UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE BIOTECNOLOGÍA

"La evolución de la red de regulación transcripcional en bacterias es extremadamente flexible"

T E S I S

Que para obtener el grado de

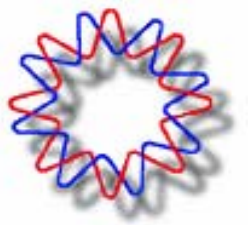
M A E S T R O E N C I E N C I A S

Presenta

BIÓL. IRMA LOZADA CHÁVEZ

Director de tesis:

DR. PEDRO JULIO COLLADO VIDES



2 0 0 6

Este proyecto se desarrolló en el laboratorio de Genómica Computacional

del Centro de Ciencias Genómicas de la UNAM

Asesorado institucionalmente por:

Dr. Pedro Julio Collado Vides	**CCG**
Dr. Ernesto Pérez Rueda	**IBT**
Dr. Alejandro Garcíarrubio Granados	**IBT**

Evaluado para obtención de grado por el comité de sinodales:

Dr. Enrique Merino Pérez	**Presidente**
Dr. Pedro Julio Collado Vides	**Secretario**
Dr. Arturo Carlos II Becerra Bracho	**Vocal**
Dr. Ernesto Pérez Rueda	**Suplente**
Dr. Guillermo Gosset Lagarda	**Suplente**

Asesorado académicamente de forma externa por:

Candidato a doctor Juan Javier Díaz Mejía	**A lo largo de todo el proyecto**
Dr. Bruno Contreras Moreira	**En la revisión crítica del artículo**
M.C. Sarath Chandra Janga	**En la estadística para redes**

Con fondos del programa de becas CONACyT, expediente número 185993.
Así como con un apoyo post-beca CONACyT derivado de los recursos del laboratorio de Genómica Computacional del CCG de la UNAM,
de donde el Dr. Pedro Julio Collado Vides es Jefe de Grupo.

Esta tesis está dedicada a

Alejandro Nabor Lozada Chávez,
Ana María Chávez Morales,
Nabor Lozada Guerrero,

y

a mi muy amada *Universidad Nacional Autónoma de México*
(a quién debo mi felicidad académica)

Agradecimientos:

Para quienes tengan antecedentes de la tesis de licenciatura, verán que mi agradecimiento hacia ustedes, no ha sido menor o diferente, sólo he incrementado mi deuda y mi fortuna por su amistad y apoyo.

Quiero expresar primeramente mi total agradecimiento a mi hermano **Alejandro** porque ha sido mi familia y mi mejor amigo desde que han sido necesarios, porque me ha impulsado siempre a esperar más de mí y porque la fortaleza y lo bueno que pueda tener yo como persona se lo debo principalmente a él.

Deseo agradecer especialmente a **Julio Collado** por haberme permitido trabajar dentro de su excelente grupo de trabajo, por haberme apoyado económicamente en el desarrollo, culminación y difusión (congresos) de esta tesis; pero sobretodo por su amistad y porque me ha enseñado, de muchas formas, que nunca es demasiado en cuanto a aprender y a concebir proyectos se trata.

Así mismo, quiero agradecer a **Arturo Becerra** y a **Luis Delaye** por la relación académica que tenemos, porque siempre han sabido guiarme sobre "el buen camino" en los tantos problemas que acogen al estudio de la evolución temprana de la vida. Les agradezco muy especialmente por la amistad que tenemos y porque siempre han creído en mí.

Agradezco a **Sandra Ramírez** por mantener la confianza, la voluntad, las ilusiones y la disponibilidad para desarrollar conmigo amistad, trabajos de investigación y de divulgación científicos. Porque Sandra representa el inicio de una nueva generación de investigadores en México que, en el área del origen de la vida, están dispuestos a abordar las preguntas necesarias, por muy difíciles y ambiciosas que parezcan, para generar propuestas en un ambiente científico abierto e interdisciplinario. Esto representa gran admiración y motivación para mis próximos objetivos académicos.

No tengo palabras para agradecer a **Javier Díaz** todo el apoyo académico que me ha brindado desde que me incursioné en la ciencia. Le agradezco que comparta conmigo su experiencia en la dureza, formalidad, pasión, intuición y regocijo que trae la ciencia cada día a las personas que como él y como yo tratamos de ejercerla. Mi calidad académica, en el más amplio sentido de la palabra, se debe a su asesoría y a su ejemplo. Muy especialmente deseo agradecer su amistad, porque de todas, la nuestra ha sido la más probada, la más compleja y la más incierta, pero definitivamente también ha sido una de las más importantes en mi vida.

Quiero agradecer a **Bruno Contreras** por su amistad, su apoyo y sumo ingenio oculto discretamente por su sencillez, porque su humildad me ha dado otro ejemplo de la felicidad obtenida por la sabiduría y la naturaleza humana.

Por otra parte, le agradezco a **Ernesto Pérez Rueda** por su asesoría académica, siempre muy acertada, y muy enfáticamente por su amistad.

Les agradezco a mis compañeros y amigos del laboratorio: (¡las chicas!) **Concepción Hernánadez**, Socorro Gama, **Verónica Jiménez**, Heladia Salgado, **Mónica Peñaloza**, **Fabiola Sánchez**, Lucia López, Irma Martínez; (¡los chicos!) Cei Abreu, Bruno Contreras, Martín Peralta, Carlos Rodríguez, Juan Segura, Luis Treviño, César Bonavides, Edgar Díaz, Sarath Chandra, Luis Muñíz, Alberto Santos, Julio Freyre, **Víctor del Moral**, **Romualdo Zayas**, por su apoyo académico y personal durante mi estancia en el laboratorio. Por compartir juntos, con igual diversión, los días de trabajo y los de festejo, logrando que me sintiera como en una gran familia. Extrañaré sobremanera este ambiente de trabajo. Particularmente quiero agradecer a **Concepción Hernández, Mónica Peñaloza** y a **Heladia Salgado** todo su apoyo incondicional en los momentos críticos, ¡que fueron muchos!

Agradezco a mi ya inseparable parte artística y humanista: **Tania Cabagne, Lucelina Nunes** y **Alberto Cabañas,** por ser mis amigos bajo historias de vida muy parecidas, porque me ayudan a explotar y mantener al artista que llevo dentro, lo cual ha hecho sentirme humanamente completa. Pero sobretodo, por compartir conmigo invaluables experiencias, secretos y una percepción de la vida.

Quiero agradecer a mis amigos: **Cynthia, Mónica Araujo, Alma Mendoza, Lilia Montoya, Lidia Leal, Maribel Hernández, Woyteck, Sarath Chandra, Ricardo Menchaca, Cristhian Ávila, Mónica Peñaloza, Cristhian Torres, Leticia Ortega, Guillermo Quevedo, Alejandro Carbajal,** con quienes comparto ciencia, juventud y muy gratos momentos que llenan mi vida de forma muy peculiar.

Le agradezco a **Mathieu Deschênes** todo su amor, los momentos invaluables que hemos compartido juntos, y por haber generado un espacio de felicidad en base a la honestidad, al entendimiento y al apoyo mútuos que permitieron la adaptación de nuestros diferentes estilos de vida y, con ello, hacer de mí una mejor persona.

Deseo agradecer a **Concepción**, **Daniel**, **Abraham**, **Moisés** y **Esaú,** a quienes amo desde que entraron en mi vida, y quienes más se han percatado del tiempo que he invertido en mi desarrollo profesional.

Muy especialmente, agradezco a quienes considero como mis familias y mis amigos, la **familia Araujo** y la **familia Díaz** que siempre han tenido un espacio especial entre su gente para mí.

Finalmente quiero agradecer a mi padre, **Nabor Lozada**, y a mi madre, **Ana María Chávez**, por mostrarme lo difícil que es la vida, así como la fortaleza que una persona puede tener cuando parece que todo está perdido. Les agradezco porque sin importar circunstancias o motivos me concibieron y cuidaron de mí bajo una familia los años que recuerdo como los más completos y felices de mi vida, mi niñez. Pero sobretodo, quiero agradecerles por buscar conmigo, siempre que es posible y necesario, los puentes que nos permitan recuperar todo el tiempo perdido.

Mi mayor agradecimiento para la **Universidad Nacional Autónoma de México** porque me ha brindado las oportunidades necesarias para generar una expectativa de la vida, a través del conocimiento, lo cual contribuye en gran medida a mi felicidad cotidiana.

“La evolución de la red de regulación transcripcional en bacterias es extremadamente flexible”

Por Irma Lozada-Chávez

ÍNDICE

Página web del proyecto: http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/

RESUMEN

A través de millones de años, la estructura y complejidad de la Red Regulatoria Transcripcional (TRN – Transcriptional Regulatory Network) en las bacterias ha sido cambiada y reorganizada para permitir a estos organismos adaptarse a casi cualquier nicho ecológico sobre la Tierra. Con el fin de entender la flexibilidad evolutiva de las TRNs en bacterias, estudiamos la conservación de las TRNs de los dos organismos modelo *Escherichia coli K12* y *Bacillus subtilis* por medio de herramientas bioinformáticas para detectar sus contrapartes en los genomas completos de Bacteria, Archaea y Eukarya a tres diferentes niveles: componentes de la TRN, pares de interacciones y regulones. Encontramos que los factores de transcripción (TF) se pierden mucho más rápido que los genes regulados (TG) a través de los *phyla*[1]. Mostramos que los reguladores globales (GR) están pobremente conservados a través del espectro filogenético. Más aún, de los TF's detectados, notamos que son menores las interacciones equivalentes que éstos guardan en las especies distintas a los modelos. De ahí, que los TFs podrían ser los principales elementos responsables para la plasticidad y *evolucionabilidad*[2] de las TRNs. También encontramos que hay sólo una pequeña fracción de interacciones regulatorias transcripcionales conservadas significativamente entre los diferentes *phyla* bacterianos y, que existe una restricción sobre los elementos de la interacción para co-ocurrir. Finalmente, nuestros resultados sugieren que la mayoría de los regulones en bacterias han cambiado rápidamente con la distancia filogenética, lo cual implica un nivel más alto de jerarquía en la flexibilidad de las TRNs. Desde nuestro análisis comparativo, generamos la hipótesis de que la regulación transcripcional de ciertos procesos celulares esenciales, tales como la síntesis de arginina, biotina y ribosa, transporte de aminoácidos y hierro, disponibilidad de fosfato, procesos de replicación y la respuesta a SOS han sido bien conservados en la evolución. A partir de este trabajo, es posible inferir que la regulación transcripcional es más flexible que el componente genético de los organismos y que su estructura juega un papel importante en la adaptación fenotípica.

[1] ***Phylum*** (1). (Plural: ***phyla***) Es un taxón usado en la clasificación de la vida como primera división de un reino, de mayor rango que una clase. Un *phyla* representa grandes grupos de organismos generalmente aceptados con ciertos rasgos evolutivos.

[2] ***Evolucionabilidad*** (2). (***Evolvability***) Es una capacidad del genotipo para generar variación fenotípica heredable. Esta capacidad tiene dos componentes: **i)** reducir la letalidad potencial de mutaciones y **ii)** reducir el número de mutaciones necesitadas para producir rasgos novedosos fenotípicamente. La *evolvability* puede haber sido seleccionada generalmente en el transcurso de la selección para robustecer procesos flexibles convenientes para el desarrollo complejo y fisiología, y específicamente seleccionada en linajes sometidos a radiaciones repetidas.

INTRODUCCIÓN

La **evolución** de los organismos está dada por la variación y la selección de sus componentes, procesos y estructura a través del tiempo. Siendo un componente y proceso básico del organismo, la **regulación transcripcional** juega un papel prominente en la expresión de la información genética. Su función primaria es permitir que la célula responda, controle y se adapte en respuesta a cambios intracelulares y extracelulares, tales como el estado nutricional, la división celular, la esporulación (cuando aplica), así como varios estreses. Una importante idea que surge con los trabajos de Michael A. Savageau y Stuart A. Kauffman desde los años 70's es que la regulación transcripcional puede ser vista como una red compleja de interacciones entre diversos tipos de moléculas, tales como proteínas, RNA, DNA y metabolitos (4-12) [ver Figura I1]. A partir de estas redes, podemos entender niveles jerárquicos más altos en la organización de la información y procesos biológicos, tal como en este caso de la regulación transcripcional. En este trabajo, tratamos de valorar la evolución de la estructura y flexibilidad de la **Red de Regulación Transcripcional** (**TRN – Transcriptional Regulatory Network**), por medio de un análisis comparativo a través de los genomas completos de los organismos desde tres distintos niveles: componentes individuales de la TRN, pares de interacciones regulatorias y regulones.

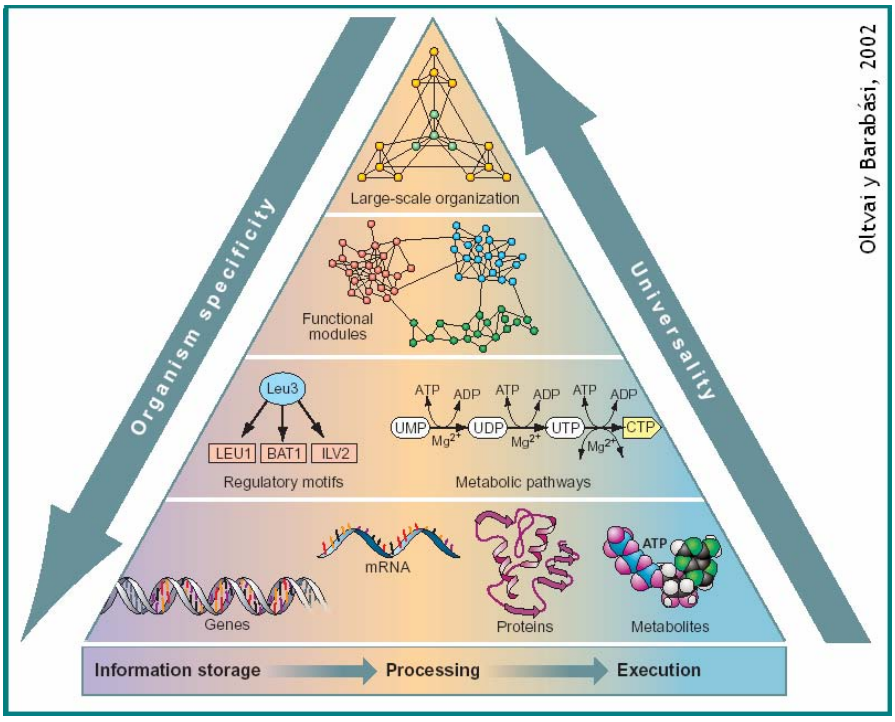


Figura I1. Representación de la organización de la información y procesos celulares en la naturaleza (13). En la base de la pirámide se muestra la dirección de la expresión genética: el genoma, transcriptoma, proteoma y metaboloma (**nivel 1**), que a su vez proporciona la especificidad de los organismos. Pueden conseguirse revelaciones sobre la lógica de la organización celular cuando observamos a la célula como una red compleja, en la cual los componentes son conectados a través de conexiones funcionales. En el nivel más bajo de estas redes, los componentes forman motivos genético-regulatorios o rutas metabólicas (**nivel 2**), los cuales llegan a ser parte de los bloques constitutivos de los módulos funcionales (**nivel 3**). Estos módulos son agrupados, generando una arquitectura jerárquica "libre de escala" (definida posteriormente en este trabajo) (**nivel 4**). Este último nivel de organización nos sugiere que se pueden encontrar principios de organización universales y aplicar a la naturaleza.

I. Generalidades sobre la regulación transcripcional en bacterias.

La **transcripción** es el proceso a través del cual una cadena de DNA (**DNA - DeoxyriboNucleic Acid**) es enzimáticamente copiada por una **RNA polimerasa DNA-dependiente** (**RNApol**) para producir una cadena de RNA (**RNA - RiboNucleic Acid**) complementario, a partir del cual se lleva a cabo la expresión de los genes[3] y la síntesis de proteínas (14) [ver Figura I2B]. La regulación de la transcripción en las bacterias sucede principalmente antes y durante el inicio de este proceso: modificando la holoenzima RNA polimerasa, haciendo re-arreglos estructurales de la región reguladora o, por medio de pequeñas moléculas de RNA que también activan o inhiben la transcripción al unirse a sitios específicos. De forma general, la regulación transcripcional es llevada a cabo por la RNA polimerasa, ayudada por elementos intrínsecos (terminadores, promotores, operadores) e extrínsecos al DNA (factores σ, factores transcripcionales) que dirigen y coordinan su actividad durante las primeras etapas del proceso (15) [ver Figura I2-A y B].

El **promotor** es uno de los elementos intrínsecos del DNA más importante para la transcripción, pues indica dónde iniciar y, a través del factor σ, funciona como punto de anclaje para la RNApol al DNA. El promotor está conformado por dos hexanucleótidos (uno localizado en la posición -10 "caja-10" y el otro en la posición -35 "caja-35") separados por una región de entre 15 y 21 pares de bases (bp – base pairs) (14) [ver Figura I2-A y B]. La secuencia de los hexanucleótidos es distinta para cada factor σ. En algunos promotores, existe una estructura rica en AT (el **elemento UP**) que incrementa la actividad del promotor, se localiza a 20 bases río arriba de la región -35 y es reconocida por el dominio C-terminal de la subunidad α de la RNApol [ver Figura I2B]. Por otro lado, los **operadores**, o sitios de unión al DNA (**bsDNA – DNA binding site**), son secuencias específicas para los factores transcripcionales de aproximadamente 20 bps, localizados dentro de la región reguladora. Los operadores se pueden ubicar río arriba (*upstream*), traslapados con el promotor, o río abajo (*downstream*) del inicio del gen (15) [ver Figura I2B].

Los **factores sigma (σ)** corresponden a una de las subunidades de la holoenzima RNA polimerasa y tienen tres funciones principales: ***a)*** garantizar el reconocimiento de las secuencias promotoras específicas; ***b)*** posicionar la holoenzima sobre un promotor blanco; y ***c)*** facilitar el desenrrollamiento del dúplex de DNA que está cercano al sitio de inicio de la transcripción. Muchas bacterias contienen múltiples factores σ que son capaces del reconocimiento de diferentes promotores en distintas condiciones ambientales (15).

Los **factores transcripcionales** (**TF – Transcription Factor**) son proteínas (monómero o multímero) que se unen a bsDNAs específicos (operadores) en ambas hebras del DNA sin abrirlo, formando una interacción por puentes de H, cuya función es activar o reprimir la transcripción de un **gen(es) blanco** (**TG – Target Gene**) [ver Figura I2B]. Los TFs llevan a cabo su actividad por la presencia o ausencia de sus señales, llamados **efectores** o **ligandos**, los cuales corresponden a distintos tipos de metabolitos o grupos químicos de óxido-reducción, que cambian la conformación o el estado químico de los TFs (16).

[3] **Gen** (14). Región de DNA que controla una característica hereditaria discreta, usualmente corresponde a un simple polipéptido o RNA estable. Esta definición incluye la unidad funcional entera, abarcando secuencias de DNA codificantes y secuencias regulatorias de DNA no codificantes.

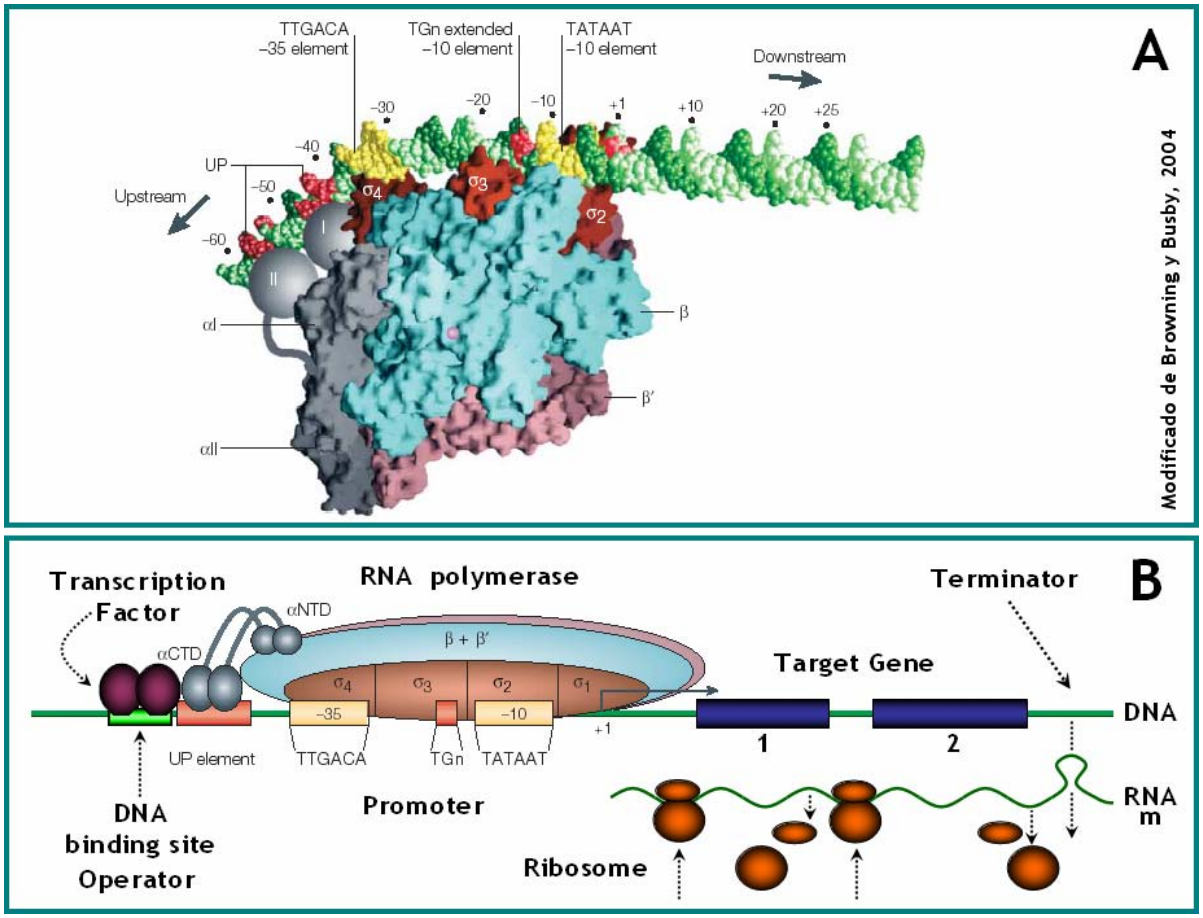


Figura I2. RNA polimerasa (RNApol) y los componentes de la transcripción. A) Un modelo basado sobre los estudios cristalográficos del anclaje inicial de la holoenzima RNApol a un promotor (14). **B)** Esquema del modelo mostrado en la parte A) (14-15), ilustrando las diferentes interacciones entre los elementos de la transcripción. Para ambos esquemas, son mostrados en amarillo las **secuencias consenso** del **promotor** para los elementos **-35 (TTGACA)** y **-10 (TATAAT)**, así como en rojo la secuencias extendido **-10 (TGn)** y el **elemento UP**. La **RNApol** está conformada por un núcleo enzimático de cuatro subunidades proteínicas (una **β**, una **β'** y dos **α**) y una subunidad de unión temporal (el factor σ). β y β' juntas forman el centro catalítico (sitio activo Mg^{2+} en magenta), β' (en rosa) se une a la hebra templado, β (en azul) se encarga de la unión de los nucleótidos, las dos α (en gris: αNTD –dominio amino y αCTD –dominio carboxilo) se requieren para interaccionar con los **factores transcripcionales** (**TFs** en púrpura) y reconocer al promotor; además, por medio del **factor σ** (sus diferentes dominios en rojo) ocurre el reconocimiento y unión al promotor.

Cuando RNApol (α2ββ') se unen al factor σ, se forma el complejo **holoenzima (α2ββ'-σ)**, que sólo existe durante las dos primeras etapas de la transcripción. La RNApol tiene afinidad general por el **DNA** (mostrado en verde), como holoenzima su afinidad hacia el promotor incrementa 10^7 veces. Para iniciar la transcripción, la holoenzima pasa por varios estados de unión al DNA: 1) **Complejo cerrado binario (RPc),** cuando la holoenzima está unida al promotor; 2) **Complejo abierto (RPa),** cuando se separan las dos cadenas de DNA, pero aún no se inicia la síntesis del **RNA mensajero** (**mRNA** en verde ondulado); 3) **Complejo abierto de inicio (RPin)**, cuando la holoenzima está como RPa y tiene unidos cuatro nucleótidos trifosfatados. Durante el RPin, la holoenzima se mantiene sintetizando mRNA, pero son fragmentos muy cortos que se abortan y vuelve a empezar hasta que puede continuar la síntesis con más nucleótidos, liberándose así del promotor y del factor σ, como RNApol.

Los elementos intrínsecos en el DNA, promotor y **operadores** (**bsDNA** en verde claro), se ubican río arriba (*upstream*) antes del inicio del gen (posición +1), en la **región "reguladora"** (cuyas posiciones se escriben con signo negativo). Ésta región coordina, con el promotor, donde se debe unir la RNApol y, con los operadores, si se debe incrementar o disminuir la transcripción de los **genes** (**TGs** en azul) río debajo (*downstream*) de su inicio. Los elementos extrínsecos al DNA, el factor σ y los TFs, actúan sobre la región reguladora; de acuerdo al efecto que tengan sobre la transcripción se clasifican como: a) **positivos o activadores** si favorecen la unión de la RNApol con su promotor o si eluden la barrera cinética para promover la estabilidad de RPc a RPa; b) **negativos o represores** si limitan el acceso de la RNApol al promotor.

II. Componentes y estructura de la TRN.

En 1961, François Jacob y Jacques Monod fueron los primeros en proponer la organización del DNA para su regulación conjunta. Propusieron que la transcripción de los genes es regulada por las interacciones específicas entre secuencias *cis* localizadas en la región reguladora y los productos de genes codificados en regiones *trans*. Las secuencias *cis* son los operadores y promotores, mientras que los productos de las *trans* son generalmente los TFs (14, 16). Así como también, desde los años 70's, los trabajos en diversas redes celulares (circuitos metabólicos, de regulación, inmunológicos, embriológicos, etc.) de Michael A. Savageau y Stuart A. Kauffman formalizaron las bases teóricas necesarias que ahora nos permiten entender diversas características de la estructura y organización de la regulación transcripcional (4-7).

De esta forma, se ha propuesto que la unidad básica de la **interacción de regulación genética** consiste de tres componentes: un **factor de transcripción** (**TF**), su **sitio de unión a DNA** (operador) (**bsDNA**) y el **gen regulado** o blanco (**TG**) (17) [ver Figura I2B]. La colección total de tales interacciones regulatorias en un organismo puede ser conceptualizada como una red, referida como la **Red de Regulación Transcripcional** (**TRN – Transcriptional Regulatory Network**). De esta forma, la TRN puede ser representada como un grafo[4] dirigido, en el cual los nodos representan los genes (TFs o TGs) y las aristas representan interacciones regulatorias dirigidas (18) [ver Figura I3].

Estructuralmente, la TRN es compleja porque los genes pueden estar regulados por más de un TF (ya que algunos genes se utilizan en más de una situación del ciclo de vida celular o en diferentes etapas de alguna condición ambiental); así como también, muchos TFs pueden controlar a más de un gen a través de uno o varios bsDNAs (17, 19-20) [ver Figura I3].

Derivado de ello, podemos delimitar varios niveles de organización en las TRNs. Al arreglo de la región reguladora y el(los) gen(es) correspondientes se le llama **unidad transcripcional** (**TU – Transcription Unit**) (14, 16). En las bacterias, los genes involucrados en una misma ruta metabólica o para un mismo proceso celular generalmente están acomodados de forma consecutiva, con una sola región reguladora formando una TU llamada **operón**, que se transcribe en un solo RNA mensajero (**mRNA – messenger RNA**) (14-15) [ver Figura I3]. A su vez, los operones se organizan como **regulones**[5] en un nivel mayor. Los regulones no son propiamente arreglos estructurales en el DNA, más bien, son operones (en distintas regiones del DNA) que se regulan conjuntamente por un solo TF en respuesta a una señal generalizada (14-16) [ver Figura I3].

[4] **Grafo** (18). Es un par de conjuntos $G = \{P, E\}$, donde P es un conjunto de N nodos (o vértices o puntos) $P_1, P_2, \ldots, P_N$ y E es un conjunto de aristas (o uniones o líneas) que conectan dos elementos de P.

[5] **Regulón** (14-16). La definición inicial delimita a un grupo de genes sujetos a regulación de uno y sólo un regulador (Maas W.K., 1964). Actualmente, a esta definición se ha denominado regulón simple, cuya diferencia con un regulón complejo radica en el número de factores de transcripción que regulan a un grupo de genes, y en un regulón complejo estricto es adicional mantener el efecto de cada regulador (activador o represor) sobre todos los genes regulados. **Nota:** en este estudio se usó la definición de regulón simple.

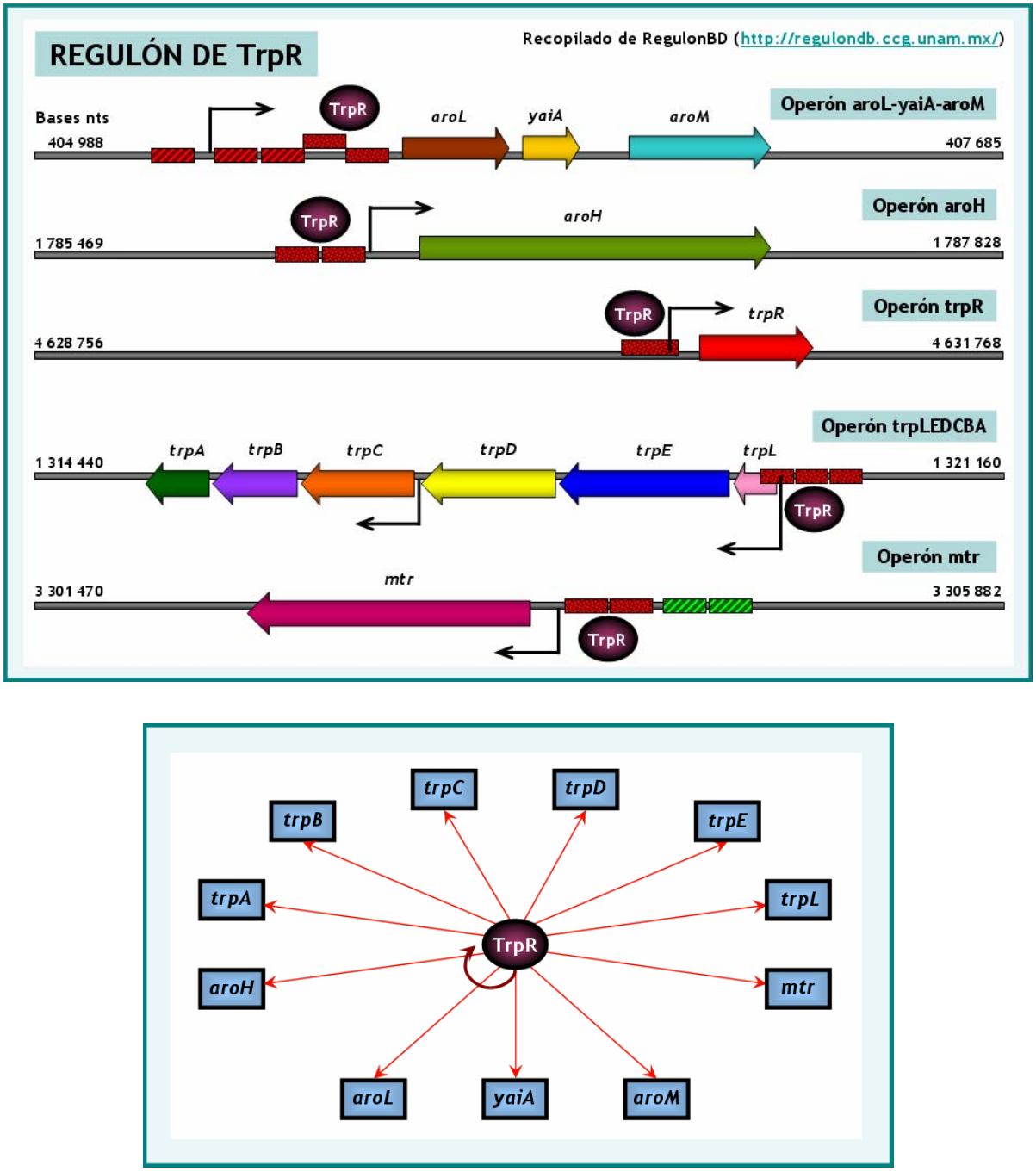


Figura I3. Unidad básica de la interacción regulatoria genética. Sobre el recuadro superior se esquematizan las interacciones de la proteína TrpR (que regula la transcripción del triptofano) en sus tres componentes: **1)** un **TF** (TrpR como un óvalo de color púrpura), **2)** su **sitio de unión** (rectángulos de color rojo – por represión - punteados) **al DNA** (las líneas grises) y, 3) el **TG(s)** (rectángulos flechados con dirección *forward* o *reverse*, según la cadena de DNA donde se localicen). El **regulón** de **TrpR** está formado por **5 operones**, cada uno de los cuales, está conformado con al menos un sitio de unión al DNA para este TF y un TG al que regule. La flecha delgada negra indica el sitio de inicio de la transcripción (+1). Los rectángulos de color rojo (por represión) y verdes (por activación) rayados diagonalmente representan los sitios de unión al DNA para el **TF TyrR** (que regula la transcripción de tirosina) que también regula a algunos de estos operones. En el recuadro inferior se representa el grafo de las interacciones del regulón de TrpR (12 interacciones en total, una de las cuales es autorregulación), donde los nodos representan el TF (óvalo de color púrpura) y los TGs que regula (cuadros de color azul), unidos por medio de aristas dirigidas (flechas) que representan la dirección y el tipo de regulación (rojo: represora) de la interacción regulatoria.

Los distintos niveles de complejidad en las TRNs han sido evaluados con ciertas características estadísticas, las cuales han revelado que la estructura jerárquica de las TRNs se aproxima a una topología *libre de escala*[6], donde la conectividad[7] de los nodos sigue una distribución del tipo *ley de potencias*[8] (7). Esto implica, para los nodos con conectividad de salida[7] (TFs), que pocos TFs tienen un gran número de interacciones (conexiones con TGs), de esta manera, dominan el comportamiento general de la red; mientras que la mayoría de los TFs presentan una conectividad baja con TGs. Así, el grupo de TFs con una alta conectividad (de salida) podrían tomar el papel de *hubs*[9] de regulación en la red. Estos *hubs* pueden ser caracterizados en las TRNs como **reguladores globales** (**GR - Global Regulators**). Aunque una categoría de GR no sólo se atribuye a la conectividad que pueda tener un TF, sino también por el número y tipo de co-reguladores que presenta, los diferentes tipos de factores σ con los que interactúa, el número de otros TFs que regula, así como la variedad de condiciones experimentales donde ejerce su control (19). Por estas razones, algunos grupos de trabajo han diferido en el número y tipo de GRs propuestos para las TRNs que se han caracterizado previamente, aún cuando la conectividad de un TF pueda ser una medida acertada sobre su condición global o no en la regulación celular.

Sin embargo, es importante reconocer que a pesar de que están disponibles abundantes datos de secuencia y genomas completos, la determinación y recopilación experimental de las TRNs ha sido limitada a pocos organismos. Aún cuando trabajos previos de genómica estructural[10] han generado predicciones de interacciones del tipo proteína-DNA (24-26), no se ha determinado una relación clara entre la presencia de un TF, su TG y el bsDNA(s), que permita la transferencia de estas interacciones entre los genomas. Entonces, resulta difícil encontrar una medida específica entre la homología a nivel de secuencia, la función y la transferencia de la interacción[11] para cualquiera dos proteínas involucradas en una interacción de regulación transcripcional (27-29). Sin embargo, varios grupos han examinado recientemente la transferencia de las interacciones regulatorias de un organismo a otro usando enfoques de genómica comparativa (27, 29). La transferencia de tales interacciones involucra la asignación de funciones a TFs y TGs, basados sobre la similitud de la secuencia de las proteínas y sobre la conservación de patrones topológicos de la TRN, tales como motivos y módulos de interacciones (28, 30).

---

[6] **Topología libre de escala** (22)**.** (***Scale free***) Es un término introducido por Barabási y Albert en 1999 que describe, a nivel genérico, cualquier grafo o red en la cual las conectividades de los vértices siguen una distribución que se aproxima a una ley de potencias (*power law*).

[7] **Conectividad** (*K*) (18). Las redes son caracterizadas por dos distribuciones de grado: la distribución de conectividad de salida, $P_{out}(k)$, significa la probabilidad que un nodo tiene *k* aristas de salida; y la distribución de conectividad de entrada, $P_{in}(k)$, es la probabilidad que *k* aristas apunten a un cierto nodo. Varios estudios (Albert y Barabási, 2002) han establecido que tanto $P_{out}(k)$ y $P_{in}(k)$ tienen distribuciones *power law*.

[8] **Distribución de tipo ley de potencias** (18). (***Power-law***) Matemáticamente, una secuencia finita $y = (y_1, y_2, \ldots, y_n)$ de números reales, asumidos sin pérdida de generalidad que son ordenados tal que $y_1 \geq y_2 \geq \ldots \geq y_n$, se dice que sigue una ley de potencias o relación no estocástica de escala si $k = cy_k^{-\alpha}$, donde *k* es (por definición) el intervalo de y_k, *c* es una constante fija y α es llamado el índice de escala. De esta forma, el que una red tenga una distribución *power law*, significa que la frecuencia de nodos (*N*) como una función de su conectividad *k* decrece como $k^{-\alpha}$, indicando una alta variabilidad en el número de conexiones, con un número pequeño de nodos altamente conectados y con un alto número de nodos poco conectados.

[9] ***Hubs*** (22). Definido normalmente como un nodo con un alto número de conexiones en una red determinada.

[10] **Genómica estructural** (23). Es la rama de la genómica orientada a la caracterización y localización de las macromoléculas derivadas de los genomas, en base a la cristalografía de rayos X, el análisis de espectroscopia de resonancia magnética nuclear –NMR, entre otros. La información originada nos permite entender las características de las macromoléculas y los mecanismos biológicos de su función a nivel atómico y molecular.

[11] **Transferencia de la interacción** (16). Anotación de una interacción, del tipo proteína-proteína o proteína-DNA, desde un organismo modelo a otro usando genómica comparativa.

III. El método Regulog para la transferencia de las interacciones regulatorias.

El enfoque ***Regulog***[12] usa datos conocidos experimentalmente para predecir interacciones proteína-DNA en otros genomas. Se predice que una interacción (TF y su TG) en una especie ocurre en otra especie si su mejor apareamiento de secuencia ha sido determinado en un grupo blanco de genomas. La presencia de uno de los componentes de la interacción regulatoria no es suficiente para transferir la anotación de la interacción, sino que es necesario que ambos, el TF y su TG, sean detectados en el organismo de interés [ver Figura I4]. Usando este enfoque, Yu *et al.* (27) han mostrado que los genes ortólogos[13] de TFs y TGs de *Saccharomyces cerevisiae* y de *Drosophila melanogaster* comparten la misma interacción regulatoria si los TFs de estos eucariontes tienen una identidad a nivel de secuencia de proteína mínima del 30 al 60%, dependiendo de la familia de proteínas.

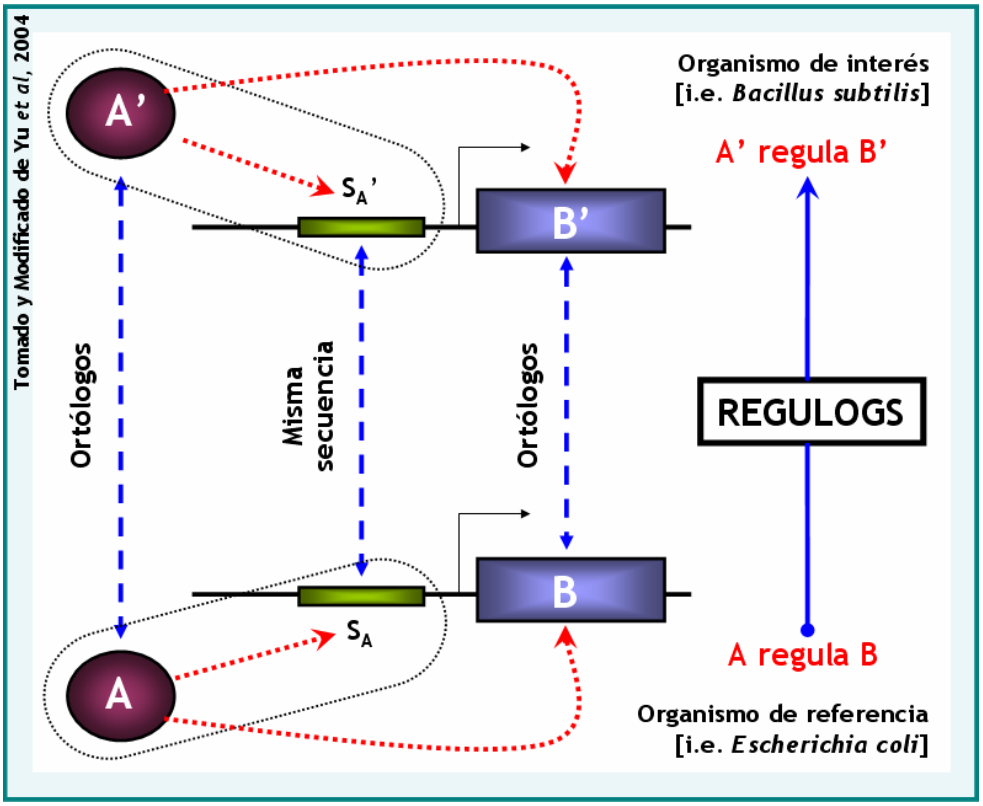


Figura I4. Representación esquemática del enfoque REGULOG. En el organismo de referencia, un TF A se une a su sitio de unión (S_A) y regula al gen B río abajo. Para llevar a cabo el mapeo de regulogs, **1)** TF A' en el organismo de interés (blanco) necesita ser el ortólogo de A. **2)** Las proteínas B y B' serían también ortólogas. **3)** La secuencia de DNA río arriba del gene B' necesita contener el mismo motivo (S_A') como S_A. La interacción entre TF A' y S_A' es el interolog proteína-DNA de A y S_A. La relación regulatoria entre A → B y A' → B' son regulogs. Sin embargo, la condición **3)** no es necesaria que se cumpla, ya que solamente TF A y A' necesitan compartir ≥ 30% de identidad.

[12] **Regulog** (27). (**Protein-DNA interolog**) Walhout *et al*. (2000) introdujeron el concepto de "interologs" que se definen como pares de ortólogos de proteínas interactuando en diferentes organismos. Yu *et al*. (2004) generalizaron este concepto para interacciones del tipo proteína-DNA, en donde se mostró que todos los TFs dentro de un cierto intervalo de identidades invariables (≥ de 30 y 60% de identidad dependiendo de la familia) comparten la misma secuencia de unión al DNA y, por lo tanto, se mostró que en general existe la misma relación regulatoria de la transcripción.

[13] **Ortólogos** (31). Son definidos como proteínas en diferentes especies que evolucionaron de un ancestro común por especiación y usualmente tienen la misma función. Ver material y métodos.

De esta forma, un ***regulog*** se define como el par de ortólogos que refleja una relación de regulación conservada entre proteínas interactuando a través de diferentes especies (16). En otras palabras, cuando una interacción proteína-DNA es transferida a través de las especies, la relación regulatoria entre el TF y su TG es también implícitamente transferida.

Yu *et al*. (2004) sugieren tres condiciones que son necesarias para transferir *regulogs*:

1. El TF A y su homólogo A' deben compartir al menos ≥ 30% de identidad de secuencia[14].

2. El TG B y su homólogo B' deben ser ortólogos.

3. La secuencia de DNA río arriba (*upstream*) de B' debe contener el mismo bsDNA que la de B.

Sin embargo, es necesario hacer notar que la tercera condición no es estrictamente necesaria para llevar a cabo la transferencia de *regulogs*. De esta forma, Yu *et al*. (2004) mostraron que el TF A y A' solamente necesitan compartir ≥ 30% de identidad. Entonces, el enfoque *Regulog* requiere mínimamente de la primera y segunda condición. Además, el uso de la información basada en los bsDNAs podría generar los siguientes problemas metodológicos:

a) La reducción drástica en el tamaño de las TRNs de referencia, ya que sólo pocos TFs en la red tienen caracterizados experimentalmente los bsDNAs para construir un perfil confiable de búsqueda de secuencias consenso en otros genomas.

b) La reducción del número de genomas que pueden ser comparados, ya que no es posible la detección confiable de los bsDNAs en organismos distantemente relacionados, lo cual generaría un sesgo sobre un análisis a gran escala.

c) Una menor cobertura en las predicciones, dado que los bsDNAs para un mismo TF están poco conservados, incluso en organismos cercanamente relacionados. A la fecha, no existe una tendencia clara entre la conservación de los TFs y la de sus bsDNAs. Dentro de los ejemplos más conocidos y contrastantes se encuentran los regulones de los TFs BirA (21) y LexA (13), donde el elemento regulador (TF) y regulado (TG) se han conservado en bacterias; así como su interacción de regulación transcripcional, pero no necesariamente por medio de los mismos bsDNA [ver Tabla I1].

[14] Formalmente, A y A' deberían ser ortólogos. Sin embargo, operacionalmente esta condición fué definida en el trabajo de Yu *et al.* (2004) por el criterio de porcentaje de identidad en las secuencias, a partir del cual se incluyeron secuencias parálogas en el análisis y con ello, una mayor cobertura en las predicciones. Sin embargo, no puede inferirse un mayor intervalo de confianza sobre las funciones de los TFs predichos. Ya que el hecho de que existan familias de TFs que presentan diferentes relaciones con el porcentaje de identidad en las secuencias refleja su diversidad regulatoria. De esta forma, familias con altos porcentajes de identidad contienen TFs que regulan muchos procesos diferentes; mientras que aquellas con bajos porcentajes de identidad contienen TFs que regulan pocos procesos diferentes (Luscombe y Thornton, 2002). En este sentido, los TFs parálogos incluidos en las predicciones de Yu *et al.* (2004) no garantizan la conservación de sus roles funcionales con respecto al TF del organismo de referencia.

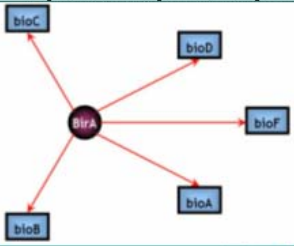
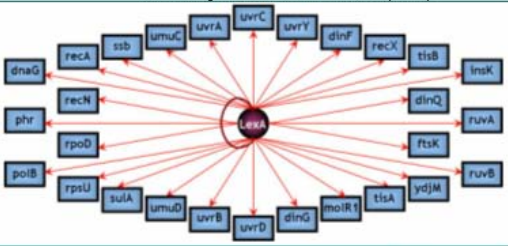
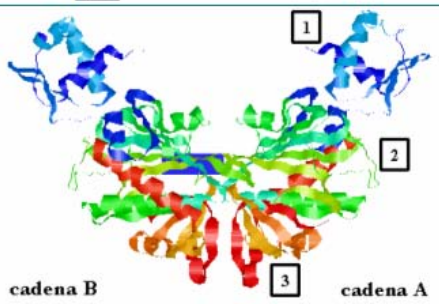
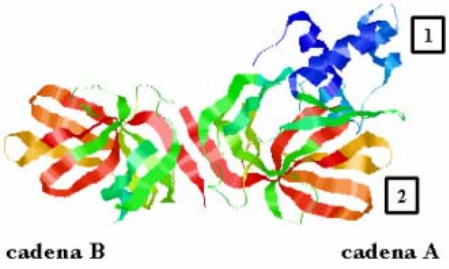
BirA [16131807 | b3973]	**Características**	**LexA** [16131869 | b4043]
EC: 6.3.4.15 Ligasa para la proteína biotina. Represor transcripcional del operón de biotina	**FUNCIÓN**	**EC: 3.4.21.88** Represor transcripcional de genes involucrados en la respuesta de daño a DNA (SOS)
	**REGULÓN**	
 • Dímero (**PDB**: **2EWN**, PFAM v10, 3D-Rasmol v2.6). 1. **PF08279**: N – terminal bsDNA (Helix-turn-helix). 2. **PF03099**: Central catalítico, contiene el sitio de unión para el biotinyl-5'-AMP (también requerido para regulación). 3. **PF02237**: C – terminal con función desconocida.	**DOMINIOS Y bsDNA(s)**	 • Dímero (**PDB**: **1JHH**, PFAM v10, 3D-Rasmol v2.6). 1. **PF01726**: N – terminal bsDNA (Helix-turn-helix). 2. **PF00717**: Central y C – terminal, auto-proteolítico (hidrólisis de unión Ala - Gly) al interactuar con RecA (Peptidasa S24).
TTGTAAACC - (16) - GGTTTACAA	***Escherichia coli***	CTGTN – (8) – ACAG
tTGTaAACC-N(14)... (16)-GGTTtACAa	**Proteobacteria**	CTGTN – (8) – ACAG (gamma γ - beta β) GAACN – (7) – GAAC ó GTTCN – (7) – GTTC (alfa α)
wwTGTtAAC-N(14)... (16)-GTTaACAww (Eubacteria - Arquea)	**No Proteobacteria**	CGAACrNryGTTyC (Firmicutes, gram +) rGTACNNNdGTwCb (Cyanobacteria)

Tabla I1. Comparación de las características regulatorias de dos regulones transcripcionales. Sobre la columna izquierda, **BirA** constituye un excelente modelo para la transferencia de interacciones pues tanto este TF, los TGs a los que regula y su bsDNA se encuentran conservados en procariontes. De forma contraria, en el caso de **LexA** (en la columna derecha) sólo conservan en procariontes este TF y los TGs a los que regula. BirA presenta dos funciones, por un lado es una ligasa para la proteína biotina y por el otro es el represor transcripcional (flechas en color rojo) del operón de biotina, mientras que LexA tiene la función de regular a los genes involucrados en la respuesta SOS de daño a DNA (KEGG, http://www.genome.jp/kegg/). Sus componentes regulados constan de 5 TGs para el caso de BirA, mientras que LexA regula a 28 TGs (Regulon DataBase, http://regulondb.ccg.unam.mx/index.html). Las estructuras terciarias presentan tres dominios PFAM conservados para BirA y dos para LexA, de los cuales, el que corresponde al número uno en ambos casos es el dominio de unión al DNA (Helix-turn-helix) (PFAM, http://www.sanger.ac.uk/Software/Pfam/). Se puede apreciar que los bsDNA para BirA, caracterizados computacional y experimentalmente, presentan una secuencia consenso conservada en eubacterias y arqueas (32); mientras que para LexA han sido reportado al menos 5 diferentes bsDNA en diversas clases y *phyla* bacterianos (33).

Más recientemente, Sharan *et al.* (34) asociaron funciones a las proteínas usando la conservación de grupos de interacciones en las redes proteína-proteína en genomas eucariontes. Ésto implica que una alta similitud de secuencia[15] no necesariamente significa que la función está conservada o que pueda ser compartida por otra secuencia de menor similitud; sino que la conservación de las interacciones al nivel de motivos o módulos en las redes, permite una mayor cobertura en la determinación de la función de una

[15] *Hits*. Generalmente, se toma el primer bit de una corrida de BLAST (el de mayor significancia estadística según el e-value calculado).

proteína (nodo) a partir de su contexto. De esta forma, los mejores apareamientos de secuencia (*hits*)[15] no siempre están presentes dentro de los grupos de proteínas conservadas. Esto proporciona una ventaja en la extensión de las predicciones porque se incrementa la tasa de detección de funciones conservadas al incluir familias de parálogos y la pérdida de genes. La alta especificidad de las predicciones realizadas por Sharan *et al*. (34), pueden ser apoyadas porque la conservación es evaluada en el contexto de una "subred" de interacción de proteínas y no independientemente para cada interacción. Sin embargo, se ha demostrado que los patrones de conservación entre interacciones proteína-proteína *versus* interacciones proteína-DNA es diferente (27), y que la lógica regulatoria transcripcional difiere radicalmente entre eucariontes y procariontes (35). Como una consecuencia, la realización de métodos de mapeo de las interacciones transcripcionales bajo la estrategia de Sharan *et al.* (34) puede ser evaluada a una gran escala, pero con bajos niveles de confianza sobre las predicciones (20, 27).

De esta forma, por medio del enfoque *Regulog* es posible reconstruir la evolución de las TRNs a través del mapeo de los componentes y de las interacciones regulatorias a lo largo del amplio número de genomas completamente secuenciados. El entendimiento sobre la evolución de las TRNs no sólo mejoraría nuestro conocimiento acerca de las restricciones biológicas de la regulación transcripcional que los diferentes organismos han adquirido para la adaptación a su ambiente, sino que también nos permitiría descifrar los principios del diseño básico sobre los que subyace su fenotipo molecular (36). A partir de los cuales, podemos reconstruir una historia regulatoria a partir del grupo de interacciones que han sido conservadas para diversos procesos celulares en las bacterias.

IV. Comparación general de los modelos: *Escherichia coli* vs *Bacillus subtilis*.

En este trabajo usamos las TRNs de dos diferentes modelos de bacterias. Uno de estos es la TRN de la bacteria gram-negativa ***Escherichia coli* K12** contenida en **RegulonDB** (**Regulon DataBase**), la cual es la mejor conocida en bacterias (21). Esta base de datos contiene la información experimental correspondiente a aproximadamente 20% de la TRN de *E. coli* (19). El segundo procarionte mejor estudiado en términos de su regulación transcripcional es la bacteria gram-positiva ***Bacillus subtilis***. Obtuvimos el grupo completo de las interacciones regulatorias de esta bacteria documentada en **DBTBS** (**DataBase of Transcriptional regulation in *Bacillus Subtilis***) (37). La Tabla I2 muestra una comparación entre estos dos organismos sobre sus principales características, así como de los elementos que conforman a sus TRNs.

Aún cuando ambas bacterias son de vida libre y requieren concentraciones similares de niveles de oxígeno y de temperatura. *E. coli* se ha adaptado a vivir dentro de su hospedero, mientras que *B. subtilis* se ha adaptado a ambientes de suelo (38-39) [ver Tabla I2-I]. Con respecto a sus características morfológicas, aunque ambos organismos presentan una forma cilíndrica; por un lado, *E. coli* presenta movilidad por medio de flagelos peritricos (que rodean su cuerpo), mientras que *B. subtillis* no posee flagelos y tampoco presenta movilidad (38-39) [ver Tabla I2-II].

Otra importante diferencia entre ambos modelos es que *B. subtilis* presenta dos distintos procesos de división celular, el de división binaria o vegetativo es compartido con *E. coli* y con todas las bacterias. Sin embargo, en estado de inanición o condiciones físico-químicas adversas en el ambiente, *B. subtilis* lleva a cabo un proceso de diferenciación celular que dará paso a la formación de esporas y de un cuerpo frutal de forma colectiva (39). La esporulación se lleva a cabo en siete estadios de diferenciación que empieza con una división celular asimétrica que produce dos tipos celulares distintos: la célula pre-espora (más pequeña) y la célula madre. Ambas células contienen el mismo material genético, pero utilizan diferentes interacciones en la expresión de los genes (40-41). De esta forma, en la célula pre-espora se sintetizará la endospora, madurará y liberará como una espora latente. La mayoría de la esporulación se lleva a cabo al nivel de la regulación transcripcional (41) y es controlada por diversos factores σ (40) [ver Tabla I2-III].

Las diferencias detectadas en la morfología, los estadios de vida, así como en los nichos ecológicos entre *E. coli* y *B. subtilis* hace que ambos organismos respondan a diferentes necesidades y requerimientos metabólicos, de energía, proliferación y de división celular, lo cual puede verse reflejado en los componentes que regulan la expresión primaria de los genes. En conjunto, el componente regulador (TFs) y el regulado (TGs) de la TRN comprenden una proporción significativa del genoma en cada organismo (23% en *Escherichia coli* y 19% en *Bacillus subtilis*) (21), lo cual constituye un componente genético considerable para la diversificación de los fenotipos bacterianos. Entender cómo evoluciona la TRN nos permite evaluar la expansión y el patrón de la diversidad genética en la regulación transcripcional, los efectos fenotípicos de la variación molecular y sus posibles consecuencias ecológicas.

Característica	*Escherichia coli* K12[a]	*Bacillus subtilis* [b]	Referencias
Linaje	Bacteria Proteobacteria Gammaproteobacteria Enterobacteriales Enterobacteriaceae Escherichia	Bacteria Firmicutes Bacillales Bacillaceae Bacillus	[a] y [b] **KEGG** (http://www.genome.ad.jp/kegg/)
Estilo de vida	Libre/Asociado hospedero	Suelos	[a] Blattner F.R., et al. (1997) [b] Kunst F., et al. (1997)
Patógeno	NO	NO	
Requerimiento de O_2	Facultativo	Facultativo	[a] y [b] **Bergey's Manual of Determinative Bacteriology** (http://www.cme.msu.edu/Bergeys/)
Niveles de temperatura	Mesófilo	Mesófilo	
Tamaño de genoma (pb)	4 639 675	4 214 630	[a] Blattner F.R., et al. (1997) [b] Kunst F., et al. (1997)
# de CDS	4 289	4 106	
# de RNAs	156	119	
% de G – C	50	43	[a] y [b] **NCBI** (http://www.ncbi.nlm.nih.gov/genomes/)
# de interacciones	1678	785	
# de TFs	119	99	[a] **RegulonDB** (http://regulondb.ccg.unam.mx) [b] **DBTBS** (http://dbtbs.hgc.jp/)
# de TGs	850	666	
# de regulones	118	93	
# de factores sigmas	7	17	
Reguladores globales	7 (CRP, FNR, IHF, FIS, ArcA, Hns, LRP)	8 (CcpA[1], AbrB[2], ComK[3], FUR[4], PhoP[5], TnrA[6], CodY[7], PurR[8])	[a] Martínez-Antonio & Collado-Vides (2003) [b]

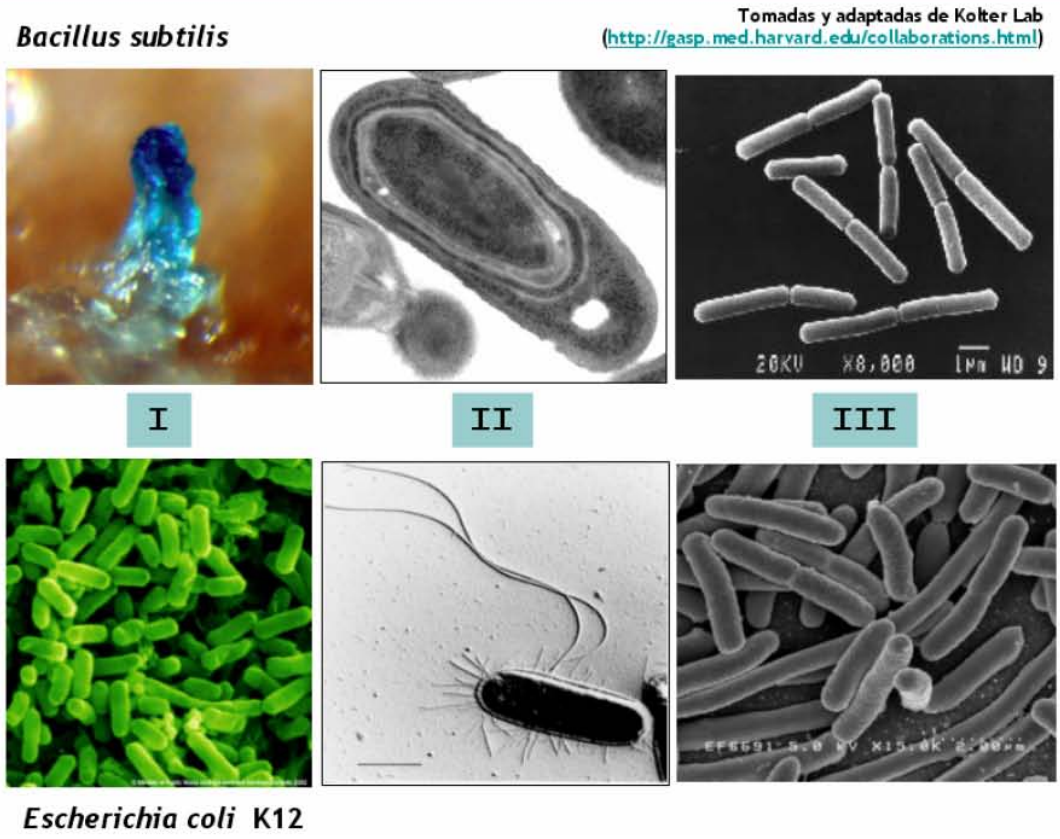


Tabla I2. Comparación de las características más sobresalientes de los modelos a usar en este trabajo. La bacteria gram-negativa *E. coli* y la gram-positiva *B. subtilis* presentan, entre otras características, estilos de vida y reguladores transcripcionales globales diferentes. [b] Referencias de los GRs de *B. subtilis* al final. En I, II y III se muestran las diversas diferencias morfológicas y de diferenciación celular entre ambos organismos.

OBJETIVO GENERAL

La motivación del proyecto estriba en conocer las restricciones y principios mínimos que la regulación transcripcional ha experimentado a lo largo de la evolución del fenotipo en las especies bacterianas.

De esta forma, usaremos una versión modificada del enfoque Regulog para poder identificar los componentes de la interacción regulatoria, pares de interacciones y regulones de las redes de regulación transcripcional de *Escherichia coli* y *Bacillus subtilis* a través de la comparación contra genomas completos de especies pertenecientes a los dominios celulares de Bacteria, Archaea y Eukarya.

OBJETIVOS PARTICULARES

- **Cuantificar la extensión (en 204 genomas secuenciados) de los componentes individuales, los pares de interacciones y grupos de interacciones (regulones) que conforman la red de regulación transcripcional desde dos modelos bacterianos.**
- **Inferir las relaciones existentes entre la segregación filogenética, el estilo de vida y la conservación de la red de regulación transcripcional en las especies bacterianas.**
- **Determinar si existe un grupo de interacciones regulatorias conservadas en las bacterias, y si éstas tienen asignadas funciones similares a las reportadas previamente en otros sistemas celulares.**
- **Tratar de entender cómo las redes de regulación transcripcional han evolucionado y, de qué forma estas redes influyen en el proceso de adaptación fenotípica de las bacterias.**

HIPÓTESIS

Dado que se ha demostrado en trabajos previos que las redes de interacciones metabólicas y las de proteína-proteína se encuentran conservadas en una gran escala filogenética (desde procariontes hasta eucariontes), esperamos que a partir del concepto de ortología, las redes de regulación transcripcional (conformadas de interacciones proteína-DNA) se encuentren conservadas en correspondencia con la distancia filogenética de los organismos bacterianos (Bacteria y Archaea).

MATERIALES Y MÉTODOS

A) Colección de proteomas. Un total de 204 genomas completamente secuenciados, incluyendo *Escherichia coli* K12 y *Bacillus subtilis* fueron obtenidos de la base de datos Kyoto Encyclopedia of Genes and Genomes (KEGG, ftp://ftp.genome.ad.jp/pub/kegg/genomes/) (42) [ver Esquema M1-I]. Ver **ANEXO 1** para detalles acerca de los 204 genomas completamente secuenciados que se usaron en este estudio.

B) Datos de las interacciones proteína-DNA. Obtuvimos las interacciones regulatorias de *E. coli* K12 de la base de datos RegulonDB version 4.0 (21), la cual recopila la información experimental extraída de la literatura. También obtuvimos las interacciones regulatorias de *B. subtilis* de la base de datos DBTBS (37). En este trabajo, sólo se consideraron aquellas interacciones regulatorias donde los TFs y los TGs codifican para un polipéptido. Así, las interacciones que involucran tRNAs y otros TGs que no codifican para un polipéptido fueron ignoradas. De igual forma, hay varios TFs que activan y reprimen al mismo TG a través de la presencia de más de un sitio de unión al DNA (e.g. CRP regula a *galE* como activador o represor a partir de dos diferentes sitios de unión a DNA); estos casos fueron considerados como interacciones redundantes y solamente una de las interacciones fue usada para representarlas en el grupo final. De esta forma, un total de 1678 interacciones regulatorias no redundantes que representan 119 TFs actuando sobre 850 TGs fueron incluidas en este trabajo para la TRN de *E. coli* [ver **ANEXO 2A**]. Mientras que un total de 785 interacciones no redundantes representando 99 TFs afectando 666 TGs fueron incluidos de la TRN de *B. subtilis* [ver **ANEXO 2B**].

C) Detección de TFs y TGs potenciales en diversos genomas. Ha sido extensamente reportado que: la duplicación de secuencias, la divergencia y la recombinación son las principales fuentes de variación funcional en la evolución de las proteínas (43-44). Sin embargo, es importante notar que dado que la definición de "función" ha sido vaga, diferentes enfoques han sido considerados en genómica comparativa para delimitarla (28, 45-46). En este trabajo, asignamos papeles funcionales a TFs y TGs en otros genomas usando la intersección de tres criterios para la detección de proteínas ortólogas: a) el mejor hit bidireccional (BDBHs - Bi-Directional Best Hits), b) la cobertura en el alineamiento BLASTP (47) y c) la conservación de dominios PFAM (48) [ver Esquema M1].

Los ortólogos son definidos como proteínas en diferentes especies que evolucionaron de un ancestro común por especiación (31) y usualmente mantienen la misma función. Las proteínas que recientemente evolucionaron de un ancestro común por duplicación previa a cualquier evento de especiación son llamados "outparalogs" y en mucho menor proporción mantienen la misma función (49). Por constraste, los "inparalogs" son aquellos que han evolucionado por duplicaciones genéticas después del evento de especiación y en mayor medida estas secuencias conservan también su función. Operacionalmente, ambas

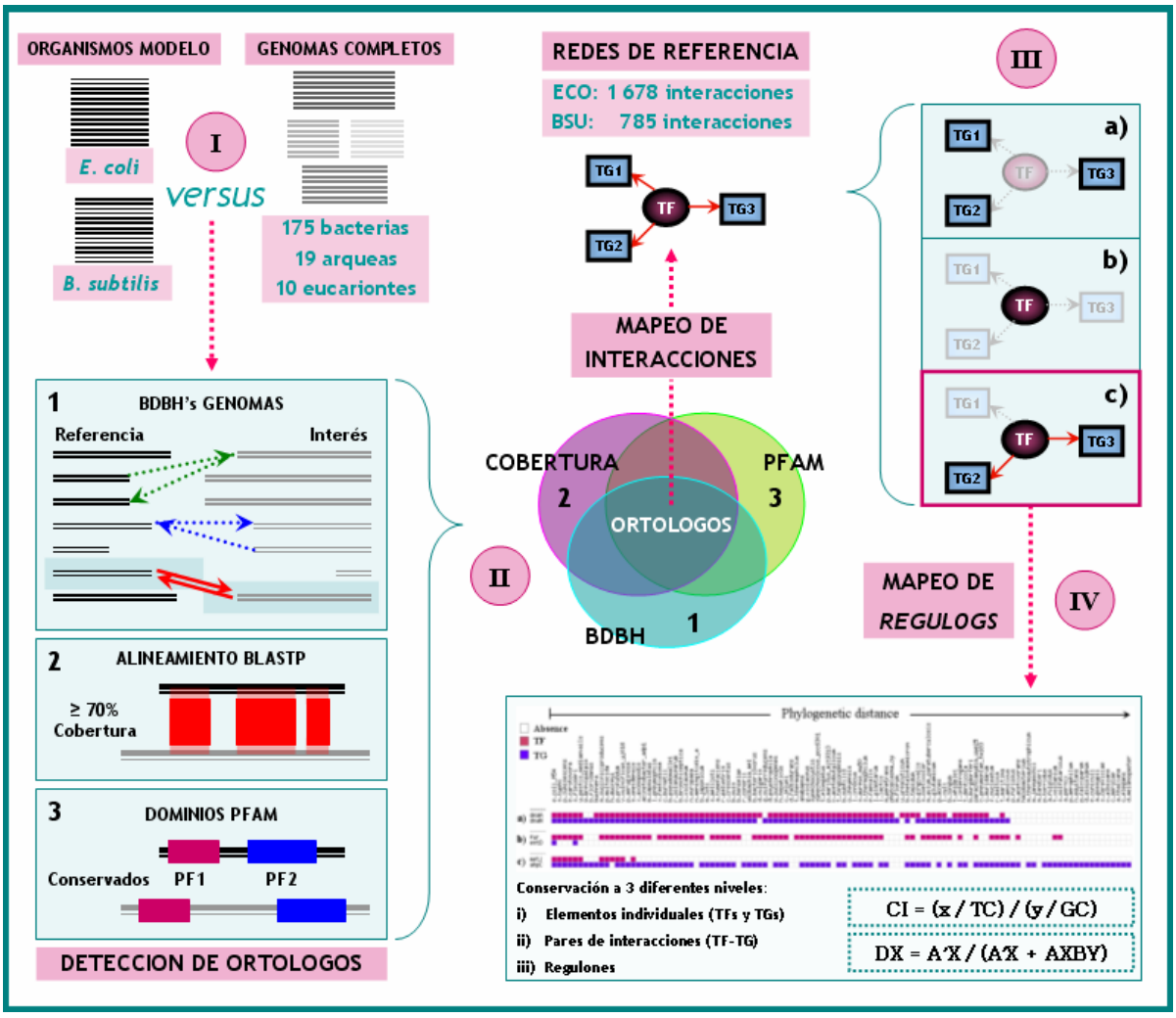
secuencias *inparalogs* y ortólogos son usualmente definidas como los homólogos de mejor apareamiento o como el mejor hit bidireccional en otro organismo (50-52). Las secuencias en el mismo genoma con más de 95% de identidad, estimadas con el programa CD-HIT (53), fueron consideradas en este trabajo como "inparalogs" e incluidas en grupos de paralogía reciente, en los cuales un solo representante (el que asigna CD-HIT) fue considerado para la búsqueda de los ortólogos, y con ello evitar menor cobertura en los BDBHs [ver Esquema M1-II.1]. Para identificar ortólogos usamos la definición de BDBHs a través de genomas depurados al 95% de identidad, con un e-value de BLASTP significativo ($\leq 10^{-3}$), usando el programa WU-BLAST (47). Sin embargo, la asignación funcional no está completa con sólo este enfoque.

Es bien conocido que los dominios de una proteína determinan su función específica y que éstos representan unidades evolutivas. Especialmente para proteínas con más de un dominio, donde el patrón de conservación funcional es más complejo. De esta forma, las proteínas tienden a compartir funciones si ellas contienen los mismos dominios en un arreglo similar (28, 54-55). Sin embargo, es importante considerar que un incremento en el número de dominios puede cambiar la función original de una proteína (56). Definimos los dominios conservados en las secuencias analizadas en este trabajo a través de la librería de Modelos basados en Cadenas de Harkov (HMM - *Hidden Markov Models*) tomados de la base de datos PFAM versión 10 (48), usando el programa HMMER 2.3.1 (57) con un e-value $\leq 10^{-3}$ [ver Esquema M1-II.3]. En adición, al menos un 70% del modelo PFAM fué cubierto por la secuencia.

Operacionalmente, identificamos ortólogos como aquellas proteínas que satisfacen las siguientes cuatro condiciones:

1. Secuencias que en el genoma de interés tienen el mejor hit bidireccional con el genoma de referencia con un e-value BLASTP significativo ($\leq 10^{-3}$).
2. Al menos el 70% de la secuencia de referencia debe estar incluida en el alineamiento BLASTP.
3. Secuencias candidatos que tengan uno o más dominios PFAM en la misma orientación y arreglo que aquellos de la secuencia de referencia y que no incrementen el tamaño total de las proteínas de referencia en más que 100 residuos.
4. Todas las secuencias que fueron incluidas previamente en los grupos de *inparalogs* fueron considerados candidatos que mantienen la función sólo si las condiciones 1, 2 y 3 son verdaderas para la secuencia representativa del grupo.

De esta forma, identificamos los ortólogos y los dominios PFAM de 119 TFs y 850 TGs de la TRN de *E. coli*, y de 99 TFs y 666 TGs de la TRN de *B. subtilis*, así como para el resto de las proteínas de los genomas completos de *E. coli* y *B. subtilis* en los genomas completos de 175 bacterias, de 19 arqueas y de 10 eucariontes. Una vez obtenidos los ortólogos, nos basamos en las interacciones de la TRN de la especie de referencia para construir matrices de presencia y ausencia de las interacciones por medio del método Regulog, donde la presencia de ambos componentes, el TF y su TG son necesarios para transferir la interacción a las especies de interés [ver Esquema M1-III y IV].



Esquema M1. Representación gráfica de la metodología utilizada para transferir interacciones. Se usó el método Regulog para transferir las interacciones regulatorias de la TRN de *E. coli* y *B. subtilis* a través de 175 genomas bacterianos, 19 arqueas y 10 eucariontes (**sección I**), para lo cual usamos la aproximación de tres enfoques genómicos en la detección de ortólogos (**sección II**): **1)** el mejor hit bidireccional (BDBHs), la reducción de inparalogs se usó como estrategia para evitar una menor cobertura de los BDBHs detectados, ya que secuencias entre los organismos de referencia e interés que son parálogos recientes y pueden desviar la bidireccionalidad de la comparación; **2)** 70% de la cobertura de la secuencia de referencia en el alineamiento BLASTP, y **3)** la obtención de dominios conservados PFAM. A partir de las interacciones conocidas para las TRNs de *E. coli* y *B. subtilis* y de los ortólogos determinados a través de estos tres enfoques, se trazaron las interacciones conocidas en todos los genomas analizados (**sección III a**, **b** y **c**) y aplicando el principio de Regulog se transfirieron las interacciones para cuyos componentes (TF y su TG) se encontraron los ortólogos correspondientes (**sección III c**) en los organismos de interés. La transferencia de tales interacciones después se analizó bajo diversos enfoques: componentes, pares de interacciones, reguladores globales y regulones, por medio del uso de métricas tales como el CI, DX, entre otras descritas más abajo (**sección IV**).

D) Tratamiento de los datos. Para facilitar el despliegue de los resultados, mostramos 110 genomas en todas las figuras, obtenidos al filtrar las cepas y especies del mismo género bacteriano, dejando como representante aquella con el número máximo de genes entre un género dado de organismos. Las relaciones filogenéticas derivadas de la distancia evolutiva desde *E. coli* y *B. subtilis* a todos los organismos fue obtenida según el procedimiento de reconstrucción filogenética previamente reportado por Brown *et al.* (58). La distancia evolutiva entre cualesquiera dos especies está relacionada a la suma de las distancias entre cada organismo y su ancestro común más cercano.

E) Conservación de ortólogos. Para normalizar el alcance de la conservación de los componentes (TFs y TGs) de la red regulatoria en comparación al genoma total, nosotros ideamos una métrica simple llamada **Índice de Conservación** (**Conservation Index - CI**) definido como:

$$\mathbf{CI = (x / TC) / (y / GC)}$$

Donde **x** es el número de ortólogos presentes en el genoma de interés del número total de componentes (**TC** = TFs o TGs) de la TRN bajo consideración, y **y** es el número total de ortólogos detectados en el genoma de interés de todos los genes que codifican proteínas (CoDing Sequences - CDS) en el genoma de referencia (**GC**), en el caso de *E. coli* involucra 4248 CDS y para *B. subtilis* 4079 CDS. De esta forma, el **CI** es una medida de conservación de los componentes de la TRN de un genoma ponderado respecto a la conservación de sus genes. Un **CI** cercano a **0** indicaría que los componentes de la red regulatoria están pobremente conservados en comparación a la conservación del genoma, mientras que un **CI** cercano a **1** sugeriría que ambos, el TF y el TG están conservados en el mismo alcance. Un **CI** por arriba de **1** nos sugiere que el componente regulatorio (TFs o TGs) desde el organismo de referencia se encuentra conservado en mayor medida que el resto de su contenido genético.

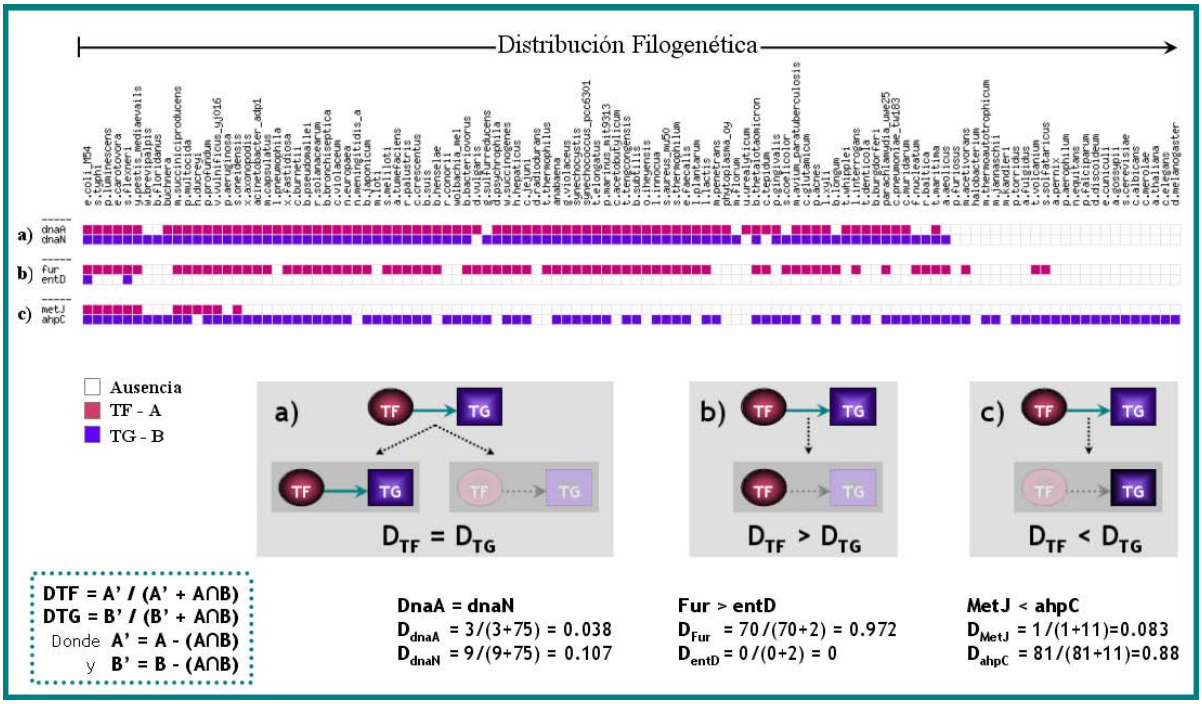
F) Persistencia de los pares de ortólogos de TF-TG a través de los genomas. Las enormes diferencias en el tamaño del genoma y el contenido de genes entre bacteria, arqueas y eucariontes, o entre organismos con diferentes estilos de vida (parásito, endosimbiótico o de vida libre) pueden introducir un sesgo cuando se calcula la frecuencia de distribución de las interacciones regulatorias conservadas entre organismos (a partir de sus ortólogos presentes). Para corregir este problema, consideramos un **factor de distancia** (**D**) para evaluar la presencia o ausencia conjunta de los TFs y sus TGs a través de los genomas:

$$\mathbf{D_X = A' / (A' + A \cap B)} \quad \mathbf{y} \quad \mathbf{D_Y = B' / (B' + A \cap B)}$$

$$\mathbf{Donde} \quad \mathbf{A' = A - (A \cap B)} \quad \mathbf{y} \quad \mathbf{B' = B - (A \cap B)}$$

Para el factor de transcripción **X-TF**, el cual regula al gene blanco **Y-TG**, **A** denota el grupo de todos los organismos dentro de los 110 genomas no redundantes en los cuales un ortólogo para X-TF es encontrado, y **B** denota el grupo de todos los organismos en los cuales un ortólogo para Y-TG es detectado. **A'** representa el subgrupo de organismos que tienen un ortólogo para X pero no para Y, y **B'** representa el subgrupo de organismos que tienen un ortólogo de Y pero no para X. **A∩B** representa el número de organismos en los cuales los ortólogos (A y B) son detectados. Como un ejemplo, consideremos el caso de un par interactuando, TF-X y TG-Y, donde la distancia **TF (D_X)** es más grande que la distancia **TG (D_Y)** porque el TF contiene un mayor número de ortólogos que su TG. Entonces, D_X puede contener un mayor número de ortólogos que D_Y, y derivado de ello, el número de ortólogos únicos en D_Y (B') debería ser más pequeño. En el límite, si la D_Y no tiene ortólogos únicos relativos a D_X, entonces la distancia D_Y sería cero. Un procedimiento similar ha sido usado por trabajos anteriores, pero enfocándose sobre el contenido de dominios en los genomas (59) [ver Esquema M2].

Las distancias de cada par de TF y TG para las TRNs completas generadas por el enfoque arriba mencionado fueron clasificadas en tres clases: **a)** un TF y TG co-ocurren y se asume que mantienen su relación regulatoria (**$D_{TF} = D_{TG}$**), **b)** el TF ocurre más que el TG (**$D_{TF} > D_{TG}$**) y **c)** el TG ocurre más que el TF (**$D_{TF} < D_{TG}$**) basados sobre umbrales predefinidos descritos en la próxima sección.



Esquema M2. Representación gráfica de las categorías evolutivas a partir de la co-ocurrencia en los componentes de la interacción (un par TF y su TG). Se calculó una métrica distancia (**D**) con la que se espera conocer una tendencia de co-ocurrencia entre los 1620 pares TF-TG en *E. coli* y 738 pares TF-TG pares en *B. subtilis* a través de los genomas analizados. Aquí mostramos el cálculo de D y la asignación a una de las tres categorías de co-ocurrencia para tres ejemplos de pares de interacción TF-TG obtenidas de la red de *E. coli*: DnaA-*dnaN*, Fur-*entD* y MetJ-*ahpC*. Se representa la distribución filogenéticas de los ortólogos del **TF** (**A en rojo**), de su **TG** (**B en morado**), y la ausencia de ortólogos (con blanco).

G) Obtención de los valores de corte (*thresholds*) para identificar las categorías de los pares de interacciones TF-TG co-ocurriendo. En ausencia de un grupo de control de interacciones regulatorias que puedan ser usadas para obtener los valores de corte (*thresholds*) para clasificar los pares TF-TG en las tres categorías y a partir de la métrica de distancia D, usamos la colección de pares de genes de la rutas metabólicas bajo la consideración de que estos últimos se encuentran más cercanos a co-ocurrir [ver **ANEXO 3A**]. Se ha observado que las rutas metabólicas se encuentran bien conservadas a través de los diversos grupos filogenéticos y son relativamente estables, en contraste de la maquinaria transcripcional (60-61).

Con la finalidad de representar mejor un valor de corte (*threshold*) para obtener pares de co-ocurrencias, el proceso de la obtención de distancias fue repetido para ambos genomas usando los pares de genes (que codifican para enzimas) de las rutas metabólicas documentadas en la base de datos KEGG (42) [ver **ANEXO 3B**]. Removimos la redundancia en los pares, ya que dos rutas diferentes pueden tener los mismos pares de genes. En resumen, para cada ruta metabólica P con N genes en ambos genomas, calculamos las distancias **D1** y **D2** de los pares de interacciones no-redundantes (**P1 - P2**) posibles **(N* (N-1) / 2).** Un promedio de las distancias D1 y D2 fue calculado para la colección completa no-redundante de pares en cada ruta del genoma de referencia. Sin embargo, esta distancia promedio D1 o D2 es un sobre-estimado del alcance de la conservación para las interacciones regulatorias. De esta forma, una medida de ***desviación estándar promedio*** fue aplicada para generar pares interactuantes co-ocurriendo significativamente. Ya que las interacciones metabólicas D1 y D2 son sinónimos poco similares a D_{TF} y D_{TG} en las interacciones regulatorias, nosotros usamos un valor de corte final (*final threshold*) de:

$$T = \{(<D1> - STD1) + (<D2> - STD2)\} / 2$$

Donde **<D1>** y **<D2>** son los valores promedio de D1 y D2 para todos los pares, y donde **STD1** y **STD2** son las desviaciones estándares de D1 y D2 [ver **ANEXO 3B**].

H) El análisis estadístico para la persistencia de las interacciones. Para evaluar la significancia estadística de la co-ocurrencia de las interacciones regulatorias en estas tres diferentes clases, comparamos contra 1 000 redes regulatorias construidas aleatoriamente para *E. coli* y *B. subtilis*. Cada red aleatoria se reconstruyó con el mismo número de interacciones como las TRNs originales, pero con una reconexión de las aristas, manteniendo así el grado de conectividad de cada nodo como en la TRN original. Es importante notar que este método de aleatorización preserva el grado de conectividad de entrada y de salida del nodo y, de esta forma, se preserva la topología de las TRNs de referencia [ver **ANEXO 4A**]. Para este propósito, se utilizó el *software* mfinder (62) generando 1 000 redes aleatorias (-r) en el modo *default* que preserva la distribución de conectividad de cada nodo (*in-degree* y *out-degree*). En el análisis completo, excluimos las interacciones donde TFs son auto-regulados (*mutual-degree*), ya que estas

interacciones podrían generar un sesgo al calcular el efecto entre la co-ocurrencia de pares TF-TG. Así, el grupo final de interacciones analizadas bajo este enfoque incluyó 1620 pares TF-TG en *E. coli* y 738 pares TF-TG en *B. subtilis*. Una vez obtenidas las 1 000 redes aleatorias se calcularon nuevamente las métricas de distancias (D_{TF} y D_{TG}) para cada red aleatoria. Para los datos generados en esta última fase, se calcularon dos medidas de significancia estadística, el Z-score y el p-value, dado que previamente se evaluó que los datos presentan una distribución normal [ver **ANEXO 4B**].

I) Regulones conservados. Para cada TF en *E. coli* y *B. subtilis*, calculamos el porcentaje de interacciones totales conservadas para cada regulón entre los genomas de interés. Para representar esta distribución, agrupamos estos porcentajes por la extensión de la TRN y la conservación del regulón a través del método *Centroid Linkage Clustering* con una métrica de distancia *Uncentered Correlation* del programa Cluster (63). Los datos agrupados representan 118 regulones en *E. coli* y 93 regulones en *B. subtilis* distribuidos a través de los genomas.

Calculamos grupos de regulones y de genomas por medio del método k-means del programa Cluster (63) para evaluar otras métricas de distancia, tales como *Euclidean* y *Kendall 's Tau*, pero no se encontró que fueran significativamente diferentes en su habilidad para formar agrupamientos que delimiten los linajes de los genomas, así como tampoco de los regulones. Iniciamos la comparación con una colección de objetos (regulones y especies: 118 regulones en *E. coli*, 93 regulones en *B. subtilis*, y 110 genomas en ambos casos) que formaron 10 grupos (*clusters*) (k1) de regulones, y 5 grupos (*clusters*) (k2) de especies que fueron calculados a partir de la aleatorización de tales grupos en 100 veces.

La suma de las distancias dentro de los clusters es usada para comparar las diferentes soluciones de los agrupamientos. La solución del agrupamiento, en conjunto con la suma más pequeña de la distancia al interior al cluster, fué reclutada para cada métrica de distancia. De esta forma, comparamos los objetos (5 clusters de regulones y 10 clusters de las especies) para cada grupo generado con la métrica de distancia *Uncentered Correlation*, con respecto a los objetos agrupados en cada cluster generado con las métricas de distancias *Euclidean* y *Kendall 's Tau*.

Finalmente, calculamos la proporción de objetos (regulones y especies) compartidos para cada cluster en las dos últimas métricas con respecto a *Uncentered Correlation* En particular, encontramos que los resultados no son sensitivos a este parámetro. Así, decidimos usar la métrica de distancia *Uncentered*, la cual es ampliamente usada para representar la distribución de amplios grupos de datos biológicos [ver análisis en la página web diseñada para este trabajo, dirección más abajo].

J) Interacciones regulatorias ancestrales. Usamos la distribución filogenética de los pares TF-TG, el principio de Parsimonia y el método DOLLOP (64) para detectar si existe un grupo de interacciones regulatorias transcripcionales compartidos entre bacterias. Es importante hacer notar, que este método para detectar las interacciones conservadas es diferente de los otros enfoques descritos previamente en este trabajo. En este método no sólo se determina el número total de genomas en los que ocurre una interacción,

sino que se evalúa la presencia o ausencia de una interacción para cada grupo filogenético definidos previamente en Brown *et al.* (58).

En la primera fase, sólo consideramos la distribución filogenética de los genomas no-redundantes (como se describe en MÉTODOS-D de tratamiento de los datos). Además, no consideramos aquellos genomas menores a 1000 CDS y a aquellos organismos cuyos estilos de vida son considerados parásitos o endosimbiontes, ya que presentan una presión selectiva para reducir sus genomas. En el próximo paso, los pares de interacciones TF-TG fueron analizados en los grupos filogenéticos para cada dominio celular a través del principio de Parsimonia, el cual detecta caracteres que posiblemente se han transferido horizontalmente y los elimina de futuros análisis. Todas aquellas interacciones que fueron detectadas en más de la mitad del número total de especies en cada clado, fueron consideradas como las interacciones ancestrales para este grupo. En un último paso, los grupos de pares TF-TG candidatos como ancestrales fueron evaluados con el programa DOLLOP disponible en PHYLIP (64), en cuya base teórica se asume que en la evolución es mucho más difícil ganar una característica compleja que perderla [ver **ANEXO 5**].

K) El material suplementario disponible *online*. El material suplementario que incluye: **a)** los datos de las interacciones regulatorias (regulones) utilizadas para *E. coli* K12 y *B. subtilis*, **b)** el análisis estadístico realizado, **c)** la evaluación de la métricas de clustering utilizadas, **d)** las tablas de funciones de los TFs para ambos organismos, **e)** las figuras de alta resolución, así como **f)** las predicciones generadas en el presente trabajo en los 204 genomas completos pueden ser consultadas y se encuentran disponibles en la siguiente página web: http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/

RESULTADOS Y DISCUSIÓN

1. Conservación de los componentes de las TRNs (TFs y TGs) entre las especies.

Basados en la información experimental de 119 TFs y 850 TGs de *Escherichia coli* K12 y 99 TFs y 666 TGs de *Bacillis subtilis*, identificamos sus contrapartes en 175 genomas de bacterias, 19 genomas de arqueas y en 10 genomas de eucariontes [ver **ANEXO 6A** y **6B**]. La Figura R1A-B muestra la distribución de los ortólogos de los componentes (TFs y TGs) de la TRN de *E. coli* y *B. subtilis* a través de 110 genomas no redundantes, representando 23 diferentes *phyla* de los tres dominios celulares basados en la reconstrucción filogenética de Brown *et al.* (58) [ver MÉTODOS-D].

Desde la perspectiva de *E. coli* [ver Figura R1A], el *phylum* más cercano incluye 76 diferentes Proteobacterias agrupadas en cinco subdivisiones (15 α, 10 β, 42 γ, 4 δ and 5 ε). El alcance de conservación en estos grupos es el más alto de todos los *phyla* analizados, donde el 30% de los TFs y TGs están conservados[16], con la excepción de organismos parásitos y endosimbiontes, los cuales comparten sólo 10% de los TFs y 20% de los TGs de la TRN de *E. coli*. En los Firmicutes, divididos en cuatro clases (10 Mollicutes, 22 Bacillales, 15 Lactobacillales y 4 Clostridia), se encontró que tienen del 20 al 30% de TGs y del 10 al 20% de TFs conservados; con la excepción de los organismos parásitos y endosimbiontes que pertenecen a los grupos de Mollicutes, Mycobacterium, Tropheryma, los cuales presentan menos del 5% de conservación para ambos componentes. Las fracciones de ortólogos en Firmicutes son similares a las detectadas en Actinobacteria. Otros *phyla* como Bacteroidetes, Fusobacteria, Planctomyces, Cyanobacteria, Deinococci, Aquificae y Thermotogae comparten del 10 al 25% de TGs y del 5 al 15% de TFs. Los *phyla* de parásitos que incluyen Chlamydiales y Spirochaetes comparten menos que 15 y 5% de TGs y TFs, respectivamente. Entre los 19 genomas de arqueas, los cuales comprenden 4 Crenarchaeota y 14 Euryarchaeota, encontramos que comparten entre el 7 y el 15% de TGs y menos que 3% de los TFs. El único parásito arqueano conocido, *Nanoarchaeum equitans*, comparte menos que el 1% de TGs y TFs de la TRN de *E. coli*.

Finalmente, 11 genomas de eucariontes, los cuales incluyen 2 Protistas, 4 Hongos, 2 Plantas, 1 Insecto y 1 Nematodo comparten entre 8 y 18% de los TGs con la excepción de los parásitos intracelulares obligados, tales como *Encephalitozoon cuniculi* que muestra solamente 2% de TGs. En contraste, se puede

[16] **NOTA:** Se usará de forma preferencial los conceptos **compartir, presencia o ausencia** de TFs, TGs e interacciones para los ortólogos detectados en los distintos genomas aquí analizados, en lugar de los conceptos **conservación y pérdida** de TFs, TGs e interacciones, ya que el uso de estas palabras implica que: **1)** para el caso de la **conservación**, se refleja un proceso de divergencia evolutiva por especiación que mantiene los ortólogos y, en este trabajo no descartamos metodológicamente la transferencia horizontal de genes de TFs ni de TGs; y **2)** para el caso de la **pérdida**, se refleja implícitamente un proceso de conservación y presencia de TFs y TGs en las especies ancestrales y, bajo el principio de evolución por Parsimonia (evolución por divergencia en el menor número de pasos), en este trabajo NO se hipotetiza que las TRNs de *E. coli* ni la de *B. subtilis* sean las ancestrales a todas las especies analizadas (Bacteria, Archaea y Eukarya). El uso de los conceptos de conservación y pérdida se aplica, en este trabajo, en aquellos casos donde las especies de comparación con las de referencia, tengan distancias filogenéticas cercanas y más preferencialmente a grupos filogenéticos completos, donde la existencia de un ancestría común en las TRNs es más probable. Así como también, se aplica para aquellos organismos donde se ha mostrado que existen fenómenos de reducción genómica masiva, resultado de estilos de vida parásito o simbionte y, que muestren la condición anterior de relación filogenética.

Figura R-1A. Conservación de los componentes de la TRN (119 TFs y 850 TGs) de *Escherichia coli* K12 (con 4 248 genes) a través de los 3 dominios celulares. Sobre el eje X se representan los 110 genomas no-redundantes ordenados por distribución filogenética [ver MÉTODOS-D y **ANEXO 1** y **6A**]. En el eje Y (sobre la izquierda) se representa el porcentaje de conservación de los elementos (TFs y TGs) de la TRN de referencia. Los valores CI (mostrados a la derecha del eje Y) representan una medida de conservación de los componentes de la TRN de un genoma con respecto a la conservación de sus genes. Un CI cercano a 0 indicaría que los componentes de la red regulatoria están pobremente conservados en comparación a la conservación del genoma, mientras que un CI cercano a 1 sugeriría que ambos, el TF y el TG están conservados en el mismo alcance. Un CI por arriba de 1 nos sugiere que el componente regulatorio (TFs o TGs) desde el organismo de referencia se encuentra conservado en mayor medida que el resto de su contenido genético.

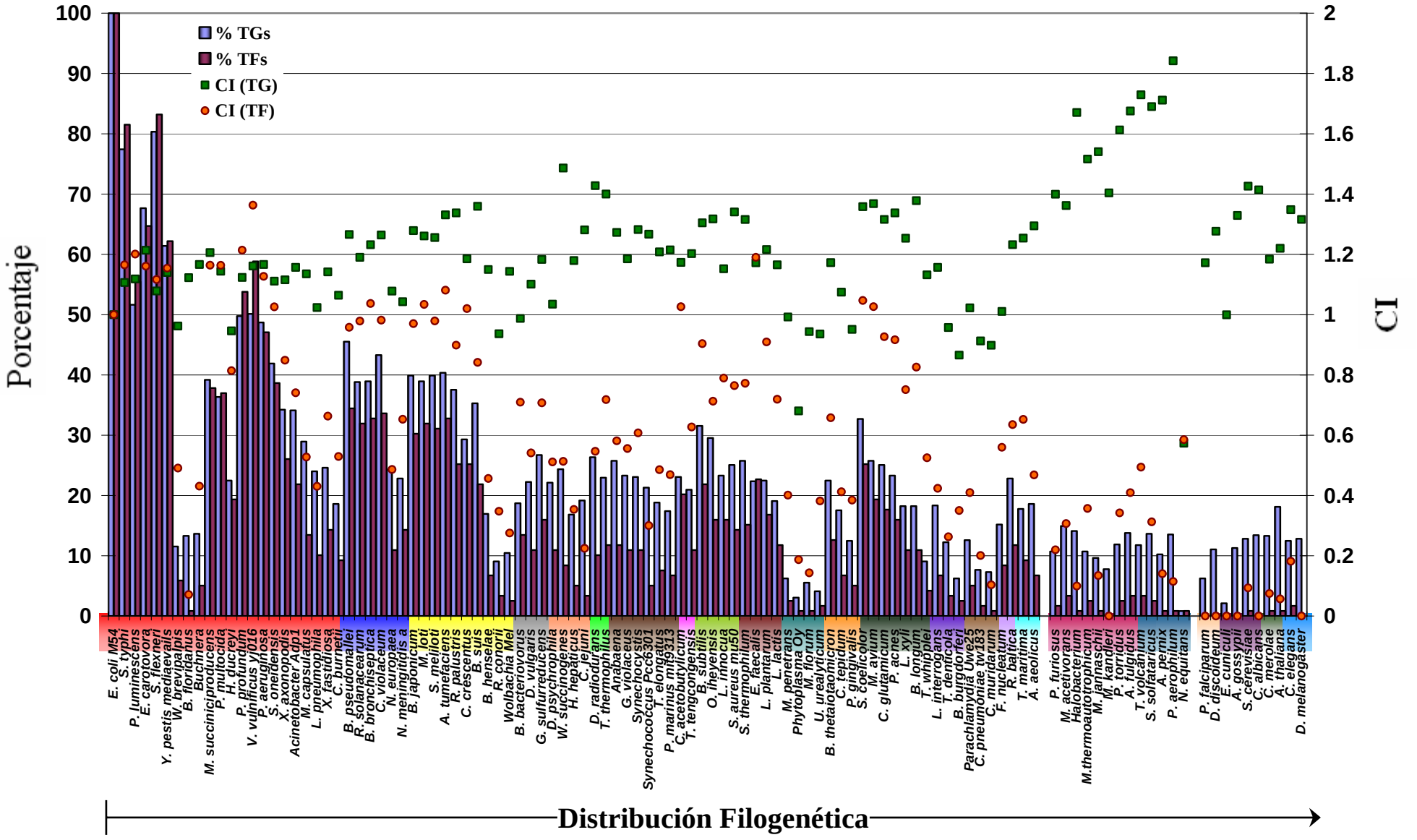
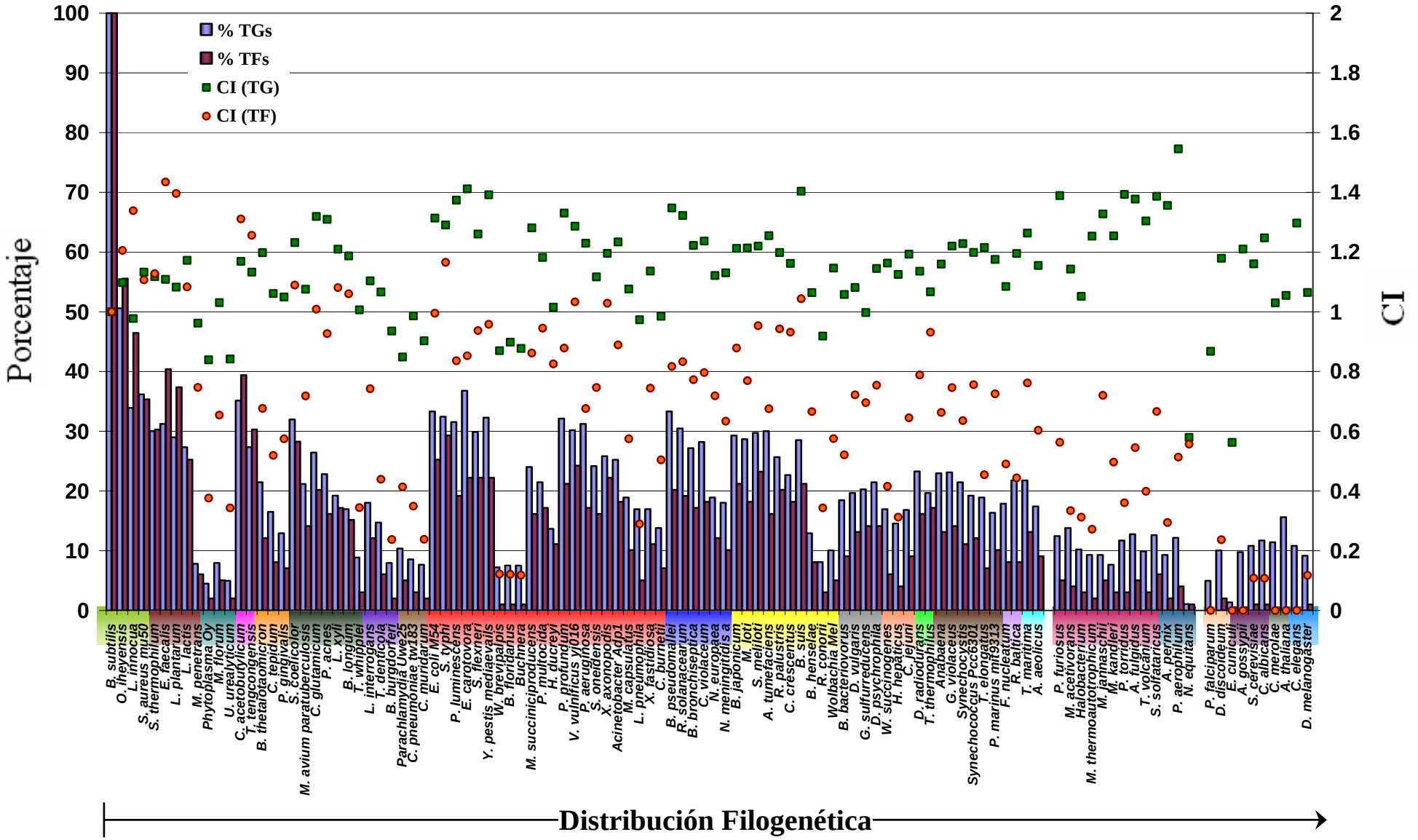


Figura R-1B. Conservación de los componentes de la TRN (99 TFs y 666 TGs) de *Bacillus subtilis* (con 4 079 genes) a través de los 3 dominios celulares. Sobre el eje X se representan los 110 genomas no-redundantes ordenados por distribución filogenética [ver MÉTODOS-D y **ANEXO 1** y **6B**]. En el eje Y (sobre la izquierda) se representa el porcentaje de conservación de los elementos (TFs y TGs) de la TRN de referencia. Los valores CI (mostrados a la derecha del eje Y) representan una medida de conservación de los componentes de la TRN de un genoma con respecto a la conservación de sus genes. Un CI cercano a 0 indicaría que los componentes de la red regulatoria están pobremente conservados en comparación a la conservación del genoma, mientras que un CI cercano a 1 sugeriría que ambos, el TF y el TG están conservados en el mismo alcance. Un CI por arriba de 1 nos sugiere que el componente regulatorio (TFs o TGs) desde el organismo de referencia se encuentra conservado en mayor medida que el resto de su contenido genético.



observar que *Saccharomyces cerevisiae*, *Cyanidioschyzon merolae*, *Arabidopsis thaliana* y *Caenorhabditis elegans* comparten menos que 1% de los ortólogos de TFs de la TRN de *E. coli*, mientras que en el resto de los genomas eucariontes no se detectaron ortólogos de TFs de esta red.

Desde la perspectiva de *B. subtilis* [ver Figura R1B], pueden observarse mayores fluctuaciones en la distribución de los ortólogos detectados entre los genomas. Más de un 25% de los TFs y TGs se encuentran conservados en los linajes de Bacillales y Lactobacillales. Los organismos parásitos y endosimbiontes en Mollicutes, Chlamydia, Spirochetes y α y γ-Proteobacterias comparten menos del 10% de los TGs y 5% de los TFs. La conservación de los TGs a través de las Proteobacterias es similar hasta *Bdellovibrio bacteriovorus* con respecto a la detectada en Bacillales, a pesar de las variaciones en las distancias filogenéticas entre estos grupos. A diferencia de la distribución de los TFs observada en *E. coli*, en *B. subtilis* se muestra una presencia casi constante en el número de TFs que se comparten en los diferentes grupos filogenéticos del dominio Bacteria. Aún cuando los porcentajes de ortólogos son similares, ello no implica que sean el mismo grupo de TFs. Más allá de los linajes bacterianos, donde encontramos que la conservación de los TFs de la TRN de *B. subtilis* disminuye rápidamente en arqueas y en eucariontes, sólo se detectó un TF compartido entre *S. cerevisiae, Candida albicans y Drosophila melanogaster* y dos TFs compartidos en *Dictyostelium discoideum.*

Desde la perspectiva de ambos genomas, se pueden observar tendencias similares en la distribución de ortólogos de los componentes de las TRNs de *E. coli* y *B. subtilis* a través de varios *phyla*. Lo primero que resalta es el rápido decremento en los ortólogos, tanto de TFs como de TGs. Inclusive, las especies filogenéticamente cercanas no presentan más de 30 a 60% de ambos componentes. Por otro lado, se observa que los TGs tienden a estar más conservados que los TFs conforme la distancia filogenética incrementa, mientras que en los linajes cercanos filogenéticamente, los TFs se encuentran más conservados que los TGs. Esto nos sugiere que la mayoría de la maquinaria regulatoria transcripcional en Bacteria puede ser linaje específico, lo cual refuerza una observación similar previa sugerida al nivel de taxa (65). Esta tendencia también nos sugiere que existen presiones de selección para mantener parte de la TRN en los organismos pertenecientes a un mismo *phyla* por medio de un grupo de reguladores comunes, mientras que el componente regulado (TGs) es el que cambiaría constantemente en estos casos.

El componente regulado (TGs) de las TRNs caracterizadas para *E. coli* corresponde al 20% y para *B. subtilis* al 16% de sus genomas completos; mientras que el componente regulador (TFs) corresponde al 3% en *E. coli* y al 2.5% en *B. subtilis*. En general, la medida de conservación (CI) [ver MÉTODOS-E y Figura R1A-B] para *E. coli* muestra que hay un incremento en la conservación de la proporción del componente regulado en los organismos en comparación al componente regulador, lo cual es mucho más notorio en los dominios Archaea y Eukarya. Desde el punto de vista de *B. subtilis*, aunque el decremento del CI en el componente regulador no es claro sino hasta linajes lejanos, la tendencia del componente regulado se puede observar más semejante a la detectada con *E. coli*. Lo anterior nos sugiere que el componente regulador ha tenido una mayor divergencia dentro de la evolución del genoma.

2. Evolución de los Reguladores Globales a través de las especies bacterianas.

Los Reguladores Globales (Global Regulators - GRs) regulan la actividad del 51% de la TRN conocida en *E. coli* (19). Estos TFs regulan grandes sistemas de operones que se ven sometidos a una regulación coordinada común, encaminada a la supervivencia de la bacteria en determinadas circunstancias ambientales extremas o, para realizar grandes ajustes metabólicos que le permitan adaptarse a cambios bruscos en las condiciones nutricionales. De esta forma, esperamos que estos GRs se encuentren conservados a lo largo del espectro filogenético. Consideramos la definición de GRs para *E. coli* de Martínez-Antonio y Collado-Vides (19), basados en el número de genes que regulan y algunos factores adicionales, tales como el número de co-reguladores y el número de condiciones en los que ejercen su actividad regulatoria. Dada la ausencia de suficiente información sobre *B. subtilis* para clasificar TFs sobre las mismas bases, consideramos como GRs a aquellos TFs que regulan el mayor número de genes en la TRN conocida (más de 20 interacciones regulatorias). Así, se conciben siete GRs en *E.coli*: CRP, FNR, IHF, FIS, ArcA, Hns y LRP, y ocho en *B. subtilis*: CcpA, AbrB, ComK, FUR, PhoP, TnrA, CodY y PurR.

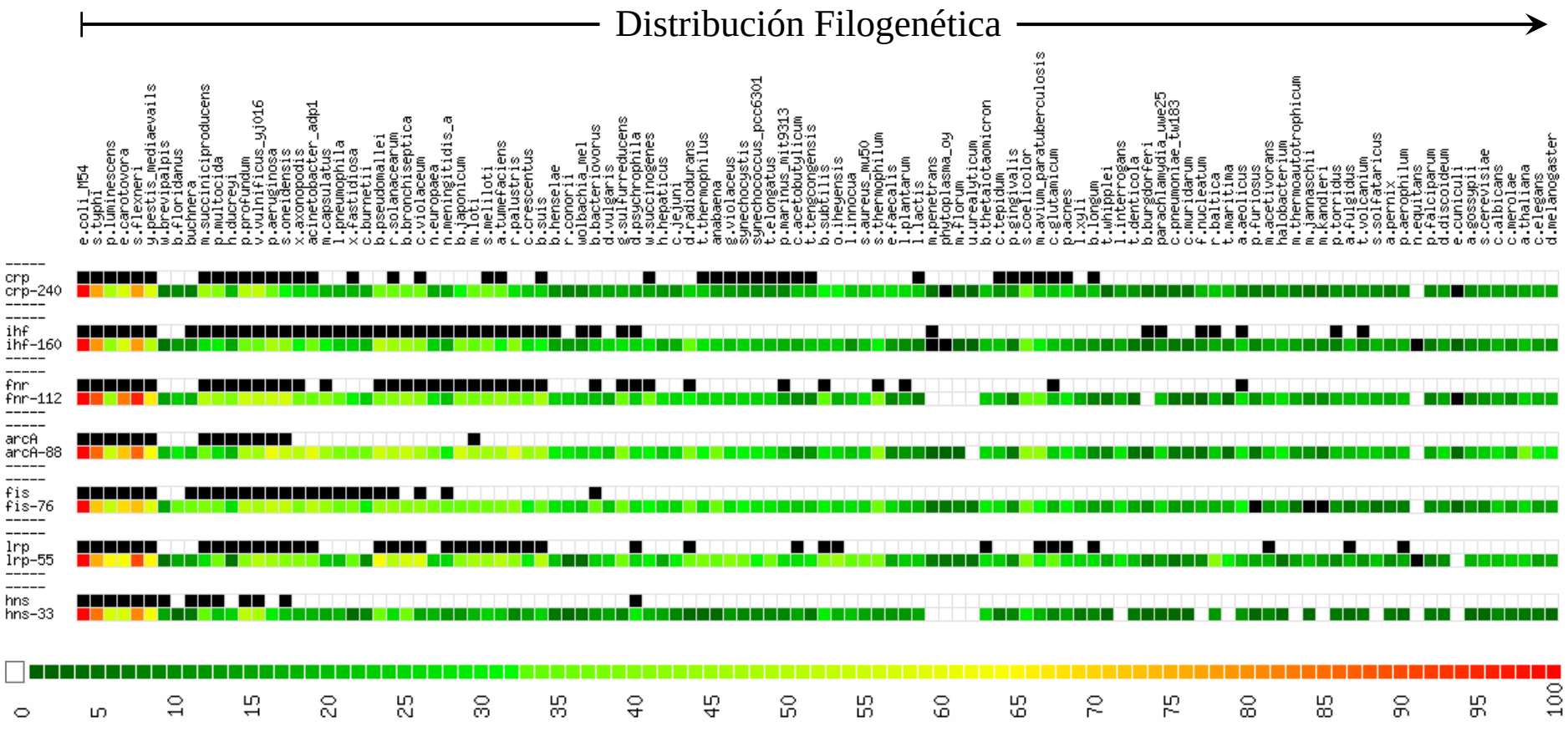
Sorpresivamente observamos que los ortólogos identificados para los GRs varían en su distribución filogenética y ninguno de ellos se identificó en eucariontes [ver Figura R2A-B]. Sin embargo, los TGs de estos GRs están conservados en los tres dominios celulares, sugiriendo que estos TGs podrían ser regulados en estos organismos por TFs parálogos, análogos o por otros mecanismos de regulación (i.e. *riboswitches*, RNAs pequeños, etc.).

Es interesante notar que ninguno de los GRs son parálogos a lo largo de toda la secuencia (no sólo al nivel de una porción de los dominios, como el dominio *Helix Turn Helix* de unión al DNA) entre *E. coli* [ver Figura R2A] y *B. subtilis* [ver Figura R2B]. Lo anterior indica que los TFs globales en organismos diferentes no comparten ancestros comunes cercanos. Esta observación podría implicar que los TFs globales evolucionan independientemente en diferentes linajes, en respuesta a los requerimientos en diferentes condiciones ambientales. De todos los GRs, sólo LRP y las subunidades de IHF (HimD o HimA) fueron detectadas en el dominio Archaea, sugiriendo que la gran parte de estos GRs se originaron de forma independiente en los linajes de Bacteria. Curiosamente, se observa que en algunos grupos de Proteobacteria, Firmicutes y Cianobacteria los ortólogos de CRP y FNR, los cuales son parálogos en *E. coli*, tienen una distribución alternativa, posiblemente indicando una sustitución de sus papeles funcionales en estos linajes. Otra posibilidad es que se haya transferido horizontalmente uno (o ambos) de los miembros de la familia CRP en diferentes organismos. Se ha reportado que los miembros de la familia CRP-FNR destacan en la respuesta de un amplio espectro a señales intracelulares y extracelulares, tales como AMP cíclico (cAMP), anoxia, estado redox, estrés oxidativo[17] y nitrosativo[18], óxido nítrico, monóxido de carbono, 2-oxoglutarato,

[17] **Estrés oxidativo** (66). Es un estado de la célula en la cual se encuentra alterada la homeostasis óxido-reducción intracelular, es decir el balance entre prooxidantes y antioxidantes. Este desbalance se produce a causa de una excesiva producción de especies reactivas de oxígeno (EROs) y/o por deficiencia en los mecanismos antioxidantes, conduciendo a daño celular.

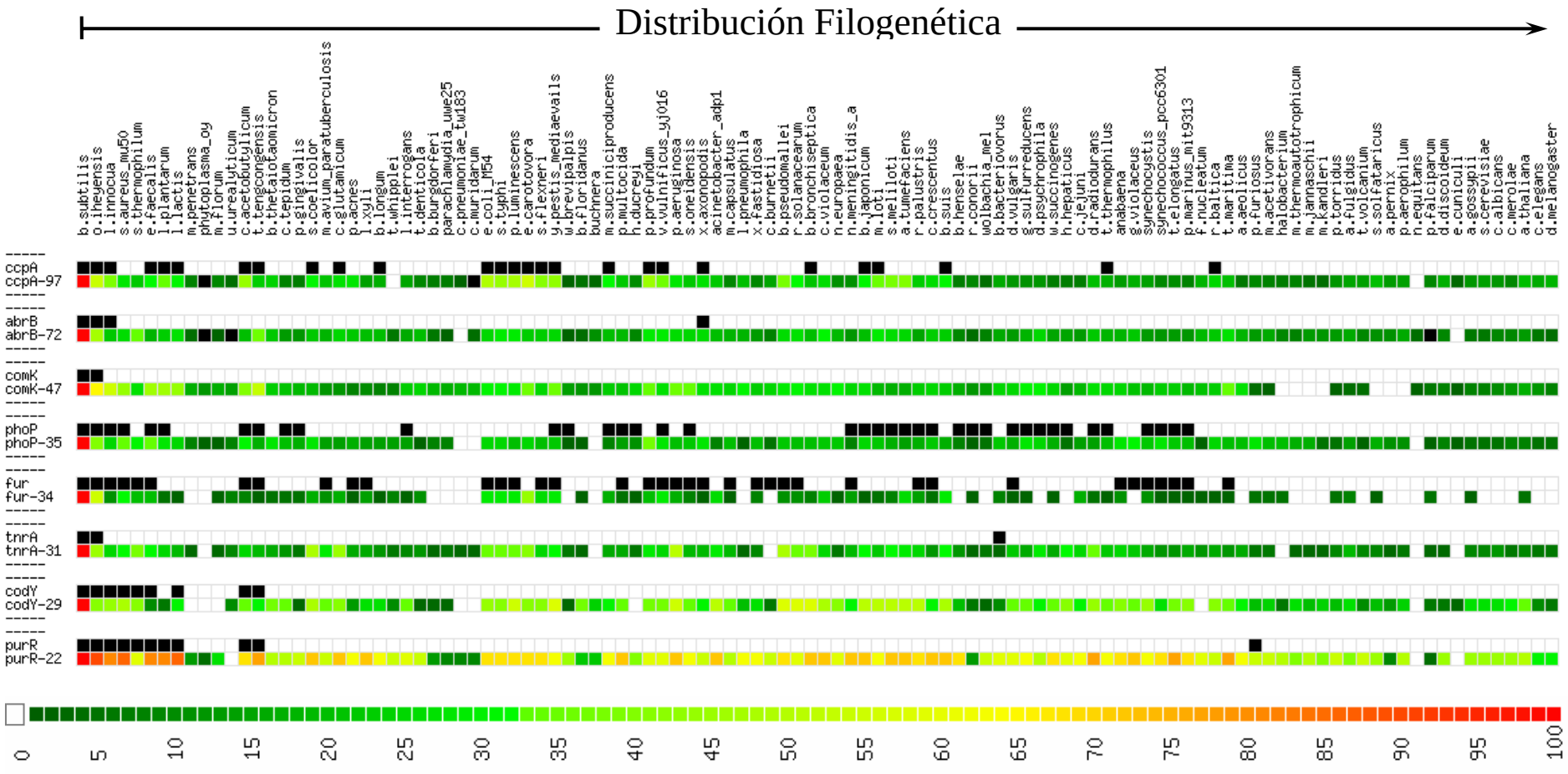
[18] En analogía al término "estrés oxidativo", Hausladen y Stambler (66) han denominado "**estrés nitrosativo**" a la excesiva o desregulada formación del radical óxido nítrico ($NO^{\cdot}$) y especies reactivas del Nitrógeno (ERNs) derivadas del mismo.

Figura R-2A. Conservación de los Reguladores Globales de la TRN (7 GRs) de *Escherichia coli* K12 a través de los 3 dominios celulares. Sobre el eje horizontal se representan los 110 genomas no-redundantes ordenados por distribución filogenética [ver MÉTODOS-D y **ANEXO 1** y **7A**]. En el eje vertical (sobre la izquierda) se representa para cada uno de los 7 GRs la fila de distribución de los ortólogos detectados para el GR (el color negro significa presencia y el blanco significa ausencia); y en la fila de abajo, la distribución del porcentaje de ortólogos detectados de los TGs a los cuales regula con respecto al total reportados en *E. coli* (GR - #TGs). Una barra de correspondencia para los porcentajes representados por colores, se muestra en la base de la figura. Las funciones para cada GR fueron obtenidas de **RegulonDB** (http://regulondb.ccg.unam.mx).



GR	#TGs	Función sólo en regulación transcripcional
CRP	**240**	Regula el censado a catabolitos
IHF	**160**	Factor hospedero de integración-recombinación en DNA
FNR	**112**	Regula la reducción de nitrato y fumarato
ArcA	**88**	Regula la respiración aeróbica
FIS	**76**	Regula la transcripción del RNA ribosomal
LRP	**55**	Regula la respuesta a leucina
HNS	**33**	Regula la transcripción del DNA en varios estreses

Figura R-2B. Conservación de los Reguladores Globales de la TRN (8 GRs) de *Bacullis subtilis* a través de los 3 dominios celulares. Sobre el eje horizontal se representan los 110 genomas no-redundantes ordenados por distribución filogenética [ver MÉTODOS-D y **ANEXO 1** y **7B**]. En el eje vertical (sobre la izquierda) se representa para cada uno de los 8 GRs la fila de distribución de los ortólogos detectados para el GR (el color negro significa presencia y el blanco significa ausencia); y en la fila de abajo, la distribución del porcentaje de ortólogos detectados de los TGs a los cuales regula con respecto al total reportados en *B. subtilis* (GR - #TGs). Una barra de correspondencia para los porcentajes representados por colores, se muestra en la base de la figura. Las funciones para cada GR fueron obtenidas de **DBTBS** (http://dbtbs.hgc.jp/).



GR	#TGs	Función sólo en regulación transcripcional
CcpA	**97**	Regula la represión a catabolitos y la excreción de exceso de carbono
AbrB	**72**	Regula el crecimiento vegetativo, la fase estacionaria, la esporulación y la represión a catabolitos
ComK	**47**	Regula la competencia en la transducción de señales
PhoP	**35**	Regula las condiciones de reserva a fosfato
FUR	**34**	Regula la biosíntesis de sideróforos y la transcripción de ferri-sideróforos
TnrA	**31**	Regula la degradación de componentes que contienen nitrógeno
CodY	**29**	Regula a los genes ComK y SrfA en presencia de casaminoácidos
PurR	**22**	Regula la transcripción de purinas, mediante la regulación dependiente del nucleótido adenina

o temperatura (67). Para acompañar su papel funcional, los miembros de la familia CRP-FNR tienen módulos sensoriales intrínsecos, lo cual permite la unión de moléculas efectoras alostéricas, o tienen grupos protéticos para la interacción con la señal. De esta forma, la adaptabilidad regulatoria y la flexibilidad estructural representada en los ortólogos de las proteínas CRP-FNR han permitido la evolución de un importante grupo de TFs versátiles fisiológicamente [ver ejemplos en el **ANEXO 8A**].

En *B. subtilis*, solamente los GRs CcpA, FUR y PhoP se presentan en genomas distantes filogenéticamente, sugiriendo un origen antiguo comparado a la distribución de sus otros GRs. Finalmente, FIS, ArcA y HNS en *E. coli* tienen una distribución restringida a Proteobacterias, mientras que AbrB, ComK, CodY y PurR en *B. subtilis* se encuentran restringidas a los linajes de Bacillales y Lactobacillales. De lo que se podría inferir que estos GRs divergieron después de la separación entre las Proteobacterias y Firmicutes.

La distribución filogenética de LRP se extiende hasta Archaea y hacemos una breve discusión al respecto. LRP es el único GR homodimérico[19] que se encuentra bien conservado a través de los linajes en Procariontes. Los homólogos de LRP han sido previamente identificados en procariontes exclusivamente (68). La amplia distribución filética de los homólogos de LRP entre Archaea y Bacteria sugiere que un regulador del tipo LRP estuvo presente en el último ancestro común (Last Common Ancester - LCA) de los dominios Bacteria y Archaea. No obstante, la distribución de reguladores del tipo LRP se ha visto que varían a través de los organismos (e.g. 20 copias en *Mesorhizobium loti*, 3 en *E. coli* y ninguna en las cepas de *Buchnera*, *Mycoplasma* y *Chlamydia*). Estos últimos son parásitos o endosimbiontes que dependen completamente de sus hospederos para el suplemento de amino ácidos y otros tipos de metabolitos. Estos organismos también tienen genomas reducidos, lo cual podrían explicar la ausencia de los miembros de la familia. A pesar de su conservación en varios *phyla*, aún en especies relacionadas cercanamente, su mecanismo de regulación global no se ha conservado, tal como se demuestra en el análisis del ortólogo LRP de *Haemophilus influenzae* (69) [ver **ANEXO 8B**].

Estas observaciones apuntan a la conclusión de que, aún en organismos del mismo *phyla*, no existen claras restricciones para mantener el mecanismo de regulación generalizada de los GRs, aún cuando los TFs estén conservados al nivel de secuencia (69) [ver **ANEXO 8C** donde se ejemplifica el caso de la sustitución de TFs (análogos) para llevar a cabo la regulación catabólica en *E. coli* (CRP) y en *B. subtilis* (CcpA)]. De esta forma, podemos inferir que el papel funcional y jerárquico de los TFs ortólogos en la TRN puede ser diferente o sustituible en cada organismo.

[19] **NOTA:** Recordemos que **IHF** es un regulador transcripcional global (GR) heterodimérico que requiere de la unión de dos subunidades para su estado activo en *E. coli*: **himA** y **himD** (21). En el presente estudio, hemos detectado la presencia de ambas subunidades diferencialmente, a partir del cual podemos puntualizar que no son similares sus patrones de distribución, incluso en algunas especies es opuesta, y la única subunidad que se detecta en Archaea es himA, mientras que himD se restringe a Bacteria.

3. Evolución de pares TF-TG en TRNs.

Se ha demostrado que los complejos de interacciones proteína-proteína, especialmente como pares, tienden a estar conservados o perdidos de una forma concertada a lo largo de diferentes proteomas (70-71). Dado que una interacción transcripcional involucra un TF y su TG, uno podría esperar que co-ocurran como pares a través del proceso evolutivo. De esta forma, implementamos una medida de distancia (D) como se describe en MÉTODOS-F para realizar un análisis de pares de co-ocurrencia (TF y su TG). De esta forma, cada par del conjunto de las interacciones regulatorias transcripcionales de *E. coli* y *B. subtilis* puede comportarse como una de la tres categorías siguientes: **a)** tanto un TF y su TG co-ocurren, **b)** cuando un TF ocurre en más especies que su TG y, **c)** cuando un TG ocurre en más especies que su TF [ver ejemplos en la Figura R3A].

Idealmente, si la interacción regulatoria está co-ocurriendo en diversas especies, uno esperaría que D_{TF} y D_{TG} tiendan a cero, pero por varias razones, tales como eventos de transferencia horizontal, pérdida o duplicación de genes, así como también posibles errores involucrados en la detección de ortólogos, uno obtendría sesgos en la distribución de pares TF-TG co-ocurriendo. Para evaluar la influencia de estos factores en la distribución de los ortólogos detectados, y para determinar un umbral (*threshold*) para la identificación de pares TF-TG co-relacionados, usamos pares de genes en rutas metabólicas de la base de datos del KEGG (42) como un control [ver MÉTODOS-G y **ANEXO 3A**]. Es bien conocido que genes en la misma ruta metabólica a menudo co-ocurren (60-61, 72-73). Basados en los umbrales determinados para cada genoma, identificamos a los pares TF-TG co-ocurriendo, e incluimos al resto de las interacciones dentro de una de las dos clases (TF > TG y TF < TG) basadas sobre el cálculo en que un D_{TF} es más alto o bajo que su D_{TG} [ver Figura R3B, MÉTODOS-G y **ANEXO 3B**].

La Tabla R1 de la Figura R3B muestra los Z-scores de conservación para cada categoría de pares TF-TG en ambos genomas, datos calculados sobre la comparación con las 1 000 TRNs generadas aleatoriamente [ver Figura R3-B, MÉTODOS-H y **ANEXO 4A-B**]. Puede observarse que hay una pequeña fracción de la TRN en ambos genomas que co-evoluciona a través de varios organismos. Sin embargo, la significancia de la co-ocurrencia desde la perspectiva de *B. subtilis* (p-value 0.0028) se observa relativamente más baja que en *E. coli* (p-value <0.0001) [ver Figura R3B-Tabla R1]. Tal observación debe ser evaluada en función de ciertos factores que no pueden verse reflejados directamente en el p-value, como el sesgo en el número de genomas secuenciados que conforman a Firmicutes comparado con el de Proteobacterias, así como la diferencia en el tamaño de las TRNs que son usadas, en cuyo caso es menor para la TRN de *B. subtilis.*

Una proporción igualmente aproximada de pares TF-TG se determinó en las categorías de TF > TG y TF < TG en ambos genomas [ver Figura R3B-Tabla R1]. Los Z-scores calculados en las categorías respectivas, sugieren que no existe una tendencia clara hacia la distribución de un TF > TG o hacia la de un TF < TG en ninguno de los dos genomas. Los Z-scores en cada caso corresponden a no más de 3-4 desviaciones estándar, excepto para el caso de las distribuciones TF < TG en *E. coli* (Z-score 3.68) [ver

Figura R3B-Tabla R1]. Esto demuestra que existe una restricción de co-ocurrencia para la mayoría de las interacciones entre un par TF y su TG.

Es importante hacer notar que la perspectiva de este análisis es diferente con respecto al análisis cuantitativo de la distribución de los elementos individuales (TFs y TGs) de la TRN de referencia. Esta perspectiva se localiza sobre el siguiente nivel de la evolución regulatoria, las interacciones. Específicamente, trata del entendimiento sobre la co-ocurrencia de los componentes que la conforman. De esta forma, las TRNs estudiadas aquí revelan que los pares de interacciones TF-TG no co-ocurren preferencialmente, en contraste a las interacciones del tipo proteína-proteína. Con excepción a una pequeña fracción significativa (que representa el 2%) de pares que co-ocurren en las TRNs analizadas (15 interacciones en ambos genomas), el análisis presente demuestra que las fuerzas de la selección natural podrían actuar de forma relativamente independiente para retener un TF y su TG. Es decir, los componentes de la interacción regulatoria transcripcional se pierden o ganan independientemente.

4. Conservación de los regulones e interacciones conservadas de las TRNs a través de las especies bacterianas.

La regulación de la actividad celular difícilmente está determinada por la participación de interacciones aisladas. La TRN tiende a estar interconectada y organizada en módulos funcionales. A su vez, existen muy pocos módulos que se encuentran separados enteramente del resto de la red (12, 30, 74). En los módulos funcionales existen patrones específicos de inter-regulación entre un TF y sus TGs para responder a diferentes condiciones y, según el enfoque de la organización de la TRN, éstos pueden ser reconocidos a distintos niveles. Derivado de los argumentos anteriores, esperamos que los grupos de interacciones caracterizadas al nivel de regulones (grupos definidos por el número total de TGs que son regulados por un solo TF) se encuentren conservados en un amplio espectro filogenético. Por otro lado, estamos interesados en conocer si existen ciertos procesos celulares que hayan mantenido su regulación transcripcional a partir de los mismos módulos funcionales desde las etapas tempranas de la divergencia bacteriana.

En la Figura R4, mostramos la conservación de interacciones regulatorias al nivel de regulones tanto para *E. coli* como para *B. subtilis* en los 110 genomas completos no redundantes. Dos enfoques de conservación son generados a partir de un análisis de agrupamiento (*clustering*) de regulones de las TRNs de referencia:

1) Un **agrupamiento horizontal** que representa la extensión de la TRN de referencia en cada uno de los 110 genomas, a partir del total de regulones que la componen [ver MÉTODOS-I].
2) Un **agrupamiento vertical** que representa la distribución de cada uno de los regulones a través de los 110 genomas [ver MÉTODOS-I].

En general, puede observarse que las TRNs de *E. coli* [ver Figura R4A] y de *B. subtilis* [ver Figura R4B] comparten un número limitado de regulones a través de los genomas analizados, aunque la conservación es mayor en linajes cercanos filogenéticamente. Comparamos la distribución de los genomas en el agrupamiento horizontal con la basada en la relación filogenética generada por el método de Brown *et al.* (58) y encontramos que varios linajes fueron agrupados apropiadamente en función de los porcentajes de conservación de los regulones.

Desde el agrupamiento horizontal, es interesante notar que los clados más cercanos a *E. coli* [ver Figura R4A], los cuales incluyen varias Proteobacterias, comparten alrededor del 40% de las interacciones regulatorias transcripcionales desde *E. coli*, excepto para organismos parásitos y endosimbiontes, los cuales fueron agrupados juntos y muestran una escasa conservación de la TRN. Las proteobacteria *Blochmannia floridanus*, el mollicute *Mesoplasma florum* y *Ureaplasma urealyticum*, las arqueas *Methanopyrus kandleri, Methanococcus jannaschi, Methanobacterium thermoautotropicum, Halobacterium* sp, *Pyrococcus furiosus, Aeropyrum pernix, Nanoarchaeum equitans* y los 10 organismos eucariontes analizados no muestran ninguna interacción regulatoria del tipo *E. coli*. Desde la perspectiva de *B. subtilis* [ver Figura R4B] los clados más cercanos comparten alrededor del 30% de las interacciones regulatorias

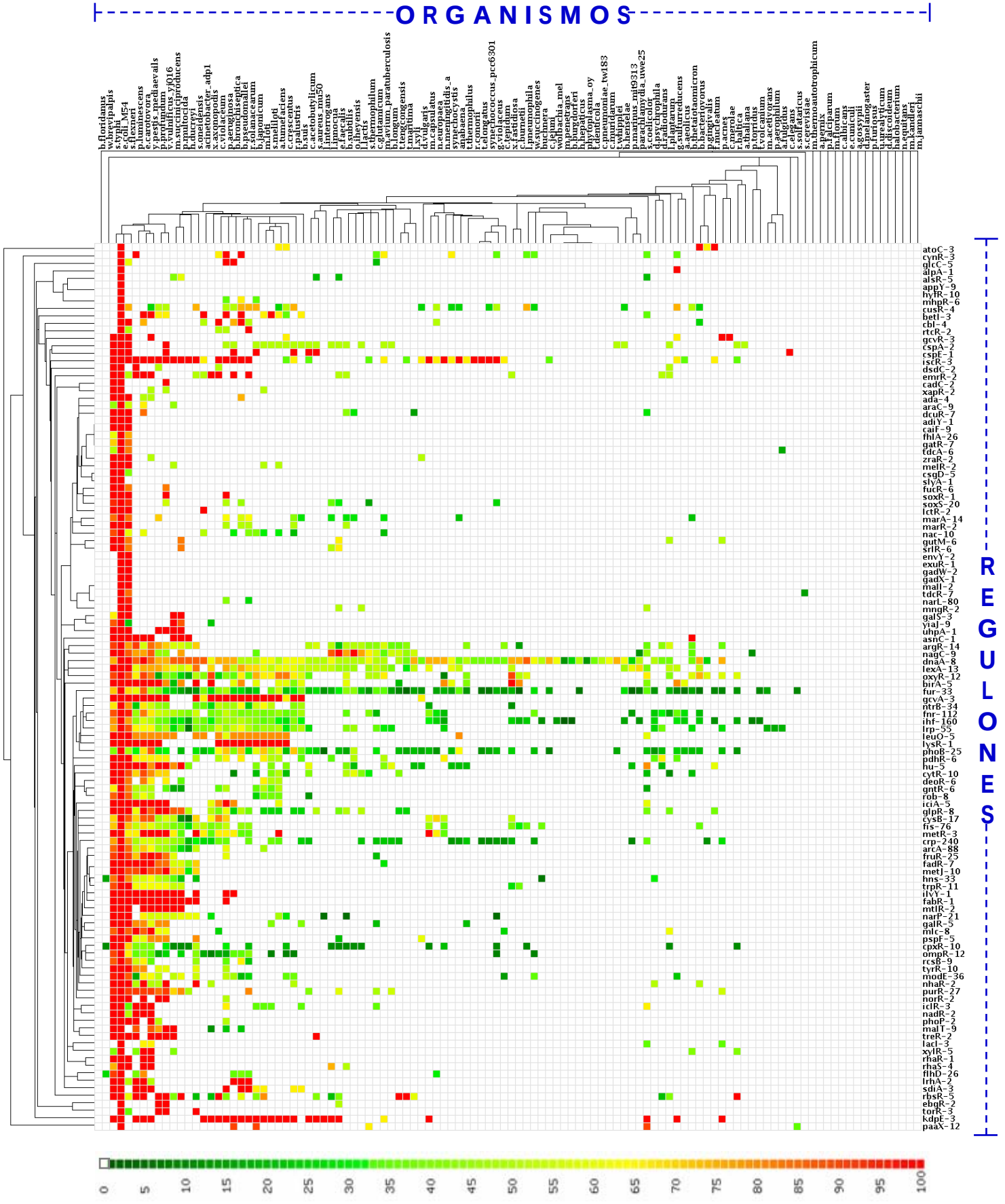


Figura R-4A. Conservación de regulones (118 en total) de la TRN de *Escherichia coli* K12 a través de 110 genomas completes. Se realizaron dos agrupamientos basados en los porcentajes de interacciones conservadas por cada regulón. El primer agrupamiento, que aquí denominamos **horizontal**, corresponde a la extensión de la TRN de referencia a través de los genomas analizados. Un segundo agrupamiento, denominado **vertical**, corresponde a la conservación que presentan los diferentes regulones que componen a la TRN a través de los genomas. Una barra de correspondencia en colores (abajo) representa el espectro de los porcentajes de interacciones detectadas para cada regulón por genoma (de 0% en blanco al 100% en rojo) [ver MÉTODOS-I y **ANEXO 9A**].

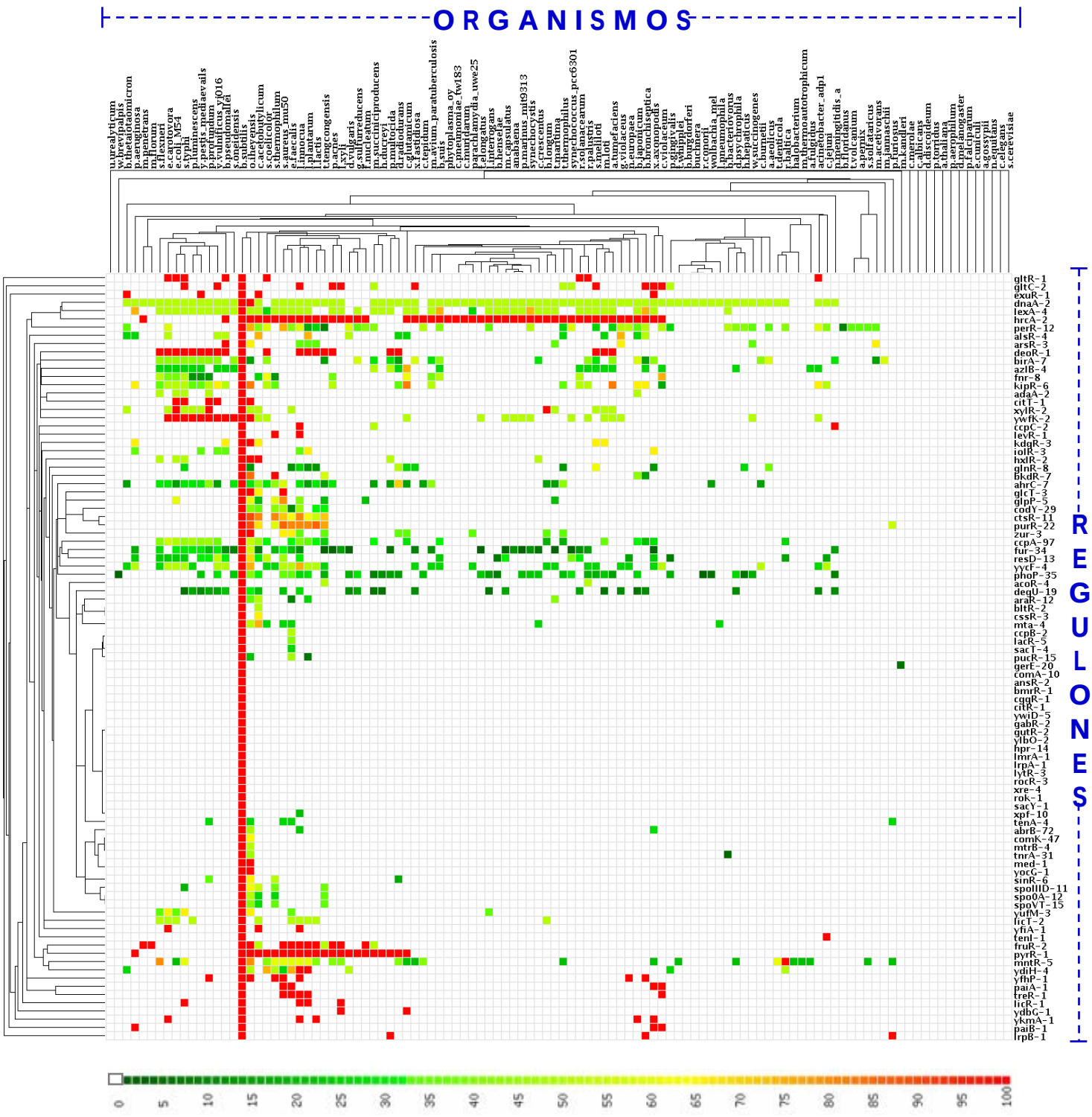


Figura R-4B. Conservación de regulones (93 en total) de la TRN de *Bacillus subtilis* a través de 110 genomas completes. Se realizaron dos agrupamientos basados en los porcentajes de interacciones conservadas por cada regulón. El primer agrupamiento, que aquí denominamos **horizontal**, corresponde a la extensión de la TRN de referencia a través de los genomas analizados. Un segundo agrupamiento, denominado **vertical**, corresponde a la conservación que presentan los diferentes regulones que componen a la TRN a través de los organismos. Una barra de correspondencia en colores (abajo) representa el espectro de los porcentajes de interacciones detectadas para cada regulón por genoma (de 0% en blanco al 100% en rojo) [ver MÉTODOS-I y **ANEXO 9B**].

transcripcionales, excepto algunos parásitos y endosimbiontes de los linajes Bacillus y Lactobacillus. Los mollicutes *U. urealyticum*, las arqueas *Picrophilus torridus, Pyrobaculum aerophilum, N. equitans* y los 10 organismos eucariontes analizados no muestran ninguna interacción regulatoria del tipo *B. subtilis*.

Las observaciones anteriores sugieren que pese a la escasa conservación de la TRNs, la regulación transcripcional refleja la historia evolutiva de la segregación de los principales grupos bacterianos, tal como se prueba con el uso de la distancia filogenética. Y por otro lado, que el patrón de conservación de las TRNs a través del espectro filogenético, provee una medida de distinción entre los estilos generales de vida, tales como organismos parásitos, simbiontes y de vida libre, así como también las diferencias que existen en la regulación transcripcional entre Bacteria y Archaea (REF).

En la agrupación vertical se puede observar que una fracción alta de los regulones de la TRN de *E. coli* [ver Figura R4A] se encuentran compartidos en un gran grupo de organismos, los cuales subyacen bajo TFs metabólicos y estructurales del tipo IscR, ArgR, AsnC, BirA, Crp, DnaA, Fnr, Fur, GlpR, Ihf, LexA, Lrp, NagC, OxyR, OmpR, PhoB, KdpE y RbsR [ver la **información disponible online** sobre la función detallada de cada uno de estos TFs y sus regulones]. Aún cuando estos regulones tengan un alto porcentaje de sus interacciones, ello no implica que las interacciones sean las mismas en los organismos donde se identifican. De esta misma forma, un regulón que posee un bajo porcentaje de sus interacciones en otros genomas, podría implicar que si bien el regulón no está muy conservado, es posible que las pocas interacciones que se mantienen en otros genomas sean las mismas.

Derivado de este razonamiento, se llevó a cabo otra estrategia para detectar las interacciones de regulación transcripcional conservadas en los organismos bacterianos. El análisis elaborado por el método DOLLOP y en base al principio de Parsimonia [ver MÉTODOS-J] sobre las predicciones de las TRNs desde *E. coli*, reveló que existe solamente un pequeño grupo de interacciones conservadas comunes entre diferentes *phyla* bacterianos, los cuales representan alrededor del 6% de la TRN de *E. coli*. Estas interacciones regulan importantes procesos celulares en *E. coli*, tales como síntesis de arginina, asparagina, biotina y ribosa, transporte de aminoácidos y hierro, disponibilidad de fosfato, procesos de replicación y el sistema de respuesta SOS, [ver **ANEXO 5** para más detalles sobre estas interacciones].

Por su parte, los regulones altamente conservados desde *B. subtilis* incluyen reguladores metabólicos y estructurales del tipo DnaA, LexA, HrcA, PerR, BirA, AzlB, YwfK, AhrC (ArgR), CcpA, FUR, ResD, YycF, PhoP, DegU y MntR [ver la **información disponible online** sobre la función detallada de cada uno de estos TFs y sus regulones]. Entre estos regulones conservados hay, al menos, dos TFs hipotéticos: YwfK y YycF cuya función es aún desconocida. Los patrones de conservación aquí generados podrían ser usados para entender y caracterizar estos TFs a través de la combinación de métodos experimentales y computacionales que puedan apoyar en la determinación de su función, tales como el contexto funcional de sus TGs, sus ortólogos en otros organismos o el análisis de sus bsDNA a través de *phylogenetic footprinting* de las regiones río arriba de los TGs en genomas cercanos.

Con la misma estrategia que para la TRN de *E. coli* [ver MÉTODOS-J], se detectó una menor proporción de interacciones conservadas entre diferentes *phyla* bacterianos desde la TRN de *B. subtilis*. Las

interacciones conservadas de la TRN de *B. subtilis* están involucradas de forma similar que en *E. coli* en la síntesis de arginina y biotina, transporte de managaneso, disponibilidad de fosfato, genes de respuesta al estrés por calor, funciones regulatorias globales, procesos de replicación y el sistema de respuesta SOS [ver **ANEXO 5** para más detalles sobre estas interacciones].

La conservación de algunos de los regulones detectados aquí, tales como ArgR, BirA y LexA han sido previamente reportados en varios *phyla* (32-33, 75). Sin embargo, el repertorio de regulones conservados que se detectan en este estudio, podría aumentar nuestro entendimiento en la conservación de específicos módulos funcionales, lo cual podría guiar a futuros estudios experimentales para caracterizarlos.

De esta forma, el análisis de conservación elaborado por grupos de interacciones, que a través del concepto de regulón pueden entenderse como un módulo funcional en la red, nos revela que la evolución de las TRNs ha experimentado grandes cambios, si bien no sabemos si también en su estructura, al menos sí en sus interacciones. Por otro lado, existen pocos pero bien conservadas interacciones que han regulado procesos estructurales, principalmente, desde etapas tempranas de la diversificación bacteriana.

DISCUSIÓN GENERAL Y CONCLUSIONES

La complejidad de las redes de regulación transcripcional en los organismos bacterianos es ampliamente afectada por su adaptación al estrés ambiental que cambia dinámicamente en los nichos ecológicos. Por ejemplo, las bacterias entéricas, bacterias de suelo y otras bacterias de vida libre viven en ambientes complejos y tienen subsistemas de respuestas de censado a señales metabólicas y extracelulares (76). En contraste, el limitado espectro ecológico y la disgregación frecuente de la población en los patógenos obligados y simbiontes han resultado en un incremento en las tasas de deriva genética y restricciones selectivas que limitan las funciones en el censado extracelular, así como el número de los genes que utilizan para responder a estrés intracelular (77-79). Nuestros resultados indican que los estilos de vida de endosimbiontes así como de parásitos estrictos comparten solamente alrededor del 10% de los componentes ortólogos, y menos del 2% de las interacciones que conforman las TRNs de *E. coli* y *B. subtilis*. La pérdida de elementos regulatorios puede reflejar una relativa constancia intracelular en el ambiente del hospedero, lo cual permite a estos organismos tener una estructura regulatoria simplificada y tentativamente poco diversa (76, 80). Acorde a nuestros resultados, la pérdida de muchos ortólogos de los factores de transcripción, y no así de los genes blancos, podría ser la principal causa de estos cambios dramáticos en la TRN, como puede verse desde el escenario de los reguladores globales de *E. coli* y *B. subtilis*, que tienen una distribución limitada aún cuando modulan directamente la expresión de aproximadamente el 51% de los genes en *E. coli* (19).

Encontramos que la conservación de las redes de regulación transcripcional de *E. coli* y *B. subtilis* (en sus diversos niveles de organización) subyace a *grosso modo* a la historia evolutiva de las especies, y que se ve reflejada en su distancia filogenética y estilo de vida (29). Nuestros resultados muestran que los factores de transcripción de una especie particular, incluso entre especies cercanas, se encuentran menos conservados que los genes blancos conforme se incrementa la distancia filogenética. Esto sugiere que la regulación transcripcional de los genes cambia más rápido en la evolución que los genes en sí. Relacionado con esto, Maslov *et al*. (81) encontraron que la tasa de diferenciación evolutiva de las interacciones es mucho más rápida que la los genes blancos y que la de sus interacciones de proteína. Adicionalmente, nuestro análisis de la co-ocurrencia de pares de interacciones regulatorias a través de los genomas indica tendencias diferentes en la conservación de pares TF-TG, mostrando significativamente que la tendencia de co-ocurrencia entre un TF y su TG es casi nula, es decir, que ambos componentes evolucionan de forma independiente en la estructura de las TRNs. No obstante esta tendencia, es claro que:

a) Cuando un TF está conservado en diferentes especies sin su correspondiente TG, esto podría implicar que el TF es requerido en la regulación de un diferente grupo de TGs que aquellos reportados en el genoma de referencia.

b) Cuando un TG es conservado y su TF se pierde, esto podría implicar que el TG es regulado por un factor análogo, homólogo o por otros tipos de regulación transcripcional, tales como riboswitches y RNAs pequeños.

c) Cuando un TF y su TG(s) se conservan en el genoma de interés, no necesariamente éstos interactúan.

Ésto sugiere que el nivel de flexibilidad que las TRNs pueden imponer sobre la evolución de la regulación transcripcional de los organismos en diferentes ambientes es alto. Las razones y mecanismos evolutivos para las tendencias observadas en el presente trabajo necesitan ser analizadas en líneas de investigación derivadas de este estudio.

A pesar de la escasa conservación de las interacciones regulatorias a través de las especies, ciertas interacciones individuales han sido bien conservadas a través de diferentes *phyla* eubacterianos. Estas interacciones regulan procesos transcripcionales esenciales en Bacteria. De hecho, muchos de estos procesos han sido bien caracterizados y están relacionados directamente o indirectamente a la maquinaria estructural, transcripcional y traduccional de la célula, explicando la razón de su conservación en amplias distancias filogenéticas. Aún cuando el tipo de regulación (represor o activador) y que el sitio de unión al DNA pueda cambiar a través de los genomas, es razonable pensar que es importante mantener la regulación de estos procesos comunes con los mismos elementos, como en el caso de los reguladores BirA y DnaA, los cuales se asumen como un resultado de la ancestría común en todas las bacterias.

La red de regulación transcripcional parece evolucionar de manera puntual, con pérdida, recambio y ganancia de elementos específicos de las interacciones individuales. Probablemente esta evolución puntual juega un papel más importante que la pérdida y la ganancia de motivos y módulos completos de interacciones. Como Teichmann y Babu (82) reportaron previamente, muchas redes se han originado por evolución convergente y no por duplicación genética de circuitos ancestrales (83). De esta forma, con la excepción de una pequeña fracción de la TRN, puede inferirse que grandes proporciones de las TRNs han evolucionado en los organismos a través de extensos cambios y re-conexiones entre sus componentes.

Desde nuestro análisis comparativo, demostramos que los elementos individuales, los pares interactuantes y los grupos de interacciones (regulones) de la regulación transcripcional no están conservados, incluso en especies relacionadas cercanamente. Esto refleja que en cada evento de especiación, en el que se requiere de la adaptación a nuevas condiciones intra y extracelulares, la regulación transcripcional es más flexible que el componente genético de los organismos para la adaptación fenotípica. De esta forma, este trabajo puede proveer una perspectiva de la flexibilidad de la red de regulación transcripcional en bacterias, lo cual contribuye al entendimiento de las bases transcripcionales de la variación natural.

REFERENCES

1. The American Heritage® Dictionary of the English Language. 4th Edition. 2000.

2. Kirschner M. and Gerhart J. (1998) **Evolvability.** *Proc. Natl. Acad. Sci.* 95:8420-8427.

3. Alberts B., et al. **Molecular Biology of the Cell.** 4th Edition. 2002. Garland Science. New York, USA.

4. Savageau M.A. (1991) **Reconstructionist molecular biology.** *New Biol.* 3, 190-7.

5. Savageau M.A. (2001) **Design principles for elementary gene circuits: Elements, methods, and examples.** *Chaos* 11, 142-159.

6. Kauffman S.A. (1973) **Control circuits for determination and transdetermination.** *Science* 181, 310-8.

7. Bagley R.J., Farmer J.D., Kauffman S.A., Packard N.H., Perelson A.S., Stadnyk I.M. (1989) **Modeling adaptive biological systems.** *Biosystems* 23, 113-37.

8. Fell, D.A. and Wagner, A. (2000) The small world of metabolism. Nat Biotechnol, 18, 1121-1122.

9. Ouzounis, C.A. and Karp, P.D. (2000) **Global properties of the metabolic map of *Escherichia coli*.** *Genome Res*, 10, 568-576.

10. Thieffry, D., Huerta, A.M., Perez-Rueda, E. and Collado-Vides, J. (1998) **From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*.** *Bioessays*, 20, 433-440.

11. Guelzim N., Bottani S., Bourgine P. and Kepes F. (2002) **Topological and causal structure of the yeast transcriptional regulatory network.** *Nat Genet*, 31, 60-63

12. Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) **Network motifs in the transcriptional regulation network of *Escherichia coli***. *Nat Genet*, 31, 64-68.

13. Oltvai Z.N. and Barabási A.L. (2002) **Systems biology. Life's complexity pyramid.** *Science* 298, 763-4.

14. Lewin, B. ***Genes VII*.** 2000. Oxford University Press. New York, USA.

15. Browning D.F. and Busby S.J.W. (2004) **The regulation of bacterial transcription initiation**. *Nat Rev Microbiol*, 2, 57-65.

16. Neidhardt F.C., et al. ***Escherichia coli* and *Salmonella*: Cellular and Molecular Biology.** Vol. 2, 2th Edition. 1996. ASM Press. Washington D.C., USA.

17. Teichmann, S.A. and Babu, M.M. (2004) **Gene regulatory network growth by duplication**. *Nat Genet*, 36, 492-496.

18. Albert R. and Albert-László B. (2002) **Statistical mechanics of complex networks**. *Rev Modern Phys*, 34, 47-97.

19. Martinez-Antonio, A. and Collado-Vides, J. (2003) **Identifying global regulators in transcriptional regulatory networks in bacteria.** *Curr Opin Microbiol*, 6, 482-489.

20. Herrgard, M.J., Covert, M.W. and Palsson, B.O. (2004) **Reconstruction of microbial transcriptional regulatory networks**. *Curr Opin Biotechnol*, 15, 70-77.

21. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C. *et al.* (2004) **RegulonDB (version 4.0): transcriptional regulation, operon organization and growth conditions in *Escherichia coli* K-12.** *Nucleic Acids Res*, 32, D303-306.

22. Barabasi A.L. and Albert R. (1999) **Emergence of scaling in random networks.** *Science*, 286, 509-512.

23. Brown T.A. **Genomes.** 1th Edition. 1999. Wiley-Liss Press. New York, USA.

24. Paillard G. and Lavery R. (2004) **Analyzing protein-DNA recognition mechanisms.** *Structure*, 12, 113-22.

25. Ahmad S., Gromiha M.M., Sarai A. (2004) **Analysis and prediction of DNA-binding proteins and their binding residues based on composition, sequence and structural information.** *Bioinformatics*,20, 477-86.

26. Zhang Y., Xi Z., Hegde R.S., Shakked Z., Crothers D.M. (2004) **Predicting indirect readout effects in protein-DNA interactions.** *Proc Natl Acad Sci U S A*, 101, 8337-41.

27. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) **Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs**. *Genome Res*, 14, 1107-1118.

28. Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) **Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores.** *J Mol Biol*, 297, 233-249.

29. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. (2004) **Structure and evolution of transcriptional regulatory networks**. *Curr Opin Struct Biol*, 14, 283-291.

30. Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U. and Margalit, H. (2004) **Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction**. *Proc Natl Acad Sci U S A*, 101, 5934-5939.

31. Fitch, W.M. (1970) **Distinguishing homologous from analogous proteins**. *Syst Zool*, 19, 99-113.

32. Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2002) **Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea.** *Genome Res*, 12, 1507-1516.

33. Erill, I., Jara, M., Salvador, N., Escribano, M., Campoy, S. and Barbe, J. (2004) **Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics**. *Nucleic Acids Res*, 32, 6617-6626.

34. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. (2005) **Conserved patterns of protein interaction in multiple species.** *Proc Natl Acad Sci U S A*, 102, 1974-1979.

35. Struhl, K. (1999) **Fundamentally different logic of gene regulation in eukaryotes and prokaryotes.** *Cell*, 98, 1-4.

36. Su Z., Olman V., Mao F. and Xu Y. (2005) **Comparative genomics analysis of NtcA regulons in cyanobacteria: regulation of nitrogen assimilation and its coupling to photosynthesis.** *Nucleic Acids Res*, 33, 5156–5171

37. Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) **DBTBS: database of transcriptional regulation in *Bacillus subtilis* and its contribution to comparative genomics.** *Nucleic Acids Res*, 32, D75-77.

38. Blattner F.R., *et al.* (1997) **The complete genome sequence of *Escherichia coli* K-12.** *Science*, 277, 1453-74.

39. Kunst F., *et al.* (1997) **The complete genome sequence of the gram-positive bacterium *Bacillus subtilis*.** *Nature*, 390:249-56.

40. Evans L., Feucht A. and Errington J. (2004) **Genetic analysis of the *Bacillus subtilis sigG* promoter, which controls the sporulation-specific transcription factor s^G.** *Microbiology* 150, 2277-2287.

41 Feucht A., Evans L., and Errington J. (2003**) Identification of sporulation genes by genome-wide analysis of the s^E regulon of *Bacillus subtilis*.** *Microbiology* 149, 3023-3034.

42. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) **The KEGG resource for deciphering the genome.** *Nucleic Acids Res*, 32, D277-280.

43. Teichmann, S.A., Murzin, A.G. and Chothia, C. (2001) **Determination of protein function, evolution and interactions by structural genomics.** *Curr Opin Struct Biol*, 11, 354-363.

44. Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) **Evolution of the protein repertoire.** *Science*, 300, 1701-1703.

45. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998) **Predicting function: from genes to genomes and back.** *J Mol Biol*, 283, 707-725.

46. Lan, N., Montelione, G.T. and Gerstein, M. (2003) **Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level.** *Curr Opin Chem Biol*, 7, 44-54.

47. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. and Gish, W. (2003) **WU-Blast2 server at the European Bioinformatics Institute.** *Nucleic Acids Res*, 31, 3795-3798.

48. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) **The Pfam protein families database.** *Nucleic Acids Res*, 28, 263-266.

49. Sonnhammer, E.L. and Koonin, E.V. (2002) **Orthology, paralogy and proposed classification for paralog subtypes.** *Trends Genet*, 18, 619-620.

50. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) **A genomic perspective on protein families**. *Science*, 278, 631-637.

51. Huynen, M.A. and Bork, P. (1998) **Measuring genome evolution**. *Proc Natl Acad Sci U S A*, 95, 5849-5856.

52. Janga, S.C. and Moreno-Hagelsieb, G. (2004) **Conservation of adjacency as evidence of paralogous operons**. *Nucleic Acids Res*, 32, 5392-5397.

53. Li, W., Jaroszewski, L. and Godzik, A. (2002) **Tolerating some redundancy significantly speeds up clustering of large protein databases.** *Bioinformatics*, 18, 77-82.

54. Hegyi, H. and Gerstein, M. (2001) **Annotation transfer for genomics: measuring functional divergence in multi-domain proteins.** *Genome Res*, 11, 1632-1640.

55. Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A. and Weiner, J., 3rd. (2005) **The evolution of domain arrangements in proteins and interaction networks.** *Cell Mol Life Sci*, 62, 435-445.

56. Wheelan, S.J., Marchler-Bauer, A. and Bryant, S.H. (2000) **Domain size distributions can predict domain boundaries.** *Bioinformatics*, 16, 613-618.

57. Eddy, S.R. (1996) **Hidden Markov models.** *Curr Opin Struct Biol*, 6, 361-365.

58. Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J. (2001) **Universal trees based on large combined protein sequence data sets.** *Nat Genet*, 28, 281-285.

59. Yang, S., Doolittle, R.F. and Bourne, P.E. (2005) **Phylogeny determined by protein domain content.** *Proc Natl Acad Sci U S A*, 102, 373-378.

60. Forst C.V. and Schulten K. (2001) **Phylogenetic Analysis of Metabolic Pathways.** *J Mol Evol*, 52, 471–489

61. Pinter R.Y., Rokhlenko O., Yeger-Lotem E. and Ziv-Ukelson M. (2005) **Alignment of metabolic pathways.** *Bioinformatics*, 21, 3401-3408.

62. Milo R., Shen-Orr S., Itzkovitz S., Kashtan N., Chklovskii D. and Alon U. (2002) **Network motifs: simple building blocks of complex networks.** *Science*, 298, 824-827.

63. de Hoon, M.J., Imoto, S., Nolan, J. and Miyano, S. (2004) **Open source clustering software**. *Bioinformatics*, 20, 1453-1454.

64. Felsentein J. (1993). **PHYLIP (Phylogeny interface package)** version 3.6a2. Distributed by the author. Department of Genetics, University of Washington, Seattle.

65. Coulson, R.M., Enright, A.J. and Ouzounis, C.A. (2001) **Transcription-associated protein families are primarily taxon-specific.** *Bioinformatics*, 17, 95-97.

66. Hausladen A. and Stambler I.S. (1999) **Nitrosative Stress.** *Methods Enzymol*, 300, 389-395.

67. Körner H., Sofía H.J. and Zumft W.G. (2003) **Phylogeny of the bacterial superfamily of Crp-Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs.** *FEMS Microbiology Reviews*, 27, 559-592.

68. Brinkman, A.B., Ettema, T.J., de Vos, W.M. and van der Oost, J. (2003) **The Lrp family of transcriptional regulators.** *Mol Microbiol*, 48, 287-294.

69. Friedberg, D., Midkiff, M. and Calvo, J.M. (2001) **Global versus local regulatory roles for Lrp-related proteins: Haemophilus influenzae as a case study.** *J Bacteriol*, 183, 4004-4011.

70. Pagel, P., Mewes, H. W. and Frishman, D. (2004) **Conservation of protein-protein interactions - lessons from ascomycota.** *Trends Genet*, 20, 72-6.

71. Aravind, L., Watanabe, H., Lipman, D. J. & Koonin, E. V. (2000). **Lineage-specific loss and divergence of functionally linked genes in eukaryotes.** *Proc Natl Acad Sci U S A*, 97, 11319-11324.

72. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) **The large-scale organization of metabolic networks.** *Nature*, 407, 651-654.

73. Date, S.V. and Marcotte, E.M. (2003) **Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages.** *Nat Biotechnol*, 21, 1055-1062.

74. Mazurie, A., Bottani, S. and Vergassola, M. (2005) **An evolutionary and functional assessment of regulatory network motifs.** *Genome Biol*, 6, R35.

75. Makarova, K.S., Mironov, A.A. and Gelfand, M.S. (2001) **Conservation of the binding site for the arginine repressor in all bacterial lineages.** *Genome Biol*, 2, RESEARCH0013.

76. Cases, I., de Lorenzo, V. and Ouzounis, C.A. (2003) **Transcription regulation and environmental adaptation in bacteria.** *Trends Microbiol*, 11, 248-253.

77. Moran, N.A. (1996) **Accelerated evolution and Muller's rachet in endosymbiotic bacteria.** *Proc Natl Acad Sci U S A*, 93, 2873-2878.

78. Itoh, T., Martin, W. and Nei, M. (2002) **Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts.** *Proc Natl Acad Sci U S A*, 99, 12944-12948.

79. Andersson, S.G. and Kurland, C.G. (1998) **Reductive evolution of resident genomes.** *Trends Microbiol*, 6, 263-268.

80. Wilcox, J.L., Dunbar, H.E., Wolfinger, R.D. and Moran, N.A. (2003) **Consequences of reductive evolution for gene expression in an obligate endosymbiont.** *Mol Microbiol*, 48, 1491-1500.

81. Maslov, S., Sneppen, K., Eriksen, K.A. and Yan, K.K. (2004) **Upstream plasticity and downstream robustness in evolution of molecular networks.** *BMC Evol Biol*, 4, 9.

82. Teichmann, S.A. and Babu, M.M. (2004) **Gene regulatory network growth by duplication.** *Nat Genet*, 36, 492-496.

83. Conant G.C. and Wagner A. (2003) **Convergent evolution of gene circuits.** *Nat Genet*,34, 264-6.

Bacterial regulatory networks are extremely flexible in evolution

Irma Lozada-Chávez, Sarath Chandra Janga and Julio Collado-Vides

Programa de Geonómica Computacional, Centro de Ciencias Geonómicas, Universidad Nacional Autónoma de México, Apdo. Postal 565-A, Av. Universidad, Cuernavaca, Morelos, 62100 México.

Addresses for correspondence: ilozada@ccg.unam.mx, sarath@ccg.unam.mx

Keywords: transcriptional regulatory network, conservation, regulon, prokaryotes, regulog, interactions

ABSTRACT

Over millions of years the structure and complexity of the transcriptional regulatory network (TRN) in bacteria has changed, reorganized and enabled them to adapt to almost every environmental niche on earth. In order to understand the plasticity of TRNs in bacteria, we studied the conservation of currently known TRNs of the two model organisms *Escherichia coli K12* and *Bacillus subtilis* across complete genomes including Bacteria, Archaea and Eukarya at three different levels: individual components of the TRN, pairs of interactions and regulons. We found that transcription factors (TFs) evolve much faster than the target genes (TGs) across phyla. We show that global regulators are poorly conserved across the phylogenetic spectrum and hence transcription factors could be the major players responsible for the plasticity and evolvability of the TRNs. We also found that there is only a small fraction of significantly conserved transcriptional regulatory interactions among different phyla of bacteria and that there is no constraint on the elements of the interaction to co-evolve. Finally our results suggest that majority of the regulons in bacteria are rapidly lost implying a high order flexibility in the TRNs. We hypothesize that during the divergence of bacteria certain essential cellular processes like the synthesis of arginine, biotine and ribose, transport of amino acids and iron, availability of phosphate, replication process and the SOS response are well conserved in evolution. From our comparative analysis, it is possible to infer that transcriptional regulation is more flexible than the genetic component of the organisms and its complexity and structure plays an important role in the phenotypic adaptation.

INTRODUCTION

Evolution is the result of variation and selection of the components and structure of organisms through time. Transcriptional regulation plays a prominent role in the expression of genetic information. Its primary role in microbial organisms is controlling the response to environmental changes, such as nutritional status and several stresses. An important idea emerging in post-genomic biology is that transcriptional regulation can be viewed as a complex network of interactions among diverse types of molecules like proteins, DNA and metabolites (1-4). In this work we try to assess the evolution of the structure and plasticity of the transcriptional regulatory network (TRN) across species at three distinct levels: individual components of the TRN, pairs of regulatory interactions and regulons (A regulon is defined as the group of all genes regulated by a transcription factor), through a comparative analysis of their conservation.

The basic unit of gene regulatory interaction consists of three components: a transcription factor (TF), its DNA binding site (operator) and the target gene (TG). Topologically, the TRN is complex because genes may be regulated by more than one TF and some TFs may control more than one gene through DNA binding site(s) (5-7). The TRN comprises a significant proportion of the genome in each organism and it constitutes a major component of the genetic basis for the evolution of diverse aspects of bacterial phenotypes. It is important to learn how the TRN evolves as it would enable us to study the molecular evolutionary ecology of regulatory diversification by examining both the extent and pattern of regulatory gene diversity, the phenotypic effects of molecular variation and their ecological consequences.

It is also important to recognize that, although abundant sequence data and complete genomes are available, the experimental determination of TRNs has been limited to a few organisms even in prokaryotes. Besides, there is no clear relationship between the presence of a TF, its TG and DNA binding site(s), and their structural and biochemical characteristics that could have been transferred between genomes. It is also difficult to evaluate a specific measure between sequence homology, function and interaction transfer for any two proteins involved

in a regulatory interaction (6,8,9). However, several groups have recently examined the transfer of regulatory interaction annotations from one organism to another using comparative genomic approaches (9,10). The transfer of such interactions involves assigning functional roles to TFs and TGs, based on protein sequence similarity and on the conservation of topological patterns of the TRN, such as motifs and modules (8,11).

The *Regulog* approach uses cross-species data to predict DNA-protein interactions across genomes. A TF and TG interaction in one species is predicted to occur in another species if their best sequence matches have been determined in the target group of genomes. The presence of just one of the components of the regulatory interaction is not enough to transfer the interaction annotation, it is necessary that both TF and its TG(s) are detected in another organism. Using this approach, Yu *et al.* (9) have shown that orthologous TFs and TGs of *Saccharomyces cerevisiae* and *Drosophila melanogaster* tend to share the same regulatory interaction if the eukaryotic TFs have minimal sequence identities of 30% to 60% depending on the protein family. More recently, Sharan *et al.* (12) associated functions to proteins using network-level conservation of protein-protein interactions in eukaryotic genomes. This implies that high sequence similarity does not necessarily mean that the function is conserved; but conservation at the level of network modules allows more confident function determination from the context. Therefore, the best matches are not always present within conserved protein clusters enforcing the notion that it is advantageous to increase the detection of conserved functions by including paralogous family expansion and contraction, and even gene loss. The high specificity of the predictions attained by Sharan *et al.* can be maintained because conservation is evaluated in the context of a protein interaction subnetwork and not independently for each interaction. However, it has been shown that the patterns of conservation between protein-protein interactions *versus* protein-DNA interactions is different (9), and that the transcriptional regulatory logic differs radically between Eukaryotes and Prokaryotes (13). As a consequence, the performance of transcriptional interaction mapping methods cannot be currently assessed at a large scale (7,9).

Given the increasing number of sequenced genomes, it is possible and quite important to have a broader perspective of the evolution of TRNs by mapping the changes in the components of the regulatory interactions, which might differ from the common reconstruction of the metabolic, structural and some transcriptional histories of the organisms. Understanding the evolution of TRNs will not only improve our insight over the biological constraints different organisms have acquired over time but also enable us to decipher the basic design principles underlying them. Besides, one can reconstruct a regulatory history from the core of the transcriptional regulatory interactions that have been shared in the cellular processes of bacteria.

We used the TRNs of two different model Bacteria. One of these is the TRN of the gram-negative bacterium *Escherichia coli* K12 contained in RegulonDB, which is probably the best known in bacteria (14). This database contains experimental information corresponding to nearly 20% of the TRN of *E. coli* (5). The second best studied Prokaryote in terms of transcriptional regulation is the gram-positive *Bacillus subtilis.* We obtained the complete set of regulatory interactions in this bacterium documented in DataBase of Transcriptional regulation in *Bacillus Subtilis* (DBTBS) (15). It is interesting to note that even though both are free living bacteria and require similar concentrations of oxygen and temperature levels, *E. coli* has adapted to thrive inside its host while *B. subtilis* has adapted to soil environments. In this work we used a modified version of the Regulog approach described above to identify the interaction pairs and regulons of these networks through a comparison against complete genomes of Bacteria, Archaea and Eukarya.

MATERIALS AND METHODS

Protein sequence collection. A total of 204 completely sequenced genomes, including *Escherichia coli* K12 and *Bacillus subtilis*, were downloaded from the Kyoto Encyclopedia of Genes and Genomes (KEGG, ftp://ftp.genome.ad.jp/pub/kegg/genomes/) (16). See Table 1 in the supplementary material for details about the 204 completely sequenced genomes used in this study.

Interaction Data. We obtained the transcriptional regulatory interactions of *E. coli* K12 from RegulonDB version 4.0 (14), which compiles experimental information extracted from the literature. We also obtained the regulatory interactions of *B. subtilis* from DBTBS (15). We only considered regulatory interactions where regulators and target genes encode a polypeptide. Hence, interactions involving tRNA and other non-polypeptide coding target genes were ignored. Likewise, there are some transcription factors that activate and repress the same target gene due to the presence of more than one DNA binding site (e.g. Crp regulating *galE* as activator or repressor depending on two different DNA binding sites); we considered them as redundant interactions and only one interaction was used to represent them in the final dataset. Therefore, a total of 1678 non-redundant regulatory interactions that represent 119 TFs acting on 850 TGs were included in this work for *E. coli*. While a total of 785 non-redundant interactions representing 99 TFs affecting 666 TGs were included from the *B. subtilis* TRN.

Detection of potential TFs and TGs across species. It has been extensively reported that i) duplication of sequences, ii) divergence and iii) recombination are major sources of functional variation in protein evolution (17,18). However, it is important to note that the definition of "function" has often been vague and different approaches have been considered in comparative genomics (19-21). In this work, we assigned functional roles to TFs and TGs in other genomes by using an intersection of three criteria for the detection of orthologous proteins: **a)** bi-directional best hits (BDBHs), **b)** coverage in the BLASTP (22) alignment and **c)** detection of PFAM (23) conserved domains.

Orthologs are defined as proteins in different species that evolved from a common ancestor by speciation (24) and usually have the same function. Proteins that recently evolved from a

common ancestor by duplication previous to any speciation event are called “outparalogs” and hence are less likely to maintain the same function (25). By contrast, “inparalogs” are defined as those which have evolved by gene duplications that happened after the speciation event and are more likely to conserve their function. Operationally, both inparalogs and orthologous sequences are usually defined as best-matching homologs or bi-directional best hits (BDBHs) in another organism (26-28). Sequences in the same genome with more than 95% identity estimated with the CD-HIT program (29) were considered in this work as “inparalogs” and grouped into clusters. To identify orthologs we use the BDBHs definition through depurated genomes at 95% identity, with a significant BLASTP E-value ($\leq 10^{-3}$) using the WU-BLAST program (22). However, functional assignment is not yet complete with this approach since identifying orthologs for transcription factors is not always straightforward.

It is well known that conserved domains inside a protein determine their specific function and that these can represent evolutionary units especially for proteins with more than one domain where the pattern of functional conservation is more complex. Therefore, proteins are more likely to share functions if they contain the same domains in a similar arrangement (20,30,31). However, it is very important to consider that an increase in the number of domains can change the original function of a protein (32). We defined the conserved domains for all sequences analyzed in this work by Hidden Markov Models (HMM) taken from PFAM version 10 (23), using the HMMER 2.3.1 program (33) with an E-value of $\leq 10^{-3}$. In addition, we required that at least 70% of the PFAM model is covered by the sequence.

Operationally we identified orthologs as those proteins that satisfy the following four conditions:

1. Sequences of the target genome that have a bi-directional best hit in the query genome with a significant BLASTP E-value ($\leq 10^{-3}$).
2. At least 70% of the query sequence is included in the BLASTP alignment.
3. Target sequences share the PFAM domains of their query counterparts. Target sequences having one or more domains which match the orientation and arrangement to that of the query sequence and do not increment the total size of the protein in more than 100 residues were also considered in the analysis.

4. All the sequences included previously in the inparalog cluster were considered candidates that maintain the function only if the conditions 1, 2 and 3 are true for the representative sequence of the cluster.

We predicted the orthologs and PFAM domains of 119 TFs and 850 TGs of the TRN of *E. coli*, 99 TFs and 666 TGs of the TRN of *B. subtilis* as well as for the rest of the proteins from the complete genomes of *E. coli* and *B. subtilis* across the complete genomes of 175 bacteria, 19 archaea and 10 eukaryotes.

Data Management. To facilitate the display of results, we only show 110 complete genomes in all the figures, obtained by filtering out strains and species of the same bacterial genus keeping the strain or species with the maximum number of genes among a given genera of organisms. The evolutionary distance from *E. coli* and *B. subtilis* to all organisms was obtained according to the evolutionary branching process previously reported by Brown *et al.* (34). The evolutionary distance between any two organisms is related to the sum of the distances between each organism and its closest common ancestor.

Conservation of orthologs. To normalize the extent of conservation of the components (TFs and TGs) of the regulatory network in comparison to the total genome, we devised a simple metric called Conservation Index (CI) defined as

$$\mathbf{CI = (}x\mathbf{ / TC) / (}y\mathbf{ / GC)}$$

Where x is the number of orthologs present in the target genome from the total number of components (TC = TFs or TGs) of the TRN under consideration, and y is the total number of orthologs detected in the target genome from all protein coding genes (GC) in the genome under consideration, which in the case of *E. coli* would stand at 4248 and 4079 for *B. subtilis*. Therefore, CI is a measure of conservation of the components of the TRN of a genome pondered respect to the conservation of its genes. A CI near to 0 would indicate that the regulatory network components are poorly conserved in comparison to the genomic conservation, while a CI close to 1 would suggest that both the TF and TG are conserved to the same extent.

Prevalence of TF-TG orthologous pairs across genomes. The huge differences in genome size and gene content across Bacteria, Archaea and Eukarya, or between parasitic, symbiotic and free living organisms, can introduce bias when calculating the frequency distribution of the shared regulatory interactions across organisms. To correct for this problem, a factor of distance (D) was considered for weighting the presence or absence of transcriptional regulators and their target genes across genomes:

$$\mathbf{D_X = A' / (A' + A \cap B) \ \ and \ \ D_Y = B' / (B' + A \cap B)}$$

$$\mathbf{Where \ \ A' = A - (A \cap B) \ \ and \ \ B' = B - (A \cap B)}$$

For the transcription factor TF-*X* which regulates a target gene TG-*Y*, A denotes the set of all organisms from 110 non-redundant genomes in which an ortholog is found for TF-*X* and B denotes the set of all organisms in which an ortholog is seen for TG-*Y*. A' represents the subset of organisms which has an ortholog for *X* but not for *Y* and B' represents the subset of organisms for which an ortholog of *Y* is found but not *X*. A∩B represents the set of organisms in which both orthologs are found. As an example, consider the case of an interacting pair, TF-*X* and TG-*Y*, where the TF distance (D_x) is higher than TG distance (D_y) because TF contains a higher number of orthologs than the TG. Clearly, D_x should contain most of the orthologs corresponding to that of the D_y, and the unique number in the D_y (B') ought to be very small. In the limit, if the D_y has no unique orthologs relative to the D_x, the distance D_y would reach zero. A similar procedure for weighting has been used by others in the past, but focusing on domain contents (35).

The distances for each pair of TF and TG for the complete TRNs generated by the above approach were classified into three classes: a) TF and TG co-occur and hence the TF is likely to regulate the TG ($D_{TF} = D_{TG}$), b) TF is more conserved than TG ($D_{TF} > D_{TG}$) and c) TG is more conserved than TF ($D_{TF} < D_{TG}$) based on pre-defined thresholds (see below and supplementary material Method 1). To evaluate the statistical significance of the conservation of the regulatory interactions in these three different classes, we compared against 1000 randomly constructed regulatory networks for *E. coli* and *B. subtilis* each composed of the same number of interactions as the original TRNs but by switching the edges while maintaining the degree of each node the same as in the known TRN. It should be noted that this method of randomization preserves the in and out degree of the node and hence

topologically resembles known TRNs. In the entire analysis we excluded the interactions where TFs are auto-regulated as they would generate a bias when calculating the co-occurrence effect of TF-TG pairs. So the final set of interactions analyzed in this approach included 1620 TF-TG pairs in *E. coli* and 738 TF-TG pairs in *B. subtilis*.

Clustering the conserved interactions. For each TF in *E. coli* and *B. subtilis*, we calculated the percentage of total interactions conserved in its regulon across genomes. To represent this distribution we clustered by the extent of TRN and regulon conservation using Centroid Linkage Clustering method with an Uncentered Correlation as distance metric from the Cluster program (36). Other distance metrics were also evaluated but were not found to be significantly different in their ability to group lineages and regulons. Clustering data represents 118 regulons in *E. coli* and 93 regulons in *B. subtilis* conserved across genomes.

RESULTS

Conservation of TFs and TGs across species

Based on experimental information from 119 TFs and 850 TGs in *Escherichia coli* K12 and 99 TFs and 666 TGs in *Bacillis subtilis,* forming the components of their respective TRNs, we predicted their counterparts in 204 complete genomes, including 175 bacteria, 19 archaea and 10 eukaryotes (see Table 1 and Figure 1 in supplementary material for details). Figure 1 shows the distribution of orthologous conservation of the components (TFs and TGs) of the TRN from *E. coli* and *B. subtilis* across 110 non-redundant genomes representing 23 different phyla of the three cellular domains based on the phylogenetic reconstruction from Brown *et al.* (34).

From the perspective of *E. coli* (Figure 1a), the closest phylum includes 76 different Proteobacteria grouped from five subdivisions (15 α, 10 β, 42 γ, 4 δ and 5 ε). The extent of conservation in these groups is the highest of all analyzed phyla, where just over 30% of both TFs and TGs were conserved with the exception of parasitic and endosymbiotic organisms, which share only 10% of TFs and 20% TGs of the TRN from *E. coli*. Firmicutes from four different classes (10 Mollicutes, 22 Bacillales, 15 Lactobacillales and 4 Clostridia) were included too, which were found to have 20-30% of conserved TGs and 10-20% of conserved TFs, with the exception of parasitic and endosymbiotic organisms from Mollicutes, Mycobacterium, Tropheryma, which present less than 5% conservation of both TFs and TGs. Similar fractions of orthologs were detected in Actinobacteria as in Firmicutes. Other phyla like Bacteroidetes, Fusobacteria, Plactomyces, Cyanobacteria, Deinococci, Aquificae and Thermotogae share 10-25% of TGs and 5-15% of TFs. Parasitic phyla that include Chlamydiae and Spirochaetes have less than 15% and 5% of conserved TGs and TFs respectively. Among the 19 archaeal genomes which comprise 4 Crenarchaeota and 14 Euryarchaeota, we found 7-15% TGs and less than 3% of the TFs. The only known archaeal parasite, *Nanoarchaeum equitans* shares less than 1% TGs and TFs. Finally, 11 eukaryotic genomes which included 2 Protists, 4 Fungi, 2 Plants, 1 Insect and 1 Nematode share between 8 and 18% of TGs with the exception of the obligate intracellular parasite *Encephalitozoon cuniculi* that shows only 2% of TGs. Only *Saccharomyces cerevisiae, Cyanidioschyzon merolae, Arabidopsis thaliana* and *Caenorhabditis elegans* contain less than 1% of TF orthologs to those in *E. coli*.

From the perspective of *B. subtilis* (Figure 1b), although there seem to be fluctuations in the distribution of orthologs across genomes, more than 25% of TFs and TGs were found conserved in the Bacillus and Lactobacillus lineages. Parasitic and endosymbionts organisms in Mollicutes, Chlamydia, Spirochete and αγ Proteobacteria share less than 10% of the TGs and 5% of TFs. Conservation of the TFs and TGs across proteobacteria seems to be roughly constant, despite variations in phylogenetic distances with respect to *B. subtilis* until *Bdellovibrio bacteriovorus*. Beyond bacterial lineages we found that the conservation of the TFs drops off rapidly with no TFs conserved in Eukarya.

Irrespective of the variations in the conservation of the TFs and TGs across various phyla from the perspective of both the genomes, we observe that TGs tend to be more conserved than TFs as the phylogenetic distance increases while in closely related lineages TFs seem to be more conserved than TGs. This suggests that the majority of the transcriptional regulatory machinery in Bacteria could be lineage specific strengthening a previous observation made at the level of taxa (37). The TGs of the experimentally characterized TRN of *E. coli* correspond to 20% of its complete genome, while the characterized TFs correspond to 3%. In general, the measure of conservation (CI) (see Methods and Figure 1) shows that there is a steady increase in the conservation of the proportion of regulated component (TGs) of the cell in comparison to the regulatory component (TFs). Another interesting observation is that there is a certain fraction of the regulated component which is conserved in all lineages irrespective of the extent of genomic conservation of genes. However, it should be noted that the conserved fraction need not necessarily correspond to the same set of genes. From the view point of *B. subtilis*, although the decrease in TF conservation with phylogenetic distance is not as clear until far off lineages, the distribution of TG conservation seem to be more like that of *E. coli*.

Evolution of Global Regulators across bacterial species

Here we consider the definition of Global Regulators (GRs) for *E. coli* from Martínez-Antonio and Collado-Vides (5), based on the number of genes they regulate and additional factors, such as the number of co-regulators and the number of conditions. Due to the absence of sufficient information for *B. subtilis* to classify TFs on the same basis, we considered those TFs as GRs which regulate the highest number of genes in the known TRN (more than 20

regulatory interactions). GRs regulate the activity of 51% of the known TRN in *E. coli* (5), so we aimed at understanding the conservation spectrum of these genes. There are seven GRs in *E. coli*: Crp, Fnr, Ihf, Fis, ArcA, Hns and Lrp and eight in *B. subtilis:* CcpA, AbrB, ComK, Fur, PhoP, TnrA, CodY and PurR.

The predicted orthologs of GRs vary in their extent of conservation across the phylogenetic spectrum, although none of them seem to be conserved in eukaryotes (see Figure 2a and b). However, their TGs are conserved in the three cellular domains suggesting that these TGs could be regulated in those organisms by analogous or paralogous TFs. It is interesting to note that none of the global transcription factors are homologously related at the sequence level between *E. coli* and *B. subtilis*, indicating that global TFs need not be conserved among phylogenetically distant genomes. This observation could imply that global TFs evolve in different lineages independently, according to the requirements in different conditions in which the organisms dwell. Of all the GRs only Lrp and the Ihf subunits (HimD or HimA) were found to occur in Archaea suggesting that most of these GRs originated in bacterial lineages. Curiously, orthologs of Crp and Fnr, which are paralogs in *E coli*, seem to have an alternating distribution beyond Proteobacteria, in Firmicutes and Cyanobacteria, possibly indicating a substitution of their roles in these lineages or horizontal gene transfer of one of the members of the Crp family from one to another. In *B. subtilis* only the GRs, CcpA, Fur and PhoP seem to show their presence in phylogenetically distant genomes, suggesting an ancient origin compared to its other GRs. Finally, Fis, ArcA and Hns in *E. coli* have a limited distribution in other bacterial species, specifically restricted to Proteobacteria, while AbrB, ComK, CodY and PurR in *B. subtilis* are restricted to Bacillus and Lactobacillus lineages. A recent work in this direction shows the poor conservation of the hubs in regulatory networks of prokaryotic genomes (38).

The case of the phylogenetic distribution of Lrp extending to Archaea, needs further discussion as it is the only monomeric GR that is well conserved across lineages. Previously, homologs of Lrp-like transcriptional regulators were identified although their presence was detected only in Prokaryotes (39). The wide phyletic distribution of Lrp homologs among Archaea and Bacteria suggests that an Lrp-type regulator was present in the last common ancestor of Bacteria and Archaea. Nevertheless the distribution of Lrp-type regulators seems to vary across organisms (e.g. 20 copies in *Mesorhizobium loti*, 3 in *E. coli* and none in the

strains of *Buchnera*, *Mycoplasma* and *Chlamydia*). The latter ones are bacterial endosymbionts and completely depend on their host for the supply of amino acids and other key metabolites. They have reduced genomes which could explain the absence of Lrp members. In spite of their conservation in various phyla, even in closely related species its global regulatory mechanism does not seem to be conserved as has been demonstrated by the analysis of the Lrp ortholog of *Haemophilus influenzae* (40). These observations point to the conclusion that even in organisms of the same phyla there is no real constraint for the conservation of the functions of GRs, although the GRs themselves might be well conserved at the level of sequence. Hence providing insight that at the level of transcriptional regulatory machinery orthologous genes in different genomes could play different functional roles.

Evolution of TF-TG pairs in TRNs

We implemented a distance (D) as described in Methods for a comparative analysis of the orthology distribution of a TF and its TG. As mentioned earlier, we studied the conservation of the transcriptional regulatory interactions across bacterial species by assigning each TF-TG pair to one of the three different categories: a) when a TF and its TG are both present or absent together, b) when a TF is conserved in more species than its TG and c) when a TG is conserved in more species than its TF (see examples in Figure 3). Ideally if the regulatory interaction is co-occurring across species, one would expect that D_{TF} and D_{TG} should both be equal to zero, but for several reasons like horizontal transfer events, loss or duplication of genes and errors involved in the detection of orthologs, one could obtain biases in the distribution of co-evolving TF-TG pairs. In order to take into account these factors and to determine a threshold for identifying co-evolving TF-TG pairs we used pairs of genes in metabolic pathways from KEGG (16) as a control (see supplementary material for details about generation of thresholds). It is known that genes in the same metabolic pathway often co-evolve and are well conserved (41,42). Based upon the thresholds determined for each genome we identified the co-evolving TF-TG pairs and then included the rest of the interactions into one of the two classes based on whether D_{TF} is higher or lower than D_{TG}.

Table 1 shows the Z-scores of conservation for each category of TF-TG pairs in both the genomes computed upon comparing with the randomly generated TRNs as described earlier (see Methods). It can be seen that there is a relatively small fraction of the TRN in both genomes which is conserved and co-evolving across genomes. However the significance of co-evolution from the perspective of *B. subtilis* seems to be smaller than in *E. coli* as seen from the p-value (see table), which might be due to the under representation of the number of genomes in Firmicutes compared to Proteobacteria or due to the difference in the size of the TRNs being used in the two genomes. A roughly equal proportion of TF-TG pairs occur in the categories of TF > TG and TF < TG in both genomes (see Figure 2 in supplementary material). The Z-scores in the respective categories suggest that there is a no clear tendency for either TF > TG or TF < TG in both the genomes, as the Z-scores in each case correspond to no more than 3-4 standard deviations except that of TF < TG in *E. coli*. This indicates that there is no constraint on the co-evolution of a TF and its TG in an interacting pair for majority of the interactions. Note that this is different from the quantitative analysis of the conservation of individual elements (TFs and TGs) conserved from the TRN as here we are interested in the co-evolution of the pairs of interactions. The above analysis suggests that there is a small but well conserved fraction of the TRN which is present in diverse phyla and that the majority of the TF-TG pairs evolve independently.

Conservation of regulons across bacterial species

In Figure 4, we show the conservation of regulatory interactions at the level of regulons for both *E. coli* and *B. subtilis* across 110 non-redundant complete genomes representing various phyla, clustered horizontally by the extent of TRN conservation across genomes and vertically by the extent of regulon conservation (see Methods). In general, it can be seen that the TRNs share few regulons across the phylogenetic spectrum, although the conservation is more in closely related lineages.

We compared the distribution of the genomes in the horizontal axis which was based on the extent of conservation of the TRNs, with that of the phylogenetic distribution generated by the method of Brown *et al.* (34) and found that several lineages were appropriately grouped, suggesting that TRN conservation can aid in segregating the major bacterial kingdoms and

that phylogenetic distance provides a measure of the extent of TRN conservation. It is interesting to note that the clades closest to *E. coli* (Figure 4a), which includes several Proteobacteria, share around 40% of the transcriptional regulatory interactions from *E. coli* except for several parasitic and endosymbiotic organisms, which were grouped together and show poor conservation of the TRN. The Proteobacteria *Blochmannia floridanus*, mollicutes *Mesoplasma florum* and *Ureaplasma urealyticum*, the Archaea *Methanopyrus kandleri, Methanococcus jannaschi, Methanobacterium thermoautotropicum, Halobacterium* sp, *Pyrococcus furiosus, Aeropyrum pernix, Nanoarchaeum equitans* and the ten analyzed eukaryotic organisms do not seem to share any regulatory interaction with *E. coli*.

From the perspective of *B. subtilis* (Figure 4b) the closest clades share around 30% of the transcriptional regulatory interactions except for some parasitic and endosymbiotic organisms from the Bacillus and Lactobacillus lineages. The Mollicute *U. urealyticum*, the Archaea *Picrophilus torridus*, *Pyrobaculum aerophilum*, *N. equitans* and the ten analyzed eukaryotic organisms do not seem to share any regulatory interaction with the TRN of *B. subtilis*.

The horizontal distribution in Figure 4 which shows the conservation of regulons in the respective TRNs points out that certain regulons are widely conserved across species, although this fraction seems to be higher from the perspective of *E. coli*. Highly conserved regulons from *E. coli's* TRN include metabolic and structural components like IscR, ArgR, AsnC, BirA, Crp, DnaA, Fnr, Fur, GlpR, Ihf, LexA, Lrp, NagC, OxyR, OmpR, PhoB, KdpE and RbsR. Within these conserved regulons there is only a small set of ancient conserved interactions among different bacterial phyla, which represent around 6% of the TRN of *E. coli*. These interactions regulate important cellular processes in *E. coli* such as synthesis of arginine, asparagine, biotin and ribose, transport of amino acids and iron, availability of phosphate, replication process and the SOS response system (see Table 2 in supplementary material for additional information about these anciently conserved interactions from both the genomes). Highly conserved regulons from *B. subtilis* include DnaA, LexA, HrcA, PerR, BirA, AzlB, YwfK, AhrC (ArgR), CcpA, Fur, ResD, YycF, PhoP, DegU and MntR. These regulons are involved in the regulation of the synthesis of arginine and biotin; transport of manganese, availability of phosphate, heat shock response genes, global regulatory functions, replication process and the SOS response system. Some of the conserved regulons found here such as ArgR, BirA and LexA have been previously reported to be found in various phyla

(6,43,44). However, this repertoire of conserved regulons should enhance our understanding of the conservation at the level of regulons and guide further experimental studies to characterize them.

For example, among these conserved regulons there are at least two hypothetical TFs: YwfK and YycF whose function is yet unknown. These conserved patterns at the level of regulons could be used to understand and characterize these TFs through a combination of experimental and computational methods thereby aiding in the determination of their function. Computationally, one can identify the function of the TF from the functional context of its regulated genes or their conserved orthologs in a way similar to what has been demonstrated earlier for several TFs from genomic context (45). Regulatory binding sites can be identified through a phylogenetic foot printing analysis of the upstream regions of putative target genes in closely related genomes. Experimental evidence added to computational predictions can elevate the quality of the predictions as has been shown in the case of LexA regulon (6).

DISCUSSION

The complexity of the transcriptional regulatory networks in bacterial organisms is largely affected by their adaptation to the dynamically changing environmental stresses that are characteristic of an organism's niche. For example, enteric bacteria, soil bacteria and other free-living bacteria live in complex environments and have correspondingly complex sensor-response-control subsystems (46). In contrast, the narrow ecological ranges and frequent population bottlenecks of obligate pathogens and symbionts have resulted in increased rates of genetic drift and reduced selective constraint on gene function and number (47-49). Our results indicate that obligate symbiotic as well as parasitic life styles share only around 10% of the orthologous components of the TRNs of *E. coli* and *B. subtilis*. The loss of regulatory elements may reflect a relative constancy in the host environment, allowing these organisms to have a simplified regulatory structure (46,50). According to our results, the loss of transcription factors more than target genes could be the main cause of these dramatic changes in the TRN. This can also be seen from the specific scenario of the conservation of global regulators of *E. coli* which have a limited biological distribution although they directly modulate the expression of about 51% of its genes (5).

As previously reported, the conservation of genes and regulatory interactions is related to the phylogenetic distance and to the life style of the organisms (10,38). Based on our results, we can see that quantitatively that transcription factors are less conserved than the target genes as phylogenetic distance increases, which could suggest that transcriptional regulation of genes changes faster through evolution than the genes themselves. Related to this, Maslov *et al.* (51) found that the rate of evolutionary differentiation of transcriptional regulatory interactions proceeds faster than that of target genes and their protein interactions. However, an analysis of the conservation of pairs of regulatory interactions across genomes indicated different tendencies in the conservation of TF-TG pairs, suggesting that TF-TG pairs often do not co-evolve in the evolution of TRNs. Nevertheless it should be clear that in the first case, when a TF is conserved in different species without its corresponding TG, it would imply that the TF is indeed involved in the regulation of a different set of TGs than those in the genome under consideration and in the second case, when a TG is conserved and its TF is lost, it could imply that the TG is regulated by an analogous or paralogous factor. Both cases suggest a level of plasticity that TRNs can impose on the evolution of genomes to different

environments. The evolutionary reasons for the observed tendencies in the conservation of TF-TG pairs need to be analyzed more specifically.

Despite poor conservation of the regulatory interactions across genomes, certain individual interactions have been well conserved across different eubacterial phyla, which could regulate essential transcriptional processes in Bacteria. Most of these processes are well characterized and are related directly or indirectly to the translational, structural and transcriptional machinery of the cell, suggesting a cause for their conserved nature across wide phylogenetic distances. Despite the type of regulation (repressor or activator) and that DNA binding site(s) can change across genomes, it is reasonable to think that it is important to maintain the regulation of these core processes through the same elements, as in the case of BirA and DnaA regulators which seem to be a result of common ancestry in all bacteria.

The transcriptional regulatory network appears to evolve in a step-wise manner, with loss and gain of individual interactions probably playing a greater role than loss and gain of whole motifs or modules of interactions. As Teichmann and Babu (52) reported previously, most network motifs have risen by convergent evolution and not by genetic duplication of ancestral circuits. Thus, with the exception of a small fraction of the TRN, it could be possible that large portions of the TRNs might have evolved through extensive changes and re-connections among the components of the network in the evolution of the species. Here we demonstrate that individual elements, interacting pairs and groups of interactions are not conserved, in fact even in closely related species. This reflects that in each speciation event to adapt to environmental changes, transcriptional regulation is more flexible than the genetic component of the organisms for phenotypic adaptation. This work should provide a perspective of the plasticity of the transcriptional regulatory network in bacteria, which could contribute to understand the transcriptional basis of natural variation.

Supplementary material including the complete set of regulons in *Escherichia coli K12* and *Bacillus subtilis* analyzed in this work and predicted regulons in complete genomes can be downloaded from:
http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/.

ACKNOWLEDGEMENTS

We would like to thank Heladia Salgado Osorio for supplying the interaction data from RegulonDB and to Kenta Nakai and Yuko Makita for providing us the data of transcriptional regulation from DBTBS. We would also like to thank J. Javier Díaz Mejía, Gabriel Moreno-Hagelsieb, Cei Abreu Goodger and Bruno Contreras Moreira for insightful discussions and comments on the manuscript. We would also like to acknowledge support from CONACyT program number 185993.

REFERENCES

1. Fell, D.A. and Wagner, A. (2000) The small world of metabolism. *Nat Biotechnol*, **18**, 1121-1122.

2. Thieffry, D., Huerta, A.M., Perez-Rueda, E. and Collado-Vides, J. (1998) From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in Escherichia coli. *Bioessays*, **20**, 433-440.

3. Ouzounis, C.A. and Karp, P.D. (2000) Global properties of the metabolic map of Escherichia coli. *Genome Res*, **10**, 568-576.

4. Shen-Orr, S.S., Milo, R., Mangan, S. and Alon, U. (2002) Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, **31**, 64-68.

5. Martinez-Antonio, A. and Collado-Vides, J. (2003) Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol*, **6**, 482-489.

6. Erill, I., Jara, M., Salvador, N., Escribano, M., Campoy, S. and Barbe, J. (2004) Differences in LexA regulon structure among Proteobacteria through in vivo assisted comparative genomics. *Nucleic Acids Res*, **32**, 6617-6626.

7. Herrgard, M.J., Covert, M.W. and Palsson, B.O. (2004) Reconstruction of microbial transcriptional regulatory networks. *Curr Opin Biotechnol*, **15**, 70-77.

8. Mazurie, A., Bottani, S. and Vergassola, M. (2005) An evolutionary and functional assessment of regulatory network motifs. *Genome Biol*, **6**, R35.

9. Yu, H., Luscombe, N.M., Lu, H.X., Zhu, X., Xia, Y., Han, J.D., Bertin, N., Chung, S., Vidal, M. and Gerstein, M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res*, **14**, 1107-1118.

10. Babu, M.M., Luscombe, N.M., Aravind, L., Gerstein, M. and Teichmann, S.A. (2004) Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, **14**, 283-291.

11. Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U. and Margalit, H. (2004) Network motifs in integrated cellular networks of transcription-regulation and protein-protein interaction. *Proc Natl Acad Sci U S A*, **101**, 5934-5939.

12. Sharan, R., Suthram, S., Kelley, R.M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R.M. and Ideker, T. (2005) Conserved patterns of protein interaction in multiple species. *Proc Natl Acad Sci U S A*, **102**, 1974-1979.

13. Struhl, K. (1999) Fundamentally different logic of gene regulation in eukaryotes and prokaryotes. *Cell*, **98**, 1-4.

14. Salgado, H., Gama-Castro, S., Martinez-Antonio, A., Diaz-Peredo, E., Sanchez-Solano, F., Peralta-Gil, M., Garcia-Alonso, D., Jimenez-Jacinto, V., Santos-Zavaleta, A., Bonavides-Martinez, C. *et al.* (2004) RegulonDB (version 4.0): transcriptional

regulation, operon organization and growth conditions in Escherichia coli K-12. *Nucleic Acids Res*, **32**, D303-306.

15. Makita, Y., Nakao, M., Ogasawara, N. and Nakai, K. (2004) DBTBS: database of transcriptional regulation in Bacillus subtilis and its contribution to comparative genomics. *Nucleic Acids Res*, **32**, D75-77.

16. Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y. and Hattori, M. (2004) The KEGG resource for deciphering the genome. *Nucleic Acids Res*, **32**, D277-280.

17. Teichmann, S.A., Murzin, A.G. and Chothia, C. (2001) Determination of protein function, evolution and interactions by structural genomics. *Curr Opin Struct Biol*, **11**, 354-363.

18. Chothia, C., Gough, J., Vogel, C. and Teichmann, S.A. (2003) Evolution of the protein repertoire. *Science*, **300**, 1701-1703.

19. Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M. and Yuan, Y. (1998) Predicting function: from genes to genomes and back. *J Mol Biol*, **283**, 707-725.

20. Wilson, C.A., Kreychman, J. and Gerstein, M. (2000) Assessing annotation transfer for genomics: quantifying the relations between protein sequence, structure and function through traditional and probabilistic scores. *J Mol Biol*, **297**, 233-249.

21. Lan, N., Montelione, G.T. and Gerstein, M. (2003) Ontologies for proteomics: towards a systematic definition of structure and function that scales to the genome level. *Curr Opin Chem Biol*, **7**, 44-54.

22. Lopez, R., Silventoinen, V., Robinson, S., Kibria, A. and Gish, W. (2003) WU-Blast2 server at the European Bioinformatics Institute. *Nucleic Acids Res*, **31**, 3795-3798.

23. Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Howe, K.L. and Sonnhammer, E.L. (2000) The Pfam protein families database. *Nucleic Acids Res*, **28**, 263-266.

24. Fitch, W.M. (1970) Distinguishing homologous from analogous proteins. *Syst Zool*, **19**, 99-113.

25. Sonnhammer, E.L. and Koonin, E.V. (2002) Orthology, paralogy and proposed classification for paralog subtypes. *Trends Genet*, **18**, 619-620.

26. Tatusov, R.L., Koonin, E.V. and Lipman, D.J. (1997) A genomic perspective on protein families. *Science*, **278**, 631-637.

27. Huynen, M.A. and Bork, P. (1998) Measuring genome evolution. *Proc Natl Acad Sci U S A*, **95**, 5849-5856.

28. Janga, S.C. and Moreno-Hagelsieb, G. (2004) Conservation of adjacency as evidence of paralogous operons. *Nucleic Acids Res*, **32**, 5392-5397.

29. Li, W., Jaroszewski, L. and Godzik, A. (2002) Tolerating some redundancy significantly speeds up clustering of large protein databases. *Bioinformatics*, **18**, 77-82.

30. Hegyi, H. and Gerstein, M. (2001) Annotation transfer for genomics: measuring functional divergence in multi-domain proteins. *Genome Res*, **11**, 1632-1640.

31. Bornberg-Bauer, E., Beaussart, F., Kummerfeld, S.K., Teichmann, S.A. and Weiner, J., 3rd. (2005) The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, **62**, 435-445.

32. Wheelan, S.J., Marchler-Bauer, A. and Bryant, S.H. (2000) Domain size distributions can predict domain boundaries. *Bioinformatics*, **16**, 613-618.

33. Eddy, S.R. (1996) Hidden Markov models. *Curr Opin Struct Biol*, **6**, 361-365.

34. Brown, J.R., Douady, C.J., Italia, M.J., Marshall, W.E. and Stanhope, M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet*, **28**, 281-285.

35. Yang, S., Doolittle, R.F. and Bourne, P.E. (2005) Phylogeny determined by protein domain content. *Proc Natl Acad Sci U S A*, **102**, 373-378.

36. de Hoon, M.J., Imoto, S., Nolan, J. and Miyano, S. (2004) Open source clustering software. *Bioinformatics*, **20**, 1453-1454.

37. Coulson, R.M., Enright, A.J. and Ouzounis, C.A. (2001) Transcription-associated protein families are primarily taxon-specific. *Bioinformatics*, **17**, 95-97.

38. Madan Babu, M., Teichmann, S.A. and Aravind, L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol*, **358**, 614-633.

39. Brinkman, A.B., Ettema, T.J., de Vos, W.M. and van der Oost, J. (2003) The Lrp family of transcriptional regulators. *Mol Microbiol*, **48**, 287-294.

40. Friedberg, D., Midkiff, M. and Calvo, J.M. (2001) Global versus local regulatory roles for Lrp-related proteins: Haemophilus influenzae as a case study. *J Bacteriol*, **183**, 4004-4011.

41. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651-654.

42. Date, S.V. and Marcotte, E.M. (2003) Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nat Biotechnol*, **21**, 1055-1062.

43. Makarova, K.S., Mironov, A.A. and Gelfand, M.S. (2001) Conservation of the binding site for the arginine repressor in all bacterial lineages. *Genome Biol*, **2**, RESEARCH0013.

44. Rodionov, D.A., Mironov, A.A. and Gelfand, M.S. (2002) Conservation of the biotin regulon and the BirA regulatory signal in Eubacteria and Archaea. *Genome Res*, **12**, 1507-1516.

45. Doerks, T., Andrade, M.A., Lathe, W., 3rd, von Mering, C. and Bork, P. (2004) Global analysis of bacterial transcription factors to predict cellular target processes. *Trends Genet*, **20**, 126-131.

46. Cases, I., de Lorenzo, V. and Ouzounis, C.A. (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol*, **11**, 248-253.

47. Moran, N.A. (1996) Accelerated evolution and Muller's rachet in endosymbiotic bacteria. *Proc Natl Acad Sci U S A*, **93**, 2873-2878.

48. Itoh, T., Martin, W. and Nei, M. (2002) Acceleration of genomic evolution caused by enhanced mutation rate in endocellular symbionts. *Proc Natl Acad Sci U S A*, **99**, 12944-12948.

49. Andersson, S.G. and Kurland, C.G. (1998) Reductive evolution of resident genomes. *Trends Microbiol*, **6**, 263-268.

50. Wilcox, J.L., Dunbar, H.E., Wolfinger, R.D. and Moran, N.A. (2003) Consequences of reductive evolution for gene expression in an obligate endosymbiont. *Mol Microbiol*, **48**, 1491-1500.

51. Maslov, S., Sneppen, K., Eriksen, K.A. and Yan, K.K. (2004) Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol*, **4**, 9.

52. Teichmann, S.A. and Babu, M.M. (2004) Gene regulatory network growth by duplication. *Nat Genet*, **36**, 492-496.

Table 1. Statistical significance of conservation for the different categories of TF-TG pairs in the TRNs of *Escherichia coli* K12 and *Bacillus subtilis.*

Escherichia coli K12			*Bacillus subtilis*		
Category	**Interactions**	**Z-score(P-value)**	**Category**	**Interactions**	**Z-score(P-value)**
TF = TG	**15**	**5.44 (< 0.0001)**	**TF = TG**	**15**	**2.99 (0.0028)**
TF > TG	**813**	**-5.06 (< 0.0001)**	**TF > TG**	**363**	**3.24 (0.0012)**
TF < TG	**759**	**3.68 (0.00023)**	**TF < TG**	**349**	**-3.60 (0.00032)**

FIGURE LEGENDS:

Figure 1. Conservation of the components of the TRN (TFs and TGs) across the three domains of life for a) *Escherichia coli* K12 and b) *Bacillus subtilis*. In X axis are 110 non-redundant genomes ordered by phylogenetic distance (see Methods). In Y-axis (to the left) is the percentage of conservation of the elements (TFs and TGs) of the TRNs. CI values (shown to the right on the Y-axis) represent a measure of conservation of the components of the TRN of a genome with respect to the conservation of its genes. Color codes on X-axis represent different phylogenetic clades as described in the supplementary material.

Figure 2. Conservation of Global Regulators (GR) and their regulons across genomes for a) *Escherichia coli* K12 and b) *Bacillus subtilis*. Note that for GRs only presence (in black) or absence (in white) is shown (upper section) while for regulons percentage of interactions conserved is shown (lower section for each GR). Genomes are arranged in increasing order of phylogenetic distance with respect to the organism of reference.

Figure 3. Classification of TF-TG pairs into three different categories. Examples of TF-TG pairs distributed in to different classes based on their co-evolution pattern: a) TFs and TGs co-evolve [*dnaA* and *dnaN* in *E. coli* where $D_{dnaA}=3/(3+75)=0.038$ and $D_{dnaN}=9/(9+75)=0.107$] b) TF is evolutionarily more conserved than TG [*fur* and *entD* in *E. coli* where $D_{fur}=70/(70+2)=0.972$ and $D_{entD}=0/(0+2)=0$] and c) TF is less conserved than TG [*metJ* and *ahpC* in *E. coli* where $D_{metJ}=1/(1+11)=0.083$ and $D_{ahpC}=81/(81+11)=0.88$].

Figure 4. Conservation of regulons across genomes clustered by the extent of TRN and regulon conservation for a) *Escherichia coli* K12 and b) *Bacillus subtilis.* The intensity of the color for each regulon in each genome indicates the percentage of total interactions in the regulon shared.

Figure 1a.

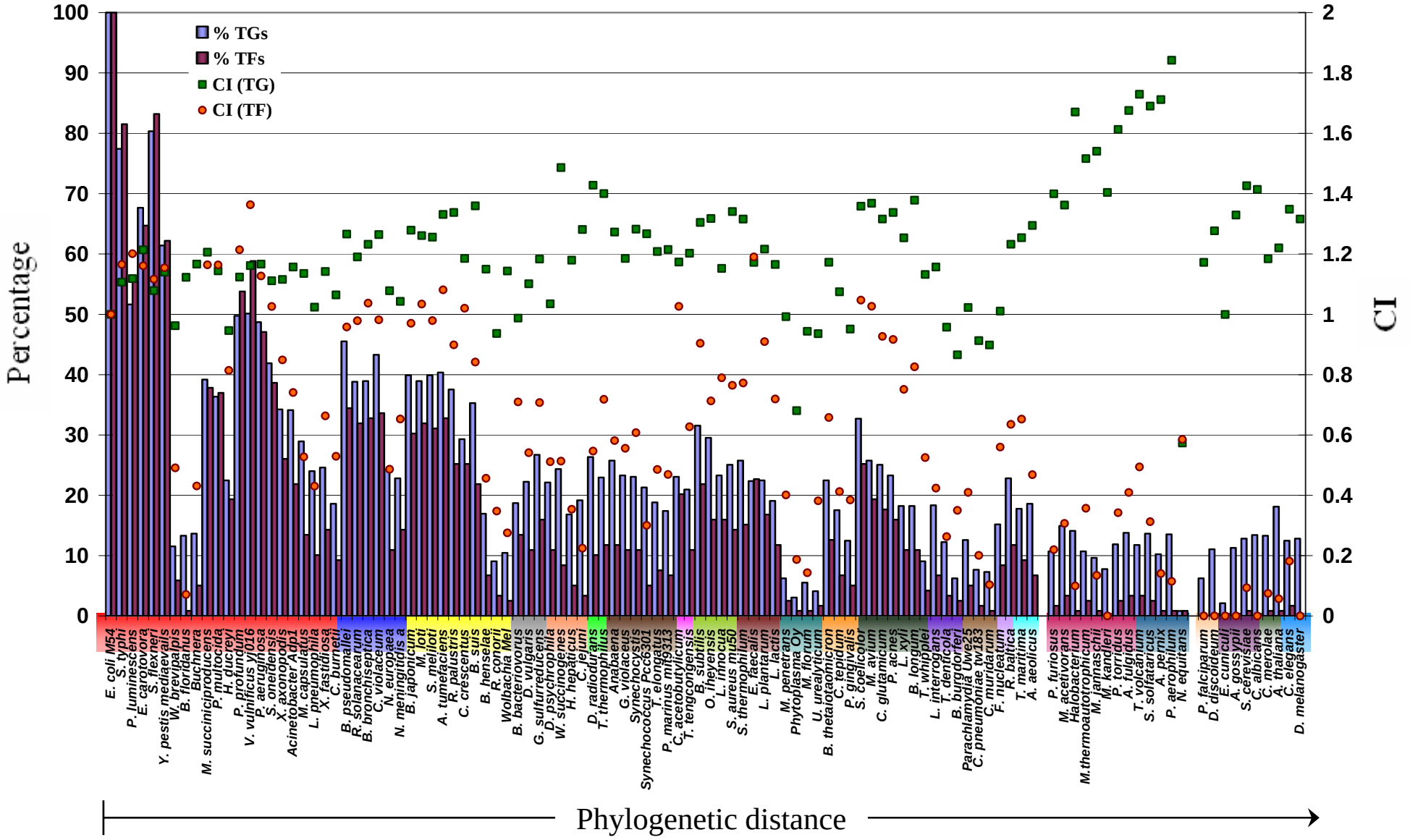

% TGs
% TFs
CI (TG)
CI (TF)
Percentage
CI
100
90
80
70
60
50
40
30
20
10
0
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
E. coli M54
S. typhi
P. luminescens
E. carotovora
S. flexneri
Y. pestis mediaevalis
W. brevipalpis
B. floridanus
Buchnera
M. succiniciproducens
P. multocida
H. ducreyi
P. profundum
V. vulnificus yj016
P. aeruginosa
S. oneidensis
X. axonopodis
Acinetobacter Ap01
M. capsulatus
L. pneumophila
X. fastidiosa
C. burnetii
B. pseudomallei
R. solanacearum
B. bronchiseptica
C. violaceum
N. europaea
N. meningitidis a
B. japonicum
M. loti
S. meliloti
A. tumefaciens
R. palustris
C. crescentus
B. suis
B. henselae
R. conorii
Wolbachia Mel
B. bacteriovorus
D. vulgaris
G. sulfurreducens
D. psychrophila
W. succinogenes
H. hepaticus
C. jejuni
D. radiodurans
T. thermophilus
Anabaena
G. violaceus
Synechocystis
Synechococcus Pcc6301
T. elongatus
P. marinus mit9313
C. acetobutylicum
T. tengcongensis
B. subtilis
O. iheyensis
L. innocua
S. aureus mu50
S. thermophilum
E. faecalis
L. plantarum
L. lactis
M. penetrans
Phytoplasma Oy
M. florum
U. urealyticum
B. thetaiotaomicron
C. tepidum
P. gingivalis
S. coelicolor
M. avium
C. glutamicum
P. acnes
L. xyli
B. longum
T. whipplei
L. interrogans
T. denticola
B. burgdorferi
Parachlamydia Uwe25
C. pneumoniae tw183
C. muridarum
F. nucleatum
R. baltica
T. maritima
A. aeolicus
P. furiosus
M. acetivorans
Halobacterium
M.thermoautotrophicum
M. jannaschii
M. kandleri
P. torridus
A. fulgidus
T. volcanium
S. solfataricus
A. pernix
P. aerophilum
N. equitans
P. falciparum
D. discoideum
E. cuniculi
A. gossypii
S. cerevisiae
C. albicans
C. merolae
A. thaliana
C. elegans
D. melanogaster
Phylogenetic distance

Figure 1b.

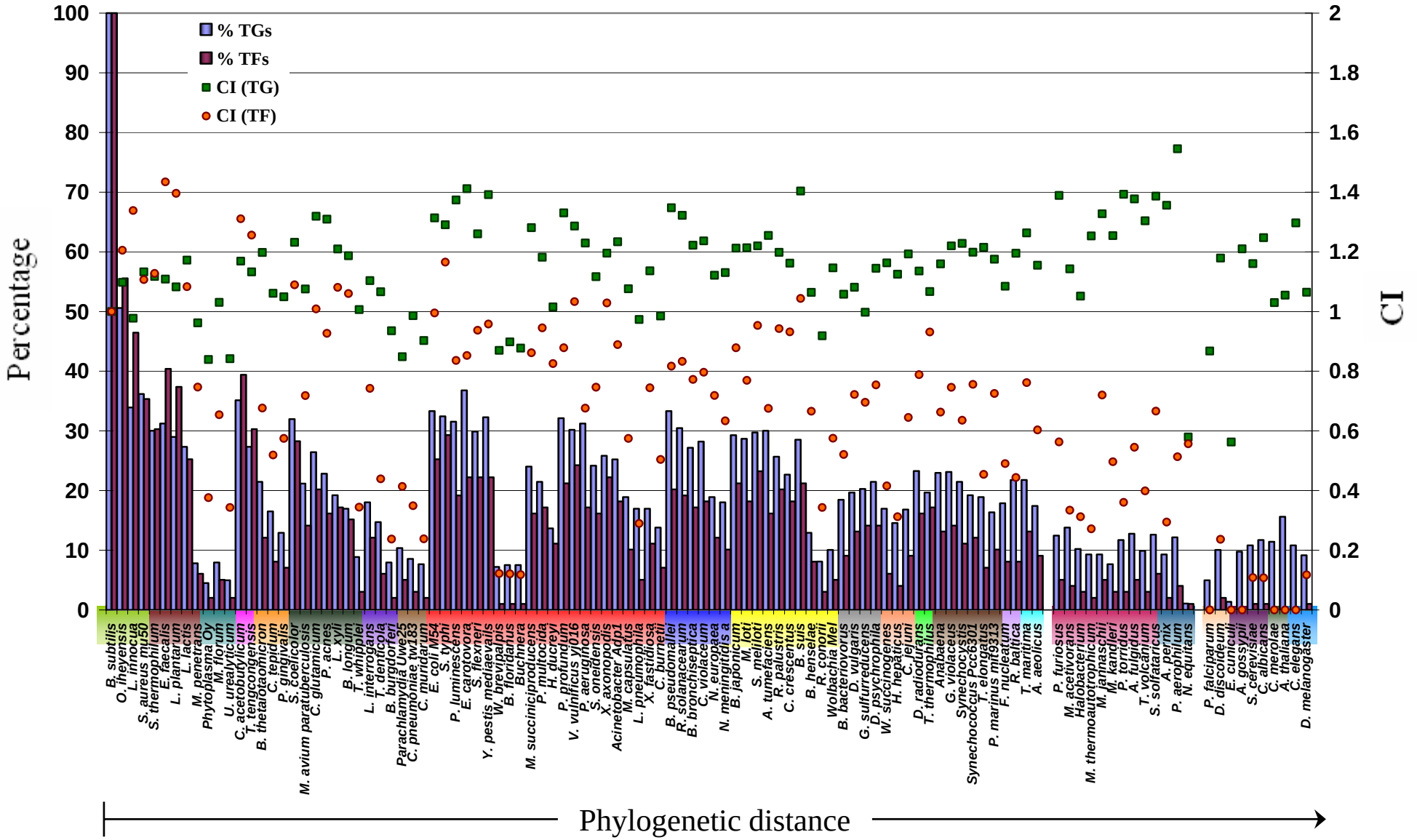

% TGs
% TFs
CI (TG)
CI (TF)
Percentage
CI
100
90
80
70
60
50
40
30
20
10
0
2
1.8
1.6
1.4
1.2
1
0.8
0.6
0.4
0.2
0
B. subtilis
O. iheyensis
L. innocua
S. aureus mu50
S. thermophilum
E. faecalis
L. plantarum
L. lactis
M. penetrans
Phytoplasma Oy
M. florum
U. urealyticum
C. acetobutylicum
T. tengcongensis
B. thetaiotaomicron
C. tepidum
P. gingivalis
S. coelicolor
M. avium paratuberculosis
C. glutamicum
P. acnes
L. xyli
B. longum
T. whipplei
L. interrogans
T. denticola
B. burgdorferi
Parachlamydia Uwe25
C. pneumoniae tw183
C. muridarum
E. coli M54
S. typhi
P. luminescens
E. carotovora
S. flexneri
Y. pestis mediaevalis
W. brevipalpis
B. floridanus
Buchnera
M. succiniciproducens
P. multocida
H. ducreyi
P. profundum
V. vulnificus yj016
P. aeruginosa
S. oneidensis
X. axonopodis
Acinetobacter Adp1
M. capsulatus
L. pneumophila
X. fastidiosa
C. burnetii
B. pseudomallei
R. solanacearum
B. bronchiseptica
C. violaceum
N. europaea
N. meningitidis a
B. japonicum
M. loti
S. meliloti
A. tumefaciens
R. palustris
C. crescentus
B. suis
B. henselae
R. conorii
Wolbachia Mel
B. bacteriovorus
D. vulgaris
G. sulfurreducens
D. psychrophila
W. succinogenes
H. hepaticus
C. jejuni
D. radiodurans
T. thermophilus
Anabaena
G. violaceus
Synechocystis
Synechococcus Pcc6301
T. elongatus
P. marinus mit9313
F. nucleatum
R. baltica
T. maritima
A. aeolicus
P. furiosus
M. acetivorans
Halobacterium
M. thermoautotrophicum
M. jannaschii
M. kandleri
P. torridus
A. fulgidus
T. volcanium
S. solfataricus
A. pernix
P. aerophilum
N. equitans
P. falciparum
D. discoideum
E. cuniculi
A. gossypii
S. cerevisiae
C. albicans
C. merolae
A. thaliana
C. elegans
D. melanogaster
Phylogenetic distance

Figure 2a.

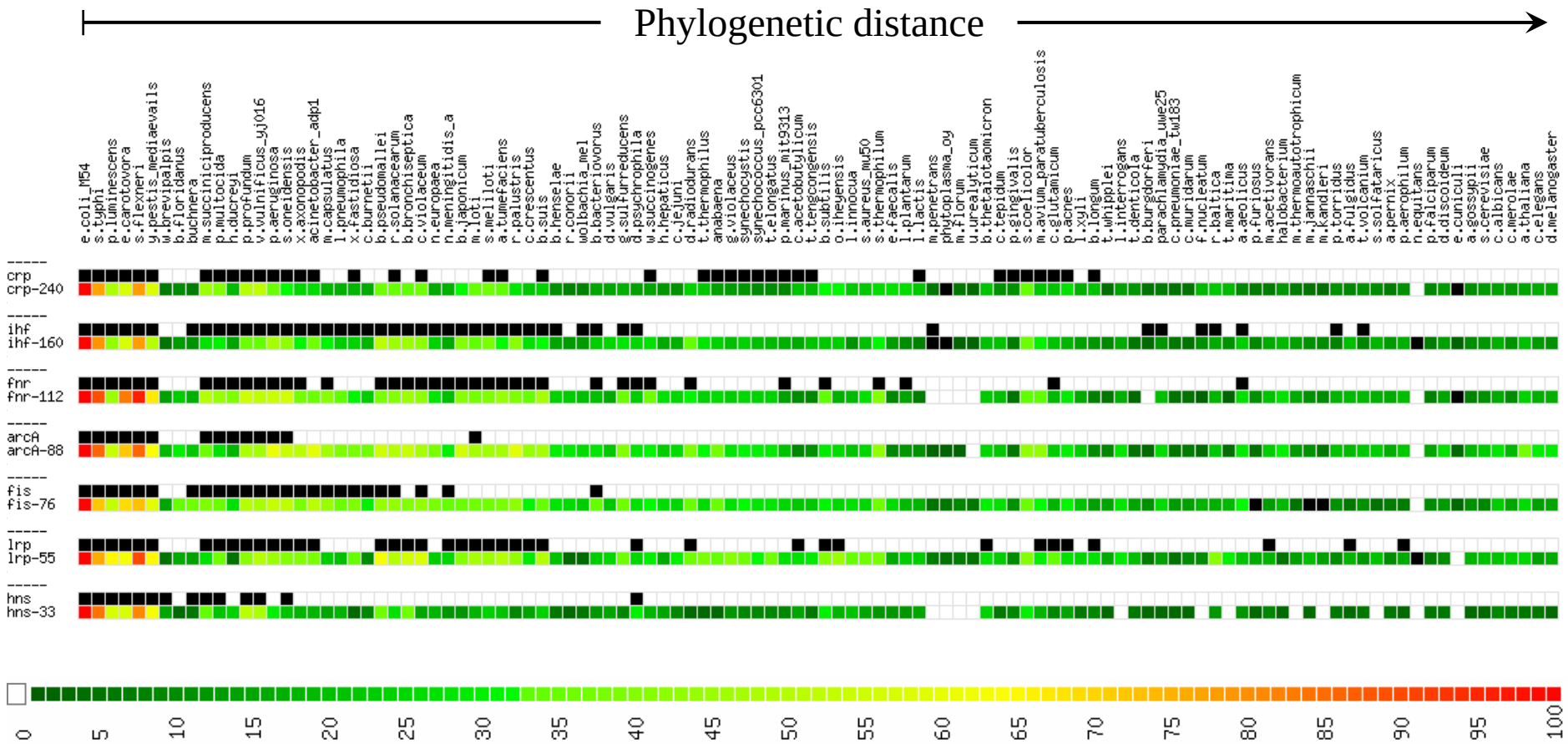
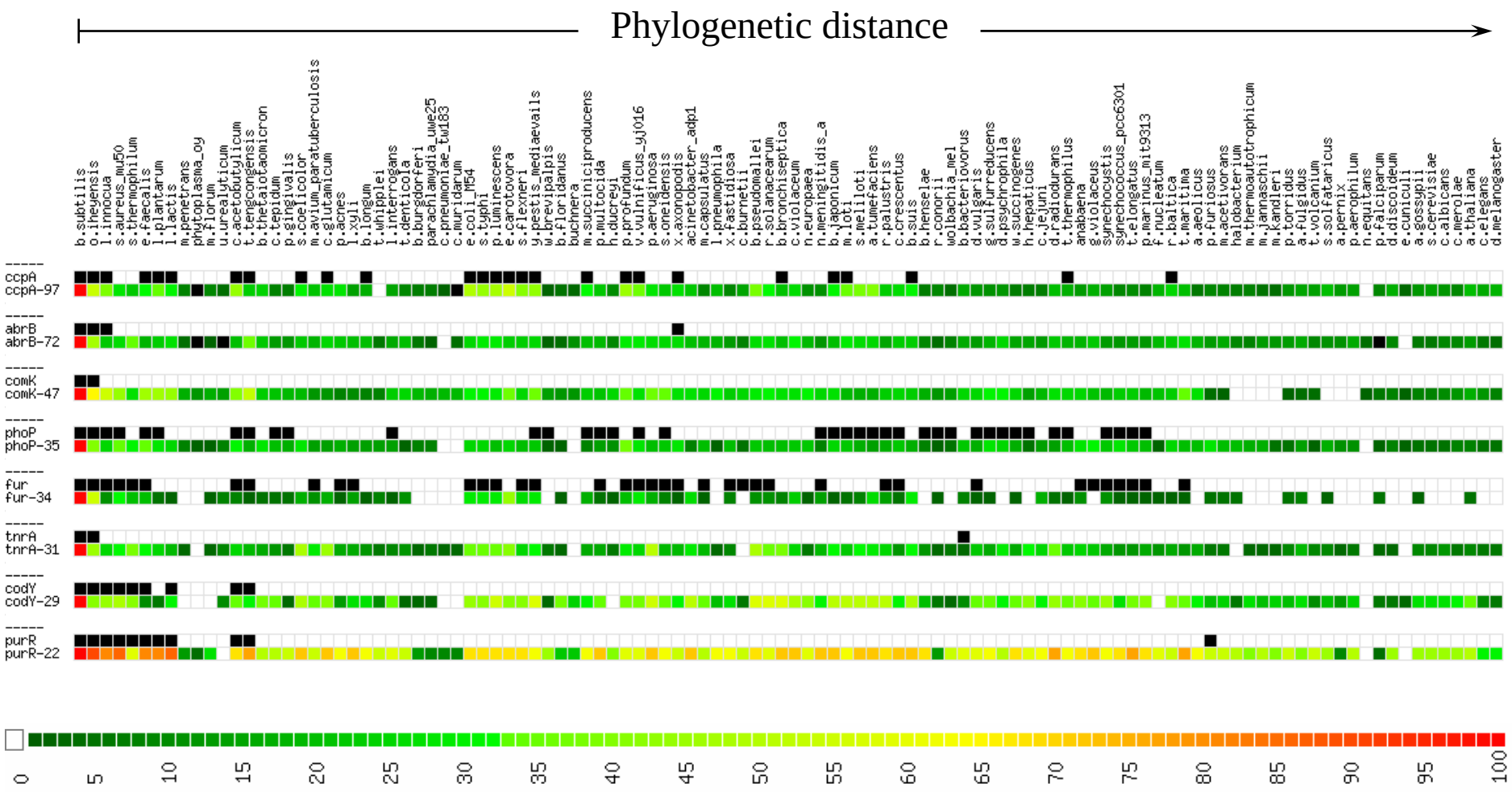

Phylogenetic distance
e.coli_M54
s.typhi
p.luminescens
e.carotovora
s.flexneri
y.pestis_mediaevails
w.brevipalpis
b.floridanus
buchnera
m.succiniciproducens
p.multocida
h.ducreyi
p.profundum
v.vulnificus_yj016
p.aeruginosa
s.oneidensis
x.axonopodis
acinetobacter_adp1
m.capsulatus
l.pneumophila
x.fastidiosa
c.burnetii
b.pseudomallei
r.solanacearum
b.bronchiseptica
c.violaceum
n.europaea
n.meningitidis_a
b.japonicum
m.loti
s.meliloti
a.tumefaciens
r.palustris
c.crescentus
b.suis
b.henselae
r.conorii
wolbachia_mel
b.bacteriovorus
d.vulgaris
g.sulfurreducens
d.psychrophila
w.succinogenes
h.hepaticus
c.jejuni
d.radiodurans
t.thermophilus
anabaena
g.violaceus
synechocystis
synechococcus_pcc6301
t.elongatus
p.marinus_mit9313
c.acetobutylicum
t.tengcongensis
b.subtilis
o.iheyensis
l.innocua
s.aureus_mu50
s.thermophilum
e.faecalis
l.plantarum
l.lactis
m.penetrans
phytoplasma_oy
m.florum
u.urealyticum
b.thetaiotaomicron
c.tepidum
p.gingivalis
s.coelicolor
m.avium_paratuberculosis
c.glutamicum
p.acnes
l.xyli
b.longum
t.whipplei
i.interrogans
t.denticola
b.burgdorferi
parachlamydia_uwe25
c.pneumoniae_tw183
c.muridarum
f.nucleatum
r.baltica
t.maritima
a.aeolicus
p.furiosus
m.acetivorans
halobacterium
m.thermoautotrophicum
m.jannaschii
m.kandleri
p.torridus
a.fulgidus
t.volcanium
s.solfataricus
a.pernix
p.aerophilum
n.equitans
p.falciparum
d.discoideum
e.cuniculi
a.gossypii
s.cerevisiae
c.albicans
c.merolae
a.thaliana
c.elegans
d.melanogaster
crp
crp-240
ihf
ihf-160
fnr
fnr-112
arcA
arcA-88
fis
fis-76
lrp
lrp-55
hns
hns-33
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

Figure 2b.


Phylogenetic distance
b.subtilis
o.iheyensis
l.innocua
s.aureus_mu50
s.thermophilum
e.faecalis
l.plantarum
l.lactis
m.penetrans
phytoplasma_oy
m.florum
u.urealyticum
c.acetobutylicum
t.tengcongensis
b.thetaiotaomicron
c.tepidum
p.gingivalis
s.coelicolor
m.avium_paratuberculosis
c.glutamicum
p.acnes
l.xyli
b.longum
t.whipplei
i.interrogans
t.denticola
b.burgdorferi
parachlamydia_uwe25
c.pneumoniae_tw183
c.muridarum
e.coli_k54
s.typhi
p.luminescens
e.carotovora
s.flexneri
y.pestis_mediaevalis
w.brevipalpis
b.floridanus
buchnera
m.succiniciproducens
p.multocida
h.ducreyi
p.profundum
v.vulnificus_yj016
p.aeruginosa
s.oneidensis
x.axonopodis
acinetobacter_adp1
m.capsulatus
l.pneumophila
x.fastidiosa
c.burnetii
b.pseudomallei
r.solanacearum
b.bronchiseptica
c.violaceum
n.europaea
n.meningitidis_a
b.japonicum
m.loti
s.meliloti
a.tumefaciens
r.palustris
c.crescentus
b.suis
b.henselae
r.conorii
wolbachia_mel
b.bacteriovorus
d.vulgaris
g.sulfurreducens
d.psychrophila
w.succinogenes
h.hepaticus
c.jejuni
d.radiodurans
t.thermophilus
anabaena
g.violaceus
synechocystis
synechococcus_pcc6301
t.elongatus
p.marinus_mit9313
f.nucleatum
r.baltica
t.maritima
a.aeolicus
p.furiosus
m.acetivorans
halobacterium
m.thermoautotrophicum
m.jannaschii
m.kandleri
p.torridus
a.fulgidus
t.volcanium
s.solfataricus
a.pernix
p.aerophilum
n.equitans
p.falciparum
d.discoideum
e.cuniculi
a.gossypii
s.cerevisiae
c.albicans
c.merolae
a.thaliana
c.elegans
d.melanogaster
ccpA
ccpA-97
abrB
abrB-72
comK
comK-47
phoP
phoP-35
fur
fur-34
tnrA
tnrA-31
codY
codY-29
purR
purR-22
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

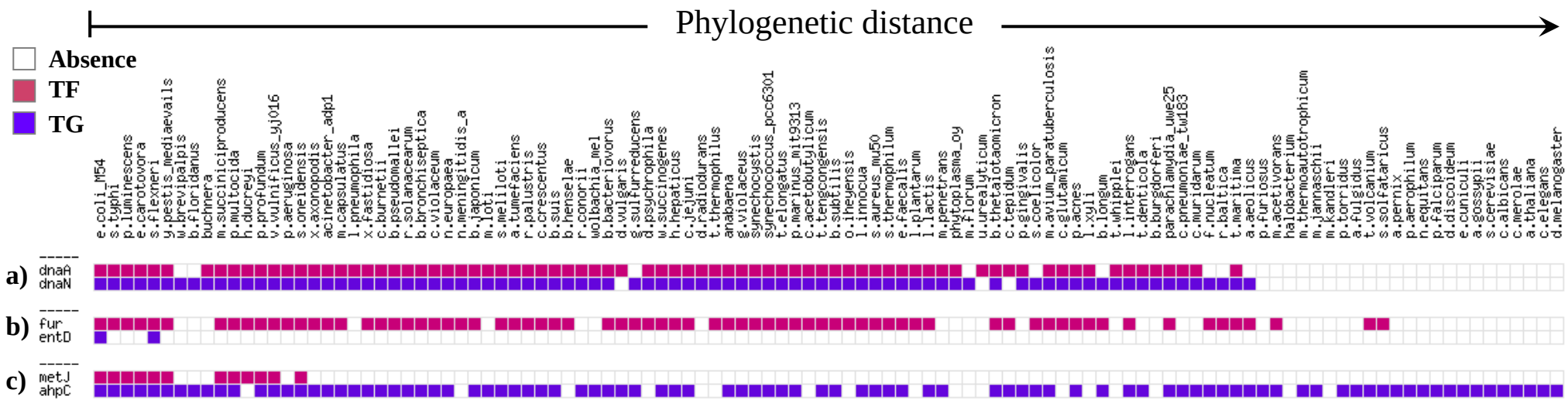
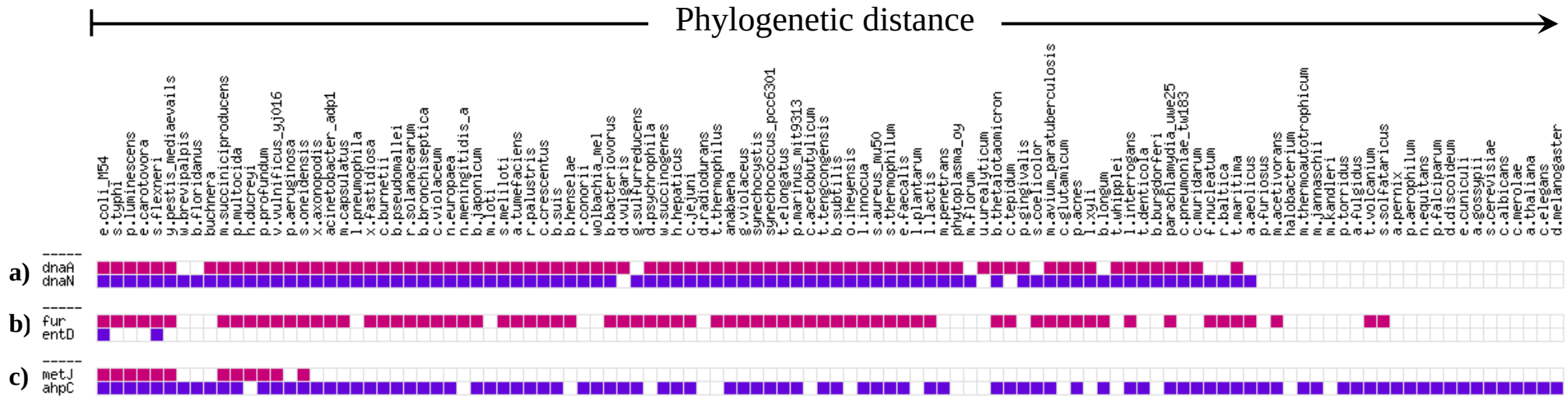
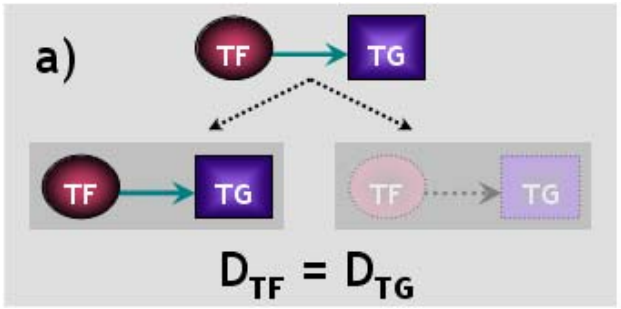
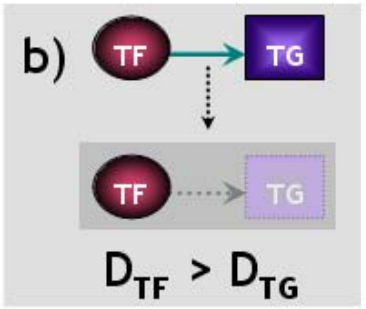

Phylogenetic distance
Absence
TF
TG
e.coli_M54
s.typhi
p.luminescens
e.carotovora
s.flexneri
y.pestis_mediaevalis
w.brevipalpis
b.floridanus
buchnera
m.succiniciproducens
p.multocida
h.ducreyi
p.profundum
v.vulnificus_yj016
p.aeruginosa
s.oneidensis
x.axonopodis
acinetobacter_adp1
m.capsulatus
l.pneumophila
x.fastidiosa
c.burnetii
b.pseudomallei
r.solanacearum
b.bronchiseptica
c.violaceum
n.europaea
n.meningitidis_a
b.japonicum
m.loti
s.meliloti
a.tumefaciens
r.palustris
c.crescentus
b.suis
b.henselae
r.conorii
wolbachia_mel
b.bacteriovorus
d.vulgaris
g.sulfurreducens
d.psychrophila
w.succinogenes
h.hepaticus
c.jejuni
d.radiodurans
t.thermophilus
anabaena
g.violaceus
synechocystis
synechococcus_pcc6301
t.elongatus
p.marinus_mit9313
c.acetobutylicum
t.tengcongensis
b.subtilis
o.iheyensis
l.innocua
s.aureus_mu50
s.thermophilum
e.faecalis
l.plantarum
l.lactis
m.penetrans
phytoplasma_oy
m.florum
u.urealyticum
b.thetaiotaomicron
c.tepidum
p.gingivalis
s.coelicolor
m.avium_paratuberculosis
c.glutamicum
p.acnes
l.xyli
b.longum
t.whipplei
l.interrogans
t.denticola
b.burgdorferi
parachlamydia_uwe25
c.pneumoniae_tw183
c.muridarum
f.nucleatum
r.baltica
t.maritima
a.aeolicus
p.furiosus
m.acetivorans
halobacterium
m.thermoautotrophicum
m.jannaschii
m.kandleri
p.torridus
a.fulgidus
t.volcanium
s.solfataricus
a.pernix
p.aerophilum
n.equitans
p.falciparum
d.discoideum
e.cuniculi
a.gossypii
s.cerevisiae
c.albicans
c.merolae
a.thaliana
c.elegans
d.melanogaster
a)
dnaA
dnaN
b)
fur
entD
c)
metJ
ahpC

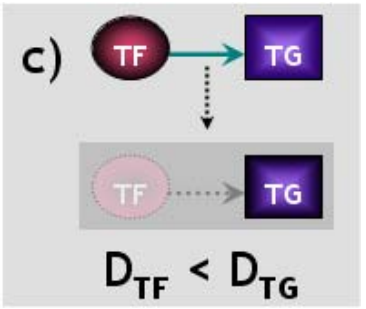
Figure 3.



- Absence
- TF - A
- TG - B







DTF = A' / (A' + A∩B)
DTG = B' / (B' + A∩B)
Where **A' = A - (A∩B)**
and **B' = B - (A∩B)**

DnaA = dnaN
D_{dnaA} = 3/(3+75) = 0.038
D_{dnaN} = 9/(9+75) = 0.107

Fur > entD
D_{Fur} = 70/(70+2) = 0.972
D_{entD} = 0/(0+2) = 0

MetJ < ahpC
D_{MetJ} = 1/(1+11)=0.083
D_{ahpC} = 81/(81+11)=0.88

Figure 4a.

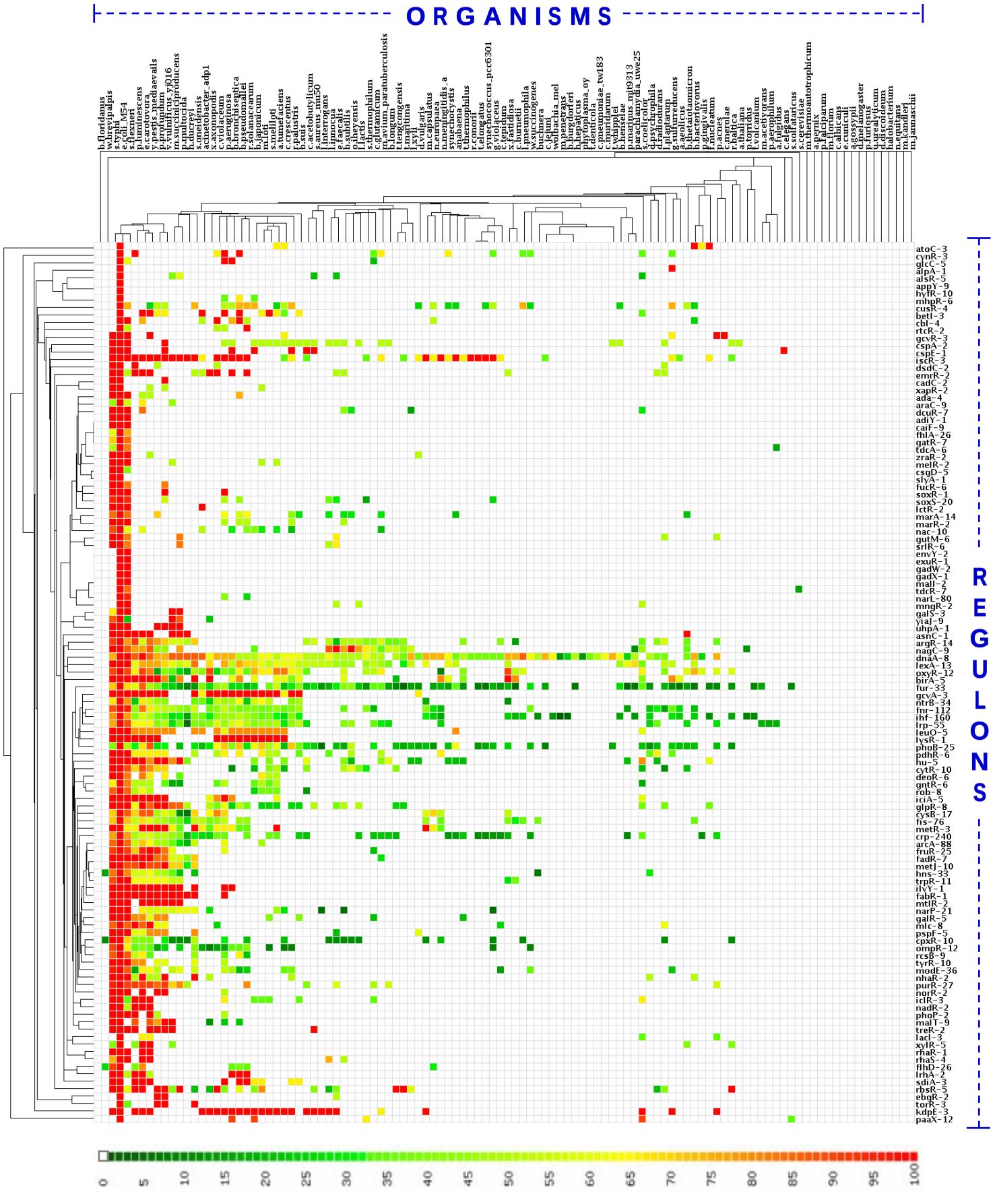
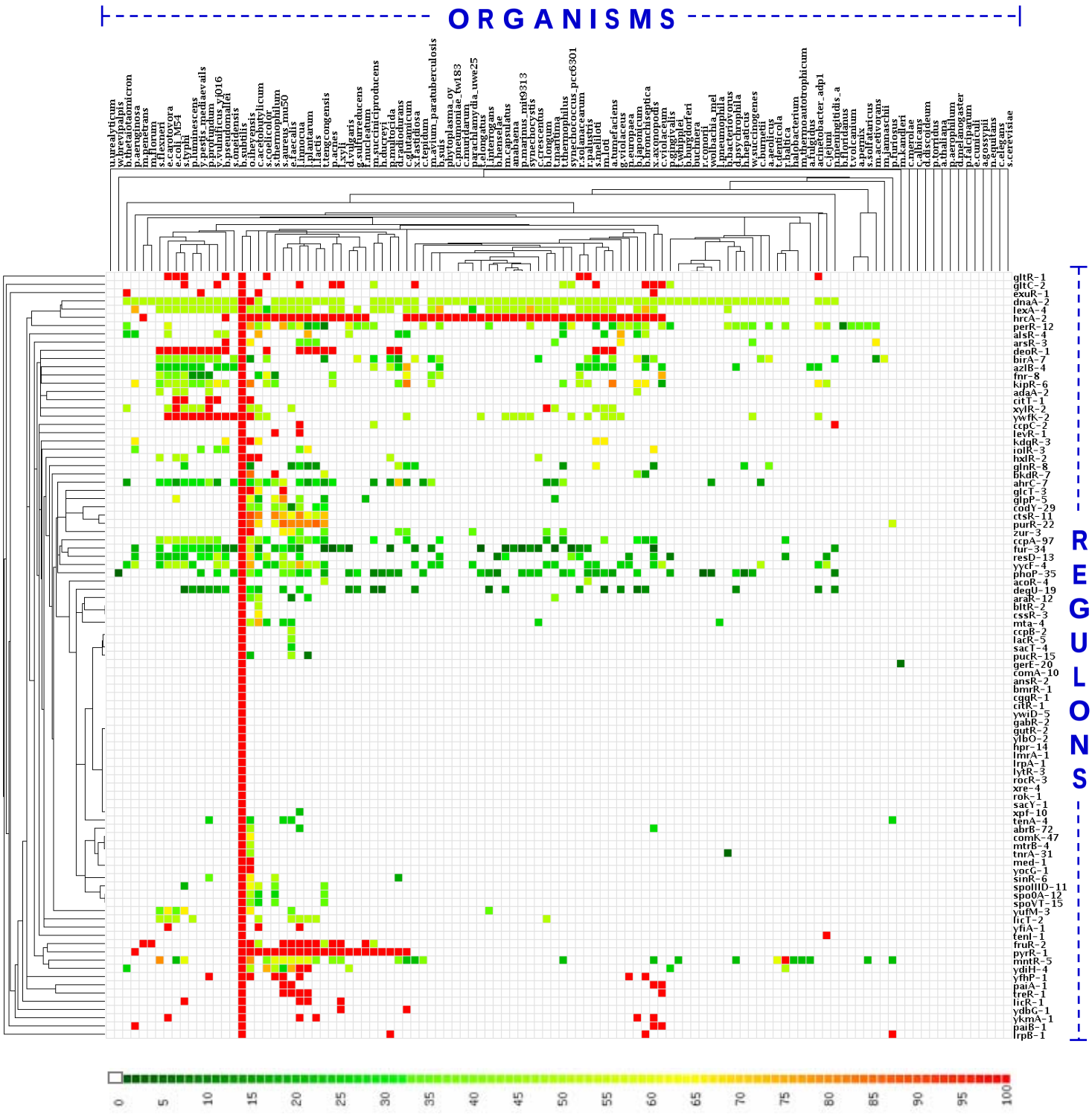

ORGANISMS
REGULONS
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

Figure 4b.


ORGANISMS
REGULONS
u.urealyticum
s.thermophilus
b.thetaiotaomicron
p.aeruginosa
m.genetranis
m.florum
s.aureus
e.carotovora
e.coli_MS4
s.typhi
p.luminescens
p.multocida_pm70
p.profundum
v.vulnificus_yj016
b.pseudomallei
s.oneidensis
h.pylori
o.iheyensis
c.acetobutylicum
s.coelicolor
s.thermophilum
s.aureus_mu50
e.faecalis
l.lactis
l.plantarum
l.johnsonii
t.tengcongensis
p.aceies
l.xyli
d.vulgaris
g.sulfurreducens
f.nucleatum
m.succiniciproducens
h.ducreyi
g.metallireducens
d.radiodurans
c.glutamicum
x.fastidiosa
c.tepidum
m.avium_paratuberculosis
l.innocua
phytoplasma_oy
c.pneumoniae_fw183
c.muridarum
p.carinii_jamydia_uwe25
l.interrogans
b.henselae
m.capsulatus
n.europaea
n.marinus_mit9313
synechocystis
c.crescentus
b.longum
b.parapertussis
t.thermophilus
synechococcus_pcc6301
r.solanacearum
r.palustris
s.flexneri
m.loti
a.tumefaciens
g.violaceus
n.europaea
b.japonicum
b.bronchiseptica
x.axonopodis
c.violaceum
f.johnsoniae
b.burgdorferi
buchnera
r.conorii
wolbachia_mel
l.pneumophila
b.bacteriovorus
d.psychrophila
h.hepaticus
w.succinogenes
c.burnetii
a.aeolicus
t.denticola
h.hepaticus
halobacterium
a.fulgidus
m.thermoautotrophicum
acinetobacter_adp1
c.jejuni
n.meningitidis_a
b.floridanus
t.volcanium
a.pernix
s.solfataricus
m.acetivorans
m.jannaschii
p.furiosus
m.kandleri
n.equitans
c.albicans
d.discoideum
p.torridus
a.thaliana
a.aeolicum
d.melanogaster
p.falciparum
e.cuniculi
a.gossypii
n.crassa
c.elegans
s.cerevisiae
gltR-1
gltC-2
exuR-1
dnaA-2
lexA-4
hrcA-2
perR-12
alsR-4
arsR-3
deoR-1
birA-7
azlB-4
fnr-8
kipR-6
adaA-2
citT-1
xylR-2
ywfK-2
ccpC-2
levR-1
kdgR-3
iolR-3
hxlR-2
glnR-8
bkdR-7
ahrC-7
glcT-3
glpP-5
codY-29
ctsR-11
purR-22
zur-3
ccpA-97
fur-34
resD-13
yycF-4
phoP-35
acoR-4
degU-19
araR-12
bltR-2
cssR-3
mta-4
ccpB-2
lacR-5
sacT-4
purR-15
gerE-20
comA-10
ansR-2
bmrR-1
cggR-1
ctlR-1
ywiD-5
gabR-2
gutR-2
ylbO-2
hpr-14
lmrA-1
lrpA-1
lytR-3
rocR-3
xre-4
rok-1
sacY-1
xpf-10
tenA-4
abrB-72
comK-47
mtrB-4
tnrA-31
med-1
yocG-1
sinR-6
spoIID-11
spo0A-12
spoVT-15
yufM-3
licT-2
yfiA-1
fenI-1
fruR-2
pyrR-1
mntR-5
ydhH-4
yfhP-1
paiA-1
treR-1
licR-1
ydbC-1
ykmA-1
paiB-1
lrpB-1
0
5
10
15
20
25
30
35
40
45
50
55
60
65
70
75
80
85
90
95
100

ANEXO 1.

(Also Supplementary material for the paper)

Additional information for the 204 complete genomes analyzed in this work. (Sources NCBI taxanomy). Those in blue are the list of non-redundant genomes used in the analysis. Each color in column 1 describes a phylogenetic clade.

PHYLO GENETIC CLADE	SPECIES NAME	ABBREVIATION	SIZE (nts)	SIZE (genes)	IN-PARALOG PROTENS	OXYGEN RESPONSE	LIFE STYLE	TEMPERATURE LEVELS	PATHOGEN
1 GAMMA (γ) PROTEO BACTERIA	**Escherichia coli K-12 MG1655**	e.coli M54	4639675	4289	51	facultative	host-associated	mesophilic	no
	Escherichia coli O157 Sakai	e.coli_o157j	5594477	5447	287	facultative	host-associated	mesophilic	yes
	Escherichia coli CFT073	e.coli_cft073	5231428	5379	116	facultative	host-associated	mesophilic	yes
	Escherichia coli O157 EDL933	e.coli_o157	5528445	5349	348	facultative	host-associated	mesophilic	yes
	Escherichia coli K-12 W3110	e.coli_j	4641433	4390	101				
	Salmonella typhi CT18	s.typhi	5133713	4767	80	facultative	host-associated	mesophilic	yes
	Salmonella typhimurium LT2	s.typhimurium	4951371	4553	37	facultative	host-associated	mesophilic	yes
	Salmonella enterica serovar typhi Ty2	s.typhi_ty2	4791961	4323	53	facultative	host-associated	mesophilic	yes
	Photorhabdus luminescens subsp. laumondii TTO1	p.luminescens	5688987	4683	190	facultative	host-associated	mesophilic	yes
	Erwinia carotovora subsp. atroseptica SCRI1043	e.carotovora	5064019	4472	20	facultative	multiple	mesophilic	yes
	Shigella flexneri 301 (serotype 2a)	s.flexneri	4828821	4180	514	facultative	host-associated	mesophilic	no
	Shigella flexneri 2457T (serotype 2a)	s.flexneri_2457t	4599354	4068	471	facultative	host-associated	mesophilic	yes
	Yersinia pestis biovar Mediaevails 91001	y.pestis_mediaevails	4803217	4142	168	facultative	multiple	mesophilic	yes
	Yersinia pestis KIM	y.pestis_kim	4701745	4090	158	facultative	multiple	mesophilic	yes
	Yersinia pestis CO92	y.pestis	4829855	4083	211	facultative	multiple	mesophilic	yes
	Yersinia pseudotuberculosis IP32953	y.pseudotuberculosis	4840899	4038	39	facultative	multiple	mesophilic	yes
	Wigglesworthia brevipalpis	w.brevipalpis	703004	611	0	facultative	host-associated	mesophilic	no
	Blochmannia floridanus	b.floridanus	705557	583	0	facultative	host-associated	mesophilic	yes
	Buchnera aphidicola APS	buchnera	655725	574	1	aerobic	host-associated	mesophilic	no
	Buchnera aphidicola Sg	b.aphidicola_sg	641454	504	0	aerobic	host-associated	mesophilic	no
	Buchnera aphidicola Bp	b.aphidicola_bp	615980	545	0	aerobic	host-associated	mesophilic	no
	Mannheimia succiniciproducens MBEL55E	m.succiniciproducens	2314078	2384	20	facultative	host-associated	mesophilic	no
	Pasteurella multocida PM70	p.multocida	2257487	2014	2	facultative	host-associated	mesophilic	yes
	Haemophilus ducreyi 35000HP	h.ducreyi	1698955	1717	28	anaerobic	host-associated	mesophilic	yes
	Haemophilus influenzae (serotype d)	h.influenzae	1830138	1709	10	facultative	host-associated	mesophilic	yes
	Photobacterium profundum	p.profundum	6403280	5413	317	anaerobic	multiple	mesophilic	yes
	Vibrio vulnificus YJ016	v.vulnificus_yj016	5260086	5028	69	facultative	aquatic	mesophilic	no
	Vibrio parahaemolyticus	v.parahaemolyticus	5165770	4832	29	facultative	aquatic	mesophilic	yes
	Vibrio vulnificus CMCP6	v.vulnificus	5126798	4537	88	facultative	aquatic	mesophilic	no
	Vibrio cholerae El Tor N16961	v.cholerae	4033464	3828	69	facultative	aquatic	mesophilic	yes
	Pseudomonas aeruginosa PA01	p.aeruginosa	6264403	5565	19	aerobic	multiple	mesophilic	no

	Pseudomonas syringae DC3000	p.syringae	6538260	5471	201	aerobic	multiple	mesophilic	yes
	Pseudomonas putida KT2440	p.putida	6181863	5350	54	aerobic	multiple	mesophilic	yes
	Shewanella oneidensis MR-1	s.oneidensis	5131416	4778	232	facultative	multiple	mesophilic	yes
	Xanthomonas axonopodis	x.axonopodis	5274174	4312	61	aerobic	host-associated	mesophilic	yes
	Xanthomonas campestris 8004	x.campestris	5148708	4181	93	aerobic	host-associated	mesophilic	yes
	Acinetobacter sp. ADP1	acinetobacter_adp1	3598621	3325	20	aerobic	multiple	mesophilic	
	Methylococcus capsulatus	m.capsulatus	3304561	2959	40				
	Legionella pneumophila Philadelphia 1	l.pneumophila	3397754	2942	19	aerobic	host-associated	mesophilic	yes
	Xylella fastidiosa 9a5c	x.fastidiosa	2731750	2831	93	aerobic	host-associated	mesophilic	yes
	Xylella fastidiosa Temecula1	x.fastidiosa_t	2521148	2036	59	aerobic	host-associated	mesophilic	yes
	Coxiella burnetii RSA 493	c.burnetii	2032674	2009	26	facultative	multiple	mesophilic	yes
2 BETA (β) PROTEO BACTERIA	**Burkholderia pseudomallei K96243**	b.pseudomallei	7247547	5728	58				
	Burkholderia mallei	b.mallei	5835527	4764	246				
	Ralstonia solanacearum	r.solanacearum	5810922	5116	131	aerobic	multiple	mesophilic	no
	Bordetella bronchiseptica RB50	b.bronchiseptica	5339179	4994	45	aerobic	host-associated	mesophilic	yes
	Bordetella parapertussis	b.parapertussis	4773551	4185	40	aerobic	host-associated	mesophilic	yes
	Bordetella pertussis	b.pertussis	4086189	3447	239	aerobic	host-associated	mesophilic	yes
	Chromobacterium violaceum	c.violaceum	4751080	4407	10	facultative	multiple	mesophilic	yes
	Nitrosomonas europaea	n.europaea	2812094	2461	117	aerobic	multiple	mesophilic	no
	Neisseria meningitidis Z2491 (serogroup A)	n.meningitidis_a	2184406	2065	51	aerobic	host-associated	mesophilic	no
	Neisseria meningitidis MC58 (serogroup B)	n.meningitidis	2272360	2025	82	aerobic	host-associated	mesophilic	no
3 ALFA (α) PROTEO BACTERIA	**Bradyrhizobium japonicum**	b.japonicum	9105828	8317	104	aerobic	host-associated	mesophilic	no
	Mesorhizobium loti	m.loti	7596297	7281	65	aerobic	multiple	mesophilic	no
	Sinorhizobium meliloti	s.meliloti	6691694	6205	114	aerobic	multiple	mesophilic	yes
	Agrobacterium tumefaciens C58 (UWash/Dupont)	a.tumefaciens	5674064	5402	18	aerobic	multiple	mesophilic	yes
	Agrobacterium tumefaciens C58 (Cereon)	a.tumefaciens_c	5673465	5301	24	aerobic	multiple	mesophilic	yes
	Rhodopseudomonas palustris CGA009	r.palustris	5459213	4814	14	facultative	multiple	mesophilic	yes
	Caulobacter crescentus CB15	c.crescentus	4016947	3737	29	aerobic	aquatic	mesophilic	no
	Brucella suis 1330	b.suis	3315175	3264	19	aerobic	host-associated	mesophilic	yes
	Brucella melitensis 16M	b.melitensis	3294931	3198	19	aerobic	host-associated	mesophilic	yes
	Bartonella henselae	b.henselae	1931047	1488	46				
	Bartonella quintana Toulouse	b.quintana	1581384	1142	5				
	Rickettsia conorii Malish 7	r.conorii	1268755	1374	1	aerobic	host-associated	mesophilic	yes
	Rickettsia typhi	r.typhi	1111496	838	2				
	Rickettsia prowazekii	r.prowazekii	1111523	834	0	aerobic	host-associated	mesophilic	yes
	Wolbachia wMel	wolbachia_mel	1267782	1195	42				
4 DELTA (δ) PROTEO BACTERIA	**Bdellovibrio bacteriovorus HD100**	b.bacteriovorus	3782950	3583	0				
	Desulfovibrio vulgaris Hildenborough	d.vulgaris	3773159	3531	41				
	Geobacter sulfurreducens	g.sulfurreducens	3814139	3445	36				
	Desulfotalea psychrophila	d.psychrophila	3659634	3236	20				

5 EPSILON PROTEO BACTERIA	**Wolinella succinogenes DSMZ 1740**	w.succinogenes	2110355	2044	21	microaerophilic	host-associated	mesophilic	yes
	Helicobacter hepaticus	h.hepaticus	1799146	1875	3	aerobic	host-associated	mesophilic	yes
	Helicobacter pylori 26695	h.pylori	1667867	1566	18	aerobic	host-associated	mesophilic	yes
	Helicobacter pylori J99	h.pylori_j99	1643831	1491	4	aerobic	host-associated	mesophilic	yes
	Campylobacter jejuni NCTC11168	c.jejuni	1641481	1634	2	microaerophilic	multiple	mesophilic	yes
6 THERMUS/ DEINOCOCCUS	**Deinococcus radiodurans R1**	d.radiodurans	3284156	3102	32	aerobic	soil	mesophilic	no
	Thermus thermophilus HB27	t.thermophilus	2127482	2210	20				
7 CYANO BACTERIA	**Anabaena sp. PCC7120**	anabaena	7211789	6131	91				
	Gloeobacter violaceus	g.violaceus	4659019	4430	33		soil	mesophilic	no
	Synechocystis sp. PCC6803	synechocystis	3573470	3264	61		aquatic	mesophilic	no
	Synechococcus sp. PCC6301	synechococcus_pcc6301	2696255	2525	5				
	Synechococcus sp. WH8102	synechococcus_wh8102	2434428	2517	7		aquatic	mesophilic	no
	Thermosynechococcus elongatus BP-1	t.elongatus	2593857	2475	78		specialized	thermophilic	yes
	Prochlorococcus marinus MIT9313	p.marinus_mit9313	2410873	2265	9		aquatic	mesophilic	
	Prochlorococcus marinus SS120	p.marinus	1751080	1882	2		aquatic	mesophilic	
	Prochlorococcus marinus MED4	p.marinus_med4	1657990	1712	5		aquatic	mesophilic	
8 CLOSTRIDIUM	**Clostridium acetobutylicum**	c.acetobutylicum	4132880	3848	6	anaerobic	multiple	mesophilic	no
	Clostridium perfringens	c.perfringens	3085740	2723	5	anaerobic	multiple	mesophilic	yes
	Clostridium tetani E88	c.tetani	2873333	2373	27	anaerobic	multiple	mesophilic	yes
	Thermoanaerobacter tengcongensis	t.tengcongensis	2689445	2588	81	anaerobic	specialized	hyperthermophilic	no
9 BACILLALES	**Bacillus subtilis**	b.subtilis	4214630	4106	4	facultative	soil	mesophilic	no
	Bacillus licheniformis DSM13	b.licheniformis_dsm13	4222645	4196	20				
	Bacillus licheniformis ATCC 14580	b.licheniformis	4222334	4161	19				
	Bacillus halodurans	b.halodurans	4202352	4066	82	facultative	multiple	mesophilic	no
	Bacillus anthracis A2012	b.anthracis_a2012	5370060	5852	7	facultative	multiple	mesophilic	yes
	Bacillus anthracis Ames 0581	b.anthracis_ames0581	5503926	5558	8	facultative	soil	mesophilic	yes
	Bacillus anthracis Ames	b.anthracis	5227293	5311	8				
	Bacillus anthracis Sterne	b.anthracis_sterne	5228663	5287	7				
	Bacillus cereus ATCC 10987	b.cereus_atcc10987	5432652	5603	16	aerobic	soil	mesophilic	yes
	Bacillus cereus ATCC 14579	b.cereus	5427083	5255	27				
	Bacillus cereus ZK	b.cereus_zk	5843235	5134	6				
	Bacillus thuringiensis 97-27	b.thuringiensis	5314794	5117	41				
	Oceanobacillus iheyensis HTE831	o.iheyensis	3630528	3496	15	aerobic	multiple	mesophilic	yes
	Listeria innocua CLIP 11262 (serotype 6a)	l.innocua	3093113	3043	38	facultative	multiple	mesophilic	yes
	Listeria monocytogenes EGD-e	l.monocytogenes	2944528	2846	3	facultative	multiple	mesophilic	yes
	Listeria monocytogenes F2365	l.monocytogenes_f2365	2905310	2821	6				
	Staphylococcus aureus Mu50	s.aureus_mu50	2903636	2748	35	facultative	host-associated	mesophilic	yes
	Staphylococcus aureus MRSA252	s.aureus_mrsa252	2902619	2656	34				
	Staphylococcus aureus MW2	s.aureus_mw2	2820462	2632	5	facultative	host-associated	mesophilic	yes
	Staphylococcus aureus N315	s.aureus_n315	2839469	2623	38	facultative	host-associated	mesophilic	yes

	Staphylococcus aureus MSSA476	s.aureus_mssa476	2820454	2579	3				
	Staphylococcus epidermidis ATCC 12228	s.epidermidis	2564615	2419	38	facultative	host-associated	mesophilic	yes
10 LACTO BACILLALES	**Streptococcus thermophilum**	s.thermophilum		3337	563	facultative	host-associated	mesophilic	yes
	Streptococcus agalactiae 2603	s.agalactiae	2160267	2124	25	facultative	host-associated	mesophilic	yes
	Streptococcus agalactiae NEM316	s.agalactiae_nem316	2211485	2094	100	facultative	host-associated	mesophilic	yes
	Streptococcus pneumoniae TIGR4	s.pneumoniae	2160837	2094	35	facultative	multiple	mesophilic	yes
	Streptococcus pneumoniae R6	s.pneumoniae_r6	2038615	2043	42	facultative	multiple	mesophilic	yes
	Streptococcus mutans	s.mutans	2030921	1960	12	facultative	host-associated	mesophilic	yes
	Streptococcus pyogenes MGAS10394 (serotype M6)	s.pyogenes_mgas10394	1899877	1886	23				
	Streptococcus pyogenes MGAS315 (serotype M3)	s.pyogenes_m3	1900521	1865	31	facultative	host-associated	mesophilic	yes
	Streptococcus pyogenes SSI-1 (serotype M3)	s.pyogenes_ssi1	1894275	1861	26	facultative	host-associated	mesophilic	yes
	Streptococcus pyogenes MGAS8232 (serotype M18)	s.pyogenes_m18	1895017	1845	28	facultative	host-associated	mesophilic	yes
	Streptococcus pyogenes SF370 (serotype M1)	s.pyogenes	1852441	1696	14	facultative	host-associated	mesophilic	yes
	Streptococcus pyogenes MGAS6180 (serotype M28)	e.faecalis	1897573	3113	23	facultative	multiple	mesophilic	yes
	Lactobacillus plantarum	l.plantarum	3348625	3009	20	facultative	host-associated	mesophilic	no
	Lactobacillus johnsonii	l.johnsonii	1992676	1821	12				
	Lactococcus lactis	l.lactis	2365589	2266	72	facultative	multiple	mesophilic	no
11 MOLLICUTES	**Mycoplasma penetrans**	m.penetrans	1358633	1037	27	facultative	host-associated	mesophilic	yes
	Mycoplasma mycoides	m.mycoides	1211703	1016	130				
	Mycoplasma pulmonis	m.pulmonis	963879	782	6	facultative	host-associated	mesophilic	yes
	Mycoplasma gallisepticum	m.gallisepticum	996422	726	7	facultative	host-associated	mesophilic	yes
	Mycoplasma pneumoniae	m.pneumoniae	816394	688	3	facultative	host-associated	mesophilic	no
	Mycoplasma mobile	m.mobile	777079	633	4				
	Mycoplasma genitalium	m.genitalium	580074	480	0	facultative	host-associated	mesophilic	yes
	Onion yellows phytoplasma OY-M	phytoplasma_oy	860631	754	79				
	Mesoplasma florum	m.florum	793224	683	0				
	Ureaplasma urealyticum	u.urealyticum	751719	611	2	facultative	host-associated	mesophilic	yes
12 CFB/ GREEN SULFUR BACTERIA	**Bacteroides thetaiotaomicron**	b.thetaiotaomicron	6293399	4778	69	anaerobic	host-associated	mesophilic	
	Bacteroides fragilis YCH46	b.fragilis	5310990	4625	24				
	Chlorobium tepidum TLS	c.tepidum	2154946	2252	8	anaerobic	specialized	thermophilic	no
	Porphyromonas gingivalis W83	p.gingivalis	2343476	1909	71	anaerobic	host-associated	mesophilic	yes
13 ACTINO BACTERIA	**Streptomyces coelicolor**	s.coelicolor	9054847	7897	128	aerobic	multiple	mesophilic	no
	Streptomyces avermitilis	s.avermitilis	9119895	7671	28	aerobic	multiple	mesophilic	no
	Mycobacterium avium paratuberculosis	m.avium_paratuberculosis	4829781	4350	41				
	Mycobacterium tuberculosis CDC1551	m.tuberculosis_cdc1551	4403837	4187	32	aerobic	host-associated	mesophilic	yes
	Mycobacterium tuberculosis H37Rv	m.tuberculosis	4411532	3869	58	aerobic	host-associated	mesophilic	yes
	Mycobacterium bovis	m.bovis	4345492	3920	25	aerobic	host-associated	mesophilic	yes

	Mycobacterium leprae	m.leprae	3268203	1605	5	aerobic	host-associated	mesophilic	yes
	Corynebacterium glutamicum (Kitasato)	c.glutamicum	3309401	2993	31	facultative	multiple	mesophilic	no
	Corynebacterium efficiens	c.efficiens	3147090	2950	81	facultative	multiple	mesophilic	no
	Corynebacterium diphtheriae	c.diphtheriae	2488635	2272	12	aerobic	multiple	mesophilic	yes
	Propionibacterium acnes	p.acnes	2560265	2297	4				
	Leifsonia xyli xyli CTCB07	l.xyli	2584158	2030	36				
	Bifidobacterium longum	b.longum	2260266	1729	8	anaerobic	host-associated	mesophilic	no
	Tropheryma whipplei Twist	t.whipplei	927303	808	4	aerobic	host-associated	mesophilic	yes
	Tropheryma whipplei TW08/27	t.whipplei_s	925938	783	5	aerobic	host-associated	mesophilic	yes
	Leptospira interrogans serovar lai	l.interrogans	4691184	4727	79	aerobic	host-associated	mesophilic	yes
	Leptospira interrogans serovar Copenhageni	l.interrogans_copenhageni	4627366	3660	32				
	Treponema denticola	t.denticola	2843201	2767	23				
	Treponema pallidum	t.pallidum	1138011	1031	2	anaerobic	host-associated	mesophilic	yes
14 SPIROCHETE	**Borrelia burgdorferi**	b.burgdorferi	1519856	1637	221	aerobic	host-associated	mesophilic	yes
	Borrelia garinii	b.garinii	986914	832	0				
	Parachlamydia sp. UWE25	parachlamydia_uwe25	2414465	2031	26				
	Chlamydophila pneumoniae TW183	c.pneumoniae_tw183	1225935	1113	0		host-associated	mesophilic	yes
	Chlamydophila pneumoniae AR39	c.pneumoniae_ar39	1229853	1110	3		host-associated	mesophilic	yes
	Chlamydophila pneumoniae J138	c.pneumoniae_j138	1226565	1069	0		host-associated	mesophilic	yes
	Chlamydophila pneumoniae CWL029	c.pneumoniae	1230230	1052	2		host-associated	mesophilic	yes
	Chlamydophila caviae	c.caviae	1181356	1005	0		host-associated	mesophilic	yes
15 CHLAMYDIA	**Chlamydia muridarum**	c.muridarum	1080451	911	0		host-associated	mesophilic	yes
	Chlamydia trachomatis serovar D	c.trachomatis	1042519	894	0		host-associated	mesophilic	yes
16 PLANCTOMYCES/ FUSOBACTERIA	**Fusobacterium nucleatum**	f.nucleatum	2174500	2067	34	anaerobic	host-associated	mesophilic	yes
	Rhodopirellula baltica	r.baltica	7145576	7325	92				
17 TERMOTOGALES/ AQUIFICACEAE	**Thermotoga maritima**	t.maritima	1860725	1846	9	anaerobic	specialized	hyperthermophilic	no
	Aquifex aeolicus	a.aeolicus	1590791	1553	5	aerobic	specialized	hyperthermophilic	no
	Pyrococcus furiosus DSM 3638	p.furiosus	1908256	2065	24	anaerobic	aquatic	hyperthermophilic	no
	Pyrococcus horikoshii OT3	p.horikoshii	1738505	2061	1	anaerobic	aquatic	hyperthermophilic	yes
18 EURYARCHAEOTA	Pyrococcus abyssi GE5	p.abyssi	1768562	1765	1	anaerobic	aquatic	hyperthermophilic	no
	Methanosarcina acetivorans C2A	m.acetivorans	5751492	4540	188	anaerobic	aquatic	mesophilic	no
	Methanosarcina mazei Goe1	m.mazei	4096345	3371	102	anaerobic	multiple	mesophilic	no
	Halobacterium sp. NRC-1	halobacterium	2571010	2605	192				
	Methanobacterium thermoautotrophicum deltaH	m.thermoautotrophicum	1751377	1869	3	anaerobic	specialized	hyperthermophilic	no
	Methanococcus jannaschii DSM2661	m.jannaschii	1739927	1770	6	anaerobic	aquatic	hyperthermophilic	no
	Methanococcus maripaludis S2	m.maripaludis	1661137	1722	0				
	Methanopyrus kandleri AV19	m.kandleri	1694969	1687	0	anaerobic	specialized	hyperthermophilic	no
	Picrophilus torridus DSM 9790	p.torridus	1545895	1535	0				

	Archaeoglobus fulgidus VC-16	a.fulgidus	2178400	2407	19	anaerobic	aquatic	hyperthermophilic	no
	Thermoplasma volcanium GSS1	t.volcanium	1584804	1526	9	facultative	specialized	thermophilic	no
	Thermoplasma acidophilum DSM 1728	t.acidophilum	1564906	1478	2	facultative	specialized	thermophilic	no
19 CRENARCHAEOTA	**Sulfolobus solfataricus**	s.solfataricus	2992245	2977	204	aerobic	specialized	hyperthermophilic	no
	Sulfolobus tokodaii	s.tokodaii	2694756	2826	53	aerobic	specialized	hyperthermophilic	no
	Aeropyrum pernix K1	a.pernix	1669695	2694	1	aerobic	specialized	hyperthermophilic	no
	Pyrobaculum aerophilum IM2	p.aerophilum	2222430	2605	17	facultative	aquatic	hyperthermophilic	no
	Nanoarchaeum equitans Kin4-M	n.equitans	490885	536	0	anaerobic	host-associated	hyperthermophilic	yes
20 PROTISTS	**Plasmodium falciparum 3D7**	p.falciparum	21847047	5265	40				
	Dictyostelium discoideum	d.discoideum		12173	489				
21 FUNGI	**Encephalitozoon cuniculi**	e.cuniculi	2497519	1996	124				
	Ashbya gossypii (Eremothecium gossypii)	a.gossypii	8764998	4726	14				
	Saccharomyces cerevisiae	s.cerevisiae	12156590	5855	235				
	Candida albicans	c.albicans		6383	200				
22 PLANTAS	**Cyanidioschyzon merolae**	c.merolae		4772	96				
	Arabidopsis thaliana	a.thaliana	119707899	28159	1716				
23 ANIMALS	**Caenorhabditis elegans**	c.elegans	100283706	21357	1755				
	Drosophila melanogaster	d.melanogaster	118377116	16130	1848				

ANEXO 2A.

Date sets of regulons used in the entire analysis from *Escherichia coli* from Regulon DataBase v4.0: http://regulondb.ccg.unam.mx/index.html

Transcription Factor Name|Genebank|Bnumber	TGs Number	Target Genes Name|Genebank|Bnumber
ada|16130150|b2213	**4**	accA|16128178|b0185 aidB|16132009|b4187 alkA|16130008|b2068 alkB|16130149|b2212
adiY|16131942|b4116	**1**	adiA|16131943|b4117
alpA|16130542|b2624	**1**	slp|16131378|b3506
alsR|16131915|b4089	**5**	alsA|16131913|b4087 alsB|16131914|b4088 alsC|16131912|b4086 alsE|16131911|b4085 alsI|16131916|b4090
appY|16128547|b0564	**9**	appA|16128946|b0980 appB|16128945|b0979 appC|16128944|b0978 hyaA|16128938|b0972 hyaB|16128939|b0973 hyaC|16128940|b0974 hyaD|16128941|b0975 hyaE|16128942|b0976 hyaF|16128943|b0977
araC|16128058|b0064	**9**	araA|16128056|b0062 araB|16128057|b0063 araD|16128055|b0061 araE|16130745|b2841 araF|16129851|b1901 araG|16129850|b1900 araH_1|b1899|b1899 araH_2|b1898|b1898 araJ|16128381|b0396
arcA|16132218|b4401	**88**	aceA|16131841|b4015 aceB|16131840|b4014 aceK|16131842|b4016 acnA|16129237|b1276 acnB|16128111|b0118 aldA|16129376|b1415 b0725|16128700|b0725 betA|16128296|b0311 betB|16128297|b0312 betI|16128298|b0313 betT|16128299|b0314 cadA|16131957|b4131 cadB|16131958|b4132 caiA|16128033|b0039 caiB|16128032|b0038 caiC|16128031|b0037 caiD|16128030|b0036 caiE|16128029|b0035 caiT|16128034|b0040 cydA|16128708|b0733 cydB|16128709|b0734 cyoA|16128417|b0432 cyoB|16128416|b0431 cyoC|16128415|b0430 cyoD|16128414|b0429 cyoE|16128413|b0428 dctA|16131400|b3528 fadA|16131691|b3845 fadB|16131692|b3846 focA|16128871|b0904 fumA|16129570|b1612 fumC|16129569|b1611 glcA|16130875|b2975 glcB|16130876|b2976 glcD|16130879|b2979 glcE|16130878|b2978 glcG|16130877|b2977 glpA|16130176|b2241 glpB|16130177|b2242 glpC|16130178|b2243 gltA|16128695|b0720 hemA|16129173|b1210 hyaA|16128938|b0972 hyaB|16128939|b0973 hyaC|16128940|b0974 hyaD|16128941|b0975 hyaE|16128942|b0976 hyaF|16128943|b0977 hyb0|16130897|b2997 hybA|16130896|b2996 hybB|16130895|b2995 hybC|16130894|b2994 hybD|16130893|b2993 hybE|16130892|b2992 hybF|16130891|b2991 hybG|16130890|b2990 icdA|16129099|b1136 lctD|16131476|b3605 lctP|16131474|b3603 lctR|16131475|b3604 lpdA|16128109|b0116 mdh|16131126|b3236 moeA|16128795|b0827 moeB|16128794|b0826 ndh|16129072|b1109 nuoA|16130223|b2288 nuoB|16130222|b2287 nuoC|16130221|b2286 nuoE|16130220|b2285 nuoF|16130219|b2284 nuoG|16130218|b2283 nuoH|16130217|b2282 nuoI|16130216|b2281 nuoJ|16130215|b2280 nuoK|16130214|b2279 nuoL|16130213|b2278 nuoM|16130212|b2277 nuoN|16130211|b2276 pflB|16128870|b0903 sdhA|16128698|b0723 sdhB|16128699|b0724 sdhC|16128696|b0721 sdhD|16128697|b0722 sodA|16131748|b3908 sucA|16128701|b0726 sucB|16128702|b0727 sucC|16128703|b0728 sucD|16128704|b0729
argR|16131127|b3237	**14**	argB|16131797|b3959 argC|16131796|b3958 argD|16131238|b3359 argE|16131795|b3957 argF|16128258|b0273 argH|16131798|b3960 argI|16132076|b4254 astA|16129701|b1747 astB|16129699|b1745 astC|16129702|b1748 astD|16129700|b1746 astE|16129698|b1744 carA|16128026|b0032 carB|16128027|b0033
asnC|16131611|b3743	**1**	asnA|16131612|b3744
atoC|16130157|b2220	**3**	atoA|16130159|b2222 atoD|16130158|b2221 atoE|16130160|b2223
betI|16128298|b0313	**3**	betA|16128296|b0311 betB|16128297|b0312 betT|16128299|b0314
birA|16131807|b3973	**5**	bioA|16128742|b0774 bioB|16128743|b0775 bioC|16128745|b0777 bioD|16128746|b0778 bioF|16128744|b0776
cadC|16131959|b4133	**2**	cadA|16131957|b4131 cadB|16131958|b4132
caiF|16128028|b0034	**9**	caiA|16128033|b0039 caiB|16128032|b0038 caiC|16128031|b0037 caiD|16128030|b0036 caiE|16128029|b0035 caiT|16128034|b0040 fixB|16128036|b0042 fixC|16128037|b0043 fixX|16128038|b0044
cbl|16129929|b1987	**4**	tauA|16128350|b0365 tauB|16128351|b0366 tauC|16128352|b0367 tauD|16128353|b0368
cpxR|16131752|b3912	**10**	csgA|16129005|b1042 csgB|16129004|b1041 csgD|16129003|b1040 csgE|16129002|b1039 csgF|16129001|b1038 csgG|16129000|b1037

		dsbA|16131701|b3860 ecfI|16131556|b3688 ppiA|16131242|b3363 yihE|16131700|b3859
crp|16131236|b3357	**240**	acnA|16129237|b1276 acnB|16128111|b0118 acs|16131895|b4069 aldA|16129376|b1415 aldB|16131459|b3588 ansB|16130858|b2957 araA|16128056|b0062 araB|16128057|b0063 araC|16128058|b0064 araD|16128055|b0061 araE|16130745|b2841 araF|16129851|b1901 araG|16129850|b1900 araH_1|b1899|b1899 araH_2|b1898|b1898 araJ|16128381|b0396 aspA|16131964|b4139 b0725|16128700|b0725 bglG|16131591|b3723 caiA|16128033|b0039 caiB|16128032|b0038 caiC|16128031|b0037 caiD|16128030|b0036 caiE|16128029|b0035 caiF|16128028|b0034 caiT|16128034|b0040 cirA|16130093|b2155 cpdB|16132035|b4213 crr|16130343|b2417 csiD|16130573|b2659 csiE|16130460|b2535 cyaA|16131658|b3806 dadA|16129152|b1189 dadX|16129153|b1190 dctA|16131400|b3528 dcuA|16131963|b4138 dcuB|16131949|b4123 deoA|16132199|b4382 deoB|16132200|b4383 deoC|16132198|b4381 deoD|16132201|b4384 dsdA|16130298|b2366 dsdX|16130297|b2365 dusB|16131148|b3260 ebgA|16130971|b3076 ebgC|16130972|b3077 envZ|16131281|b3404 epd|16130828|b2927 fadL|16130277|b2344 fis|16131149|b3261 fixB|16128036|b0042 fixC|16128037|b0043 fixX|16128038|b0044 flhC|16129843|b1891 flhD|16129844|b1892 focA|16128871|b0904 fucA|16130707|b2800 fucI|16130709|b2802 fucK|16130710|b2803 fucO|16130706|b2799 fucP|16130708|b2801 fucR|16130712|b2805 fucU|16130711|b2804 fumB|16131948|b4122 fur|16128659|b0683 gadA|16131389|b3517 gadX|16131388|b3516 galE|16128727|b0759 galK|16128725|b0757 galM|16128724|b0756 galS|16130089|b2151 galT|16128726|b0758 gcd|16128117|b0124 glgA|16131303|b3429 glgC|16131304|b3430 glgP|16131302|b3428 glgS|16130945|b3049 glnA|16131710|b3870 glnG|16131708|b3868 glnL|16131709|b3869 glpA|16130176|b2241 glpB|16130177|b2242 glpC|16130178|b2243 glpD|16131300|b3426 glpE|16131299|b3425 glpF|16131765|b3927 glpG|16131298|b3424 glpK|16131764|b3926 glpQ|16130174|b2239 glpR|16131297|b3423 glpT|16130175|b2240 gltA|16128695|b0720 gntK|16131309|b3437 gntT|16131291|b3415 gntU_1|b3436|b3436 gntU_2|b3435|b3435 guaA|16130432|b2507 guaB|16130433|b2508 gutM|16130613|b2706 gutQ|16130615|b2708 gyrA|16130166|b2231 hlyE|16129145|b1182 hupA|16131830|b4000 hupB|16128425|b0440 idnD|16132089|b4267 idnO|16132088|b4266 idnR|16132086|b4264 idnT|16132087|b4265 ilvB|16131541|b3671 ilvN|16131540|b3670 ivbL|16131542|b3672 lacA|16128327|b0342 lacY|16128328|b0343 lacZ|16128329|b0344 lamB|16131862|b4036 lpdA|16128109|b0116 lyxK|16131451|b3580 malE|16131860|b4034 malF|16131859|b4033 malG|16131858|b4032 malI|16129578|b1620 malK|16131861|b4035 malM|16131863|b4037 malP|16131293|b3417 malQ|16131292|b3416 malS|16131442|b3571 malT|16131294|b3418 malX|16129579|b1621 malY|16129580|b1622 manX|16129771|b1817 manY|16129772|b1818 manZ|16129773|b1819 mdh|16131126|b3236 melA|16131945|b4119 melB|16131946|b4120 melR|16131944|b4118 mglA|16130087|b2149 mglB|16130088|b2150 mglC|16130086|b2148 mhpA|16128332|b0347 mhpB|16128333|b0348 mhpC|16128334|b0349 mhpD|16128335|b0350 mhpE|16128337|b0352 mhpF|16128336|b0351 mpl|16132055|b4233 mtlA|16131470|b3599 mtlD|16131471|b3600 mtlR|16131472|b3601 nagA|16128653|b0677 nagB|16128654|b0678 nagC|16128652|b0676 nagD|16128651|b0675 nagE|16128655|b0679 nmpC|16128536|b0553 nupC|16130325|b2393 nupG|16130865|b2964 ompA|16128924|b0957 ompR|16131282|b3405 osmY|16132194|b4376 oxyR|16131799|b3961 paaA|16129349|b1388 paaB|16129350|b1389 paaC|16129351|b1390 paaD|16129352|b1391 paaE|16129353|b1392 paaF|16129354|b1393 paaG|16129355|b1394 paaH|16129356|b1395 paaI|16129357|b1396 paaJ|16129358|b1397 paaK|16129359|b1398 paaZ|16129348|b1387 pflB|16128870|b0903 pgk|16130827|b2926 ppiA|16131242|b3363 proP|16131937|b4111 ptsG|16129064|b1101 ptsH|16130341|b2415 ptsI|16130342|b2416 putP|16128981|b1015 rbsA|16131617|b3749 rbsB|16131619|b3751 rbsC|16131618|b3750 rbsD|16131616|b3748 rbsK|16131620|b3752 rhaA|16131743|b3903 rhaB|16131744|b3904 rhaD|16131742|b3902 rhaR|16131746|b3906 rhaS|16131745|b3905 rhaT|16131747|b3907 rpoH|16131333|b3461 rpoS|16130648|b2741 sdhA|16128698|b0723 sdhB|16128699|b0724 sdhC|16128696|b0721 sdhD|16128697|b0722 serA|16130814|b2913 sgbE|16131454|b3583 sgbH|16131452|b3581 sgbU|16131453|b3582 sohB|16129233|b1272 speC|16130866|b2965 srlA|16130609|b2702 srlB|16130611|b2704 srlD|16130612|b2705 srlE|16130610|b2703 srlR|16130614|b2707 sucA|16128701|b0726 sucB|16128702|b0727 sucC|16128703|b0728 sucD|16128704|b0729 tdcA|16131011|b3118 tdcB|16131010|b3117 tdcC|16131009|b3116 tdcD|16131008|b3115 tdcE|16131007|b3114 tdcF|16131006|b3113 tdcG|16131005|b3112 tnaA|16131576|b3708 tnaB|16131577|b3709 tnaL|16131575|b3707 treB|16132062|b4240 treC|16132061|b4239 tsx|16128396|b0411 ubiG|16130167|b2232 udp|16131680|b3831 uhpT|16131536|b3666 uxuA|16132143|b4322 uxuB|16132144|b4323 yhfA|16131235|b3356 yiaJ|16131445|b3574 yiaK|16131446|b3575 yiaL|16131447|b3576 yiaM|16131448|b3577 yiaN|16131449|b3578 yiaO|16131450|b3579 yjcG|16131893|b4067 yjcH|16131894|b4068
csgD|16129003|b1040	**5**	csgA|16129005|b1042 csgB|16129004|b1041 csgE|16129002|b1039 csgF|16129001|b1038 csgG|16129000|b1037
cspA|16131427|b3556	**2**	gyrA|16130166|b2231 hns|16129198|b1237
cspE|16128606|b0623	**1**	cspA|16131427|b3556
cusR|16128554|b0571	**4**	cusA|16128558|b0575 cusB|16128557|b0574 cusC|16128555|b0572 cusF|16128556|b0573
cynR|16128323|b0338	**3**	cynS|16128325|b0340 cynT|16128324|b0339 cynX|16128326|b0341

cysB|16129236|b1275	**17**	cbl|16129929|b1987 cysA|16130348|b2422 cysC|16130657|b2750 cysD|16130659|b2752 cysH|16130669|b2762 cysI|16130670|b2763 cysJ|16130671|b2764 cysK|16130340|b2414 cysM|16130347|b2421 cysN|16130658|b2751 cysP|16130350|b2425 cysU|16130349|b2424 cysW|b2423|b2423 tauA|16128350|b0365 tauB|16128351|b0366 tauC|16128352|b0367 tauD|16128353|b0368
cytR|16131772|b3934	**10**	deoA|16132199|b4382 deoB|16132200|b4383 deoC|16132198|b4381 deoD|16132201|b4384 nupC|16130325|b2393 nupG|16130865|b2964 ppiA|16131242|b3363 rpoH|16131333|b3461 tsx|16128396|b0411 udp|16131680|b3831
dcuR|16131950|b4124	**7**	dctA|16131400|b3528 dcuB|16131949|b4123 frdA|16131979|b4154 frdB|16131978|b4153 frdC|16131977|b4152 frdD|16131976|b4151 fumB|16131948|b4122
deoR|16128808|b0840	**6**	deoA|16132199|b4382 deoB|16132200|b4383 deoC|16132198|b4381 deoD|16132201|b4384 nupG|16130865|b2964 tsx|16128396|b0411
dnaA|16131570|b3702	**8**	aldA|16129376|b1415 dnaN|16131569|b3701 guaA|16130432|b2507 guaB|16130433|b2508 nrdA|16130169|b2234 nrdB|16130170|b2235 recF|16131568|b3700 rpoH|16131333|b3461
dsdC|16130296|b2364	**2**	dsdA|16130298|b2366 dsdX|16130297|b2365
ebgR|16130970|b3075	**2**	ebgA|16130971|b3076 ebgC|16130972|b3077
emrR|16130596|b2684	**2**	emrA|16130597|b2685 emrB|16130598|b2686
envY|16128549|b0566	**2**	ompC|16130152|b2215 ompF|16128896|b0929
exuR|16130989|b3094	**1**	exuT|16130988|b3093
fadR|16129150|b1187	**7**	fabA|16128921|b0954 fadA|16131691|b3845 fadB|16131692|b3846 fadD|16129759|b1805 fadL|16130277|b2344 iclR|16131844|b4018 uspA|16131367|b3495
fhlA|16130638|b2731	**26**	fdhF|16131905|b4079 hycA|16130632|b2725 hycB|16130631|b2724 hycC|16130630|b2723 hycD|16130629|b2722 hycE|16130628|b2721 hycF|16130627|b2720 hycG|16130626|b2719 hycH|16130625|b2718 hydN|16130620|b2713 hyfA|16130|b2481 hyfB|16130407|b2482 hyfC|16130408|b2483 hyfD|16130409|b2484 hyfE|16130410|b2485 hyfF|16130411|b2486 hyfG|16130412|b2487 hyfH|16130413|b2488 hyfI|16130414|b2489 hyfJ|16130415|b2490 hypA|16130633|b2726 hypB|16130634|b2727 hypC|16130635|b2728 hypD|16130636|b2729 hypE|16130637|b2730 hypF|16130619|b2712
fis|16131149|b3261	**76**	acnB|16128111|b0118 acs|16131895|b4069 adhE|16129202|b1241 alaT|b3853|b3853 alaW|b2397|b2397 aldB|16131459|b3588 argX|b3796|b3796 bglG|16131591|b3723 cysG|16131246|b3368 dusB|16131148|b3260 guaA|16130432|b2507 guaB|16130433|b2508 gyrA|16130166|b2231 gyrB|16131567|b3699 hns|16129198|b1237 hupA|16131830|b4000 hupB|16128425|b0440 infB|16131060|b3168 leuP|b4369|b4369 leuX|b4270|b4270 lpdA|16128109|b0116 marA|16129490|b1531 marB|16129491|b1532 marR|16129489|b1530 mazE|16130690|b2783 mazF|16130689|b2782 metT|b0673|b0673 mtlA|16131470|b3599 mtlD|16131471|b3600 mtlR|16131472|b3601 ndh|16129072|b1109 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 nrdA|16130169|b2234 nrdB|16130170|b2235 nrfA|16131896|b4070 nrfB|16131897|b4071 nrfC|16131898|b4072 nrfD|16131899|b4073 nrfE|16131900|b4074 nrfF|16131901|b4075 nrfG|16131902|b4076 nuoA|16130223|b2288 nuoB|16130222|b2287 nuoC|16130221|b2286 nuoE|16130220|b2285 nuoF|16130219|b2284 nuoG|16130218|b2283 nuoH|16130217|b2282 nuoI|16130216|b2281 nuoJ|16130215|b2280 nuoK|16130214|b2279 nuoL|16130213|b2278 nuoM|16130212|b2277 nuoN|16130211|b2276 nusA|16131061|b3169 pheU|b4134|b4134 pnp|16131056|b3164 proL|b2189|b2189 proM|b3799|b3799 rbfA|16131059|b3167 rpsO|16131057|b3165 serT|b0971|b0971 sra|16129439|b1480 thrT|b3979|b3979 thrV|b3273|b3273 thrW|b0244|b0244 tpr|16129192|b1229 trmA|16131803|b3965 truB|16131058|b3166 tyrT|b1231|b1231 tyrU|b3977|b3977 yhbC|16131062|b3170 yjcG|16131893|b4067 yjcH|16131894|b4068
flhD|16129844|b1892	**26**	fliA|16129869|b1922 fliL|16129891|b1944 fliM|16129892|b1945 fliN|16129893|b1946 fliO|16129894|b1947 fliP|16129895|b1948 fliQ|16129896|b1949 fliR|16129897|b1950 fliY|16129867|b1920 fliZ|16129868|b1921 glpA|16130176|b2241 glpB|16130177|b2242 glpC|16130178|b2243 hydN|16130620|b2713 hypF|16130619|b2712 mdh|16131126|b3236 mglA|16130087|b2149 mglB|16130088|b2150 mglC|16130086|b2148 nrfA|16131896|b4070 nrfB|16131897|b4071 nrfC|16131898|b4072 nrfD|16131899|b4073 nrfE|16131900|b4074 nrfF|16131901|b4075 nrfG|16131902|b4076
fnr|16129295|b1334	**112**	acnA|16129237|b1276 adhE|16129202|b1241 ansB|16130858|b2957 arcA|16132218|b4401 aspA|16131964|b4139 b0725|16128700|b0725 caiF|16128028|b0034 ccmA|16130138|b2201 ccmB|16130137|b2200 ccmC|16130136|b2199 ccmD|16130135|b2198 ccmE|16130134|b2197 ccmF|16130133|b2196 ccmH|16130131|b2194 cydA|16128708|b0733 cydB|16128709|b0734 cydC|16128853|b0886 cydD|16128854|b0887 cyoA|16128417|b0432 cyoB|16128416|b0431 cyoC|16128415|b0430 cyoD|16128414|b0429 cyoE|16128413|b0428 cysG|16131246|b3368 dcuA|16131963|b4138 dcuB|16131949|b4123 dmsA|16128861|b0894 dmsB|16128862|b0895 dmsC|16128863|b0896 dsbE|16130132|b2195 fdnH|16129434|b1475 fdnI|16129435|b1476 focA|16128871|b0904 frdA|16131979|b4154 frdB|16131978|b4153 frdC|16131977|b4152

		frdD|16131976|b4151 fumB|16131948|b4122 glpA|16130176|b2241 glpB|16130177|b2242 glpC|16130178|b2243 hemA|16129173|b1210 hlyE|16129145|b1182 hypA|16130633|b2726 hypB|16130634|b2727 hypC|16130635|b2728 hypD|16130636|b2729 hypE|16130637|b2730 icdA|16129099|b1136 moaA|16128749|b0781 moaB|16128750|b0782 moaC|16128751|b0783 moaD|16128752|b0784 moaE|16128753|b0785 moeA|16128795|b0827 moeB|16128794|b0826 napA|16130143|b2206 napB|16130140|b2203 napC|16130139|b2202 napD|16130144|b2207 napF|16130145|b2208 napG|16130142|b2205 napH|16130141|b2204 narG|16129187|b1224 narH|16129188|b1225 narI|16129190|b1227 narJ|16129189|b1226 narK|16129186|b1223 narL|16129184|b1221 narX|16129185|b1222 ndh|16129072|b1109 nikA|16131348|b3476 nikB|16131349|b3477 nikC|16131350|b3478 nikD|16131351|b3479 nikE|16131352|b3480 nikR|16131353|b3481 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 nrfA|16131896|b4070 nrfB|16131897|b4071 nrfC|16131898|b4072 nrfD|16131899|b4073 nrfE|16131900|b4074 nrfF|16131901|b4075 nrfG|16131902|b4076 nuoA|16130223|b2288 nuoB|16130222|b2287 nuoC|16130221|b2286 nuoE|16130220|b2285 nuoF|16130219|b2284 nuoG|16130218|b2283 nuoH|16130217|b2282 nuoI|16130216|b2281 nuoJ|16130215|b2280 nuoK|16130214|b2279 nuoL|16130213|b2278 nuoM|16130212|b2277 nuoN|16130211|b2276 pflB|16128870|b0903 sdhA|16128698|b0723 sdhB|16128699|b0724 sdhC|16128696|b0721 sdhD|16128697|b0722 sodA|16131748|b3908 sucA|16128701|b0726 sucB|16128702|b0727 sucC|16128703|b0728 sucD|16128704|b0729 yeiL|16130101|b2163 yfiD|16130504|b2579
fruR|16128073|b0080	**25**	aceA|16131841|b4015 aceB|16131840|b4014 aceK|16131842|b4016 adhE|16129202|b1241 crr|16130343|b2417 cysG|16131246|b3368 eda|16129803|b1850 edd|16129804|b1851 epd|16130828|b2927 fruA|16130105|b2167 fruB|16130107|b2169 fruK|16130106|b2168 icdA|16129099|b1136 mtlA|16131470|b3599 mtlD|16131471|b3600 mtlR|16131472|b3601 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 pckA|16131280|b3403 pgk|16130827|b2926 ppsA|16129658|b1702 ptsH|16130341|b2415 ptsI|16130342|b2416 pykF|16129632|b1676
fucR|16130712|b2805	**6**	fucA|16130707|b2800 fucI|16130709|b2802 fucK|16130710|b2803 fucO|16130706|b2799 fucP|16130708|b2801 fucU|16130711|b2804
fur|16128659|b0683	**33**	cirA|16130093|b2155 entA|16128579|b0596 entB|16128578|b0595 entC|16128576|b0593 entD|16128566|b0583 entE|16128577|b0594 entS|16128574|b0591 exbB|16130904|b3006 exbD|16130903|b3005 fecA|16132112|b4291 fecB|16132111|b4290 fecC|16132110|b4289 fecD|16132109|b4288 fecE|16132108|b4287 fecI|16132114|b4293 fecR|16132113|b4292 fepA|16128567|b0584 fepB|16128575|b0592 fepC|16128571|b0588 fepD|16128573|b0590 fepG|16128572|b0589 fhuA|16128143|b0150 fhuB|16128146|b0153 fhuC|16128144|b0151 fhuD|16128145|b0152 fhuE|16129065|b1102 nrdE|16130589|b2675 nrdF|16130590|b2676 nrdH|16130587|b2673 nrdI|16130588|b2674 sodA|16131748|b3908 tonB|16129213|b1252 ybdB|16128580|b0597
gadW|16131387|b3515	**2**	gadA|16131389|b3517 gadX|16131388|b3516
gadX|16131388|b3516	**1**	gadA|16131389|b3517
galR|16130741|b2837	**5**	galE|16128727|b0759 galK|16128725|b0757 galM|16128724|b0756 galS|16130089|b2151 galT|16128726|b0758
galS|16130089|b2151	**3**	mglA|16130087|b2149 mglB|16130088|b2150 mglC|16130086|b2148
gatR_1|b2087|b2087	**7**	gatA|16130032|b2094 gatB|16130031|b2093 gatC|16130030|b2092 gatD|16130029|b2091 gatR_2|b2090|b2090 gatY|16130034|b2096 gatZ|16130033|b2095
gcvA|16130715|b2808	**3**	gcvH|16130806|b2904 gcvP|16130805|b2903 gcvT|16130807|b2905
gcvR|16130404|b2479	**3**	gcvH|16130806|b2904 gcvP|16130805|b2903 gcvT|16130807|b2905
glcC|16130880|b2980	**5**	glcA|16130875|b2975 glcB|16130876|b2976 glcD|16130879|b2979 glcE|16130878|b2978 glcG|16130877|b2977
glnL|16131709|b3869	**34**	amtB|16128436|b0451 argT|16130245|b2310 astA|16129701|b1747 astB|16129699|b1745 astC|16129702|b1748 astD|16129700|b1746 astE|16129698|b1744 cbl|16129929|b1987 ddpA|16129446|b1487 ddpB|16129445|b1486 ddpC|16129444|b1485 ddpD|16129443|b1484 ddpF|16129442|b1483 ddpX|16129447|b1488 glnA|16131710|b3870 glnG|16131708|b3868 glnH|16128779|b0811 glnK|16128435|b0450 glnP|16128778|b0810 glnQ|16128777|b0809 hisJ|16130244|b2309 hisM|16130242|b2307 hisP|16130241|b2306 hisQ|16130243|b2308 nac|16129930|b1988 potF|16128822|b0854 potG|16128823|b0855 potH|16128824|b0856 potI|16128825|b0857 ycdG|16128972|b1006 yhdW|16131156|b3268 yhdX|16131157|b3269 yhdY|16131158|b3270 yhdZ|16131159|b3271
glpR|16131297|b3423	**8**	glpA|16130176|b2241 glpB|16130177|b2242 glpC|16130178|b2243 glpD|16131300|b3426 glpF|16131765|b3927 glpK|16131764|b3926 glpQ|16130174|b2239 glpT|16130175|b2240
gntR|16131310|b3438	**6**	eda|16129803|b1850 edd|16129804|b1851 gntK|16131309|b3437 gntT|16131291|b3415 gntU_1|b3436|b3436 gntU_2|b3435|b3435
gutM|16130613|b2706	**6**	gutQ|16130615|b2708 srlA|16130609|b2702 srlB|16130611|b2704 srlD|16130612|b2705 srlE|16130610|b2703 srlR|16130614|b2707
himA|16129668|b1712	**160**	aceA|16131841|b4015 aceB|16131840|b4014 aceK|16131842|b4016 acs|16131895|b4069 adiA|16131943|b4117 amiA|16130360|b2435 caiA|16128033|b0039 caiB|16128032|b0038 caiC|16128031|b0037 caiD|16128030|b0036 caiE|16128029|b0035 caiT|16128034|b0040

		carA|16128026|b0032 carB|16128027|b0033 cysG|16131246|b3368 dmsA|16128861|b0894 dmsB|16128862|b0895 dmsC|16128863|b0896 dppA|16131416|b3544 dppB|16131415|b3543 dppC|16131414|b3542 dppD|16131413|b3541 dppF|16131412|b3540 dps|16128780|b0812 dusB|16131148|b3260 ecpD|16128133|b0140 envZ|16131281|b3404 fimA|16132135|b4314 fimC|16132137|b4316 fimD|16132138|b4317 fimF|16132139|b4318 fimG|16132140|b4319 fimH|16132141|b4320 fimI|16132136|b4315 fis|16131149|b3261 flhC|16129843|b1891 flhD|16129844|b1892 focA|16128871|b0904 gcd|16128117|b0124 glcA|16130875|b2975 glcB|16130876|b2976 glcD|16130879|b2979 glcE|16130878|b2978 glcG|16130877|b2977 glnH|16128779|b0811 glnP|16128778|b0810 glnQ|16128777|b0809 gltA|16128695|b0720 hemA|16129173|b1210 hemF|16130361|b2436 himD|16128879|b0912 htrE|16128132|b0139 hycA|16130632|b2725 hycB|16130631|b2724 hycC|16130630|b2723 hycD|16130629|b2722 hycE|16130628|b2721 hycF|16130627|b2720 hycG|16130626|b2719 hycH|16130625|b2718 hypA|16130633|b2726 hypB|16130634|b2727 hypC|16130635|b2728 hypD|16130636|b2729 hypE|16130637|b2730 ilvA|16131630|b3772 ilvD|16131629|b3771 ilvE|16131628|b3770 ilvG_1|b3767|b3767 ilvG_2|b3768|b3768 ilvL|16131626|b3766 ilvM|16131627|b3769 lyxK|16131451|b3580 mtr|16131053|b3161 narG|16129187|b1224 narH|16129188|b1225 narI|16129190|b1227 narJ|16129189|b1226 narK|16129186|b1223 ndh|16129072|b1109 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 nmpC|16128536|b0553 nrfA|16131896|b4070 nrfB|16131897|b4071 nrfC|16131898|b4072 nrfD|16131899|b4073 nrfE|16131900|b4074 nrfF|16131901|b4075 nrfG|16131902|b4076 nuoA|16130223|b2288 nuoB|16130222|b2287 nuoC|16130221|b2286 nuoE|16130220|b2285 nuoF|16130219|b2284 nuoG|16130218|b2283 nuoH|16130217|b2282 nuoI|16130216|b2281 nuoJ|16130215|b2280 nuoK|16130214|b2279 nuoL|16130213|b2278 nuoM|16130212|b2277 nuoN|16130211|b2276 ompC|16130152|b2215 ompF|16128896|b0929 ompR|16131282|b3405 osmE|16129693|b1739 osmY|16132194|b4376 paaA|16129349|b1388 paaB|16129350|b1389 paaC|16129351|b1390 paaD|16129352|b1391 paaE|16129353|b1392 paaF|16129354|b1393 paaG|16129355|b1394 paaH|16129356|b1395 paaI|16129357|b1396 paaJ|16129358|b1397 paaK|16129359|b1398 paaZ|16129348|b1387 pflB|16128870|b0903 phoU|16131592|b3724 pspA|16129265|b1304 pspB|16129266|b1305 pspC|16129267|b1306 pspD|16129268|b1307 pspE|16129269|b1308 pstA|16131594|b3726 pstB|16131593|b3725 pstC|16131595|b3727 pstS|16131596|b3728 rtcA|b3420|b3420 rtcB|16131295|b3421 sgbE|16131454|b3583 sgbH|16131452|b3581 sgbU|16131453|b3582 sodA|16131748|b3908 sodB|16129614|b1656 sucA|16128701|b0726 sucB|16128702|b0727 sucC|16128703|b0728 sucD|16128704|b0729 tdcA|16131011|b3118 tdcB|16131010|b3117 tdcC|16131009|b3116 tdcD|16131008|b3115 tdcE|16131007|b3114 tdcF|16131006|b3113 tdcG|16131005|b3112 tyrP|16129857|b1907 yeiL|16130101|b2163 yiaJ|16131445|b3574 yiaK|16131446|b3575 yiaL|16131447|b3576 yiaM|16131448|b3577 yiaN|16131449|b3578 yiaO|16131450|b3579 yjcG|16131893|b4067 yjcH|16131894|b4068
hns|16129198|b1237	**33**	adiA|16131943|b4117 bglG|16131591|b3723 bolA|16128420|b0435 cadA|16131957|b4131 cadB|16131958|b4132 csiD|16130573|b2659 csiE|16130460|b2535 cydA|16128708|b0733 cydB|16128709|b0734 cysG|16131246|b3368 chiA|16131217|b3338 flhC|16129843|b1891 flhD|16129844|b1892 fliA|16129869|b1922 fliY|16129867|b1920 fliZ|16129868|b1921 gadA|16131389|b3517 gadX|16131388|b3516 hlyE|16129145|b1182 mukB|16128891|b0924 mukE|16128890|b0923 mukF|16128889|b0922 nhaA|16128013|b0019 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 proV|16130591|b2677 proW|16130592|b2678 proX|16130593|b2679 rcsA|16129898|b1951 smtA|16128888|b0921 sodB|16129614|b1656 stpA|16130583|b2669
hupA|16131830|b4000	**5**	galE|16128727|b0759 galK|16128725|b0757 galM|16128724|b0756 galT|16128726|b0758 seqA|16128663|b0687
hyfR|16130416|b2491	**10**	hyfA|16130|b2481 hyfB|16130407|b2482 hyfC|16130408|b2483 hyfD|16130409|b2484 hyfE|16130410|b2485 hyfF|16130411|b2486 hyfG|16130412|b2487 hyfH|16130413|b2488 hyfI|16130414|b2489 hyfJ|16130415|b2490
iciA|16130817|b2916	**5**	dnaA|16131570|b3702 dnaN|16131569|b3701 nrdA|16130169|b2234 nrdB|16130170|b2235 recF|16131568|b3700
iclR|16131844|b4018	**3**	aceA|16131841|b4015 aceB|16131840|b4014 aceK|16131842|b4016
ilvY|16131631|b3773	**1**	ilvC|16131632|b3774
iscR|16130456|b2531	**3**	iscA|16130453|b2528 iscS|16130455|b2530 iscU|16130454|b2529
kdpE|16128670|b0694	**3**	kdpA|16128674|b0698 kdpB|16128673|b0697 kdpC|16128672|b0696
lacI|16128330|b0345	**3**	lacA|16128327|b0342 lacY|16128328|b0343 lacZ|16128329|b0344
lctR|16131475|b3604	**2**	lctD|16131476|b3605 lctP|16131474|b3603
leuO|16128070|b0076	**5**	leuA|16128068|b0074 leuB|16128067|b0073 leuC|16128066|b0072 leuD|16128065|b0071 leuL|16128069|b0075
lexA|16131869|b4043	**13**	dinF|16131870|b4044 dnaG|16130962|b3066 oraA|16130605|b2698 phrB|16128683|b0708 recA|16130606|b2699 rpoD|16130963|b3067 rpsU|16130961|b3065 sulA|16128925|b0958 umuC|16129147|b1184 umuD|16129146|b1183 uvrA|16131884|b4058 uvrB|16128747|b0779 uvrD|16131665|b3813
lrhA|16130224|b2289	**2**	flhC|16129843|b1891 flhD|16129844|b1892

lrp|16128856|b0889	**55**	aidB|16132009|b4187 csiD|16130573|b2659 dadA|16129152|b1189 dadX|16129153|b1190 fimA|16132135|b4314 fimC|16132137|b4316 fimD|16132138|b4317 fimF|16132139|b4318 fimG|16132140|b4319 fimH|16132141|b4320 fimI|16132136|b4315 gcvH|16130806|b2904 gcvP|16130805|b2903 gcvT|16130807|b2905 gltB|16131102|b3212 gltD|16131103|b3213 gltF|16131104|b3214 ilvA|16131630|b3772 ilvD|16131629|b3771 ilvE|16131628|b3770 ilvG_1|b3767|b3767 ilvG_2|b3768|b3768 ilvH|16128071|b0078 ilvI|b0077|b0077 ilvL|16131626|b3766 ilvM|16131627|b3769 kbl|16131488|b3617 leuA|16128068|b0074 leuB|16128067|b0073 leuC|16128066|b0072 leuD|16128065|b0071 leuL|16128069|b0075 livF|16131326|b3454 livG|16131327|b3455 livH|16131329|b3457 livJ|16131332|b3460 livK|16131330|b3458 livM|16131328|b3456 lysU|16131955|b4129 malT|16131294|b3418 ompC|16130152|b2215 ompF|16128896|b0929 oppA|16129204|b1243 oppB|16129205|b1244 oppC|16129206|b1245 oppD|16129207|b1246 oppF|16129208|b1247 osmC|16129441|b1482 osmY|16132194|b4376 sdaA|16129768|b1814 serA|16130814|b2913 serC|16128874|b0907 stpA|16130583|b2669 tdh|16131487|b3616 yeiL|16130101|b2163
lysR|16130743|b2839	**1**	lysA|16130742|b2838
malI|16129578|b1620	**2**	malX|16129579|b1621 malY|16129580|b1622
malT|16131294|b3418	**9**	lamB|16131862|b4036 malE|16131860|b4034 malF|16131859|b4033 malG|16131858|b4032 malK|16131861|b4035 malM|16131863|b4037 malP|16131293|b3417 malQ|16131292|b3416 malS|16131442|b3571
marA|16129490|b1531	**14**	acrA|16128447|b0463 acrB|16128446|b0462 fpr|16131762|b3924 fumC|16129569|b1611 inaA|16130172|b2237 marB|16129491|b1532 marR|16129489|b1530 nfo|16130097|b2159 pqiA|16128917|b0950 pqiB|16128918|b0951 putA|16128980|b1014 slp|16131378|b3506 sodA|16131748|b3908 zwf|16129805|b1852
marR|16129489|b1530	**2**	marA|16129490|b1531 marB|16129491|b1532
melR|16131944|b4118	**2**	melA|16131945|b4119 melB|16131946|b4120
metJ|16131776|b3938	**10**	ahpC|16128588|b0605 ahpF|16128589|b0606 metA|16131839|b4013 metB|16131777|b3939 metC|16130906|b3008 metF|16131779|b3941 metI|16128191|b0198 metL|16131778|b3940 metN|16128192|b0199 metQ|16128190|b0197
metR|16131677|b3828	**3**	glyA|16130476|b2551 metA|16131839|b4013 metH|16131845|b4019
mhpR|16128331|b0346	**6**	mhpA|16128332|b0347 mhpB|16128333|b0348 mhpC|16128334|b0349 mhpD|16128335|b0350 mhpE|16128337|b0352 mhpF|16128336|b0351
mlc|16129552|b1594	**8**	crr|16130343|b2417 malT|16131294|b3418 manX|16129771|b1817 manY|16129772|b1818 manZ|16129773|b1819 ptsG|16129064|b1101 ptsH|16130341|b2415 ptsI|16130342|b2416
mngR|16128705|b0730	**2**	mngA|16128706|b0731 mngB|16128707|b0732
modE|16128729|b0761	**36**	ccmA|16130138|b2201 ccmB|16130137|b2200 ccmC|16130136|b2199 ccmD|16130135|b2198 ccmE|16130134|b2197 ccmF|16130133|b2196 ccmH|16130131|b2194 dmsA|16128861|b0894 dmsB|16128862|b0895 dmsC|16128863|b0896 dsbE|16130132|b2195 hycA|16130632|b2725 hycB|16130631|b2724 hycC|16130630|b2723 hycD|16130629|b2722 hycE|16130628|b2721 hycF|16130627|b2720 hycG|16130626|b2719 hycH|16130625|b2718 moaA|16128749|b0781 moaB|16128750|b0782 moaC|16128751|b0783 moaD|16128752|b0784 moaE|16128753|b0785 modA|16128731|b0763 modB|16128732|b0764 modC|16128733|b0765 napA|16130143|b2206 napB|16130140|b2203 napC|16130139|b2202 napD|16130144|b2207 napF|16130145|b2208 napG|16130142|b2205 napH|16130141|b2204 narL|16129184|b1221 narX|16129185|b1222
mtlR|16131472|b3601	**2**	mtlA|16131470|b3599 mtlD|16131471|b3600
nac|16129930|b1988	**10**	asnC|16131611|b3743 codA|16128322|b0337 codB|16128321|b0336 gabD|16130575|b2661 gabP|16130577|b2663 gabT|16130576|b2662 gltB|16131102|b3212 gltD|16131103|b3213 gltF|16131104|b3214 nupC|16130325|b2393
nadR|16132207|b4390	**2**	nadB|16130499|b2574 pncB|16128898|b0931
nagC|16128652|b0676	**9**	glmS|16131597|b3729 glmU|16131598|b3730 manX|16129771|b1817 manY|16129772|b1818 manZ|16129773|b1819 nagA|16128653|b0677 nagB|16128654|b0678 nagD|16128651|b0675 nagE|16128655|b0679
narL|16129184|b1221	**80**	adhE|16129202|b1241 caiF|16128028|b0034 ccmA|16130138|b2201 ccmB|16130137|b2200 ccmC|16130136|b2199 ccmD|16130135|b2198 ccmE|16130134|b2197 ccmF|16130133|b2196 ccmH|16130131|b2194 cydC|16128853|b0886 cydD|16128854|b0887 cysG|16131246|b3368 dcuB|16131949|b4123 dmsA|16128861|b0894 dmsB|16128862|b0895 dmsC|16128863|b0896 dsbE|16130132|b2195 fdnH|16129434|b1475 fdnI|16129435|b1476 focA|16128871|b0904 frdA|16131979|b4154 frdB|16131978|b4153 frdC|16131977|b4152 frdD|16131976|b4151 fumB|16131948|b4122 hyaA|16128938|b0972 hyaB|16128939|b0973 hyaC|16128940|b0974 hyaD|16128941|b0975 hyaE|16128942|b0976 hyaF|16128943|b0977 hyb0|16130897|b2997 hybA|16130896|b2996 hybB|16130895|b2995 hybC|16130894|b2994 hybD|16130893|b2993 hybE|16130892|b2992 hybF|16130891|b2991 hybG|16130890|b2990 moeA|16128795|b0827 moeB|16128794|b0826 napA|16130143|b2206 napB|16130140|b2203 napC|16130139|b2202 napD|16130144|b2207 napF|16130145|b2208 napG|16130142|b2205 napH|16130141|b2204

		narG|16129187|b1224 narH|16129188|b1225 narI|16129190|b1227 narJ|16129189|b1226 narK|16129186|b1223 nirB|16131244|b3365 nirC|b3367|b3367 nirD|16131245|b3366 nrfA|16131896|b4070 nrfB|16131897|b4071 nrfC|16131898|b4072 nrfD|16131899|b4073 nrfE|16131900|b4074 nrfF|16131901|b4075 nrfG|16131902|b4076 nuoA|16130223|b2288 nuoB|16130222|b2287 nuoC|16130221|b2286 nuoE|16130220|b2285 nuoF|16130219|b2284 nuoG|16130218|b2283 nuoH|16130217|b2282 nuoI|16130216|b2281 nuoJ|16130215|b2280 nuoK|16130214|b2279 nuoL|16130213|b2278 nuoM|16130212|b2277 nuoN|16130211|b2276 pflB|16128870|b0903 torA|16128963|b0997 torC|16128962|b0996 torD|16128964|b0998
narP|16130130|b2193	**21**	ccmA|16130138|b2201 ccmB|16130137|b2200 ccmC|16130136|b2199 ccmD|16130135|b2198 ccmE|16130134|b2197 ccmF|16130133|b2196 ccmH|16130131|b2194 dsbE|16130132|b2195 hyaA|16128938|b0972 hyaB|16128939|b0973 hyaC|16128940|b0974 hyaD|16128941|b0975 hyaE|16128942|b0976 hyaF|16128943|b0977 napA|16130143|b2206 napB|16130140|b2203 napC|16130139|b2202 napD|16130144|b2207 napF|16130145|b2208 napG|16130142|b2205 napH|16130141|b2204
nhaR|16128014|b0020	**2**	nhaA|16128013|b0019 osmC|16129441|b1482
norR|16130616|b2709	**2**	norV|16130617|b2710 norW|16130618|b2711
ompR|16131282|b3405	**12**	bolA|16128420|b0435 csgD|16129003|b1040 csgE|16129002|b1039 csgF|16129001|b1038 csgG|16129000|b1037 fadL|16130277|b2344 flhC|16129843|b1891 flhD|16129844|b1892 nmpC|16128536|b0553 ompC|16130152|b2215 ompF|16128896|b0929 sra|16129439|b1480
oxyR|16131799|b3961	**12**	agn43|16129941|b2000 ahpC|16128588|b0605 ahpF|16128589|b0606 dps|16128780|b0812 fur|16128659|b0683 katG|16131780|b3942 sufA|16129640|b1684 sufB|16129639|b1683 sufC|16129638|b1682 sufD|16129637|b1681 sufE|16129635|b1679 sufS|16129636|b1680
paaX|16129360|b1399	**12**	paaA|16129349|b1388 paaB|16129350|b1389 paaC|16129351|b1390 paaD|16129352|b1391 paaE|16129353|b1392 paaF|16129354|b1393 paaG|16129355|b1394 paaH|16129356|b1395 paaI|16129357|b1396 paaJ|16129358|b1397 paaK|16129359|b1398 paaZ|16129348|b1387
pdhR|16128106|b0113	**6**	aceE|16128107|b0114 aceF|16128108|b0115 lctD|16131476|b3605 lctP|16131474|b3603 lctR|16131475|b3604 lpdA|16128109|b0116
phoB|16128384|b0399	**25**	b4103|16131929|b4103 phnC|16131932|b4106 phnD|16131931|b4105 phnE|16131930|b4104 phnF|16131928|b4102 phnG|16131927|b4101 phnH|16131926|b4100 phnI|16131925|b4099 phnJ|16131924|b4098 phnK|16131923|b4097 phnL|16131922|b4096 phnM|16131921|b4095 phnN|16131920|b4094 phnO|16131919|b4093 phnP|16131918|b4092 phoA|16128368|b0383 phoE|16128227|b0241 phoH|16128984|b1020 phoR|16128385|b0400 phoU|16131592|b3724 psiF|16128369|b0384 pstA|16131594|b3726 pstB|16131593|b3725 pstC|16131595|b3727 pstS|16131596|b3728
phoP|16129093|b1130	**2**	phoQ|16129092|b1129 treR|16132063|b4241
pspF|16129|b1303	**5**	pspA|16129265|b1304 pspB|16129266|b1305 pspC|16129267|b1306 pspD|16129268|b1307 pspE|16129269|b1308
purR|16129616|b1658	**27**	codA|16128322|b0337 codB|16128321|b0336 cvpA|16130248|b2313 gcvH|16130806|b2904 gcvP|16130805|b2903 gcvT|16130807|b2905 glnB|16130478|b2553 glyA|16130476|b2551 guaA|16130432|b2507 guaB|16130433|b2508 prsA|16129170|b1207 purB|16129094|b1131 purC|16130401|b2476 purD|16131835|b4005 purE|16128507|b0523 purF|16130247|b2312 purH|16131836|b4006 purK|16128506|b0522 purL|16130482|b2557 purM|16130424|b2499 purN|16130425|b2500 pyrC|16129025|b1062 pyrD|16128912|b0945 speA|16130839|b2938 speB|16130838|b2937 ubiX|16130246|b2311 ycfC|16129095|b1132
rbsR|16131621|b3753	**5**	rbsA|16131617|b3749 rbsB|16131619|b3751 rbsC|16131618|b3750 rbsD|16131616|b3748 rbsK|16131620|b3752
rcsB|16130154|b2217	**9**	b2060|16130000|b2060 ftsA|16128087|b0094 ftsZ|16128088|b0095 osmC|16129441|b1482 rcsA|16129898|b1951 wcaA|16129999|b2059 wcaB|16129998|b2058 wza|16130002|b2062 wzb|16130001|b2061
rhaR|16131746|b3906	**1**	rhaS|16131745|b3905
rhaS|16131745|b3905	**4**	rhaA|16131743|b3903 rhaB|16131744|b3904 rhaD|16131742|b3902 rhaT|16131747|b3907
rob|16132213|b4396	**8**	fumC|16129569|b1611 inaA|16130172|b2237 marA|16129490|b1531 marB|16129491|b1532 marR|16129489|b1530 nfo|16130097|b2159 sodA|16131748|b3908 zwf|16129805|b1852
rtcR|16131296|b3422	**2**	rtcA|b3420|b3420 rtcB|16131295|b3421
sdiA|16129863|b1916	**3**	ftsA|16128087|b0094 ftsQ|16128086|b0093 ftsZ|16128088|b0095
slyA|16129600|b1642	**1**	hlyE|16129145|b1182
soxR|16131889|b4063	**1**	soxS|16131888|b4062
soxS|16131888|b4062	**20**	acrA|16128447|b0463 acrB|16128446|b0462 fldA|16128660|b0684 fldB|16130797|b2895 fpr|16131762|b3924 fumC|16129569|b1611 fur|16128659|b0683 inaA|16130172|b2237 marA|16129490|b1531 marB|16129491|b1532 marR|16129489|b1530 mdaA|16128819|b0851 nfo|16130097|b2159

		pqiA|16128917|b0950 pqiB|16128918|b0951 rimK|16128820|b0852 sodA|16131748|b3908 ybjC|16128818|b0850 ybjN|16128821|b0853 zwf|16129805|b1852
srlR|16130614|b2707	**6**	gutM|16130613|b2706 gutQ|16130615|b2708 srlA|16130609|b2702 srlB|16130611|b2704 srlD|16130612|b2705 srlE|16130610|b2703
tdcA|16131011|b3118	**6**	tdcB|16131010|b3117 tdcC|16131009|b3116 tdcD|16131008|b3115 tdcE|16131007|b3114 tdcF|16131006|b3113 tdcG|16131005|b3112
tdcR|16131012|b3119	**7**	tdcA|16131011|b3118 tdcB|16131010|b3117 tdcC|16131009|b3116 tdcD|16131008|b3115 tdcE|16131007|b3114 tdcF|16131006|b3113 tdcG|16131005|b3112
torR|16128961|b0995	**3**	torA|16128963|b0997 torC|16128962|b0996 torD|16128964|b0998
treR|16132063|b4241	**2**	treB|16132062|b4240 treC|16132061|b4239
trpR|16132210|b4393	**11**	aroH|16129660|b1704 aroL|16128373|b0388 aroM|16128375|b0390 mtr|16131053|b3161 trpA|16129221|b1260 trpB|16129222|b1261 trpC|16129223|b1262 trpD|16129224|b1263 trpE|16129225|b1264 trpL|16129226|b1265 yaiA|16128374|b0389
tyrR|16129284|b1323	**10**	aroF|16130522|b2601 aroG|16128722|b0754 aroL|16128373|b0388 aroM|16128375|b0390 aroP|16128105|b0112 mtr|16131053|b3161 tyrA|16130521|b2600 tyrB|16131880|b4054 tyrP|16129857|b1907 yaiA|16128374|b0389
uhpA|16131539|b3669	**1**	uhpT|16131536|b3666
xapR|16130331|b2405	**2**	xapA|16130333|b2407 xapB|16130332|b2406
xylR|16131440|b3569	**5**	xylA|16131436|b3565 xylB|16131435|b3564 xylF|16131437|b3566 xylG|16131438|b3567 xylH|16131439|b3568
yiaJ|16131445|b3574	**9**	lyxK|16131451|b3580 sgbE|16131454|b3583 sgbH|16131452|b3581 sgbU|16131453|b3582 yiaK|16131446|b3575 yiaL|16131447|b3576 yiaM|16131448|b3577 yiaN|16131449|b3578 yiaO|16131450|b3579
yijC|16131801|b3963	**1**	fabA|16128921|b0954
zraR|16131834|b4004	**2**	zraP|16131832|b4002 zraS|16131833|b4003

ANEXO 2B.

Date sets of regulons used in the entire analysis from *Bacillus subtilis* from DBTBS v4.0: http://dbtbs.hgc.jp/

Transcription Factor Name|Genebank|BGnumber	TGs Number	Target Genes Name|Genebank|BGnumber
abrB|16077105|BG10100	**72**	albA|16080789|BG12471 albB|16080790|BG12470 albC|16080791|BG12469 albD|16080792|BG12468 albE|16080793|BG12467 albF|16080794|BG12466 albG|16080795|BG12465 aprE|16078094|BG10190 argB|16078186|BG10193 argC|16078184|BG10191 argD|16078187|BG10194 argF|16078190|BG10197 argJ|16078185|BG10192 asnH|16081043|BG11116 carA|16078188|BG10195 carB|16078616|BG10196 citB|16078863|BG10478 comK|16078106|BG11059 dppA|16078357|BG10842 dppB|16078358|BG10843 dppC|16078359|BG10844 dppD|16078360|BG10845 dppE|16078361|BG10846 ftsA|16078592|BG10231 ftsZ|16078593|BG10232 hpr|16078063|BG10659 hutG|16080989|BG11099 hutH|16080986|BG10667 hutI|16080988|BG11100 hutM|16080990|BG11101 hutP|16080985|BG10666 hutU|16080987|BG10668 kapB|16080198|BG10746 kinB|16080197|BG10745 pbpE|16080497|BG10390 racX|16080496|BG10391 rbsA|16080647|BG10879 rbsB|16080649|BG10881 rbsC|16080648|BG10880 rbsD|16080646|BG10878 rbsK|16080645|BG10877 rbsR|16080644|BG10876 rok|16078488|BG13307 sbo|16080788|BG12671 sboX|50812300|BG14182 sigH|16077166|BG10159 sigW|16077241|BG11573 sinI|16079516|BG10753 sinR|16079517|BG10754 spo0E|16078428|BG10769 spoVG|16077117|BG10112 ybbM|16077242|BG11574 ycnK|16077464|BG12047 yknW|16078498|BG13243 yknX|16078499|BG13244 yknY|16078500|BG13245 yknZ|16078501|BG13246 yoaW|16078938|BG13493 yocH|16078981|BG13521 yrhI|16079771|BG12298 yrhJ|16079770|BG12299 yvlA|16080566|BG14116 yvlB|16080565|BG14117 yvlC|16080564|BG14118 yvlD|16080563|BG14119 yvqH|16080365|BG14137 yxaM|16081044|BG11115 yxbA|16081041|BG11351 yxbB|16081033|BG11352 yxbC|16081039|BG11353 yxbD|16081038|BG11354 yxnB|16081042|BG14166
acoR|16077877|BG11790	**4**	acoA|16077873|BG12558 acoB|16077874|BG12559 acoC|16077875|BG12560 acoL|16077876|BG12561
adaA|16077249|BG10166	**2**	adaB|16077250|BG10167 alkA|16077248|BG10165
ahrC|16079481|BG10309	**7**	argC|16078184|BG10191 rocA|16080830|BG10622 rocB|16080829|BG10623 rocC|16080828|BG10624 rocD|16081086|BG10722 rocE|16081085|BG10933 rocF|16081084|BG10932
alsR|16080655|BG10470	**4**	alsD|16080653|BG10472 alsS|16080654|BG10471 lctP|16077375|BG12001 ldh|16077374|BG19003
ansR|16079416|BG10299	**2**	ansA|16079415|BG10300 ansB|16079414|BG10301
araR|16080450|BG11913	**12**	abfA|16079924|BG11900 abnA|16079933|BG11901 araA|16079932|BG11904 araB|16079931|BG11905 araD|16079930|BG11906 araE|16080449|BG11907 araL|16079929|BG11908 araM|16079928|BG11909 araN|16079927|BG11910 araP|16079926|BG11911 araQ|16079925|BG11912 xsa|16079903|BG11985
arsR|16079634|BG11301	**3**	arsB|16079632|BG11303 arsC|16079631|BG11304 yqcK|16079633|BG11302
azlB|16079725|BG11914	**4**	azlC|16079724|BG11915 azlD|16079723|BG11916 brnQ|16079722|BG11918 yrdK|16079721|BG12286
birA|16079301|BG11206	**7**	bioA|16080075|BG11524 bioB|16080072|BG11525 bioD|16080073|BG11526 bioF|16080074|BG11527 bioI|16080071|BG11528 bioW|50812281|BG11529 ytbQ|16080070|BG11787
bkdR|16079466|BG11721	**7**	bcd|16079464|BG11723 bkdA1|16079461|BG10307 bkdA2|16079460|BG10306 bkdB|16079459|BG10305 buk|16079463|BG11724 lpd|16079462|BG11725 ptb|16079465|BG11722
bltR|16079711|BG10904	**2**	blt|16079712|BG10905 bltD|16079713|BG10906
bmrR|50812267|BG10304	**1**	bmr|16079457|BG10303
ccpA|16080026|BG10376	**97**	ackA|16079999|BG10813 acoA|16077873|BG12558 acoR|16077877|BG11790 acsA|16080020|BG10370 acuA|16080021|BG10369 acuB|16080022|BG10368 acuC|16080023|BG10367 alsS|16080654|BG10471 amyE|16077373|BG10473 araA|16079932|BG11904 araB|16079931|BG11905 araE|16080449|BG11907 bglH|50812306|BG10935 bglP|16080978|BG10934 bglS|16080958|BG10476 ccpC|16078478|BG13297 citC|16079965|BG10856 citM|16078996|BG10273 citZ|16079966|BG10855 dctP|16077514|BG12075 dctR|16077513|BG12074 dctS|16077512|BG12073 deoR|16080994|BG10982 dra|16080993|BG10983 exuR|16078302|BG13211 exuT|16078301|BG13210 fbaB|16081018|BG11125 galK|16080871|BG10581 galT|16080870|BG10582 glpF|16077993|BG10186 glpK|16077994|BG10187 gntK|16081058|BG10649 gntP|16081059|BG10650 gntR|16081057|BG10648 gntZ|16081060|BG10651 hutP|16080985|BG10666 idh|16081021|BG10669 iolB|16081026|BG11118 iolC|16081025|BG11119 iolD|16081024|BG11120 iolE|16081023|BG11121 iolF|16081022|BG11122

		iolH|16081020|BG11123 iolI|16081019|BG11124 kdgA|16079268|BG11396 lcfA|16079908|BG11946 levD|16079761|BG10316 levE|16079760|BG10317 levF|16079759|BG10318 levG|16079758|BG10319 licA|16080908|BG11349 licB|16080910|BG11347 licC|16080909|BG11348 licH|16080907|BG11350 licT|16080959|BG10474 malA|16077885|BG11839 malP|16077887|BG11848 mdh|16079964|BG11386 mmgA|16079473|BG11319 mmgB|16079472|BG11320 mmgC|16079471|BG11321 mmgD|16079470|BG11322 mmgE|16079469|BG11323 mmsA|16081027|BG11117 msmX|16080932|BG11954 nupC|16080992|BG10984 pdp|50812307|BG10985 pta|16080818|BG10634 rbsR|16080644|BG10876 sacC|16079757|BG10320 scoA|16080950|BG11153 scoB|16078463|BG11154 treA|16077848|BG11010 treP|16077847|BG11009 treR|16077849|BG11011 uxaA|16078304|BG13213 uxaB|16078303|BG13212 uxaC|16078295|BG13204 uxuA|16078299|BG13208 xylA|16078823|BG10806 xylB|16078824|BG10807 xynB|16078821|BG11987 xynP|16078820|BG12262 ydhO|16077650|BG12192 yfiA|16077886|BG11847 yflN|16077829|BG12949 yjmB|16078296|BG13205 yjmC|16078297|BG13206 yjmD|16078298|BG13207 yjmF|16078300|BG13209 yobO|16078963|BG13507 yqiQ|16079468|BG11720 yxiE|16080976|BG11134 yxjC|50812305|BG11152 yxjF|16080948|BG11155 yxkF|16080933|BG12543 yxkJ|16080928|BG12546
ccpB|16081139|BG10045	**2**	gntR|16081057|BG10648 xylA|16078823|BG10806
ccpC|16078478|BG13297	**2**	citB|16078863|BG10478 citZ|16079966|BG10855
cggR|16080448|BG14085	**1**	gapA|16080447|BG10827
citR|16078008|BG10853	**1**	citA|16078009|BG10854
citT|16077826|BG12577	**1**	citM|16078996|BG10273
codY|16078680|BG10968	**29**	acsA|16080020|BG10370 citB|16078863|BG10478 comK|16078106|BG11059 comS|16077419|BG11045 dppA|16078357|BG10842 gabP|16077698|BG11328 hag|16080589|BG10655 hutP|16080985|BG10666 ilvA|16079236|BG10673 ilvB|16079883|BG10670 ilvC|16079881|BG10672 ilvD|16079246|BG11532 ilvN|16079882|BG10671 leuA|16079880|BG11948 leuB|16079879|BG10675 leuC|16079879|BG11949 leuD|16079877|BG11950 phrC|16077446|BG11959 ptb|16079465|BG11722 rapC|16077445|BG11966 srfAA|16077417|BG10168 srfAB|16077418|BG10169 srfAC|16077420|BG10170 srfAD|16077421|BG10171 ureA|16080719|BG11981 ureB|16080718|BG11982 ureC|16080717|BG11983 ybgE|16077308|BG12749 ypmP|16079235|BG11621
comA|16080219|BG10381	**10**	comFA|16080600|BG10395 comFB|16080599|BG10396 comFC|16080598|BG10397 degQ|16080223|BG10378 phrA|16078309|BG10653 phrE|16079637|BG11521 rapA|16078308|BG10652 rapC|16077445|BG11966 rapE|16079636|BG11299 srfAA|16077417|BG10168
comK|16078106|BG11059	**47**	addA|16078127|BG10466 addB|16078126|BG10465 comC|16079859|BG10323 comEA|16079613|BG10480 comEB|16079612|BG10481 comEC|16079611|BG10482 comFA|16080600|BG10395 comGA|16079529|BG10483 comGB|16079528|BG10484 comGC|16079527|BG10485 comGD|16079526|BG10486 comGE|16079525|BG10487 comGF|16079524|BG10488 comGG|16079523|BG10489 cspB|16077975|BG10824 degR|16079253|BG10699 flgK|16080594|BG10401 flgL|16080593|BG11936 flgM|16080596|BG10399 gidA|16081153|BG10059 gidB|16081152|BG10058 glcR|16080683|BG12503 nin|16077411|BG10839 nucA|50812179|BG10838 radC|16079856|BG10325 rapH|16077751|BG11031 recA|16078757|BG10721 rok|16078488|BG13307 rpsF|16081143|BG10049 rpsR|50812313|BG10047 sbcD|50812209|BG10467 smf|16078674|BG11006 ssb|16078513|BG10048 thdF|16081154|BG10060 ybdK|16077270|BG12724 yhcH|16077973|BG11586 yhjB|16078109|BG13069 yneA|16078849|BG11820 yneB|16078850|BG11821 ynzC|16078851|BG13466 yqzE|16079522|BG13771 yvrP|16080382|BG14152 yvyF|16080597|BG10398 yvyG|16080595|BG10400 ywpH|16080684|BG12502 yyaA|16081151|BG10057 yyaF|16081144|BG10050
cssR|16080354|BG14131	**3**	cssS|16080355|BG14132 htrA|16078355|BG12608 yvtA|50812291|BG14155
ctsR|16077151|BG10145	**11**	clpC|16077154|BG10148 clpE|16078434|BG12578 clpP|16080507|BG19016 clpX|16079874|BG11387 lonA|16079872|BG10338 mcsA|16077152|BG10146 mcsB|16077153|BG10147 sms|16077155|BG10149 trxA|16079902|BG10348 yacK|16077156|BG10150 ysxC|16079871|BG10339
degU|16080602|BG10393	**19**	aprE|16078094|BG10190 comC|16079859|BG10323 comEA|16079613|BG10480 comFA|16080600|BG10395 comGA|16079529|BG10483 comK|16078106|BG11059 degQ|16080223|BG10378 degR|16079253|BG10699 mecA|16078217|BG10680 sacB|16080498|BG10388 sacX|16080892|BG10560 sacY|16080893|BG10559 srfAA|16077417|BG10168 wapA|16080974|BG10797 yveA|16080500|BG12429 yveB|16080499|BG12430 yxjI|16080945|BG12539 yxjJ|16080944|BG12540 yxxG|16080973|BG10798
deoR|16080994|BG10982	**1**	dra|16080993|BG10983
dnaA|16077069|BG10065	**2**	dnaN|16077070|BG10066 sda|50812270|BG14183
exuR|16078302|BG13211	**1**	uxaC|16078295|BG13204
fnr|16080784|BG11343	**8**	hemZ|16081159|BG12999 narG|16080781|BG11081 narH|16080780|BG11082 narI|16080778|BG11084 narJ|16080779|BG11083 narK|16080785|BG11342 ywcJ|16080857|BG10594 ywiD|16080782|BG11345
fruR|16078502|BG12589	**2**	fruA|16078504|BG11938 fruK|16078503|BG12588
fur|16079409|BG11766	**34**	dhbA|16080253|BG11019 dhbB|16080250|BG11241 dhbC|16080252|BG11242 dhbE|16080251|BG11020 dhbF|50812288|BG11243 feuA|16077231|BG10835

		feuB|16077230|BG10836 feuC|16077229|BG10837 fhuD|16080386|BG10828 nasE|16077398|BG11097 ybbA|16077228|BG11565 ybbB|16077232|BG10834 yclN|16077448|BG12034 yclO|16077449|BG12035 yclP|16077450|BG12036 yclQ|16077451|BG12037 ydbN|16077520|BG12081 ydhU|16077656|BG12198 yfhA|16077913|BG12876 yfhC|16077915|BG12878 yfiZ|16077912|BG12902 yfkM|16077852|BG12929 yfmC|16077819|BG12954 yhfQ|16078097|BG13061 ykuN|16078479|BG13298 ykuO|16078480|BG13299 ykuP|16078481|BG13300 yoaJ|16078923|BG13481 ypbR|16079261|BG11604 yuiI|16080254|BG13974 yusV|16080346|BG14034 ywbL|16080879|BG10573 ywjA|16080776|BG11306 yxeB|16081012|BG11878
gabR|16077457|BG12042	**2**	gabD|16077459|BG12044 gabT|16077458|BG12043
gerE|16079893|BG10355	**20**	cgeA|16079036|BG11193 cgeB|16079037|BG11194 cgeC|16079035|BG11195 cgeD|16079034|BG11196 cgeE|16079033|BG11197 cotA|16077697|BG10490 cotB|16080658|BG10491 cotD|16079278|BG10493 cotV|16078243|BG10496 cotW|16078242|BG10497 cotX|16078241|BG10500 cotY|16078240|BG10498 cotZ|16078239|BG10499 cwlH|16079624|BG11633 ftsY|16078658|BG11539 spoIIIC|16079692|BG10919 spoIVCB|16079629|BG10459 sspG|50812290|BG14173 yoaN|16078927|BG13484 yurS|16080317|BG14005
glcT|16078452|BG12593	**3**	ptsG|16078453|BG10198 ptsH|16078454|BG10200 ptsI|50812224|BG10201
glnR|16078808|BG10424	**8**	glnA|16078809|BG10425 nasA|16077402|BG11093 nasB|16077401|BG11094 nasC|16077400|BG11095 nasD|16077399|BG11096 nasE|16077398|BG11097 nasF|16077397|BG11098 ureA|16080719|BG11981
glpP|16077992|BG10185	**5**	glpD|16077995|BG10188 glpF|16077993|BG10186 glpQ|16077282|BG10646 glpT|16077283|BG10645 yhxA|16077991|BG10184
gltC|16078907|BG10810	**2**	gltA|16078906|BG10811 gltB|16078905|BG12594
gltR|16079720|BG11942	**1**	gltA|16078906|BG10811
gutR|16077681|BG10178	**2**	gutB|16077682|BG10177 gutP|16077683|BG12795
hpr|16078063|BG10659	**14**	appA|16078203|BG11087 appB|16078204|BG11088 appC|16078205|BG11089 appD|16078201|BG11085 appF|16078202|BG11086 aprE|16078094|BG10190 nprE|16078534|BG10448 oppA|16078208|BG10771 oppB|16078209|BG10772 oppC|16078210|BG10773 oppD|16078211|BG10774 oppF|16078212|BG10775 sinI|16079516|BG10753 yclF|16077435|BG12027
hrcA|16079603|BG10662	**2**	groEL|16077670|BG10423 groES|50812190|BG10422
hxlR|16077416|BG11184	**2**	hxlA|16077415|BG11183 hxlB|16077414|BG11182
iolR|16081028|BG11364	**3**	iolS|16081029|BG11363 mmsA|16081027|BG11117 ydjK|16077690|BG12802
kdgR|16079270|BG11398	**3**	kdgA|16079268|BG11396 kdgK|16079269|BG11397 kdgT|16079267|BG11399
kipR|16077477|BG11214	**6**	kipA|50812187|BG14193 kipI|50812186|BG11231 ycsF|50812183|BG11227 ycsG|50812184|BG11228 ycsI|50812185|BG11230 ycsK|16077478|BG11232
lacR|16080470|BG12435	**5**	lacA|16080466|BG12439 yvfK|16080469|BG12436 yvfL|16080468|BG12437 yvfM|16080467|BG12438 yvfO|50812295|BG12440
levR|16079762|BG10677	**1**	levD|16079761|BG10316
lexA|16078848|BG10678	**4**	dinB|16077630|BG10539 tagC|16080630|BG10453 uvrA|16080570|BG10502 uvrB|16079901|BG10349
licR|16080911|BG11346	**1**	licB|16080910|BG11347
licT|16080959|BG10474	**2**	bglP|16080978|BG10934 bglS|16080958|BG10476
lmrA|16077337|BG12612	**1**	lmrB|16077336|BG12613
lrpA|16077572|BG12122	**1**	glyA|16080743|BG10944
lrpB|16077573|BG12123	**1**	glyA|16080743|BG10944
lytR|16080618|BG10404	**3**	lytA|16080617|BG10405 lytB|16080616|BG10406 lytC|16080615|BG10407
med|50812218|BG13126	**1**	comK|16078106|BG11059
mntR|16079508|BG11702	**5**	ydaR|16077503|BG12065 ytgA|16080129|BG13851 ytgB|16080128|BG13852 ytgC|16080127|BG13853 ytgD|16080126|BG13854
mta|16080713|BG12482	**4**	blt|16079712|BG10905 bmr|16079457|BG10303 htpG|16081033|BG11359 ydfK|16077612|BG12158
mtrB|16079334|BG10278	**4**	pabA|16077143|BG10138 trpE|16079325|BG10287 ycbK|16077323|BG11166 yhaG|16078065|BG12983
paiA|16080268|BG10695	**1**	nprE|16078534|BG10448
paiB|16080267|BG10696	**1**	nprE|16078534|BG10448
perR|16081061|BG12227	**12**	ahpC|16081061|BG11385 ahpF|16081062|BG11204 fur|16079409|BG11766 hemA|16079869|BG10340 hemB|16079865|BG10344 hemC|16079867|BG10342 hemD|16079866|BG10343 hemL|16079864|BG10345 hemX|16079868|BG10341 katA|16077947|BG10849 mrgA|16080351|BG10864 ykvW|16078449|BG13325
phoP|16079963|BG10363	**35**	glpQ|16077282|BG10646 phoA|16078006|BG10183 phoB|16077641|BG10697 phoD|16077331|BG11174 phoR|16079962|BG10364 pstA|16079552|BG11377 pstBA|16079551|BG11378 pstBB|16079550|BG11379 pstC|16079553|BG11376 pstS|16079554|BG11375 resA|16079372|BG10531 resB|16079371|BG10532

		resC|16079370|BG10533 resD|16079369|BG10534 resE|16079368|BG10535 tagA|16080628|BG10450 tagB|16080629|BG10451 tagD|16080627|BG10449 tagE|16080626|BG10724 tagF|16080625|BG10725 tuaA|16080614|BG12688 tuaB|16080613|BG12689 tuaC|16080612|BG12690 tuaD|16080611|BG12691 tuaE|16080610|BG12692 tuaF|16080609|BG12693 tuaG|16080608|BG12694 tuaH|16080607|BG12695 ydhF|16077640|BG12183 yhaX|50812207|BG13000 yhbH|16077963|BG11238 yjbC|16078214|BG13132 yjbD|16078215|BG13133 ykoL|16078398|BG13257 yttP|16080015|BG13927
pucR|16080295|BG13983	**15**	gde|16078382|BG13240 pucA|16080304|BG13992 pucB|16080303|BG13991 pucC|16080302|BG13990 pucD|16080301|BG13989 pucE|16080300|BG13988 pucH|16080294|BG13982 pucJ|16080296|BG13984 pucK|16080297|BG13985 pucL|16080298|BG13986 pucM|16080299|BG13987 ureA|16080719|BG11981 yurG|16080305|BG13993 yurH|16080306|BG13994 ywoE|16080700|BG12492
purR|16077115|BG10110	**22**	folD|16079487|BG11711 glyA|16080743|BG10944 guaC|16080266|BG12392 nusB|16079488|BG11710 pbuG|16077704|BG12811 pbuX|16079264|BG11080 purA|16081094|BG10002 purB|16077717|BG10707 purC|16077713|BG10703 purD|16077721|BG10711 purE|16077712|BG10700 purH|16077720|BG10710 purK|16077711|BG10701 purL|16077716|BG10705 purM|16077718|BG10708 purN|16077719|BG10709 purQ|16077715|BG10706 purS|16077714|BG10704 pyrP|16078612|BG10992 xpt|16079265|BG11079 yabJ|16077116|BG10111 ytiP|16080051|BG13864
pyrR|16078611|BG10712	**1**	pyrP|16078612|BG10992
resD|16079369|BG10534	**13**	ctaA|16078551|BG10213 ctaB|16078552|BG10214 fnr|16080784|BG11343 hemZ|16081159|BG12999 hmp|16078369|BG11418 ldh|16077374|BG19003 nasD|16077399|BG11096 phoP|16079963|BG10363 qcrA|16079313|BG11325 qcrB|16079312|BG11326 qcrC|16079311|BG11327 resA|16079372|BG10531 sbo|16080788|BG12671
rocR|16081087|BG10723	**3**	rocA|16080830|BG10622 rocD|16081086|BG10722 rocG|16080831|BG10621
rok|16078488|BG13307	**1**	comK|16078106|BG11059
sacT|16080858|BG10593	**4**	sacA|16080855|BG10596 sacP|50812303|BG10595 sacX|16080892|BG10560 ywdA|16080854|BG10597
sacY|16080893|BG10559	**1**	sacX|16080892|BG10560
sinR|16079517|BG10754	**6**	aprE|16078094|BG10190 comK|16078106|BG11059 rok|16078488|BG13307 spo0A|16079478|BG10765 spoIIAA|16079404|BG10296 spoIIGA|16078595|BG10234
spo0A|16079478|BG10765	**12**	abrB|16077105|BG10100 argC|16078184|BG10191 dltA|16080901|BG10551 kinA|16078463|BG10204 kinC|16078513|BG10989 rbsR|16080644|BG10876 sinI|16079516|BG10753 spo0F|16080766|BG10411 spoIIAA|16079404|BG10296 spoIIE|16077132|BG10127 spoIIGA|16078595|BG10234 yqxM|16079520|BG11076
spoIIID|16080695|BG10408	**11**	bofA|16077091|BG10087 cotA|16077697|BG10490 cotD|16079278|BG10493 cotX|16078241|BG10500 gerE|16079893|BG10355 spoIID|16080728|BG10766 spoIIIAA|16079499|BG10540 spoIVCA|16079630|BG10458 spoIVCB|16079629|BG10459 spoVD|16078581|BG10222 spoVE|16078585|BG10226
spoVT|16077124|BG10119	**15**	bofC|16079827|BG11917 csgA|16077276|BG11504 dacF|16079405|BG10295 gerAA|16080358|BG10385 gerBA|16080633|BG10640 gerD|16077223|BG10644 gpr|16079608|BG10438 sigG|16078597|BG10236 spoIVB|16079479|BG10311 spoVAA|16079401|BG10892 sspA|16080009|BG10786 sspB|16078040|BG10787 sspD|16078411|BG10788 sspE|16077932|BG10789 ycxE|16077460|BG11066
tenA|16078230|BG10791	**4**	aprE|16078094|BG10190 nprE|16078534|BG10448 phoA|16078006|BG10183 sacB|16080498|BG10388
tenI|16078231|BG10792	**1**	tenA|16078230|BG10791
tnrA|16078396|BG11805	**31**	alsT|16078873|BG11798 gabP|16077698|BG11328 glnH|16079797|BG11486 glnM|16079798|BG11487 glnP|16079799|BG11488 glnQ|16079796|BG11485 glnR|16078808|BG10424 gltA|16078906|BG10811 nasA|16077402|BG11093 nasB|16077401|BG11094 nasD|16077399|BG11096 nrgA|16080704|BG10869 oppA|16078208|BG10771 pel|16077823|BG10840 ptb|16079465|BG11722 pucJ|16080296|BG13984 ureA|16080719|BG11981 yccC|16077338|BG12755 ycsF|50812183|BG11227 ykzB|16078397|BG13330 yodF|16079016|BG13535 yttA|16080088|BG13924 ywdI|16080846|BG10605 ywdJ|16080845|BG10606 ywdK|16080844|BG10607 ywlF|16080745|BG10942 ywlG|16080744|BG10943 ywrD|16080663|BG12523 yxkC|16080936|BG12541 yycB|16081100|BG10007 yycC|16081099|BG10006
treR|16077849|BG11011	**1**	treP|16077847|BG11009
xpf|16078321|BG10998	**10**	xkdE|16078324|BG11540 xkdF|16078325|BG11541 xkdG|16078326|BG11542 xkdH|16078327|BG11543 xkdI|16078328|BG11544 xkdJ|16078329|BG11545 xkdK|16078330|BG11546 xkdM|16078331|BG11547 xtmA|16078322|BG10999 xtmB|16078323|BG11000
xre|16078316|BG10994	**4**	xkdB|16078317|BG10995 xkdC|16078318|BG10996 xkdD|16078319|BG10997 xtrA|16078320|BG11559
xylR|16078822|BG11986	**2**	xylA|16078823|BG10806 xynP|16078820|BG12262
ydbG|16077513|BG12074	**1**	dctP|16077514|BG12075

ydiH|16077664|BG12205	**4**	cydA|16080927|BG11925 cydB|16080926|BG11926 cydC|16080925|BG11927 cydD|16080924|BG11928
yfhP|16077928|BG12890	**1**	yfhQ|16077929|BG12891
yfiA|16077886|BG11847	**1**	malA|16077885|BG11839
ykmA|16078380|BG13239	**1**	yklA|16078379|BG13238
ylbO|16078573|BG13367	**2**	cgeA|16079036|BG11193 cotY|16078240|BG10498
yocG|16078980|BG13520	**1**	des|16078978|BG13518
yufM|16080205|BG12348	**3**	maeN|16080210|BG12353 yflS|16077824|BG12951 ywkA|16080758|BG11312
ywfK|16080817|BG10635	**2**	yvgQ|16080396|BG14099 yvgR|16080397|BG14100
ywiD|16080782|BG11345	**5**	alsS|16080654|BG10471 hemN|16079604|BG11395 hemZ|16081159|BG12999 ldh|16077374|BG19003 nasD|16077399|BG11096
yycF|16081093|BG10001	**4**	ftsA|16078592|BG10231 tagD|16080627|BG10449 ykvT|16078446|BG13322 yocH|16078981|BG13521
zur|16079565|BG11668	**3**	ycdH|16077354|BG12763 yciA|16077403|BG12019 yciC|16077405|BG12021

ANEXO 3A. (Was translate to online web: http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/)

Collection of non-redundant pairs of genes of the metabolic pathways defined according to the KEGG database in *E.coli* and *B. subtilis*.

ANEXO 3B.

Obtaining thresholds for identifying co-evolving TF-TG pairs in the TRNs by comparing against co-occuring pairs of genes in metabolic pathways.

An average of distances D1 and D2 was calculated for the complete non-redundant collection of pairs (P1-P2) in each genomes' pathways. However this average distance D1 or D2 is an over estimate of the extent of conservation for regulatory interactions so a (Average-standard deviation) was employed to generate significantly conserved interacting pairs. Since in metabolic interactions D1 and D2 are synonymous unlike Dtf and Dtg in regulatory interactions we used a final threshold of

$$T = \{(\langle D1 \rangle - STD1) + (\langle D2 \rangle - STD2)\} / 2$$

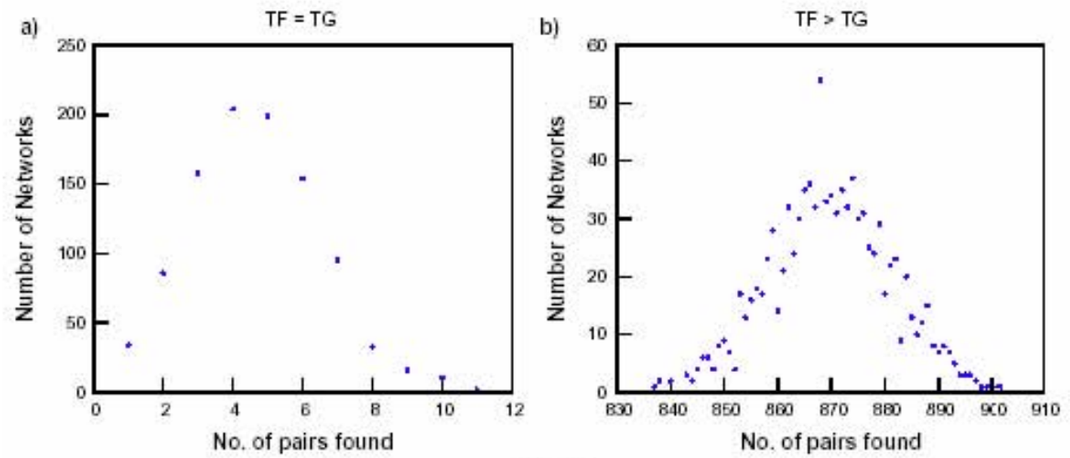
Where <D1> and <D2> are the average values of D1 and D2 for all pairs and STD1 and STD2 are the standard deviations of D1 and D2.
To best represent a threshold to obtain highly conserved pairs. The process was repeated for both the genomes using the respective known pathways documented in Kegg database.

Mesarues	KEGG in *Escherichia coli*	KEGG in *Bacillus subtilis*
AVG 1	0.472009789418263	0.480096384161564
AVG 2	0.517286194952669	0.519944735263887
STD 1	0.281274994523379	0.321133581894223
STD 2	0.279065684017891	0.310716106014003
FINAL SC 1	0.190734794894885	0.158962802267342
FINAL SC 2	0.238220510934779	0.209228629249884
THRESHOLD	0.214477652914832	0.184095715758613

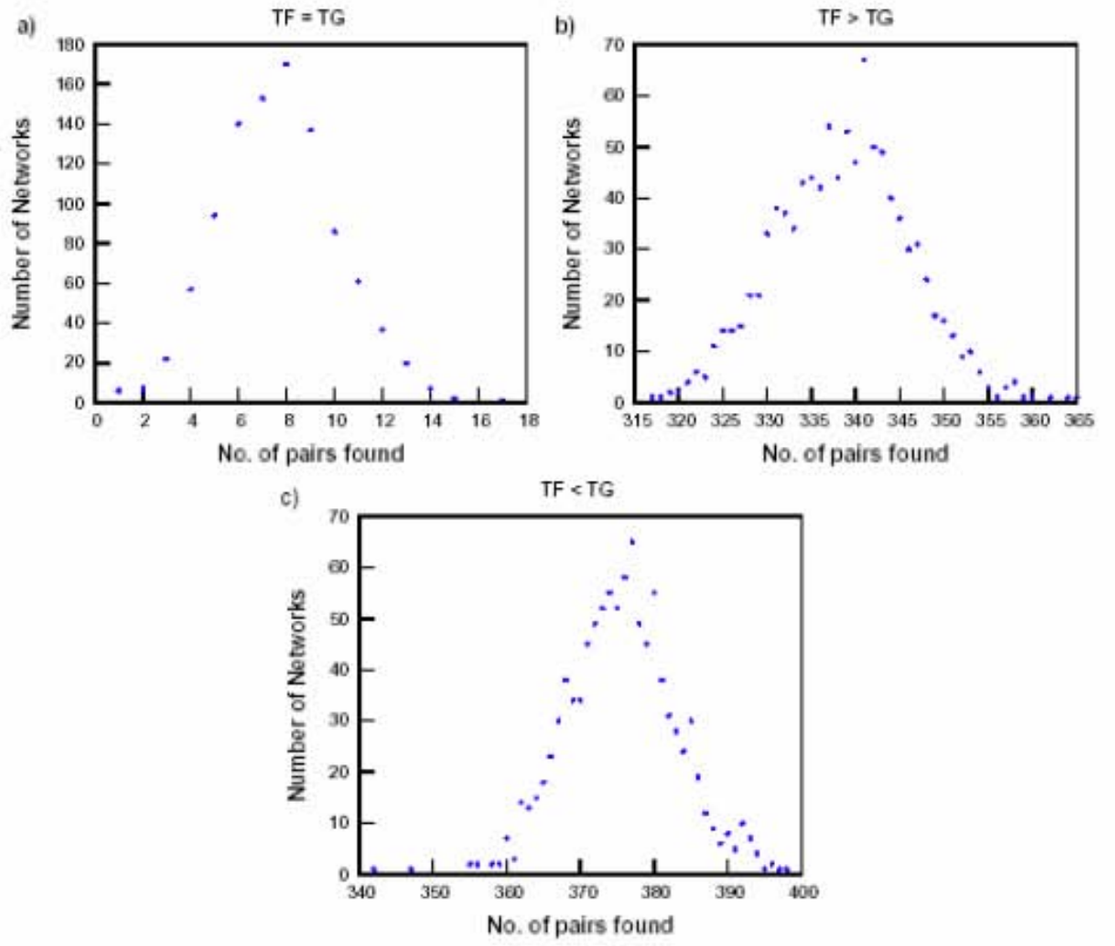
ANEXO 4B.

Figures showing that the distributions of the quantities (TF=TG, TF> TG and TF < TG) in the randomized networks (based on 1000 random networks) are normal. In each case we found that the distribution can be easily approximated to normal and hence tests used for normal distributions were employed.

Escherichia coli



Bacillus subtilis



Escherichia coli K12 *Threshold*: 0.2145			*Bacillus subtilis* *Threshold*: 0.1841		
Categoría	**Interacciones**	**Z-score (P-value)**	**Categoría**	**Interacciones**	**Z-score (P-value)**
TF = TG	**15**	**5.44 (< 0.0001)**	**TF = TG**	**15**	**2.99 (0.0028)**
TF > TG	**813**	**-5.06 (< 0.0001)**	**TF > TG**	**363**	**3.24 (0.0012)**
TF < TG	**759**	**3.68 (0.00023)**	**TF < TG**	**349**	**-3.60 (0.00032)**

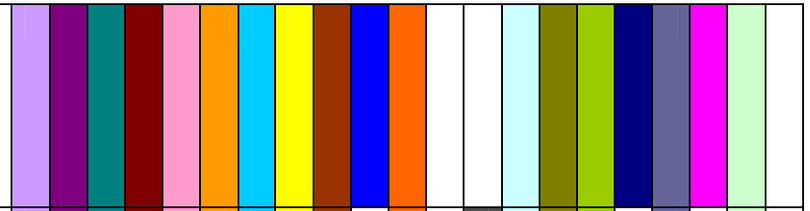
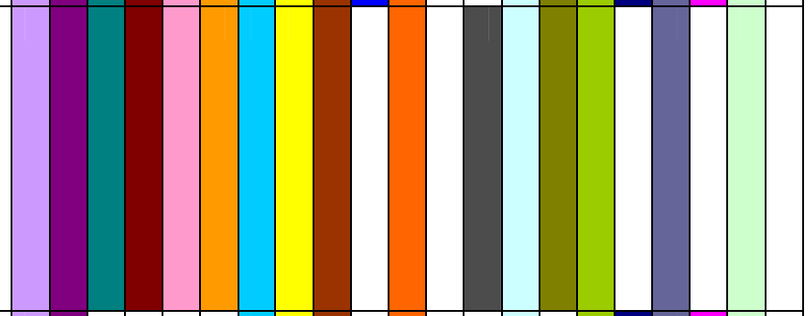
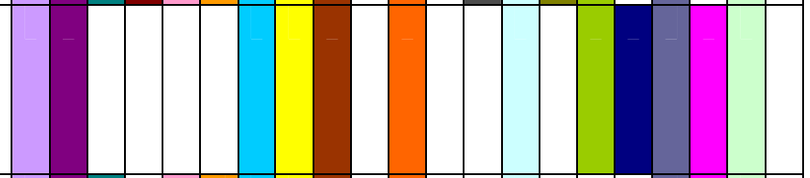
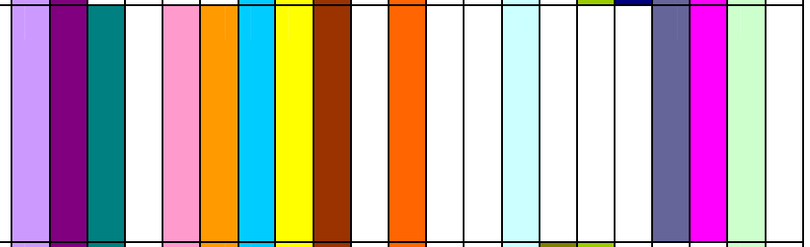
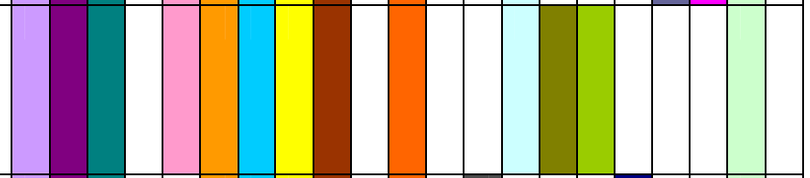
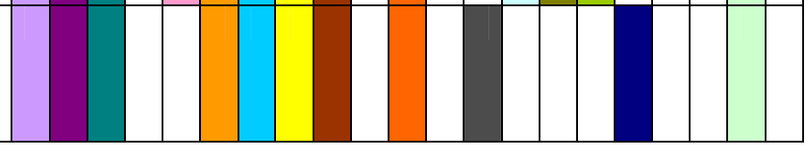
ANEXO 4A. (Was translate to online web**:** http://www.ccg.unam.mx/Computational_Genomics/TRNS/conservation/**)**

Figure that respresent the methodology to reconstruct random TRNs from *E.coli* and *B. subtilis*.

ANEXO 5.

Conserved interactions across bacteria from the perspective of the TRNs of *E. coli* and *B. subtilis*.

TF	FUNCTION (from REGULON DataBase and DBTBS)	NO. TG's	CONSERVATED TG's	BIOLOGICAL DISTRIBUTION
				GPE GP BP EP DP AP BF LF CF MF ACT FU PLA CHL SPI CYA GS NGS DEI HYP A
FUR*	Fur is a negative transcriptional regulator that is involved in the regulation of a large number of operons that encode enzymes involved in **iron transport** (fecANCDE, fecIR, and fhuACDB) including several outer-membrane proteins. It is activated by Mn2+ and belongs to the Fur family of transcriptional regulators.	**34**	fecD, fecE, nrdE, sodA	
IHF*	IHF consist of two subunits, they are the products of the himA and himD genes. This protein introduces sharp bends in the **DNA** and is required for site-specific **recombination, DNA replication, plus transcriptional and translational control.** IHF is essential in the **lysogenic life cycle of bacteriophage lambda.** The protein belongs to the HI-HN-S family.	**162**	acs, carA, carB, dps, dusB, gltA, hemA, hemF, ilvD, ilvE, ilvG_1, ilvG_2, nuoE, nuoF, nuoH, nulo, nuoM, nuoN, phoU, pstA, pstB, pstC, sodB, sucA, sucB, sucC, sucD	
LRP*	This regulator participates in controlling several genes involved in the high-affinity branched-chain **amino acid transport system**. It is mediator of the leucine response. Lrp affects several operons. The protein belongs to the AsnC family.	**61**	dadA, gabD, gcvH, gcvP, gcvT, gltB, gltD, ilvA, ilvD, ilvE, ilvG_1, ilvG_2, ilvH, ilvI, kbl, legua, leuB, leuC, leuD, livF, livG, Lib, lysU, oppD, oppF, serA, serC	
AsnC*	This protein controls the expression of two genes (asnA and asnC) involved in **asparagine biosynthesis**.It controls its own synthesis and negatively controls the expression of gidA at a post-transcriptional level. It is a dual transcriptional regulator that belongs to the AsnC family.	**2**	ansA	
NagC*	This regulator participates in controlling several genes coding for **catabolic enzymes**. NagC represses the nagE gene and the manXYZ, nagBACD operons. However it exerts a dual regulatory effect on the glmUS operon The site length where the protein binds comes from a single site (medline number 9814342) The protein belongs to the NagC/XylR family.	**10**	glmS, glmU, nagA, nagB	

DnaA*	It plays an important role in **initiating chromosomal replication** to bind sequences in the chromosomal origin. In addition, it binds to ATP and to acidic phospholipids. DnaA can inhibit its gene expression as well as other genes such as guaA, dam, rpoH, ftsA, and mioC. It belongs to the DnaA family of transcriptional regulators.	**9**	aldA, dnaN, guaA, guaB, nrdA, nrdB, recF, rpoH	
phoB*	Phosphate regulon transcriptional regulatory protein PhoB is a member of the two-component regulatory system PhoR/PhoB. The PhoR/PhoB system is involved in the regulation of the phosphate regulon gene expression. Under conditions of **phosphate limitation** the PhoB protein is phosphorylated by phospho-PhoR. In this phosphorylated state phospho-PhoB acts as a transcriptional activator of the Pho regulon. These controls are regulated by the carbon and energy source. The protein belongs to the two-component family.	**27**	PhoB, asr, phoA, phoU, pstA, pstB, pstC, pstS	
ArgR*	This protein negatively controls the expression **of arginine biosynthesis** in addition to the carAB operon. ArgR is also essential in the resolution of plasmid ColE1 multimers during replication through cer-mediates site-specific recombination. It belongs to the ArgR family.	**16**	argB, argC, argH, argI, ArgR, carA, carB	
BirA	The birA gene codes for a multifunctional protein, possessing both enzymatic and regulatory activities. This protein negatively controls the expression of the **biotin-operon** (bioBFCD and bioA) and the enzyme that synthesizes the co repressor.The enzymatic functions include the synthesis of the enzyme-bound biotinyl-5'-adenylate (bio-5'-AMP), and the transfer of the biotin.	**5**	bioA, bioB	
LexA*	This regulator participates in controlling several genes involved in the **SOS response**. The recA protein interacts with LexA causing an autocatalytic cleavage, which disrupts the DNA-binding capacity of LexA. The protein belongs to the LexA family.	**14**	LexA, dnaG, recA, rpsU, ssb, uvrA, uvrB, uvrD	
RbsR	This regulator participates in controlling several genes involved in **ribose metabolism**. RbsR repress the transcription of the rbsDACBK operon. When D-ribose binds to RbsR the protein becomes inactive. The protein belongs to the GalR/LacI family.	**5**	rbsA, rbsB, rbsC, rbsD, rbsK	

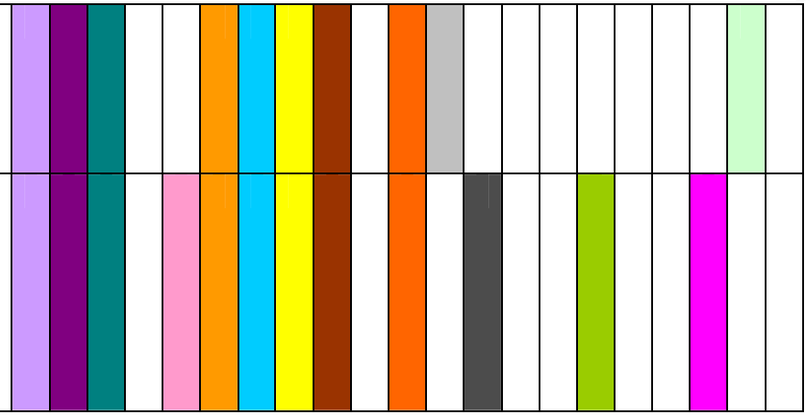
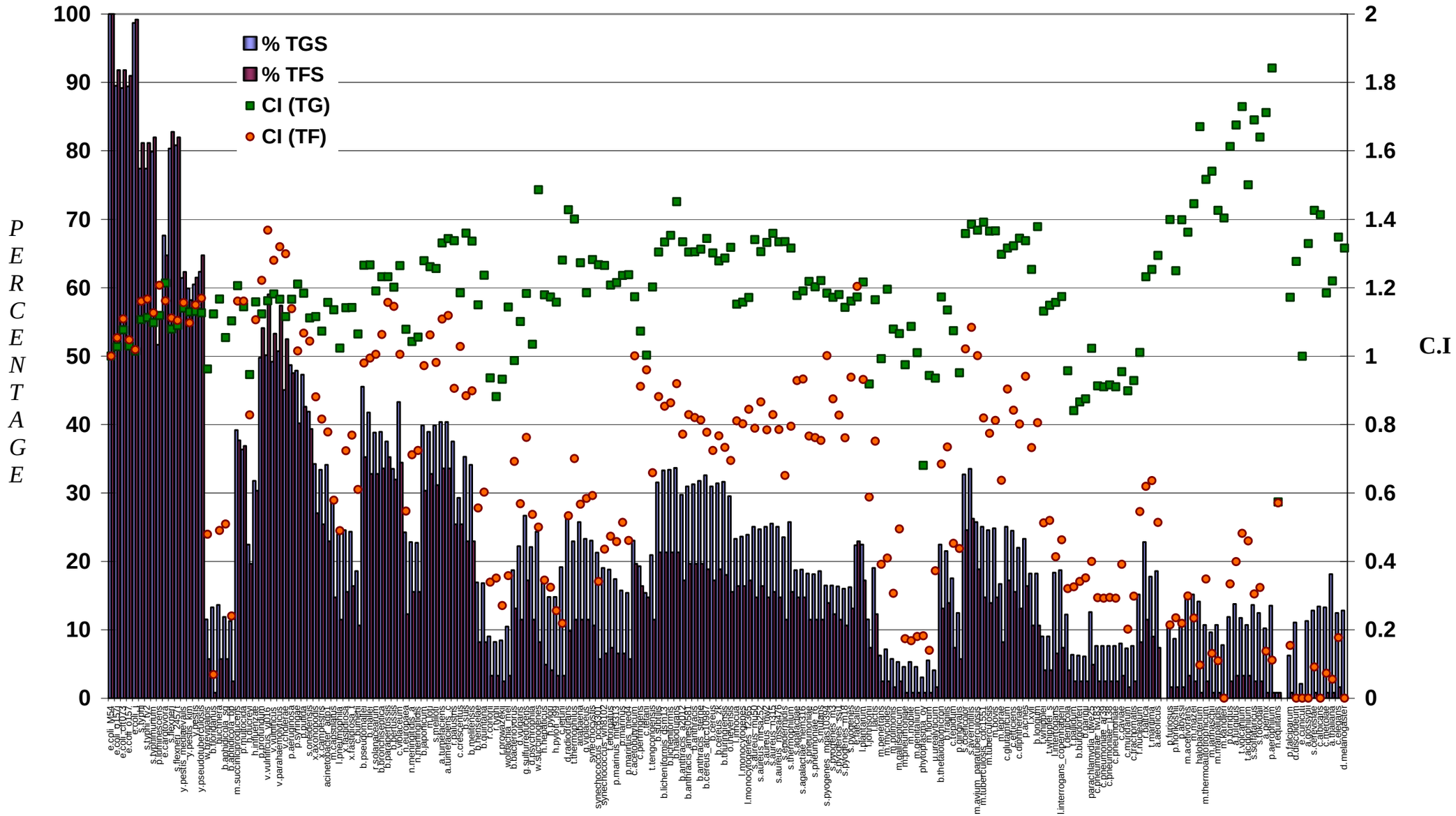
GlpR	It is the repressor specifically bound to control regions of the glpD, glpFK, glpTQ, and glpACB operons. The binding of DNA by the repressor was diminished in the presence of sn-glycerol 3-phosphate. It belongs to the DeoR family of transcriptional regulators.	**8**	GlpD, glpF, glpK	
OxyR*	This regulator participates in controlling several genes involved in the **response to oxidative stress** and the production of surface proteins that control the colony morphology and auto-aggregation ability. It regulates intracellular hydrogen peroxide levels by activating genes such dps, fur, and katG. OxyR is negatively autoregulated. The protein belongs to the LysR family.	**13**	ahpC, dps, fur, sufB, sufC, sufS	

Table. Biological distribution and additional information about the regulatory interactions well conserved in bacteria. * means that transcription factor is regulated by itself (autoregulation). **GPE:** gamma proteobateria enterobacteria, **GP:** other gamma proteobacteria, **BP:** beta proteobacteria, **EP**: epsilon proteobacteria, **DP: delta proteobacteria, AP:** alfa proteobacteria, **BF:** bacillales, **LF:** lactobacillales, **CF:** clostridia, **MF:** mollicutes, **ACT:** actinobacteria, **FU:** fuseobacteria, **PLA:** planctomyces**, CHL:** chlamydias, **SPI:** spirochetes, **CYA:** cyanobacteria, **GS:** green sulfur bacteria, **NGS:** non green sulfur bacteria, **DEI:** deinococcus-thermus, **HYP:** hyperthermophilic bacteria, **A:** archaea.

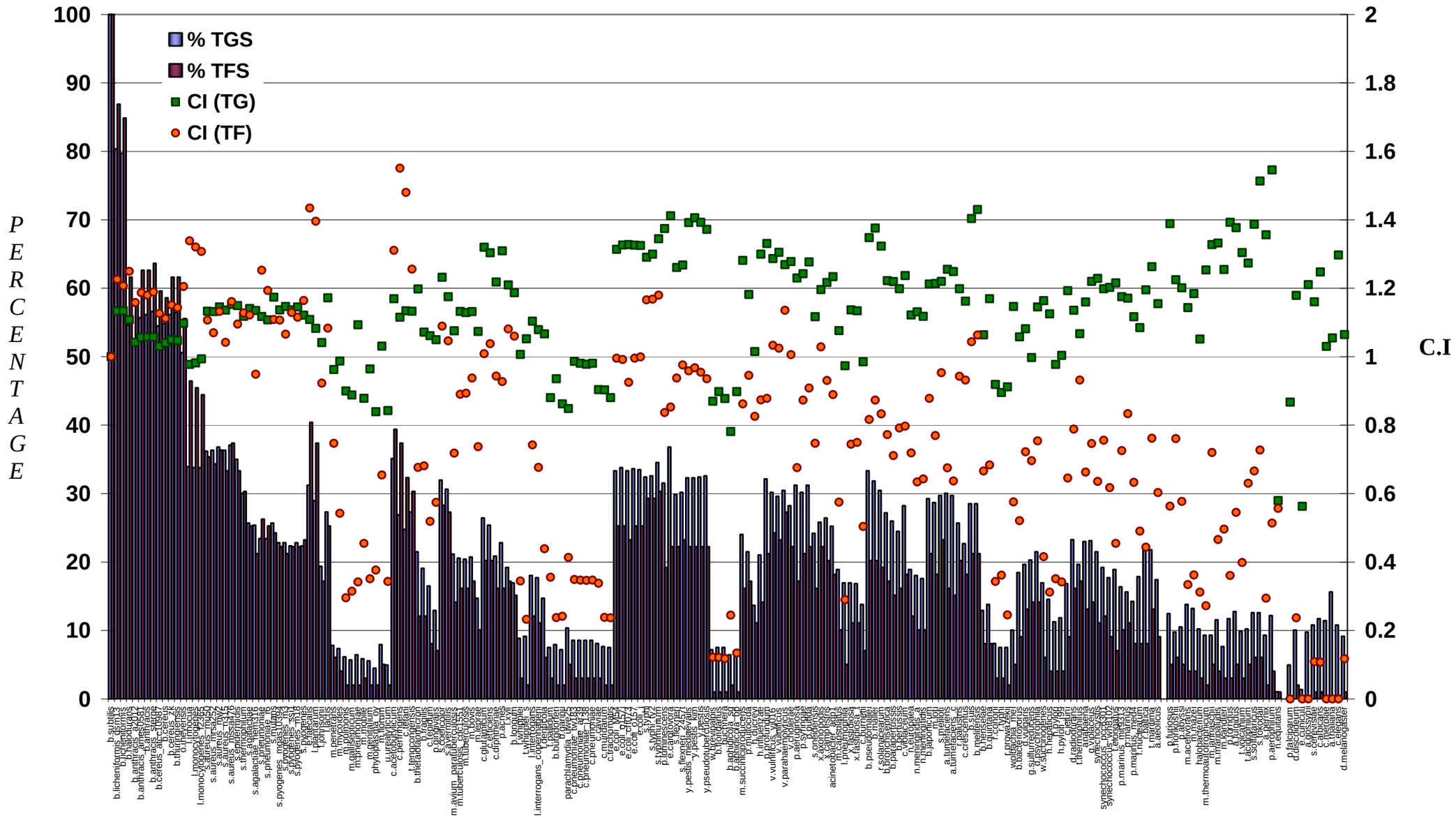
ANEXO 6A.

Distribution of TF and TG orthologs for *Escherichia coli K12* and their C_{TF} - C_{TG} index values in 204 complete genomes.



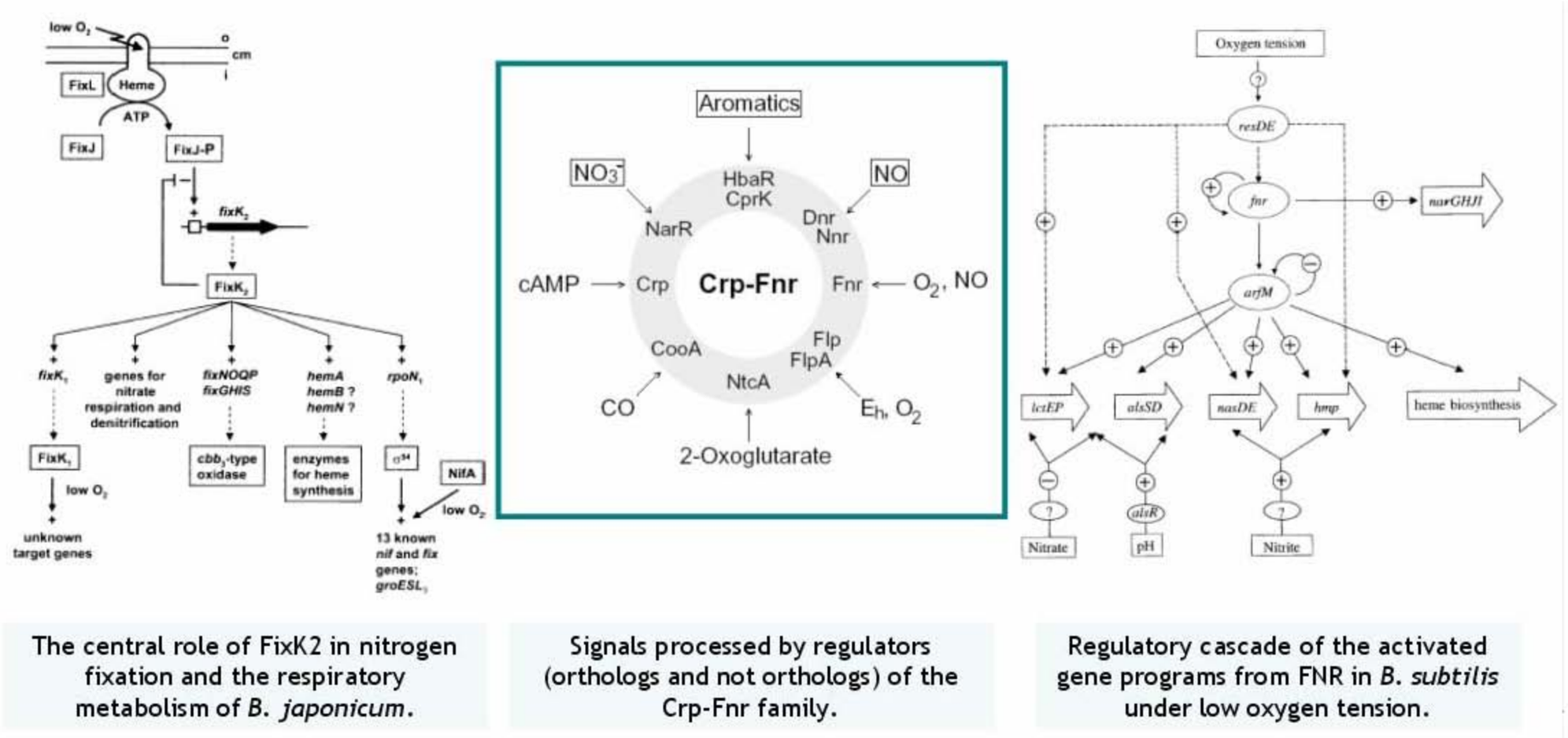
ANEXO 6B.

Distribution of TF and TG orthologs for *Bacillus subtilis* and their C_{TF} - C_{TG} index values in 204 complete genomes.



ANEXO 8A.

The regulatory adaptability and structural flexibility represented in the Crp-Fnr scaffold has led to the evolution of an important group of physiologically versatile transcription factors (Körner *et al.*, 2003).



The central role of FixK2 in nitrogen fixation and the respiratory metabolism of *B. japonicum*.

Signals processed by regulators (orthologs and not orthologs) of the Crp-Fnr family.

Regulatory cascade of the activated gene programs from FNR in *B. subtilis* under low oxygen tension.

ANEXO 8B.

Is the global regulatory role associated with *E. coli* Lrp limited to enteric bacteria?

To probe this question Friedberg *et al.* (2001) investigated LrfB, an Lrp-related protein from *Haemophilus influenzae* that shares 75% sequence identity with *E. coli* Lrp (highest sequence identity among 42 sequences compared). A strain of *H. influenzae* having an *lrfB* null allele grew at the wild-type growth rate but with a filamentous morphology. A comparison of two-dimensional (2D) electrophoretic patterns of proteins from parent and mutant strains showed only two differences (comparable studies with *lrp*1 and *lrp E. coli* strains by others showed 20 differences). The abundance of LrfB in *H. influenzae*, estimated by Western blotting experiments, was about 130 dimers per cell (compared to 3,000 dimers per *E. coli* cell). LrfB expressed in *E. coli* replaced Lrp as a repressor of the *lrp* gene but acted only to a limited extent as an activator of the *ilvIH* operon. Thus, although LrfB resembles Lrp sufficiently to perform some of its functions, its low abundance is consonant with a more local role in regulating but a few genes, a view consistent with the results of the 2D electrophoretic analysis. Friedberg *et al.* (2001) speculate that an Lrp having a global regulatory role evolved to help enteric bacteria adapt to their ecological niches and that it is unlikely that Lrp-related proteins in other organisms have a broad regulatory function.

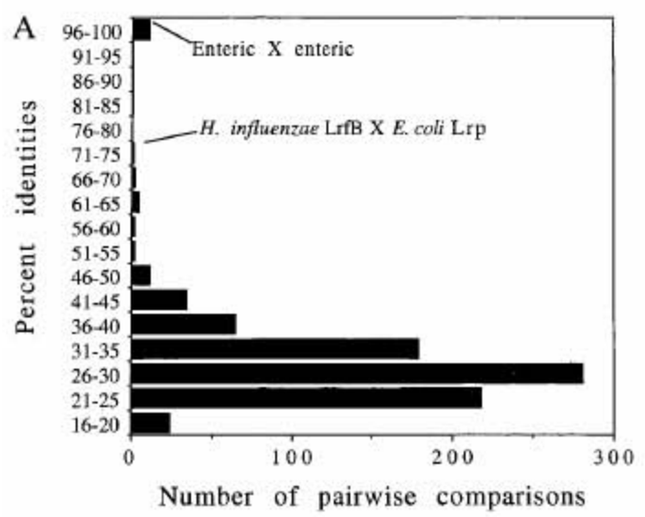


FIG. 1. Comparison of Lrp-related proteins from a number of microorganisms. (A) Amino acid identities for pairwise combinations of sequences. Amino acid sequences of the following Lrp-related proteins were compared in pairwise combinations using the Genetics Computer Group GAP alignment program (12). (B) Highly conserved positions in Lrp-related proteins. The Genetics Computer Group Pileup program (12) was used to align sequences identified in the legend above by a tilde. The sequence of *E. coli* Lrp is shown, with asterisks designating amino acids that are completely conserved among the 23 proteins. Cases in which a position is limited to only two amino acids are also shown. The helix-turn-helix region is identified with an overline.

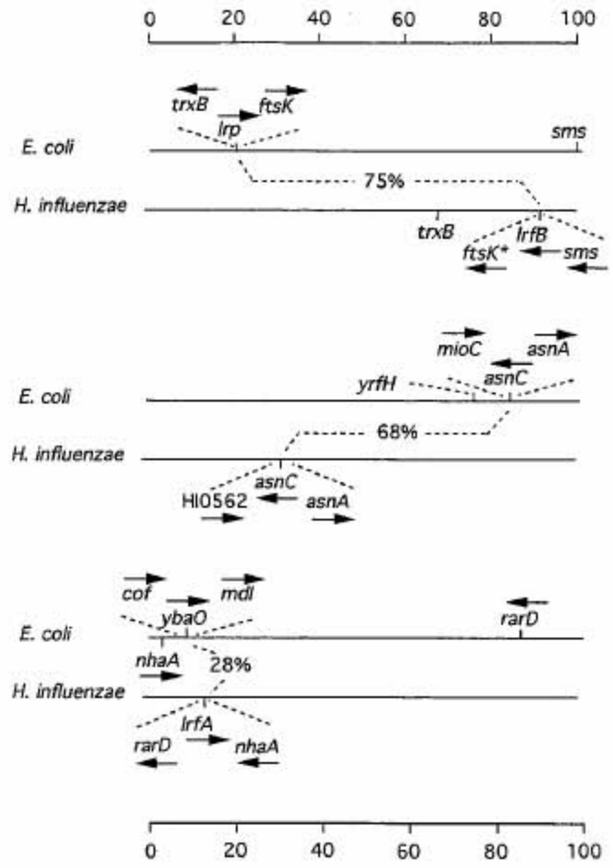


FIG. 2. Genes related to *lrp* in *E. coli* and *H. influenzae*. Each horizontal line represents the genome of *E. coli* or *H. influenzae* on a scale of 100 units. Arrows represent direction of transcription. The *lrp*-related genes *lrp, asnC, ybaO, lrfA* (HI0224), and *lrfB* are shown together with flanking genes. The percent predicted amino acid identities of proteins encoded by pairs of *lrp*-related genes is indicated. *ftsK** (HI1592, HI1593, HI1594, and HI1595) is related to *E. coli ftsK* as follows: HI1595 is weakly related to the N-terminal amino acids of FtsK (unfiltered BLAST score, 110), and HI1592 is highly related to the C terminus of FtsK (BLAST score, 1358). Orthologs are homologous genes in different organisms that encode proteins with the same function and that have evolved by direct, vertical descent. Two genes are considered to be orthologs if their predicted amino acid sequences fulfill the following criteria (37): they show higher similarity to each other than to other proteins in either organisms, they show a higher similarity to each other than to homologs from phylogenetically more distant organisms, and they align throughout most of their lengths. In addition, an orthologous relationship between two genes is strengthened if the two are flanked on one or both sides by orthologous gene pairs.

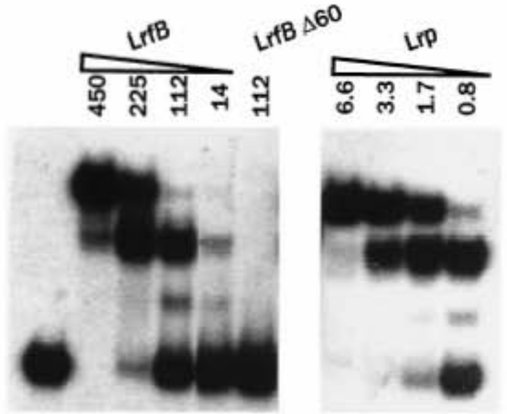


FIG. 3. Binding of Lrp and LrfB to sites upstream of the *ilvIH* promoter as visualized by gel retardation. A ^{32}P-labeled 276-bp DNA fragment from a region upstream of the *E. coli ilvIH* promoter containing six binding sites for Lrp was mixed with no protein (lane 1), purified 6×His-LrfB (lanes 2 to 5), purified 6×His-LrfBΔ60 (lacking 60 C-terminal amino acids) (lane 5), and purified 6×His-Lrp (lanes 7 to 10), and samples were fractionated by electrophoresis. The concentrations (nanomolar) of proteins are indicated. For the experiment with Lrp, the faster-moving band contains Lrp bound to sites 1 and 2, and the slower-moving band contains Lrp bound to sites 1, 2, 3, 4, 5, and 6 (38).

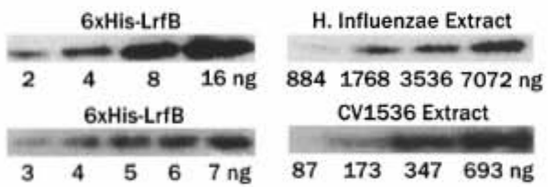


FIG. 5. Amount of LrfB in *H. influenzae* (A) and in *E. coli* strain CV1536 (B) as determined by Western blotting. A standard curve prepared using 6×His-LrfB was employed to estimate that 0.073% of the total protein in *H. influenzae* is LrfB. The number of dimers per cell was calculated assuming that the total protein per unit volume is the same for *E. coli* and *H. influenzae*. The following parameters were used in the calculation: total protein per *E. coli* cell, 0.156 pg (28); volume of an *E. coli* cell, 0.87×10^{-12} ml, calculated from the weight of a cell together with the densities of the components of a cell (28) (a similar number is derived from assuming that the dimensions of an *E. coli* cell are 0.75 by 2 μm); volume of an *H. influenzae* cell, 0.636×10^{-12} ml, calculated assuming dimensions of 0.3 by 1 μm (27); LrfB dimer molecular mass, 37.8 kDa.

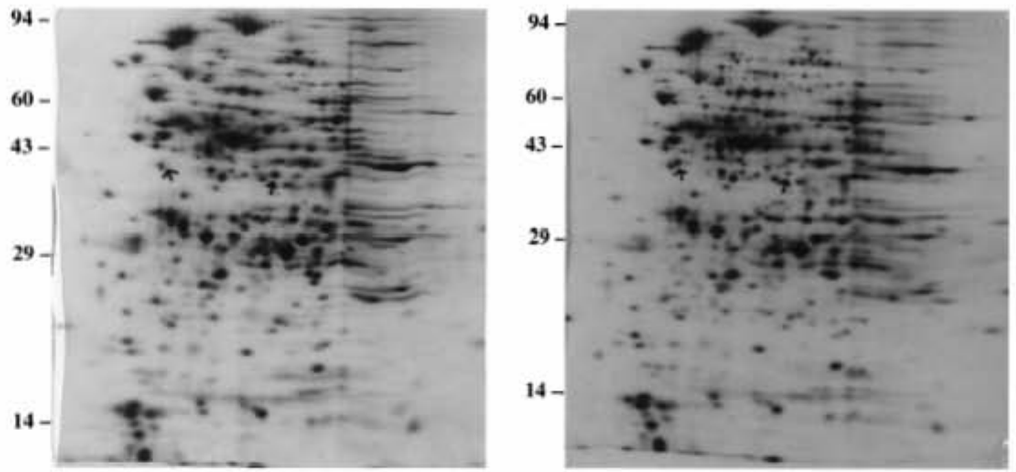


FIG. 4. 2D gel analysis of proteins in extracts of wild-type (A) and *lrfB::kan* (B) strains of *H. influenzae*. For the isoelectric focusing dimension, the anode is on the left. Cells were grown in MIc medium and pulse-labeled with [^{35}S]methionine. Arrows indicate the positions of the two spots that were reproducibly different in the two strains.

TABLE 2. LrfB activates expression from the *E. coli ilvIH* promoter and represses expression from the *lrp* promoter

Strain[a]	*lrp* or *lrfB* allele on:		Promoter-*lacZ* fusion	Addition to medium[b]	Sp act of β-galactosidase[c]	Relative sp act (%)
	Chromosome	Plasmid				
CV975	*lrp*$^+$		P*ilvIH*		359 ± 20	100
CV975	*lrp*$^+$		P*ilvIH*	Leu	37.4 ± 2.6	10
CV1535	*lrp-35*::Tn*10*		P*ilvIH*		3.0 ± 2.1	0.83
CV1536	*lrp-35*::Tn*10*	*lrfB*	P*ilvIH*		43.3 ± 9.1	12
CV1536	*lrp-35*::Tn*10*	*lrfB*	P*ilvIH*	Leu	6.6 ± 0.34	1.8
CV1534	*lrp-35*::Tn*10*		P*lrp*		774 ± 53	100
CV1528	*lrp-35*::Tn*10*	*lrfB*	P*lrp*		<0.5	<0.06

[a] Strains CV975, CV1535, and CV1536 were grown to log phase in sSSA minimal medium containing 50 μg of ampicillin per ml (except for CV975). Strains CV1534 and CV1528 were grown in LB medium containing 100 μg of ampicillin per ml.
[b] When present, leucine was at 50 μg/ml.
[c] Specific activity is in Miller units. Data are averages ± standard deviations of at least two different experiments, each performed in duplicate.

ANEXO 8C.

Role of CcpA in Regulation of the Central Pathways of Carbon Catabolism in *Bacillus subtilis.*

The *Bacillus subtilis* two-dimensional (2D) protein index contains almost all glycolytic and tricarboxylic acid (TCA) cycle enzymes, among them the most abundant housekeeping proteins of growing cells. Therefore, a comprehensive study on the regulation of glycolysis and the TCA cycle was initiated (Tobisch *et al.*, 1999. Whereas expression of genes encoding the upper and lower parts of glycolysis (*pgi*, *pfk*, *fbaA*, and *pykA*) is not affected by the glucose supply, there is an activation of the glycolytic *gap* gene and the *pgk* operon by glucose. This activation seems to be dependent on the global regulator CcpA, as shown by 2D polyacrylamide gel electrophoresis analysis as well as by transcriptional analysis. Furthermore, a high glucose concentration stimulates production and excretion of organic acids (overflow metabolism) in the wild type but not in the *ccpA* mutant. Finally, CcpA is involved in strong glucose repression of almost all TCA cycle genes. In addition to TCA cycle and glycolytic enzymes, the levels of many other proteins are affected by the *ccpA* mutation. Tobisch *et al.* (1999), suggest (i) that *ccpA* mutants are unable to activate glycolysis or carbon overflow metabolism and (ii) that CcpA might be a key regulator molecule, controlling a superregulon of glucose catabolism.

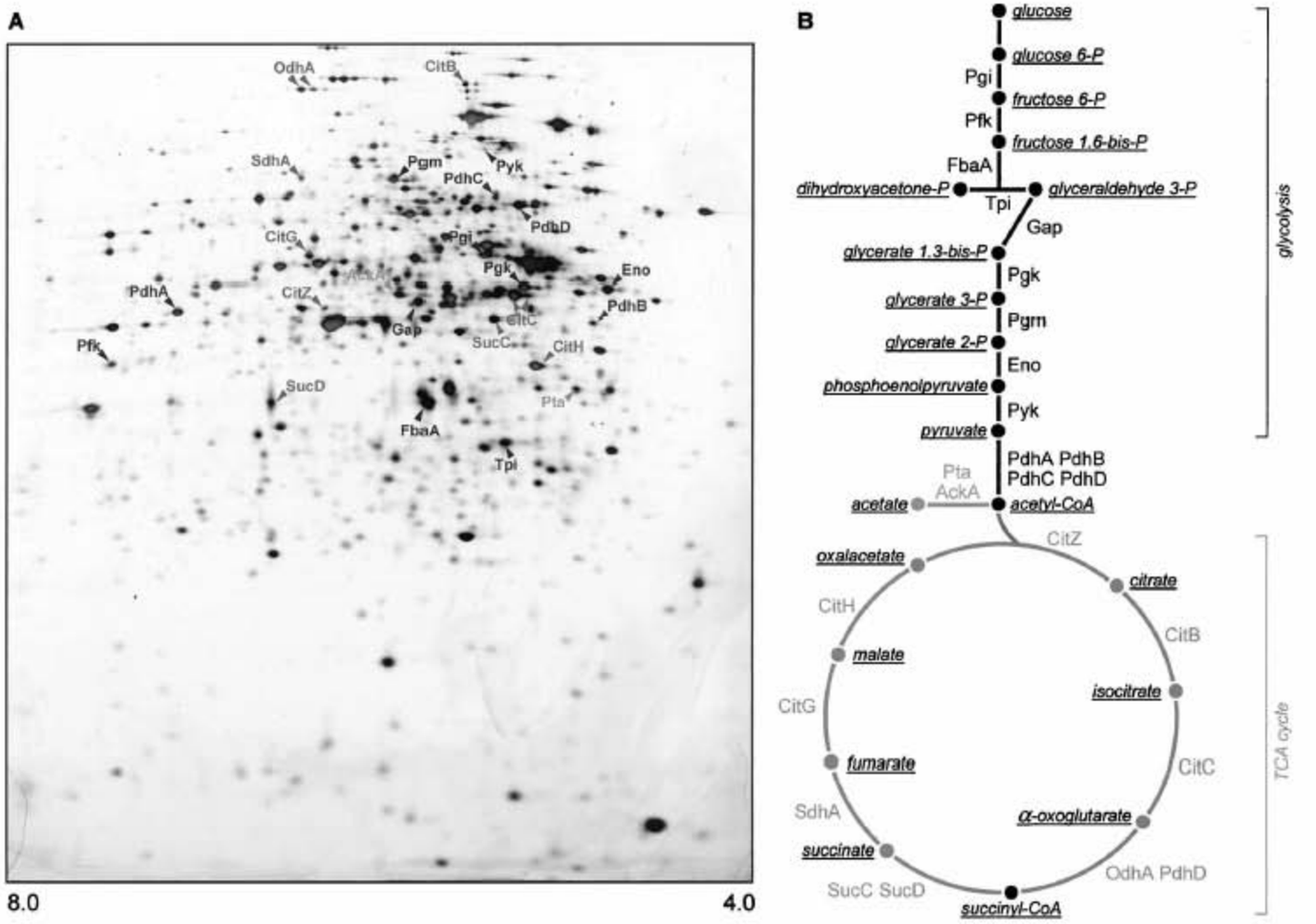
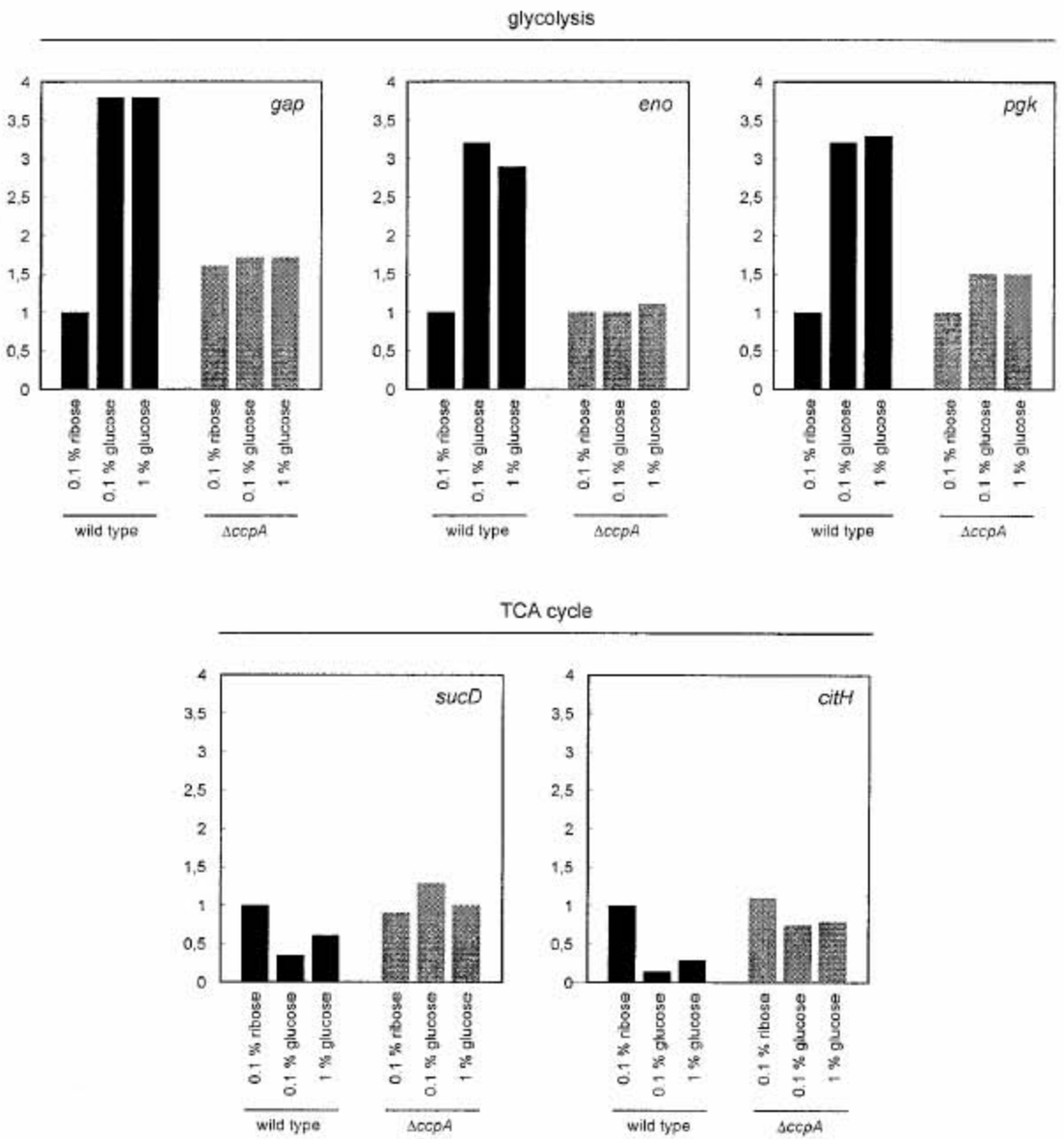


FIG. 1. Main pathways of carbohydrate metabolism. (A) Silver-stained 2D gel showing identified proteins which are involved in glycolysis (Pgi, Pfk, FbaA, Tpi, Gap, Pgk, Pgm, Eno), pyruvate dehydrogenesis (PdhA, PdhB, PdhC, PdhD), the TCA cycle (CitZ, CitB, CitC, OdhA, PdhD, SucC, SucD, SdhA, CitG, CitH), and overflow metabolism (Pta, AckA). (B) Schematic representation of glycolysis and the TCA cycle. Enzymes and metabolites are indicated. Note that not all the proteins of these pathways are identified on 2D gels.



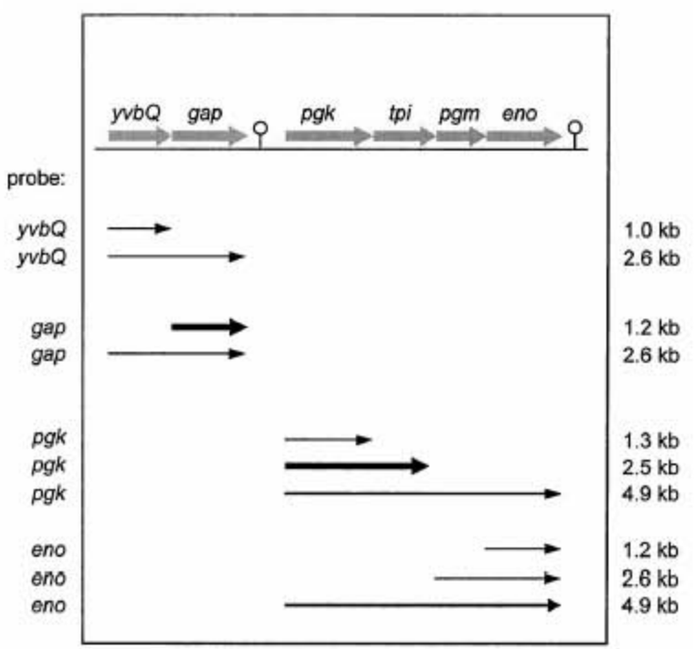


FIG. 4. Quantitative analysis of mRNAs of glycolytic and TCA cycle enzymes. Different concentrations of RNA (1 and 0.5 μg) from *B. subtilis* IS58 and BGW2 were blotted onto nylon membranes and hybridized with probes specific for the genes indicated. The mRNA amount obtained for IS58 in the presence of 0.1% ribose was set at 1.

Role of CRP in Regulation of the Central Pathways of Carbon Catabolism in *Escherichia coli.*

Even though transcriptional regulation plays a key role in establishing the metabolic network, the extent to which it actually controls the in vivo distribution of metabolic fluxes through different pathways is essentially unknown. Based on metabolism-wide quantification of intracellular fluxes, Perrenoud and Sauer (2005) systematically elucidated the relevance of global transcriptional regulation by ArcA, ArcB, Cra, Crp, Cya, Fnr, and Mlc for aerobic glucose catabolism in batch cultures of *Escherichia coli*. Knockouts of ArcB, Cra, Fnr, and Mlc were phenotypically silent, while deletion of the catabolite repression regulators Crp and Cya resulted in a pronounced slow-growth phenotype but had only a nonspecific effect on the actual flux distribution. Knockout of ArcA-dependent redox regulation, however, increased the aerobic tricarboxylic acid (TCA) cycle activity by over 60%. Like aerobic conditions, anaerobic derepression of TCA cycle enzymes in an ArcA mutant significantly increased the in vivo TCA flux when nitrate was present as an electron acceptor. The in vivo and in vitro data demonstrate that ArcA-dependent transcriptional regulation directly or indirectly controls TCA cycle flux in both aerobic and anaerobic glucose batch cultures of *E. coli*. This control goes well beyond the previously known ArcA-dependent regulation of the TCA cycle during microaerobiosis.

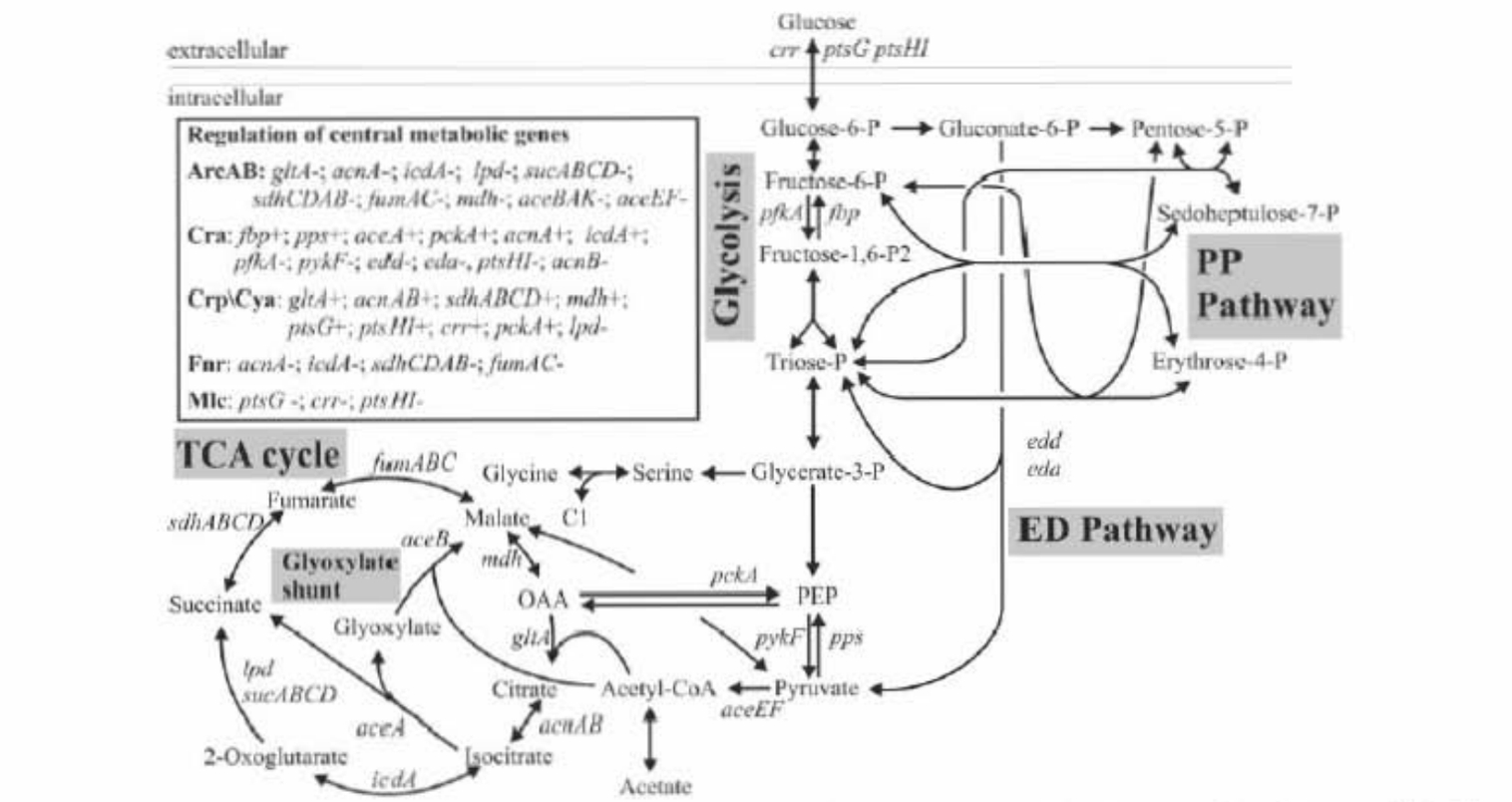


FIG. 1. Biochemical reaction network for central carbon metabolism in *E. coli*. The arrowheads indicate the assumed reaction reversibility. The inset provides an overview of central metabolic genes that are regulated by the global regulators which we investigated. Negative transcriptional regulation and positive transcriptional regulation are indicated by minus signs and plus signs that follow the gene abbreviations, respectively. Only regulated genes are indicated in the network for clarity.

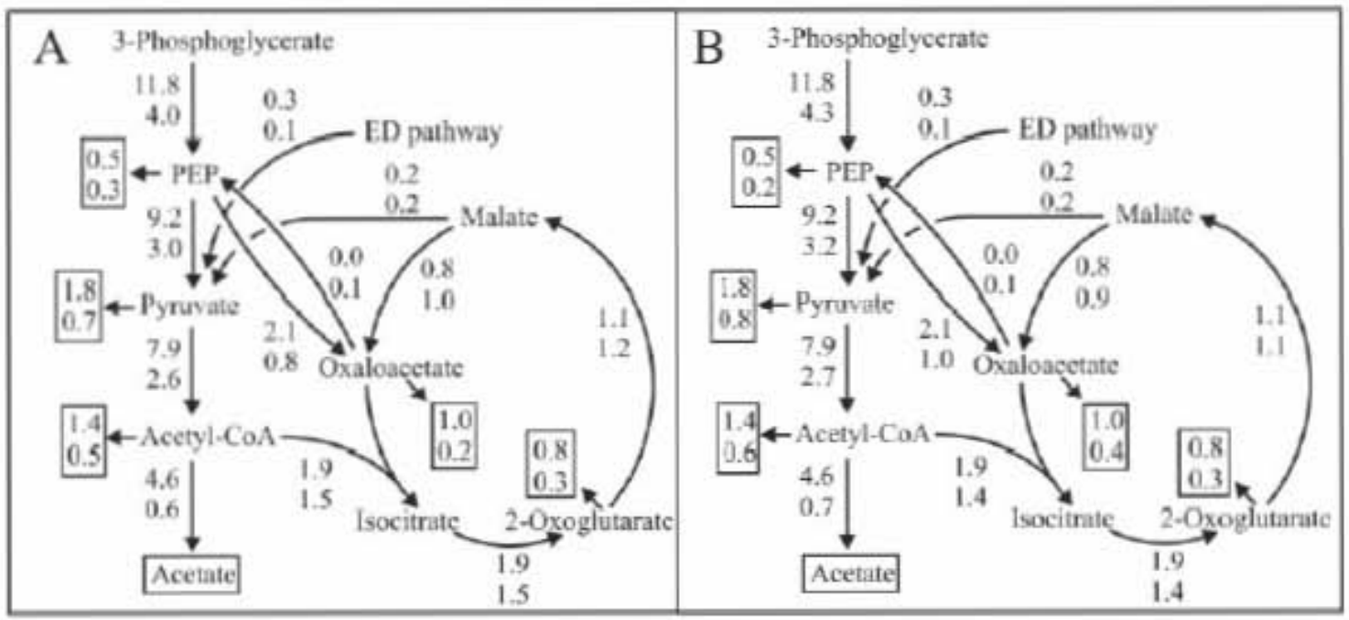


FIG. 4. Metabolic flux distribution in the Crp mutant (bottom values) ard the parent strain (top values) (A) and in the Cya mutant (bottom values) and the parent strain (top values) (B) during exponential aerobic growth on glucose. For clarity only the lower part of metabolism is shown. Net molar fluxes (expressed in millimoles per gram per hour) were determined by ^{13}C-constrained flux analysis from two separate experiments with 100% [1-^{13}C]glucose and with a mixture of 20% [U-^{13}C]glucose and 80% unlabeled glucose. In all cases the standard deviation was less than 0.2 mmol g^{-1} h^{-1}. Fluxes to biomass building blocks and extracellular products are enclosed in boxes.