtained from the Rfam database (42), and residues are color coded accordingly. (B) Model of the regulatory alternatives for the T-box mechanism. During the transcription of a leader region by RNA polymerase (red ovals), the nascent RNA folds into a structure competent for binding of the cognate tRNA at two sites. The binding of uncharged tRNA (top) to both the specifier sequence and the antiterminator bulge stabilizes the antiterminator (green RNA segment), preventing the formation of the terminator. This allows transcription to proceed into the downstream-regulated coding sequence (blue box). Charged tRNA (represented by Tyr attached to the 3' end of the tRNA) can interact with the specifier sequence but cannot interact with the antiterminator; a failure to stabilize the antiterminator allows the formation of the terminator helix (red RNA segment), and transcription is terminated before the downstream coding region can be transcribed. Conserved elements of T-box RNAs are stem I (black), stem II (orange), the stem IIA/stem IIB pseudoknot (light blue), and stem III (purple).

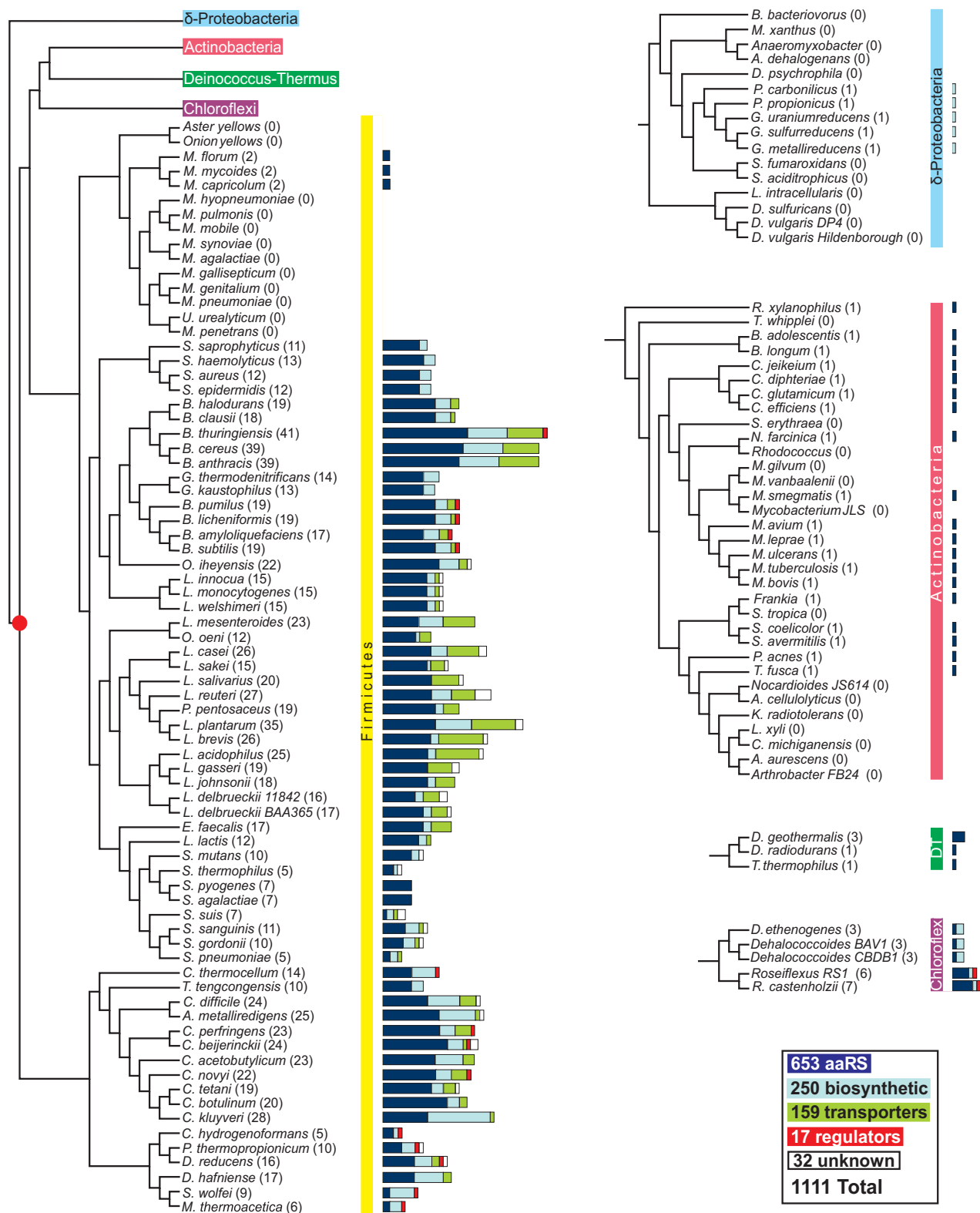


FIG. 2. Distribution of T-box regions in different phylogenetic taxa. The phylogenetic tree for organisms relevant to our study was constructed based on the phylogenetic distances of aligned sequences from the concatenation of 31 proteins in 191 species, as previously described (18). Alignments were generated using the program MUSCLE (28), and phylogenetic reconstruction was performed using the PROTDIST program of the PHYLIP phylogeny inference package program (version 3.57c; J. Felsenstein, University of Washington, Seattle). Operons were predicted based on an analysis of intergenic distances, as described previously (76). Horizontal bar lengths are drawn to scale, reflecting the number of operons regulated by a T-box sequence; these are classified into one of the following groups: aminoacyl-tRNA synthetases (dark blue), amino acid biosynthetic genes (light blue), genes coding for regulatory proteins (red), transporter genes (green), and genes of unknown function (white). The most parsimonious scenario would place the initially evolved T-box regulatory sequence in a common ancestor of the *Firmicutes*, the *Actinobacteria*, the *Chloroflexi*, and the *Deinococcus-Thermus* (DT) group. The postulated origin is represented by a red dot in the tree. Names of *Firmicutes* are as follows: *M. hypopneumoniae*, *Mycoplasma hypopneumoniae*;



signal that leads to the activation of transcription of the *B. subtilis tyrS* gene, encoding TyrRS. Regulation was achieved by modulating the readthrough of the leader RNA transcription terminator. Similar findings were reported for other *B. subtilis* aaRS genes, including *thrS*, *thrZ*, *leuS*, and *valS* (20, 69, 82, 106), and T-box leader regions were identified upstream of additional aaRS genes, including *pheS*, *tyrZ*, and *trpS*, in *Bacillus* sp. (43). Subsequent analyses uncovered *serS*, *ileS*, *glyQS*, *alaS*, and *hisS-aspS* as additional T-box-regulated aaRS genes in the *B. subtilis* genome (17).

Our genomic analyses of the aaRS gene family revealed that the aaRS responsible for charging each of the 20 amino acids is regulated by the T-box mechanism in at least one firmicute species (Fig. 2 and 3). This mechanism appears to be the most commonly used regulatory mechanism for this family of genes in this group of organisms. T-box sequences were identified for members of both class I and class II aaRS enzymes. As previously mentioned, members of the two classes of aaRS enzymes are not homologous and are grouped based on the topology of the ATP binding domain. Class I proteins contain a Rossmann fold, while class II enzymes possess an unrelated  $\beta$ -sheet arrangement (113, 114). Class I and class II aaRS enzymes are present in all organisms, and each type of tRNA is aminoacylated exclusively by a member of one of the two classes of aaRS, with the exception of tRNA<sup>Lys</sup>, which is aminoacylated by LysRS enzymes of both classes (113). A class II LysRS is present in all members of the *Firmicutes* and is encoded by a gene residing within a supraoperon containing genes involved in folate biosynthesis. This supraoperon does not contain a T-box sequence, and in *B. subtilis*, it is transcriptionally regulated in a growth phase-dependent manner (25). In addition to this class II LysRS, a gene encoding a class I LysRS (which is preceded by a T-box regulatory sequence) is present in *Bacillus cereus*, *Bacillus thuringiensis*, and *Clostridium beijerinckii* (Fig. 3).

Generally, T-box-regulated aaRS genes are located in monocistronic transcriptional units, with three major exceptions: (i) when they encode different polypeptides in a heterodimeric enzyme, like *glyQS*; (ii) when they are associated with biosynthetic genes, such as *cysS*, which is located within the cysteine biosynthetic operon in some members of the *Firmicutes*; and (iii) when two aaRS enzymes for different amino acids are encoded in the same operon, such as the *hisS-aspS* operon, encoding HisRS and AspRS (Fig. 3). The coexpression of biosynthesis genes with aaRS genes of the same amino acid class represents an efficient use of the T-box regulatory mechanism, as both sets of genes respond in concert with the same tRNA. In contrast, the cotranscription of aaRS genes of different amino acid classes represents a potential regulatory problem, as the expression of each gene would be expected to

respond individually to its cognate uncharged tRNA. A possible solution is apparent in some *hisS-aspS* aaRS operons in which the specifier loop of the T-box element contains the sequence GACAC, which has Asp (GAC) and His (CAC) codons that overlap by one nucleotide. The predicted RNA secondary structure places GAC in the most appropriate position for interaction with the tRNA anticodon, suggesting that the transcription of both *hisS* and *aspS* would depend on the accumulation of uncharged tRNA<sup>Asp</sup>. This could cause a deficiency in the sensing of tRNA<sup>His</sup>, resulting in a failure to increase the synthesis of HisRS when increased charging of tRNA<sup>His</sup> is required. However, in *B. halodurans*, *Bacillus licheniformis*, and *Clostridium thermocellum*, the stem I region could adopt an alternative secondary structure in which the His (CAC) codon is accessible for interactions with tRNA<sup>His</sup>, allowing the transcription of the operon to respond to each of these uncharged tRNA species. This hypothesis requires experimental validation.

A second example in which aaRS genes for two different amino acids are regulated by one T-box RNA sequence is found in several *Clostridium* species (*Clostridium acetobutylicum*, *C. beijerinckii*, *C. difficile*, *C. perfringens*, and *C. tetani*). In these organisms, *cysS* and *proS* are in the same transcriptional unit, which appears to be regulated by a T-box sequence responding only to tRNA<sup>Pro</sup>; this is predicted to create an imbalance favoring the sensing of tRNA<sup>Pro</sup> (Fig. 3). In *C. tetani*, the Cys biosynthetic genes are also cotranscribed in a *cysS-proS* operon with a Pro specifier sequence (Fig. 3), making this potential imbalance even more significant, as both Cys biosynthesis and tRNA<sup>Cys</sup> charging would therefore be predicted to be regulated in response to the availability of Pro rather than Cys. It is unclear whether there is an additional regulatory strategy (e.g., at the level of transcription initiation or translation) that overcomes this potential problem. Another example of regulation of an operon by an unexpected tRNA is found in the *yrG-serAS* operon in *Pelotomaculum thermopropionicum*, which includes genes involved in the biosynthesis of Ser as well as the SerRS gene yet is regulated by a Gly (GGC) T box. This is likely to be explained by the efficient interconversion of glycine and serine in bacterial cells (see below).

Although, in general, there is only one copy of each aaRS gene per genome, there are a few organisms in which multiple copies of an aaRS gene are present. These include *tyrS/tyrZ* and *thrS/thrZ* in *B. subtilis*, each of which is regulated by the T-box mechanism. *Bacillus anthracis*, *B. cereus*, and *B. thuringiensis* have two genes encoding AspRS, one of which is monocistronic, whereas the other is present in a *hisS-aspS* operon. These two aaRS genes are both predicted to be regulated by the T-box mechanism using a leader RNA with an Asp speci-

*M. pulmonis*, *Mycoplasma pulmonis*; *M. mobile*, *Mycoplasma mobile*; *M. synoviae*, *Mycoplasma synoviae*; *M. agalactiae*, *Mycoplasma agalactiae*; *M. gallisepticum*, *Mycoplasma gallisepticum*; *M. genitalium*, *Mycoplasma genitalium*; *M. pneumoniae*, *Mycoplasma pneumoniae*; *U. urealyticum*, *Ureaplasma urealyticum*; *M. penetrans*, *Mycoplasma penetrans*. Names of Deltaproteobacteria are as follows: *B. bacteriovorus*, *Bdellovibrio bacteriovorus*; *M. xanthus*, *Myxococcus xanthus*; *A. dehalogenans*, *Anaeromyxobacter dehalogenans*; *D. psychrophila*, *Desulfotalea psychrophila*; *S. fumaroxidans*, *Syntrophobacter fumaroxidans*; *S. aciditrophicus*, *Syntrophus aciditrophicus*; *L. intracellularis*, *Lawsonia intracellularis*; *D. sulfuricans*, *Desulfobacterium sulfuricans*; *D. vulgaris*, *Desulfobacterium vulgaris*. Names of Actinobacteria are as follows: *T. whipplei*, *Tropheryma whipplei*; *M. gilvum*, *Mycobacterium gilvum*; *M. vanbaalenii*, *Mycobacterium vanbaalenii*; *S. tropica*, *Salinispora tropica*; *A. cellulolyticus*, *Acidothermus cellulolyticus*; *K. radiotolerans*, *Kineococcus radiotolerans*; *L. xyli*, *Leifsonia xyli*; *C. michiganensis*, *Clavibacter michiganensis*; *A. aurescens*, *Arthrobacter aurescens*. Additional names are listed in the legend to Fig. 3.

Alanine	<i>A. met, B. amy, B. ant, B. cer, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, D. haf, E. fae, G. kau, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. lac, L. mon, L. mes, L. pla, L. sal, L. reu, L. wel, O. ihe, O. oen, P. pen, S. aur, S. epi, S. hae, S. pyo, S. sap, S. the</i>	Methionine	<i>A. met, B. ant, B. cer, B. cla, B. hal, B. thu, C. ace, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, D. red, O. ihe, T. ten</i>
	<i>S. aga, S. gor, S. pne, S. san</i>		<i>C. per</i>
Arginine	<i>B. ant, B. cer, B. cla, B. hal, B. lic, B. sub, B. thu, C. ace, C. per, C. tet, G. kau, L. aci, L. inn, L. joh, L. lac, L. mon, L. pla, O. ihe, S. aur, S. epi</i>	Phenylalanine	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. klu, C. nov, C. the, C. per, C. tet, D. haf, D. red, E. fae, G. kau, G. the, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. lac, L. mes, L. mon, L. pla, L. sak, L. sal, L. reu, L. wel, M. the, O. ihe, O. oen, P. the, P. pen, S. aga, S. aur, S. epi, S. hae, S. mut, S. pyo, S. sap, S. the, S. wol, T. ten</i>
	<i>A. met, B. ant, B. cer, B. hal, B. thu, C. ace, C. bei, C. bot, C. nov, C. per, C. tet, E. fae, L. aci, L. bre, L. cas, L. gas, L. joh, L. pla, L. sak, L. pla, L. sal, L. reu, P. pen</i>		<i>S. gor, S. san</i>
Asparagine	<i>A. met, B. ant, B. cer, B. pum, B. thu, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, C. the, L. bre, P. pen, S. mut, S. san</i>	Proline	<i>B. ant, B. cer, B. pum, B. thu, C. bei, C. bot, C. klu, C. nov, C. per, O. ihe</i>
	<i>L. del, L. pla, L. reu</i>		<i>C. ace, C. bei, C. dif, C. per</i>
Aspartate	<i>A. met, B. ant, B. cer, B. thu, C. ace, C. bot, L. reu</i>	Serine	<i>C. tet</i>
	<i>B. ant, B. cer, B. cla, B. lic, B. sub, B. pum, B. thu, D. haf, E. fae, G. kau, S. aur, S. epi, T. ten</i> [see <i>C. ace, C. dif, C. per</i> and <i>C. tet</i> in proline]		<i>A. met</i>
Cysteine	<i>O. ihe</i>	Threonine	<i>A. met, B. amy, B. ant, B. cer, B. hal, B. lic, B. pum, B. sub, B. thu, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, E. fae, G. kau, G. the, L. aci, L. bre, L. cas, L. del, L. inn, L. gas, L. joh, L. lac, L. mes, L. mon, L. pla, L. reu, L. sal, L. wal, O. ihe, P. pen, S. aga, S. aur, S. epi, S. gor, S. hae, S. mut, S. pne, S. pyo, S. san, S. sap, T. ten</i>
	<i>L. sak</i>		<i>C. bot, E. fae, L. pla</i>
	<i>C. the</i>		<i>D. red, S. wol</i>
	<i>A. met</i>		<i>C. ace</i>
Glutamate	<i>C. per</i>	Tryptophan	<i>B. ant, B. cer, B. thu</i>
	<i>C. bei</i>		<i>B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. per, D. haf, G. kau, G. the, L. aci, L. bre, L. cas, L. inn, L. joh, L. lac, L. mon, L. pla, L. reu, L. sak, L. sal, L. wal, O. ihe, P. pen, S. pyo, S. the</i>
Glutamine	<i>C. bei</i>	Tyrosine	<i>B. ant, B. cer, B. thu</i>
	<i>B. amy, B. cla, B. hal, B. lic, B. pum, B. sub, D. red, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. lac, L. mon, L. pla, L. sak, L. sal, L. reu, L. wal, O. ihe, O. oen, P. the, P. pen, S. pyo, S. the</i>		<i>B. ant, B. cer, B. cla, B. pum, B. thu</i>
	<i>A. met, B. ant, B. cer, B. thu, C. ace, C. bei, C. bot, C. klu, C. nov, C. per, C. tet, G. kau, G. the, S. aur, S. epi, R. SM1, R. cas, D. geo, D. rad</i>		<i>B. lic, C. bei, C. bot, C. dif, L. sak, M. the, S. hae</i>
	<i>E. fae, S. mut, S. san, S. sui</i>		<i>B. sub</i>
	<i>S. pne</i>		<i>S. mut, S. sui, T. ten, R. SM1, R. cas</i>
	<i>S. aga</i>		<i>O. oen</i>
Histidine/Aspartate	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. lic, B. sub, B. thu, C. the, D. haf, D. red, G. kau, G. the, L. aci, L. cas, L. del BAA365, L. gas, L. joh, L. inn, L. lac, L. mes, L. mon, L. pla, L. sak, L. sal, L. wal, O. ihe, P. the, P. pen, S. aur, S. epi, S. hae, S. mut, S. sap, T. ten</i>	Valine	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. nov, C. the, E. fae, G. the, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. mon, L. pla, L. reu, L. sak, L. sal, L. wal, O. ihe, O. oen, P. pen, S. aur, S. epi, S. hae, S. sap</i>
	<i>B. hal</i>		<i>S. aga, S. mut, S. pne, S. pyo</i>
Histidine	<i>L. reu</i>	Leucine	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. nov, C. per, C. tet, G. kau, G. the, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. mon, L. pla, L. reu, L. sak, L. sal, L. wal, O. ihe, O. oen, P. pen, S. aur, S. epi, S. hae, S. sap, T. ten</i>
	<i>S. gor</i>		<i>B. ant, B. cer, B. thu, O. ihe</i>
Isoleucine	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, C. the, D. haf, D. red, E. fae, G. kau, G. the, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. lac, L. mes, L. mon, L. pla, L. sak, L. sal, L. wal, M. cap, M. flo, M. myc, O. ihe, O. oen, P. the, P. pen, S. aga, S. aur, S. epi, S. hae, S. gor, S. mut, S. pyo, S. san, S. sap, S. the, T. ten, B. ado, B. lon, C. dip, C. glu, C. jei, F. aln, Frankia, M. avi, M. bov, M. lep, M. ulc, M. sme, M. tub, N. far, P. acn, R. xyl, S. ave, S. coe, T. fus, D. eth, D. BAV1, D. CBDB1, R. SM1, R. cas, D. geo, T. the</i>	Lysine	<i>A. met, B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. sub, B. thu, C. ace, C. bei, C. bot, C. dif, C. klu, C. nov, C. per, C. tet, D. haf, D. red, E. fae, G. kau, G. the, L. aci, L. bre, L. cas, L. del, L. gas, L. inn, L. joh, L. lac, L. mes, L. mon, L. pla, L. sak, L. sal, L. reu, L. wal, O. ihe, O. oen, P. pen, S. aur, S. epi, S. hae, S. sap, T. ten</i>
	<i>C. hyd, P. the, S. wol</i>		<i>B. cer, B. thu, C. bei</i>

T box attenuation mechanism

FIG. 3. Aminoacyl-tRNA synthetase genes regulated by the T-box mechanism. From the set of T-box-regulated genes identified in our study, operons containing aaRS genes were grouped according to the amino acid class of the aaRS. Operons containing more than one different aaRS gene are shown under the amino acid category matching the predicted specifier sequence. In the exceptional case of *leuS* in *C. hydrogeniformans*, *P. thermopropionicum*, and *S. wolfei*, the T-box sequence, drawn in green, contains a tRNA gene. Organism nomenclature is as follows for members of the Firmicutes: *A. met*, “*Alkaliphilus metalliredigens*”; *B. amy*, *Bacillus amyloliquefaciens*; *B. cla*, *Bacillus clausii*; *B. cer*, *Bacillus cereus*; *B. hal*, *Bacillus halodurans*; *B. lic*, *Bacillus licheniformis*; *B. pum*, *Bacillus pumilus*; *B. sub*, *Bacillus subtilis*; *B. ste*, *Bacillus stearothermophilus*; *B. thu*, *Bacillus thuringiensis*; *C. ace*, *Clostridium acetobutylicum*; *C. bei*, *Clostridium beijerinckii*; *C. bot*, *Clostridium botulinum*; *C. dif*, *Clostridium difficile*; *C. hyd*, *Clostridium hydrogeniformans*; *C. klu*, *Clostridium kluyveri*; *C. nov*, *Clostridium novyi*; *C. per*, *Clostridium*

fier sequence. These organisms (among others) also encode a third AspRS-related aaRS that mischarges tRNA<sup>Asn</sup> with Asp to generate Asp-tRNA<sup>Asn</sup>, which is subsequently converted to Asn-tRNA<sup>Asn</sup> by an amidotransferase enzyme complex. This mischarging AspRS gene is preceded by a T-box leader that is predicted to respond to uncharged tRNA<sup>Asn</sup>, consistent with its function (Fig. 3).

aaRS genes are generally regulated by a single T-box element, although examples with tandem T-box arrangements have been identified (Fig. 3). An extreme example is represented by the *B. subtilis thrZ* gene, encoding an isozyme of ThrRS (80). The leader region of the *thrZ* operon contains three T-box elements as direct repeats. This arrangement results in tighter regulation, since the binding of three molecules of uncharged tRNA<sup>Thr</sup> is required for transcription to proceed through all three leader region terminators (38, 82). The *thrS* gene, which encodes the major ThrRS isoenzyme, is regulated by a single T-box element and therefore can be expressed when uncharged tRNA<sup>Thr</sup> first begins to accumulate; ThrRS activity increases the pool of charged tRNA<sup>Thr</sup>, and as a consequence, the transcription of *thrZ* is rarely induced under normal growth conditions. A similar arrangement is found in *B. clausii* and *Bacillus pumilus*, where the *thrZ* gene is regulated by tandem T-box sequences. This T-box distribution is also found in the leader regions of some operons containing biosynthetic genes (see below).

The coupling of transcription of each aaRS gene with the charging of its cognate tRNA appears to be metabolically beneficial. Reduced levels of an aaRS relative to its uncharged tRNA substrates decrease the extent of charging of the corresponding tRNAs, which promotes the stalling of the translational machinery and eventually triggers mRNA degradation and the stringent response. The increased level of expression of each aaRS when charging of the cognate tRNA is low allows the maintenance of protein synthesis. However, the synthesis of excess levels of an individual aaRS relative to its uncharged

tRNA substrate increases the risk of mischarging of a noncognate tRNA. The ability of the T-box mechanism to monitor the ratio between its substrate (uncharged tRNA) and its product (charged tRNA) allows the synthesis of each aaRS to precisely match the physiological requirements of the cell.

The phylogenetic distribution of T-box sequences revealed by our analysis agrees with data from the recent study by Vitreschak et al. (107) showing that the number of aaRS genes regulated by a T box in gram-positive bacteria is highly variable, ranging from high among the *Bacillaceae*, where most aaRS genes are regulated by a T-box sequence, to low in the *Actinomycetes*, where the T-box regulation of aaRS genes is uncommon (Fig. 2). This variability may reflect differences in the evolutionary history of each organism as well as differences in their environmental niches.

### T-BOX REGULATION OF AMINO ACID BIOSYNTHETIC GENES

The biosynthesis of amino acids is energetically costly; accordingly, the expression of amino acid biosynthetic genes is generally highly regulated. The mechanisms used to regulate the expression of these genes have evolved to respond to changes in the intracellular levels of their free amino acids and/or the relative levels of their corresponding nonaminoacylated and aminoacylated tRNAs (118). In gram-negative bacteria, this regulation often utilizes regulatory proteins that control transcription initiation in response to amino acid availability as well as leader regions that mediate transcription attenuation via leader peptide coding region translation, which is sensitive to the availability of specific charged tRNAs (116, 117). In gram-positive bacteria, regulatory proteins are also used to sense amino acids (29, 97, 98). For example, in *B. subtilis* and its close relatives, the biosynthetic genes of the *trp* operon are regulated by tryptophan (Trp) and the TRAP RNA binding protein, which regulates transcription termination (5,

---

*perfringens*; C. tet, *Clostridium tetani*; C. the, *Clostridium thermocellum*; D. haf, *Desulfotobacterium hafniense*; D. red, *Desulfotomaculus reducens*; E. fae, *Enterococcus faecalis*; G. kau, *Geobacillus kaustophilus*; G. the, *Geobacillus thermodenitrificans*; L. aci, *Lactobacillus acidophilus*; L. bre, *Lactobacillus brevis*; L. cas, *Lactobacillus casei*; L. del, both *Lactobacillus delbrueckii* subsp. *bulgaricus* strains; L. del 11842, *Lactobacillus delbrueckii* subsp. *bulgaricus* strain ATCC 11842; L. del BAA365, *Lactobacillus delbrueckii* subsp. *bulgaricus* strain ATCC BAA365; L. gas, *Lactobacillus gasseri*; L. inn, *Listeria innocua*; L. joh, *Lactobacillus johnsonii*; L. lac, *Lactococcus lactis*; L. mes, *Leuconostoc mesenteroides*; L. mon, *Listeria monocytogenes*; L. pla, *Lactobacillus plantarum*; L. reu, *Lactobacillus reuteri*; L. sak, *Lactobacillus sakei*; L. sal, *Lactobacillus salivarius*; L. wel, *Listeria welshimeri*; M. cap, *Mycoplasma capricolum*; M. flo, *Mesoplasma florum*; M. myc, *Mycoplasma mycoides*; M. the, *Moorella thermoacetica*; O. ihe, *Oceanobacillus iheyensis*; O. oen, *Oenococcus oeni*; P. pen, *Pediococcus pentosaceus*; P. the, *Pelotomaculum thermopropionicum*; S. aga, *Streptococcus agalactiae*; S. aur, *Staphylococcus aureus*; S. epi, *Staphylococcus epidermidis*; S. gor, *Streptococcus gordonii*; S. hae, *Staphylococcus haemolyticus*; S. mut, *Streptococcus mutans*; S. pne, *Streptococcus pneumoniae*; S. pyo, *Streptococcus pyogenes*; S. san, *Streptococcus sanguinis*; S. sap, *Staphylococcus saprophyticus*; S. sui, *Streptococcus suis*; S. the, *Streptococcus thermophilus*; S. wol, *Syntrophomonas wolfei*; T. ten, *Thermoanaerobacter tengcongensis*. Organism nomenclature is as follows for members of the *Actinobacteria*: B. ado, *Bifidobacterium adolescentis*; B. lon, *Bifidobacterium longum*; C. eff, *Corynebacterium efficiens*; C. dip, *Corynebacterium diphtheriae*; C. glu, *Corynebacterium glutamicum*; C. jei, *Corynebacterium jeikeium*; M. avi, *Mycobacterium avium*; M. bov, *Mycobacterium bovis*; M. lep, *Mycobacterium leprae*; M. sme, *Mycobacterium smegmatis*; M. tub, *Mycobacterium tuberculosis*; M. ulc, *Mycobacterium ulcerans*; N. far, *Nocardia farcinica*; P. can, *Propionibacterium acnes*; R. xyl, *Rubrobacter xylanophilus*; S. ave, *Streptomyces avermitilis*; S. coe, *Streptomyces coelicolor*; T fus, *Thermobifida fusca*. Organism nomenclature is as follows for members of the *Fusobacteria*: F. nuc, *Fusobacterium nucleatum*. Organism nomenclature is as follows for members of the *Deinococcus-Thermus* group: D. geo, *Deinococcus geothermalis*; D. rad, *Deinococcus radiodurans*; T. the, *Thermus thermophilus*. Organism nomenclature is as follows for members of the *Chlorobi*: C. tep, *Chlorobium tepidum*; C. aur, *Chloroflexus aurantiacus*; C. hut, *Cytophaga hutchinsonii*. Organism nomenclature is as follows for members of the *Chloroflexi*: D. BAV1, "*Dehalococcoides*" sp. strain BAV1; D. CBDB1, *Dehalococcoides* sp. strain CBDB1; D. eth, "*Dehalococcoides ethenogenes*"; R. SM1, *Roseiflexus castenholzii*. Organism nomenclature is as follows for members of the *Proteobacteria*: G. sul, *Geobacter sulfurreducens*; G. met, *Geobacter metallireducens*; G. ura, *Geobacter uraniumreducens*; P. car, *Pelobacter carbinolicus*; P. pro, *Pelobacter propionicus*. Operon predictions and the color code used for the different types of regulated genes are described in the legend of Fig. 2.



40). In contrast, in many other members of the *Firmicutes*, the transcriptional regulation of the *trp* operon is mediated by a T-box RNA that responds to uncharged tRNA<sup>Trp</sup>. The regulation of biosynthetic operons by the T-box mechanism might have evolved in response to metabolic demands, as revealed by our initial genomic analysis of the *trp* biosynthetic operons of gram-positive bacteria (52) and extended in the present study to genes concerned with other amino acid biosynthetic pathways.

#### Regulation of Serine and Glycine Biosynthetic Genes by the T-Box and *gcvT* Riboswitches

The *serA* gene encodes the enzyme that catalyzes the first reaction in the serine (Ser) biosynthetic pathway. The mechanism of regulation of *serA* in most members of the *Firmicutes* is unknown, although in *B. clausii*, *B. halodurans*, *C. acetobutylicum*, and *C. tetani*, a Ser T-box sequence is located in the *serA* regulatory region (Fig. 4). In *B. clausii* and *B. halodurans*, a Ser (AGC) codon is present in the specifier loop of the *serA* T-box RNA, while in the other two organisms, the Ser (UCC) codon is present. In *C. acetobutylicum* and *C. tetani*, *serA* is in an operon that also encodes a probable serine-pyruvate/aspartate aminotransferase. This operon also contains the *serS* gene, encoding SerRS, in *C. acetobutylicum*. In *B. subtilis*, the *serA* paralog *yoaD* is regulated by the S-box riboswitch, which responds to S-adenosylmethionine (SAM) rather than a tRNA (45).

The *serC* and *serB* genes encode the enzymes that catalyze the second and third reactions in this pathway, respectively. No information on their regulation is currently available, but our analyses suggest that in *C. thermocellum* and *Desulfotobacterium hafniense*, the *serC* gene is regulated by a Ser (UCC) T-box RNA. In *Lactococcus lactis* and *Streptococcus mutans*, *serB* is located in the *his* operon and appears to be regulated by a T-box sequence that responds to tRNA<sup>His</sup>. Ser can also be synthesized from pyruvate by a one-step enzymatic reaction; the gene encoding this enzyme (designated *yurG* in *B. subtilis*) is cotranscribed with *serA* in *C. acetobutylicum*, *Clostridium botulinum*, *Clostridium kluyveri*, *Clostridium novyi*, *C. tetani*, *Desulfotomaculum reducens*, and *Syntrophomonas wolfei*. This gene is regulated by a single Ser (UCC) T box in all of the clostridia mentioned above except *C. kluyveri*, where regulation is controlled by tandem Ser (UCC) T boxes (Fig. 4).

Ser can also be synthesized from glycine by GlyA (46). The *glyA* gene is commonly regulated by the Gly-responsive *gcvT* riboswitch, which results in increased expression when glycine is abundant (1, 7, 70). Interconversion of Gly and Ser by GlyA is the primary pathway used for Gly synthesis in bacteria (46, 99). This pathway also produces 5,10-methylenetetrahydrofolate, a major contributor of the one-carbon unit in the formation of methionine (Met), purines, and thymine (75).

The only Gly T boxes identified so far that regulate biosynthetic operons are found in *Moorella thermoacetica* (*yurG-serA*) and in *P. thermopropionicum* (*yurG-serAS*) (Fig. 5). Although YurG and SerA are directly involved in Ser biosynthesis, the biological relevance of the regulation of these operons in response to uncharged tRNA<sup>Gly</sup> accumulation might be that an increase in the pool of Ser (which can be converted to Gly) would also increase the pool of glycine.

#### Pathways for Synthesis of the Sulfur-Containing Amino Acids Methionine and Cysteine Are Regulated by S-Box and T-Box Riboswitches

The biosynthesis of Met and cysteine (Cys) can utilize a number of alternate pathways and is regulated by a wide variety of mechanisms, including the T-box and S-box mechanisms. The SAM-responsive S-box riboswitch also regulates the expression of the *metK* gene, encoding SAM synthetase, reflecting the importance of Met not only as an amino acid but also as a precursor of SAM. A third type of riboswitch, the S<sub>MK</sub> box, regulates *metK* expression in lactic acid bacteria including *Enterococcus*, *Streptococcus*, and *Lactococcus* spp. in response to SAM (33). In contrast, DNA binding transcription factors such as MtaR, MetR, and CmbR are utilized in streptococci (67, 86), in common with the regulatory pattern found in *E. coli*.

Homoserine is a key intermediate in the biosynthesis of Met and Cys, and it also participates in the biosynthesis of Gly and threonine (Thr) (46). Homoserine is derived from aspartate (Asp) by the action of homoserine dehydrogenase, which is encoded by the *hom* gene. Since the product of the *hom* gene participates in multiple biosynthetic pathways, it is not surprising to find that this gene is regulated by several different mechanisms. For example, in *B. clausii*, *B. halodurans*, *D. hafniense*, and *Thermoanaerobacter tengcongensis*, the expression of the *hom* gene is regulated by an S-box riboswitch in response to SAM, while in *C. difficile*, the *hom* gene is regulated by tandem Thr-responsive T boxes (Fig. 5) (46). In *B. anthracis*, *B. cereus*, and *B. thuringiensis*, there are two paralogous copies of the *hom* gene. One of these *hom* genes is cotranscribed with the *metY* and *metA* genes and is regulated by an S-box riboswitch, while the other is cotranscribed with the *thrB* and *thrC* genes and is regulated by tandem T boxes, each with a Thr specifier sequence (Fig. 5).

In *B. subtilis*, homoserine is acetylated by MetA followed by a reaction with Cys to form cystathionine. The *metA* gene is regulated by the S-box mechanism in a number of organisms, including *C. difficile* and *Staphylococcus* sp., while the regulatory mechanism in *B. subtilis* remains unknown. In contrast, in *Lactobacillus plantarum*, the *metA* gene is cotranscribed with a gene encoding O-acetylhomoserine (thiol)-lyase in an operon regulated by the T-box mechanism using a Met specifier sequence. Other organisms (including *E. coli*) utilize MetA to convert homoserine to O-succinylhomoserine (41, 46). Further steps are catalyzed by cystathionine  $\gamma$ -synthase (YjcI/MetI) and cystathionine  $\beta$ -lyase (YjcJ/MetJ). The genes encoding these two enzymes are cotranscribed in *B. subtilis* and are regulated by a SAM-responsive S-box riboswitch (45, 46), while in *L. plantarum*, *Leuconostoc mesenteroides*, *Oenococcus oeni*, and *Staphylococcus* sp., *yjcI* and *yjcJ* are regulated by a T box with a Met specifier sequence (Fig. 4). Previously, the identification of Met T boxes in biosynthetic genes had been restricted to the *Lactobacillaceae* (37, 107), but we can now extend this regulation to the staphylococci.

The final step in the Met biosynthetic pathway is catalyzed by methionine synthase. There are two classes of this enzyme, the B<sub>12</sub>-dependent class, encoded by *metH*, and the B<sub>12</sub>-independent class, encoded by *metE*. The *metE* gene is regulated by a T box with a Met specifier sequence in *Lactobacillus casei*



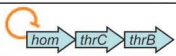


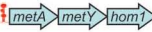








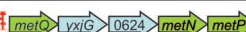











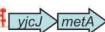




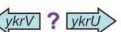
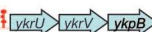
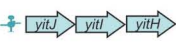




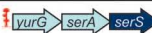
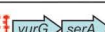
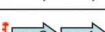
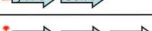


	<i>B. subtilis</i> regulation	T box regulation						
Methionine		 <i>L. pla, S. aur, S. epi</i>						
		 <i>L. pla</i>						
	 <i>B. cla, C. ace</i>	 <i>L. cas</i>						
	<table border="1" data-bbox="337 480 736 623"> <tr> <td>Riboswitches</td> <td>T box attenuation mechanism</td> </tr> <tr> <td>Posttranslational Regulation</td> <td>S box attenuation mechanism</td> </tr> <tr> <td>Unknown</td> <td>Feedback inhibition</td> </tr> </table>	Riboswitches	T box attenuation mechanism	Posttranslational Regulation	S box attenuation mechanism	Unknown	Feedback inhibition	 <i>L. pla</i>
		Riboswitches	T box attenuation mechanism					
		Posttranslational Regulation	S box attenuation mechanism					
		Unknown	Feedback inhibition					
			 <i>L. cas</i>					
		 <i>L. reu</i>						
		 <i>L. mes</i>						
		 <i>E. fae, L. bre, L. pla, L. reu, P. pen</i>						
	 <i>B. cla, B. lic, L. inn, L. mon, O. ihe</i>	 <i>L. aci, L. gas, L. joh</i>						
		 <i>L. sal</i>						
		 <i>O. oen</i>						
	 <i>L. mes</i>							
	 <i>L. del 11842, L. reu</i>							
	 <i>L. del BAA 365</i>							
	 <i>L. del BAA 365</i>							
	 <i>L. reu</i>							
 <i>B. ant, B. cer, B. thu, G. kau</i>	 <i>S. aur, S. epi, S. hae, S. sap</i>							
	 <i>L. mes</i>							
	 <i>O. oen</i>							
	 <i>L. pla</i>							
	 <i>L. cas</i>							
	 <i>L. pla</i>							
	 <i>L. reu</i>							
	 <i>L. mes</i>							
Serine		 <i>B. cla, B. hal</i>						
		 <i>C. ace</i>						
		 <i>D. red, S. wol</i>						
		 <i>C. klu</i>						
		 <i>C. bot, C. nov, C. tet</i>						
		 <i>D. haf</i>						
		 <i>C. the</i>						
								

FIG. 4. Variety of mechanisms used in regulating methionine and serine biosynthetic genes of *B. subtilis* and other bacteria. The regulatory mechanisms and operon arrangements found in *B. subtilis* (left column) are compared with those of T-box-regulated operons in other members of the *Firmicutes* (right column). Where genes of other organisms share the same regulatory mechanism as *B. subtilis*, the names of these organisms are indicated in the *B. subtilis* column. The graphic representation of each type of regulatory element is indicated in the red box in each subfigure. No attempt was made to identify pathways exhibiting feedback inhibition of enzyme activity; only those reported in the literature for *B. subtilis* are indicated. The regulatory proteins shown represent their corresponding binding sites in the operon. Genes that have not been annotated were labeled based on their corresponding COG numbers (i.e., a gene that belongs to COG1878 is drawn as an arrow containing the number “1878”). Organism abbreviations and color codes are described in the legends of Fig. 2 and 3.

and *L. plantarum* (Fig. 4), while it is regulated by a SAM-responsive S box in *B. subtilis*, *C. acetobutylicum* (45, 46, 56), *B. clausii*, *Listeria innocua*, and *Listeria monocytogenes*. The *metH* gene is present in a few members of the *Firmicutes*, and it is regulated by an S-box riboswitch in *C. acetobutylicum* and *Oceanobacillus iheyensis* as well as in *Deinococcus geothermalis*, a member of the *Deinococcus-Thermus* group. The orthologous *yxjG* and *yxjH* genes, which are related to *metE*, are regulated by an S box in *B. subtilis* (45). Orthologs of *yxjG/yxjH* are regulated by T-box elements with a Met specifier sequence in *Enterococcus faecalis* (45) and many members of the *Lactobacillales*. The conversion of homocysteine to Met also requires methylenetetrahydrofolate reductase, encoded by the S-box-regulated *yitJ* gene in *B. subtilis*. *yitJ* is cotranscribed with *metE* in *L. innocua* and *L. monocytogenes* and is regulated by the S-box mechanism, but in *L. mesenteroides*, *yitJ* is monocistronic and is regulated by a Met-responsive T-box element (Fig. 4). Met can also be synthesized by the recycling of methylthioadenosine, which is generated as a by-product of polyamine biosynthesis, in a number of organisms including *B. subtilis* (46, 77, 95). Genes involved in this pathway in *Bacillus* sp. are regulated by SAM via the S-box riboswitch (45).

Homocysteine can also be recycled via the activated methyl cycle (22). This pathway employs the LuxS protein, which recycles the toxic intermediate *S*-adenosylhomocysteine to yield homocysteine as well as 4,5-dihydroxy-2,3-pentanedione (22). 4,5-Dihydroxy-2,3-pentanedione is then spontaneously rearranged into autoinducer-2, which is a key molecule in quorum sensing (92). In some members of the *Lactobacillaceae*, *luxS* is regulated by a Met-specific T-box RNA (Fig. 4). This is the first reported example of a gene involved in quorum sensing regulated by a T-box riboswitch.

As noted above, the Cys biosynthetic gene *cysE* and the cysteinyl-tRNA synthetase gene *cysS* are cotranscribed and regulated by a T-box sequence with a Cys (UGC) codon in many members of the *Firmicutes* (35) (Fig. 5). Cys can also be synthesized by the reduction of sulfate compounds such as thiosulfate, which can be converted to *S*-sulfocysteine (46). This reaction is carried out by cysteine synthase, the product of *cysK*, which also participates in the conversion of Ser to Cys. The *cysK* gene is regulated in response to tRNA<sup>Cys</sup> by the T-box mechanism in *C. acetobutylicum*, *C. beijerinckii*, *C. botulinum*, *C. kluyveri*, *C. perfringens*, *L. plantarum*, and *Staphylococcus epidermidis* (Fig. 5). The transcription of *cysK* is regulated by the CymR DNA binding protein in *B. subtilis* (29).

Some bacterial species, including *B. subtilis*, can also synthesize Cys by the reverse *trans*-sulfuration pathway using Met as a precursor (71, 86). The two genes involved in these reactions are *yrhA* (encoding cystathionine  $\beta$ -synthase) and *yrhB* (encoding cystathionine  $\gamma$ -lyase), which are commonly found in the same transcriptional unit. In *C. acetobutylicum* (46), this operon is regulated by the T-box mechanism using a Cys specifier sequence, consistent with its role in Cys biosynthesis (Fig. 5). It is interesting that *yrhB* is closely related to the *yjcI* and *yjcJ* genes, which encode enzymes that catalyze the conversion of Cys to homocysteine and which respond to Met (or SAM) accumulation rather than to an increase in the Cys level.

### The Branched-Chain Amino Acids Isoleucine, Leucine, and Valine and Their Relationship to the Pantothenate Pathway

Pyruvate is the common precursor of the branched-chain amino acids (BCAAs) isoleucine (Ile), leucine (Leu), and valine (Val). The biosynthetic pathways for these amino acids share the first four enzymes, encoded by *ilvC*, *ilvD*, *ilvE*, and *ilvB-ilvN*, the products of which form a heterodimeric enzyme complex. In the *Firmicutes*, these genes are rarely organized in a single transcription unit: *ilvB*, *ilvN*, and *ilvC* are commonly found within the *ilvBNC-leuABCD* operon, while *ilvD* and *ilvE* are usually monocistronic. In addition, *B. anthracis*, *B. cereus*, and *B. thuringiensis* contain paralogous copies of the *ilvB*, *ilvN*, and *ilvC* genes that are cotranscribed within the *ilvEBNCDA* operon. These operons are regulated by tandem T-box riboswitches, both of which contain an Ile (AUC) specifier sequence (Fig. 6). Therefore, their expression would require a more substantial decrease in tRNA<sup>Ile</sup> charging since two molecules of uncharged tRNA<sup>Ile</sup> are necessary to promote the readthrough of both terminators. This operon could be considered as a backup when tRNA<sup>Ile</sup> charging is critically low, by analogy with the regulation of the *B. subtilis* ThrRS genes (38, 82).

In *B. subtilis* and its closest relatives, the *ilv* and *leu* genes are organized in a single operon, *ilvBNC-leuABCD*, which is preceded by a T-box leader that responds to uncharged tRNA<sup>Leu</sup> (Fig. 6). This could result in an imbalance during growth under Leu-rich conditions if Ile and/or Val were limiting. A second level of regulation uses the CodY DNA binding protein to repress *ilvB* promoter activity in response to the availability of Ile and Val. This ensures that the intracellular levels of all three BCAAs have an impact on the regulation of this operon (96). The *ilvBNC-leuABCD* operon in *B. subtilis* is also regulated by TnrA, which responds to nitrogen limitation (32), and CcpA, which is active in glucose-grown cells and prevents repression by CodY (34, 36, 109).

*C. beijerinckii* has two *ilv* transcription units belonging to the T-box regulon, *ilvH-leuACDB* (Ile specifier sequence) and *ilvBH* (Val specifier sequence). In *C. thermocellum*, *leuA* is regulated by tandem Leu T boxes (Fig. 6). The genome of this organism also contains a Leu T-box-regulated transcriptional unit annotated as specifying an uncharacterized cyclic AMP-dependent synthase and ligase; we suggest that either this gene has an unknown role in BCAA metabolism or it is incorrectly annotated. Finally, *C. kluyveri* has a large number of operons predicted to be regulated by T boxes that respond to BCAAs; most of these genes correspond to canonical BCAA biosynthetic genes, but others represent unexpected examples of T-box-regulated genes and will be discussed below.

The regulation of BCAA biosynthesis by the T-box mechanism is not restricted to the *Firmicutes*. The *leuA* genes of *Geobacter* and *Pelobacter* spp. are regulated by a Leu T box (Fig. 6). In contrast to the conclusions reported by Vitreschak et al. (107), we suggest that the DNA region responsible for this regulatory mechanism could have been acquired by horizontal gene transfer (HGT) since this would be the more parsimonious scenario considering the great phylogenetic distances between the *Firmicutes* and the *Deltaproteobacteria* (see below) (Fig. 2).

We also identified genes in *L. plantarum*, *L. reuteri*, and *L.*

	<i>B. subtilis</i> regulation	T box regulation										
Alanine	?	<i>C. hyd, C. the, C. nov, C. per, D. red, P. the, S. wol</i>										
	 Predicted to occur in all Firmicutes, except <i>C. difficile</i>	<i>C. dif</i>										
Arginine	 Predicted to occur in all Firmicutes, except <i>C. difficile</i>	<i>C. dif</i>										
		<i>C. dif</i>										
Asparagine	None	<i>L. del, L. pla, L. reu</i>										
		None										
		None										
		<i>C. klu</i>										
	?	<i>C. ace</i>										
Aspartate		<i>C. hyd, C. per</i>										
Cysteine	 <i>B. amy, B. ant, B. cer, B. cla, B. hal, B. lic, B. pum, B. thu, E. fae, G. kau, G. the, L. inn, L. mon, L. wel, O. ihe, S. aur, S. epi, S. hae, S. sap, T. ten</i>	<i>M. the</i>										
	<table border="1"> <tr> <td>Riboswitches</td> <td>  T box attenuation mechanism   gcvT element (glycine riboswitch)         </td> </tr> <tr> <td>DNA Binding Proteins</td> <td>  ArgR binding site   CymR binding site   Sporulation Sigma Factor binding site         </td> </tr> <tr> <td>Posttranslational Regulation</td> <td>  Feedback inhibition         </td> </tr> <tr> <td>Other</td> <td>  Constitutive expression   Expression is restricted to stationary phase   Expression depends on sulfate levels         </td> </tr> <tr> <td>Unknown</td> <td>?</td> </tr> </table>	Riboswitches	T box attenuation mechanism gcvT element (glycine riboswitch)	DNA Binding Proteins	ArgR binding site CymR binding site Sporulation Sigma Factor binding site	Posttranslational Regulation	Feedback inhibition	Other	Constitutive expression Expression is restricted to stationary phase Expression depends on sulfate levels	Unknown	?	<i>L. sak</i>
	Riboswitches	T box attenuation mechanism gcvT element (glycine riboswitch)										
	DNA Binding Proteins	ArgR binding site CymR binding site Sporulation Sigma Factor binding site										
	Posttranslational Regulation	Feedback inhibition										
	Other	Constitutive expression Expression is restricted to stationary phase Expression depends on sulfate levels										
	Unknown	?										
		<i>D. haf</i>										
		<i>C. tet</i>										
		<i>A. met</i>										
		<i>C. the</i>										
		<i>C. ace, S. wol</i>										
		<i>C. tet, C. nov</i>										
		<i>L. pla, C. ace, C. per, S. epi</i>										
		<i>C. bei, C. bot, C. klu, C. per</i>										
	<i>C. per</i>											
	<i>C. bot</i>											
	<i>C. ace, C. klu</i>											
	<i>C. ace</i>											
	<i>L. pla</i>											
	<i>C. per</i>											
	<i>S. sui</i>											
Glycine	 The vast majority of Firmicutes	None										
	?	<i>M. the</i>										
Threonine		<i>P. the</i>										
		<i>B. ant, B. cer, B. thu</i>										
		<i>C. klu</i>										
		<i>L. pla, S. aur, S. epi</i>										
		<i>C. dif</i>										
	<i>C. dif</i>											

FIG. 5. Variety of mechanisms used in regulating leucine, isoleucine, valine, and histidine biosynthetic genes of *B. subtilis* and other bacteria. The color code used for the different types of regulated genes and abbreviations of organisms are described in the legend of Fig. 2. The graphic representation of each type of regulatory element is described in the legend of Fig. 4.



	<b>B. subtilis regulation</b>	<b>T box regulation</b>								
<b>Leucine</b>	<p><i>B. hal</i>, <i>B. lic</i></p>	<p><i>B. amy</i>, <i>B. cla</i>, <i>B. pum</i>, <i>G. kau</i>, <i>G. the</i></p> <p><i>D. red</i></p> <p><i>A. met</i>, <i>C. dif</i></p> <p><i>C. klu</i></p> <p><i>S. wol</i></p> <p><i>C. klu</i></p> <p><i>B. pum</i></p> <p><i>C. klu</i>, <i>C. the</i></p> <p><i>C. ace</i>, <i>C. klu</i>, <i>D. haf</i>, <i>P. pro</i>, <i>G. sul</i>, <i>G. met</i>, <i>G. ura</i></p> <p><i>P. car</i></p>								
	<p>? <i>vnaJ</i> → <i>fadK</i> → <i>vnaH</i></p>	<p><i>C. the</i></p>								
	None	<p><i>S. wol</i></p>								
	None									
<b>Isoleucine</b>	<p><i>B. hal</i>, <i>B. lic</i></p> <table border="1" style="margin-left: 20px;"> <tr> <td>Riboswitches</td> <td>T box attenuation mechanism</td> </tr> <tr> <td>DNA Binding Proteins</td> <td>  TnrA   CodY   CcpA                 </td> </tr> <tr> <td>Posttranslational Regulation</td> <td>  Feedback inhibition                 </td> </tr> <tr> <td>Unknown</td> <td>?</td> </tr> </table>	Riboswitches	T box attenuation mechanism	DNA Binding Proteins	TnrA CodY CcpA	Posttranslational Regulation	Feedback inhibition	Unknown	?	<p><i>B. ant</i>, <i>B. cer</i>, <i>B. thu</i></p> <p><i>D. red</i></p> <p><i>B. ant</i>, <i>B. cer</i>, <i>B. thu</i></p> <p><i>T. ten</i></p> <p><i>O. ihe</i></p> <p><i>C. bei</i></p> <p><i>A. met</i></p> <p><i>C. nov</i></p> <p><i>C. klu</i></p> <p><i>C. klu</i></p> <p><i>A. met</i></p>
	Riboswitches	T box attenuation mechanism								
	DNA Binding Proteins	TnrA CodY CcpA								
	Posttranslational Regulation	Feedback inhibition								
	Unknown	?								
	None	<p><i>L. pla</i>, <i>L. reu</i></p>								
	<p>? <i>lvkA</i> → <i>panE</i></p>	<p><i>L. pla</i>, <i>L. reu</i>, <i>L. sal</i></p>								
	None	<p><i>C. klu</i>, <i>S. wol</i></p>								
	<p>? <i>ysiA</i> → <i>ysiB</i> → <i>etfB</i> → <i>etfA</i></p>	<p><i>C. klu</i></p>								
	None	<p><i>S. sui</i></p>								
<b>Valine</b>	<p><i>B. hal</i>, <i>B. lic</i></p>	<p><i>C. ace</i></p> <p><i>A. met</i>, <i>C. bei</i></p> <p><i>C. ace</i></p>								
	<p>? <i>lvkA</i> → <i>panE</i></p>	<p><i>L. pla</i></p>								
	<p>? <i>nusB</i> → <i>folD</i></p>	<p><i>C. klu</i></p>								
	None									
<b>Histidine</b>	<p><i>B. hal</i>, <i>B. lic</i></p>	<p><i>L. mes</i>, <i>L. pla</i></p> <p><i>L. cas</i></p> <p><i>L. lac</i>, <i>S. mut</i></p> <p><i>S. gor</i></p> <p><i>S. san</i></p>								

*salivarius* that participate in pantothenate biosynthesis and are predicted to be regulated by BCAA T boxes (Fig. 6). This finding is particularly interesting considering that the biosyntheses of Val, Leu, and pantothenate share  $\alpha$ -ketoisovalerate as a common precursor (110). One possible explanation for the use of this type of regulation for the pantothenate (*pan*) biosynthetic genes would include the sensing of a metabolic stage requiring an increased synthesis of the BCAAs and their precursors.

#### Histidine Biosynthesis: Possible Consequences of a Weak tRNA-T-Box Interaction

In almost all bacteria for which sequence information is available, including all members of the *Firmicutes*, the genes for histidine (His) biosynthesis are clustered in a single operon (4). The first reaction in the pathway, the condensation of ATP with 5-phosphoribosyl-1-pyrophosphate to form *N'*-5'-phosphoribosyl-ATP, is catalyzed by the enzyme *N'*-5'-phosphoribosyl-ATP transferase (4). This reaction is performed by an octameric enzyme complex composed of polypeptides encoded by the *hisG* and *hisZ* genes, which often are the first two genes in the operon (11, 23, 24). The majority of these *his* genes are arranged in the order *hisZGDBHAFIE*. The *his* operon is only rarely regulated by the T-box mechanism; most examples are found in members of the *Lactobacillales*, including *L. lactis* (24) and *S. mutans*, both of which show the unusual *hisCZGCSerB-hisB-ymdC-hisHAFIK* gene order (Fig. 6). In contrast, the *his* operons of *L. casei*, *L. plantarum*, and *L. mesenteroides* are regulated by T-box RNAs with a His (CAC) specifier sequence despite the presence of the *hisZGDBHAFIEC* gene organization typical of *his* operons that are not regulated by the T-box mechanism. The rarity of T-box regulation for the genes of this pathway, as well as for HisRS genes, may be due to the fact that the acceptor arm of tRNA<sup>His</sup> has only three unpaired nucleotides (CCA) at its 3' end; tRNA<sup>His</sup> therefore lacks one of the unpaired nucleotides involved in the interaction of tRNAs with the antiterminator bulge. The absence of the fourth position of pairing could potentially result in a lower efficiency of tRNA-dependent antitermination. In gram-negative bacteria, the *his* biosynthetic genes are also clustered within a *his* operon, and the transcription of this operon is regulated by transcription attenuation in response to the accumulation of uncharged tRNA<sup>His</sup>, which is sensed by the translation of sequential *his* codons in a leader peptide coding region (115). The activity of the HisG enzyme is also highly regulated by feedback inhibition (4).

#### Aromatic Amino Acid Biosynthesis: Prediction of Tight Regulation by Tandem T Boxes

Chorismate is the common precursor of all three aromatic amino acids. It can yield anthranilate, which is used to synthesize Trp, or prephenate, which is the common precursor of

phenylalanine (Phe) and tyrosine (Tyr). Several of the proteins participating in chorismate biosynthesis are encoded by genes that are located in the *aro-trp* supraoperon in *B. subtilis* and its closest relatives (40, 59, 74). In contrast, in the other members of the *Firmicutes*, these genes are generally not clustered with the biosynthetic *trp* genes, and they exhibit a very diverse chromosomal organization (Fig. 7) (52, 53).

*aroA* encodes 3-deoxy-D-arabino-heptulosonate 7-phosphate synthase, which catalyzes the first step in chorismate formation. This gene is regulated by the T-box mechanism (sensing tRNA<sup>Tyr</sup>, tRNA<sup>Phe</sup>, or tRNA<sup>Trp</sup>) or by another mechanism that is yet to be determined. Examples of *aroA* genes that are regulated by a T-box sequence are found in *B. anthracis*, *B. cereus*, and *B. thuringiensis* (Tyr specifier sequence); *C. thermocellum* and *M. thermoacetica* (Phe specifier sequence); and *T. tengcongensis* and *Dehalococcoides* sp. (Trp specifier sequence) (Fig. 7). Alternatively, examples of *aroA* genes that are regulated by Trp-activated TRAP RNA binding can be found in *S. wolfei* and *Carboxydotherrmus hydrogenoformans*. Since chorismate is the precursor of all three aromatic amino acids, a regulatory mechanism that senses only a single amino acid could result in imbalanced regulation if only one of the aromatic amino acids is available. To deal with this imbalance, some organisms have paralogous copies of *aroA* genes that may respond to other aromatic amino acids. This expectation is supported by the genomic context of the *aroA* paralogs, which are clustered with *phe* biosynthetic genes in *S. wolfei* and *C. hydrogenoformans* and with *tyr* biosynthetic genes in *C. hydrogenoformans*, *C. thermocellum*, and *M. thermoacetica*. The number of *aroA* genes per organism varies, from one in *Bacillus* sp. to five in *S. wolfei*. In *S. wolfei*, one copy of *aroA* is cotranscribed with *pheA* and is regulated by a Phe (UUC) T box (Fig. 7), while a second copy of *aroA* is cotranscribed with the *trp* biosynthetic genes and is regulated by Trp-activated TRAP.

AroF catalyzes the last common step in the aromatic pathway, the reaction that results in the synthesis of chorismate. In *B. anthracis*, *B. cereus*, and *B. thuringiensis*, the gene that encodes this enzyme is part of the *aroF-hisC-tyrA-aroE* operon, which, like *aroA*, is regulated by a T-box sequence that responds to tRNA<sup>Tyr</sup>. As noted above, this could result in a deficiency in the sensing of Trp and Phe. The regulation of expression of the *aroF* gene has not yet been described for the remaining members of the *Firmicutes*, with the exception of *B. subtilis*, where *aroF* is in the *aro-trp* supraoperon.

Chorismate can be converted to prephenate by the aromatic aminotransferase AroH. Prephenate can be converted to tyrosine and phenylalanine by the enzymatic reactions carried out by the PheA/TyrA, HisC, and HisH enzymes. The transcription of the genes encoding these enzymes is regulated by different schemes among the *Firmicutes*. For example, in *B. anthracis*, *B. cereus*, and *B. thuringiensis*, the transcription of *tyrA* is regulated by a Tyr (UAC) T-box RNA, whereas in *D. hafniense* and *S. wolfei*, *pheA* is regulated by a Phe (UUC)

FIG. 6. Variety of mechanisms used in regulating alanine, arginine, asparagine, aspartate, cysteine, glycine, and threonine biosynthetic genes of *B. subtilis* and other bacteria. The color code used for the different types of regulated genes and abbreviations of organisms are described in the legend of Fig. 2. The graphic representation of each type of regulatory element is described in the legend of Fig. 4.

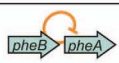
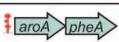




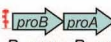
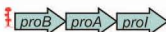


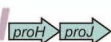




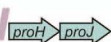
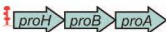






















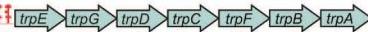


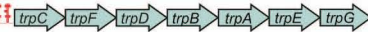
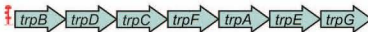












	<b>B. subtilis regulation</b>	<b>T box regulation</b>														
<b>Phenylalanine</b>		 <i>P. the</i> , <i>S. wol</i>														
	None	 <i>D. haf</i>														
		 <i>P. the</i>														
		 <i>D. red</i>														
<b>Proline</b>	 <i>B. amy</i> , <i>B. ant</i> , <i>B. cer</i> , <i>B. lic</i> , <i>B. thu</i> , <i>G. the</i> , <i>O. ihe</i>	 <i>B. cla</i> , <i>B. hal</i>														
	 <i>B. amy</i> , <i>B. lic</i> , <i>B. pum</i>	 <i>A. met</i>														
		 <i>A. met</i>														
		 <i>G. kau</i> , <i>G. the</i>														
		 <i>O. ihe</i>														
		 <i>D. haf</i>														
		 <i>D. red</i>														
None	None															
<b>Tryptophan</b>	<table border="1" data-bbox="250 1073 697 1410"> <tr> <td>Riboswitches</td> <td>T box attenuation mechanism</td> </tr> <tr> <td>RNA Binding Protein</td> <td> Tryptophan RNA Binding Protein (TRAP) binding site</td> </tr> <tr> <td>Posttranslational Regulation</td> <td> Feedback inhibition</td> </tr> <tr> <td rowspan="3">Other</td> <td> Regulation under osmotic stress conditions</td> </tr> <tr> <td> Regulation during vegetative growth</td> </tr> <tr> <td> Expressed during sporulation</td> </tr> <tr> <td>Unknown</td> <td> Expression depends on sulfate levels</td> </tr> <tr> <td>Unknown</td> <td>?</td> </tr> </table> <p> <i>B. amy</i>, <i>B. cla</i>, <i>B. hal</i>, <i>B. lic</i>, <i>B. pum</i>, <i>B. ste</i>, <i>G. kau</i></p>	Riboswitches	T box attenuation mechanism	RNA Binding Protein	 Tryptophan RNA Binding Protein (TRAP) binding site	Posttranslational Regulation	 Feedback inhibition	Other	 Regulation under osmotic stress conditions	 Regulation during vegetative growth	 Expressed during sporulation	Unknown	 Expression depends on sulfate levels	Unknown	?	 <i>A. met</i> , <i>C. ace</i> , <i>C. bei</i> , <i>C. klu</i> , <i>L. pla</i> , <i>S. gor</i> , <i>S. mut</i> , <i>S. pne</i> , <i>S. san</i>
		Riboswitches	T box attenuation mechanism													
		RNA Binding Protein	 Tryptophan RNA Binding Protein (TRAP) binding site													
		Posttranslational Regulation	 Feedback inhibition													
		Other	 Regulation under osmotic stress conditions													
			 Regulation during vegetative growth													
			 Expressed during sporulation													
		Unknown	 Expression depends on sulfate levels													
		Unknown	?													
		 <i>T. ten</i>														
		 <i>B. ant</i> , <i>B. cer</i> , <i>B. thu</i> , <i>O. ihe</i> , <i>S. aur</i> , <i>S. epi</i> , <i>S. hae</i> , <i>S. sap</i> , <i>L. mon</i> , <i>L. inn</i> , <i>L. wel</i>														
		 <i>D. BAV1</i> , <i>D. CBDB1</i> , <i>D. eth</i>														
		 <i>S. mut</i> , <i>S. the</i>														
		 <i>D. haf</i>														
 <i>L. mes</i>																
 <i>L. cas</i>																
 <i>L. lac</i>																
 <i>C. the</i>																
 <i>C. nov</i>																
 <i>C. klu</i> , <i>C. nov</i> , <i>D. haf</i> , <i>S. gor</i> , <i>S. san</i> , <i>D. BAV1</i> , <i>D. CBDB1</i> , <i>D. eth</i> , <i>R. SM1</i> , <i>R. cas</i>																
 <i>S. san</i>																
 <i>L. cas</i>																
 <i>C. dif</i>																
 <i>S. sui</i>																
 <i>B. hal</i> , <i>B. lic</i>	 <i>B. ant</i> , <i>B. cer</i> , <i>B. thu</i>															
None	 <i>B. ant</i> , <i>B. cer</i> , <i>B. thu</i>															

FIG. 7. Variety of mechanisms used in regulating phenylalanine, proline, tryptophan, and tyrosine biosynthetic genes of *B. subtilis* and other bacteria. The color code used for the different types of regulated gene and abbreviations of organisms are described in the legend of Fig. 2. The



T-box RNA, and in *L. casei*, *L. lactis*, *L. plantarum*, and *S. mutans*, *hisC* and *hisH* (which encode the enzymes that catalyze some of the last steps of the His, Tyr, and Phe pathways) are located within the *his* operon and are regulated by a His (CAC) T-box RNA (Fig. 6 and 7).

In *B. anthracis*, *B. cereus*, and *B. thuringiensis*, there is an additional route for the synthesis of Tyr from Phe via the participation of the bipterin compound. This reaction is performed by enzymes encoded in the *pah-dcoH* operon, which is regulated by tandem Tyr (UAC) T-box RNAs that are predicted to respond to tRNA<sup>Tyr</sup> (Fig. 7).

After the common aromatic pathway diverges, seven enzymatic reactions lead to Trp biosynthesis. The regulation of this pathway is particularly important due to the high energetic cost of Trp synthesis. The *trp* operon of the gram-negative bacterium *E. coli* has been thoroughly analyzed, and it has been established that the regulatory mechanisms used sense both tRNA<sup>Trp</sup> charging and the intracellular level of free Trp (reviewed in reference 118). These two signals are also sensed by the gram-positive bacterium *B. subtilis* but by regulatory mechanisms different from those used by *E. coli* (74, 118). In *B. subtilis*, free Trp inhibits *trp* operon expression by activating the TRAP protein, which binds to *trp* leader RNA and promotes transcription termination (39, 40). The level of charged tRNA<sup>Trp</sup> is sensed in *B. subtilis* by the *rtpA-ycbK* operon (15, 16, 104, 105). The first gene of this operon encodes an anti-TRAP (AT) protein, which can inhibit Trp-activated TRAP and prevent it from terminating transcription in the *trp* operon (105). The expression of the AT operon is transcriptionally regulated by uncharged tRNA<sup>Trp</sup> via the T-box mechanism (104, 105). In *B. subtilis*, AT synthesis is also regulated translationally by a leader peptide-coding region containing three Trp codons (15, 16). In this case, whenever the charged tRNA<sup>Trp</sup> level is insufficient to allow the rapid translation of the three leader peptide Trp codons, the ribosome synthesizing the leader peptide stalls, exposing the *rtpA* SD region for an efficient initiation of translation of the *rtpA* coding region. When there is sufficient Trp to allow the ribosome synthesizing the leader peptide to reach its stop codon, this ribosome blocks the adjacent *rtpA* SD sequence, inhibiting AT synthesis (16). Although AT has been found only in *B. subtilis*, *Bacillus amylo-liquefaciens*, and *B. licheniformis*, the distribution of TRAP is more widespread (reviewed in reference 53). It is not clear whether organisms that use TRAP but not AT have an alternate mechanism for sensing the level of charged tRNA<sup>Trp</sup>.

In contrast to the regulatory mechanisms used by *B. subtilis* and its close relatives, the *trp* biosynthetic genes of the vast majority of the *Firmicutes* are generally concerned solely with Trp biosynthesis and are organized as a single operon, which is regulated by a T-box element in response to the accumulation of uncharged tRNA<sup>Trp</sup> (52, 53) (Fig. 7). It is not known if these organisms can sense free Trp as a regulatory signal. The *trp* biosynthetic genes of *C. thermocellum* are organized into two different operons. This constitutes an interesting example of

the coordinate regulation of the Trp biosynthetic pathway genes by two different regulatory mechanisms: TRAP, sensing L-Trp, regulates the *trpEGDCF* operon, and a T box, sensing tRNA<sup>Trp</sup>, regulates the *trpBA* operon (Fig. 7). The regulatory elements of these two operons were incorrectly assigned in our previous genome analysis of the evolution and regulation of the *trp* biosynthetic genes (74).

Tandem T boxes are found in the *trp* operons of several members of the *Firmicutes* (52, 53). The presence of tandem T boxes implies that these organisms require the accumulation of a higher relative level of uncharged tRNA<sup>Trp</sup> to allow an appreciable expression of the *trp* operon, since the binding of multiple tRNA<sup>Trp</sup> molecules is needed for transcriptional readthrough (52). Interestingly, the *trp* operon of these species is a discrete Trp biosynthetic unit containing only the seven Trp pathway genes, unlike the *aro* supraoperon, which contains only six of these genes in combination with other aromatic amino acid biosynthetic genes. This gene organization is consistent with a tight, specific response to tRNA<sup>Trp</sup>.

Several organisms appear to lack crucial portions of the Trp biosynthetic pathway. The facultative pathogen *C. novyi* has only the *trpB* gene, which encodes a polypeptide that catalyzes the last reaction in Trp biosynthesis. Although in this organism, the *trp* pathway is incomplete, this gene is in an operon regulated by a Trp T box (Fig. 7). It seems likely that this organism acquires indole from its host and converts it to Trp by the action of TrpB. The lactic acid bacterium *L. casei* lacks the *trpE* and *trpG* genes, which is consistent with its habitat in the human gut and mouth, where it presumably can find an adequate supply of Trp precursors. The *trpDCFBA* operon of this organism is regulated by a Trp T box (Fig. 7).

T-box elements were also found upstream of the *trp* genes in members of the group *Chloroflexus*. In *Dehalococoides* sp., T-box regulation was observed in the *trp* operon as well as in a *trpB* paralog, which encodes an alternative tryptophan synthetase beta subunit (53). On the other hand, in *Roseiflexus* sp., only the *trpB* paralog gene is regulated by a Trp T box, and the mechanism used for the regulation of the *trp* operon is unknown. Several members of the *Firmicutes* including *C. kluyveri*, *C. novyi*, *D. hafniense*, *Streptococcus gordonii*, and *Streptococcus sanguinis* also have a *trpB* paralog that is regulated by a Trp T box (Fig. 7).

#### Biosynthetic Genes for Aspartate and Asparagine, Key Precursors of Many Other Amino Acids

The Krebs cycle intermediate oxaloacetate is the common precursor of Asp and asparagine (Asn). Oxaloacetate is converted to Asp by an aspartate aminotransferase, the product of the *aspB* gene. In *B. subtilis*, the expression of *aspB* does not respond to the presence of Asp and appears to be constitutive (8, 63). All low-G+C gram-positive bacteria have multiple genes encoding proteins similar to AspB (8), although most of these genes have not been characterized biochemically. The

---

graphic representation of each type of regulatory element is described in the legend of Fig. 4. TRAP, in addition to transcriptionally regulating the *trp* operon in *B. subtilis* and its closest relatives, can also regulate *trpE* translation by binding to the *trpE* leader RNA and promoting the formation of a secondary structure that sequesters the SD sequence, inhibiting translation initiation (73).

*aspB* genes and the majority of their paralogs are not regulated by a T-box mechanism, although in *C. hydrogenoformans* and *C. perfringens*, *aspB* is cotranscribed with *asnC* (which encodes a putative transcriptional regulator) in an operon preceded by an Asp T-box sequence (Fig. 5). Since Asp is a key metabolite in the synthesis of many other amino acids, such as Asn, lysine (Lys), Thr, Met, and Ile, it is not surprising that the genes responsible for its synthesis are often expressed constitutively.

Asp can also be synthesized by the action of asparaginase, encoded by the genes of the *ansAB* operon (8, 101). This operon is expressed during vegetative growth in *B. subtilis* but does not contribute significantly to Asp synthesis (8). Its regulation does not depend on the T-box mechanism, but instead, it is subject to strong repression by the Asn-responsive transcriptional regulator AsnR during the early stages of sporulation (102).

In the *Firmicutes*, Asn can be synthesized from Asp, either in its free form or after it is charged onto tRNA<sup>Asn</sup> (by AspRS), or from glutamine (Gln). In the first strategy, Asn is synthesized from Asp by an asparagine synthetase, AsnA (8). In *Lactobacillus delbrueckii* subsp. *bulgaricus* (66), *L. plantarum*, and *L. reuteri*, *asnA* is cotranscribed with *asnS* (encoding AsnRS) and is regulated by an Asn T box, whereas in *B. anthracis*, *B. cereus*, *B. thuringiensis*, *C. perfringens*, *C. tetani*, *Lactobacillus acidophilus*, *Lactobacillus brevis*, *Lactobacillus gasserii*, *Lactobacillus johnsonii*, and *Pediococcus pentosaceus*, *asnA* is monocistronic and regulated by an Asn T-box element (Fig. 5).

The second route to Asn synthesis is by the transamidation of Asp. Using this mechanism, a nondiscriminating AspRS charges Asp not only onto tRNA<sup>Asp</sup> but also onto tRNA<sup>Asn</sup>. Subsequently, a tRNA-dependent Asp-tRNA<sup>Asn</sup> amidotransferase converts Asp to Asn to form Asn-tRNA<sup>Asn</sup> (61, 62). This heterotrimeric enzyme, encoded by the *gatCAB* genes, also carries out the transamidation of Glu-tRNA<sup>Gln</sup> to Gln-tRNA<sup>Gln</sup> (61). In *C. acetobutylicum*, the *gatCAB* genes are cotranscribed with the gene encoding the nondiscriminatory AspRS in a transcription unit regulated by an Asn T box, a regulatory pattern that is consistent with the role of this operon in the generation of Asn-tRNA<sup>Asn</sup> (Fig. 5).

Asn can also be synthesized from Gln by a glutamine amidotransferase. Three enzymes of this class, AsnB, AsnH, and AsnO, have been found in *B. subtilis* (8, 120), and none of these is regulated by the T-box mechanism; in contrast, *asnB* in *C. kluyveri* is regulated by an Asn T box (Fig. 5). In *B. subtilis*, *asnB* and *asnH* are expressed almost constitutively, although the level of *asnH* expression decreases somewhat in response to the accumulation of excess Asn. *asnO* expression is restricted to stationary phase and is dependent on the  $\sigma^E$  sporulation sigma factor (8, 120). In other members of the *Firmicutes*, the mechanism of regulation of these genes has not yet been described.

### Alanine Biosynthesis Involves T-Box Regulation of Operons Containing Biosynthetic and Regulatory Genes

Alanine (Ala) is synthesized by the transamination of pyruvate. In several members of the *Clostridium* group, the aminotransferase gene *alaT* is cotranscribed with *alaR*, a putative transcriptional regulator of the Lrp/AsnC family (8) (Fig. 5) (see below). The *alaRT* operon in these organisms contains a T-box sequence with an unusual Ala specifier sequence (GCA

or GCC), in contrast to most other tRNA<sup>Ala</sup>-regulated T-box genes that contain a GCU specifier sequence. The common use of GCU is an exception to the preference for codons ending in C for most tRNA classes. We note that in several of these organisms, *alaT* is incorrectly annotated as *aspB* (*alaT* and *aspB* are 38% identical). The identification of the specifier sequence of the regulatory sequence, and recognition of the similarity to *B. subtilis* *alaR-alaT* (8), permits us to assign these genes to the alanine biosynthetic pathway (in agreement with data described previously by Vitreschak et al.) (107).

### Threonine Biosynthesis

In all bacteria, the biosynthesis of Thr from Asp involves five steps that are catalyzed by enzymes encoded by the *thrD*, *asd*, *hom*, *thrB*, and *thrC* genes (8, 78, 79). As previously mentioned, the *hom* gene also participates in the biosynthesis of Met (46), and this gene is regulated either by the S-box riboswitch mechanism in response to SAM or by a T-box mechanism that responds to tRNA<sup>Thr</sup> or tRNA<sup>Met</sup>. S-box-regulated *hom* genes are found in *B. clausii*, *B. halodurans*, *D. hafniense*, and *T. tengcongensis*, while *hom* genes with Thr T-box regulation are in *C. difficile* (46), *C. kluyveri*, *B. anthracis*, *B. cereus*, and *B. thuringiensis*. In the last three organisms, *hom* is cotranscribed with *thrB* and *thrC*, and the operon is preceded by two tandem T boxes, each responding to tRNA<sup>Thr</sup> (Fig. 5). In addition, Hom activity can be repressed by feedback inhibition in response to the presence of methionine, isoleucine, and possibly threonine, as has been shown for *B. subtilis* (119).

### Proline Biosynthesis

In *B. subtilis*, proline (Pro) is synthesized from Gln in three enzymatic steps (8). The first reaction is carried out by ProB (and its paralog, ProJ), which catalyzes the conversion of Gln to  $\gamma$ -glutamyl phosphate. The second reaction involves the synthesis of  $\Delta^1$ -pyrroline 5-carboxylate by the action of ProA. The two genes encoding these proteins are cotranscribed, and their transcription is regulated by the T-box mechanism using a Pro (CCC or CCU) specifier sequence in *B. clausii*, *B. halodurans*, *B. licheniformis*, *B. subtilis* (17), *B. anthracis*, *B. cereus*, *B. thuringiensis*, and *D. hafniense* (Fig. 7). The third reaction in Pro synthesis is catalyzed by  $\Delta^1$ -pyrroline 5-carboxylate reductase, encoded by *proC*. In addition to ProC, three other proteins (ProI, ProG, and ProH) carry out a ProC-like function. The activity of any of these enzymes is sufficient for Pro biosynthesis (8). In *B. clausii*, *B. halodurans*, and *D. hafniense*, *proA* and *proB* are organized with *proI* in a single operon, whereas in *B. anthracis*, *B. cereus*, and *B. thuringiensis*, the *proBA* operon is transcribed divergently from *proC* (3, 9). *proI* is monocistronic in some members of the *Firmicutes* and is regulated by a Pro (CCU) T box in *B. licheniformis*, *B. pumilus*, and *B. subtilis* (Fig. 7). The *proHJ* operon is transcribed under osmotic stress conditions from a  $\sigma^A$ -type promoter in *B. subtilis*. Its transcription guarantees the synthesis of a higher intracellular level of Pro, presumably for use as an osmoprotectant (12). The *comER* gene, a homolog of *proG*, is expressed during vegetative growth (54) and also during sporulation from a  $\sigma^E$ -type promoter (30). The complex regulation of the *pro* genes in the *Firmicutes* contrasts with their constitutive expres-



sion in *E. coli* (8) and may relate to the dual role of Pro in protein synthesis and osmoprotection in the *Firmicutes*.

### Regulation of Arginine Biosynthesis in the *Firmicutes* Is Mediated Predominantly by a DNA Binding Transcriptional Repressor Protein

The *B. subtilis* genes that are responsible for synthesizing arginine (Arg) from glutamate (Glu) are organized into two operons, *argCJBD-carAB-argF* and *argGH*. Both operons are commonly regulated by the ArgR DNA binding transcriptional repressor protein, which uses Arg as its corepressor (8). *argR/ahrC*-like genes have been detected in the genomes of all low-G+C gram-positive bacteria except *C. difficile* (8). In this organism, Arg (AGA) T-box riboswitches were identified in the regulatory regions of the *argCJBMF*, *argG*, and *argH* transcriptional units (Fig. 5). It is unclear why *C. difficile* is unique in its use of the T-box mechanism for the regulation of Arg biosynthesis. Analysis of additional genomes may reveal other organisms in this group.

### Amino Acid Biosynthetic Pathway Genes That Are Not Regulated by the T-Box Mechanism

Analysis of 559 fully sequenced bacterial genomes failed to detect T-box regulation for operons involved in the biosynthetic pathways for Lys, Gln, and Glu. In these pathways, gene regulation is performed by a broad repertoire of regulatory mechanisms that include (i) DNA binding transcriptional regulatory proteins, e.g., the GltC repressor used for regulating operons involved in Glu biosynthesis (10); (ii) metabolite binding riboswitches, e.g., the Lys-responsive L box used for regulating operons of Lys biosynthesis (48, 85); and (iii) posttranslational regulation by feedback inhibition, as occurs in the regulation of synthesis of glutamine synthetase, a key enzyme in the Gln biosynthetic pathway (26, 93). It is possible that examples of T-box-regulated genes for these pathways will be identified as additional genome sequences become available.

## REGULATION OF AMINO ACID TRANSPORTER GENES

Many genes encoding proteins involved in amino acid transport have been identified in operons regulated by the T-box mechanism. As these transport proteins increase the intracellular pool of their respective amino acids, regulation in response to tRNA charging provides an attractive mechanism for coupling the expression of a transporter gene to intracellular pools of its substrate. In the present study, we identified 34 different families of orthologous genes (COGs) encoding amino acid transporters that are regulated by a T-box sequence. T-box regulation of transporter genes appears to occur exclusively in the *Firmicutes* (Fig. 2 and 8).

Most of the transporters whose synthesis is regulated by the T-box mechanism are annotated as BCAA transporters or as members of the ABC transporters. The majority of the transporter gene operons possess only one T-box sequence, although there are a few examples with tandem T-box sequences (Fig. 8). These transporter genes are found in either monocistronic or polycistronic operons, where the genes presumably encode different subunits of the same transmembrane protein complex. Transporter genes are occasionally cotranscribed with

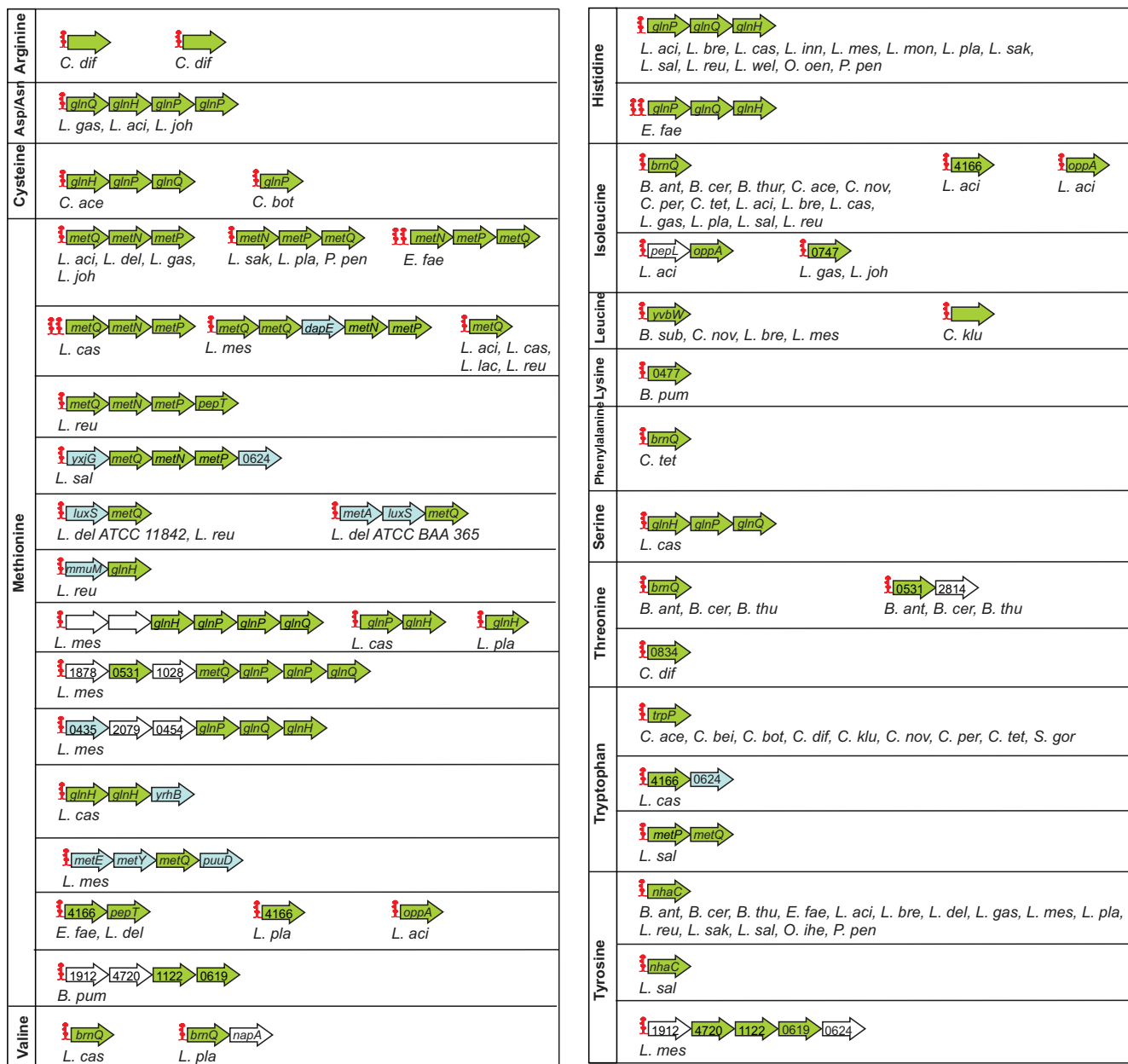
biosynthetic genes. This has been observed for genes of the Met biosynthetic pathway in some members of the *Lactobacillales*. For example, the *metQ* transporter gene is cotranscribed with *metA*, which encodes a homoserine *trans*-succinylase, in *L. delbrueckii*. This operon also contains a gene that encodes the LuxS S-adenosyl homocysteine recycling protein (Fig. 5).

Analysis of the amino acid sequence of a transporter protein does not always provide sufficient information to allow the prediction of its substrate. The identification of a T-box sequence (and its corresponding specifier sequence) upstream of a transporter gene therefore provides valuable information not only on which uncharged tRNA regulates the expression of the gene but also for the prediction of the amino acid likely to be transported. For example, the *yvbW* gene of *B. subtilis*, annotated as a “hypothetical protein,” is related to the “gamma-aminobutyrate permease” group of genes. This gene has a T-box leader sequence with a Leu (CUC) specifier sequence. Consistent with the prediction that *yvbW* encodes a transporter of Leu or a related compound, *yvbW* expression was shown to be induced upon Leu limitation (87). Some recent reports have annotated the specificity of a large group of amino acid transporters that had previously been poorly characterized and were classified only as “hypothetical proteins” or “BCAA permeases” (107, 111). We further characterized the T-box regulation of these groups of amino acid transporters and observed that ~70% of these genes are regulated by tRNA<sup>Ile</sup>, while ~15% respond to tRNA<sup>Thr</sup>; the remainder are regulated by tRNA<sup>Leu</sup>, tRNA<sup>Val</sup>, and tRNA<sup>Phe</sup>. This suggests that while most of these genes are likely to encode BCAA permeases, other members of this group may transport other substrates.

The predictive power of a T-box leader sequence can be used to identify incorrect gene annotations. For example, an orthologous operon of uncharacterized proteins present in members of the *Lactobacillales* contains genes related to ABC-type transporters. These operons are regulated by a T-box element with a His (CAC) specifier sequence, suggesting that the product is a His transporter. In *L. plantarum*, the genes of this operon are annotated in the GenBank database as Gln transporters (*glnPQH*), which is likely to be incorrect. We also identified a number of other transporter genes misannotated as *glnPQH* homologs; we predict from their T-box leader specifier sequences that they are involved in the transport of Asp, Asn, Phe, Cys, Met, and Ser (Fig. 8). Correcting errors of this type is important, as further annotation is likely to lead to the repeated misannotation of related genes in other genomes.

### Shared Regulatory Mechanisms for Biosynthetic and Transporter Genes

Comparisons of the mechanisms employed in regulating gene expression have revealed that in many cases, a common regulatory mechanism is used (e.g., involving a T-box, L-box, or SAM riboswitch or the TRAP RNA binding protein) by an organism to regulate both the corresponding biosynthetic genes and amino acid transporter genes (Fig. 9). For example, in *B. subtilis* and closely related species, both the *trp* operon and the *trpP* tryptophan transporter gene are regulated by the TRAP RNA binding protein (90). However, in *C. acetobutylicum*, *C. beijerinckii*, *C. kluyveri*, *C. novyi*, and *S. gordonii*, the *trp* biosynthetic operon and *trp* transporter gene are regulated by



**T box attenuation mechanism**

FIG. 8. Genes involved in amino acid transport that are regulated by the T-box mechanism. The common designation for each class of transporter gene is shown inside each arrow. Genes that have not been annotated were labeled based on their corresponding COG numbers (i.e., a gene that belongs to COG4166 is drawn as an arrow with the number “4166”). Note that genes are named in accordance with the GenBank annotation and might not represent the real specificity of the transporter as revealed by the identification of the specifier codon in our T-box analysis. Organism abbreviations and gene color codes are described in the legends of Fig. 2 and 3.

the T-box mechanism. A second example can be found in *C. difficile*, which regulates both Arg biosynthetic and transporter genes by the T-box mechanism. In contrast, in other organisms, the T-box mechanism is used only for the ArgRS gene, and the biosynthetic and transporter genes are regulated by the ArgR transcriptional repressor. Similarly, in a large number of members of the *Firmicutes*, both the Lys biosynthetic operon and the Lys transport operon are regulated by a lysine-responsive

L-box riboswitch (1, 48, 85), and Met biosynthetic and transport operons are regulated by a SAM-responsive S-box riboswitch (37, 58, 86) (Fig. 9).

Due to the prevalence of the T box as a regulatory element in the *Firmicutes*, the most common use of shared regulation by biosynthetic and transporter genes involves the T-box mechanism. All of these examples may reflect the need of the organism to acquire the missing amino acid by coordinating biosyn-

	Lysine		Methionine		Tryptophan	
	Biosynthesis	Transport	Biosynthesis	Transport	Biosynthesis	Transport
<i>Clostridia</i>						
<i>B. subtilis</i> group						
<i>B. cereus</i> group						
<i>Lactobacilli</i>						
<i>Staphylococci</i>						
<i>Streptococci</i>						

Riboswitches		S box attenuation mechanism
		S <sub>MK</sub> box attenuation mechanism [1]
		T box attenuation mechanism
		L box (also called LYS element)
RNA Binding Protein		Tryptophan RNA Binding Protein (TRAP)
Protein Transcription Factors		MtaR/MetR/CmbR [2]

FIG. 9. Common strategy in regulating amino acid transporter genes and biosynthetic genes. Both amino acid transporters and biosynthetic enzymes can fulfill the need of an organism for certain amino acids. This results in a tendency to coordinate the expression of the genes for these classes of proteins by using a shared regulatory mechanism. As shown for three amino acid-related genes, this tendency is specific for each phylogenetic clade. The regulatory elements were identified using our Riboswitch Web server (RibEx) (1) and previously reported data (2). [1], the S<sub>MK</sub> riboswitch regulates *metK* genes in lactic acid bacteria, including *Enterococcus*, *Streptococcus*, and *Lactococcus* spp. (see reference 33). [2], in streptococci, unlike other members of the *Firmicutes*, methionine biosynthesis and transport are controlled by protein transcription factors, in this case, MtaR, MetR, and CmbR (see reference 67). Organism names and gene color codes are described in the legends of Fig. 2 and 3.

thesis with transport. It is also possible that T-box riboswitches that regulate different classes of genes in the same organism in response to the same effector tRNA may be differentially sensitive to the level of that uncharged tRNA and thereby allow differential regulation. Differential responses to SAM pools (which correlate with variability in the affinity of the riboswitch RNA for SAM) were observed for S-box-regulated genes in *B. subtilis*; this results in the expression of genes encoding a Met transporter when SAM pools are relatively high, whereas the expression of genes involved in Met biosynthesis remain repressed until SAM pools drop to a lower level (103). A similar variability in affinity for the cognate uncharged tRNA could allow an increased synthesis of the corresponding transporter prior to the induction of the genes for the full biosynthetic pathway. The possibility of differential regulation using the T-box mechanism requires further experimental study.

#### REGULATION OF SYNTHESIS OF REGULATORY PROTEINS

The expression of several operons encoding regulatory proteins is regulated by the T-box regulatory mechanism. In these examples, T-box-mediated control has a broader effect, as it influences the expression of all of the genes responding to each regulatory protein. Almost all regulatory genes predicted to be controlled by a T-box sequence are cotranscribed with an aaRS gene or with genes involved in amino acid synthesis or transport (Fig. 10). Thus, in *L. innocua*, *L. monocytogenes*, and *Listeria welshimeri*, a gene annotated as a “regulator of competence-specific genes” is cotranscribed with *serS* under the control of a Ser T-box sequence. The function of this putative regulatory protein is unknown, but it is likely that it plays a role in the regulation of serine metabolism. Similarly, in *O. oeni*, a gene annotated as a putative regulator is cotranscribed with *pheST* (encoding PheRS) and is regulated by a Phe T box. Examples of regulatory proteins that are cotranscribed with

biosynthetic genes are found in the Ala T-box-regulated *alaRT* operon in members of the *Clostridium* group. The *alaR* gene encodes a transcriptional regulator of the Lrp/AsnC family, members of which are involved in modulating a variety of metabolic functions including the catabolism and anabolism of certain amino acids (13). The second gene of the operon, *alaT*, encodes an alanine transaminase, which can synthesize Ala from pyruvate (these genes are often misannotated as *asnC aspB*) (see above). In *B. pumilus*, an AraC family transcriptional regulator is cotranscribed with the biosynthetic gene *leuA* and is regulated by a Leu (CUC) T-box region (Fig. 10). Members of the AraC family regulate genes involved predominantly in carbon source catabolism, the stress response, and virulence (60). An AraC family protein was reported to affect the regulation of a Val biosynthetic operon in *Pseudomonas aeruginosa* (100), which is consistent with our finding of an AraC-like regulator involved in the biosynthesis of Leu. In *Roseiflexus* sp., a monocistronic gene annotated as a “putative transcriptional regulator of the MerR family” is regulated by a Phe (UUC) T box. In this case, the presence of the T-box regulatory sequence provides the only clue relating this gene to Phe metabolism.

As noted above, the T-box mechanism plays an important secondary role in the regulatory events that modulate the synthesis of Trp in *B. subtilis*. A Trp T-box sequence regulates the transcription of *rtpA*, the gene encoding the AT regulatory protein, which modulates the activity of the major Trp-activated regulatory protein, TRAP, in response to the accumulation of uncharged tRNA<sup>Trp</sup> (Fig. 10) (105). AT synthesis is also influenced by the translation of the three Trp codons in the coding region for a 10-residue regulatory leader peptide located in the *at* operon (16). Hence, a T-box regulatory region that senses the level of uncharged tRNA<sup>Trp</sup> helps to define the overall fate of the regulatory cascade that modulates *trp* operon expression. The *rtpA* gene, encoding the AT regulatory protein, is cotranscribed with the *ycbK* gene, which encodes a presumed tryptophan transport protein, providing another ex-

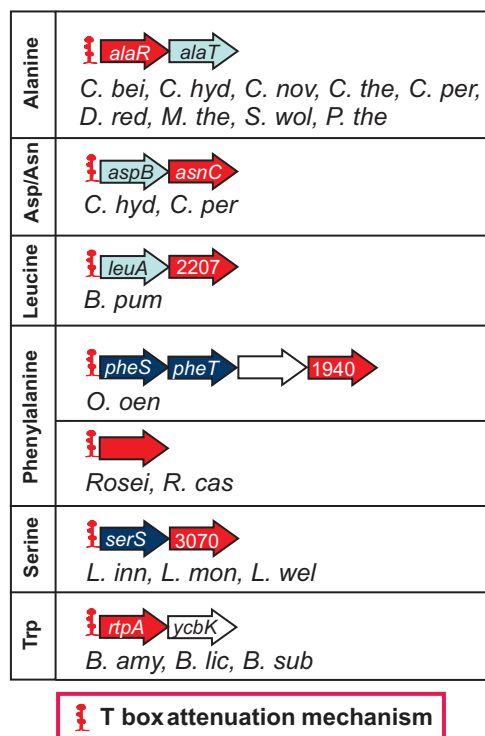


FIG. 10. Regulatory genes controlled by the T-box mechanism. Organism abbreviations and color codes are described in the legends of Fig. 2 and 3. COG1940 is annotated as a “negative regulator of the xylose operon”; this annotation does not correspond to the function deduced from its Phe T-box specificity. COG3070 corresponds to a “regulator of competence-specific genes”; its true function is unknown, but it is predicted to be related to its Ser specifier sequence and the fact that is cotranscribed with *serS*. COG2207 is the family of AraC transcriptional regulators. The *Roseiflexus* (*Rosei*) sp. regulatory genes do not belong to any COG family but are annotated in GenBank as “putative transcriptional regulators, MerR family.”

ample of the coordinate regulation of amino acid biosynthesis and transport (91).

#### OTHER IMPORTANT FEATURES OF THE T-BOX MECHANISM

##### *ileS* Is the Gene Most Widely Regulated by the T-Box Mechanism

In our computer analysis of 559 genome sequences, *ileS*, encoding IleRS, was the gene most often predicted to be regulated by the T-box mechanism. Our search identified T-box-regulated *ileS* genes in 57 members of the *Firmicutes*, 19 members of the *Actinobacteria*, 5 members of the *Chloroflexi*, and 2 members of the *Deinococcus-Thermus* group. The presence of a T-box region preceding *ileS* is most evident in some groups of the pathogenic *Firmicutes* (organisms that often lack genes for amino acid biosynthesis and rarely use the T-box regulatory mechanism) and in members of the *Actinobacteria*, which have predominantly only one T-box-regulated gene per genome. In each of these groups, the few T-box-regulated genes include *ileS* (Fig. 2 and 3).

#### Over- and Underrepresentation of T-Box Regions in Genomes

Among the organisms for which complete genomes are currently available, *B. thuringiensis* appears to employ the T-box mechanism most widely (Fig. 2). We identified 40 putative T-box-regulated operons in its genome; these represent more than 1% of its transcriptional units. The closely related species *B. anthracis* and *B. cereus* also appear to use the T box frequently, with 38 putative T boxes in their genomes. A likely explanation for the abundance of T-box-regulated genes in these organisms is their expanded capacity for amino acid and peptide uptake. *B. thuringiensis* and its closest relatives appear to encode a large number of ABC-type peptide binding proteins and BCAA transporters. This property may reflect the adaptation of these organisms to a protein-rich environment, as exists in decaying animal matter (84).

#### Single versus Tandem T-Box Elements

A T-box regulatory region is generally present as a single unit. However, as noted above, there are several examples where tandem T-box sequences are evident, most commonly upstream of biosynthetic operons. Within this group, the operons encoding the enzymes of the *trp* pathway illustrate this bias, since of 42 organisms with T-box-regulated *trp* operons, 12 have tandem T boxes (Fig. 7). This preference is also observed in operons used for BCAA biosynthesis although to a lesser extent. As mentioned previously, the *thrZ* gene of *B. subtilis*, encoding a paralog of ThrRS, is the only known example of a leader sequence with three tandem T-box elements; the presumed tight repression of expression of this gene ensures that the *thrS*-encoded major ThrRS is active under normal conditions, with *thrZ* expression occurring only when the levels of Thr-tRNA<sup>Thr</sup> drop very low (38, 82).

#### T-Box Sequences Containing a tRNA Gene

Our genome searches for T-box sequences revealed that in certain members of the *Clostridia* (*S. wolfei*, *P. thermopropionicum*, and *C. hydrogenoformans*), *leuS* is preceded by a Leu T-box region, which contains a tRNA<sup>Ala</sup> gene embedded within the T-box RNA sequence. This tRNA-encoding gene appears to be integrated within stem III, preceding the antiterminator structure. The tRNA presumably specified by this region is predicted to be functional since it contains all of the conserved features of tRNA<sup>Ala</sup>, including those nucleotides of the tRNA anticodon and the G-U pair in the acceptor stem that are required for tRNA<sup>Ala</sup> charging by AlaRS. We propose that this tRNA could be excised from the RNA by posttranscriptional processing after the terminator-antiterminator decision is made. This processing event would therefore have no effect on the transcription termination mechanism, although it could affect the stability of the leader transcript and/or the readthrough transcript. Alternatively, the excision of the tRNA before RNA polymerase reaches the termination site could prevent the appropriate interaction with the regulatory tRNA, resulting in termination. Future experimental studies will be required to determine if this type of tRNA sequence insertion in a T-box region has any effect on the T-box regulatory response.



## EVOLUTIONARY ORIGIN OF T-BOX ELEMENTS

The analysis of the phylogenetic distribution of T boxes shown in Fig. 2 suggests that the T-box mechanism could have arisen in a common ancestor of the *Firmicutes*, the *Chloroflexi*, *Deinococcus-Thermus* group, and the *Actinobacteria*. However, a plausible explanation for the acquisition of the T-box mechanism by organisms in the *Deltaproteobacteria* is by HGT, given that very few (and closely related) members of the *Deltaproteobacteria* have a T box in their genomes, and there are many phylogenetic clades between the *Deltaproteobacteria* and the aforementioned phyla in which the T-box mechanism is absent.

*Geobacter* sp. and its relatives represent one group of the *Deltaproteobacteria* in which T-box sequences have been identified. The *leuA* genes of *Geobacter* sp. and *Pelobacter* sp. are monocistronic and exhibit 50% amino acid sequence similarity with the monocistronic and T-box-regulated *leuA* of the firmicute *C. acetobutylicum*. Since these organisms share the same ecological niche, it is likely that a common ancestor of these *Geobacter/Pelobacter* spp. could have acquired the *leuA* gene and its associated T-box element by HGT. It is noteworthy that *C. acetobutylicum* and *G. sulfurreducens* are syntrophic organisms, where *C. acetobutylicum* provides acetate, a carbon compound required for growth by *G. sulfurreducens* (21). The close contact of these organisms in their natural habitat increases the probability of HGT.

Examples of likely HGT are not restricted to organisms of different phyla but can also be traced within the *Firmicutes*. In *Streptococcus thermophilus* and *S. mutans*, small open reading frames encoding proteins of ~100 amino acids in length are located upstream of the *trpEGDCFBA* operons. These putative coding sequences exhibit statistically significant similarity to the T-box-regulated *pheA* chorismate mutase gene of *D. hafniense* and *S. wolfei* but show no significant similarity to the non-T-box-regulated *pheA* genes of their respective genomes. We speculate that this small open reading frame along with its T-box regulatory region could have been horizontally transferred and inserted upstream of the *trpEGDCFBA* operons of *S. thermophilus* and *S. mutans*.

## EXPECTED INSIGHTS ON T-BOX REGULATION FROM ANALYSES OF NEW GENOME SEQUENCES

Despite the large number of sequenced genomes currently available, we predict that new findings regarding the distribution of T-box regulatory regions will emerge as additional genome sequences become available for analysis. To support this prediction, we highlight deductions that are based on analyses of the recently sequenced genome of *C. kluyveri*.

*C. kluyveri* appears to have exceptional metabolic capabilities, including extremely active sulfur metabolism and the ability to grow anaerobically on ethanol and acetate as sole carbon and energy sources (94). We identified 12 transcription units regulated by T-box sequences predicted to respond to BCAAs. Two of these correspond to aaRS genes (*leuS* and *ileS*), and one encodes a Leu transporter. Of the remaining nine transcriptional units, six contain known BCAA biosynthetic genes (*ilvE*, *ilvCB*, *ilvC*, *ilvI*, and two copies of *leuA*). One copy of *leuA* and the *ilvI* gene are regulated by tandem T boxes, which is indicative of tighter regulation (Fig. 6). We also identified an

Ile T-box-regulated operon annotated as the *porCDAB* operon, which encodes proteins similar to the subunits of pyruvate:ferredoxin oxidoreductase (POR) (Fig. 6). This enzyme catalyzes the thiamine pyrophosphate-dependent oxidative decarboxylation of pyruvate to form acetyl-CoA and CO<sub>2</sub> (64), a reaction not apparently related to BCAA biosynthesis. Val and Ile are synthesized via the same five-step pathway except for the first step, which, for Ile biosynthesis, is a thiamine pyrophosphate-dependent reaction (normally carried out by *IlvA*) that resembles those carried out by the proteins encoded in the *por* operon (108). Based on the identification of the Ile T-box sequence upstream of the *porCDAB* operon in *C. kluyveri*, we propose that the enzymes encoded by this operon are involved in the first step of Ile biosynthesis. *S. wolfei* represents a similar situation, in which a *porCDAB* operon regulated by an Ile T box could be involved in Ile biosynthesis. We note that *S. wolfei* has another copy of the *porCDAB* operon that is regulated by a Leu T box; the possible relationship to Leu biosynthesis is unclear, since the proposed *porCDAB*-encoded step is apparently specific to Ile biosynthesis (Fig. 6).

*C. kluyveri* also has two copies of the *etfBA* operon, which encodes the subunits of an electron transfer flavoprotein that participates in a cycle for the reduction of 5-crotonyl-CoA to 5-butyryl-CoA (94). In this cycle, seven NADH molecules are oxidized to NAD<sup>+</sup>. One of these *etf* operons is regulated by an Ile T-box sequence (Fig. 6). We do not completely understand the physiological significance of the regulation of this electron transfer module in response to Ile availability. We note that Ile is one of the major regulators of basic metabolic processes in *B. subtilis* via its interaction with the CodY regulatory protein. One possibility is that reduced levels of Ile may signal the activation of electron flow toward amino acid biosynthesis. Another possibility is that the above-mentioned POR complex has a high requirement for electron donors, especially in organisms that have limited redox potential, and the *etfBA*-encoded electron transfer flavoprotein could serve as the electron donor.

In addition to the above-mentioned T-box-regulated operons, *C. kluyveri* also has a BCAA-responsive T-box operon that encodes the product of *fhsA* (formate-tetrahydrofolate [THF] ligase), *fchA* (methenyl-THF cyclohydrolase), and *folD* (THF dehydrogenase), a participant in the biosynthesis of THF. This operon is regulated by a Val T box (Fig. 6), consistent with the observation that *valS* is cotranscribed with *folC* in the majority of the *Firmicutes* (Fig. 3). As mentioned above, THF is an important cofactor that donates a one-carbon unit during the synthesis of Met, purines, thymine, and pantothenic acid. Pantothenic acid biosynthesis branches off from the Val pathway at keto-valine. The first enzyme of this pathway is a hydroxymethyltransferase, which uses methylene THF as a cofactor. THF may be synthesized when the Val pools are low to ensure the availability of THF for the biosynthesis of pantothenic acid and Val.

## CONCLUSIONS

The discovery of the T-box regulatory mechanism (43, 44, 56, 57) and our prediction of its widespread use in regulating genes involved in amino acid charging, biosynthesis, and transport in gram-positive bacteria demonstrate the apparent importance of sensing the extent of charging of specific tRNAs in regulating the expression of genes involved in amino acid me-

tabolism. Early bacterial evolution was undoubtedly influenced by the fact that amino acids are expensive to synthesize and that they can be used for a variety of essential processes in addition to protein synthesis. Therefore, in regulating the expression of genes concerned with amino acid biosynthesis, charging, and transport, it was also essential to sense the availability of individual aminoacylated tRNAs, as the level of a charged or uncharged tRNA is a more accurate measure of whether an additional amino acid must be provided to maintain protein synthesis. Aminoacyl-tRNA synthetases usually recognize two features of their tRNA substrates, the anticodon sequence and the acceptor end sequence; it was therefore evolutionarily efficient for each T-box regulatory sequence to be able to detect these same features. The T-box mechanism illustrates the logical elegance of genome changes and developments that must have occurred during evolution to allow a single type of molecule (e.g., tRNA) to serve multiple functions.

In the present study, comparative genomics was used to identify all of the bacterial genes of amino acid metabolism that are predicted to be regulated by the T-box mechanism. From our findings, we conclude that the T-box mechanism is the most prominent RNA-based regulatory mechanism known to be employed by members of the *Firmicutes*. This mechanism is used to a lesser extent by members of the *Chloroflexi*, the *Deinococcus-Thermus* group, and the *Actinobacteria*. Based on the distribution of T-box sequences in bacterial genomes, we hypothesize that this regulatory system originated in a common ancestor of members of these phyla and that its use expanded in the *Firmicutes*, followed by HGT to a very few members of the *Deltaproteobacteria*.

Sequence analysis of T-box regions allows the prediction of probable recent events in T-box evolution. A notable example is provided by the Leu T-box sequences upstream of *leuS* in some members of the *Clostridia*. These T-box sequences contain an apparently intact gene that specifies tRNA<sup>Ala</sup> inserted into a portion of stem III that is highly variable and insensitive to mutation. The presence of this tRNA gene in only a small subclass of *leuS* T-box sequences in a related group of organisms suggests that its insertion is likely to have been a recent event that preserved both tRNA and leader RNA function because of a post-transcriptional processing event that releases the mature tRNA from either the terminated or readthrough transcript.

The list of genes known to be involved in amino acid biosynthesis that are regulated by the T-box mechanism was increased to include genes in the Ala and Gly pathways. In addition, the T-box regulation of genes involved in the biosynthesis of Ser, Pro, Arg, Met, Cys, Ile, Leu, Val, Tyr, Phe, Trp, His, Asn, Asp, and Thr was confirmed. We also predict that the *luxS* gene, which is involved in quorum sensing and Met biosynthesis, is regulated by the T-box mechanism in some members of the *Lactobacillales*. For most T-box-regulated genes, such as those involved in tRNA charging, amino acid transport, and amino acid biosynthesis, the relationship between the specificity of the T-box sequence and the corresponding regulated genes is obvious. In some instances, novel biochemical relationships between the regulated genes and their corresponding metabolic pathway were revealed; *C. kluyveri* provides an interesting example.

It is likely that the genes encoding aminoacyl-tRNA synthetases, amino acid biosynthetic proteins, amino acid trans-

port proteins, and key regulatory proteins may require differential responses to modulations in tRNA aminoacylation. The arrangement of T-box leader sequences as single, double, or triple copies expands the regulatory range for this mechanism, as the presence of tandem copies requires the independent binding of additional uncharged tRNA molecules to promote transcription readthrough. Regulation at the translational level, as predicted for the *Actinobacteria*, may also result in variability in the sensitivity of the system. It is also likely that individual transcriptional units may utilize regulatory mechanisms in addition to the T-box mechanism to ensure tighter regulation or a response to multiple regulatory signals, as has been observed for the *B. subtilis ilv-leu* operon (97).

Our analyses revealed an imbalance in tRNA sensing during the regulation of expression of operons containing multiple aaRS genes or biosynthetic genes involved in pathways common to more than one amino acid. This potential regulatory imbalance may be the consequence of (i) the phylogenetic origin of the operon, as noted by Vitreschak et al. (107); (ii) an incomplete or transitory stage in the evolutionary process leading to operon organization or regulation; or (iii) the ecological niche of the organism. In these cases, other regulatory mechanisms may act in concert with the T-box mechanism to ensure an appropriate physiological response. Alternatively, the organism may contain additional copies of individual genes that are subject to a different regulatory response.

In contrast to metabolite binding riboswitches that generally recognize several unique features of a single ligand, the T-box system allows a specific response of each regulated transcriptional unit to a deficiency of a single charged tRNA. A change in the regulatory response can be achieved by minor changes in the specifier sequence (and antiterminator bulge) of each T-box sequence to allow the recognition of a new uncharged tRNA class. This high degree of flexibility, without a loss of specificity, is probably responsible for the abundant use of the T-box mechanism in regulating gene expression in the *Firmicutes*. Also notable in terms of the global impact of this mechanism on cell physiology is the use of the T-box system to regulate the synthesis of proteins involved in the regulation of other gene families, including an enzyme involved in quorum sensing.

#### ACKNOWLEDGMENTS

We acknowledge the contributions of our students and postdoctoral research associates, past and present, to the work described in this review. We also thank Ricardo Ciria, Christian Eduardo Martínez, and Arturo Ocaziz for computer support and Shirley Ainsworth for bibliographical assistance. We thank Roy Jensen, Mariana Peimbert, Rosa María Gutiérrez-Ríos, Ceí Abreu-Goodger, Aubin Arroyo, and Dmitry Rodionov for their critical comments.

This work was supported by CONACyT grants 60127-Q and SALUD-2007-C01-68992 to E.M., NIH grant R01GM47823 to T.M.H., and NSF grant MCB-0615390 to C.Y.

#### REFERENCES

1. Abreu-Goodger, C., and E. Merino. 2005. RibEx: a Web server for locating riboswitches and other conserved bacterial regulatory elements. *Nucleic Acids Res.* **33**:W690–W692.
2. Abreu-Goodger, C., N. Ontiveros-Palacios, R. Ciria, and E. Merino. 2004. Conserved regulatory motifs in bacteria: riboswitches and beyond. *Trends Genet.* **20**:475–479.
3. Ahn, K. S., and R. G. Wake. 1991. Variations and coding features of the sequence spanning the replication terminus of *Bacillus subtilis* 168 and W23 chromosomes. *Gene* **98**:107–112.
4. Alifano, P., R. Fani, P. Lio, A. Lazcano, M. Bazzicalupo, M. S. Carlomagno,

- and C. B. Bruni. 1996. Histidine biosynthetic pathway and genes: structure, regulation, and evolution. *Microbiol. Rev.* **60**:44–69.
5. Babitzke, P., and P. Gollnick. 2001. Posttranscription initiation control of tryptophan metabolism in *Bacillus subtilis* by the *trp* RNA-binding attenuation protein (TRAP), anti-TRAP, and RNA structure. *J. Bacteriol.* **183**: 5795–5802.
  6. Bailey, T. L., and M. Gribskov. 1998. Combining evidence using P-values: application to sequence homology searches. *Bioinformatics* **14**:48–54.
  7. Barrick, J. E., K. A. Corbino, W. C. Winkler, A. Nahvi, M. Mandal, J. Collins, M. Lee, A. Roth, N. Sudarsan, I. Jona, J. K. Wickiser, and R. R. Breaker. 2004. New RNA motifs suggest an expanded scope for riboswitches in bacterial genetic control. *Proc. Natl. Acad. Sci. USA* **101**:6421–6426.
  8. Belitsky, B. R. 2002. Biosynthesis of amino acids of the glutamate and aspartate families, alanine, and polyamines, p. 203–231. In A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
  9. Belitsky, B. R., J. Brill, E. Bremer, and A. L. Sonenshein. 2001. Multiple genes for the last step of proline biosynthesis in *Bacillus subtilis*. *J. Bacteriol.* **183**:4389–4392.
  10. Belitsky, B. R., L. V. Wray, Jr., S. H. Fisher, D. E. Bohannon, and A. L. Sonenshein. 2000. Role of TnrA in nitrogen source-dependent repression of *Bacillus subtilis* glutamate synthase gene expression. *J. Bacteriol.* **182**: 5939–5947.
  11. Bovee, M. L., K. S. Champagne, B. Demeler, and C. S. Franklyn. 2002. The quaternary structure of the HisZ-HisG N-1-(5'-phosphoribosyl)-ATP transferase from *Lactococcus lactis*. *Biochemistry* **41**:11838–11846.
  12. Bremer, E. 2002. Adaptation to changing osmolarity, p. 385–391. In A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
  13. Brinkman, A. B., T. J. G. Ettema, W. M. de Vos, and J. van der Oost. 2003. The Lrp family of transcriptional regulators. *Mol. Microbiol.* **48**:287–294.
  14. Carter, C. W., Jr. 2008. Whence the genetic code? Thawing the 'frozen accident.' *Heredity* **100**:339–340.
  15. Chen, G., and C. Yanofsky. 2003. Tandem transcription and translation regulatory sensing of uncharged tryptophan tRNA. *Science* **301**:211–213.
  16. Chen, G., and C. Yanofsky. 2004. Features of a leader peptide coding region that regulate translation initiation for the anti-TRAP protein of *B. subtilis*. *Mol. Cell* **13**:703–711.
  17. Chopin, A., V. Biaudet, and S. D. Ehrlich. 1998. Analysis of the *Bacillus subtilis* genome sequence reveals nine new T-box leaders. *Mol. Microbiol.* **29**:662–664.
  18. Ciccarelli, F. D., T. Doerks, C. von Mering, C. J. Creevey, B. Snel, and P. Bork. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311**:1283–1287.
  19. Ciria, R., C. Abreu-Goodger, E. Morett, and E. Merino. 2004. GeConT: gene context analysis. *Bioinformatics* **20**:2307–2308.
  20. Condon, C., H. Putzer, and M. Grunberg-Manago. 1996. Processing of the leader mRNA plays a major role in the induction of *thrS* expression following threonine starvation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **93**:6992–6997.
  21. Cord-Ruwisch, R., D. R. Lovley, and B. Schink. 1998. Growth of *Geobacter sulfurreducens* with acetate in syntrophic cooperation with hydrogen-oxidizing anaerobic partners. *Appl. Environ. Microbiol.* **64**:2232–2236.
  22. De Keersmaecker, S. C., K. Sonck, and J. Vanderleyden. 2006. Let LuxS speak up in AI-2 signaling. *Trends Microbiol.* **14**:114–119.
  23. Delorme, C., S. D. Ehrlich, and P. Renault. 1992. Histidine biosynthesis genes in *Lactococcus lactis* subsp. *lactis*. *J. Bacteriol.* **174**:6571–6579.
  24. Delorme, C., S. D. Ehrlich, and P. Renault. 1999. Regulation of expression of the *Lactococcus lactis* histidine operon. *J. Bacteriol.* **181**:2026–2037.
  25. de Saizieu, A., P. Vankan, C. Vockler, and A. P. van Loon. 1997. The *trp* RNA-binding attenuation protein (TRAP) regulates the steady-state levels of transcripts of the *Bacillus subtilis* folate operon. *Microbiology* **143**:979–989.
  26. Deuel, T. F., and S. Prusiner. 1974. Regulation of glutamine synthetase from *Bacillus subtilis* by divalent cations, feedback inhibitors, and L-glutamine. *J. Biol. Chem.* **249**:257–264.
  27. Eddy, S. R. 2002. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* **3**:18.
  28. Edgar, R. C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**:113.
  29. Even, S., P. Burguiere, S. Auger, O. Soutourina, A. Danchin, and I. Martin-Verstraete. 2006. Global control of cysteine metabolism by CymR in *Bacillus subtilis*. *J. Bacteriol.* **188**:2184–2197.
  30. Fawcett, P., P. Eichenberger, R. Losick, and P. Youngman. 2000. The transcriptional profile of early to middle sporulation in *Bacillus subtilis*. *Proc. Natl. Acad. Sci. USA* **97**:8063–8068.
  31. Reference deleted.
  32. Fisher, S. H., and M. Debarbouille. 2002. Nitrogen source utilization and its regulation, p. 181–192. In A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
  33. Fuchs, R. T., F. J. Grundy, and T. M. Henkin. 2006. The S<sub>MK</sub> box is a new SAM-binding RNA for translational regulation of SAM synthetase. *Nat. Struct. Mol. Biol.* **13**:226–233.
  34. Fujita, Y., Y. Miwa, A. Galinier, and J. Deutscher. 1995. Specific recognition of the *Bacillus subtilis gnt cis*-acting catabolite-responsive element by a protein complex formed between CcpA and seryl-phosphorylated HPr. *Mol. Microbiol.* **17**:953–960.
  35. Gagnon, Y., R. Breton, H. Putzer, M. Pelchat, M. Grunberg-Manago, and J. Lapointe. 1994. Clustering and co-transcription of the *Bacillus subtilis* genes encoding the aminoacyl-tRNA synthetases specific for glutamate and for cysteine and the first enzyme for cysteine biosynthesis. *J. Biol. Chem.* **269**:7473–7482.
  36. Galinier, A., J. Haiech, M. C. Kilhoffer, M. Jaquinod, J. Stulke, J. Deutscher, and I. Martin-Verstraete. 1997. The *Bacillus subtilis crh* gene encodes a HPr-like complex involved in carbon catabolite repression. *Proc. Natl. Acad. Sci. USA* **94**:8439–8444.
  37. Gelfand, M. S. 2006. Evolution of transcriptional regulatory networks in microbial genomes. *Curr. Opin. Struct. Biol.* **16**:420–429.
  38. Gendron, N., H. Putzer, and M. Grunberg-Manago. 1994. Expression of both *Bacillus subtilis* threonyl-tRNA synthetase genes is autogenously regulated. *J. Bacteriol.* **176**:486–494.
  39. Gollnick, P., P. Babitzke, A. Antson, and C. Yanofsky. 2005. Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu. Rev. Genet.* **39**:47–68.
  40. Gollnick, P., P. Babitzke, E. Merino, and C. Yanofsky. 2002. Aromatic amino acid metabolism in *Bacillus subtilis*, p. 233–244. In A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
  41. Greene, R. C. 1996. Biosynthesis of methionine, p. 542–560. In F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology, 2nd ed. ASM Press, Washington, DC.
  42. Griffiths-Jones, S., A. Bateman, M. Marshall, A. Khanna, and S. R. Eddy. 2003. Rfam: an RNA family database. *Nucleic Acids Res.* **31**:439–441.
  43. Grundy, F. J., and T. M. Henkin. 1993. tRNA as a positive regulator of transcription antitermination in *B. subtilis*. *Cell* **74**:475–482.
  44. Grundy, F. J., and T. M. Henkin. 1994. Conservation of a transcription antitermination mechanism in aminoacyl-tRNA synthetase and amino acid biosynthesis genes in gram-positive bacteria. *J. Mol. Biol.* **235**:798–804.
  45. Grundy, F. J., and T. M. Henkin. 1998. The S box regulon: a new global transcription termination control system for methionine and cysteine biosynthesis genes in gram-positive bacteria. *Mol. Microbiol.* **30**:737–749.
  46. Grundy, F. J., and T. M. Henkin. 2002. Synthesis of serine, glycine, cysteine, and methionine, p. 245–254. In A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and its closest relatives: from genes to cells. ASM Press, Washington, DC.
  47. Grundy, F. J., and T. M. Henkin. 2003. The T box and S box transcription termination control systems. *Front. Biosci.* **8**:D20–D31.
  48. Grundy, F. J., S. C. Lehman, and T. M. Henkin. 2003. The L box regulon: lysine sensing by leader RNAs of bacterial lysine biosynthesis genes. *Proc. Natl. Acad. Sci. USA* **100**:12057–12062.
  49. Grundy, F. J., T. R. Moir, M. T. Haldeman, and T. M. Henkin. 2002. Sequence requirements for terminators and antiterminators in the T box transcription antitermination system: disparity between conservation and functional requirements. *Nucleic Acids Res.* **30**:1646–1655.
  50. Grundy, F. J., W. C. Winkler, and T. M. Henkin. 2002. tRNA-mediated transcription antitermination in vitro: codon-anticodon pairing independent of the ribosome. *Proc. Natl. Acad. Sci. USA* **99**:11121–11126.
  51. Grundy, F. J., M. R. Yousef, and T. M. Henkin. 2005. Monitoring uncharged tRNA during transcription of the *Bacillus subtilis glyQS* gene. *J. Mol. Biol.* **346**:73–81.
  52. Gutierrez-Preciado, A., R. A. Jensen, C. Yanofsky, and E. Merino. 2005. New insights into regulation of the tryptophan biosynthetic operon in gram-positive bacteria. *Trends Genet.* **21**:432–436.
  53. Gutierrez-Preciado, A., C. Yanofsky, and E. Merino. 2007. Comparison of tryptophan biosynthetic operon regulation in different gram-positive bacterial species. *Trends Genet.* **23**:422–427.
  54. Hahn, J., G. Inamine, Y. Kozlov, and D. Dubnau. 1993. Characterization of *comE*, a late competence operon of *Bacillus subtilis* required for the binding and uptake of transforming DNA. *Mol. Microbiol.* **10**:99–111.
  55. Henkin, T. M. 2008. Riboswitch RNAs: using RNA to sense cellular metabolism. *Genes Dev.* **22**:3383–3390.
  56. Henkin, T. M. 1994. tRNA-directed transcription antitermination. *Mol. Microbiol.* **13**:381–387.
  57. Henkin, T. M., B. L. Glass, and F. J. Grundy. 1992. Analysis of the *Bacillus*



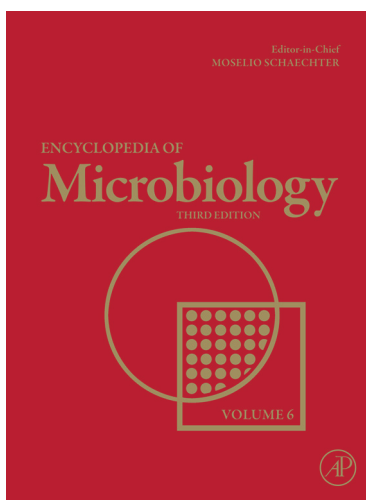
- subtilis tyrS* gene: conservation of a regulatory sequence in multiple tRNA synthetase genes. *J. Bacteriol.* **174**:1299–1306.
58. Henkin, T. M., and F. J. Grundy. 2006. Sensing metabolic signals with nascent RNA transcripts: the T box and S box riboswitches as paradigms. *Cold Spring Harb. Symp. Quant. Biol.* **71**:231–237.
  59. Henner, D. J., and C. Yanofsky. 1993. Biosynthesis of aromatic amino acids, p. 269–280. *In* R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology, and molecular genetics. ASM Press, Washington, DC.
  60. Ibarra, J. A., E. Perez-Rueda, L. Segovia, and J. L. Puente. 2008. The DNA-binding domain as a functional indicator: the case of the AraC/XylS family of transcription factors. *Genetica* **133**:65–76.
  61. Ibba, M., H. D. Becker, C. Stathopoulos, D. L. Tumbula, and D. Soll. 2000. The adaptor hypothesis revisited. *Trends Biochem. Sci.* **25**:311–316.
  62. Ibba, M., and D. Soll. 2000. Aminoacyl-tRNA synthesis. *Annu. Rev. Biochem.* **69**:617–650.
  63. Iijima, T., M. D. Diesterhaft, and E. Freese. 1977. Sodium effect of growth on aspartate and genetic analysis of a *Bacillus subtilis* mutant with high aspartate activity. *J. Bacteriol.* **129**:1440–1447.
  64. Ikeda, T., T. Ochiai, S. Morita, A. Nishiyama, E. Yamada, H. Arai, M. Ishii, and Y. Igarashi. 2006. Anabolic five subunit-type pyruvate:ferredoxin oxidoreductase from *Hydrogenobacter thermophilus* TK-6. *Biochem. Biophys. Res. Commun.* **340**:76–82.
  65. Janga, S. C., W. F. Lamboy, A. M. Huerta, and G. Moreno-Hagelsieb. 2006. The distinctive signatures of promoter regions and operon junctions across prokaryotes. *Nucleic Acids Res.* **34**:3980–3987.
  66. Kim, S. I., J. E. Germond, D. Pridmore, and D. Soll. 1996. *Lactobacillus bulgaricus* asparagine synthetase and asparaginyl-tRNA synthetase: coregulation by transcription antitermination? *J. Bacteriol.* **178**:2459–2461.
  67. Kovaleva, G. Y., and M. S. Gelfand. 2007. Transcriptional regulation of the methionine and cysteine transport and metabolism in streptococci. *FEMS Microbiol. Lett.* **276**:207–215.
  68. Reference deleted.
  69. Luo, D., J. Leautey, M. Grunberg-Manago, and H. Putzer. 1997. Structure and regulation of expression of the *Bacillus subtilis* valyl-tRNA synthetase gene. *J. Bacteriol.* **179**:2472–2478.
  70. Mandal, M., M. Lee, J. E. Barrick, Z. Weinberg, G. M. Emilsson, W. L. Ruzzo, and R. R. Breaker. 2004. A glycine-dependent riboswitch that uses cooperative binding to control gene expression. *Science* **306**:275–279.
  71. Marcos, A. T., K. Kosalkova, R. E. Cardoza, F. Fierro, S. Gutierrez, and J. F. Martin. 2001. Characterization of the reverse transsulfuration gene *mecB* of *Acremonium chrysogenum*, which encodes a functional cystathionine-gamma-lyase. *Mol. Gen. Genet.* **264**:746–754.
  72. Martinez-Guerrero, C. E., R. Ciria, C. Abreu-Goodger, G. Moreno-Hagelsieb, and E. Merino. 2008. GeConT 2: gene context analysis for orthologous proteins, conserved domains and metabolic pathways. *Nucleic Acids Res.* **36**:W176–W180.
  73. Merino, E., P. Babitzke, and C. Yanofsky. 1995. *trp* RNA-binding attenuation protein (TRAP)-*trp* leader RNA interactions mediate translational as well as transcriptional regulation of the *Bacillus subtilis trp* operon. *J. Bacteriol.* **177**:6362–6370.
  74. Merino, E., R. A. Jensen, and C. Yanofsky. 2008. Evolution of bacterial *trp* operons and their regulation. *Curr. Opin. Microbiol.* **11**:78–86.
  75. Moat, A. G., J. W. Foster, and M. P. Spector. 2002. Biosynthesis and metabolism of amino acids, p. 503–544. *In* A. G. Moat, J. W. Foster, and M. P. Spector (ed.), *Microbial physiology*. Wiley-Liss, Inc., New York, NY.
  76. Moreno-Hagelsieb, G., and J. Collado-Vides. 2002. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18**(Suppl. 1):S329–S336.
  77. Murphy, B. A., F. J. Grundy, and T. M. Henkin. 2002. Prediction of gene function in methylthioadenosine recycling from regulatory signals. *J. Bacteriol.* **184**:2314–2318.
  78. Parsot, C. 1986. Evolution of biosynthetic pathways: a common ancestor for threonine synthase, threonine dehydratase and D-serine dehydratase. *EMBO J.* **5**:3013–3019.
  79. Parsot, C., and G. N. Cohen. 1988. Cloning and nucleotide sequence of the *Bacillus subtilis hom* gene coding for homoserine dehydrogenase. Structural and evolutionary relationships with *Escherichia coli* aspartokinase-homoserine dehydrogenases I and II. *J. Biol. Chem.* **263**:14654–14660.
  80. Putzer, H., A. A. Brakhage, and M. Grunberg-Manago. 1990. Independent genes for two threonyl-tRNA synthetases in *Bacillus subtilis*. *J. Bacteriol.* **172**:4593–4602.
  81. Putzer, H., C. Condon, D. Brechemier-Baey, R. Brito, and M. Grunberg-Manago. 2002. Transfer RNA-mediated antitermination in vitro. *Nucleic Acids Res.* **30**:3026–3033.
  82. Putzer, H., N. Gendron, and M. Grunberg-Manago. 1992. Co-ordinate expression of the two threonyl-tRNA synthetase genes in *Bacillus subtilis*: control by transcriptional antitermination involving a conserved regulatory sequence. *EMBO J.* **11**:3117–3127.
  83. Putzer, H., S. Laalami, A. A. Brakhage, C. Condon, and M. Grunberg-Manago. 1995. Aminoacyl-tRNA synthetase gene regulation in *Bacillus subtilis*: induction, repression and growth-rate regulation. *Mol. Microbiol.* **16**:709–718.
  84. Read, T. D., S. N. Peterson, N. Tourasse, L. W. Baillie, I. T. Paulsen, K. E. Nelson, H. Tettelin, D. E. Fouts, J. A. Eisen, S. R. Gill, E. K. Holtzapple, O. A. Okstad, E. Helgason, J. Rilstone, M. Wu, J. F. Kolonay, M. J. Beanan, R. J. Dodson, L. M. Brinkac, M. Gwinn, R. T. Deboy, R. Madpu, S. C. Daugherty, A. S. Durkin, D. H. Haft, W. C. Nelson, J. D. Peterson, M. Pop, H. M. Khouri, D. Radune, J. L. Benton, Y. Mahamoud, L. X. Jiang, I. R. Hance, J. F. Weidman, K. J. Berry, R. D. Plaut, A. M. Wolf, K. L. Watkins, W. C. Nierman, A. Hazen, R. Cline, C. Redmond, J. E. Thwaite, O. White, S. L. Salzberg, B. Thomason, A. M. Friedlander, T. M. Koehler, P. C. Hanna, A. B. Kolsto, and C. M. Fraser. 2003. The genome sequence of *Bacillus anthracis* Ames and comparison to closely related bacteria. *Nature* **423**:81–86.
  85. Rodionov, D. A., A. G. Vitreschak, A. A. Mironov, and M. S. Gelfand. 2003. Regulation of lysine biosynthesis and transport genes in bacteria: yet another RNA riboswitch? *Nucleic Acids Res.* **31**:6748–6757.
  86. Rodionov, D. A., A. G. Vitreschak, A. A. Mironov, and M. S. Gelfand. 2004. Comparative genomics of the methionine metabolism in gram-positive bacteria: a variety of regulatory systems. *Nucleic Acids Res.* **32**:3340–3353.
  87. Rollins, S. M. 2002. The mRNA/tRNA interaction promoting T box transcriptional antitermination. Ph.D. thesis. The Ohio State University, Columbus.
  88. Rollins, S. M., F. J. Grundy, and T. M. Henkin. 1997. Analysis of *cis*-acting sequence and structural elements required for antitermination of the *Bacillus subtilis tyrS* gene. *Mol. Microbiol.* **25**:411–421.
  89. Salgado, H., G. Moreno-Hagelsieb, T. F. Smith, and J. Collado-Vides. 2000. Operons in *Escherichia coli*: genomic analyses and predictions. *Proc. Natl. Acad. Sci. USA* **97**:6652–6657.
  90. Sarsero, J. P., E. Merino, and C. Yanofsky. 2000. A *Bacillus subtilis* gene of previously unknown function, *yhaG*, is translationally regulated by tryptophan-activated TRAP and appears to be involved in tryptophan transport. *J. Bacteriol.* **182**:2329–2331.
  91. Sarsero, J. P., E. Merino, and C. Yanofsky. 2000. A *Bacillus subtilis* operon containing genes of unknown function senses tRNA<sup>Trp</sup> charging and regulates expression of the genes of tryptophan biosynthesis. *Proc. Natl. Acad. Sci. USA* **97**:2656–2661.
  92. Schauder, S., K. Shokat, M. G. Surette, and B. L. Bassler. 2001. The LuxS family of bacterial autoinducers: biosynthesis of a novel quorum-sensing signal molecule. *Mol. Microbiol.* **41**:463–476.
  93. Schreier, H. J. 1993. Biosynthesis of glutamine and glutamate and assimilation of ammonia, p. 281–298. *In* A. L. Sonenshein, J. A. Hoch, and R. Losick (ed.), *Bacillus subtilis* and other gram-positive bacteria: biochemistry, physiology, and molecular genetics. ASM Press, Washington, DC.
  94. Seedorf, H., W. F. Fricke, B. Veith, H. Bruggemann, H. Liesegang, A. Strittmatter, M. Miethke, W. Buckel, J. Hinderberger, F. Li, C. Hagemeyer, R. K. Thauer, and G. Gottschalk. 2008. The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. *Proc. Natl. Acad. Sci. USA* **105**:2128–2133.
  95. Sekowska, A., and A. Danchin. 1999. Identification of *yyrU* as the methylthioadenosine nucleosidase gene in *Bacillus subtilis*. *DNA Res.* **6**:255–264.
  96. Shivers, R. P., and A. L. Sonenshein. 2004. Activation of the *Bacillus subtilis* global regulator CodY by direct interaction with branched-chain amino acids. *Mol. Microbiol.* **53**:599–611.
  97. Shivers, R. P., and A. L. Sonenshein. 2005. *Bacillus subtilis* *ilvB* operon: an intersection of global regulons. *Mol. Microbiol.* **56**:1549–1559.
  98. Smith, M. C., A. Mountain, and S. Baumberg. 1986. Cloning in *Escherichia coli* of a *Bacillus subtilis* arginine repressor gene through its ability to confer structural stability on a fragment carrying genes of arginine biosynthesis. *Mol. Gen. Genet.* **205**:176–182.
  99. Stauffer, G. V. 1996. Biosynthesis of serine, glycine, and one-carbon units, p. 506–513. *In* F. C. Neidhardt, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella*: cellular and molecular biology. ASM Press, Washington, DC.
  100. Steele, M. I., D. Lorenz, K. Hatter, A. Park, and J. R. Sokatch. 1992. Characterization of the *mmsAB* operon of *Pseudomonas aeruginosa* PAO encoding methylmalonate-semialdehyde dehydrogenase and 3-hydroxyisobutyrate dehydrogenase. *J. Biol. Chem.* **267**:13585–13592.
  101. Sun, D., and P. Setlow. 1993. Cloning and nucleotide sequence of the *Bacillus subtilis ansR* gene, which encodes a repressor of the *ans* operon coding for L-asparaginase and L-aspartase. *J. Bacteriol.* **175**:2501–2506.
  102. Sun, D. X., and P. Setlow. 1991. Cloning, nucleotide sequence, and expression of the *Bacillus subtilis ans* operon, which codes for L-asparaginase and L-aspartase. *J. Bacteriol.* **173**:3831–3845.
  103. Tomsic, J., B. A. McDaniel, F. J. Grundy, and T. M. Henkin. 2008. Natural variability in S-adenosylmethionine (SAM)-dependent riboswitches: S-box elements in *Bacillus subtilis* exhibit differential sensitivity to SAM in vivo and in vitro. *J. Bacteriol.* **190**:823–833.



104. **Valbuzzi, A., P. Gollnick, P. Babitzke, and C. Yanofsky.** 2002. The anti-trp RNA-binding attenuation protein (anti-TRAP), AT, recognizes the tryptophan-activated RNA binding domain of the TRAP regulatory protein. *J. Biol. Chem.* **277**:10608–10613.
105. **Valbuzzi, A., and C. Yanofsky.** 2001. Inhibition of the *B. subtilis* regulatory protein TRAP by the TRAP-inhibitory protein, AT. *Science* **293**:2057–2059.
106. **Vander Horn, P. B., and S. A. Zahler.** 1992. Cloning and nucleotide sequence of the leucyl-tRNA synthetase gene of *Bacillus subtilis*. *J. Bacteriol.* **174**:3928–3935.
107. **Vitreschak, A. G., A. A. Mironov, V. A. Lyubetsky, and M. S. Gelfand.** 2008. Comparative genomic analysis of T-box regulatory systems in bacteria. *RNA* **14**:717–735.
108. **Voet, D., and J. G. Voet.** 1995. *Biochemistry*, 2nd ed., p. 764–776. John Wiley & Sons, Inc., Hoboken, NJ.
109. **Voskuil, M. I., and G. H. Chambliss.** 1996. Significance of HPr in catabolite repression of  $\alpha$ -amylase. *J. Bacteriol.* **178**:7014–7015.
110. **Webb, M. E., A. G. Smith, and C. Abell.** 2004. Biosynthesis of pantothenate. *Nat. Prod. Rep.* **21**:695–721.
111. **Wels, M., K. T. Groot, M. Kleerebezem, R. J. Siezen, and C. Francke.** 2008. An in silico analysis of T-box regulated genes and T-box evolution in prokaryotes, with emphasis on prediction of substrate specificity of transporters. *BMC Genomics* **9**:330.
112. **Winkler, W. C., F. J. Grundy, B. A. Murphy, and T. M. Henkin.** 2001. The GA motif: an RNA element common to bacterial antitermination systems, rRNA, and eukaryotic RNAs. *RNA* **7**:1165–1172.
113. **Woese, C. R., G. J. Olsen, M. Ibba, and D. Soll.** 2000. Aminoacyl-tRNA synthetases, the genetic code, and the evolutionary process. *Microbiol. Mol. Biol. Rev.* **64**:202–236.
114. **Wolf, Y. I., L. Aravind, N. V. Grishin, and E. V. Koonin.** 1999. Evolution of aminoacyl-tRNA synthetases—analysis of unique domain architectures and phylogenetic trees reveals a complex history of horizontal gene transfer events. *Genome Res.* **9**:689–710.
115. **Yanofsky, C.** 1981. Attenuation in the control of expression of bacterial operons. *Nature* **289**:751–758.
116. **Yanofsky, C.** 1988. Transcription attenuation. *J. Biol. Chem.* **263**:609–612.
117. **Yanofsky, C.** 2000. Transcription attenuation: once viewed as a novel regulatory strategy. *J. Bacteriol.* **182**:1–8.
118. **Yanofsky, C.** 2004. The different roles of tryptophan transfer RNA in regulating *trp* operon expression in *E. coli* versus *B. subtilis*. *Trends Genet.* **20**:367–374.
119. **Yeggy, J. P., and D. P. Stahly.** 1980. Sporulation and regulation of homoserine dehydrogenase in *Bacillus subtilis*. *Can. J. Microbiol.* **26**:1386–1391.
120. **Yoshida, K., Y. Fujita, and S. D. Ehrlich.** 1999. Three asparagine synthetase genes of *Bacillus subtilis*. *J. Bacteriol.* **181**:6081–6091.
121. **Yousef, M. R., F. J. Grundy, and T. M. Henkin.** 2005. Structural transitions induced by the interaction between tRNA(Gly) and the *Bacillus subtilis* *ghyQS* T box leader RNA. *J. Mol. Biol.* **349**:273–287.

**Provided for non-commercial research and educational use.  
Not for reproduction, distribution or commercial use.**

This article was originally published in the *Encyclopedia of Microbiology* published by Elsevier, and the attached copy is provided by Elsevier for the author's benefit and for the benefit of the author's institution, for non-commercial research and educational use including without limitation use in instruction at your institution, sending it to specific colleagues who you know, and providing a copy to your institution's administrator.



All other uses, reproduction and distribution, including without limitation commercial reprints, selling or licensing copies or access, or posting on open internet sites, your personal or institution's website or repository, are prohibited. For exceptions, permission may be sought for such use through Elsevier's permissions site at:

<http://www.elsevier.com/locate/permissionusematerial>

A G-Preciado. Genome Sequence Databases: Types of Data and Bioinformatic Tools. *Encyclopedia of Microbiology*. (Moselio Schaechter, Editor), pp. 211-236 Oxford: Elsevier.

## Genome Sequence Databases: Types of Data and Bioinformatic Tools

A G-Preciado, M Peimbert, and E Merino, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

© 2009 Elsevier Inc. All rights reserved.

Defining Statement

Introduction

Literature Databases

Gateways to Databases and Bioinformatics Tools

Sequence Databases

Searching for Sequence Similarity

Protein Databases

Integrated Databases

Protein Structure

Resources for Genome-Scale Analysis

Metabolomics

Metagenomics

Taxonomy and Phylogeny

Resources for the Analysis of Gene Expression

Resources for Proteomics

Resources for Gene Regulation Analysis

MBDBs and Analysis Programs of RNA Regulatory Elements

Dedicated Integration Systems for Molecular Biology Databases

Future of Biological Databases

Further Reading

### Glossary

**curation** Curation is the process of examining, testing and selecting information before its incorporation into a database.

**exhaustive algorithm** Exhaustive algorithm iteratively produces the entire solution space for a given problem, checks to see if the problem is solved, and continues until a correct solution is generated, at which point the optimal solution is returned.

**heuristic algorithm** Heuristic algorithm is an algorithm that makes educated guesses to solve a problem. This results in a faster solution without necessarily an optimal solution.

**HMM** A Markov model is a statistical model in which the probability of an event depends on the immediately

previous event. In a HMM (for hidden Markov model) the parameters of the process are 'hidden' and the challenge is to determine them from the observable data.

**OTU** OTU stands for Operational Taxonomic Unit.

**protein domain** Protein domain is a protein evolution unit. It can be the entire protein or a compact part of protein structure.

**sequence motif** Sequence motif is a characteristic nucleotide or amino acid sequence that is conserved in a group of sequences. In most cases, it has a biological function, such as the catalytic site of a protein, or a DNA-binding site.

### Abbreviations

**BBH** bidirectional best hit

**BLAST** Basic Local Alignment and Search Tool

**CAMERA** Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

**CDD** conserved domain database

**CGH** comparative genomic hybridization

**CIBEX** Centre for Information Biology Gene Expression

**CMR** Comprehensive Microbial Resource

**COG** Cluster of Orthologous Groups of Proteins

**DBTBS** Database of Transcriptional regulation in *B. subtilis*

**DDBJ** DNA Data Bank of Japan

**EMBL-EBI** European Molecular Biology Laboratory-European Bioinformatics Institute

**GEO** Gene Expression Omnibus

**GO** Gene Ontology

<b>GOLD</b>	Genomes Online Database	<b>ORF</b>	Open Reading Frame
<b>HGT</b>	Horizontal gene transfer	<b>OTU</b>	Operational Taxonomic Unit
<b>HMM</b>	hidden Markov model	<b>PAML</b>	Phylogenetic analysis using maximum likelihood
<b>IMG/M</b>	Integrated Microbial Genomes/ Metagenomes	<b>PDB</b>	Protein Data Bank
<b>INFERNAL</b>	inference of RNA Alignment	<b>PHYLIP</b>	phylogeny Inference Package
<b>iTOL</b>	Interactive Tree Of Life	<b>PIR</b>	Protein Information Resource
<b>KEGG</b>	The Kyoto Encyclopaedia of Genes and Genomes	<b>PRF</b>	Protein Research Foundation
<b>LSU</b>	large subunit sequence	<b>PRIDE</b>	Proteomics Identification Database
<b>MAMMOTH</b>	Matching molecular models obtained from theory	<b>PSI-BLAST</b>	Position-Specific Iterated BLAST
<b>MAST</b>	Motif Alignment and Search Tool	<b>PSSM-based</b>	Position Specific Scoring Matrix-based
<b>MBDBs</b>	Molecular Biology Databases	<b>RFLPs</b>	restriction fragment length polymorphisms
<b>MCMC</b>	Markov chain Monte Carlo	<b>rRNA</b>	ribosomal RNA
<b>MEME</b>	Multiple EM for Motif Elicitation	<b>RSA</b>	Regulatory Sequence Analysis
<b>MeSH</b>	Medical Subject Heading	<b>SCFGs</b>	stochastic context-free grammars
<b>ML</b>	Maximum Likelihood	<b>SCOP</b>	Structure Classification of Proteins
<b>MP</b>	Maximum Parsimony	<b>SIDD</b>	stress-induced duplex destabilization
<b>MSA</b>	Multiple Sequence Alignment	<b>SMART</b>	Simple Modular Architecture Research Tool
<b>MStA</b>	Multiple Structure Alignments	<b>SMD</b>	Stanford Microarray Database
<b>MUSCLE</b>	multiple Sequence Comparison by Log-Expectation	<b>SSU</b>	small subunit sequence
<b>NCBI</b>	National Center for Biotechnology Information	<b>TF</b>	transcription factor
<b>ncRNAs</b>	noncoding RNAs	<b>TRs</b>	transcriptional regulator
<b>NJ</b>	Neighbor Joining	<b>UniProt</b>	Universal Protein Resource
<b>OMSSA</b>	Open Mass Spectrometry Search Algorithm	<b>UPGMA</b>	Unweighted Pair Group Method with Arithmetic mean
<b>OPD</b>	Open Proteomics Database	<b>WGS</b>	Whole Genome Shotgun

## Defining Statement

The vast and diverse data generated from the recent large-scale genomic projects has no precedent. For optimal use, it has been compiled and organized in different kinds of Molecular Biology Databases. This article reviews some of the most important databases and the software developed for their analyses.

## Introduction

In the past decade, large scale projects have generated a vast amount of molecular biological data that has been deposited and organized into different Molecular Biology Databases (MBDBs). These MBDBs include information on genomics, proteomics, transcriptomics, interactomics, and metabolomics among many others. More than 1000 different MBDBs are publicly or

commercially available and have become an essential element of every day scientific activity, making it possible to relate data to a specific biological problem and to assist the scientist to guide their research. MBDBs users can easily find answers to questions such as: which gene codes for an enzyme that performs a particular reaction in a specific organism? How and to what extent is such a gene regulated? What is the three dimensional structure of its corresponding enzyme? What other enzymes participate in the same metabolic pathway? Which scientific articles are related to this gene, protein, or pathway? (Figure 1). Furthermore, since most of the homologous proteins – proteins that have evolved from a common ancestor – are structurally and commonly functionally related, MBDBs users can easily identify, by simple sequence comparisons, other organisms carrying homologous genes and, in general, extend the aforementioned questions to these new set of organisms to find common properties and

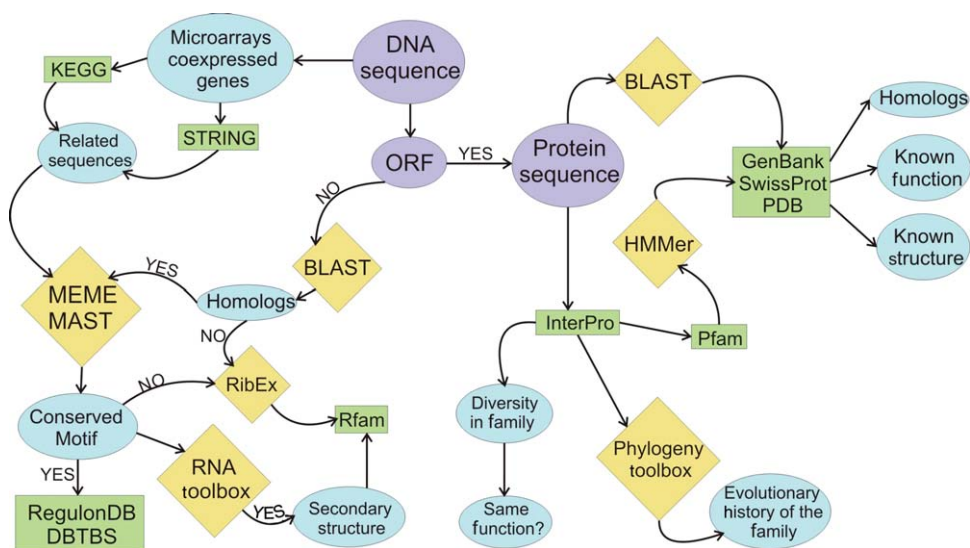


Figure 1 General flow pathway for the use of molecular biology databases.

generate general and properly supported conclusions. In fact, due to the relevance of MBDBs to the scientific community, one of the most important journals, *Nucleic Acids Research* devotes every year a freely available issue to biological databases and another one to papers describing web-based software resources. In this article, we review the main biological databases and the public domain software that has been developed to analyze them.

### Literature Databases

The NCBI's PubMed database includes citations from life science journals for biomedical articles, most of them with abstracts and many with links to the full-text articles. It is heavily linked to other core Entrez databases, such as Nucleotide, Protein, Gene, Structure, and PubChem where it provides a crucial bridge between the data of molecular biology and the scientific literature. PubMed records are also linked to one another within Entrez as 'related articles' on the basis of computationally detected similarities using the Medical Subject Heading (MeSH) terms and the text of titles and abstracts. Also, it includes digitized back content of many journals, going back in some cases to the 1800s or early 1900s. ISI Web of Knowledge<sup>SM</sup> platform covers literature databases of sciences, social sciences, arts, and humanities; ISI Web of Knowledge<sup>SM</sup> also includes abstracts and links to the full-text articles. This platform can be very useful especially when literature outside life science field is required.

### Gateways to Databases and Bioinformatics Tools

#### NCBI

The National Center for Biotechnology Information (NCBI) maintains 31 databases. The internet address of some of these and other important databases, and web servers are listed in Table 1. Entrez is an integrated database retrieval system that enables text searching using simple Boolean queries. Entrez searches rapidly across all NCBI databases and returns the counts of matching records in each database, including DNA and protein sequences (GenBank and Proteins, respectively), NCBI taxonomy, genomes, population sets, gene expression data, gene-oriented sequence clusters in UniGene, protein structures, alignment-based protein domains and the biomedical literature via PubMed, and online books. Results can be saved in a local file, shown in the browser as plain text. Results may be also sent to the Entrez clipboard where they may be recalled later using My NCBI. In addition, PubMed results and those from other databases may be emailed directly from Entrez or exported. Entrez's My NCBI allows users to store personal configuration options, such as search filters and document delivery providers. My NCBI also saves searches and can automatically email updated search results.

#### EMBL-EBI

The European Molecular Biology Laboratory-European Bioinformatics Institute (EMBL-EBI) server houses more than 50 databases and 50 bioinformatics tools. EMBL-EBI databases include DNA and protein sequences (EMBL and UniProt, respectively), protein structures, gene

**Table 1** Table of molecular biology databases and software for their analysis

**Literature databases**

Nucleic Acid Research Journal	<a href="http://nar.oxfordjournals.org/">http://nar.oxfordjournals.org/</a>
PubMed	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
ISI Web of Knowledge <sup>SM</sup>	<a href="http://www.isiknowledge.com/">http://www.isiknowledge.com/</a>

**Gateways to databases and bioinformatic tools**

NCBI	<a href="http://www.ncbi.nlm.nih.gov/">http://www.ncbi.nlm.nih.gov/</a>
EMBL-EBI	<a href="http://www.ebi.ac.uk/">http://www.ebi.ac.uk/</a>

**Nucleotide sequence databases**

Entrez Nucleotide	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
EMBL Nucleotide Sequence Database	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
DNA Data Bank of Japan	<a href="http://www.ddbj.nig.ac.jp/">http://www.ddbj.nig.ac.jp/</a>

**Protein sequence databases**

Entrez Protein	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
UniProt	<a href="http://beta.uniprot.org">http://beta.uniprot.org</a>
SwissProt	<a href="http://ca.expasy.org">http://ca.expasy.org</a>

**Searching for sequence similarity**

BLAST	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
MEME	<a href="http://meme.sdsc.edu">http://meme.sdsc.edu</a>
MAST	<a href="http://meme.sdsc.edu">http://meme.sdsc.edu</a>
Oligo-analysis	<a href="http://rsat.ulb.ac.be">http://rsat.ulb.ac.be</a>
ClustalW	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
T-Coffee	<a href="http://www.ch.embnet.org">http://www.ch.embnet.org</a>
MUSCLE	<a href="http://phylogenomics.berkeley.edu">http://phylogenomics.berkeley.edu</a>
HMMER	<a href="http://hmmer.janelia.org/">http://hmmer.janelia.org/</a>

**Protein databases**

COGs	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
PROSITE	<a href="http://ca.expasy.org">http://ca.expasy.org</a>
PRINTS	<a href="http://www.bioinf.manchester.ac.uk">http://www.bioinf.manchester.ac.uk</a>
ProDom	<a href="http://prodom.prabi.fr/">http://prodom.prabi.fr/</a>
Pfam	<a href="http://www.sanger.ac.uk">http://www.sanger.ac.uk</a>
SMART	<a href="http://smart.embl.de/">http://smart.embl.de/</a>
TIGRFAMs	<a href="http://www.tigr.org">http://www.tigr.org</a>
PIRSF	<a href="http://pir.georgetown.edu">http://pir.georgetown.edu</a>
InterPro	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
CDD	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>

**Protein structure**

PDB	<a href="http://www.rcsb.org">http://www.rcsb.org</a>
SCOP	<a href="http://scop.mrc-lmb.cam.ac.uk">http://scop.mrc-lmb.cam.ac.uk</a>
CATH	<a href="http://www.cathdb.info/">http://www.cathdb.info/</a>

**Protein structure visualization**

RasMol	<a href="http://www.umass.edu/microbio/rasmol/">http://www.umass.edu/microbio/rasmol/</a>
DeepView	<a href="http://ca.expasy.org">http://ca.expasy.org</a>

**Protein structure alignments**

MAMMOTH	<a href="http://ub.cbm.uam.es">http://ub.cbm.uam.es</a>
Dali	<a href="http://ekhidna.biocenter.helsinki.fi">http://ekhidna.biocenter.helsinki.fi</a>

**Protein structure prediction**

Swiss-Model	<a href="http://swissmodel.expasy.org/">http://swissmodel.expasy.org/</a>
MODELLER	<a href="http://www.salilab.org/modeller/">http://www.salilab.org/modeller/</a>

**Fold recognition**

Phyre	<a href="http://www.sbg.bio.ic.ac.uk">http://www.sbg.bio.ic.ac.uk</a>
PSIPRED	<a href="http://bioinf.cs.ucl.ac.uk">http://bioinf.cs.ucl.ac.uk</a>
Gene3D	<a href="http://gene3d.biochem.ucl.ac.uk">http://gene3d.biochem.ucl.ac.uk</a>

**Protein classification based on ontology**

The Gene Ontology	<a href="http://www.geneontology.org/">http://www.geneontology.org/</a>
-------------------	---

**Resources for genome-scale analysis**

Entrez genome	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
TaxPlot	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
CMR	<a href="http://www.tigr.org">http://www.tigr.org</a>
GOLD	<a href="http://www.genomesonline.org">http://www.genomesonline.org</a>
Genome Reviews	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
Microbes Online	<a href="http://www.microbesonline.org/">http://www.microbesonline.org/</a>
Integr8	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
UCSC Genome Browser	<a href="http://genome.ucsc.edu/">http://genome.ucsc.edu/</a>

(Continued)

**Table 1** (Continued)

<b>Metabolomics</b>	
KEGG	<a href="http://www.genome.jp">http://www.genome.jp</a>
EcoCyc	<a href="http://ecocyc.org/">http://ecocyc.org/</a>
<b>Metagenomics</b>	
IMG/M	<a href="http://img.jgi.doe.gov/m">http://img.jgi.doe.gov/m</a>
CAMERA	<a href="http://camera.calit2.net/">http://camera.calit2.net/</a>
<b>Taxonomy databases</b>	
UniProt Taxonomy	<a href="http://beta.uniprot.org">http://beta.uniprot.org</a>
NCBI Taxonomy	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
<b>Phylogenetic analysis algorithms</b>	
MrBayes	<a href="http://mrbayes.csit.fsu.edu/">http://mrbayes.csit.fsu.edu/</a>
PAUP*	<a href="http://paup.csit.fsu.edu/">http://paup.csit.fsu.edu/</a>
PHYLIP	<a href="http://evolution.genetics.washington.edu">http://evolution.genetics.washington.edu</a>
PAML	<a href="http://abacus.gene.ucl.ac.uk">http://abacus.gene.ucl.ac.uk</a>
TreeView	<a href="http://taxonomy.zoology.gla.ac.uk">http://taxonomy.zoology.gla.ac.uk</a>
<b>Universal tree of life</b>	
iTOL	<a href="http://itol.embl.de/">http://itol.embl.de/</a>
The ARB Project	<a href="http://www.arb-home.de/">http://www.arb-home.de/</a>
Silva	<a href="http://www.arb-silva.de/">http://www.arb-silva.de/</a>
European Ribosomal RNA database	<a href="http://bioinformatics.psb.ugent.be/webtools/rRNA/">http://bioinformatics.psb.ugent.be/webtools/rRNA/</a>
Ribosomal Database Project II	<a href="http://rdp.cme.msu.edu/">http://rdp.cme.msu.edu/</a>
Greengenes	<a href="http://greengenes.lbl.gov/">http://greengenes.lbl.gov/</a>
<b>Gene expression</b>	
Stanford Microarray Data base	<a href="http://genome-www5.stanford.edu/">http://genome-www5.stanford.edu/</a>
GEO	<a href="http://www.ncbi.nlm.nih.gov">http://www.ncbi.nlm.nih.gov</a>
<i>E. coli</i> GenExpDB	<a href="http://chase.ou.edu">http://chase.ou.edu</a>
ArrayExpress	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
CIBEX	<a href="http://cibex.nig.ac.jp">http://cibex.nig.ac.jp</a>
<b>Resources for molecular structure and proteomics</b>	
Swiss-2DPAGE	<a href="http://ca.expasy.org">http://ca.expasy.org</a>
OMSSA	<a href="http://pubchem.ncbi.nlm.nih.gov">http://pubchem.ncbi.nlm.nih.gov</a>
Mascot	<a href="http://www.matrixscience.com/">http://www.matrixscience.com/</a>
PRIDE	<a href="http://www.ebi.ac.uk">http://www.ebi.ac.uk</a>
OPD	<a href="http://bioinformatics.icmb.utexas.edu">http://bioinformatics.icmb.utexas.edu</a>
<b>Resources for gene regulation analysis</b>	
RegulonDB	<a href="http://regulondb.ccg.unam.mx">http://regulondb.ccg.unam.mx</a>
DBTBS	<a href="http://dbtbs.hgc.jp/">http://dbtbs.hgc.jp/</a>
Neural Network Promoter Prediction	<a href="http://www.fruitfly.org">http://www.fruitfly.org</a>
WebSIDD	<a href="http://www.genomecenter.ucdavis.edu/benham/sidd/">http://www.genomecenter.ucdavis.edu/benham/sidd/</a>
Regulatory Sequence Analysis Tools	<a href="http://rsat.scmbb.ulb.ac.be">http://rsat.scmbb.ulb.ac.be</a>
Predicted Attenuators in Bacteria	<a href="http://cmgm.stanford.edu/~merino">http://cmgm.stanford.edu/~merino</a>
RibEx	<a href="http://ribex.ibt.unam.mx">http://ribex.ibt.unam.mx</a>
Gene Context Tool	<a href="http://gecont.ibt.unam.mx">http://gecont.ibt.unam.mx</a>
<b>MBDBs and analysis programs of RNA regulatory elements</b>	
Rfam	<a href="http://www.sanger.ac.uk/Software/Rfam">http://www.sanger.ac.uk/Software/Rfam</a>
INFERNAL	<a href="http://infernal.janelia.org/">http://infernal.janelia.org/</a>
MFOLD	<a href="http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html">http://bioweb.pasteur.fr/seqanal/interfaces/mfold-simple.html</a>
Vienna RNA Package	<a href="http://www.tbi.univie.ac.at">http://www.tbi.univie.ac.at</a>
HotKnots	<a href="http://www.cs.ubc.ca/labs/beta/Software/HotKnots/">http://www.cs.ubc.ca/labs/beta/Software/HotKnots/</a>
RNAMST	<a href="http://bioinfo.csie.ncu.edu.tw">http://bioinfo.csie.ncu.edu.tw</a>
SCARNA	<a href="http://www.scarna.org">http://www.scarna.org</a>
STRAL	<a href="http://www.biophys.uni-duesseldorf.de/stral">http://www.biophys.uni-duesseldorf.de/stral</a>
MARNA	<a href="http://www.bioinf.uni-freiburg.de/Software/MARNA/">http://www.bioinf.uni-freiburg.de/Software/MARNA/</a>
FOLDALIGN	<a href="http://foldalign.ku.dk/">http://foldalign.ku.dk/</a>
STEMLOC	<a href="http://biowiki.org">http://biowiki.org</a>
Dynalign	<a href="http://rna.urmc.rochester.edu">http://rna.urmc.rochester.edu</a>
PSTAG	<a href="http://pstag.dna.bio.keio.ac.jp/">http://pstag.dna.bio.keio.ac.jp/</a>
Mifold	<a href="http://www.lcb.uu.se/~evaf/Mifold/">http://www.lcb.uu.se/~evaf/Mifold/</a>
RNASHAPES	<a href="http://bibiserv.techfak.uni-bielefeld.de">http://bibiserv.techfak.uni-bielefeld.de</a>
KnetFold	<a href="http://knetfold.abcc.ncifcrf.gov/">http://knetfold.abcc.ncifcrf.gov/</a>
COVE	<a href="http://selab.wustl.edu/software/cove">http://selab.wustl.edu/software/cove</a>
RNAmine	<a href="http://rnamine.ncrna.org">http://rnamine.ncrna.org</a>

(Continued)



Table 1 (Continued)

**Dedicated integration systems for molecular biology databases**

DBGET	<a href="http://www.genome.jp">http://www.genome.jp</a>
SRS	<a href="http://srs.ebi.ac.uk">http://srs.ebi.ac.uk</a>
STRING	<a href="http://string.embl.de/">http://string.embl.de/</a>

expression data, molecular interactions, several kinds of alignments, literature, and so on. The server has different browsers; EB-eye performs searches in all the databases. Once you know the proper database entries, data retrieval can be performed easily by EBI Dbfetch; up to 200 entries can be retrieved. EBI-tools comprise mainly sequence, structure, and expression analysis tools.

**Sequence Databases****Nucleotide Sequence Databases**

NCBI GenBank, EMBL Nucleotide Sequence Database (EMBL), and the DNA Data Bank of Japan (DDBJ) are comprehensive databases that contain publicly available nucleotide sequences. The three organizations synchronize their data on a daily basis to ensure worldwide coverage. GenBank/EMBL/DDBJ data are submitted by the scientific community, primarily from large-scale sequencing projects. Each sequence entry includes a concise description of the sequence, the scientific name of the source organism and bibliographic references. Records cannot be updated, corrected, or amended without the permission of the original submitter. GenBank/EMBL/DDBJ records include individual genes, Whole Genome Shotgun (WGS), RNA, high-throughput cDNA, synthetic sequences, and environmental sequencing (for which the source organism is unknown). Due to its completeness and standing as a primary data provider, GenBank/EMBL/DDBJ is the initial source for many MBDBs.

**Protein Sequence Databases**

NCBI GenPept and TrEMBL used to be the comprehensive protein sequence databases; they were produced by translating GenBank and EMBL, respectively. Now, the protein sequences can be found at NCBI Protein and Universal Protein Resource (UniProt). NCBI Protein is compiled from a variety of sources, including Swiss-Prot, PIR (Protein Information Resource), PRF (Protein Research Foundation), PDB (Protein Data Bank), and translations from annotated coding regions in GenBank/EMBL/DDBJ. UniProt fused TrEMBL, Swiss-Prot, and PIR.

Swiss-Prot (more properly known as UniProtKB/Swiss-Prot) is a manually annotated protein sequence database with information extracted from literature and curator-evaluated computational analysis. Although

Swiss-Prot is at least ten times smaller than UniProt or Proteins, it is a high quality database.

**Searching for Sequence Similarity**

The nucleotide and protein sequences collected in the sequence databases come from very heterogeneous organisms which might have diverged hundreds of millions of years ago. Regardless of this enormous period of time, these organisms share remarkably important similarities, since some of their genes and proteins were present in their last common ancestor. Such genes/proteins are said to be homologous. Homologous genes/proteins can be commonly identified in sequence databases by the comparisons of their nucleotide or amino acid sequences. There are different approaches to perform a sequence search for homologous proteins: the election of the most convenient depends on many factors, such as the size of the database; the expected similarity of the query with their potential targets; the available computer resources; and the speed of the search and the required accuracy of the results among others. Here, we summarize the main search approaches and their corresponding publicly available software.

**Genomic Searches Using Pairwise Comparison**

The simplest approach to perform a database search considers the comparison of a pair of sequences, commonly known as pairwise comparison. This is based on the alignment of two sequences; it can denote that the two sequences are similar either globally or locally. If the similarity is global, similarity may reflect that they are closely related and, hence, they may have the same function. If the similarity is local, this may indicate evolutionary constraints restricted to these particular regions. In order to perform extremely fast and accurate sequence similarity searches of a particular sequence (commonly known as the query sequence) against nucleotide or amino acid databases, the NCBI has developed the BLAST (Basic Local Alignment and Search Tool) set of programs that can be used locally or via their web server. BLAST programs look for small sets of continuous characters or 'words' in the sequences of the database that corresponds to fragments of the sequence query. The length of the words can be specified by the user and depends on the kind of database; commonly, 3 for amino



acid databases and 11 for nucleotide databases. BLAST assumes that the significant alignments contain highly similar pairs of aligned words. The similarity between any pair of words is scored using substitution matrices, such as BLOSUM and PAM, which express the probabilities that one amino acid can be replaced by another in a set of homologous proteins. If the similarity between two words has a score greater than a specific predetermined value, the aligned word is called a 'hit' and is further considered in the analysis. After the identification of all hits, BLAST tries to extend each hit in both directions to connect neighbor hits in a bigger alignment. Insertions and deletions are not considered during this stage of analysis. The resulting extended aligned regions are further considered only if the corresponding alignment scores are greater than a predetermined cut-off value. Finally, BLAST performs a new alignment between the query sequence and the database sequence allowing gaps to extend the regions of the sequence similarity. Each alignment of the search is scored and assigned to a measure of statistical significance called the expectation value (E-value, i.e., the number of alignments with at least the same score that would have been expected to occur in the database by chance) which is used to sort and limit the alignments reported to the user. BLAST takes into account the amino acid composition of the query sequence in the estimation of statistical significance. This composition-based statistical treatment, used in conventional protein BLAST searches, tends to reduce the number of false-positive database hits.

The aforementioned heuristic approach taken by BLAST importantly speeds the process of searching for similar sequences as much as 50 times, in comparison to exhaustive algorithms, that search for the best alignment solution. This speed characteristic of BLAST is particularly important considering the enormous sizes of the commonly used databases such as GenBank or Swiss-Prot.

Considering the type of query sequence and database, the following alternatives in a BLAST search are available: (1) the query and the databases are nucleotide sequences (blastn); (2) the query and the databases are amino acid sequences (blastp); (3) the query is a nucleotide sequence that is translated into its six reading frames to be compared with an amino acid database (blastx); and (4) the query is an amino acid sequence and is compared with a nucleotide database dynamically translated into their six reading frames (tblastn). The Web BLAST output service offers a new Tree View option to create a dendrogram that clusters sequences according to their distances from the query sequence. This display is helpful for organizing the presence of aberrant or unusual sequences or natural groupings of related sequences such as members of a gene family or homologues from other species in the BLAST output. In addition to the

aforementioned alternatives, a comparison of two DNA or protein sequences that produces a dot-plot representation of the alignments can be obtained using BLAST2Sequences.

For genomic searches, the MegaBLAST program can be used. This program was designed to find nearly exact matches and operates up to ten times faster than the standard nucleotide BLAST. MegaBLAST is the default search program for NCBI's Genomic BLAST pages. Genomic BLAST may be used to search the genomic sequence of an organism. Its searches can be displayed within their genomic context using the Map Viewer to show the location of neighboring genes and nearby genomic landmarks.

For finding distant relatives of a protein, a much more sensitive BLAST version called PSI-BLAST (Position-Specific Iterated BLAST) can be used. This program initially performs a standard BLAST search to identify closely related proteins which are then used to elaborate a consensus profile of the protein family that reflects the specific tendency of certain residues to be present in particular positions of the sequence. The profile is used as a new query to perform a new search against the database. As a result of this new search, new members of the family could be identified and considered in addition to the original set of proteins to construct a new and more representative profile. This process can be repeated iteratively until the user considers that all the members in the database have been identified.

### Genomic Searches Using Profiles

Structurally relevant residues as well as catalytic active sites of enzymes have a tendency to be conserved in a family of homologous proteins. These conserved regions or sequence patterns occur repeatedly in a group of related protein or DNA sequences and are known as motifs. Motifs can be represented by a profile matrix where the frequencies for each amino acid at each position are evaluated from the conserved region's residue distribution rather than from a more general distribution. Motifs can be used to search, in a database, for other proteins of the family. This is particularly relevant to identify distantly related proteins that might not present evident similarity across their entire sequences, but conserve the essential residues of the group. Motifs are not exclusive to protein sequences; they can also be used to represent relevant residues in a family of RNA (e.g., tRNA and catalytic RNA) or DNA (e.g., regulatory protein-binding sites) sequences. One of the advantages of the sequence searches based on motifs over the pairwise methods is their ability to use the characteristics of the whole family of sequences during the database search. One of the most common programs used to identify motifs in a set of nucleotide or amino acid related

sequences is MEME (Multiple EM for Motif Elicitation). This program finds, in a set of unaligned sequences, all the motifs that are statistically over-represented. The most worth-mentioning options of MEME allow the user to select: (1) if the motif should be found as one or more occurrences per sequence; (2) the minimum and maximum size of the motif; (3) the minimum statistical significance of the motif; and (4) the maximum number of motifs to find. MEME represents each motif as a scoring matrix that expresses the probability of each possible residue at each position in the pattern. This matrix can be used to search databases for homologue or related sequences using the MAST (Motif Alignment and Search Tool) program. MAST calculates the statistical significance of the matches of a group of motifs characteristic of a protein or a DNA sequence family in a target sequence. For each motif, MAST finds the position in the sequence that best matches, and it represents the statistical significance of the match as a p-value (i.e., the probability of observing a match with a score at least as good when the motif is compared to a random sequence). In order to refine a motif, the resulting sequences obtained by MAST can be selected in conjunction with the seed sequences to perform a new MEME–MAST cycle. This cycle can be repeated until no more new sequences are obtained, resulting in a refined motif. Individual MEME motifs do not contain gaps; instead patterns with gaps are represented by MEME by two or more separate motifs. These individual motifs can be integrated into a more complete model using the program Meta-MEME that combines the set of motifs into a single one using hidden Markov models (HMMs) (see below).

Oligo-analysis is a second program that can be used to perform sequence searches based on motifs. It was originally created to uncover binding sites for transcription factors in *Saccharomyces cerevisiae*, but it works very nicely on any organism. In contrast with heuristic methods, oligo-analysis is an exhaustive algorithm. Its range of detection is however limited to relatively simple patterns: short motifs with a highly conserved core. These features seem to be shared by a good number of regulatory sites in yeast. The oligo-analysis program is contained in the RSA (Regulatory Sequence Analysis) Tools website (see below). Nicely, oligo-analysis has an option to create random sequences; hence, it provides a negative control to assure that the patterns obtained are statistically significant. Analogous to the previously described MAST program, it possesses a tool named pattern matching, which searches for the motif in the genome of interest, and it will draw a new feature-map with the newly predicted sites. Hence, a cyclic process can also be used as the one described for MEME and MAST. In contrast to the MEME and MAST programs, oligo-analysis works better if all the analyzed sequences belong to a single organism, because it adjusts the frequencies of the nucleotide

composition of the genome to the search of over-represented motifs. MEME and MAST also possess this option, but they work better than oligo-analysis especially with large motifs and with a set of orthologous genes.

A comparison of these and other sequence search methods was reviewed by Tompa and colleagues in 2005.

### Genomic Searches Using Multiple Sequence Alignment and Hidden Markov Models

Conserved motifs can also be identified by the comparison of more than two sequences at a time in a process called Multiple Sequence Alignment (MSA). Since the computational process of MSA is much more complex than the simple pairwise alignments, most of the MSA programs use heuristic approaches. One of the most common heuristic MSA protocols is ClustalW, which considers the progressive pairwise alignments on successively less closely related sequences. It can be used via web servers or locally. A second alternative of a progressive alignment program is T-Coffee. This program calculates pairwise alignments by combining the direct alignment of the pair with indirect alignments that aligns each sequence of the pair to a third sequence. T-Coffee is slower than ClustalW, but usually generates more accurate alignments, especially if the sequences to be aligned are distantly related. Finally, MUSCLE (Multiple Sequence Comparison by Log-Expectation) is a third commonly used MSA program. MUSCLE is claimed to perform better and faster MSA than ClustalW or T-Coffee, since the distance measure that it uses to assess the relatedness of two sequences is updated between iteration stages.

MSA are considered in molecular biology as an important primary source of information for different studies, such as phylogenetic analysis (see below) or homology database searches. In the former case, sequence motifs can be identified directly from the conserved regions of the MSA and used to construct models that represent the essential regions of the nucleotide or protein group of sequences. One efficient protocol to simultaneously consider the different motifs from a MSA is by HMMs. A Markov model is a statistical model in which the probability of an event depends on the immediately previous event. In a HMM, the parameters of the process are 'hidden' and the challenge is to determine them from the observable data. In the case of biological sequences, the observed events are represented by the presence of certain types of residues (i.e., nucleotide or amino acid) in a specific column of the alignment and the 'hidden' events are the biological properties of the process, for example, the secondary structure of a protein or the probability of a given DNA sequence to be part of a gene. One of the most popular analysis packages that uses HMMs to perform database searches is HMMER which includes, among others, programs to build the HMM from a multiple

sequence alignment (hmmbuild), to calibrate the model and determine the parameters for more sensitive searches (hmmcalibrate) and to search, in a sequence database, for sequences that match an HMM (hmmsearch).

There are several databases of relevant precompiled HMM of conserved protein families or conserved domains such as Pfam, SMART, TIGRFAMs, and PIRSF which are considered below.

## Protein Databases

### The Cluster of Orthologous Groups of Proteins Database

The NCBI's Clusters of Orthologous Groups of proteins (COGs) database has been designed as an attempt to classify proteins from completely sequenced genomes on the basis of the orthology concept. Orthologues are direct evolutionary counterparts related by vertical descent as opposed to paralogues which are genes within the same genome related by duplication. Typically, orthologous proteins have the same domain architecture and the same function.

The COGs reflect one-to-one relationships. COGs have been identified on the basis of all-against-all sequence comparison of the proteins encoded in complete genomes using the gapped BLAST program. The construction procedure is based on the simple notion that any group of at least three proteins from distant genomes that are more similar to each other than they are to any other proteins from the same genomes are most likely to belong to an orthologous family. This prediction holds even if the absolute level of sequence similarity between the proteins in question is relatively low, and thus the COG approach accommodates both slowly-evolving and fast-evolving genes. Moreover, this allows COGs to accommodate the possibility that a single (or multiple) protein(s) from one genome may be related to several paralogues in a second genome (one-to-many or many-to-many relationships).

The most straightforward application of the COGs is the prediction of functions of individual proteins or protein sets. This is done by fitting proteins into a COG using the COGNITOR program. The COG website offers automatic means to isolate all COGs with a particular phylogenetic pattern (i.e., a pattern of species that are represented or not represented in a given COG), for example, those that are found only in pathogenic bacteria. More generally, the COG system is a convenient platform for a variety of evolution-oriented analyses of protein families.

The COG website contains the following principal types of data: (1) list of all COGs organized by the (predicted) functional category and hyperlinked to (2) individual COG pages; each COG page shows the respective phylogenetic pattern and is hyperlinked to: (a) pictorial

representations of BLAST search outputs for each member of the COG, (b) a multiple alignment of the COG members produced automatically using the ClustalW program (see above), and (c) a cluster dendrogram generated using the BLAST scores (see above) as the measure of similarity between proteins; (3) the COGNITOR page, where a protein sequence can be pasted, searched against the database of proteins from complete genomes, and assigned to a COG; (4) a phylogenetic pattern search tool; and (5) a matrix of co-occurrence of genomes in COGs.

Moreover, a web page that contains additional structural and functional information on the COG as a whole and individual members is now associated with each COG. These pages include systematic classification of the COG members under the current classification systems for enzyme or transporters (if applicable); indications of which COG members (if any) have been genetically and biochemically characterized; information on the domain architecture of the proteins comprising the COG and the three dimensional structure of the domains if known or predictable; a succinct summary of the common structural and functional features of the COG members; and peculiarities of individual COG members.

### Protein Functional Classification and Protein Signatures

Databases with signatures diagnostic for protein families, domains, or functional sites are important tools for the computational functional classification of newly determined sequences that lack biochemical characterization. Computational annotation of protein function is generally obtained via sequence similarity: once a close neighbor with known function has been identified, its annotation is copied to the sequence with known function. This strategy works very well in functionally homogeneous families.

In some cases the sequence of an unknown protein is too distantly related to any protein to detect its resemblance by pairwise sequence alignment (see above). However, relationships can be revealed by the occurrence of a particular motif in its sequence. These motifs, typically around 10–20 amino acids in length, arise because specific residues and regions thought or proved to be important to the biological function of a group of proteins are conserved in both sequence and structure during evolution. These biologically significant regions or residues are generally: enzyme catalytic sites; prosthetic group attachment sites; amino acids involved in binding a metal ion; cysteines involved in disulphide bonds; and regions involved in binding a molecule or another protein. Some families are defined not just by one motif but by the co-occurrence of two or more motifs of low specificity.

The increasing amount of genomic sequences that need to be annotated has led to proliferation of protein domain families and protein domain signature databases.

Protein domains are units of molecular evolution, usually associated with particular aspects of molecular function such as catalysis or binding. In general, they represent discrete units of three dimensional (3D) structures. Most proteins are built up in a modular fashion from two or more domains fused together. The identification of functionally characterized domains in protein sequences may give the first clues as to their molecular and cellular function.

### **PROSITE**

PROSITE is an annotated collection of motif descriptors dedicated to the identification of protein families and domains. The motif descriptors used in PROSITE are either patterns or profiles, which are derived from multiple alignments of homologous sequences. PROSITE patterns are short sequence motifs, while PROSITE profiles are position specific score matrices. Profiles characterize protein domains over their entire length, and they are more sensitive than patterns. Profiles and patterns have complementary qualities. Patterns, confined to small regions with high sequence similarity, are often powerful predictors of protein functions such as enzymatic activities. Profiles covering complete domains are more suitable for predicting protein structural properties.

### **PRINTS**

PRINTS database houses a collection of protein fingerprints. Fingerprints are groups of conserved sequence motifs that together provide diagnostic signatures for protein families. The tools available for searching PRINTS are (1) a BLAST server, for searches against sequences matched in PRINTS database and (2) the FingerPRINTS, which can suit for searches against fingerprints in the database – this affords greater specificity than the BLAST implementation. PRINTS has a hierarchical structure which allows associations to be traced from subfamily to superfamily relations. This is relevant to putative distantly related clan members that share no significant sequence similarity.

### **ProDom**

ProDom is a comprehensive database of protein domain families generated from the global comparison of all available sequences. ProDom families are built by an automated process based on a recursive use of PSI-BLAST homology searches. The ProDom website allows querying of ProDom in a variety of ways, such as accession number, ProDom families, related databases, and keywords. The output is either information on a given domain family or cartoons displaying the domain arrangements of all proteins matching the query. One can also compare a sequence of interest via BLAST to the ProDom database. ProDom will suggest a possible domain arrangement for any query protein. When 3D

structures are available for target domains, the output is directly linked to both SWISS-MODEL and Geno3D servers.

### **Pfam**

Pfam is a database of protein domains and families. It contains curated multiple sequence alignments for each family and corresponding profile HMMs (see above). Pfam families are divided into two categories, Pfam-A and Pfam-B. Each Pfam-A family consists of a curated seed alignment containing a small set of representative members of the family, a profile HMM built from the seed alignment and an automatically generated full alignment which contains all detectable protein sequences belonging to the family. Pfam-B entries are automatically generated from the ProDom database and are represented by a single alignment. The use of representative seed alignments for Pfam-A families allow efficient and sustainable manual curation of alignments and annotation, while the automatic generation of full alignments and Pfam-B clusters ensures that Pfam is a comprehensive classification of protein families that scales effectively with the growth of the sequence database.

Within Pfam, several metagenomic datasets have been included. These new datasets contain many novel protein sequences, which are currently unannotated. This section, within Pfam, enables the community to assess our current understanding of the domain composition found in such environmental datasets. Moreover, this will provide a potential source of new Pfam families and/or allow verification of families where there are few representatives.

### **SMART**

The Simple Modular Architecture Research Tool (SMART) is an online resource used for protein domain identification and the analysis of protein domain architectures. The basic data of SMART are high-quality manually derived alignments of protein domain families. As HMMs (see above), SMART alignments allow the identification of protein domain in sequence databases. Protein sequences can be scanned for the presence of important catalytic amino acids. Absence of one of these amino acids very likely results in loss of catalytic activity. The data provide a framework for understanding the evolution and function of genes and proteins throughout the living world. Its genomic perspective allows further cross-referencing with protein–protein interaction maps, making SMART an invaluable tool for systems biologists to interpret pathways and networks.

### **TIGRFAMs**

TIGRFAMs is a collection of manually curated protein families consisting of HMMs (see above), multiple sequence alignments, commentaries, Gene Ontology (GO) assignments, literature references, and pointers



related to TIGRFAMs, Pfam, and InterPro models. TIGRFAMs contains models of full-length proteins and shorter regions at the level of superfamilies, subfamilies, and equivalogues, where equivalogues are sets of homologous proteins conserved with respect to function since their last common ancestor. The models in the TIGRFAMs database have been built specifically to aid in automated annotation of microbial genes, particularly by focusing on the creation of equivalogue family models. TIGRFAMs uses the term equivalogue to describe the relationship of proteins conserved in function since their last common ancestor, where both orthology and horizontal gene transfer may be part of the evolutionary history.

### **PIRSF**

PIRSF is a network classification system that accommodates a flexible number of levels from superfamily to subfamily to reflect varying degrees of sequence conservation. Members of a PIRSF homeomorphic family share full-length sequence similarity with a common domain architecture (homeomorphic) and have a common evolutionary origin (monophyletic). PIRSF HMMs are designed to cover the full length of a protein sequence, and thus to include all domains within the sequence. In this way, PIRSF homeomorphic families tend to encompass one or more of the existing InterPro domain entries and show the domain composition of UniProt sequences. Classification based on full-length protein allows annotation of both generic biochemical and specific biological functions, identification of domain and family relationships, and classification of multidomain proteins.

## **Integrated Databases**

### **InterPro: A Database of Protein Families**

The EMBL InterPro incorporates the major protein signature databases into a single resource. These include PROSITE, which uses regular expressions and profiles; PRINTS, which uses Position Specific Scoring Matrix-based (PSSM-based) fingerprints; ProDom, which uses automatic sequence clustering; and Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, and Gene3D, all of which use HMMs (see above). Protein signatures from these databases that describe the same family or domain, in terms of sequence positions and protein coverage are integrated into single InterPro entries, to which are added annotation and cross-references. Proteins with 3D structures modeled by MODBASE and SWISS-MODEL have links to the structure predictions from the matched graphical views. These links complement the experimentally determined structures in the PDB.

InterPro unites all these databases capitalizing on their individual strengths, producing a single entity that is far greater than the sum of its parts. A primary application of InterPro is the annotation and functional classification of uncharacterized sequences. The EBI is using InterPro for enhancing the automated annotation of TrEMBL. InterPro has also proven its usefulness for whole proteome analysis with comparative genome analysis in several organisms. InterPro has provided a useful tool for protein sequence analysis and characterization.

### **Conserved Domain Database**

The NCBI conserved domain database (CDD) contains PSI-BLAST-derived Position Specific Score Matrices representing domains taken from the SMART, Pfam, and domain alignments derived from COGs. CDD attempts to collate the set of protein domains characterized so far and to organize related domain models in a hierarchical fashion, meant to reflect major ancient gene duplication events and subsequent functional diversification.

Whenever possible CDD hits are linked to structures which, coupled with a multiple sequence alignment of representatives of the domain hit, are equipped with advanced alignment-building tools that use the PSI-BLAST and threading algorithms. In alignment curation, information from 3D structure and structure superposition is considered when possible to define structurally conserved cores.

The Conserved Domain Architecture Retrieval Tool allows searches of protein databases on the basis of a conserved domain and returns the domain architectures of database proteins containing the query domain.

## **Protein Structure**

### **Protein Data Bank**

PDB includes all the public 3D structures for proteins, nucleic acids, and carbohydrates. PDB records contain atomic ( $x, y, z$ ) coordinates of macromolecules, a brief macromolecular description, the authors of the structure, the biological source, the protein or nucleotide sequence, literature references, and some relevant experimental data. Most 3D structure data are obtained from X-ray crystallography and NMR spectroscopy; they provide a wealth of information on the biological function, on mechanisms linked to the function, and on the evolutionary history of macromolecules and relationships between them. PDB interface provides search and retrieval interfaces and cross references to other structural databases.

There are some domain structure classification databases. Structure Classification of Proteins (SCOP) contains a structural and evolutionary classification of

the proteins in the PDB. In SCOP, a multidomain protein is split up into its constituent domains, which are then considered separately. SCOP is a hierarchical classification system, which utilizes four levels: class, fold, superfamily, and family. class and fold lack evolutionary relationship evidence. CATH is another structure domain hierarchical classification system. The main difference with SCOP is that CATH is a semiautomatic classification while SCOP is manually done.

### Protein Structure Visualization

PDB files are long text files; atom coordinates are not comprehensible by reading these files. Moreover, PDB files do not include connectivity data. For simplification, protein structures can be represented in many ways depending on the information to be conveyed (e.g., wire models for comparisons, ribbon models to highlight secondary structures, ball and stick models to detail, and surface models for electrostatic potentials). There are several PDB visualization tools that transform the coordinates into virtual 3D structures. One of the most popular free software available is RasMol, whose drawback is that one must master its command-line language. An amateur-friendly, free visualization program is DeepView (Swiss Pdb-Viewer), which has links to many bioinformatics resources.

### Protein Structure Alignments

Long distance evolutionary relationships can be detected by protein structure similarities. This is useful when no detectable protein sequence homologues are available for the gene of interest, but instead, the structure of the gene of interest is known. If this is the case, structural homologues can be found. MAMMOTH (Matching molecular models obtained from theory) is a pairwise comparison method. MAMMOTH takes a PDB file as query and searches in a PDB files database, like SCOP. Using a heuristic algorithm, it calculates a structural similarity score based on the likelihood of obtaining an alignment between two proteins or between two conformations of the same protein by chance.

Moreover, Multiple Structure Alignments (MStA) can also be performed using Dali and MAMMOTH-mult server. Dali searches the PDB for those structures similar to the query. For this server, the query can be either a coordinate file or a PDB ID. Dali can also perform a two-structure alignment. On the contrary, MAMMOTH-mult server can be used in two ways: it can either multiple align a target protein against a given SCOP superfamily or align among them a set of input proteins. Both servers will return an MStA via e-mail.

### Protein Structure Prediction

In 1962, C.B. Anfinsen demonstrated for ribonuclease that protein structure is encoded in the protein sequence. Since then, many efforts have been made for protein structure prediction based on the amino acid sequence. Nowadays, some programs based on homologous structures are good predictors. Some programs for structure determination, based on physicochemical and statistical properties, have been developed successfully, but these programs still need a user with protein structure and bioinformatics expertise. Homology based programs work much better.

Swiss-Model is a fully automated protein structure homology server. The entry is a protein sequence and the output is an atom coordinate file and the names of the structure templates. This server divides the prediction into five stages. First, it finds protein homologues with known structure by sequence comparison. Second, it selects those templates with more than 25% identity in tracks of more than 20 residues; some sequences do not have homologues with known structure, so for these cases Swiss-Model cannot give an answer. Third, it makes some input files. Fourth, it generates the model by threading the new amino acid sequence into the known backbone. Fifth, it refines the model by energy minimization to avoid clashes and holes.

MODELLER is another program for protein structure prediction based on homology; it is more powerful than Swiss-Model, since many parameters can be changed. Nevertheless, MODELLER is not fully automated and it must be installed in your own computer.

### Fold Recognition Tools

Another approach to know the 3D structure of a sequence is fold recognition. Fold recognition is used when the query protein is distantly related to protein(s) with known structure; many of these programs are based on homology detection by HMM. The structures generated by fold recognition used to be less accurate than those from prediction programs, and should be used carefully. PSIPRED and PHYRE are protein recognition programs that run via web pages.

### Fold Recognition Databases

The Gene3D database is focused on providing structural annotation for protein sequences without structural representatives, including the complete proteome sets. The structural annotation is generated using HMMs based on the CATH domain families. Gene3D maps CATH domain families to protein sequences. This is a similar task to that carried out by SUPERFAMILY for SCOP. The Gene3D website includes a BLAST search

facility that will identify the likely family that the query sequence belongs to.

### Protein Classification Based on Ontology

The GO project has defined specific terms and vocabulary to describe common properties of genes and proteins. GO is a structured network consisting of defined terms and relationships between them that describe three attributes of gene products: their molecular function, biological process, and cellular component. There are three separate aspects to this effort: first, the development and maintenance of the ontologies themselves; second, the annotation of gene products, which entails making associations between the ontologies and the genes and gene products in the collaborating databases; and third, development of tools that facilitate the creation, maintenance, and use of ontologies.

### Resources for Genome-Scale Analysis

The NCBI Entrez Genome provides access to over 370 complete microbial genomic sequences. Specialized viewers and BLAST pages are also available for viruses. Genomes are chosen from alphabetical listing or a phylogenetic tree and can be examined at increasing levels of details ranging from a graphical overview of an entire genome to the level of a single gene. At the level of a genome or chromosome, a coding region display gives the locations of coding regions, the lengths, names, and GenBank identification numbers of the protein products. An RNA genes view lists the locations and names for ribosomal and transfer RNA genes. A summary of COG functional groups is also presented. At the level of a single gene, links are provided to sequence neighbors for the implied protein with links to the COGs database.

For complete microbial genomes, precomputed BLAST neighbors for protein sequences, including their taxonomic distribution and links to 3D structures, are given in TaxTables and PDBTables, respectively. The Entrez Genome Project database is supported by providing an overview of the status of complete and in-progress large-scale sequencing, assembly, annotation, and mapping projects. For bacterial organisms, Genome Project indexes a number of characteristics of interest to biologists, such as organism morphology and motility; environmental requirements, such as salinity, temperature, and pH range; oxygen requirements; and pathogenicity. The database allows genome sequence centers to register their project early in the sequencing process so that project data can be linked to other NCBI-hosted data at the earliest opportunity.

NCBI's TaxPlot plots similarities in the proteomes of two organisms to that of a reference organism for more than 580 bacterial and archaeal genomes.

### Comprehensive Microbial Resource

Comprehensive Microbial Resource (CMR) contains robust annotation of all complete microbial genomes and allows a wide variety of data retrieval. Retrievals can be based on protein properties such as molecular weight or hydrophobicity, GC-content, functional role assignments and taxonomy. The CMR also has special web-based tools divided into Genome Tools and Comparative Tools. Genome Tools include a list of the available genomes; a list of all genes in a selected genome; a list of genes ordered by categories; detailed information on each of the DNA molecules found in the organism (chromosomes, plasmids) including the topology (linear or circular), length, A, T, G, C percentage and number of genes; information on the characteristics of organisms derived from genomic data and literature sources; KEGG pathway displays; lists of all the intergenic regions for a selected DNA molecule as well as lists all the interRNA regions for a selected DNA molecule. The Genome Tools also possess different graphical displays to view the genome, the genetic context of selected genes, and a circular image of a selected DNA molecule, including representation of all genes on the molecule as well as all tRNAs and rRNAs. One can also retrieve sequence or a list of genes between a pair of coordinates for the selected DNA molecule, search all proteins in a genome for a given motif, display restriction digest information for a genome, display a computer model of a 2-dimensional gel for a selected organism, display the %GC for a set 'window' of nucleotides across the entire DNA molecule, show codon usage within a genome, and use the computer program Primer3 to find primers for a selected gene or organism.

Moreover, the Comparative Tools include Protein Homology Tools, which show the number of proteins in a reference genome that have hits up to 15 selected comparison genomes, the display of orthologue information across genomes, and the number of protein hits a reference genome has in common with all of the genomes in the CMR based on blast searches. It can also compare the number of protein hits between a reference and comparison organism in a scatter plot and compare the %GC content between a reference and comparison organism. It can align genomes using MUMmer to align any two DNA molecules in the database based on exact DNA sequence matches. It uses NUCmer to compare two closely related DNA molecules and PROmer to compare the protein sequences between two DNA molecules in the database.

### The Genomes Online Database

The Genomes Online Database (GOLD) is a World Wide Web resource for comprehensive access to information regarding complete and ongoing genome projects, as well as metagenomes and metadata, around the world. According to GOLD, nearly 700 microbial genomes have been published at the time of writing with over 3000 other projects ongoing and in the process of being launched.

### Genome Reviews

The Genome Reviews database provides an up-to-date, standardized and comprehensively annotated view of the genomic sequence of organisms with completely deciphered genomes. Currently, Genome Reviews contains the genomes of archaea, bacteria, bacteriophage, and selected eukaryotes. Genome Reviews is available as a MySQL relational database or a flat file format derived from the EMBL Nucleotide Sequence Database.

### Microbes Online

Microbes Online is a publicly available suite of web-based comparative genomic tools, which include operon and regulon predictions, a multispecies genome browser, a multispecies GO browser, a comparative KEGG metabolic pathway viewer, a Bioinformatics Workbench for in-depth sequence analysis, and Gene Carts that allow users to save genes of interest for further study while they browse. An additional interface for genome annotation is provided.

### Integr8

The Integr8 web portal provides easy access to integrated information about deciphered genomes and their corresponding proteomes. Available data includes DNA sequences (from databases including the EMBL Nucleotide Sequence Database, Genome Reviews, and Ensembl); protein sequences (from databases including the UniProt Knowledgebase and IPI); statistical genome and proteome analysis (performed using InterPro, CluSTr, and GOA); and information about orthology, paralogy, and synteny.

### UCSC Genome Browser

The UCSC Genome Browser provides support for genome-centric exploration of archaeal genomes enriched with data from computational analysis and experimental studies.

Microbial genomes can be explored using a variety of analysis tools provided by resources that often further enrich the data in archival or curated public resources, some of which are described in this article.

### Metabolomics

#### The Kyoto Encyclopedia of Genes and Genomes

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a comprehensive set of metabolic pathways. KEGG is displayed as graphic maps, than can be viewed as a network of genes (enzymes) or as a network of compounds. KEGG is constructed from four databases: PATHWAY contains 354 reference pathways manually drawn, GENES is a collection of the genes and proteins from complete genomes, LIGAND is a compilation of chemical compounds including drugs approved in United States and Japan, and BRITE is a collection of hierarchies and binary relations of the other databases. KEGG includes genes of 574 bacteria and 49 archaea (KEGG version 45; 8 January 2008). KEGG graphic interface is nicely presented. Reference pathway maps show general organization metabolism in various species; these maps are boxes (proteins) and circles (compounds) connected by arrows. Color representations indicate presence of those proteins in a particular organism. Atlas tool displays global maps that can be colored by the user to highlight any genes or compounds. KEGG can be browsed directly over the pathway graphs, or by NCBI, UniProt, or KEGG identifiers. Lists of orthologous genes are provided as well as several cross-links to other databases. KEGG can be used for modeling and simulation, search, and retrieval. It also includes sections of genetic information processing and environmental transduction (e.g., two-component systems).

#### EcoCyc and BioCyc: An Encyclopedia of Genes and Metabolism

EcoCyc is a comprehensive database resource for *Escherichia coli*. It contains curated information of genome, transcription regulation, membrane transporters, and metabolism. Chromosome maps can be browsed by gene name, or nucleotide number, to see graphical representation of that particular region. Pathways can also be browsed to obtain maps. Fully bibliographic information is obtained. BioCyc is the extended version for other genomes and metagenomes. Genomes can be selected for comparative analysis. Since these databases rely on available literature, EcoCyc is the more interesting of these databases.

### Metagenomics

'Metagenomics' describes the functional and sequence-based analysis of the collective microbial genomes contained in an environmental sample. Current microbiological culturing techniques are inadequate for studying the vast majority of microorganisms. Consequently, many organisms



remain underrepresented in the main sequence databases. Recently, with the advent of better sequencing technologies, large samples from environments such as the sea have been sequenced directly, thereby avoiding the need for culturing. Sequencing using this approach gives rise to many sequences from a diverse set of organisms, albeit at low read coverage and with no knowledge of the source organisms. For example, these advances have enabled the adaptation of shotgun sequencing to metagenomic samples. A 2004 metagenomic study of the Sargasso Sea found DNA from nearly 2000 different species including 148 types of bacteria never seen before, obtaining over 1 million kilobase of nonredundant sequence. Metagenomics is reviewed in 'Metagenomics'.

### Structure and Function of Microbial Communities

Metagenomics concerns the extraction, cloning, and analysis of the entire genetic complement of a habitat. Metagenome analysis is expected to provide a comprehensive picture of the gene functions and metabolic capacity of microbial communities. Several statistical tools for describing and comparing microbial communities have been developed by Patrick Schloss and Jo Handelsman. Some of these tools include DOTUR, which calculates an estimate of the richness and diversity in a community; SONS, which defines the structure and memberships of two communities; LIBSHUFF, which is a statistical test to compare community structures and determines whether two samples are drawn from the same population or whether one is a subset of the other; and TreeClimber, which describes gene flow from a given phylogeny, that is, determines whether the differences between two communities arose due to random variation or whether lineages from one community had become more dominant through negative or positive selection pressures.

### Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis

Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) is a web resource for metagenomic research. CAMERA's debut coincides with the publication of the Venter Global Ocean Sampling expedition's extensive dataset cataloguing of over 6 million new genes from uncultured marine microbes.

CAMERA's aim is to create a rich, distinctive data repository and bioinformatics tools resource that will address many of the unique challenges of metagenomics and enable researchers to unravel the biology of environmental microorganisms. CAMERA's database includes environmental metagenomic and genomic sequence

data, associated environmental parameters ('metadata'), precompiled search results, and software tools to support powerful cross-analysis of environmental samples.

### Integrated Microbial Genomes/Metagenomes

Integrated Microbial Genomes/Metagenomes (IMG/M) is an experimental metagenome data management and analysis system, which provides tools and viewers for analyzing both metagenomes and isolated genomes individually or in a comparative context. Comparative analysis of the metagenomes in the context of available reference isolated genomes could potentially reveal large-scale patterns of biochemical interactions and habitat-specific correlations in the host environment that might otherwise be missed.

Functional roles of genes can be characterized in the context of pathways, whereby pathways are associated with genes via gene products that can function as enzymes catalyzing individual reactions of metabolic pathways. Similar to isolated microbes, the metabolic capacity of a whole microbiome can be characterized by analyzing the metabolic maps inferred from the gene content and distribution of its composite genome. Comparative data analysis plays an important role in understanding the biology of isolated microbial genomes. Similar to isolated genomes, the analysis of metagenomes in the comparative context of other genomes is substantially more efficient than analyzing each metagenome in isolation. Microbiome samples can be compared in terms of presence and abundance of certain gene families or of certain metabolic pathways. These analyses help to infer the metabolic capabilities of the component organisms in the community, and thus identifying the key members of the microbiome that perform community-essential tasks and pinpoint the metabolic interactions within the microbiome and between the microbiome and its host environment.

IMG integrates bacterial, archaeal, and selected eukaryotic genomic data collected from multiple data sources. In addition to the isolated genomes, IMG/M includes metagenome sequences generated from an acid mine drainage biofilm, an agricultural soil sample, three isolated deep sea 'whale fall' carcasses, and two enhanced biological phosphorus removing sludge samples.

The data model for the IMG/M data warehouse allows integrating primary genomic sequence information, computationally predicted and curated gene models, precomputed sequence similarity relationships, and functional annotations and pathway information in a coherent biological context. Isolated organisms are identified via their taxonomic lineage (domain, phylum, class, order, family, genus, species, and strain). For each genome, the primary DNA sequence and its organization in scaffolds and/or contigs are recorded. Genomic features,

such as predicted coding sequences and some functional RNAs are also recorded. Protein coding genes are further characterized in terms of molecular function and participation in pathways. Proteins are grouped into families based on sequence similarity. Pathways, reactions, and compounds are included from KEGG and LIGAND. Additional functional annotations according to GO terms are provided by EBI Genome Reviews, while COG provides clusters of orthologous groups of genes (see above). Orthologue and paralogue gene relationships for isolated microbial organisms are computed based on bidirectional best hit (BBH) with clusters formed using a Markov Clustering method. Isolate organisms are characterized in terms of phenotypes (e.g., morphology and geochemistry), ecotype (including geographical coordinates), and disease.

A Phylogenetic Profiler tool allows comparing the gene content between isolate genomes or metagenomes, by defining a profile for the genes in terms of presence or absence of homologues in other entities. Similar to isolate genomes, differences in gene content between metagenomes can be correlated with a specific phenotype or environment, while comparison of the gene content within the metagenome helps inferring the metabolic capabilities of the component populations and identifying the organisms that may be responsible for community-essential tasks.

The Occurrence Profile tools allow examining profiles of genes and functions across metagenomes and isolate organisms. This might give insights of the evolutionary history of the selected gene and may potentially be functionally linked, or co-regulated in a pathway. The Functional Occurrence Profile tools, such as COG Profile, Pfam Profile, and Enzyme Profile, show the occurrence profiles for functional characterizations such as COGs, Pfam families, or enzymes involved in pathways metagenomes and genomes. This tool is especially useful for analysis of datasets obtained from the communities with high species diversity, where little or no sequence assembly can be achieved; for such datasets identification of predominant families allows users to infer habitat-specific biological traits.

IMG/M provides support for the exploration and comparative analysis of metagenomes and their component population in the context of other metagenomes and isolate genomes.

## **Taxonomy and Phylogeny**

### **UniProt Taxonomy Database**

The UniProt taxonomy database integrates taxonomy data compiled in the NCBI database and data specific to the UniProt Knowledgebase. Organisms are classified in a hierarchical tree structure. This taxonomy database

contains every node (taxon) of the tree. UniProtKB taxonomy data are manually curated; next to manually verified organism names, it provides a selection of external links, organism strains, and viral host information.

### **NCBI Taxonomy Database**

The NCBI taxonomy database indexes named organisms that are represented in the databases with at least one nucleotide or protein sequence. The Taxonomy Browser can be used to view the taxonomic position or retrieve data from any of the principal Entrez databases for a particular organism or group. Entrez Taxonomy displays include custom taxonomic trees representing user-specified subsets of the full NCBI taxonomy.

### **Phylogeny Prediction**

Phylogeny and molecular evolution have provided a huge toolbox to study the evolutionary history of organisms, allowing inferences of phylogenetic relations (ancestor–descendent) of protein domains, genes, and organisms. The resulting phylogenetic hypotheses are crucial to make phylogeny predictions or inferences. It also allows estimation of evolutionary forces (such as selection, genetic drift, migration, and recombination) in protein domains, genes, genomes, and populations. Phylogeny can make important hypotheses such as the universal tree of life and distinction between orthologues and paralogues as well as the important inferences of their function, for example, the likely change of function between paralogue proteins.

Phylogeny's objective is to trace an ancestor–descendant relation of organisms through different taxonomic levels. The markers used to build a phylogeny are contained in the (DNA or protein) sequences, and these include restriction fragment length polymorphisms (RFLPs), genomic fingerprints, among others. The sequences that contain this information must be aligned with the previously described bioinformatic programs like ClustalW, T-Coffee, and MUSCLE.

In general, DNA sequences give a finer resolution of the evolutionary history of an organism since a great variability exists in the substitution rate within DNA sequences, for example, comparing coding regions and intergenic regions, catalytic residues versus noncatalytic residues, structural domains versus nonstructural domains, third positions versus first and second positions of codons in coding sequences, and stems versus loops of rRNAs and tRNAs. Moreover, different genes evolve at different rates; viral genes evolve very fast in contrast to the slow evolutionary rate of 16S rRNAs.

Horizontal gene transfer (HGT) and homoplasy represent a problem and a limitation of phylogenies. Homoplasy occurs when characters are similar, but are

not derived from a common ancestor. There are several types of homoplasy: parallel evolution, which is the independent evolution to reach the same final state, from the same ancestral state; convergent evolution, which is the independent evolution to reach the same final state, from a different ancestral state; and secondary loss, which is a reversion to the ancestral state.

A phylogenetic tree is a mathematical structure used to represent the evolutionary history among a group of sequences or organisms. Phylogenetic inference requires a precise selection of the method to use from all the available ones, given a set of sequences. The aim of phylogenetic inference is to obtain the best estimate of an evolutionary history based on the incomplete and noisy information contained in the sequences.

One of the most commonly used methods to construct a phylogenetic tree is based on distances between sequences coming from a multiple sequence alignment. The distance values are arranged as a distance matrix whose values depend on the evolutionary model selected and could be used to calculate the tree by the Unweighted Pair Group Method with Arithmetic mean (UPGMA) and Neighbor Joining (NJ) methods. The clustering methods, UPGMA and NJ, reconstruct the tree from a distance matrix. These are very fast methods, but very sensitive to certain parameters such as the order in which Operational Taxonomic Unit (OTUs) are added to the tree. This is because the distance matrix is built pairwise, that is, a distance measure is chosen to quantify the differences between a pair of items. NJ and UPGMA are good only to have a quick idea of how your tree looks, but the resulting tree will not be robust.

Alternatively to distance methods, trees can be constructed using discrete methods such as Maximum Parsimony (MP), Maximum Likelihood (ML), and Bayesian. These methods consider each site (column) in the alignment directly. The MP and ML methods follow a different optimization criterion, which allows a better selection of the resulting topology from millions of topologies that are to be analyzed. The big limitation of these optimization criteria is that they are computationally costly.

MP assumes an implicit evolutionary model that prefers the resulting phylogeny with the minimum substitutions needed. Parsimony informative sites are those that partition sequences into at least two groups each with at least two members. MP uses the branch and bound optimization algorithm. Exhaustive search and branch and bound methods guarantee finding the best tree. However, exhaustive search methods do not work for anything more than ten sequences on existing single-processor computers. It is important to consider that MP often gives multiple equally parsimonious trees making it difficult to choose among them; it underestimates branch lengths and it does not take account of multiple substitutions at a given site.

ML tree reconstruction is an explicit statistical technique based on the likelihood framework. ML makes several assumptions of substitution models. The most typical are (1) the probability of a change is independent of the prior history of the site (a Markov Model; see above), (2) substitution probabilities do not change with time or over the tree (a homogeneous Markov process), and (3) change is time reversible. All sites are informative because a site that has the same base for two sequences tells us something about the time separating the two molecules.

There are several ML advantages: it is mathematically rigorous and performs well in computer simulations, it takes into account multiple/hidden substitutions, and there is a large support of statistical theory for likelihood estimation and inference and extensions to Bayesian analysis. However, there are disadvantages: ML may be inconsistent if the model of evolution is miss-specified, it is computationally tedious and intensive, and it is not immediately intuitive.

### Rooted and Unrooted Trees

A rooted phylogenetic tree is a tree that has a unique node that corresponds to the most recent common ancestor of all the elements in the tree. The strategy that is normally used to root a tree is the inclusion of an outgroup. This outgroup should be close enough to the rest of the sequences in order to infer phylogenetic relationships, but far enough to be a clear outgroup. On the other contrary, an unrooted tree is commonly used to show the relatedness of the leaf nodes without making assumptions about common ancestry. Obtaining a tree is seldom the end of an analysis. It is usually the beginning. There are several statistical tests that one may want to perform concerning the quality of the phylogenetic tree and the data about the process of evolution and about patterns of evolution. Several methods have been proposed that attach numerical values to nodes in trees that are intended to provide some measure of the strength of support for that node. The most popular of these is the bootstrap. Bootstrapping is a modern statistical technique that uses computer-intensive random resampling of data to determine sampling error or confidence intervals for some estimated parameter. In bootstrap phylogenies, characters are resampled with replacement to create many bootstrap replicate datasets. Each bootstrap replicate dataset is analyzed. The agreement among the resulting trees is summarized with a majority-rule consensus tree. The frequencies of occurrence of groups, bootstrap proportions, are a measure of support for those groups. High BPs (e.g., >80%) are indicative of strong support for a particular clade. Usually, 1000 or more bootstrap pseudoreplicates are generated. The bootstrap is totally reliant on the accuracy of the tree-building method. One can get

good bootstrap support for the wrong group if the tree-building method is inappropriate.

## Phylogenetic Analysis Algorithms

### **MrBayes**

MrBayes is a program for the Bayesian inference of phylogenies from nucleic acid sequences, protein sequences, and morphological characters. Bayesian inference of phylogeny is based on a quantity called the posterior probability distribution of trees, which is the probability of a tree conditioned on the observations. The conditioning is accomplished using Bayes's theorem. The posterior probability distribution of trees is impossible to calculate analytically; instead, MrBayes uses a simulation technique called Markov chain Monte Carlo (MCMC) to approximate the posterior probabilities of trees.

The output is several files with the parameters that were sampled by the MCMC algorithm. MrBayes can summarize the information in these files for the user. It can also use a hierarchical Bayesian framework to infer sites that are under natural selection. It allows for rate variation among sites and a variety of models of sequence evolution. Testing the resulting tree of MrBayes differs from the other programs. The numbers in the branches are not bootstrap values, but probabilities *a posteriori* for each clade. Since these probabilities are always higher than the bootstrap probabilities, it cannot be considered equally. Moreover, we have to take into account the fact that MCMC starts from a random position, so it will take some time before it reaches the general vicinity of values from the target distribution. This period is called 'burn-in' period, and any nonrepresentative samples taken during this period should be discarded. The easiest way to determine how long to allow for a burn-in is to plot a parameter of interest to determine if it has plateaued. This information is contained in the parameter output file of MrBayes.

### **PAUP\***

PAUP\* originally meant Phylogenetic Analysis Using Parsimony. PAUP\* is a major analytical tool in phylogenetic analysis. It makes available a very wide variety of analytical methods in a single environment and can be operated via window/mouse, command-line, or scripts. It includes parsimony, distance matrix, invariants, maximum likelihood methods, and many indices and statistical tests. Unfortunately PAUP\* is a commercial program (available from Sinauer Associates), although it is quite a good value.

### **PHYLIP**

PHYLIP (the Phylogeny Inference Package) is one of the most important packages of programs for inferring phylogenies. It is available free over the Internet, and written to work on as many different kinds of computer systems as

possible. Methods that are available in the package include parsimony, distance matrix, and likelihood methods, including bootstrapping and consensus trees. Data types that can be handled include molecular sequences (protein, DNA, and RNA sequences), gene frequencies, restriction sites and fragments, distance matrices, and discrete characters.

### **Phylogenetic analysis using maximum likelihood**

Phylogenetic analysis using maximum likelihood (PAML) is a package of programs for phylogenetic analyses of DNA or protein sequences using Maximum Likelihood. The programs can estimate branch lengths in a phylogenetic tree and parameters in the evolutionary model such as the transition/transversion rate ratio, the gamma parameter for variable substitution rates among sites, rate parameters for different genes, and synonymous and nonsynonymous substitution rates. PAML can also test evolutionary models, calculate substitution rates at particular sites, reconstruct ancestral nucleotide or amino acid sequences, simulate DNA and protein sequence evolution, compute distances based on the synonymous and nonsynonymous changes, and of course do phylogenetic tree reconstruction by ML and Bayesian Markov Chain Monte Carlo methods.

### **TreeView**

TreeView is a program for displaying and manipulating trees. It can draw rooted and unrooted trees, display bootstrap values, and edit trees by moving branches, collapsing them, and rerooting. It provides a simple way to view the contents of a NEXUS, PHYLIP, Hennig86, ClustalW, or other format tree file, and allows the user to create publication quality trees.

## Universal Tree of Life

The amount and diversity of species with at least partial sequence information is rapidly increasing and the tree of life is constantly being redrawn. Phylogenetic trees represent a backbone of various other biological studies and it is therefore essential to have the state-of-art tools for their display, customization, and interpretation. In the following section we will describe some of these.

### **iTOL**

Interactive Tree of Life (iTOL) is a web based tool for the display, manipulation, and annotation of phylogenetic trees. Branches can be pruned or collapsed, and any node can be used to reroot the tree. Various types of data, such as genome size or protein domain repertoires, can be mapped onto the tree. iTOL can automatically determine taxonomic classes of all internal nodes and assign proper scientific names to leaves. iTOL is the first visualization tool that supports the display of HGTs. Export to several bitmap and vector graphics formats are supported.



## ARB

The ribosomal RNA (rRNA) molecule has been considered the 'gold-standard' for the investigation of the phylogeny and ecology of microorganisms. The rapidly increasing number of available rRNA sequences led to the development of ARB. ARB is an integrated package of cooperating software tools for data handling and analysis that fulfils the necessity of rRNA-based identification systems. A central database of processed (aligned) sequences and any type of additional data linked to the respective sequence entries is structured according to phylogeny or other user-defined criteria.

It provides tools for building up databases of RNA sequences, aligning them, and searching, editing, modifying, profiling, and constructing trees. ARB uses its own RNA sequence database, which is a manually curated and quality checked dataset for ribosomal RNA genes. These datasets are maintained in collaboration with the 'arbsilva' project and can be obtained from the ARB Silva database site. For phylogenies, it uses programs from PHYLIP and fastDNAmI, as well as its own ARB Neighbor-Joining program. ARB also incorporates a variety of other sequence analysis software. It can handle large numbers of sequences and has sophisticated tree drawing and manipulation. ARB is distributed as executables for a variety of versions of UNIX.

## The SILVA system

This is a system implemented to provide a central comprehensive web resource to update quality-controlled databases of aligned rRNA sequences from the Bacteria, Archaea and Eukarya domains. SILVA serves as the main data source for ARB. In addition to the ARB approach, there are currently three projects offering access to a set of curated rRNA sequence and alignment databases: the European rRNA Databank, the Ribosomal Database Project, and the greengenes project. All four projects offer at least one 16S rRNA dataset, but vary in the amount of sequences, quality checks, alignments, and update procedures. However, the ARB project is the only platform that actively incorporates homologous small (SSU) as well as large (LSU) subunit sequences from all three domains of life, the Bacteria, Archaea (16S/23S), and Eukarya (18S/28S).

## Resources for the Analysis of Gene Expression

When a microarray is used to assay gene expression, the resulting data must be analyzed to tell which genes are up-regulated, down-regulated or do not present a change in the given experiment. These changes, derived from the fluorescence intensity of each probe in the array, must be translated into numerical values. These measured values

are sometimes irreproducible. Moreover, these values must be normalized in order to be compared with the other conditions in the experiment, or with other arrays.

There are three widely used techniques that can be used to normalize gene-expression data from single array hybridization: (1) total intensity normalization, (2) normalization using regression techniques, and (3) normalization using ratio statistics. All of these techniques assume that all (or most) of the genes in the array should have an average expression ratio equal to one. The normalization factor used to adjust the data to compensate for experimental variability and to normalize the fluorescence signals from the two samples being compared.

Once significant signals are obtained, depending on the experiment's objective, the next step is to cluster the signals from genes with similar expression. Various clustering techniques can be applied to the identification of patterns in gene-expression data. Most of the cluster analysis techniques are hierarchical. These differ in the manner in which distances are calculated between the growing clusters and the remaining members of the dataset, including other clusters. Clustering algorithms include, but are not limited to, the following: (1) Single-linkage clustering, which tends to produce clusters that are 'loose' because clusters can be joined if any two members are close together. (2) Complete-linkage clustering, which tends to produce very compact clusters of elements and the clusters are often very similar in size. (3) Average-linkage clustering. There are, in fact, various methods for calculating averages; the most common is the UPGMA. In this method, the two clusters with the lowest average distance are joined together to form a new cluster (see above) called (4) weighted pair-group average, which is identical to UPGMA, except that in the computations, the size of the respective clusters is used as a weight. Hence, this method (rather than UPGMA) should be used when the cluster sizes are suspected to be greatly uneven: (5) within-groups clustering, which is similar to UPGMA except that clusters are merged and a cluster average is used for further calculations rather than the individual cluster elements. This tends to produce tighter clusters than UPGMA: (6) Ward's method, which produces the smallest possible increase in the sum of square errors.

When the microarray datasets are ready to be uploaded in several publicly available databases, it is desirable that they meet certain criteria such as MIAME and preferably a MAGE-TAB format. MIAME describes the Minimum Information About a Microarray Experiment that is needed to enable the interpretation of the results of the experiment unambiguously and potentially to reproduce the experiment. MIAME does not specify a particular format; however, obviously the data are more useful if they are encoded in a way that the essential information specified by MIAME can be

accessed easily. Most of the databases described below support the use of the MAGE-TAB format, which is based on spreadsheets or MAGE-ML. The public repositories ArrayExpress at the EBI (UK), GEO at NCBI (US), Stanford Microarray Database (SMD) at Stanford (US), and CIBEX at DDBJ (Japan) are designed to accept, hold, and distribute MIAME compliant microarray data.

The information contained in these gene expression arrays can serve different purposes in understanding the biology of the selected organism; in particular, they are commonly used for the prediction of regulatory motifs of co-regulated genes. The regulatory regions of these genes are usually used as input to programs such as MEME, oligo-analysis, and RibEx, which were described above. In general, the results obtained from these programs may give insights regarding the discovery of new regulatory sites.

### **The Stanford Microarray Database**

The SMD is a research tool and archive that allows hundreds of researchers worldwide to store, annotate, analyze, and share data generated by microarray technology. SMD supports most major microarray platforms and is MIAME-supportive. The primary mission of SMD is to be a research tool that supports researchers from the point of data generation to data publication and dissemination, but it also provides unrestricted access to analysis tools and public data from 300 publications.

SMD stores gene expression as well as array CGH (comparative genomic hybridization) and chIP-chip (chromatin immunoprecipitation on array) experiments. SMD supports multiple microarray platforms (spotted cDNA or oligonucleotide analysis, Affymetrix, Agilent, Combimatrix, and Nimblegen arrays). It has a data pipeline that can communicate published data directly to ArrayExpress and GEO.

Defining a set of microarrays of interest is the first step for an analysis process. SMD provides two different search forms as effective tools to locate microarrays of interest. Once the microarrays are selected, users decide whether they intend to do an analysis that is applicable to one array at a time or to a group of arrays. Since microarray data are known to be sensitive to several experimental factors, it is important to be able to assess the quality of the data collected from a microarray and to normalize the data appropriately. SMD has several tools that can be used for the assessment of microarray quality. Background correction and normalization methods are often used to correct experimental biases in microarray data, and SMD accordingly provides access to several normalization and background correction methods using the marray and limma packages, respectively, from BioConductor.

### **Gene Expression Omnibus**

The NCBI's Gene Expression Omnibus (GEO) is a data repository and retrieval system for microarray and other forms of high-throughput molecular abundance data generated by the scientific community. In addition to gene expression data, GEO accepts array CGH data, ChIP-chip data, SNP array data, and some proteomic data types. The GEO repository accepts MIAME-compliant data submissions.

### ***E. coli* GenExpDB**

The University of Oklahoma Bioinformatics Core hosts the *E. coli* Gene Expression Database. *E. coli* GenExpDB was designed from the biologist's perspective, with a tool box-style, integrated and interactive display interface and simplified warehouse architecture; it has a wide variety of microarrays for different conditions.

Expression-Ring displays gene expression ratio data in a blue/yellow heat map of concentric rings, one for each experiment, with all displayed Transcription Factor (TF) genes labeled on outside ring(s) and with hyperbolas connecting to all target genes regulated by each TF.

### **ArrayExpress**

ArrayExpress is a public database for high-throughput functional genomics data. ArrayExpress consists of two parts: the ArrayExpress Repository, which is a MIAME supportive public archive of microarray data, and the ArrayExpress Data Warehouse, which is a database of gene expression profiles selected from the repository and consistently reannotated. Archived experiments can be queried by experiment attributes, such as keywords, species, array platform, authors, journals, or accession numbers. Gene expression profiles can be queried by gene names and properties, such as GO terms, and gene expression profiles can be visualized.

### **Centre for Information Biology Gene Expression Database**

The gene expression database Centre for Information Biology Gene Expression Database (CIBEX), with a data retrieval system in compliance with MIAME, a standard that the MGED Society has developed for comparing data produced in microarray experiments in different laboratories worldwide. CIBEX serves as a public repository for a wide range of high-throughput experimental data in gene expression research, including microarray-based experiments measuring mRNA, serial analysis of gene expression (SAGE tags), and mass spectrometry proteomic data.

## Resources for Proteomics

### Swiss-2DPAGE

Swiss-2DPAGE database stores data from two-dimensional polyacrylamide gel electrophoresis including reference maps, images, mapping procedures, and cross-references. Data can be searched by protein name, protein accession number, by experimental pI/MW range, and by clicking on the spot.

### Open Mass Spectrometry Search Algorithm

The NCBI's Open Mass Spectrometry Search Algorithm (OMSSA) is an efficient search engine for identifying tandem MS (MS/MS) peptide spectra by searching libraries of known protein sequences. It allows up to 2000 spectra to be analyzed in a single session using either BLAST 'nr' or refseq\_protein sequence libraries for comparison. Standalone versions of OMSSA for several popular computer platforms that accept larger batches of spectra and allows searches of custom sequence libraries are available.

### Mascot: A Search Engine that Uses Mass Spectrometry Data to Identify Proteins from Primary Sequence Databases

Additionally to OMSSA, there is another search engine called Mascot. This tool uses mass spectrometry data to identify proteins from primary sequence databases. Mascot is unique in that it integrates all of the proven methods of searching: (1) peptide mass fingerprint, in which the only experimental data are peptide mass values; (2) sequence query, in which peptide mass data are combined with amino acid sequence and composition information; (3) MS/MS ion search, which uses uninterpreted MS/MS data from one or more peptides.

### The Proteomics Identification Database

Proteomics Identification Database (PRIDE) is a centralized, standards compliant, public data repository for proteomics data. This database contains information from experiments, identified proteins, identified peptides, unique peptides, and spectra. PRIDE offers a web-based query interface, a user-friendly data upload facility, and a documented application programming interface for direct computational access. It supports identification from both MS-based and gel-based techniques. Processed peak list arising from MS, MS/MS, and higher MS levels are supported. PRIDE retrieves the complete set of protein identifications for a publication, along with the supporting peptide identifications and hyperlinks to further

information. PRIDE also finds all relevant datasets for a particular protein of interest.

### The Open Proteomics Database

Open Proteomics Database (OPD) stores and disseminates MS-based proteomics data. The data residing in OPD represent diverse proteomics samples – some interpreted, some uninterpreted, some on simple but defined samples to be used for training algorithms, and some on highly complex samples, such as whole-cell lysates from different organisms. In all, proteomics data from *E. coli*, *Mycobacterium smegmatis*, *S. cerevisiae*, and *Homo sapiens* are represented with roughly 400 000 total mass spectra, cataloguing the expression of several thousand proteins overall. All data are freely accessible with the intent that computational groups interested in studying the many computational problems posed by proteomics will have a source of protein mass spectra and expression data.

## Resources for Gene Regulation Analysis

As was reviewed in 'Posttranscriptional regulation' and 'Transcriptional regulation', regulation of gene expression can occur at multiple levels, transcription initiation being the most common way of regulation, but also can take place at the posttranscriptional level. In any case, the main regulatory elements are localized upstream of operons; however, regulatory sites are sometimes found inside or at the end of the transcription unit. Our current knowledge of different genes, operons, and regulatory mechanisms is quite variable. For a few model organisms, such as *E. coli* or *Bacillus subtilis*, the regulatory mechanisms are very well characterized for the vast majority of their genes; whereas for the majority of other organisms, their regulatory elements have been poorly characterized or not characterized at all. Therefore, regulatory databases of model organisms are of great value to infer the gene regulation in other phylogenetically related organisms. In any case, these kinds of databases constitute a useful source to guide and design experimental work.

### RegulonDB: A Database for Transcriptional Regulation in *E. coli*

RegulonDB is the internationally recognized reference database of *E. coli* K-12 offering curated knowledge of the regulatory network and operon organization. It is currently the largest electronically encoded database of the regulatory network of any free-living organism. It is a model of the complex regulation of transcription initiation or regulatory network of the cell, as well as a model of the organization of the genes in transcription units, operons, and simple and complex regulons. Continuous

curation of the original scientific literature provides the evidence behind every single object and feature. This knowledge is complemented by comprehensive computational predictions across the complete genome. Literature-based and predicted data are clearly distinguished in the database. RegulonDB public releases are synchronized with those of EcoCyc, since RegulonDB's curation supports both databases. The complex biology of regulation is simplified in a navigation scheme based on three major streams: genes, operons, and regulons. Regulatory knowledge is directly available in every navigation step. Displays combine graphic and textual information and are organized allowing different levels of detail and biological context. This knowledge is the backbone of an integrated system for the graphic display of the network, graphic, and tabular microarray comparisons with curated and predicted objects, as well as predictions across bacterial genomes and predicted networks of functionally related gene products.

With RegulonDB, the user can get mechanistic information about the different transcription units and their regulatory elements, such as promoters and their sigma factor types, genes and their ribosome binding sites, terminators, binding site of specific transcriptional regulator (TRs) as well as their organization into regulatory phrases, active and inactive conformations of TRs and regulons simple and complex.

### **DBTBS: A Database for Transcriptional Regulation in *B. subtilis***

The counterpart of RegulonDB is DBTBS (Database of Transcriptional Regulation in *B. subtilis*), a reference database of transcriptional regulation in the other model organism, *B. subtilis*, summarizing the experimentally characterized transcription factors, their recognition sequences, and the genes they regulate. The goal of this database is to help elucidate its complete gene regulatory network. The construction of the DBTBS aims to compare the results of systematic experiments with the rich source of individual experimental results accumulated so far. The DBTBS database contains a collection of experimentally validated gene regulatory relations and the corresponding transcription factor binding sites upstream of *B. subtilis* genes as well as experimentally validated *B. subtilis* operons and their terminators. Its current version is constructed by surveying the scientific literature and contains the information of binding factors and gene regulatory relations. For each promoter, all of its known *cis*-elements are listed according to their positions, while these *cis*-elements are aligned to illustrate the consensus sequence for each transcription factor. All probable transcription factors coded in the genome are classified using Pfam motifs.

Given the increase in the number of fully sequenced bacterial genomes, DBTBS has extended its usability in comparative regulatory genomics. A new section on the conservation of the upstream regulatory sequences among homologous genes in 40 Gram-positive bacterial species as well as on the presence of overrepresented hexameric motifs that may have regulatory functions was created.

### **Predictive Web Pages on Gene Regulation**

In addition to the two aforementioned regulatory databases that compile data from molecular biology experimental work, there are some other databases and web pages dealing with the *in silico* prediction of regulatory elements. Although the accuracy of the computer predictions is obviously not as solid as the data coming from the experimental analysis, predictive analysis is very important for the design of working hypotheses. Some examples of predictive regulatory databases and web sites are as follows.

#### **Neural Network Promoter Prediction**

Computer prediction of promoter elements is one of the most commonly used analyses of experimental scientists. Neural Network Promoter Prediction is a web page that predicts both prokaryote and eukaryote promoters based on neural networks that have been trained to recognize promoter elements using a pruning iterative procedure that deletes those weights in the network that add the lowest predictive value to the overall promoter prediction. This pruned neural network gives clues about the importance of specific positions in the different types of promoter elements by studying their relative weights.

#### **WebSIDD**

WebSIDD is a web-based service designed to predict locations and extents of the stress-induced duplex destabilization (SIDD) that occur in a double-stranded DNA sequence, on which a specified level of super-helical stress has been imposed. The algorithm calculates the approximate equilibrium statistical mechanical distribution of a population of identical molecules among the accessible states. Its output is the calculated transition probability and destabilization energy of each base pair in the sequence. The structural and energy parameters used in the calculation are all determined experimentally. This method has illuminated the roles of SIDD properties in the regulation of diverse biological processes, such as transcription initiation (promoter prediction) and termination. The prediction of promoter sequences is much more accurate if it considers the prediction of SIDD and sequence-dependent motifs finder algorithms are taken simultaneously.



### **Regulatory sequence analysis tools**

This site offers a series of tools dedicated to the detection of regulatory signals in noncoding sequences that are grouped into the following main blocks of analysis: (1) sequence retrieval, (2) regulatory pattern discovery (string-based pattern discovery, matrix-based pattern discovery; see above, 'Oligo-analysis'), and (3) regulatory pattern matching that includes the genome-scale pattern matching to scan entire genomes for genes having a particular regulatory motif. Interestingly, this web site offers a computer application to predict regulatory motifs from clusters of co-expressed genes based on microarray data. Each one of the Regulatory Sequence Analysis Tools (RSA Tools) is presented as a form to fill. For each form, a manual page provides detailed information about the parameters. The RSA Tools web site offers a set of clear tutorials to get familiarized with their tools.

### **Predicted transcription attenuation in bacteria**

Gene regulation by transcription termination–antitermination, often called transcription attenuation, is a strategy commonly used by bacteria to sense a specific metabolic signal and enables a response that directs RNA polymerase to either terminate transcription or transcribe the downstream genes of an operon (for a review of these mechanisms see 'Posttranscriptional Regulation').

The decision whether to terminate transcription is often based on the selective arrangement of one of the two mutually exclusive RNA secondary structures in the nascent transcript, the antiterminator and the terminator. Transcription attenuation web page compiles a computer-based predict transcription attenuators for fully sequence genomes. The computer predictions are based on the search of potential alternative RNA-hairpin structures in the leader sequence that precedes a particular gene or operon. The predicted transcription attenuators in this database are clustered by organisms or by COG classification.

### **RibEx: a web server for the prediction of riboswitches and other conserved regulatory elements**

This web tool clusters the intergenic region of orthologous genes by an iterative process; conserved motifs across phylogenetically distant organisms are identified. These motifs correspond to reported riboswitches and other likely regulatory systems that appear to depend on conserved RNA structures. A riboswitch is a part of an mRNA leader sequence capable of binding, with great specificity and affinity, a signal molecule without the intervention of any protein factor. One part of the riboswitch can fold to form either of two alternative hairpin structures, one of which functions as an intrinsic transcriptional terminator or a secondary structure that blocks gene translation. The binding of the riboswitch to the

metabolite controls the downstream gene expression by selecting between these two alternative conformations. RibEx allows the visual inspection of these conserved motifs and riboswitches in any sequence given by the user.

### **GeCont: a web server to analyze the genome context of orthologous genes**

In bacteria, the coordinate transcription of functionally related genes that belong to the same pathway or process is commonly accomplished by the operon structure. Based on this property, gene function can sometimes be inferred by the inspection of the function of its neighboring genes. This idea is particularly true if the gene context is conserved among many other genomes. GeCont is a web server designed to show the genomic context of a particular gene and their orthologous counterparts, based on COG classification, in the set of fully sequenced organisms.

### **Public Available Software for the Analysis of Gene Regulation**

#### **Consensus**

In addition to the aforementioned MEME/MAST and HMMER programs that identifies and searches for conserved motifs in a set of given sequences, the Consensus program has been used to identify regulatory sequences. The operating principle of Consensus assumes that regulatory motifs can be represented by weight matrices. For this purpose, Consensus uses a greedy algorithm that searches for the matrix with maximum information content. It first finds the pair of sequences that share a motif with greatest information content, then finds a third sequence that can be added to the previously identified motif resulting in the motif with the greatest information content, and so on.

### **MBDBs and Analysis Programs of RNA Regulatory Elements**

During the past few years, new roles of RNA molecules in the control of RNA expression have been elucidated. As a result of the continuous efforts done in this field, important databases have been created. For instance, The Wellcome Trust Sanger Institute in collaboration with Janelia Farm have created the Rfam database, which is a large collection of multiple sequence alignments and covariance models covering many common noncoding RNA families.

### **Rfam and INFERNAL: A Database for RNA Families and Analysis Software**

Rfam aims to facilitate the identification and classification of new members of known sequence families, and

distributes annotation of noncoding RNAs (ncRNAs) in over 200 complete genome sequences. A small number of families are essential in all three kingdoms of life with large numbers of smaller families specific for certain taxa. The Rfam database is, thus, a comprehensive collection of ncRNAs, whose products are components of some of the most important cellular machineries, such as the ribosome, the spliceosome, and the telomerase. The known repertoire of ncRNA cellular functions is expanding rapidly. Ribozymes catalyze a range of reactions, such as self-cleavage of hepatitis delta virus transcripts and 5' maturation of tRNAs by the ubiquitous RNase P. Riboswitches are in *cis*-regulatory sequences capable of regulating gene expression by directly sensing a metabolite without the intervention of a protein. Examples of some riboswitches and other RNA regulatory elements are described in Posttranscriptional Regulation.

Like Pfam for protein-coding genes, ncRNA sequences can be grouped into families, and much can be learnt about structure and function from multiple sequence alignments of such families. Unlike proteins, ncRNAs often conserve a base-paired secondary structure with low primary sequence similarity. The combined secondary structure and primary sequence profile of a MSA of ncRNAs can be captured by statistical models, called profile stochastic context-free grammars (SCFGs), analogous to profile HMMs of protein alignments.

This database comprises a covariance model for each RNA family, represented by a MSA and profile SCFGs available through its web page. Each family and its model or profile can be downloaded, and then, in conjunction with the INFERNAL software, it is possible to search for any one of the RNA families in a particular sequence or genome.

INFERNAL (Inference of RNA Alignment) is an implementation of a special case of profile SCFGs called covariance models. A CM is like a sequence profile, but it scores a combination of sequence consensus and RNA secondary structure consensus; so in many cases, it is more capable of identifying RNA homologues that conserve their secondary structure more than their primary sequence.

The INFERNAL software package makes consensus RNA secondary structure profiles and uses them to create new structure-based multiple sequence alignments. To make a profile, you need to have a multiple sequence alignment of an RNA sequence family, and the alignment must be annotated with a consensus RNA secondary structure. The program *cmbuild* takes an annotated multiple alignment as input, and outputs a profile. A profile of a known model, for instance the T box riboswitch, can be downloaded from the Rfam database. Once a profile is obtained (either built from the user's sequences or downloaded from Rfam), it can be used to search for homologue sequences in a database using the program *cmsearch*. The profile can also be used by the program

*cmalign* to align a set of unaligned sequences. The previous procedure allows the user to build a hand-curated representative alignment of RNA sequence family, and then use this profile to automatically align any number of sequences to that seed profile. This is the strategy used to maintain the Rfam database of RNA multiple alignments and profiles.

### Prediction of Secondary Structure

The development of computational tools for the interconnection of sequence and structural information to annotate and discover ncRNAs faces two main limitations: the requirements in computational resources and our understanding of RNA structural and evolutionary rules. As the secondary structure is the main energetic component of RNA architecture, it produces strong constraints for the tertiary structure and its definition constitutes a first and essential step.

Mfold and RNAfold from the Vienna package predict an RNA structure which guarantees a convergence to the minimal free-energy structure. In contrast, HotKnots does not guarantee a convergence to the minimal free-energy structure, its algorithm is based on the widely used free-energy minimization, and by using a heuristic approximation and an extended free-energy model is able to compute the predicted RNA structure, including pseudoknots, in a reasonable amount of time. Alternatively, the RNAMST web server searches for possible RNA structures, with previously user-defined constraints, against several databases, such as Rfam. *cmfinder* can also be used to search structural homologues in databases.

Multiple alignments of RNA sequences are rarely available. Several programs can be used to achieve this task. Comparative analysis of a family of homologous sequences enables derivation of covariations within an RNA family and thus the identification of the locations of regular helices. However, the identification of the conserved core of the secondary structure within a MSA is a difficult and iterative task, usually performed by hand using human expertise. Many tools enable realization of this structural alignment automatically. Some of these tools predict secondary structure for each individual sequence and perform an alignment of these derived structures, such as SCARNA, STRAL, and MARNA. The main drawback is that folding of any single RNA sequence might not produce the biologically active RNA structure. Other tools propose to predict the secondary structure and align RNA sequences simultaneously. These are called Sankoff-like methods, and FOLDALIGN, STEMLOC, and Dynalign are examples of this. Most of these programs limit their computational costs by doing a pairwise alignment (SCARNA, STEMLOC, FOLDALIGN, and Dynalign), while others (STRAL) make use of heuristic methods or limit sequence length (MARNA). Another strategy is to align unfolded RNA

sequences according to a reference molecule with a secondary structure that has been well described using HMMs, which is used by the program PSTAG.

Hence, if a multiple alignment of RNA sequences is available, Mifold proposes a Matlab package to exploit mutual information measure in order to identify covering sites. RNAalifold, from the Vienna package and for which a web interface is available, predicts the consensus secondary structure for the RNA alignment by minimizing the overall free energy and uses information on compensatory mutations between the sequences to improve secondary structure predictions. An exception to this is RNacast, now integrated into the RNashapes analysis package, which can predict an RNA consensus structure from a set of unaligned sequences. RNacast uses the concept of RNA families, which share a common structure leading to searching in a reduced space and hence decreasing the computing requirements. Finally, KNetFold is a machine learning approach, which measures the distance between each column from an alignment compared to the Rfam alignments. Hence, this program can estimate which columns are paired. The previously described cmfinder uses the Vienna RNA package, and takes as input a set of unaligned sequences and searches the most conserved secondary structures whose length and number of stem-loops are within a user-defined range. The selected subset can be used to construct an initial multiple alignment, which is improved using the probabilistic framework provided by COVE. RNamine searches for a frequently appearing pattern of stems. This defined pattern of stems can be searched between multiple RNA families. RNamine, as well as cmfinder, can find a conserved structure in a subset of input sequences.

As previously described for computational approaches for overrepresented sequence motifs, such as MEME and MAST, cyclic processes also apply for the prediction of secondary structures and for finding structural homologues. Once a structural signature is available for a given RNA family, the next logical step is to make a refinement and expansion of the ncRNA family using a genome-scale homology search. An example of this can be done with cmbuild-cmfinder cycles, where from a set of input sequences an RNA profile can be constructed and then be used to search within a given database. The new sequences found can be used for a new set of input sequences for cmbuild, and so on.

The computational tools developed to discover secondary structure from sequence and to discover structural information from RNA families face two main limitations: the requirements in computational resources and our understanding of the RNA structural and evolutionary rules. Among the most promising initiatives in this field is the creation of a consortium dedicated to the construction of a common, dynamic, and controlled vocabulary (or

ontology) that should capture all the RNA concepts and their relations, inspired in the previously described GO.

## Dedicated Integration Systems for Molecular Biology Databases

As could be observed in this review, MBDBs are very heterogeneous and their distribution is widespread. Important efforts have been taken for the integration of Biological Databases such as Entrez, DBGET, and more importantly SRS that is an indexing and retrieval tool for flat file data libraries, such as the EMBL, SwissProt, or PROSITE (see above). For this purpose, SRS has developed a special language called ODD that recognizes the different library formats and organizations and extracts other data needed during retrieval. SRS has a friendly web interface that allows easy inspection of retrieval of the entries.

A second example of a dedicated integrative system is called STRING aimed at predicting direct (physical) and indirect (functional) protein-protein interactions based on the quantitative integration of data coming from (1) genomic context, (2) high-throughput experiments, (3) gene co-expression, and (4) reported knowledge. These predictions are very important to relate the molecular functions of individual proteins in a more general and integrated context.

## Future of Biological Databases

The exponential growth of most of the aforementioned databases makes clear that their size in the near future will become an important problem. Specialized software dedicated to the maintenance and efficient updating of the information will be required. In addition, the highly diverse types of data that biologists require, such as microarray expression data, metabolic, and protein-protein interaction networks or protein structure, just to mention some of them, creates a crucial challenge to combine and integrate all of the MBDBs by cross-references, with a similar aim as the aforementioned SRS database. Furthermore, in order to fully exploit these database resources, data mining and new knowledge discovery algorithms will need to be developed.

MBDBs are highly diverse, although at some point, most of them are interconnected. Their use depends on the kind of question that has to be solved, and almost always there is more than one pathway to conduct a study. In any case, a DNA sequence is commonly one of the simplest units of information that a user might have as a starting point. Regarding its nature, the relevance of the sequence is commonly considered as a coding region, Open Reading Frame (ORF), or as a DNA/RNA regulatory element. This is the main branch point for many studies.

In the case of coding sequences, a common objective is the protein function assignment. To this end, the DNA sequence is initially translated and used as a query to perform either pairwise searches using BLAST or motif based searches using HMM algorithms such as HMMER. In any case, the common aim of a search is the identification of homologue counterparts, since they might share a similar biological/biochemical function. In the first case, the BLAST search can be done against large databases (e.g., GenBank) or specialized databases (e.g., Swiss-Prot or PDB). In addition to this kind of search, the user might ask if the sequence has conserved motifs previously identified (e.g., protein families). This is particularly important to identify distantly related proteins that might not present evident similarity across their entire sequences, but in smaller discrete regions. These conserved motifs usually correspond to catalytic or prosthetic sites, binding domains, or structure determinants. The user can identify these motifs using the web services of InterPro that incorporates the major protein signature databases, such as PROSITE, PRINTS, ProDom, Pfam, SMART, TIGRFAMs, PIRSF, SUPERFAMILY, and Gene3D, among others. Interestingly, more than one motif can be considered in a database search using HMM programs such as HMMER.

For the analysis of noncoding sequences, such as regulatory elements, a similar strategy can be followed, since the principles of these database searches are the same. This analysis involves specialized databases, web services, and computer programs that consider the DNA/RNA nature of the sequence, for example, the Watson–Crick kind of interactions between nucleic acid molecules. In the case of regulatory elements, it is important to consider that gene regulation in bacteria mainly takes place at transcription initiation or at posttranscriptional events by RNA elements located in the 5' region of the transcription units. Therefore, a common approach to identify potential regulatory sites or elements considers the analysis of over-represented motifs in a set of 5' upstream regions of co-expressed genes, commonly identified by microarray experiments. The resulting motifs can be compared against known regulatory sites in databases of model organisms such as *E. coli* (RegulonDB) or *B. subtilis* (DBTBS) or against previously identified RNA regulatory elements described in databases such as Rfam or RibEx.

Regardless of the kind of analysis selected, an issue that the user should keep in mind is that MBDBs are important sources of information to formulate or verify scientific hypotheses. Searching for related data on

previously reported scientific literature in databases such as Entrez or PubMed may give new insights to the working project. Incorporating this knowledge into the user's results produces a more complete dataset that can be used as an input to restart the process. The new dataset may be refined in comparison with the previous one, generating, with each cycle, a better and more comprehensive scientific model.

**See also:** Metagenomics; Posttranscriptional Regulation; Transcriptional Regulation

## Further Reading

- Altschul SF, Madden TL, Schaffer AA, *et al.* (1997) Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research* 25: 3389–3402.
- Bailey TL and Gribskov M (1998) Combining evidence using p-values: Application to sequence homology searches. *Bioinformatics* 14: 48–54.
- Baxevas AD and Ouellette BFF (2001) *Bioinformatics*. United States of America: Wiley-Interscience.
- DeLong EF and Karl DM (2005) Genomic perspectives in microbial oceanography. *Nature* 437: 336–342.
- Eddy SR (2004) How do RNA folding algorithms work? *Nature Biotechnology* 22: 1457–1458.
- Eddy SR (2004) What is a hidden Markov model? *Nature Biotechnology* 22: 1315–1316.
- Eddy SR (2004) What is Bayesian statistics? *Nature Biotechnology* 22: 1177–1178.
- Finn RD, Tate J, Mistry J, *et al.* 2008. The Pfam protein families database. *Nucleic Acids Research* 36: D281–D288.
- Kopp J and Schwede T (2006) The SWISS-MODEL repository: New features and functionalities. *Nucleic Acids Research* 34: D315–D318.
- Kouranov A, Xie L, de la CJ, *et al.* (2006) The RCSB PDB information portal for structural genomics. *Nucleic Acids Research* 34: D302–D305.
- Larkin MA, Blackshields G, Brown NP, *et al.* (2007) Clustal W and Clustal X version 2.0. *Bioinformatics*.
- Lee D, Redfern O, and Orengo C (2007) Predicting protein function from sequence and structure. *Nature Reviews Molecular Cell Biology* 8: 995–1005.
- Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, and Bork P (2006) SMART 5: Domains in the context of genomes and networks. *Nucleic Acids Research* 34: D257–D260.
- Mulder NJ, Apweiler R, Attwood TK, *et al.* (2005) InterPro, progress and status in 2005. *Nucleic Acids Research* 33: D201–D205.
- Petsko GA and Ringe D (2004) *Protein Structure and Function*. London, UK: Sinauer Associates, Incorporated.
- Tompa M, Li N, Bailey TL, *et al.* (2005) Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* 23: 137–144.
- Wheeler DL, Barrett T, and Benson DA (2007) Database resources of the national center for biotechnology information. *Nucleic Acids Research* 35: D5–D12.
- Whitfield EJ, Pruess M, and Apweiler R (2006) Bioinformatics database infrastructure for biotechnology research. *Journal of Biotechnology* 124: 629–639.



# Elucidating metabolic pathways and digging for genes of unknown function in microbial communities: the riboswitch approach

A. Gutiérrez-Preciado and E. Merino

Department of Molecular Microbiology, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

## Abstract

In the current post-genomic era, only 3% of all genes have been annotated based on experimental evidence. Even though functions can readily be predicted for many genes, 25% of these are likely to be wrong. The most widely used methods for function prediction rely on sequence similarity, which might be misleading in many cases. Other methods such as genomic context or phylogenetic profiles have been developed to increase gene annotation accuracy; nevertheless these are only efficient when complete genome sequences are available. Here we propose a new approach based on riboswitch identification. Riboswitches are highly conserved regulators of gene expression located in the 5' untranslated region of certain genes. When transcribed they adopt three-dimensional structures that recognize their ligands with great affinity and specificity. This specificity is a key issue for our method, allowing functional assignment with great accuracy.

**Keywords:** Gene function, gene regulation, genome annotation, riboswitches, T box

**Original Submission:** 22 November 2011; **Revised Submission:** 20 March 2012; **Accepted:** 20 March 2012

Editors: A. Moya, Rafael Cantón, and D. Raoult

*Clin Microbiol Infect* 2012; **18** (Suppl. 4): 35–39

**Corresponding author:** Enrique Merino, Instituto de Biotecnología, UNAM, Av. Universidad #2001, Col. Chamilpa, C.P. 62210, Cuernavaca, Morelos, México  
**E-mail:** merino@ibt.unam.mx

## Introduction: function assignment of recently sequenced genes

A common task in life sciences is the assignment of a biological function to a recently sequenced gene. Massive genomic and metagenomic sequencing projects have unveiled a large collection of genes whose function remains to be determined. On average, one-third of the genes from a given genome have poorly understood or unknown functions. The most common method for assigning a biological function to a new gene is through homology inference. In the best scenario, a similarity search (e.g. using BLAST) will find a clear homologue with a known function. In these cases, it is likely that both genes have the same or a similar function. However, many proteins have high sequence similarity despite performing different functions (e.g. paralogues TrpE and PabA [1]). The converse is also true, where proteins with

low similarity can present the same structure and function (e.g. AroE and YdiB [2]). Only 3% of today's sequences are annotated based on experimental evidence and it has been estimated that over 25% of existing sequences are annotated incorrectly [3].

## The riboswitch approach

Knowing how a particular gene is regulated can provide insights on its overall nature, the metabolic pathway in which it participates and even, in some cases, the conditions in which it is expressed. Since gene regulation in bacteria mainly takes place at the transcriptional level, identification of regulatory elements in the upstream region of transcription units is of crucial relevance. The first models of bacterial regulation were dependent on regulatory proteins, the binding sites of which tend to be short and with a low degree of conservation. Thus, gene function assignment based on the prediction of regulatory elements has been poorly explored. Recently, a new family of regulatory elements has gained importance: riboswitches. In contrast to regulatory proteins, riboswitches are not free molecules; they are part of the

transcription unit that carries the genes that will be regulated, either by activating or repressing their expression. Since 2001, a total of 14 riboswitch families have been described and experimentally characterized. Most of them bind small metabolites with high affinity and specificity, such as amino acids, vitamins or nucleotides. By definition, riboswitches recognize their target molecules in complete absence of proteins, thanks to their complex RNA structure that imposes a high level of conservation at both the three-dimensional level and the underlying nucleotide sequence [4]. These features contribute to placing riboswitches as excellent candidates to be identified with great accuracy in comparative genomic studies. Furthermore, a significant number of genes in a genome are regulated by a riboswitch. For example in *Bacillus subtilis*, 110 out of 4105 genes are regulated by a riboswitch.

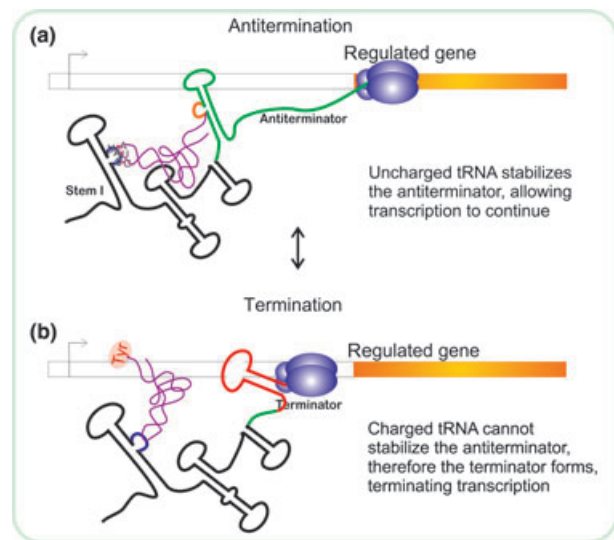
As a representative example of the scope of our riboswitch approach, we present an analysis of gene function assignment using the T box, which was the first class of riboswitch to be identified. In contrast to most riboswitches, which recognize small metabolites, the T box recognizes specific tRNAs and distinguishes between their charged/uncharged states. This is due to Watson–Crick interactions that occur between the T box and the anticodon on the tRNA (Fig. 1). For each operon that is regulated by a T box, a specific tRNA is recognized and the relative levels of its charged/uncharged state determine if transcription will proceed (Fig. 1) [5,6]. Many genes that are under the control of this regulatory element are involved in increasing the intracellular levels of a specific amino acid, e.g. amino acid biosynthetic genes, transporter genes and regulatory genes. Proteins that aminoacetylate a specific tRNA can also be regulated by the T box mechanism [11].

## Aims and impact

Given the specific recognition by the T box riboswitch for each of the different tRNAs, the aim of this project was to develop a new method to infer the biochemical or metabolic function of genes in metagenomic sequences. The method can also be applied to any other set of genes for which functional evidence is lacking (e.g. complete genomes or whole-genome shotgun projects).

## Methods

The main steps of our method for functional inference are as follows.



**FIG. 1.** Model of the T box regulatory mechanism. Structural model of the *Bacillus subtilis* *tyrS* T box leader RNA, as originally described by Henkin *et al.* [16]. The standard T box leader RNA arrangement consists of three major elements, stem I, stem II and stem III, plus the stem IIA/stem IIB pseudoknot and the competing terminator and antiterminator structures; the drawing is simplified. The Specifier loop, shown in blue, is an internal bulge in stem I and contains the Specifier sequence (zoomed in (a)): UAC residues complementary to the anticodon sequence of tRNA<sup>Tyr</sup>. The antiterminator structure (green) has a bulge (orange) that interacts with the unpaired residues at the acceptor end of an uncharged tRNA. During the transcription of the leader region by RNA polymerase (purple ovals), the nascent RNA folds into a structure competent for binding of the cognate tRNA at two sites. The binding of uncharged tRNA (a) to both the Specifier sequence and the antiterminator bulge stabilizes the antiterminator (green RNA segment), preventing the formation of the terminator. This allows transcription to proceed into the downstream coding sequence (orange box). Charged tRNA (b) (represented by Tyr attached to the 3' end of the tRNA) can interact with the Specifier sequence but cannot interact with the antiterminator; a failure to stabilize the antiterminator allows the formation of the terminator helix (red RNA segment), and transcription is terminated before the downstream coding region can be transcribed.

1 *Operon predictions.* Completely sequenced genomes as well as metagenomes were used as input. The transcription units (operons) in each of the genomic and metagenomic sequences were predicted based on the intergenic distances and on the functional relationships of the protein products of contiguous genes, obtained from the STRING database (as described previously [7]).

2 *T box identification.* The upstream region of each operon was used to identify the T box regulatory element by two complementary bioinformatic approaches: sequence

conservation (using the MEME and MAST programs [8,9]) and secondary structure conservation (using the CMSEARCH program from the INFERNAL package [10]). *Ad hoc* Perl scripts were created to identify which tRNA can be bound by each T box. This is performed by locating the stem I and then the Specifier loop in which the codon–anticodon interactions indicate the T box specificity [11].

- 3** *Function assignment of the regulated genes.* This analysis was done using BLAST to match the genes against COG [12], KEGG [13] and other functional databases. The description of the genes in NCBI was also taken into account.
- 4** *Identification of relevant cases for the formulation of metabolic models.* As previously indicated, the T box riboswitch was originally found regulating aminoacyl tRNA synthetases. Interestingly, these are not the only genes regulated by this mechanism; different organisms use T boxes to regulate the expression of genes involved in amino acid biosynthesis by non-canonical pathways in response to a particular metabolic demand. When this is the case, the model suggested by T box regulation must consider the congruence of the type of tRNA that is being sensed and the metabolic function of the regulated genes.

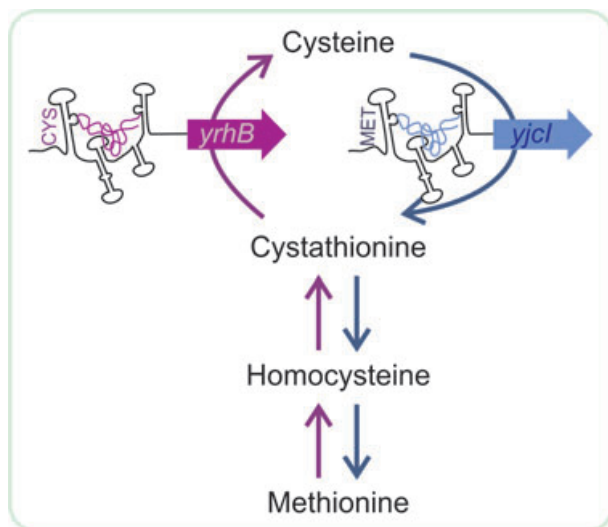
## Results

We performed our analysis on the genomes of 248 fully sequenced Firmicutes, encompassing a total of 771 528 genes. From this set of genes, we identified 7034 genes (0.91%) that are regulated by a T box and can be divided in four different categories in accordance with their annotated functions, as follows.

- 1** Genes that have a clear functional annotation that is consistent with T-box-regulated genes, according to the literature. These include aminoacyl tRNA synthetases and genes involved in amino acid biosynthetic pathways (for a review, see [11]). We found 4805 genes in this category.
- 2** Genes that have only a general annotation. We found 857 genes in this category, most of them corresponding to ‘transporters’. The identification of a T box element upstream of these genes strongly suggests that they are participating in amino acid uptake. Moreover, we can predict the particular amino acid for each transporter by considering the tRNA that would be specifically sensed by its upstream T box. A representative example that illustrates this category is the *yvbW* gene in *Bacillus subtilis*.

Our analysis revealed that this gene codes for a permease that is regulated by a T box specific for tRNA<sup>Leu</sup> (also called a Leu T box). Hence, we predict that this gene codes for a leucine transporter. Experimental evidence supporting our conclusion has been obtained by the Henkin laboratory [14]. Using a similar approach we could identify a specific amino acid in nearly 90% of the cases of these poorly annotated genes.

- 3** Genes without functional annotation. These genes are commonly annotated as ‘hypothetical proteins’. An interesting example is found in the methionine and cysteine biosynthetic pathway of some Firmicutes. The components of this route have been experimentally determined in *B. subtilis*, where an interconversion step between methionine and cysteine takes place and is performed by a pair of paralogous genes: *yjcl* and *yrhB*. The product of *yjcl* synthesizes methionine from cystathionine, which is synthesized from cysteine. On the other hand, the product of *yrhB* synthesizes cysteine from cystathionine. Yjcl and YrhB are 50% identical which makes it difficult to distinguish them through sequence comparison. However, their corresponding genes are regulated by different riboswitches in many Firmicutes, and this makes it possible to predict which amino acid will be synthesized (Fig. 2).
- 4** Genes with annotation that is inconsistent with a T box regulation or that are likely to be annotated incorrectly. Since the great majority of genes have been annotated by non-supervised methods based on sequence similarity criteria without experimental characterization, the functional annotations of some of them are incorrect. This mistaken annotation is commonly passed from one gene to another when new sequence comparisons are performed. Examples showing incorrectly annotated genes or canonical genes that could be performing a novel function in some bacteria include those of the *por* and *etf* operons in anaerobic *Clostridia*. The *por* operon codes for proteins similar to the subunits of the pyruvate : ferredoxin oxidoreductase (POR), which catalyzes a thiamine pyrophosphate-dependent oxidative decarboxylation of pyruvate to form acetyl-CoA and CO<sub>2</sub>. This reaction, normally occurring in central metabolism is, biochemically speaking, highly similar to that of the second step of Ile biosynthesis, which is usually done by *IlvB*. The *por* operon in these *Clostridia* appears to be regulated by an Ile T box. Hence, this POR complex, responding to Ile deficiencies, could be substituting for the reaction usually catalysed by *IlvB* and participating in Ile biosynthesis using different substrates, providing metabolic versatility (Fig. 3). An analogous example is the *etf* operon, also in

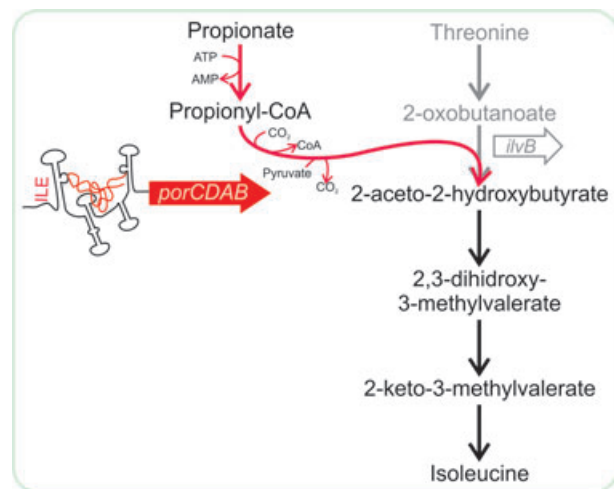


**FIG. 2.** Cysteine and methionine interconversion. YrhB and Yjcl are 50% identical at the amino acid sequence level, which makes it difficult to determine through sequence comparison which gene makes cysteine and which makes cystathionine. However, these genes are regulated by riboswitches in the Firmicutes, which makes it possible to differentiate which amino acid will be synthesized. For instance, in *Bacillus subtilis*, *yjcl* is regulated by an S box, a riboswitch that can sense S-adenosylmethionine intracellular levels. Taking this information into account, it is possible to distinguish the direction of the enzymatic reaction towards methionine, which is in accordance with the reported experimental evidence. In other Firmicutes, this gene is not regulated by an S box but by a T box. Considering that we are able to predict the Watson–Crick interactions of the T box and the tRNA that will be sensed, we have been able to discriminate between tRNA<sup>Cys</sup> or tRNA<sup>Met</sup> and hence predict the direction of the enzymatic reaction that will be performed by the regulated gene [11].

the anaerobic *Clostridia*. These organisms have limited redox potential [15] and use electron transfer flavoproteins to increase it through the cyclic reduction of 5-crotonyl-CoA to 5-butyryl-CoA generating NADH. These *Clostridia* tend to have duplicated *etf* operons, with one copy being regulated by an Ile T box, probably activating the electron flow towards amino acid biosynthesis [11].

## Conclusions

The typical strategies for gene function assignment, compared in [17], are particularly useful for completely sequenced genomes and for genes with annotated homologues. Our gene function assignment approach, based on ri-



**FIG. 3.** The role of the *por* operon in isoleucine biosynthesis. The *por* operon is regulated by a T box responding to the intracellular levels of Ile. This suggests that this protein complex might participate in the first two steps of Ile biosynthesis using different substrates instead of catalysing the canonical reaction that links glycolysis to the Krebs cycle of pyruvate decarboxylation to form acetyl-CoA and CO<sub>2</sub>.

boswitches, is an important alternative to classical methods and is particularly well suited when presented with metagenomic sequences or partially sequenced genomes. It also provides more physiological information since it is based on the relationship of gene expression, metabolic requirements and gene function. With our method, the role of particular genes in metabolic pathways can be predicted even when their homologues have no function, or when they simply have no discernable homologues. Hence, riboswitch prediction contributes to mapping genes to the metabolic context of the cell or community of cells, giving us a bypass from raw sequences to a more comprehensive view of gene function. In this study, we have used the T box as a representative example of function assignment based on riboswitch identification. Nevertheless, the great specificity of all riboswitches for their corresponding target metabolites (nucleotides, vitamins, amino acids, co-factors, etc.) can be exploited for gene function annotation.

## Acknowledgements

We would like to thank Ricardo Ciria for the maintenance of our group database, and Cei Abreu-Goodger and Marel Chengé for insightful discussions on this paper. This work was supported by CONACyT grant 154817 and PAPPIT grant IN203211 to E.M.



## Transparency Declaration

---

We declare no conflicts of interest.

## References

---

1. Xie G, Keyhani NO, Bonner CA, Jensen RA. Ancient origin of the tryptophan operon and the dynamics of evolutionary change. *Microbiol Mole Biol Rev* 2003; 67: 303–342.
2. Michel G, Roszak AW, Sauve V et al. (2003) Structures of shikimate dehydrogenase AroE and its Paralog YdiB. A common structural framework for different activities. *J Biol Chem* 2003; 278: 19463–19472.
3. Brown DP, Krishnamurthy N, Sjolander K. Automated protein subfamily identification and classification. *PLoS Comput Biol* 2007; 3: e160.
4. Batey RT, Gilbert SD, Montange RK. Structure of a natural guanine-responsive riboswitch complexed with the metabolite hypoxanthine. *Nature* 2004; 432: 411–415.
5. Yousef MR, Grundy FJ, Henkin TM. Structural transitions induced by the interaction between tRNA(Gly) and the *Bacillus subtilis* glyQS T box leader RNA. *J Mol Biol* 2005; 349: 273–287.
6. Grundy FJ, Henkin TM. The T box and S box transcription termination control systems. *Front Biosci* 2003; 8: D20–D31.
7. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res* 2010; 38: e130.
8. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 1994; 2: 28–36.
9. Bailey TL, Gribskov M. Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 1998; 14: 48–54.
10. Eddy SR. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an RNA secondary structure. *BMC Bioinformatics* 2002; 3: 18.
11. Gutierrez-Preciado A, Henkin TM, Grundy FJ, Yanofsky C, Merino E. Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiol Mol Biol Rev* 2009; 73: 36–61.
12. Tatusov RL, Koonin EV, Lipman DJ. A genomic perspective on protein families. *Science* 1997; 278: 631–637.
13. Ogata H, Goto S, Sato K, Fujibuchi W, Bono H, Kanehisa M. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res* 1999; 27: 29–34.
14. Rollins SM. *The mRNA/tRNA interaction promoting T box transcriptional antitermination*. PhD Thesis. Columbus, OH: Ohio State University, 2002.
15. Seedorf H, Fricke WF, Veith B et al. The genome of *Clostridium kluyveri*, a strict anaerobe with unique metabolic features. *Proc Natl Acad Sci U S A* 2008; 105: 2128–2133.
16. Henkin TM, Glass BL, Grundy FJ. Analysis of the *Bacillus subtilis* tyrS gene: conservation of a regulatory sequence in multiple tRNA synthetase genes. *J Bacteriol* 1992; 174: 1299–1306.
17. Korber JO, Jensen LJ, von Mering C, Bork P. Analysis of Genomic Context: Prediction of Functional associations from conserved bidirectionally transcribed gene pairs. *Nat. Biotechnol.* 2004; 22(7): 911–917.