



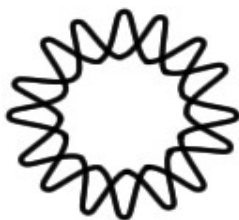
**UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO
INSTITUTO DE BIOTECNOLOGÍA**

Programa de Doctorado en Ciencias Bioquímicas

**Divergencia de los sistemas de regulación en secuencias
parálogas en bacterias: una perspectiva genómica**

**TESIS
QUE PARA OBTENER EL GRADO DE
DOCTOR EN CIENCIAS BIOQUÍMICAS
PRESENTA**

M. C. Mario Alberto Martínez Núñez



Cuernavaca, Morelos

Julio del 2011



Universidad Nacional
Autónoma de México

Dirección General de Bibliotecas de la UNAM

Biblioteca Central



UNAM – Dirección General de Bibliotecas
Tesis Digitales
Restricciones de uso

DERECHOS RESERVADOS ©
PROHIBIDA SU REPRODUCCIÓN TOTAL O PARCIAL

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

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

Índice

1. Introducción	1
1.1. Parálogos y función	3
1.2. Parálogos y regulación	5
1.2.1. Regulación transcripcional en procariotes	5
1.2.2. Red de regulación transcripcional, TRN	7
1.2.3. TRN y duplicación	8
2. Hipótesis	11
3. Objetivos	11
4. Metodología	12
4.1. Identificación de genes parálogos	12
4.2. Identificación de elementos de regulación basados en TFs	12
4.3. Identificación de elementos de regulación en genes parálogos basados en moéculas de RNA	12
4.4. Documentación de funciones de genes parálogos	13
5. Resultados	15
5.1. Divergencia en el tipo de regulación (activación/represión) del TF sobre sus genes parálogos blanco	17
5.2. Divergencia en sistemas de regulación distintos a TFs: factores σ	21
5.3. Divergencia en sistemas de regulación distintos a TFs: sistemas de dos componentes	23
5.4. Divergencia en sistemas de regulación distintos a TFs: RNAs (<i>cis</i> y <i>trans</i>) e interacciones proteína-RNA	25
5.5. Divergencia en sistemas de regulación distintos a TFs: super-enrollamiento	27
5.6. Distintos tipos de elementos de regulación modulan a los parálogos	28
6. Conclusiones	30

7. Perspectivas	31
8. Suplementos	44
8.1. Figura	44
8.2. Artículo	48

Índice de figuras

1.	Regulación transcripcional	6
2.	Modelo de Teichmann-Babu	9
3.	Metodología	14
4.	Regulación de parálogos por TFs	16
5.	Regulación de parálogos por tipo de activación de TFs	19
6.	Parálogos y KEGG	22
7.	Regulación de parálogos por factores σ	23
8.	Regulación de parálogos por sistemas de dos componentes	24
9.	Regulación de parálogos por RNA en <i>cis</i> y <i>trans</i>	26
10.	Regulación de parálogos por RNA-proteínas	27
11.	Regulación de porinas parálogas	29
12.	Elementos por grupo en <i>E. coli</i>	44
13.	Elementos por grupo en <i>B. subtilis</i>	46
14.	Curva de distribución normal	47

Índice de cuadros

1. Valores de interacciones observadas y simuladas	18
2. Frecuencias de parálogos por grupos	45

Abstract

Extensive genomic studies on gene duplication in model organisms such as Escherichia coli and Saccharomyces cerevisiae have recently been undertaken. In these models, it is commonly considered that a duplication event may include a transcription factor (TF), a target gene, or both. Following a gene duplication episode, varying scenarios have been postulated to describe the evolution of the regulatory network. However, in most of these, the TFs have emerged as the most important and in some cases the only factor shaping the regulatory network as the organism responds to a natural selection process, in order to fulfil its metabolic needs. Recent findings concerning the regulatory role played by elements other than TFs have indicated the need to reassess these early models. Thus, we performed an exhaustive review of paralogous gene regulation in E. coli and Bacillus subtilis based on published information, available in the NCBI PubMed database and in well-established regulatory databases. Our survey reinforces the notion that despite TFs being the most prominent components shaping the regulatory networks, other elements are also important. These include small RNAs, riboswitches, RNA-binding proteins, sigma factors, protein-protein interactions and DNA supercoiling, which modulate the expression of genes involved in particular metabolic processes or induce a more complex response in terms of the regulatory networks of paralogous genes in an integrated interplay with TFs.

Resumen

*Estudios extensos se han llevado a cabo sobre la duplicación genética en organismos modelos como *Escherichia coli* y *Saccharomyces cerevisiae*. En estos modelos, se considera comúnmente que un evento de duplicación puede incluir un factor de transcripción (TF), un gene blanco, o ambos. Después de un episodio de duplicación, se han postulado distintos escenarios para describir la evolución de la red de regulación. Sin embargo, en la mayoría de ellos, los TFs se han convertido en el elemento más importante y en algunos casos el único elemento que configura la red de regulación como respuesta del organismo a un proceso de selección natural, con el fin de satisfacer sus necesidades metabólicas. Hallazgos recientes sobre la función reguladora de otros elementos distintos a los TFs, han mostrado la necesidad de adecuar los primeros modelos. Por lo tanto se realizó una revisión exhaustiva de la regulación de genes parálogos en *E. coli* y *Bacillus subtilis* basada en información disponible en la base de datos PubMed del NCBI y en bases de datos sobre regulación. Nuestro estudio refuerza la idea de que a pesar de que los TFs son el componente más importante que configuran la red de regulación, otros elementos son también importantes. Estos incluyen a RNAs pequeños, riboswitches, proteínas de unión a RNA, factores sigma, interacciones proteína-proteína y super-enrollamiento del DNA, los cuales que modulan la expresión de genes implicados en procesos metabólicos particulares o induciendo una respuesta más compleja en términos de las redes de regulación de genes parálogos en una interacción integrada con los TFs*

1. Introducción

El conjunto de genes de un organismo evoluciona a través del surgimiento de funciones o de mecanismos de regulación nuevos, que permiten una mejor adecuación de la célula a las fluctuaciones del medio interno y externo. Eventos como las mutaciones en los nucleótidos de las secuencias de los genes, pueden originar una nueva función en sus productos polipeptídicos, por ejemplo, al modificar los sitios de reconocimiento o de unión de la proteína hacia su sustrato o a otra proteína. A través de la transferencia horizontal de genes (HGT) es también posible que el repertorio de genes de los organismos aumente, adquiriéndose genes con funciones nuevas provenientes de otros organismos. La transferencia de genes ocurre con mayor frecuencia en organismos procariotes y filogenéticamente cercanos, disminuyendo la probabilidad de integración del gene foráneo en el genoma del hospedero mientras más alejados se encuentren [84], [4]. La duplicación genética es otro mecanismo para la generación de nuevos genes, a través de la cual se pueden dar innovaciones tanto en el aspecto funcional de las proteínas como en la regulación de los mismos. Gracias a la secuenciación y análisis de genomas de los tres dominios de la vida (eubacteria, arqueobacteria y eucariota), se ha visto la prevalencia e importancia de los procesos de duplicación genética. A los genes surgidos por procesos de duplicación se les llama parálogos. El porcentaje de genes parálogos presentes en los organismos va desde un 17% para bacterias como *Helicobacter pylori*, hasta un 65% en organismos eucariotes como la planta *Arabidopsis thaliana* [99].

Después de la duplicación, el destino más frecuente para los parálogos es que una de las copias sufra eventos de mutación, los cuales se acumulan y convierten a la copia en un pseudo-gene. La acumulación de mutaciones en los genes interrumpen su estructura y función, lo cual lleva a los duplicados a una posterior eliminación del genoma. Los pseudo-genes también pueden ocurrir en genes parálogos que han sido seleccionados y fijados en la población, pero que pasan por periodos actuales de relajación en las restricciones funcionales [99], [79]. Distintos modelos han sido postulados para dar una explicación del destino de las copias que permanecen en el genoma, como por ejemplo el de la neo-funcionalización, la sub-funcionalización, diversificación de la selección, entre otros. Dichos modelos pueden ser agrupados en dos, dependiendo el tipo de selección invocada durante la etapa de fijación de los duplicados.

Los modelos como la neo-funcionalización y la sub-funcionalización invocan una selección neutral en las copias después de duplicadas, dicha selección neutral permite a los parálogos permanecer en el genoma para posteriormente derivar una función nueva (neo-funcionalización), o bien, la función original del gene parental ser compartida por las copias (sub-funcionalización). Los modelos agrupados en el segundo grupo como el de la diversificación de la selección, invocan a la selección positiva en la etapa inmediata a la duplicación. Dicha selección positiva actúa sobre los cambios en las secuencias de las copias, hasta que deriva en la generación de una nueva función o una especialización mayor del parólogo [45], [33], [21]. En los modelos anteriormente descritos, la función del gene original es subdividida entre las copias, o bien, se genera una nueva función. Sin embargo, los parálogos pueden permanecer en el genoma conservando la misma función que el gene original. La retención de la función del gene parental por parte de la copia, puede ser llevada a cabo por un mecanismo de selección purificadora o bien por un mecanismo de recombinación entre las copias llamado evolución concertada, la cual permite que las secuencias permanezcan homogéneas. La selección purificadora es invocada en el caso en el que la célula requiere una demanda alta de los productos codificados por los parálogos, a lo que se conoce con el nombre de dosificación genética. La selección negativa ejercida sobre los parálogos, asegura que las secuencias de los genes permanezcan sin cambios y conservando la misma función, ante la demanda de poder contar con un número elevado de proteína de cierto gene. Entre los ejemplos de genes que presentan el fenómeno de dosificación genética más conocidos se encuentran los genes ribosomales, cuyo producto peptídico es necesario en altas concentraciones para poder llevar a cabo el proceso de traducción de todos los genes del organismo. Un segundo ejemplo de genes que presentan el fenómeno de dosificación genética son diversos genes que codifican para proteínas transportadoras, como es el caso del gene *cadA*, que es duplicado para que la bacteria pueda adaptarse a ambientes de aguas contaminadas con altas concentraciones de cadmio [12], [50]. En un estudio de la familia del gene de la ubiquitina de 28 especies de plantas, hongos y animales, Nei *et al.* [65] encontraron que la selección purificadora tiene un papel más preponderante en mantener la misma función entre los parálogos que la evolución concertada. Así mismo, Sanchez-Perez *et al.* [75] han descrito una clase de genes parálogos, llamados eco-parálogos, que mantienen la misma función como respuesta ante diferentes condiciones ambientales, y no por un fenómeno de dosificación genética. Los eco-parálogos

llevan a cabo la misma función celular pero bajo diferentes condiciones ecológicas. Un ejemplo lo representan las proteínas parálogas glutamato deshidrogenasa de la bacteria halófila *Salinibacter ruber*, las cuales tienen dos distintos intervalos de actividad en sal: una de las copias muestra su máxima actividad enzimática en ausencia de sal, mientras que la otra copia requiere de altas concentraciones salinas para ser activa [75]. De esta forma, la redundancia genética en la célula puede responder a la necesidad de contar con altas concentraciones de proteína de cierto gene, de adaptación a variaciones ambientales (eco-parálogos), o a la necesidad de innovación funcional (divergencia funcional), haciendo más complejo el genoma de los organismos.

1.1. Parálogos y función

Los genes y sus proteínas pueden ser agrupadas en familias funcionales, donde los miembros que pertenecen a una misma familia, llevan a cabo funciones similares. La presencia de genes parálogos en las familias incrementa el número de miembros en ellas, y frecuentemente en genomas procariotes hay de dos a 30 duplicados por familia [74]. El estudio de 127 genomas de eubacterias en el 2004 por Pushker *et al.* [ídem], y el estudio de 200 genomas procariotes realizado por Bratlie *et al.* en el 2010, muestran una correlación entre el aumento de genes parálogos en las familias y el aumento del tamaño del genoma en los organismos [15]. Estudios realizados por Conant y Wagner [20] con genes parálogos de *Saccharomyces cerevisiae* y *E. coli*, mostraron que la duplicación de genes se presentaba con mayor frecuencia en ciertas familias funcionales. Analizando la distribución de los parálogos en diferentes categorías funcionales de acuerdo a la base de datos Gene Ontology (GO), encontraron una sobre-representación de duplicados en las categorías del metabolismo de energía y transportadores, en ambos genomas. La sobre-representación de parálogos en ciertas familias funcionales, también ha sido reportada por Pushker *et al.* [74] para *E. coli*, coincidiendo con Conant y Wagner (2002) en que las familias de transportadores, en especial de metabolitos, contienen un mayor número de duplicados. En estudios similares a los mencionados anteriormente, Bratlie *et al.* [15] utilizando la clasificación funcional de la base de datos de familias de proteínas ortólogas COG, encontró que los parálogos de 200 genomas procariotes, se distribuyen de manera preferencial en las categorías funcionales de

producción y conversión de energía (categoría COG C), motilidad celular (N), transporte de iones inorgánicos (P), transducción de señales (T) y mecanismos de defensa (V). Gevers *et al.* utilizando también la clasificación de la base de datos COGs, encontraron una sobre-representación de genes parálogos en las categorías funcionales de metabolismo de aminoácidos, transcripción (K), mecanismos de defensa (V) y conversión de energía (C) [29]. La sobre-representación de los parálogos en ciertas categorías funcionales, puede reflejar una impronta histórica de los ambientes diversos con los que se han encontrado estos organismos durante su historia evolutiva. Muchas bacterias habitan ambientes diversos, los cuales pueden ser estresantes para el organismo, por ejemplo, al estar contaminados, al presentar pocos nutrientes o con altas o bajas temperaturas. La duplicación de genes que han ocurrido con mayor frecuencia en ciertas categorías funcionales, como por ejemplo los transportadores de iones inorgánicos, permiten a la célula la toma de metales que se encuentran en baja disponibilidad en el medio, favoreciendo una mejor adecuación de la célula al ambiente [20], [15].

Los parálogos dentro de las familias pueden divergir en la función, fenómeno conocido como divergencia funcional. La divergencia de la función en los parálogos cubre un espectro variado, donde los cambios pueden ser en el reconocimiento del sustrato o sitio de unión con otras proteínas, en la reacción bioquímica o en el producto generado por las copias [78]. La familia de las proteínas cinasas representan un ejemplo de variación en el reconocimiento del sustrato, pero conservando la misma reacción bioquímica [63]. En *E. coli* los genes parálogos *trpE* y *pabB* codifican para las enzimas antranilato sintasa y para-aminobenzoato sintasa, respectivamente. Estas dos enzimas reconocen al corismato como sustrato, pero los productos de sus reacciones son distintos y participan en vías bioquímicas diferentes [44], [40]. En cambio las enzimas parálogas PurK y PurT llevan a cabo reacciones distintas, la primera es una carboxilasa, mientras que la segunda es una formil-transferasa [63]. El surgimiento de funciones nuevas en los duplicados, se encuentra asociado no sólo a los cambios en la región codificante de los genes, si no que también pueden ocurrir cambios en las regiones de regulación de los parálogos. La regulación de los genes parálogos en las bacterias se encuentra sujeta también a la divergencia.

1.2. Parálogos y regulación

La adecuación de los parálogos no ocurre sólo a nivel de la función, si no que también suceden cambios en la regulación de los mismos, aumentando la complejidad del organismo. La evolución de los parálogos dentro de los genomas, incluye a los mecanismos de regulación que modulan la expresión de los genes. Dicha modulación diferencial permite una respuesta adecuada de la célula a las fluctuaciones del ambiente externo y de los flujos metabólico internos.

1.2.1. Regulación transcripcional en procariotes

Aunque la regulación puede llevarse a cabo a diferentes niveles (transcripcional, traduccional o post-traduccional), en bacterias la regulación se efectúa preferentemente al nivel del inicio de transcripción. La transcripción en las células procariotes inicia cuando los factores σ interactúan con la RNA polimerasa (RNAP) para el reconocimiento de sus correspondientes secuencias promotoras. El inicio de la transcripción en la fase de crecimiento exponencial de la bacteria *E. coli*, se lleva a cabo principalmente por el factor σ^{70} , llamado también esencial (“housekeeping”). Para el caso de la bacteria *Bacillus subtilis* el factor σ equivalente se nombra como σ^A . Los otros factores σ se encuentran asociados a grupos de genes requeridos durante condiciones específicas, como por ejemplo la respuesta a estrés por calor que es mediada en *E. coli* por el factor σ^{32} [41], o en el caso de la transición entre fases de crecimiento, cuando en fase estacionaria el principal factor σ que interactúa con la RNAP es σ^{38} [42]. De esta manera a través de los factores σ se tiene la oportunidad de mantener una expresión genética basal de los genes esenciales, así como el redireccionamiento de la maquinaria transcripcional hacia la expresión de genes de respuesta ante determinado estímulo o fase de crecimiento celular [95], [46], [60]. Otro elemento de la regulación de la maquinaria transcripcional son las proteínas de unión al DNA, llamadas factores transcripcionales (TFs), los cuales pueden regular de manera positiva o negativa la transcripción de sus correspondientes genes blanco (Figura 1.A). Los TFs afectan la expresión de los genes bloqueando o favoreciendo el reconocimiento de la RNA polimerasa (RNAP) al promotor. En términos generales, los TFs están compuestos de dos dominios, uno de unión al DNA y otro de regulación. En el 80 % de los TFs conocidos, el plegamiento predominante que se presenta en

el dominio de unión al DNA es el de hélice-giro-hélice (HTH), mientras que en el porcentaje restante se han identificado en menor proporción otros plegamientos como los dedos de zinc, las hojas- β y el plegamiento de hélice-asa-hélice, el cual se encuentra restringido a las gamma-proteobacterias [70], [71]. El dominio de regulación de los TFs puede interactuar con otras proteínas o metabolitos ante determinadas señales internas o cambios ambientales [16], [48], [56], consiguiendo de esta forma sensor los estados celulares de la bacteria.

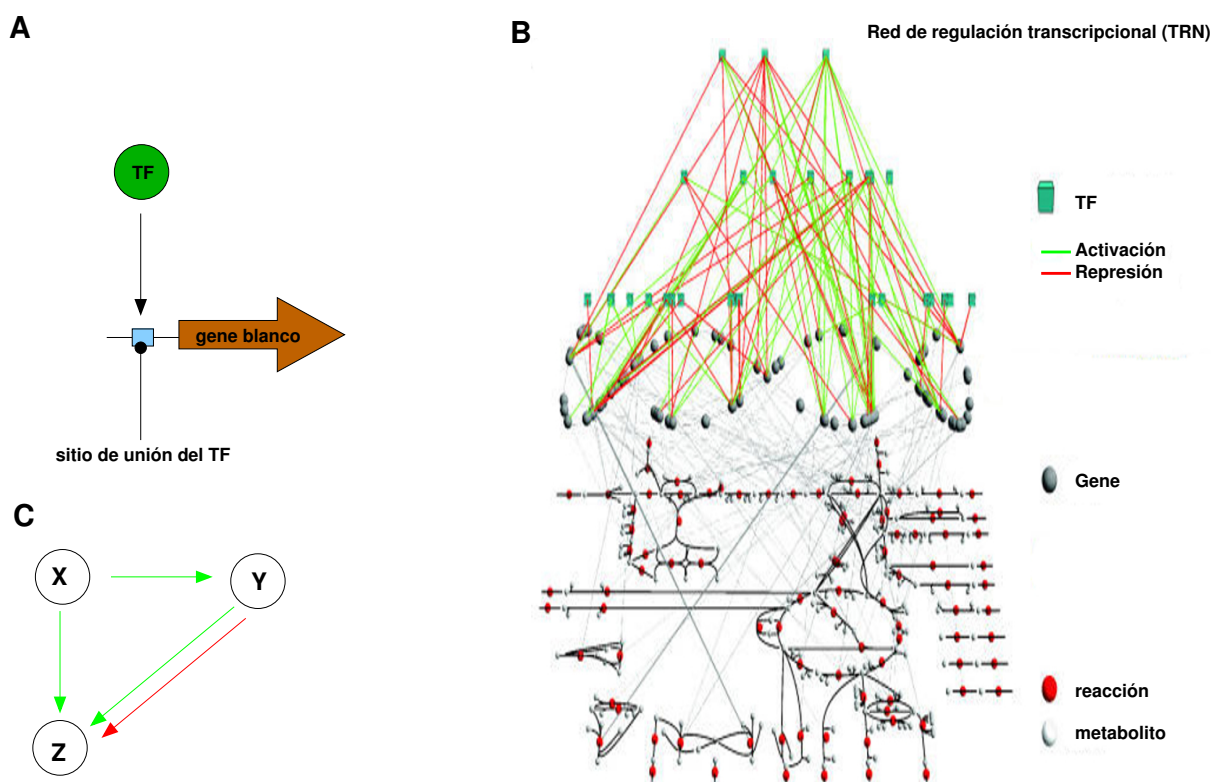


Figura 1: Elementos de la regulación transcripcional y su representación como una red. **(A)** Elementos que componen la regulación transcripcional: TF, sitio de unión del TF y el gene blanco. **(B)** Representación de las regulaciones transcripcionales como una red, llamada red de regulación transcripcional (TRN). Figura modificada de [83]. **(C)** Motivo de regulación “feed-forward”, en el cual la acción de X sobre Z puede ser de manera directa, o indirecta a través de Y. TF: factor transcripcional.

1.2.2. Red de regulación transcripcional, TRN

Los elementos de regulación pueden establecer entre sí relaciones de distintos tipos, por ejemplo el que un gene blanco sea regulado por más de un TF, el que un TF regule a más de un blanco, o que un TF se auto-regule. Los TFs, sus genes blanco y las interacciones transcripcionales entre ellos han sido conceptualizados como una red, por lo que se le ha nombrado red de regulación transcripcional (TRN) [101], [88] (Fig. 1.B). En dicha red los TFs y sus blancos representan los nodos, mientras que las interacciones entre ellos representan los enlaces o aristas [87], [5]. Estudios sobre la regulación transcripcional, desde la perspectiva de redes, han mostrado ciertas características en común, como por ejemplo que su arquitectura es libre de escala (muchos nodos con pocas conexiones y pocos nodos con muchas conexiones), revelando la existencia de reguladores globales (hubs), como la proteína de represión catabólica por fuentes de carbono CRP, la cual regula a 359 genes blanco [10], [36]. También es frecuente encontrar en la estructura de la TRN la presencia de circuitos de regulación con interconexiones recurrentes, llamadas motivos. El motivo de regulación más frecuente es el “feed-forward”, en el cual hay 3 elementos involucrados: X, Y y Z (Fig. 1.C). El diseño del motivo “feed-forward” permite dos tipos de dinámica, una cuando X, Y son ambos activadores de Z, con lo que el circuito actúa como un filtro ante señales transitorias. Si X regula a Z de manera positiva y Y regula a Z negativamente, entonces el circuito puede mantener la expresión de Z dentro de un intervalo deseado [80], [24]. Las TRN han sido estudiadas también desde una perspectiva dinámica, analizando su comportamiento ante cambios de condiciones específicas. En el caso de *S. cerevisiae* utilizando datos de expresión genética para cinco diferentes condiciones, se ha visto que la TRN presenta cambios en su arquitectura, como que los TFs adquieran o pierdan conexiones, con lo cual se modifica la topología de la red. Una consecuencia de la modificación de la arquitectura es que algunos TFs que son “hubs” en una condición, no lo son en otra, mientras que algunos TFs “hubs” permanecen como tales en cualquier condición [53], [9]. En el caso de *E. coli*, se han identificado unidades de procesamiento de señales ambientales, llamadas orígenes (“origin”), las cuales son sub-redes dentro de la TRN. Los orígenes se encuentran agrupados alrededor de un TF, el cual se encuentra especializado para la detección de cierta señal. La respuesta transcripcional de la célula depende del tipo de per-

turbación sensada, lo cual desencadena la activación de diferentes sub-redes (origones) bajo distintas condiciones ambientales. En *Mycobacterium tuberculosis* se han identificado distintos grupos de sub-redes dentro de la TRN, las cuales se activan en diferentes etapas de adaptación del patógeno en su hospedero. Con lo cual también en genomas procariotes la TRN presenta una dinámica que va modificando la estructura de la red, dependiendo de las condiciones en que se encuentre la célula [7], [8], [4].

1.2.3. TRN y duplicación

Las relaciones de regulación que se observan en la TRN son un producto de un proceso evolutivo, en el cual distintos mecanismos como las mutaciones, la transferencia horizontal y las duplicaciones, moldean la estructura y generan la diversidad de la red. Los eventos de duplicación de elementos de regulación como los TFs, tiene como consecuencia la generación de una red con interacciones complejas. El aumento en el número de TFs en genomas grandes con respecto a los de menor tamaño, no implica un aumento en la diversidad de las familias de TFs. Lo que se ha observado es un aumento en el número de miembros por familia, muchos de los cuales han surgido por duplicación [87], [74]. La duplicación de TFs permiten una adaptación más versátil de la divergencia funcional adquirida de la duplicación de genes estructurales. Estudios realizados con datos de microarreglos para evaluar la expresión de genes parálogos en la levadura *S. cerevisiae*, muestran que una alta proporción de genes duplicados tienen una rápida divergencia en su expresión. 40 % de los parálogos evolutivamente cercanos muestran una ausencia de co-expresión, mientras que el 80 % de las copias con tasas de sustituciones sinónimas (K_S) mayores a 1.5 no son co-expresados. Estos resultados son consistentes con la idea de que la divergencia en la co-expresión aumenta con el tiempo evolutivo [35]. Después de ocurrida la duplicación, la divergencia de la expresión y la regulación de las copias se incrementa, habiendo una evolución asimétrica en los parálogos, donde la expresión de una copia evoluciona rápidamente, mientras que la otra mantiene su perfil de expresión ancestral [34]. Los cambios en los patrones de expresión están asociados a los cambios en los mecanismos de regulación que modulan la actividad de las copias. En estudios hechos en las TRNs utilizando los genomas de *S. cerevisiae*, *Drosophila melanogaster* y *H. pylori*, han reportado una disminución de los sitios de unión de los TFs que son compartidos entre

genes parálogos, observándose una pérdida de 3% de TFs comunes por cada 1% de divergencia en la identidad de la secuencia de aminoácidos de las copias [68], [58]. Los TFs desempeñan un papel importante en las redes de regulación de los genes parálogos, siendo los elementos de regulación más destacados y con frecuencia los únicos tomados en cuenta en los estudios realizados con las TRNs. A partir del estudio de las TRNs de *S. cerevisiae* y *E. coli*, Teichmann y Babu [87] han propuesto un modelo evolutivo ampliamente aceptado acerca de la evolución de la regulación en genes duplicados. En dicho modelo, la pérdida o ganancia de interacciones dentro de la red de regulación puede ocurrir a través de tres escenarios: el primero es la duplicación de un TF, el segundo la duplicación de un gene blanco y el tercero la duplicación tanto del TF como el de su gene blanco (Fig. 2).

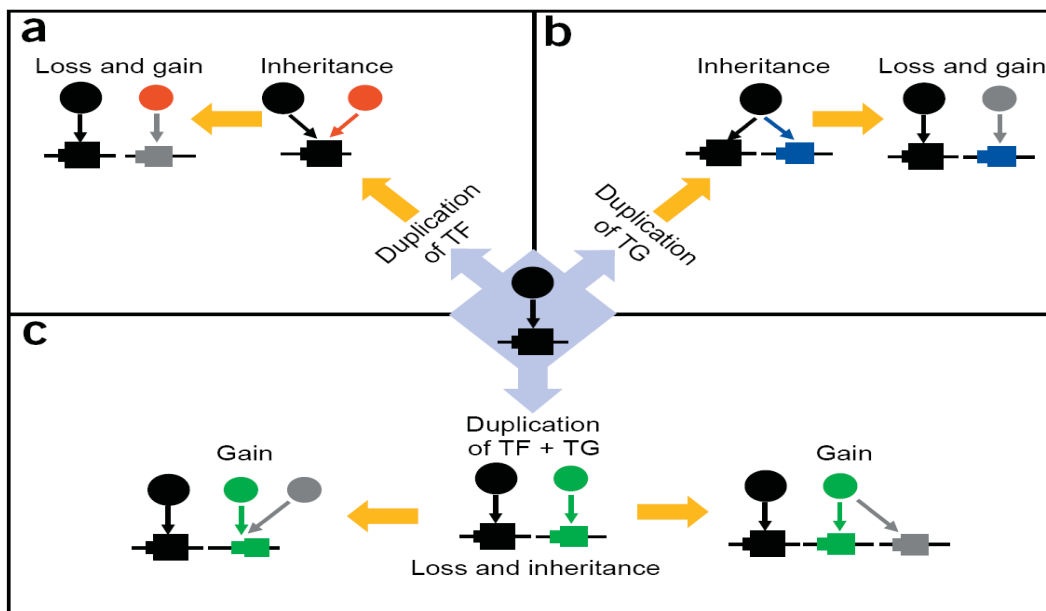


Figura 2: Modelo de crecimiento por duplicación y sus consecuencias para la evolución de la red. La unidad básica de regulación se muestra en el centro: el factor de transcripción (TF), el gene blanco (TG) y el sitio de unión del TF. Los paneles describen los posibles eventos de duplicación de la unidad básica de regulación y su subsecuente divergencia resultando en una nueva interacción de regulación. (a) Duplicación del TF. (b) Duplicación del gene blanco. (c) Duplicación del TF y del gene blanco. Figura tomada de [87].

Si bien los TFs son los elementos de regulación más abundantes en la célula, no son los únicos en participar en la dinámica y la conformación de las redes de regulación. La divergencia en los mecanismos de regulación en los parálogos no queda circunscrita únicamente a TFs y sus interacciones proteína-

DNA, pudiendo abarcar otros sistemas de regulación tales como los mediados por RNA (en *cis* y *trans*), RNA-proteínas o interacciones proteína-proteína. El modelo de evolución de la regulación de genes parálogos de Teichman y Babu (2004) puede ser generalizado a otros sistemas de regulación donde los elementos reguladores sean distintos a los TFs, así como a distintos niveles de regulación aparte del transcripcional, obteniendo un panorama más completo y un mejor conocimiento de la evolución de las redes de regulación de genes parálogos. El presente trabajo examinó de manera sistemática la divergencia en la regulación de las secuencias parálogas de los genomas de *E. coli* y *B. subtilis* para identificar mecanismos de regulación adicionales a los operados por TFs, cuyo aporte a la TRN le proporcionen mayor versatilidad.

2. Hipótesis

- La divergencia en la regulación de genes parálogos se presenta también sistemas de regulación tales como los mediados por moléculas de RNAs (*cis* y *trans*), por interacciones del tipo proteína-RNA o proteína-proteína, aparte de los modulados por TFs.

3. Objetivos

- Identificación de genes parálogos en los genomas de *E. coli* y *B. subtilis*
- Identificación de mecanismos de regulación por TFs en las secuencias parálogas
- Identificación de mecanismos de regulación en genes parálogos mediados por otros elementos distintos a los TFs

4. Metodología

4.1. Identificación de genes parálogos

Para el estudio de la divergencia en la regulación de genes parálogos se utilizaron los genomas de *E. coli* y *B. subtilis*. La identificación de proteínas parálogas se llevó a cabo bajo los criterios descritos en otros trabajos, donde los alineamientos pareados entre las secuencias de aminoácidos presentan al menos un 30% de identidad, sobre al menos un 60% de las longitudes de las secuencias comparadas. Así como también el tener un valor de corte de E-val de 10^{-5} [15], [59], [74]. Por medio de una comparación todos-contra-todos (“all-against-all”) entre las secuencias de proteínas de un mismo genoma y usando el programa BLASTP [3], se identificaron a 477 grupos de genes parálogos en el genoma de *E. coli*, en tanto que en el genoma de *B. subtilis* se identificaron a 483 grupos de genes parálogos.

4.2. Identificación de elementos de regulación basados en TFs

Para indagar sobre los mecanismos de regulación presentes en los grupos antes definidos, se consultaron tanto base de datos de regulación de organismos modelo, como la literatura disponible sobre sistemas de regulación. Para el análisis de los mecanismos de regulación mediados por factores de transcripción (TF) en el genoma de *B. subtilis* se utilizó la base de datos DBTBS [81] en su versión 5, la cual compila información de datos experimentales. Mientras que el análisis de los TFs en *E. coli* se realizó utilizando la base de datos RegulonDB [26] en su versión 6.3, la cual es parcialmente curada manualmente.

4.3. Identificación de elementos de regulación en genes parálogos basados en moéculas de RNA

La búsqueda de elementos de regulación operados por sistemas de RNA, se llevó a cabo utilizando modelos de covarianza (CM) depositados en la base de datos Rfam (version 9.1) [32], la cual contiene modelos de 603 familias de RNAs estructurales, con propiedades de regulación. Los modelos de co-

varianza modelan de manera simultánea la relación que guarda la secuencia del RNA con respecto a su estructura, caracterizando de esta forma a una familia de RNAs, cuyo plegamiento sea importante para la función. Utilizando el programa Infernal [23], se buscaron en las regiones 5' río arriba de los genes parálogos identificados, secuencias de regulación que se adecuaron a los CMs utilizados, cumpliendo con los valores de corte impuestos para cada familia. Otros sistemas de regulación como sRNA o sistemas de dos componentes, fueron documentados realizando una búsqueda y compilación de información en la literatura científica utilizando la base de datos de referencias PubMed [76].

4.4. Documentación de funciones de genes parálogos

La documentación sobre las funciones de los genes y su participación en rutas metabólicas se realizó con la base de datos KEGG [49], la cual contiene información sobre las rutas bioquímicas de los distintos organismos, así como los genes identificados en ellas.

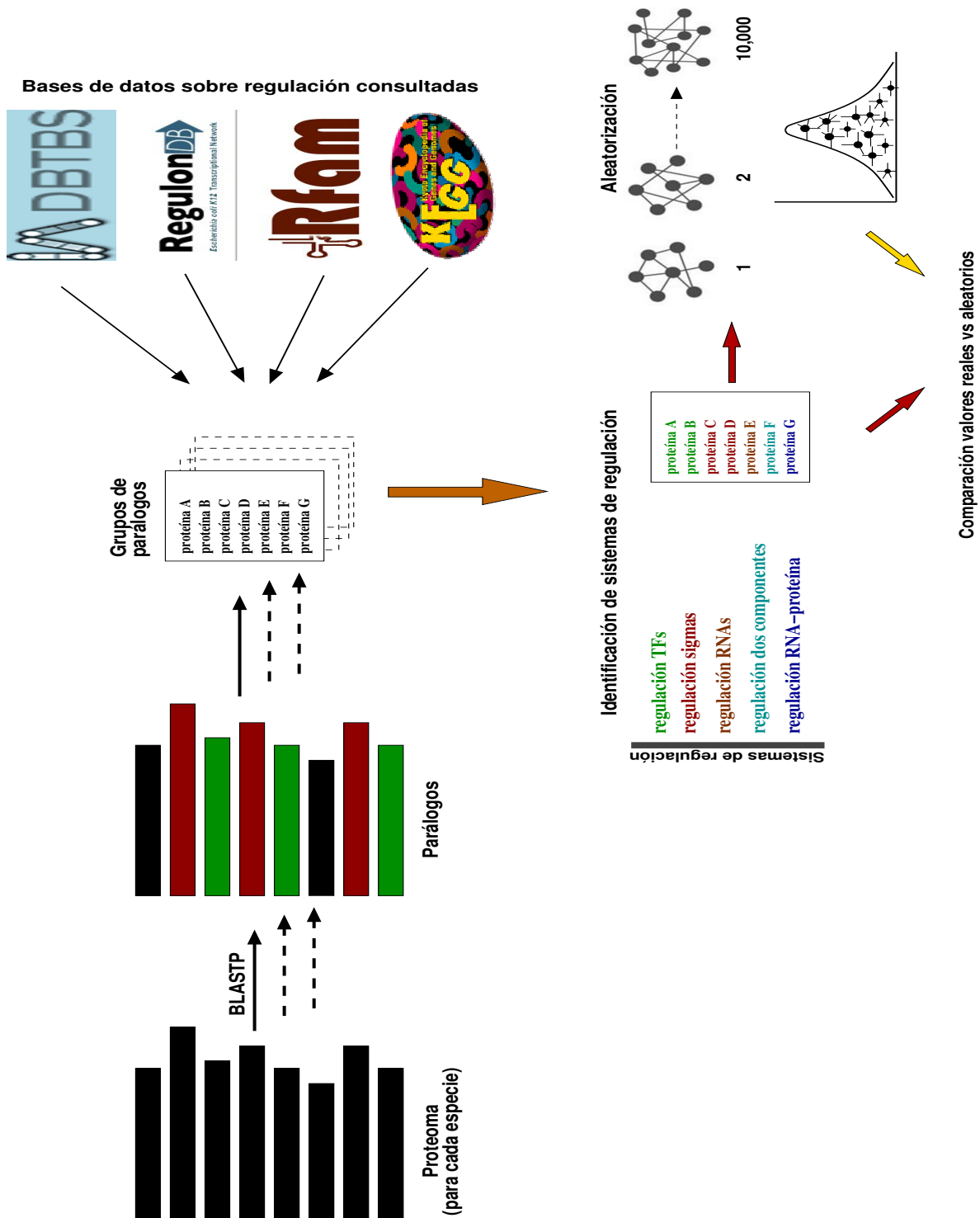


Figura 3: Metodología utilizada para la identificación de proteínas parálogas y sus mecanismos de regulación en los genomas de *E. coli* y *B. subtilis*. A partir de las secuencias de proteínas de cada genoma, se identificaron a los grupos de parálogos utilizando el programa BLASTP. La regulación presente en cada grupo de parálogos fue obtenida de distintas bases de datos sobre regulación. Se clasificaron a los sistemas de regulación de acuerdo al elemento de regulación: TFs, factores σ , RNA (*cis* y *trans*), sistemas de dos componentes y RNA-proteína. Posteriormente se realizó una aleatorización de las relaciones observadas en cada tipo de sistema de regulación. Se llevaron a cabo 10,000 simulaciones para cada tipo de sistema, los valores medios obtenidos de las aleatorizaciones fueron comparados contra los valores reales observados, en términos de desviaciones estándar.

5. Resultados

De los 477 grupos de parálogos identificados en *E. coli*, el 92 % de ellos está constituido de dos a cinco parálogos cada uno, mientras que el restante 8 % de grupos se encuentra conformado de seis o más parálogos, siendo 60 el número máximo de copias encontradas en un grupo (Fig. suplementaria 12, Cuadro suplementario 2). Los grupos con el mayor número de duplicados (60, 28, 22), contienen proteínas parálogas cuyas asignaciones funcionales de acuerdo a la base de datos COG, son, transportadores del tipo ABC y reguladores transcripcionales de unión al DNA. 291 grupos se encuentran regulados por al menos un TF, y en 248 grupos de parálogos, los TFs que regulan a las copias pueden ser distintos. Un ejemplo lo representan los genes parálogos *zntA* y *copA* de *E. coli*, los cuales codifican para proteínas transportadoras de zinc y cobre, respectivamente. La regulación de *zntA* y *copA* se lleva a cabo por los TFs parálogos ZntR y CueR, ZntR regula la expresión de *zntA* y CueR la de *copA* (Fig. 4). Los iones metálicos son requeridos en cantidades traza por las células, por lo que su disponibilidad en bajas o en altas concentraciones, provocan alteraciones en el metabolismo celular. La homeostasis del zinc en *E. coli* se mantiene por medio de la activación de un grupo de genes, cuyos productos codifican para proteínas de transporte y exportación. Cuando en el interior de la célula hay un exceso de zinc, el TF ZntR activa la transcripción de *zntA*, cuya proteína lleva a cabo la exportación del exceso de zinc del citoplasma al periplasma [97], [69]. En el caso de la homeostasis del cobre, la proteína exportadora CopA es la encargada de expulsar el exceso de cobre en el citoplasma. El gene *copA* se encuentra regulado por el TF parálogo CueR, el cual sensa los niveles de cobre en el citoplasma de la bacteria. La activación transcripcional de *copA*, permite que su producto polipeptídico exporte el exceso de cobre del citoplasma al periplasma, donde la proteína CueO se encarga de oxidarlo [96], [17], [85].

Para el caso de *B. subtilis*, de los 483 grupos de parálogos que se definieron, el 92 % presentan de dos a cinco duplicados, mientras que el restante 2 % presentan de seis hasta 57 parálogos por grupo (Fig. suplementaria 13). Los grupos con el mayor número de copias (57, 29, 28) tienen asignaciones funcionales de acuerdo a las categorías COG de transportadores tipo ABC, deshidrogenasas de cadena corta y reguladores transcripcionales de unión al DNA. En 225 grupos se encontró regulación

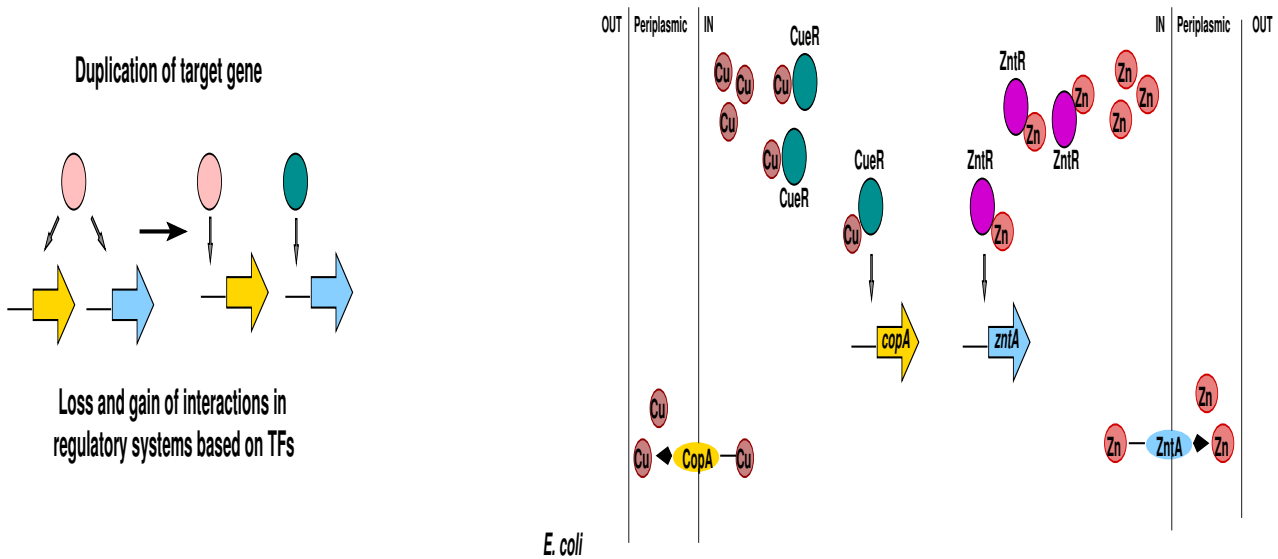


Figura 4: Genes parálogos donde la divergencia de la regulación se debe a la presencia de distintos TFs actuando en las copias. Los genes parálogos *zntA* y *copA* codifican para los transportadores de zinc y cobre, respectivamente. La regulación de las copias es ejercida por los TFs parálogos ZntR y CueR. Figura modificada de [57].

en al menos una de las copias por TFs. Para evaluar la ocurrencia de la divergencia en la regulación de los genes parálogos, se procedió a determinar la posibilidad de igualar el número de interacciones de regulación observadas a través del azar. Para lo cual se contrastó la ocurrencia de la divergencia en las redes de parálogos estudiadas, contra las generadas de manera aleatoria, y observar si el azar era capaz de reproducir los datos obtenidos. Para determinar la significancia estadística de que diferentes TFs se encuentren regulando a un par de genes parálogos, se procedió a realizar una evaluación similar a la realizada por Teichmann y Babu (2004) [87], para el análisis de los resultados encontrados en sus estudios de TRN. Las interacciones de regulación divergente observadas fueron aleatoriamente barajadas, para lo cual se realizaron 10,000 simulaciones conservando el número de relaciones, así como los mismos elementos (TFs, genes blanco). Posteriormente se compararon las medias de las interacciones simuladas contra el número de interacciones observadas; el mismo procedimiento se realizó tanto para *E. coli* como para *B. subtilis*. En términos de las desviaciones estándar (SD), el número de interacciones observadas en *E. coli* se encuentra a 19.5 SD de la media de las interacciones divergentes de las 10,000 simulaciones (Cuadro 1). Para el caso de *B. subtilis* el número de las interacciones observadas

en la TRN estuvo a 19 SD de la media (Cuadro 1) de las 10,000 aleatorizaciones correspondientes a su red. Los datos alejados de la media, en unidades de desviaciones estándar, implican eventos estadísticamente significativos, los cuales no se encuentran solamente debido al azar en una distribución normal. El área bajo la curva a la derecha de 3 SD equivalen al 0.5% de datos y la probabilidad de obtener al azar datos tan alejados de la media, como los obtenidos de 19 SD, es prácticamente nula (Fig. suplementaria 14). Los resultados obtenidos corroboran lo reportado en otros estudios para los genes parálogos ([34], [68], [58]), donde lo que se observa es una divergencia en los mecanismos de regulación que modulan a las copias. Las relaciones de regulación surgidas por procesos de duplicación son las que predominan en las TRN, reportándose sólo una pequeña fracción de relaciones de regulación evolucionando por recombinación o por innovación [87], [4].

5.1. Divergencia en el tipo de regulación (activación/represión) del TF sobre sus genes parálogos blanco

El modelo propuesto por Teichmann y Babu no considera en su esquema el tipo de regulación que ejerce el TF sobre su blanco, si es positiva o negativa, y si se conserva o cambia; Teichmann y Babu sólo consideran el surgimiento o pérdida de la interacción entre los parálogos y su regulador. En 43 grupos de genes parálogos de los 291 regulados por TFs en *E. coli*, el TF que regula a las copias se conserva y ejerce una actividad de regulación dual sobre las copias: a una lo regula de manera positiva, mientras que a la otra de manera negativa. Este tipo de escenario evolutivo se observa en la regulación de los genes *gntK* e *idnK* de *E. coli*, involucrados en la síntesis de 6-fosfogluconato para las rutas Entner-Doudoroff y de las pentosas fosfato, respectivamente. Tanto *gntK* como *idnK*, se encuentran regulados por los mismos TFs, los cuales son CRP, GntR y IdnR. La regulación que ejerce IdnR sobre las copias es diferencial, mientras que en *gntK* el efecto ejercido por IdnR es la de reprimir la transcripción, en el parálogo *idnK* la acción de IdnR es la de ser un activador de la transcripción [11], [47] (Fig. 5).

En el caso de *B. subtilis*, de los 225 grupos de parálogos regulados por al menos un TF, sólo 8 grupos están regulados por TFs cuya actividad sobre las copias es dual. Un ejemplo que ilustra la

Cuadro 1: Valores de las simulaciones de las interacciones de los distintos tipos de sistemas de regulación, descritos para *E. coli* (eco) y *B. subtilis* (bsu). i-real: interacciones reales observadas; \bar{x} i-azar: media de 10,000 interacciones al azar simuladas; NA: No Aplica.

	i-real	\bar{x} i-azar	SD	distancia
Divergencia en TFs				
eco	531	239.82	14.91	19.52
bsu	189	45.32	7.55	19.02
Divergencia en tipo de regulación				
eco	136	94.66	11.26	3.67
bsu	16	8.57	3.79	1.96
Factores σ, anti-σ, anti-anti-σ				
eco	70	7.01	3.51	17.97
bsu	91	39.72	7.55	6.79
Interacciones proteína-proteína				
eco	16	0.23	0.67	23.46
bsu	12	0.29	0.75	15.53
RNAs en <i>cis</i> y <i>trans</i>				
eco	NA	NA	NA	NA
bsu	16	0.88	1.3	11.64

divergencia en la modulación en *B. subtilis*, son los genes parálogos *tnrA* y *glnR*. Los productos de *tnrA* y *glnR* son los factores transcripcionales TnrA y GlnR, respectivamente, y controlan la expresión de los genes involucrados en la disponibilidad de nitrógeno. Los dominios de unión al DNA de ambos TFs son similares, variando en el dominio de la transducción de las señales, por lo que cada proteína responde ante diferentes condiciones de crecimiento. TnrA activa y reprime la transcripción de sus blancos cuando hay una limitante en el nitrógeno disponible, en tanto que que GlnR reprime la transcripción de sus blancos ante un exceso de nitrógeno. Los genes *tnrA* y *glnR* se auto-regulan, además de presentar una regulación cruzada. GlnR reprime a *tnrA* y a su propio gene, mientras que

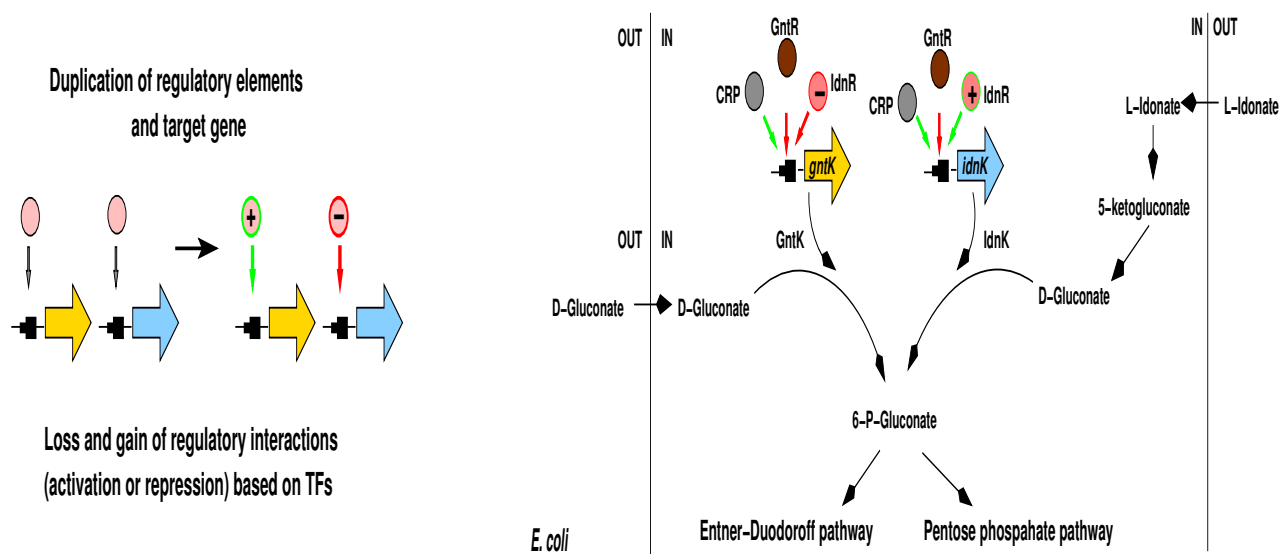


Figura 5: Genes parálogos donde la divergencia de la regulación se debe al cambio de la actividad reguladora sobre las copias. El TF IdnR ejerce una regulación dual en los genes duplicados *gntK* y *idnK*, activando en uno la transcripción y reprimiéndola en el otro. Figura modificada de [57].

en el caso de TnrA, regula de manera dual a sus blancos: reprime a el gene *glnR* y activa a *tnrA* [14], [98]. El porcentaje de grupos de parálogos con regulación diferencial fue menor en *B. subtilis* con un valor de 3.6%, mientras que en *E. coli* fue de 14.8%, lo que nos estaría indicando un mayor aporte a la diversidad de la TRN en éste último organismo por divergencia en el tipo de interacción (+/-). En *E. coli* la actividad de regulación dual se encuentra con mayor frecuencia asociado a grupos de parálogos regulados por TFs globales como ArcA, CRP, FIS, FNR, Hns, IHF y Lrp [55]. En contraste, sólo algunos de los TFs globales de *B. subtilis* como AbrB, CcpA, SpoOA y TnrA [61] presentan una actividad dual de represión/activación sobre los parálogos, mientras que otros actúan siempre como represores (CodY, Fur, LexA y YrzC) o como activadores (ComK). Solamente una porción muy pequeña de TFs locales presentan una actividad dual (+/-) sobre sus parálogos blancos. Es el caso del regulador NtrC de *E. coli*, el cual reprime la transcripción del transportador de glutamina *glnH*, mientras que activa la transcripción del transportador de histidina/lisina/arginina/ornitina *hisJ* y el transportador de lisina/arginina/ornitina *argT* [102]. Para evaluar la significancia estadística de que un elemento de regulación del tipo TF, se encuentre modulando de manera divergente (activando a uno

y reprimiendo al otro) a genes parálogos, se procedió a realizar una evaluación similar a la realizada anteriormente a los grupos con divergencia en la regulación por TFs. Las interacciones de regulación dual observadas fueron aleatoriamente barajadas, posteriormente se compararon las medias de las interacciones simuladas contra el número de interacciones observadas. el mismo procedimiento se realizó tanto para *E. coli* como para *B. subtilis*. En términos de las desviaciones estándar, el número de interacciones observadas en *E. coli* se encuentra a 3.7 SD (Cuadro 1) de la media de las interacciones duales de las 10,000 simulaciones. Para el caso de *B. subtilis* el número de las interacciones duales observadas en la TRN estuvo a 2 SD (Cuadro 1) de la media de las 10,000 aleatorizaciones correspondientes a su red. Cuando los reguladores globales tanto de *E. coli* y de *B. subtilis* son removidos y se realizan de nuevo las aleatorizaciones, las distancias se incrementan 6.5 y 3.4 SD, respectivamente. Estos resultados remarcan la plasticidad evolutiva de las redes de regulación, no sólo a través de la adquisición o pérdida de una interacción, si no también con la modificación en el tipo de modulación (positiva/negativa) ejercida sobre los parálogos. Nuestros resultados sugieren que la divergencia en la modulación es una de las primeras etapas para la innovación a un nivel bioquímico en las copias, tal vez como un paso hacia la modificación de toda una ruta. Aunque los parálogos pueden retener la misma función, la expresión diverge de manera rápida entre las copias después del evento de duplicación, implicando que modificaciones en la expresión genética preceden a potenciales cambios en la función. Diversos autores mencionan la contribución que tiene la divergencia funcional de los parálogos para la innovación a nivel bioquímico [25], [34], [100]. Para corroborar nuestras observaciones, analizamos la divergencia metabólica de los genes parálogos, a través de la identificación de las diferencias en la reacción bioquímica o ruta metabólica de las copias. Observamos para *E. coli* que 198 (42%) grupos de parálogos, tienen asignados una reacción enzimática o una ruta metabólica de acuerdo a la base de datos KEGG [49]. Del conjunto de 198 grupos con un KEGG asignado, 53 (27%) de ellos comparten la misma ruta metabólica y realizan la misma reacción enzimática; 64 (32%) presentaron parálogos en diferentes rutas metabólicas; 26 (13%) grupos estuvieron conformados con genes parálogos que se encuentran en la misma ruta metabólica, pero ocupando diferentes pasos dentro de la misma; 55 (28%) grupos presentaron ID-KEGGS formados parcialmente por alguna de las categorías descritas antes (Fig. 6). Con respecto a los 485 grupos de parálogos de *B. subtilis*, se encontraron que 178 (36%)

de ellos tienen asignados una reacción o una ruta metabólica en el KEGG. Del universo de 178 grupos, 79 (44 %) de ellos comparten la misma ruta metabólica y realizan la misma reacción enzimática; 52 (29 %) presentaron parálogos en diferentes rutas metabólicas; 11 (6 %) grupos estuvieron conformados con genes parálogos que se encuentran en la misma ruta metabólica, pero ocupando diferentes pasos dentro de la misma; y 36 (20 %) grupos presentaron parálogos formados parcialmente por elementos en distintas rutas o pasos metabólicos (Fig. 6). La mayor divergencia tanto a nivel metabólico como en el tipo de regulación (+/-), sucede en *E. coli*. La rapidez con que ocurren los cambios metabólicos que originan la divergencia en la ruta o en el paso, es menor en *B. subtilis*. Más del 70 % de grupos de parálogos en *E. coli* han evolucionado para ser reclutados en vías metabólicas diferentes, o bien, estando en la misma ruta, las copias se encuentran realizando pasos metabólicos distintos. En *B. subtilis* la mayoría de los grupos (44 %) parálogos se encuentran conservando la ruta metabólica.

5.2. Divergencia en sistemas de regulación distintos a TFs: factores σ

Con respecto a la divergencia en la regulación por factores σ , 100 (21 %) de los 477 grupos de parálogos en *E. coli* y 144 (30 %) de los 485 grupos parálogos de *B. subtilis* están regulados por factores σ alternativos, o bien anti- σ y anti-anti- σ . Se procedió a realizar una comparación de los datos observados contra las interacciones obtenidas por simulación, tal como se describió antes para los TFs con regulación dual. El valor de las interacciones de regulación por σ alternativos en *E. coli* y *B. subtilis* estuvieron a 18 y 6.8 SD, respectivamente, de la media obtenida para las aleatorizaciones correspondientes (Cuadro 1). El escenario de evolución para genes parálogos basado en factores σ , se ilustra con σ^B y σ^F de *B. subtilis*. σ^B se encuentra involucrada en la respuesta a estrés y σ^F en la esporulación temprana, la red de regulación involucra también a factores anti- σ y anti-anti- σ , haciendo más compleja la evolución de los mecanismos de regulación. En un primer momento, la actividad de σ^B y σ^F sobre sus genes blanco es inhibida por la unión de los factores anti- σ RsbW y SpoIIAB, respectivamente. Posteriormente los factores parálogos anti- σ son negativamente regulados por los factores anti-anti- σ RsbV y SpoIIAA, finalmente la transcripción de los genes que codifican para los factores parálogos anti-anti- σ dependen de los factores σ parálogos σ^A y σ^F [41] (Fig. 7).

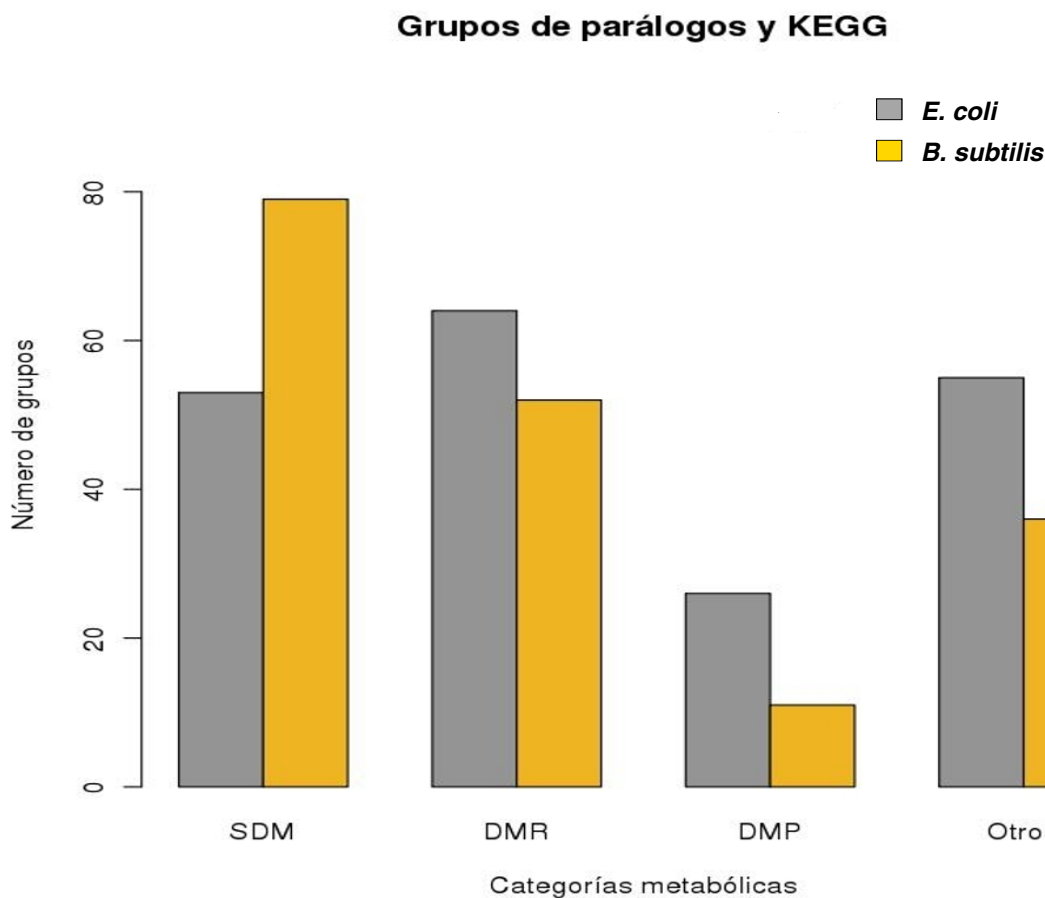


Figura 6: Grupos de genes parálogos y su anotación en el KEGG [49]. Las categorías en las cuales se agruparon son: SDM (sin divergencia en la ruta metabólica y la función), DMR (divergencia en la ruta metabólica), DMP (divergencia en el paso metabólico). En la categoría de otros hay una combinación de divergencia en la ruta y/o en el paso metabólico.

La red de regulación de los factores σ , representa un ejemplo de la extensión del modelo de Teichmann y Babu [87], donde los elementos de regulación divergentes no son TFs. La evolución de los motivos estructurales de las TRN se ve enriquecida por la duplicación de los elementos de regulación que participan en el mismo, se ha observado que los motivos del tipo “feed-forward” incorporan a genes duplicados dentro de sus circuitos, como sucede con los parálogos σ^F y σ^E de *B. subtilis* los cuales son parte de motivos “feed-forward” distintos [72]. Los eventos de duplicación enriquecen los motivos estructurales, lo cual permite una reconfiguración de las relaciones dentro de la TRN, permitiendo al organismo una mejor adaptación a su nicho [28].

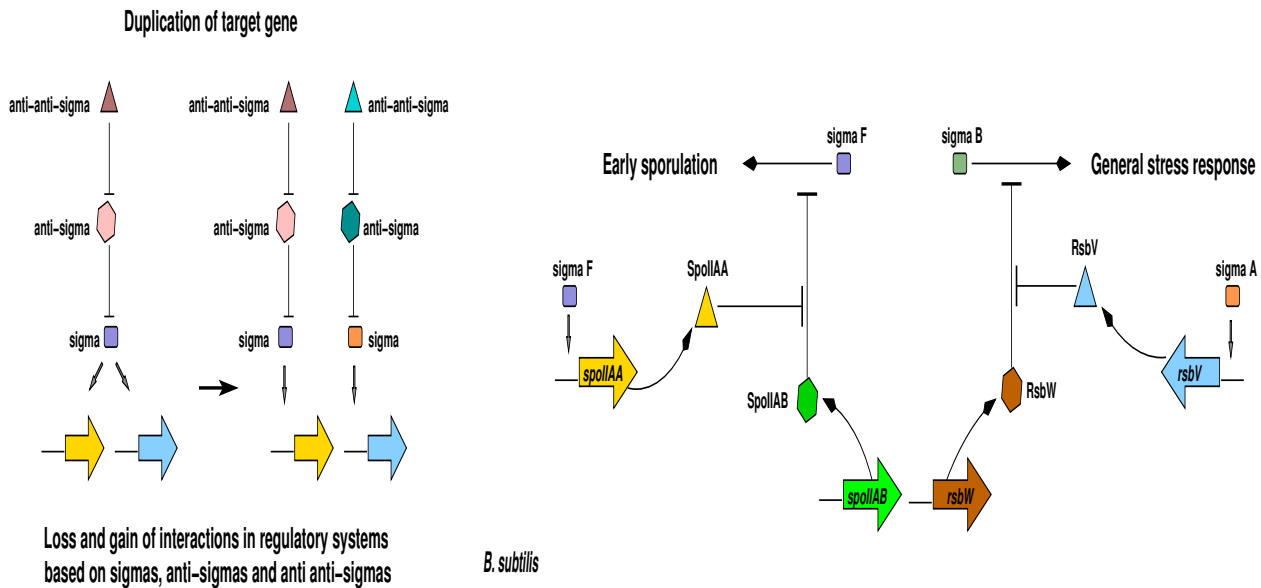


Figura 7: Genes parálogos donde la divergencia de la regulación se presenta en sistemas mediados por factores σ , anti- σ y anti-anti- σ . Figura modificada de [57].

5.3. Divergencia en sistemas de regulación distintos a TFs: sistemas de dos componentes

Las redes de regulación que evolucionan por duplicación de genes, no se encuentran reguladas exclusivamente a nivel del inicio de la transcripción. La actividad de proteínas parálogas puede verse modificada también a nivel post-traducciona, a través de interacciones proteína-proteína. Este tipo de interacción está representada por los sistemas de dos componentes de transducción de señales. Muchos de estos sistemas están constituidos por una proteína histidina-cinasa unida a la membrana de la célula, la cual sensa señales específicas del ambiente y cataliza la transferencia de un grupo fosfato a su correspondiente regulador de respuesta, el cual induce la activación de sus blancos. Los sistemas de dos componentes, han surgido en su mayoría por procesos de duplicación y diversificación de los dominios que componen a las proteínas [92]. El dominio sensor de la histidina-cinasa, evoluciona de manera más rápida que el dominio de unión al DNA de los reguladores de respuesta. La variación de los dominios de regulación permiten una divergencia en la especificidad de la regulación, lo cual

amplía el intervalo de señales ante las que puede responder la célula. La duplicación y la subsecuente recombinación de los sistemas de dos componentes, permiten el desarrollo de la diversidad genética [19], [2], [90]. Un ejemplo de regulación por sistemas de dos componentes, son los reguladores de respuesta parálogos GlnG y AtoC de *E. coli*. GlnG y AtoC están involucrados en la activación transcripcional de genes de respuesta a la limitación de nitrógeno y el catabolismo de los ácidos grasos de cadena corta, respectivamente [51]. La actividad de regulación de GlnG y AtoC depende de sus correspondientes histidina-cinasas parálogos GlnL y AtoS [52], [66] (Fig. 8).

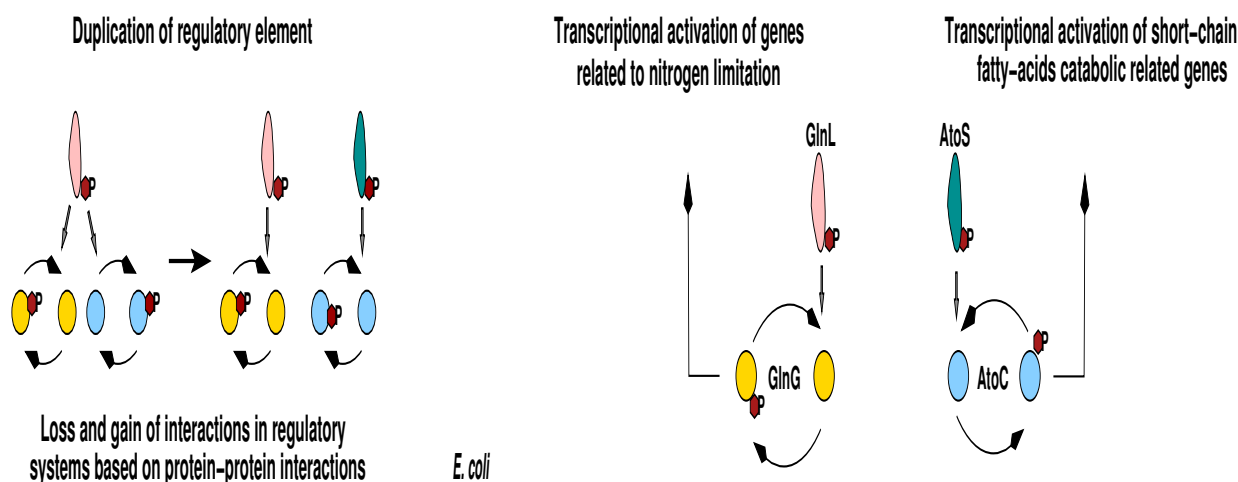


Figura 8: Genes parálogos donde la divergencia de la regulación se presenta en sistemas mediados por sistemas de dos componentes. Figura modificada de [57].

Del conjunto de genes parálogos analizados, se encontro que nueve grupos de parálogos tanto en *E. coli* como en *B. subtilis* están regulados por sistemas de dos componentes. A pesar de que el número de grupos con sistemas de regulación proteína-proteína representa una fracción pequeña del conjunto de datos, las interacciones observadas se encuentran a 23.5 y 15.5 SD con respecto de la media de las aleatorizaciones en *E. coli* y *B. subtilis*, respectivamente. El papel que desempeñan las interacciones proteína-proteína en las redes de genes parálogos, tiene una relevancia jerárquica para toda la red, ya que se encuentran regulando muchos otros genes o interactuando con otras proteínas para formar cascadas de regulación transcripcional. Dichas cascadas permiten la amplificación de la respuesta [62], ante determinadas señales externas. Es el caso de los sistemas de dos componentes PhoP/PhoQ, Rs-

tA/RstB y PhoB/PhoR, los cuales responden a las condiciones ácidas del medio en *E. coli* [67], y cuyos reguladores de respuesta son parálogos.

5.4. Divergencia en sistemas de regulación distintos a TFs: RNAs (*cis* y *trans*) e interacciones proteína-RNA

La importancia de la molécula de RNA en la regulación de los genes ha sido documentada en años recientes. En bacterias, el mecanismo de regulación por RNA puede actuar de dos formas, como un elemento en *cis*, por ejemplo los riboswitches, los cuales se localizan en la región líder 5' del gene que regulan [94], [93]. O bien en *trans*, regulando a su mRNA blanco a través del apareamiento de sus bases, resultando en cambios en la traducción o estabilidad [30]. Usando los modelos de co-varianza de la base de datos Rfam [27], se procedió a buscar elementos de regulación por RNA. En *E. coli* los genes parálogos *rpoH* y *rpoS* se regulan a través del RNA. El gene *rpoH* codifica para el factor σ^H y está regulado por un elemento en *cis*, el cual se encuentra en la región codificante. A bajas temperaturas, el plegamiento del termo-regulador sobrelapa y secuestra al sitio de unión del ribosoma (RBS), inhibiendo la traducción del mRNA de *rpoH*. A temperaturas de 42°C, la conformación del plegamiento que bloquea la traducción cambia, permitiendo que el RBS sea reconocido por el ribosoma e inicie la traducción [64]. En el caso del gene parálogo *rpoS*, la regulación depende de un grupo de RNA pequeños actuando en *trans*, en respuesta a distintos tipos de factores de estrés como bajas temperaturas (DsrA), choque osmótico (RprA) y oxidativo (OxyS). RprA y DsrA tienen un efecto desestabilizador sobre la estructura que bloquea el acceso al RBS, permitiendo que se lleve a cabo la traducción, en tanto que OxyS tiene un efecto negativo en la traducción por un mecanismo no conocido [30] (Fig. 9).

Se identificaron 28 y 52 grupos de genes parálogos correspondientes a *E. coli* y *B. subtilis*, con regulación por moléculas de RNA. Se realizaron simulaciones para las relaciones observadas donde un RNA participa en la regulación de dos o más parálogos, encontrando interacciones de este tipo sólo en *B. subtilis*, con un valor alejado en 11.6 SD de la media de su aleatorización. En general los elementos de regulación en *cis*, como los riboswitches, están presentes en mayor medida en los genes

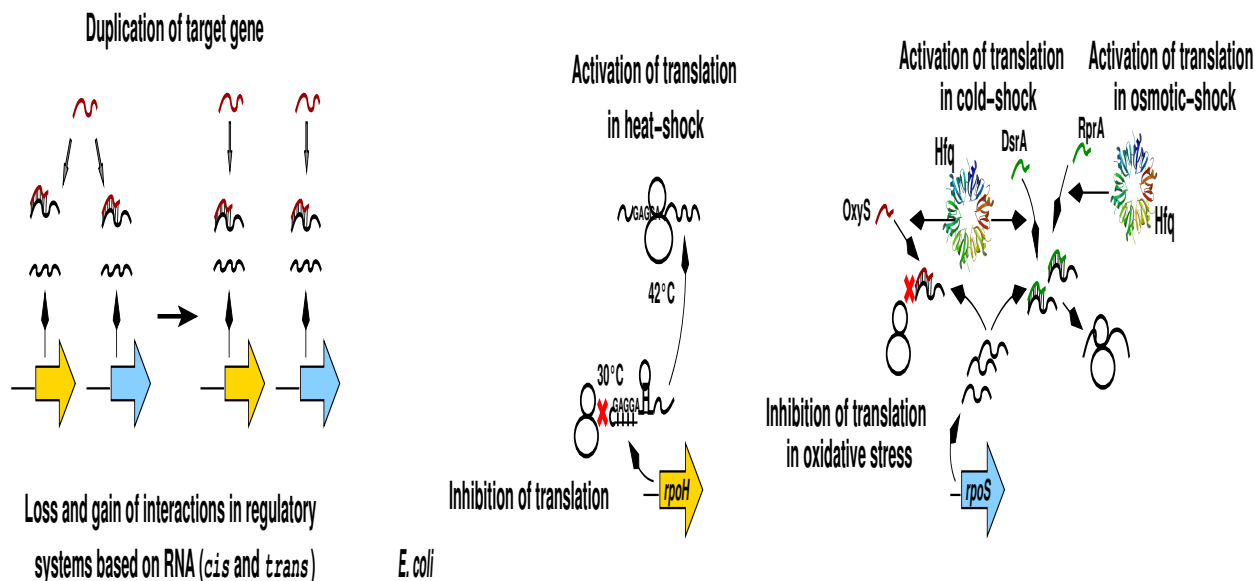


Figura 9: Genes parálogos donde la divergencia de la regulación se presenta en sistemas mediados por RNAs en *cis* y *trans*. Figura modificada de [57].

parálogos del Firmicute *B. subtilis*. En tanto que la regulación de genes parálogos por RNAs en *trans* es particularmente importante en *E. coli* y otras bacterias Gram-negativas. Hay una tendencia en los RNAs en *trans* de regular genes que codifican para proteínas de membrana externa, las cuales están involucradas en la respuesta a estrés. Algunos de los sRNA se han generado por duplicación, como el caso del RNA OmrA de *E. coli*, el cual es expresado en la fase estacionaria tardía, y su parálogo OmrB, el cual se expresa de manera transitiva durante la fase estacionaria temprana [37]. Las interacciones proteína-RNA también pueden ser incluidas en las redes de regulación generadas por duplicación. Un ejemplo de regulación por interacciones proteína-RNA son los genes parálogos *bglP* y *sacP* de *B. subtilis*, los cuales codifican para las permeasas que transportan salicina y sacarosa, respectivamente. La regulación transcripcional de las copias se lleva a cabo por las proteínas parálogas de unión al RNA LicT y SacT, las cuales se unen a anti-terminadores que se localizan en las regiones río arriba de los genes *bglP* y *sacP* [77] (Fig. 10). Dicha unión lleva a cabo la disrupción de estructuras terminadoras que bloquean el inicio de la transcripción de *bglP* y *sacP*. Se llevó a cabo la evaluación de las interacciones observadas RNA-proteína. Dicha evaluación se realizó sólo en *B. subtilis*, ya que en *E. coli* no se observó la existencia de dicho mecanismo. El número de interacciones observadas de RNA-proteína

fue de 12, dicho valor se encuentra a 11.6 SD de la media de sus simulaciones.

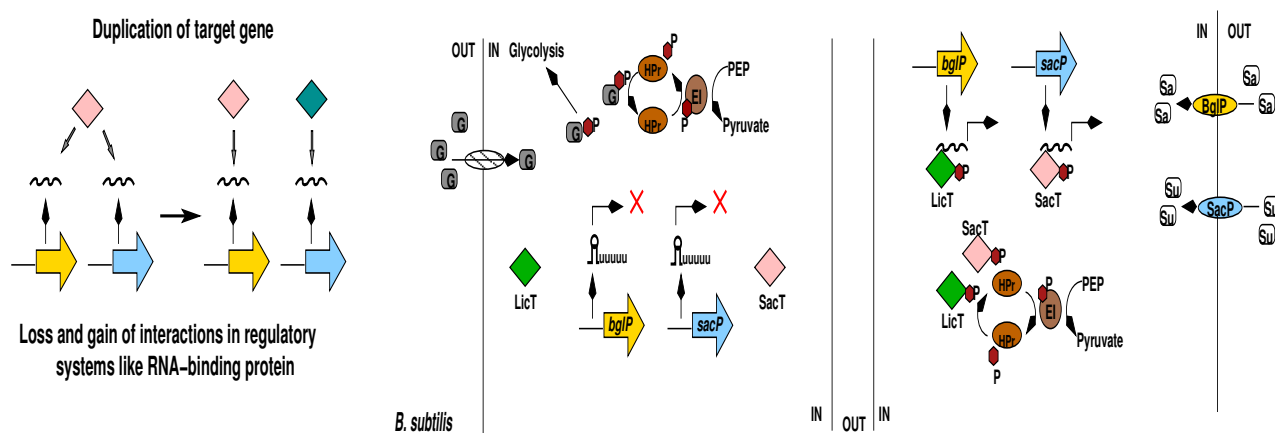


Figura 10: Genes parálogos donde la divergencia de la regulación se presenta en sistemas mediados por RNA-proteínas.

Figura modificada de [57].

5.5. Divergencia en sistemas de regulación distintos a TFs: super-enrollamiento

Otro mecanismo de regulación que actúa en la modulación de los genes parálogos es el super-enrollamiento, el cual ha sido descrito como un regulador global de la expresión genética. La expresión coordinada de los genes involucra distintos niveles jerárquicos, como el control individual de genes u operones, o bien operones organizados en estructuras mayores como los regulones. Se ha propuesto que los niveles más altos en la jerarquía de la regulación genética, son modulados por sistemas de super-enrollamiento del DNA [38]. El grado de super-enrollamiento del DNA puede ser modulado por la acción de las topo-isomerasas. Dicha modulación surge como respuesta a una variedad de condiciones ambientales o nutricionales como lo son la temperatura, osmolaridad, estrés oxidativo, cambios en el crecimiento de medio aerobio a anaerobio, entre otros [38], [31], [18], [43]. El super-enrollamiento del DNA ajusta los niveles basales de expresión de todos aquellos genes cuya transcripción dependa de promotores que son directamente sensibles al enrollamiento del DNA [38]. Es el caso para algunos genes cuya transcripción por σ^{38} , dependen del super-enrollamiento que ocurre ante las condiciones de crecimiento [13]. El sistema de regulación por super-enrollamiento forma parte también de las redes

de regulación de genes parálogos. De acuerdo a nuestros datos existen en *E. coli* 15 grupos de genes parálogos cuya regulación está siendo directamente modulada por super-enrollamiento, mientras que en *B. subtilis* no se observó este efecto. Tres de los grupos regulados por super-enrollamiento del DNA, son reguladores transcripcionales que a su vez regulan a otros genes. El efecto inicial de regulación por super-enrollamiento, se expande en la red de regulación a través de los reguladores transcripcionales. Un ejemplo es el regulador OxyR, el cual está involucrado en la respuesta a estrés oxidativo. El bajo número de interacciones observadas donde dos o más parálogos estuvieran regulados por super-enrollamiento, no permitió realizar las aleatorizaciones antes descritas.

5.6. Distintos tipos de elementos de regulación modulan a los parálogos

La respuesta a las diferentes señales extracelulares y metabólicas por parte de la red de regulación de genes parálogos, involucra la integración de distintos tipos de elementos reguladores, actuando a diferentes niveles. Un buen ejemplo sobre la integración de elementos y mecanismos de regulación se encuentra en las porinas parálogos OmpC y OmpF de *E. coli*, involucradas en la respuesta a estrés por osmolaridad, temperatura y oxidativo [37], [73]. Las distintas capas y tipos de elementos de regulación de las copias comprende: i) regulación en el inicio de la transcripción por 11 TFs que modulan directamente a *ompC* y *ompF*; ii) regulación al inicio de la traducción, efectuada por un grupo de RNAs pequeños; iii) control post-transcripcional por la chaperona de RNA StpA sobre el RNA de *micF*, induciendolo a degradación; y iv) modificaciones post-traduccionales de la actividad de la proteína, como en el caso de la activación de OmpR, por medio de su fosforilación por la histidina-cinasa EnvZ. Un sólo elemento de regulación puede actuar a distintos niveles, como por ejemplo, H-NS el cual regula de manera directa a *ompC* reprimiendo su transcripción y de manera indirecta la traducción de *ompF*, reprimiendo la transcripción del RNA antisentido *micF* (Fig. 11). El TF OmpR presenta una actividad de regulación dual sobre su blanco. OmpR es un regulador de respuesta del sistema compuesto por OmpR-EnvZ, el cual sensa los cambios de osmolaridad. El estado activo de OmpR depende de la fosforilación que lleva a cabo EnvZ, a baja osmolaridad, la proteína fosforilada OmpR-P se encuentra en bajas concentraciones, y se lleva a cabo la activación transcripcional de

ompF. Mientras que a alta osmolaridad, las concentraciones de OmpR-P son altas, y la proteína se une a sitios adicionales de baja afinidad en *ompF*, bloqueando la transcripción del gene [73].

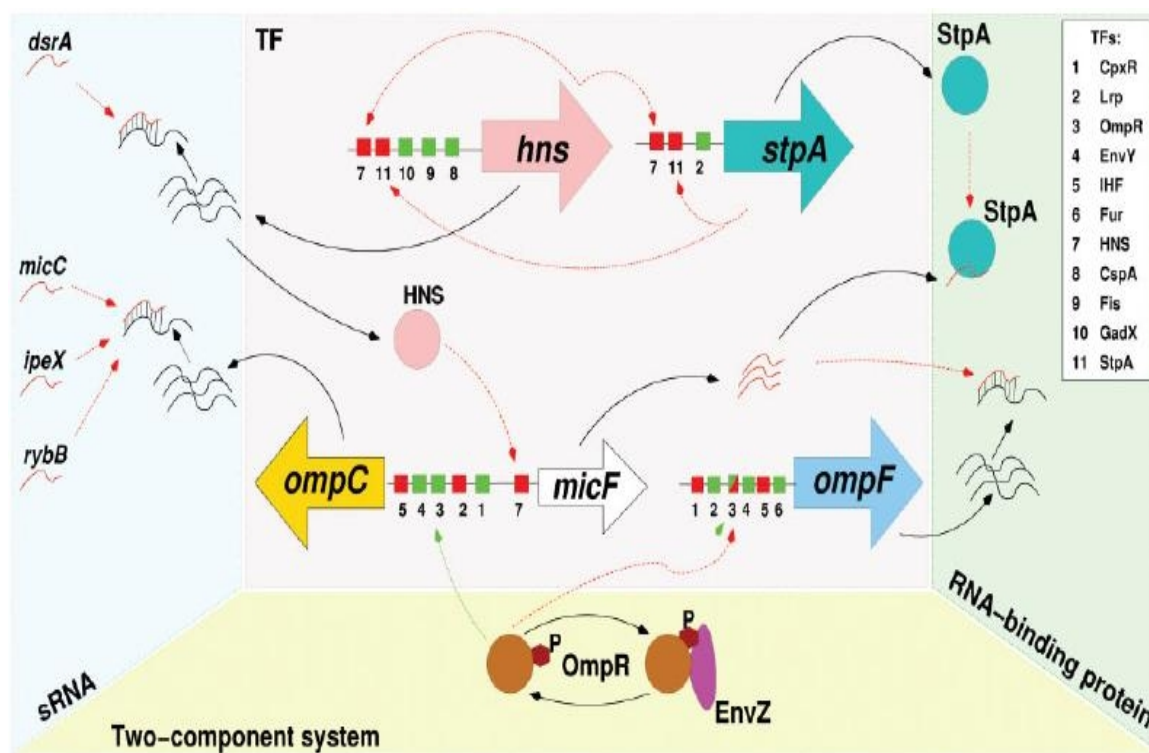


Figura 11: Ejemplo de integración de la red de regulación de genes parálogos, basado en diferentes elementos de regulación. Se esquematiza la red de regulación de las porinas parálogas *OmpC* y *OmpF*. Los genes y sus correspondientes productos polipeptídicos están con el mismo color. Figura tomada de [57].

6. Conclusiones

La compilación y análisis de los elementos de regulación de genes parálogos en *E. coli* y *B. subtilis*, nos ha llevado a extender el modelo original de evolución de redes de regulación, los cuales tomaban sólo a los TFs en cuenta, a un modelo de tipo más general, el cual incluye otros reguladores tales como los RNA pequeños (*trans*), riboswitches (RNA en *cis*), proteínas de pegado a RNA, factores σ , sistemas de dos componentes y super-enrollamiento del DNA. A pesar de que los TFs son los elementos más extensamente usados en las redes de regulación, la utilización de otros elementos de regulación ha resultado en un incremento significativo de la versatilidad de la red. Nuestros análisis son consistentes con observaciones previas [21] acerca de la relevancia de los genes parálogos en las redes de regulación, para los procesos de adaptación de un organismo en respuesta a los cambios ambientales. Los procesos de duplicación genética proporcionan no sólo el material para la evolución de los genes, si no que aportan también complejidad a la evolución de los genomas [39]. La regulación de los genes parálogos además de poder divergir [68], [58], [35], puede integrar diferentes sistemas de modulación, los cuales actúan en distintas etapas del procesamiento de la información genética. Tal es el caso de la regulación de las porinas OmpC y OmpF, las cuales se encuentran reguladas no sólo a nivel de la transcripción, si no que los procesos de regulación suceden también a nivel post-transcripcional, mediados por elementos como sRNA e interacciones RNA-proteína (Fig. 11).

Nuestros resultados también destacan la importancia que tiene la modulación de un TF sobre sus genes blanco parálogos, donde a una de las copias la puede regular de manera positiva, mientras que a la otra copia de manera negativa. Este tipo de mecanismo permite incrementar la diversidad de circuitos de regulación dentro de la red, sin la necesidad de ganar o perder una interacción a través del reclutamiento de TFs nuevos o la pérdida de los sitios de unión del TF en los parálogos. La divergencia en el tipo de regulación (+/-) que se ejerce sobre genes parálogos, puede ocurrir como una primera etapa para una innovación bioquímica en las copias. La mayoría de grupos de parálogos que presentan divergencia en el tipo de modulación (+/-), presentan también divergencia metabólica, tanto en *E. coli* como en *B. subtilis*. Dicha divergencia metabólica hace que los parálogos se encuentren en diferentes rutas bioquímicas, o bien, si se encuentran en la misma ruta, las copias ocupan pasos distintos dentro de

la misma. De esta manera los procesos de divergencia en la regulación de los parálogos, se acompañan de cambios en las funciones metabólicas de las copias.

7. Perspectivas

La divergencia en los mecanismos de regulación en genes parálogos [57], preceden los cambios que suceden entre las copias a nivel de función. El proceso de duplicación es un mecanismo por el cual familias de proteínas homólogas pueden expandirse y diversificarse, generando subfamilias al interior del grupo. Las subfamilias de genes parálogos producto de las duplicaciones, son una fuente para la generación de divergencia funcional [1], [54], [91]. La correcta identificación de subfamilias dentro de las familias de proteínas homólogas, nos ayudará a definir mejor las funciones llevadas a cabo por cada subgrupo, lo cual redundará en una mejor transferencia de anotaciones funcionales. Un ejemplo donde se observa la divergencia funcional entre subfamilias de proteínas, es el de la familia de la fosfatidilglicerofosfato sintasa (PgsA) la cual es la encargada de catalizar la conversión de CDP-diacilglicerol a fosfatidil-glicerofosfato. En la base de datos de proteínas ortólogas (COG) [86], la familia de la PgsA se encuentra agrupada en el COG0558, por lo que se considera que todas las secuencias en la familia llevan a cabo la misma función. PgsA forma parte de una familia de proteínas que utilizan CDP-DAG como precursor para diferentes compuestos, entre las que se encuentran la fosfatidilcolina sintasa (Pcs), fosfatidilserina sintasa (Pss) y la fosfatidilinositol sintasa (Pis) [82].

Utilizando las relaciones de bidireccionalidad que se establecen a nivel de secuencias entre las proteínas del COG0558, se dividió a la familia en dos subgrupos empleando el programa Markov Cluster Algorithm (MCL) (<http://micans.org/mcl/>), [89]. El análisis de los dos grupos obtenidos, nos mostró que uno de ellos estaba constituido por secuencias que contienen un motivo de aminoácidos específico para la familia PgsA, mientras que el otro subgrupo presentó secuencias con el motivo de la familia de proteínas Pis. Debido al incremento de genomas secuenciados en su totalidad y la necesidad de contar con herramientas más precisas de anotación funcional, desarrollaremos un servidor web para poner a disposición del usuario las subfamilias encontradas por nuestras divisiones con MCL. Dicho servidor pretende servir como herramienta de análisis de subfamilias de proteínas surgidas por duplicación,

donde a parte de divergir en la regulación, pueden divergir en su función. Como parte de las estrategias a seguir se encuentran las siguientes:

- División de familias de proteínas ortólogas utilizando los Bbt y el programa MCL
- Caracterización de las subfamilias utilizando perfiles hechos con MEME [6] y HMMER [22]
- Elaboración de un servidor web que ponga a disposición del usuario las subfamilias encontradas.

Referencias

- [1] ABHIMAN, S., AND SONNHAMMER, E. L. L. Funshift: a database of function shift analysis on protein subfamilies. *Nucleic Acids Res* 33, Database issue (Jan 2005), D197–D200.
- [2] ALM, E., HUANG, K., AND ARKIN, A. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLoS Comput Biol* 2, 11 (Nov 2006), e143.
- [3] ALTSCHUL, S. F., MADDEN, T. L., SCHAFFER, A. A., ZHANG, J., ZHANG, Z., MILLER, W., AND LIPMAN, D. J. Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic Acids Res* 25, 17 (Sep 1997), 3389–3402.
- [4] BABU, M. M. Early career research award lecture. structure, evolution and dynamics of transcriptional regulatory networks. *Biochem Soc Trans* 38, 5 (Oct 2010), 1155–1178.
- [5] BABU, M. M., TEICHMANN, S. A., AND ARAVIND, L. Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358, 2 (Apr 2006), 614–633.
- [6] BAILEY, T. L., AND ELKAN, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2 (1994), 28–36.
- [7] BALAZSI, G., BARABASI, A.-L., AND OLTVAI, Z.Ñ. Topological units of environmental signal processing in the transcriptional regulatory network of escherichia coli. *Proc Natl Acad Sci U S A* 102, 22 (May 2005), 7841–7846.
- [8] BALAZSI, G., HEATH, A. P., SHI, L., AND GENNARO, M. L. The temporal response of the mycobacterium tuberculosis gene regulatory network during growth arrest. *Mol Syst Biol* 4 (2008), 225.
- [9] BALAZSI, G., AND OLTVAI, Z.Ñ. Sensing your surroundings: how transcription-regulatory networks of the cell discern environmental signals. *Sci STKE* 2005, 282 (May 2005), pe20.
- [10] BARABASI, A.-L., AND OLTVAI, Z.Ñ. Network biology: understanding the cell’s functional organization. *Nat Rev Genet* 5, 2 (Feb 2004), 101–113.

- [11] BAUSCH, C., RAMSEY, M., AND CONWAY, T. Transcriptional organization and regulation of the l-idonic acid pathway (gntII system) in *Escherichia coli*. *J Bacteriol* 186, 5 (Mar 2004), 1388–1397.
- [12] BERGTHORSSON, U., ANDERSSON, D. I., AND ROTH, J. R. Ohno's dilemma: evolution of new genes under continuous selection. *Proc Natl Acad Sci U S A* 104, 43 (Oct 2007), 17004–17009.
- [13] BORDES, P., CONTER, A., MORALES, V., BOUVIER, J., KOLB, A., AND GUTIERREZ, C. Dna supercoiling contributes to disconnect sigma accumulation from sigma-dependent transcription in *Escherichia coli*. *Mol Microbiol* 48, 2 (Apr 2003), 561–571.
- [14] BRANDENBURG, J. L., WRAY, L. V., BEIER, L., JARMER, H., SAXILD, H. H., AND FISHER, S. H. Roles of pucr, glhr, and tnra in regulating expression of the *Bacillus subtilis* ure p3 promoter. *J Bacteriol* 184, 21 (Nov 2002), 6060–6064.
- [15] BRATLIE, M. S., JOHANSEN, J., SHERMAN, B. T., HUANG, D. W., LEMPICKI, R. A., AND DRABLOS, F. Gene duplications in prokaryotes can be associated with environmental adaptation. *BMC Genomics* 11 (2010), 588.
- [16] BROWNING, D. F., AND BUSBY, S. J. The regulation of bacterial transcription initiation. *Nat Rev Microbiol* 2, 1 (Jan 2004), 57–65.
- [17] CHANGELA, A., CHEN, K., XUE, Y., HOLSCHEIN, J., OUTTEN, C. E., O'HALLORAN, T. V., AND MONDRAGON, A. Molecular basis of metal-ion selectivity and zeptomolar sensitivity by cuer. *Science* 301, 5638 (Sep 2003), 1383–1387.
- [18] CHEUNG, K. J., BADARINARAYANA, V., SELINGER, D. W., JANSE, D., AND CHURCH, G. M. A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res* 13, 2 (Feb 2003), 206–215.
- [19] COMMICHAU, F. M., AND STULKE, J. Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression. *Mol Microbiol* 67, 4 (Feb 2008), 692–702.

- [20] CONANT, G. C., AND WAGNER, A. Genomehistory: a software tool and its application to fully sequenced genomes. *Nucleic Acids Res* 30, 15 (Aug 2002), 3378–3386.
- [21] CONANT, G. C., AND WOLFE, K. H. Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9, 12 (Dec 2008), 938–950.
- [22] EDDY, S. R. Profile hidden markov models. *Bioinformatics* 14, 9 (1998), 755–763.
- [23] EDDY, S. R. A memory-efficient dynamic programming algorithm for optimal alignment of a sequence to an rna secondary structure. *BMC Bioinformatics* 3 (Jul 2002), 18.
- [24] ENTUS, R., AUFDERHEIDE, B., AND SAURO, H. M. Design and implementation of three incoherent feed-forward motif based biological concentration sensors. *Syst Synth Biol* 1, 3 (Aug 2007), 119–128.
- [25] EVANGELISTI, A. M., AND WAGNER, A. Molecular evolution in the yeast transcriptional regulation network. *J Exp Zool B Mol Dev Evol* 302 (2004), 392–411.
- [26] GAMA-CASTRO, S., JIMENEZ-JACINTO, V., PERALTA-GIL, M., SANTOS-ZAVALA, A., PENALOZA-SPINOLA, M. I., CONTRERAS-MOREIRA, B., SEGURA-SALAZAR, J., MUNIZ-RASCADO, L., MARTINEZ-FLORES, I., SALGADO, H., BONAVIDES-MARTINEZ, C., ABREU-GOODGER, C., RODRIGUEZ-PENAGOS, C., MIRANDA-RIOS, J., MORETT, E., MERINO, E., HUERTA, A. M., TREVINO-QUINTANILLA, L., AND COLLADO-VIDES, J. Regulondb (version 6.0): gene regulation model of escherichia coli k-12 beyond transcription, active (experimental) annotated promoters and textpresso navigation. *Nucleic Acids Res* 36, Database issue (Jan 2008), D120–D124.
- [27] GARDNER, P. P., DAUB, J., TATE, J. G., NAWROCKI, E. P., KOLBE, D. L., LINDGREEN, S., WILKINSON, A. C., FINN, R. D., GRIFFITHS-JONES, S., EDDY, S. R., AND BATEMAN, A. Rfam: updates to the rna families database. *Nucleic Acids Res* 37, Database issue (Jan 2009), D136–D140.

- [28] GELFAND, M. S. Evolution of transcriptional regulatory networks in microbial genomes. *Curr Opin Struct Biol* 16, 3 (Jun 2006), 420–429.
- [29] GEVERS, D., VANDEPOELE, K., SIMILLON, C., AND DE PEER, Y. V. Gene duplication and biased functional retention of paralogs in bacterial genomes. *Trends Microbiol* 12, 4 (Apr 2004), 148–154.
- [30] GOTTESMAN, S., MCCULLEN, C. A., GUILLIER, M., VANDERPOOL, C. K., MAJDALANI, N., BENHAMMOU, J., THOMPSON, K. M., FITZGERALD, P. C., SOWA, N. A., AND FITZGERALD, D. J. Small rna regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol* 71 (2006), 1–11.
- [31] GRAU, R., GARDIOL, D., GLIKIN, G. C., AND DE MENDOZA, D. Dna supercoiling and thermal regulation of unsaturated fatty acid synthesis in bacillus subtilis. *Mol Microbiol* 11, 5 (Mar 1994), 933–941.
- [32] GRIFFITHS-JONES, S., MOXON, S., MARSHALL, M., KHANNA, A., EDDY, S. R., AND BATEMAN, A. Rfam: annotating non-coding rnas in complete genomes. *Nucleic Acids Res* 33, Database issue (Jan 2005), D121–D124.
- [33] GU, X. Evolution of duplicate genes versus genetic robustness against null mutations. *Trends Genet* 19, 7 (Jul 2003), 354–356.
- [34] GU, X., ZHANG, Z., AND HUANG, W. Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A* 102, 3 (Jan 2005), 707–712.
- [35] GU, Z., NICOLAE, D., LU, H. H.-S., AND LI, W. H. Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18, 12 (Dec 2002), 609–613.
- [36] GUELZIM, N., BOTTANI, S., BOURGINE, P., AND KEPES, F. Topological and causal structure of the yeast transcriptional regulatory network. *Nat Genet* 31, 1 (May 2002), 60–63.
- [37] GUILLIER, M., GOTTESMAN, S., AND STORZ, G. Modulating the outer membrane with small rnas. *Genes Dev* 20, 17 (Sep 2006), 2338–2348.

- [38] HATFIELD, G. W., AND BENHAM, C. J. Dna topology-mediated control of global gene expression in *escherichia coli*. *Annu Rev Genet* 36 (2002), 175–203.
- [39] HE, X., AND ZHANG, J. Gene complexity and gene duplicability. *Curr Biol* 15, 11 (Jun 2005), 1016–1021.
- [40] HE, Z., LAVOIE, K. D. S., BARTLETT, P. A., AND TONEY, M. D. Conservation of mechanism in three chorismate-utilizing enzymes. *J Am Chem Soc* 126, 8 (Mar 2004), 2378–2385.
- [41] HECKER, M., PANE-FARRE, J., AND VOLKER, U. Sigb-dependent general stress response in *bacillus subtilis* and related gram-positive bacteria. *Annu Rev Microbiol* 61 (2007), 215–236.
- [42] HENGGE-ARONIS, R. Signal transduction and regulatory mechanisms involved in control of the sigma(s) (σ^s) subunit of rna polymerase. *Microbiol Mol Biol Rev* 66, 3 (Sep 2002), 373–95.
- [43] HSIEH, L. S., BURGER, R. M., AND DRLICA, K. Bacterial dna supercoiling and [atp]/[adp] changes associated with a transition to anaerobic growth. *J Mol Biol* 219, 3 (Jun 1991), 443–450.
- [44] IKEDA, M. Towards bacterial strains overproducing l-tryptophan and other aromatics by metabolic engineering. *Appl Microbiol Biotechnol* 69, 6 (Feb 2006), 615–626.
- [45] INNAN, H., AND KONDRASHOV, F. The evolution of gene duplications: classifying and distinguishing between models. *Nat Rev Genet* 11, 2 (Feb 2010), 97–108.
- [46] ISHIHAMA, A. Prokaryotic genome regulation: multifactor promoters, multitarget regulators and hierarchic networks. *FEMS Microbiol Rev* 34, 5 (Sep 2010), 628–645.
- [47] IZU, H., ADACHI, O., AND YAMADA, M. Gene organization and transcriptional regulation of the *gntrku* operon involved in gluconate uptake and catabolism of *escherichia coli*. *J Mol Biol* 267, 4 (Apr 1997), 778–793.
- [48] JANGA, S. C., AND COLLADO-VIDES, J. Structure and evolution of gene regulatory networks in microbial genomes. *Res Microbiol* 158, 10 (Dec 2007), 787–794.

- [49] KANEHISA, M., GOTO, S., KAWASHIMA, S., OKUNO, Y., AND HATTORI, M. The kegg resource for deciphering the genome. *Nucleic Acids Res* 32, Database issue (Jan 2004), D277–D280.
- [50] KONDRASHOV, F. A., AND KONDRASHOV, A. S. Role of selection in fixation of gene duplications. *J Theor Biol* 239, 2 (Mar 2006), 141–151.
- [51] KYRIAKIDIS, D. A., AND TILIGADA, E. Signal transduction and adaptive regulation through bacterial two-component systems: the escherichia coli atosc paradigm. *Amino Acids* 37, 3 (Sep 2009), 443–458.
- [52] LIOLIOU, E. E., MIMITOU, E. P., GRIGOROUDIS, A. I., PANAGIOTIDIS, C. H., PANAGIOTIDIS, C. A., AND KYRIAKIDIS, D. A. Phosphorylation activity of the response regulator of the two-component signal transduction system atos-atoc in e. coli. *Biochim Biophys Acta* 1725, 3 (Oct 2005), 257–268.
- [53] LUSCOMBE, N. M., BABU, M. M., YU, H., SNYDER, M., TEICHMANN, S. A., AND GERSTEIN, M. Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* 431, 7006 (Sep 2004), 308–312.
- [54] LYNCH, M., AND FORCE, A. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154, 1 (Jan 2000), 459–473.
- [55] MARTINEZ-ANTONIO, A., AND COLLADO-VIDES, J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* 6 (2003), 482–489.
- [56] MARTINEZ-ANTONIO, A., JANGA, S. C., SALGADO, H., AND COLLADO-VIDES, J. Internal-sensing machinery directs the activity of the regulatory network in escherichia coli. *Trends Microbiol* 14, 1 (Jan 2006), 22–27.
- [57] MARTINEZ-NUNEZ, M. A., PEREZ-RUEDA, E., GUTIERREZ-RIOS, R. M., AND MERINO, E. New insights into the regulatory networks of paralogous genes in bacteria. *Microbiology* 156, Pt 1 (Jan 2010), 14–22.

- [58] MASLOV, S., SNEPPEN, K., ERIKSEN, K. A., AND YAN, K.-K. Upstream plasticity and downstream robustness in evolution of molecular networks. *BMC Evol Biol* 4 (Mar 2004), 9.
- [59] MIRA, A., PUSHKER, R., AND RODRIGUEZ-VALERA, F. The neolithic revolution of bacterial genomes. *Trends Microbiol* 14, 5 (May 2006), 200–206.
- [60] MITTENHUBER, G. An inventory of genes encoding rna polymerase sigma factors in 31 completely sequenced eubacterial genomes. *J Mol Microbiol Biotechnol* 4, 1 (Jan 2002), 77–91.
- [61] MORENO-CAMPUZANO, S., JANGA, S., AND PEREZ-RUEDA, E. Identification and analysis of dna-binding transcription factors in bacillus subtilis and other firmicutes a genomic approach. *BMC Genomics* 7 (2006), 147.
- [62] MUTALIK, V. K., SHAH, P., AND VENKATESH, K. V. Allosteric interactions and bifunctionality make the response of glutamine synthetase cascade system of escherichia coli robust and ultrasensitive. *J Biol Chem* 278, 29 (Jul 2003), 26327–26332.
- [63] NAHUM, L. A., AND RILEY, M. Divergence of function in sequence-related groups of escherichia coli proteins. *Genome Res* 11, 8 (Aug 2001), 1375–1381.
- [64] NARBERHAUS, F., WALDMINGHAUS, T., AND CHOWDHURY, S. Rna thermometers. *FEMS Microbiol Rev* 30, 1 (Jan 2006), 3–16.
- [65] NEI, M., ROGOZIN, I. B., AND PIONTKIVSKA, H. Purifying selection and birth-and-death evolution in the ubiquitin gene family. *Proc Natl Acad Sci U S A* 97, 20 (Sep 2000), 10866–10871.
- [66] NINFA, A. J., AND MAGASANIK, B. Covalent modification of the glng product, nri, by the glnl product, nrii, regulates the transcription of the glngl operon in escherichia coli. *Proc Natl Acad Sci U S A* 83, 16 (Aug 1986), 5909–5913.
- [67] OGASAWARA, H., HASEGAWA, A., KANDA, E., MIKI, T., YAMAMOTO, K., AND ISHIHAMA, A. Genomic selex search for target promoters under the control of the phoqp-rstba signal relay cascade. *J Bacteriol* 189, 13 (Jul 2007), 4791–4799.

- [68] PAPP, B., PAL, C., AND HURST, L. D. Evolution of cis-regulatory elements in duplicated genes of yeast. *Trends Genet* 19, 8 (Aug 2003), 417–422.
- [69] PENNELLA, M. A., AND GIEDROC, D. P. Structural determinants of metal selectivity in prokaryotic metal-responsive transcriptional regulators. *Biometals* 18, 4 (Aug 2005), 413–428.
- [70] PEREZ-RUEDA, E., AND COLLADO-VIDES, J. Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol* 53, 3 (Sep 2001), 172–179.
- [71] PEREZ-RUEDA, E., AND JANGA, S. C. Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol* 27, 6 (Jun 2010), 1449–1459.
- [72] PRAGAI, Z., ALLENBY, N. E. E., O’CONNOR, N., DUBRAC, S., RAPOPORT, G., MSADEK, T., AND HARWOOD, C. R. Transcriptional regulation of the phopr operon in bacillus subtilis. *J Bacteriol* 186, 4 (Feb 2004), 1182–1190.
- [73] PRATT, L. A., HSING, W., GIBSON, K. E., AND SILHAVY, T. J. From acids to osmz: multiple factors influence synthesis of the ompf and ompc porins in escherichia coli. *Mol Microbiol* 20, 5 (Jun 1996), 911–917.
- [74] PUSHKER, R., MIRA, A., AND RODRIGUEZ-VALERA, F. Comparative genomics of gene-family size in closely related bacteria. *Genome Biol* 5, 4 (2004), R27.
- [75] SANCHEZ-PEREZ, G., MIRA, A., NYIRO, G., PASIĆ, L., AND RODRIGUEZ-VALERA, F. Adapting to environmental changes using specialized paralogs. *Trends Genet* 24, 4 (Apr 2008), 154–158.
- [76] SAYERS, E. W., BARRETT, T., BENSON, D. A., BOLTON, E., BRYANT, S. H., CANESE, K., CHETVERNIN, V., CHURCH, D. M., DICUCCIO, M., FEDERHEN, S., FEOLO, M., GEER, L. Y., HELMBERG, W., KAPUSTIN, Y., LANDSMAN, D., LIPMAN, D. J., LU, Z., MADDEN, T. L., MADEJ, T., MAGLOTT, D. R., MARCHLER-BAUER, A., MILLER, V., MIZRACHI, I., OSTELL, J., PANCHENKO, A., PRUITT, K. D., SCHULER, G. D., SEQUEIRA, E., SHERRY, S. T.,

- SHUMWAY, M., SIROTKIN, K., SLOTTA, D., SOUVOROV, A., STARCHENKO, G., TATUSOVA, T. A., WAGNER, L., WANG, Y., WILBUR, W. J., YASCHENKO, E., AND YE, J. Database resources of the national center for biotechnology information. *Nucleic Acids Res* 38, Database issue (Jan 2010), D5–16.
- [77] SCHILLING, O., HERZBERG, C., HERTRICH, T., VORSMANN, H., JESSEN, D., HUBNER, S., TITGEMEYER, F., AND STULKE, J. Keeping signals straight in transcription regulation: specificity determinants for the interaction of a family of conserved bacterial rna-protein couples. *Nucleic Acids Res* 34, 21 (2006), 6102–6115.
- [78] SERRES, M. H., KERR, A. R. W., MCCORMACK, T. J., AND RILEY, M. Evolution by leaps: gene duplication in bacteria. *Biol Direct* 4 (2009), 46.
- [79] SHAKHNOVICH, B. E., AND KOONIN, E. V. Origins and impact of constraints in evolution of gene families. *Genome Res* 16, 12 (Dec 2006), 1529–1536.
- [80] SHEN-ORR, S. S., MILO, R., MANGAN, S., AND ALON, U. Network motifs in the transcriptional regulation network of escherichia coli. *Nat Genet* 31, 1 (May 2002), 64–68.
- [81] SIERRA, N., MAKITA, Y., DE HOON, M., AND NAKAI, K. Dbtbs: a database of transcriptional regulation in bacillus subtilis containing upstream intergenic conservation information. *Nucleic Acids Res* 36, Database issue (Jan 2008), D93–D96.
- [82] SOHLENKAMP, C., LOPEZ-LARA, I. M., AND GEIGER, O. Biosynthesis of phosphatidylcholine in bacteria. *Prog Lipid Res* 42, 2 (Mar 2003), 115–162.
- [83] SONNENSCHNIG, N., GEERTZ, M., MUSKHELISHVILI, G., AND HUTT, M.-T. Analog regulation of metabolic demand. *BMC Syst Biol* 5 (2011), 40.
- [84] SOREK, R., ZHU, Y., CREEVEY, C. J., FRANCINO, M. P., BORK, P., AND RUBIN, E. M. Genome-wide experimental determination of barriers to horizontal gene transfer. *Science* 318, 5855 (Nov 2007), 1449–1452.

- [85] STOYANOV, J. V., HOBMAN, J. L., AND BROWN, N. L. Cuer (ybbi) of escherichia coli is a merr family regulator controlling expression of the copper exporter copA. *Mol Microbiol* 39, 2 (Jan 2001), 502–511.
- [86] TATUSOV, R. L., KOONIN, E. V., AND LIPMAN, D. J. A genomic perspective on protein families. *Science* 278, 5338 (Oct 1997), 631–637.
- [87] TEICHMANN, S. A., AND BABU, M. M. Gene regulatory network growth by duplication. *Nat Genet* 36, 5 (May 2004), 492–496.
- [88] THIEFFRY, D., HUERTA, A. M., PEREZ-RUEDA, E., AND COLLADO-VIDES, J. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in escherichia coli. *Bioessays* 20, 5 (May 1998), 433–440.
- [89] VAN DONGEN, S. *Graph Clustering by Flow Simulation*. PhD thesis, University of Utrecht, 2000.
- [90] WANNER, B. L. Is cross regulation by phosphorylation of two-component response regulator proteins important in bacteria? *J Bacteriol* 174, 7 (Apr 1992), 2053–2058.
- [91] WARD, R., AND DURRETT, R. Subfunctionalization: How often does it occur? how long does it take? *Theor Popul Biol* 66, 2 (Sep 2004), 93–100.
- [92] WHITWORTH, D. E., AND COCK, P. J. A. Evolution of prokaryotic two-component systems: insights from comparative genomics. *Amino Acids* 37, 3 (Sep 2009), 459–466.
- [93] WINKLER, W. C. Riboswitches and the role of noncoding rnas in bacterial metabolic control. *Curr Opin Chem Biol* 9, 6 (Dec 2005), 594–602.
- [94] WINKLER, W. C., AND BREAKER, R. R. Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* 59 (2005), 487–517.
- [95] WOSTEN, M. M. Eubacterial sigma-factors. *FEMS Microbiol Rev* 22, 3 (Sep 1998), 127–150.

- [96] YAMAMOTO, K., AND ISHIHAMA, A. Transcriptional response of escherichia coli to external copper. *Mol Microbiol* 56, 1 (Apr 2005), 215–227.
- [97] YAMAMOTO, K., AND ISHIHAMA, A. Transcriptional response of escherichia coli to external zinc. *J Bacteriol* 187, 18 (Sep 2005), 6333–6340.
- [98] ZALIECKAS, J. M., WRAY, L. V., AND FISHER, S. H. Cross-regulation of the bacillus subtilis glnra and tnra genes provides evidence for dna binding site discrimination by glnr and tnra. *J Bacteriol* 188, 7 (Apr 2006), 2578–2585.
- [99] ZHANG, J. Evolution by gene duplication: an update. *Trends in Ecology & Evolution* 18 (2003), 292–298.
- [100] ZHANG, J., ROSENBERG, H. F., AND NEI, M. Positive darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci* 95 (1998), 3708–3713.
- [101] ZHOU, D., AND YANG, R. Global analysis of gene transcription regulation in prokaryotes. *Cell Mol Life Sci* 63, 19-20 (Oct 2006), 2260–2290.
- [102] ZIMMER, D. P., SOUPENE, E., LEE, H. L., WENDISCH, V. F., KHODURSKY, A. B., PETER, B. J., BENDER, R. A., AND KUSTU, S. Nitrogen regulatory protein c-controlled genes of escherichia coli: scavenging as a defense against nitrogen limitation. *Proc Natl Acad Sci* 97 (2000), 14674–14679.

8. Suplementos

8.1. Figura

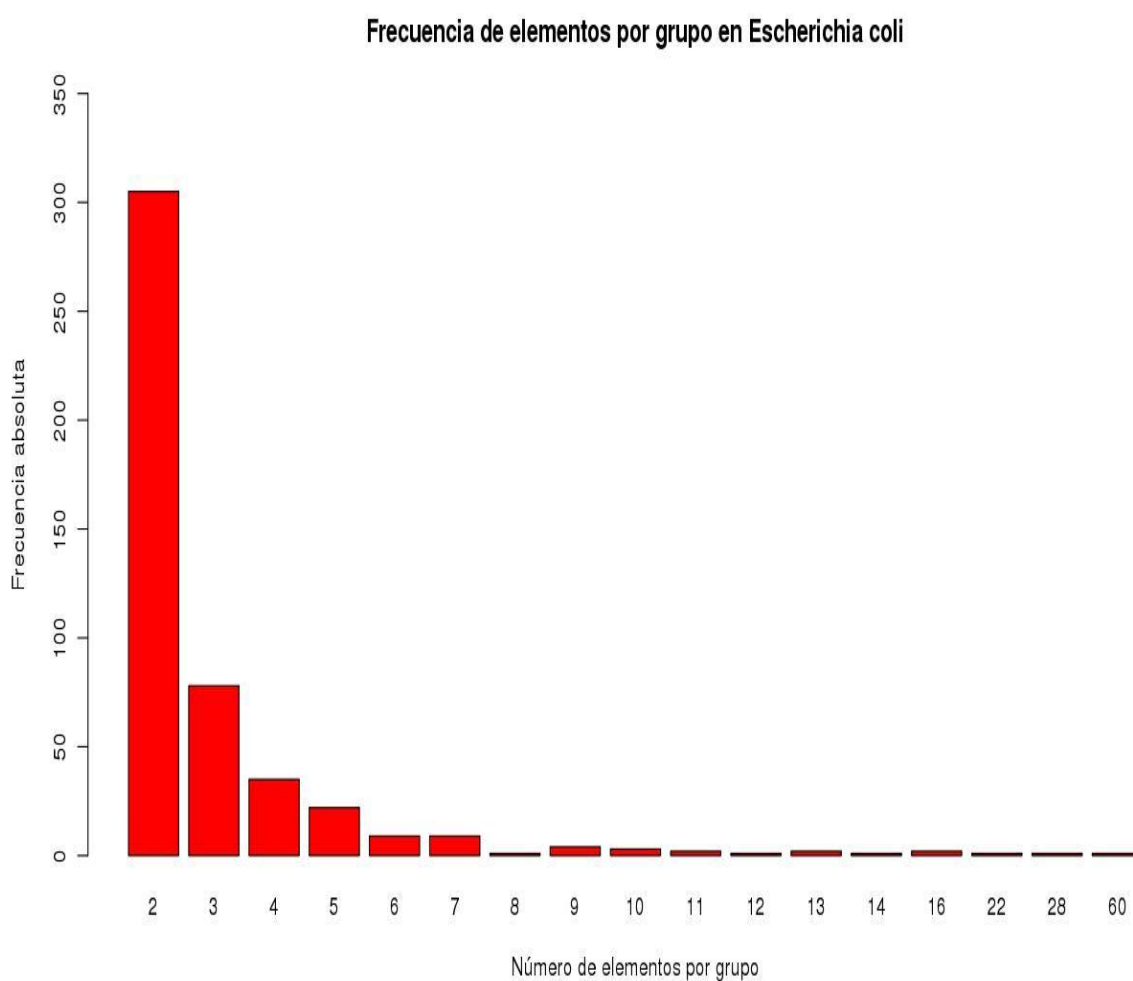


Figura 12: Frecuencia absoluta del número de elementos (parálogos) por grupo en *E. coli*. El 92% de los grupos están conformados de dos a cinco parálogos, mientras que el restante 8% presentan de seis a más copias por grupo.

Cuadro 2: Valores de las frecuencias absolutas de los números de parálogos en cada uno de los 477 grupos definidos en *E. coli* y en los 485 grupos definidos en *B. subtilis*.

Número de parálogos	Frecuencia en <i>E. coli</i>	Frecuencia en <i>B. subtilis</i>
2	305	311
3	78	94
4	35	31
5	22	10
6	9	7
7	9	12
8	1	4
9	4	4
10	3	1
11	2	4
12	1	1
13	2	0
14	1	0
15	0	1
16	2	1
19	0	1
22	1	0
28	1	1
29	0	1
57	0	1
60	1	0

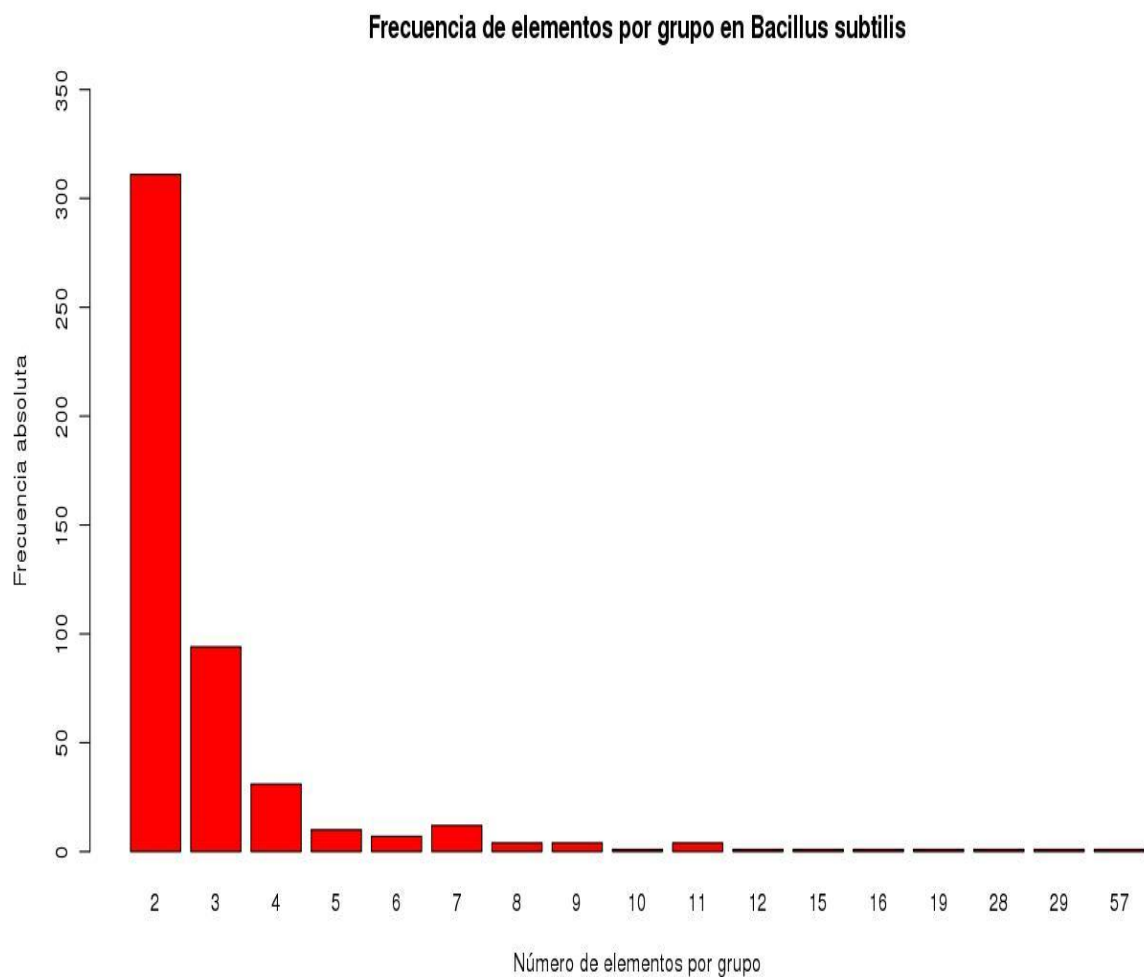


Figura 13: Frecuencia absoluta del número de elementos (parálogos) por grupo en *B. subtilis*.

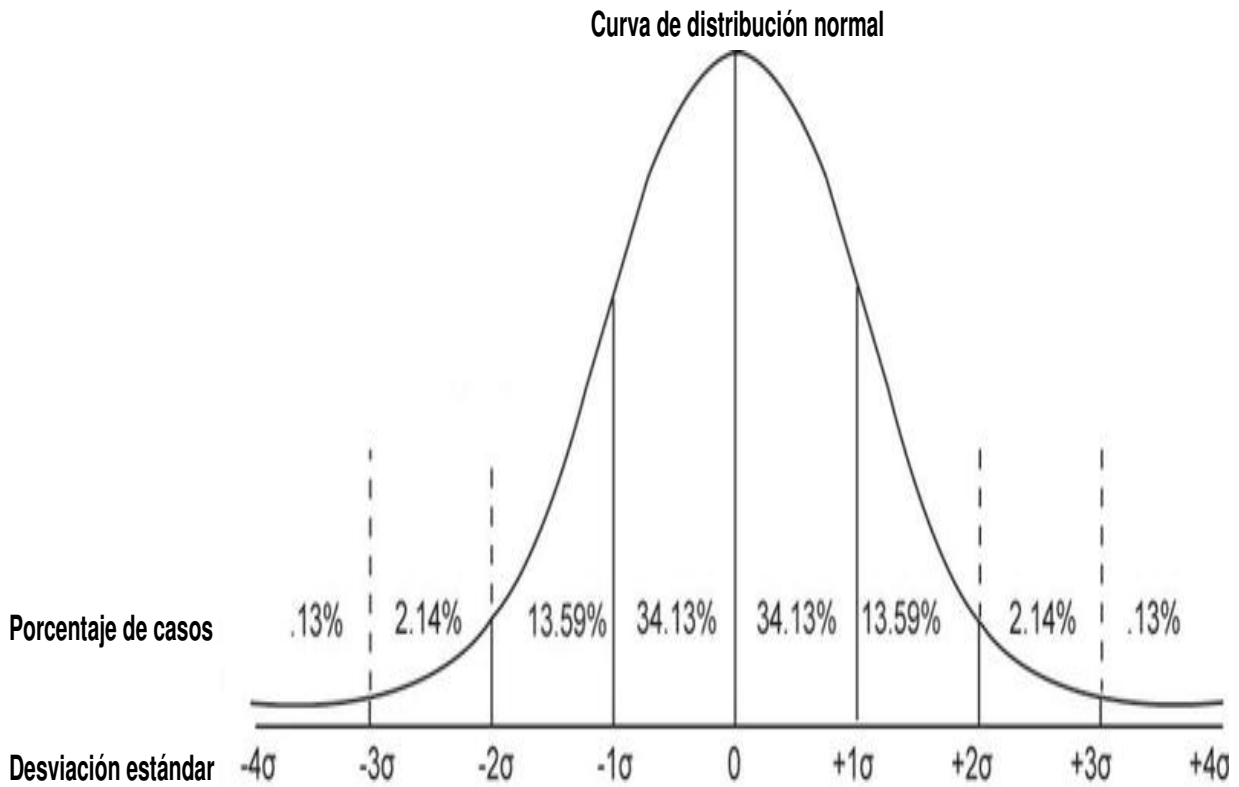


Figura 14: Curva de distribución normal. Se muestra el área bajo la curva correspondiente a 1 y 4 desviaciones estándar, así como el porcentaje de casos que se encuentran en ellas.

8.2. Artículo

New insights into the regulatory networks of paralogous genes in bacteria. 2010. Martínez-Núñez Mario Alberto, Pérez-Rueda E, Gutierrez-Rios RM, Merino E. *Microbiology* 156, Pt 1 (Jan 2010), 14-22.

Review

New insights into the regulatory networks of paralogous genes in bacteria

Mario A. Martínez-Núñez,¹ Ernesto Pérez-Rueda,²
Rosa María Gutiérrez-Ríos¹ and Enrique Merino¹

Correspondence

Enrique Merino

merino@ibt.unam.mx

¹Departamento de Microbiología Molecular, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

²Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, Mexico

Extensive genomic studies on gene duplication in model organisms such as *Escherichia coli* and *Saccharomyces cerevisiae* have recently been undertaken. In these models, it is commonly considered that a duplication event may include a transcription factor (TF), a target gene, or both. Following a gene duplication episode, varying scenarios have been postulated to describe the evolution of the regulatory network. However, in most of these, the TFs have emerged as the most important and in some cases the only factor shaping the regulatory network as the organism responds to a natural selection process, in order to fulfil its metabolic needs. Recent findings concerning the regulatory role played by elements other than TFs have indicated the need to reassess these early models. Thus, we performed an exhaustive review of paralogous gene regulation in *E. coli* and *Bacillus subtilis* based on published information, available in the NCBI PubMed database and in well-established regulatory databases. Our survey reinforces the notion that despite TFs being the most prominent components shaping the regulatory networks, other elements are also important. These include small RNAs, riboswitches, RNA-binding proteins, sigma factors, protein–protein interactions and DNA supercoiling, which modulate the expression of genes involved in particular metabolic processes or induce a more complex response in terms of the regulatory networks of paralogous genes in an integrated interplay with TFs.

Introduction

Gene duplication is one of the main sources of functional divergence in organisms (Babu *et al.*, 2004; Conant & Wolfe, 2008; Gelfand, 2006; Lynch & Conery, 2000; Teichmann & Babu, 2004). In order to fulfil an organism's metabolic needs, selection processes modify the regulation of the paralogous gene copies, taking advantage of a repertoire of new functions. Duplication events may include genes coding for transcription factors (TFs), permitting a more versatile adaptation of the functional diversity gained from the duplication of structural genes. Different aspects of the evolution of the regulatory networks of paralogous genes have been examined, including the co-evolution of the upstream regulatory regions and their corresponding TFs; the likely consequences of gain, loss, and replacement of TFs in the regulatory networks of paralogous genes (Gelfand, 2006; Teichmann & Babu, 2004); and also the topological and dynamic properties of the regulatory networks (Babu *et al.*, 2006; Balaji *et al.*, 2007; Luscombe *et al.*, 2004). This review

will consider the fascinating repertoire of additional mechanisms other than TFs, which help regulate the expression of paralogous genes in *Escherichia coli* and *Bacillus subtilis*. This analysis is derived from an extensive compilation of the regulatory information available from the NCBI PubMed database and deposited in the *E. coli* RegulonDB Release 6.3 (Gama-Castro *et al.*, 2008), *B. subtilis* DBTBS (Sierro *et al.*, 2008) and Rfam version 9.1 (Gardner *et al.*, 2009) databases.

TFs are essential elements in the regulatory networks of paralogous genes

TFs are the most prominent and usually the only regulatory element taken into account in any of the aforementioned studies because of the important role they play in the regulatory networks of paralogous genes. Teichmann & Babu (2004) formulated a widely accepted evolutionary model portraying the regulation of duplicated genes. In this model, the loss and gain of regulatory interactions may occur following the duplication of either a TF or a target gene, or following the duplication of both a TF and a target gene (Fig. 1a). The Teichmann and Babu model does not

Four supplementary tables are available with the online version of this paper.

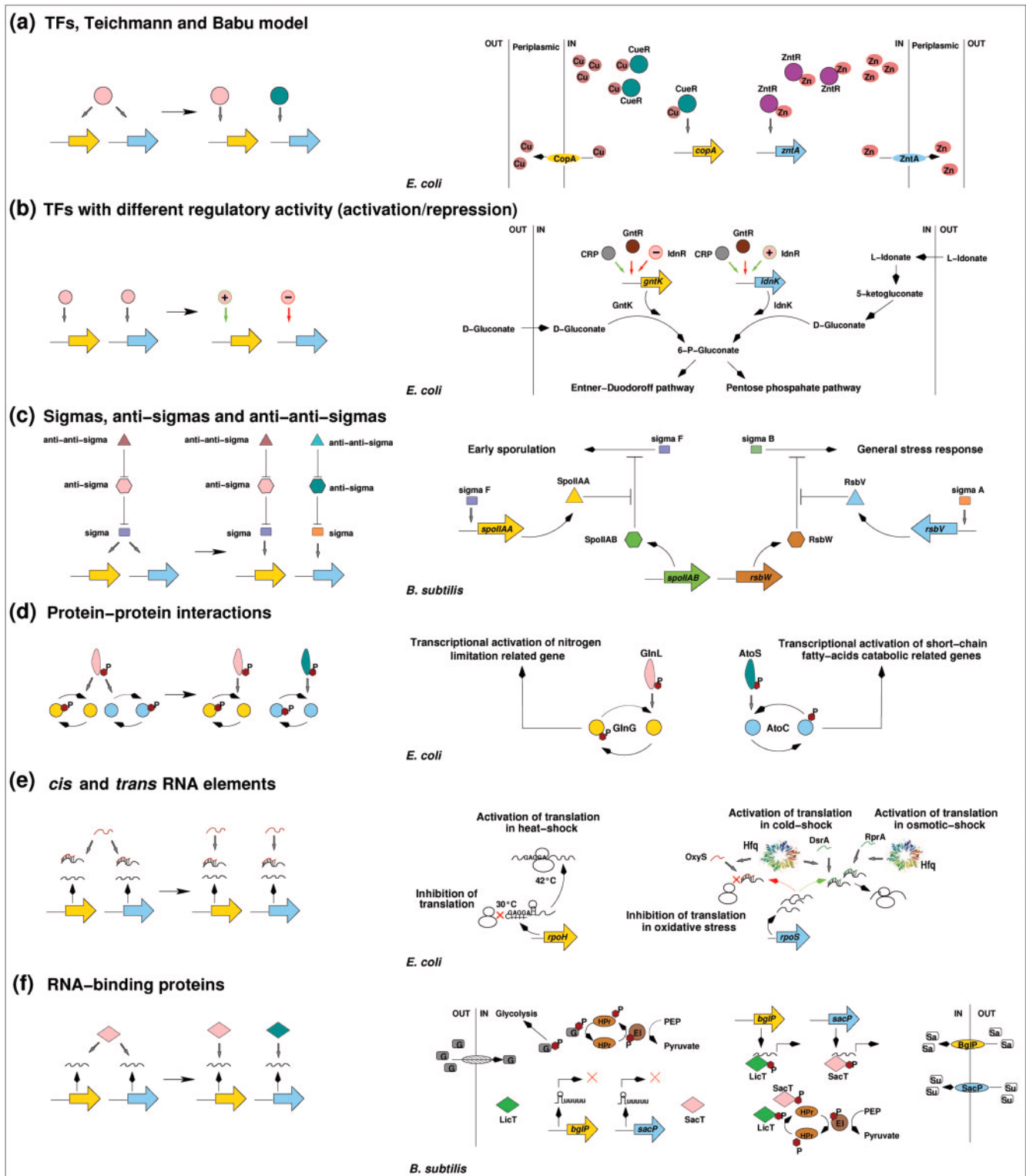
question whether the regulatory effect (either activation or repression) of the TFs on the paralogous genes is conserved or changed in any of these instances; it only considers the regulatory interactions. An example of a regulatory conserving effect is found in the *E. coli* paralogous genes *zntA* and *copA* (Group 393, Supplementary Tables S1 and S3), coding for zinc and copper transporters, which are regulated by the paralogous TFs ZntR and CueR, respectively (Yamamoto & Ishihama, 2005a, b) (Fig. 1a). We also identified examples of regulatory systems where the TF is maintained but a different effect is gained (either activation or repression) in one of the paralogous genes. This evolutionary scenario can be observed in the regulation of the *E. coli* *gntK* and *idnK* gluconate kinase genes, involved in the 6-phosphogluconate synthesis which takes place in the Entner–Doudoroff and pentose phosphate pathways, respectively. Although these genes are regulated by the same TFs, CRP, GntR and IdnR, the last TF represses the transcription of *gntK*, whereas it activates the transcription of *idnK* (Bausch *et al.*, 2004; Izu *et al.*, 1997) (see Fig. 1b and Group 449 of Supplementary Tables S1 and S3).

In order to quantify the divergence in the regulation of paralogous genes in *E. coli* and *B. subtilis*, BLAST comparisons (Altschul *et al.*, 1990) of their corresponding protein sequence sets were carried out. Based on these comparisons, we identified 477 groups of paralogous genes in *E. coli*. From this set, 291 groups of paralogous genes were found to be regulated by at least one TF, and for 43 of these (14.8%) the TFs were shown to have a dual activation/repression effect on their corresponding paralogous target genes (Supplementary Table S3). In *B. subtilis* this tendency was considerably less. Out of the 483 groups of paralogous genes, we found that 225 groups were regulated by at least one TF, and eight of these (3.6%) presented at least one TF with a dual activation/repression regulatory activity (Supplementary Table S4). In *E. coli*, this dual regulatory activity is more frequently found in the groups of paralogous genes regulated by global TFs: ArcA, CRP, FIS, FNR, Hns, IHF and Lrp (Martinez-Antonio & Collado-Vides, 2003) (see Supplementary Table S3). In contrast, only some of the *B. subtilis* global TFs such as AbrB, CcpA, Spo0A and TnrA (Moreno-Campuzano *et al.*, 2006) exhibit a dual activation/repression activity, whereas others either act always as repressors (CodY, Fur, LexA and YrzC), or always as activators (ComK) (see Supplementary Table S4). Finally, a smaller proportion of local TFs show dual activation/repression regulatory activity on paralogous genes. This is the case for the *E. coli* NtrC transcription regulator, which represses the transcription of the glutamine transporter subunit *glnH* gene and activates the transcription of the histidine/lysine/arginine/ornithine transporter *hisJ* and the lysine/arginine/ornithine transporter *argT* genes (Zimmer *et al.*, 2000) (Group 27 of Supplementary Table S1).

In order to test the statistical significance of finding a regulatory element (TFs, sigma factors, protein–protein

interactions, DNA supercoiling, RNAs and RNA-binding proteins) that regulates two or more paralogous genes in a regulatory network, we carried out a computer simulation procedure, similar to the one performed by Teichmann & Babu (2004), where the regulatory interactions were randomly shuffled. Having performed the randomization process 10 000 times, we compared the mean values of these simulations against the number of interactions in the real regulatory networks of *E. coli* and *B. subtilis* (531 and 189 interactions, respectively). These real interactions were found to be significantly different from the average interactions of their corresponding randomized *E. coli* and *B. subtilis* datasets (19.5 and 19 standard deviations, SDS, respectively). The statistical significance of two or more paralogous genes being regulated by a similar TF, but having a dissimilar regulation (activation vs repression), was also evaluated. In this case, the distance of the real values from the mean of randomly generated networks in *E. coli* and *B. subtilis* was 3.7 and 2 SDS, respectively. Interestingly, when this analysis was performed excluding the global regulators of *E. coli* (ArcA, CRP, FIS, FNR, Hns, IHF and Lrp) (Martinez-Antonio & Collado-Vides, 2003) and *B. subtilis* (AbrB, ComK, TnrA, CodY, Spo0A and CcpA) (Moreno-Campuzano *et al.*, 2006), both distances increased significantly, to 6.5 and 3.4 SDS, respectively. These results highlight the evolutionary plasticity of the regulatory networks: not only as a result of the duplication of TFs' interactions on the regulatory network as previously proposed by Teichmann & Babu (2004), but also as a result of the divergent effect of the TF interactions, activating or repressing the transcription of paralogous genes.

Divergence of paralogous genes may include the modification or acquisition of new regulatory mechanisms, changes in gene dosage (Gu *et al.*, 2002), subdivision of ancestral functions and the evolution of new functions (Conant & Wolfe, 2008). Our results suggest that modulation is one of the first steps towards evolutionary innovation at a biochemical level, perhaps as a step towards the modification of the entire pathway. In this regard, diverse authors have proposed that while paralogues may retain similar functionality, gene expression rapidly diverges immediately after duplication events (Gu *et al.*, 2005, 2002; Li *et al.*, 2005), implying that alterations in gene expression precede potential changes in function. Furthermore, it has been postulated that paralogous genes which diverge in function contribute to evolutionary innovation at a biochemical level (Evangelisti & Wagner, 2004; Gu *et al.*, 2005, 2002; Zhang *et al.*, 1998). In order to corroborate these observations, we analysed the metabolic divergence of paralogous genes. By identifying the differences in the assigned biochemical reaction or metabolic pathway, we were able to show that 198 (42%) of the paralogous groups in *E. coli* have a known metabolic function, according to the KEGG database (Gu *et al.*, 2005, 2002; Kanehisa *et al.*, 2008). From this set, we identified 53 groups (27%) that possess members which share the same metabolic pathway



and perform the same enzymic reactions; 64 groups (32 %) which are formed exclusively from proteins that belong to different pathways; and 26 groups (13 %) which are entirely constituted of proteins that are part of the same pathway, although they take different metabolic steps.

Intermediate cases of groups partially formed by any of the above categories were found on 55 occasions (28 %). Slightly different tendencies were observed in the 485 groups of paralogous genes in *B. subtilis*. Finally, duplication events may not be restricted to individual

Fig. 1. Regulatory elements involved in the regulation of the paralogous gene copies. (a) Simplification of the model of Teichmann & Babu (2004), where a new TF is gained in order to regulate one of the paralogous gene copies. TFs are represented by circles. (b) Extension of the Teichmann and Babu model where the differential regulation of the paralogous genes depends on the different regulatory activity (either activation or repression) of a TF on its gene targets. The activation and repression activities of the TFs are represented by circles with green and red perimeters, respectively. (c) Regulatory networks of paralogous genes dependent on sigma, anti-sigma and anti-anti-sigma factors. These factors are represented by rectangles, hexagons and triangles, respectively. The inhibitory activity of these elements is represented by T-like lines. (d) Regulatory networks of paralogous genes dependent on protein–protein interactions. In bacteria, two-component signal-transduction systems are one of the most common types of regulatory networks. The histidine kinases of the system are represented by ovals, and the response regulators by circles. (e) Regulatory networks of paralogous genes dependent on *cis*- and *trans*-RNA elements. *cis*-RNA elements are represented by stem-and-loop structures, whereas different small RNAs are represented by short undulating lines. (f) Regulatory networks of paralogous genes dependent on RNA-binding proteins (represented by rhomboids).

genes, but may also involve large parts of metabolic pathways, as is the case with the leucine, arginine and lysine biosynthesis pathways, where several enzymes are involved in related chemical reactions, with analogous but different substrates (Fondi *et al.*, 2007).

Alternative sigma, anti-sigma and anti-anti-sigma factors may also be part of the regulatory network of paralogous genes

Transcription regulation by alternative sigma factors, anti-sigma factors and anti-anti-sigma factors is a common strategy used by various organisms in order to accomplish different cellular processes, including (i) the regulation of specific biogenesis processes, such as flagellar biosynthesis in *E. coli*, which is regulated by σ^{28} (Chilcott & Hughes, 2000); (ii) the response to a metabolic requirement, as occurs when *E. coli* grows in nitrogen-deprived conditions and gene transcription depends on σ^{54} (Reitzer & Schneider, 2001); (iii) the response to different kinds of stress factors, as in heat shock stress in *E. coli*, which is regulated by σ^{32} , or in the case of general stress in *B. subtilis*, which is regulated by σ^B (Hecker *et al.*, 2007); (iv) synchronized transcription at a specific growth stage, as occurs in the stationary phase of *E. coli*, where transcription is regulated by σ^{38} (Hengge-Aronis, 2002); and (v) coordinated cell differentiation processes, such as *B. subtilis* forespore formation, whose regulation depends on genes transcribed by σ^E , σ^F , σ^G , σ^H and σ^K (Kroos & Yu, 2000). Although in general each of the sigma factors recognizes clearly distinguishable promoter sequences and thus transcribes different groups of genes, in *E. coli* the stationary-phase σ^{38} and its paralogous counterpart, the housekeeping σ^{70} , may share overlapping sets of target genes (Typas *et al.*, 2007). The presence of *cis*-acting elements located in the promoter region, as well as the activity of *trans*-acting proteins, establishes whether a promoter is recognized by RNA polymerase containing σ^{38} or σ^{70} . This transcription regulation based on σ^{38} and σ^{70} constitutes a clear example of an extension of the Teichmann and Babu model, where the regulatory network grows by the duplication of regulatory elements other than

TFs, which modulate the expression of common gene targets.

According to our compiled data referring to the regulatory networks of paralogous genes, there are 100 out of 477 (21 %) groups in *E. coli* and 144 out of 483 (30 %) groups in *B. subtilis* that are regulated by alternative sigma, anti-sigma or anti-anti-sigma factors (Supplementary Tables S1 and S2). In order to evaluate the statistical significance of finding a sigma factor which regulates two or more paralogous genes, we repeated the simulation procedure performed in the TF analysis, using the sigma factor interaction as input data. Our statistical analysis shows that these sigma factor interactions on paralogous genes in *E. coli* and *B. subtilis* were highly significant, as they registered 18 and 6.8 SDS above the mean shown by those obtained in the corresponding randomized regulatory networks. The regulatory model of paralogous genes based on sigma factors is illustrated by the stress response and the early sporulation stage of *B. subtilis*, which both depend on the paralogous sigma factors σ^B and σ^F . The regulatory network of these systems also involves a set of paralogous anti-sigma and anti-anti-sigma factors and represents an excellent example of an extension of the Teichmann and Babu model, where the regulatory elements are not TFs. In a first instance, the transcriptional role of σ^B and σ^F on their target genes is inhibited by the binding of the anti-sigma factors RsbW and SpoIIAB (Group 232, Supplementary Tables S2 and S4), respectively. In a second instance, these paralogous anti-sigma factors are negatively regulated by the anti-anti-sigma factors RsbV and SpoIIAA (Group 313, Supplementary Tables S2 and S4). Finally, the transcription of the genes that encode these paralogous anti-anti-sigma factors depends on the paralogous sigma factors σ^A and σ^F (Hecker *et al.*, 2007) (Fig. 1c).

The impact of evolutionary forces on the topological structures known as motifs has been analysed for different kinds of networks (Shen-Orr *et al.*, 2002); for example biological regulatory networks (Kashtan & Alon, 2005; Shen-Orr *et al.*, 2002) and paralogous gene networks (Teichmann & Babu, 2004). It has been reported that duplication of an entirely feed-forward motif (a topological

structure where a TF regulates a second TF, and both TFs simultaneously regulate a target gene) has not been observed in the regulatory networks of model organisms, although single genes generated by duplication could be part of a new feed-forward or other kind of motif. For example, in *B. subtilis* the paralogues σ^F and σ^E are part of different feed-forward motifs. In the first case, σ^F forms a feed-forward motif with the anti-anti-sigma factor SpoIIAA and the anti-sigma factor SpoIIAB, as σ^F regulates the expression of SpoIIAA and SpoIIAB, whereas SpoIIAA modulates the expression of SpoIIAB. In a similar manner, the second feed-forward loop is formed by σ^E , PhoP and PhoR, which are involved in phosphate uptake, post-exponential growth phase and other stress responses (Pragai *et al.*, 2004). Duplication events, or mutations in the duplicated regulatory genes or in the regulatory target sites, can generate new feed-forward motifs useful for the rewiring of the regulatory networks of the paralogous genes, which favours the adaptation process of the organism as it responds to niche adaptation (Gelfand, 2006).

Protein–protein interactions modulate the activity of paralogous proteins

The regulatory networks evolved by gene duplication are not exclusively regulated at the level of transcription initiation. The activity of paralogous proteins can also be significantly post-translationally modified by protein–protein interactions. This kind of regulation is mainly represented by the two-component signal-transduction systems. Most of these systems consist of a membrane-bound histidine kinase that senses a specific environmental signal and catalyses the transfer of a phosphate group to its corresponding response regulator, inducing its activation. An example of this regulation is found in *E. coli* by the paralogous TFs GlnG and AtoC, involved in the transcriptional activation of genes in response to nitrogen limitation and the catabolism of short-chain fatty acids, respectively. The regulatory activity of these TFs depends on their corresponding paralogous histidine kinases GlnL and AtoS (Lioliou *et al.*, 2005; Ninfa & Magasanik, 1986) (see Fig. 1d and Group 8 of Supplementary Tables S1 and S3). Another well-characterized example of the regulation of paralogous proteins by protein–protein interaction is observed in the chemotactic response of *E. coli*. In this case CheR and CheB cause the post-translational regulation of the methyl-accepting chemotaxis proteins Trg, Tap, Tar, Aer and Tsr, paralogous proteins that undergo changes in their methylation state in response to chemical stimuli, thus establishing the swimming pattern (Ferris & Minamino, 2006) (See Group 41 of Supplementary Tables S1 and S3). From our set of paralogous genes, we found nine groups regulated by protein–protein interactions in *E. coli* and nine groups in *B. subtilis* (Supplementary Tables S1 and S2). The numbers of post-translational interactions that involve one regulator modulating the activity of two or more paralogous genes in *E. coli* and *B. subtilis* were 16 and

12, respectively. Despite the fact that these interactions represent only a small fraction of their corresponding regulatory networks, we found these values to be statistically significant, as they are 23.5 and 15.5 SDs above the means of the corresponding *E. coli* and *B. subtilis* randomly generated networks. It is worth mentioning that these protein–protein interactions in the networks of paralogous genes have great hierarchical relevance in terms of the overall regulatory networks, since the modified proteins commonly correspond to regulatory proteins that in turn regulate the transcription of many other genes or interact among these to form a cascade of transcriptional regulation, as in the cases of PhoP/PhoQ, RstA/RstB and PhoB/PhoR, two-component systems that respond to acidic conditions in *E. coli* (Ogasawara *et al.*, 2007).

The preceding examples demonstrating the regulatory role of the two-component system proteins provide evidence of the relevance of protein–protein interactions in the evolution of the regulatory networks of paralogous genes in *E. coli* and *B. subtilis*. It has been documented that the sensor domain of the histidine kinases is subject to a faster evolutionary process than that of the DNA-binding domain of the response-regulator proteins (Alm *et al.*, 2006). The plasticity of these elements in terms of their capacity to change their regulatory specificity as they respond to the adaptive process may include gene duplication and specialization (Commichau & Stulke, 2008). Furthermore, it has been proposed that two-component proteins acquired through duplication may undergo subsequent domain shuffling, permitting the development of novel genetic diversity. This may happen more quickly than that manifested by two-component systems acquired by horizontal gene transfer, as these proteins are more likely to retain their original functions (Alm *et al.*, 2006; Wanner, 1992; Yosef *et al.*, 2009). To sum up, massive duplication followed by shuffling, faster evolution in the sensor protein domain, and probably horizontal gene transfer, all have a significant influence on the evolution of protein–protein interactions and on the architecture of regulatory networks in bacteria.

RNA and RNA-binding proteins are important factors involved in shaping the regulatory networks of paralogous genes

The important role played by RNA molecules in gene regulation has been recently documented. In bacteria, RNA molecules can regulate either *in cis*, e.g. as riboswitches that are located in the 5' untranslated leader region (Winkler & Breaker, 2005) or *in trans*, acting as diffusible products which act by base pairing with their target mRNAs, resulting in changes in their translation or stability (Gottesman *et al.*, 2006). We identified the participation of RNA elements in 28 and 52 groups of paralogous genes in *E. coli* and *B. subtilis*, respectively (Supplementary Tables S1 and S2), by means of computational analysis, using the co-variance models from the Rfam database (Gardner *et*

al., 2009). As an example of the RNA molecules recognized, Fig. 1(e) shows the regulation of the *E. coli* *rpoH* and *rpoS* paralogous genes (Group 374 of Supplementary Table S1), which code for the σ^H and σ^S sigma factors, respectively. The translation of *rpoH* is under the regulation of a *cis*-acting RNA element located in the coding sequence of the *rpoH* gene. At low temperatures, this thermo-regulator folds into a secondary structure which overlaps and sequesters the ribosome-binding site (RBS), inhibiting *rpoH* translation (Narberhaus *et al.*, 2006) (Fig. 1e). In contrast, the regulation of *rpoS* depends on a set of *trans*-acting small RNAs, in response to different kinds of stress factors, such as low temperature (DsrA), osmotic shock (RprA) and oxidative shock (OxyS). Interestingly, RprA and DsrA have a positive effect on *rpoS* translation by destabilizing a secondary structure that sequesters the RBS, whereas OxyS has a negative effect on translation, by an unidentified mechanism (Gottesman *et al.*, 2006) (Fig. 1e).

In general, the number of RNA molecules acting *in cis* as riboswitches on the regulatory networks of paralogous genes is greater in the Firmicute *B. subtilis* than in the Proteobacterium *E. coli*. Indeed, there are *B. subtilis* riboswitches that do not exhibit a counterpart in *E. coli*, which is the case for the T-box element, regulating different kinds of genes involved in amino acid biosynthesis, transport and charging onto their corresponding tRNAs (Gutierrez-Preciado *et al.*, 2005, 2009). On the other hand, the regulation of paralogous genes by *trans*-acting RNA elements is particularly important in *E. coli* and other Gram-negative bacteria, where it has been related to the response to different stress factors such as the aforementioned low temperature (DsrA), osmotic shock (RprA) and oxidative shock (OxyS), in addition to the stress caused by iron limitation (RyhB) and the accumulation of glucose phosphate (SgrR) (reviewed by Gottesman, 2005). In fact, there is an important bias of these small RNAs to regulate paralogous gene coding for outer-membrane proteins involved in stress response, previously described (see Groups 19, 58 and 278 of Supplementary Tables S1 and S3). Interestingly, some of these small RNAs have been generated by duplication, as is the case of *E. coli* OmrA RNA, which is accumulated in the late stationary phase, whereas its paralogous counterpart, OmrB, is transiently expressed in the early stationary phase (Guillier *et al.*, 2006). In higher organisms such as rice, it has been proposed that some pseudogenes may have acquired novel regulatory roles as antisense small RNAs, after being duplicated (Guo *et al.*, 2009). Duplication of regulatory RNA elements is not restricted to *trans*-acting RNAs, but may include *cis*-acting riboswitches. Such is the case with tandem T-boxes located in the regulatory regions of the tryptophan biosynthetic operons in *Bacillus anthracis*, *Bacillus cereus*, *Bacillus halodurans* and other Gram-positive bacteria. It has been proposed that these tandem riboswitches may expand the range of regulation of their respective transcription units (Gutierrez-Preciado *et al.*, 2005).

RNA-binding proteins can also be included in the regulatory networks of genes generated by duplication. An example of this is found in the carbon catabolite repression system in *B. subtilis*, which makes possible the preferential utilization of easily metabolizable carbon sources, such as glucose. Fig. 1(f) shows the regulation of the paralogous genes *bglP* and *sacP*, which code for the permeases responsible for the transport of the β -glucoside salicin and sucrose, respectively (Group 36 of Supplementary Table S2). The transcriptional regulation of these genes is based on the paralogous RNA-binding proteins LicT and SacT, which can bind and disrupt potential ribonucleic acid anti-terminators located in the upstream regions of the *bglP* and *sacP* genes (Schilling *et al.*, 2006). As previously carried out for other regulatory elements, a statistical analysis was undertaken for the RNA regulatory interactions of *E. coli* and *B. subtilis*. Despite the participation of RNA elements in different paralogous groups, we only found 12 examples of an RNA element regulating two or more paralogous genes in *B. subtilis*. This value registers 11.6 SDS above the mean of their corresponding randomized regulatory networks.

The regulatory networks of paralogous genes are influenced by DNA supercoiling

Coordinate gene expression involves several hierarchical levels that comprise local control of individual genes or operons, multiple operons within a regulon and multiple regulons within a stimulon. It has been proposed that the highest level of hierarchical gene regulation is mediated by DNA supercoiling, which serves as a global regulator of gene expression (Hatfield & Benham, 2002). Although chromosomal DNA has slight negative supercoiling, the degree of supercoiling can be modulated by the action of DNA topoisomerases, in response to a wide variety of altered nutritional and environmental conditions, including temperature (Grau *et al.*, 1994), osmolarity (Cheung *et al.*, 2003; Grau *et al.*, 1994), oxidative stress (Weinstein-Fischer *et al.*, 2000), aerobic to anaerobic growth shift (Hsieh *et al.*, 1991) and growth stage in either *E. coli* (Cheung *et al.*, 2003) or *B. subtilis* (Bird *et al.*, 1992). DNA supercoiling adjusts the basal levels of expression of all genes whose transcription depends on promoters that are directly sensitive to supercoiling (Hatfield & Benham, 2002) or indirectly, acting in combination with some TFs or sigma factors. This is the case for some genes, where transcription by σ^{38} depends on the superhelical densities, occurring as a function of growth conditions (Bordes *et al.*, 2003). This fine-tuning of regulation by DNA supercoiling is also found to form part of the regulatory networks of paralogous genes. According to our data survey, in *E. coli* there are 15 groups of paralogous genes where regulation is directly modulated by DNA supercoiling. Interestingly, three of these groups correspond to DNA-binding transcriptional regulators which in turn exert a regulatory control on other genes within the network, thus expanding the regulatory response in a hierarchical manner. One of

these groups includes the OxyR regulator involved in the oxidative stress response (Group 4, See Supplementary Tables S1 and S3). In addition, we found groups of paralogous genes coding for transporters of amino acids and other types of metabolite, as well as genes coding for prophage proteins (Supplementary Tables S1 and S3). It is interesting that all the genes whose expression is reported to be modulated by DNA supercoiling belong to different paralogous groups; thus no statistical analysis was carried out to determine the probability of finding this regulatory element acting on two or more different gene targets.

Different kinds of regulatory elements simultaneously participate in the regulatory networks of paralogous genes

The response to different external and metabolic signals on the part of the regulatory networks of paralogous genes involves the integration of multiple kinds of regulatory elements, acting at different levels. An excellent example of this kind of regulatory integration is found in the *E. coli* paralogous porins OmpC and OmpF (Group 58 of Supplementary Table S3) in response to medium osmolarity, temperature and oxidative stress (Guillier *et al.*, 2006; Pratt *et al.*, 1996). Fig. 2 summarizes the different regulatory layers and types of regulatory elements of this

system, which comprise: (i) regulation at transcription initiation by at least 11 TFs that act directly on *ompC* and *ompF*, or indirectly regulate the transcription of other regulatory elements of the system; (ii) regulation at translation initiation, accomplished by a set of small antisense RNAs; (iii) post-transcriptional control of the RNA chaperone StpA on *micF* RNA, by the induction of its degradation; and (iv) post-translational modification of protein activity, as is the case in OmpR activation, by means of its phosphorylation by the histidine kinase EnvZ.

Furthermore, Fig. 2 shows that a single regulatory element may act at different levels on more than one target gene. For example, H-NS may directly regulate *ompC* expression by blocking its transcription and indirectly regulate *ompF* translation by repressing the transcription of the *micF* antisense RNA. Interestingly, a specific TF may have a dual activity on its target genes and may act as a transcription activator or repressor, as a result of its interaction with other regulatory elements of the system. This is the case with OmpR, where the active state depends on its phosphorylation by EnvZ, in response to osmolarity. At a low concentration of OmpR-P (low osmolarity), transcription of *ompF* is activated. At high osmolarity, the concentration of OmpR-P increases and the protein binds to additional low-affinity sites, blocking the transcription of *ompF* (Pratt *et al.*, 1996).

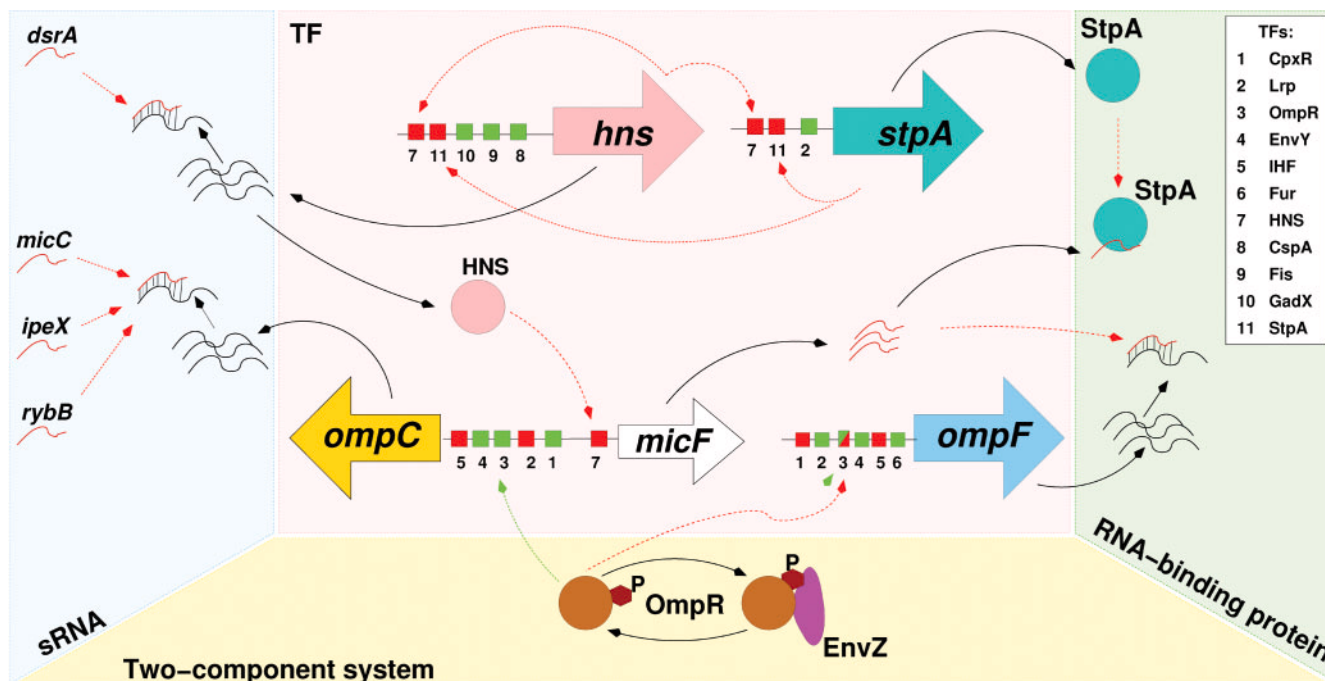


Fig. 2. Integrative example of a regulatory network of paralogous genes based on different kinds of regulatory elements. The regulatory network of the OmpC and OmpF paralogous porins is shown. Genes and their corresponding polypeptide products are depicted in the same colour. Dashed arrows represent regulatory interactions and are coloured according to their either positive (green) or negative (red) effect on expression. Small boxes in the intergenic sequences are coloured according to the regulatory effect of their corresponding TFs. The RNA gene products are represented by short, undulating lines and the protein products by ovals.

Concluding remarks

The compilation and analysis of regulatory elements in paralogous genes of model organisms has led us to extend the original model of the evolution of regulatory networks, which took only TFs into account, to a model of more general type, which includes other regulators such as small RNAs, riboswitches, RNA-binding proteins, sigma factors, protein–protein interactions and DNA supercoiling. Despite the fact that TFs are the most extensively used element in regulatory networks, the extended repertoire of other regulatory mechanisms has resulted in a significant increase in the versatility of the network, accurately modulating the organism's gene expression. Our analysis of the model organisms *E. coli* and *B. subtilis* is consistent with previous observations (Conant & Wolfe, 2008) regarding the relevance of the regulatory networks of paralogous genes in the process of adaptation of an organism in response to changing environments.

Acknowledgements

We thank Cei Abreu-Goodger and Zuemy Rodriguez-Escamilla for their critical reading of the manuscript, Ricardo Ciria for computer support and Shirley Ainsworth for bibliographical assistance. This work was supported by CONACyT grants 60127-Q and SALUD-2007-C01-68992 to E. M. and DGAPA grant IN217508 to E. P.-R.

References

- Alm, E., Huang, K. & Arkin, A. (2006). The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. *PLOS Comput Biol* 2, e143.
- Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol* 215, 403–410.
- Babu, M. M., Luscombe, N. M., Aravind, L., Gerstein, M. & Teichmann, S. A. (2004). Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol* 14, 283–291.
- Babu, M. M., Teichmann, S. A. & Aravind, L. (2006). Evolutionary dynamics of prokaryotic transcriptional regulatory networks. *J Mol Biol* 358, 614–633.
- Balaji, S., Babu, M. M. & Aravind, L. (2007). Interplay between network structures, regulatory modes and sensing mechanisms of transcription factors in the transcriptional regulatory network of *E. coli*. *J Mol Biol* 372, 1108–1122.
- Bausch, C., Ramsey, M. & Conway, T. (2004). Transcriptional organization and regulation of the L-idoic acid pathway (GntII system) in *Escherichia coli*. *J Bacteriol* 186, 1388–1397.
- Bird, T., Burbulys, D., Wu, J. J., Strauch, M. A., Hoch, J. A. & Spiegelman, G. B. (1992). The effect of supercoiling on the in vitro transcription of the *spoIIA* operon from *Bacillus subtilis*. *Biochimie* 74, 627–634.
- Bordes, P., Conter, A., Morales, V., Bouvier, J., Kolb, A. & Gutierrez, C. (2003). DNA supercoiling contributes to disconnect σ^S accumulation from σ^S -dependent transcription in *Escherichia coli*. *Mol Microbiol* 48, 561–571.
- Cheung, K. J., Badarinarayana, V., Selinger, D. W., Janse, D. & Church, G. M. (2003). A microarray-based antibiotic screen identifies a regulatory role for supercoiling in the osmotic stress response of *Escherichia coli*. *Genome Res* 13, 206–215.
- Chilcott, G. S. & Hughes, K. T. (2000). Coupling of flagellar gene expression to flagellar assembly in *Salmonella enterica* serovar typhimurium and *Escherichia coli*. *Microbiol Mol Biol Rev* 64, 694–708.
- Commichau, F. M. & Stulke, J. (2008). Trigger enzymes: bifunctional proteins active in metabolism and in controlling gene expression. *Mol Microbiol* 67, 692–702.
- Conant, G. C. & Wolfe, K. H. (2008). Turning a hobby into a job: how duplicated genes find new functions. *Nat Rev Genet* 9, 938–950.
- Evangelisti, A. M. & Wagner, A. (2004). Molecular evolution in the yeast transcriptional regulation network. *J Exp Zool B Mol Dev Evol* 302, 392–411.
- Ferris, H. U. & Minamino, T. (2006). Flipping the switch: bringing order to flagellar assembly. *Trends Microbiol* 14, 519–526.
- Fondi, M., Brilli, M., Emiliani, G., Paffetti, D. & Fani, R. (2007). The primordial metabolism: an ancestral interconnection between leucine, arginine, and lysine biosynthesis. *BMC Evol Biol* 7 (Suppl 2), S3.
- Gama-Castro, S., Jimenez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Penaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muniz-Rascado, L., Martinez-Flores, I. & other authors (2008). RegulonDB (version 6.0): gene regulation model of *Escherichia coli* K-12 beyond transcription, active (experimental) annotated promoters and Textpresso navigation. *Nucleic Acids Res* 36, D120–D124.
- Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S. & other authors (2009). Rfam: updates to the RNA families database. *Nucleic Acids Res* 37, D136–D140.
- Gelfand, M. S. (2006). Evolution of transcriptional regulatory networks in microbial genomes. *Curr Opin Struct Biol* 16, 420–429.
- Gottesman, S. (2005). Micros for microbes: non-coding regulatory RNAs in bacteria. *Trends Genet* 21, 399–404.
- Gottesman, S., McCullen, C. A., Guillier, M., Vanderpool, C. K., Majdalani, N., Benhammou, J., Thompson, K. M., FitzGerald, P. C., Sowa, N. A. & FitzGerald, D. J. (2006). Small RNA regulators and the bacterial response to stress. *Cold Spring Harb Symp Quant Biol* 71, 1–11.
- Grau, R., Gardiol, D., Glikin, G. C. & de Mendoza, D. (1994). DNA supercoiling and thermal regulation of unsaturated fatty acid synthesis in *Bacillus subtilis*. *Mol Microbiol* 11, 933–941.
- Gu, Z., Nicolae, D., Lu, H. H. & Li, W. H. (2002). Rapid divergence in expression between duplicate genes inferred from microarray data. *Trends Genet* 18, 609–613.
- Gu, X., Zhang, Z. & Huang, W. (2005). Rapid evolution of expression and regulatory divergences after yeast gene duplication. *Proc Natl Acad Sci U S A* 102, 707–712.
- Guillier, M., Gottesman, S. & Storz, G. (2006). Modulating the outer membrane with small RNAs. *Genes Dev* 20, 2338–2348.
- Guo, X., Zhang, Z., Gerstein, M. B. & Zheng, D. (2009). Small RNAs originated from pseudogenes: cis- or trans-acting? *PLOS Comput Biol* 5, e1000449.
- Gutierrez-Preciado, A., Jensen, R. A., Yanofsky, C. & Merino, E. (2005). New insights into regulation of the tryptophan biosynthetic operon in Gram-positive bacteria. *Trends Genet* 21, 432–436.
- Gutierrez-Preciado, A., Henkin, T. M., Grundy, F. J., Yanofsky, C. & Merino, E. (2009). Biochemical features and functional implications of the RNA-based T-box regulatory mechanism. *Microbiol Mol Biol Rev* 73, 36–61.
- Hatfield, G. W. & Benham, C. J. (2002). DNA topology-mediated control of global gene expression in *Escherichia coli*. *Annu Rev Genet* 36, 175–203.

- Hecker, M., Pane-Farre, J. & Volker, U. (2007). SigB-dependent general stress response in *Bacillus subtilis* and related gram-positive bacteria. *Annu Rev Microbiol* **61**, 215–236.
- Hengge-Aronis, R. (2002). Signal transduction and regulatory mechanisms involved in control of the σ^S (RpoS) subunit of RNA polymerase. *Microbiol Mol Biol Rev* **66**, 373–395 (table).
- Hsieh, L. S., Burger, R. M. & Drlica, K. (1991). Bacterial DNA supercoiling and [ATP]/[ADP]. Changes associated with a transition to anaerobic growth. *J Mol Biol* **219**, 443–450.
- Izu, H., Adachi, O. & Yamada, M. (1997). Gene organization and transcriptional regulation of the *gntRku* operon involved in gluconate uptake and catabolism of *Escherichia coli*. *J Mol Biol* **267**, 778–793.
- Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S. & other authors (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Res* **36**, D480–D484.
- Kashtan, N. & Alon, U. (2005). Spontaneous evolution of modularity and network motifs. *Proc Natl Acad Sci U S A* **102**, 13773–13778.
- Kroos, L. & Yu, Y. T. (2000). Regulation of sigma factor activity during *Bacillus subtilis* development. *Curr Opin Microbiol* **3**, 553–560.
- Li, W. H., Yang, J. & Gu, X. (2005). Expression divergence between duplicate genes. *Trends Genet* **21**, 602–607.
- Lioliou, E. E., Mimitou, E. P., Grigoroudis, A. I., Panagiotidis, C. H., Panagiotidis, C. A. & Kyriakidis, D. A. (2005). Phosphorylation activity of the response regulator of the two-component signal transduction system AtoS-AtoC in *E. coli*. *Biochim Biophys Acta* **1725**, 257–268.
- Luscombe, N. M., Babu, M. M., Yu, H., Snyder, M., Teichmann, S. A. & Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. *Nature* **431**, 308–312.
- Lynch, M. & Conery, J. S. (2000). The evolutionary fate and consequences of duplicate genes. *Science* **290**, 1151–1155.
- Martinez-Antonio, A. & Collado-Vides, J. (2003). Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr Opin Microbiol* **6**, 482–489.
- Moreno-Campuzano, S., Janga, S. C. & Perez-Rueda, E. (2006). Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes – a genomic approach. *BMC Genomics* **7**, 147.
- Narberhaus, F., Waldminghaus, T. & Chowdhury, S. (2006). RNA thermometers. *FEMS Microbiol Rev* **30**, 3–16.
- Ninfa, A. J. & Magasanik, B. (1986). Covalent modification of the *glnG* product, NRI, by the *glnL* product, NRII, regulates the transcription of the *glnALG* operon in *Escherichia coli*. *Proc Natl Acad Sci U S A* **83**, 5909–5913.
- 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**, 4791–4799.
- Pragai, Z., Allenby, N. E., O'Connor, N., Dubrac, S., Rapoport, G., Msadek, T. & Harwood, C. R. (2004). Transcriptional regulation of the *phoPR* operon in *Bacillus subtilis*. *J Bacteriol* **186**, 1182–1190.
- Pratt, L. A., Hsing, W., Gibson, K. E. & Silhavy, T. J. (1996). From acids to *osmZ*: multiple factors influence synthesis of the OmpF and OmpC porins in *Escherichia coli*. *Mol Microbiol* **20**, 911–917.
- Reitzer, L. & Schneider, B. L. (2001). Metabolic context and possible physiological themes of σ^{54} -dependent genes in *Escherichia coli*. *Microbiol Mol Biol Rev* **65**, 422–444.
- Schilling, O., Herzberg, C., Hertrich, T., Vorsmann, H., Jessen, D., Hubner, S., Titgemeyer, F. & Stulke, J. (2006). Keeping signals straight in transcription regulation: specificity determinants for the interaction of a family of conserved bacterial RNA-protein couples. *Nucleic Acids Res* **34**, 6102–6115.
- Shen-Orr, S. S., Milo, R., Mangan, S. & Alon, U. (2002). Network motifs in the transcriptional regulation network of *Escherichia coli*. *Nat Genet* **31**, 64–68.
- Sierro, N., Makita, Y., de Hoon, M. & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* **36**, D93–D96.
- Teichmann, S. A. & Babu, M. M. (2004). Gene regulatory network growth by duplication. *Nat Genet* **36**, 492–496.
- Typas, A., Becker, G. & Hengge, R. (2007). The molecular basis of selective promoter activation by the σ^S subunit of RNA polymerase. *Mol Microbiol* **63**, 1296–1306.
- Wanner, B. L. (1992). Is cross regulation by phosphorylation of two-component response regulator proteins important in bacteria? *J Bacteriol* **174**, 2053–2058.
- Weinstein-Fischer, D., Elgrably-Weiss, M. & Altuvia, S. (2000). *Escherichia coli* response to hydrogen peroxide: a role for DNA supercoiling, topoisomerase I and Fis. *Mol Microbiol* **35**, 1413–1420.
- Winkler, W. C. & Breaker, R. R. (2005). Regulation of bacterial gene expression by riboswitches. *Annu Rev Microbiol* **59**, 487–517.
- Yamamoto, K. & Ishihama, A. (2005a). Transcriptional response of *Escherichia coli* to external copper. *Mol Microbiol* **56**, 215–227.
- Yamamoto, K. & Ishihama, A. (2005b). Transcriptional response of *Escherichia coli* to external zinc. *J Bacteriol* **187**, 6333–6340.
- Yosef, N., Kupiec, M., Rupp, E. & Sharan, R. (2009). A complex-centric view of protein network evolution. *Nucleic Acids Res* **37**, e88.
- Zhang, J., Rosenberg, H. F. & Nei, M. (1998). Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc Natl Acad Sci U S A* **95**, 3708–3713.
- Zimmer, D. P., Soupene, E., Lee, H. L., Wendisch, V. F., Khodursky, A. B., Peter, B. J., Bender, R. A. & Kustu, S. (2000). Nitrogen regulatory protein C-controlled genes of *Escherichia coli*: scavenging as a defense against nitrogen limitation. *Proc Natl Acad Sci U S A* **97**, 14674–14679.