

**CENTRO DE INVESTIGACIÓN Y DE ESTUDIOS
AVANZADOS DEL INSTITUTO POLITÉCNICO
NACIONAL**

Unidad Irapuato

**ESTUDIO DE LAS REDES DE REGULACIÓN
TRANSCRIPCIONAL DE BACTERIAS**

Tesis que presenta:

M. C. Edgardo Galán Vásquez

Para obtener el grado de

DOCTOR EN CIENCIAS

En la especialidad de

BIOTECNOLOGÍA DE PLANTAS

Director de tesis:

Dr. Agustino Martínez Antonio

Irapuato, Guanajuato

2016

“Every object that biology studies is a systems of systems”

François Jacob (1974)

DEDICO ESTA TESIS

A mis familiares, Demetrio Galán Martínez, Eva Vásquez Enríquez, Jocelyn Galán Vásquez, Felipa Enríquez Peralta y Fausto Galán, por todo su amor, apoyo y confianza, todo lo que hoy soy es gracias a ustedes.

A Cinthia Soberanes, por tu amor, ánimo y apoyo incondicional que me has brindado durante este maravilloso tiempo juntos.

A la familia Galán y familia Vásquez, por brindarme todo su cariño y apoyo durante estos años lejos de ustedes (Team Oaxaca).

AGRADEZCO

Al Consejo Nacional de Ciencia y Tecnología (CONACYT) por la beca otorgada durante el desarrollo de esta tesis doctoral.

Al Centro de Investigación y de Estudios Avanzados del IPN (CINVESTAV) Unidad Irapuato por brindarme la oportunidad de realizar mis estudios de doctorado.

Al Dr. Agustino Martínez Antonio, por su guía, apoyo y paciencia durante el desarrollo de este proyecto, pero sobre todo por darme la facilidad de desarrollar todas las ideas que surgieron durante el transcurso de este proyecto.

A mis sinodales la Dra. Laila Pamela Partida Martínez, el Dr. Luis Delaye Arredondo, el Dr. Cei Abreu Goodger y el Dr. Edgardo Ugalde Saldaña, por las ideas y aportes que me brindaron durante el desarrollo de todo el doctorado.

Al Dr. Ismael Sánchez Osorio, por todo la ayuda en las diferentes fases de desarrollo de este proyecto.

A la Dra. Cinthia Valentina Soberanes Gutiérrez, por las horas de discusión y aportes en este proyecto.

A mis amigos y compañeros del Cinvestav, M.C. David Israel Cruz Gómez, Dr. Julio Massange, M.C. Cynthia Paola Rangel, M.C. David Velázquez, Ana Lilia Hernández y Dora Anguiano, por la ayuda en el desarrollo de este proyecto.

ÍNDICE GENERAL

| | |
|--|-----------|
| RESUMEN | 3 |
| ABSTRACT | 4 |
| ÍNDICE DE FIGURAS | 5 |
| ÍNDICE DE TABLAS | 6 |
| 1. INTRODUCCIÓN | 7 |
| 1.1 RED DE REGULACIÓN TRANSCRIPCIONAL..... | 7 |
| 1.1.1 <i>Reconstrucción de redes de regulación transcripcional</i> | 11 |
| 1.2 ESTUDIO DE REDES DE REGULACIÓN TRANSCRIPCIONAL | 13 |
| 1.2.1 <i>Evolución de las redes de regulación transcripcional</i> | 14 |
| 1.3 BACTERIAS CON REDUCCIÓN GENÓMICA | 17 |
| 1.4 CONTRIBUCIONES DE ESTE TRABAJO | 19 |
| 2. OBJETIVOS | 21 |
| 3. MÉTODOS Y DATOS | 22 |
| 3.1 RECONSTRUCCIÓN DE LAS REDES DE REGULACIÓN TRANSCRIPCIONAL..... | 22 |
| 3.2 ANÁLISIS DE LOS FACTORES DE TRANSCRIPCIÓN CONSERVADOS | 28 |
| 3.3 ANÁLISIS DE LA INFLUENCIA DE LAS RELACIONES FILOGENÉTICAS EN LA CONSERVACIÓN DE TIPOS DE REGULADORES..... | 32 |
| 3.4 IDENTIFICACIÓN DE REGULADORES GLOBALES | 33 |
| 3.5 ANÁLISIS DE REGRESIÓN MULTIVARIABLE | 35 |
| 4. RESULTADOS Y DISCUSIÓN | 38 |
| 4.1 CONSERVACIÓN DE ACTIVADORES, REPRESORES Y DUALES | 38 |
| 4.2 GENES EN LAS REDES DE REGULACIÓN CONSERVADOS EN ENDOSIMBIONTES..... | 45 |
| 4.3 INTERPRETACIÓN BIOLÓGICA DE LA CONSERVACIÓN DE FACTORES DE TRANSCRIPCIÓN EN EL CONTEXTO DE REDUCCIÓN GENÓMICA | 51 |
| 4.4 MODELO CONCEPTUAL DE FACTORES DE TRANSCRIPCIÓN CONSERVADOS EN GENOMAS REDUCIDOS | 59 |
| 5. CONCLUSIONES | 64 |

| | |
|----------------------------------|-----------|
| 6. REFERENCIAS | 66 |
| 7. ANEXOS | 76 |
| 7.1 ARTÍCULO 1 | 76 |
| 7.2 ARTÍCULO 2 | 77 |
| 7.3 ARTÍCULO 3 | 78 |
| 7.4 ARTÍCULO 4 | 79 |
| 7.5 MATERIAL SUPLEMENTARIO | 80 |

RESUMEN

La descripción de las redes de regulación transcripcional ha sido fundamental en la comprensión de los principios operacionales que utilizan los organismos para responder y adaptarse a diversas condiciones ambientales. Si bien, el estudio de la topología y dinámica de estas redes ha sido objeto de diversos trabajos, la evolución de su topología como resultado de la adaptación de los organismos a diferentes condiciones ambientales ha recibido poca atención.

En este trabajo, se estudió la evolución de las redes de regulación en bacterias desde una perspectiva de reducción genómica, que se manifiesta como la pérdida de genes a diferentes niveles. La red de regulación transcripcional de la bacteria *Escherichia coli K-12 MG1655* se utilizó como referencia para reconstruir las redes de regulación de 113 γ -proteobacterias, los cuales fueron clasificados en cuatro grupos de acuerdo a su estilo de vida: 94 genomas de organismos de vida libre, 6 genomas de simbioses recién restringidos a huéspedes, 11 genomas de simbioses obligados a largo plazo y 2 genomas de simbioses con genomas diminutos.

El resultado fue que el tipo de regulación ejercido por cada tipo de factor transcripcional correlaciona con su grado de conservación, donde los reguladores duales son los más conservados, seguido por los represores y los activadores, los cuales se pierden más rápidamente durante la reducción genómica. Por otra parte, todo indica que la conservación preponderante de los reguladores duales puede ser debido a su papel como reguladores globales y como proteínas asociadas al nucleóide.

Se sugiere que en los diferentes niveles de reducción genómica se ejerce una forma de regulación transcripcional diferenciada, en la que la regulación transcripcional estándar sucede en los organismos de vida libre y conforme se pierden genes termina con la inexistencia de regulación mediada por factores de transcripción en los organismos con los genomas más pequeños. Finalmente, se propone cómo la reducción genómica puede afectar los nodos e interacciones en las redes de regulación transcripcional.

ABSTRACT

The description of transcriptional regulatory networks has been pivotal for understanding the operating principles under which organisms respond and adapt to varying conditions. While the study of the topology and dynamics of these networks has been the subject of considerable work, the evolution of their topology, as a result of adaptation of organisms to different environmental conditions, has received little attention.

This work studies the evolution of transcription factor networks in bacteria from the perspective of genome reduction, which manifests itself as the loss of genes to different degrees. The transcriptional regulatory network of *Escherichia coli* was used as a reference to compare smaller genomes of 113 γ -proteobacteria, which they were classified into four groups according to their lifestyle: 94 free-living organism genomes, 6 recently host-restricted symbionts genomes, 11 long-term obligate symbionts genome and 2 tiny-genome symbionts.

The type of regulatory action exerted by transcription factors in bacteria was found to correlate well with their degree of conservation during genome reduction, dual regulators are more conserved than repressors and activators in conditions of extreme reduction. Moreover, the preponderant conservation of dual regulators might be due to their concomitant role as both global regulators and nucleoid-associated proteins. Finally, these results were integrated into a conceptual model that extends previous studies on the evolution of transcriptional regulatory networks.

ÍNDICE DE FIGURAS

| | |
|--|----|
| FIGURA 1. ESTRATEGIA DE RECONSTRUCCIÓN DE INTERACCIONES ENTRE PARES DE GENES EN LAS REDES DE REGULACIÓN. | 24 |
| FIGURA 2. PRUEBA DE HIPÓTESIS PARA CORROBORAR LA SIGNIFICANCIA DE CONSERVACIÓN DE LOS FACTORES DE TRANSCRIPCIÓN EN OTROS GENOMAS. | 30 |
| FIGURA 3. FILOGENIA DE LAS <i>γ</i> -PROTEOBACTERIAS USADAS EN ESTE TRABAJO..... | 35 |
| FIGURA 4. IDENTIFICACIÓN DE REGULADORES GLOBALES | 36 |
| FIGURA 5. EJEMPLO DE REDES DE REGULACIÓN TRANSCRIPCIONAL | 38 |
| FIGURA 6. DISTRIBUCIÓN DE LOS TAMAÑOS DE LOS GENOMAS UTILIZADOS EN ESTE ESTUDIO DE EVOLUCIÓN DE REDES..... | 39 |
| FIGURA 7. CONTRASTES FILOGENÉTICOS DE FACTORES DE TRANSCRIPCIÓN CONSERVADOS EN LOS 113 GENOMAS..... | 41 |
| FIGURA 8. PROBABILIDADES CALCULADAS POR LA DISTRIBUCIÓN HIPERGEOMÉTRICA PARA CADA UNO DE LOS GENOMAS..... | 42 |
| FIGURA 9. CONTRASTES FILOGENÉTICOS DE LAS FRECUENCIAS RELATIVAS DE LA CONSERVACIÓN DE LOS FACTORES DE TRANSCRIPCIÓN | 44 |
| FIGURA 10. CLASIFICACIÓN DE LOS GENES REGULADOS POR ONTOLOGÍA DE PROCESO BIOLÓGICOS PRESENTES EN LAS REDES DE REGULACIÓN TRANSCRIPCIONAL RECONSTRUIDAS | 48 |
| FIGURA 11. PROBABILIDADES CALCULADAS DESDE LA DISTRIBUCIÓN HIPERGEOMÉTRICA PARA REGULADORES GLOBALES Y NAPs | 50 |
| FIGURA 12. MODELO DE EVOLUCIÓN DE LAS REDES DE REGULACIÓN TRANSCRIPCIONAL BAJO LA INFLUENCIA DE REDUCCIÓN GENÓMICA..... | 61 |
| FIGURA 13. EFECTO DE LA REDUCCIÓN GENÓMICA EN LAS REDES DE REGULACIÓN TRANSCRIPCIONAL. | 63 |

ÍNDICE DE TABLAS

| | |
|---|----|
| TABLA 1. GENOMAS DE BACTERIAS UTILIZADOS EN ESTE TRABAJO, CLASIFICADOS DE ACUERDO A SU ESTILO DE VIDA. | 25 |
| TABLA 2. LOS REGULADORES GLOBALES DE <i>E. COLI</i> | 49 |
| TABLA 3. COEFICIENTE DE REGRESIÓN PARCIAL ESTANDARIZADAS | 51 |
| TABLA 4. LOS 10 ACTIVADORES MÁS CONSERVADOS EN LAS REDES DE REGULACIÓN TRANSCRIPCIONAL DE LAS γ -PROTEOBACTERIAS..... | 52 |
| TABLA 5. LOS 10 REPRESORES MÁS CONSERVADOS EN LAS REDES DE REGULACIÓN TRANSCRIPCIONAL DE LAS γ -PROTEOBACTERIAS | 54 |
| TABLA 6. LOS 10 REGULADORES DUALES MÁS CONSERVADOS EN LAS REDES DE REGULACIÓN TRANSCRIPCIONAL DE LAS γ -PROTEOBACTERIAS..... | 56 |

1. INTRODUCCIÓN

En esta sección se establecen los conceptos básicos de las redes de regulación transcripcional y las principales perspectivas de estudio de esta área. En la sección 1.1 se describen las funciones de la red de regulación y los métodos para reconstruir redes. En la sección 1.2 se abordan las diferentes aproximaciones en el estudio de las redes de regulación, centrándose en los estudios de evolución. En la sección 1.3 se proporciona una descripción de los organismos endosimbiontes, sus características e importancia. Finalmente, en la sección 1.4 se describen las aportaciones de este trabajo.

1.1 RED DE REGULACIÓN TRANSCRIPCIONAL

La transcripción es el proceso por el cual los genes son decodificados del ADN al ARN (mensajero, ribosomal, transferencia y pequeños). Este proceso es altamente conservado en los tres dominios de la vida, desde bacterias hasta eucariontes (Ptashne M. 2005). Y es realizado por la enzima ARN polimerasa, la cual en el caso de bacterias está compuesta de múltiples subunidades (α , β , β' , ω , σ). Las bacterias y arqueas sólo tienen una ARN polimerasa, mientras que los eucariontes cuentan con tres ARN polimerasas (I, II y III), miembros de una familia de proteínas de múltiples subunidades altamente conservadas. El proceso de transcripción en sí, es llevado a cabo por el denominado núcleo enzimático (α , β , β' , ω), que es capaz de realizar la unión de los ribonucleótidos (Ebright RH. 2000). Sin embargo, para reconocer el promotor e iniciar la transcripción en sitios específicos, es necesario que la ARN polimerasa se una a una subunidad sigma (hay 7 en *Escherichia coli*) y que se forme la holoenzima (Murakami KS, *et al.* 2002).

La subunidad sigma puede ser considerada como el principal regulador en *trans* del inicio de la transcripción, las proteínas de la familia de sigma⁷⁰ (6 en *E. coli*) tienen tres funciones principales: 1) asegurar el reconocimiento de la secuencia

del promotor al identificar las cajas -10 (TTGACA), -10 extendido (TGn) y -35 (TATAAT) (Browning DF y Busby SJW. 2004); 2) posicionar la holoenzima de la ARN polimerasa en la zona del promotor; y 3) facilitar el desenrollamiento del dúplex de ADN, necesario para iniciar la transcripción (Wösten MMSM. 1998).

Además de los factores sigma, existen otras proteínas que regulan el inicio de la transcripción. Éstas son denominadas factores de transcripción y fueron descritas por François Jacob y Jacques Monod como moléculas de dos cabezas, de las cuales una corresponde al dominio de unión al ADN y la otra es un dominio alostérico al que se une un metabolito o señal efectora (Martínez-Antonio A y Collado-Vides J. 2003). Los factores de transcripción regulan la eficiencia de unión de la ARN polimerasa al promotor, por lo tanto pueden promover o inhibir esta interacción. Los factores de transcripción pueden ser clasificados de acuerdo a su actividad reguladora en: activadores (cuando promueven la unión de la ARN polimerasa al promotor), represores (cuando inhiben la unión de la ARN polimerasa a la zona promotora) o duales (cuando en una condición o fase de desarrollo, activan la transcripción y en otra condición la inhiben) (Janga SC, *et al.* 2009).

Los activadores interactúan de tres formas con el promotor: los denominados de clase I, se unen en la región río arriba del elemento -35 y reclutan a la ARN polimerasa por medio de la interacción con el carboxilo terminal de la subunidad α . La unión del activador al carboxilo y amino terminal de la subunidad α son flexibles, por lo cual los activadores que usan este tipo de regulación pueden unirse a varios sitios río arriba del promotor. Los de clase II, se unen al promotor sobrelapando el elemento -35, lo que provoca el reclutamiento de la ARN polimerasa por medio de la interacción con el dominio cuatro de la subunidad sigma. En la clase III, los activadores se unen entre los elementos -10 y -35, lo que provoca un cambio de conformación en el ADN, esto facilita el reconocimiento de los elementos reguladores por la ARN polimerasa (Browning DF y Busby SJW. 2004, Lee DJ, *et al.* 2012).

De la misma manera que los activadores, los represores pueden interactuar de tres formas con el promotor: los de clase I, realizan un impedimento estérico, debido a que el represor se une al ADN cerca de los elementos -10 y -35 del promotor impidiendo que éstos sean reconocidos por el factor sigma de la ARN polimerasa. En la clase II, múltiples moléculas del represor se unen en regiones distantes del promotor y la represión es dada por un doblamiento del ADN que impide la unión de la ARN polimerasa. Finalmente en la clase III, el represor actúa como un anti-activador e impide que el activador reclute a la ARN polimerasa, bloqueando el reconocimiento del promotor (Rojo F. 2001, Browning DF y Busby SJW. 2004).

Para el caso de los reguladores duales, estos pueden funcionar de cualquiera de las tres maneras de activación y represión, algunos ejemplos en *E. coli* son: CRP “cAMP receptor protein” también llamado proteína activadora de catabolitos (Zheng D, *et al.* 2004), puede activar la transcripción por interacción con la región río arriba del elemento -35 del promotor del operón *lac* y reprimir la transcripción por impedimento estérico en el operón *gal* (Kolb A, *et al.* 1993). FNR “Fumarate and Nitrate Reduction” que regula la transición del crecimiento aeróbico a anaeróbico (Kang Y, *et al.* 2005), activa la transcripción interactuando con el carboxilo terminal de la subunidad α o sobrelapando el elemento -35 del promotor y reprime la transcripción por la interacción con regiones río arriba del elemento -35 como en el operón *ndh* (Wing HJ, *et al.* 1995). Otros factores de transcripción que pueden actuar como reguladores duales son: ArcA “Anoxic Redox Control” el cual reprime genes involucrados en metabolismo respiratorio y activa genes involucrados en el metabolismo fermentativo (Gunsalus RP y Park SJ. 1994). Fur “Ferric Uptake Regulation” regula la transcripción de genes involucrados en la homeostasis del hierro (Hantke K. 2001). Fis “Factor for Inversion Stimulation” (Browning DF, *et al.* 2010). IHF “Integration Host Factor” (Dillon SC y Dorman CJ. 2010). H-NS “Histone-like Nucleoid Structuring protein” (Weng X y Xiao J. 2014). NarL “Nitrate/nitrite response regulator” regula genes relacionados al transporte de

electrones y fermentación en respuesta a altas concentraciones de nitrito (Under G y Bongaerts J. 1997). Lrp “Leucine-responsive regulatory protein” regula genes involucrados en biosíntesis de aminoácidos y catabolismo (Ernsting BR, *et al.* 1992). FlhDC es el regulador principal de la biosíntesis del flagelo (Stafford GP, *et al.* 2005).

A finales de la década de los 90s se describieron proteínas que estaban asociadas principalmente al nucleóide y se les denominó “Nucleoid Associated Proteins” o NAPs, por su abreviación en inglés. Estas proteínas pueden afectar la estructura local y global del nucleóide por la inducción de curvaturas, uniones, enrollamiento y agrupación del ADN (McLeon SM y Johnson RC. 2001, Browning DF, *et al.* 2010, Dilllon SC y Dorman CJ. 2010, Weng X y Xiao J. 2014). Y pueden regular la expresión de una gran variedad de genes de manera directa o indirecta y muchas veces en coordinación con reguladores específicos. En *E. coli* se han descrito cinco principales NAPs: HU, IHF, H-NS, Fis y StpA.

HU interactúa con el ADN de manera no específica, pero tiene preferencia a unirse a regiones distorsionadas, tales como curvas o cruces de cuatro hebras del ADN. De tal forma que HU funciona en la gestión de la recombinación y la topología del ADN. Adicionalmente, HU induce la formación de bucles en el ADN que facilitan su flexibilidad, que es importante para la regulación de genes y la conservación de la arquitectura del cromosoma (Dilllon SC y Dorman CJ. 2010).

IHF se une a secuencias conservadas del ADN induciendo vueltas en forma de “U”, centradas en el sitio de unión. Es capaz de reclutar al factor sigma⁵⁴ para que se una la ARN polimerasa en algunos promotores. La unión con el ADN puede influenciar la transcripción al facilitar el contacto entre las proteínas reguladoras y la ARN polimerasa. IHF también afecta el inicio de la replicación cromosomal, por medio de la interacción con el origen de replicación (*oriC*), de la misma forma que HU y Fis (Dilllon SC y Dorman CJ. 2010, McLeon SM y Johnson RC. 2001).

Fis: contribuye a muchos procesos como transcripción del ADN, replicación y recombinación. Además, reprime el inicio de la transcripción principalmente por impedimento estérico y la activa por interacción directa con la ARN polimerasa. Sin embargo, también puede unirse a sitios distantes del promotor provocando cambios conformacionales del ADN (Browning DF, *et al.* 2010).

H-NS y su parólogo StpA pueden restringir el superenrollamiento del ADN. Por su parte, H-NS es capaz de reprimir la transcripción en respuesta a señales del ambiente (como temperatura) o por unión cooperativa con el ADN superenrollado y además tiene la habilidad de formar puentes de ADN/H-NS/ADN. Por otro lado, StpA es mejor caracterizada por ser una proteína de unión a ARN o a chaperonas de ARN (Dillon SC y Dorman CJ. 2010, Weng X y Xiao J. 2014).

1.1.1 RECONSTRUCCIÓN DE REDES DE REGULACIÓN TRANSCRIPCIONAL

El conjunto de interacciones entre los factores de transcripción y los genes regulados pueden conceptualizarse para su estudio en forma de una red. En esta red de regulación transcripcional, los factores de transcripción y los genes regulados son representados por los nodos y las interacciones físicas que existen entre los elementos son representadas por aristas que conectan a cada par de nodos. Para computar y estudiar una red de manera formal, ésta se representa en forma de un grafo ($G=(V,E)$), el cual es un objeto matemático consistente de un conjunto de nodos (V) y un conjunto de aristas (E) (Junker HB y Schreiber F. 2008).

Idealmente las redes de regulación se construyen de manera experimental, definiendo cada interacción entre un par de nodos por medio de métodos como: *Electrophoretic Mobility Shift Assay* (EMSA) por medio de la unión de proteínas purificadas, mutación dirigida de sitios de unión a factores de transcripción, *Chromatin-Immunoprecipitation-Chip* analizado en *chip* o microarreglo (*ChIP-chip*), *chromatin-immunoprecipitation-sequencing* analizado por secuenciación masiva

(*ChIP-seq*), *DNA adenine methyl transferase identification* (DamID) o *protein binding universal DNA microarrays* (PBMs). Estos métodos permiten identificar las interacciones entre los factores de transcripción y el promotor de los genes regulados. Esta información existe en mayor medida para organismos modelos en los cuales se han reconstruido las redes de regulación transcripcional. Sin embargo, los organismos menos estudiados no cuentan con la información necesaria para reconstruir estas redes de regulación de forma experimental.

Por lo anterior, se han desarrollado tres aproximaciones para inferir las redes de regulación en organismos menos estudiados experimentalmente. El primer método es basado en una plantilla o referencia, considerando el principio de que los factores de transcripción ortólogos regulan la expresión de genes ortólogos. Este método inicia con una red conocida y se transfiere la información de las interacciones de los genes identificados como ortólogos, entre el organismo de interés y el organismo utilizado como referencia (Yu H, *et al.* 2004, Babu MM, *et al.* 2009). La segunda aproximación, denominada de ingeniería reversa, usa los niveles de expresión de los genes a través de diferentes condiciones, con base en esto se identifican conjuntos de genes con perfiles de expresión similares que pueden ser co-regulados por el mismo conjunto de factores de transcripción (Gardner TS, *et al.* 2003, Qian J, *et al.* 2003). Finalmente, el tercer método se basa en la predicción de elementos *cis*-reguladores. En esta aproximación se hace uso de los sitios de unión de los factores de transcripción identificados experimentalmente para hacer inferencia sobre nuevas interacciones reguladoras. De tal forma que, las regiones promotoras del genoma de interés son escaneadas con las secuencias consenso de los factores de transcripción conocidos, para identificar posibles sitios de unión y se infiere que los genes río abajo son regulados por este factor de transcripción (Wang T y Stormo GD. 2005).

Sin embargo, dependiendo del tipo de reconstrucción utilizada, éste repercute en su nivel de resolución. En el caso de la reconstrucción basada en una platilla, mientras más lejanos son los organismos filogenéticamente, se espera que la

reconstrucción sea menos eficiente. Por otro lado, la reconstrucción con datos de expresión se ve comprometida por efectos de regulación indirectos como la existencia de co-expresión pero la inexistencia de co-regulación de ciertos genes, además de los efectos de regulación post-transcripcionales. Finalmente, la reconstrucción por medio de predicción de elementos *cis*, conlleva a un alto número de falsos positivos, debido a que los sitios de unión son fragmentos de ADN pequeños (~10 pb). Para minimizar estas desventajas se ha propuesto una metodología que une las tres aproximaciones anteriores, permitiendo obtener una red de regulación más exacta (Imam S, *et al.* 2015).

1.2 ESTUDIO DE REDES DE REGULACIÓN TRANSCRIPCIONAL

Las redes de regulación transcripcional se han estudiado desde tres principales perspectivas: La primera analiza la estructura de la red, en la cual se ha identificado que las redes de regulación adoptan una arquitectura jerárquica y muestran una topología de escala libre, la cual se caracteriza por la presencia de pocos nodos altamente conectados (llamados *hubs* o globales) y muchos nodos poco conectados. Además muestran una conservación de subestructuras denominadas motivos o módulos, lo que contribuye a la robustez inherente de esta topología (Blais A y Dynlacht BD. 2005).

La segunda aproximación analiza la dinámica o el cambio de expresión de los genes a través del tiempo. Existen dos formas básicas para modelar las redes de regulación: el primero es llamado modelo lógico, que describe cuantitativamente la red de regulación y permite la comprensión básica de la red bajo diferentes condiciones; el segundo modelo es llamado continuo y permite comprender y manipular el comportamiento de la red a través del tiempo, con escalas de tiempo más precisas y a concentraciones moleculares exactas (Ivanov I y Dougherty ER. 2005, Karlebach G y Shamir R. 2008).

Finalmente, la tercera aproximación analiza los cambios que sufre una red a nivel de ganancia y pérdida de nodos e interacciones, además de las reconexiones que pudieran existir en la red. En la siguiente sección se explicará más sobre este tema.

1.2.1 EVOLUCIÓN DE LAS REDES DE REGULACIÓN TRANSCRIPCIONAL

Para entender cómo es que las redes de regulación transcripcional evolucionan, es necesario comprender los principios básicos que delinear la evolución de los factores de transcripción y sus interacciones con los genes regulados individualmente. La disponibilidad de genomas, aunado al desarrollo de experimentos a gran escala (*high-throughput*), han facilitado el estudio de la historia evolutiva de las redes de regulación transcripcional. De tal forma que se ha propuesto que las redes de regulación deben seguir dos principales perspectivas de evolución: la primera que considera la ganancia de nodos e interacciones, y la segunda que considera la pérdida de nodos e interacciones (Babu MM. 2010).

La evolución de las redes de regulación es influenciada por las mutaciones que sufre un genoma. Mutaciones como sustitución de un simple nucleótido pueden afectar la función de una o pocas bases (Wittkopp P y Kalay G. 2011), mientras otras mutaciones como duplicación (McAdams HH, *et al.* 2004), expansión por transposones o transferencia horizontal de genes, pueden generar un gran segmento de material genético (Teichmann SA y Babu MM. 2004). Estos eventos pueden afectar las interacciones reguladoras, a nivel de los elementos *cis*, introduciendo mutaciones puntuales en las regiones río arriba de los genes regulados, afectando los sitios de unión de factores de transcripción. A nivel de los factores que actúan en *trans*, se pueden introducir cambios en secuencias pre-existentes y generar nuevos dominios de unión a ADN o genes completos (Babu MM. 2010).

Estos cambios a nivel de la secuencia de ADN generan la ganancia o pérdida de interacciones y nodos en la red. En la perspectiva de crecimiento de las redes reguladoras se ha observado que las dos principales fuerzas que dirigen esta ganancia de genes son la duplicación y la transferencia horizontal de genes. En la duplicación de genes, se genera una nueva copia de un gen regulado, de un factor de transcripción o de ambos. Después de la duplicación ocurre un estado de redundancia funcional dado por ambas copias, esto puede dar lugar a una relajación de la presión de selección, permitiendo que ambas copias acumulen un gran número de mutaciones (Zhang J. 2003). Al final, estos genes duplicados tienen tres posibles destinos: I) pseudo-funcionalización, donde una de las copias se pierde, II) sub-funcionalización, donde ambos genes son mantenidos realizando parte de las funciones originales, ya que los dos en conjunto realizan la función ancestral, y III) neo-funcionalización, donde una de las copias realiza la función ancestral y la segunda copia adquiere una nueva función.

En las redes de *E. coli* y de *Saccharomyces cerevisiae* se identificó que más de dos terceras partes de las interacciones reguladoras han evolucionado como consecuencia de la duplicación de genes. Además, más de la mitad de las interacciones en las redes fueron heredadas desde duplicaciones ancestrales de factores de transcripción y de genes regulados. Así mismo, se determinó que solo una fracción de interacciones reguladoras evoluciona por recombinación genética o innovación (Teichmann SA y Babu MM. 2004). A nivel topológico de la red de regulación se encontró que los genes que forman parte de un motivo estructural de la red pueden surgir como consecuencia de duplicación, pero las interacciones que forman los motivos tienden a ser ganados o se han adquirido como consecuencia de la reconexión (Conant GC y Wagner A. 2003). A nivel de la estructura global de la red, la topología de escala libre no parece ser una consecuencia directa de la duplicación de genes (Teichmann SA y Babu MM. 2004). Estas observaciones son consistentes con la posibilidad de que la estructura de escala libre puede evolucionar por selección natural (Lynch M. 2007).

Por su parte, la adición de nodos por medio de la transferencia horizontal de genes, requiere la incorporación de un fragmento de ADN desde otros organismos, usualmente otra bacteria. Esto puede ocurrir de tres maneras: la primera denominada conjugación de ADN, que es dada por el contacto entre bacterias a través de los *pili* formados por una de ellas; la segunda denominada transducción, que es realizada por un vector viral tal como un bacteriófago; la tercera es llamada transformación celular, que utiliza la habilidad de algunas bacterias para tomar ADN del medio ambiente. Después de que un gen es incorporado dentro de una bacteria, si es transcripcionalmente activo y no es letal, éste puede ser adecuadamente regulado por la red de regulación existente. La probabilidad de integrar correctamente un gen transferido en una red de regulación existente, se espera que disminuya a medida que aumenta la distancia filogenética entre el organismo de origen del ADN y el organismo receptor (Koonin EV, *et al.* 2001).

En este contexto, se propuso que los genes transferidos y retenidos contribuyen a un mejor desempeño del organismo receptor y que principalmente se adicionan en las partes externas de la red de regulación y preferentemente genes regulados en lugar de factores de transcripción (Lagomarsino MC, *et al.* 2007). Modelos computacionales que simulan la adición de un nodo o conexión en una red teórica, han mostrado que la red incrementa su evolvabilidad (capacidad adaptativa), lo cual le permite tanto ganar nuevas funciones como responder a nuevas condiciones del ambiente (Aldana M. *et al.* 2007, Isalan M. *et al.* 2008).

Por otro lado, la perspectiva de reducción genómica se ha estudiado por medio de la reconstrucción de redes completas, principalmente por la búsqueda de homólogos. Se ha identificado que los genes reguladores son los que cambian más rápido en comparación con los genes regulados (Babu MM, *et al.* 2006, Lozada-Chávez I, *et al.* 2006, Price NM, *et al.* 2007), sugiriendo que esto permite a los organismos evolucionar su conjunto de reguladores de una manera más eficiente para responder a las condiciones ambientales. Se identificó que a nivel topológico global, los organismos con estilos de vida similares conservan

interacciones y motivos similares. La conservación de los factores de transcripción es independiente de su conectividad, mientras que el medio ambiente parece ser la principal fuerza que dirige la ganancia y pérdida de factores de transcripción e interacciones reguladoras. Sin embargo, estos trabajos no consideran el efecto de la pérdida de fragmentos de ADN y específicamente tampoco la pérdida de nodos e interacciones en redes de regulación reales.

1.3 BACTERIAS CON REDUCCIÓN GENÓMICA

La pérdida de nodos dentro de una red de regulación puede ser estimada a través del análisis de organismos que sufren el proceso de reducción genómica, en el cual los genomas reducen su tamaño con respecto a su ancestro. Este tipo de reducción de tamaño es más significativo en bacterias, como el caso de simbioses en los cuales la reducción genómica es extensa.

El término simbiosis se refiere a la relación ecológica cercana entre dos (o más) especies, capaz de reportar beneficios a todos (mutualismo) o algunos de los organismos implicados, con o sin causar daño a una de las especies involucradas (parasitismo o comensalismo, respectivamente) (Reyes-Prieto M, *et al.* 2014).

Una alta proporción de las relaciones simbióticas estudiadas a la fecha involucran un organismo eucarionte y bacterias. Los eucariontes proporcionan un medio rico de nutrientes a la comunidad bacteriana, un nicho estable y protección. Por otro lado, las bacterias proporcionan metabolitos esenciales (principalmente aminoácidos), los cuales no pueden ser sintetizados por los organismos eucariontes, ni tampoco obtenidos por medio de su alimentación. Cuando esta relación es obligada, es decir, que un organismo no puede sobrevivir fuera del otro se llama “simbiosis obligada”. Cuando en esta relación simbiótica una de las partes, generalmente un simbionte procarionte, vive dentro de una célula eucariótica se llama endosimbiosis. En el caso de insectos, estos endosimbiontes viven dentro de células especializadas llamadas bacteriocitos, los cuales pueden

formar órganos especializados (bacteriomas), que son localizados dentro del abdomen del insecto (Moran NA y Bennett GM. 2014).

En este estilo de vida restringido dentro de un huésped, la transmisión de los endosimbiontes se da de manera materna de generación en generación, y se espera que sea inexistente el intercambio genético por transferencia horizontal desde otras bacterias. Además, se piensa que existe un decremento en la selección purificadora combinada con un incremento de la deriva genética, lo que lleva a la acumulación de mutaciones levemente deletéreas en un proceso conocido como trinquete de Muller, que conduce a una eventual inactivación y pérdida de genes no esenciales. Todos estos factores llevan a las bacterias endosimbiontes a tener genomas muy reducidos (Delaye L, *et al.* 2010, McCutcheon JP y Moran NA. 2012).

Existen tres principales hipótesis de cómo puede darse el proceso de reducción genómica: la primera es denominada hipótesis de “simplificación” (*streamlining* en inglés), en la cual los genomas pequeños son favorecidos por la selección natural como un camino hacia la economía celular. Esto ocurre principalmente en organismos que viven bajo deficiencia de nutrientes. De acuerdo a esta hipótesis, la selección natural conduce también a un bajo contenido de las bases nucleotídicas guanina (G) y citosina (C), lo que permitiría requerir menor cantidad de fósforo y nitrógeno para la síntesis de ADN. Además, las bacterias son mas pequeñas de tal manera que pudieran requerir menores cantidades de ADN (Mira A, *et al.* 2001). La segunda hipótesis es denominada “hipótesis del incremento de la tasa mutacional”, la cual sugiere que la reducción genómica es un subproducto del aumento de la tasa de mutaciones, que puede ser ventajoso en la colonización de nuevos nichos (Dufresne A, *et al.* 2005, Marais GAB, *et al.* 2008). La tercera hipótesis denominada “reina negra”, propone que la pérdida de genes y la reducción genómica son eventos adaptativos dependientes de la comunidad. De tal forma que se necesitan tres condiciones para que esto ocurra, i) la existencia de un bien público, el cual actúa como moneda de cambio, ii) un organismo que

sea capaz de producir el bien público y, iii) un beneficiario, que es un organismo que utiliza el bien público (Morris JJ, *et al.* 2012, Martínez-Cano DJ, *et al.* 2015).

La regulación transcripcional en los endosimbiontes disminuye durante el proceso de reducción genómica, posiblemente debido a que los genes reguladores no son esenciales bajo las condiciones de simbiosis. Sin embargo, se ha observado que los factores de transcripción conservados en bacterias endosimbiontes siguen siendo funcionales, como es el caso de los genes de biosíntesis de metionina en *Buchnera aphidicola SP*, donde se determinó que existe una respuesta transcripcional de los genes regulados por MetR (Moran NA, *et al.* 2005). Sin embargo, al ser modelos de estudio poco estudiados, no se tiene una caracterización completa de los procesos reguladores conservados en estos organismos endosimbiontes.

1.4 CONTRIBUCIONES DE ESTE TRABAJO

El estudio de la regulación transcripcional desde una perspectiva de redes de regulación, cobró gran relevancia en los últimos años debido a la posibilidad de integrar y estudiar los datos desde una perspectiva de genomas completos. Esto ha permitido comprender el fenómeno más a fondo y vislumbrar las implicaciones que tienen las interacciones reguladoras en la respuesta integral de un organismo ante las condiciones a las que está expuesto.

Una pregunta que ha surgido en el desarrollo de los análisis de redes, ha sido ¿cómo han evolucionado las redes de regulación?. En este trabajo se plantea una aproximación analizando organismos que sufren pérdida masiva de genes y su comparación al organismo modelo *E. coli*, con el fin de identificar si el fenómeno de reducción genómica puede actuar como una fuerza que dirija la evolución de las redes de regulación desde una perspectiva de reducción de redes.

Se utilizó la red de regulación transcripcional de *E. coli* *k-12* *MG1655* (Bacteria, Proteobacteria, Gammaproteobacteria, Enterobacteriales, Enterobacteriaceae, *Escherichia*, *E. coli*) como una plantilla para reconstruir las redes de regulación de 113 genomas de gamma-proteobacterias. Se identificaron los elementos conservados en las redes de regulación reconstruidas empleando una aproximación de genómica comparativa. Además, se utilizó una combinación de análisis de correlación con contrastes independientes filogenéticamente y se observó que en los organismos que presentan reducción genómica existe una pérdida gradual de factores de transcripción, pero existe una conservación preferencial de reguladores duales, sobre los represores y activadores.

Se identificó también que la conservación de reguladores duales se debe principalmente a su naturaleza global y a la función estructural que ejercen sobre el nucleóide. Todos estos resultados fueron integrados dentro de un modelo conceptual sobre la evolución de redes de regulación desde una perspectiva de reducción genómica.

2. OBJETIVOS

General

Estudiar la evolución de las redes de regulación transcripcional en bacterias que experimentan reducción genómica.

Específicos

- 1.- Reconstruir las redes de regulación transcripcional de las γ -proteobacterias.
- 2.- Analizar las redes de regulación transcripcional por medio de contrastes filogenéticos independientes.
- 3.- Analizar las contribuciones a la conservación diferencial de los reguladores debido a su carácter de regulación dual y por su papel como proteínas asociadas al nucleoide.
- 4.- Proponer un modelo de la evolución de las redes de regulación transcripcional en bacterias, desde una perspectiva de reducción genómica.

3. MÉTODOS Y DATOS

3.1 RECONSTRUCCIÓN DE LAS REDES DE REGULACIÓN TRANSCRIPCIONAL

Se seleccionó un conjunto de 113 genomas de las γ -proteobacterias, con el objetivo de cubrir un amplio rango de tamaños de genomas desde 0.159 Mbp (correspondiente a *Candidatus Carsonella Ruddii PV*) hasta 4.639 Mbp (correspondiente a *E. coli K-12 MG1655*). La selección de genomas analizados se realizó minimizando la redundancia de genomas de la misma especie, además de intentar minimizar el rango de tamaños entre cada par de genomas que no excedieran los 400 Kbp. Las secuencias de las proteínas para cada uno de los genomas analizados fueron descargadas de la base de datos del NCBI (<ftp://ftp.ncbi.nlm.nih.gov/genomes/genbank/bacteria/>).

Las redes de regulación transcripcional de los 113 genomas fueron reconstruidas siguiendo la aproximación para la obtención de *Regulogs* (Yu H, *et al.* 2004). Esta técnica de genómica comparativa tiene como objetivo encontrar interacciones reguladoras en organismos que carecen de información experimental. Para reconstruir las redes de regulación, se utilizó como base la información existente de una red de regulación transcripcional conocida experimentalmente. Los *Regulogs* infieren las interacciones reguladoras basadas en la suposición de que los factores de transcripción ortólogos generalmente también regulan la transcripción de genes ortólogos. De tal forma que, si en el organismo de referencia se tiene un factor de transcripción (A) que interactúa con un gen regulado (B) y éstos tienen ortólogos A' y B' en el organismo de inferencia, entonces la interacción reguladora se considera conservada en el genoma de interés (Figura 1) (Yu H, *et al.* 2004). En este estudio se utilizó la red de regulación transcripcional de *E. coli* como plantilla, misma que fue obtenida de la base de datos RegulonDB (Salgado H, *et al.* 2013). La red de regulación conocida de *E.*

coli está compuesta por 1,784 nodos y 4,058 interacciones, como resultado de la interacción entre 196 factores de transcripción y 1,588 genes regulados.

La identificación de los genes ortólogos se realizó utilizando el método de *bidirectional best hit*. Primero se identificaron los genes homólogos entre dos genomas, el homólogo de un gen en un genoma de interés es el gen que representa el mejor resultado en el genoma de referencia. Se considera bidireccional si el gen del genoma de referencia devuelve como mejor homólogo al mismo gen en el genoma de interés. Un *bidirectional best hit* representa entonces, a dos genes en distintos genomas con la mayor similitud de entre todos los demás genes en ambos genomas, lo que puede indicar que los genes proceden de un ancestro común. Los ortólogos fueron aceptados si ellos tenían un *E*-valor $< 1e^{-6}$, una identidad de secuencia $\geq 30\%$ y una longitud de alineamiento $> 60\%$ del total del tamaño en ambas proteínas (Moreno-Hegelsieb G y Latimer K. 2008).

El acercamiento de *Regulogs* tiene dos principales suposiciones: la primera es asumir que los factores de transcripción conservan su actividad reguladora a través de los diferentes genomas (es decir, como activadores, represores y duales). Esto a su vez, implica que se conservan las estructuras de las proteínas y sus sitios de unión al ADN. La segunda es dada por el método de reconstrucción, con lo cual suponemos que si se conserva el factor de transcripción y los genes regulados, entonces las interacciones reguladoras son conservadas. Esto implicaría que no hay mutaciones en los sitios de unión del factor de transcripción al ADN, lo cual puede llevar a la pérdida de la interacción reguladora o mutaciones que cambien la proteína y el sitio de unión al ADN. Estas suposiciones buscan ser reducidas al considerar sólo a organismos filogenéticamente cercanos como es el caso de las γ -proteobacterias, lo que conduce a conservar las proteínas y las regiones reguladoras.

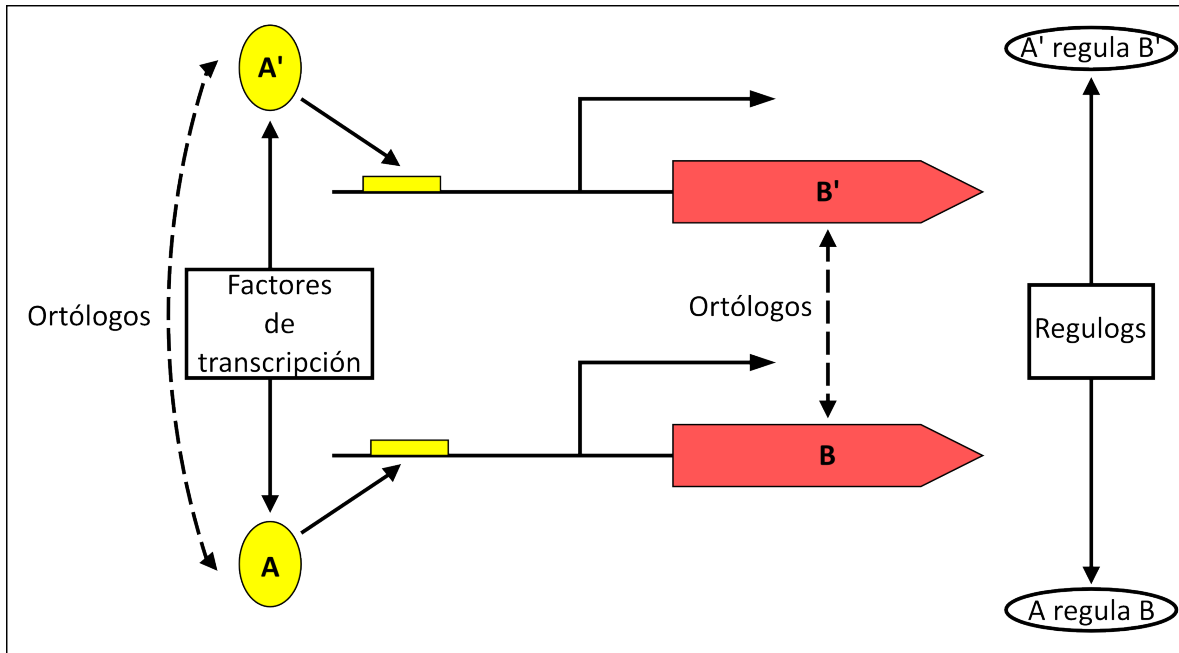


Figura 1. Estrategia de reconstrucción de interacciones entre pares de genes en las redes de regulación.

Para reconstruir las redes de regulación transcripcional se utilizó una aproximación de *Regulogs*, identificando los ortólogos de los factores de transcripción y de los genes regulados. Se consideró que si un factor de transcripción regula un gen regulado en el genoma de referencia y éstos tienen ortólogos en el genoma de interés, entonces la interacción reguladora es conservada.

Los 113 genomas se dividieron en cuatro categorías, siguiendo los lineamientos del grupo de Nancy Moran (McCutcheon JP y Moran NA. 2012). El primer grupo incluye a bacterias de vida libre con un tamaño de genoma menor al de *E. coli*; el segundo grupo comprende a las bacterias clasificadas como simbiotes recién restringidos a huéspedes; el tercero contiene a simbiotes obligados a largo plazo; y el cuarto contiene simbiotes con genomas diminutos, que exhiben reducción genómica extrema (con genoma < 300 Kb) (Tabla 1).

Tabla 1. Genomas de bacterias utilizados en este trabajo, clasificados de acuerdo a su estilo de vida.

| Bacteria | Tamaño del genoma (Mpb) | Número de genes | Estilo de vida |
|--|-------------------------|-----------------|----------------|
| <i>Escherichia coli</i> K 12 substr MG1655 | 4.638 | 4499 | Vida libre |
| <i>Yersinia enterocolitica</i> palearctica 105 5R r | 4.621 | 4206 | Vida libre |
| <i>Shewanella loihica</i> PV 4 | 4.602 | 3993 | Vida libre |
| <i>Alteromonas macleodii</i> English str Channel 673 | 4.601 | 3979 | Vida libre |
| <i>Alteromonas macleodii</i> str English Channel 615 | 4.582 | 3644 | Vida libre |
| <i>Escherichia coli</i> BW2952 | 4.578 | 4262 | Vida libre |
| <i>Cellvibrio japonicus</i> Ueda107 | 4.576 | 3812 | Vida libre |
| <i>Stenotrophomonas maltophilia</i> R551 3 | 4.573 | 4134 | Vida libre |
| <i>Pseudomonas stutzeri</i> A1501 | 4.567 | 4209 | Vida libre |
| <i>Psychromonas ingrahamii</i> 37 | 4.559 | 3863 | Vida libre |
| <i>Aeromonas veronii</i> B565 | 4.551 | 4170 | Vida libre |
| <i>Pseudomonas stutzeri</i> ATCC 17588 LMG 11199 | 4.547 | 4314 | Vida libre |
| <i>Shewanella denitrificans</i> OS217 | 4.545 | 3905 | Vida libre |
| <i>Stenotrophomonas maltophilia</i> JV3 | 4.544 | 4177 | Vida libre |
| <i>Cronobacter sakazakii</i> SP291 | 4.518 | 4286 | Vida libre |
| <i>Alteromonas macleodii</i> str Deep ecotype | 4.480 | 4121 | Vida libre |
| <i>Alteromonas macleodii</i> str Black Sea 11 | 4.480 | 3858 | Vida libre |
| <i>Salmonella bongori</i> NCTC 12419 | 4.460 | 4054 | Vida libre |
| <i>Alteromonas macleodii</i> str Ionian Sea U7 | 4.442 | 3931 | Vida libre |
| <i>Providencia stuartii</i> MRSN 2154 | 4.402 | 4196 | Vida libre |
| <i>Shewanella amazonensis</i> SB2B | 4.306 | 3785 | Vida libre |
| <i>Simiduia agarivorans</i> SA1 DSM 21679 | 4.300 | 3887 | Vida libre |
| <i>Ferrimonas balearica</i> DSM 9799 | 4.279 | 3947 | Vida libre |
| <i>Vibrio fischeri</i> ES114 | 4.273 | 3985 | Vida libre |
| <i>Cronobacter sakazakii</i> ES15 | 4.268 | 4018 | Vida libre |
| <i>Vibrio cholerae</i> MJ 1236 | 4.236 | 3910 | Vida libre |
| <i>Acinetobacter baumannii</i> TCDC AB0715 | 4.218 | 4010 | Vida libre |
| <i>Pseudomonas stutzeri</i> DSM 10701 | 4.174 | 3888 | Vida libre |
| <i>Shimwellia blattae</i> DSM 4481 NBRC 105725 | 4.158 | 4017 | Vida libre |
| <i>Acinetobacter oleivorans</i> DR1 | 4.152 | 3963 | Vida libre |

| Bacteria | Tamaño del genoma (Mpb) | Número de genes | Estilo de vida |
|--|-------------------------|-----------------|----------------|
| <i>Legionella longbeachae</i> NSW150 | 4.149 | 3739 | Vida libre |
| <i>Nitrosococcus halophilus</i> Nc 4 | 4.145 | 4087 | Vida libre |
| <i>Thioflavicoccus mobilis</i> 8321 | 4.137 | 3783 | Vida libre |
| <i>Glaciecola nitratreducens</i> FR1064 | 4.134 | 3720 | Vida libre |
| <i>Listonella anguillarum</i> M3 | 4.117 | 3882 | Vida libre |
| <i>Proteus mirabilis</i> HI4320 | 4.099 | 3862 | Vida libre |
| <i>Erwinia pyrifoliae</i> Ep1 96 | 4.072 | 3852 | Vida libre |
| <i>Erwinia tasmaniensis</i> Et1 99 | 4.067 | 3804 | Vida libre |
| <i>Marinobacter</i> sp BSs20148 | 4.063 | 3944 | Vida libre |
| <i>Halomonas elongata</i> DSM 2581 | 4.061 | 3554 | Vida libre |
| <i>Vibrio anguillarum</i> 775 | 4.052 | 3838 | Vida libre |
| <i>Vibrio cholerae</i> O1 biovar El Tor str N16961 | 4.033 | 3693 | Vida libre |
| <i>Thioalkalivibrio nitratreducens</i> DSM 14787 | 4.002 | 3880 | Vida libre |
| <i>Marinobacter hydrocarbonoclasticus</i> ATCC 49840 | 3.989 | 3843 | Vida libre |
| <i>Thalassolituus oleivorans</i> MIL 1 | 3.920 | 3732 | Vida libre |
| <i>Marinomonas posidonica</i> IVIA Po 181 | 3.899 | 3652 | Vida libre |
| <i>Acinetobacter calcoaceticus</i> PHEA 2 | 3.862 | 3674 | Vida libre |
| <i>Xanthomonas albilineans</i> GPE PC73 | 3.852 | 3266 | Vida libre |
| <i>Pseudoalteromonas haloplanktis</i> TAC125 | 3.850 | 3619 | Vida libre |
| <i>Proteus mirabilis</i> BB2000 | 3.846 | 3558 | Vida libre |
| <i>Erwinia amylovora</i> CFBP1430 | 3.833 | 3803 | Vida libre |
| <i>Edwardsiella tarda</i> EIB202 | 3.804 | 3719 | Vida libre |
| <i>Morganella morganii</i> subsp <i>morganii</i> KT | 3.799 | 3624 | Vida libre |
| <i>Vibrio cholerae</i> LMA3984 4 | 3.738 | 3442 | Vida libre |
| <i>Chromohalobacter salexigens</i> DSM 3043 | 3.696 | 3409 | Vida libre |
| <i>Allochromatium vinosum</i> DSM 180 | 3.669 | 3366 | Vida libre |
| <i>Frateuria aurantia</i> DSM 6220 | 3.603 | 3288 | Vida libre |
| <i>Nitrosococcus oceani</i> ATCC 19707 | 3.522 | 3185 | Vida libre |
| <i>Legionella pneumophila</i> subsp <i>pneumophila</i> | 3.492 | 3327 | Vida libre |
| <i>Tolomonas auensis</i> DSM 9187 | 3.471 | 3292 | Vida libre |
| <i>Thioalkalivibrio sulfidophilus</i> HL EbGr7 | 3.464 | 3372 | Vida libre |
| <i>Pseudoxanthomonas spadix</i> BD a59 | 3.452 | 3202 | Vida libre |
| <i>Pseudoxanthomonas suwonensis</i> 11 1 | 3.419 | 3172 | Vida libre |
| <i>Nitrosococcus watsonii</i> C 113 | 3.373 | 3245 | Vida libre |
| <i>Methylococcus capsulatus</i> Bath | 3.304 | 3052 | Vida libre |

| Bacteria | Tamaño del genoma (Mpb) | Número de genes | Estilo de vida |
|---|-------------------------|-----------------|--|
| <i>Alkalilimnicola ehrlichii</i> MLHE 1 | 3.275 | 2940 | Vida libre |
| <i>Acidithiobacillus caldus</i> SM 1 | 3.237 | 3235 | Vida libre |
| <i>Acidithiobacillus ferrivorans</i> SS3 | 3.207 | 3335 | Vida libre |
| <i>Methylophaga nitratireducenticrescens</i> | 3.137 | 3096 | Vida libre |
| <i>Alcanivorax borkumensis</i> SK2 | 3.120 | 2806 | Vida libre |
| <i>Psychrobacter cryohalolentis</i> K5 | 3.101 | 2581 | Vida libre |
| <i>Acidithiobacillus ferrooxidans</i> ATCC 23270 | 2.982 | 3303 | Vida libre |
| <i>Kangiella koreensis</i> DSM 16069 | 2.852 | 2697 | Vida libre |
| <i>Idiomarina loihiensis</i> GSL 199 | 2.839 | 2717 | Vida libre |
| <i>Methylophaga frappieri</i> | 2.745 | 2748 | Vida libre |
| <i>Xylella fastidiosa</i> 9a5c | 2.731 | 2905 | Vida libre |
| <i>Mannheimia haemolytica</i> D174 | 2.702 | 2814 | Vida libre |
| <i>Gallibacterium anatis</i> UMN179 | 2.694 | 2587 | Vida libre |
| <i>Halorhodospira halophila</i> SL1 | 2.678 | 2493 | Vida libre |
| <i>Cycloclasticus zancles</i> 7 ME | 2.655 | 2623 | Vida libre |
| <i>Psychrobacter arcticus</i> 273 4 | 2.650 | 2211 | Vida libre |
| <i>Halothiobacillus neapolitanus</i> c2 | 2.582 | 2465 | Vida libre |
| <i>Thiomicrospira crunogena</i> XCL 2 | 2.427 | 2259 | Vida libre |
| <i>Bibersteinia trehalosi</i> USDA ARS USMARC 192 | 2.407 | 2325 | Vida libre |
| <i>Pasteurella multocida</i> 36950 | 2.349 | 2202 | Vida libre |
| <i>Haemophilus parasuis</i> ZJ0906 | 2.324 | 2289 | Vida libre |
| <i>Actinobacillus succinogenes</i> 130Z | 2.319 | 2199 | Vida libre |
| <i>Mannheimia succiniciproducens</i> MBEL55E | 2.314 | 2449 | Vida libre |
| <i>Aggregatibacter actinomycetemcomitans</i> D7S 1 | 2.309 | 2334 | Vida libre |
| <i>Haemophilus somnus</i> 2336 | 2.263 | 2065 | Vida libre |
| <i>Actinobacillus pleuropneumoniae</i> serovar 3 str JL03 | 2.242 | 2147 | Vida libre |
| <i>Coxiella burnetii</i> Dugway 5J108 111 | 2.212 | 2362 | Vida libre |
| <i>Aggregatibacter actinomycetemcomitans</i> ANH9381 | 2.136 | 2142 | Vida libre |
| <i>Haemophilus influenzae</i> Rd KW20 | 1.830 | 1690 | Vida libre |
| <i>Haemophilus ducreyi</i> 35000HP | 1.698 | 1838 | Vida libre |
| <i>Sodalis glossinidius</i> str morsitans | 4.292 | 2607 | Simbiontes recién restringidos a huéspedes |
| <i>Candidatus Hamiltonella defensa</i> 5AT | | | Simbiontes recién restringidos a huéspedes |
| <i>Acyrtosiphon pisum</i> | 2.169 | 2196 | Simbiontes recién restringidos a huéspedes |
| <i>Francisella tularensis holarctica</i> LVS | 1.895 | 1744 | Simbiontes recién restringidos a huéspedes |

| Bacteria | Tamaño del genoma (Mpb) | Número de genes | Estilo de vida |
|---|-------------------------|-----------------|--|
| <i>Francisella tularensis mediasiatica</i> FSC147 | 1.893 | 1453 | Simbiontes recién restringidos a huéspedes |
| <i>Francisella tularensis tularensis</i> FSC198 | 1.892 | 1607 | Simbiontes recién restringidos a huéspedes |
| <i>Dichelobacter nodosus</i> VCS1703A | 1.389 | 1334 | Simbiontes recién restringidos a huéspedes |
| <i>Serratia symbiotica</i> str <i>Cinara cedri</i> Secondary endosymbiont of | 1.762 | 711 | Simbiontes obligados a largo plazo |
| <i>Ctenarytaina eucalypti</i> Secondary endosymbiont of | 1.441 | 963 | Simbiontes obligados a largo plazo |
| <i>Heteropsylla cubana</i> | 1.121 | 620 | Simbiontes obligados a largo plazo |
| Candidatus <i>Vesicomysocius okutanii</i> HA | 1.022 | 975 | Simbiontes obligados a largo plazo |
| <i>Wigglesworthia glossinidia</i> str endosymbiont of <i>glossina brevipalpis</i> | 0.703 | 641 | Simbiontes obligados a largo plazo |
| <i>Baumannia cicadellinicola</i> str Hc | | | Simbiontes obligados a largo plazo |
| <i>Homalodisca coagulata</i> | 0.686 | 657 | Simbiontes obligados a largo plazo |
| <i>Buchnera aphidicola</i> APS | | | Simbiontes obligados a largo plazo |
| <i>Acyrtosiphon pisum</i> | 0.655 | 608 | Simbiontes obligados a largo plazo |
| <i>Buchnera aphidicola</i> Sg <i>Schizaphis graminum</i> | 0.641 | 581 | Simbiontes obligados a largo plazo |
| <i>Buchnera aphidicola</i> Bp <i>Baizongia pistaciae</i> | 0.618 | 542 | Simbiontes obligados a largo plazo |
| <i>Buchnera aphidicola</i> <i>Cinara tujafilina</i> | 0.444 | 394 | Simbiontes obligados a largo plazo |
| <i>Buchnera aphidicola</i> BCc | 0.422 | 392 | Simbiontes obligados a largo plazo |
| Candidatus <i>Portiera aleyrodidarum</i> BT B | 0.351 | 296 | Simbiontes con genomas diminutos |
| Candidatus <i>Carsonella ruddii</i> PV | 0.159 | 213 | Simbiontes con genomas diminutos |

3.2 ANÁLISIS DE LOS FACTORES DE TRANSCRIPCIÓN CONSERVADOS

De los aproximadamente 300 factores de transcripción propuestos en *E. coli*, 196 han sido descritos experimentalmente y son parte de la red de regulación transcripcional en esta bacteria. Estos 196 fueron clasificados de acuerdo a la actividad reguladora que ejercen en: activadores, represores y duales (6 de los factores de transcripción no contienen información sobre su manera de regular la

transcripción, por lo que no se incluyeron en ninguna de las tres clasificaciones) (Keseler IM, *et al.* 2013). Con la finalidad de identificar si existe una conservación preferencial de algún tipo de regulación en las bacterias al disminuir el tamaño de los genomas, se identificaron los 190 factores de transcripción dentro de los 113 genomas y se reconstruyeron sus redes de regulación.

Por otro lado, para determinar si la conservación de los factores de transcripción era debida o no al azar, se utilizó un análisis de distribución hipergeométrica. Este permite estimar la probabilidad de encontrar más de x factores de transcripción de cada tipo de regulación en cada uno de los genomas. Se calculó la función de distribución acumulativa hipergeométrica, de la cual se obtuvieron tres p -valores para cada uno de los tipos de regulación (activadores, represores y duales). Se utilizó el lenguaje R con el siguiente comando: `phyper(x-1, m, n-m, k, lower.tail=FALSE)` donde x es el número de activadores, represores o duales en cada red reconstruida; m es el número total de activadores, represores o duales en la red de *E. coli*; n es el número total de factores de transcripción ortólogos en la red reconstruida y k es el número total de factores de transcripción ortólogos en la red reconstruida (Figura 2). Un valor menor de 0.05 es considerado como significativo, con lo cual se concluye que cada tipo de factores de transcripción (activadores, represores y duales) son conservados más frecuentemente de lo esperado aleatoriamente en cada uno de los genomas. El procesamiento de datos y computación de parámetros también fue realizado en R (www.r-project.org/).

Con el propósito de verificar si los análisis para la conservación de cada tipo de reguladores eran significativos desde un punto de vista global, se realizó un meta-análisis. Éste permite la integración de los resultados dispersos de múltiples estudios, que ponen a prueba la misma hipótesis con el fin de integrarlos mediante un estimador global. De manera que sea posible obtener una conclusión más fuerte, además de que también existe una mayor probabilidad de que el resultado

sea correcto en comparación con el obtenido de los análisis independientes. Esto puede ser realizado combinando estimadores o p -valores (Chernick MR. 2011).

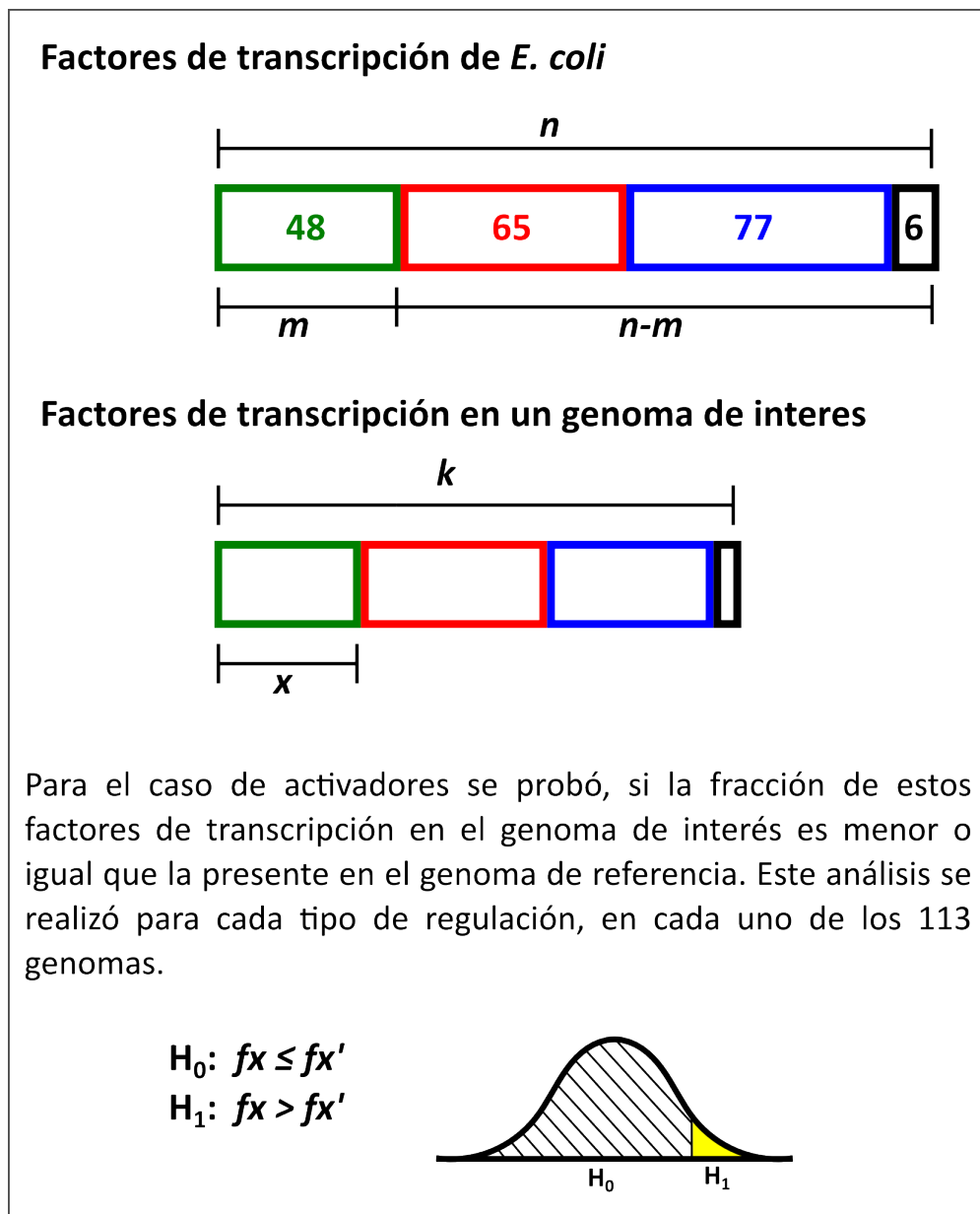


Figura 2. Prueba de hipótesis para corroborar la significancia de conservación de los factores de transcripción en otros genomas.

Se utilizó la prueba de distribución hipergeométrica para determinar si la reducción del número de genes en los genomas tenía una correlación con la conservación de los factores de transcripción. Se puso a prueba la hipótesis de que: la proporción de un tipo de reguladores en un genoma con un menor número de factores de transcripción

conserva la misma o menor proporción del tipo de factores de transcripción que en nuestro genoma de referencia.

El método de Fisher es una de las pruebas más utilizadas en este tipo de análisis, el cual fue desarrollado con el propósito de combinar p -valores estadísticos desde varios análisis independientes con la misma hipótesis (H_0) usando una distribución X^2 . Específicamente, la prueba supone que p_i es el p -valor desde el i ésimo estudio, de tal forma que tenemos:

$$X^2 = -2 \sum_{i=1}^K \ln(p_i)$$

La ventana de soluciones tiene una distribución X^2 con $2K$ grados de libertad, donde k es el número de pruebas que se combinan. Sus p -valores p_i siguen una distribución uniforme en el intervalo $[0,1]$ bajo la hipótesis nula. Al tomar el logaritmo natural negativo de un valor distribuido uniformemente, $-\ln(p_i)$ sigue una distribución exponencial. La escala da un valor que sigue una distribución exponencial por un factor de dos, y produce una cantidad que sigue una distribución de chi-cuadrada con dos grados de libertad. Finalmente, la suma de los K valores de chi-cuadrada independientes (cada uno con dos grados de libertad), sigue una distribución X^2 con $2K$ grados de libertad. Una ventaja al emplear el método de Fisher es que no se necesitan los datos de cada estudio y no es necesario que los análisis realizados en cada estudio sean iguales (Fisher RA. 1934).

El método de Fisher se utilizó para determinar si existe una significancia general en la conservación de cada uno de los tipos de factores de transcripción. Para ello se combinaron los p -valores de los 113 análisis estadísticos para cada tipo de factor de transcripción y se aplicó el método de Fisher, dichos análisis también se realizaron en R (Chen D-G y Peace KE. 2013).

3.3 ANÁLISIS DE LA INFLUENCIA DE LAS RELACIONES FILOGENÉTICAS EN LA CONSERVACIÓN DE TIPOS DE REGULADORES

Para definir la forma en que las fracciones de cada tipo de regulador (activadores, represores o duales) varía con la reducción en el tamaño del genoma, se calculó el coeficiente de correlación de Pearson, que permite medir el grado y la dirección de la relación lineal entre estas variables. Sin embargo, los resultados no pueden ser considerados estadísticamente independientes uno del otro, debido a que los genomas en el estudio por definición parten de una estructura filogenética (todas son γ -proteobacterias), es decir tienen un mismo ancestro común. Esta dependencia causada por la similitud filogenética puede generar correlaciones significativas entre el número de factores de transcripción y el tamaño del genoma cuando no exista una relación entre ellos. Para corregir estos efectos, se incorporó la información filogenética en los análisis estadísticos usando contrastes independientes de Felsenstein (Felsenstein J. 1985).

En el método propuesto por Felsenstein, se calculan las diferencias ponderadas entre los valores de los rasgos asociados a los pares de nodos, en cada punto de bifurcación en las ramas del árbol de filogenia conocida, asumiendo que el patrón de evolución sigue un camino aleatorio en el tiempo. Este acercamiento resulta en contrastes que son independientes pero idénticamente distribuidos, por lo tanto, pueden ser usados en análisis estadísticos convencionales. Se requieren tres tipos de información para usar el método de Felsenstein: 1) datos para dos o más rasgos fenotípicos para una serie de especies, 2) la relación cladística de estas especies y 3) la longitud de rama filogenética en unidades de varianza esperada de cambios (Felsenstein J. 1985, Garland T, *et al.* 1992, Garland T, *et al.* 1999).

Se utilizó el módulo de PDAP:PDTREE (Garland T, *et al.* 1999, Garland T y Ives AR. 2000) del programa Mesquite en su versión 3.02 (Maddison WP y Maddison DR. 2012) para calcular los contrastes independientes filogenéticamente estandarizados para los siguientes rasgos: tamaño del genoma, número de

activadores, número de represores, número de duales, porcentaje de activadores, porcentaje de represores, porcentaje de duales, porcentaje de globales y porcentaje de NAPs, además de calcular las regresiones lineales a través del origen (Garland T, *et al.* 1992). El árbol filogenético se realizó con el programa MrBayes-3.2, que implementa un método de inferencia bayesiano (Holder M y Lewis PO. 2003, Ronquist F, *et al.* 2012). Para reconstruir el árbol, se empleó el gen que codifica para el ARN 16S ribosomal de los 113 genomas del estudio, además del genoma de *E. coli* que se utilizó como referencia (Figura 3). Después de calcular los contrastes filogenéticos, se realizó un análisis de significancia usando un *p*-valor de 0.05 para determinar si las correlaciones obtenidas entre los contrastes calculados pudieran ser atribuidos sólo al azar.

3.4 IDENTIFICACIÓN DE REGULADORES GLOBALES

La red de regulación transcripcional contiene nodos altamente conectados llamados reguladores globales o *hubs*, que contribuyen a la cohesión estructural y robustez de la red y a su coherencia funcional (Lima-Mendez G y van Helden J. 2009). Previamente se han descrito los criterios operativos para la identificación de reguladores globales, estos incluyen: i) el número de genes regulados, ii) el número de factores de transcripción regulados, iii) el número de factores de transcripción con los que co-regula, iv) el número de factores sigma con los que co-regula, v) la heterogeneidad de las clases funcionales de sus genes regulados (Martínez-Antonio A y Collado-Vides J. 2003). Posteriormente se propuso una formalización de estos criterios mediante una ecuación (ver eq. 1) que asocia, para cada factor de transcripción *x*, un valor $G(x) \in [0,1]$ que indica su actividad global en el contexto de una red de regulación (Galán-Vásquez E. *et al.* 2011):

$$G(x) = \frac{1}{4} \left(\frac{TFR(x)}{N_{TF} + N_{SF} - 1} + \frac{GR(x)}{N_G} + \frac{SF(x)}{N_{SF}} + \frac{CR(x)}{N_{TF} - 1} \right) \quad (1)$$

Figura 3. Filogenia de las γ -proteobacterias usadas en este trabajo

Árbol filogenético obtenido con las 113 bacterias consideradas en este estudio utilizando el gen que codifica el 16s ribosomal. Las bacterias de vida libre son marcadas en color anaranjado, los simbioses recién restringidos a huéspedes son marcados en color amarillo, los simbioses obligados a largo plazo en color violeta y los simbioses con los genomas diminutos en color azul, en color rojo se muestra *E. coli K-12 MG1655*.

Donde N_{TF} es el número de factores de transcripción en la red de regulación, N_{SF} es el número de factores sigma en la red y N_G es el número de genes regulados en la red. Por otro lado, $TFR(x)$ representa el número de factores de transcripción regulados por x , $GR(x)$, el número de genes regulados por x , $SF(x)$ es el número de factores sigma que co-regulan con x , $CR(x)$ el número de factores de transcripción con los que co-regula x (Figura 4). La métrica descrita por la ecuación (1) define una jerarquía de reguladores globales en la red de regulación, convencionalmente se definen los 10 reguladores más globales, los cuales se identificaron para cada una de las redes de regulación reconstruidas.

3.5 ANÁLISIS DE REGRESIÓN MULTIVARIABLE

Como la conservación de la fracción de cada tipo de factor de transcripción puede ser influenciado no solo por el tamaño del genoma sino también por la naturaleza global de los reguladores, así como el que ellos pueden actuar como NAPs, se evaluó el efecto de estas variables adicionales por medio de regresiones lineales multivariantes. En estas ecuaciones se define como variable dependiente a la fracción relativa de cada tipo de factor de transcripción y se incluyen como variables independientes (predictores) a: el tamaño del genoma, naturaleza global de los reguladores y su función como NAPs. Los coeficientes de regresión parcial obtenidos desde la regresión multivariable permiten distinguir la contribución de cada variable de interés en la conservación de cada tipo de factor de transcripción cuando el efecto de los predictores adicionales se mantiene constante.

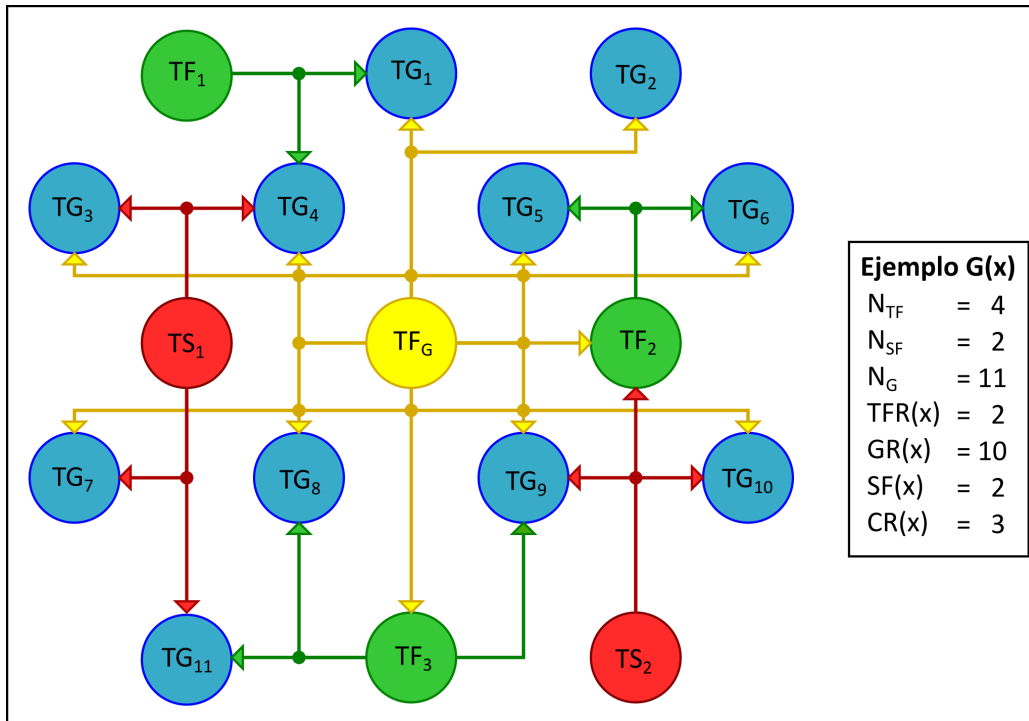


Figura 4. Identificación de reguladores globales

Ejemplo de la identificación del regulador global TF_G (nodo amarillo), en una red que contiene 17 nodos y 28 interacciones, donde TF_G , regula a 10 genes no reguladores (nodos en azul), regula a 2 factores de transcripción (nodos en verde), co-regula con 2 factores sigma (nodos en rojo) y 3 factores de transcripción.

Los coeficientes ($\beta_0, \beta_1, \beta_2, \beta_3$) se identificaron de acuerdo a la ecuación de multivariadas (2), usando una rutina construida en R, implementando el método de mínimos cuadrados:

$$Y_{FT} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \varepsilon \quad (2)$$

Y_{TF} es la predicción de mínimos cuadrados de la fracción relativa correspondiente a un tipo particular de factor de transcripción y las variables X_1, X_2, X_3 representan a los predictores (tamaño del genoma, reguladores globales, función NAPs, respectivamente) que contribuyen a la variable dependiente. El error ε es asumido para tener una distribución normal con media cero y varianza

constante. β_0 es la intersección con Y_{TF} cuando todas variables independientes son cero. Los otros coeficientes de regresión parciales ($\beta_1, \beta_2, \beta_3$) son las pendientes de cada uno de los predictores de Y_{TF} , manteniendo a las otras variables fijas (Alexopoulos EC. 2010). Debido a que los coeficientes son expresados en diferentes unidades, las β 's no pueden ser directamente usadas para comparar la contribución de cada predictor en el valor de Y_{TF} . Para evitar este problema, se transformaron las variables independientes en desviaciones estándar o medidas Z y se obtuvieron los correspondientes coeficientes de regresión parcial (b_1, b_2, b_3).

Después de transformar las β 's en b 's la magnitud relativa de los coeficientes b (considerando sus p -valores) se utilizó para comparar las magnitudes relativas de las relaciones entre Y_{TF} y cada predictor, mientras todos los otros predictores en la ecuación de regresión se mantuvieron constantes. De esta manera se determinó qué predictor tiene la relación más fuerte con la fracción de cada tipo de regulador (Davis JH. 2011).

4. RESULTADOS Y DISCUSIÓN

4.1 CONSERVACIÓN DE ACTIVADORES, REPRESORES Y DUALES

En este trabajo se estudió la evolución de las redes de regulación transcripcional en bacterias, tomando como referencia la red de *E. coli* K-12 MG1655. A partir de esta red, se reconstruyeron 113 redes de regulación para igual número de genomas de γ -proteobacterias, las cuales tienen menos genes que *E. coli*. La reconstrucción se realizó siguiendo una aproximación de genómica comparativa, identificando los factores de transcripción y genes regulados ortólogos a los de *E. coli* en cada uno de los otros 113 genomas. Las bacterias estudiadas incluyen 19 genomas de simbioses, los cuales están bajo el fenómeno de reducción genómica, lo que hace de ellos modelos biológicos convenientes para el estudio de la evolución de redes de regulación en el contexto de reducción genómica (Figura 5 y 6).

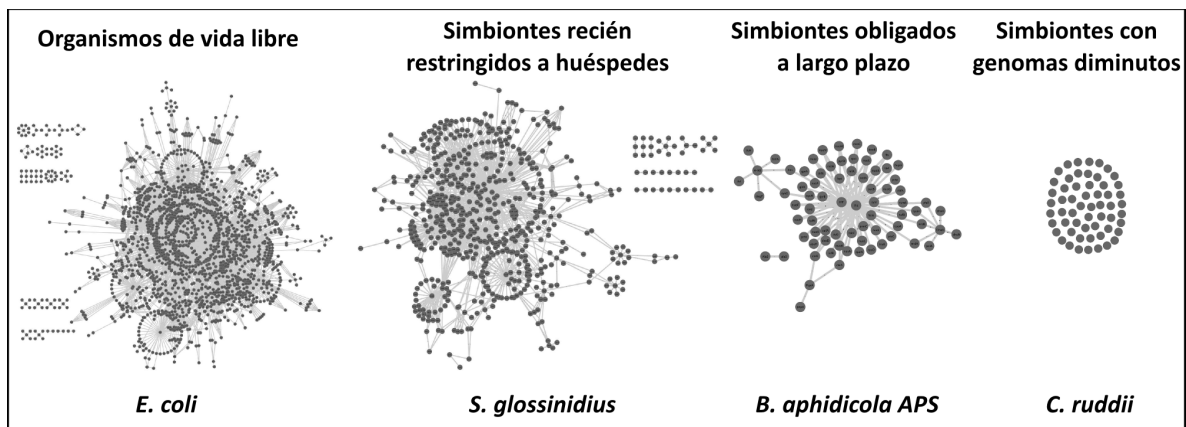


Figura 5. Ejemplo de redes de regulación transcripcional

Se muestran los grafos de las redes de regulación transcripcional más representativas en los niveles de clasificación estudiados en el presente trabajo, 94 genomas de organismos de vida libre, 6 simbioses recién restringidos a huéspedes, 11 simbioses obligados a largo plazo, y 2 simbioses con genomas diminutos (las redes fueron dibujadas en Cytoscape (Shannon P, *et al.* 2003)).

Estudios previos han demostrado que la maquinaria reguladora es el conjunto de nodos que más rápido evoluciona en las redes de regulación transcripcional (Babu MM, *et al.* 2006, Lozada-Chávez I, *et al.* 2006, Price NM, *et al.* 2007). Una primera aproximación para determinar la evolución en las redes, fue clasificar de acuerdo a su actividad reguladora a los 196 factores de transcripción caracterizados experimentalmente y que forman parte de la red de regulación de *E. coli* en: activadores, represores y duales. Además, se determinó la presencia o ausencia de estos reguladores en los 113 genomas de las γ -proteobacterias estudiadas y de esta forma se identificó si existe una conservación preferencial por el tipo de regulación transcripcional a través de los genomas analizados.

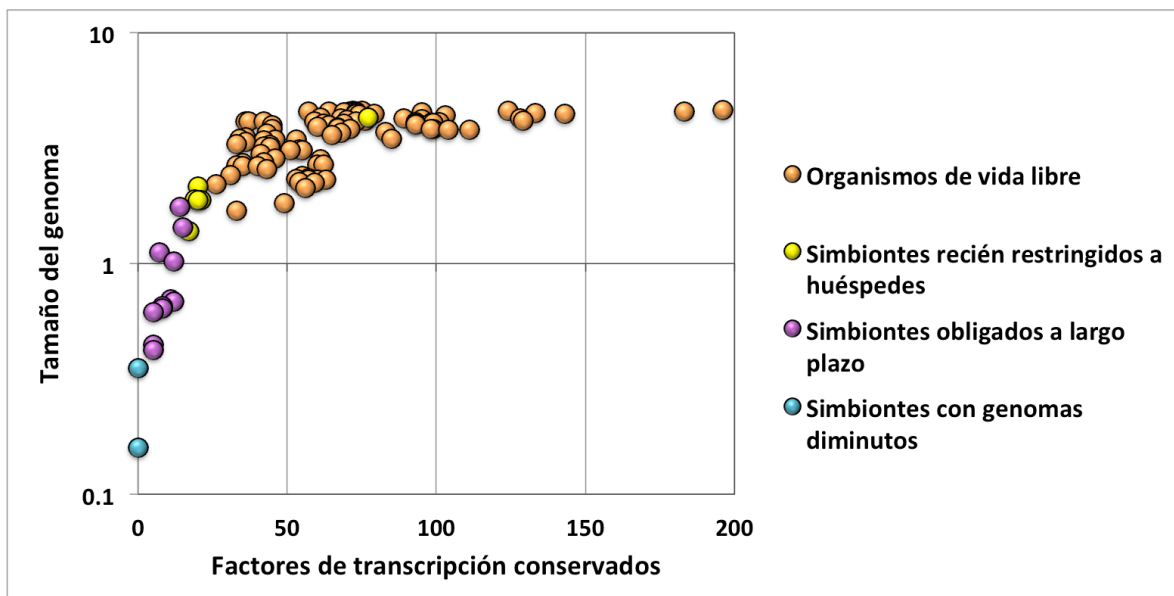


Figura 6. Distribución de los tamaños de los genomas utilizados en este estudio de evolución de redes.

Se utilizaron 113 genomas con tamaño de genoma menor que el de *E. coli*, en los cuales se identificaron los factores de transcripción conservados. Estos genomas fueron clasificados de acuerdo a su estilo de vida, en organismos de vida libre (círculos naranjas), organismos simbiontes recién restringidos a huéspedes (círculos amarillos), organismos simbiontes obligados a largo plazo (círculos morados) y organismos simbiontes con genomas diminutos (círculos azules).

Para medir el grado de correlación entre el tipo de regulación y el tamaño del genoma, primero se corrigieron las dependencias filogenéticas de los datos y después se realizó un análisis estadístico como se describe en la sección de métodos y datos. Se obtuvieron coeficientes de correlación de Pearson de $r = 0.555$ con un p -valor = 1.53×10^{-10} , entre el número de activadores y el tamaño de los genomas; $r = 0.711$ con un p -valor = 2.20×10^{-16} , entre el número de represores y el tamaño de los genomas; y $r = 0.637$ con un p -valor = 2.53×10^{-14} , entre el número de duales y el tamaño de los genomas (Figura 7).

El resultado de estos análisis indicó que, en general, el número de factores de transcripción disminuyen a medida que los tamaños de los genomas son más pequeños. Sin embargo, estas correlaciones pueden resultar de una pérdida generalizada o ausencia aleatoria de genes. Para determinar si esta pérdida de reguladores sigue una distribución aleatoria o no, se realizó un análisis de significancia del número de factores de transcripción obtenidos en cada uno de los 113 genomas y para cada uno de los tipos de reguladores. Se utilizó la distribución hipergeométrica para determinar la probabilidad de encontrar más de x reguladores de cada tipo, de tal forma que si la fracción del tipo de reguladores en cada genoma es similar o menor al que presenta *E. coli*, esto correspondería a que en los genomas mas pequeños los reguladores se pierden de manera aleatoria sin importar su tipo de regulación. Aplicando este análisis, se obtuvo que el número de factores de transcripción correspondiente a activadores fue menor que lo esperado aleatoriamente en la mayoría de genomas, por consiguiente el perfil de pérdida no fue significativamente desviado de lo esperado por la pérdida generalizada de genes. El número de represores conservados en los 113 genomas tampoco resultó ser significativamente diferente y los datos se ajustaron a lo esperado por la pérdida aleatoria de genes en la mayoría de genomas. Sin embargo, en el caso de la presencia de los reguladores duales la mayoría de genomas obtuvo un resultado estadísticamente significativo, que indica que este tipo de factores de transcripción son más conservados que lo esperado ante una

pérdida generalizada de genes. Se empleó un p -valor < 0.05 como corte para determinar si los datos eran significativos en cada prueba (Figura 8).

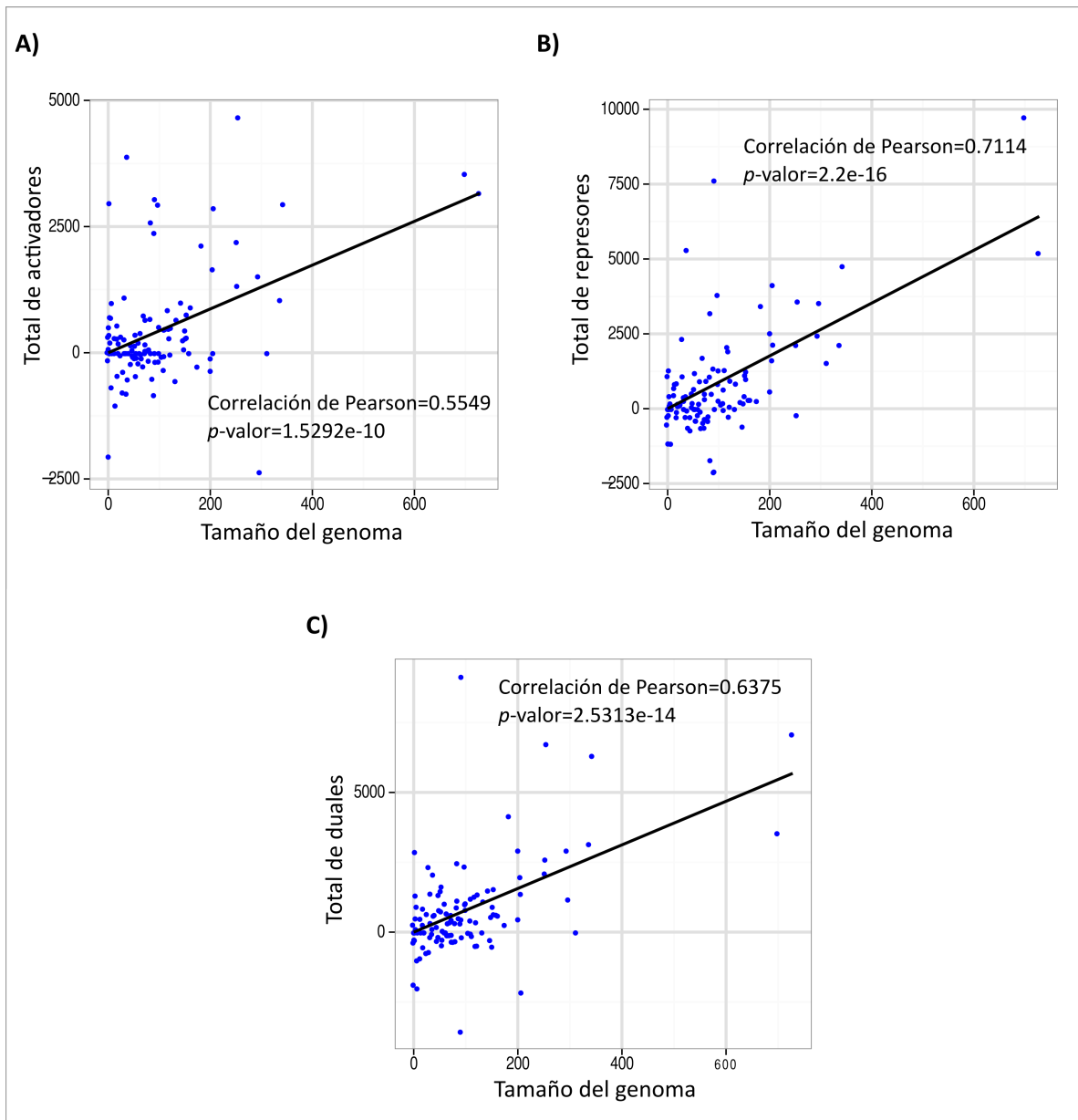


Figura 7. Contrastes filogenéticos de factores de transcripción conservados en los 113 genomas

Resultados de contrastes independientes filogenéticamente, para identificar la correlación entre la frecuencia del tipo de regulación contra el tamaño de los genomas (contrastos positivizados). Presentamos la regresión lineal a través del origen y la correlación de Pearson con su respectivo p -valor para: A) Los activadores conservados,

B) Represores conservados y C) Duales conservados. Cada punto representa los contrastes calculados con el método de Felsenstein.

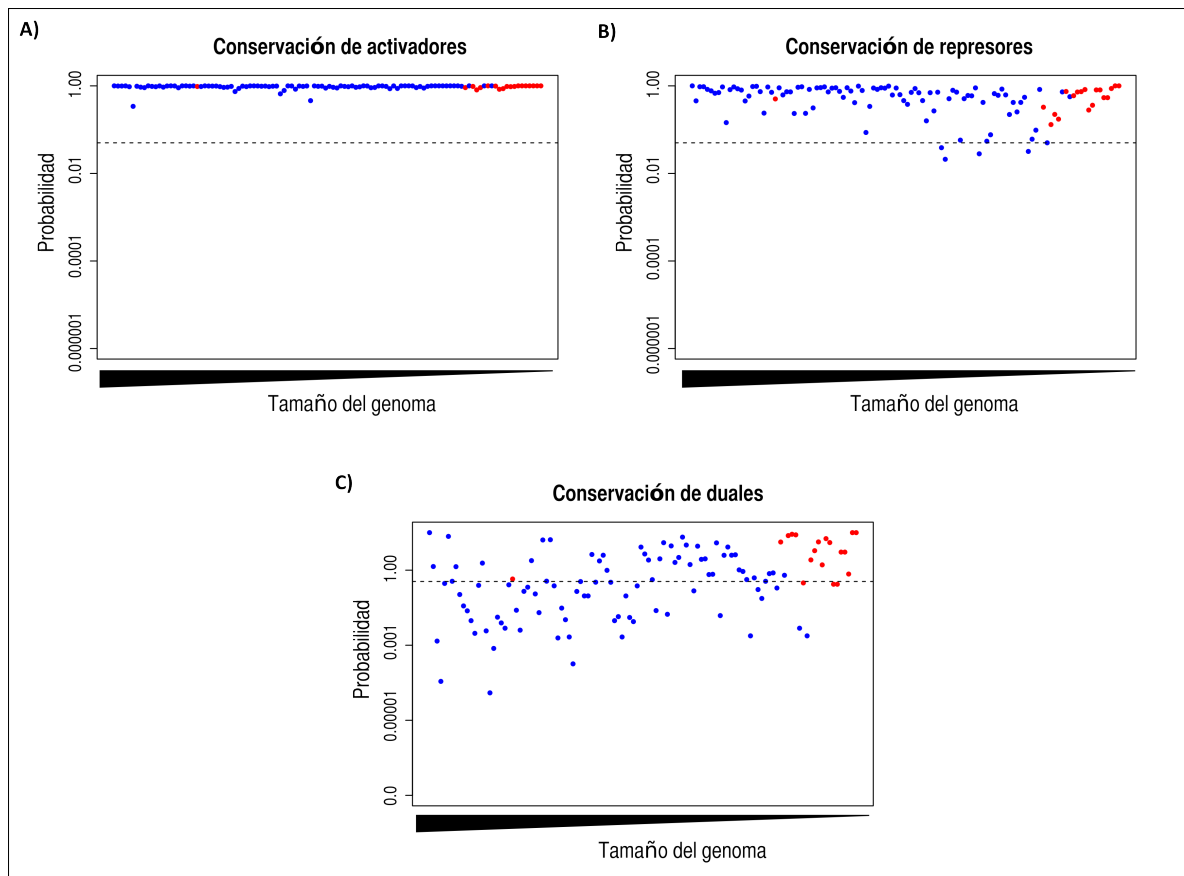


Figura 8. Probabilidades calculadas por la distribución hipergeométrica para cada uno de los genomas

Representación de la presencia de cada tipo de reguladores, si son conservados más de lo esperado aleatoriamente. A) Conservación de activadores, B) conservación de represores, C) conservación de duales. En los 113 genomas representados por los puntos en las gráficas, los de color azul corresponden a organismos de vida libre mientras que los de color rojo representan a los simbiotes. Los genomas están ordenados por tamaño de mayor a menor (de izquierda a derecha). La línea punteada marca el valor de corte de significancia de 0.05.

Para determinar si los análisis anteriores son estadísticamente significativos a nivel global de cada tipo de factores de transcripción, se realizó un análisis con el método de Fisher, el cual nos permite combinar un conjunto de pruebas independientes, las cuales tiene la misma hipótesis. Se utilizaron los p -valores de las 113 pruebas independientes para cada tipo de factores de transcripción y de esta forma se comprobó si existía una significancia estadística. En el caso de los activadores, se obtuvo un p -valor = 0.99, lo que nos indica nuevamente que no tiene una significancia estadística y lo mismo ocurre en el caso de los represores con un p -valor = 0.99. En cambio, los reguladores duales muestran un p -valor = 2.91×10^{-50} , lo que indica que tienen una fuerte significancia estadística, es decir, que se conservan más que lo esperado aleatoriamente.

Para evaluar más a fondo la desviación en la conservación de los factores de transcripción duales a través del proceso de reducción genómica, se realizó un análisis similar al de factores de transcripción totales utilizando los contrastes filogenéticos. En este caso, se utilizaron las frecuencias relativas correspondientes a cada tipo de reguladores, en lugar del número total de factores de transcripción por genoma. Resultó una correlación positiva para los activadores y los represores con un coeficiente de correlación de $r = 0.254$ (p -valor = 0.01) y $r = 0.225$ (p -valor = 0.02) respectivamente, y una correlación negativa para los reguladores duales $r = -0.321$ (p -valor = 4.90×10^{-4}) (Figura 9).

Los análisis revelan que del total de factores de transcripción, los activadores y los represores tienden a estar más presentes en los organismos con genomas más grandes (organismos de vida libre), mientras que la proporción de reguladores duales es mayor en organismos con genomas pequeños (como los endosimbiontes obligados). Estos resultados sugieren que existe una selección evolutiva en la conservación del tipo de factores de transcripción durante el proceso de reducción de genomas, siendo los reguladores duales el tipo más preservado en genomas que exhiben mayor reducción genómica, hasta cierto límite (ver mas adelante).

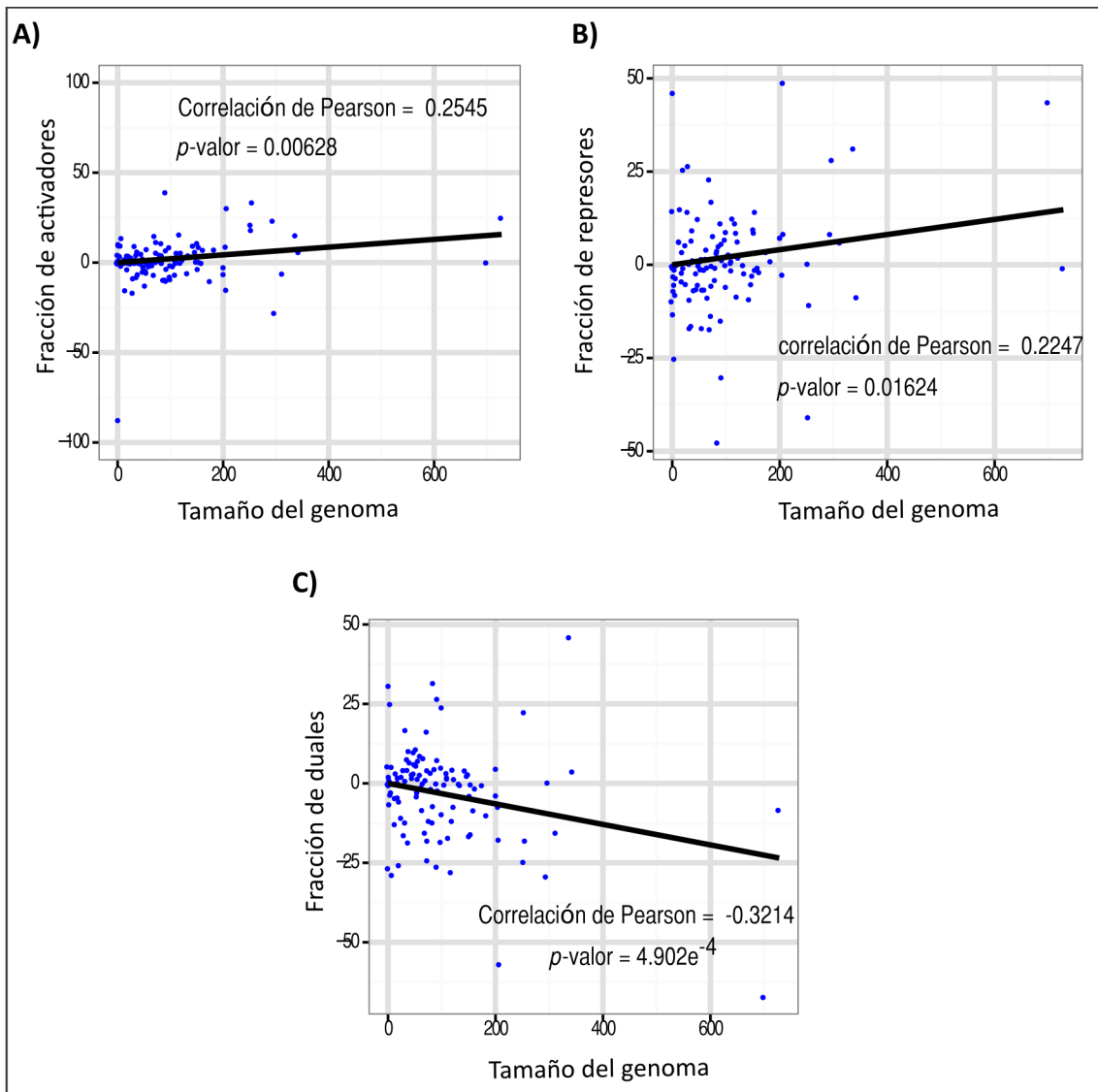


Figura 9. Contrastes filogenéticos de las frecuencias relativas de la conservación de los factores de transcripción

Contrastes independientes filogenéticamente para identificar la correlación entre la frecuencia del tipo de regulación contra el tamaño de los genomas (contrastes positivizados). Se muestra la regresión lineal a través del origen y la correlación de Pearson con su respectivo p -valor para: A) Activadores conservados, B) Represores conservados y C) Duales conservados. Cada punto representa los contrastes calculados con el método de Felsenstein.

4.2 GENES EN LAS REDES DE REGULACIÓN CONSERVADOS EN ENDOSIMBIOTES

Existe un marcado contenido en el número de genes asociados al estilo de vida de un organismo, los endosimbiontes presentan genomas de menos de ~1000 Kpb. En el presente estudio este grupo de organismos comprende un total de 13 genomas. Comparados con los organismos de vida libre, hay pocos factores de transcripción presentes en las bacterias simbiotes, posiblemente porque son perdidos en el proceso de reducción genómica, debido a que se ha demostrado que no son genes esenciales. Los principales factores de transcripción conservados son: AlaS que es un miembro de la familia de las aminoacil-tRNA sintetetasas. Se conoce que este regulador en *E. coli* actúa como un autorepresor del gen *alas* (Cusack S, *et al.* 1991), su actividad reguladora depende de la concentración de alanina en la célula. Otro regulador conservado es PepA, el cual es una enzima multifuncional de unión a ADN, involucrada en el metabolismo de arginina y prolina (Minh PN, *et al.* 2009). BirA es miembro de la familia de los *biotin repressors-like*, es una proteína bifuncional que exhibe actividad de ligasa de biotina y también actúa como represor de la expresión del operón de biotina (Eisenberg MA, *et al.* 1982). AccB es un miembro de las proteínas *biotinyl-lipoyl-carrier*, es un regulador del metabolismo central (Cronan JE. 2001). Todos estos factores de transcripción pueden ser clasificados como reguladores del metabolismo bacteriano.

Por otro lado, las proteínas asociadas a nucleóide como HU, IHF, Fis y H-NS son también bien conservadas en los organismos endosimbiontes. Este tipo de factores de transcripción ejercen una regulación a nivel de la transcripción, junto con una regulación dada por la estructuración general del nucleóide. Estas proteínas se unen de manera poco específica a regiones ricas en A+T (Macvanin M y Adhya S. 2012). HU es un regulador miembro de la familia *IHF-like*, codificado por los genes *hupA* y *hupB*, HU puede ser activo como heterodímero o homodímero, induce curvaturas en el ADN, condensa el ADN en una fibra y

también interactúa con una sola de las hebras del ADN (Swinger KK y Rice PA. 2004, Macvanin M y Adhya S. 2012). IHF es también un miembro de la familia de unión a ADN *IHF-like* codificada por *ihfA* y *ihfB*, introduce giros en forma de U en el ADN. Su función principal es doblar fuertemente las hebras de ADN y por lo tanto funciona en conjunto con otros NAPs y otros factores de transcripción (Swinger KK y Rice PA. 2004, Prieto AI, *et al.* 2011). H-NS es un miembro de la familia H-NS *histone-like*, es capaz de condensar y superenrollar el ADN (Dillon SC y Dorman, 2010). Fis es un miembro de la familia *Fis-like*, que tiene un papel importante en la organización y mantenimiento de la estructura del nucleóide a través de la unión directa con el ADN y a través de la modulación de girasas y topoisomerasas (Dillon SC y Dorman, 2010, Duprey A, *et al.* 2014).

Adicionalmente, otros factores de transcripción relacionados con la regulación de iones metálicos son también conservados como lo son Fur y Zur, que son miembros de la familia *Fur-like*. Fur regula la expresión de genes involucrados en mantener la homeostasis celular de hierro y Zur a genes involucrados en el sistema de transporte ABC que capta el zinc extracelular (Panina EM, *et al.* 2003, Chen Z, *et al.* 2007). IscR regula la expresión de genes involucrados en la formación del clúster hierro-sulfuro y para la formación de bio-películas (Giel IL, *et al.* 2006). Finalmente, también se conservan factores de transcripción que regulan a genes involucrados en diferentes procesos, tales como: DnaA, que regula el inicio del proceso de replicación celular de una manera dependiente de la suficiencia de ATP y por la interacción directa con su sitio de unión al ADN en el origen de replicación del nucleóide bacteriano (Donczew R, *et al.* 2014). BolA es miembro de la familia de reguladores *BolA-like*, y regula a genes de respuesta a condiciones de estrés entre los que están genes que funcionan en la fase estacionaria (Freire P, *et al.* 2009). SlyA pertenece a una familia de proteínas de bajo peso molecular y regula la expresión de una hemolisina críptica, interviene en los procesos de división celular y de adaptación de la estructura celular (Wyborn NR, *et al.* 2004).

Por otro lado, los genes que no son reguladores y que son más conservados se relacionan a funciones en procesos del metabolismo celular central tales como: metabolismo de aminoácidos, obtención de energía, metabolismo central e intermedio y degradación de macromoléculas. Otro conjunto de genes altamente conservados pertenece a la categoría de procesos celulares que incluye genes involucrados en el metabolismo del ADN y del ARN, así como, involucrados en la estructura de la membrana celular, ribosomas, división celular, ciclo celular, adaptación al estrés y genes involucrados en el transporte de metabolitos exógenos (Figura 10). Se ha propuesto que las bacterias simbiotas retienen este tipo de genes para establecer una adecuada relación con su huésped, de esta forma le suministran aminoácidos y precursores esenciales para el desarrollo del huésped (Moran NA y Bennett GM. 2014).

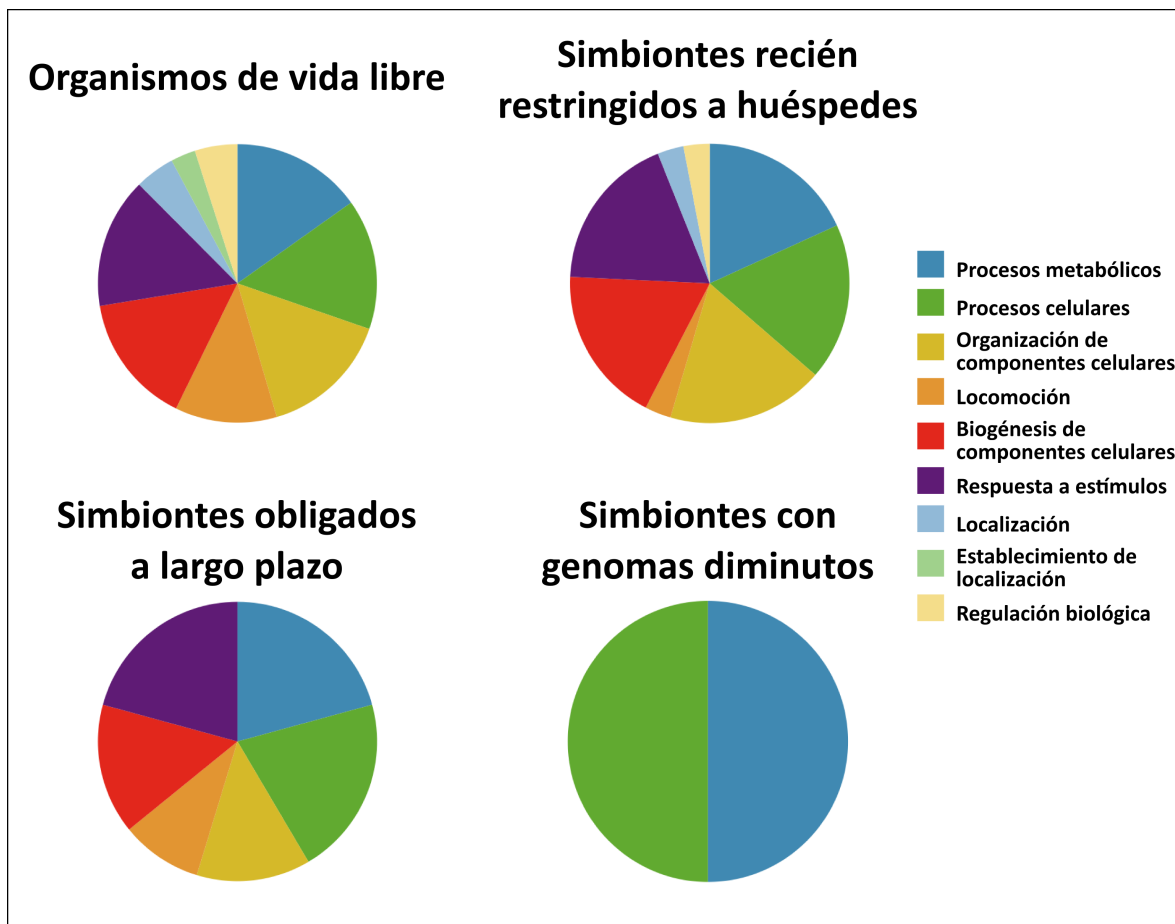


Figura 10. Clasificación de los genes regulados por ontología de proceso biológicos presentes en las redes de regulación transcripcional reconstruidas

Se utilizaron las primeras categorías (categorías parentales) de la ontología de proceso biológico de *Gene Ontology* para clasificar los genes no regulados de las redes de regulación transcripcional reconstruidas. Se observó una conservación de genes que intervienen principalmente en procesos metabólicos y procesos celulares a través de los diferentes niveles de reducción genómica. Se muestran las distribuciones de cada categoría conservada en genomas representativos de cada nivel de reducción, iniciando con los organismos de vida libre (*E. coli*), simbioses recién restringidos a huéspedes (*Sodalis glossinidius*), simbioses obligados a largo plazo (*Buchnera aphidicola* APS) y simbioses con genomas diminutos (*Candidatus Carsonella ruddii*).

Una característica importante de las redes, son los nodos altamente conectados denominados *hubs* o reguladores globales en el contexto de redes de regulación transcripcional. Los diez reguladores globales de la red de *E. coli* son los siguientes: Crp, IHF, FNR, Fis, H-NS, CpxR, ArcA, Fur, NsrR y Lrp (Tabla 2). Estos reguladores también son altamente conservados dentro de las redes de regulación reconstruidas, principalmente aquellos que son clasificados como duales y/o como proteínas asociadas a nucleóide. Para determinar si los reguladores globales y los asociados al nucleóide son más conservados que lo esperado aleatoriamente, se utilizó el análisis de distribución hipergeométrica (Figura 11). Analizando los resultados con el método de Fisher se obtuvo un p -valor = 7.25×10^{-127} y p -valor = 8.25×10^{-49} para los reguladores globales y las NAPs respectivamente, lo que indica que es estadísticamente significativo y se sugiere que son conservados más de lo esperado aleatoriamente.

Tabla 2. Los reguladores globales de *E. coli*

| Reguladores globales | Genes regulados | Co-regula | Sigma | TFs regulados | Valor de G |
|-----------------------------|------------------------|------------------|--------------|----------------------|-------------------|
| CRP | 495 | 12 | 7 | 46 | 0.49 |
| FNR | 296 | 12 | 6 | 18 | 0.30 |
| IHF | 219 | 12 | 6 | 9 | 0.30 |
| Fis | 225 | 12 | 6 | 9 | 0.25 |
| H-NS | 179 | 12 | 6 | 21 | 0.20 |
| ArcA | 172 | 12 | 6 | 7 | 0.18 |
| Fur | 129 | 11 | 7 | 6 | 0.18 |
| CpxR | 63 | 12 | 6 | 2 | 0.18 |
| NsrR | 83 | 3 | 6 | 7 | 0.16 |
| Lrp | 103 | 9 | 6 | 4 | 0.15 |

Por tal motivo, enseguida se planteó responder a la pregunta ¿los factores de transcripción se conservan por su efecto dual o importa también que sean reguladores globales o proteínas asociadas a nucleóide? Para responder esto, se utilizó un análisis de regresión lineal multivariable, para identificar cuál de las variables tienen mayor influencia en la conservación de cada tipo de factor de transcripción.

A partir de los coeficientes de regresión parcial estandarizados resultantes en cada una de las regresiones lineales, se infirió el efecto que tienen los predictores en la conservación de los activadores, represores y duales. Para el caso de los activadores y represores, se identificó que la influencia de las variables predictoras (es decir, tamaño del genoma, naturaleza global y función como NAP) fue despreciable, como puede observarse en la Tabla 3. Por otro lado, los coeficientes de regresión parcial estandarizados para los reguladores duales indicaron que la naturaleza global de estos reguladores (con una $b = 0.49$ y p -valor = 2.00×10^{-4}) puede explicar su conservación preponderante.

Interesantemente, una gran proporción de reguladores duales (globales), tienen un papel como proteínas asociadas a nucleóide en endosimbiontes. Al realizar la correlación de Pearson entre los reguladores globales y las NAP, se identificó una fuerte relación con un coeficiente de correlación igual a 0.75. Esto indica que tales variables no son estrictamente independientes, porque son redundantes y aportan la misma información. De tal forma que, si se remueve alguna de ellas (pero no ambas) en la ecuación de regresión multivariable, no se tiene un efecto notable en los valores predichos de la variable dependiente (fracción de reguladores duales).

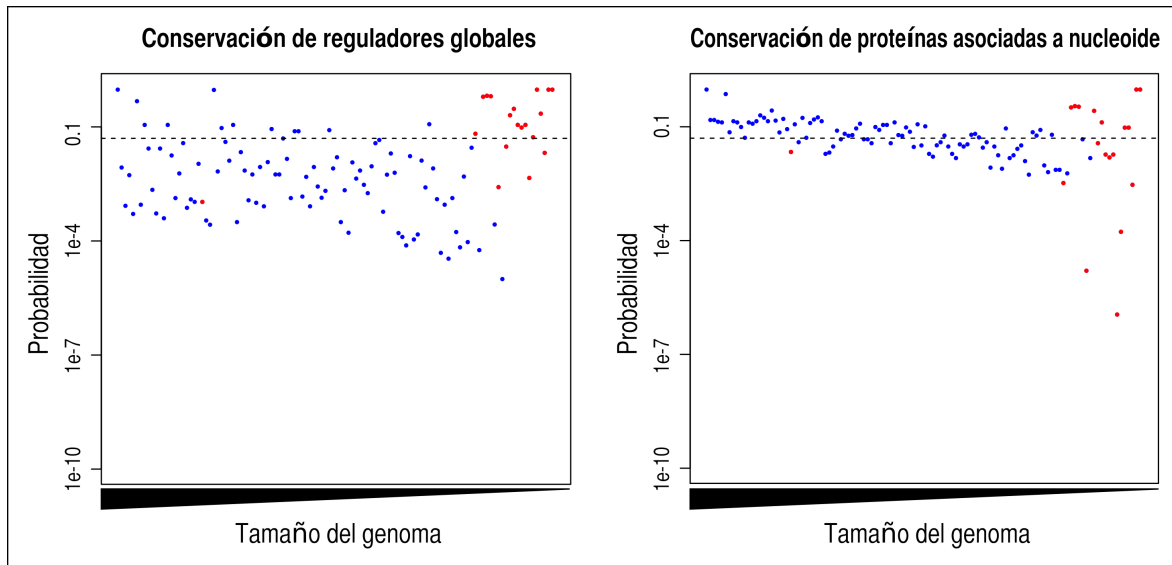


Figura 11. Probabilidades calculadas desde la distribución hipergeométrica para reguladores globales y NAPs

Análisis de la conservación de cada tipo de regulador. A) Conservación de reguladores globales, B) Conservación de NAPs. Los 113 genomas son representados por puntos; los de color azul corresponden a organismos de vida libre mientras que los puntos en color rojo representan a los simbiotes. Los genomas están ordenados de mayor a menor tamaño de genoma (de izquierda a derecha). La línea punteada marca el valor de corte de 0.05.

Tabla 3. Coeficiente de regresión parcial estandarizadas

| Variables independientes | Activadores | | Represores | | Duales | |
|--------------------------|----------------------|-----------------|----------------------|-----------------|----------------------|--------------------|
| | Coeficiente <i>b</i> | <i>p</i> -valor | Coeficiente <i>b</i> | <i>p</i> -valor | Coeficiente <i>b</i> | <i>p</i> -valor |
| Tamaño del genoma | 0.21 | 0.03 | 0.04 | 0.68 | -0.15 | 0.10 |
| Reguladores globales | -0.10 | 0.50 | 0.12 | 0.42 | 0.49 | 2x10 ⁻⁴ |
| Función como NAPs | -0.09 | 0.54 | -0.09 | 0.55 | -0.07 | 0.61 |

* Valores significativos (menores de 0.05)

4.3 INTERPRETACIÓN BIOLÓGICA DE LA CONSERVACIÓN DE FACTORES DE TRANSCRIPCIÓN EN EL CONTEXTO DE REDUCCIÓN GENÓMICA

En este apartado se hace una interpretación biológica de la conservación de los factores de transcripción de acuerdo a las funciones asociadas para cada tipo de regulador. En el caso de los activadores, éstos generalmente dependen de un co-inductor para responder a una señal intra o extracelular (por ejemplo: la respuesta adaptativa, respuesta a antibióticos, factores de virulencia, etc.) (Pérez-Rueda E y Collado-Vides 2000, Martínez-Antonio A, *et al.* 2006, Balderas-Martínez YI, *et al.* 2013). La pérdida de estos factores de transcripción, pueden no afectar sustancialmente la expresión y sobre todo la función de sus genes regulados. Esto puede ser dado a la estructura no restrictiva del nucleóide y al contenido rico en A+T de los procariontes con genomas pequeños, lo que puede facilitar el acceso de la maquinaria transcripcional a la región promotora sin necesidad de activadores. De tal forma que exista una transcripción basal de los genes que

pierden su activador (Struhl K. 1999). Adicionalmente, estudios recientes proveen fuertes evidencias de que la ARN polimerasa pasa la mayor parte del tiempo unida de manera no-específica al ADN (alrededor del 85% del tiempo) por lo cual se sugiere que siempre está localizando un sitio promotor (Stracy M, *et al.* 2015). Esto permite que los genes regulados positivamente, puedan mostrar cierto nivel de expresión y por lo tanto, conservar su función a pesar de la pérdida de sus activadores (ver Tabla 4 para ejemplos de activadores).

Tabla 4. Los 10 activadores más conservados en las redes de regulación transcripcional de las γ -proteobacterias

| Nombre del factor de transcripción | Genes regulados en la red de <i>E. coli</i> | Número de genomas en los que se conserva (de 113) | Función celular (generalmente descrita en <i>E. coli</i>) |
|---|--|--|--|
| QseB | 4 | 64 (56.6%) | Regula la transcripción de genes involucrados en la biogénesis del flagelo. Es miembro del sistema de dos componentes QseBC (Sperandio V, <i>et al.</i> 2002). |
| SlyA | 1 | 63 (55.7%) | Incrementa la expresión de la hemolisina E mediante antagonizar el efecto represivo de H-NS (Wyborn NR, <i>et al.</i> 2004). |
| ZntR | 1 | 61 (53.9%) | Regula genes involucrados en el sistema de transporte de Zn, Cd y Pb (Binet MR y Poole RK. 2000). |
| CusR | 6 | 51 (45.1%) | Regula genes relacionados al sistema de expulsión de cobre y |

| | | | |
|------|----|------------|--|
| | | | plata, bajo condiciones de crecimiento anaeróbico (Franke S, <i>et al.</i> 2001). |
| YehT | 1 | 51 (45.1%) | Miembro del sistema de dos componentes YehU/YehT, involucrado en el control de la fase estacionaria (Kraxenberger T, <i>et al.</i> 2012). |
| YqhC | 2 | 45 (39.8%) | Regula la expresión de genes involucrados en la tolerancia al furfural y otros componentes ambientales tóxicos (Turner PC, <i>et al.</i> 2011). |
| ArgP | 14 | 43 (38.0%) | Regula la transcripción de genes involucrados en el sistema de transporte de arginina y de genes involucrados en la replicación del ADN (Thony B, <i>et al.</i> 1991). |
| RstA | 10 | 37 (32.7%) | Controla genes involucrados en tolerancia a ácido, para la formación de fimbria, curli y para la respiración anaerobia (Ogasawara H, <i>et al.</i> 2007). |
| KdpE | 4 | 37 (32.7%) | Regula los genes implicados en el sistema de captación de potasio (Ballal A, <i>et al.</i> 2007). |
| CadC | 3 | 36 (31.8%) | Es un activador sensible a metales que regula la expresión de genes implicados en la síntesis y secreción de |

cadaverina a altas concentraciones de lisina (Kuper C y Jung K. 2005).

En cuanto a los represores, encontramos que la proporción de este tipo de factores de transcripción por genoma, es ligeramente mayor a la fracción de activadores en el curso de la reducción genómica (~15% más). Desde un punto de vista biológico, la pérdida de represores puede llevar a generar un gasto innecesario de fuentes celulares, debido a que los genes que pierden esta represión tendrían ahora una expresión constitutiva. Además, la conservación de la represión transcripcional puede ser una ventaja biológica sobre la activación debido a que el control homeostático es usualmente mantenido a través de la regulación negativa, el cual es un principio del metabolismo (Gerosa L, *et al.* 2013). Por otro lado, se conoce que la autoregulación negativa puede acelerar los tiempos de respuesta en los sistemas celulares (Alon U. 2007, Madar D, *et al.* 2011), y la respuesta dinámica de este tipo de control contribuye al mantenimiento homeostático sin considerar el estado fisiológico de los organismos (ver Tabla 5 para ejemplos) (Klumpp P, *et al.* 2009).

Tabla 5. Los 10 represores más conservados en las redes de regulación transcripcional de las γ -proteobacterias

| Nombre del factor de transcripción | Genes regulados en la red de <i>E. coli</i> | Número de genomas en los que se conserva (de 113) | Función celular (normalmente descrita en <i>E. coli</i>) |
|---|--|--|---|
| AlaS | 1 | 110 (97.3%) | Reprime a los genes de la alanil-tRNA sintetasa (Cusack S, <i>et al.</i> 1991). |
| PepA | 3 | 108 (95.5%) | Es una peptidasa que se une al |

| | | | |
|------|----|-------------|---|
| | | | ADN y también regula síntesis de genes involucrados en la síntesis de carbamoil fosfato (Minh PN, <i>et al.</i> 2009). |
| BirA | 5 | 104 (92.0%) | Reprime la transcripción de genes necesarios para la síntesis de biotina (Eisenberg MA, <i>et al.</i> 1982). |
| LexA | 56 | 81 (71.6%) | Reprime a genes involucrados en la respuesta celular a daño del ADN y también inhibe la replicación de ADN cuando hay daños celulares (Fernández De Henestrosa AR, <i>et al.</i> 2000). |
| PutA | 2 | 78 (69.0%) | Regula la transcripción de genes involucrados en la degradación de prolina (Vinod MP, <i>et al.</i> 2002). |
| Zur | 6 | 75 (66.3%) | Regula a genes involucrados en el sistema de captación de Zinc (Panina EM, <i>et al.</i> 2003). |
| ArsR | 3 | 69 (61.0%) | Reprime genes involucrados en la resistencia a metales como arsénico y antimonita (Carlin A, <i>et al.</i> 1995). |
| FabR | 2 | 65 (57.5%) | Regula la expresión de genes esenciales para la síntesis de ácidos grasos mono-insaturados (Marrakchi H, <i>et al.</i> 2002). |
| NsrR | 83 | 58 (51.3%) | Regula genes involucrados en la protección celular contra óxido |

| | | | |
|------|---|------------|---|
| AcrR | 4 | 56 (49.5%) | nítrico (Rankin LD, <i>et al.</i> 2008). Regula la expresión de genes implicados en el transporte de multidrogas (Ma D, <i>et al.</i> 1996). |
|------|---|------------|---|

Por otro lado, la mayor proporción de factores de transcripción duales conservados en los genomas a medida que se hacen más pequeños, puede deberse al hecho de que estos reguladores permiten un control más flexible (positivo y negativo), de sus genes regulados. Además, se ha observado que los factores de transcripción duales, también son usualmente reguladores globales y sus genes regulados generalmente pertenecen a más de una clase funcional (ver Tabla 6) (Huynen MA, *et al.* 2005, Dillon SC y Dorman CJ. 2010). Se sugiere también que los reguladores globales orquestan la actividad de la red de regulación y confieren a los organismos la habilidad de mantener la adecuación biológica (*fitness*) con una maquinaria reguladora mínima.

Tabla 6. Los 10 reguladores duales más conservados en las redes de regulación transcripcional de las γ -proteobacterias

| Nombre del factor de transcripción | Genes regulados en la red de <i>E. coli</i> | Número de genomas en los que se conserva (113) | Función celular (descrita en <i>E. coli</i>) |
|------------------------------------|---|--|--|
| DnaA | 12 | 108 (95.5%) | Regula el inicio de la replicación del ADN, por la interacción con el origen de replicación (Donczew R, <i>et al.</i> 214). |
| HU | 9 | 107 (94.6%) | Proteína codificada por <i>hupA</i> y <i>hupB</i> , que regula la organización del nucleóide introduciendo curvaturas en el ADN (Swinger |

| | | | |
|------|-----|-------------|---|
| IHF | 219 | 105 (92.9%) | KK y Rice PA. 2004, Macvanin M y Adhya S. 2012). Es un regulador global que ayuda en el mantenimiento de la arquitectura del ADN, introduciendo giros en forma de U (Swinger KK y Rice PA. 2004, Prieto AI, <i>et al.</i> 2011). |
| Fur | 129 | 101 (89.3%) | Regula la transcripción de genes involucrados en la captación de hierro (Chen Z, <i>et al.</i> 2007). |
| BolA | 2 | 100 (88.4%) | Está involucrado en la morfogénesis de las bacterias, haciéndolas más pequeñas y redondas para conferir protección bajo condiciones de estrés (Freire P, <i>et al.</i> 2009). |
| Fis | 225 | 100 (88.4%) | Modula procesos celulares tales como transcripción, replicación cromosomal, inversión de ADN y transposición de ADN (Dillon SC y Dorman, 2010, Duprey A, <i>et al.</i> 2014). |
| NrdR | 9 | 97 (85.8%) | Regula la expresión de operones que codifican para reductasas ribonucleótidas, de acuerdo con la abundancia de deoxi-ribonucleósidos trifosfatos generados desde ribonucleótidos (McKethan BL y Spiro S. 2013). |
| IscR | 32 | 93 (82.3%) | Regula la expresión de genes |

| | | | |
|------|-----|------------|--|
| | | | involucrados en una ruta de ensamble del cluster hierro – sulfuro, enzimas de respiración anaeróbica y formación de bio-películas (Wu Y y Outten FW. 2009). |
| OxyR | 33 | 92 (81.4%) | Regula genes involucrados en la respuesta a estrés oxidativo, en particular en niveles elevados de peróxido (Anjem A, <i>et al.</i> 2009). |
| CRP | 495 | 86 (76.1%) | Involucrado en la regulación de genes del catabolismo de fuentes de carbono, osmoregulación, formación de bio-películas, virulencia, asimilación de nitrógeno, toma de hierro, etc. (Zheng D, <i>et al.</i> 2004). |

Los resultados anteriores en conjunto, pueden ser interpretados dentro del contexto de la teoría de la demanda de la regulación de genes propuesta por M. Savageau (Savageau MA. 1997). Esta teoría afirma que el tipo de regulación requerido por un gen está en función de la demanda de la proteína que codifica durante el ciclo de vida de una bacteria. De tal forma que si una proteína es requerida constantemente, el gen que la codifica debería ser regulado positivamente. En el caso contrario, si una proteína es requerida sólo esporádicamente, el gen que la codifica debería ser regulado negativamente (Savageau MA.1997). En este contexto, la conservación de un tipo particular de regulación puede ser elegida para optimizar el uso del regulador (Savageau MA.1998). Por lo tanto, debe existir un balance entre el costo y el beneficio a nivel de la adecuación biológica (*fitness*) para conservar un conjunto de genes

activados o reprimidos. Con los resultados obtenidos, se puede considerar que en las bacterias endosimbiontes este balance podría ser influenciado por el medio o condiciones en las que viven dichos organismos, los cuales son relativamente estables, conduciendo a la pérdida diferencial de factores de transcripción.

4.4 MODELO CONCEPTUAL DE FACTORES DE TRANSCRIPCIÓN CONSERVADOS EN GENOMAS REDUCIDOS

Las bacterias simbiotes exhiben cambios genómicos asociados al proceso de reducción del número de genes, hecho que sin duda está asociado a que los organismos simbiotes van adquiriendo paulatinamente un estilo de vida restrictivo pero constante en el interior de otro organismo (Moran NA y Bennett GM. 2014). Un resultado interesante de este y otros estudios, es que hay una retención mayor de genes no reguladores en comparación con los genes reguladores en estos organismos. Adicionalmente, como se mencionó en las secciones anteriores, existe una tendencia para retener una mayor proporción de reguladores duales que de activadores y represores, en los genomas que presentan mayor reducción del número de genes.

Considerando todos estos aspectos, se propone un modelo conceptual para la evolución de las redes de regulación transcripcional asociado a la reducción genómica (Figura 12). Para ello, primero hay que considerar que para el estudio es conveniente delimitar al menos tres niveles de reducción genómica y que pueden ser definidos de acuerdo a su contenido de genes (McCutcheon JP y Moran NA. 2012). El estado temprano de reducción de una red de regulación es representado por organismos simbiotes recién restringidos a un huésped (como son las bacterias *S. glossinidius* y *Serratia symbiotica*). Éstos evolucionaron de organismos de vida libre por medio de la pérdida de genes. Los genomas en este nivel de reducción, son caracterizados por conservar una variedad de factores de transcripción, lo cual les permite responder a condiciones del medio o del huésped

de una manera similar a lo que ocurre en los organismos de vida libre, aunque quizás de una forma más limitada.

En la etapa intermedia de reducción genómica (~1000 a ~300 Kpb), se encuentran endosimbiontes obligados que han evidenciado reducción genómica avanzada (por ejemplo: *Baumannia cicadellinicola* y *Buchnera aphidicola*). Éstos residen en células especializadas de huéspedes como los insectos. Estas bacterias conservan una mayor proporción de NAPs con respecto a su número de reguladores totales. Por lo tanto, en dichos organismos, la regulación genética puede operar principalmente a nivel de regulación global por medio de la reestructuración del nucleoide. En esta misma etapa, se observó la pérdida de todos los genes regulados de un regulón, conservando solamente el factor de transcripción. Esto ocurre cuando el factor de transcripción realiza más de una función celular. Por ejemplo: BolA que puede realizar funciones reguladoras y catalíticas, actuando como una sistema enzimático multifuncional en organismos que carecen de sensores de dos componentes (Briza L, *et al.* 2013) y HU que además de estar involucrado en funciones reguladoras actúa como una proteína de organización del ADN (Dillon SC y Dorman CJ. 2010).

En el estado más extremo de reducción genómica, se encuentra a los endosimbiontes diminutos como *Candidatus Portiera aleyrodidarum* y *Candidatus Carsonella ruddii*. Estos organismos no conservan factores de transcripción conocidos. Además, dichas bacterias están en el límite de lo que se puede considerar organismos, para parecer ya más organelos. Por lo cual, si la regulación opera en estos organismos muy probablemente sólo sea a nivel de la topología intrínseca del ADN o por señales fisiológicas que actúan directamente sobre la ARN polimerasa (Scott M y Hwa T. 2011).

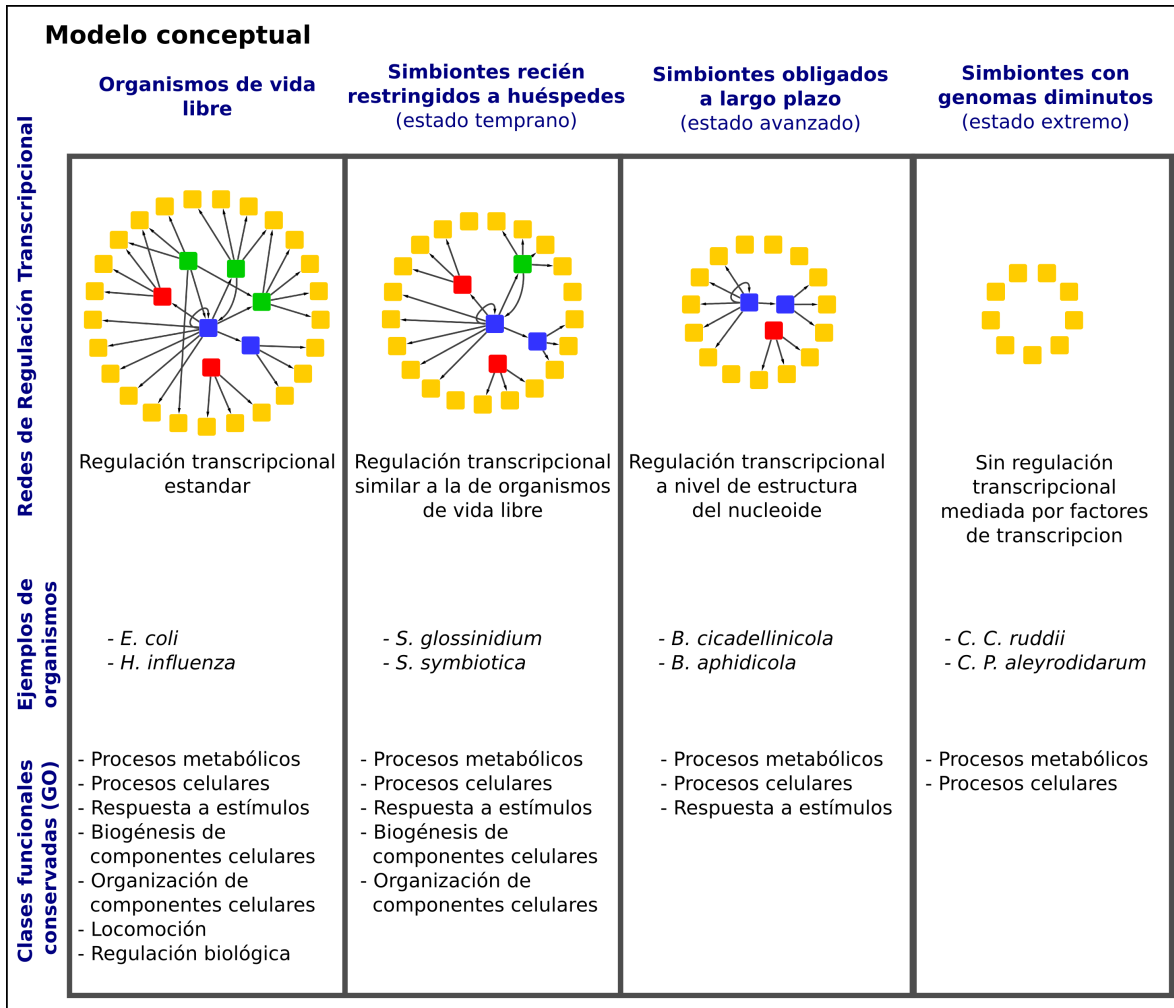


Figura 12. Modelo de evolución de las redes de regulación transcripcional bajo la influencia de reducción genómica.

Se utilizó la clasificación de grupos de bacterias con respecto a su reducción genómica propuesta por el grupo de N. Moran (McCutcheon JP y Moran NA. 2012). Se inicia con una red de regulación como la existente en *E. coli*; es decir, compuesta por activadores, represores y duales, típica de organismos de vida libre. En el primer nivel de reducción genómica, los simbiontes empiezan a ser restringidos a un huésped, estos organismos pierden rápidamente factores de transcripción y otros genes no reguladores; sin embargo el proceso de regulación transcripcional aún es similar al de organismos de vida libre. En el segundo estado de reducción genómica, los organismos se vuelven simbiontes obligados. En este estado la mayoría de los factores de transcripción son perdidos: se propone que en estas bacterias, la regulación transcripcional puede ser mediada sólo a nivel de estructura del ADN, por medio de la reestructuración de proteínas asociadas al nucleóide. En el último estado, los organismos exhiben una reducción genómica extrema

y ya no conservan factores de transcripción, lo que sugiere que la regulación al inicio de la transcripción es ausente, al menos la mediada por factores de transcripción.

Finalmente, se propone que la reducción genómica está implícitamente asociada con la evolución de la red de regulación. Esto puede ser por medio de la pérdida de fragmentos de ADN, los cuales conducen a la eliminación de nodos e interacciones dentro de la red de regulación. Se sugieren tres escenarios de la manera en que se da la eliminación de nodos: En el primero, los fragmentos de ADN eliminados contienen uno o varios factores de transcripción, lo cual lleva a la pérdida de los reguladores y a la desregulación de otros genes que estaban bajo la regulación de los factores de transcripción perdidos y que se conservan en el organismo, éste es el escenario más probable, debido a que existe una retención mayor de genes no reguladores en comparación con los factores de transcripción en estos organismos (Figura 13-A). En el segundo escenario, los fragmentos de ADN eliminados contienen uno o varios genes regulados pero no reguladores, de tal forma que el regulón remanente del factor transcripción tiende a disminuir en tamaño (Figura 13-B). En el tercer escenario, los fragmentos de ADN eliminados contienen uno o varios factores de transcripción y genes regulados que contribuyen a la disminución de las redes de regulación transcripcional (Figura 13-C).

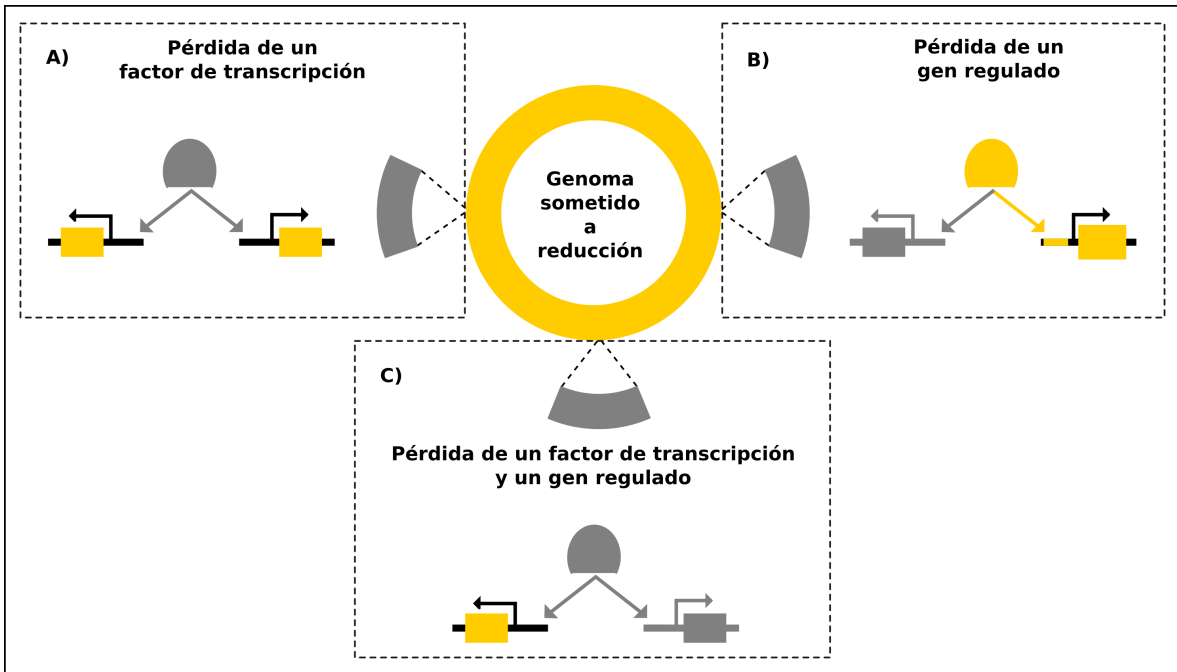


Figura 13. Efecto de la reducción genómica en las redes de regulación transcripcional.

Partiendo de un genoma sujeto a reducción, podemos observar tres escenarios dependiendo del tipo de genes contenidos en los fragmentos de ADN perdidos (representados en color gris): en el primero, el fragmento perdido contiene sólo reguladores, en el segundo contiene sólo genes regulados y en el tercero contiene a ambos tipos de genes.

5. CONCLUSIONES

Las redes de regulación pueden evolucionar desde dos perspectivas generales: la primera de ganancia de genes, que resulta en la adición de nodos e interacciones, y la segunda de pérdida, que resulta en la eliminación de nodos e interacciones. La primera, es principalmente el resultado de la duplicación y transferencia horizontal de genes, lo cual permite aumentar la capacidad adaptativa de los organismos. Por otro lado, en la perspectiva de pérdida, se ha observado que los factores de transcripción son menos conservados que los genes regulados y se predice una respuesta más limitada de los organismos. Sin embargo, no se han descrito las fuerzas que dirigen la reducción de las redes de regulación.

Por medio de genómica comparativa, en el presente trabajo se identificó que la pérdida de factores de transcripción no es generalizada sino que puede ser preferencial, de tal manera que al final los reguladores duales son los más conservados con respecto a los represores y activadores. Además, los factores de transcripción duales posiblemente son conservados también por su naturaleza de reguladores globales y principalmente por ser proteínas asociadas a nucleóide.

También se planteó un modelo conceptual que propone el proceso de reconfiguración de las redes de regulación transcripcional en cada una de las etapas de reducción genómica. De manera que, en la etapa inicial se conserva una regulación similar a la existente en los organismos de vida libre, en la etapa intermedia, la regulación es principalmente mediada sólo a nivel estructural y en la etapa de mayor reducción, la regulación transcripcional ya no es mediada por factores de transcripción.

También se propuso cómo el proceso de reducción genómica puede estar asociado con el mecanismo de reconfiguración de las redes de regulación, dando

lugar a nuevas configuraciones de las redes de regulación por medio de la pérdida de nodos e interacciones.

El estudio de la reducción de los genomas y de sus redes de regulación transcripcional es muy importante desde la perspectiva de biología sintética, debido a que facilita el diseño, construcción y funcionamiento de genomas mínimos. Estos resultados enmarcan la posibilidad de que la regulación genética puede ser posible en ausencia de factores de transcripción, de tal forma que la regulación transcripcional puede ser dada a otros niveles como por ejemplo: el nivel estructural del ADN y la fisiología de la célula, que dependen de ciertos metabolitos y señales intracelulares.

De tal manera que la regulación por medio de factores de transcripción, sólo puede representar una parte de la regulación global del sistema y posiblemente ser la regulación más dispensable de los organismos, sobretodo en los que requieren de pocos genes. En cambio, en organismos mas grandes y complejos, como los de vida libre, la regulación por factores de transcripción es necesaria para contribuir a la adaptación y competitividad de los organismos en el ambiente.

6. REFERENCIAS

Alexopoulos EC. (2010) Introduction to multivariate regression analysis. Hippokratia. 14(Suppl 1): 23-28.

Alon U. (2007) Network motifs: theory and experimental approaches. Nat Rev Genet. 8 (6): 450-461.

Aldana M, Balleza E, Kauffman S, Resendiz O. (2007) Robustness and evolvability in genetic regulatory networks. J Theor Biol. 245 (3): 433-448.

Anjem A, Varghese S, Imlay JA. (2009) Manganese import is a key element of the OxyR response to hydrogen peroxide in *Escherichia coli*. Mol Microbiol. 72 (4): 844-58.

Babu MM, Teichmann SA, Aravind L. (2006) Evolutionary dynamics of prokaryotic transcriptional regulatory networks. J Mol Biol. 358 (2): 614-633.

Babu MM, Lang B, Aravind L. (2009) Methods to reconstruct and compare transcriptional regulatory networks. Methods Mol Biol. 541: 163-180.

Babu MM. (2010) Structure, evolution and dynamics of transcriptional regulatory networks. Biochim Soc Trans. 38 (5): 115-178.

Balderas-Martínez YI, Savageau M, Salgado H, Pérez-Rueda E, Morett E, Collado-Vides J. (2013) Transcription factors in *Escherichia coli* prefer the Holo conformation. Plos One. 8 (6): 1-9.

Ballal A, Basu B, Apte SK (2007). The Kdp-ATPase system and its regulation. J Biosci. 32 (3): 559-568.

Binet MR, Poole RK. (2000) Cd(II), Pb(II) and Zn(II) ions regulate expression of the metal-transporting P-type ATPase ZntA in *Escherichia coli*. FEBS Lett. 473 (1): 67-70.

Blais A, Dynlacht BD. (2005) Constructing transcriptional regulatory networks. Genes Dev. 19 (13): 1499-1511.

Briza L, Calevro F, Charles H. (2013) Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of *Buchnera*. BMC Genomics. 14 (73): 1-15.

Browning DF, Busby SJW. (2004) The regulation of bacterial transcription initiation. *Nature Reviews Microbiology*. *Nat Rev Microbiol.* 2 (1): 1-9.

Browning DF, Grainger DC, Busby SJW. (2010) Effects of nucleoid-associated proteins on bacterial chromosome structure and gene expression. *Curr Opin Microbiol.* 13: 773-780.

Carlin A, Shi W, Dey S, Rosen BP. (1995) The ars operon of *Escherichia coli* confers arsenical and antimonial resistance. *J Bacteriol.* 177 (4): 981-986.

Chen D-G, Peace KE. (2013) *Applied meta-analysis with R*. Taylor & Francis Group. Boca Raton.

Chen Z, Lewis KA, Shultzaberger RK, Lyakhov IG, Zheng M, Doan B, Storz G, Schneider TD. (2007) Discovery of Fur binding site clusters in *Escherichia coli* by information theory models. *Nucleic Acids Res.* 35 (20): 6762-6777.

Chernick MR. (2011) *The essentials of biostatistics for physicians, nurses and clinicians*. WILEY. Hoboken, New Jersey.

Conant GC, Wagner A. (2003) Convergent evolution of gene circuits. *Nat Genet.* 34 (3): 264-266.

Cronan JE. (2001) The biotinyl domain of *Escherichia coli* acetyl-CoA carboxylase. Evidence that the "thumb" structure is essential and that the domain functions as a dimer. *J Biol Chem.* 276 (40): 37355-37364.

Cusack S, Hartlein M, Leberman R. (1991) Sequence, structural and evolutionary relationships between class 2 aminoacyl-tRNA synthetases. *Nucleic Acids Res.* 19 (13): 3489-3498.

Davis JH. (2011) *Statistics for compensation: A practical guide to compensation analysis*. 1st ed. John Wiley & Sons, Inc.

Delage L, Gil R, Pereto J, Latorre A, Moya A. (2010) Life with a few genes: A survey on naturally evolved reduced genomes. *Open Evol J.* 4: 12-22.

Dillon SC, Dorman CJ. (2010) Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol.* 8 (3): 185-195.

Donczew R, Zakrzewska-Czerwinska J, Zawilak-Pawlik A. (2014) Beyond DnaA: The role of DNA topology and DNA methylation in bacterial replication initiation. *J Mol Biol.* 426 (12): 2269-2282.

Dufresne A, Garczarek L, Partensky F. (2005) Accelerated evolution associated with genome reduction in a free-living prokaryote. *Genome Biology*. 6 (2): R14.

Duprey A, Reverchon S, Nasser W. (2014) Bacterial virulence and Fis: adapting regulatory networks to the host environment. *Cell*. 22 (2): 92-99.

Ebright RH. (2000) RNA Polymerase: Structural similarities between bacterial RNA polymerase and eukaryotic RNA polymerase II. *J Mol Biol*. 304: 687-698.

Eisenberg MA, Prakash O, Hsiung SC. (1982) Purification and properties of the biotin repressor. A bifunctional protein. *J Biol Chem*. 257 (24): 15167-15173.

Felsenstein J. (1985) Phylogenies and the comparative method. *Amer Nat*. 125: 1-15.

Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R. (2000) Identification of additional genes belonging to the LexA regulon in *Escherichia coli*. *Mol Microbiol* 35 (6): 1560-1572.

Fisher RA (1932) *Statistical methods for research workers*. 4 th edition. London: Oliver and Boyd.

Franke S, Grass G, Nies DH. (2001) The product of the ybdE gene of the *Escherichia coli* chromosome is involved in detoxification of silver ions. *Microbiology*. 147 (Pt 4): 965-972.

Freire P, Moreira RN, Arraiano CM. (2009) BolA inhibits cell elongation and regulates MreB expression levels. *J Mol Biol*. 385 (5): 1345-1351.

Galán-Vásquez E, Luna B, Martínez-Antonio A. (2011) The regulatory network of *Pseudomonas aeruginosa*. *Microb Inform Exp*. 1 (1): 1-11.

Gardner TS, di Bernado D, Lorenz D, Collin JJ. (2003) Inferring genetic networks and identifying compound mode of action via expression profiling. *Science*. 301 (5629): 102-105.

Garland T, Harvey PH, Ives AR. (1992) Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst Biol*. 41: 18-32.

Garland T, Midford PE, Ives AR. (1999) An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *Amer Zool*. 39: 374-388.

Garland T, Ives AR. (2000) Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Amer Zool.* 155 (3): 346-364.

Gerosa L, Kochanowski K, Heinemann M, Sauer U. (2013) Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol Syst Biol.* 9 (658): 1-11.

Giel JL, Rodionov D, Liu M, Blattner FR, Kiley PJ. (2006) IscR-dependent gene expression links iron-sulphur cluster assembly to the control of O-regulated genes in *Escherichia coli*. *Mol Microbiol.* 60 (4): 1058-1075.

Holder M, Lewis PO. (2003) Phylogeny estimation: traditional and bayesian approaches. *Nat Rev Genet.* 4 (4): 275-284.

Huynen MA, Spronk CA, Gabaldon T, Snel B. (2005) Combining data from genomes, Y2H and 3D structure indicates that BolA is a reductase interacting with a glutaredoxin. *FEBS Lett.* 579 (3): 591-596.

Imam S, Noguera DR, Donohue TJ. (2015) An integrated approach to reconstructing genome-scale transcriptional regulatory networks. *Plos Computational Biology.* 11 (2): e10041103.

Isalan M, Lemerle C, Michalodimitrakis K, Horn C, Beltrao P, Raineri E, Garriga-Canut M, Serrano L. (2008) Evolvability and hierarchy in rewired bacterial gene networks. *Nature.* 452 (7189): 840-845.

Ivanov I, Dougherty ER. (2006) Modeling genetic regulatory networks: continuous or discrete. *J Biol Syst.* 14 (2): 219.

Janga SC, Salgado H, Martínez-Antonio A. (2009) Transcriptional regulation shapes the organization of genes on bacterial chromosomes. *Nucleic Acids Res.* 37 (11): 3680-3688.

Junker HB, Schreiber F. (2008) Analysis of biological networks. 1st ed. Hoboken. New Jersey: John Wiley & Sons, Inc. pp. 368.

Karlebach G, Shamir R. (2008) Modelling and analysis of gene regulatory networks. *Nat Rev Mol Cell Biol.* 9: 770-780.

Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, Fulcher C, Huerta AM, Kothari A, Krummenacker

M, Latendresse M, Muñiz-Rascado L, Ong Q, Paley S, Schröder I, Shearer AG, Subhraveti P, Travers M, Weerasinghe D, Weiss V, Collado-Vides J, Gunsalus RP, Paulsen I, Karp PD. (2013) EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 41: D605–D612.

Klumpp S, Zhang Z, Hwa T. (2009) Growth rate-dependent global effects on gene expression in bacteria. *Cell.* 139 (7): 1366-1375.

Kolb A, Busby S, Buc H, Garges s, Adhya S. (1993) Transcriptional regulation by cAMP and its receptor protein. *Annu Rev Biochem.* 62: 749-795.

Koonin EV, Makarova KS, Aravind L. (2001) Horizontal gene transfer in prokaryotes: quantification and classification. *Annu Rev Microbiol.* 55: 709-742.

Lagomarsino MC, Bassetti B, Isambert H. (2007) Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proc Natl Acad Sci USA.* 104 (13): 5516-5520.

Kraxenberger T, Fried L, Behr S, Jung K. (2012) First insights into the unexplored two-component system YehU/YehT in *Escherichia coli*. *J Bacteriol.* 194 (16): 4272-4284.

Kuper C, Jung K. (2005) CadC-mediated activation of the cadBA promoter in *Escherichia coli*. *J Mol Microbiol Biotechnol* 10(1);26-39.

Lima-Mendez G, van Helden J. (2009) The powerful law of the power law and other myths in network biology. *Mol Biosyst.* 5 (12): 1482-1493.

Lozada-Chávez I, Janga SC, Collado-Vides J. (2006) Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 34 (12): 3434-3445.

Lynch M. (2007) The evolution of genetic networks by non-adaptive processes. *Nat Rev Gen.* 8 (10): 803-813.

Ma D, Alberti M, Lynch C, Nikaido H, Hearst JE. (1996) The local repressor AcrR plays a modulating role in the regulation of acrAB genes of *Escherichia coli* by global stress signals. *Mol Microbiol.* 19 (1): 101-112.

Macvanin M, Adhya S. (2012) Architectural organization in *E. coli* nucleoid. *Biochim Biophys Acta.* 1819 (7): 830-835.

Madar D, Dekel E, Bren A, Alon U. (2011) Negative auto-regulation increases the input dynamic-range of the arabinose system of *Escherichia coli*. BMC Syst Biol. 5 (111): 1-9.

Maddison WP, Maddison DR. (2015) Mesquite: a modular system for evolutionary analysis (Version 3.02). Available: <http://mesquiteproject.org>.

Marais GAB., Calteau A., Tenaillon O. (2008) Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica. 134: 205-210.

Marrakchi H, Zhang YM, Rock CO. (2002) Mechanistic diversity and regulation of Type II fatty acid synthesis. Biochem Soc Trans. 30 (Pt 6): 1050-1055.

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

Martínez-Antonio A, Janga SC, Salgado H, Collado-Vides J. (2006) Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. Trends Microbiol. 14 (1): 22-27.

Martínez-Cano DJ, Reyes-Prieto M, Martínez-Romero E, Partida-Martínez LP, Latorre A, Moya A, Delaye L. (2015) Evolution of small prokaryotic genomes. Front Microbiol. 5 (742): 1-23.

McAdams HH, Srinivasan B, Arkin AP. (2004) The evolution of genetic regulatory systems in bacteria. Nat Rev Genet. 5 (3): 169-178.

McCutcheon JP, Moran NA. (2012) Extreme genome reduction in symbiotic bacteria. Nat Rev Microbiol. 10 (1): 13-26.

McKethan BL, Spiro S. (2013) Cooperative and allosterically controlled nucleotide binding regulates the DNA binding activity of NrdR. Mol Microbiol. 90 (2): 278-89.

McLeod SM, Johnson RC. (2001) Control of transcription by nucleoid proteins. Curr Opin Microbiol. 4: 152-159.

Minh PN, Devroede N, Massant J, Maes D, Charlier D. (2009). Insights into the architecture and stoichiometry of *Escherichia coli* PepA*DNA complexes involved in transcriptional control and site-specific DNA recombination by atomic force microscopy. Nucleic Acids Res. 37 (5): 1463-1476.

Mira A., Ochman H., Moran NA. (2001) Deletion bias and the evolution of bacterial genomes. *TRENDS in Genetics*. 17 (10): 589-596.

Moran NA, Dunbar HE, Wilcox JL. (2005) Regulation of transcription in a reduced bacterial genome: nutrient-provisioning genes of the obligate symbionts *Buchnera aphidicola*. *J Bacteriol*. 187 (12): 4229-4237.

Moran NA, Bennett GM. (2014) The tiniest tiny genomes. *Annu Rev Microbiol*. 68: 195-215.

Moreno-Hagelsieb G, Latimer K. (2008) Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics*. 24 (3): 319-324.

Morris JJ., Lenski RE., Zinser ER. (2012) The black queen hypothesis: Evolution of dependencies through adaptive gene loss. *mBio*. 3 (2): 1-7.

Murakami KS, Masuda S, Campbell EA, Muzzin O, Darst SA. (2002) Structural basis of transcription initiation: An RNA polymerase holoenzyme-DNA complex. *Science*. 296 (5571): 1285-1290.

NCBI ftp server [<http://www.ncbi.nlm.nih.gov/FTP/>]

Ogasawara H, Hasegawa A, Kanda E, Miki T, Yamamoto K, Ishihama A. (2007) Genomic SELEX search for target promoters under the control of the PhoQP-RstBA signal relay cascade. *J Bacteriol*. 189 (13): 4791-4799.

Panina EM, Mironov AA, Gelfand MS. (2003) Comparative genomics of bacterial zinc regulons: enhanced ion transport, pathogenesis, and rearrangement of ribosomal proteins. *Proc Natl Acad Sci U S A*. 100 (17): 9912-9917.

Pérez-Rueda E, Collado-Vides J. (2000) The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res*. 28 (8): 1838-1847.

Price MN, Dehal PS, Arkin AP. (2007) Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLoS Comput Biol*. 3 (9): 1739-1750.

Prieto AI, Kahramannoglou C, Ali RM, Fraser GM, Seshasayee ASN, Luscombe M. (2011) Genomic analysis of DNA binding and gene regulation by homologous nucleoid-associated protein IHF and HU *Escherichia coli* K12. *Nucleic Acids Res*. 40 (8): 3524-3537.

Ptashne M. (2005) Regulation of transcription: from lambda to eukaryotes. *Trends Biochem Sci.* 30 (6): 275-279.

Qian J, Lin J, Luscombe NM, Yu H, Gerstein M. (2003) Prediction of regulatory network: Genome-wide identification of transcription factor targets from gene expression data. *Bioinformatics.* 19 (15): 1917-1927.

Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, Weiss V, Solano-Lira H, Martínez-Flores I, Medina-Reviera A, Salgado-Osorio G, Alquicira-Hernández S, Alquicira-Hernández K, López-Fuentes A, Porrón-Sotelo L, Huerta AM, Bonavides-Matínez C, Balderas-Martínez YI, Pannier L, Olvera M, Labastida A, Jiménez-Jacinto V, Vega-Alvarado L, Moral-Chávez V, Hernández-Alvarez A, Morett E, Collado-Vides J. (2013) RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phases, cross-validated gold standards and more. *Nucleic Acids Res.* 41: D203–D213.

Savageau MA. (1977) Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci USA.* 74 (12): 5647-5651.

Savageau MA. (1998) Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of *Escherichia coli*. *Genetics.* 149 (4): 1677-1691.

Scott M, Hwa T. (2011) Bacterial growth laws and their applications. *Curr Opin Biotechnol.* 22 (4): 559-565.

M, Lesterlin C, Garza de Leon F, Uphoff S, Zawadzki P, Kapanidis AN. (2015) Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proc Natl Acad Sci U.S.A.* 112 (32): E4390-E4399.

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. (2003) Cytoscape: A software environment for integrated model of biomolecular interaction networks. *Genome Res.* 13: 2498-2504.

Sperandio V, Torres AG, Kaper JB. (2002) Quorum sensing *Escherichia coli* regulators B and C (QseBC): a novel two-component regulatory system involved in the regulation of flagella and motility by quorum sensing in *E. coli*. *Mol Microbiol* 43 (3): 809-821.

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

Swinger KK, Rice P. (2004) IHF and HU: flexible architects of bent DNA. *Curr Opin Struct Biol*. 14 (1): 28-35.

R Development core team, 2008, <http://www.r-project.org/>.

Rankin LD, Bodenmiller DM, Partridge JD, Nishino SF, Spain JC, Spiro S. (2008) *Escherichia coli* NsrR regulates a pathway for the oxidation of 3-nitrotyramine to 4-hydroxy-3-nitrophenylacetate. *J Bacteriol*. 190 (18): 6170-7.

Reyes-Prieto M, Latorre A, Moya A. (2014) Scanty microbes, the “symbionelle” concept. *Environmental Microbiology*. 16 (2): 335-338.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, Larget B, Liu L, Suchard MA, Huelsenbeck JP. (2012) MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Sys Biol*. 61 (3): 539-542.

Teichmann SA, Babu MM. (2004) Gene regulatory network growth by duplication. *Nat Genet*. 36 (5): 492-496.

Thony B, Hwang DS, Fradkin L, Kornberg A. (1991) *iciA*, an *Escherichia coli* gene encoding a specific inhibitor of chromosomal initiation of replication in vitro. *Proc Natl Acad Sci U S A*. 88 (10): 4066-4070.

Turner PC, Miller EN, Jarboe LR, Baggett CL, Shanmugam KT, Ingram LO. (2011) YqhC regulates transcription of the adjacent *Escherichia coli* genes *yqhD* and *dkgA* that are involved in furfural tolerance. *J Ind Microbiol Biotechnol*. 38 (3): 431-439.

Vinod MP, Bellur P, Becker DF. (2002) Electrochemical and functional characterization of the proline dehydrogenase domain of the PutA flavoprotein from *Escherichia coli*. *Biochemistry*. 41 (20): 6525-6532.

Wang T, Stormo GD. (2005) Identifying the conserved network of cis-regulatory sites of a eukaryotic genome. *Proc Natl Acad Sci USA*. 102 (48): 17400-17405.

Weng X, Xiao J. (2014) Spatial organization of transcription in bacterial cells. *Cell*. 30 (7): 287-297.

Wing HJ, Williams SM, Busby SJ. (1995) Spacing requirements for transcription activation by *Escherichia coli* FNR protein. *J Bacteriol.* 177 (23): 6704-6710.

Wittkopp PJ, Kalay G. (2011) Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 13 (1): 59-69.

Wösten MMSM. (1998) Eubacterial sigma-factors. *FEMS Microbiol Rev.* 22 (3): 127-50.

Wu Y, Outten FW. (2009) IscR controls iron-dependent biofilm formation in *Escherichia coli* by regulating type I fimbria expression. *J Bacteriol.* 191 (4): 1248-1257.

Wyborn NR, Stapleton MR, Norte VA, Roberts RE, Grafton J, Green J. (2004) Regulation of *Escherichia coli* hemolysin E expression by H-NS and Salmonella SlyA. *J Bacteriol.* 186 (6): 1620-1628.

Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, Bertin N, Chung S, Vidal M, Gerstein M. (2004) Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* 14 (6): 1107-1118.

Zhang J. (2003) Evolution by gene duplication: an update. *Trends Ecol. Evol.* 18 (6): 292-298.

Zheng D, Constantinidou C, Hobman JL, Minchin SD. (2004) Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Res.* 32 (19): 5874-5893.

7. ANEXOS

7.1 ARTÍCULO 1

Description of gene regulatory networks in three bacterial models: *Escherichia coli*, *Bacillus subtilis* and *Pseudomonas aeruginosa*

Edgardo Galán-Vásquez, Beatriz Luna, Agustino Martínez-Antonio. Recent Res Devel Microbiology. 2012. 12: 19-39.

Aportes del trabajo

En este artículo se describen las principales métricas obtenidas en la caracterización topológica de las redes de regulación transcripcional en las redes de regulación de *E. coli*, *B. subtilis* y *P. aeruginosa*: la distribución de grados, el coeficiente de clusterización, conectividad, autoregulaciones, caminos, circuitos, motivos, y reguladores. El estudio se realizó, con el objeto de determinar la similitud entre estas redes de regulación. Todas las redes se ajustan a una estructura de escala libre, la activación es el tipo de regulación predominante en todas las redes, además de que la autorepresión es dominante en *E. coli* y *B. subtilis*, pero no en *P. aeruginosa*. A pesar de que las redes de regulación de cada bacteria difieren en cantidad de nodos e interacciones, la mayoría de las propiedades topológicas permanecen constantes.

7.2 ARTÍCULO 2

Structural comparison of biological networks based on dominant vertices

Beatriz Luna , **Edgardo Galán-Vásquez**, Agustino Martínez-Antonio. *Molecular BioSystems*. 2013. 9: 1765-1773.

Aportes del trabajo

La teoría de grafos ha permitido estudiar la estructura de las redes, en este contexto se han propuesto diferentes medidas como son: el coeficiente de clusterización, la centralidad de un nodo, la distribución de grados, los motivos de la red, nodos maestros, etc. Sin embargo, ninguna de estas métricas permite determinar la similitud global entre redes de regulación de diferentes tamaños, por tal motivo el objetivo de este trabajo fue introducir una pseudo-distancia que permitiera esta comparación entre dos redes de regulación, para ello se utilizó el concepto de vértices dominantes, con lo cual se compararon las redes de regulación de *E. coli*, *B. subtilis*, *P. aeruginosa*, *M. tuberculosis*, *S. aureus* y *C. glutamicum*. Además se realizó una comparación con redes aleatorias. Se obtuvo que las redes de regulación en dichas bacterias son caracterizadas a diferentes niveles de detalle, que tienen diferente tamaño y son descritos en distintos aspectos biológicos en cada organismo. A pesar de esto, las características generales de las redes son recuperadas con la pseudo-métrica propuesta, permitiendo diferenciar las redes reales de las redes aleatorias.

7.3 ARTÍCULO 3

Regulatory switches for hierarchical use of carbon sources in *E. coli*

Ruth S. Pérez-Alfaro, Moisés Santillán, **Edgardo Galán-Vásquez**, Agustino Martínez-Antonio. 2014. Network Biology. 4 (3): 95-108.

Aportes del trabajo

Las bacterias están sometidas a diversos cambios del medio en el que habitan, una forma de responder rápidamente a dichos cambios es proporcionada por la red de regulación transcripcional. Permitiendo al organismo percibir las condiciones externas e internas de la célula y de esta forma activar o reprimir circuitos reguladores que dan una respuesta adecuada a las condiciones ambientales. En este contexto, se estudiaron *switches* reguladores en la bacteria *E. coli* involucrados en el uso de 3 fuentes de carbono diferentes a glucosa. Por medio del uso de fusiones transcripcionales con el gen reportero *gfpmut2* y los promotores reconocidos por los factores de transcripción que integran los circuitos reguladores. Se probó el orden preferencial de uso de los cuatro azúcares, en el cual se obtuvo que existe una preferencia jerárquica en el uso de cada fuente de carbono, de tal forma que la bacteria *E. coli* prefiere utilizar dichos azúcares en el siguiente orden: glucosa > arabinosa > sorbitol > galactosa.

7.4 ARTÍCULO 4

Transcription factors exhibit differential conservation in bacteria with reduced genomes

Edgardo Galán-Vásquez, Ismael Sánchez-Osorio, agustino Martínez-Antonio.
Plos One. 2016. 11 (1): e0146901. doi:10.1371/journal.pone.0146901

Aportes del Trabajo

La descripción de las redes de regulación han ayudado a la comprensión de los principios bajo los cuales responden y se adaptan los organismos. Mientras el estudio topológico y dinámico de estas redes ha sido objeto de diversos estudios, la investigación de su topología como resultado de la adaptación de los organismos ha diferentes condiciones ambientales, tiene poca atención. En este trabajo se estudió la evolución de las redes de regulación transcripcional de bacterias desde una perspectiva de reducción genómica y se encontró que existe una conservación preponderante de factores de transcripción dependiendo de su actividad regulatoria, además, se planteó un modelo conceptual de la regulación transcripcional en los diferentes niveles de reducción genómica.

7.5 MATERIAL SUPLEMENTARIO

Tabla S1. Factores de transcripción identificados en las redes de regulación reconstruidas.

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|--|---|-------------|------------|--------|--------------|
| <i>Escherichia coli</i> K 12 substr MG1655 | 196 | 48 | 65 | 77 | 6 |
| <i>Yersinia</i> <i>enterocolitica</i> <i>paleartica</i> 105 5R r | 124 | 24 | 42 | 53 | 5 |
| <i>Shewanella loihica</i> PV 4 | 75 | 12 | 20 | 40 | 3 |
| <i>Alteromonas</i> <i>macleodii</i> English str Channel 673 | 72 | 10 | 19 | 41 | 2 |
| <i>Alteromonas</i> <i>macleodii</i> str English Channel 615 | 71 | 13 | 21 | 34 | 3 |
| <i>Escherichia coli</i> BW2952 | 183 | 46 | 60 | 71 | 6 |
| <i>Cellvibrio japonicus</i> Ueda107 | 57 | 9 | 18 | 28 | 2 |
| <i>Stenotrophomonas</i> <i>maltophilia</i> R551 3 | 73 | 14 | 23 | 33 | 3 |
| <i>Pseudomonas</i> <i>stutzeri</i> A1501 | 71 | 14 | 19 | 35 | 3 |
| <i>Psychromonas</i> <i>ingrahamii</i> 37 | 64 | 5 | 25 | 33 | 1 |
| <i>Aeromonas veronii</i> B565 | 95 | 18 | 29 | 46 | 2 |
| <i>Pseudomonas</i> <i>stutzeri</i> ATCC 17588 LMG 11199 | 71 | 13 | 19 | 37 | 2 |
| <i>Shewanella</i> <i>denitrificans</i> OS217 | 69 | 10 | 20 | 37 | 2 |
| <i>Stenotrophomonas</i> <i>maltophilia</i> JV3 | 73 | 14 | 22 | 35 | 2 |
| <i>Cronobacter</i> <i>sakazakii</i> SP291 | 133 | 26 | 45 | 56 | 6 |
| <i>Alteromonas</i> <i>macleodii</i> str Deep | 79 | 9 | 26 | 41 | 3 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|--|---|-------------|------------|--------|--------------|
| <i>ecotype</i> | | | | | |
| <i>Alteromonas macleodii str Black Sea 11</i> | 73 | 11 | 19 | 42 | 1 |
| <i>Salmonella bongori NCTC 12419</i> | 143 | 32 | 42 | 66 | 3 |
| <i>Alteromonas macleodii str Ionian Sea U7</i> | 74 | 10 | 23 | 38 | 3 |
| <i>Providencia stuartii MRSN 2154</i> | 103 | 14 | 37 | 50 | 2 |
| <i>Shewanella amazonensis SB2B</i> | 77 | 13 | 21 | 40 | 3 |
| <i>Simiduia agarivorans SA1 DSM 21679</i> | 61 | 8 | 19 | 30 | 4 |
| <i>Ferrimonas balearica DSM 9799</i> | 68 | 12 | 19 | 35 | 2 |
| <i>Vibrio fischeri ES114</i> | 89 | 13 | 29 | 45 | 2 |
| <i>Cronobacter sakazakii ES15</i> | 128 | 25 | 41 | 57 | 5 |
| <i>Vibrio cholerae MJ 1236</i> | 95 | 17 | 30 | 44 | 4 |
| <i>Acinetobacter baumannii TCDC AB0715</i> | 70 | 11 | 26 | 31 | 2 |
| <i>Pseudomonas stutzeri DSM 10701</i> | 76 | 14 | 21 | 37 | 4 |
| <i>Shimwellia blattae DSM 4481 NBRC 105725</i> | 129 | 28 | 38 | 59 | 4 |
| <i>Acinetobacter oleivorans DR1</i> | 73 | 14 | 27 | 28 | 4 |
| <i>Legionella longbeachae NSW150</i> | 36 | 5 | 10 | 19 | 2 |
| <i>Nitrosococcus halophilus Nc 4</i> | 37 | 8 | 14 | 14 | 1 |
| <i>Thioflavococcus mobilis 8321</i> | 42 | 8 | 11 | 22 | 1 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|---|---|-------------|------------|--------|--------------|
| <i>Glaciecola nitratreducens FR1064</i> | 59 | 9 | 16 | 33 | 1 |
| <i>Listonella anguillarum M3</i> | 93 | 18 | 26 | 45 | 4 |
| <i>Proteus mirabilis HI4320</i> | 101 | 17 | 32 | 49 | 3 |
| <i>Erwinia pyrifoliae Ep1 96</i> | 98 | 16 | 29 | 49 | 4 |
| <i>Erwinia tasmaniensis Et1 99</i> | 99 | 17 | 29 | 51 | 2 |
| <i>Marinobacter sp BSs20148</i> | 62 | 10 | 19 | 31 | 2 |
| <i>Halomonas elongata DSM 2581</i> | 69 | 11 | 23 | 33 | 2 |
| <i>Vibrio anguillarum 775</i> | 93 | 18 | 27 | 44 | 4 |
| <i>Vibrio cholerae O1 biovar El Tor str N16961</i> | 93 | 17 | 29 | 44 | 3 |
| <i>Thioalkalivibrio nitratreducens DSM 14787</i> | 45 | 6 | 16 | 20 | 3 |
| <i>Marinobacter hydrocarbonoclastic us ATCC 49840</i> | 64 | 15 | 15 | 31 | 3 |
| <i>Thalassolituus oleivorans MIL 1</i> | 60 | 13 | 18 | 27 | 2 |
| <i>Marinomonas posidonica IVIA Po 181</i> | 67 | 9 | 27 | 29 | 2 |
| <i>Acinetobacter calcoaceticus PHEA 2</i> | 67 | 10 | 24 | 31 | 2 |
| <i>Xanthomonas albilineans GPE PC73</i> | 45 | 9 | 12 | 23 | 1 |
| <i>Pseudoalteromonas haloplanktis TAC125</i> | 71 | 11 | 21 | 37 | 2 |
| <i>Proteus mirabilis</i> | 99 | 19 | 29 | 48 | 3 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|---|---|-------------|------------|--------|--------------|
| <i>BB2000</i> | | | | | |
| <i>Erwinia amylovora</i> <i>CFBP1430</i> | 98 | 17 | 29 | 49 | 3 |
| <i>Edwardsiella tarda</i> <i>EIB202</i> | 111 | 28 | 30 | 51 | 2 |
| <i>Morganella</i> <i>morganii subsp</i> <i>morganii KT</i> | 104 | 18 | 34 | 50 | 2 |
| <i>Vibrio cholerae</i> <i>LMA3984 4</i> | 83 | 15 | 24 | 42 | 2 |
| <i>Chromohalobacter</i> <i>salexigens DSM</i> <i>3043</i> | 68 | 11 | 22 | 33 | 2 |
| <i>Allochromatium</i> <i>vinosum DSM 180</i> | 43 | 8 | 15 | 18 | 2 |
| <i>Frateuria aurantia</i> <i>DSM 6220</i> | 65 | 11 | 23 | 28 | 3 |
| <i>Nitrosococcus</i> <i>oceanii ATCC 19707</i> | 36 | 6 | 11 | 17 | 2 |
| <i>Legionella</i> <i>pneumophila subsp</i> <i>pneumophila</i> | 34 | 6 | 9 | 18 | 1 |
| <i>Tolomonas auensis</i> <i>DSM 9187</i> | 85 | 12 | 27 | 42 | 4 |
| <i>Thioalkalivibrio</i> <i>sulfidophilus HL</i> <i>EbGr7</i> | 46 | 7 | 16 | 21 | 2 |
| <i>Pseudoxanthomona</i> <i>s spadix BD a59</i> | 53 | 9 | 21 | 21 | 2 |
| <i>Pseudoxanthomona</i> <i>s suwonensis 11 1</i> | 42 | 4 | 13 | 24 | 1 |
| <i>Nitrosococcus</i> <i>watsonii C 113</i> | 36 | 6 | 14 | 15 | 1 |
| <i>Methylococcus</i> <i>capsulatus Bath</i> | 33 | 5 | 10 | 16 | 2 |
| <i>Alkalilimnicola</i> <i>ehrllichii MLHE 1</i> | 44 | 2 | 20 | 20 | 2 |
| <i>Acidithiobacillus</i> <i>calvus SM 1</i> | 42 | 5 | 20 | 15 | 2 |
| <i>Acidithiobacillus</i> | 44 | 8 | 15 | 18 | 3 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|---|---|-------------|------------|--------|--------------|
| <i>ferrivorans SS3</i> | | | | | |
| <i>Methylophaga nitratireducenticres cens</i> | 54 | 10 | 16 | 25 | 3 |
| <i>Alcanivorax borkumensis SK2</i> | 55 | 7 | 17 | 28 | 3 |
| <i>Psychrobacter cryohalolentis K5</i> | 51 | 6 | 22 | 21 | 2 |
| <i>Acidithiobacillus ferrooxidans ATCC 23270</i> | 41 | 6 | 14 | 19 | 2 |
| <i>Kangiella koreensis DSM 16069</i> | 46 | 9 | 15 | 21 | 1 |
| <i>Idiomarina loihiensis GSL 199</i> | 61 | 8 | 20 | 29 | 4 |
| <i>Methylophaga frappieri</i> | 42 | 8 | 11 | 21 | 2 |
| <i>Xylella fastidiosa 9a5c</i> | 35 | 3 | 17 | 14 | 1 |
| <i>Mannheimia haemolytica D174</i> | 60 | 6 | 21 | 32 | 1 |
| <i>Gallibacterium anatis UMN179</i> | 62 | 7 | 26 | 27 | 2 |
| <i>Halorhodospira halophila SL1</i> | 33 | 3 | 15 | 14 | 1 |
| <i>Cycloclasticus zancles 7 ME</i> | 35 | 6 | 11 | 16 | 2 |
| <i>Psychrobacter arcticus 273 4</i> | 40 | 6 | 13 | 18 | 3 |
| <i>Halothiobacillus neapolitanus c2</i> | 43 | 8 | 12 | 21 | 2 |
| <i>Thiomicrospira crunogena XCL 2</i> | 31 | 4 | 10 | 16 | 1 |
| <i>Bibersteinia trehalosi USDA ARS USMARC 192</i> | 55 | 5 | 21 | 27 | 2 |
| <i>Pasteurella multocida 36950</i> | 57 | 4 | 20 | 32 | 1 |
| <i>Haemophilus parasuis ZJ0906</i> | 53 | 4 | 20 | 26 | 3 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|---|---|-------------|------------|--------|--------------|
| <i>Actinobacillus succinogenes</i> 130Z | 60 | 7 | 21 | 30 | 2 |
| <i>Mannheimia succiniciproducens</i> MBEL55E | 63 | 8 | 21 | 32 | 2 |
| <i>Aggregatibacter actinomycetemcomitans</i> D7S 1 | 57 | 2 | 25 | 28 | 2 |
| <i>Haemophilus somnus</i> 2336 | 54 | 4 | 23 | 26 | 1 |
| <i>Actinobacillus pleuropneumoniae</i> serovar 3 str JL03 | 59 | 6 | 24 | 28 | 1 |
| <i>Coxiella burnetii</i> Dugway 5J108 111 | 26 | 3 | 7 | 15 | 1 |
| <i>Aggregatibacter actinomycetemcomitans</i> ANH9381 | 56 | 3 | 24 | 27 | 2 |
| <i>Haemophilus influenzae</i> Rd KW20 | 49 | 4 | 15 | 28 | 2 |
| <i>Haemophilus ducreyi</i> 35000HP | 33 | 0 | 11 | 21 | 1 |
| <i>Sodalis glossinidius</i> str morsitans Candidatus | 77 | 14 | 26 | 36 | 1 |
| <i>Hamiltonella defensa</i> 5AT | 20 | 3 | 8 | 8 | 1 |
| <i>Acyrtosiphon pisum</i> | | | | | |
| <i>Francisella tularensis holarctica</i> LVS | 19 | 2 | 9 | 6 | 2 |
| <i>Francisella tularensis mediasiatica</i> FSC147 | 21 | 4 | 9 | 6 | 2 |
| <i>Francisella tularensis tularensis</i> FSC198 | 20 | 3 | 9 | 6 | 2 |
| <i>Dichelobacter nodosus</i> VCS1703A | 17 | 3 | 5 | 8 | 1 |

| Organismo | Ortólogos totales de los factores de transcripción | Activadores | Represores | Duales | Desconocidos |
|---|---|-------------|------------|--------|--------------|
| <i>Serratia symbiotica</i> <i>str Cinara cedri</i> Secondary endosymbiont of | 14 | 0 | 4 | 9 | 1 |
| <i>Ctenarytaina</i> <i>eucalypti</i> Secondary endosymbiont of | 15 | 1 | 5 | 8 | 1 |
| <i>Heteropsylla cubana</i> Candidatus | 7 | 1 | 2 | 3 | 1 |
| <i>Vesicomysocius</i> <i>okutanii</i> HA | 12 | 1 | 3 | 7 | 1 |
| <i>Wigglesworthia</i> <i>glossinidia str</i> endosymbiont of | 11 | 1 | 5 | 4 | 1 |
| <i>glossina brevipalpis</i> Baumannia | | | | | |
| <i>cicadellinicola str</i> Hc <i>Homalodisca</i> <i>coagulata</i> | 12 | 1 | 5 | 5 | 1 |
| <i>Buchnera aphidicola</i> APS <i>Acyrtosiphon</i> <i>pisum</i> | 8 | 0 | 2 | 6 | 0 |
| <i>Buchnera aphidicola</i> Sg <i>Schizaphis</i> <i>graminum</i> | 8 | 0 | 2 | 6 | 0 |
| <i>Buchnera aphidicola</i> Bp <i>Baizongia</i> <i>pistaciae</i> | 5 | 0 | 2 | 3 | 0 |
| <i>Buchnera aphidicola</i> <i>Cinara tujafilina</i> | 5 | 0 | 2 | 3 | 0 |
| <i>Buchnera aphidicola</i> BCc | 5 | 0 | 1 | 4 | 0 |
| Candidatus <i>Portiera</i> <i>aleyrodidarum</i> BT B | 0 | 0 | 0 | 0 | 0 |
| Candidatus <i>Carsonella ruddii</i> PV | 0 | 0 | 0 | 0 | 0 |

Tabla S2. Clasificación de los factores de transcripción identificados experimentalmente en la red de regulación de *E. coli* y con tipo de regulación (los 190).

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| AccB | Desconocido | 103 | 2 | Single hybrid motif |
| AcrR | Represor | 56 | 4 | Homeodomain-like |
| Ada | Dual | 42 | 4 | Homeodomain-like |
| AdiY | Activador | 11 | 8 | Homeodomain-like |
| AgaR | Represor | 15 | 11 | Winged helix DNA-binding domain |
| AidB | Represor | 44 | 1 | Acyl-CoA dehydrogenase NM domain-like |
| AlaS | Represor | 110 | 1 | ThrRS/AlaRS common domain |
| AllR | Represor | 5 | 9 | Winged helix DNA-binding domain |
| AllS | Activador | 5 | 3 | Winged helix DNA-binding domain |
| AlsR | Represor | 0 | 6 | Homeodomain-like |
| AppY | Activador | 14 | 10 | Homeodomain-like |
| AraC | Dual | 32 | 11 | Homeodomain-like |
| ArcA | Dual | 49 | 172 | CheY-like |
| ArgP | Activador | 43 | 14 | Winged helix DNA-binding domain |
| ArgR | Dual | 51 | 37 | Winged helix DNA-binding domain |
| ArsR | Represor | 69 | 3 | Winged helix DNA-binding domain |
| AscG | Represor | 18 | 5 | lambda repressor-like DNA-binding domains |
| AsnC | Represor | 44 | 4 | Winged helix DNA-binding domain |
| AtoC | Activador | 28 | 4 | Homeodomain-like |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| BaeR | Activador | 35 | 8 | CheY-like |
| BasR | Activador | 18 | 11 | CheY-like |
| BetI | Represor | 38 | 4 | Homeodomain-like |
| BglJ | Activador | 4 | 4 | C-terminal effector domain of the bipartite response regulators |
| BirA | Represor | 104 | 5 | Winged helix DNA-binding domain |
| BluR | Represor | 31 | 4 | Putative DNA-binding domain |
| BolA | Dual | 100 | 2 | BolA-like |
| CadC | Activador | 36 | 3 | C-terminal effector domain of the bipartite response regulators |
| CaiF | Activador | 6 | 10 | |
| Cbl | Activador | 15 | 9 | Winged helix DNA-binding domain |
| CdaR | Activador | 27 | 10 | |
| ChbR | Dual | 13 | 6 | Homeodomain-like |
| ComR | Represor | 45 | 1 | Homeodomain-like |
| CpxR | Dual | 64 | 63 | CheY-like |
| Cra | Dual | 40 | 78 | lambda repressor-like DNA-binding domains |
| CreB | Dual | 22 | 5 | CheY-like |
| CRP | Dual | 86 | 495 | Winged helix DNA-binding domain |
| CsgD | Dual | 27 | 23 | C-terminal effector domain of the bipartite response regulators |
| CsiR | Represor | 23 | 5 | Winged helix DNA-binding domain |
| CspA | Activador | 32 | 2 | Nucleic acid-binding proteins |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| CueR | Dual | 63 | 7 | Putative DNA-binding domain |
| CusR | Activador | 51 | 6 | CheY-like |
| CynR | Dual | 23 | 4 | Winged helix DNA-binding domain |
| CysB | Dual | 86 | 24 | Winged helix DNA-binding domain |
| CytR | Represor | 34 | 13 | lambda repressor-like DNA-binding domains |
| Dan | Activador | | 4 | Winged helix DNA-binding domain |
| DcuR | Activador | 12 | 9 | Winged helix DNA-binding domain |
| DeoR | Represor | 18 | 6 | Winged helix DNA-binding domain |
| DhaR | Dual | 4 | 4 | Homeodomain-like |
| DicA | Dual | 26 | 8 | lambda repressor-like DNA-binding domains |
| DinJ | Represor | 18 | 3 | |
| DnaA | Dual | 108 | 12 | P-loop containing nucleoside triphosphate hydrolases |
| DpiA | Dual | 19 | 11 | Winged helix DNA-binding domain |
| DsdC | Dual | 15 | 3 | Winged helix DNA-binding domain |
| EbgR | Represor | 4 | 2 | lambda repressor-like DNA-binding domains |
| EnvR | Represor | 12 | 2 | Homeodomain-like |
| EnvY | Activador | 1 | 2 | Homeodomain-like |
| EvgA | Activador | 5 | 18 | C-terminal effector domain of the bipartite response regulators |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| ExuR | Represor | 17 | 8 | Winged helix DNA-binding domain |
| FabR | Represor | 65 | 2 | Homeodomain-like |
| FadR | Dual | 48 | 18 | Winged helix DNA-binding domain |
| FeaR | Activador | 15 | 1 | Homeodomain-like |
| FhIA | Activador | 18 | 30 | Homeodomain-like |
| Fis | Dual | 100 | 225 | Homeodomain-like |
| FlhDC | Dual | 14 | 80 | FhC-like |
| FliZ | Desconocido | 13 | 20 | lambda integrase-like, N-terminal domain |
| FNR | Dual | 85 | 296 | Winged helix DNA-binding domain |
| FucR | Activador | 11 | 7 | Winged helix DNA-binding domain |
| Fur | Dual | 101 | 129 | Winged helix DNA-binding domain |
| GadE | Activador | 1 | 36 | C-terminal effector domain of the bipartite response regulators |
| GadW | Dual | 9 | 14 | Homeodomain-like |
| GadX | Dual | 16 | 28 | Homeodomain-like |
| GalR | Dual | 30 | 10 | lambda repressor-like DNA-binding domains |
| GalS | Dual | 19 | 10 | lambda repressor-like DNA-binding domains |
| GatR | Represor | 2 | 6 | Winged helix DNA-binding domain |
| GcvA | Dual | 69 | 5 | Winged helix DNA-binding domain |
| GlcC | Dual | 6 | 7 | Winged helix DNA-binding domain |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| GlpR | Represor | 54 | 9 | Winged helix DNA-binding domain |
| GlrR | Activador | 35 | 1 | Homeodomain-like |
| GntR | Represor | 41 | 12 | lambda repressor-like DNA-binding domains |
| GutM | Activador | 14 | 7 | |
| GutR | Represor | | 7 | Winged helix DNA-binding domain |
| H-NS | Dual | 34 | 179 | H-NS histone-like proteins |
| HcaR | Dual | 11 | 6 | Winged helix DNA-binding domain |
| HdfR | Dual | 44 | 5 | Winged helix DNA-binding domain |
| HipA | Desconocido | 39 | 2 | |
| HipB | Represor | 29 | 2 | lambda repressor-like DNA-binding domains |
| HU | Dual | 107 | 9 | IHF-like DNA-binding proteins |
| HyfR | Activador | 34 | 12 | Homeodomain-like |
| HypT | Dual | 19 | 13 | Winged helix DNA-binding domain |
| IclR | Represor | 29 | 4 | Winged helix DNA-binding domain |
| IdnR | Dual | 4 | 7 | lambda repressor-like DNA-binding domains |
| IHF | Dual | 105 | 219 | IHF-like DNA-binding proteins |
| IlvY | Dual | 49 | 2 | Winged helix DNA-binding domain |
| IscR | Dual | 93 | 32 | Winged helix DNA-binding domain |
| KdgR | Represor | 28 | 2 | Winged helix DNA-binding domain |
| KdpE | Activador | 37 | 4 | CheY-like |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| LacI | Represor | 6 | 3 | lambda repressor-like DNA-binding domains |
| LeuO | Dual | 57 | 20 | Winged helix DNA-binding domain |
| LexA | Represor | 81 | 56 | Winged helix DNA-binding domain |
| LldR | Represor | 12 | 3 | Winged helix DNA-binding domain |
| LrhA | Dual | 40 | 4 | Winged helix DNA-binding domain |
| Lrp | Dual | 81 | 103 | Winged helix DNA-binding domain |
| LsrR | Represor | 19 | 9 | Winged helix DNA-binding domain |
| LysR | Dual | 19 | 2 | Winged helix DNA-binding domain |
| Mall | Represor | 11 | 3 | lambda repressor-like DNA-binding domains |
| MalT | Activador | 31 | 10 | C-terminal effector domain of the bipartite response regulators |
| MarA | Dual | 6 | 38 | Homeodomain-like |
| MarR | Represor | 31 | 3 | Winged helix DNA-binding domain |
| MatA | Dual | 0 | 8 | C-terminal effector domain of the bipartite response regulators |
| MazE | Represor | 10 | 3 | AbrB/MazE/MraZ-like |
| MazF | Represor | 11 | 3 | Cell growth inhibitor/plasmid maintenance toxic component |
| McbR | Dual | 10 | 4 | Winged helix DNA-binding |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| | | | | domain |
| MeiR | Dual | 17 | 3 | Homeodomain-like |
| MetJ | Represor | 52 | 15 | Ribbon-helix-helix |
| MetR | Dual | 82 | 5 | Winged helix DNA-binding domain |
| MhpR | Activador | 1 | 6 | Winged helix DNA-binding domain |
| Mlc | Represor | 29 | 10 | Winged helix DNA-binding domain |
| MlrA | Activador | 23 | 8 | Putative DNA-binding domain |
| MngR | Represor | 20 | 3 | Winged helix DNA-binding domain |
| MntR | Dual | 17 | 5 | Winged helix DNA-binding domain |
| ModE | Dual | 43 | 46 | Winged helix DNA-binding domain |
| MprA | Represor | 54 | 6 | Winged helix DNA-binding domain |
| MqsA | Represor | 22 | 4 | lambda repressor-like DNA-binding domains |
| MtlR | Represor | 27 | 3 | MtlR-like |
| MurR | Dual | 17 | 3 | Homeodomain-like |
| Nac | Dual | 18 | 20 | Winged helix DNA-binding domain |
| NadR | Represor | 32 | 4 | lambda repressor-like DNA-binding domains |
| NagC | Dual | 40 | 35 | Winged helix DNA-binding domain |
| NanR | Dual | 9 | 8 | Winged helix DNA-binding domain |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| NarL | Dual | 25 | 121 | C-terminal effector domain of the bipartite response regulators |
| NarP | Dual | 50 | 49 | C-terminal effector domain of the bipartite response regulators |
| NemR | Represor | 50 | 3 | Homeodomain-like |
| NhaR | Activador | 32 | 7 | Winged helix DNA-binding domain |
| NikR | Represor | 16 | 6 | ACT-like |
| NorR | Dual | 26 | 3 | Homeodomain-like |
| NrdR | Dual | 97 | 9 | |
| NsrR | Represor | 58 | 83 | Winged helix DNA-binding domain |
| NtrC | Dual | 0 | 44 | Homeodomain-like |
| OmpR | Dual | 65 | 17 | C-terminal effector domain of the bipartite response regulators |
| OxyR | Dual | 92 | 33 | Winged helix DNA-binding domain |
| PaaX | Represor | 8 | 12 | Winged helix DNA-binding domain |
| PdhR | Dual | 45 | 42 | Winged helix DNA-binding domain |
| PepA | Represor | 108 | 3 | Zn-dependent exopeptidases |
| PgrR | Represor | 25 | 4 | Winged helix DNA-binding domain |
| PhoB | Dual | 85 | 60 | CheY-like |
| PhoP | Dual | 54 | 55 | CheY-like |
| PrpR | Dual | 10 | 5 | PrpR receptor domain-like |
| PspF | Dual | 35 | 7 | Homeodomain-like |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| PurR | Represor | 37 | 31 | Lambda repressor-like DNA-binding domains |
| PutA | Represor | 78 | 2 | ADLDH-like |
| PuuR | Represor | 37 | 7 | lambda repressor-like DNA-binding domains |
| QseB | Activador | 64 | 4 | CheY-like |
| RbsR | Dual | 38 | 9 | lambda repressor-like DNA-binding domains |
| RcdA | Desconocido | 14 | 9 | Homeodomain-like |
| RcnR | Represor | 11 | 3 | |
| RcsA | Activador | 12 | 16 | C-terminal effector domain of the bipartite response regulators |
| RcsB | Dual | 20 | 48 | C-terminal effector domain of the bipartite response regulators |
| RelB | Represor | 14 | 3 | |
| RelE | Desconocido | 31 | 3 | RelE-like |
| RhaR | Activador | 22 | 2 | Homeodomain-like |
| RhaS | Desconocido | 38 | 6 | Homeodomain-like |
| Rob | Activador | 23 | 26 | Homeodomain-like |
| RstA | Activador | 37 | 10 | C-terminal effector domain of the bipartite response regulators |
| RtcR | Activador | 6 | 2 | P-loop containing nucleoside triphosphate hydrolases |
| RutR | Dual | 41 | 17 | Homeodomain-like |
| SdiA | Dual | 30 | 5 | C-terminal effector domain of the bipartite response regulators |
| SgrR | Dual | 21 | 9 | Winged helix DNA-binding domain |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---|
| SlyA | Activador | 63 | 1 | Winged helix DNA-binding domain |
| SoxR | Dual | 31 | 3 | Putative DNA-binding domain |
| SoxS | Dual | 16 | 38 | Homeodomain-like |
| StpA | Dual | 30 | 5 | H-NS histone-like proteins |
| TdcA | Activador | 5 | 7 | Winged helix DNA-binding domain |
| TdcR | Activador | 1 | 7 | |
| TorR | Dual | 13 | 12 | CheY-like |
| TreR | Represor | 20 | 2 | lambda repressor-like DNA-binding domains |
| TrpR | Represor | 43 | 12 | TrpR-like |
| TyrR | Dual | 53 | 12 | Homeodomain-like |
| UhpA | Activador | 22 | 1 | C-terminal effector domain of the bipartite response regulators |
| UidR | Represor | 44 | 4 | Homeodomain-like |
| UlaR | Represor | 19 | 7 | Winged helix DNA-binding domain |
| UxuR | Represor | 25 | 8 | Winged helix DNA-binding domain |
| XapR | Activador | 5 | 2 | Winged helix DNA-binding domain |
| XylR | Activador | 22 | 6 | Homeodomain-like |
| YafQ | Represor | 16 | 3 | RelE-like |
| YdeO | Dual | 4 | 5 | Homeodomain-like |
| YefM | Represor | 33 | 2 | YefM-like |
| YehT | Activador | 51 | 1 | CheY-like |
| YeiL | Activador | 1 | 1 | Winged helix DNA-binding domain |
| YiaJ | Represor | 11 | 10 | Winged helix DNA-binding |

| Regulador | Tipo de regulación | Genomas conservados | Genes regulados (<i>E. coli</i>) | Familia evolutiva |
|------------------|---------------------------|----------------------------|---|---------------------------------|
| | | | | domain |
| YoeB | Represor | 30 | 2 | RelE-like |
| YpdB | Activador | 17 | 1 | CheY-like |
| YqhC | Activador | 45 | 2 | Homeodomain-like |
| YqjI | Represor | 16 | 2 | Winged helix DNA-binding domain |
| ZntR | Activador | 61 | 1 | Putative DNA-binding domain |
| ZraR | Activador | 12 | 3 | Homeodomain-like |
| Zur | Represor | 75 | 6 | Winged helix DNA-binding domain |



Research Signpost
37/661 (2), Fort P.O.
Trivandrum-695 023
Kerala, India

Review Article

Recent Res. Devel. Microbiology, 12(2012): 19-39 ISBN: 978-81-308-0467-5

2. Description of gene regulatory networks in three bacterial models: *Escherichia coli*, *Bacillus subtilis* and *Pseudomonas aeruginosa*

Edgardo Galán-Vásquez¹, Beatriz Luna² and Agustino Martínez-Antonio¹

¹Departamento de Ingeniería Genética, Cinvestav, Km. 9.6 Libramiento Norte Carr. Irapuato-León 36821 Irapuato Gto. México; ²Scientific Computing and Applied Mathematics Laboratory, Universidad del Papaloapan, Av. Ferrocarril s/n. 68400 Loma Bonita, Oaxaca, México

Abstract. Bacterial organisms are important biological models for the study of diverse aspects of biology from molecular to synthetic biology. However to better redesign and engineer simple unicellular organisms it is imperative to know the internal structure and dynamics of their genetic regulatory networks. In this work we take advantage of the health of knowledge we have about the gene regulation of bacterial model systems to describe and compare the structure of the gene regulatory networks in *E. coli*, *B. subtilis* and *P. aeruginosa*. We report that even when these networks cover in distinct amplitude the total of genes in each organism they conserve general topological properties, for instance: degree distribution, clustering coefficient distribution and connectivity; and also relevant aspects about their functional

Correspondence/Reprint request: Dr. Agustino Martínez-Antonio, Departamento de Ingeniería Genética Cinvestav, Km. 9.6 Libramiento Norte Carr. Irapuato-León 36821 Irapuato Gto. México.
E-mail: amartinez@ira.cinvestav.mx

organization, for instance: activation as the most common form of regulation over their genes, and auto-repression for transcription factors, except for *P. aeruginosa* where self-activation dominates. We found the global regulators of the three networks, which affect as 50% of genes in the network of *E. coli* (CRP) to the 15% in the case of *P. aeruginosa* (LasR). We also describe additional characteristics of the networks such as hierarchy, cycles, motifs and biological modules among others. We hope this study helps us to achieve a global picture of gene regulation in these important bacterial models.

Introduction

Organisms are constituted by thousand of molecular entities of diverse nature such as nucleic acids, proteins, organic compounds, metals, lipids, etc. To function, organisms need to coordinate the activities of these thousands of compounds, among them and by perceiving and responding to the multiple clues from the environment. The origin of the myriad of compounds to make a cell function could be genome-encoded and non-encoded. The genome-encoded information is all those genetic information proper of each population of organisms or species. A fraction of the encoded information in turn encodes for a regulatory machinery, which directs the elaboration of cellular machinery depending of the environmental conditions. On the other hand, the non-encoded information is composed of all the physicochemical components being part of the organism but that the cellular machinery takes out from the environment and uses and modifies for their proper functioning.

Simplest organisms like bacteria normally have thousands of genes encoded in their genomes. Depending on the environmental conditions bacteria express a limited repertoire of genes in different patterns. DNA-binding proteins known as transcription and sigma factors compose the regulatory machinery that controls gene expression at the transcriptional level. Sigma factors direct the RNA polymerase holoenzyme to specific promoters to start the transcription of RNA using the coding strand of DNA as a template. Sigma factors can start the transcription themselves but in most cases they require accessory regulatory proteins: the transcription factors (TF), which allow to better modulate the transcription of genes since they work as regulatory switches depending on the environmental conditions. The natural balance among the components of the regulatory machinery (transcription units, transcription and sigma factors) is around an order of magnitude among them; i.e. thousands, hundreds, dozens, respectively [1].

The multiple biological elements, whose activities sustain life, can be represented into a mathematical network, where vertices symbolize biochemical components, and edges symbolize their interactions. In particular, gene

regulatory networks are constructed having experimental evidence that indicates that the product of a gene (a regulatory protein) controls the transcription of their own gene (self-regulation) or of additional genes (encoding for regulatory or non-regulatory products). This abstract representation allows us to obtain the regulatory networks, whose internal structure can be described employing graph theory tools, and interpreted from a biological point of view. In this way the genetic components of each organism can be represented within a network from which we can compute measures of complexity in order to characterize it and compare the functional structures of biological systems. This analysis provides information about which parts of the networks are common structures, and which are particularities of every organism. The study of gene regulatory networks might help scientist to put into context the regulatory interactions between a TF and their target genes. This network context can stimulate the consideration of additional elements that impact the activity of a gene, which are difficult to appreciate outside this context.

From the mathematical point of view, regulatory networks inspire the creation of new metrics that allow to characterize and to compare these networks in their different levels of organization.

In this work we compared the three best-known transcriptional regulatory networks, which respectively corresponds to the most studied bacterial systems. It is important to lay stress on no one autonomous living system that has been fully characterized to a molecular level. However the partial knowledge we have about the functional activities of these organisms have allowed us to relate genetic network structures and their regulatory dynamics with the corresponding phenotypes. In the following sections we shall present information about the genetic regulatory networks in three different ways; first we will describe the general information of the networks, then we will show the topological and functional organization, and finally we will compare and discuss these results from a biological point of view.

1. Bacteria models

We chose to study the following bacterial models, since they are the best-known bacterial species to the genetical, biochemical, developmental, and pathogenical levels. This study was possible due to the work of many researchers around the world dedicated to discern the multiple interactions happening in each organism. This information has made possible that in the last decade the study of transcriptional regulatory networks has had

significant progresses. Nowadays, significant data is known about the genetic regulatory networks of bacterial models, which permit us to describe and compare them for the first time.

Escherichia coli

This Gram-negative bacterium is the bacterial model of study par excellence. This bacterium was important to establish the basis of modern genetics and genomics. It is also an important model in biotechnology, since it has served as the host organism for the development of vectors and recombinant DNA technology. The first complete DNA sequence of *E. coli* genome was published in 1997 (strain K-12 MG1655). It consists of 4'639'221 base pairs, containing 4'400 annotated protein-coding genes [2]. The total number of predicted transcription factors is around 300, and 7 sigma factors [3]. There are several databases containing molecular information about this bacterium, especially the *E. coli* database portal (<http://www.uni-giessen.de/ecoli/IECA/index.php>), RegulonDB [4], and Ecocyc [5] for transcriptional regulation and metabolic information respectively.

Bacillus subtilis

Is the model of study for Gram-positive bacteria, it is the best-studied soil microorganism in relation with their biochemistry, physiology and genetics. This bacterium is well known for its ability to differentiate from vegetative growing cells into metabolically inactive spores that allow it to survive for long times under extreme environmental conditions [6]. Its genome is constituted of 4'214'810 base pairs and comprises 4'422 protein-coding genes [7]. The distinctive strain of the specie is the called 168. The estimated number of transcription factors is around 275, and the number of sigma factor is 14 [8]. The main databases that contain gene regulatory information are Subtiwiki (http://subtiwiki.uni-goettingen.de/wiki/index.php/Main_Page), and DBTBS [9].

Pseudomonas aeruginosa

It is a metabolically versatile Gram-negative bacterium. It expresses a wide range of virulence factors. This also allows *P. aeruginosa* to grow in soil and marine habitats, as well as on plant and animal tissues, this fact constitutes this bacterium as an opportunist pathogen. It is also a significant source of bacteremia in burn victims, urinary-tract infections, hospital-acquired pneumonia; and predominant cause of morbidity and mortality in cystic fibrosis patients. All these make *P. aeruginosa* the most studied bacterial

model regarding the control of pathogenic determinants. The genome sequence of *P. aeruginosa* strain PAO1 was reported in 2000 [10], since then numerous database and genomic resources have been implemented to study their molecular and pathogenic biology [11,12,13]. The information analyzed here was taken from the first comprehensive regulatory network, published in 2011 [14]. The additional database PseudoCAP is dedicated to the genetic information about this bacterium [15].

2. The gene regulatory networks (GRN)

The data about the genetic regulatory network, and more specifically about transcriptional regulation, was taken from the database REGULONDB version 7.0 for *E. coli* [4], it consists of 1'565 genes (of the 4'400 predicted in the genome this bacteria, 35.56% of total genes) and 3'530 regulatory interactions. The total number of genes in the network includes 176 transcription factors, 7 sigma factors and 1'382 target genes (**Figure 1A**).

The transcriptional regulatory network of *B. subtilis* was obtained from the database DBTBS release 5 [9], it consists of 1'659 genes (of the 4'422 predicted -37.5%-) and 3'121 regulatory interactions. The total number of genes in the network includes 106 transcription factors, 16 sigma factors (six of these defined with extra-cytoplasmic functions –ECF-), 6 anti-sigma factors and 1'531 target genes (**Figure 1B**).

Finally, the transcriptional regulatory network of *P. aeruginosa* was taken from [14]; it consists of 690 genes (of the 5'570 predicted -12.4%-) and 1'020 regulatory interactions. The total number of genes in this network includes 76 transcription factors, 14 sigma factors (nine of these defined with extra-cytoplasmic functions –ECF-), 7 anti-sigma factor, and 593 target genes (**Figure 1C**).

The sizes of the networks represent clearly the interest of genetic bacteriologists to study gene regulation in these bacteria. The better-characterized bacterium is *E. coli* where the network is reconstructed from pairwise regulatory interactions. In the case of *B. subtilis* a great part of data is derived of high-throughput methodologies and bioinformatics studies. *P. aeruginosa* data was taken from pairwise regulatory interactions also. High-throughput studies could help to increase this kind of information in bacteria such as *P. aeruginosa* rapidly, but the level of detail obtained with studies of pairwise interactions, as those from *E. coli*, is difficult to achieve with high-throughput methodologies.

In the following sections we shall describe diverse aspects of these genetic regulatory networks.

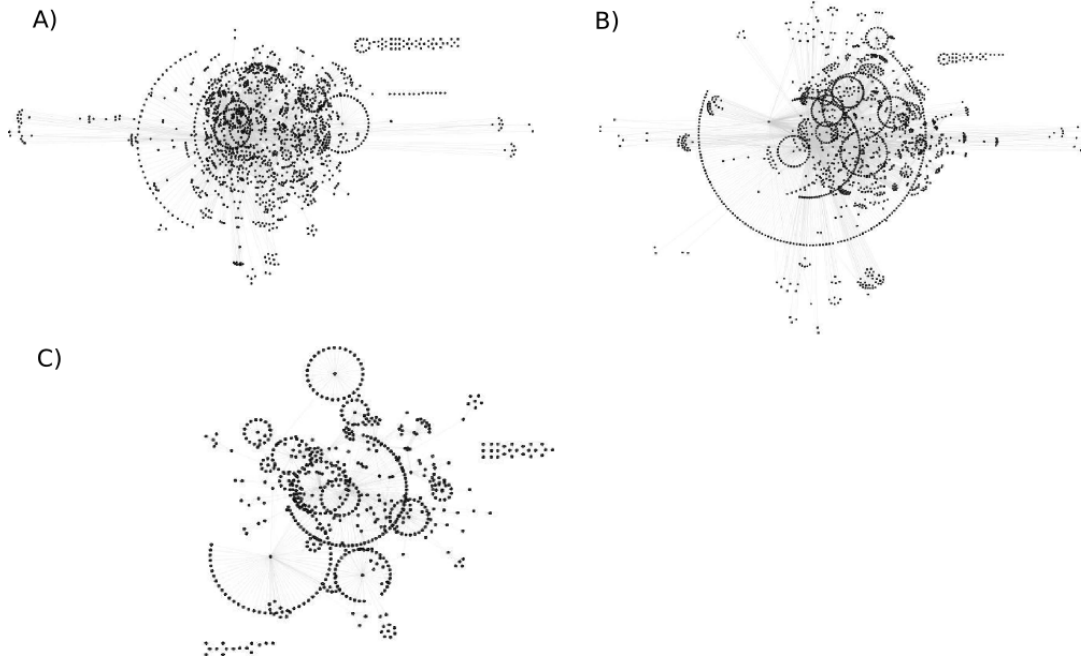


Figure 1. Gene –transcriptional- regulatory networks (TRNs) considered in this work. A) TRN of *E. coli*, B) TRN of *B. subtilis* and C) TRN of *P. aeruginosa*. Networks were drawn with the Cytoscape software [16]. Black nodes represent genes and gray edges represent the regulatory interactions.

3. Topological description of the gene regulatory networks

In this part of the study we shall describe the structural organization of the regulatory networks. For this purpose we will consider a network N with a set of vertices V , and a set of edges A , where every edge (u, v) connects two vertices u and v in the network. Network can be directed, meaning that there is only one sense of the interaction, from u (the tail) to v (the head), or undirected, where there is no direction of the interaction between any two vertices [17].

Degree distribution

The degree k_v of a vertex v in a network N is defined as the number of edges, which connect the vertex with the rest of vertices in the network. In this case, since we consider a directed network, where every edge has an orientation, we can consider also the input (k_v^{in}) and output (k_v^{out}) degree, which are defined as the number of edges with head and tail in v , respectively [17]. In biological terms this means the number of TF regulating to a target gene (*in*) and the number of genes regulated by a TF (*out*), respectively.

The degree distribution $P(k)$ gives the probability that a selected vertex has degree k . This value allows us to distinguish between different types of networks, being common to find a power law distribution in biological networks, that is to say, in most cases we can approximate the distribution as $P(k) = Ak^{-\gamma}$. This type of distribution possesses well-defined characteristics, for instance, the existence of hubs (highly connected vertices), which are connected to vertices with low degree. Common values of biological networks are $2 < \gamma < 3$, meaning that we can observe a hierarchy among hubs [18].

For *E. coli* and *B. subtilis*, we found $A=0.97$ and $A=0.86$ respectively, and $2 < \gamma < 3$ for the input degree distributions in both cases, which indicates that even when the networks cover different percentages of the genes in each organism their overall distribution have the same arrangement (**Figure 2**). The network of *P. aeruginosa* has the same behavior; we found $A=0.8856$ and $2 < \gamma < 3$ for the input degree distribution.

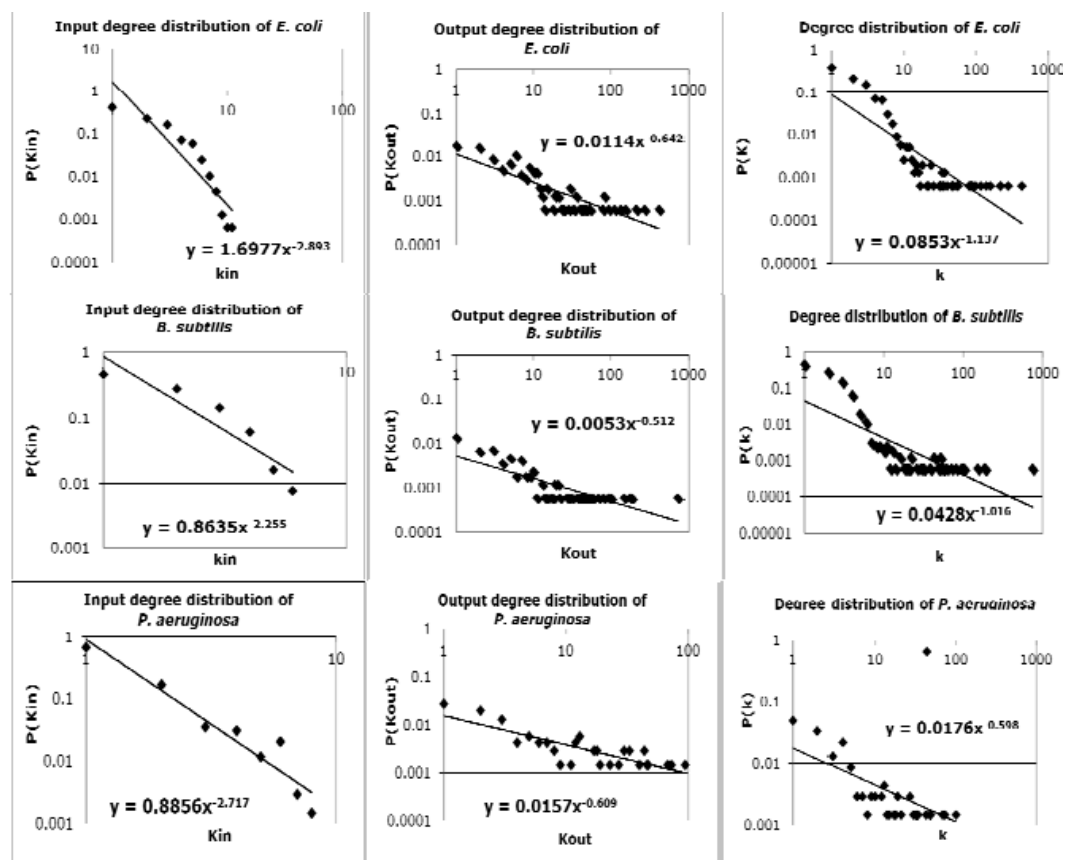


Figure 2. Degree distributions. In the X-axis we show the different kin , $kout$ and k found in the three networks. In the Y-axis we show the probability that a selected vertex has the given degree. Graphs are presented in logarithmic scales.

As it is known output and overall degree distributions show different distributions compared with the input degree distribution, some theories indicate that these degree distributions are better adjusted by a Poisson distribution, since analyses fail to confirm the power law distribution, meaning that the variability between the degree distributions rules out the hope to discover any universal law that describe all distributions [19].

Clustering coefficient

The clustering coefficient C_i of the vertex v_i is calculated as: $C_i = 2E_i / (k_i)(k_i - 1)$, where E_i is the number of edges between the neighbors of v_i , in other words this quantity evaluates the probability that two vertices with a common neighbor are also connected, and it relates to the local cohesiveness of a network in the way of local modules [20].

The results (**Figure 3**) are shown considering the clustering coefficient as function of the degree k . For many real networks it is known that $C(k) = Bk^{-\beta}$ with β closer to 1, which indicates that a hierarchy is present in the network [18].

The three TRNs exhibit a decreasing value of $C(k)$ with respect to the degree k , in such a way that in small groups or modules of genes the elements are well connected, but as the group increases in size the elements are progressively less connected. This reflects the modular organization of regulatory networks (see below).

The TRNs show $B=1.430$ and $\beta=1.067$, $B=2.247$ and $\beta=1.109$, and $B=1.989$ and $\beta=1.096$ for *E. coli*, *B. subtilis* and *P. aeruginosa* respectively, which indicates again a similar arrangement in the clustering of their gene networks.

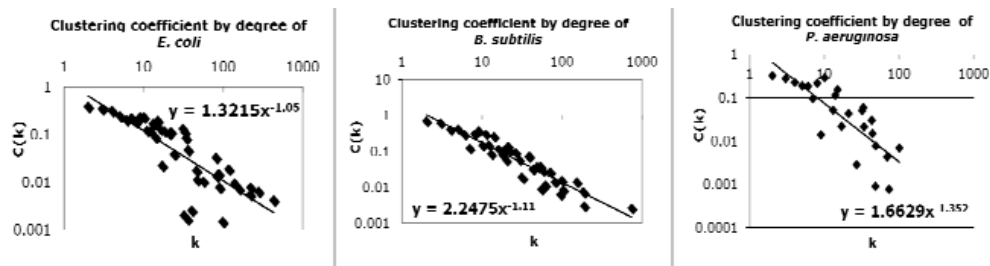


Figure 3. Clustering coefficient by degree in the TRNs of *E. coli*, *B. subtilis* and *P. aeruginosa*. In the X-axes we represent the degrees for each network. In the Y-axes we show the clustering coefficient. Graphs are shown in logarithmic scales.

Connectivity

If we consider an undirected network, a connected component is a subgraph of the network in which it exists a path (a finite sequence of vertices and arrows) from u to v , for any two vertices u and v in the network [17].

In the case of *E. coli* there are 16 connected components, with one giant component of 1'501 genes (95% genes). The network of *B. subtilis* contains 9 connected components, with one giant component of 1'623 genes (97% genes). For the TRN of *P. aeruginosa* there are 12 connected components, with one giant component containing 650 genes (90% genes), (**Figure 4**).

As we can observe, along the analysis of these topological characteristics, there are well recognized similarities among these regulatory networks. This fact indicates that there is a common topology on regulatory networks and that we are finally working with organisms from a same phylogenetic origin. On the other hand, we can observe that the main regulatory components are not necessarily the same ones in these networks but the general structure is maintained. There are studies pending on the dynamics of regulatory pathways inside these networks, which will allow us to gain knowledge in order to obtain a complete description of these networks and their components.

4. Functional organization of the TRNs

Self-repression is the dominating activity of auto-regulatory genes

TFs in addition to its DNA-binding domain normally have a domain for binding signals effectors of different nature (sugars, amino acids, metals, etc.). This additional domain permits them to function as regulatory switches. Since TF can have effect over a long number of genes, it is normal that they self-regulate

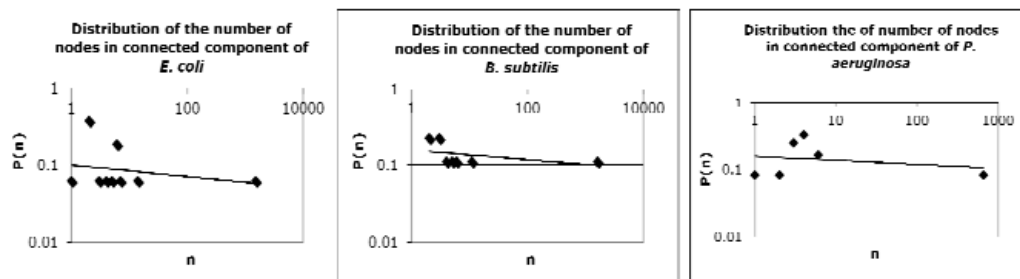


Figure 4. Distribution of the number of connected components in the three TRNs. In the X-axes we show the number of vertices by component. In the Y-axes the relative frequency in which this number appears is represented. Graphs are presented in logarithmic scales.

Table 1. Self-regulations of TF in the TRNs of bacterial models (in parenthesis are shown the percentages of kinds of self-regulations).

| | <i>E. coli</i> | <i>B. subtilis</i> | <i>P. aeruginosa</i> |
|----------------------------------|----------------|--------------------|----------------------|
| TF in the network | 176 | 106 | 76 |
| Total of TF with auto-regulation | 112 | 69 | 29 |
| Positive auto-regulation | 31 (28%) | 27 (39%) | 16 (55%) |
| Negative auto-regulation | 73 (65%) | 41 (60%) | 13 (45%) |
| Dual auto-regulation | 8 (7%) | 1 (1%) | 0 |

as an efficient form and respond quickly to the environmental conditions. There are three forms in which a regulatory gene (TF) can self-regulate: activation, repression or both. The **Table 1** shows the type of self-regulation for the TFs of regulatory networks we are studying. We can observe that self-repression dominates in the cases of *E. coli* and *B. subtilis*. This kind of self-regulation on TF make biological sense since negative feedbacks are necessary to keep homeostasis and this is a function that most of TF in bacteria perform. In the case of *P. aeruginosa* there is a slight overrepresentation of self-activation for their TFs [14]. There is a study pending on the regulatory network in this last bacterium to know if this is a general feature of the regulatory network in this organism or if on the contrary, it is a feature of the part of the network controlling pathogenic and virulence factors, until now this has been the most characterized part of the network in *P. aeruginosa*. The study of pathogeny and virulence in *P. aeruginosa* is in sharp contrast with the other two organisms, where most of the characterized networks are related to diverse physiological aspects including central metabolism.

Activation is the dominant mode of regulation of genes in the networks

There are three forms in which a TF affects the expression of a target gene: activation, repression or both (activation and repression). We compute into the three networks the form in which every TF is affecting their target genes. In **Table 2** we show that activation is the dominant form of regulation into the three networks.

The fact that activation dominates the regulation of genes might be a necessary condition for the correct execution of large transcriptional programs since it needs the sequential activity of several genes. This also implies

Table 2. Regulatory interactions in the TRNs of bacterial models (in parenthesis are shown the percentages of regulation modes).

| Regulatory interactions | <i>E. coli</i> | <i>B. subtilis</i> | <i>P. aeruginosa</i> |
|----------------------------------|----------------|--------------------|----------------------|
| Total of regulatory interactions | 3530 | 3121 | 1020 |
| Positive regulation | 1886 (53%) | 2030 (65%) | 779 (76%) |
| Negative regulation | 1459 (41%) | 977 (31%) | 218 (21%) |
| Dual regulation | 182 (5%) | 8 (2%) | 11 (1%) |
| Unknown mode of regulation | 3 | 106 | 12 |

the existence of regulatory paths that are prone to run transcriptional programs. These aspects need more study and characterization into the transcriptional networks.

Hierarchy (path lengths) into the regulatory networks

A path (W) in graph theory is a finite sequence of vertices and arrows, from u different to v , where the internal vertices and edges are also distinct. Biologically, a path describes a regulatory hierarchical cascade. It means that the regulatory information, start in a TF, and can go through different TFs until it ends in a target gene.

It has been thought that prokaryotes have short regulatory cascades since they lack developmental programs where long regulatory paths are normally required. **Figures 5** and **6** show the maximal length paths as if aligned to the longest path in each organism. The average of path lengths is of 3.19 for *E. coli*, 2.98 for *B. subtilis* and 4.08 for *P. aeruginosa*. The maximal path length is 14 steps for *E.coli* (flagella formation and iron regulation), 12 steps for *B. subtilis* (devoted to the control of heat-shock), and 11 steps for *P. aeruginosa* (involved in the production of the virulence factor Exotoxin A). Large regulatory cascades make sense since some bacterial processes are considered as developmental process, which implies transitions across more than one cellular phenotype, which might imply the execution of large regulatory programs.

In the **Figure 5** we can appreciate how TFs in the hierarchical steps are self-regulating their activities. Most self-repressing TFs are located in the first steps as the case of *E. coli* and *B. subtilis*, which is not the case of *P. aeruginosa*. From **Figures 5** and **6** we can appreciate that genes are unevenly distributed, being the more uniformly distributed as in the case of *E.coli*, since for *B. subtilis* and *P. aeruginosa* the highest percentages of genes are respectively in the second and third level. All these arrangements indicate that regulatory genes in a level are

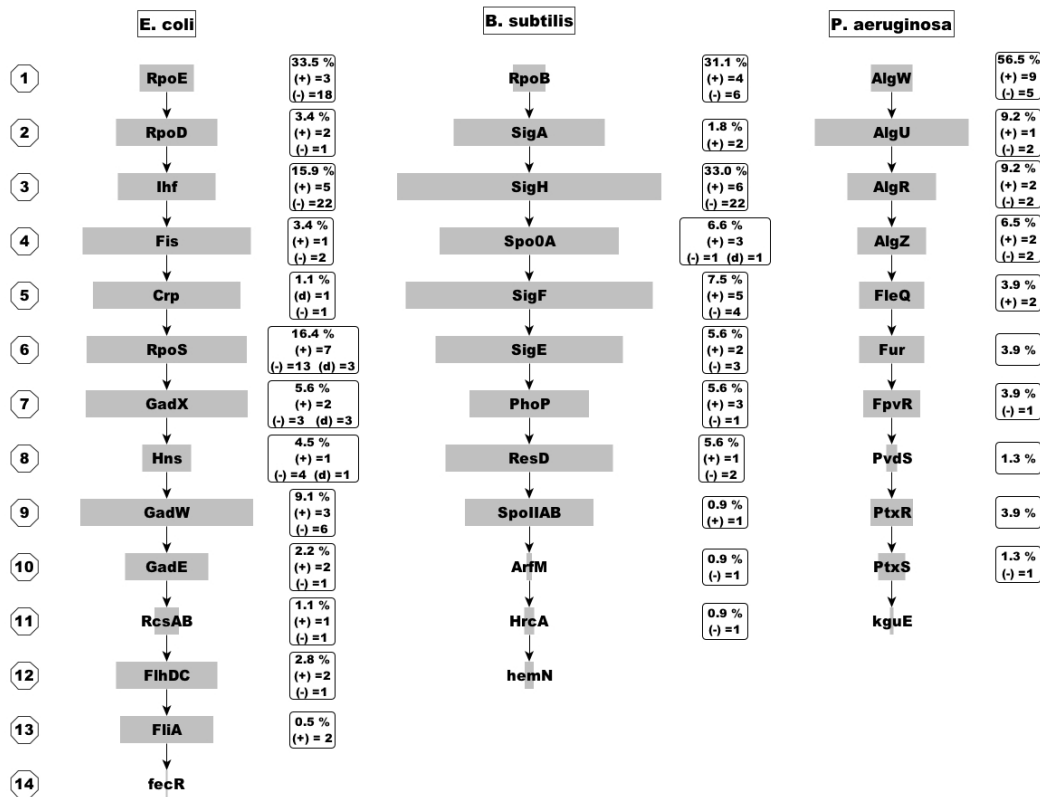


Figure 5. Length of regulatory paths into the three regulatory networks. In each organism is exemplified the longest regulatory paths. The width of boxes in each level indicates the percentages of the total genes in each step in these paths. Data in the boxes aside indicates the percentages of TF with their modes of self-regulations in each level.

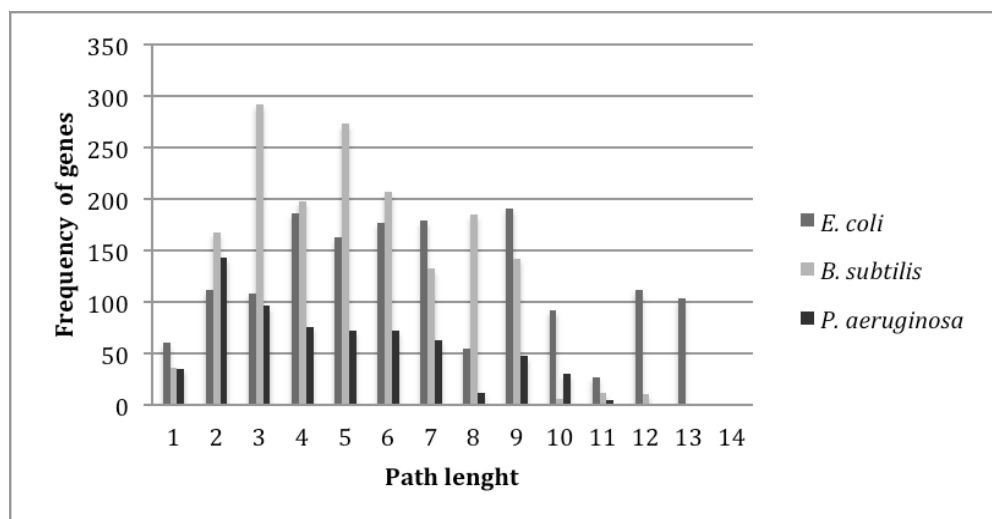


Figure 6. Distribution of genes along the longest regulatory paths in each organism.

regulating to genes in the subsequent levels, lowers the hierarchy. Since TFs regulate most of their genes in a first step (first neighbors) it is easy to locate the position of global regulators (see below) in these networks.



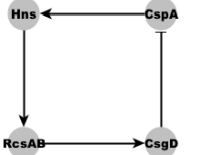
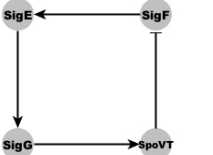
Multi-element regulatory circuits (cycles) into the regulatory networks

A cycle is a self-enclosed path W , that is to say, that its origin u and terminus v are the same vertex. Multi-element regulatory circuits imply that more than one vertex (gene) is part of a feedback circuit. These elements are important in biology, since feedback loops with multiple elements present higher robustness than the one-element feedback loops. Since several TFs are contained in cycles this implies the feedback systems could drive positive or negative functions (the sign of a cycle is computed as the product of the sign of their interactions) by sensing and responding to the presence of multiple signal effectors.

In the networks analyzed here we found cycles from between two and four elements, the most common being those of 2 elements (**Table 3**).

Motifs in biological networks are defined as patterns of interconnections occurring at numbers that are significantly higher than those in random networks








Table 3. Multi-element regulatory circuits found in the regulatory networks of bacterial models. The numbers in parenthesis indicate the frequency in which these cycles are found in reconstructed random networks (1'000).

| Feedback multi-element cycles | Examples | <i>E. coli</i> | <i>B. subtilis</i> | <i>P. aeruginosa</i> |
|-------------------------------|---|----------------|--------------------|----------------------|
| Positive |  | 8 (1.59) | 7 (1.12) | 1 (0.94) |
| Negative |  | 6 (0.89) | 1 (0.79) | 6 (0.50) |
| Negative |  | 1 | 0 | 0 |
| Negative |  | 0 | 1 | 0 |

Abundance of network motifs

We looked for motifs with three and four vertices, and pick out the more representatives. Then we generated 1000 random networks with the same number of vertices and the same number of edges for every one of the transcriptional regulatory networks of bacterial models obtaining the significant values of each of these selected motifs with respect to the random networks. Among the most representative motifs in the regulatory networks we found the feed-forward loop (FFL), and the Bi-fan (**Table 4**).

Table 4. Network motifs most representatives into the TRN of bacteria.

| Motifs | Examples | <i>E. coli</i> | <i>B. subtilis</i> | <i>P. aeruginosa</i> |
|--------------------------|---|------------------------|------------------------|-----------------------|
| FFL coherent Type 1 |  | 316 (4.89) | 416 (2.51) | 89 (1.65) |
| FFL coherent Type 2 |  | 204 (0.40) | 153 (0.34) | 17 (0.54) |
| FFL incoherent Type 1 |  | 287 (1.42) | 431 (0.99) | 11 (0.58) |
| FFL incoherent Type 4 |  | 127 (1.39) | 16 (0.95) | 17 (0.54) |
| Bi-fan Type 1 |  | 14446 (4.11) | 14210 (1.65) | 3832 (1.13) |
| Bi-fan Type 2 |  | 11003 (2.42) | 3462 (1.23) | 171 (0.43) |
| Bi-fan Type 3 |  | 9742 (0.34) | 12041 (0.22) | 214 (0.05) |

A FFL network motif consists of three genes: a regulator X , which regulates to another regulator Y and together regulate to the gene z . We found particularly abundant those FFL of the coherent type 1, where all the regulatory interactions are positive. On the other hand, the FFL incoherent of type 1, where X regulates Y and z positively, but Y regulates z negatively, is the most abundant FFL motif for *B. subtilis*. It is well known that the coherent FFL acts as a sign-sensitive delay and a persistence detector. On the other hand the incoherent FFL is a pulse generator and a response accelerator [22].

Another common motif in the regulatory network is the Bifan (where two TF; X and Y regulate to the genes w and z) [21]. The Bi-fan motif is identified as a building block of dense arrays of overlapping regulation, which performs hard-wired combinatorial decisions governed by the input signals of two co-regulating TFs [23]. In the three networks the most common type of Bifan motif is that where both regulators activate to their target genes (**Table 4**).

Global regulators into the three bacterial regulatory networks

The most influencing TFs in a regulatory network are called “global regulators”. They are defined by a series of operative properties, including: i) they should regulate a large number of genes; ii) they should regulate other sigma’s and regulatory genes; iii) they should co-regulate together with many TFs and, iv) their target genes should have promoters using more than one kind of sigma factor [24]. Computing these criteria we identified the top ten global regulators in each of the transcriptional regulatory networks of bacterial models (**Tables 5A-C**) using the following equation already reported in [14].

$$G = \frac{1}{4} \left(\frac{TFR}{N_{TF} + N_{SF} - 1} + \frac{GR}{N_G} + \frac{SF}{N_{SF}} + \frac{CR}{N_{TF} - 1} \right)$$

Where, N_{TF} indicates the total number of TFs (in the known network in each case), N_G is the number of target genes, and N_{SF} is the number of sigma factors used by the promoters of genes in the whole network. Additionally, TFR and GR , represent the number of TFs and target genes regulated by each TF, respectively. SF represents the distinct sigma factors used by the promoters of genes regulated by each TF; and CR represents the number of TFs each TF co-regulates with. We obtain a G coefficient for each TF, which indicates the relative global activity of a TF in the network.

Table 5A. Top ten most influencing regulators in the TRN of *E. coli*.

| TF | TFR | GR | SF | CR | G coefficient |
|------|-----|-----|----|-----|---------------|
| CRP | 45 | 383 | 5 | 106 | 0.461 |
| IHF | 10 | 204 | 4 | 65 | 0.286 |
| FNR | 13 | 264 | 4 | 52 | 0.282 |
| FIS | 10 | 208 | 3 | 54 | 0.235 |
| HNS | 15 | 121 | 2 | 51 | 0.186 |
| CpxR | 4 | 51 | 4 | 18 | 0.183 |
| ArcA | 6 | 151 | 2 | 44 | 0.169 |
| NsrR | 7 | 76 | 3 | 21 | 0.160 |
| Fur | 7 | 76 | 2 | 34 | 0.143 |
| Lrp | 3 | 92 | 2 | 34 | 0.140 |

Table 5B. Top ten most influencing regulators in the TRN of *B. subtilis*.

| TF | TFR | GR | SF | CR | G coefficient |
|-------|-----|-----|----|----|---------------|
| CcpA | 13 | 176 | 7 | 36 | 0.318 |
| ComK | 6 | 175 | 6 | 23 | 0.247 |
| Spo0A | 5 | 31 | 7 | 7 | 0.207 |
| AbrB | 12 | 136 | 3 | 20 | 0.171 |
| SpoVT | 3 | 34 | 6 | 2 | 0.166 |
| CodY | 2 | 54 | 4 | 17 | 0.154 |
| LexA | 0 | 54 | 5 | 3 | 0.141 |
| TnrA | 3 | 57 | 3 | 17 | 0.131 |
| ResD | 1 | 52 | 3 | 13 | 0.116 |
| Phop | 1 | 38 | 3 | 10 | 0.107 |

Table 5C. Top ten most influencing regulators in the TRN of *P. aeruginosa*.

| TF | TFR | GR | SF | CR | G coefficient |
|------|-----|----|----|----|---------------|
| LasR | 6 | 89 | 3 | 20 | 0.155 |
| Fur | 14 | 54 | 3 | 6 | 0.114 |
| MexT | 1 | 46 | 3 | 15 | 0.107 |
| Vfr | 4 | 3 | 4 | 12 | 0.099 |
| AlgR | 1 | 29 | 3 | 14 | 0.097 |
| Anr | 2 | 42 | 2 | 12 | 0.086 |
| Ihf | 1 | 29 | 2 | 14 | 0.085 |
| PtxR | 1 | 11 | 3 | 12 | 0.082 |
| RhlR | 2 | 25 | 2 | 13 | 0.082 |
| AlgW | 1 | 1 | 3 | 7 | 0.062 |

We can observe that in each network there is a different influence of the global regulators; for the case of *E. coli* and *B. subtilis* the most influencing regulators are controlling several aspects of the central metabolism whereas for *P. aeruginosa* the

most influencing regulator is that of the main mechanism of quorum-sensing in this bacterium. The most influencing regulator in *E. coli* is affecting the expression of 46% of genes in their regulatory network, 31% for the case of *B. subtilis*, whereas their most global regulator affects only 15% of genes in *P. aeruginosa*.

5. Biological processes most represented in the TRNs

Identification of biological modules into the regulatory networks

As we advance in the description of the organization of biological networks it has been revealed that they have a modular organization; that is to say, the whole network seems to be subdivided in relatively autonomous, highly interconnected components. The genes of these modules are related mostly to the execution of defined biological functions [25]. In a previous study in *E. coli* [26], the authors assayed several computational methods trying to find the best metric that recovers functional modules as those curated manually. They concluded that the simple metric of $1/D^2$ was the best computational tool to recover the most of biological modules. In this metric D represents the minimum length of all the possible paths between vertices u and v .

We followed this procedure to identify functional modules in the three regulatory networks. Additionally we checked out these results with the modules reconstructed by reading the function performed by each of the regulators in scientific articles and grouping this information in different functional categories according to their biological functions (**Figures 7 A-C**).

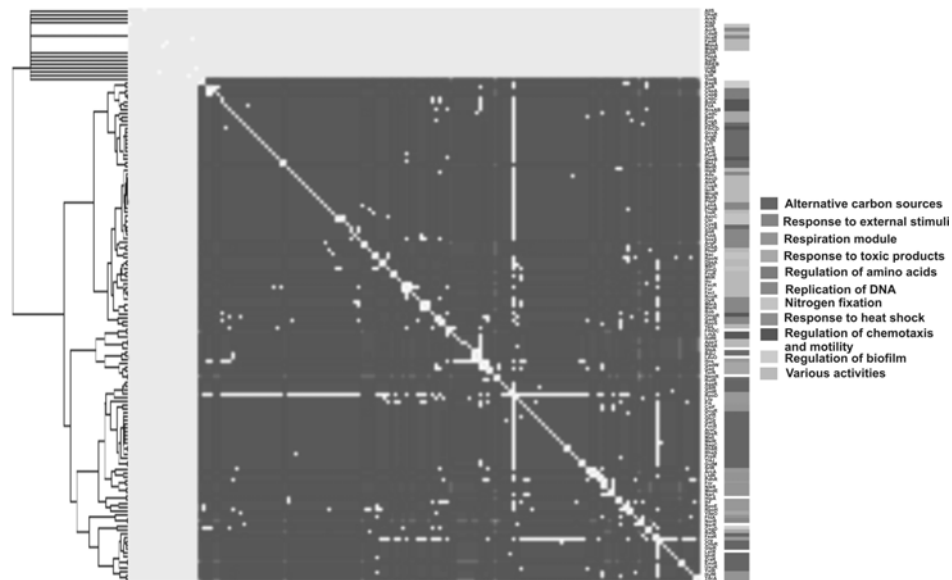


Figure 7A. Functional modules of genes into the *E. coli* regulatory network.

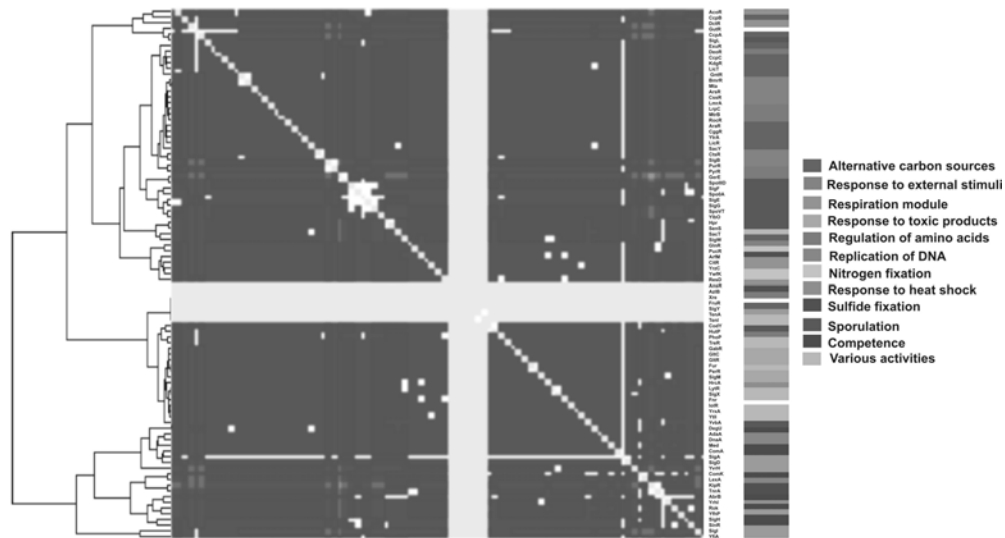


Figure 7B. Functional modules of genes into the *B. subtilis* regulatory network.

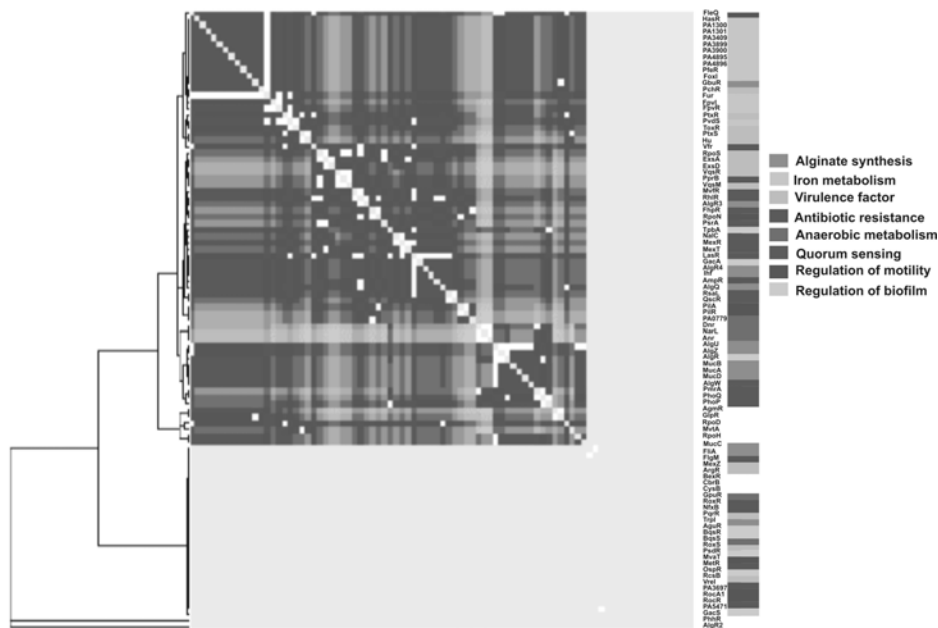


Figure 7C. Functional modules of genes into the regulatory network of *P. aeruginosa*.

Conclusions

In this study we have given a macromolecular description of regulatory networks in three bacterial models. We have shown that even when these networks cover in distinct amplitude the total amount of genes in each organism they conserve their general topological properties. Regarding their functional organization these networks conserve activation as the form of

regulation most common of their genes but their TFs are mostly auto-repressed. Our knowledge of genetic regulatory networks in bacterial models reflect the different aspects of biology that are being studied in each organism, *E. coli* is the most studied in all aspects of its physiology, which confirm its status as model organism, whereas pathogenic and virulence regulation is the only physiology studied in *P. aeruginosa*. We hope this general description of regulatory networks in bacterial models can foster the study of additional physiology in them to better study and compare the rewiring of regulatory networks in bacteria.

Acknowledgements

We thank Cheryl Lynn Gad for her critical comments to the ms. This work was supported by CONACYT grant 102854 given to AM-A. EG-V has a PhD fellow from CONACYT.

References

1. Pérez-Rueda, E., Janga, S. C., Martínez-Antonio, A. 2009. *Mol. Biosyst.* 5, 1494-1501.
2. Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B., Shao, Y. 1997, *Science.* 277, 1453-1462.
3. Pérez-Rueda, E., Collado-Vides, J. 2000. *Nucl. Acids Res.* 28, 1838-1847.
4. Gama-Castro, S., Salgado, H., Peralta-Gil, M., Santos-Zavaleta, A., Muñoz-Rascado, L., Solano-Lira, H., Jimenez-Jacinto, V., Weiss, V., García-Sotelo, J. S., López-Fuentes, A., Porrón-Sotelo, L., Alquicira-Hernández, S., Medina-Rivera, A., Martínez-Flores, I., Alquicira-Hernández, K., Martínez-Adame, Ruth., Bonavides-Martínez, C., Miranda-Ríos, J., Huerta, A. M., Mendoza-Vargas, A., Collado-Torres, L., Taboada, B., Vega-Alvarado, L., Olvera, M., Olvera, L., Grande, R., Morett, E., Collado-Vides, J. 2010, *Nucl. Acids Res.* 1-8.
5. Keseler, M. I., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñoz-Rascado, L., Bonavides-Martínez, C., Parey, S., Krummenacker, M., Altman, T., Kaipa, P., Spauding, A., Pacheco, L. M., Fulcher, C., Sarker, M., Shearer, G. A., Mackie, A., Paulsen, I., Gunsalus, R. P., Karp, P. D. 2011. *Nucl. Acids. Res.* 39, D583-D590.
6. Lopez, D., Vlamakis, H., Kolter, R. 2009, *FEMS Microbiol Rev.* 33, 152-163.
7. Kunst, F., Ogasawara, N., Moszer, I., Albertini, A. M., Alloni, G., Azevedo, V., Bertero, M. G., Bessieres, P., Bolotin, A., Borchert, S., Borriss, R., Boursier, L., Brans, A., Braun, M., Brignerll, S. C., Bron, S., Brouillet, S., Bruschi, C. V., Caldwell, B., Capuano, V., Carter, N. M., Choi, S. K., Codani, J. J., Connerton, I. F., Cummings, N. J., Daniel, R. A., Denicot, F., Devine, M., Düsterhöft, A.,

- Enrlich, S. D., Emmerson, P. T., Entian, K. D., Errington, J., Fabret, C., Ferrari, E., Foulger, D., Fritz, C., Fujita, M., Fijuta, Y., Fuma, S., Galizzi, A., Galleron, N., Chim, S. J., Glser, P., Goffeau, A., Golightly, E. J., Grandi, G., Guiseppin, G., Guy, B. J., Haga, K., Haiech, J., Harwood, C. R., Hénaut, A., Hilbert, H., Holsappel, S., Hosono, S., Hullo, M. F., Itaya, M., Jones, L., Joris, B., Karatama, D., Kasahara, Y., Klaer-Blanchard, M., Klein, C., Kobayashi, Y., Koetter, P., Koningstein, G., Krogh, S., Kumano, M., Kurita, K., Lapidis, A., Lardinois, S., Lauber, J., Lazarevic, V., Lee, S. M., Levine, A., Liu, H., Masuda, S., Mauël, C., Medique, C., Medina, N., Mellado, R. P., Mizuno, M., Moestl, D., Nakai, S., Noback, M., Noone, D., O'reilly, M., Ogawa, K., Ogiwara, A., et al. 1997, *Nature*. 390, 249-256.
8. Takahiro, I., Ken-ichi, Y., Terai, G., Yasutaro, F., Kenta, N. 2001, *Nucl. Acids Res.* 29, 278-280.
 9. Sierro, N., Makita, Y., De Hoon, M., Nakai, K. 2008. *Nucl. Acids Res.* 36, D93-D96.
 10. Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., Brinkman, F. S., Hufnagle, W. O., Kowalik, D. J., Lagrou, M., Garber, R. L., Goltry, L., Tolentino, E., Westbrook-Wadman, S., Yuan, Y., Brody, L. L., Coulter, S. N., Folger, K. R., Kas, A., Larbig, K., Lim, R., Smith, K., Spencer, D., Wong, G. K., Wu, Z., Paulsen, I. T., Reizer, J., Saier, M. H., Hancock, R. E., Lory, S., Olson, M. V. 2000, *Nature*. 406, 959-964.
 11. Klein, J., Leupold, S., Münch, R., Pommerenke, C., Johl, T., Kärst, U., Jansch, L., Jahn, D., Retter, I. 2008, *Nucl. Acids Res.* 36, W460-W464.
 12. Choi, C., Münch, R., Leupold, S., Klein, J., Siegel, I., Thielen, B., Benkert, B., Kucklick, M., Schobert, M., Barthelmes, J., Ebeling, C., Haddad, I., Scheer, M., Grote, A., Hiller, K., Bunk, B., Schreiber, K., Retter, I., Schomburg, D., Jahn, D. 2007, *Nucl. Acids Res.* 35, D533-D537.
 13. Winsor, G., Van, T., Lo, R., Bhavjinder, K., Whiteside, M., Hancock, R., Brinkman, S. 2009, *Nucl. Acids Res.* 2009, 37, D483-D488.
 14. Galán-Vásquez, E., Luna B., Martínez-Antonio, A., 2011. *Microbial Informatics and Experimentation*. 1:3, 1-11.
 15. Winsor, G. L., Lo, R., Ho Sui, S. J., Ung, K. S., Huang, S., Cheng, D., Ching, W. K., Hancock, R. E., Brinkman, F. S. 2005. *Nucl. Acids Res.* 33 D338-43.
 16. Shannon, P., Markiel, A., Ozier, O., Baliga, N., Wang, J., Ramage, D., Amin, N., Schwikowski, B., Ideker, T. 2003. *Genome Res.* 13, 2498-2504.
 17. Bondy, J. A., Murty, U. S. R. 1976, *Graph theory with applications*. Elsevier Science Publishing Co., Inc.
 18. Barabási, A. L., Oltvai, Z. N. 2004, *Nature Rev.* 5, 101.
 19. Lima-Mendez, G., van Helden, J. 2009. *Mol. BioSyst.* 5, 1482-1493.
 20. Björn, H., J., Falk, S. 2008, *Analysis of biological networks*. Wiley inter-science. USA.
 21. Milo, R., Shen-Orr, S., Itzkovitz, S., Kashtan, N., Chklovskii, D., Alon, U. 2002, *Science*. 298, 824.
 22. Uri Alon. 2007, *Nature Rev. Genetic.* 8, 450.
 23. Kashtan, N., Itzkovitz, S., Milo R., Alon U. 2004, *Physical Review*. E 70, 031909.

24. Martínez-Antonio, A., Collado-Vides, J., 2003. *Current Opinion In Microbiology*. 6, 482-489.
25. Warner, G. P., Pavlicev, M., Cheverud, J. M. 2007. *Nature Rev.* 8, 921-931.
26. Resendis-Antonio, O., Freyre-González, J., Menchaca-Méndez, R., Gutiérrez-Ríos, R., Martínez-Antonio, A., Avila-Sánchez, C., Collado-Vides J. 2005, *TRENDS in Genetics*. 21, 16.

PAPER

Structural comparison of biological networks based on dominant vertices†

Cite this: *Mol. Biosyst.*, 2013, **9**, 1765

Beatriz Luna,^{*a} Edgardo Galán-Vásquez,^b Edgardo Ugalde^c and Agustino Martínez-Antonio^{*b}

It is a current practice to organize biological data in a network structure where vertices represent biological components and arrows represent their interactions. A great diversity of graph theoretical notions, such as clustering coefficient, network motifs, centrality, degree distribution, etc., have been developed in order to characterize the structure of these networks. However, none of the existent characterizations allow us to determine global similarity among networks of different sizes. It is the aim of the present paper to introduce a mathematical tool to compare networks not only with regard to their topological structure, but also in their dynamical capabilities. For this reason we aim to propose a pseudo-distance between networks, built around the notions of determination and dominance, concepts recently introduced in the context of regulatory dynamics on networks. We use our proposed pseudo-distance to compare networks from the following bacteria: *E. coli*, *B. subtilis*, *P. aeruginosa*, *M. tuberculosis*, *S. aureus* and *C. glutamicum*. We also use this pseudo-distance to compare these real bacterial networks with equivalent homogeneous, scale-free and geometric three dimensional random networks. We found that even when bacterial networks are characterized with different levels of detail, have different sizes and represent different aspects of the organisms, the proposed pseudo-distance captures all these characteristics, and indicates how similar they are or not from random networks.

Received 21st February 2013,
Accepted 9th May 2013

DOI: 10.1039/c3mb70077a

www.rsc.org/molecularbiosystems

1 Introduction

We consider gene expression as the process by which genes coded in DNA are transcribed into RNA, then, this messenger RNA is used by the translation machinery to synthesize proteins. Proteins are necessary in several functions performed by organisms. Frequently, the protein produced by a certain gene is responsible for the activation or inhibition of another gene; that is to say, it promotes (enhances) or suppresses (restricts) its expression. The overall gene interactions architecture is encoded in a network where vertices symbolize genes and edges represent their regulatory interactions. Therefore, regulatory networks represent a complex set of highly interconnected processes that govern the rate at which the genes in a cell are expressed.¹ Network representation allows working with a large quantity

of data, which can be quantified and be analyzed by means of the tools supplied by graph theory. Indicators like degree distribution, clustering coefficient, cycles distribution, and network motifs^{1–3} have been used to characterize and understand the topological complexity of a given network. In order to understand the relation between topology and function it is essential to compare the structure of networks performing similar functions. However, the lack of information and the size differences between networks with similar dynamical behavior makes this type of global comparison difficult. Recent studies have dealt with this problem proposing different approaches, for instance:

- Comparison of the interconnections of different sub-graphs into each network.⁴
- Identification of interacting patterns of regulation and genetic sequence similarities between nodes.⁵
- Recognition of modular network components, followed by projection of these modules onto networks of other species.⁶
- Quantification of the structural differences between two networks using the normalized graph spectra.⁷

On the other hand, comparative analyses have been also proposed for metabolic pathways considering:

- The pathway topology and the number of enzymes they contain.⁸

^a *Scientific Computing and Applied Mathematics Laboratory, Universidad del Papaloapan, Av. Ferrocarril s/n 68400 Loma Bonita, Oaxaca, México. E-mail: bcluna@unpa.edu.mx*

^b *Departamento de Ingeniería Genética, Cinvestav, Km. 9.6 Libramiento Norte Carr. Irapuato León, México. E-mail: amartinez@ira.cinvestav.mx*

^c *Instituto de Física, Av. Manuel Nava 6, zona Universitaria, 78290 San Luis Potosí, SLP, México*

† Electronic supplementary information (ESI) available. See DOI: 10.1039/c3mb70077a

- The phylogenetic tree built from the structural information inherent in the metabolic pathways.⁹
- The compression of vertices and edges.¹⁰

More recently a technique based on the determination of rewiring rates has been proposed to compare genetic interacting networks, metabolic pathway networks, transcription, phosphorylation and social coauthorship networks.¹¹

The motivation behind this study is twofold: first, to provide an alternative way to compare available regulatory networks; second, to understand the information that comparisons can yield. Regarding the first point, until recently only the *E. coli* network had been reconstructed from the literature, however, new high-throughput methodologies allow the massive reconstruction of regulatory networks from bacteria and other organisms. Focusing on bacteria, under supposition that they have a common origin, one way to compare them might be based on phylogenetic analysis, which depends on orthologous genes, however, this approach is limited, since the phylogenetic divergency increases as the bacterial genus. Therefore it is necessary to appeal to methods developed from systems biology theory. From this perspective, we focus on the analysis of the elements of the regulatory genome, the transcription factors, particularly on those crucial elements for regulation. Regarding the second point, this analysis allows us to discern between real biological networks and random networks, but most important, it allows us to decipher the salient elements/arrangements on biological networks associated with interesting traits: *i.e.*, pathogenic *vs.* non-pathogenic, adaptation to different life conditions and intracellular *vs.* free/living styles, among others.

We propose a global measure, based on the concepts of determination and dominancy, which have been recently introduced and explored. As shown by Luna and Ugalde,¹² the global state of the network can be reconstructed from the history of a sub-collection of vertices they call dominant vertices through a series of steps following a hierarchical organization of the network. In the same study, dominant vertices and vertex determination were used as indicators of the structural complexity of the network. Here, we define a pseudo-distance between networks by using the hierarchy imposed by the dominancy and vertex determination. We use this pseudo-distance to compare bacterial regulatory networks and additionally to compare these real biological networks with random network models, as performed in ref. 10 and 13.

In the next section we present some preliminary concepts and the bacteria whose transcriptional networks will be examined. In Section 3 we give the definition of our main analytical tool and we describe the computational method that was applied to the six bacterial regulatory networks and equivalent random networks. Section 4 contains results and discussion. Finally, Section 5 is devoted to the final remarks and conclusions.

2 Preliminaries

2.1 Dominant vertices

The dynamics of regulatory networks is usually modeled by means of piecewise-contractive dynamical systems.^{14,15} For these kind of

dynamical models, it is possible to distinguish a set of vertices in a network that allows recovery of the global configuration of the network based just on the history of those vertices. Conversely, one can determine the future configurations the network will have by imposing the successive states of the systems on the dominant vertices. Every nonempty directed network admits a dominant set. The rigorous definition of the dominant vertices and the statement of their properties, were widely introduced by Luna and Ugalde.¹²

Here we use the word network as a synonym of finite directed graph. It consists of a finite set of vertices, denoted by V , and a finite collection of ordered pairs of vertices, which we identify with arrows and denote by A . We only consider simple networks, which means that given two vertices $u, v \in V$ there exists at most an arrow going from u to v , which corresponds to the ordered pair (u, v) . The input set of a vertex v is the collection $I(v)$ of all vertices included in an arrow ending at v , *i.e.*, $I(v) := \{u \in V : (u, v) \in A\}$. The input degree of $v \in V$ is simply the cardinality of $I(v)$. Similarly, the output set of v is the set $O(v)$ of all vertices included in an arrow starting at v , *i.e.*, $O(v) := \{u \in V : (v, u) \in A\}$. The output degree of $v \in V$ is nothing but the cardinality of $O(v)$.

A regulatory network is a dynamical system involving several interconnected units, whose dynamic rule is determined by the interconnected topology. It is therefore a dynamical system over a network. The interacting units (which in our case correspond to gene products) have a numerical value (which in our case corresponds to its relative concentration) at each time step. The time can be considered continuous, in which case the system is modeled by a system of piecewise affine differential equations, or can be considered discrete, in which case the pertinent model is a piecewise affine coupled map network. To fix these ideas, let us take the discrete time case. Here the global configuration of the system is recorded at times $t_0 < t_1 < \dots < t_n < \dots$, which by simplicity, we identify with the integer numbers progression. After the enumeration of the interacting units, (v_1, v_2, \dots, v_N) , the configuration of the system at any time t can be encoded in an N -dimensional array $(x_1^t, x_2^t, \dots, x_N^t)$, where N is just the number of interacting units, and $x_i^t \in \mathbb{R}$ denotes the level of activity of the i -th unit at time t . The dynamic rule governing the state of the system $(x_1^{t+1}, x_2^{t+1}, \dots, x_N^{t+1})$ at time $t + 1$, from its configuration at time t , reads, for vertex v_j , as

$$x_j^{t+1} = \alpha_j x_j^t + D_j(x_k^t : v_k \in I(v_j)), \quad (1)$$

where $\alpha_j \in (0, 1)$ accounts for the degradation rate of v_j , while D_j is a function which depends on the activity levels of the vertices in the set of inputs of the referred vertex. The function D_j takes a finite number of values, which can be related to the levels of activation affecting the vertex v_j . This function is analogous to the so-called regulatory phrase of Boolean and logical networks,¹⁶ this analogy allows to obtain similar results from these models. The dynamics so defined is such that the asymptotic activity level of the unit v_j is completely determined by the sequence of activating levels affecting this unit in the course of time. This characteristic allows the reconstruction of

the asymptotic dynamics of the whole system from solely the history of a well-chosen collection of vertices. The reconstruction goes as follows: let $V_0 \subset V$ be the collection of dominant vertices, and suppose that the activity level x_k^t is known for all the course of time t at every one of the dominant vertices $v_k \in V_0$. Now, consider the set $\partial V_0 := \{v \in V: I(v) \subset V_0\}$, i.e., the collection of vertices whose input set is a sub-collection of the dominant vertices. The dynamics of the system is such that, for each $v_k \in \partial V_0$ and all time τ sufficiently large, x_j^τ depends only on the sequence $(x_i^t: v_i \in V_0)_{i=0}^{\tau-1}$, i.e., x_j^τ can be determined, up to a negligible error, by the history of the dominant vertices up to time $\tau - 1$. Now, the history of the interacting units in the set $V_1 = V_0 \cup \partial V_0$ containing the dominant vertices and vertices determined by them, determines the state of vertices in the set $\partial V_1 := \{v \in V: I(v) \subset V_1\}$ and defines an enlarged set $V_2 = V_1 \cup \partial V_1$. We can continue this enlargement procedure, based on vertex determination, until no new vertices can be added. In order for V_0 to be dominant it is necessary that at the end of this iterative process all the vertices from the network are included, i.e., if there exists a nested sequence of vertex sets, $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, such that $V_{i+1} = V_i \cup \partial V_i$, $i = 1, \dots, d$. Here,

$$\partial V_i := \{v \in V: I(v) \subset V_i\} \quad (2)$$

for each $V_i \subset V$. The number d of steps needed to determine all the vertices from the dominant seed is the depth of the network with respect to the chosen dominant set V_0 , here referred to as hierarchical levels or dominance levels. It is worth noting that a given network can have several sets of dominant vertices. A trivial choice is $V_0 = V$. Obviously, a valid choice would be a dominant set of minimal cardinality. Luna and Ugalde¹² proposed an algorithm aimed to identify dominant sets of minimal cardinality, as well as the hierarchy of sets it determines. In what follows we will make extensive use of that algorithm. It is important to notice that this paper is dedicated to a structural comparison, for that reason we do not make use of the dynamical properties of the dominant set, however given its dynamical characteristics, already proved, it is possible to perform a dynamical comparative analysis.

2.2 Connected components

A network (V, A) is said to be simply connected if any couple of vertices $u, v \in V$ can be joined through an undirected path, i.e., if there are vertices $u \equiv u_1, u_2, \dots, u_\ell \equiv v \in V$ such that for each $1 \leq i \leq \ell - 1$, either (u_i, u_{i+1}) or (u_{i+1}, u_i) belong to the arrow set A . If (V, A) satisfies the stronger condition requiring that any two vertices can be joined through a directed path, then the network is said to be strongly connected. It is not difficult to prove that any network can be decomposed in a unique way into a collection of connected sub-networks called connected components (CC). This concept will be mentioned later.

2.3 Random networks

Formally speaking, a random network is a probability distribution over a specific set of networks, for instance, the set of all networks on N vertices, or the set of all finite networks, or the set of all the infinite networks with bounded input and

output degrees, etc. These objects were introduced by P. Erdős and A. Rényi.¹⁷ These authors were interested in the connectivity properties of random homogeneous graphs. Roughly speaking, they considered distributions over the set of networks on N vertices, giving the same or almost the same probability to all the simple symmetric networks having a fixed number of input–output degrees.

A class of random networks, intended as models of complex evolving structures was introduced by A. Barabási and R. Albert.² A random process that incorporates continuous growth and preferential attachment usually defines these structures, known as scale free networks. There are several ways to implement this kind of construction, leading to the similar statistical properties.¹⁸ One can for instance start with the complete undirected graph on a small number of vertices, then, at each time step, add a new vertex forming new edges (bidirectional arrows) with randomly chosen vertices in the preexisting network. In order to ensure the preferential attachment, a vertex is chosen to form a new edge with a probability proportional to its degree. The process would continue until a desired number of vertices are obtained.

More recently, random geometric networks were introduced as models of biological networks.^{4,19} We use the three dimensional model, where every vertex of a set is randomly placed at a unit cube, and vertices are connected if the Euclidean distance between them is less than a parameter r .

2.4 Description of the transcriptional networks here studied

In this study we consider six transcriptional regulatory networks pertaining to six bacteria: *E. coli*, whose information was obtained from RegulonDB;²⁰ *B. subtilis*, obtained from DBTBS;²¹ *P. aeruginosa*;²² *M. tuberculosis*;²³ *C. glutamicum* obtained from the CoryneREGNet database;²⁴ and *S. aureus*.²⁵ Below we briefly describe the principal attributes for each of these networks, also shown in Table 1.

- *Escherichia coli* (EC) is a Gram-negative bacterium and the bacterial model of study par excellence. The genome sequence of strain K-12 MG1655 was published in 1997; it has 4.6 Million base pairs (Mbp) and contains 4400 predicted genes.²⁶ The predicted transcription factors (TFs) are around 300, and 7 sigma factors.²⁷ The strain of reference is a benign commensal in the gut of mammals but some other strains could be pathogenic.

- *Bacillus subtilis* (BS) is the model of a Gram-positive bacterium. It is well known for its capability to produce, from vegetative growing cells, metabolically inactive spores.²⁸ The distinctive

Table 1 Components of the transcriptional regulatory networks of bacteria here analyzed

| Bacteria | Genes on the network | Interactions | TF genes | Target genes | σ factors | CC |
|----------|----------------------|--------------|----------|--------------|------------------|----|
| EC | 1565 | 3530 | 176 | 1382 | 7 | 18 |
| BS | 1659 | 3121 | 106 | 1531 | 16 | 9 |
| PA | 690 | 1020 | 76 | 593 | 14 | 12 |
| MT | 1624 | 3212 | 83 | 1531 | 10 | 6 |
| CG | 574 | 806 | 71 | 499 | 4 | 27 |
| SA | 558 | 671 | 46 | 512 | 0 | 21 |

strain is the so named 168, its genome has 4.2 Mbp and contains 4422 predicted genes.²⁹ The estimated number of TFs is around 275, and the number of sigma factors is about 14.³⁰ This bacterium normally inhabits soil.

- *Pseudomonas aeruginosa* (PA) is a metabolically versatile Gram-negative bacterium. It expresses a wide range of virulence factors that allow it to be an opportunist pathogen of plants and animals.³¹ The genome sequence of strain PAO1 has 6.2 Mbp that contains 5570 predicted genes.³² The estimated number of TFs is around 475, and the number of sigma factors is about 26.²⁷

- *Mycobacterium tuberculosis* (MT) strain H37Rv is a pathogenic bacteria of the genus *Mycobacterium* and the causative agent of most cases of tuberculosis in humans, one of the most infectious diseases.³³ Its genome sequence was released in 1998, it consists of 4.41 Mbp and contains approximately 4047 predicted genes. The number of predicted regulatory proteins is approximately 190 TFs and 13 sigma factors.³⁴

- *Corynebacterium glutamicum* (CG) strain ATCC 13032 is a Gram-positive bacterium, and it is well known as the organism for the commercial production of monosodium glutamate. Its genome was published in 2003, it consists of 3.3 Mbp and contains 3002 predicted genes.³⁵ The number of predicted regulatory proteins is 182 TFs and 7 sigma factors.³⁶

- *Staphylococcus aureus* (SA) strain N315 is a Gram-positive bacterium and it is an important human pathogen. It has been intensively studied in virulence, biofilm formation and central metabolism. Its genome sequence was published in 2001, it consists of 2.8 Mbp and contains 2593 predicted genes.³⁷ The number of predicted regulatory proteins is 120 for TFs and 4 for sigma factors.³⁸

3 Methods

We began our analyses of bacterial networks by using the algorithm mentioned in ref. 12 (which we briefly describe below), then we computed the sets of dominant vertices of minimal cardinality, and we determined the associated hierarchical levels of these sets for each one of the six bacterial networks mentioned above. After that, we used the pseudo-metric M , defined below, to compare these networks.

In the second part of the study, we generated 100 homogeneous random networks, 100 scale free networks and 100 geometric random graphs with the same number of nodes and on average the same number of arrows of every one of the six bacterial networks considered. By using this random sample we estimated the mean M -distance for each bacterial network, as well as the variance of this distance. The results of these numerical experiments are presented in Section 4.

3.1 Computation of dominant sets

We determine dominant sets by using the following algorithm: first, we select all the vertices with input degree equal to zero, together with the vertices with the highest output-degree/input-degree ratio. Then we compute the hierarchy of sets of vertices so determined by this collection. If the hierarchy does not

include all the vertices, then we restart the iteration by adding vertices to the starting dominant set until all the vertices are included in the hierarchy (still considering the highest output/input-degree ratio criterion). At the end of this process some dominant vertices are eliminated in order to obtain a dominant set of minimal cardinality; elimination could be performed by different methods, for instance randomly or by ordering the vertices in the dominant set. It is recommended to apply the same algorithm to every set of networks to compare. In our case we chose the ordering criterion. In this way, we obtain a nested sequence of sets $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, such that $V_{\ell+1} = V_\ell \cup \partial V_\ell$, $\ell = 1, \dots, d$.¹²

3.2 Hierarchy obtained by dominant vertices

The nested sequence of sets $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, defines a hierarchy of V_0 . What we call the dominant set of vertices corresponds to the first level of the hierarchy. Then for each $1 \leq \ell \leq d$, the determined set ∂V_ℓ constitutes the $\ell + 1$ -th level of the hierarchy. The depth of this hierarchy corresponds to the number d of levels needed to reach all the vertices in the network, starting from the dominant set.

In order to illustrate these notions, consider the network depicted in Fig. 1. Here, a dominant set of minimal cardinality is $V_0 = V_1 = \{1, 4\}$, and the hierarchy determined for that set is the following: $\partial V_1 = \{2\}$, $V_2 = \{1, 4, 2\}$, $\partial V_2 = \{3, 5\}$, $V_3 = \{1, 2, 3, 4, 5\} = V$. The depth in this network is $d = 2$.

3.3 A pseudo-distance based on dominancy

The hierarchical decomposition of networks obtained from dominant sets of minimal cardinality provides us with a tool to compare networks of different sizes and topologies, but sharing global dynamical features, as follows. Given two networks $G := (V, \mathcal{A})$ and $G' := (V', \mathcal{A}')$ with hierarchical decomposition $V_0 \equiv V_1 \subset V_2 \subset \dots \subset V_d \subset V_{d+1} \equiv V$, and $V'_0 \equiv V'_1 \subset V'_2 \subset \dots \subset V'_d \subset V'_{d+1} \equiv V'$, define

$$m(G, G') = \left| \frac{\#V_0}{\#V} - \frac{\#V'_0}{\#V'} \right| + \sum_{\ell=1}^D \left| \frac{\#\partial V_\ell}{\#V} - \frac{\#\partial V'_\ell}{\#V'} \right| \quad (3)$$

where $D := \max(d, d')$ and $\#S$ denotes the cardinality of any set S . It is worth noting that m depends on the chosen dominant set of vertices and that in general this choice is not unique. Nevertheless, for dominant sets of minimal cardinality the characteristics of the associated decomposition are very similar. In our case, since our algorithm chooses a particular dominant set of minimal cardinality, the pair function m is well defined.

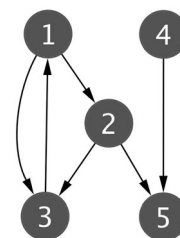


Fig. 1 Illustrative network. Its hierarchy is shown according to dominant vertices.

It is symmetric, non-negative, and satisfies the triangle inequality. It can vanish for non-identical networks, and for this reason is not a distance but a pseudo-distance. It is not difficult to exhibit networks $G \neq G'$ for which $m(G, G') = 0$. In this case the networks G and G' have decompositions with the same number of levels and the same proportion of vertices at each level. This pseudo-distance allows for comparing networks regardless the respective number of vertices, arrows and connected components. It gives a global idea of the network structure.

A refinement of the pseudo-distance m is the following

$$M(G, G') = \frac{1}{2} \left(m(G, G') + \left| \frac{d+1}{\#V} - \frac{d'+1}{\#V'} \right| \right) \quad (4)$$

which weights in the pseudo-distance m as the relative number of levels in each decomposition. Clearly $M(G, G') = 0$ implies $m(G, G') = 0$, and whenever $m(G, G') > 0$, the pseudo-distance M magnifies the effect of networks having relatively many levels.

3.4 Random networks

As mentioned above, besides the comparison between the six bacterial networks, we also compare each one of them to random networks with the same number of vertices and approximately the same number of arrows. We proceed as follows:

- To generate homogeneous random networks we estimate the connection probability p by using the mean output degree, *i.e.*, for a given bacterial network (V, A) , we estimate $p = \#A/(\#V^2 - \#V)$. We obtain values $p = 0.001$ in the case of *E. coli*, *B. subtilis*, and *M. tuberculosis*, and $p = 0.002$ for *P. aeruginosa*, *C. glutamicum*, and *S. aureus*.

- For scale-free networks we compute the output degree distributions for each bacterial network, then we randomly connect vertices in order to obtain a network with the same output degree distribution as the real bacterial networks. The input degree distribution turns out to be Poissonian in this kind of network.

- To generate random geometric networks we consider every given bacterial network (V, A) . For every element of V we place randomly one point in the unit cube. For every pair of vertices u, v , an arrow (u, v) , or (v, u) , or both are placed in the cube with probability 0.4, 0.4 and 0.2 respectively, if $d < r$, where d is the Euclidean distance. Values of r were considered such that the number of arrows in the real network is approximately the same as that in real bacterial networks. For this study we used $r = 0.085$ in the case of *E. coli* and *M. tuberculosis*, $r = 0.098$ for *P. aeruginosa*, $r = 0.1$ for *C. glutamicum*, and $r = 0.094$ for *S. aureus*.

4 Results and discussion

4.1 Comparison of bacteria networks

By using the pseudo-distances m and M , we compare the six bacterial regulatory networks. The main results are shown in Table 2. As explained before, numbers close to zero indicate a high similarity between the network structures, according to the hierarchy provided by dominant vertices of the compared networks. Both pseudo-distances give similar results in most of cases.

Table 2 Comparison of the six bacterial networks by using the pseudo-distances m and M

| | EC | BS | PA | MT | CG | SA |
|----|----|--------------------------|--------------------------|--------------------------|--------------------------|--------------------------|
| EC | 0 | $m = 0.35$ $M = 0.17$ | $m = 0.53$ $M = 0.27$ | $m = 0.59$ $M = 0.29$ | $m = 0.86$ $M = 0.43$ | $m = 1.12$ $M = 0.56$ |
| BS | | 0 | $m = 0.56$ $M = 0.28$ | $m = 0.44$ $M = 0.22$ | $m = 0.64$ $M = 0.32$ | $m = 0.87$ $M = 0.44$ |
| PA | | | 0 | $m = 0.59$ $M = 0.30$ | $m = 0.69$ $M = 0.35$ | $m = 0.86$ $M = 0.43$ |
| MT | | | | 0 | $m = 0.30$ $M = 0.15$ | $m = 0.57$ $M = 0.29$ |
| CG | | | | | 0 | $m = 0.44$ $M = 0.22$ |
| SA | | | | | | 0 |

Table 3 Percentage of vertices by level for each bacterial network

| EC | BS | PA | MT | CG | SA |
|-------|-------|-------|-------|-------|-------|
| 4.5% | 2.3% | 7.3% | 2.8% | 8.3% | 5.5% |
| 20% | 34.3% | 32.0% | 48.8% | 52.7% | 75.0% |
| 34.9% | 33.6% | 11.3% | 24.1% | 19.6% | 10.9% |
| 9.4% | 8.3% | 11.1% | 10.2% | 16.0% | 8.4% |
| 13.2% | 4.4% | 13.9% | 9.7% | 3.1% | |
| 5.3% | 7.7% | 5.0% | 1.5% | | |
| 10.1% | 6.6% | 7.3% | 2.7% | | |
| 1.4% | 2.5% | 1.7% | | | |
| 0.74% | | 7.2% | | | |
| 0.06% | | 2.7% | | | |

As we can observe, *M. tuberculosis* and *C. glutamicum* attain the smallest M -distance. Although these networks are significantly different in size, they have similar global structures. The second smallest M -distance is obtained from the comparison between *E. coli* and *B. subtilis* networks. They have 4.53% and 2.35% of dominant vertices, respectively (Table 3). The largest M -distance, which indicates the greatest dissimilarity, was obtained by comparing *E. coli* and *S. aureus*.

Fig. 2 shows the network hierarchies determined by the dominant vertices. This characterization of the networks gives a global picture allowing a qualitative comparison between networks. Note the higher similarity between the regulatory

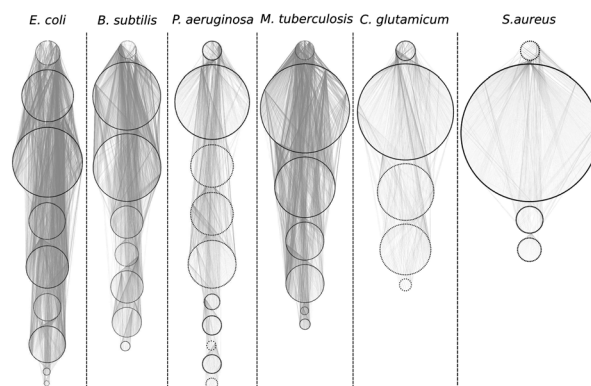


Fig. 2 Hierarchical distribution of vertices on bacterial networks. We show the hierarchies arranged according to the dominant vertices for the six bacterial transcriptional regulatory networks. Each circle size is proportional to the number of genes per level.

networks of *M. tuberculosis* and *C. glutamicum*, and *E. coli* and *B. subtilis*.

4.2 Comparison between real-bacteria and random networks

In Tables 4 and 5 we show the results of the comparison between bacterial networks and their corresponding random homogeneous, scale-free and geometric random networks. In all cases, homogeneous random networks have a larger set of dominant vertices, between 20% and 30% of all the vertices, and a higher number of levels. In contrast, scale-free random networks have a depth comparable to the corresponding real bacterial networks, but as the homogeneous random networks maintain a comparatively much larger set of dominant vertices with respect to the real ones. On the other hand geometric random networks have the largest set of dominant vertices with around 50%, but depth is not as big as in homogeneous random networks.

In Fig. 3 we show an example of the homogeneous (HR), scale free (SF), and geometric (GEO) random networks corresponding

Table 4 Comparison among bacterial regulatory networks and their corresponding random networks

| | EC | BS | PA |
|--------------------------------------|--------|--------|--------|
| Bacterial | | | |
| Percentage of dominant vertices | 4.60% | 3.90% | 7.39% |
| Depth | 9 | 7 | 9 |
| Homogeneous | | | |
| Mean percentage of dominants | 19.37% | 20.62% | 25.48% |
| Mean depth | 159.36 | 141.44 | 541.92 |
| Mean <i>m</i> -distance to bacterial | 1.4 | 1.43 | 1.12 |
| Mean <i>M</i> -distance to bacterial | 0.7 | 0.75 | 0.59 |
| Scale free | | | |
| Mean percentage of dominants | 10.41% | 13.89% | 23.00% |
| Mean depth | 14.65 | 10.8 | 10.04 |
| Mean <i>m</i> -distance to bacterial | 0.77 | 0.75 | 0.66 |
| Mean <i>M</i> -distance to bacterial | 0.39 | 0.38 | 0.33 |
| Geometric | | | |
| Mean percentage of dominants | 39.8% | 39.66% | 44.33% |
| Mean depth | 10.9 | 11.69 | 7.43 |
| Mean <i>m</i> -distance to bacterial | 0.78 | 0.81 | 0.81 |
| Mean <i>M</i> -distance to bacterial | 0.39 | 0.40 | 0.41 |

Table 5 Comparison among bacterial regulatory networks and their corresponding random networks

| | MT | CG | SA |
|--------------------------------------|--------|--------|--------|
| Bacterial | | | |
| Percentage of dominant vertices | 2.83% | 8.36% | 5.55% |
| Depth | 6 | 4 | 3 |
| Homogeneous | | | |
| Mean percentage of dominants | 20.13% | 27.37% | 33.63% |
| Mean depth | 152.89 | 43.97 | 27.89 |
| Mean <i>m</i> -distance to bacterial | 1.46 | 1.26 | 1.26 |
| Mean <i>M</i> -distance to bacterial | 0.77 | 0.66 | 0.65 |
| Scale free | | | |
| Mean percentage of dominants | 12.86% | 24.00% | 30.00% |
| Mean depth | 10.08 | 8.09 | 5.89 |
| Mean <i>m</i> -distance to bacterial | 0.87 | 0.82 | 1.00 |
| Mean <i>M</i> -distance to bacterial | 0.43 | 0.41 | 0.50 |
| Geometric | | | |
| Mean percentage of dominants | 39.47% | 46.63% | 50.89% |
| Mean depth | 11.24 | 6.73 | 5.99 |
| Mean <i>m</i> -distance to bacterial | 0.80 | 0.80 | 0.99 |
| Mean <i>M</i> -distance to bacterial | 0.40 | 0.40 | 0.50 |

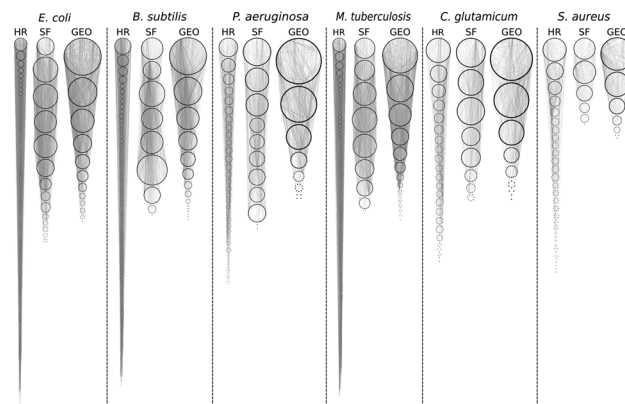


Fig. 3 Hierarchy of random networks. Homogeneous (HR), scale free (SF), and geometric (GEO) networks corresponding to each real bacterial network.

to each bacterial network: *E. coli* (HR $d = 136$, SF $d = 13$, GEO $d = 11$), *B. subtilis* (HR $d = 98$, SF $d = 8$, GEO $d = 14$), *P. aeruginosa* (HR $d = 34$, SF $d = 9$, GEO $d = 7$), *M. tuberculosis* (HR $d = 114$, SF $d = 8$, GEO $d = 14$), *C. glutamicum* (HR $d = 22$, SF $d = 7$, GEO $d = 7$) and *S. aureus* (HR $d = 27$, SF $d = 5$, GEO $d = 5$).

In Fig. 4 we show the proportion of vertices by level in the hierarchical decomposition of each bacterial network, and the mean proportion of vertices in their corresponding homogeneous, scale-free and geometric random networks. This hierarchical decomposition allows us to clearly identify some salient topological characteristics on each kind of network and distinguish it between real bacterial networks from homogeneous, scale-free and geometric random networks.

4.3 Biological analysis

The bacterial networks here studied have different numbers of vertices, and in general, each network represents the description of different biological aspects, and at a different depth, on each bacterium. This is due to the motivations behind the research studies of these organisms. *E. coli* is possibly the bacteria where the study of different aspects of the biological organism are better balanced, in contrast with *P. aeruginosa* and *B. subtilis* where virulence and sporulation are the biological processes better described. *S. aureus* is the bacterium whose network more poorly describes its biology. Having said this, we warn our lectors that these are not the final results and interpretations of the comparison of the regulatory networks in these organisms but we think this is a beginning to achieve that objective. Next we will describe the most representative biological aspects of each network, considering for this discussion, the biological role of the products of the genes regulated for the transcriptional and sigma factors on each level of dominance.

The first level in the TRN of *E. coli* is enriched with regulators for the transport and catabolism of carbon sources as well as for metal homeostasis and sensory systems of two-components. The second level is dominated by regulators of amino acids metabolism and for the response to exogenous stresses. The third level is enriched with regulators controlling the central metabolism. Regulators for exogenous response are

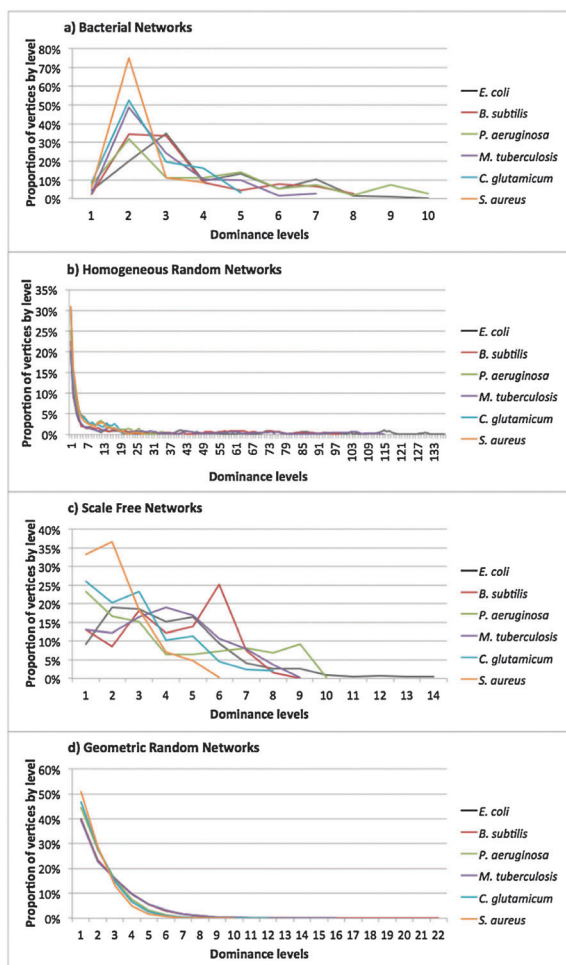


Fig. 4 Distribution of vertices by level on real and random networks. In the X axis we show the level of dominance according to the hierarchies, in the Y axis we show the proportion of vertices by level of dominance. The real bacterial networks (a), random homogeneous networks (b), scale-free random networks (c), and geometric random networks (d).

most representative in the fourth and fifth levels whereas regulators for response to endogenous stresses, motility and biofilm are located in the lowest hierarchical levels (ESI[†]). From the seven global regulators in *E. coli*,³⁹ 3 are positioned at the first level (*FNR*, *FIS* and *H-NS*), 3 in the second level (*CRP*, *ArcA* and *Lrp*) and *IHF*, the global regulator mostly expressed in the stationary phase, is located until the fourth level. The distribution of sigma factors is uniform; *RpoD* and *FecI* at the first level, *RpoE* and *RpoS* at the third, *RpoN* at the fourth, and *RpoH* and *FlaA* at the fifth and seventh levels. The most regulating and regulated TFs are shown in the ESI[†]. In a gross description the most regulating TFs are positioned on the upper part of the network hierarchy and correspond to the so called global regulators whose regulated genes make a great diversity of biological functions. On the contrary, the most regulated TFs are located on the lower part of the network hierarchy and correspond to regulators controlling the best defined biological processes; namely, response to drugs and endogenous stresses and the master regulators for biofilm and flagella formation.

For *B. subtilis*, the first level of the network hierarchy is enriched with regulators for response to exogenous stresses, sensory systems of two-components, regulators for amino acids metabolism and for the use of carbon sources. The second level is the most populated with regulators and is dominated by those regulators controlling responses to endogenous and exogenous stresses, the metabolism of amino acids and carbon sources, those regulating the secretion of proteolytic enzymes, competence and response to metals are also present on this level. In the third level the presence of anti-sigma factors and regulators of aerobic/anaerobic respiration is notable. The only biological process which has regulators in all the seven hierarchical levels is that for sporulation; being those which control the late spore coat and late mother cell on the lowest levels. The most regulating and regulated TFs are shown in the ESI[†]. Similarly to *E. coli*, the most regulating TFs are in the upper part of the network whereas the most regulated ones are in the lowest level of hierarchy. Remarkably, these last regulators are controlling genes for competence and sporulation processes.

In *P. aeruginosa* the first level is dominated by regulators controlling the alginate biosynthesis, virulence and response to external stresses, regulators of respiration, motility and biofilm formation are also concentrated on this level. At the second level there are regulators for response to external stresses and for quorum sensing. The regulators of siderophores are concentrated on the seventh level, being the regulators of the synthesis of pyoverdine and exotoxin A on the lowest levels. The regulators for alginate biosynthesis and quorum are located on the first three hierarchical levels whereas those for virulence, siderophores and anti-sigma factors are more distributed at all levels. The most regulating and regulated TFs are shown in the ESI[†]; unusually in this organism the most regulating gene is also one of the most regulated (*lasR*).

In *M. tuberculosis* the first level is dominated by regulators for response to exogenous stresses, diverse sigma factors, two-component sensory systems and for response to metals. On the second level there are regulators for response to exogenous stresses and metals. On the third level there are regulators for response to endogenous stress and for the production of virulence factors. In the ESI[†] are enumerated the most regulating and regulated TFs in *M. tuberculosis*. The most regulating genes are located in the upper part of the network.

In *C. glutamicum* the first level is enriched with regulators for response to external stresses, for the use of carbon sources and response to metals, the metabolism of amino acids, cofactors and nucleosides are also well represented at this level. The second level is mostly populated with regulators of amino acid metabolism. The ESI[†] shows the most regulating and regulated TFs in this bacterium.

In *S. aureus* the use of carbon sources, metals response and the metabolism of amino acids are controlling most of the regulators, as well as for the biosynthesis of cofactors, nucleosides and the production of virulence factors. In the second level there are regulators of carbon sources and amino acids metabolism. The ESI[†] shows the more regulating and regulated TFs in this bacterium.

A gross interpretation of the function of the dominant set of vertices in all six bacteria here analyzed is that they control fundamental aspects for survivability, that is to say, those functions related with the interaction with the surroundings by means of sensory systems of two-components, transport and efflux of small molecules and metals. In the middle part of the networks there are regulators for the control of the central metabolism and for the biosynthesis of cofactors (*e.g. biotin*) and nucleosides, it is also common to find regulators controlling the response to endogenous stresses as well as those produced by the bacterial metabolism (*e.g. organic acids*). The lowest part of the networks are populated with genes of the so called development processes in bacteria; namely, biofilm and flagella formation⁴⁰ and for the synthesis of virulence and competitive factors (*siderophores, exotoxins*).

The regulatory networks of *M. tuberculosis* and *C. glutamicum* which have the most similar global topologies are also the most related phylogenetically since both of them are Actinobacteria. It is adventurous to conclude more considering they are some of the least well-described regulatory networks and the livelihood of these organisms seems to be different (*pathogen vs. glutamate producer*); we should wait until the description of their regulatory networks are more complete. The regulatory networks of *E. coli* and *B. subtilis*, which are also the largest regulatory networks of the six here analyzed, have the second smallest *M*-distance. The known information contained in these networks is similar, but the network of *B. subtilis* is dominated by sporulation processes on all the levels. The regulatory network of *P. aeruginosa* is the least similar to the other bacterium networks, and this could be explained because in this bacterium the most studied biological processes are clearly biased to the virulence and pathogenic processes setting aside central biological processes.²²

5 Conclusions

The correct comparison of biological networks could have important implications in the understanding, at structural and functional levels, of diverse phenomena like evolution, diseases, modularity and robustness in organisms. This goal is important not only for biological networks but also for other types of complex networks, for instance social or web networks. Even for graph theory, where it has been tackled.

In this study we propose the pseudo-distance *M* to compare networks with different numbers of vertices, and arrows. This metric allows for the comparisons of networks retaining important structural and dynamical characteristics. Given several networks, we can determine the two of them having more similar global structures by using *M*.

The pseudo-distance *M* is based on the hierarchy provided by the concept of dominant vertices. Other hierarchies have been proposed in the literature, or implemented in computer programs related to networks analysis. However, the hierarchy obtained by the algorithm here used fulfils certain dynamical properties and provides insights into the biological network organization.

Other methods have been proposed to compare networks, however the comparison based on the hierarchy provided by

dominant vertices has the following advantages: it is performed by an algorithm of low complexity (polynomial time), in contrast with other methods that require the search of subgraphs, which is an NP complete problem;⁴ it allows the comparison of networks with different sizes and numbers of connected components;⁴¹ it is a global measure, converse to several heuristics measures which are focused on local properties; it is applied to directed networks, which is not allowed for most of the proposed methods;⁴¹ it returns only a number, such that networks are similar as the number closes to zero.

Consistent with other studies,^{4,13} we found that SF and GEO networks are closer to real biological networks than ER networks in a similar order to that established in those studies. Moreover, pseudo-distance makes clear that existing network models are far from real biological networks.

About biological networks here analyzed: *E. coli*, *B. subtilis*, *P. aeruginosa*, *M. tuberculosis*, *C. glutamicum* and *S. aureus*, we can manifest that these bacteria have distinctive life styles, however the information we have on their regulatory networks (close to a third of the total of their predicted number of genes) reveal a topology with approximately the same number of dominant vertices and depth, containing most of their vertices in the second and third levels of their hierarchies. The proposed hierarchies also capture the relationships in the known information, since regulatory networks are still in construction.

Lastly, the pseudo-distance *M* has proven to be effective for the comparison of real biological networks. It allows us to distinguish relevant biological features of the bacteria here analyzed and show them to be different from random networks.

Acknowledgements

We thank Irma Tristan, Zaira Luna and Marcelino Ramírez for their critical comments on the manuscript. This project was supported by NPTC project PROMEP 103.5/11/4818 given to B.L., and Conacyt grant 103686 given to A.M.-A. and Julio Collado-Vides.

References

- 1 B. H. Junker and F. Schreiber, *Analysis of biological networks*, Wiley Interscience, 1st edn, 2008.
- 2 B. Albert-László and N. O. Zoltán, *Nat. Rev.*, 2004, 101–114.
- 3 U. Alon, *Nat. Rev.*, 2007, 450–461.
- 4 N. Przulj, *Bioinformatics*, 2006, **23**, e177–e183.
- 5 J. Berg and M. Lässig, *BioSystems*, 2006, 186–196.
- 6 S. Erten, X. Li, G. Bebek, J. Li and M. Koyutürk, *BMC Bioinf.*, 2009, 1–14.
- 7 A. Banerjee, *BioSystems*, 2012, **107**, 186–196.
- 8 A. Mano, T. Tuller, O. Béjác and R. Y. Pinter, *BMC Bioinf.*, 2010, **11**, 1–10.
- 9 M. Heyman and A. Singh, *Bioinformatics*, 2003, **19**, i138–i149.
- 10 M. Hayashida and T. Akutsu, *BMC Syst. Biol.*, 2010, **4**, 1–11.
- 11 C. Shou, N. Bhardwaj, H. Y. K. Lam, Y. Koon-Kiu, P. M. Kim, M. Snyder and M. B. Gerstein, *PLoS Comput. Biol.*, 2011, **7**, 1–14.

- 12 B. Luna and E. Ugalde, *Phys. D*, 2008, **237**, 2685–2695.
- 13 W. Hayes, K. Sun and N. Przulj, *Bioinformatics*, 2013, **29**, 483–491.
- 14 H. D. Jong, *J. Comput. Biol.*, 2002, **9**, 67–103.
- 15 R. Coutinho, B. Fernandez, R. Lima and A. Meyroneinc, *J. Math. Biol.*, 2006, **52**, 524–570.
- 16 D. Thieffry and R. Thomas, *Bull. Math. Biol.*, 1995, **57**, 277–297.
- 17 P. Erdős and A. Rényi, *Publ. Math.*, 1959, **6**, 290–297.
- 18 R. Durrett, *Random Graph Dynamics*, Cambridge Series in Statistical and Probabilistic Mathematics, 1st edn, 2007.
- 19 M. Penrose, *Geometric Random Graphs*, 2003, vol. 5.
- 20 S. Gama-Castro, H. Salgado, M. Peralta-Gil, A. Santos-Zavaleta, L. Muñoz-Rascado, H. Solano-Lira, V. Jimenez-Jacinto, V. Weiss, J. S. García-Sotelo, A. López-Fuentes, L. Porrón-Sotelo, S. Alquicira-Hernández, A. Medina-Rivera, I. Martínez-Flores, K. Alquicira-Hernández, R. Martínez-Adame, C. Bonavides-Martínez, J. Miranda-Ríos, A. M. Huerta, A. Mendoza-Vargas, L. Collado-Torres, B. Taboada, L. Vega-Alvarado, M. Olvera, L. Olvera, R. Grande, E. Morett and J. Collado-Vides, *Nucleic Acids Res.*, 2011, D98–D105.
- 21 N. Sierro, Y. Makita, M. Hoon and K. Nakai, *Nucleic Acids Res.*, 2008, **36**, D93–D96.
- 22 E. Galán-Vásquez, B. Luna and A. Martínez-Antonio, *Microb. Inf. Exp.*, 2011, 1–11.
- 23 J. Sanz, J. Navarro, A. Arbués, C. Martín, P. C. Marijuán and Y. Moreno, *PLoS One*, 2011, **6**, 1–9.
- 24 J. Pauling, R. Röttger, A. Tauch, V. Azevedo and J. Baumbach, *Nucleic Acids Res.*, 2011, **40**, D610–D614.
- 25 D. A. Ravcheev, *J. Bacteriol.*, 2011, **193**, 3228–3240.
- 26 F. Blattner, G. Plunkett III, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D. Glasner, C. K. Rode, G. F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau and Y. Shao, *Science*, 1997, **277**, 1453–1462.
- 27 E. Pérez-Rueda and J. Collado-Vides, *Nucleic Acids Res.*, 2000, **28**, 1838–1847.
- 28 D. Lopez, H. Vlamakis and R. Kolter, *FEMS Microbiol. Rev.*, 2009, **33**, 152–163.
- 29 F. Kunst, N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessières, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. Brouillet, V. Bruschi, B. Caldwell, V. Capuano, N. Carter, S. Choi, J. Codani, I. Connerton, N. J. Cummings, R. A. Daniel, F. Denizot, M. Devine, A. Dusterhoft, S. D. Ehrlich, P. T. Emmerson, K. D. Entian, J. Errington, C. Fabret, E. Ferrari, D. Foulger, C. Fritz, M. Fujita, Y. Fujita, D. Fuma, A. Galizzi, N. Galleron, S. Y. Chim, P. Glaser, A. Goffeau, E. J. Golightly, G. Grandi, G. Guiseppe, B. J. Guy, K. Haga, J. Haieah, C. R. Harwood, A. Henaut, H. Hilbert, S. Holsappel, S. Hosono, M. F. Hullo, M. Itaya, L. Jones, B. Joris, D. Karamata, Y. Kasahara, M. Klaerr-Blanchard, C. Klein, Y. Kobayashi, P. Koetter, G. Koningstein, S. Krogh, M. Jumano, K. Kurita, A. Lapidus, S. Lardinois, J. Lauber, V. Lazarevic, S. M. Lee, A. Levine, H. Liu, S. Masuda, C. Manuel, C. Medigue, N. Medina, R. P. Mellado, M. Mizuno, D. Moestl, S. Nakai, M. Noback, D. Noone, M. O'Reilly, K. Ogawa, A. Ogiwara, B. Oudega, S. H. Park, V. Parro, T. M. Poji, D. Portelle, S. Porwollik, A. M. Prescott, E. Presecan, P. Pujic, B. Pernelle, G. Rapoport, M. Rey, S. Reynolds, M. Rieger, C. Rivolta, E. Rocha, B. Roche, M. Rose, Y. Sadaie, T. Sato, E. Tacconi, T. Takagi, H. Takahashi, K. Takemaru, M. Takeuchi, A. Tamakoshi, T. Tanaka, P. Terpstra, A. Tognoni, V. Tosato, S. Uchiyama, M. Vandenberg, F. Vannier, A. Vassarotti, A. Viari, R. Wambutt, E. Wedler, H. Wedler, T. Weitzenegger, P. Winters, A. Wipat, H. Yamamoto, K. Yamane, K. Yasumoto, K. Yata, K. Yoshida, H.-F. Yoshikawa, E. Zumstein, H. Yoshikawa and A. Danchin, *Nature*, 1997, **390**, 249–256.
- 30 I. Takahiro, Y. Ken-ichi, T. Goro, F. Yasutaro and N. Kenta, *Nucleic Acids Res.*, 2001, **29**, 278–280.
- 31 J. B. Lyczak, C. L. Cannon and G. B. Pier, *Microbes Infect.*, 2000, **2**, 1051–1060.
- 32 C. K. Stover, X. Q. Pham, A. L. Erwin, S. D. Mizoguchi, P. Warrener, M. J. Hickey, F. S. L. Brinkman, W. O. Hufnagle, D. J. Kowalik, M. Lagrou, R. L. Garber, L. Goltry, E. Tolentino, S. Westbrook-Wadman, Y. Yuan, L. L. Brody, S. N. Coulter, K. R. Folger, A. Kas, K. Larbing, R. Lim, K. Smith, D. Spencer, G. K. Wong, Z. Wu, L. T. Paulsen, J. Reizer, M. H. Saier, R. E. W. Hancock, S. Lory and M. V. Olson, *Nature*, 2000, **406**, 959–964.
- 33 I. Smith, *Clin. Microbiol. Rev.*, 2003, **16**, 463–496.
- 34 S. T. Cole, R. Brosch, J. Parkhill, T. Garnier, C. Churcher, D. Harris, S. V. Gordon, K. Eiglmeier, S. Gas, C. E. Barry, 3rd, F. Tekaiia, K. Badcock and D. Basham, *Nature*, 1998, 537–544.
- 35 J. Kalinowski, B. Bathe, D. Bartels, N. Bischoff, M. Bott, A. Burkovski, N. Dusch, L. Eggeling, B. Eikmanns, L. Gaigalat, A. Goesmann, M. Hartmann, K. Huthmacher, R. Krämer, B. Linke, A. C. McHardy, F. Meyer, B. Möckel, W. Pfefferle, A. Pühler, D. A. Rey, C. Rückert, O. Rupp, H. Sahm, V. F. Wendisch, I. Wiegräbe and A. Tauch, *J. Biotechnol.*, 2003, **4**, 5–25.
- 36 T. M. Venancio and L. Aravind, *J. Biol.*, 2009, **29**, 1–5.
- 37 M. Kuroda, *Lancet*, 2001, 1225–1240.
- 38 J. S. Nielsen, M. H. G. Christiansen, M. Bonde, S. Gottschalk, D. Frees, L. R. Thomsen and B. H. Kallipolitis, *Arch. Microbiol.*, 2011, **193**, 23–34.
- 39 A. Martínez-Antonio and J. Collado-Vides, *Curr. Opin. Microbiol.*, 2003, **6**, 482–489.
- 40 A. Martínez-Antonio, S. C. Janga and D. Thieffry, *J. Mol. Biol.*, 2008, **381**, 238–247.
- 41 G. Jurman, S. Riccadonna, R. Visintainer and C. Furlanello, *arXiv preprint arXiv:1109.0220*, 2011.

Regulatory switches for hierarchical use of carbon sources in *E. coli*

Ruth S. Pérez-Alfaro¹, Moisés Santillán², Edgardo Galán-Vásquez¹, Agustino Martínez-Antonio¹

¹Departamento de Ingeniería Genética, Centro de Investigación y de Estudios Avanzados del IPN, Unidad Irapuato, Km. 9.6 Libramiento Norte Carr. Irapuato-León 36821 Irapuato, Guanajuato, México

²Centro de Investigación y de Estudios Avanzados del IPN, Unidad Monterrey, Vía del Conocimiento 201, 66600 Apodaca NL, México

E-mail: amartinez@ira.cinvestav.mx

Received 2 May 2014; Accepted 5 June 2014; Published online 1 September 2014



Abstract

In this work we study the preferential use of carbon sources in the bacterium *Escherichia coli*. To that end we engineered transcriptional fusions of the reporter gene *gfpmut2*, downstream of transcription-factor promoters, and analyzed their activity under several conditions. The chosen transcription factors are known to regulate catabolic operons associated to the consumption of alternative sugars. The obtained results indicate the following hierarchical order of sugar preference in this bacterium: glucose > arabinose > sorbitol > galactose. Further dynamical results allowed us to conjecture that this hierarchical behavior might be operated by at least the following three regulatory strategies: 1) the coordinated activation of the corresponding operons by the global regulator catabolic repressor protein (CRP), 2) their asymmetrical responses to specific and unspecific sugars and, 3) the architecture of the associated gene regulatory networks.

Keywords *Escherichia coli*; transcription factors; carbon sources; hierarchical use.

Network Biology
ISSN 2220-8879
URL: <http://www.iaees.org/publications/journals/nb/online-version.asp>
RSS: <http://www.iaees.org/publications/journals/nb/rss.xml>
E-mail: networkbiology@iaees.org
Editor-in-Chief: Wenjun Zhang
Publisher: International Academy of Ecology and Environmental Sciences

1 Introduction

The way bacteria use different carbon sources (Monod, 1942) has been studied for a long time, in which *Escherichia coli* has been the favorite model organism. We learned from the very beginning that glucose is the carbon source supporting the fastest growth on this bacterium (Walker et al., 1934). This sugar is also the preferred one if bacteria are exposed to a mixture of carbon sources. It seems that *E. coli* uses carbon sources on the basis of “best food served first”. The molecular mechanisms behind this operating principle are various, the best-known ones are: inducer exclusion, local or dedicated transcriptional regulation, global transcriptional regulation, small RNAs, and catabolite repression.

Inducer exclusion (Jones-Mortimer and Kornberg, 1974; Chen et al., 2013) takes place when, in the presence of glucose or other PTS sugars, the unphosphorylated EIIA^{Glc} (part of the PTS system) binds to and

stabilizes the resting state of non PTS-sugar transporters, inhibiting the transport (and use) of alternative carbon sources.

Local or dedicated transcriptional regulation operates at the initiation of gene transcription of sugars catabolic operons. These operons are normally subject to repression by at least one specific regulator, whose derepression occurs when the corresponding specific sugar is available and binds to it. This binding causes the effector-repressor complex to unbind from the operator zone, which is a necessary condition for the corresponding operon to become active (Jacob and Monod, 1961; Sellitti et al., 1987).

Global transcriptional regulation. The complementary condition for the transcription of sugar catabolic genes is given by the activity of the global regulator CRP (catabolic repressor protein or cAMP regulatory protein). CRP becomes active when bound by cyclic adenosine monophosphate (cAMP). The cAMP-CRP complex is then capable of recruiting RNA polymerase to promoter zones of catabolic operons so their transcription is started if no repressor is present. Hence, a condition for the transcription of catabolic operons is that high cAMP levels are present. High cAMP levels are in general achieved in the absence of glucose although, as mentioned below, it could be the result or a wider physiological status (Gottesman, 1984; Martínez-Antonio and Collado-Vides, 2003).

Small RNAs (sRNA). Arguably, the best known sRNA is the multi-target Spot42, which inhibits the translation of at least 14 genes, mostly related to the use of non-PTS sugars. Spot42 is activated by cAMP-CRP and together form a coherent feed-forward loop to avoid use of non-PTS sugars when the preferred sugars are available (Beisel and Storz, 2001; Wright et al., 2013).

Catabolite repression (Magasanik, 1961; Görke and Stülke, 2008), a physiological concept so-named by Boris Magasanik as a generalization of the “glucose effect” described many years earlier (Cohn, 1957). It was derived after observing the repression, when glucose is present, of catabolic enzymes specific for carbon and nitrogen metabolism. This phenomenon was related to cAMP levels that increase when poor carbon sources are present in the milieu (Epstein et al., 1975). cAMP is synthesized by the CyaA enzyme, which is activated by phosphorylated EIIA^{glu} but requires an additional unidentified factor (Park et al., 2006). It was postulated that a derived catabolite of carbon sources (the repressor catabolite) is the responsible to trigger cAMP synthesis. Only recently, a high-throughput proteome analysis (studying carbon, nitrogen and sulfur sources metabolism) in *E. coli* revealed that cAMP levels are diminished by α -ketoacids (mainly by oxaloacetate) through the inhibition of adenylate cyclase, the enzyme responsible for cAMP synthesis (You et al., 2013). This explains how this central metabolite is balancing the overall bacterial physiology throughout the nitrogen/carbon metabolism (Rabinowitz and Silhavy, 2013).

Here we present a study that copes with the activities of the promoters of specific catabolic regulators, which in addition to self-regulation, respond to the global regulator CRP (points 2 and 3 above). These locals and the global regulator operate together to regulate transcriptional initiation in *E. coli* catabolic operons for the transport and use of carbon sources other than glucose.

We tackle the question of how bacteria decide to consume alternative carbon sources, focusing in L-arabinose, D-sorbitol and D-galactose. The regulation, transport and first catabolic steps in the metabolism of these sugars are depicted in Fig. 1. As we can see, not only the corresponding genes are activated by CRP, but they are also repressed by specific transcription factors. We investigate in this work the promoter activities of these specific regulators. Importantly, all of them use the transported sugars as signal effectors to modulate their activities. The signal sugar binds to the repressor and unbinds it from the operator zone, thus allowing transcription of the corresponding genes. Finally, all the promoters here analyzed require of the housekeeping σ^{70} to be transcribed so this is not a variable to consider in this study.

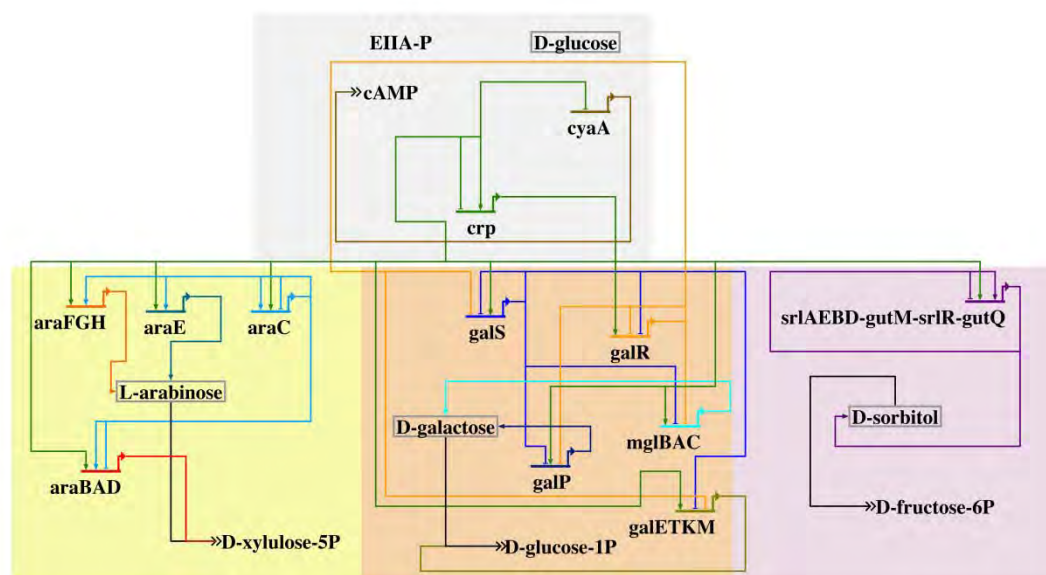


Fig. 1 Regulatory network controlling the use of the carbon sources employed in this study. It shows the module corresponding to the global regulator CRP. cAMP, a co-activator of CRP, is synthesized via CyaA when glucose is absent. The arabinose module includes the dual regulator AraC, which transcriptionally regulates the arabinose transporter genes AraE (low affinity) and AraFGH (high affinity). AraC also regulates the genes of enzymes isomerase (AraA), ribulokinase (AraB) and epimerase (AraD), which metabolize arabinose to D-xylulose 5-P. The galactose module has two repressors, GalR and GalS, in different transcription units. In absence of galactose they repress the genes for galactose transporters GalP (low affinity) and MglBAC (high affinity), as well as those for the enzymes GalK (galactokinase), GalT (uridiltransferase), and galM (epimerase), which metabolize galactose to glucose 1-P. The sorbitol module is also regulated by two transcription factors, SrlR and GutM, encoded in the same operon, which also includes genes for high affinity transporter (SrlAEB) and for the enzymes SrlD (dehydrogenase) and GutQ (isomerase), which transform sorbitol to fructose 6-P.

2 Material and Methods

2.1 Strains

In all our experiments we employ *Escherichia coli* K-12 MG1655 strain and derivatives harboring the different transcriptional fusions shown in Table 1. Most of the used transcriptional fusions were taken from a collection reported previously (Zaslaver et al., 2006). However, we rebuilt the transcriptional fusions for *gutM* and *crp* promoters in order to include regulatory sites for transcription factors not comprised in fusions from the collection. We realized the necessity of such regulatory sites by inspecting the transcription-factor binding sites reported in RegulonDB (Salgado et al., 2013). These last fusions were engineered by amplifying (through PCR and specific primers) the corresponding regulatory regions, cloning the resulting DNA fragments on pUA66 with the aid of the BamHI and XhoI restriction sites, and verifying the construction by means of DNA sequencing.

2.2 Bacterial growth

For strain maintenance we routinely used LB medium and for experimental tests we used M9 medium, supplemented with sugars as indicated. Also when indicated, we added kanamycin (Km) 50 $\mu\text{g ml}^{-1}$. Pre-inoculates were grown overnight in 5 ml of LB medium at 37 °C with agitation (200 rpm). Next, the cultures were diluted 1 : 100 in 150 μl of fresh M9 media in micro-titer plates of 96 wells and incubated for 12

h with agitation (250 rpm) at 37 °C. We supplemented M9 with 0.4% or 0.03% of glucose, and 0.2% of one or two alternative sugars as specified. We followed bacterial growth, by measuring OD595nm, and fluorescence (535 nm) every hour in a Perkin Elmer Victor X3 plate multi-lector.

Table 1 Regulatory regions employed on the transcriptional fusions.

| Promoter fusions | <i>E. coli</i> chromosome coordinates | Designed primers* 5'-3' | Region size | Cloning vector | Reference |
|-----------------------|---------------------------------------|---|-------------|----------------|-----------------------|
| <i>araCp::gfpmut2</i> | 69973-70452 | | 479bp | pUA66 | Zaslaver et al., 2006 |
| <i>crpp::gfpmut2</i> | 3483776-3484200 | F:tgatgactc <u>gaggcggatt</u> c R:tgccaatgagacag <u>ggatc</u> ca | 424bp | pUA66 | This study |
| <i>galSp::gfpmut2</i> | 2239619-2239844 | | 225bp | pUA139 | Zaslaver et al., 2006 |
| <i>galRp::gfpmut2</i> | 2973960-2974698 | | 738bp | pUA66 | Zaslaver et al., 2006 |
| <i>gutMp::gfpmut2</i> | 2823533-2823932 | F:cttgctgctc <u>gaggcggca</u> a R:ccatcc <u>gatccacacctc</u> tccgc | 399bp | pUA66 | This study |
| <i>srlRp::gfpmut2</i> | 2826905-2827074 | | 169bp | pUA66 | Zaslaver et al., 2006 |

*Underlined nucleotides define restriction sites for XhoI and BamHI endonucleases on forward and reverse primers.

2.3 Data acquisition and processing

The raw numerical data obtained from the Victor X3 plate multi-lector consisted of discrete measurements of optical density (OD) and fluorescence (GFP) versus time along the growth curves, with a sampling frequency of 1 hr⁻¹. Although enough to provide an overview of the time evolution of variables OD and GFP, such sampling frequency is too low to perform more refined quantitative analyses. For that, it is necessary to find a function that fits the experimental data. Since the generalized logistic function is a widely used sigmoid function for growth modeling we decided to employ it. In all cases we found that it fits both the growth curves and the GFP profiles with correlation factors higher than 0.99. The functions used to fit the OD and GFP profiles are:

$$OD(t) = a_1 + \frac{k_1 - a_1}{(1 + q_1 e^{-b_1 t})^{1/v_1}}, \quad (1)$$

$$GFP(t) = a_2 + \frac{k_2 - a_2}{(1 + q_2 e^{-b_2 t})^{1/v_2}}, \quad (2)$$

in which a_i , b_i , k_i , q_i , and v_i ($i = 1, 2$) are fitting parameters. Zaslaver et al. (2006) and Martínez-Antonio et al. (2012) have argued that promoter activity is proportional to $(dGFP(t) / dt) / OD(t)$. Thus, after finding the best

fitting parameters we differentiated function (2) and divided the result by eq. (1) to compute the promoter activity in each case.

For every experimental condition and for every transcriptional function we periodically measured the values of optical density and green fluorescence in triplicate, computed the corresponding average values, and respectively fitted to Eqs. (1) and (2), and computed the promoter activity level as explained above.

In our experiments we could observe that the *crp*, *galR* and *srlR* promoters were unresponsive under all the tested conditions (data not shown). The absence of *crp* promoter activity might be explained because it is the most global regulator in *E. coli*. Not only *crp* regulates itself, but it is also subject to dual regulation by another global regulator: FIS (factor for inversion stimulation). In the case of *galR* and *srlR*, the reason why they present constant low expression levels may be that they are constitutively expressed; up to date no regulator is known for these genes. Due this unresponsiveness and for the sake of clarity we excluded the results corresponding to these promoters in the fore coming sections.

3 Results

3.1 Different carbon sources support the growth of *E. coli* differentially

Our first objective was to analyze how the different carbon sources under study support the growth of *E. coli*. For this, we followed the progression of *E. coli* cultures growing in M9 minimal medium added with L-arabinose, D-sorbitol and D-galactose, both separately and in dual combinations. The growth profiles show in Fig. 2A confirm that glucose is by far the sugar that best supports *E. coli* growth. The hierarchical order of sugars in terms of their capacity to sustain cell growth is as follows: glucose > arabinose > sorbitol > galactose. When combinations of two alternative sugars were used, the bacterial growth rate almost equated that of glucose during the exponential growth phase. On the other hand, with all the sugar combinations, the maximal bacterial population density surpassed that of glucose alone. The decreasing order of alternative sugar combinations in terms of the exponential growth rate they are capable of sustaining is: arabinose+sorbitol > arabinose+galactose > galactose+sorbitol. In these experiments, glucose was set at a limiting amount (0.03%) from the very beginning to clearly distinguish the time at which *E. coli* starts using alternative carbon sources (Fig. 2A). We observed a differential growth of cultures with not limitation as compared with those limited on glucose as early as 3.5 hours after the start of the experiment. However, a careful observation on the alternative-sugar catabolic-operon promoter activity reveals that they become active after 2 hours of the experiment beginning (see below).

3.2 Glucose limitation triggers foraging alternatives

Our second objective was to study the dynamics of the alternative-sugar catabolic-operon promoters under glucose exhaustion conditions. Specifically, we were interested in the following scenarios: 1) when glucose is limiting from the culture at the very beginning and, 2) when glucose is exhausted after a normal period of bacterial growth. For that purpose we engineered specific reporters for relevant transcription factors (Table 1). These reporters were built by transcriptionally fusing each promoter to gene *gfpmut2*, and promoter activity was estimated by measuring fluorescence along the bacterial growth curves (Zaslaver et al., 2006). We made sure that the presence of the vector and genetic constructions were not detrimental for *E. coli* growth before the assays.

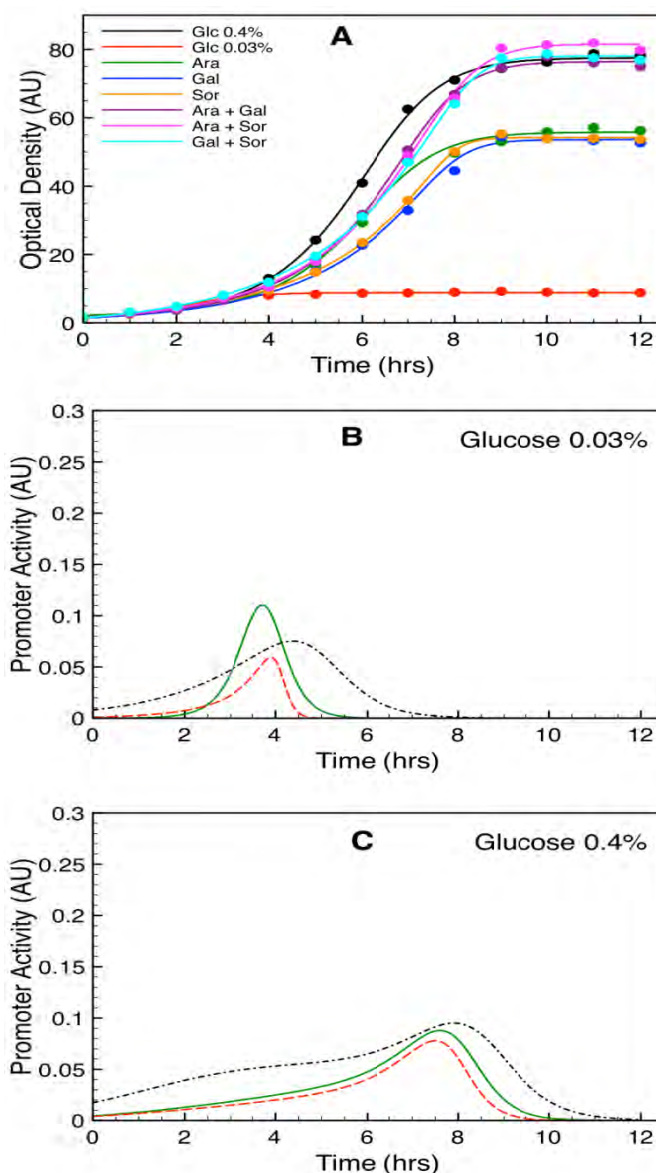


Fig. 2 Carbon source consumption by *Escherichia coli*. **A**) Wild-type *E. coli* growth curves (experimental data and best fitting generalized logistic functions) while cultured with different carbon-source concentrations and combinations. **B**) Transcription-factor promoter activities under glucose limitation conditions (M9 + 0.03% glucose). **C**) Transcription-factor promoter activities during a normal course of *E. coli* growth (M9 + 0.4% glucose). The color code in all graphs is as follows: *araC*, black; *gutM*, green; *galS*, red.

The results of these experiments can be summarized as follows. The time when alternative promoters are maximally active depends on the time at which glucose is exhausted. When glucose is limiting from the beginning of the experiment, the alternative-sugar catabolic-operon promoters start activating as soon as 2 hours after the experiment start, they reach their maximal of activity around the hour 4, and their activity starts declining thereafter (Fig. 2B). Contrarily, if there is a considerable amount of glucose at the culture beginning, the alternative-sugar catabolic-operon promoters become active only after glucose has been presumably exhausted, with the exemption of *araC* promoter that shows some activity during all the experiment. The three examined promoters reach their maximal activity about 8 hours after the beginning of the experiment (Fig. 2C).

The promoter behaviors reported in Figs. 2B and 2C, in absence of specific sugars in the milieu, can be explained by the activity of the master regulator CRP which only becomes active and turns up its target genes (among others the ones corresponding to the here studied transcription factors) when glucose is exhausted, and supposedly, when cAMP production is increased. Interestingly, no matter how fast glucose is exhausted, the studied promoters start showing some sign of activity about 2 hours after the cultures' start. Finally, under conditions of high initial glucose levels, not only all promoters reach their maximal activity at roughly the same time, but also their maximal activity levels are quite similar. When the initial glucose concentration is low, the maximal levels are dissimilar, although they are reached at similar times. However, it is important to emphasize that promoter activity is inversely proportional to bacterial density and that bacterial density (estimated by means of optical density measurements) is very low when glucose levels are initially low. All this implies that the obtained maximal promoter activity levels are not as reliable as those corresponding to high initial glucose concentration. Having this in mind it is remarkable that the maximal promoter activity levels have the same order of magnitude in all cases, see Figs. 2B and 2C.

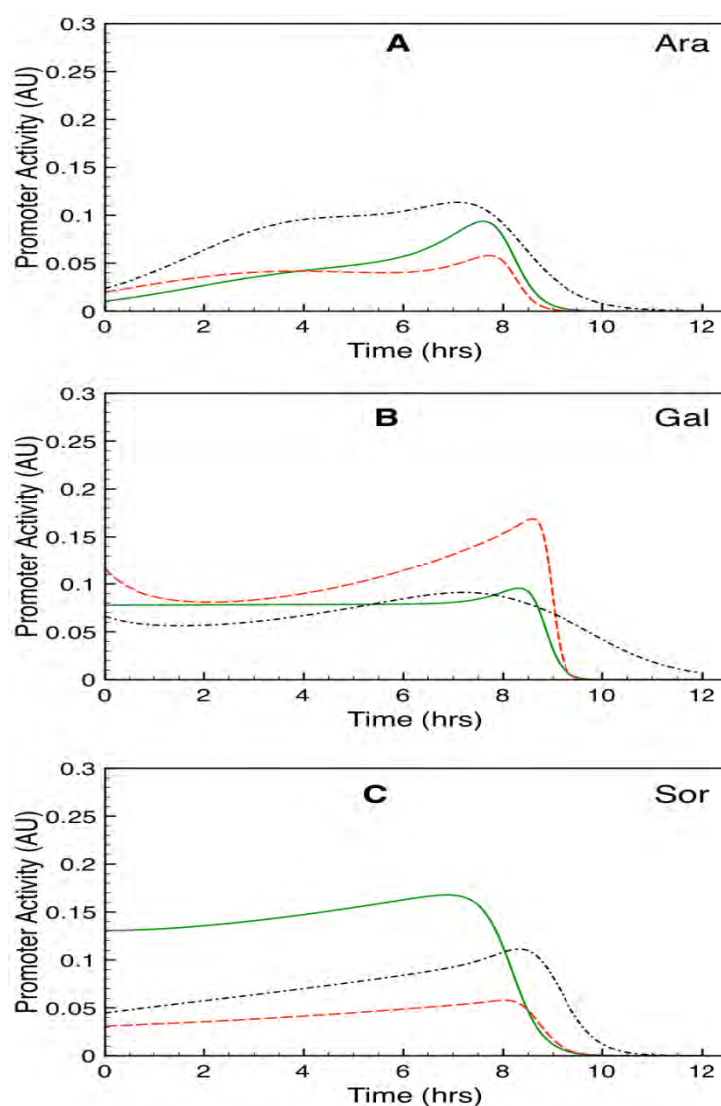


Fig. 3 Transcription-factor promoter activities as response to the presence of alternative carbon sources. **A)** Promoter activities in the presence of L-arabinose. **B)** Promoter activities in the presence of D-galactose. **C)** Promoter activities in the presence of D-sorbitol. The alternative sugars were supplemented at 0.2% (M9 + 0.03% glucose + 0.2% each alternative sugar). The color code in all panels is as follows: *araC*, black; *gutM*, green; *galS*, red.

3.3 Promoter specificity for sugar effectors

We further investigated the specificity of promoter response to different sugars. To that end we grew *E. coli* in M9 medium with a limiting amount of glucose plus different carbon sources, and measured the activity of the previously mentioned transcription-factor promoters. The results are reported in Fig. 3 in which the promoter activities are plotted vs. time under different conditions.

We found that each transcription-factor promoter primarily responds to the sugar whose catabolic operon it controls. However the individual effects vary in amplitude as follows. The largest effect is that of sorbitol on *gutM*, whose maximal activity is about 2.4 times larger the one caused by glucose exhaustion solely; then we have the effect of galactose on *galS* that shows a 1.8 fold increased activity; and the smallest response is that of *araC* to arabinose, with an increase of 1.3 in the promoter activity.

We could also observe a cross-regulation effect between signals (sugars) and the expression of transcription factors. This may allow transcription factors to display asymmetrical responses to specific and unspecific sugars. We can see that arabinose has a slight positive effect on *gutM* promoter, while it inhibits the expression level of *galS* promoter by about 40%. Regarding sorbitol, it increases by about 40% the expression level of promoter *araC*, and inhibits by about 30% the expression of promoter *galS*. Finally, galactose has no noticeable effect on *araC* promoter, but increases by about 20% the expression level of *gutM* promoter. The above-discussed results are summarized in Fig. 4.

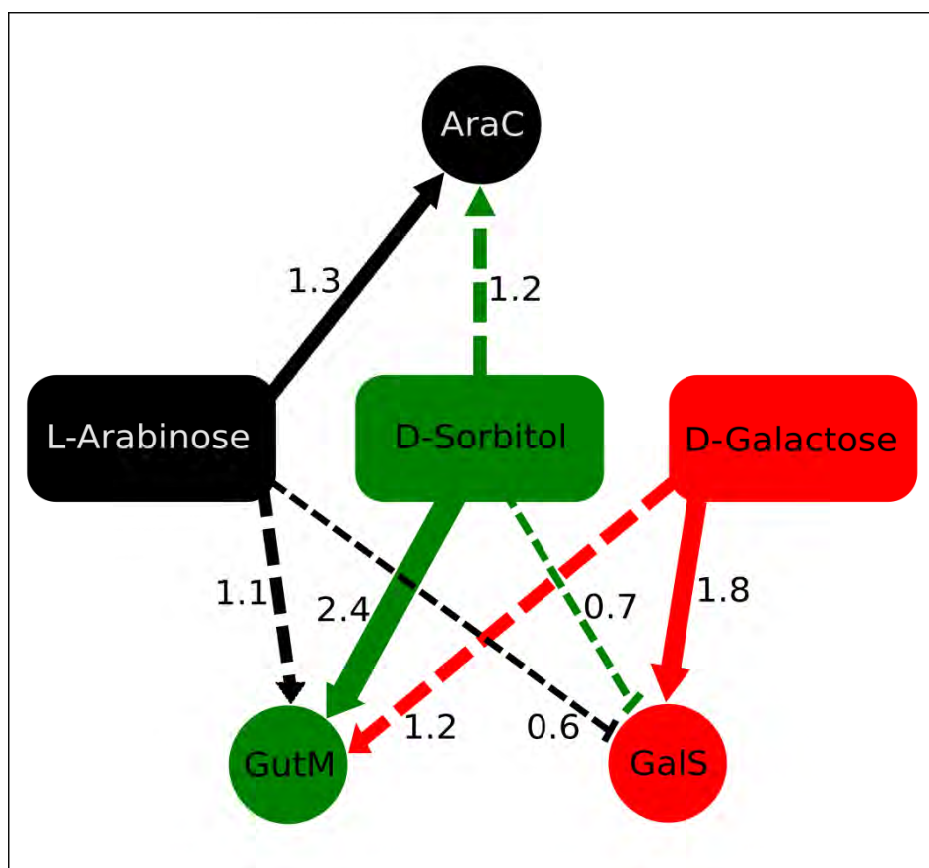


Fig. 4 Asymmetrical response of transcription factor promoters to sugar signals. The values next to the arrows indicate the percent of response of each promoter to every one of the three alternative sugars, as compared to the effect caused by glucose exhaustion solely. Continuous lines indicate the response of transcription factor to their effector sugar, while dashed lines indicate cross-response to the presence of other sugars.

Note from Fig. 4 that the promoter that is most responsive to sugars other than its own one is *araC*, followed by *gutM* and *galS* promoters. This behavior is interesting because a partial turn on of promoters due to non specific stimulus is an indicator of a putative conditioned behavior already observed in sugars consumption in *E. coli* (Tagkopoulos et al., 2008). Our results suggest that the promoter most prone to conditional behavior is that of *araC*.

Finally, the sugar that most enhances the activity of transcription-factor promoters other than the one specific to it is galactose. This suggests that, in agreement with (Liu et al., 2005), the worse a carbon source that sustains bacterial growth, the more it positively affects the activity of alternative-sugar catabolic operons.

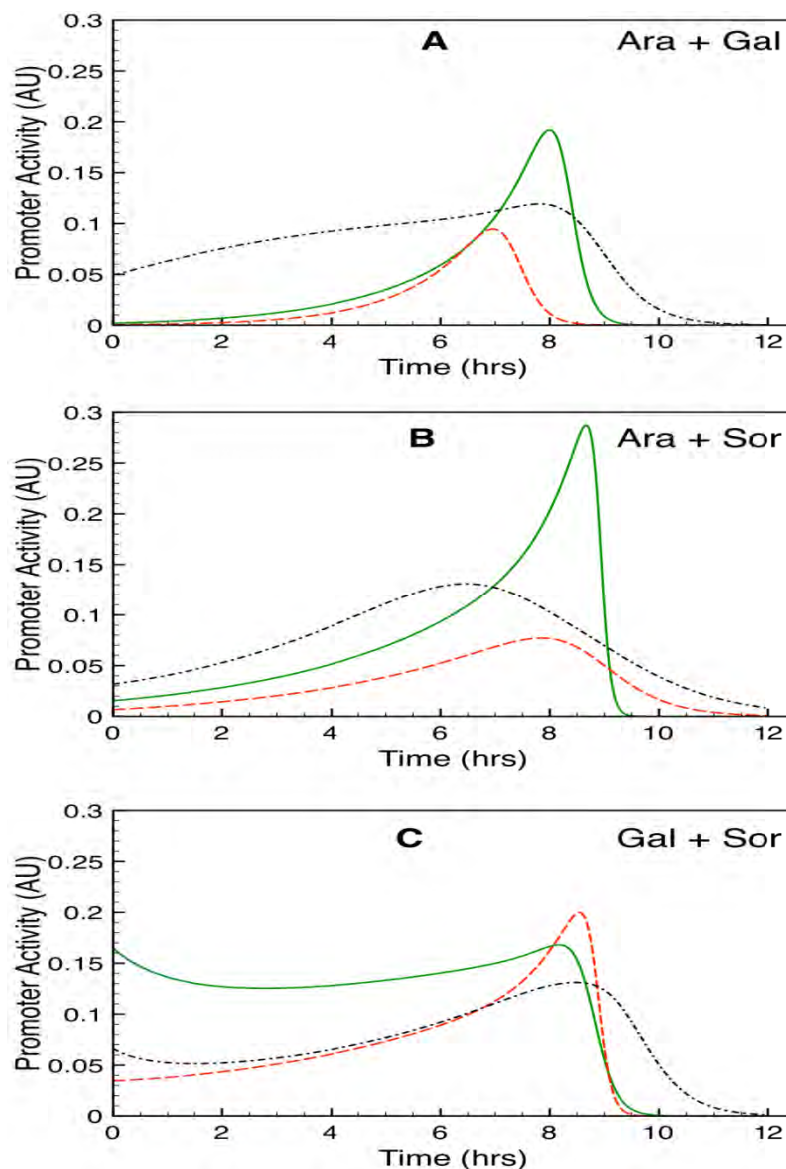


Fig. 5 Transcription-factor promoter activities in the presence of pairs of alternative sugars (M9 + 0.03% glucose + 0.2% sugar 1 + 0.2% sugar 2). **A)** Promoters activities in the presence of arabinose and galactose. **B)** Promoter activities in the presence of arabinose and sorbitol. **C)** Promoter activities in the presence of galactose and sobitol. The color code in all panels is as follows: *araC*, black; *gutM*, green; *galS*, red.

3.4 Promoter activities reflect carbon-use hierarchy

We finally performed experiments in which bacteria were grown in the presence of a limiting quantity of glucose (0.03%) and a mix of two alternative sugars (0.2% each). The rationale behind this experiment was that once glucose is exhausted, *E. coli* should be forced to consume one of the two alternative sugars present in the milieu, and that this decision might be evidenced by the activity of the promoter associated to the transcription factor regulating the catabolic operon of the sugar of choice. The results of these experiments are reported in Fig. 5.

Notice that whenever bacteria are cultured in the presence of arabinose, promoter *araC* becomes active before the other two. The explanation of this observation is straightforward under the assumption that arabinose is preferred by bacteria over sorbitol and galactose.

Note from the arabinose + galactose experiment (Fig. 5A) that the positive influence of galactose upon *galS* is capable of completely counteracting the negative influence of arabinose. This is consistent with the supposition that arabinose is consumed before galactose by bacteria. It is also interesting that the rather small positive effects that both arabinose and galactose individually have on *gutM* boost each other to render a combined over-expression of more than 100%. A detailed observation of the curves in Fig. 5B (arabinose + sorbitol experiment), reveals that their amplitudes completely agree with the interaction scheme in Fig. 4.

In Fig. 5C (sorbitol + galactose experiment) we can see that the negative effect of sorbitol on *galS* is fully counteracted by the positive effect of galactose. However, the maximum activity of promoter *galS* is posterior to that of promoter *gutM*. This suggests that sorbitol is consumed before than galactose by bacteria. Furthermore, the maximal expression levels correspond to what one would expect from the interaction scheme in Fig. 4.

In summary, our results suggest that the investigated alternative sugars are consumed in the following order: arabinose, sorbitol and galactose. Moreover, the maximum expression levels in the experiments with two alternative sugars agree with the interaction scheme reported in Fig. 4, except for the expression of promoter *gutM* in the arabinose + galactose experiment. It seems that arabinose and galactose synergically make promoter *gutM* increase its expression level by more than 100%.

3.5 CRP as a global coordinator for carbon metabolism

All the transcription-factor operator regions here analyzed include a DNA-binding site for CRP, in addition to being self-regulated. In the previous sections we studied the contribution of specific catabolite signals to the activity of their local regulators. Thus, to have a complete picture it is necessary to test the effect of CRP (and indirectly that of cAMP) on the promoter activities of these local regulators. To this end we used a CRP mutant strain (Baba et al., 2006) as a receptor of the transcriptional fusions analyzed before, and measured growth and promoter activities when glucose is limiting at the beginning (0.03%) and at the end of the culture (0.4%), see Fig. 6. A first observation is that deletion of *crp*, although not essential, has negative effects on the bacterial growth rate (Figs. 6B vs 6A). Note that the culture final OD decreases as compared with that of the strain with an intact *crp* gene (comparable on 0.4% glucose). This could be explained by taking into consideration that CRP is a global coordinator of *E. coli* physiology, which regulates more, that 30% of all the genes with known regulation in this bacterium. However, the negative effect on growth is more pronounced in the strains harboring the transcriptional fusions of *gfp* with *galS* and *gutM* promoters than that with *araC*. We do not have a consistent explanation for this observation.

Regarding the promoter activities in the absence of *crp*, when glucose is depleted at the beginning of the culture (0.03%), the promoters activities changed as follows, as compared with the intact-*crp* strain: the *araCp* activity profile changed neither its amplitude nor the time at which the maximum value was achieved, yet the profile is now narrower; the maximal *gutMp* activity level was doubled, although it was retarded by more than

one hour; the *galSp* activity profile became wider and its maximal level decreased by about 50% (Figs. 6D vs 6C).

On the other hand, when glucose is exhausted at the end of the culture (0.4%) all the promoter activities are diminished and retarded except that of *araCp* (Figs. 6F vs 6E). Together, these observations indicate that CRP contributes to the fitness and performance of bacterial growth and coordinates the response of alternative regulatory machinery for the use of alternative carbon sources in *E. coli*, although it seems not to be essential for the transcriptional response of local repressors.

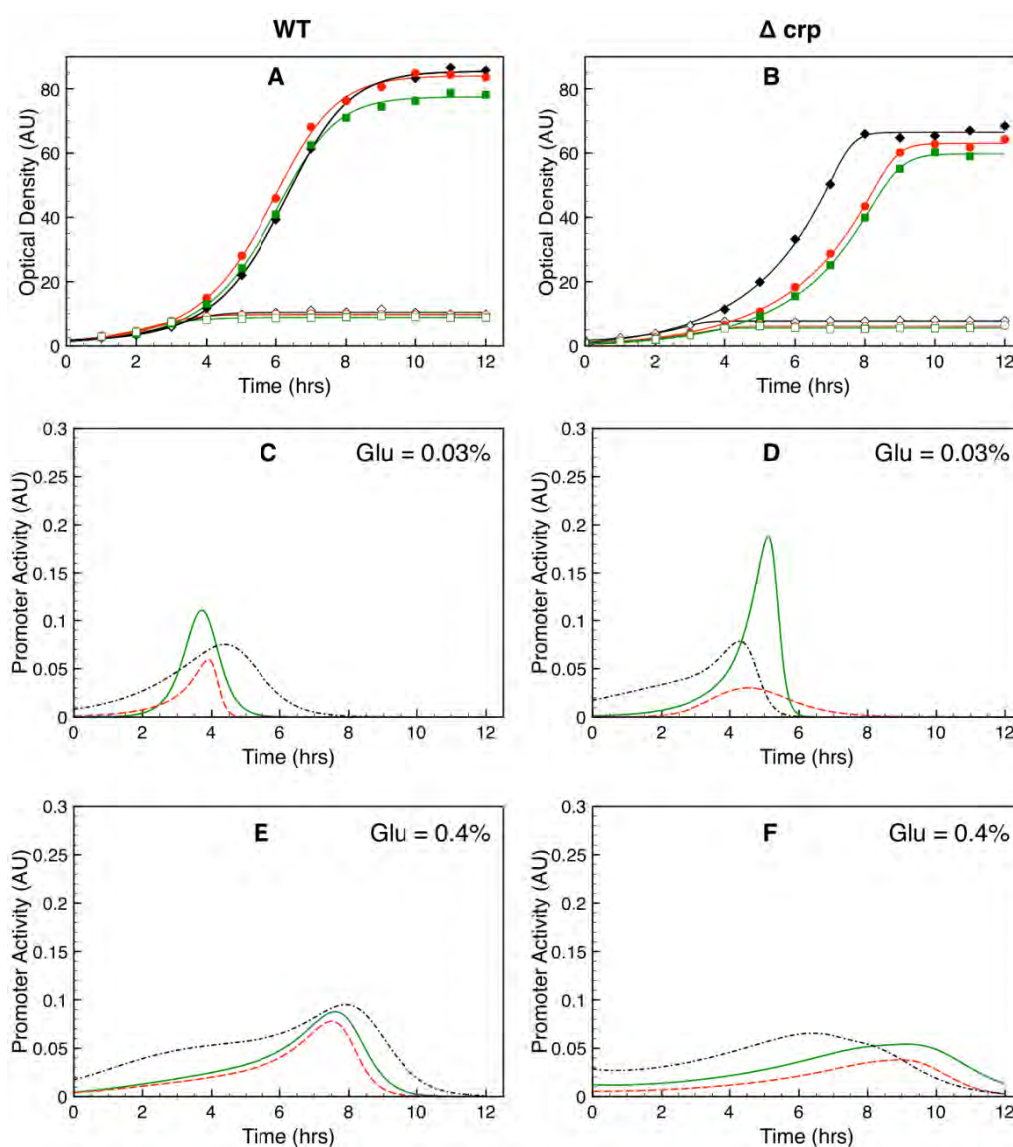


Fig. 6 Effect of CRP on the promoter activity of various catabolic repressors. Growth of bacteria harboring the transcriptional fusions on wild type background (A) and on *crp* deletion (B). Solid (empty) symbols correspond to a glucose concentration of 0.4% (0.03%). Promoter activities with glucose limited at the beginning of the culture on WT (C) and on Δcrp mutant (D). Promoter activities without glucose limitation at the beginning of the culture on WT (E) and on *crp* deletion (F). The color code in all panels is as follows: *araC*, black; *gutM*, green; *galS*, red. (C) and (E) are the same as Figs. 3B and 3C but are repeated here for the sake of clarity.

3.6 Could the architecture of regulatory circuits be responsible of this hierarchical behavior?

Given the displayed activities of specific regulators for alternative sugar consumption we decided to analyze

the operator regions of the corresponding promoters, Fig. 7. It has been proposed that CRP recruits the *E. coli* RNA polymerase differentially to distinct promoters (Parkinson et al., 1996; Lee et al., 2012). As a result, CRP has different modes of activation. Briefly, in Class I activation, CRP is bound to an upstream site of the -35 element of promoters and contacts RNA polymerase throughout their α CTD subunit. In Class II activation, CRP binds to a site that overlaps the -35 element and contacts RNA polymerase throughout the domain 4 of σ^{70} subunit. Of the promoters here studied *araCp* correspond to the class I whereas *galS* and *gutMp* correspond to the class II (Fig. 7). In addition, there are some other differences on the promoters' architecture. For instance, *araCp* regulation involves a DNA loop by interaction of AraC dimers (Gallegos, 1997) and the binding sites for CRP and AraC fall inside this loop. In the case of *gutMp*, although there is evidence of regulation by GutM and SrlR, the DNA binding sites for these regulators have not been identified. However, it seems that SrlR repress the transcription of *gutMp* and GutM positively auto-regulates it (Yamada and Saier, 1988). Regarding *galSp*, it has a DNA binding site that seems to be the target of two repressors: GalS and GalR. These regulators bind such a site with different affinities because they share a conserved N-terminal domain that determines the affinity for the DNA-binding region (Geanakopoulos and Adhya, 1997). In addition to the different promoter architectures, the self-regulatory circuits are different for each of these regulatory systems: AraC is subject to both positive and negative self-regulation; GutM seems to have dual regulation, positive self-regulation and negative regulation by a different specific regulator, and GalS is repressed both by itself and by and related apparently constitutive repressor. All of this stresses the necessity of additional experiments to determine whether promoter or regulatory architectures are capable of explaining the observed promoter activities.

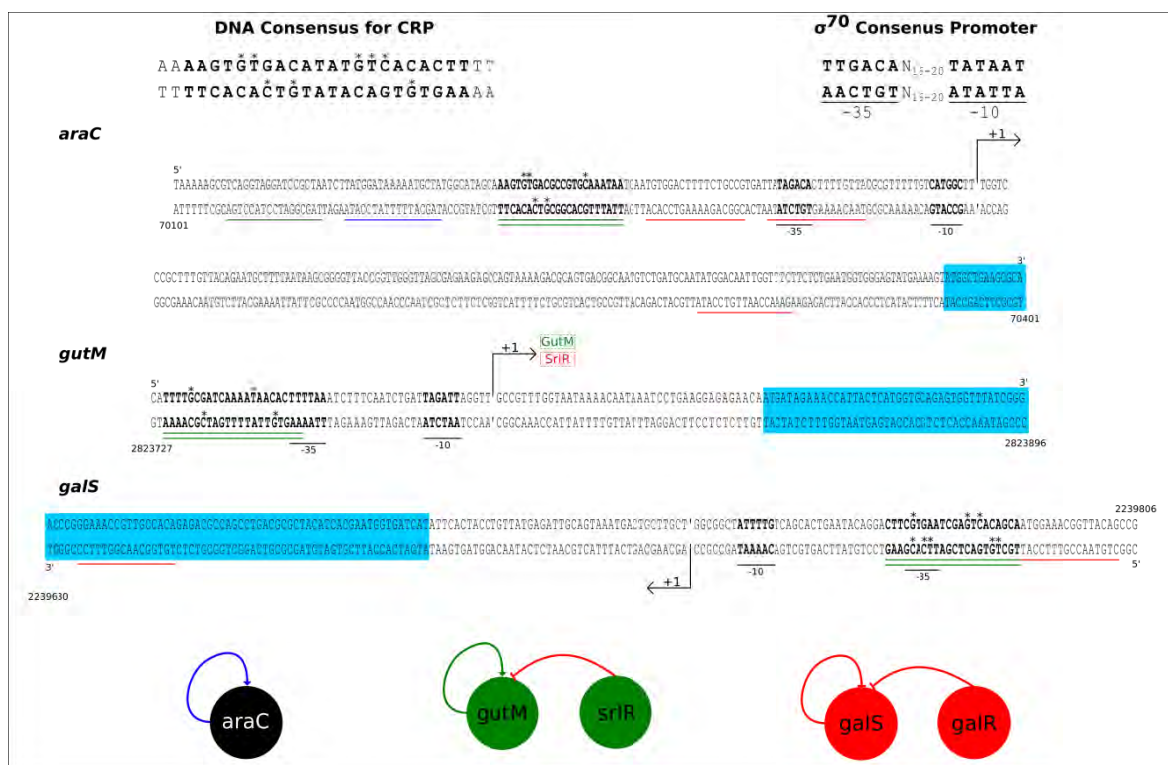


Fig. 7 Binding sites of CRP and σ^{70} in the studied promoters. The consensus binding sites for CRP and σ^{70} are shown. Nucleotides marked with asterisks in the consensus for CRP are those that interact with the CRP protein (Parkinson et al., 1996). Numbers at the end of each sequence denote genome positions of promoters. The initial translated amino acids of all regulatory proteins are shown in cyan. The DNA self-regulatory binding sites for each transcription factor are represented with single underlines. Double underlines are employed to denote the DNA binding sites for CRP. The -10 and -35 elements where the σ^{70} subunit of RNA polymerase binds are also marked. Colored underlines denote the type of regulation: green for activation, red for repression and blue for dual regulation. Finally, simplified schemes of regulatory switches for promoters are shown.

4 Conclusions

It has been known for a long time that *E. coli* preferably consumes glucose over other carbon sources. However, we lack a complete knowledge of how this is achieved and regulated at the molecular level. In this work we present a proof of principle that permits to track the hierarchical use of carbon sources by following bacterial growth and promoter activities of the regulatory proteins that respond to specific sugars. We were able to identify the following order for the preferential use of carbon sources by *E. coli*: glucose > arabinose > sorbitol > galactose. A detailed analysis of regulator promoters for the corresponding catabolic operons indicates that this behavior can be due to at least three factors: 1) the coordinated activation of local regulators by the global regulator CRP, 2) the asymmetrical responses of transcription factors for specific and unspecific sugars and, 3) the architecture of promoters and operon-regulatory circuits. However, many questions remain open regarding the control mechanisms leading to this hierarchical behavior. Answering them will require a large amount of both experimental and mathematical modeling work. Finally, *E. coli* can consume more carbon sources than the ones here studied. It is still pending to test them to have a more complete scheme regarding the preferential use of carbon source in this bacterium.

Acknowledgments

We thank Alejandro Hernández-Morales and Susana Ruiz for their assistance in the construction of transcriptional fusions. This work was supported by CONACYT grant 103686 including an undergraduate fellow to R.S.P.-A. A.M.-A. conceived the study. R.S.P.-A. performed experiments. M.S., R.S.P.-A. and E. G.-V. analyzed data and prepared figures. A.M.-A. and M.S. drafted the manuscript. All the authors approved the manuscript.

References

- Baba T, Ara T, Hasegawa M, et al. 2006. Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Molecular Systems Biology*, 2: 0008
- Beisel CL, Storz G. 2011. The base-pairing RNA spot 42 participates in a multi output feedforward loop to help enact catabolite repression in *Escherichia coli*. *Molecular Cell*, 41: 286-297
- Chen S, Oldham ML, Davidson AL, et al. 2013. Carbon catabolite repression of the maltose transporter revealed by X-ray crystallography. *Nature*, 499: 364-368
- Cohn M. 1957. Contributions of studies on the beta-galactosidase of *Escherichia coli* to our understanding of enzyme synthesis. *Bacteriology Review*, 21: 140-168
- Epstein W, Rothman-Denes LB, Hesse J. 1975. Adenosine 3':5'-cyclic monophosphate as mediator of catabolite repression in *Escherichia coli*. *Proceedings of the National Academy of Sciences of USA*, 72: 2300-2304
- Gallegos MT, Schleif R, Bairoch A, et al. 1997. AraC/XylS family of transcriptional regulators. *Microbiology and Molecular Biology Reviews*, 61: 393-410
- Geanacopoulos M, Adhya S. 1997. Functional characterization of roles of GalR and GalS as regulators of the gal regulon. *Journal of Bacteriology*, 179: 228-234
- Görke B, Stülke J. 2008. Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nature Reviews Microbiology*, 6: 613-624
- Gottesman S. 1984. Bacterial regulation: global regulatory networks. *Annual Review of Genetics*, 18: 415-441

- Jacob F, Monod J. 1961. Genetic regulatory mechanisms in the synthesis of proteins. *Journal of Molecular Biology*, 3: 318-356
- Jones-Mortimer MC, Kornberg HL. 1974. Genetic control of inducer exclusion by *Escherichia coli*. *FEBS Letters*, 48: 93-95
- Lee DJ, Minchin SD, Busby SJ. 2012. Activating transcription in bacteria. *Annual Review of Microbiology*, 66: 125-152
- Liu M, Durfee T, Cabrera JE, et al. 2005. Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *Journal of Biological Chemistry*, 280: 15921-15927
- Longabaugh WJR, Davidson EH, Bolouri H. 2009. Visualization, documentation, analysis, and communication of large-scale gene regulatory networks. *Biochimica et Biophysica Acta*, 1789: 363-374
- Magasanik B. 1961. Catabolite repression. *Cold Spring Harbor Symposium on Quantitative Biology*, 26: 249-256
- Martínez-Antonio A, Collado-Vides J. 2003. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 6: 482-489
- Martínez-Antonio A, Velázquez-Ramírez DA, Mondragón-Sánchez J, et al. 2012. Hierarchical dynamics of a transcription factors network in *E. coli*. *Molecular BioSystem*, 8: 2932-2936
- Monod J. 1942. *Recherchessur la Croissance des Cultures Bactériennes*. Hermann et Cie, Paris, France
- Park YH, Lee BR, Seok YJ, et al. 2006. In vitro reconstitution of catabolite repression in *Escherichia coli*. *Journal of Biological Chemistry*, 281: 6448-6454
- Parkinson G, Wilson C, Gunasekera A, et al. 1996. Structure of the CAP-DNA complex at 2.5 Å resolution: a complete picture of the protein-DNA interface. *Journal of Molecular Biology*, 260: 395-408
- Rabinowitz JD, Silhavy TJ. 2013. Systems biology: metabolite turns master regulator. *Nature*, 500: 283-284
- Salgado H, Peralta-Gil M, Gama-Castro S, et al. 2013. RegulonDB v8. 0: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Research*, 41: D203-D213
- Sellitti MA, Pavco PA, Steege DA. 1987. Lac repressor blocks in vivo transcription of lac control region DNA. *Proceedings of the National Academy of Sciences of USA*, 84: 3199-3203
- Tagkopoulos I, Liu YC, Tavazoie S. 2008. Predictive behavior within microbial genetic networks. *Science Signaling*, 320: 1313
- Walker HH, Winslow CE, Mooney MG. 1934. Bacterial cell metabolism under anaerobic conditions. *Journal of General Physiology*, 17: 349-357
- Wright PR, Richter AS, Papenfort K, et al. 2013. Comparative genomics boosts target prediction for bacterial small RNAs. *Proceedings of the National Academy of Sciences of USA*, 110: E3487-96
- Yamada M, Saier MH Jr. 1988. Positive and negative regulators for glucitol (gut) operon expression in *Escherichia coli*. *Journal of Molecular Biology*, 203: 569-583
- You C, et al. 2013. Coordination of bacterial proteome with metabolism by cyclic AMP signalling. *Nature*, 500: 301-306
- Zaslaver A, Bren A, Ronen M, et al. 2006. A comprehensive library of fluorescent transcriptional reporters for *Escherichia coli*. *Nature Methods*, 3: 623-628

RESEARCH ARTICLE

Transcription Factors Exhibit Differential Conservation in Bacteria with Reduced Genomes

Edgardo Galán-Vásquez, Ismael Sánchez-Osorio, Agustino Martínez-Antonio*

Center for Research and Advanced Studies of the National Polytechnic Institute, Campus Irapuato, Genetic Engineering Department, Cinvestav, Km. 9.6 Libramiento Norte Carr. Irapuato-León 36821, Irapuato Gto, México

* amartinez@ira.cinvestav.mx



OPEN ACCESS

Citation: Galán-Vásquez E, Sánchez-Osorio I, Martínez-Antonio A (2016) Transcription Factors Exhibit Differential Conservation in Bacteria with Reduced Genomes. PLoS ONE 11(1): e0146901. doi:10.1371/journal.pone.0146901

Editor: Marc Robinson-Rechavi, University of Lausanne, SWITZERLAND

Received: August 11, 2015

Accepted: December 23, 2015

Published: January 14, 2016

Copyright: © 2016 Galán-Vásquez et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the paper and its Supporting Information files.

Funding: The authors received no specific funding for this work.

Competing Interests: The authors have declared that no competing interests exist.

Abstract

The description of transcriptional regulatory networks has been pivotal in the understanding of operating principles under which organisms respond and adapt to varying conditions. While the study of the topology and dynamics of these networks has been the subject of considerable work, the investigation of the evolution of their topology, as a result of the adaptation of organisms to different environmental conditions, has received little attention. In this work, we study the evolution of transcriptional regulatory networks in bacteria from a genome reduction perspective, which manifests itself as the loss of genes at different degrees. We used the transcriptional regulatory network of *Escherichia coli* as a reference to compare 113 smaller, phylogenetically-related γ -proteobacteria, including 19 genomes of symbionts. We found that the type of regulatory action exerted by transcription factors, as genomes get progressively smaller, correlates well with their degree of conservation, with dual regulators being more conserved than repressors and activators in conditions of extreme reduction. In addition, we found that the preponderant conservation of dual regulators might be due to their role as both global regulators and nucleoid-associated proteins. We summarize our results in a conceptual model of how each TF type is gradually lost as genomes become smaller and give a rationale for the order in which this phenomenon occurs.

Introduction

Transcription is an essential molecular process through which cells respond and adapt to changing environmental conditions, such as different kinds of stress and nutrient deficiencies [1]. The regulation of this dynamic process is usually carried-out by proteins called transcription factors (TFs), which bind to upstream regions (called promoters) of target genes (TGs) and promote or inhibit the synthesis of RNA molecules (and the subsequent production of proteins). TFs can be classified as activators, repressors, or dual regulators according to whether they promote, prevent, or exert both regulatory actions on the transcription of genes,

respectively [2]. Although other molecular regulatory mechanisms might be operating at other stages of the gene expression process, such as DNA-methylation, RNA interference, etc., transcriptional regulation mediated by TFs is the predominant type of control in gene expression [3].

When the product of a regulated gene is a TF that regulates its own expression or the expression of other genes, the resulting regulatory interactions and the corresponding genes can be conceptualized as edges and nodes in a transcriptional regulatory network (TRN) whose topology dictates a hierarchy between regulators and target genes [4]. According to the number of genes they regulate, TFs can be ranked and hence defined as global regulators, when they act as hubs (highly connected nodes) in a network; and local regulators, when they regulate few genes, usually an operon or genes represented as terminal nodes in a regulatory network [5].

TRNs organize the responses of organisms to particular conditions and, despite their diversity, they share several topological features, including conserved network motifs, similar hierarchies, and scale-free structures [4, 6, 7]. These general patterns have been uncovered thanks to the availability of whole genome sequences of several organisms and to advances in detailed experimental techniques for the detection of protein-DNA interactions. In spite of the fact that network topologies obtained from computational methods alone are incomplete (or may contain links not supported to date by physical evidence) [8], analyses conducted mainly in free-living bacteria have revealed that their elements tend to change considerably, with the set of nodes corresponding to regulatory genes undergoing radical changes compared with the non-regulatory ones, which are more conserved among genomes [9–11].

On the one hand, the above alterations in the elements of a network may involve mutations that occur at a genome level, such as single nucleotide substitutions or those produced by the action of transposons, which affect one or a few nucleotides and can lead to the creation or deletion of DNA-binding sites on promoter regions [12]. On the other hand, network alterations also include changes that arise from gene duplication or horizontal gene transfer events that add large DNA fragments (containing one or more genes) to the genomes [13, 14]. Through these evolution-driven processes some genes and interactions are gained while others are lost. The appraisal of this phenomenon raises some intriguing questions concerning the evolution of TRNs (*i.e.* the gain or loss of nodes and interactions in a network throughout time) [14]. For example: How does the structure of TRNs evolve in different organisms? Is there a tendency in the organization of regulatory networks for organisms undergoing massive loss of their genes? Is a particular type of regulatory gene favored during evolution such that the regulators they code for are more conserved than others? If so, what cellular functions are they regulating?

Some of the above questions have been partially addressed by studying the TRNs of model organisms, revealing interesting facts. For example, Teichmann and Babu [15] found that more than two-thirds of the known interactions in the TRNs of *E. coli* and *Saccharomyces cerevisiae* evolved as a consequence of gene duplication. Over one-half of these regulatory interactions were inherited from ancestral duplications of TFs and target genes. Only a small portion of the remaining fraction of regulatory interactions (not evolved by duplication) evolved by gene recombination or innovation. In the same line, Lagomarsino *et al.* [16] discovered that horizontally transferred genes (which impact on network growth) mostly contribute to cellular fitness under very particular conditions, but they are not essential in unstressful environments. These genes generally locate at the bottom of the network hierarchy in the TRN of *E. coli*. In another related work, Molina and van Nimwegen [17] studied how the average number of regulatory sites per intergenic region and the average number of sites regulated by a particular TF vary with genome size (across 105 free-living bacteria). They found that the structure of TRNs varies substantially with genome size. More precisely, they concluded that small genomes code

for few TFs (each binding to a large number of target sites) while large genomes contain many TFs (each binding to a few sites), in which case TFs seem to be more specialized.

While relevant for the understanding of network evolution, the studies described previously have not considered the loss of nodes or interactions in regulatory networks, perhaps due to the scarcity of whole genome data (particularly that of endosymbionts and obligate pathogens) until the last decade. In this work, we investigated the evolution of TRNs by considering organisms undergoing loss of their genes, with a particular interest in the way in which genes involved in regulatory functions are lost according to their type. As it is not trivial to determine from the genome size alone whether an organism has undergone reduction or growth of its genome, we selected as the subject of study those bacteria in the γ -proteobacteria class with genomes smaller than that of *E. coli*, 19 of which are thought to have evolved into obligate pathogens and endosymbionts. These 19 organisms, which are restricted to live inside other organisms [18], are believed to have been subjected to a Muller's ratchet process, a phenomenon that results in the accumulation of slightly deleterious mutations [19]. These mutations in turn bring about genomes with low G+C content, accelerated rate of nucleotide substitution, and loss of adaptive codon bias [19, 20].

E. coli is phylogenetically related to endosymbiotic γ -proteobacteria. This is an advantage for our purposes, since the TRN of *E. coli* is the most studied and well characterized among all bacteria. Hence, we used this regulatory network as a template to reconstruct the networks of the bacteria studied in this work. We followed a comparative genomics approach to identify conserved elements in the networks of the different organisms and applied a combination of correlation analysis with phylogenetically independent contrasts [21] to correct for statistical dependencies induced by phylogenetic relationships (see [Methods](#), section 2). The picture that have emerged from our study is that there is an order in the way TFs are conserved: on the one hand, activators (which represent the major proportion of TFs in large genomes), followed by repressors, become less abundant as genome size decreases, eventually disappearing in conditions of extreme genome reduction; on the other hand, dual regulators become more preponderant as genomes get smaller, being the last TF type lost under extreme genome reduction. Moreover, we observed that such TFs are global regulators and some of them are nucleoid-associated proteins (NAPs) involved in the control of nucleoid structure. All our findings were integrated into a conceptual model that portrays network evolution in organisms undergoing genome reduction and describe the order in which this phenomenon happens. Finally, we highlight the essential ideas arising from our study and discuss their implications in the understanding of gene regulation in reduced genomes.

Methods and Data

To study the evolution of TRNs from the perspective of reduction, we selected a set of 113 genomes belonging to organisms phylogenetically related to *E. coli* and reconstructed the corresponding transcriptional networks using a comparative genomics approach (section 2.1). After assessing the preferential conservation of TFs in terms of their regulatory nature, by formulating hypothesis tests for each type of TF in the 113 genomes (as described in section 2.2), we applied the method of phylogenetic contrasts (see section 2.3) to correct for dependencies in our data and calculate correlations between each type of regulation (*i.e.*, activation, repression, and dual regulation) and genome size. To account for the influence of other variables in the observed correlations, we considered global regulators -which we had to first identify using a metric developed in a previous work (see section 2.4)- and NAPs as additional factors. We subsequently performed multivariate linear regressions, defining as criterion variables the relative fractions of each TF type, and as predictors the remaining variables of interest, specifically:

genome size, global nature of regulators, and their function as NAPs (section 2.5). From the regression coefficients obtained in the latter analysis, we could infer the contribution of each factor of interest on the conservation of each type of regulator.

2.1 Reconstruction of Transcriptional Regulatory Networks

We selected a set of 113 genomes of γ -proteobacteria whose sizes range from 0.159 Mbp (corresponding to *Candidatus Carsonella Ruddii PV*) to 4.639 Mbp (corresponding to *E. coli K-12 MG1655*), trying as much as possible to reduce the redundancy of genomes that belong to a same species and keeping the difference in genome size between pairs of contiguous genomes below 400 Kbp. From these 113 genomes, which are smaller than the genome of *E. coli*, 19 belong to symbionts. Information on complete genome sequences was downloaded from the NCBI genomes website [22]. The TRNs of the 113 genomes were reconstructed by the conventional Regulog approach [23]. This comparative genomics technique aims at finding conserved transcriptional regulatory interactions in organisms where knowledge of their TRNs is missing, using for this purpose information about the interactions in a known (reference) regulatory network. Regulog infers regulatory interactions based on the assumption that orthologous TFs generally regulate the transcription of orthologous target genes (TGs). For example, if a TF and its TG in *E. coli* have orthologous *TF'* and *TG'* in other organism, then a regulatory interaction is inferred as conserved in the target genome [23]. Thus, using the well-characterized transcriptional network of *E. coli* as a template (composed of 1,784 nodes and 4,058 interactions between TFs and their TGs) [24], we identified orthologous TFs and TGs in each target organism via the classic bidirectional best-hit method [25] (see [S1 Table](#)).

For each protein in the transcriptional network of *E. coli*, we did a Blast search against the genome of each of the 113 γ -proteobacteria selected in this study ([Fig 1](#)). The best hit obtained (for each genome) was in turn used as a query sequence in a Blast search against the genome of *E. coli*. We define orthologous proteins when the best hit in the last step (when the search parameters are reversed) corresponded to the protein in the genome of *E. coli* originally used in the first query. Orthologs were accepted if they had an *e*-value $< 1e-6$, sequence identity $> 30\%$, and alignment length $> 60\%$ of the individual proteins.

Our approach to the reconstruction of TRNs involves two basic assumptions. The first is that TFs retain their type of regulatory function (*i.e.*, activation, repression and dual regulation) throughout all the studied genomes, and that the corresponding TGs conserve their TF binding sites. The second assumption is that if the sequences corresponding to TFs and TGs are conserved, then regulatory interactions are conserved too. As a consequence, the impact those mutations of genes and promoter sequences have on the regulatory interactions are not considered. The fact that the 113 genomes compared in our analysis belong to closely related organisms makes our network inferences more reliable, as we circumvent to a certain extent some of the limitations inherent in our method of reconstruction; however, the results and conclusions derived from this method are valid to the extent that the underlying assumptions do not contradict experimental evidence. For example, it is known that the assumptions described above become inadequate as the evolutionary distance increases and that some orthologs with high sequence similarity may not be functionally conserved, in which case the reconstruction approach would be generating false positives.

The 113 genomes were divided into four categories as defined by Moran's group [18]: the first included bacteria classified as free-living, with genomes smaller than that of *E. coli*; the second contains bacteria identified as host-restricted endosymbionts or pathogens; the third contains obligate pathogens or endosymbionts; and the fourth contains endosymbionts exhibiting extreme genome reduction (see [S2 Table](#)). Additionally, from information obtained from



Fig 1. Phylogeny of the γ -proteobacteria studied in this work. The tree lists the 113 bacteria considered in this study. Free-living bacteria are in orange, host-restricted symbionts and pathogens in yellow, obligate symbionts and pathogens in violet, tiny-genomes symbionts in blue, and *E. coli* K-12 MG1655 in red. The phylogeny was constructed using the software MrBayes-3.2 based on the 16S rRNA genes.

doi:10.1371/journal.pone.0146901.g001

EcoCyc [26], 190 TFs out of the 196 known to regulate at least one gene in the TRN of *E. coli* were classified as activators, repressors, or dual regulators, according to the type of net effective regulation they exert over their TGs (6 of these TFs do not have information about their type of regulatory action). We also classified and determined the biological functions of all the orthologous genes, found in previous steps, using the parental categories of Gene Ontology [27, 28], (see S3 Table).

2.2 Analysis of Transcription Factor Conservation

To decide if reduction of genome size favors the conservation of a particular type of TF (*i.e.*, activator, repressor, or dual regulator), we performed statistical tests for each of the 113 studied genomes, defining as null hypotheses those statements indicating a reduction or no variation in the relative genome fraction of each type of regulator with respect to the corresponding fraction in the genome of *E. coli*. Accordingly, the alternative hypotheses were formulated as the logical complements to such statements. More precisely, the general form of the tests for the case of activators is as follows:

$$H_{0|i}^a : f_i^a \leq f_r^a$$

$$H_{1|i}^a : f_i^a > f_r^a$$

Where $H_{0|i}^a$ (with the index i ranging from the 1st to 113th genome) denotes the null hypothesis that the fraction of activators in the i^{th} genome (f_i^a) is less or equal to the corresponding fraction of activators in the reference genome (f_r^a). An identical form of the test applies to repressors and dual regulators. When the null hypotheses are rejected, the corresponding alternative hypotheses ($H_{1|i}^a, H_{1|i}^r$ or $H_{1|i}^d, 1 \leq i \leq 113$, as the case may be) -that there is in fact a preferential conservation of a particular type of regulator- are supported. If, on the contrary, the relative fraction of a regulator in a genome i is not significantly higher than the fraction in the genome of *E. coli*, then we fail to reject the corresponding null hypothesis and, in consequence, failed to support the alternative hypothesis.

To compute the p -values and determine the statistical significance (at the 0.05 level) of the tests, we used a hypergeometric distribution for each kind of TF in the 113 genomes. The reason for this is that such distribution describes the probability that given a reference genome (in this case that of *E. coli*) coding for N TFs in total, from which K TFs are of a particular type (activator, repressor, or dual regulator), k elements from the latter appear in another genome that contains n orthologous. We generated three hypergeometric distributions for each genome used in this study (one for each type of regulation) with defining parameters n and k ; where n , representing the number of orthologous TFs, is fixed in each genome and k , which varies from 0 to n , is the number of TFs of the particular regulatory type (activator, repressor, or dual regulator), which corresponds to a relative genome fraction whose probability of occurrence we want to compute. The processing of data and computations of parameters were performed in R (see S4 Table) [29].

To assess an overall statistical significance from all the 113 tests performed for each type of TF, we combined the p -values obtained from the set of independent tests corresponding to activators, repressors, and dual regulators, respectively, using Fisher's method [30, 31]. A requirement for the application of this kind of meta-analysis is that the same undelaying form of the hypothesis test is used in the independent, more elementary tests, which is valid for the present study.

2.3 Analysis of the Influence of Phylogenetic Relationships

To discern how the genome fraction of each type of regulator (*i.e.*, activator, repressor, dual) varies with the reduction in genome size, we calculated the Pearson correlation coefficient, which measures the degree and direction of linear relationships between these variables. Because the genomes being studied belong to closely related organisms that are part of a hierarchically structured phylogeny, our data points cannot be considered as statistically independent from one another. This dependence (caused by phylogenetic similarity) may generate significant correlations between the number of TFs and genome size when no causal link exists between them. To correct for such effects, we incorporated phylogenetic information into our statistical analysis using Felsenstein's independent contrasts [21]. This method computes weighted differences (called contrasts) between the trait values associated to pairs of nodes at each bifurcation point in a known phylogeny, assuming evolution follows a random walk pattern in time. The resulting contrasts are, in principle, independent and normally distributed, and thus can be used in conventional statistical analyses. A detailed account of the method can be found in the references [21, 32, 33].

We used the PDAP:PDTREE module [33, 34] of Mesquite software version 3.02 [35] to calculate standardized phylogenetically independent contrasts for the following variables: genome size, frequency of activators, frequency of repressors, frequency of dual regulators, fraction of activators, fraction of repressors, fraction of dual regulators, fraction of global regulators and fraction of NAPs (see S5 Table), and subsequently determined the correlations and regressions through the origin. This approach demands that contrasts in the X-axis be "positivized" and that the original dataset has a normal distribution [32], a requirement that has been met in this work. The phylogenetic tree necessary for the calculation of contrasts was constructed with the program MrBayes-3.2, which implements a Bayesian inference method [36, 37]. To construct the tree, we employed the 16s ribosomal gene sequences of all the 114 bacterial genomes (including *E. coli*) used in this study and performed 2 runs with 1×10^7 generations, a 25% burn-in, and sampling every 1000th iteration.

After calculating the phylogenetic contrasts, as described in the above paragraphs, we performed significance tests using a *p*-value of 0.05 to determine whether the obtained correlations between the calculated contrasts could be attributed to chance. Because the correlations turned out to be significant, we accepted the correlations as measuring statistically dependent variation and showing an underlying relationship between the variables.

2.4 Identification of Global Regulators

TRNs contain highly connected nodes called global regulators or regulatory hubs, which contribute to the structural cohesion and robustness of the networks and to their functional coherence [38]. In a previous work [5], we described the operational criteria for the identification of a global regulator, which included: i) number of regulated TGs; ii) number of TFs to which it co-regulates, iii) number of coupled sigma factors in the regulation of TGs; and iv) number of TFs that it regulates. From the definitions of these parameters, we proposed in Galan E., *et al.* [39] an equation (see Eq 1) that associates, to every transcription factor *x*, a value $G(x) \in [0,1]$ indicating its global activity in the context of a regulatory network:

$$G(x) = \frac{1}{4} \left(\frac{TFR(x)}{N_{TF} + N_{SF} - 1} + \frac{GR(x)}{N_G} + \frac{SF(x)}{N_{SF}} + \frac{CR(x)}{N_{TF} - 1} \right) \quad (1)$$

Where N_{TF} is the number of TFs, N_{SF} the number of sigma factors, and N_G the number of non-regulatory genes in the network. On the other hand, $TFR(x)$ represents the number of TFs regulated by *x*, $GR(x)$ the number of non-regulatory genes regulated by *x*, $SF(x)$ the number of

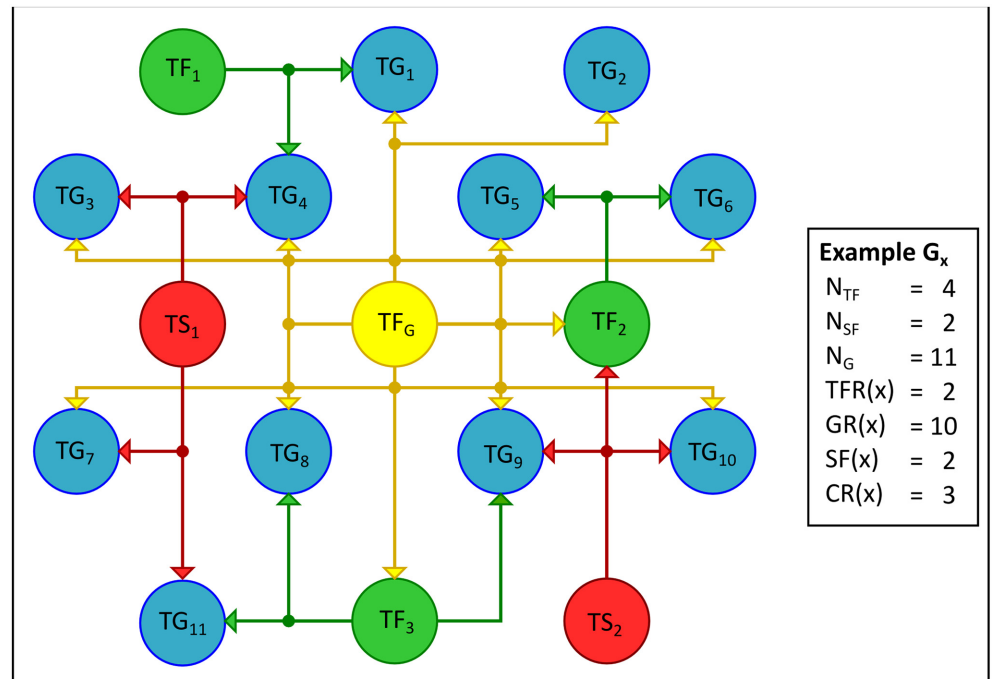


Fig 2. Identification of global regulators. This figure exemplifies the calculation of the metric for a global regulator TF_G (yellow node) in a hypothetical network consisting of 17 nodes and 26 interactions, in which TF_G regulates 10 TGs (blue nodes) and 2 TFs (green nodes); in addition, it co-regulates along with 2 sigma factors (red nodes) the nodes labeled as TG_3 , TG_4 , TG_7 , TG_9 , TG_{10} , and TF_2 ; and along with 3 TFs, the following target genes: TG_1 , TG_4 , TG_5 , TG_6 , TG_8 , and TG_9 . From the metric described by Eq (1), we obtain a value of $G(TF_G) = 0.8272$.

doi:10.1371/journal.pone.0146901.g002

sigma factors required by x to regulate its TGs, and finally, $CR(x)$ the number of TFs co-regulating with x other TGs. The metric described by Eq (1) defines a hierarchy of global regulators in the TRN of *E. coli* and permits the identification of conserved global regulators in each of the reconstructed TRNs (see Fig 2 for an example).

2.5 Multivariate Regression Analysis

As the conservation of the fractions of each TF type may be influenced not only by genome size but also by the global nature of regulators, as well as the fact that they act as NAPs, we assessed the effects of these additional variables by performing multivariate linear regressions. In these equations we defined as dependent variables the relative fractions of each TF type (also known as criterion variables), and include as independent variables (or predictors) the factors whose influence we want to measure; specifically, genome size, global nature of regulators, and their function as NAPs. The partial regression coefficients obtained from the multiple regressions allowed us to discern the contribution of each variable of interest on the conservation of each type of regulator when the effect of the remaining predictors is held constant.

In summary, we found the coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ of the multivariate Eq (2), using a built-in R routine implementing a least squares method.

$$Y_{TF} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \epsilon \tag{2}$$

Y_{TF} is the least squares prediction of the relative fraction corresponding to a particular type of TF and the variables X_1, X_2, X_3 represent the predictors (genome size, global regulators, and function as NAPs, respectively) whose contribution to the criterion variable, Y_{TF} , is addressed in this analysis. The error ε is assumed to have a normal distribution with mean 0 and constant variance σ^2 . β_0 is the Y intercept of the hyperplane defined by Eq (2) and is interpreted as the predicted value of Y_{TF} when all the predictors are zero. The other partial regression coefficients ($\beta_1, \beta_2, \beta_3$) are the corresponding slopes of Y_{TF} on each of the predictors, keeping the remaining variables fixed [40]. Being coefficients expressed in different units, the β 's cannot be directly used to compare the contribution of each predictor on the value of Y_{TF} . To circumvent this problem, we transformed the independent variables (X_1, X_2, X_3) into standard deviates or “Z measures” and obtained the corresponding standardized partial regression coefficients (b_1, b_2, b_3).

After transforming the β 's into b 's, we used the relative magnitude of b coefficients (considering their p -values) to compare the strength of the relationship between Y_{TF} and each particular predictor, when all the other predictors in the regression equation are kept constant. Thus, we could infer which predictor had the strongest influence on the conservation of each type of regulator [41].

3 Results and Discussion

3.1 Conservation of Activators, Repressors and Dual Regulators

In this work, we studied the evolution of TRNs in bacteria taking as a reference the regulatory network of *E. coli*. From this network, we reconstructed 113 TRNs of γ -proteobacteria with genomes smaller than that of *E. coli*. This was done following a comparative genomics approach (as explained in section 2.1). The studied bacteria included 19 obligate pathogens and endosymbionts, which are convenient biological models for the study of regulatory network evolution in the context of extreme genome reduction (Fig 1). We classified 190 transcription factors in the TRN of *E. coli* according to their regulatory activity into activators, repressors, and dual regulators. The presence (or absence) of these regulators was identified in all of the 113 γ -proteobacteria.

To measure the degree of correlation between the type of regulation and genome size, we corrected for phylogenetic dependencies in our data and performed statistical correlation analyses as described in section 2.3. We obtained a coefficient of $r = 0.56$ with p -value = 1.53×10^{-10} , for the correlation between the number of activators and genome size; $r = 0.71$ with p -value = 2.2×10^{-16} , for the number of repressors and genome size; and $r = 0.64$ with p -value = 2.53×10^{-14} , for the number of dual TFs and genome size. These coefficients seemed to indicate a whole picture in which the numbers of TFs decrease as genome size become smaller, in agreement with one of the scaling laws found in [42]. However, since such correlation may be the result of a generalized, random gene loss, we performed significance tests (at the level of 0.05) on the relative fractions of TFs obtained from each of the 113 genomes and for each type of regulation (refer to section 2.2 for further details). After applying Fisher's method [30, 31] (see section 2.2) to combine the resulting independent p -values for each type of regulator, the relative fractions corresponding to activators (p -value = 0.99) and repressors (p -value = 0.99) in the 113 genomes turned out to be not significant (at the 5% level). Nevertheless, the fraction of dual regulators produced a statistically significant result (p -value = 2.90×10^{-50}), which indicates that small genomes favor the conservation of dual TFs (see S4 Table).

To further assess the preference in the conservation of dual TFs as genomes become smaller, a similar analysis to the one described in the previous paragraph was performed, but in this case using the relative frequencies corresponding to each TF type, instead of the net number of

TFs per genome. We found positive correlations for activators and repressors with correlation coefficients of $r = 0.26$ (p -value = $6.28e-3$) and $r = 0.23$ (p -value = $1.62e-2$) respectively, and a negative correlation for dual regulators $r = -0.32$ (p -value = $4.90e-4$) (see Fig 3). Our data and statistical analysis (refer to section 2.3), revealed that from the whole of TFs, activators and repressors tend to be present in a major proportion in large genomes (*e.g.*, those corresponding to free-living organisms) while the fraction of dual regulators become more prevalent in small genomes (*e.g.*, those belonging to obligate endosymbionts). This result suggests certain evolutionary preference in the conservation of TFs during the process of genome reduction, being dual TFs the type of regulator most preserved in genomes exhibiting advanced reduction.

Because we realized that genome size, global regulator nature, and function as NAP, are variables that may affect the conservation of TFs, we performed a multivariable linear regression analysis (as explained in detail in section 2.5) to identify which of these variables have the highest influence on the conservation of each TF type. From the standardized partial regression coefficients computed for each linear regression, we inferred the effect that each predictor has on the conservation of activators, repressors, and dual regulators, respectively. For the case of activators and repressors, we found that the influence of the predictor variables (*i.e.*, genome size, global regulator nature, and function as NAP) was negligible. As can be observed in Table 1, the standardized partial regression coefficients in these cases were very small, with only one of them being significant. In contrast to these results, the standardized partial regression coefficients for dual regulators indicated that the global nature of such regulators (with a $b = 0.49$ and p -value = $2.00e-4$) might explain their preponderant conservation.

Interestingly, a major portion of conserved, dual, global regulators had a role as nucleoid organizing proteins in endosymbionts. When computing the Pearson correlation between global regulators and NAPs, we found a strong relationship (with a correlation coefficient equal to 0.75). Thus, we concluded that such variables were not strictly independent, but were in a sense redundant as they provided similar information. In fact, removing any of them (but not both) from the corresponding multivariable regression equation did not have a notable effect on the predicted value of the dependent variable (fraction of dual regulators).

3.2 Biological Interpretation of TF Conservation in the Context of Genome Reduction

The above findings have a biological explanation in the light of the functions associated to each kind of regulator. In the case of activators, which generally depend on co-inducers to respond to intra- or extracellular signals (such as adaptive response, antibiotics, virulence factors, etc.) [43–45], their loss might not substantially affect the expression of their TGs, because the non-restrictive structure of nucleoids in prokaryotes with small genomes facilitates the access of the transcriptional machinery to promoter regions, thus making basal transcription of genes possible [46]. In addition to this, recent studies have provided strong evidence that RNA polymerases spend most of their search time for a binding site (around 85% of the total time) nonspecifically bound to DNA [47], which gives a large probability that positively regulated genes may still exhibit some level of expression in spite of the loss of their activators.

In the case of repressors, we found that their proportion becomes increasingly larger than the fraction of activators as genomes get progressively smaller. The biological justification of this result might rely on the fact that the loss of repressors can lead to unnecessary expense of cellular resources due to basal expression of TGs. Besides, the conservation of transcriptional repression might give some slight biological advantage over activation because homeostatic control is usually implemented through negative regulation, a recurrent motif in metabolic and gene regulatory processes [48]. Moreover, it is known that negative auto-regulation (when

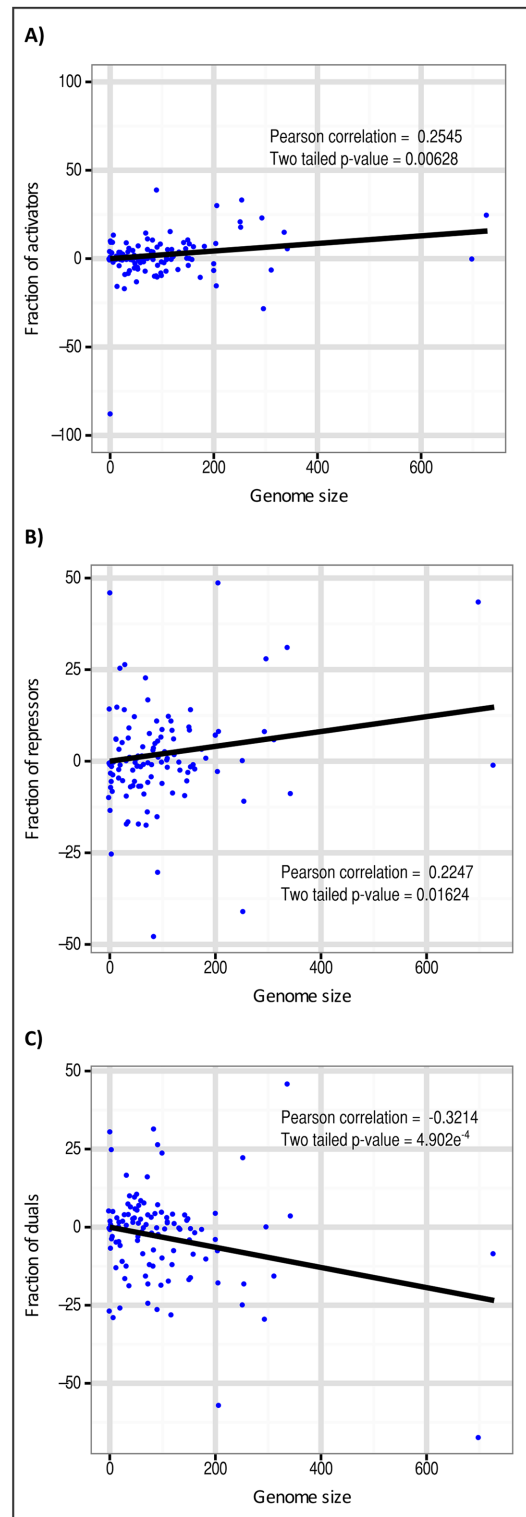


Fig 3. Phylogenetic contrasts of conserved transcription factors. Regression line through the origin and the Pearson correlation with corresponding *p*-value for A) Conservation of activators, B) Conservation of repressors and C) Conservation of dual regulators. Each point represents the contrast between the fraction of each TF type versus genome size (contrasts positized as recommended by Garland [32]).

doi:10.1371/journal.pone.0146901.g003

Table 1. Standardized partial regression coefficients.

| Independent variables | Activators TFs | | Repressors TFs | | Dual TFs | |
|-----------------------|----------------------|-----------------|----------------------|-----------------|----------------------|-----------------|
| | Coefficient <i>b</i> | <i>p</i> -value | Coefficient <i>b</i> | <i>p</i> -value | Coefficient <i>b</i> | <i>p</i> -value |
| Genome size | 0.21 | 0.03 | 0.04 | 0.68 | -0.15 | 0.10 |
| Global regulators | -0.10 | 0.50 | 0.12 | 0.42 | 0.49 | 2.00e-4 |
| Function as NAPs | -0.09 | 0.54 | -0.09 | 0.55 | -0.07 | 0.61 |

doi:10.1371/journal.pone.0146901.t001

appropriately tuned) can speed up responses in cellular systems [49, 50] and this dynamical response may contribute to the maintenance of cellular homeostasis regardless the physiological state of the organism [51].

Finally, the preponderance of dual TFs in small genomes might be explained from the fact that these regulators permit more flexible control (positive and negative) over their target genes. In addition, from our observations that dual TFs are usually global regulators and their regulated genes fall in more than one functional class [52, 53], we speculate that dual regulators confer to organisms the ability of maintaining cellular fitness with a minimal regulatory machinery.

The above conjectures can be framed into the demand theory of gene regulation proposed by Savageau [54]. This theory states that the type of regulation required for a gene is a function of the demand of the protein it encodes for during the life-cycle of bacteria; that is, if a protein is required most of the time, their encoding gene should be positively regulated; on the contrary, if a gene product is required only sporadically, their encoding gene should be negatively regulated [54]. In this context, the conservation of a particular type of regulation might be chosen to optimize the use of the regulator [55]. Thus, there should be a balance between the fitness cost and benefit of conserving a set of regulated (or unregulated) genes. We hypothesize that in endosymbiont bacteria this balance is influenced by the relatively stable environment or condition in which they live, which leads to differential loss of their TFs.

3.3 Conceptual Model of TF Conservation in Reduced Genomes

Endosymbiont bacteria exhibit genomic changes associated to the process of genome reduction, leading a free-living organism to acquire a host-restricted lifestyle [18, 56]. Studies conducted in organisms with reduced genomes have been possible until recently when genome sequences of many endosymbiont bacteria have been reported.

As a result of our comparative analysis using these organisms, we have given additional support to previous findings in free-living organisms of a tendency to retain more non-regulatory genes than genes coding for TFs. In a complementary manner, we have also discovered a tendency of dual regulators to be more conserved than activators and repressors in the last stages of genome reduction.

From our analysis, we have developed a conceptual model of TRN evolution driven by genome reduction, which we summarize in Fig 4. Three levels of genome reduction have been already defined according to the changes organisms undergo in their gene content [18]. In the model, the early stage of network reduction is represented by lately evolved, free-living organisms, which turned into host-restricted endosymbionts and intracellular pathogens (e.g., *Sodalis glossinidius* and *Serratia symbiotica*). The genomes in this stage are characterized by the conservation of a large number of TFs that respond to environmental conditions in a manner almost similar to free-living organisms. At the advanced stage, long-term obligate endosymbionts (e.g., *Baumannia cicadellinicola* and *Buchnera aphidicola*) that live in more restricted

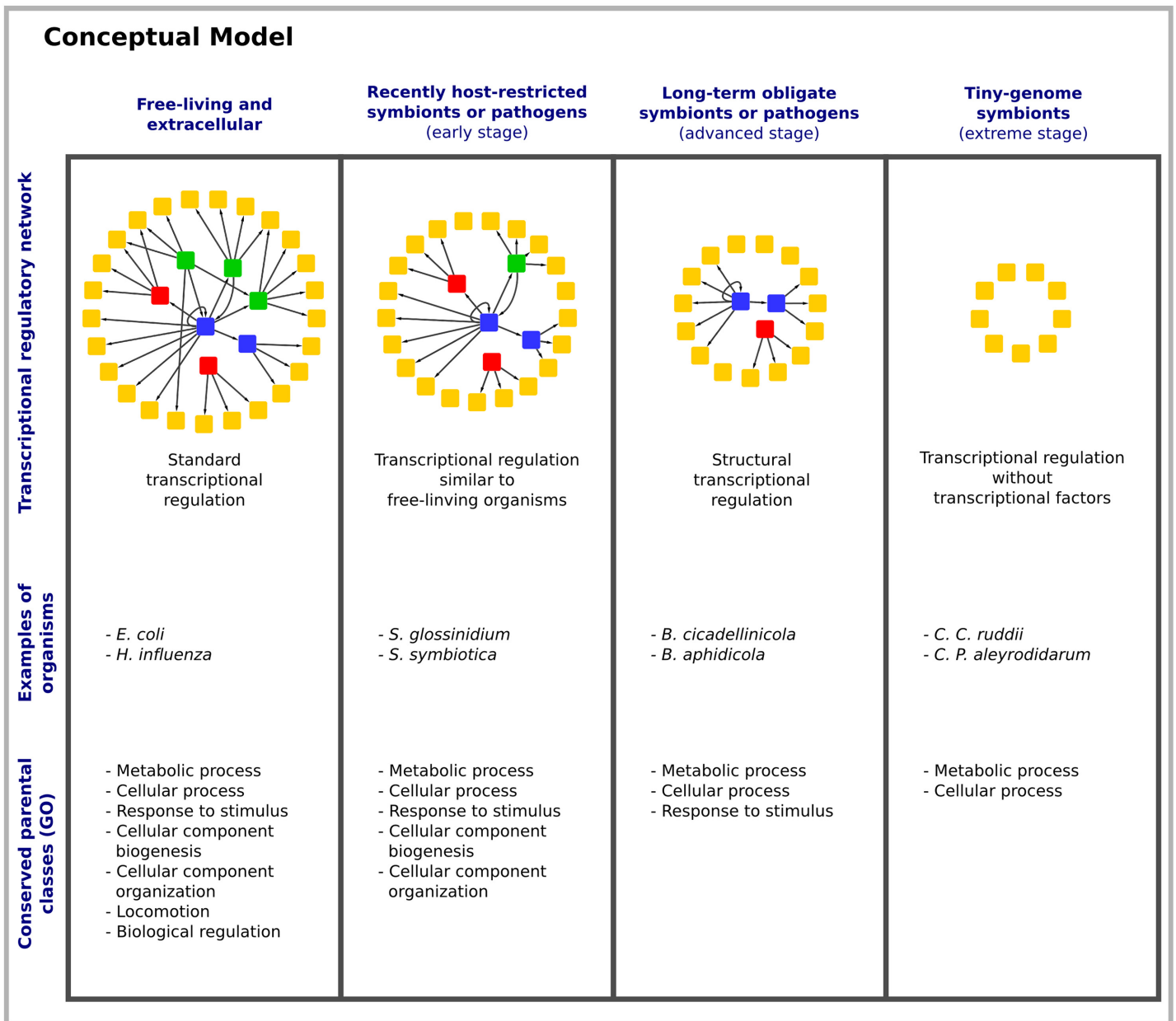


Fig 4. Conceptual model of transcriptional regulatory network evolution under the influence of genome reduction. For this model we used the classification proposed by N. Moran [18]. We start with a standard regulatory network (composed by activators, repressors, and dual regulators) corresponding to a free-living organism, in this case *E. coli*. In the first (early) stage of genome reduction, organisms are recently host-restricted symbionts or pathogens. These organisms have lost different TFs and TGs; however, the whole process of transcriptional regulation is similar to that of free-living organisms. In the second (advanced) stage of genome reduction, organisms turn into obligate symbionts or pathogens. In this stage, the majority of TFs in the TRN has been lost. We surmise that, in these bacteria, transcriptional regulation is exerted mainly at the structural level of DNA. In the last (extreme) stage, organisms exhibit extreme genome reduction and they no longer conserve TFs. Examples of organisms in each group are included, as well as the conserved parental classes of their gene-products according to the gene ontology classification.

doi:10.1371/journal.pone.0146901.g004

habitats (generally inside insect-specialized cells) conserve a major proportion of NAPs; thus, in these organisms gene regulation may be operating mainly at a global level through nucleoid organization. In this same stage, the scenario in which all target genes regulated by a TF are lost (but not the TF itself) may happen. In such situation, we speculate that TFs assume more

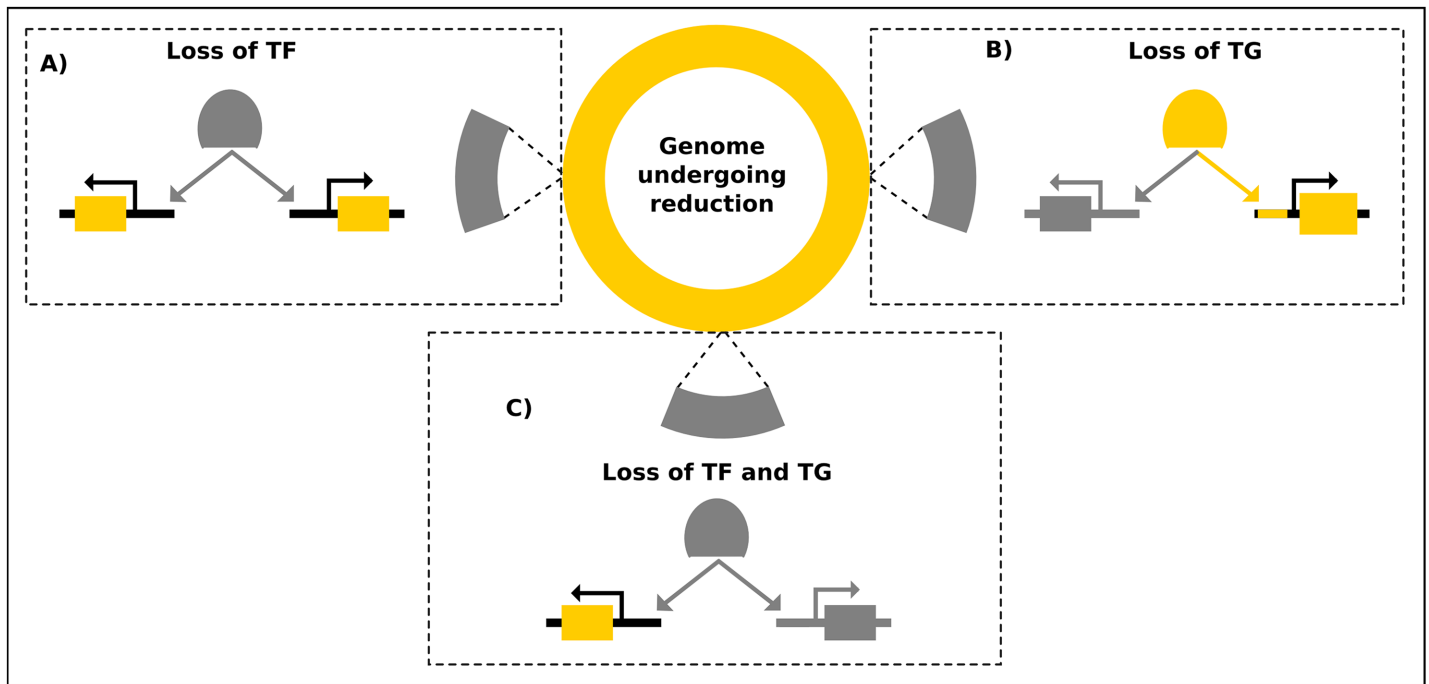


Fig 5. Effect of genome reduction in the transcriptional regulatory network. Behind the loss of genes (and interactions) is the genome reduction phenomenon acting as a force that directs the evolution of TRNs. Three scenarios might occur when a DNA fragment is lost in a genome: A) The fragment contains only TFs, B) The fragment contains only TGs, C) The fragment contains both TFs and TGs.

doi:10.1371/journal.pone.0146901.g005

than one role in the cell. This is the case, for example, of BolA, which can perform both regulatory and catalytic functions, acting as a single multifunctional enzymatic system in organisms lacking two-component sensors [57]. Another example is given by HU, which apart from being involved in regulatory functions acts as a DNA organizing protein [52]. Finally, in the extreme stage, tiny-endosymbionts (e.g., *Candidatus Portiera aleyrodidarum* and *Candidatus Carsonella ruddii*) no longer retain any type of TFs, these bacteria are considered on the edge of being organelles, thus if gene regulation operates in these organisms it might be essentially at the level of intrinsic DNA topology or by physiological conditions [58], but not at the level of transcription initiation assisted by transcription factors.

Conclusions

Regulatory networks in bacteria evolve throughout time following patterns of growth or reduction. That is, by means of several well-known molecular events (such as, horizontal gene transfer, point mutations, gene duplication, etc.) [12, 14], the number of TFs and target genes in a genome is increased (or decreased) leading to the acquisition (or loss) of nodes and interactions in TRNs. Earlier studies of evolution in TRNs have primarily focused on network growth and the comparison of genomes from free-living organism as the underlying methodology. The central outcomes that have arisen from these studies are that gene duplication and horizontal gene transfer are forces that direct the growth of a network [15, 16], and that TRNs of organisms change more quickly than target genes [9–11, 17]. In this work, we investigated the complementary aspect of TRN evolution under a network reduction perspective by comparing genome sequences of free-living organisms and phylogenetically-related symbionts. In these organisms, the loss of DNA fragments containing sequences coding for TFs, TGs, or both, can lead to evolution of regulatory networks (Fig 5), which have received little attention until now.

We found (after reconstructing the corresponding TRNs and performing correlation analyses) that the conservation of TFs in organisms with reduced genomes is directly related to the kind of regulation they exert, with dual regulators being conserved in a larger proportion than activators and repressors in conditions of genome reduction.

Since most of bacteria exhibiting genome reduction are endosymbionts, we speculate that the lack of pressure to respond to exogenous signals makes the presence of activators dispensable in the first place. This is in agreement with our results, in which activators were the least conserved kind of regulator. As for repressors, we believe the parsimonious conditions inside a bacteriocyte make strong regulation of genes unnecessary. In agreement with our analysis, the loss of repressors followed the loss of activators. Finally, our finding that dual TFs are the most conserved regulatory elements in the transcriptional networks of these organisms might be explained in terms of their simultaneous role as global regulators and NAPs. This suggests that when the major part of the regulatory machinery is lost, gene regulation is essentially carried out at the level of structural organization of the nucleoid, pointing out the role that the conformation of DNA plays in the control of gene expression in its most intrinsic nature.

Bacteria with extreme genome reduction (*e.g.*, with less than 170 Kbp) no longer retain known transcription factors. It remains as a mystery how these organisms regulate the expression of their genes. A possibility is that gene regulation is ultimately exerted by the physiological state of the organism (*i.e.*, availability of ribosomes, RNA polymerases, pool of amino acids and nucleotides, etc.) [58, 59]. The investigation of this subject may have profound implications in the understanding of gene regulation at its most fundamental level and the determination of the smallest functional regulatory systems required in the design of minimal genomes, a topic of high relevance in synthetic biology.

Supporting Information

S1 Table. Transcriptional regulatory networks of the 113 bacteria analyzed here.

(XLSX)

S2 Table. General data of the genomes used in this work.

(XLSX)

S3 Table. Classification of biological function of all the orthologous genes (to *E. coli*) using GO.

(XLSX)

S4 Table. Computed probabilities from hypergeometric distribution for each type of regulators in each bacterial genome.

(XLSX)

S5 Table. Phylogenetically independent contrasts analyses.

(XLSX)

Acknowledgments

The authors thank Laila Partida-Martínez, Luis Delaye, Cei Abreu-Goodger and Edgardo Ugalde for their helpful advice. We also thank the reviewers of the manuscript for their valuable suggestions.

Author Contributions

Conceived and designed the experiments: EGV ISO AMA. Performed the experiments: EGV ISO AMA. Analyzed the data: EGV ISO AMA. Contributed reagents/materials/analysis tools: EGV ISO AMA. Wrote the paper: EGV ISO AMA.

References

1. Gottesman S. Bacterial regulation: global regulatory networks. *Ann Rev Genet.* 1984; 18: 415–41. PMID: [6099091](#)
2. Reznikoff WS, Siegele DA, Cowing DW, Gross CA. The regulation of transcription initiation in bacteria. *Ann Rev Genet.* 1985; 19: 355–387. PMID: [3936407](#)
3. Latchman DS. *Gene Regulation.* 5th ed. Taylor and Francis group: 2005.
4. Junker BH, Schreiber F. *Analysis of biological networks.* 1st ed. Wiley Interscience; 2008.
5. Martinez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol.* 2003; 6: 482–489. PMID: [14572541](#)
6. Albert R. Scale-free networks in cell biology. *J Cell Sci.* 2005; 118: 4947–4957. PMID: [16254242](#)
7. Milo R, Shen-Orr S, Itzkovitz S, Kashtan N, Chklovskii D, Alon U. Network motifs: simple building blocks of complex networks. *Science.* 2002; 298: 824–827. PMID: [12399590](#)
8. Babu MM, Lang B, Aravind L. Methods to reconstruct and compare transcriptional regulatory networks. *Methods Mol Biol.* 2009; 541: 163–180. doi: [10.1007/978-1-59745-243-4_8](#) PMID: [19381525](#)
9. Babu MM, Teichmann SA, Aravind L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol.* 2006; 358: 614–633. PMID: [16530225](#)
10. Lozada-Chávez I, Janga SC, Collado-Vides J. Bacterial regulatory networks are extremely flexible in evolution. *Nucleic Acids Res.* 2006; 34: 3434–3445. PMID: [16840530](#)
11. Price MN, Dehal PS, Arkin AP. Orthologous transcription factors in bacteria have different functions and regulate different genes. *PLOS Comput Biol.* 2007; 3: 1739–1750. PMID: [17845071](#)
12. Wittkopp PJ, Kalay G. Cis-regulatory elements: molecular mechanisms and evolutionary processes underlying divergence. *Nat Rev Genet.* 2011; 13: 59–69. doi: [10.1038/nrg3095](#) PMID: [22143240](#)
13. McAdams HH, Srinivasan B, Arkin AP. The evolution of genetic regulatory systems in bacteria. *Nat Rev Genet.* 2004; 5: 169–178. PMID: [14970819](#)
14. Babu MM. Structure, evolution and dynamics of transcriptional regulatory networks. *Biochem Soc Trans.* 2010; 38: 115–178.
15. Teichmann SA, Babu MM. Gene regulatory network growth by duplication. *Nat Genet.* 2004; 36: 492–496. PMID: [15107850](#)
16. Lagomarsino MC, Jona P, Bassetti B, Isambert H. Hierarchy and feedback in the evolution of the *Escherichia coli* transcription network. *Proc Natl Acad Sci U.S.A.* 2007; 104: 5516–5520. PMID: [17372223](#)
17. Molina N, van Nimwegen E. Universal patterns of purifying selection at noncoding positions in bacteria. *Genome Res.* 2008; 18: 148–160. PMID: [18032729](#)
18. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol.* 2012; 10: 13–26.
19. Moran NA. Accelerated evolution and Muller's ratchet in endosymbiotic bacteria. *Proc Natl Acad Sci U.S.A.* 1996; 93: 2873–2878. PMID: [8610134](#)
20. Delaye L, Gil R, Peretó J, Latorre A, Moya A. Life With a Few Genes: A Survey on Naturally Evolved Reduced Genomes. *Open Evol J.* 2010; 4: 12–22.
21. Felsenstein J. Phylogenies and the comparative method. *Amer Nat.* 1985; 125: 1–15.
22. NCBI ftp server. Available: <http://www.ncbi.nlm.nih.gov/FTP/>.
23. Yu H, Luscombe NM, Lu HX, Zhu X, Xia Y, Han JD, et al. Annotation transfer between genomes: protein-protein interologs and protein-DNA regulogs. *Genome Res.* 2004; 14: 1107–1118. PMID: [15173116](#)
24. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muñiz-Rascado L, García-Sotelo JS, et al. RegulonDB v8.0: omics data sets, evolutionary conservation, regulatory phases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013; 41: D203–D213. doi: [10.1093/nar/gks1201](#) PMID: [23203884](#)
25. Moreno-Hagelsieb G, Latimer K. Choosing BLAST options for better detection of orthologs as reciprocal best hits. *Bioinformatics.* 2008; 24: 319–324. PMID: [18042555](#)

26. Keseler IM, Mackie A, Peralta-Gil M, Santos-Zavaleta A, Gama-Castro S, Bonavides-Martínez C, et al. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res.* 2013; 41: D605–D612. doi: [10.1093/nar/gks1027](https://doi.org/10.1093/nar/gks1027) PMID: [23143106](https://pubmed.ncbi.nlm.nih.gov/23143106/)
27. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet.* 2000; 25: 25–29. PMID: [10802651](https://pubmed.ncbi.nlm.nih.gov/10802651/)
28. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources. *Nat Protoc.* 2009; 4: 44–57. doi: [10.1038/nprot.2008.211](https://doi.org/10.1038/nprot.2008.211) PMID: [19131956](https://pubmed.ncbi.nlm.nih.gov/19131956/)
29. R Development core team, 2008. Available: <http://www.r-project.org/>.
30. Fisher RA. *Statistical methods for research workers.* 5th Edition. Oliver and Boyd, Edinburgh, London; 1934. pp.103–105.
31. Chernick MR. *The essentials of biostatistics for physicians, nurses and clinicians.* New Jersey: John Wiley & Sons Ltd; 2011.
32. Garland T, Harvey PH, Ives AR. Procedures for the analysis of comparative data using phylogenetically independent contrasts. *Syst Biol.* 1992; 41: 18–32.
33. Garland T, Midford PE, Ives AR. An introduction to phylogenetically based statistical methods, with a new method for confidence intervals on ancestral values. *Amer Zool.* 1999; 39: 374–388.
34. Garland T, Ives AR. Using the past to predict the present: confidence intervals for regression equations in phylogenetic comparative methods. *Amer Zool.* 2000; 155: 346–364.
35. Maddison WP, Maddison DR. Mesquite: a modular system for evolutionary analysis (Version 3.02). 2015. Available: <http://mesquiteproject.org>.
36. Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, Höhna S, et al. MrBayes 3.2: Efficient Bayesian phylogenetic inference and model choice across a large model space. *Sys Biol.* 2012; 61: 539–542.
37. Holder M, Lewis PO. Phylogeny estimation: traditional and bayesian approaches. *Nat Rev Genet.* 2003; 4: 275–284. PMID: [12671658](https://pubmed.ncbi.nlm.nih.gov/12671658/)
38. Lima-Mendez G, van Helden J. The powerful law of the power law and other myths in network biology. *Mol BioSyst.* 2009; 5: 1482–1493. doi: [10.1039/b908681a](https://doi.org/10.1039/b908681a) PMID: [20023717](https://pubmed.ncbi.nlm.nih.gov/20023717/)
39. Galán-Vásquez E, Luna B, Martínez-Antonio A. The regulatory network of *Pseudomonas aeruginosa*. *Microb Inform Exp.* 2011; 1: 1–11.
40. Alexopoulos EC. Introduction to multivariate regression analysis. *Hippokratia*, 2010; 14: 23–28. PMID: [21487487](https://pubmed.ncbi.nlm.nih.gov/21487487/)
41. Davis JH. *Statistics for compensation: A practical guide to compensation analysis.* 1st ed. John Wiley & Sons, Inc; 2011.
42. van Nimwegen E. Scaling law in the functional content of genomes. *Trends Genet.* 2003; 19: 479–484. PMID: [12957540](https://pubmed.ncbi.nlm.nih.gov/12957540/)
43. Pérez-Rueda E, Collado-Vides J. The repertoire of DNA-binding transcriptional regulators in *Escherichia coli* K-12. *Nucleic Acids Res.* 2000; 28: 1838–1847. PMID: [10734204](https://pubmed.ncbi.nlm.nih.gov/10734204/)
44. Balderas-Martínez YI, Savageau M, Salgado H, Pérez-Rueda E, Morett E, Collado-Vides J. Transcription factors in *Escherichia coli* prefer the *Holo* conformation. *Plos One.* 2013; 8: 1–9.
45. Martínez-Antonio A, Janga SC, Salgado H, Collado-Vides J. Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. *Trends Microbiol.* 2006; 14: 22–27. PMID: [16311037](https://pubmed.ncbi.nlm.nih.gov/16311037/)
46. Struhl K. Fundamentally different logic of gene regulation in Eukaryotes and Prokaryotes. *Cell.* 1999; 98: 1–4. PMID: [10412974](https://pubmed.ncbi.nlm.nih.gov/10412974/)
47. Stracy M, Lesterlin C, Garza de Leon F, Uphoff S, Zawadzki P, Kapanidis AN. Live-cell superresolution microscopy reveals the organization of RNA polymerase in the bacterial nucleoid. *Proc Natl Acad Sci U.S.A.* 2015; 32: E4390–E4399.
48. Gerosa L, Kochanowski K, Heinemann M, Sauer U. Dissecting specific and global transcriptional regulation of bacterial gene expression. *Mol Syst Biol.* 2013; 9: 1–11.
49. Alon U. Network motifs: theory and experimental approaches. *Nat Rev Genet.* 2007; 8: 450–461. PMID: [17510665](https://pubmed.ncbi.nlm.nih.gov/17510665/)
50. Madar D, Dekel E, Bren A, Alon U. Negative auto-regulation increases the input dynamic-range of the arabinose system of *Escherichia coli*. *BMC Syst Biol.* 2011; 5: 1–9.
51. Klumpp S, Zhang Z, Hwa T. Growth rate-dependent global effects on gene expression in bacteria. *Cell.* 2009; 139: 1366–1375. doi: [10.1016/j.cell.2009.12.001](https://doi.org/10.1016/j.cell.2009.12.001) PMID: [20064380](https://pubmed.ncbi.nlm.nih.gov/20064380/)

52. Dillon SC, Dorman CJ. Bacterial nucleoid-associated proteins, nucleoid structure and gene expression. *Nat Rev Microbiol.* 2010; 8: 185–195. doi: [10.1038/nrmicro2261](https://doi.org/10.1038/nrmicro2261) PMID: [20140026](https://pubmed.ncbi.nlm.nih.gov/20140026/)
53. Huynen MA, Spronk CA, Gabaldon T, Snel B. Combining data from genomes, Y2H and 3D structure indicates that BoIA is a reductase interacting with a glutaredoxin. *FEBS Lett.* 2005; 579: 591–596. PMID: [15670813](https://pubmed.ncbi.nlm.nih.gov/15670813/)
54. Savageau MA. Design of molecular control mechanisms and the demand for gene expression. *Proc Natl Acad Sci U.S.A.* 1977; 74: 5647–5651. PMID: [271992](https://pubmed.ncbi.nlm.nih.gov/271992/)
55. Savageau MA. Demand theory of gene regulation. II. Quantitative application to the lactose and maltose operons of *Escherichia coli*. *Genetics*, 1998; 149: 1677–1691. PMID: [9691028](https://pubmed.ncbi.nlm.nih.gov/9691028/)
56. Moran NA, Bennett GM. The tiniest tiny genomes. *Annu Rev Microbiol.* 2014; 68: 195–215. doi: [10.1146/annurev-micro-091213-112901](https://doi.org/10.1146/annurev-micro-091213-112901) PMID: [24995872](https://pubmed.ncbi.nlm.nih.gov/24995872/)
57. Briza L, Calevro F, Charles H. Genomic analysis of the regulatory elements and links with intrinsic DNA structural properties in the shrunken genome of *Buchnera*. *BMC Genomics.* 2013; 14: 1–15.
58. Scott M, Hwa T. Bacterial growth laws and their applications. *Curr Opin Biotechnol.* 2011; 22: 559–565. doi: [10.1016/j.copbio.2011.04.014](https://doi.org/10.1016/j.copbio.2011.04.014) PMID: [21592775](https://pubmed.ncbi.nlm.nih.gov/21592775/)
59. Berthoumieux S, de Jong H, Baptist G, Pinel C, Ranquet C, Ropers D, et al. Shared control of gene expression in bacteria by transcription factors and global physiology of the cell. *Mol Syst Biol.* 2013; 9 (634): 1–11.