UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN CIENCIAS BIOMÉDICAS
CENTRO DE CIENCIAS GENÓMICAS

IDENTIFICACIÓN Y ANÁLISIS DE LOS DOMINIOS DE INTERACCIÓN A LIGANDO EN FACTORES DE TRANSCRIPCIÓN

TESIS
QUE PARA OPTAR POR EL GRADO DE:
DOCTOR EN CIENCIAS

PRESENTA:
M. EN B. NANCY RIVERA GÓMEZ

DIRECTOR DE TESIS:
DR. ERNESTO PÉREZ RUEDA
INSTITUTO DE BIOTECNOLOGÍA

COMITÉ TUTOR:
DR. PABLO VINUESA
CENTRO DE CIENCIAS GENÓMICAS

Dr. LORENZO SEGOVIA FORCELLA
INSTITUTO DE BIOTECNOLOGÍA

CUERNAVACA, MORELOS. 2013

Quisiera externar mi agradecimiento al Consejo Nacional
de Ciencia y Tecnología (CONACYT) por la beca
proporcionada con No 164293 y a la Dirección General
de Estudios de Posgrado de la UNAM por los
fondos proporcionados para la realización de este proyecto
de doctorado y los apoyos proporcionados para asistir a congresos
que me otorgaron ambas dependencias.

También quisiera agradecer al comité revisor
Dr. José Guillermo Dávila Ramos
Dra. Ma. de Lourdes Girard Cuesy
Dr. Enrique Merino Pérez
Dr. Pedro Miramontes Vidal
Dr. Ernesto Pérez Rueda

Esta tesis se realizó en el Departamento de
Ingeniería Celular y Biocatálisis
en el Instituto de Biotecnología de la
Universidad Nacional Autónoma de México

Con todo mi amor y cariño para

Sara Regina y José Luis

AGRADECIMIENTOS

De entre todas las personas que amo y quiero inmensamente existen dos que son lo más importante para mí, son mi mundo y sin ellos esto sería un título más. Ellos son mi apoyo y mi inspiración. Gracias Pepe por tu amor y cariño, por estar siempre ahí, porque a pesar de los malos ratos siempre hemos sabido salir del bache, por todas esas veces que no dejaste que me rindiera. Gracias a ti mi gorrión, mi solecito mañanero, por tu sonrisa, por ser la personita que cambio mi mundo, porque nunca te rindes, por llenarme de inspiración, gracias Sarita, por ser mi hija. Los amo.

A mi familia:

Antes que nada quisiera agradecerle a mis padres Mariana y Abraham que me dieron la mejor herencia que se le puede dar a un hijo *Educación,* pero sobre todo su apoyo incondicional para poder realizarme como persona.

A mi hermana Abde que siempre ha estado ahí apoyándome, aunque no siempre estemos de acuerdo, también a su esposo Alfredo Quiroz por enseñarme algunas cosas de programación y a mi pequeña Ximena por todas sus bellas sonrisas

A mi cuñada chula, que nos regaló un año de su vida para ayudarnos a cuidar a nuestro bello tesoro y por toda su dedicación, por ser mi amiga.

A mis suegros Julia y Efren, por abrirme las puertas de su casa y su corazón, por ser excelentes personas y brindarme su apoyo incondicional y estar siempre ahí cuando los hemos necesitado, pero también por considerarme su hija.

A Carmen y Tomas y a su familia (de ambos), porque siempre han estado ahí apoyándome y brindándonos su amistad incondicional

A mis amigos y compañeros:

Quisiera agradecerle y aunque creo que nunca dejare de hacerlo a en primer lugar por aceptarme como su alumna (nuevamente), por darme esa oportunidad, pero sobre todo por su apoyo (en muchos sentidos), comprensión y por todo ese conocimiento que me ha transmitido a Ernesto Pérez Rueda.

A mis dos amiguitas consentidas que adoro Karlita y America.

A los chicos del laboratorio, Adriana, Valerie, Fernando, Jesús, Dago, Mario, Miguel, Iris, por sus sugerencias, pero sobre todo por su amistad. A los Drs Lorenzo y Claudia por su apoyo y sugerencias.

A todos los del laboratorio Espin-Merino, con los que me toco compartir muchos momentos muy gratos y llenos de diversión.

A Mari Tabche, Jose y Sol por ser tan lindas personas y brindarnos su amistad sincera y franca.

A Gladys, por todo su apoyo tanto administrativo como personal, pero sobre todo por escucharme siempre

Y a mí comité tutor, los Drs Pablo Vinuesa, Lorenzo Segovia y Miguel A. por todas sus sugerencias y su apoyo.

Y a todos aquellos que olvide mencionar pero que son una parte importante y que de alguna manera contribuyeron a que esto se hiciera.

Gracias.

ÍNDICE

1. RESUMEN
2. ABSTRACT
3. INTRODUCCIÓN
   - 3.1 Regulación de la transcripción
   - 3.2 Factores transcripcionales
   - 3.3 Los Dominios de unión a ADN o DBD
   - 3.4 La estructura hélice vuelta hélice
   - 3.5 Familias y Superfamilias de estructuras de unión al ADN
   - 3.6 Superfamilia *winged helix-turn-helix*
   - 3.7 Los DBDs y sus dominios adicionales
4. JUSTIFICACIÓN
5. HIPÓTESIS
6. OBJETIVO GENERAL
7. OBJETIVOS PARTICULARES
8. ESTRATEGIA EXPERIMENTAL
9. RESULTADOS
   - 9.1 Análisis de las Familias de wHTH en bacterias modelo
   - 9.2 Identificación de Dominios adicionales (PaDos) asociados a los wHTH
   - 9.3 Asociación de Familias de reguladores y sus dominios adicionales.
   - 9.4 MANUSCRITO publicado en la Revista *Microbiology*.
   - 9.5 Distribución de las superfamilias en bacterias con genomas completos
   - 9.6 *Relación entre las superfamilias y sus PaDos*
10. DISCUSIÓN Y CONCLUSIONES
11. PERSPECTIVAS
12. MATERIAL ADICIONAL (CAPITULO LIBRO)
13. MATERIAL SUPLEMENTARIO
14. REFERENCIAS

INDICE DE FIGURAS

Figura 1. Dogma central de la biología molecular

Figura 2. Estructura del HTH y del wHTH

Figura 3. Diagrama de flujo del método

Figura 4. Distribución de las familias de FTs en los organismos modelo

Figura 5. Distribución de los dominios adicionales en los organismos modelo

Figura 6. Diversidad de familias con wHTH observadas y sus PaDos

Figura 7. Distribución de las superfamilias en los diferentes taxa analizados.

Figura 8. Distribución de las superfamilias con respecto al total de FTs en cada organismo.

Figura 9. Relación entre superfamilias y PaDos.

INDICE DE TABLAS

Tabla 1. Factores sigma identificadas en las bacterias *E. coli* K-12 y *B. subtilis* 168.

Tabla 2. Estructuras de unión al ADN descritos en los reguladores transcripcionales.

Tabla 3. Clasificación de los FTs en términos de Superfamilias en *E. coli* K-12.

Tabla 4. Propiedades de algunos miembros de la familia LysR

ABREVIATURAS

AbrB/MazE/MraZ-like	AbrB/MazE/MraZ-like
Acyl-CoA N- (Nat)	Acyl-CoA N-acyltransferases (Nat)
ADN	Ácido desoxirribonucleico
AF2008	Hypothetical protein AF2008
Arginine C-T	C-terminal domain of arginine repressor
ArgR	Arginine repressor (ArgR), N-terminal DNA-binding domain
ARN	Ácido Ribonucleico
ARNpol	ARN Polimerasa
ArsR	ArsR-like transcriptional regulators
Biotin	Biotin repressor-like
Bipartite	C-terminal effector domain of the bipartite response regulators
cAMP	cAMP-binding domain-like
CAP	CAP C-terminal domain-like
CBS-domain	CBS-domain
Class II biotin	Class II aaRS and biotin synthetases
CTD	C-Terminal Domain
DBDs	DNA Binding Domain
Dimeric	Dimeric alpha+beta barrel
E/Iie	Transcription factor E/IIe-alpha, N-terminal domain
F93	Hypothetical protein F93
FadR C-T	Fatty acid responsive transcription factor FadR, C-terminal domain
FlhD	Flagellar transcriptional activator FlhD
FTs	Factores de Transcripción
FUR	FUR-like
GAF	GAF domain-like
Glucocorticoid (DBD)	Glucocorticoid receptor-like (DNA-binding domain)
GntR	GntR-like transcriptional regulators
Hemolysin	Hemolysin expression modulating protein HHA
Homeodomain-like	Homeodomain-like
HPr kN-T	HPr kinase/phoshatase HprK N-terminal domain
HrcA	Heat-inducible transcription repressor HrcA, N-terminal domain
HTH	Helix Turn Helix
IclR	Transcriptional regulator IclR, N-terminal domain
IIA-Man	IIA domain of mannose transporter, IIA-Man
Iron	Iron-dependent repressor protein, dimerization domain
Iron	Iron-dependent repressor protein
KorB DBD-like	KorB DNA-binding domain-like
lambda repressor-like	lambda repressor-like DNA-binding domains
LexA	LexA/Signal peptidase
LexA	LexA repressor, N-terminal DNA-binding domain
Lrp	Lrp/AsnC-like transcriptional regulator N-terminal domain
LysR	LysR-like transcriptional regulators
MarR	MarR-like transcriptional regulators
Mj223	DNA-binding protein Mj223
ModE	N-terminal domain of molybdate-dependent transcriptional regulator ModE
MOP	MOP-like
NAD(P)	NAD(P)-binding Rossmann-fold domain
NagB	NagB/RpiA/CoA transferase-like
NTD	N-Terminal Domain

Nucleic acid-binding	Nucleic acid-binding proteins
PaDos	Partner Domains
PBP I	Periplasmic binding protein-like I
PBP II	Periplasmic binding protein-like II
Penici	Penicillinase repressor
Phos/ani t.p	Phoshotransferase/anion transport protein
P-loop	P-loop containing nucleoside triphosphate hydrolases
PLP	PLP-dependent transferases
PRTase-like	PRTase-like
PTS-reg	PTS-regulatory domain, PRD
PTS-sys	PTS system, Lactose/Cellobiose specific IIB subunit (Pfam 02302)
PurR	N-terminal domain of Bacillus PurR
Putative DBD	Putative DNA-binding domain
Rex	Transcriptional repressor Rex, N-terminal domain
Ribbon-helix-helix	Ribbon-helix-helix
Ribokinase-like	Ribokinase-like
RNAm	ARN mensajero
Rrf2	Transcriptional regulator Rrf2 (Pfam 02082)
Thio/thiol	Thioesterase/thiol ester dehydrase-isomerase
TM1602 C-T	Putative transcriptional regulator TM1602, C-terminal domain
Tran-Repr C-T	C-terminal domain of transcriptional repressors
TrpR-like	TrpR-like
wHTH	Winged Helix Turn Helix
Winged	Winged helix DNA-binding domain
Z-DNA	Z-DNA binding domain
σ	Sigma

1. RESUMEN

La capacidad de las bacterias para contender a los diversos cambios ambientales depende de su repertorio de genes y de su capacidad para regular su expresión. En este proceso regulatorio, los factores de transcripción (FT) tienen un papel fundamental, ya que afectan a la expresión génica de forma positiva y/o negativa dependiendo de la posición de su sitio blanco y de la unión al ligando. En este trabajo, se presenta un análisis comparativo de la superfamilia de FTs que presentan el *winged-Helix-Turn-Helix* (wHTH) en 428 bacterias con genomas completos. En este sentido, se evaluó el repertorio de wHTHs en términos de sus dominios asociados o *Partner Domains* (PaDos), que se presentan en altas proporciones y ampliamente distribuidas como ocurre con los wHTH´s. Con base en la distribución de los wHTH´s y sus PaDos, se definieron tres grandes grupos de familias: i) monolíticos, aquellas familias asociadas a un solo PaDo: ii) promiscuas, familias que presentan una amplia diversidad de PaDos: y iii) monodominio, aquellas familias de tamaño pequeño que están constituidas exclusivamente por el dominio de unión al DNA. Basado en estos análisis, describimos que los PaDos tienen un papel muy importante en la diversificación de respuestas en las bacterias, probablemente contribuyendo a su complejidad regulatoria. Por otro lado, los PaDos permitirían una gran flexibilidad para la regulación transcripcional debido a su capacidad para detectar diversos estímulos a través de una variedad de compuestos de unión a ligando. Estudios han demostrado asociación especifica de DBDs y sus correspondientes PaDos y pocos intercambios entre ellos.

2. ABSTRACT

The ability of bacteria to deal with diverse environmental changes depends on their repertoire of genes and their ability to regulate their expression. In this process, DNA-binding transcription factors (TFs) have a fundamental role because they affect gene expression positively and/or negatively depending on operator context and ligand-binding status. Here, we show an exhaustive analysis of winged helix–turn–helix domains (wHTHs), a class of DNA-binding TFs. These proteins were identified in high proportions and widely distributed in bacteria, representing around half of the total TFs identified so far. In addition, we evaluated the repertoire of wHTHs in terms of their partner domains (PaDos), identifying a similar trend, as with TFs, i.e. they are abundant and widely distributed in bacteria. Based on the PaDos, we defined three main groups of families: (i) monolithic, those families with little PaDo diversity, such as LysR; (ii) promiscuous, those families with a high PaDo diversity; and (iii) monodomain, with families of small sizes, such as MarR. These findings suggest that PaDos have a very important role in the diversification of regulatory responses in bacteria, probably contributing to their regulatory complexity. Thus, the TFs discriminate over longer regions on the DNA through their diverse DNA-binding domains. On the other hand, the PaDos would allow a great flexibility for transcriptional regulation due to their ability to sense diverse stimuli through a variety of ligand-binding compounds. Posterior studies have shown specific association of DBDs and their corresponding PaDos. and few intermingling between them.

3. INTRODUCCIÓN

Las bacterias responden y se adaptan a una gran diversidad de condiciones ambientales, tales como cambios de temperatura, cambios de pH, disponibilidad de nutrientes y a la competencia con otras bacterias, entre otros. Hoy en día se dispone con una gran cantidad de información experimental que ha permitido entender la diversidad de procesos celulares que permiten a las bacterias contender con dichos retos ambientales, así como a diversos mecanismos fisiológicos y metabólicos que implica el mantenimiento celular. En general, para que se lleven a cabo los procesos celulares es necesario que exista la transferencia de información genética, contenida dentro de una molécula de Ácido desoxirribonucleico o DNA. Este proceso se puede resumir en tres procesos fina y específicamente regulados, tales como la síntesis de DNA (Replicación), el paso de DNA a RNA (Transcripción), y la síntesis de proteínas a partir del RNAm (traducción). En general, al proceso de transferencia de información genética se le denomina “Dogma Central de la Biología Molecular” (Figura 1), con sus notables excepciones. En este modelo, se plantea como se llevan a cabo los procesos de la transferencia de la información genética, así como la expresión de los genes; y es este último proceso, un momento crucial en la célula, ya que es donde se decide qué y cuales genes tendrán que ser expresados (Crick, 1970). En este trabajo nos centraremos en el proceso de la transcripción o de la síntesis del RNAm a partir de un templado de DNA y de los mecanismos de regulación asociados a él.

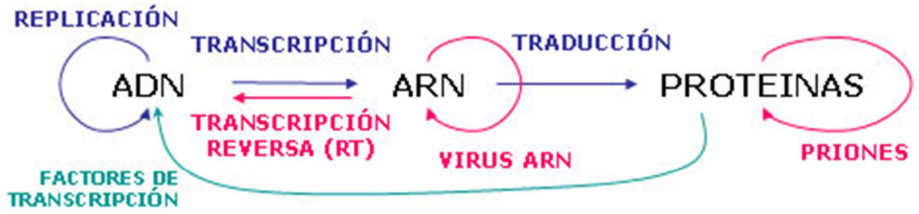


Figura 1. Dogma central de la biología molecular.

3.1. *Regulación de la transcripción*

El mecanismo de la transcripción es probablemente el más finamente regulado en la expresión de los genes. En este proceso participan un gran número de complejos proteicos que determinan el nivel de expresión de un gen de acuerdo a los requerimientos celulares. En general, el mecanismo involucra a la RNA polimerasa (RNApol) como el principal complejo enzimático responsable de sintetizar el RNA mensajero (RNAm). La RNApol es un complejo proteico formado por dos subunidades α, β y β’, así como por la subunidad ω, en las que cada una de ellas realiza una función importante para la formación de un nuevo RNAm. El sitio activo de la enzima está formado por las subunidades β y β' donde se encuentra asociado tanto el DNA blanco como el RNA sintetizado. Las dos subunidades α son proteínas idénticas y están formadas por dos dominios estructurales independientes, un dominio grande denominado αNTD que es el responsable de ensamblar las subunidades β y β' y un dominio pequeño αCTD que se une a regiones conservadas en el DNA y que participa en el reconocimiento de la secuencia promotora[*]. Finalmente, una subunidad pequeña ω que no tiene un papel directo en la transcripción, pero que funciona como chaperona para asistir en el plegamiento de la subunidad β' (Browning, y otros, 2004). En resumen, a todo el complejo proteico previamente descrito se le denomina *core* y para que se inicie la síntesis el RNAm es necesario que se forme un complejo proteico denominado holoenzima, donde el *core* de la RNApol se asocia a una sexta subunidad disociable, conocida como factor σ y que es el responsable del reconocimiento específico de las secuencias promotoras (ver Tabla 1) (Busby, y otros, 2002; Borukhov, y otros, 2002). Esta interacción proteica (*core* – factor sigma) da por resultado la formación de la holoenzima. La función de los factores sigma es la de incrementar la afinidad de la RNApol hacia un conjunto de secuencias específicas, denominadas promotores. Por ejemplo, cuando una bacteria es expuesta a estrés por calor, hay una expresión diferencial de aquellos genes que presentan promotores que son reconocidos específicamente por el factor σ^{32} (ver tabla 1) y así contender con el incremento de temperatura. De esta manera, el factor σ presenta al menos tres funciones básicas, asegurar el reconocimiento de los promotores específicos, dirigir a la RNApol a sus genes blanco para la formación del complejo cerrado y facilitar el relajamiento de la doble cadena de DNA cercano al sitio del inicio de la transcripción (formación del complejo abierto) (Rodriguez, y otros, 2006) para que se de inicio a la sintesis del

[*] **Promotor**: Una región discreta de DNA de alrededor de 40 pares de bases localizado río arriba del inicio de la traducción y que es requerido para el inicio de la síntesis del ARNm. En las bacterias la ARNpol y los factores σ se unen en esta región. Los promotores también pueden contener sitios regulatorios, tales como operadores o sitios activadores que son reconocidos por los factores de la transcripción (Schreiter, y otros, 2007)

RNAm.

Los factores σ se encuentran clasificados, con base en sus secuencias de aminoácidos, en dos grandes familias evolutivas, la familia σ^{70} y la familia σ^{54} (Borukhov, y otros, 2002). A su vez, la familia σ^{70} se clasifica en cuatro grupos filogenéticos con base en su estructura y función, donde cada grupo regula genes asociados a funciones específicas (ver tabla 1), desde aquellos de mantenimiento o *housekeeping*, hasta aquellos que responden a señales específicas, tales como estrés por calor (Paget, y otros, 2003). Además, cada organismo presenta su propio conjunto de factores sigma específicos para cada función, por lo tanto el número y el tipo de factores sigma que se han identificado varía dependiendo del organismo. Por ejemplo, en la bacteria *Bacillus subtilis* se han identificado 19 diferentes factores sigma, de los cuales varios de ellos están asociados al proceso de la esporulación (Tabla 1) (Haldenwang, 1995).

Las regiones promotoras son elementos que contribuyen a la regulación de la transcripción ya que es donde se posiciona la holoenzima para formar el complejo cerrado y el completo abierto para dar inicio a la síntesis del RNA mensajero. Actualmente se conocen una variedad de secuencias conservadas que son reconocidas por los diferentes factores sigma, al igual que los elementos reconocidos por el factor σ^{70} que se encuentran *upstream* del inicio de la transcripción o +1. Este tipo de promotores están organizados por dos hexámeros, localizados hacia el -35 (TTGACA) y hacia el -10 (TATAAT) con respecto al +1 (indicado como el inicio de la transcripción. Un subconjunto de promotores dependientes del σ^{70} carecen de una buena correspondencia con la secuencia -35 consenso, no obstante, son reconocidos por la RNApol. Estos se conocen como el -10 extendido y contiene la secuencia TGn en las posiciones -15/-14 y esto hace que la caja -10 sea más larga, facilitando el reconocimiento por la holoenzima. Por otro lado, se ha observado una región conocida como elementos *UP (Upstream)*, situada entre el -40 y el -60, rica en AT y que presenta sitios de unión para una o ambas subunidades α, incrementando el equilibrio inicial entre la RNApol y el DNA (Hinton, y otros, 2005; Browning, y otros, 2004; Hsu, 2002; Gourse, y otros, 2000; Paul, y otros, 2004; Kumar, y otros, 1993)

Tabla 1. Factores sigma identificadas en las bacterias *E. coli* K-12 y *B. subtilis* 168.

FACTOR SIGMA	GEN	FUNCIÓN	SECUENCIA DE RECONOCIMIENTO		
***E. coli* K-12**			**-35**		**-10**
σ^{70}	*rpoD*	General	TTGACA	16-18	TATAAT
σ^{32}	*rpoH*	Choque térmico	CCCTTGAA	13-15	CCCGATNT
σ^{E}	*rpoE*	Choque térmico	ND	ND	ND
σ^{54}	*rpoN*	Nitrógeno	CTGGNA (-24)	6	TTGCA (-12)
σ^{F}	*fliA*	Flagelar	CTAAA	15	GCCGATAA
$\sigma^{38(s)}$	*rpoS*	Fase estacionaria y disminución de nutrientes	Región rica en GC		TATACT
B. subtilis					
$\sigma^{A}(\sigma^{43}, \sigma^{55})$	*sigA, rpoD*	General/esporulación temprana	TTGACA	17-19	TATAAT
σ^{B} (σ^{37})	*sigB*	Respuesta general a estrés	RGGXTTRA	11-15	GGGTAT
σ^{C} (σ^{32})	ND	Expresión de genes postexporencial	AAATC	14-15	TAZTGYTTZTA
σ^{D} (σ^{28})	*sigD, flaB*	Quimiotaxis/flagelar/auto-lisina	TAAA	15-16	GCCCGATAT
σ^{H} (σ^{30})		Expresión de genes postexporencial: competencia y esporulación temprana	RWAGGAXXT	14	HGAAT
σ^{L}	*sigL*	Expresión de enzimas degradativas	TGGCAC	5	TTGCANNN
σ^{E} (σ^{29})	*sigE, spoIIGB*	Esporulación temprana	ZHATAXX	14	CATACAHT
σ $(\sigma^{SPOIIAC})$	*sigF, spoIIAC*	Esporulación temprana	GCATR	15	GGHRARHTX
σ^{G}	*sigG, sopIIIG*	Esporulación tardía	GHATR	18	GGHRARHTX
$\sigma^{K}(\sigma^{27})$		Esporulación tardía	GA	17	CATANNNTA

Designación de nucleótidos: H: A o C: N: A, G, C o T; R: A o G; W: A, G o C; X: A O T; Y: C o T; Z: T o G. tomado de (Lewin, 2001; Haldenwang, 1995; Doi, y otros, 1986; Hengge , 2009). Adicionalmente, se han identificado los factores sigma H, I, M. V, W, X, Y, Z y Y en *B. subilis* y σ^{19} (*fecI*) en *E. coli* K-12 que regula el transporte de citrao férrico (Braun, y otros, 2005). ND, No Determinado.

3.2. Factores Transcripcionales

En el mecanismo de la transcripción también participan otras proteínas que no forman parte de la holoenzima, pero que determinan de manera regulada y diferencial la expresión de los genes. A estas proteínas se les denomina **F**actores **T**ranscripcionales o FTs. Dependiendo de las condiciones ambientales en las que se encuentre la célula, ya sea crecimiento, proliferacón, asimilación de nutrientes o respuesta a estrés, la expresión de los genes serán controlados por estas proteínas, que se unen a sitios cercanos al +1 o inicio de la transcripción.

Cuando un FT se une a una región reguladora, la expresión génica puede ser activada o reprimida dependiendo de la posición del sitio con respecto al promotor. En general, se han descrito diversos modelos que explican cómo actúan los represores y los activadores. Cuando un FT actúa como activador, promueve la afinidad de la holoenzima hacia el promotor. Los activadores se unen a menudo río arriba del -35, usualmente cercanos al -61, -71, -81 ó -91 y funcionan a través de la interacción directa con el dominio C-terminal de la subunidad α de la RNApol. Por otra parte, los FTs que actúan como represores se unen a una región situada entre los elementos -35 y -10, bloqueando el acceso a la RNApol, es decir, no se permite la formación del complejo cerrado.

Se estima que tres bacterias modelo, como lo son *E. coli* K-12, *Bacillus subtilis* 168 y *Corynebacterium glutamicum* ATCC 13032 cuentan con alrededor de 300, 240 y 160 FTs respectivamente, de los que aproximadamente la mitad han sido caracterizados experimentalmente. Desde un punto de vista funcional, se ha descrito que en la bacteria *E. coli* K-12 el 35% de los FTs (con evidencias experimentales) actúan como activadores, 43% como represores y el 22% de FTs presentan ambas actividades (Browning, y otros, 2004; Busby, y otros, 2002; Janga, y otros, 2007; Pérez-Rueda, y otros, 2000; Bernard, y otros, 2004).

Adicionalmente y con base en el número de genes que regulan los FTs, estos pueden ser clasificados como globales o locales. Siendo un regulador global se define como aquel que tiene la capacidad de regular gran número de genes más allá de su diversidad funcional, controlar una compleja cascada de regulación, afectar directa o indirectamente varias vías celulares y actuar en promotores de diferentes factores sigma (Martínez-Antonio, y otros, 2006). De acuerdo a esta definición en la bacteria *E. coli* K-12 se han propuesto siete reguladores globales (CRP, FNR, IHF, FIS ArcA, NarL y Lrp), que regulan cerca del 51% de sus genes, en *B. subtilis* se han descrito siete reguladores globales, mientras que en *C. glutamicum* solo se ha descrito experimentalmente un regulador global (Moreno-Campuzano, y otros, 2006; Pérez-Rueda, y otros, 2000; Brinkrolf, y otros, 2007).

3.3. Los Dominios de unión al DNA o DBD

Las proteínas están organizadas en uno o más dominios[†] estructurales a excepción de algunas proteínas desordenadas. En este contexto, los FTs son proteínas que presentan por lo general dos dominios estructurales. Por una parte, un dominio que es el encargado del reconocimiento del ligando o la señal ambiental, mientras que el otro dominio realiza contactos específicos con sus sitios blanco en el DNA. A este dominio se le conoce como, Dominio de unión a DNA o DBD (por sus siglas en inglés ***D****NA* ***B****inding* ***D****omain*) y ha sido objeto de diversos estudios comparativos, ya que son considerados como pieza clave para clasificar y designar a los FTs en grupos evolutivamente relacionados. Estos dominios tienen una gran afinidad por sitios específicos en el DNA, en donde penetran en el surco mayor del DNA. Entre las dos moléculas existen interacciones generales, que estabilizan el complejo proteína-DNA, tales como los puentes de hidrógeno. Finalmente, diversos análisis estructurales revelan que la mayoría de los FTs en los procariontes actúan como homodímeros que reconocen secuencias palindrómicas o seudopalindrómicas en el DNA (Martínez-Antonio, y

[†] **Dominio proteico**: es una región compacta en la estructura de la proteína que a menudo, es un segmento continuo de una secuencia de aminoácidos, que generalmente es capaz de plegarse independientemente de manera estable en solución. Cerca de la mitad de todos los dominios descritos actualmente se encuentran entre 50 y 150 residuos de aminoácidos. En la naturaleza existe un límite en el repertorio de dominios que son duplicados y combinados por diferentes vías para formar el conjunto de proteínas de un organismo. Cada dominio presenta su propia función y están altamente conservados. En general, son considerados como las unidades independientes y evolutivas de las proteínas. Actualmente existen diferentes bases de datos para clasificar dominios, tales como CATH y SCOP. En CATH por ejemplo, los dominios se dividen en clases, arquitecturas, topologías, superfamilias y familias, mientras que en SCOP están divididos por clases, plegamientos (que incluye topologías y arquitecturas), familias y superfamilias. La clasificación de las estructuras de las proteínas en las bases de datos están basadas en relaciones evolutivas y en los principios que rigen su estructura tridimensional (Murzin, y otros, 1995; Wang, y otros, 2009; Orengo, y otros, 1997; Apic, y otros, 2001; Thorton, y otros, 1999).

otros, 2006; Marmorstein, y otros, 2003; Apic, y otros, 2001; Pérez-Rueda, y otros, 2001).

Actualmente se han identificado varias estructuras asociadas a los DBDs, estos difieren en su estructura terciaria, así como en la forma que interactúan con DNA y su especificidad, tales como los dedos de zinc (ZnF), el hélice-loop-hélice (HLH*),* las hojas beta antiparalelas y el hélice-vuelta-hélice (HTH) (ver Tabla 2). La estructura de unión al DNA mejor caracterizada y estudiada corresponde al HTH que a continuación se describe (Pérez-Rueda, y otros, 2000).

Tabla 2: Estructuras de unión al DNA descritos en los reguladores transcripcionales.

Estructura	Características	Distribución
Hélice vuelta hélice (HTH)	Estructura formada por dos α-hélices de aproximadamente 10 residuos cada una, conectadas por una "vuelta" de alrededor 3 a 5 residuos de aminoácidos. La hélice del C-terminal es de reconocimiento específico y contacta los nucleótidos localizados en el surco mayor del DNA.	Bacterias, Arqueas Eucariotes
Dedos de Zinc	Utiliza una o más moléculas de zinc como su componente estructural. La estructura consiste de α-hélices y β-plegadas unidas por zinc en donde las α-hélices son las que interactúan con el DNA	Eucariotes Arqueas
β-Plegada antiparalela	Es un grupo de reguladores, donde la unión al surco mayor es leída por dos β-Plegadas	Bacterias

Tomado de (Alberts, y otros, 1994; Steitz, 1990)

3.4. La estructura hélice vuelta hélice

El hélice vuelta hélice (HTH) constituye uno de los dominios proteicos más abundantes y ampliamente distribuidos en los procariontes. Este dominio ha sido objeto de múltiples estudios filogenéticos que apuntan a un probable origen monofilético (Rosinski, y otros, 1999). El dominio que contiene al HTH consiste de un *core* de tres hélices que forman un paquete helicoidal derecho con una configuración parcialmente abierta (Ver Figura 2A). La vuelta, define a este dominio y está situada entre la segunda y la tercera hélices y típicamente no tolera inserciones o distorsiones; sin embargo, el *loop* entre la primera y la segunda hélice tiene mucho mayor variabilidad y puede alojar variantes dependiendo del tipo de dominio HTH. En la configuración parcialmente abierta se forma un hueco o *cleft* entre la tercera y la primera hélice que actúa como un sitio que favorece la evolución de elementos estructurales adicionales que se empaquetan dentro de interacciones hidrofóbicas. La mayor parte de las extensiones al dominio HTH son elementos estructurales que al parecer, han evolucionado para generar una configuración más cerrada por interacciones con el hueco. La tercera hélice, es conocida como la hélice de reconocimiento, que utiliza sus aminoácidos de la cadena lateral para el reconocimiento y unión sito-especifico con el DNA, insertándose en el surco mayor del DNA. También la segunda hélice se localiza cercana al DNA de modo que la cadena principal y la cadena lateral del N-terminal son capaces de hacer contactos no específicos con el eje fosfodiéster. Sin embargo, los residuos individuales que participan en el contacto con el DNA pueden variar ampliamente en el plegamiento. Adicionalmente, los contactos secundarios con el DNA pueden ser mediados por otras partes de la estructura o extensiones más allá de HTH. Los contactos entre el HTH y el DNA incluyen puentes de hidrógeno y enlaces salinos y contactos de *van der Waals* (Brennan, y otros, 1989; Aravind, y otros, 2005; Huffman, y otros, 2002).

3.5. *Familias y Superfamilias de estructuras de unión al DNA*

Los análisis relacionados a los FTs se han enfocado, entre otras características, en las secuencias de sus dominios y motivos de unión a DNA. En *E. coli* K-12 los 300 FTs (conocidos y predichos) son agrupados en diferentes familias[‡], en donde el HTH es la estructura predominante (Pérez-Rueda, y otros, 2000) y el número de miembros que conforman a cada familia es variable. Adicionalmente, los FTs han sido clasificados en 11 superfamilias[§]. (Tabla 3), todas estas familias excepto la *Nucleic acid binding domain* son proteínas que contienen un HTH (Babu, y otros, 2003).

En un análisis realizado en 90 genomas completos de Arqueas y Bacterias, se identificaron 75 familias. La distribución y abundancia de estas familias depende en gran medida del tamaño del genoma así como de su estilo de vida. Una de las principales características de estas familias es que la gran mayoría de ellas presenta el HTH en su DBD y estas son agrupadas dentro de tres categorías, el clásico HTH, proteínas con wHTH y una miscelánea de estructuras de HTH. Familias asociadas a otros DBD con estructuras como HLH, dedos de zinc o β-*plegada antiparalela* se identificaron en bajas proporciones. Asimismo, se ha

[‡] De acuerdo con la clasificación de SCOP las proteínas son agrupadas dentro de una **familia** de dominios si todas las proteínas presentan ≥ 30% de residuos idénticos y segundo secuencias con baja identidad pero que su estructura y función sean similares. Muchas familias presentan una clasificación jerárquica mediante la cual se identifican muchos familiares cercanos, por ejemplo una secuencia con una similitud alta (≥ 40% de identidad) son agrupadas dentro de una misma familia. Estos frecuentemente muestran propiedades funcionales comunes.

[§] **Superfamilias**: Proteínas tienen baja identidad de secuencia pero cuyas estructuras y en muchos casos, características funcionales sugieren un probable origen común (Thorton, y otros, 1999; Murzin, y otros, 1995).

reportado que a lo largo de la evolución, estas familias se han distribuido en los diferentes clados taxónomicos, muchas de ellas son grandes familias en las que sus miembros se han duplicado. En contraste, existen familias pequeñas en las que la mayoría de sus miembros se encuentran asociados a organismos específicos o regulan funciones muy particulares; así mismo encontramos aquellas que han sido adquiridas a través de eventos de transferencia horizontal (Pérez-Rueda, y otros, 2000; Collado-Vides, y otros, 2004).

Tabla 3. Clasificación de los FTs en términos de Superfamilias en *E. coli* K-12.

Superfamilia	**Número de familias**	**Ejemplos**
Winged hélix	22	LexA, AsnC, MarR, GntR, LysR, Fur, CRP/FNR, TrmB
Lambda represor-like	6	LacI
C-terminal effector domain	5	PhoB, LuxR
Homeodomain-like	13	AraC/XylS, TetR
IHF-like DNA-binding proteins	1	IHF
Met repressor like	1	MetR
Putative DNA binding protein	1	MeR
Flagellar transcriptional activator FlhD	1	FlhD
Trp repressor	1	TrpR
Nucleic acid binding domain	1	Cold-shock (Csd)
FIS-like	1	EBP

Modificado de Babu y Teichmann, 2003 SCOP, (Babu, y otros, 2003).

3.6. Superfamilia winged helix-turn-helix

La superfamilia winged HTH es la más abundante y diversa de las estructuras de unión al DNA en los Procariotes, ya que alrededor del 50% de los FTs pertenecen a esta superfamilia. El wHTH contiene un HTH seguido de uno o dos β-*hairpin* (conocido como ala o *winged*). Esta ala exhibe una considerable flexibilidad para la unión y reconocimiento de su sitio blanco en el DNA. El motivo contiene dos alas, con una estructura de *loop* extendido de tres α-hélices y tres β-plegadas. El orden es H1-B1-H2-T-H3-B2-W1-B3-W2 (donde la H representa una hélice, la B son las β-plegadas, T la vuelta o "*turn"* y W es la ala o "*wing"*). La H2 y H3 forman el HTH. En las proteínas con wHTH, la H3 es la hélice de reconocimiento que típicamente hace contactos específicos en el surco mayor del DNA. La longitud de la vuelta que conecta a las dos hélices del motivo HTH varía en las diferentes proteínas, algunos dominios exhiben una variedad de β-plegadas, que se intercalan entre la primera y la segunda hélices. Existen versiones con cuatro β-plegadas; sin embargo, las versiones con dos y tres hojas plegadas son las más comunes en familias de DBDs. Adicionalmente, se ha observado que el *winged* le proporciona estabilidad al dominio y al contacto con el DNA principalmente con el surco menor (Figura 2b) (Huffman, y otros, 2002; Aravind, y otros, 2005).

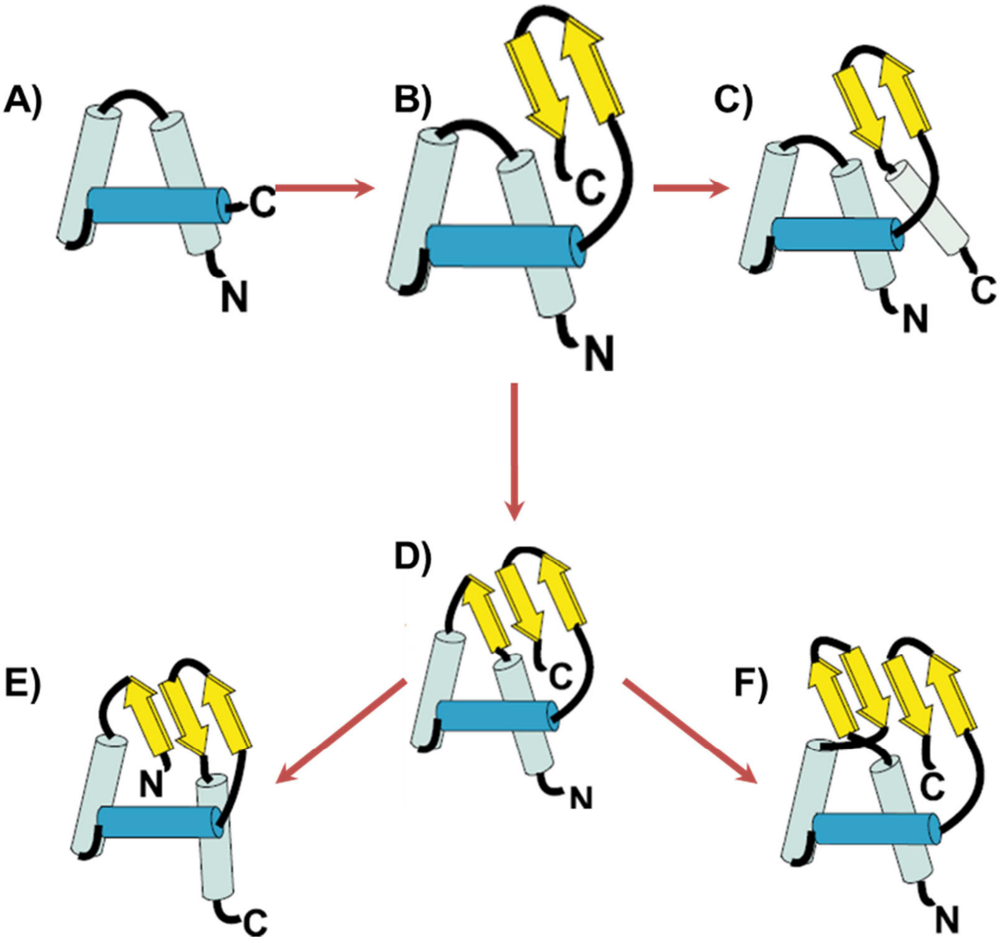


Figura 2: Estructura del HTH y de wHTH. Las flechas amarillas indican las beta-plegadas, los cilindros las alfa-hélices; los cilindros azul intenso indica la hélice de reconocimiento. A) se muestra un HTH típico formado por tres hélices. B) un wHTH, donde se observan dos β-plegadas hacia el C-terminal unido a las tres hélices, encontrado en las familias LRP/AsnC, ArgR, y LexA. C) una variante del wHTH donde se observa una cuarta hélice hacia el C-terminal, presente en las familias MarR, GntR, LysR, Fur y BirA. D) un wHTH con una tercera β-plegada presente en las familias BirA, ArsR y H1. E) versión con tres β-plegadas, familia MerR. F) versión con cuatro β-plegadas presente en la familia CRP. Tomado de Aravind et al 2005.

3.7. Los DBDs y sus dominios adicionales

Los dominios de las proteínas determinan su función y sus relaciones evolutivas. Se ha descrito la existencia de un límite de familias de dominios que se duplican y combinan por distintos procesos para formar el repertorio de proteínas en un genoma (Apic, 2001). Pocas familias se combinan con muchos tipos de dominios y muchas familias tienen una o pocas combinaciones de socios. En general es aceptado que las proteínas evolucionan gracias a los procesos de duplicación, divergencia y recombinación de dominios. Sí pensamos en la recombinación de los miembros de dos o tres superfamilias, las posibilidades de combinación de dominios entre sus repertorios de familias obtendríamos varios cientos de miles de combinaciones. Sin embargo, el repertorio de combinaciones que son observadas en las proteínas indica que todas las combinaciones están bajo una fuerte selección existiendo menos combinaciones de las esperadas (Apic, y otros, 2001; Vogel, y otros, 2004).

Los FTs son proteínas que no están exentos de estos procesos evolutivos ya que en su mayoría muestran uno o más dominios distintivos asociados al DBD presentando distintas *Arquitecturas de Dominios.* Los dominios adicionales al DBD o PaDos (por sus siglas en inglés, *Partner Domains*), están involucrados en la formación de complejos multiméricos, así como al reconocimiento de metabolitos y otras moléculas de señalización, tales como metales, aminoácidos, azúcares, nucleótidos y vitaminas, que provienen del exterior o bien son producto de una vía metabólica. En los FTs, se han identificado la fusión de DBDs con dominios globulares adicionales en el mismo polipéptido. Estos dominios globulares vinculados al DBD muestran una amplia diversidad y apuntan a una inmensa variedad de contextos funcionales en los que los distintos DBDs están presentes. Dentro de una misma familia podemos encontrar una amplia variedad de arquitecturas de dominios, lo que lleva a una ausencia de similitud significativa entre las regiones que participan en la unión del efector u oligomerización.

En general, estas observaciones han dejado de lado sistemáticamente a los PaDos durante el establecimiento de las familias, aunque estos tengan un papel importante en el proceso de la regulación. Así, parece ser que algunas familias de dominios son intrínsecamente más versátiles, es decir, se les encuentran asociados con diferentes dominios, en este caso los DBDs. Por el contrario existen dominios que solo se encuentran asociados a uno o dos familias de DBDs. Esto conlleva a distintas implicaciones funcionales que explican como miembros de una misma familia pueden regular genes con distintas funciones; tal es caso de las familias CRP y GntR, en las que sus miembros participan en la regulación de multiples funciones (Beker, y otros, 2001; Rigali, y otros, 2004) .

De esta manera, resulta relevante evaluar la contribución de los PaDos en su asociación con los DBDs, así como preguntar si estos PaDos permiten la plasticidad de respuesta a los cambios ambientales en los distintos organismos (Aravind, y otros, 2005; Rigali, y otros, 2002; Martínez-Antonio, y otros, 2006; Salgado, y otros, 2007).

Actualmente, los FTs han sido ampliamente estudiados en términos de sus DBDs y se han realizado análisis para determinar cómo han evolucionado y divergido en las distintas superfamilias; sin embargo, se sabe muy poco acerca de la diversidad de sus dominios adicionales (PaDos) y cuál es su contribución tanto evolutiva como a nivel funcional en el repertorio de los FTs. En este sentido, Rigali et al 2002 identificaron cuatro subfamilias en la familia GntR con base en el PaDo y que correlacionan con las funciones de los genes regulados (Rigali, y otros, 2002).

Adicionalmente, Babu y Teichmann en 2003 identifican diversos tipos de dominios (PaDos) en los FTs en *E. coli*, tales como aquellos que reconocen a pequeñas moléculas, involucrados en interacciones proteína-proteína, dominios enzimáticos y con función desconocida (Babu, y otros, 2003).

4. JUSTIFICACION

Con base en lo descrito previamente, consideramos que el análisis de los PaDos en conjunto con el dominio de unión al DNA en bacterias no ha sido del todo considerado, por lo que este trabajo contribuirá de manera significativa a ampliar nuestro conocimiento de los FTs desde una perspectiva integral (PaDo – DBD), así como para entender sus implicaciones funcionales y/o evolutivas. Consideramos que el repertorio de FTs está siendo significativamente modificado por la presencia de los PaDos que le proveen versatilidad en el reconocimiento de las moléculas señal. Además, de que este trabajo puede ser extendido a otras superfamilias de unión al DNA, a otras familias funcionales e incluso hacia otros dominios celulares.

5. HIPÓTESIS

Los PaDos le proveen versatilidad a los factores de transcripción en el reconocimiento de las moléculas señal y en procesos de multimerización

6. OBJETIVO GENERAL

- Analizar la distribución de los PaDos en el repertorio de la superfamilia *winged Helix-Turn-Helix* en genomas bacterianos

7. OBJETIVOS PARTICULARES

- Identificar los FTs con wHTH
- Identificar los dominios adicionales o *Partner Domains* en bacterias completamente secuenciadas
- Evaluar la distribución y abundancia de los FTs
- Evaluar la distribución y abundancia de los PaDos.

8. ESTRATEGIA EXPERIMENTAL

En una primera etapa se identificaron los FTs con wHTH en las bacterias *E. coli* K-12 (Gama-Castro, y otros, 2008)*,* *B. subtilis* 168 (Sierro, y otros, 2008) y *C. glutamicum* (Baumbach, y otros, 2008), utilizando diferentes fuentes de información y herramientas bioinformáticas. Se identificaron 295 secuencias de proteínas asociadas a FTs que poseen wHTH a partir de la base de datos SUPERFAMILY (version 2009) (Gough, y otros, 2001). De esta colección de FTs, 107 pertenecen a *E. coli*, 118 de *B. subtilis* y 70 a *C. glutamicum*. Cada grupo de FTs fue clasificado en términos de familias, utilizando las asignaciones depositadas en la base de datos de Superfamily y PFAM (Finn, y otros, 2009). Paralelamente se asignaron las superfamilias a las que pertenecen los dominios adicionales para cada organismo. En resumen, se identificaron 176 dominios adicionales que se encuentran distribuidos en 27 superfamilias, 15 asociadas a *E. coli*, 23 a *B. subtilis* y 11 a *C. glutamicum.* Figura 3.

En una segunda etapa se recuperaron los reguladores de 670 genomas bacterianos con base a los FTs recopilados previamente y con base en matrices de peso asignadas (HMM[**]), así como con los FTs depositados en la base de datos DBD (Wilson, y otros, 2008). En este paso se obtuvo el repertorio total de FTs para cada organismo y se evaluó la contribución de los FTs con wHTH. Posteriormente, se identificó el repertorio de las familias wHTH y sus respectivos

[**] Hidden Markov model (HMM): un modelo estadístico que representa un sistema del proceso de Markov, que genera una secuencia observable, tales como una secuencia de aminoácidos. Se utiliza para representar un alineamiento múltiple de una familia de dominios y para encontrar parientes lejanos de una familia a través de un alineamiento de HMM (Moore, y otros, 2008)

PaDos. Para fines de la discusión de los resultados, sólo se consideraron 428 organismos no redundantes, es decir se excluyeron 242 cepas de especies idénticas. Así, los organismos analizados corresponden a 19 diferentes Phyla: Acidobacteria, Actinobacteria, Aquificales, Bacterioidetes, Chlamydia, Chlorobi, Chloroflexi, Cyanobacteria, Deinococcus-Thermus, Dictyoglomi, Elusimicrobia, Fusobacteria, Plantomycetes, Spirochaetes, Tenericutes, Thermotogae, Verrucomicrobia, Firmicutes y Proteobacteria. Debido a que la gran mayoría de los organismos secuenciados actualmente pertenecen a estos dos últimos Phyla se subdividieron en términos de sus clases, para Firmicutes en Clostridia y Bacillus y las Proteobacterias en Alpha, Beta, Gama, Épsilon y Delta

Para evaluar la distribución y abundancia de los DBD y sus PaDos correspondientes, estos se analizaron utilizando herramientas de *agrupamiento* utilizando el programa Mev (Multiexperiment viewer: http://www.tm4.org/mev) (Saeed, y otros, 2003). Para este análisis se construyeron tres matrices que determinan la presencia y ausencia de las familias, los PaDos y la relación entre ambos. Para ello evaluamos la presencia de al menos un miembro de la familia en un genoma representativo, independientemente de su abundancia, siguiendo una fórmula de abundancia relativa en un Phyla donde se evalúa de la siguiente manera: (número de FTs identificados o Pados) / (el número organismos representativos en el phyla); los valores de este resultado van de 1 (presencia) a 0 (ausencia) con estos datos se construyeron matrices de familias y PaDos que representan el promedio de su presencia en un phyla.

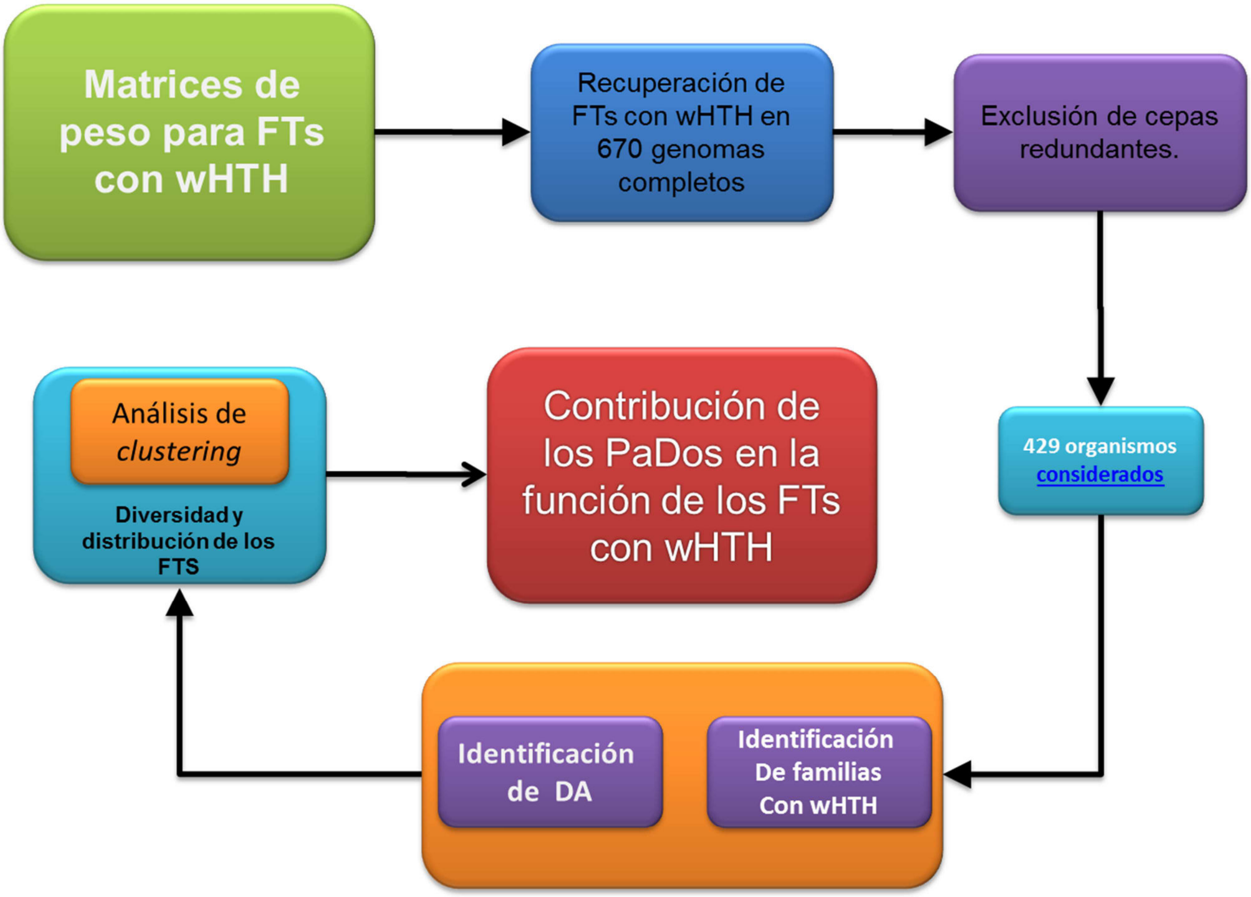

Matrices de peso para FTs con wHTH
Recuperación de FTs con wHTH en 670 genomas completos
Exclusión de cepas redundantes.
429 organismos considerados
Identificación de DA
Identificación De familias Con wHTH
Análisis de *clustering*
Diversidad y distribución de los FTS
Contribución de los PaDos en la función de los FTs con wHTH

Figura 3. Diagrama de flujo del método: Estrategia Experimental.

9. RESULTADOS

9.1. Análisis de las Familias de wHTH en bacterias modelo

Basados en modelos HMMs se identificaron 23 familias diferentes en los tres organismos modelo, 16 en *E. coli,* 20 en *B. subtilis* y 16 en *C. glutamicum* (Figura 4). De estas familias, 13 son comunes a los tres organismos, tales como LysR-like, GntR-like y Lrp/AsnC-like, entre otras, sugiriendo que existen procesos metabólicos comunes que son regulados por dichas familias de FTs.

Asimismo, se identificaron tres familias comunes en sólo dos organismos, la familia *Transcription factor E/IIe-alpha* en *E. coli-B. subtilis*; la familia *Hypothetical protein F93* en *E. coli-C. glutamicum* y la familia *Heat-inducible transcription repressor HrcA, N-terminal domain* en *B. subtilis-C. glutamicum*, lo que sugiere posibles regulación de procesos comunes entre dichos organismos.

Por otra parte, se observaron siete familias específicas a los diferentes organismos, tales como *ModE,* y *Rrf2 (Pfam 02082)* en *E. coli*; *PurR, Z-DBD, Rex, N-terminal domain y DNA-binding protein Mj223* en *B. subtilis* y *Penicillinase repressor* en *C. glutamicum.* Consideramos que estas familias estarían regulando procesos específicos asociados a los ambientes donde habitan dichos organismos.

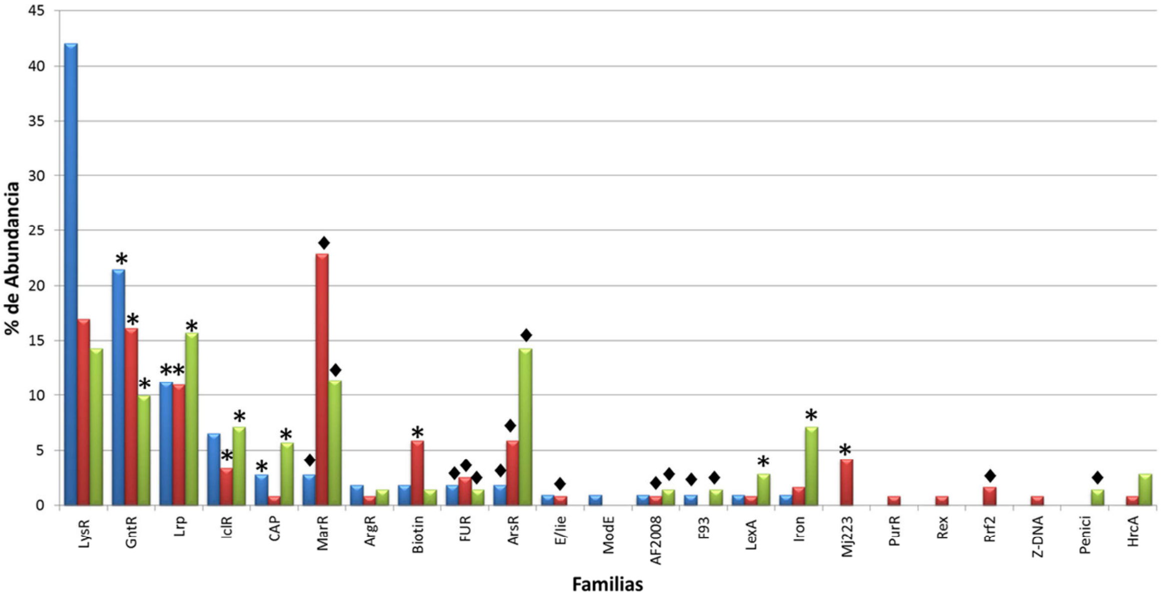


Figura 4: Distribución de las familias de FTs en los organimos modelo. En Azul están las familias identificadas en *E. coli*, en Rojo las de *B. subtilis* y en Verde las asociadas a *C. glutamicum*. Los asteriscos indican las familias que son tanto monodominio como multidominio, y los diamantes indican las familias que son exclusivamente monodominio. En el eje de las X, se presentan las familias identificadas. En el eje de las Y, se muestra el % de abundancia de dichas familias.

9.2. Identificación de Dominios adicionales (PaDos) asociados a los wHTH

En el análisis asociado a los FTs (previamente descrito) identificamos que el 30% de las proteínas asociadas a regular la expresión genética en *E. coli* son monodominio, 55% en *B. subtilis* y 54% en *C. glutamicum*.

En el figura 5 se muestra el porcentaje de dominios asociados a FTs con wHTH. Los datos están ordenados con base en la abundancia de los dominios por genoma en orden creciente (teniendo los menos abundantes a la izquierda y los más abundantes a la derecha). Existen PaDos que se encuentran mayormente representados y estos a su vez están asociados a las familias más representadas. En total identificamos en este primer análisis 27 PaDos diferentes. Donde 23 se encuentran en *B. subtilis* (siendo este el mayormente asociado a diversos dominios), 15 en *E. coli* y 11 en *C. glutamicum.* Adicionalmente, se identificaron dominios presentes en una sola especie, para *E. coli* (3) y *B. subtilis* (11), mientras que para *C. glutamicum* no se encontraron dominios exclusivos. También se encontraron 4 dominios que están asociados a dos organismos, dos en *E. coli-B. subtilis*, 1 *E. coli-C. glutamicum* y 1 *B. subtilis-C. glutamicum.* Finalmente, nueve dominios comunes a los tres organismos fueron identificados, tales como *PBP-II*, *NagB, GAF* y *Dimeric*.

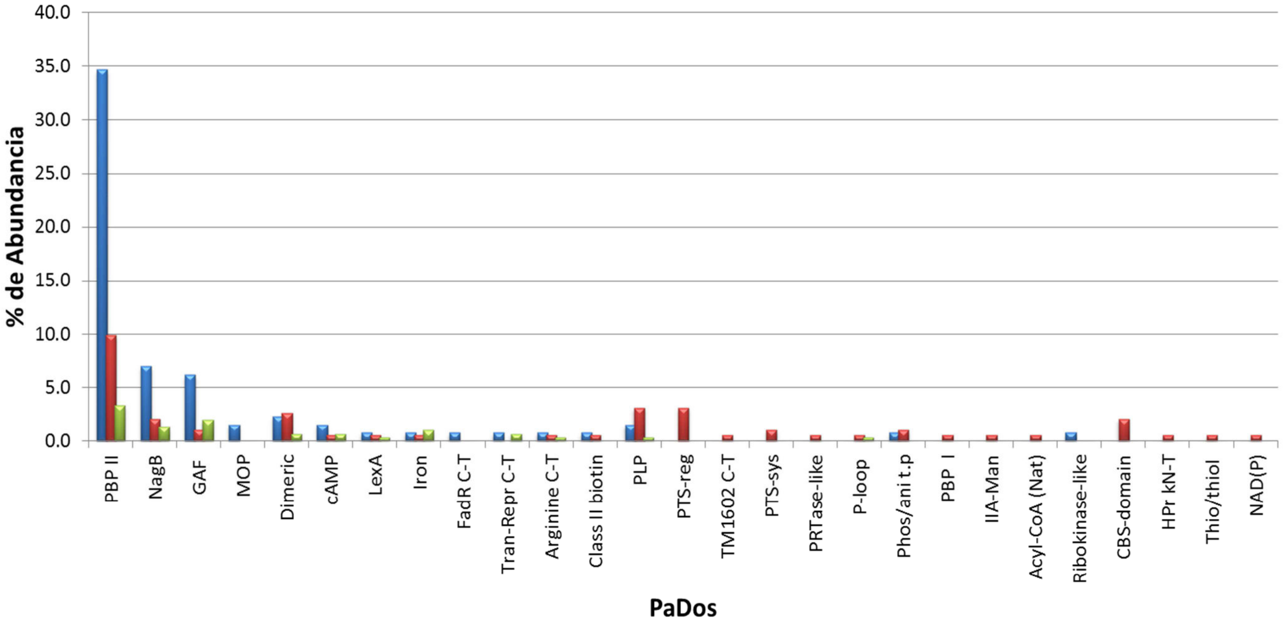


Figura 5: distribución de los dominios adicionales. En Azul se representan los grupos asociados a *E. coli*, en Rojo a *B. subtilis* y en Verde a *C. glutamicum*.

9.3. Asociación de Familias de reguladores y sus dominios adicionales (PaDos)

En la figura 6 se evaluó la relación entre los diferentes dominios adicionales y las familias de FTs. Los dominios adicionales se encuentran asociados a familias con mayor número de miembros (Figura 6D) y las familias pequeñas se encuentran asociadas con pocos dominios o con ningún dominio. También se observa que algunas de las familias se encuentran asociadas con el mismo tipo de dominio estructural en los tres organismos, como es el caso de LysR, GntR, Lrp, CAP, Iron, y ArgR en *B. subtilis* y *C. glutamicum*, lo que sugiere que este tipo de familias estarían regulando procesos comunes en los organismos donde se han identificado.

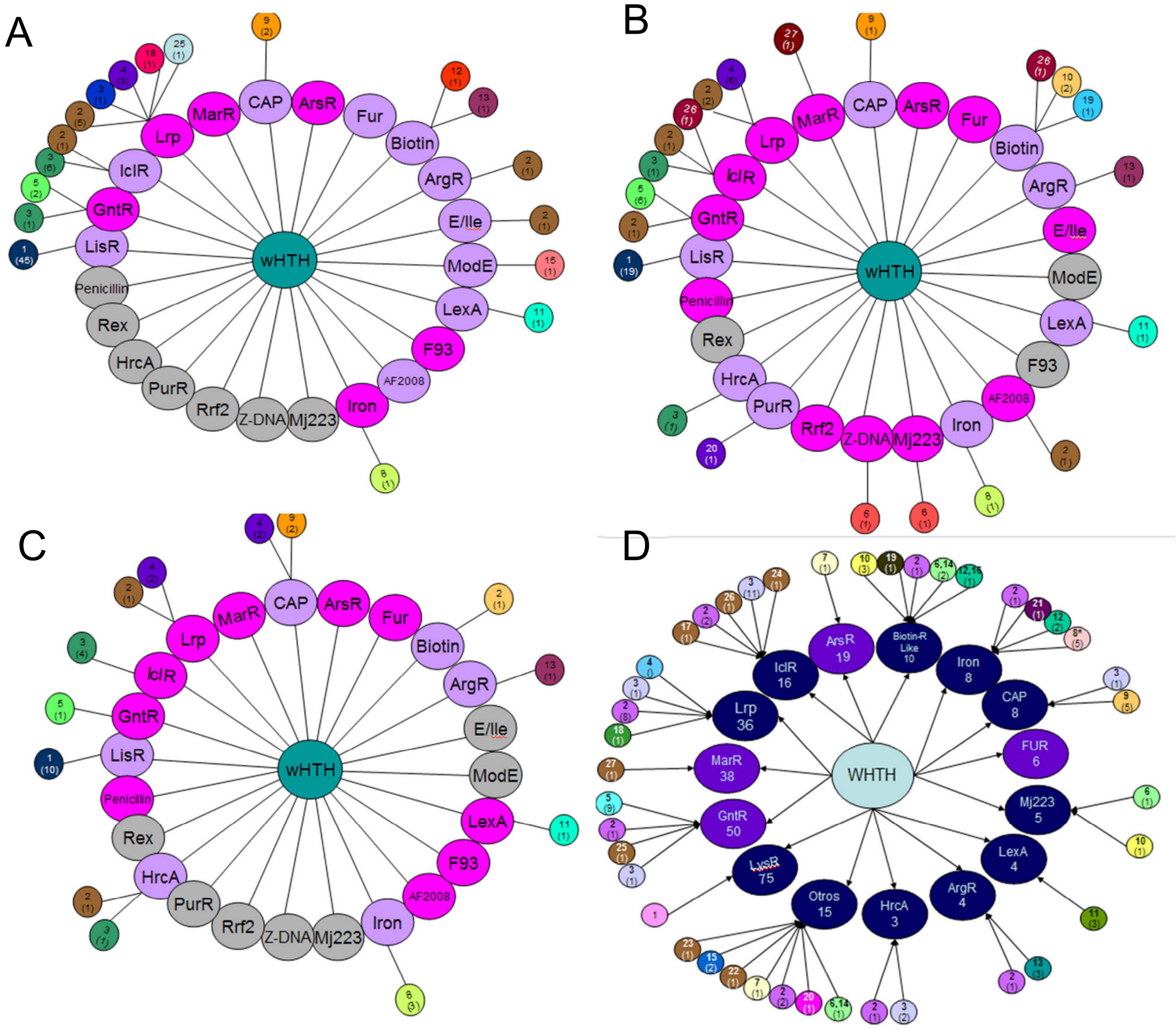


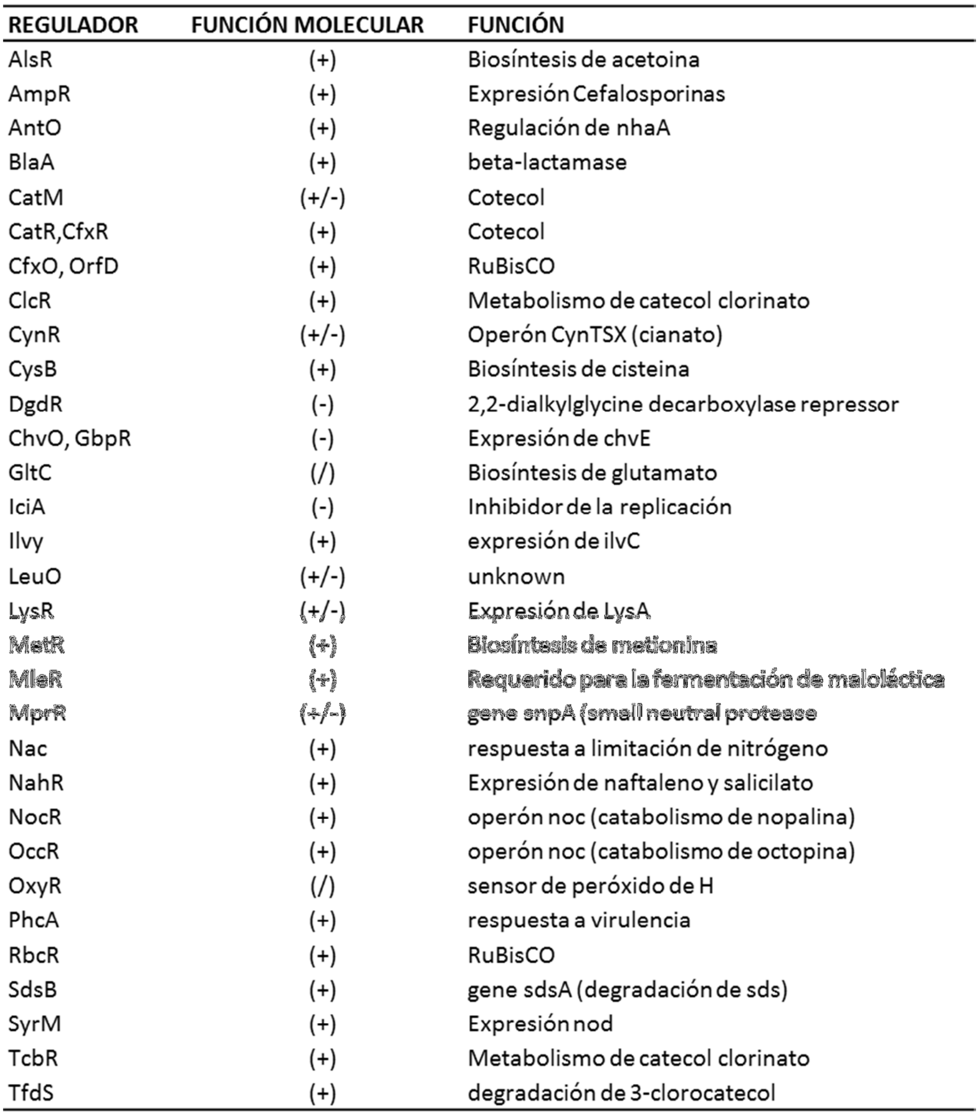
Figura 6. Diversidad de familias con wHTH observadas y sus PaDos. A) *E. coli*, B) *B. subtilis* y C) *C. glutamicum*. En rosa se muestran aquellas familias monodominio, en gris las familias que no están presentes en el organismo. Los círculos pequeños representan los diferentes dominios asociados a cada familia el número indica el dominio, el número entre paréntesis el número de miembros. En D) podemos observar una perspectiva general de los tres organismos, así como el número de miembros por familia.

En resumen, en esta primera etapa se analizó de manera individual y general los FTs asociados a la superfamilia wHTH de *E. coli*, *B. subtilis* y *C. glutamicum*, así como sus dominios adicionales. Basado en estos análisis, se identificaron 23 diferentes familias y 27 dominios adicionales. Adicionalmente, se identificaron relaciones entre los tres organismos modelo a nivel de sus FTs con wHTH y sus dominios adicionales, aunque debido a que se analizaron tres organismos, estas observaciones serian válidas en estos bacterias y en organismos filogenéticamente cercanos.

Así, en el repertorio de familias se identificaron diversos dominios adicionales sobre-representados, como es el caso de PBP II o *Periplasmic Binding Protein-like type II* que representa el mas del 30% del total de los dominios adicionales identificados. Lo interesante de este dominio es que se encuentra preferencialmente asociado a miembros de la familia LysR, haciéndola una familia con poca diversidad de dominios adicionales. Este tipo de dominios presenta una estructura similar a los dominios de proteínas que se unen al periplasma y su función principal es la unión de ligandos. Así mismo, se identificaron 13 dominios que en conjunto con PBP II representan el 94% de los dominios adicionales. Sin embargo, la familia LysR representa una de las familias más abundante en *E. coli* y esta se encuentra involucrada en distintos procesos como se muestra en la Tabla 4 (schell, 1993).

Debido a que nuestro interés es evaluar la contribución de los dominios adicionales en el repertorio de los reguladores con wHTH desde una perspectiva genómica, un análisis exhaustivo se llevó a cabo en 428 organismos no redundantes. Este enfoque nos permitió entender cómo están organizados estos FTs en términos de su distribución y abundancia y cuál podría ser su papel en el mecanismo de la regulación de la expresión genética. Observamos que las relaciones entre dominios de unión a DNA y los dominios adicionales o PaDos se mantienen relativamente constantes al analizar un mayor número de organismos. Por ejemplo, familias como LysR se encuentra asociada en su mayoría a un solo tipo de PaDo (PBP II); o familias como MarR o Fur en las que observamos que la mayoría de las proteínas no presentan un PaDo asociado.

Tabla 4. Propiedades de algunos miembros de la familia LysR



REGULADOR	FUNCIÓN MOLECULAR	FUNCIÓN
AlsR	(+)	Biosíntesis de acetoina
AmpR	(+)	Expresión Cefalosporinas
AntO	(+)	Regulación de nhaA
BlaA	(+)	beta-lactamase
CatM	(+/-)	Cotecol
CatR,CfxR	(+)	Cotecol
CfxO, OrfD	(+)	RuBisCO
ClcR	(+)	Metabolismo de catecol clorinato
CynR	(+/-)	Operón CynTSX (cianato)
CysB	(+)	Biosíntesis de cisteina
DgdR	(-)	2,2-dialkylglycine decarboxylase repressor
ChvO, GbpR	(-)	Expresión de chvE
GltC	(/)	Biosíntesis de glutamato
IciA	(-)	Inhibidor de la replicación
Ilvy	(+)	expresión de ilvC
LeuO	(+/-)	unknown
LysR	(+/-)	Expresión de LysA
MetR	(+)	Biosíntesis de metionina
MleR	(+)	Requerido para la fermentación de maloléctica
MprR	(+/-)	gene snpA (small neutral protease
Nac	(+)	respuesta a limitación de nitrógeno
NahR	(+)	Expresión de naftaleno y salicilato
NocR	(+)	operón noc (catabolismo de nopalina)
OccR	(+)	operón noc (catabolismo de octopina)
OxyR	(/)	sensor de peróxido de H
PhcA	(+)	respuesta a virulencia
RbcR	(+)	RuBisCO
SdsB	(+)	gene sdsA (degradación de sds)
SyrM	(+)	Expresión nod
TcbR	(+)	Metabolismo de catecol clorinato
TfdS	(+)	degradación de 3-clorocatecol

Tomado de Schell 1993

9.4. MANUSCRITO

En esta sección se describen los resultados del análisis en 428 genomas bacterianos.

Rivera-Gomez N, Segovia L, Perez-Rueda E. 2011. The transcriptional machinery in bacteria is modified by the repertoire of Partner Domains associated to Transcription Factors. *Microbiology.* 157, 2308-2318.

DOI 10.1099/mic.0.050617-0

Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria

Nancy Rivera-Gómez, Lorenzo Segovia and Ernesto Pérez-Rueda

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

Correspondence
Ernesto Pérez-Rueda
erueda@ibt.unam.mx

Received 19 April 2011
Revised 26 May 2011
Accepted 27 May 2011

The ability of bacteria to deal with diverse environmental changes depends on their repertoire of genes and their ability to regulate their expression. In this process, DNA-binding transcription factors (TFs) have a fundamental role because they affect gene expression positively and/or negatively depending on operator context and ligand-binding status. Here, we show an exhaustive analysis of winged helix–turn–helix domains (wHTHs), a class of DNA-binding TFs. These proteins were identified in high proportions and widely distributed in bacteria, representing around half of the total TFs identified so far. In addition, we evaluated the repertoire of wHTHs in terms of their partner domains (PaDos), identifying a similar trend, as with TFs, i.e. they are abundant and widely distributed in bacteria. Based on the PaDos, we defined three main groups of families: (i) monolithic, those families with little PaDo diversity, such as LysR; (ii) promiscuous, those families with a high PaDo diversity; and (iii) monodomain, with families of small sizes, such as MarR. These findings suggest that PaDos have a very important role in the diversification of regulatory responses in bacteria, probably contributing to their regulatory complexity. Thus, the TFs discriminate over longer regions on the DNA through their diverse DNA-binding domains. On the other hand, the PaDos would allow a great flexibility for transcriptional regulation due to their ability to sense diverse stimuli through a variety of ligand-binding compounds.

INTRODUCTION

Bacteria are unicellular organisms that have a ubiquitous distribution. In recent years, more than 1000 organisms from diverse phylogenetic divisions have been completely sequenced, showing that the genomic organization resulting in the contemporary systems is a product of diverse evolutionary events, such as gene expansion, gene loss and lateral gene transfer. In this regard, two important factors have been identified as being responsible for the plasticity of the genomes, i.e. the gene repertoire and the regulatory mechanisms (Bengtsson, 2004; Lynch & Conery, 2003; Lynch, 2006). In general, gene regulation at the transcription initiation level in bacteria is primarily mediated by sigma (σ) factors and by DNA-binding transcription factors (TFs) (Browning & Busby, 2004). σ Factors provide the specificity for promoter recognition and DNA melting needed for transcription initiation (Gruber & Gross, 2003; Ishihama, 2000; Wösten, 1998), although they perform these functions only when bound to the RNA polymerase. In contrast, TFs affect gene expression positively and/or negatively depending on operator context and ligand-binding status (Martínez-Antonio *et al.*, 2006; Miroslavova & Busby, 2006; Wall *et al.*, 2004).

Comparative bacterial genome analyses have shown that TFs may vary considerably in abundance and distribution (Aravind *et al.*, 2005; Levine & Tjian, 2003; Madan Babu *et al.*, 2006). In this regard, diverse studies have suggested that the abundance of TFs increases with increasing organismal complexity (Brown *et al.*, 2002; Changizi, 2001; Levine & Tjian, 2003; van Nimwegen, 2003; West & Brown, 2005), and the different proportions of these regulatory elements suggest that interplay among them defines the complexity of a regulatory network.

Although TFs are related to a wide diversity of functions, including differentiation, DNA restoration and cellular maintenance, and information on a large number of functional descriptions has accumulated, many questions remain unanswered, mainly associated with bacterial regulatory network organization and the repertoire of TFs. In this regard, various authors have described that bacteria share common principles of gene regulation across large phylogenetic distances, such as in *Escherichia coli* and *Bacillus subtilis* (Janga & Pérez-Rueda, 2009; Moreno-Campuzano *et al.*, 2006).

In this work, we identified the DNA-binding TFs that belong to the winged helix–turn–helix (wHTH) domain in

Abbreviations: DBD, DNA-binding domain; PaDo, partner domain; TF, transcription factor; wHTH, winged helix–turn–helix domain.

Three supplementary tables and three supplementary figures are available with the online version of this paper.

668 sequenced bacterial genomes representing a large diversity of divisions, lifestyles and genome sizes. wHTH domains have been classified as extensions of HTH domains, which are characterized by the presence of a third α-helix and an adjacent β-sheet and are central components in DNA binding. The recognition helix binds as in the regular HTH motifs, and the extra secondary structural elements provide additional contacts with the DNA backbone (Brennan & Matthews, 1989a, b). This structure has been identified in almost all micro-organisms, from Bacteria to Archaea, and is composed of diverse families, such as the catabolite gene activator (CAP) family, the heat shock and E2F/DP TFs, and the Ets domain family, among others (Brennan, 1993). In addition to the DNA-binding domain (DBD), TFs usually contain additional domains, called partner domains (PaDos), that are associated with diverse functions, such as protein–protein interactions, ligand-binding and/or catalytic activities (Madan Babu & Teichmann, 2003). This kind of structural organization is in agreement with previous reports which suggested that about two-thirds of proteins in prokaryotes are multidomain proteins (Tordai *et al.*, 2005).

Thus, TFs are two-headed proteins, with a DBD and a PaDo. DBDs have been widely used to classify TFs into families (Kummerfeld & Teichmann, 2006; Pérez-Rueda *et al.*, 2004; Pérez-Rueda & Janga, 2010). In contrast, few analyses are available on the additional domains, or PaDos, despite their importance in the regulatory response. PaDos have been associated with diverse functions, such as allosteric regulation of TFs across binding to a wide variety of functional compounds, in protein–protein interactions, or with enzymic properties (Madan Babu & Teichmann, 2003), and they are a fundamental link to environmental conditions and the functional conformational changes in the regulators (Taraban *et al.*, 2008).

In order to evaluate the abundance and distribution across all bacterial taxonomic divisions, the repertoire of wHTH TFs was analysed in terms of their domain organization. We evaluated the PaDos that are involved in DNA-binding activity, because relatively few of them have been explored in regulatory families, such as in the GntR family, for which four subfamilies have been identified that correlate with the functions of the regulated genes (Rigali *et al.*, 2002, 2004). From a global perspective, diverse types of PaDos, such as those associated with small-molecule binding, protein–protein binding and enzymic domains, and those with unknown function, have been identified in the regulatory network of *E. coli* K-12 (Madan Babu & Teichmann, 2003). The results obtained here provide insights into the functional and evolutionary constraints imposed on the expansion patterns of the TFs with a wHTH in bacteria. We believe that an improved understanding of the evolution of the transcriptional regulatory machinery across bacterial genomes will improve our knowledge about the evolutionary constraints that play a role in the formation of regulatory networks.

METHODS

Genome sequences. The complete list of genomes evaluated was obtained from the website ftp://ftpncbi.nlm.nih.gov/genomes/bacteria. We considered annotated genes as those with open reading frames that encode predicted protein sequences (the proteome) in all bacteria.

Identification of TFs. In order to identify the repertoire of TFs in the sequenced bacterial genomes, we used a combination of information sources and bioinformatics tools. As a first step, we identified and evaluated 295 wHTH TFs in three bacterial models, *E. coli* K-12, *B. subtilis* and *Corynebacterium glutamicum*, from three different databases, RegulonDB v 6.0 (Gama-Castro *et al.*, 2008), DBTBS v 5.0 (Sierro *et al.*, 2008) and CoryneRegnet v 4.0 (Baumbach, 2007), and their domain assignations were obtained from the Superfamily database (version 25 April 2010). From these, 107 TFs belong to *E. coli*, 118 to *B. subtilis* and 70 to *C. glutamicum*. These TFs clustered into families according to their PFAM assignations (Finn *et al.*, 2008), and their domain organization was defined, leaving 176 PaDos, clustered in 27 superfamilies, with 15 in *E. coli*, 23 in *B. subtilis* and 11 in *C. glutamicum.*

In the second stage, TFs associated with the wHTH domain were identified in 668 complete bacterial genomes, based on specific Hidden Markov model searches and from the regulators deposited in the DBD and Superfamily databases (Kummerfeld & Teichmann, 2006). For the purposes of our study, 428 nonredundant organisms were considered. The organisms classified were from 19 phyla: *Acidobacteria*, *Actinobacteria*, *Aquificales*, *Bacterioidetes*, *Chlamydia*, *Chlorobi*, *Chloroflexi*, *Cyanobacteria*, *Deinococcus*, *Thermus*, *Dictyoglomi*, *Elusimicrobia*, *Fusobacteria*, *Plantomycetes*, *Spirochaetes*, *Tenericutes*, *Thermotogae*, *Verrucomicrobia* and *Firmicutes*, and also *Proteobacteria*. Because of the large diversity and numerous organisms that have been completely sequenced and are associated with *Firmicutes* and *Proteobacteria*, their classes were also considered: *Bacillus* and *Clostridium* for *Firmicutes*, and *Alpha*-, *Beta*-, *Delta*-, *Epsilon*- and *Gammaproteobacteria* (Supplementary Table S1, available with the online version of this paper). In this work, we refer to nonredundant genomes as representative bacterial species, as previously defined by Janga & Moreno-Hagelsieb (2004). In brief, if there are diverse strains of the same species, a representative genome is considered; however, the order of elimination follows the importance of certain species as model organisms (such as *E. coli* K-12 and/or *B. subtilis*), and then the order of importance follows the highest number of genes having orthologues across phyla. For instance, *Mycobacterium avium* strain 104 is representative of diverse *Mycobacterium* strains (*M. avium paratuberculosis*, *M. bovis*, *M. bovis* BCG Pasteur 1173P2, *M. leprae* and *M. smegmatis* MC2 155) (Supplementary Table S1).

Clustering of families of regulatory factors and PaDos. In order to evaluate the distribution and abundance of TF families and their corresponding PaDos across 428 nonredundant bacterial genomes, a hierarchical complete linkage clustering algorithm was applied with correlation uncentred as the similarity measure. Analyses were performed using the program Mev (multiexperiment viewer; http://www.tm4.org/mev). In order to determine the relative abundance of the families of TFs and their associated PaDos by phylum, we calculated the fraction of genomes in the group that had at least one member versus the number of representative organisms. Thus, the following formula was considered: relative abundance by phylum=(total no. of TFs or PaDos identified)/(total no. of representative organisms by phylum). Thus, a value of 1 corresponds to presence and 0 represents absence. Because our aim was to evaluate the taxonomical distribution of TFs and PaDos, 24 taxonomical divisions were considered.

RESULTS

The proportion of wHTHs contributes significantly to the total repertoire of TFs in bacteria

In order to gain insights into the commonalities and differences in gene regulation between bacterial species from the perspective of TFs, we compared the repertoires of TFs identified in 428 nonredundant bacterial sequenced genomes by using diverse bioinformatics tools. From this analysis, we found that the wHTH comprises 48 % of the total TFs identified in bacteria, being the most abundant superfamily of DNA-binding structures described so far in this cellular domain (Table 1 and Supplementary Table S1). Indeed, alternative DNA-binding structures have been identified in minor proportions, such as the lambda repressor, homeodomain-like domain and C-terminal effector domain. This result not only correlates with the fact that the DBDs are associated with TFs, and in particular wHTHs are among the most ancient domains, probably derived from a relatively small set of folds (Aravind & Koonin, 1999; Madan Babu & Teichmann, 2003; Pérez-Rueda & Collado-Vides, 2001), but also shows that the wHTHs have been highly successful domains in nature.

Based on this apparent overrepresentation of wHTH proteins in the repertoire of TFs, we evaluated their abundance in the context of genome size. In this regard, it was previously described that TFs follow a quadratic distribution, i.e. large genomes contain a high proportion of TFs and, vice versa, small genomes have a small repertoire of TFs (Pérez-Rueda *et al.*, 2004, 2009). Thus, we asked whether wHTH abundance correlates with the number of open reading frames. From this analysis, we found that wHTHs follow a similar distribution to the total repertoire of TFs, suggesting that this superfamily contributes significantly to this distribution (see Fig. 1). Indeed, the wHTH corresponds to around 50 % of the total of TFs per genome and in some organisms more than 70 % of the total repertoire of TFs corresponds to wHTH proteins, such as in *Mycoplasma genitalium*, *Prochlorococcus marinus* NATL1A, *Bordetella bronchiseptica* RB50 and *Thermosynechococcus elongatus* BP-1. Based on these results, we consider that the abundance of TFs and wHTHs may be associated with organismal diversity, i.e. organisms with free-living lifestyles and a large genome size contain a major proportion of transcriptional regulators, compared with parasitic or symbiont organisms with their small genome sizes, as described previously (Pérez-Rueda *et al.*, 2009). For instance, diverse free-living bacteria, such as *Burkholderia* sp. 383, with its large genome size, contain a high proportion of wHTH TFs. Recent studies suggest that organisms with free lifestyles require a large proportion of regulatory proteins as a consequence of an increase in the genome size, which also increases the number of putative regulatory interactions (Croft *et al.*, 2003). Alternatively, organisms with smaller genome sizes, such as *Mycoplasma* species, contain a small proportion of wHTH TFs, between one and six wHTH TFs. This proportion of TFs correlates with the fact that symbionts and/or parasitic obligate bacteria have substantially reduced genome sizes. Indeed, in some organisms with a substantial reduction in the gene repertoire, we were not able to identify TFs or wHTH TFs, such as in the *Buchnera*, *Rickettsia* and *Wolbachia* genera,

Table 1. Superfamilies associated with DNA-binding TFs in bacteria

Superfamily	Total TFs (%)	No. of families	Total TFs with PaDos	No. of different PaDos
wHTH DBD	30642 (46.33)	39	17304	79
Homeodomain-like	17110 (25.87)	10	8452	75
Lambda repressor-like DBD	6437 (9.73)	6	2875	49
C-terminal effector domain of the bipartite response regulators	5057 (7.64)	3	3747	39
Putative DBD	2085 (3.15)	2	299	23
IHF-like	1782 (2.69)	1	0	0
Nucleic acid-binding proteins	1294 (1.95)	1	21	3
Glucocorticoid receptor-like (DBD)	473 (0.71)	2	473	2
Ribbon–helix–helix	434 (0.65)	4	64	6
AbrB/MazE/MraZ-like	248 (0.37)	3	38	5
KorB DBD-like	230 (0.34)	1	176	5
TrpR-like	107 (0.16)	1	–	–
ACT-like	68 (0.10)	1	63	1
Flagellar transcriptional activator FlhD	64 (0.096)	1	–	–
Haemolysin expression modulating protein H	35 (0.052)	1	–	–
A DBD in eukaryotic transcription factors	25 (0.037)	1	8	2
DBD	24 (0.036)	1	18	1
p53-like transcription factors	6 (0.009)	1	5	4
DBD of *Mlu*I-box binding protein MBP1	5 (0.008)	1	–	–

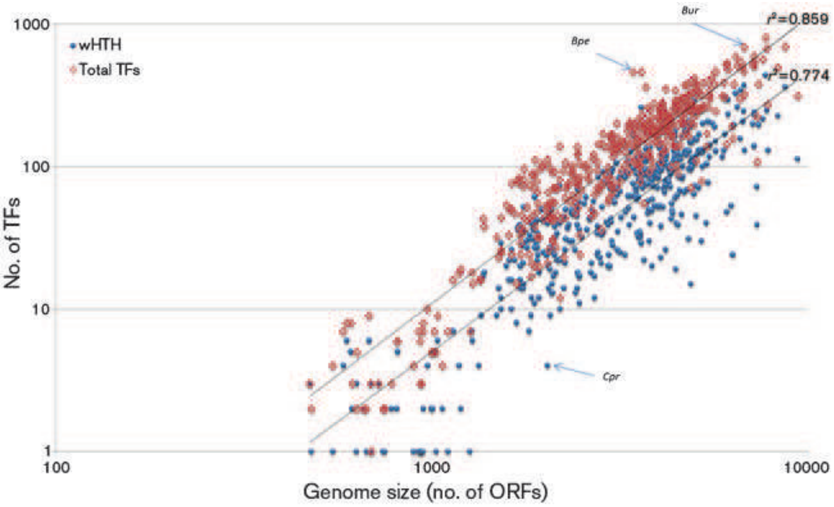


Fig. 1. Distribution of wHTH TFs and total TFs. *Bordetella pertussis* (*Bpe*), *Burkholderia* sp. 383 (*Bur*) and '*Candidatus* Protochlamydia amoebophila' (*Cpr*) are included in this illustration as reference points. On the *x* axis (log scale), genomes are sorted by size from smallest to largest. On the *y* axis (log scale) are the corresponding numbers of TFs. A linear regression was calculated using the Pearson correlation (r^2) between the number of genes and the total number of TFs. Each dot represents a bacterial genome; wHTHs (blue) and total TFs (red) are indicated.

suggesting that they exhibit alternative regulatory mechanisms beyond TFs. Although the wHTH contributes significantly to the total TFs, probably following a similar path of duplication events to the rest of the genes in Bacteria (as illustrated in Fig. 1), we identified organisms such as *Bordetella pertussis* and '*Candidatus* Protochlamydia amoebophila' in which the wHTH does not represent the most abundant superfamily. In fact, in these organisms the homeodomain-like superfamily is the most abundant DBD associated with TFs, suggesting alternative means for regulation of gene expression.

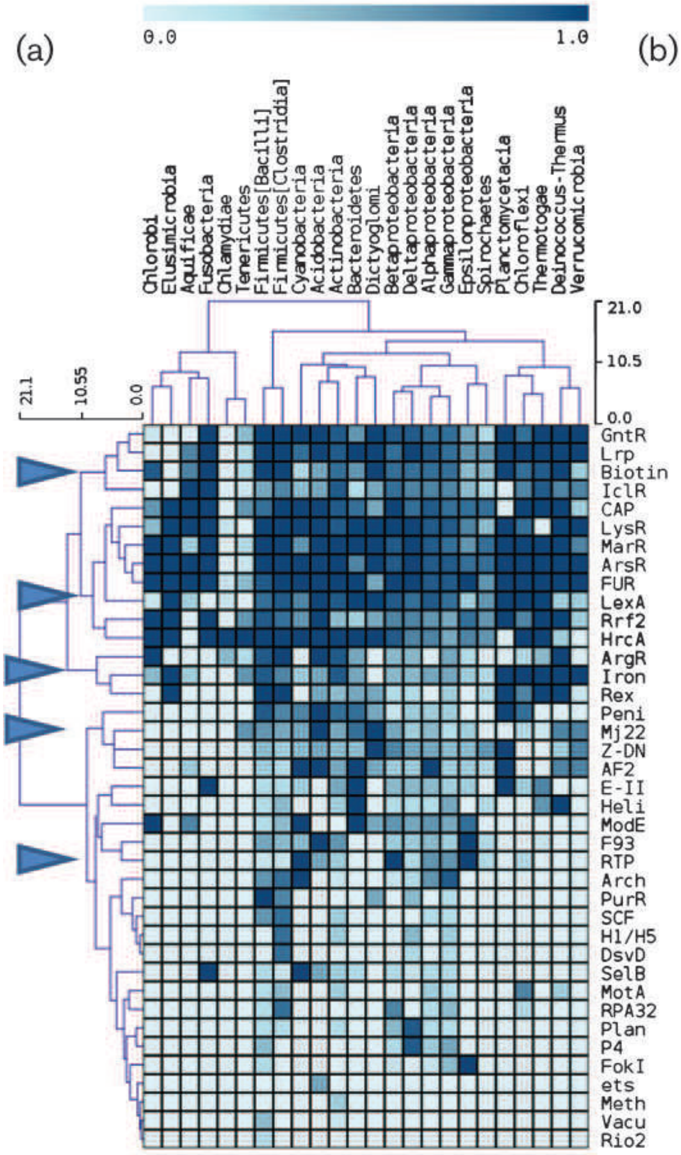
Another important question that emerges from the distribution of wHTHs is whether the abundance of these kinds of proteins in all the genomes also reflects a large proportion of families. In this sense, we found that the abundance of wHTH proteins is derived preferentially by duplication events more than by the existence of different families. Therefore, we classified the repertoire of wHTH TFs into 39 different families, and their distribution among all bacteria was evaluated (Supplementary Table S2, available with the online version of this paper). From this analysis, we describe in the next two sections the most evident results.

(i) We found that the abundance of TFs in larger genomes does not necessarily involve diversity in the repertoire of families, but it does suggest an increase in the size of the family, i.e. whereas there is a large proportion of TFs as a consequence of genome size, the number of different families per genome is almost constant, from 1 to 26, with an average of 14 families in bacteria. However, in some cases we identified a large number of families, such as for the bacterium *Saccharopolyspora erythreaea*, a filamentous soil microbe used for industrial-scale production of the antibiotic erythromycin (Oliynyk *et al.*, 2007), for which 26 different wHTH families were identified, the maximum number of families per organism. A probable identification of abundant families might suggest diverse duplication events in Bacteria, whereas small families would suggest gene loss, lateral gene transfer or invention *de novo* (Supplementary Table S1).

(ii) Seven families include 80 % of the total TFs with wHTHs: ArsR, MarR, LysR, biotin, CAP, GntR and Lrp (Fig. 2). From these families, LysR represents an interesting evolutionary group, because it contains the most abundant family of TFs identified so far as well as a large proportion of proteins with dual activity (repressor and activator proteins). This family is mainly responsible for the regulation of basic and ancient physiological processes, such as amino acid biosynthesis, associated with the last common ancestor of Bacteria, Archaea and Eukarya (Hernández-Montes *et al.*, 2008). In contrast, small families were also identified, such as iron, ArgR and LexA, proposed to be essential under standard growth conditions and in maintaining DNA integrity in *E. coli.* Based on these data, a family's abundance not only suggests ancient evolutionary events in Bacteria but also reveals a limit in the number of wHTH families in all bacteria. Thus, it seems that there is a limit of expansion for all families in bacteria, independent of the genome size and an increase in the number of duplication events associated with each family, suggesting that the TF repertoire in bacteria is associated mainly with events of duplication, recombination and lateral gene transfer. An interesting observation is that in cell division in *Chlamydiae*, during which massive gene loss events have been identified, the HrcA family of TFs was exclusively identified. HrcA is a small family that contains a large proportion of negative regulators of class I heat-shock genes (*grpE–dnaK–dnaJ* and *groELS* operons) that prevent heat-shock induction of these operons (Fischer *et al.*, 2002). We suggest that proteins of this family may be associated with alternative functions beyond heat-shock induction, playing an important role in the adaptation of parasitic bacteria to their hosts.

A universal pattern of distribution is observed in bacterial wHTH families

In order to evaluate the distribution of the 39 TF families in bacteria, we analysed their distribution across 24 taxonomical divisions (Fig. 2; see also Supplementary Tables S1



(b)

Family	No. of proteins	Proportion
GntR	4039	0.131
Lrp	2548	0.083
Biotin	1020	0.033
IclR	984	0.032
CAP	1186	0.039
LysR	9736	0.317
MarR	3852	0.125
ArsR	2080	0.068
FUR	967	0.031
LexA	478	0.016
Rrf2	449	0.015
HrcA	358	0.012
ArgR	343	0.011
Iron	423	0.014
Rex	138	0.004
Peni	339	0.011
Mj22	275	0.009
Z-DN	329	0.011
AF2	143	0.005
E-II	169	0.006
Heli	70	0.002
ModE	210	0.007
F93	193	0.006
RTP	46	0.001
Arch	32	0.001
PurR	67	0.002
SCF	10	0.0003
H1/HS	7	0.0002
DsvD	7	0.0002
SelB	64	0.002
MotA	19	0.001
RPA32	12	0.0004
Plan	81	0.003
P4	12	0.0004
FokI	27	0.001
ets	1	0.00003
Meth	1	0.00003
Vacu	7	0.0002
Rio2	1	0.00003

Fig. 2. Clustering of the co-occurrence of wHTH TFs in all the bacterial genomes. (a) A hierarchical centroid linkage clustering algorithm was applied with Manhattan metric distance as the similarity measure and complete linkage metric distance (Eisen *et al.*, 1998). In the upper section, the names of the 24 bacterial divisions are indicated. The names of the 39 families are also shown. The clusters are indicated with arrows. The families and their relative abundance levels are indicated.

and S2). The relative abundance of families was calculated by phylum, with a value of 0 representing absence and 1 representing presence. Based on this analysis we identified a cluster of eight families widely distributed in all the bacterial divisions: the LexA, LysR, FUR, ArsR, MarR, CAP, HrcA and Rrf2 families. An interesting finding was that the seven most abundant families described previously were included in this cluster, together with families that are less abundant. All these families regulate a plethora of functions, such as amino acid metabolism (LysR), carbon source assimilation (CAP, Körner *et al.*, 2003; Maddocks & Oyston, 2008), resistance to diverse stresses (MarR, Ellison & Miller, 2006; LexA, Shimoni *et al.*, 2009; HrcA, Ellison & Miller, 2006), cysteine biosynthesis and benzoate degradation (Rrf2, Even *et al.*, 2006; Peres & Harwood, 2006) and metal assimilation (FUR, Pennella & Giedroc, 2005) and resistance (ArsR, Busenlehner *et al.*, 2003). It is probable that all these families could have been present in the last common ancestor of Bacteria and as a consequence the bacterial cenancestor would have a high capability to contend with diverse, challenging environments and also be a self-supporting system.

In a second cluster, families with a large distribution pattern, except in small organisms, were identified: GntR, Lrp, biotin and IclR families. These families regulate biotin synthesis, carbon source assimilation and amino acid biosynthesis, among other processes.

A third cluster with a low distribution pattern was identified and included the ArgR, Iron and Rex families. These families regulate genes associated with carbon source assimilation, arginine biosynthesis, iron uptake and responses to changes in the cellular NADH/NAD$^+$ redox state, respectively, and include few members per organism.

Finally, a high diversity of families with erratic distribution patterns were included in diverse clusters, such as Rio2, associated with *Bacillales*, Ets, associated with *Actinobacteria*, and MetH, exclusively associated with *Actinobacteria*. In addition, we identified probable families with lateral gene transfer events, including Vacu, which is associated with *Bacillales* and *Actinobacteria*. This finding suggests that diverse lineage-specific TFs are involved in specific and important processes, such as sporulation in bacilli, or in some specific amino acid biosynthesis routes. It is interesting that the absence of TFs for several important amino acid biosynthesis routes in *B. subtilis* and other *Firmicutes* may have been complemented by the invention of novel regulatory mechanisms, such as transcription attenuation (Gollnick *et al.*, 2005; Merino & Yanofsky, 2005; Rodionov *et al.*, 2004). Indeed, a large diversity of regulatory mechanisms beyond TFs was recently described, including antisigma factors, RNAs and protein–protein interactions, among others (Martínez-Núñez et *al.*, 2010).

Multidomain proteins are highly abundant in bacterial TFs

At the present time, the considerable diversity of sequenced bacterial genomes available to the scientific community offers an invaluable source of information for evaluating the abundance and diversity of the repertoire of regulatory proteins controlling gene expression at the level of transcription initiation. These proteins can be analysed to evaluate their influence on bacterial adaptation and responses to environmental stimuli. Therefore, in order to evaluate the contributions of these domains in the total set of proteins identified as TFs, the domain repertoire beyond the DBD was analysed. From this analysis we identified different groups based on domain architectures, i.e. 57 % of the TFs exhibited more than one structural domain (the DBD and PaDos), whereas 43 % of the total repertoire was associated only with the DBD (Table 2 and Supplementary Table S3, available with the online version of this paper). The monodomain proteins can be further subdivided into two categories: the first one includes proteins for which more than 94 % of the protein is occupied by the DBD and the second category includes proteins for which the DBD covers only 50 % of the sequence. These latter proteins may exhibit additional domains not identified using structural data. Therefore, in Fig. 3 we present the distribution of multidomain TFs in all the bacterial genomes. From this illustration, it is evident that the abundance of multidomain TFs follows a similar distribution to total TFs, i.e. their abundance correlates with the genome size. For instance, organisms with small genomes may contain a lower proportion of multidomain proteins than do larger genomes. This result reinforces the notion that small genomes are associated with stable environments, where a limited number of TFs are necessary to regulate gene expression. In contrast, larger genomes contain a great proportion of multidomain TFs, suggesting that these domains contribute to functional adaptations to environmental changes. Based on this analysis, the diversity and abundance of TF families and their PaDos would contribute significantly to the regulatory diversity.

Therefore, 79 PaDos were identified in the whole collection of wHTH TFs associated with bacteria, and in order to evaluate the distribution of these PaDos in bacteria, 22 divisions were analysed. From this analysis we identified diverse groups. The most representative clusters are shown in Fig. 4 (see also Supplementary Figs S1 and S2, available with the online version of this paper). The first group contains four different PaDos, such as periplasmic-binding protein-like II (PBP II), cAMP-binding domain-like, GAF domain-like and LexA/signal peptidase domains. PBP II and cAMP domains are associated with the LysR and CAP families, whereas GAF is associated with diverse families, such as as IclR, HrcA, Plan, FUR and others. The LexA/signal peptidase domain is associated with the LexA family. It is important to mention that the first three PaDos are highly abundant in all bacteria, representing 70 % of the total PaDos, and they are also related to large TF families. These findings suggest that these PaDos are very successful in all the bacteria and are intimately related to large families of TFs.

The second cluster includes eight different PaDos: dimeric α- and β-barrel (dimeric), PLP-dependent transferases (PLP), NagB/RpiA/CoA transferase-like (NagB), C-terminal domain of transcriptional repressors, C-terminal domain of arginine repressor, class II aaRS and biotin synthetases, iron, and NAD(P)-binding Rossmann fold domains. From these domains, dimeric, PLP and NagB have been identified as highly abundant, representing 15 % of total PaDos. An interesting observation for these domains is that they exhibit a similar distribution pattern to the PaDos included in the first cluster, except they are absent in parasites, symbionts and, in general, in small genomes, suggesting probable gene loss events. The third cluster is integrated by MOP-like, *S*-adenosyl-L-methionine-dependent methyltransferases and acyl-CoA *N*-acyltransferases (Nat), which have been mainly identified in the divisions *Proteobacteria* and *Acidobacteria*. In the subsequent clusters, we identified diverse PaDos, including the PRTase-like, ribokinase-like and CBS domains and the putative transcriptional regulator TM1602, C-terminal domain, constrained to *Firmicutes* and *Fusobacteria*. Finally, we found the rhodanese/cell cycle control phosphatase, fatty acid-responsive transcription factor FadR C-terminal, SIS, phosphorytosine, Bet v1-like, and nucleotidyltransferase domains clustered together.

Table 2. Monodomain and multidomain families identified in this work

Family name	Multidomain TFs (%)*	No. of different PaDos*	Most abundant PaDo	Proportion of PaDos (%)*
DsvD	0	0	–	0
Ets	0	0	–	0
H1/H5	0	0	–	0
Meth	0	0	–	0
RTP	0	0	–	0
Vacu	0	0	–	0
Biot	51	24	Class_II	21
ArsR	14	22	PBP II	27
Mj22	6	13	GAF	29
E-II	47	15	cAMP	36
MarR	5	22	Acyl-CoA	37
MotA	42	5	Actin-like_ATPase	40
Iron	83	16	Iron	41
SelB	39	8	GAF	41
Heli	61	9	GAF	43
SCF	30	2	NAD(P)	50
FUR	1	5	PLP	50
Arch	56	4	GAF	56
Lrp	67	28	Dimeric	58
Plan	49	6	GAF	59
P4	100	4	P-loop	62
Rrf2	2	2	PLP	63
Z-DNA	54	9	GAF	64
RPA32	50	3	PLP	67
F93	2	2	Actin-like_ATPase	67
AF2	25	7	NagB	70
ArgR	92	7	Arginine_repressor	75
ModE	70	3	MOP	79
GntR	23	22	PLP	79
HrcA	97	9	GAF	81
IclR	88	20	GAF	83
Peni	5	2	SIS domain	88
CAP	92	15	cAMP	89
LexA	73	9	LexA/Signal	92
PurR	99	4	PRT	94
Rex	99	2	NAD(P)	99
Rio2	100	1	NAD(P)	100
LysR	99	7	PBP II	99
FokI	63	1	NagB	100

*0 represents monodomain families.

These domains have been identified in only low proportions, mainly in *Proteobacteria.*

In summary, we identified six PaDos that represent 84 % of the total domains identified in bacterial TFs (PBP II, GAF, dimeric, PLP, NagB and cAMP domains), many of which are universally distributed in bacteria and intimately associated with specific families. It is also interesting that these PaDos are found in abundant families, reinforcing their probable roles in basic physiological proceses, such as PBP II being exclusively associated with the largest family of TFs identified so far, LysR (see Supplementary Fig. S2). This finding suggests that PaDos and wHTH domains probably coevolved, based on their pattern distributions. Alternatively, we identified some PaDos as probably lineage specific, such as the PTS-reg, TM1602, PTS-sys, PRTase-like, CBS, HPr kN-T and Thio/thiol domains. These latter domains were exclusively identified in *Firmicutes.* We suggest that these PaDos may be the consequence of invention *de novo*, because of their specific distribution in *Firmicutes.* In addition, to reinforce these previous observations, we asked if the PaDos were specifically associated with wHTH TFs. To obtain further insights into the specific and general associations of these domains and wHTHs, we evaluated all DNA-binding structures reported so far and classified them as homeodomain-like, lambda

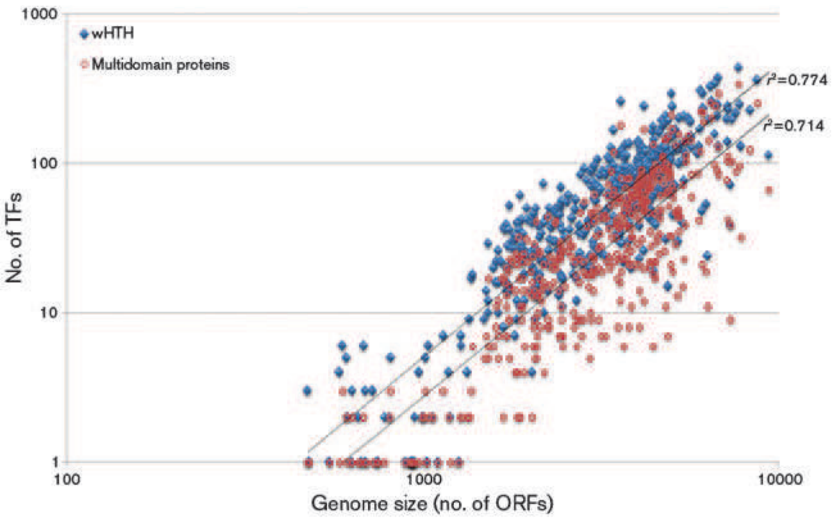


Fig. 3. Distribution of multidomain TFs in bacterial genomes. On the *x* axis (log scale), genomes are sorted by size from smallest to largest. On the *y* axis (log scale) are the corresponding numbers of TFs. A linear regression was calculated using the Pearson correlation (r^2) between the number of genes and the total number of TFs. Each dot represents a bacterial genome; wHTHs (blue) and multidomain proteins (red) are indicated.

repressor or nucleic acid-binding domains (Table 1) in order to identify the distributions of these domains. Based on this analysis, we determined that almost 40 % of the total PaDos associated with wHTHs are exclusive to this class of structural domains, suggesting that these domains have been preferentially recruited by wHTHs. A similar finding has been identified in other superfamilies, for instance, 30 % of the total PaDos are lamba specific, and 33 % of the total PaDos are homeodomain-like specific (see Supplementary Fig. S3, available with the online version of this paper).

The diversity of PaDos defines three main groups of families

In order to evaluate the diversity of TFs in terms of their structural domains, we defined three main groups of families: (i) monodomain families, those families that exhibit only the DBD; (ii) monolithic families, in which most of the protein members exhibit the DBD and a PaDo, usually in the same domain; and (iii) promiscuous families, those families with a large diversity of domains. We next describe each of these categories in more detail.

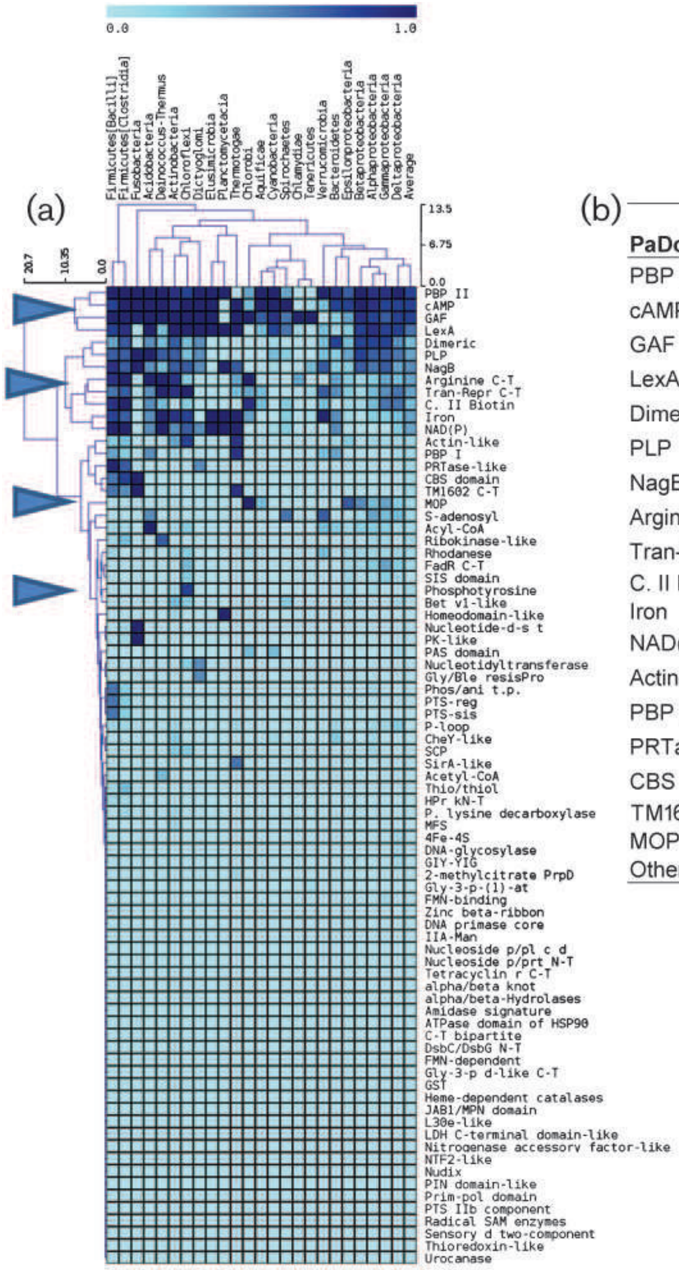
(i) Monodomain families. Twelve families were considered monodomain, including MarR, FurR and ArsR (Table 2). An interesting observation is that most of the families included in this category contain proteins of small sizes, around 150 amino acids in length, where the DBD covers most of the sequence.

(ii) Multidomain families. Multidomain families can be further subdivided into two groups, based on the presence of multiple domains: (ii) monolithic, where at least 80 % of their members exhibit a predominant PaDo associated with the DBD [such as the LysR family, in which the PBP II is present in 99 % of its members (Table 2 and Supplementary Figs S1 and S2)]; and (iii) promiscuous families, such as GntR or Lrp, for which diverse domains are associated with the DBD. Therefore, the diversity in the repertoire of regulatory proteins with wHTHs is influenced by the organization and combination with the PaDos, and the families can be classified into three main groups. There are monodomain families, where the multimerization and ligand-binding sites are included in the DBD, and multidomain families, which can be divided into two groups. In monolithic families, the DBD has undergone a similar evolution process to the PaDos with few recombination events, such as the LysR family. Indeed, the domain organization is also conserved, where the DBD is located in the N terminus whereas the PLP II is located in the C terminus. Both monodomain and monolithic families include the most abundant families identified in the repertoire of TFs. Finally, we identified a large proportion of families that do not represent the most abundant families but that include a large diversity of PaDos.

Conclusions

What are the roles of the diversity and distribution of TFs in the regulatory plasticity of bacteria? The answer to this question is important in order to determine the contribution of these regulatory families in the evolution of these organisms and in the context of their responses to diverse environmental stimuli. Thus, based on the repertoire of DNA-binding TFs associated with the wHTH domains in 668 completely sequenced bacterial genomes, representing adaptative designs for different lifestyles, we have attempted to gain an understanding of the relationships between the DBD and PaDo distribution patterns involved in the modelling of transcriptional regulatory networks.

We have shown that TF families expand or contract in a lineage-specific manner to adapt to the varied environmental needs of the organisms. Similar trends have been observed in previous comparative studies on TF families in plants versus animals and at the level of taxa (Coulson



(b)

PaDo	No. of PaDos	Proportion
PBP II	9799	0.543
cAMP	1169	0.065
GAF	1503	0.083
LexA	332	0.018
Dimeric	1029	0.057
PLP	914	0.051
NagB	763	0.042
Arginine C-T	238	0.013
Tran-Repr C-T	233	0.013
C. II Biotin	173	0.010
Iron	212	0.012
NAD(P)	161	0.009
Actin-like	235	0.013
PBP I	56	0.0031
PRTase-like	144	0.008
CBS	102	0.006
TM1602 C-T	51	0.002824859
MOP	186	0.010
Others	801	0.048

Fig. 4. Clustering of the co-occurrence of PaDos in bacterial genomes. (a) A hierarchical centroid linkage clustering algorithm was applied with Manhattan metric distance as the similarity measure and complete linkage (Eisen *et al.*, 1998). PaDo abundance levels are indicated. In the upper section, the names of the 24 bacterial divisions are indicated (as in Fig. 2). The names of the 79 PaDos are also shown. (b) The PaDos and their relative abundance levels are presented (see also Supplementary Table S3).

et al., 2001). A more general perspective regarding lineage-specific expansion of protein families and the implications for the diversification of organisms has also been described for eukaryotic species (Lespinet *et al.*, 2002). In this regard, the abundance levels of families and their domain organizations were evaluated in bacteria. From this analysis, we identified that TF families have preferentially suffered diverse expansion events more than invention *de novo*, i.e. in all organisms evaluated in our study there was a similar distribution of families, but the members were different among the families. Alternatively, PaDo families are present at between 1 and 26 per organism, contributing significantly to the diversity of the regulatory machinery. Finally, three main categories of families were defined on the basis of their domain architecture: monodomain, monolithic and promiscuous. We suggest that the interplay of all these elements, TF abundance, recombination of DBDs, and PaDos, and also duplication events, allows bacteria to adapt to changing environmental conditions and shows that they are models of regulatory networks.

ACKNOWLEDGEMENTS

This work was supported by the Mexican Science and Technology Research Council (CONACYT) with a doctoral scholarship (164293) awarded to N. R.-G. E. P.-R. was financed by grants from the DGAPA-UNAM (grant nos IN-217508 and IN-209511). We would like to thank Miguel Angel Ramirez, Tomas Villaseñor and Yalbi Martinez for critical reading of the manuscript.

REFERENCES

Aravind, L. & Koonin, E. V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. *Nucleic Acids Res* **27**, 4658–4670.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. (2005). The many faces of the helix-turn-helix domain: transcription regulation and beyond. *FEMS Microbiol Rev* **29**, 231–262.

Baumbach, J. (2007). CoryneRegNet 4.0 – a reference database for corynebacterial gene regulatory networks. *BMC Bioinformatics* **8**, 429.

Bengtsson, B. O. (2004). Modelling the evolution of genomes with integrated external and internal functions. *J Theor Biol* **231**, 271–278.

Brennan, R. G. (1993). The winged-helix DNA-binding motif: another helix–turn–helix takeoff. *Cell* **74**, 773–776.

Brennan, R. G. & Matthews, B. W. (1989a). The helix–turn–helix DNA binding motif. *J Biol Chem* **264**, 1903–1906.

Brennan, R. G. & Matthews, B. W. (1989b). Structural basis of DNA-protein recognition. *Trends Biochem Sci* **14**, 286–290.

Brown, J. H., Gupta, V. K., Li, B. L., Milne, B. T., Restrepo, C. & West, G. B. (2002). The fractal nature of nature: power laws, ecological complexity and biodiversity. *Philos Trans R Soc Lond B Biol Sci* **357**, 619–626.

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

Busenlehner, L. S., Pennella, M. A. & Giedroc, D. P. (2003). The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev* **27**, 131–143.

Changizi, M. A. (2001). Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. *J Theor Biol* **211**, 277–295.

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

Croft, L. J., Lercher, M. J., Gagen, M. J. & Mattick, J. S. (2003). Is prokaryotic complexity limited by accelerated growth in regulatory overhead. *Genome Biol* **5**, P2.

Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A* **95**, 14863–14868.

Ellison, D. W. & Miller, V. L. (2006). Regulation of virulence by members of the MarR/SlyA family. *Curr Opin Microbiol* **9**, 153–159.

Even, S., Burguière, P., Auger, S., Soutourina, O., Danchin, A. & Martin-Verstraete, I. (2006). Global control of cysteine metabolism by CymR in *Bacillus subtilis*. *J Bacteriol* **188**, 2184–2197.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H. R., Ceric, G., Forslund, K., Eddy, S. R. & other authors (2008). The Pfam protein families database. *Nucleic Acids Res* **36** (Database issue), D281–D288.

Fischer, B., Rummel, G., Aldridge, P. & Jenal, U. (2002). The FtsH protease is involved in development, stress response and heat shock control in *Caulobacter crescentus*. *Mol Microbiol* **44**, 461–478.

Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zavaleta, A., Peñaloza-Spinola, M. I., Contreras-Moreira, B., Segura-Salazar, J., Muñiz-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** (Database issue), D120–D124.

Gollnick, P., Babitzke, P., Antson, A. & Yanofsky, C. (2005). Complexity in regulation of tryptophan biosynthesis in *Bacillus subtilis*. *Annu Rev Genet* **39**, 47–68.

Gruber, T. M. & Gross, C. A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu Rev Microbiol* **57**, 441–466.

Hernández-Montes, G., Díaz-Mejía, J. J., Pérez-Rueda, E. & Segovia, L. (2008). The hidden universal distribution of amino acid biosynthetic networks: a genomic perspective on their origins and evolution. *Genome Biol* **9**, R95.

Ishihama, A. (2000). Functional modulation of *Escherichia coli* RNA polymerase. *Annu Rev Microbiol* **54**, 499–518.

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

Janga, S. C. & Pérez-Rueda, E. (2009). Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. *Comput Biol Chem* **33**, 261–268.

Körner, H., Sofia, H. J. & Zumft, W. G. (2003). Phylogeny of the bacterial superfamily of Crp–Fnr transcription regulators: exploiting the metabolic spectrum by controlling alternative gene programs. *FEMS Microbiol Rev* **27**, 559–592.

Kummerfeld, S. K. & Teichmann, S. A. (2006). DBD: a transcription factor prediction database. *Nucleic Acids Res* **34** (Database issue), D74–D81.

Lespinet, O., Wolf, Y. I., Koonin, E. V. & Aravind, L. (2002). The role of lineage-specific gene family expansion in the evolution of eukaryotes. *Genome Res* **12**, 1048–1059.

Levine, M. & Tjian, R. (2003). Transcription regulation and animal diversity. *Nature* **424**, 147–151.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. *Annu Rev Microbiol* **60**, 327–349.

Lynch, M. & Conery, J. S. (2003). The origins of genome complexity. *Science* **302**, 1401–1404.

Madan Babu, M. & Teichmann, S. A. (2003). Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. *Nucleic Acids Res* **31**, 1234–1244.

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

Maddocks, S. E. & Oyston, P. C. (2008). Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology* **154**, 3609–3623.

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

Martínez-Núñez, M. A., Pérez-Rueda, E., Gutiérrez-Ríos, R. M. & Merino, E. (2010). New insights into the regulatory networks of paralogous genes in bacteria. *Microbiology* **156**, 14–22.

Merino, E. & Yanofsky, C. (2005). Transcription attenuation: a highly conserved regulatory strategy used by bacteria. *Trends Genet* **21**, 260–264.

Miroslavova, N. S. & Busby, S. J. (2006). Investigations of the modular structure of bacterial promoters. *Biochem Soc Symp* (73), 1–10.

Moreno-Campuzano, S., Janga, S. C. & Pérez-Rueda, E. (2006). Identification and analysis of DNA-binding transcription factors in *Bacillus subtilis* and other Firmicutes – a genomic approach. *BMC Genomics* **7**, 147.

Oliynyk, M., Samborskyy, M., Lester, J. B., Mironenko, T., Scott, N., Dickens, S., Haydock, S. F. & Leadlay, P. F. (2007). Complete genome sequence of the erythromycin-producing bacterium *Saccharopolyspora erythraea* NRRL23338. *Nat Biotechnol* **25**, 447–453.

Pennella, M. A. & Giedroc, D. P. (2005). Structural determinants of metal selectivity in prokaryotic metal-responsive transcriptional regulators. *Biometals* **18**, 413–428.

Peres, C. M. & Harwood, C. S. (2006). BadM is a transcriptional repressor and one of three regulators that control benzoyl coenzyme A reductase gene expression in *Rhodopseudomonas* palustris. *J Bacteriol* **188**, 8662–8665.

Pérez-Rueda, E. & Collado-Vides, J. (2001). Common history at the origin of the position–function correlation in transcriptional regulators in archaea and bacteria. *J Mol Evol* **53**, 172–179.

Pérez-Rueda, E. & Janga, S. C. (2010). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. *Mol Biol Evol* **27**, 1449–1459.

Pérez-Rueda, E., Collado-Vides, J. & Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. *Comput Biol Chem* **28**, 341–350.

Pérez-Rueda, E., Janga, S. C. & Martínez-Antonio, A. (2009). Scaling relationship in the gene content of transcriptional machinery in bacteria. *Mol Biosyst* **5**, 1494–1501.

Rigali, S., Derouaux, A., Giannotta, F. & Dusart, J. (2002). Subdivision of the helix–turn–helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *J Biol Chem* **277**, 12507–12515.

Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B. & Titgemeyer, F. (2004). Extending the classification of bacterial transcription factors beyond the helix–turn–helix motif as an alternative approach to discover new *cis*/*trans* relationships. *Nucleic Acids Res* **32**, 3418–3426.

Rodionov, D. A., Dubchak, I., Arkin, A., Alm, E. & Gelfand, M. S. (2004). Reconstruction of regulatory and metabolic pathways in metal-reducing delta-proteobacteria. *Genome Biol* **5**, R90.

Shimoni, Y., Altuvia, S., Margalit, H. & Biham, O. (2009). Stochastic analysis of the SOS response in *Escherichia coli*. *PLoS ONE* **4**, e5363.

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** (Database issue), D93–D96.

Taraban, M., Zhan, H., Whitten, A. E., Langley, D. B., Matthews, K. S., Swint-Kruse, L. & Trewhella, J. (2008). Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. *J Mol Biol* **376**, 466–481.

Tordai, H., Nagy, A., Farkas, K., Bányai, L. & Patthy, L. (2005). Modules, multidomain proteins and organismic complexity. *FEBS J* **272**, 5064–5078.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. *Trends Genet* **19**, 479–484.

Wall, M. E., Hlavacek, W. S. & Savageau, M. A. (2004). Design of gene circuits: lessons from bacteria. *Nat Rev Genet* **5**, 34–42.

West, G. B. & Brown, J. H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. *J Exp Biol* **208**, 1575–1592.

Wösten, M. M. (1998). Eubacterial sigma-factors. *FEMS Microbiol Rev* **22**, 127–150.

Edited by: D. W. Ussery

9.5. Distribución de las superfamilias en bacterias con genomas completos

En esta sección, discutimos recientes resultados asociados a la expansión del análisis de los PaDos hacia 17 superfamilias de unión al DNA identificadas en los FTs. En los organismos analizados se identificaron entre 8 y 10 superfamilias por organismo, donde los Tenericutes o Chlamydias que son organismos parásitos y de genomas pequeños presentan un menor número de superfamilias; mientras que organismos como *Burkholderia vietnamiensis* G4 presentan 14 de las 17 superfamilias.

En la figura 7, se presenta un análisis por división taxonómica, donde los organismos con mayor número de superfamilias son las Proteobacterias de la división gama. Adicionalmente, podemos observar que algunas superfamilias están ampliamente distribuidas, tales como la winged HTH, *homeodomain-like*, *lambda repressor-like*, *Bipartite*, *Putative DBD* y la *Nucleic acid-binding*, que se identifican en más del 90% de los organismos analizados, sugiriendo un origen evolutivo muy antiguo. Estas seis superfamilias tienen una distribución en los genomas bacterianos similar a la que siguen los wHTH, es decir aumentan con respecto al tamaño del genoma, ver figura 8. El resto de las superfamilias presenta una distribución menor en comparación con el incremento del tamaño del genoma, debido a que se encuentran en pocas copias.

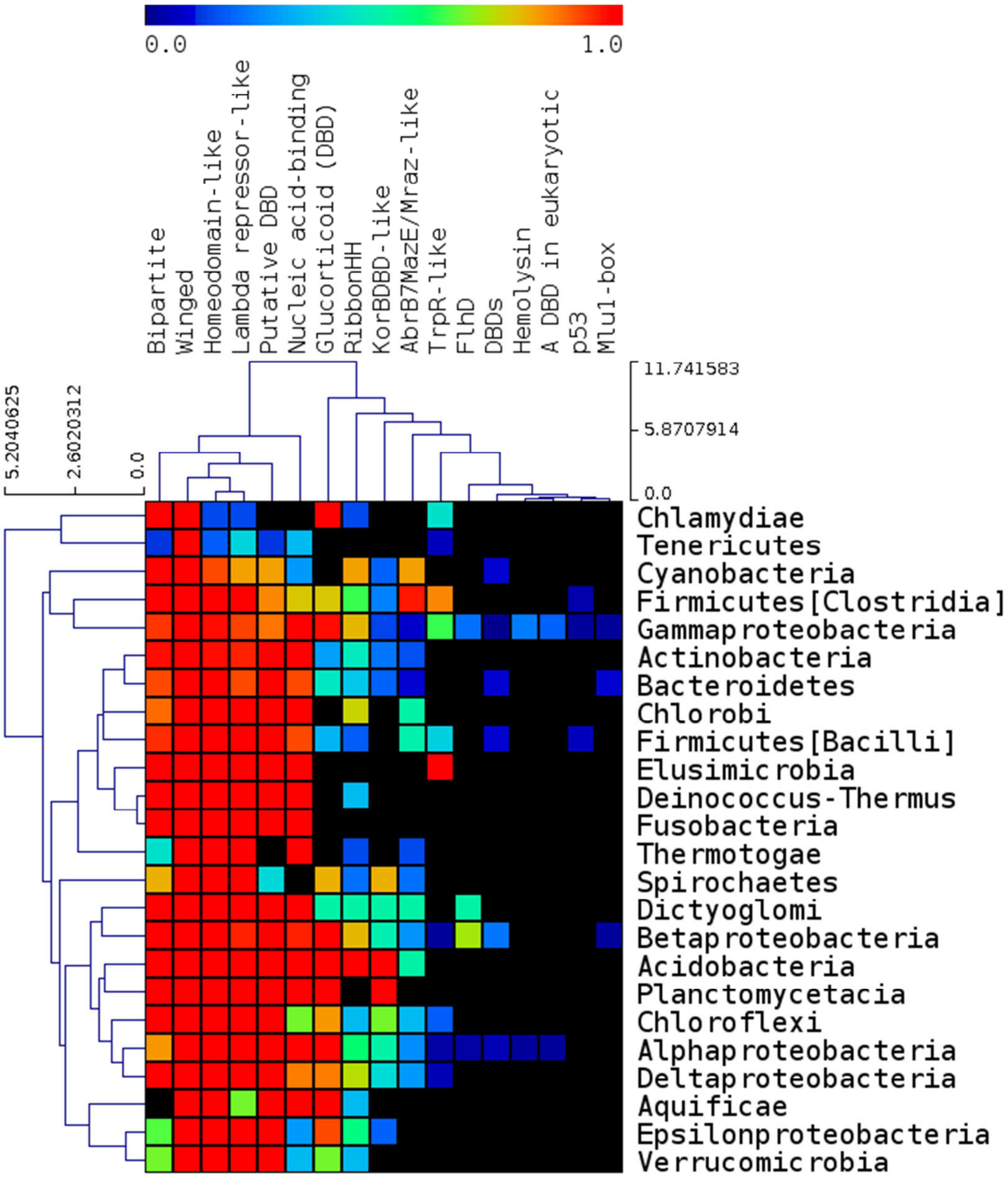


Figura 7. Distribución de las superfamilias en los diferentes taxa analizados.

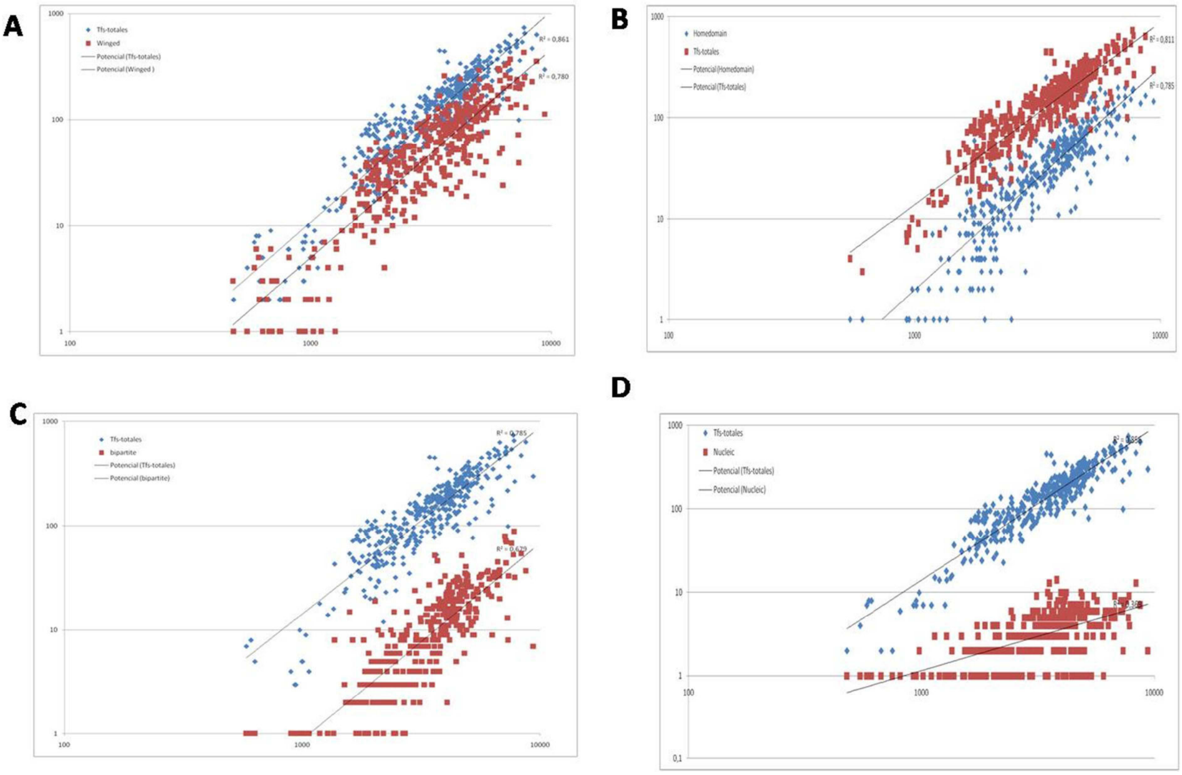


Figura 8. Distribución de las superfamilias con respecto al total de FTs en cada organismo. En la figura se encuentra graficado el tamaño del genoma (eje de las x) y el número de FTs (eje de las y). A) Superfamilia wHTH, B) Homeodomain-like, C) Bipartite y D) Nucleic.

9.6. Relación entre las superfamilias y sus PaDos

Las superfamilias previamente descritas han reclutado diferentes dominios de interacción con ligando y la relación en términos de proporción entre ambos dominios varía considerablemente. Como podemos observar en la figura 9, algunas superfamilias presentan una amplia diversidad de PaDos con respecto al número de familias, inclusive, mayor que la que presenta la superfamilia wHTH. En este contexto, es probable que las superfamilias con menor número de miembros, sus funciones sean compensadas con un incremento en el número de dominios, como se observa en la superfamilia *Bipartite* en la que solo se identifican tres familias relacionadas a 39 diferentes PaDos. Sin embargo, en este momento no se puede concluir si esta proporción está dada por el estilo de vida o por las funciones que realizan los miembros de cada una de una las superfamilias, por lo que continuamos investigando esta observación.

Por otro lado, identificamos PaDos específicos hacia cada superfamilia, tales como NagB, dimeric, Mop-like que están asociados a wHTH o ParB asociado a KorB, DnaK con Glucocorticoide; o PaDos que a pesar de estar compartidos entre varias superfamilias se asocian preferencialmente a una superfamilia, como es el caso de PBP I-lambda, PBP II-wHTH o TRP-like-Bipartite, sugiriendo que eventos específicos de reclutamiento podrían contribuir de manera sustancial a incrementar la diversidad de dominios y funciones reguladas por los FTs.

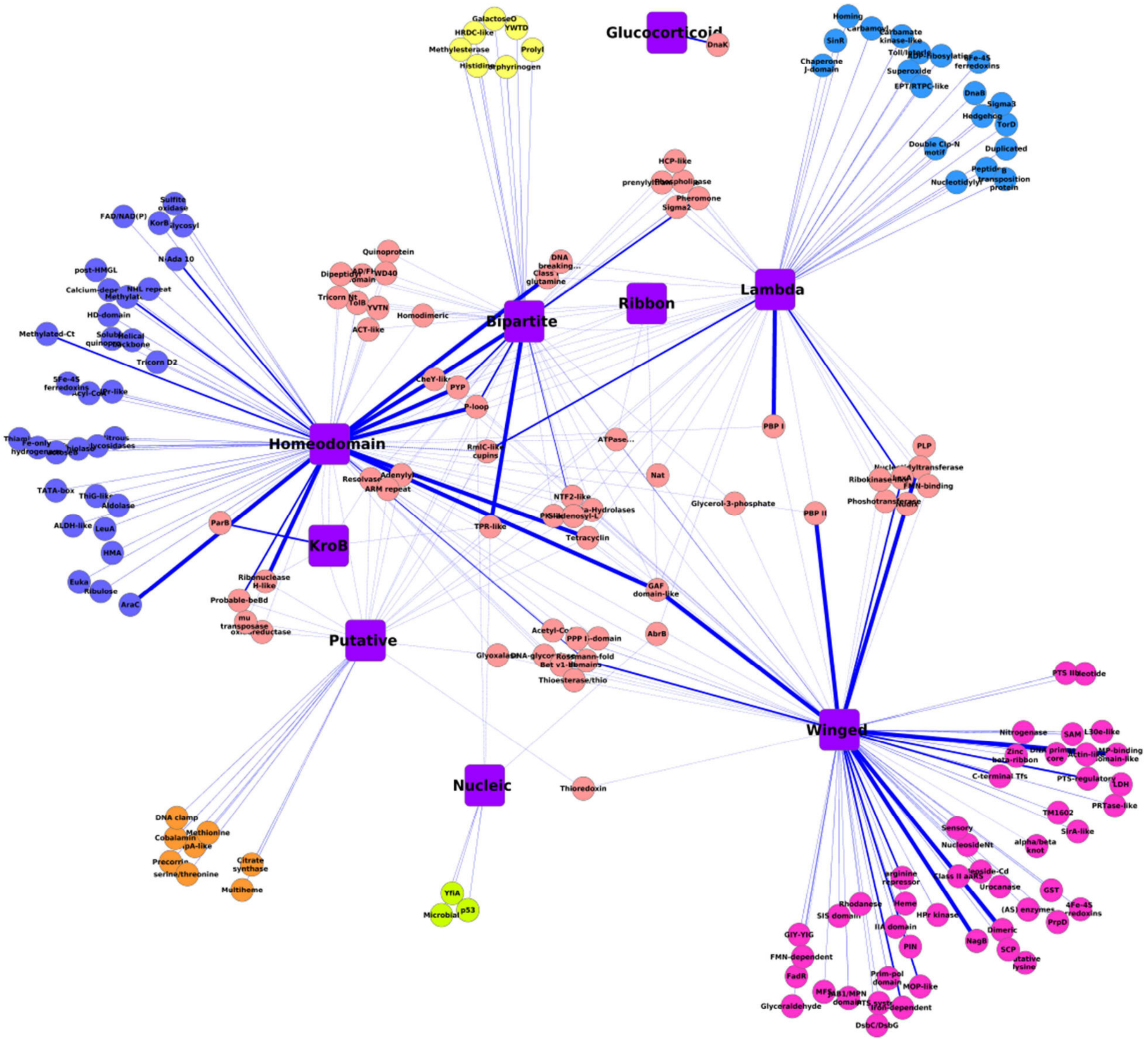


Figura 9. Relación entre superfamilias y PaDos. Las superfamilias se muestran en los rectángulos purpura, los círculos indican los PaDos, el color de cada circulo determina si pertenece a una superfamilia (magenta, azul marino, verde, naranja, amarillo, azul claro) o se asocia a diferentes superfamilias (rosa claro); las líneas delgadas representan una menor ocurrencia entre una superfamilia y un PaDo, por lo tanto, las líneas gruesas una mayor.

10. DISCUSIÓN Y CONCLUSIONES

El cambio de los hábitats y la evolución que recae en los organismos que los habitan, promueve el surgimiento de organismos cada vez más complejos, dando como resultado la generación de nuevas proteínas, que permiten la adaptación de estos organismos al ambiente que los contiene. Los mecanismos determinantes que producen estas nuevas proteínas son: la duplicación de los genes de viejas proteínas, la divergencia de estas secuencias para producir estructuras modificadas cuya utilidad conduce a su selección y en muchos casos la combinación con otros genes para modificar además sus propiedades. En la gran mayoría de los casos, la duplicación y la combinación de dominios involucra secuencias de DNA que codifican para uno o más dominios completos. Las proteínas pequeñas presentan solo un dominio con una función particular. Las combinaciones de dos o más dominios forman las grandes proteínas con funciones más sofisticadas (Moore, y otros, 2008). La teoría evolutiva postula que los homólogos observados en cada familia de proteínas son el producto final de la constante duplicación y procesos de divergencia derivados de una secuencia ancestral. A pesar de que la evolución en las bacterias también ha seguido otros procesos (pérdida de material genético o transferencia horizontal de genes entre especies), la duplicación y la diversificación de genes parecen ser los mayores factores que generan el incremento del tamaño y complejidad de los grandes genomas. Por otro lado, se ha observado que las superfamilias que se distribuyen en bacterias se pueden dividir en dos clases; aquellas que son universales, que están presentes en un número significante de especies; y por el contrario superfamilias específicas que solo están presentes en pocas especies. Esta distinción está relacionada a las diferencias funcionales y evolutivas de cada clase (Ranea, y otros, 2004).

En este trabajo, se identificó un amplio repertorio de proteínas que pertenecen a la superfamilia wHTH de reguladores transcripcionales. Así mismo se identificaron sus correspondientes dominios adicionales, los que podrían estar participando en diversos procesos biológicos. La superfamilia wHTH, es una de las más abundantes con respecto al repertorio de FTs que se encuentran en los diferentes organismos, aproximadamente la mitad de ellos pertenece a esta superfamilia. Por ejemplo, en el caso de *E. coli* se ha identificado que estás proteínas están asociadas a la regulación de alrededor del 50% de todos sus genes.

Adicionalmente, se identificó que tanto que los DBDs como los PaDos aumentan en relación con el tamaño del genoma. Es posible que tanto la ausencia como la presencia de determinadas familias y de PaDos estén relacionadas con el estilo de vida de los organismos. Así encontramos familias que se considerarían como universales, tales como LysR, MarR, GntR, entre otras, ya que se encuentran en más del 60% de los organismos analizados, familias específicas, tales como Meth en Actinobacteria o Rio2 en Firmicutes (Clostridia) .

La amplia gama de PaDos identificados sugiere que es posible que estos estén participando conjuntamente con el DBD en el reconocimiento de diferentes moléculas y en permitir la diversidad de regulación en una célula. Algunos de los PaDos identificados se encuentran estrechamente relacionados con un determinado DBD, lo que sugiere que estos han evolucionado de manera conjunta.

Se ha observado que unas pocas familias son muy versátiles en su combinación con sus dominios adicionales y algunas de estas familias también son las más abundantes en los genomas. La razón de la abundancia y la versatilidad de estas familias estaría ligada muy probablemente a su función. Por ejemplo, la energía para el movimiento y las reacciones en la célula a menudo son proporcionadas por el *P-loop nucleotide triphosphate hidrolases*, los dominios de esta familia hidrolizan ATP o GTP y pueden actuar como quinasas y transferasas por sí mismos o en combinación con diferentes familias. Los dominios Rossman son similares, estos proporcionan energía de oxidación o reducción a través de la oxidación o reducción del cofactor NAD(P)(H). La transcripción y la traducción están estrechamente reguladas por proteínas que presentan motivos de unión a ácidos nucleicos como los WH o dominios RING *finger domain*, combinados con otros dominios responsables de la especificidad de la regulación. Aunque estas familias de dominios no se encuentran en combinación con otras familias clave (Apic, y otros, 2001).

11. PERSPECTIVAS

En el presente trabajo se analizaron las proteínas asociadas a la superfamilia wHTH que participan directamente en la regulación transcripcional. De manera histórica estas proteínas han sido ampliamente estudiadas a nivel de sus dominios de unión al DNA; sin embargo, la mayoría de los estudios han dejado de lado sus dominios adicionales. Como vimos estos dominios tienen un papel importante en la plasticidad de la regulación proveyendo a los FTs de posibles funciones nuevas o especializadas. Algunos de estos reguladores presentan arquitecturas de dominios que las hacen características, es decir, que han co-evolucionado con sus respectivos PaDos. Sin embargo aún quedan preguntas por resolver, que a manera de perspectivas, se trataran de resolver en proyectos posteriores, tales como:

- Si, la distribución tanto de las familias como los PaDos está relacionada directamente con los estilos de vida de los organismos analizados.

- ¿Cómo es que familias monolíticas están asociadas a múltiples funciones?. Tal es caso de familias como LysR o CRP. Y en este sentido se realizó un análisis preliminar en la familia LysR donde se identificaron varios grupos que presentaban motivos diferentes en la secuencia de sus PaDos. Se sabe que la estructura de las proteínas homologas puede conservarse, a pesar de que a nivel de secuencia existan diferencias entre ellas y particularmente las proteínas parálogas en las que sus duplicados pueden no mantener la función, como posiblemente ocurrió con familias como LysR. Por lo tanto, un estudio exhaustivo podría explicar cómo estas proteínas monolíticas pueden estar involucradas en la regulación de múltiples funciones.

- Analizar de manera exhaustiva el resto de las Superfamilas, analizando su distribución en los distintos genomas, saber si existen superfamilias específicas para grupos de organismos

- Analizar a fondo los PaDos de las distintas superfamilias, identificar en que grupo de organismos, así como, en que FTs se encuentran los PaDos específicos y los que se encuentran altamente compartidos entre superfamilias.

- Por último, clasificar a los PaDos de manera general en términos de sus funciones, es decir, en que procesos se encuentran involucrados, tales como biosíntesis, transporte, reconocimiento proteína-proteína, reconocimiento de metales, entre otros.

12. MATERIAL ADICIONAL

Capitulo de libro:

Perez-Rueda E, Rivera-Gomez N, Martinez-Nuñez MA, Tenorio-Salgado S. 2012. Evolution of DNA-binding Transcription Factors and Regulatory Networks in Prokaryotes in: Filloux,A.A.M. Evolution of regulatory networks in bacteria Horizon Scientific Press. 333-346

Evolution of DNA-binding Transcription Factors and Regulatory Networks in Prokaryotes 13

Ernesto Perez-Rueda, Nancy Rivera-Gomez, Mario Alberto Martinez-Nuñez and Silvia Tenorio-Salgado

Abstract

The capabilities of organisms to contend with environmental changes depend on their repertoire of genes and their ability to regulate their expression. DNA-binding transcription factors have a fundamental role in this process, because they regulate transcription positively or negatively as a consequence of environmental signals. In this chapter we briefly describe some of the most recent findings on regulatory network evolution from the perspective of DNA-binding transcription factors. We explore diverse elements associated with the evolution of regulatory networks, such as gene duplication, where new interactions can emerge together with their upstream and downstream binding sites. The chapter is divided into sections covering the evolution of transcription factors and their domains, their evolution, and a global analysis. Hypotheses concerning a comprehensive picture of how regulatory networks have evolved in prokaryotes and the role of transcription factors in this organization are discussed.

Introduction

Brian Goodwin has suggested that organisms are more than the sum of their parts. In this regard, we cannot infer the phenotype of one organism by knowing the genes associated with it, because a change in a single gene is not enough to cause a change in the complete phenotype (Goodwin, 1994). Therefore, if we could obtain such information, we could understand how the structure is made. In this direction, organismal development is intimately related to genetic regulation since, for example, organisms or metabolic responses require the concerted action of many regulatory proteins. von Hippel (1998) described in an elegant manuscript that 'regulatory mechanisms developed in all organisms appear to be almost infinite in number, but the basic principles on which they operate are relatively few.' It is plausible that the regulatory elements described in all the organisms change depending on their context; however, they will act in a similar fashion to allow or block gene expression.

In all organisms, it is well known that gene regulation occurs predominantly at the level of transcription initiation, and transcription factors (TFs) play an important role, because they determine when a gene is expressed or repressed, according to the environmental conditions (Martinez-Antonio *et al.*, 2006). Given the importance of this kind of proteins, many authors have evaluated their presence and abundance in diverse organisms. From these studies, it has been observed that the number of TFs increases from a few hundred in archaea and bacteria, such as *Pyrococcus horikoshii*, *Bacillus subtilis* and/or *Escherichia coli* K12, to over 3000 in *Homo sapiens* (Levine and Tjian, 2003; Perez-Rueda *et al.*, 2004; Perez-Rueda and Janga, 2010). This increment correlates with the hypothesis of genome maturation, (Lane and Martin, 2010) where it is proposed that it is necessary for a greater number of regulatory elements to regulate a greater number of genes. Consequently, the number of genetic circuits or regulatory networks that arises also increases (Bhardwaj *et al.*, 2010). Therefore, minor changes in single genes may propagate along such networks and may produce, in the end, quite drastic effects on gene expression in response to external stimuli and change related to development. But how do these changes occur, considering that cellular differentiation, for example, sporulation, requires the concerted interplay between sigma factors, TFs, and their binding sites?

In this chapter we summarize some of the most recent insights from studies on regulatory network evolution from the perspective of DNA-binding TFs, considering that the evolution of regulatory networks requires at least two main mechanisms, gene duplication and gene transfer. Thus, new interactions may emerge together with their upstream and downstream binding sites. We break the subject into sections, covering the evolution of TFs and their domains, the promoters, their evolution, and a global analysis. We finish with some conjectures that attempt to provide a comprehensive picture about how regulatory networks have evolved in prokaryotes and the role of TFs in this organization.

Elements involved in the regulatory process

The regulation of transcription initiation in bacteria is primarily mediated by sigma (σ) factors, which provide most of the specificity for the promoter recognition and DNA melting needed for transcription initiation (Gruber and Gross, 2003; Ishihama, 2000; Wosten, 1998). Indeed, σ factors perform these functions only when bound to the RNA polymerase (RNAP). On the other hand, DNA-binding TFs (Browning and Busby, 2004) affect gene expression, in a wider context, by blocking or allowing the access of the RNAP to the promoter, depending on the operator context and ligand-binding status (Wall *et al.*, 2004; Martinez-Antonio *et al.*, 2006; Miroslavova and Busby, 2006; Janga and Collado-Vides, 2007). Usually, most of the gene transcription in exponentially growing bacteria is initiated by the RNAP carrying a housekeeping σ factor, similar to the *E. coli* σ^{70} or *B. subtilis* σ^{A}. Alternative σ factors typically redirect the transcriptional machinery or RNAP towards a subset of genes required under specific conditions, such as the stress response or growth transitions, among others (Wosten, 1998; Ishihama, 2000; Gruber and Gross, 2003). TFs represent a class of proteins devoted to sensing and binding signals to regulate the response to specific compounds (Martinez-Antonio *et al.*, 2006; Goelzer *et al.*, 2008). In Fig. 13.1 we present a simple regulatory network composed of at least three basic components:

1. the DNA-binding TF, which can be self-regulated;
2. the regulatory region on the DNA, where the

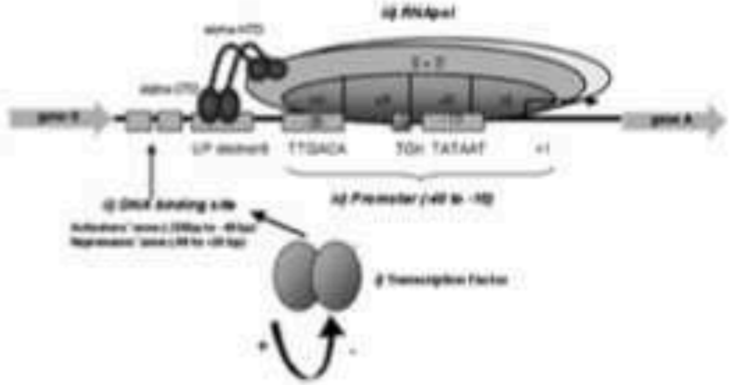


Figure 13.1 Schematic drawing of the basic construction principle of an elementary unit of a hypothetical genetic regulatory network. (i) DNA-binding TF. This protein can activate or repress gene expression. In addition, it can be positive or negatively self-regulated. (ii) A DNA-binding site, which usually is located between the −60 and +20 positions relative to the transcription start. (iii) An RNAP that consists of a protein complex necessary to start the mRNA synthesis. (iv) A sequence promoter the RNAP recognizes specifically over DNA to start the mRNA synthesis.

TF binds and by which the transcription start is modulated;

3 the promoter-binding site, where RNAP binding starts RNA synthesis.

As we will discuss in detail later in this chapter, these links are mostly 'one-to-many,' that is, each TF regulates more than one gene and most genes are controlled by some, albeit generally few, TFs. Every TF is itself regulated by another one. This combinatory perspective gives rise to regulatory networks. Each of these elements can be broken down into regulons. For example, the arabinose regulons in *E. coli* K12, composed of more than 10 different genes involved in the assimilation of arabinose is regulated by two different transcription factors, Crp and AraC (Salgado *et al.*, 2006), where Crp can be associated with a plethora of additional functions. The TF itself typically contains a DNA-binding domain and other regulatory domains, such as multimerization domains that mediate interactions with other proteins or metabolites in order to react to physiological or environmental changes.

Promoter and regulatory region

Transcription starts when the σ factor interacts with the RNAP to recognize its specific sequence promoter (Fig. 13.1). This promoter recognition stage imposes the existence of at least one σ factor per organism, which typically belongs to the σ^{70} family in bacteria (Paget and Helmann, 2003). In this context, bacterial systems could switch between different transcriptional programmes based exclusively on their repertoire of σ factors. Nonetheless, the transcriptional programmes mediated solely via σ factors would be restricted, as a result of their limited repertoire and the small collection of ligands they can recognize, such as guanosine tetraphosphate (ppGpp) (Jores and Wagner, 2003). As a consequence, σ factors exhibit a restricted ability for directly coupling responses to environmental conditions with gene transcription. In addition, σ factors have a constrained DNA-binding region in terms of the lengths and diversity of sequences they recognize, as they need to be structurally coupled to the RNAP on the promoter zone. These DNA restricted zones of action divide the universe of σ factor families into promoters recognized by σ^{70} family and those recognized by σ^{54} family, for instance, the binding zones correspond to about bp −10 to −35 for σ^{70} and bp −12 to −24 for σ^{54}, relative to the transcription start site in the bacterium *E. coli K12* (Gralla, 1996; Lloyd *et al.*, 2001).

On the other hand, TFs define a different regulatory level than do σ factors. These proteins exhibit diverse structural and functional domains, with one of them associated with binding DNA specifically, whereas the second one is devoted to sensing and binding one or more signals from endogenous and/or exogenous sources (Martinez-Antonio *et al.*, 2006). For example, *E. coli* K-12 TyrR binds to three aromatic amino acids and ATP (Pittard *et al.*, 2005); TrmB of the archaeon *Pyrococcus furiosus* binds to three different compounds (Lee *et al.*, 2003, 2005; Perez-Rueda and Janga, 2010). In addition, TFs have the ability to associate combinatorially not only with σ factors but also with a number of other TFs and DNA-binding sites (Adhya, 2003; Barnard *et al.*, 2004), thus allowing the rewiring of a transcriptional network depending on the environmental conditions. For instance, *sodA*, a gene encoding superoxide dismutase in *E. coli*, is regulated by up to eight different TFs responsible for various cellular responses, including Fur (ferric uptake regulation protein), Arc (aerobic respiratory control), and Fnr (fumarate nitrate reduction/regulator of anaerobic respiration) (Compan and Touati, 1993; Salgado *et al.*, 2006). Therefore, the diversity of sequences recognized by TFs is enormous and can occur anywhere from a few bases downstream of the promoter zone to up to hundreds of bases upstream of the transcription start site (Fig. 13.1) (Collado-Vides *et al.*, 1991; Madan Babu and Teichmann, 2003b). For instance, the *E. coli* K-12 global regulator CRP (catabolic repressor protein) can associate with four of the six possible σ factors and coregulate more than 50 different TFs (Salgado *et al.*, 2006). CcpA in *B. subtilis*, a global regulatory protein involved in catabolite repression, may act as a positive regulator of genes involved in excretion of carbon excess and can associate with three different sigma factors (σ^{A}, σ^{L} and σ^{B}) and more than 10 different TFs (Makita *et al.*, 2004;

Moreno-Campuzano *et al.*, 2006). In summary, TFs constitute a class of proteins whose space of action is more flexible than σ factors, not only in sensing diverse environmental and endogenous stimuli but also in recognizing a wide range of binding-site sequences over a larger zone on the DNA around the transcription start site.

Evolution of the repertoire of TFs

DNA-binding domains

The structures of more than 30 prokaryotic DNA-binding proteins have now been determined, and hundreds of amino acid sequences are known for many more. In general, the DNA-binding domains associated with TFs are among the most ancient domains, and they have been proposed as derived from a relatively small set of folds (Aravind and Koonin, 1999; Perez-Rueda and Collado-Vides, 2001; Madan Babu and Teichmann, 2003). These domains have been used to classify the TFs in terms of families (Perez-Rueda *et al.*, 2004). From these studies diverse principles have emerged regarding TFs; for example, the most abundant DNA-binding domain in prokaryotes is the helix–turn–helix (HTH), identified in more than 80% of TFs (Perez-Rueda and Collado-Vides, 2001; Perez-Rueda and Janga, 2011) (Fig. 13.2). Most of the archaeal and bacterial HTHs appear to have undergone a common general evolutionary pathway. The HTH might then have been a motif with general nucleic acid-binding functions that appeared early in evolution (Aravind and Koonin, 1999; Roy *et al.*, 2002) and whose subsequent radiation in the archaeal and bacterial lineages might have involved considerable loss and acquisition events. Additionally, lineage-specific duplications resulted in the accumulation of particular families in microbial species, such as the LysR family, whose members have been abundantly identified in almost all organisms. This hypothesis is consistent with the notion that a genome evolves from a set of precursor genes to a mature size by gene duplications and increasing modifications (Yanai *et al.*, 2000). Alternative DNA-binding structures, such as helix–loop–helix motifs, zinc-fingers, and σ-sheet DNA-binding structures, have been also identified, although in lower proportions, and their distributions are constrained to specific organisms. For instance, the β-sheet proteins have been identified almost exclusively in *Gammaproteobacteria*. The distribution of the RNA-binding domains (associated with cold shock proteins) suggests that they might have been acquired after the prokaryotes and eukaryotes split, probably by lateral gene transfer from eukaryotes, based on the high diversity identified in this cellular domain.

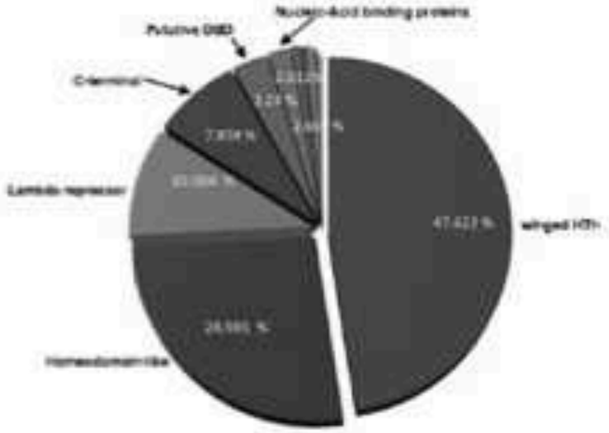


Figure 13.2 Distribution of DNA-binding domains of TFs in bacteria and archaea as defined in the Superfamily database (Madera *et al.*, 2004). The winged HTH represents 47.6% of the total repertoire of DNA-binding domains, being the most abundant structure. The lambda-repressor DNA-binding domain is present at the second highest abundance, as 26% of the repertoire. In minor proportions occur alternative DNA-binding domains, such as the homeodomain-like, with 10.0%, the C-terminal effector domain of the bipartite response regulators, at 7.8%, the putative DNA-binding domain, at 3.24%, and the nucleic acid-binding proteins, at 2.01%. In a low fraction, corresponding to 2.6% of the total DNA-binding domains, are AbrB/MazE/MraZ-like, KorB DNA-binding domain-like, TrpR-like, ACT-like, flagellar transcriptional activator FlhD, haemolysin expression-modulating protein H, a DNA-binding domain in eukaryotic TFs, DNA-binding domain, p53-like TFs, and the DNA-binding domain of the Mlu1 box-binding protein MBP1.

Abundance of TFs correlates with genome size in prokaryotes

It has been documented that organisms respond and adapt to diverse environmental conditions as a consequence of their gene repertoire and regulatory mechanisms, among other elements

(Lynch and Conery, 2003; Bengtsson, 2004; Lynch, 2006). Recent studies have shown that the evolutionary events associated with regulatory proteins, such as their expansion and contraction, contribute significantly in shaping the gene repertoire and genome size of the different lineages of prokaryotes (Perez-Rueda *et al.*, 2004; Minezaki *et al.*, 2005; Oguiza *et al.*, 2005; Rodionov, 2007). Based on comparative genomics, it has been shown that transcription factors increase in quadratic proportion with respect to genome size (van Nimwegen, 2003; Cordero and Hogeweg, 2007; Molina and van Nimwegen, 2008). In particular, this proportion is more significant when the repertoire of TFs is compared with the proportion of σ factors, being roughly ten times higher (hundreds of TFs vs. tens of σ factors) when the general profiles in all the genomes analysed are considered, suggesting a proportion on the order of 1 σ factor:10 TFs:100 annotated open reading frames per genome, although some genomes behave exceptionally. This observation suggests that a possible functional relationship between TFs and prokaryote lifestyles influences the observed trend. A plausible hypothesis is that the abundance of TFs increases with an increase in an organisms' complexity (Brown *et al.*, 2002; Changizi, 2001; Levine and Tjian, 2003; van Nimwegen, 2003; West and Brown, 2005) as a consequence of different evolutionary events, such as gene expansion, gene loss, and lateral gene transfer, among others (Levine and Tjian, 2003; Aravind *et al.*, 2005; Madan Babu *et al.*, 2006). In addition, the necessity to regulate responses to variable environments could also contribute to the abundance of TFs.

Lifestyles explain the abundance of σ factors and TFs in larger genomes

In previous sections we suggested that regulatory complexity should increase in larger genomes and might be associated with bacterial lifestyles, as the environment should influence the bacterial genome structure and function. Thus, to understand how the complexity of gene regulation depends on the number of TFs as a function of increasing genome size and how they are associated with the lifestyles, in previous work (Perez-Rueda *et al.*, 2009) we classified in four global lifestyle classes all the bacterial organisms. These included extremophiles, intracellular bacteria, pathogens, and free-living bacteria. From this analysis, it was identified that the increment of regulatory complexity in TFs contributes significantly to the regulatory complexity of prokaryotes belonging to different lifestyle groups. These results agree with previous observations that suggest that a few regulatory elements identified in small genomes would compensate for the regulation of the entire genome with an increase in the number of DNA-binding sites per element, in contrast to the large number of elements identified in large genomes, which control a smaller proportion of DNA-binding sites on average (Molina and van Nimwegen, 2008). In addition, a larger proportion of genes in small genomes are organized in operons, simplifying the transcriptional machinery necessary for gene expression, in contrast to large genomes, which have reduced number of genes in operons, which would influence the proportion of TFs in those organisms (Cherry, 2003), suggesting that complex lifestyles require a higher proportion of TFs and transcription units to better orchestrate a response to changing conditions.

Abundance of TFs does not correlate with diversity of families, and large families are not the most widely distributed

An appealing hypothesis is that the high diversity of TF families contributes significantly to the regulatory plasticity. In line with this hypothesis, the repertoire of TFs identified in bacteria has been classified into families to evaluate their distribution and abundance in all the prokaryotes. This analysis showed a reduced diversity of families in small genomes, with an increasing proportion in larger ones, especially in pathogens and free-living organisms. The diversity of families reaches a maximum in genomes with around 5000 open reading frames. The higher number of TFs in larger genomes does not necessarily imply diversity of families beyond this plateau, but instead an increase in the size of some families of TFs. Congruent with this observation, the average number of TFs per family increases linearly, with a few families of TFs expanding disproportionately (Janga and Perez-Rueda, 2009; Perez-Rueda *et* *al.*, 2009). These families comprise LysR and TetR, which represent about 25% of the total set of TFs identified. Members of these two families increase abruptly in larger genomes and coincide with the plateauing of the diversity of families in these bacterial genomes. Another feature associated with large families is that they are not widely distributed among bacteria despite their role in controlling important processes, such as cell–cell communication (LuxR), response to external conditions by two-component systems (OmpR), sensing, uptake, and metabolism of external food sources (GntR and LysR), and antibiotic resistance (TetR). Alternatively, families with few copies per genome, such as DnaA, LexA, and IHF, which have been proposed to be essential under standard growth conditions in *E. coli* and in maintaining DNA and nucleoid integrity, (Gerdes *et al.*, 2003; Yamazaki *et al.*, 2008) might be considered universal in bacteria, because they have been identified in at least 80% of the genomes, suggesting gene loss events in bacteria in which they are absent.

In summary, a TF family's abundance and distribution should be associated with the following evolutionary events in bacteria: (i) small families widely distributed among bacteria might be related to ancestral functions beyond transcriptional regulation, such as DNA organization, nucleoid integrity, or DNA salvage; (ii) large families might be associated with the regulation of dispensable or emergent processes in bacterial evolution, such as those involved in quorum sensing, belonging to the members of the LuxR family, which are widely identified in bacteria. Indeed, the evolution of this mechanism in bacteria has been proposed to be one of the early steps in the development of multicellularity (Miller and Bassler, 2001) and may be correlated with bacterial specialization.

Evolution of partner domains

DNA-binding TFs usually make contact with their DNA targets as homodimers or homotetramers, as is the case for the lactose repressor (Lewis, 2005). In this regard, there are diverse questions concerning whether multimerization precedes DNA binding. Therefore, the TFs can act as activators or repressors as a consequence of their multimerization state. Amoutzias *et al.* (2004) concluded that the ancestral TFs were probably the homodimerizing ones and that these proliferated through a series of single-gene duplications. Recent evidence supports this hypothesis, as investigators have described a high abundance of small-sized TFs, or proteins that contain a dimerization domain, but no DNA-binding domain in archaeal genomes (Perez-Rueda and Janga, 2011). The main consequence of gene duplications, mainly involving TFs, is to give rise to a complex interaction network. To the best of our knowledge, there are no studies so far that have linked genetic networks and their regulation with other networks, such as metabolism, although many effector domains are protein interaction domains. These data could help in understanding how the ligand-binding domains have been recruited to regulate gene expression. However, previous studies (Madan Babu and Teichmann, 2003; Aravind *et al.*, 2005) suggested that in the winged HTH superfamily of TFs, the partner domains contribute to the structural differentiation of duplicated genes. Therefore, the partner domains are associated with diverse functions, such as regulating allosterically the function of TFs across binding to a wide variety of functional compounds, in protein–protein interactions, or with enzymatic properties (Madan Babu and Teichmann, 2003), and they are fundamental to linking environmental conditions and the functional conformational changes in the regulators (Taraban *et al.*, 2008). Because there are few exhaustive analyses describing the partner domain repertoire in bacteria, their functional and evolutionary diversity must be evaluated (Madan Babu and Teichmann, 2003; Rivera-Gomez *et al.*, 2011). From this perspective, one might expect that the high diversity of TF families and their associated partner domains contributes significantly to the regulatory plasticity, as we previously mentioned; however, further studies are necessary. Thus, the diversity associated with dimerization domains has allowed the TFs to interact with many different partners and achieve specificity for the target gene and physiological conditions. On the other hand, multimerization domains may have been facilitated by the acquisition of a second protein interaction domain during evolution (Kaufmann *et al.*, 2005).

Evolution of promoters

Unlike coding regions, the evolution of upstream regulatory regions cannot be easily evaluated by means of sequence alignments and comparisons. Indeed, the structure, organization, and function of promoters differ widely from that of the coding sequences, resulting in totally different evolution rates and consequences of sequence variation (Rodriguez-Trelles *et al.*, 2003). In general, promoter organization underlies more variation than the coding sequences, and it is mainly influenced by DNA structure (Olivares-Zavaleta *et al.*, 2006) such as supercoiling (Martinez-Nunez *et al.*, 2010). Bacterial promoter sequences, or upstream regulatory regions, contain TF-binding sites. Several TFs often interact, depending on signals and physiological or environmental requirements. This gives rise to activating or repressing complexes (Koike *et al.*, 2004). Ishihama described that among the set of promoters under the control of the same σ factor, the level of transcription varies depending on the culture conditions or the growth phase. For that reason, it is important to evaluate the diversity of promoters and their regulation associated with all genes in an organism, for instance, in *E. coli* K-12, approximately 50% of the promoters are under the control of one specific regulator, whereas the other 50% of genes are regulated by more than two TFs. Likewise, the promoters for the genes involved in the construction of cell structures are controlled by environmental conditions and specific signals, and each is in turn monitored by a different TF. The binding sites of all these multiple factors are located in a single promoter. The most typical examples of the multifactor promoter system are the promoters for the genes encoding the master regulator for flagellum formation in *E. coli* K12, FlhCD, and the master regulator for biofilm formation, CsgD. The complexities of these promoters reflect the opposite behaviours of bacterial survival, i.e. planktonic growth as single cells (FlhCD) and biofilm formation as a bacterial community under stressful conditions in nature (Soutourina *et al.*, 1999; Ogasawara *et al.*, 2010).

Coevolution of regulatory elements

Since some proteins tend to work together in a functional context, analyses of distributions of different families in the function of the σ factors have been recently performed. Hence, the cooccurrence of the regulatory protein families (TFs and σ factors) in all the prokaryotes was evaluated. From this analysis, it was found that the distribution of the σ^{54} factor and IHF and EBP families are correlated (Fig. 13.3), supporting the functional interdependence discussed above and probable coevolution in which members and mechanisms have been preserved along the course of evolution in bacteria. A second cluster that includes σ^{70}, the ECF family of σ factors, and other highly abundant families (more than 15 members per genome) responsible for regulating diverse mechanisms of stress responses (MarR), antibiotic resistance (TetR), osmotic responses (OmpR), and the quorum-sensing response (LuxR), among other

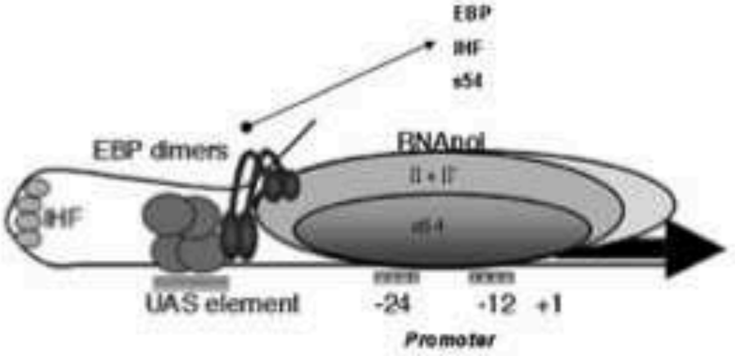


Figure 13.3 Coevolution of σ factors and TF families. A similar occurrence distribution pattern was observed for IHF, EBP, and σ^{54} families, suggesting a functional interdependence between these families. A coregulatory mode of action for these regulatory proteins is also shown. Figure modified from Perez-Rueda *et al.* (2009).

processes, was also found to be clustered, suggesting a strong functional relationship among these σ and TF families. These clusters, in addition, give insights into the functional interdependence between regulatory proteins from different families, which could help in the characterization of regulators in poorly studied organisms.

Evolution of regulatory networks

The regulation of TFs plays a key role in morphological diversity. Simple modifications within the upstream regulation region of a TF can explain both minor and major changes between species, without involving any disruption of gene structure. Therefore, evolution of regulatory regions is thought to be a major source of diversity. (Lozada-Chavez *et al.*, 2008; Perez and Groisman, 2009a,b) Duplication events of TFs are another evolutionary source that can allow diversity, permitting a more versatile adaptation of the functional divergence gained from the duplication of structural genes. Different aspects of the evolution of the regulatory networks have been examined, including the coevolution of the upstream regulatory regions and their corresponding TFs, the likely consequences of gain, loss, and replacement of TFs in the regulatory networks of duplicated genes (Teichmann and Babu, 2004; Gelfand, 2006), and also the topological and dynamic properties of the regulatory networks (Luscombe *et al.*, 2004; Madan Babu *et al.*, 2006; Balaji *et al.*, 2007).

Role of duplication events in regulatory networks

Duplication events of regulatory genes provide new interactions for the transcriptional regulatory network. These new interactions are the raw material for the generation of divergence in gene expression, which should happen in most of the copies that have remained in the genome (Teichmann and Babu, 2004). 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. 13.4). The evolutionary plasticity of the regulatory networks is not only the result of the duplication of TF interactions within a regulatory network, as previously proposed by Teichmann and Babu (2004), but also the result of the divergent effects of the TF interactions in activating or repressing the transcription of duplicated genes, as suggested by Martinez-Nuñez *et al.* (2010). Indeed, examples have been recently identified of regulatory systems where the TF is maintained but a different regulatory role is gained (either activation or repression) in one of the duplicated genes. This evolutionary scenario can be observed in the regulation of the *E. coli* *gntK* and *idnK* gluconate kinase genes, which are involved in 6-phosphogluconate synthesis in the Entner–Doudoroff and pentose phosphate pathways, respectively. Although the same TFs, CRP, GntR, and IdnR, regulate all these genes, IdnR represses the transcription of *gntK*, whereas it activates the transcription of *idnK* (Bausch *et al.*, 2004; Salgado *et al.*, 2006).

This form of diversification in regulation allows plasticity of the transcriptional regulatory network without the need to increase the number of interactions within it, if not only by varying the type of regulation (positive or negative) exerted by the TFs on their targets. It is possible 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 metabolic pathway (Martinez-Nunez *et al.*, 2010).

Concluding remarks

As a consequence of the abundance of data which have become recently available, the idea has emerged to conceptualize genetic networks as directed graphs with nodes corresponding to the TFs, linked by edges to their target genes. The previously mentioned regulatory elements (TFs, σ factors, upstream binding sites, downstream binding sites, and promoters) can be combined at a higher level, into so–called network motifs (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002). The impacts of evolutionary forces on the topological structures of regulatory networks known as motifs have been exhaustively analysed by diverse authors, for example, biological regulatory networks (Milo *et al.*, 2002; Shen-Orr *et al.*, 2002) and duplicated gene networks (Teichmann and Babu, 2004). In

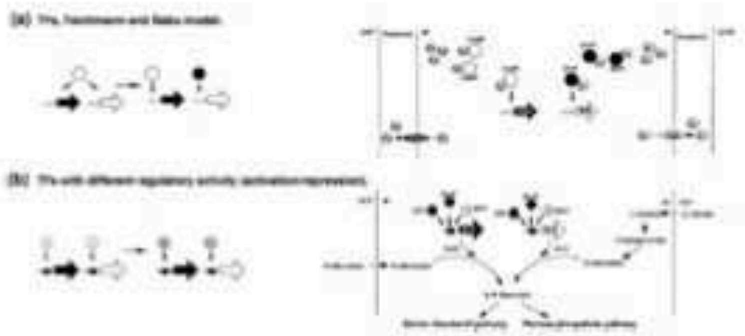


Figure 13.4 Regulatory elements involved in the regulation of the duplicated genes. (a) Simplification of the model of Teichmann and Babu (Teichmann and Babu, 2004), where a new TF is gained to regulate one of the duplicated genes. TFs are shown as circles. (b) Extension of the Teichmann and Babu model. The differential regulation of the duplicated genes depends on the activation or repression mechanism associated with the TF on its target genes. Figure modified from Martínez-Núñez *et al.* (2010).

these studies, the authors reported that duplication of an entirely feed-forward motif (a topological structure in which 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 duplicated 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-σ factor SpoIIAA and the anti-σ factor SpoIIAB, as σ^F regulates SpoIIAA and SpoIIAB expression, 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, the 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 duplicated genes, which would favour the adaptation process of the organism as it responds to changes in its niche (Gelfand, 2006). Probably the most basic motif is the autoregulatory loop: a TF that regulates its own expression (Fig. 13.1).

Generally, these motifs have been considered basic architectures in the regulatory networks, because they often overlap (Browning and Busby, 2004). In this context, diverse authors have analysed the structure of both genetic and protein–protein interaction networks (Yeger-Lotem and Margalit, 2003; Yeger-Lotem *et al.*, 2004). These analyses have shown that certain topologies of small subnets are statistically very much over-represented (Shen-Orr *et al.*, 2002). Conant and Wagner introduced the notion of common ancestry for gene circuits or motifs, where two motifs share a common ancestor if every pair of genes in the two circuits is derived from a common ancestor; all pairs in the circuits must be duplicated genes (Conant and Wagner, 2003). They found that no pairs of motifs with identical topology had common ancestry, and they concluded that their emergence is the result of convergent evolution and not duplication of one or a few ancestral circuits, suggesting that convergent evolution was more likely to be important in module topology than for protein sequences.

A third level of network organization consists of transcriptional modules (Babu *et al.*, 2004). Modules represent collections of TFs that are expressed under distinct experimental or environmental conditions (Ihmels *et al.*, 2002) and are largely controlled by one (or very few) regulators, as was shown by hierarchical clustering of expression profiles (Segal *et al.*, 2003). For instance, the *E. coli* K-12 global regulator CRP can regulate the expression of more than 20 different TFs (Thieffry *et al.*, 1998; Gama-Castro *et al.*, 2008). Exhaustive analyses of global features of genetic networks and protein interaction networks have revealed a scale-free topology (Barabasi and Albert, 1999; Barabasi and Oltvai, 2004); in other words, there are few genes, or so-called hubs, that control many others, and many genes have only a few links. These hubs can be defined as global regulators, such as Crp *in E. coli* K-12. However, regulatory networks are dynamic: it was shown that large-scale topological changes have occurred in the *E. coli*, *B. subtilis*, and *Saccharomyces cerevisiae* genomes and that although a few TFs serve as permanent hubs, they act transiently only under certain conditions (Luscombe *et al.*, 2004). It is also worth noting that more complex organisms have a higher number of regulatory genes per target gene (van Nimwegen, 2003) suggesting that it is mostly the evolution by duplication and diversification of transcription factors and of their interactions that increases organismic complexity as a whole. Over the last few years, the availability of large numbers of experimental and theoretical data has significantly enhanced our understanding of evolution of complex networks and, at the same time, enabled us to transfer knowledge from better-studied model organisms (such as *E. coli* and *B. subtilis*) to those for which fewer data are available. Basically, the ancestral genetic networks we observe today were probably a small group of DNA-binding domains that, while conserving their structure, diverged into a large variety of TFs. More recently, most proteins, among them TFs, underwent many cycles of domain rearrangements (Amoutzias *et al.*, 2005). Additional dimerization and sensor domains were gained and lost at different times. Further, they evolved across a series of single-gene duplications, thus generating networks of regulatory genes that arrange into these modules. These events may be quite recent and lineage specific, as we have learned from the uneven distribution of some TF families (Perez-Rueda *et al.*, 2004). A growing number of findings suggest that structurally similar or even identical motifs can arise repeatedly and thus represent a simple level of convergent evolution. More complex modules, which may also have preferentially arisen through a series of single-gene duplications, would give rise to similar topologies. The evolution of promoter regions is less well understood, although it is of great importance. Genotypic changes at this level are probably among the main reasons why, despite minor interorganismic differences at the level of proteins, major changes in the topologies of genetic networks during development induce wide morphological differences and diversity in contemporary organisms.

In summary the mechanisms of generating diverse networks can be associated with diverse evolutionary forces, such as gene duplication, gene loss, changes in the regulatory mechanisms (regulatory role modulation), acquisition of new activities, modular rearrangements, and finally, functional divergence. Therefore, we believe that with the availability of more information, we will be able to understand in a more comprehensive fashion the evolutionary dynamics associated with regulatory networks.

Acknowledgements

We gratefully acknowledge to Alejandro Huerta-Saquero, Antonio Ibarra, and the anonymous referee for their helpful comments in the preparation of the manuscript. This work was supported by DGAPA grant IN209511 to EP-R. NR-G, MAM-N, and ST-S were financed by a PhD fellowship from CONACyT.

References

Amoutzias, G.D., Robertson, D.L., and Bornberg-Bauer, E. (2004). The evolution of protein interaction networks in regulatory proteins. Comp. Funct. Genomics 5, 79–84.

Amoutzias, G.D., Weiner, J., and Bornberg-Bauer, E. (2005). Phylogenetic profiling of protein interaction networks in eukaryotic transcription factors reveals focal proteins being ancestral to hubs. Gene 347, 247–253.

Aravind, L., and Koonin, E.V. (1999). DNA-binding proteins and evolution of transcription regulation in the archaea. Nucleic Acids Res. 27, 4658–4670.

Aravind, L., Anantharaman, V., Balaji, S., Babu, M.M., and Iyer, L.M. (2005). The many faces of the helix–turn–helix domain: transcription regulation and beyond. FEMS Microbiol. Rev. *29*, 231–262.

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

Balaji, S., Babu, M.M., and 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.

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

Barabasi, A.L., and Oltvai, Z.N. (2004). Network biology: understanding the cell's functional organization. Nat. Rev. Genet. 5, 101–113.

Bausch, C., Ramsey, M., and Conway, T. (2004). Transcriptional organization and regulation of the L-idonic acid pathway (GntII system) in *Escherichia coli*. J. Bacteriol. *186*, 1388–1397.

Bengtsson, B.O. (2004). Modelling the evolution of genomes with integrated external and internal functions. J. Theor. Biol. *231*, 271–278.

Bhardwaj, N., Carson, M.B., Abyzov, A., Yan, K.K., Lu, H., and Gerstein, M.B. (2010). Analysis of combinatorial regulation: scaling of partnerships between regulators with the number of governed targets. PLoS Comput Biol 6, e1000755.

Brown, J.H., Gupta, V.K., Li, B.L., Milne, B.T., Restrepo, C., and West, G.B. (2002). The fractal nature of nature: power laws, ecological complexity and biodiversity. Philos. Trans. R. Soc. Lond. B Biol. Sci. *357*, 619–626.

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

Changizi, M.A. (2001). Universal scaling laws for hierarchical complexity in languages, organisms, behaviors and other combinatorial systems. J. Theor. Biol. *211*, 277–295.

Cherry, J.L. (2003). Genome size and operon content. J. Theor. Biol. *221*, 401–410.

Conant, G.C., and Wagner, A. (2003). Asymmetric sequence divergence of duplicate genes. Genome Res. *13*, 2052–2058.

Cordero, O.X., and Hogeweg, P. (2007). Large changes in regulome size herald the main prokaryotic lineages. Trends Genet. *23*, 488–493.

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., Salgado, H., *et al.* (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*, D120–124.

Gelfand, M.S. (2006). Evolution of transcriptional regulatory networks in microbial genomes. Curr. Opin. Struct. Biol. *16*, 420–429.

Gerdes, S.Y., Scholle, M.D., Campbell, J.W., Balazsi, G., Ravasz, E., Daugherty, M.D., Somera, A.L., Kyrpides, N.C., Anderson, I., Gelfand, M.S., *et al.* (2003). Experimental determination and system level analysis of essential genes in *Escherichia coli* MG1655. J. Bacteriol. *185*, 5673–5684. Goelzer, A., Bekkal Brikci, F., Martin-Verstraete, I., Noirot, P., Bessieres, P., Aymerich, S., and Fromion, V. (2008). Reconstruction and analysis of the genetic and metabolic regulatory networks of the central metabolism of *Bacillus subtilis*. BMC Syst. Biol. *2*, 20.

Goodwin, B. (1994). How the Leopard Changed its Spots: The Evolution of Complexity (Scribner). <Please provide place of publication>

Gralla, J.D. (1996). Activation and repression of *E. coli* promoters. Curr. Opin. Genet. Dev. *6*, 526–530.

Gruber, T.M., and Gross, C.A. (2003). Multiple sigma subunits and the partitioning of bacterial transcription space. Annu. Rev. Microbiol. *57*, 441–466.

von Hippel, P.H. (1998). An integrated model of the transcription complex in elongation, termination, and editing. Science *281*, 660–665.

Ihmels, J., Friedlander, G., Bergmann, S., Sarig, O., Ziv, Y., and Barkai, N. (2002). Revealing modular organization in the yeast transcriptional network. Nat. Genet. *31*, 370–377.

Ishihama, A. (2000). Functional modulation of *Escherichia coli* RNA polymerase. Annu. Rev. Microbiol. *54*, 499–518.

Janga, S.C., and Collado-Vides, J. (2007). Structure and evolution of gene regulatory networks in microbial genomes. Res. Microbiol. *158*, 787–794.

Janga, S.C., and Perez-Rueda, E. (2009). Plasticity of transcriptional machinery in bacteria is increased by the repertoire of regulatory families. Comput. Biol. Chem. 33, 261–268.

Jores, L., and Wagner, R. (2003). Essential steps in the ppGpp-dependent regulation of bacterial ribosomal RNA promoters can be explained by substrate competition. J. Biol. Chem. *278*, 16834–16843.

Kaufmann, K., Melzer, R., and Theissen, G. (2005). MIKC-type MADS-domain proteins: structural modularity, protein interactions and network evolution in land plants. Gene *347*, 183–198.

Koike, H., Ishijima, S.A., Clowney, L., and Suzuki, M. (2004). The archaeal feast/famine regulatory protein: potential roles of its assembly forms for regulating transcription. Proc. Natl. Acad. Sci. U.S.A. *101*, 2840–2845.

Lane, N., and Martin, W. (2010). The energetics of genome complexity. Nature *467*, 929–934.

Levine, M., and Tjian, R. (2003). Transcription regulation and animal diversity. Nature 424, 147–151.

Lewis, M. (2005). The lac repressor. C. R. Biol. *328*, 521–548.

Lloyd, G., Landini, P., and Busby, S. (2001). Activation and repression of transcription initiation in bacteria. Essays Biochem. 37, 17–31.

Lozada-Chavez, I., Angarica, V.E., Collado-Vides, J., and Contreras-Moreira, B. (2008). The role of DNA-binding specificity in the evolution of bacterial regulatory networks. J. Mol. Biol. *379*, 627–643.

Luscombe, N.M., Babu, M.M., Yu, H., Snyder, M., Teichmann, S.A., and Gerstein, M. (2004). Genomic analysis of regulatory network dynamics reveals large topological changes. Nature *431*, 308–312.

Lynch, M. (2006). Streamlining and simplification of microbial genome architecture. Annu. Rev. Microbiol. *60*, 327–349.

Lynch, M., and Conery, J.S. (2003). The origins of genome complexity. Science *302*, 1401–1404.

Madan Babu, M., and Teichmann, S.A. (2003). Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res. *31*, 1234–1244.

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

Madera, M., Vogel, C., Kummerfeld, S.K., Chothia, C., and Gough, J. (2004). The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res. 32, D235–239.

Martinez-Antonio, A., Janga, S.C., Salgado, H., and Collado-Vides, J. (2006). Internal-sensing machinery directs the activity of the regulatory network in *Escherichia coli*. Trends Microbiol. *14*, 22–27.

Martinez-Nunez, M.A., Perez-Rueda, E., Gutierrez-Rios, R.M., and Merino, E. (2010). New insights into the regulatory networks of paralogous genes in bacteria. Microbiology 156, 14–22.

Miller, M.B., and Bassler, B.L. (2001). Quorum sensing in bacteria. Annu. Rev. Microbiol. *55*, 165–199.

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

Minezaki, Y., Homma, K., and Nishikawa, K. (2005). Genome-wide survey of transcription factors in prokaryotes reveals many bacteria-specific families not found in archaea. DNA Res. *12*, 269–280.

Miroslavova, N.S., and Busby, S.J. (2006). Investigations of the modular structure of bacterial promoters. Biochem. Soc. Symp. 1–10.

Molina, N., and van Nimwegen, E. (2008). Universal patterns of purifying selection at noncoding positions in bacteria. Genome Res *18*, 148–160.

van Nimwegen, E. (2003). Scaling laws in the functional content of genomes. Trends Genet. *19*, 479–484.

Ogasawara, H., Yamada, K., Kori, A., Yamamoto, K., and Ishihama, A. (2010). Regulation of the *Escherichia coli* csgD promoter: interplay between five transcription factors. Microbiology *156*, 2470–2483.

Oguiza, J.A., Kiil, K., and Ussery, D.W. (2005). Extracytoplasmic function sigma factors in *Pseudomonas syringae*. Trends Microbiol. *13*, 565–568.

Olivares-Zavaleta, N., Jauregui, R., and Merino, E. (2006). Genome analysis of *Escherichia coli* promoter sequences evidences that DNA static curvature plays a more important role in gene transcription than has previously been anticipated. Genomics *87*, 329–337.

Paget, M.S., and Helmann, J.D. (2003). The sigma70 family of sigma factors. Genome Biol *4*, 203.

Perez, J.C., and Groisman, E.A. (2009a). Evolution of transcriptional regulatory circuits in bacteria. Cell *138*, 233–244.

Perez, J.C., and Groisman, E.A. (2009b). Transcription factor function and promoter architecture govern the evolution of bacterial regulons. Proc. Natl. Acad. Sci. U.S.A. *106*, 4319–4324.

Perez-Rueda, E., and Collado-Vides, J. (2001). Common history at the origin of the position-function correlation in transcriptional regulators in archaea and bacteria. J. Mol. Evol. *53*, 172–179.

Perez-Rueda, E., and Janga, S.C. (2010). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. Mol. Biol. Evol. 27, 1449–1459.

Perez-Rueda, E., and Janga, S.C. (2011). Identification and genomic analysis of transcription factors in archaeal genomes exemplifies their functional architecture and evolutionary origin. Mol. Biol. Evol. 27, 1449–1459.

Perez-Rueda, E., Collado-Vides, J., and Segovia, L. (2004). Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea. Comput. Biol. Chem. *28*, 341–350.

Perez-Rueda, E., Janga, S.C., and Martinez-Antonio, A. (2009). Scaling relationship in the gene content of transcriptional machinery in bacteria. Mol. Biosyst. 5, 1494–1501.

Pragai, Z., Allenby, N.E., O'Connor, N., Dubrac, S., Rapoport, G., Msadek, T., and Harwood, C.R. (2004). Transcriptional regulation of the phoPR operon in *Bacillus subtilis*. J. Bacteriol. *186*, 1182–1190.

Rivera-Gomez, N., Segovia, L., and Perez-Rueda, E. (2011). The diversity and distribution of TFs and their partner domains play an important role in the regulatory plasticity in bacteria. Microbiology. <Please provide volume number and page range>

Rodionov, D.A. (2007). Comparative genomic reconstruction of transcriptional regulatory networks in bacteria. Chem. Rev. *107*, 3467–3497.

Rodriguez-Trelles, F., Tarrio, R., and Ayala, F.J. (2003). Evolution of cis-regulatory regions versus codifying regions. Int. J. Dev. Biol. 47, 665–673.

Roy, S.W., Fedorov, A., and Gilbert, W. (2002). The signal of ancient introns is obscured by intron density and homolog number. Proc. Natl. Acad. Sci. U.S.A. *99*, 15513–15517.

Salgado, H., Gama-Castro, S., Peralta-Gil, M., Diaz-Peredo, E., Sanchez-Solano, F., Santos-Zavaleta, A., Martinez-Flores, I., Jimenez-Jacinto, V., Bonavides-Martinez, C., Segura-Salazar, J., *et al.* (2006). RegulonDB (version 5.0): *Escherichia coli* K-12 transcriptional regulatory network, operon organization, and growth conditions. Nucleic Acids Res. *34*, D394–397.

Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., and Friedman, N. (2003). Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nat. Genet. *34*, 166–176.

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

Soutourina, O., Kolb, A., Krin, E., Laurent-Winter, C., Rimsky, S., Danchin, A., and Bertin, P. (1999). Multiple control of flagellum biosynthesis in *Escherichia coli*: role of H-NS protein and the cyclic AMP-catabolite activator protein complex in transcription of the flhDC master operon. J. Bacteriol. *181*, 7500–7508.

Taraban, M., Zhan, H., Whitten, A.E., Langley, D.B., Matthews, K.S., Swint-Kruse, L., and Trewhella, J. (2008). Ligand-induced conformational changes and conformational dynamics in the solution structure of the lactose repressor protein. J. Mol. Biol. *376*, 466–481.

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

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

Wall, M.E., Hlavacek, W.S., and Savageau, M.A. (2004). Design of gene circuits: lessons from bacteria. Nat. Rev. Genet. 5, 34–42.

West, G.B., and Brown, J.H. (2005). The origin of allometric scaling laws in biology from genomes to ecosystems: towards a quantitative unifying theory of biological structure and organization. J. Exp. Biol. *208*, 1575–1592.

Wosten, M.M. (1998). Eubacterial sigma-factors. FEMS Microbiol. Rev. 22, 127–150.

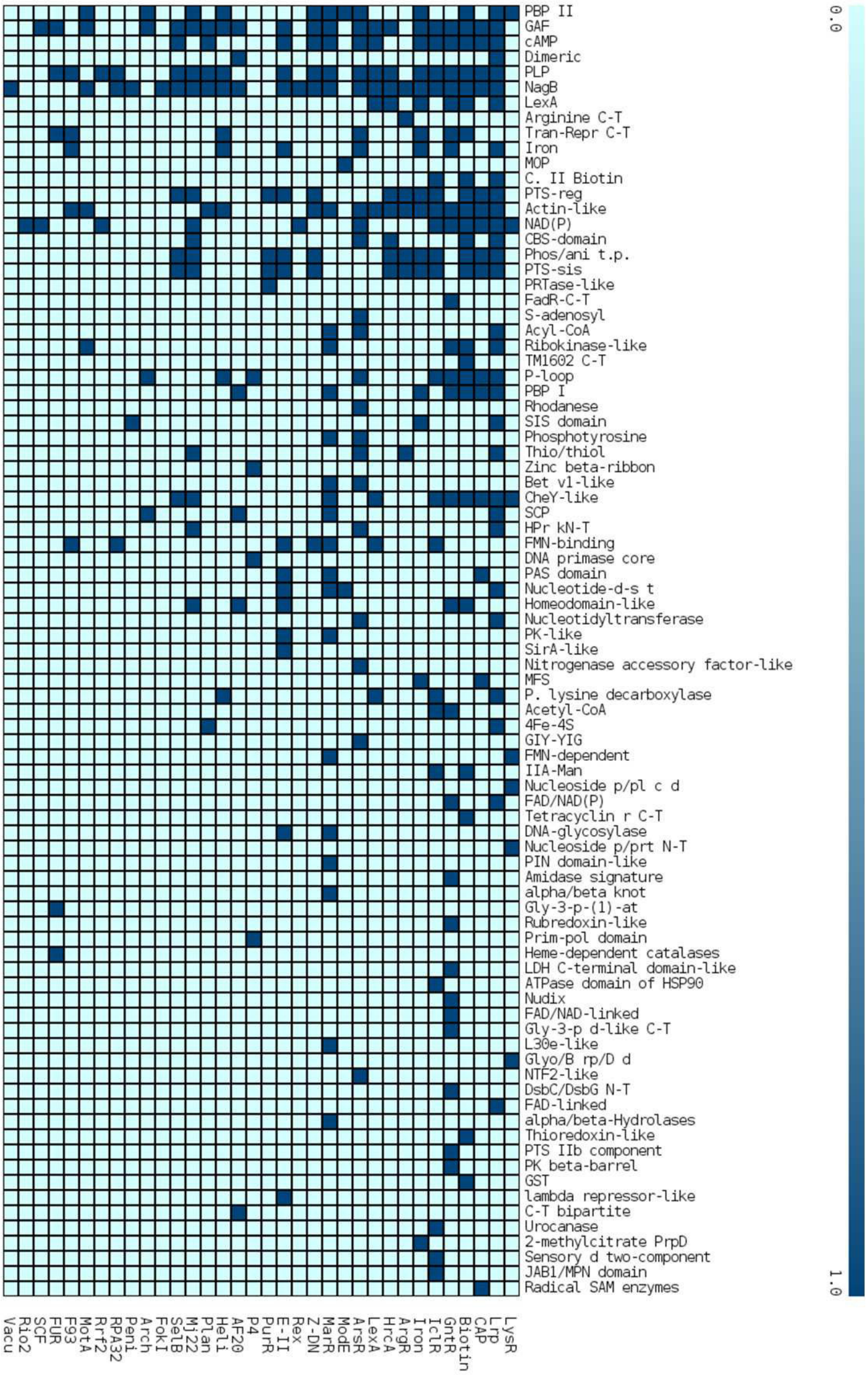
Yamazaki, Y., Niki, H., and Kato, J. (2008). Profiling of *Escherichia coli* Chromosome database. Methods Mol. Biol. *416*, 385–389.

Yanai, I., Camacho, C.J., and DeLisi, C. (2000). Predictions of gene family distributions in microbial genomes: evolution by gene duplication and modification. Phys. Rev. Lett. *85*, 2641–2644.

Yeger-Lotem, E., and Margalit, H. (2003). Detection of regulatory circuits by integrating the cellular networks of protein–protein interactions and transcription regulation. Nucleic Acids Res. *31*, 6053–6061.

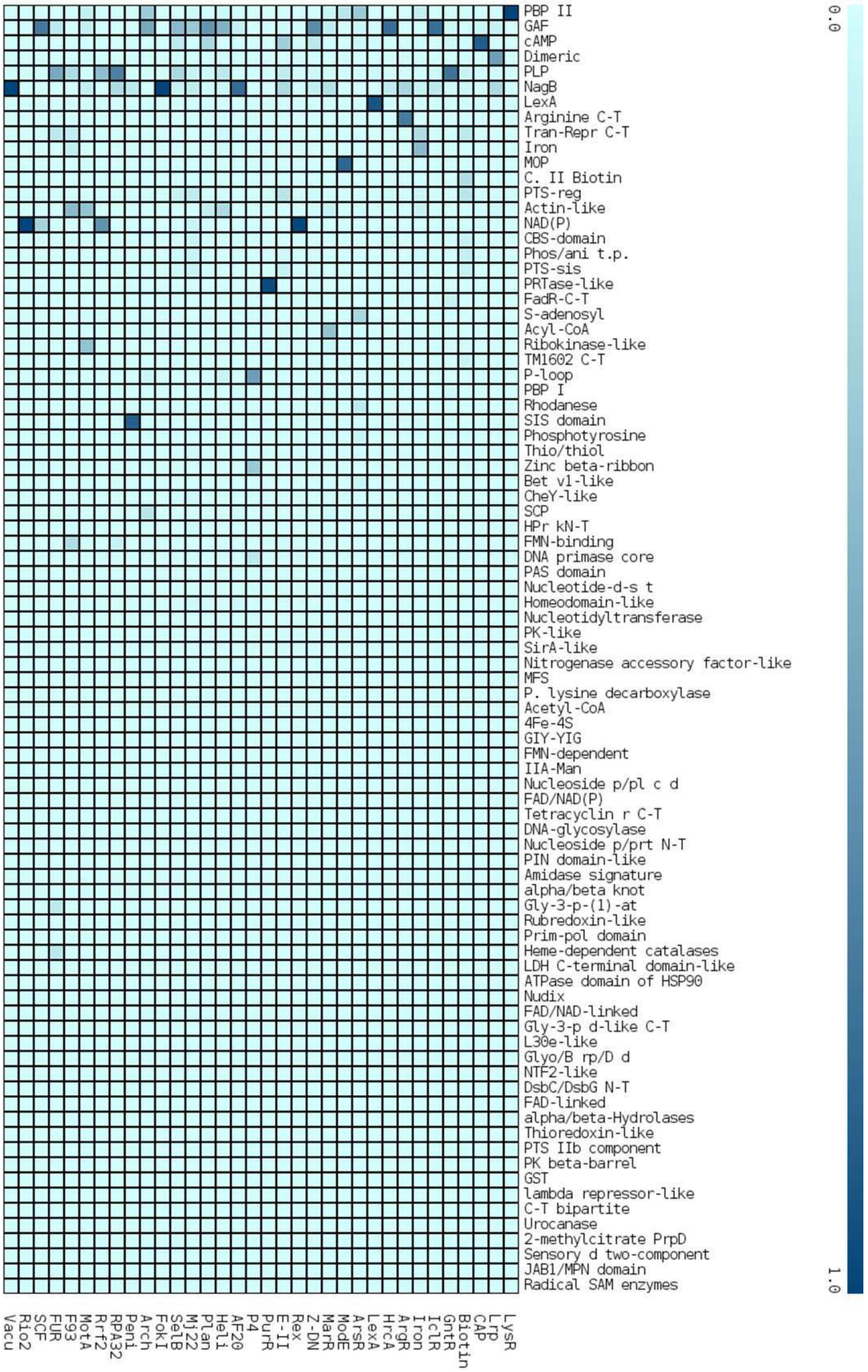
Yeger-Lotem, E., Sattath, S., Kashtan, N., Itzkovitz, S., Milo, R., Pinter, R.Y., Alon, U., and Margalit, H. (2004). Network motifs in integrated cellular networks of transcription-regulation and protein–protein interaction. Proc. Natl. Acad. Sci. U.S.A. *101*, 5934–5939.

13. MATERIAL SUPLEMENTARIO



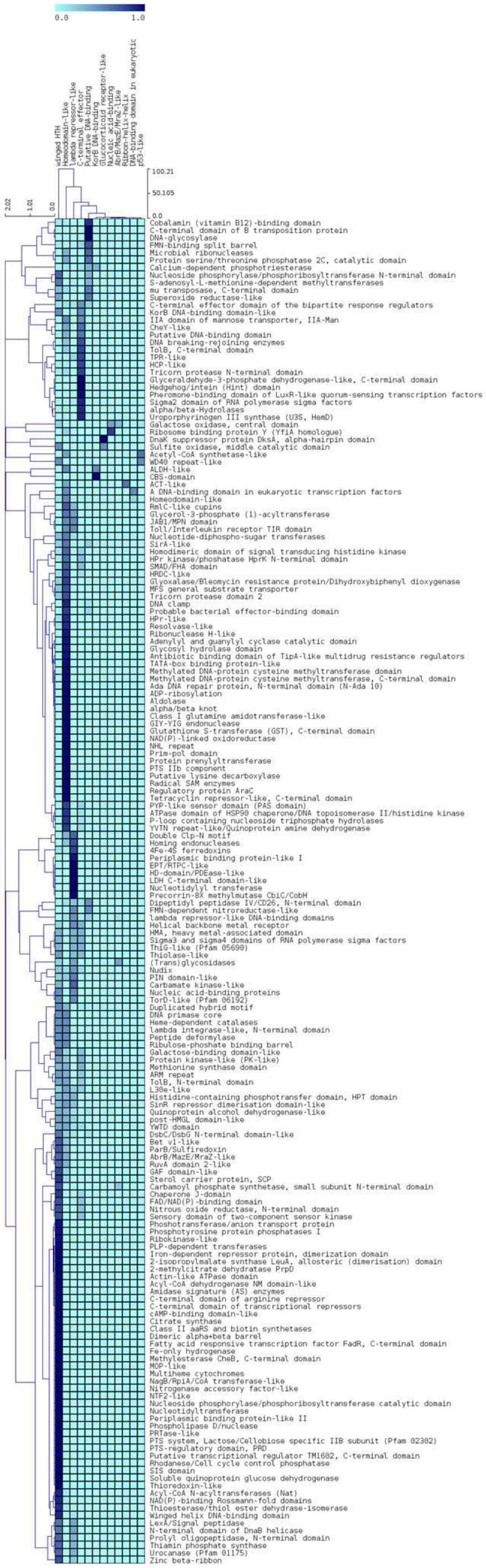
Supplementary Fig. S1. Presence of partner domains (PaDos) in families with winged helix–turn–helix (wHTH). The figure shows the distribution of PaDos in all families, for instance the LysR family exhibits diverse PaDos, although they could have different levels of abundance. 0, absence; 1, presence.

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.



Supplementary Fig. S2. Distribution and abundance of PaDos in families with wHTH. The figure shows a bias in the association of families and their corresponding PaDos, for instance, the LysR family has a strong association with the PBP II domain. 0, absence; 1, presence.

Rivera-Gomez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.



Supplementary Fig. S3. Distribution of PaDos in DNA-binding transcription factors. Winged DNA-binding domain, lambda repressors, C-terminal domain, putative DNA-binding, nucleic acid binding, ribbon-helix–helix, AbrB/MazE, KorB, and p53-like superfamilies were considered. Therefore, from this analysis we found a specific association between the PaDos and their corresponding DNA-binding structure.

Rivera-Gomez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

Supplementary Table S1. The complete set of 668 bacterial genomes analysed in this work

Phylum	Class	Bacterial Name	No. of ORFs	Total TFs	No. of wHTHs	No. of families	No. of PaDos	Unique PaDos
Acidobacteria		**Acidobacteria bacterium Ellin345**	4777	175	69	17	26	9
	Solibacteres	**Solibacter usitatus Ellin6076**	7826	273	130	21	32	13
Actinobacteria	Actinobacteria							
		Acidothermus cellulolyticus 11B	2157	70	34	14	12	9
		Arthrobacter sp. FB24	4159	245	128	20	58	14
		Bifidobacterium adolescentis ATCC 15703	1632	76	16	12	7	6
		Bifidobacterium longum	1908	93	27	14	17	8
		Bifidobacterium longum NCC2705	1727	76	26	14	16	8
		Clavibacter michiganensis ssp. sepedonicus	2940	238	61	15	20	10
		Corynebacterium diphtheriae NCTC 13129	2272	70	31	16	15	9
		Corynebacterium efficiens YS-314	3026	118	58	15	23	9
		Corynebacterium glutamicum ATCC 13032 Bielefeld	3057	130	71	17	33	11
		Corynebacterium glutamicum ATCC 13032 Kitasato	2993	127	69	17	32	11
		Corynebacterium glutamicum R	3052	146	80	17	38	11
		Corynebacterium jeikeium K411	2104	71	26	12	13	8
		Corynebacterium urealyticum DSM 7109	2022	54	24	12	13	8
		Frankia alni ACN14a	3707	362	133	22	39	11
		Frankia sp. CcI3	4499	195	63	19	17	11
		Frankia sp. EAN1pec	7191	505	139	21	42	13
		Kineococcus radiotolerans SRS30216	4480	286	110	18	46	12
		Kocuria rhizophila DC2201	2357	101	49	15	22	10
		Leifsonia xyli ssp. xyli CTCB07	2030	112	40	13	16	8
		Mycobacterium abscessus	4920	331	115	18	52	12
		Mycobacterium avium 104	5120	254	71	16	29	11
		Mycobacterium avium ssp. paratuberculosis K-10	4350	225	62	15	23	7
		Mycobacterium bovis AF2122/97	3918	149	53	19	21	10
		Mycobacterium bovis BCG Pasteur 1173P2	3949	149	53	19	21	10
		Mycobacterium gilvum PYR-GCK	5241	316	100	18	35	14
		Mycobacterium leprae TN	1605	33	16	11	8	7
		Mycobacterium marinum M	5423	262	77	18	23	10

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Mycobacterium smegmatis MC2 155	6717	502	207	21	90	14
		Mycobacterium sp. KMS	5460	343	117	21	39	15
		Mycobacterium sp. MCS	5391	340	114	21	37	15
		Mycobacterium tuberculosis CDC1551	4189	145	50	19	23	12
		Mycobacterium tuberculosis F11	3941	148	53	19	21	10
		Mycobacterium tuberculosis H37Ra	4034	147	53	19	21	10
		Mycobacterium tuberculosis H37Rv	3988	148	53	19	21	10
		Mycobacterium ulcerans Agy99	4160	189	66	19	20	10
		Mycobacterium vanbaalenii PYR-1	5979	390	130	22	42	14
		Nocardia farcinica IFM 10152	5683	468	151	24	60	15
		Nocardioides sp. JS614	4645	214	100	20	43	14
		Propionibacterium acnes KPA171202	2297	74	37	14	22	13
		Renibacterium salmoninarum ATCC 33209	3507	174	89	17	42	16
		Rhodococcus jostii RHA1	7211	543	230	23	104	14
		Rubrobacter xylanophilus DSM 9941	3140	139	83	19	34	15
		Salinispora arenicola CNS-205	4917	279	99	20	33	13
		Salinispora tropica CNB-440	4536	262	85	18	30	13
		Streptomyces avermitilis MA-4680	7580	563	216	22	89	15
		Streptomyces coelicolor A3(2)	7768	676	248	24	97	16
		Streptomyces griseus ssp. griseus NBRC 13350	7136	523	199	22	77	16
		Thermobifida fusca YX	3110	137	63	22	27	12
		Tropheryma whipplei TW08/27	783	6	2	2	2	2
		Tropheryma whipplei Twist	808	6	2	2	2	2
Aquificae	Aquificae							
		Aquifex aeolicus VF5	1529	25	10	8	5	3
		Hydrogenobaculum sp. Y04AAS1	1629	24	10	7	5	3
		Sulfurihydrogenibium sp. YO3AOP1	1721	33	14	10	7	6
Bacteroidetes	Bacteroidia							
		Bacteroides fragilis NCTC 9343	4184	131	26	11	12	7
		Bacteroides fragilis YCH46	4577	129	25	11	11	6
		Bacteroides thetaiotaomicron VPI-5482	4816	163	31	11	11	6
		Bacteroides vulgatus ATCC 8482	4066	121	23	11	8	7
		Parabacteroides distasonis ATCC 8503	3850	110	20	12	11	8
		Porphyromonas gingivalis ATCC 33277	2090	30	15	9	6	4

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Porphyromonas gingivalis W83	1909	33	15	9	6	4
	Flavobacteria							
		Flavobacterium johnsoniae UW101	5017	229	73	15	33	9
		Flavobacterium psychrophilum JIP02/86	2412	47	13	9	7	4
		Gramella forsetii KT0803	3584	90	32	14	12	7
		Candidatus Sulcia muelleri GWSS	227	10	0	0	0	0
	Sphingobacteria							
		Cytophaga hutchinsonii ATCC 33406	3785	91	30	12	8	6
		Salinibacter ruber DSM 13855	2801	71	23	14	7	5
		Candidatus Amoebophilus asiaticus 5a2	1334	16	4	4	2	2
Chlamydiae	Chlamydiae							
		Candidatus Protochlamydia amoebophila UWE25	2031	21	4	4	2	2
		Chlamydia muridarum Nigg	940	3	1	1	1	1
		Chlamydia trachomatis 434/Bu	874	4	1	1	1	1
		Chlamydia trachomatis A/HAR-13	911	4	1	1	1	1
		Chlamydia trachomatis D/UW-3/CX	895	4	1	1	1	1
		Chlamydia trachomatis L2b/UCH-1/proctitis	874	4	1	1	1	1
		Chlamydophila abortus S26/3	932	3	1	1	1	1
		Chlamydophila caviae GPIC	998	5	2	2	2	2
		Chlamydophila felis Fe/C-56	1005	5	2	2	2	2
		Chlamydophila pneumoniae AR39	1112	4	2	2	2	2
		Chlamydophila pneumoniae CWL029	1052	4	2	2	2	2
		Chlamydophila pneumoniae J138	1069	4	2	2	2	2
		Chlamydophila pneumoniae TW-183	1113	4	2	2	2	2
Chlorobi	Chlorobia							
		Chlorobaculum parvum NCIB 8327	2043	26	14	9	6	5
		Chlorobium chlorochromatii CaD3	2002	17	2	2	0	0
		Chlorobium phaeobacteroides BS1	2469	43	18	9	4	4
		Chlorobium phaeobacteroides DSM 266	2650	50	18	10	10	7
		Chlorobium phaeovibrioides DSM 265	1753	36	14	9	5	4
		Chlorobium tepidum TLS	2245	23	14	10	8	7
		Chloroherpeton thalassium ATCC 35110	2710	47	17	11	9	7
		Pelodictyon luteolum DSM 273	2083	57	13	9	5	4
		Pelodictyon phaeoclathratiforme BU-1	2707	64	12	9	6	5

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Prosthecochloris aestuarii DSM 271	2263	44	21	9	8	6
Chloroflexi	Chloroflexi							
		Chloroflexus aurantiacus J-10-fl	3853	99	49	17	26	10
		Herpetosiphon aurantiacus ATCC 23779	4976	198	64	17	21	15
		Roseiflexus castenholzii DSM 13941	4330	99	43	16	22	10
		Roseiflexus sp. RS-1	4517	90	39	17	22	10
		Vibrio harveyi ATCC BAA-1116	5919	313	117	15	86	15
	Dehalococcoidetes							
		Dehalococcoides ethenogenes 195	1580	54	20	8	6	6
		Dehalococcoides sp. CBDB1	1458	56	31	10	5	5
	Gloeobacteria							
		Gloeobacter violaceus PCC 7421	4430	139	60	16	22	7
		Acaryochloris marina MBIC11017	6254	192	53	18	19	4
		Anabaena variabilis ATCC 29413	5043	109	45	16	19	6
		Cyanothece sp. ATCC 51142	5211	79	30	12	13	4
		Microcystis aeruginosa NIES-843	6312	157	24	10	11	4
		Nostoc punctiforme PCC 73102	6086	144	48	14	21	7
		Nostoc sp. PCC 7120	5366	139	40	13	15	6
		Prochlorococcus marinus AS9601	1921	9	7	5	3	3
		Prochlorococcus marinus MIT 9211	1854	9	8	6	3	3
		Prochlorococcus marinus MIT 9215	1982	8	7	5	3	3
		Prochlorococcus marinus MIT 9301	1906	10	8	6	3	3
		Prochlorococcus marinus MIT 9303	2997	20	16	9	6	4
		Prochlorococcus marinus MIT 9312	1810	10	8	6	3	3
		Prochlorococcus marinus MIT 9313	2269	20	15	9	5	4
		Prochlorococcus marinus MIT 9515	1905	10	8	6	3	3
		Prochlorococcus marinus NATL1A	2193	12	10	7	4	3
		Prochlorococcus marinus NATL2A	2162	11	10	7	4	3
		Prochlorococcus marinus ssp. marinus CCMP1375	1883	11	10	7	4	3
		Prochlorococcus marinus ssp. pastoris CCMP1986	1717	13	10	7	3	3
		Synechococcus elongatus PCC 6301	2527	36	27	12	9	4
		Synechococcus elongatus PCC 7942	2612	36	27	12	9	4
		Synechococcus sp. CC9311	2892	24	14	8	5	4
		Synechococcus sp. CC9605	2645	21	15	8	6	4

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Synechococcus sp. CC9902	2306	19	12	8	4	4
		Synechococcus sp. JA-2-3Ba(2-13)	2862	32	23	12	9	3
		Synechococcus sp. JA-3-3Ab	2760	48	22	11	8	3
		Synechococcus sp. PCC 7002	2824	36	21	9	9	4
		Synechococcus sp. RCC307	2534	23	13	9	5	4
		Synechococcus sp. WH 7803	2533	22	13	9	5	4
		Synechococcus sp. WH 8102	2519	21	14	10	5	4
		Synechocystis sp. PCC 6803	3179	73	25	11	14	4
		Thermosynechococcus elongatus BP-1	2476	24	17	9	7	3
		Trichodesmium erythraeum IMS101	4451	103	19	10	6	3
Deinococcus-Thermus	Deinococci							
		Deinococcus geothermalis DSM 11300	2330	57	30	14	16	10
		Deinococcus radiodurans R1	2997	82	43	18	18	11
		Thermus thermophilus HB27	1982	43	25	17	14	9
		Thermus thermophilus HB8	1973	41	26	18	14	9
Dictyoglomi	Dictyoglomia							
		Lactobacillus gasseri ATCC 33323	1755	73	36	14	15	7
		Methylibium petroleiphilum PM1	3819	187	83	15	61	8
Elusimicrobia	Elusimicrobia							
		Elusimicrobium minutum Pei191	1529	26	12	10	6	6
Firmicutes	Bacilli							
		Bacillus amyloliquefaciens FZB42	3693	190	108	19	51	20
		Bacillus anthracis Ames	5328	258	129	23	44	17
		Bacillus anthracis Ames Ancestor	5208	258	129	23	44	17
		Bacillus anthracis Sterne	5289	267	132	23	44	17
		Bacillus cereus ATCC 10987	5602	264	133	23	48	17
		Bacillus cereus ATCC 14579	5234	258	126	22	49	17
		Bacillus cereus E33L	5134	276	143	24	51	17
		Bacillus cereus ssp. cytotoxis NVH 391-98	3833	158	77	20	31	14
		Bacillus clausii KSM-K16	4096	284	123	22	55	17
		Bacillus halodurans C-125	4065	226	97	20	42	19
		Bacillus licheniformis ATCC 14580	4196	235	119	22	49	20
		Bacillus pumilus SAFR-032	3678	192	94	22	47	18
		Bacillus subtilis ssp. subtilis 168	4176	219	111	18	54	19

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Firmicutes)	(Bacilli)	**Bacillus thuringiensis ser. konkukian 97-27**	5117	249	129	23	47	17
		Bacillus weihenstephanensis KBAB4	5155	256	130	22	48	17
		Enterococcus faecalis V583	3112	160	70	22	43	20
		Exiguobacterium sibiricum 255-15	3007	131	74	19	24	13
		Geobacillus kaustophilus HTA426	3497	142	62	21	32	19
		Geobacillus thermodenitrificans NG80-2	3392	141	62	19	32	18
		Lactobacillus acidophilus NCFM	1864	70	36	13	17	8
		Lactobacillus brevis ATCC 367	2185	138	73	17	21	10
		Lactobacillus casei ATCC 334	2748	117	60	19	30	15
		Lactobacillus casei BL23	3015	143	75	20	41	15
		Lactobacillus delbrueckii ssp. bulgaricus ATCC 11842	1529	44	29	15	17	9
		Lactobacillus delbrueckii ssp. bulgaricus ATCC BAA-365	1715	49	28	13	15	8
		Lactobacillus fermentum IFO 3956	1843	89	40	17	21	12
		Lactobacillus helveticus DPC 4571	1610	50	22	12	11	8
		Lactobacillus johnsonii NCC 533	1821	75	33	14	14	9
		Lactobacillus plantarum WCFS1	3100	201	106	18	50	17
		Lactobacillus sakei ssp. sakei 23K	1879	111	61	19	24	16
		Lactobacillus salivarius UCC118	1717	76	39	17	20	13
		Lactococcus lactis ssp. cremoris MG1363	2434	137	53	19	23	13
		Lactococcus lactis ssp. cremoris SK11	2384	131	48	19	20	12
		Lactococcus lactis ssp. lactis Il1403	2321	116	47	18	24	14
		Leuconostoc citreum KM20	1702	83	36	17	16	11
		Leuconostoc mesenteroides ssp. mesenteroides ATCC 8293	1970	99	42	18	22	14
		Listeria innocua Clip11262	2968	163	88	19	47	18
		Listeria monocytogenes 4b F2365	2821	162	89	19	52	17
		Listeria monocytogenes EGD-e	2846	170	92	18	56	20
		Listeria welshimeri ser. 6b SLCC5334	2774	144	84	18	47	18
		Lysinibacillus sphaericus C3-41	4584	233	112	23	48	17
		Oceanobacillus iheyensis HTE831	3500	153	71	19	39	19
		Oenococcus oeni PSU-1	1691	80	37	13	17	10
		Pediococcus pentosaceus ATCC 25745	1755	96	52	18	23	13
		Staphylococcus aureus RF122	2509	100	57	18	26	16
		Staphylococcus aureus ssp. aureus COL	2612	105	63	18	32	16

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Firmicutes)	(Bacilli)	Staphylococcus aureus ssp. aureus JH1	2747	109	61	18	32	16
		Staphylococcus aureus ssp. aureus JH9	2697	108	61	18	32	16
		Staphylococcus aureus ssp. aureus MRSA252	2650	111	62	18	33	16
		Staphylococcus aureus ssp. aureus MSSA476	2571	103	61	18	31	16
		Staphylococcus aureus ssp. aureus Mu3	2690	108	61	18	32	17
		Staphylococcus aureus ssp. aureus Mu50	2696	108	61	18	32	17
		Staphylococcus aureus ssp. aureus MW2	2624	103	59	18	27	16
		Staphylococcus aureus ssp. aureus N315	2583	103	61	18	32	17
		Staphylococcus aureus ssp. aureus NCTC 8325	2891	107	63	18	32	16
		Staphylococcus aureus ssp. aureus Newman	2614	104	62	18	28	16
		Staphylococcus aureus ssp. aureus USA300	2560	109	64	18	30	16
		Staphylococcus aureus ssp. aureus USA300_TCH1516	2657	100	57	15	27	13
		Staphylococcus epidermidis ATCC 12228	2419	80	49	18	25	16
		Staphylococcus epidermidis RP62A	2493	86	55	19	28	19
		Staphylococcus haemolyticus JCSC1435	2692	122	57	17	18	12
		Staphylococcus saprophyticus ssp. saprophyticus ATCC 15305	2446	124	69	16	27	13
		Streptococcus agalactiae 2603V/R	2124	92	38	15	26	16
		Streptococcus agalactiae A909	1996	86	39	15	29	19
		Streptococcus agalactiae NEM316	2094	87	41	16	30	16
		Streptococcus gordonii Challis subCH1	2051	101	46	16	29	16
		Streptococcus mutans UA159	1960	105	50	18	31	17
		Streptococcus pneumoniae CGSP14	2206	78	35	14	23	14
		Streptococcus pneumoniae D39	1914	67	34	14	25	15
		Streptococcus pneumoniae G54	2114	71	34	13	21	13
		Streptococcus pneumoniae Hungary19A-6	2155	72	33	13	21	13
		Streptococcus pneumoniae R6	2042	75	36	14	25	15
		Streptococcus pneumoniae TIGR4	2105	77	36	13	24	15
		Streptococcus pyogenes M1 GAS	1696	75	38	17	31	19
		Streptococcus pyogenes Manfredo	1745	81	40	18	29	17
		Streptococcus pyogenes MGAS10270	1986	91	42	17	31	19
		Streptococcus pyogenes MGAS10394	1886	78	41	17	33	20
		Streptococcus pyogenes MGAS10750	1979	86	41	17	32	19
		Streptococcus pyogenes MGAS2096	1898	84	41	17	30	18
		Streptococcus pyogenes MGAS315	1865	87	40	18	32	20

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Firmicutes)	(Bacilli)	Streptococcus pyogenes MGAS5005	1865	78	42	18	33	20
		Streptococcus pyogenes MGAS6180	1894	89	41	17	32	19
		Streptococcus pyogenes MGAS8232	1839	82	39	17	28	16
		Streptococcus pyogenes MGAS9429	1877	78	40	17	33	20
		Streptococcus pyogenes SSI-1	1859	87	40	18	32	20
		Streptococcus sanguinis SK36	2270	106	50	17	24	13
		Streptococcus suis 05ZYH33	2186	81	40	16	15	9
		Streptococcus suis 98HAH33	2185	89	41	15	16	10
		Streptococcus thermophilus CNRZ1066	1915	55	24	14	14	10
		Streptococcus thermophilus LMD-9	1709	53	22	13	13	10
		Streptococcus thermophilus LMG 18311	1889	60	24	14	13	9
	Clostridia							
		Alkaliphilus metalliredigens QYMF	4625	206	83	18	34	18
		Alkaliphilus oremlandii OhILAs	2836	145	52	17	22	13
		Caldicellulosiruptor saccharolyticus DSM 8903	2679	113	49	17	22	14
		Candidatus Desulforudis audaxviator MP104C	2157	46	22	16	14	11
		Carboxydothermus hydrogenoformans Z-2901	2620	81	43	17	24	14
		Clostridium acetobutylicum ATCC 824	3671	184	87	18	42	17
		Clostridium beijerinckii NCIMB 8052	5020	372	157	22	71	22
		Clostridium botulinum A ATCC 19397	3550	198	80	19	37	16
		Clostridium botulinum A ATCC 3502	3572	187	80	19	37	16
		Clostridium botulinum A Hall	3401	188	79	19	36	16
		Clostridium botulinum A3 Loch Maree	3654	211	88	18	41	16
		Clostridium botulinum B Eklund 17B	3425	143	68	19	41	21
		Clostridium botulinum B1 Okra	3652	202	87	19	37	16
		Clostridium botulinum F Langeland	3631	201	82	19	37	16
		Clostridium difficile 630	3738	259	113	22	54	16
		Clostridium kluyveri DSM 555	3838	217	90	21	32	16
		Clostridium novyi NT	2315	77	39	17	20	14
		Clostridium perfringens 13	2660	97	53	20	29	16
		Clostridium perfringens ATCC 13124	2876	105	58	19	29	16
		Clostridium perfringens SM101	2546	100	55	19	26	16
		Clostridium phytofermentans ISDg	3902	279	86	19	32	15
		Clostridium tetani E88	2377	87	47	16	23	15

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Clostridium thermocellum ATCC 27405	3189	93	44	19	19	14
		Desulfitobacterium hafniense DCB-2	4883	71	15	7	7	7
		Desulfitobacterium hafniense Y51	5060	344	154	19	66	15
		Desulfotomaculum reducens MI-1	3276	127	48	19	22	12
		Finegoldia magna ATCC 29328	1631	50	26	16	12	9
		Heliobacterium modesticaldum Ice1	3000	72	41	20	19	13
		Moorella thermoacetica ATCC 39073	2463	106	56	19	26	13
		Pelotomaculum thermopropionicum SI	2919	89	41	18	15	11
		Symbiobacterium thermophilum IAM 14863	3338	111	51	19	24	12
		Syntrophomonas wolfei ssp. wolfei Goettingen	2504	69	32	16	15	12
		Thermoanaerobacter pseudethanolicus ATCC 33223	2243	94	44	17	44	21
		Thermoanaerobacter sp. X514	2349	90	47	15	40	19
		Thermoanaerobacter tengcongensis MB4	2588	85	36	16	28	18
Fusobacteria	Fusobacteria							
		Fusobacterium nucleatum ssp. nucleatum ATCC 25586	2063	49	27	13	14	9
Planctomycetes	Planctomycetacia							
		Rhodopirellula baltica SH 1	7325	107	39	15	9	6
Proteobacteria	Alpha							
		Agrobacterium tumefaciens C58	4616	307	161	15	80	11
		Anaplasma marginale St. Maries	948	7	1	1	1	1
		Anaplasma phagocytophilum HZ	1264	7	1	1	1	1
		Azorhizobium caulinodans ORS 571	4717	298	177	17	112	11
		Bartonella bacilliformis KC583	1283	15	6	4	2	2
		Bartonella henselae Houston-1	1488	32	9	8	5	5
		Bartonella quintana Toulouse	1142	16	7	6	3	3
		Bartonella tribocorum CIP 105476	2069	55	9	7	5	5
		Beijerinckia indica ssp. indica ATCC 9039	3569	200	90	18	58	11
		Bradyrhizobium japonicum USDA 110	8317	497	227	16	123	8
		Bradyrhizobium sp. BTAi1	7393	377	196	19	103	9
		Bradyrhizobium sp. ORS278	6717	307	166	20	91	10
		Brucella abortus bv. 1 9-941	3084	149	88	16	45	8
		Brucella abortus S19	3000	154	90	16	45	8
		Brucella canis ATCC 23365	3251	146	85	16	45	8

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Alpha)	**Brucella melitensis 16M**	3198	152	89	16	44	8
		Brucella melitensis biov. Abortus 2308	3034	148	89	16	44	8
		Brucella ovis ATCC 25840	2890	137	77	16	38	8
		Brucella suis 1330	3271	144	84	16	43	7
		Brucella suis ATCC 23445	3241	146	84	16	42	8
		Burkholderia ambifaria MC40-6	6423	589	326	23	257	10
		Burkholderia pseudomallei 1106a	7174	371	193	21	142	10
		Candidatus Pelagibacter ubique HTCC1062	1354	16	9	8	2	2
		Caulobacter crescentus CB15	3737	172	75	17	25	9
		Caulobacter sp. K31	5061	316	121	19	59	11
		Dinoroseobacter shibae DFL 12	3584	153	80	17	44	9
		Ehrlichia canis Jake	925	7	1	1	1	1
		Ehrlichia chaffeensis Arkansas	1105	7	1	1	1	1
		Ehrlichia ruminantium Gardel	950	8	2	1	1	1
		Ehrlichia ruminantium Welgevonden	888	8	2	1	1	1
		Ehrlichia ruminantium Welgevonden	958	8	2	1	1	1
		Erythrobacter litoralis HTCC2594	3011	74	30	12	14	5
		Gluconacetobacter diazotrophicus PAl 5	3778	180	85	15	53	9
		Gluconobacter oxydans 621H	2432	76	45	17	28	8
		Granulibacter bethesdensis CGDNIH1	2437	75	39	10	24	5
		Hyphomonas neptunium ATCC 15444	3505	155	59	15	25	8
		Jannaschia sp. CCS1	4212	228	120	18	62	9
		Magnetospirillum magneticum AMB-1	4559	141	68	16	32	9
		Maricaulis maris MCS10	3063	141	56	18	19	6
		Mesorhizobium loti MAFF303099	6743	468	245	19	130	12
		Mesorhizobium sp. BNC1	4064	220	133	19	75	11
		Methylobacterium extorquens PA1	4829	199	96	16	63	10
		Methylobacterium radiotolerans JCM 2831	5686	245	128	15	84	10
		Methylobacterium sp. 4-46	6609	298	157	15	98	11
		Neorickettsia sennetsu Miyayama	932	6	1	1	1	1
		Nitrobacter hamburgensis X14	3804	118	40	10	21	6
		Nitrobacter winogradskyi Nb-255	3122	128	35	12	20	6
		Novosphingobium aromaticivorans DSM 12444	3324	156	62	18	37	7
		Ochrobactrum anthropi ATCC 49188	4424	291	163	18	92	11

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Alpha)	Orientia tsutsugamushi Boryong	1182	10	1	1	0	0
		Orientia tsutsugamushi Ikeda	1967	41	2	2	0	0
		Parvibaculum lavamentivorans DS-1	3636	190	71	16	23	7
		Rhizobium etli CFN 42	4035	261	136	16	70	10
		Rhizobium etli CIAT 652	4343	267	144	18	83	10
		Rhizobium leguminosarum bv. viciae 3841	4694	373	194	16	108	11
		Rhodobacter sphaeroides 2.4.1	3857	176	81	19	46	8
		Rhodobacter sphaeroides ATCC 17025	3111	103	49	14	28	9
		Rhodobacter sphaeroides ATCC 17029	4024	182	88	20	47	8
		Rhodopseudomonas palustris BisA53	4878	166	69	17	39	9
		Rhodopseudomonas palustris BisB18	4886	207	96	19	65	9
		Rhodopseudomonas palustris BisB5	4397	162	80	16	44	8
		Rhodopseudomonas palustris CGA009	4838	233	108	18	59	9
		Rhodopseudomonas palustris HaA2	4683	200	102	16	62	11
		Rhodopseudomonas palustris TIE-1	5246	242	111	17	58	9
		Rhodospirillum rubrum ATCC 11170	3791	206	88	17	50	9
		Rickettsia akari Hartford	1258	8	1	1	0	0
		Rickettsia bellii OSU 85-389	1475	21	3	3	0	0
		Rickettsia bellii RML369-C	1429	22	2	2	0	0
		Rickettsia canadensis McKiel	1090	6	0	0	0	0
		Rickettsia conorii Malish 7	1374	11	2	2	0	0
		Rickettsia felis URRWXCal2	1400	14	2	2	0	0
		Rickettsia massiliae MTU5	968	13	1	1	0	0
		Rickettsia prowazekii Madrid E	835	6	0	0	0	0
		Rickettsia rickettsii Iowa	1384	13	2	2	0	0
		Rickettsia rickettsii Sheila Smith	1343	13	2	2	0	0
		Rickettsia typhi Wilmington	838	6	0	0	0	0
		Roseobacter denitrificans OCh 114	3946	169	88	17	45	9
		Silicibacter pomeroyi DSS-3	3810	249	136	18	80	10
		Silicibacter sp. TM1040	3030	167	96	15	62	9
		Sinorhizobium medicae WSM419	3529	212	115	16	60	9
		Sinorhizobium meliloti 1021	3359	207	106	15	53	9
		Sphingobium japonicum UT26S	4118	272	121	17	91	20
		Sphingomonas wittichii RW1	4850	324	148	16	90	7
		Sphingopyxis alaskensis RB2256	3165	143	67	16	28	7

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Alpha)	**Wolbachia endosymbiont of Drosophila melanogaster**	1195	19	2	2	1	1
		Wolbachia endosymbiont TRS of Brugia malayi	805	3	0	0	0	0
		Wolbachia pipientis	1275	3	0	0	0	0
		Xanthobacter autotrophicus Py2	4746	240	130	15	74	11
		Yersinia pseudotuberculosis PB1/+	4461	264	144	19	75	7
		Zymomonas mobilis ssp. mobilis ZM4	1736	46	28	11	19	6
	Beta							
		Acidovorax avenae ssp. citrulli AAC00-1	4709	251	127	15	95	8
		Acidovorax sp. JS42	4007	198	103	15	74	9
		Aromatoleum aromaticum EbN1	4124	128	52	14	31	5
		Azoarcus sp. BH72	3989	180	83	14	64	8
		Bordetella avium 197N	3381	196	127	16	89	8
		Bordetella bronchiseptica RB50	4994	418	293	19	212	9
		Bordetella parapertussis 12822	4185	351	242	19	170	9
		Bordetella pertussis Tohama I	3436	462	167	18	119	9
		Bordetella petrii DSM 12804	5027	389	240	21	169	8
		Burkholderia ambifaria AMMD	6565	615	336	23	268	9
		Burkholderia cenocepacia AU 1054	6477	645	351	22	279	10
		Burkholderia cenocepacia HI2424	6763	685	372	22	294	10
		Burkholderia cenocepacia MC0-3	7008	715	383	23	301	10
		Burkholderia mallei ATCC 23344	5024	310	167	21	119	10
		Burkholderia mallei NCTC 10229	5509	305	166	21	119	10
		Burkholderia mallei NCTC 10247	6015	307	169	21	121	10
		Burkholderia mallei SAVP1	5188	259	146	21	106	10
		Burkholderia multivorans ATCC 17616	6121	528	294	21	221	9
		Burkholderia phymatum STM815	5421	395	207	21	142	9
		Burkholderia pseudomallei 1710b	6345	377	196	21	143	10
		Burkholderia pseudomallei 668	7116	374	194	21	143	9
		Burkholderia pseudomallei K96243	5728	380	197	21	143	10
		Burkholderia sp. 383	7717	806	438	22	342	10
		Burkholderia thailandensis E264	5651	379	190	19	143	8
		Burkholderia vietnamiensis G4	6484	510	261	20	203	9
		Burkholderia xenovorans LB400	8702	697	361	22	252	10
		Chromobacterium violaceum ATCC 12472	4407	234	128	15	89	10

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Beta)	**Cupriavidus taiwanensis**	5377	171	104	17	72	8
		Dechloromonas aromatica RCB	4171	148	72	15	48	8
		Delftia acidovorans SPH-1	6040	524	307	19	222	10
		Herminiimonas arsenicoxydans	3495	117	67	15	42	7
		Janthinobacterium sp. Marseille	3697	217	114	18	69	7
		Leptothrix cholodnii SP-6	4363	197	95	15	69	7
		Methylobacillus flagellatus KT	2753	94	34	14	23	8
		Neisseria gonorrhoeae FA 1090	2002	44	20	10	8	4
		Neisseria gonorrhoeae NCCP11945	2662	45	21	10	8	4
		Neisseria meningitidis 053442	2020	37	17	9	7	3
		Neisseria meningitidis MC58	2063	45	19	9	9	4
		Neisseria meningitidis Z2491	1909	47	19	9	9	4
		Nitrosomonas europaea ATCC 19718	2461	80	17	9	9	4
		Nitrosomonas eutropha C91	2444	62	18	9	11	5
		Nitrosospira multiformis ATCC 25196	2757	72	24	13	16	6
		Polaromonas naphthalenivorans CJ2	4084	205	109	17	80	7
		Polaromonas sp. JS666	4817	321	207	20	144	13
		Polynucleobacter necessarius ssp. asymbioticus QLW-P1DMWA-1	2077	43	26	13	13	5
		Polynucleobacter necessarius ssp. necessarius STIR1	1508	23	14	10	7	4
		Ralstonia eutropha H16	6629	542	342	25	242	10
		Ralstonia eutropha JMP134	5846	482	294	20	209	9
		Ralstonia metallidurans CH34	3604	460	260	19	181	8
		Ralstonia solanacearum GMI1000	3437	195	97	15	67	7
		Rhodoferax ferrireducens T118	4170	198	107	16	67	9
		Shewanella woodyi ATCC 51908	4880	251	107	14	81	13
		Thiobacillus denitrificans ATCC 25259	2827	74	33	15	19	8
		Verminephrobacter eiseniae EF01-2	4908	267	148	16	96	9
	Delta							
		Anaeromyxobacter dehalogenans 2CP-C	4346	159	55	13	35	10
		Anaeromyxobacter sp. Fw109-5	4466	138	40	14	18	9
		Arthrobacter aurescens TC1	4041	252	141	23	66	15
		Bacillus thuringiensis Al Hakam	4736	248	121	23	54	17
		Bdellovibrio bacteriovorus HD100	3587	78	34	9	23	4
		Clavibacter michiganensis ssp. michiganensis	2983	168	70	14	24	9

		NCPPB 382						
(Proteobacteria)	(Delta)	**Desulfococcus oleovorans Hxd3**	3265	102	24	11	7	7
		Desulfotalea psychrophila LSv54	3116	72	26	10	9	4
		Desulfovibrio desulfuricans ssp. desulfuricans G20	3780	127	50	16	22	10
		Desulfovibrio vulgaris DP4	2941	99	39	13	18	7
		Desulfovibrio vulgaris Hildenborough	3380	100	38	13	17	7
		Geobacter lovleyi SZ	3606	117	29	13	19	10
		Geobacter metallireducens GS-15	3520	107	34	14	17	7
		Geobacter sulfurreducens PCA	3447	106	44	14	22	8
		Geobacter uraniireducens Rf4	4358	168	49	15	20	9
		Lactobacillus reuteri DSM 20016	1900	79	42	18	22	13
		Lawsonia intracellularis PHE/MN1-00	1183	18	4	4	2	2
		Myxococcus xanthus DK 1622	7316	224	72	19	39	12
		Neisseria meningitidis FAM18	1917	44	20	9	9	4
		Pelobacter carbinolicus DSM 2380	3352	91	39	15	20	11
		Pelobacter propionicus DSM 2379	3576	102	35	14	16	9
		Saccharopolyspora erythraea NRRL 2338	7197	596	240	26	112	15
		Sorangium cellulosum So ce 56	9381	314	113	18	67	10
		Syntrophobacter fumaroxidans MPOB	4064	107	38	17	14	8
		Syntrophus aciditrophicus SB	3168	79	25	10	7	5
	Epsilon							
		Arcobacter butzleri RM4018	2259	60	26	12	9	4
		Campylobacter concisus 13826	1929	20	12	8	4	2
		Campylobacter curvus 525.92	1931	32	18	9	6	3
		Campylobacter fetus ssp. fetus 82-40	1719	24	16	11	7	4
		Campylobacter hominis ATCC BAA-381	1682	15	8	7	2	1
		Campylobacter jejuni RM1221	1838	17	9	8	2	2
		Campylobacter jejuni ssp. doylei 269.97	1731	13	6	5	1	1
		Campylobacter jejuni ssp. jejuni 81-176	1653	15	9	7	1	1
		Campylobacter jejuni ssp. jejuni 81116	1626	14	8	7	1	1
		Campylobacter jejuni ssp. jejuni NCTC 11168	1623	16	10	8	2	2
		Helicobacter acinonychis Sheeba	1613	7	4	3	0	0
		Helicobacter hepaticus ATCC 51449	1876	20	11	9	2	1
		Helicobacter pylori 26695	1573	5	2	2	0	0

(Proteobacteria)	(Delta)	Helicobacter pylori HPAG1	1531	5	2	2	0	0
		Helicobacter pylori J99	1488	5	2	2	0	0
		Helicobacter pylori Shi470	1568	5	2	2	0	0
		Nitratiruptor sp. SB155-2	1843	21	12	9	4	2
		Sulfurimonas denitrificans DSM 1251	2096	31	13	9	7	3
		Sulfurovum sp. NBC37-1	2438	51	14	9	7	4
		Wolinella succinogenes DSM 1740	2043	43	20	11	8	3
	Gamma							
		Acinetobacter baumannii ACICU	3667	229	117	16	77	6
		Acinetobacter baumannii ATCC 17978	3351	154	65	16	36	6
		Acinetobacter baumannii AYE	3607	231	121	16	80	6
		Acinetobacter baumannii SDF	2913	119	56	11	39	5
		Acinetobacter sp. ADP1	3307	179	90	15	61	6
		Actinobacillus pleuropneumoniae L20	2012	64	29	11	24	9
		Actinobacillus pleuropneumoniae ser. 3 JL03	2036	61	28	11	23	9
		Actinobacillus pleuropneumoniae ser. 7 AP76	2131	71	32	12	25	10
		Actinobacillus succinogenes 130Z	2079	92	50	16	41	12
		Aeromonas hydrophila ssp. hydrophila ATCC 7966	4121	234	108	13	83	12
		Aeromonas salmonicida ssp. salmonicida A449	4085	236	96	14	73	12
		Alcanivorax borkumensis SK2	2755	107	38	13	23	8
		Alkalilimnicola ehrlichei MLHE-1	2865	80	35	15	18	6
		Baumannia cicadellinicola Hc	595	8	6	3	3	2
		Buchnera aphidicola APS	564	4	0	0	0	0
		Buchnera aphidicola Bp	504	2	0	0	0	0
		Buchnera aphidicola Cc	357	2	0	0	0	0
		Buchnera aphidicola Sg	546	4	1	1	1	1
		Candidatus Blochmannia floridanus	583	7	4	2	1	1
		Candidatus Blochmannia pennsylvanicus BPEN	610	8	5	4	2	2
		Candidatus Carsonella ruddii PV	182	1	0	0	0	0
		Candidatus Ruthia magnifica Cm	976	10	4	4	1	1
		Candidatus Vesicomyosocius okutanii HA	937	8	3	3	0	0
		Cellvibrio japonicus Ueda107	3754	121	47	15	28	10
		Chromohalobacter salexigens DSM 3043	3319	195	110	16	73	10

(Proteobacteria)	(Gamma)	**Citrobacter koseri ATCC BAA-895**	4980	257	106	18	81	14
		Colwellia psychrerythraea 34H	4910	240	110	18	89	13
		Coxiella burnetii Dugway 5J108-111	1993	33	11	9	7	5
		Coxiella burnetii RSA 331	1930	23	7	7	4	4
		Coxiella burnetii RSA 493	1817	24	7	7	4	4
		Dehalococcoides sp. BAV1	1371	43	17	9	3	3
		Dichelobacter nodosus VCS1703A	1280	18	7	6	2	2
		Enterobacter sakazakii ATCC BAA-894	4256	226	98	16	71	16
		Enterobacter sp. 638	4115	273	116	17	84	16
		Erwinia tasmaniensis Et1/99	3427	173	76	14	62	13
		Escherichia coli 536	4619	261	106	16	79	16
		Escherichia coli APEC O1	4428	245	106	16	78	16
		Escherichia coli ATCC 8739	4199	247	107	17	83	18
		Escherichia coli CFT073	5338	271	111	17	82	17
		Escherichia coli E24377A	4749	261	104	16	74	14
		Escherichia coli HS	4378	239	104	16	75	15
		Escherichia coli K-12 subW3110	4229	246	107	16	79	15
		Escherichia coli K12	4145	243	107	16	79	15
		Escherichia coli K12 subDH10B	4126	236	100	16	76	15
		Escherichia coli O157:H7 EDL933	5298	275	103	17	77	17
		Escherichia coli O157:H7 Sakai	5229	284	104	17	77	17
		Escherichia coli SMS-3-5	4743	272	121	17	91	20
		Escherichia coli UTI89	5021	256	109	16	79	16
		Francisella novicida U112	1719	29	20	8	11	3
		Francisella philomiragia ssp. philomiragia ATCC 25017	1911	32	22	8	11	4
		Francisella tularensis ssp. holarctica	1754	82	15	7	9	3
		Francisella tularensis ssp. holarctica FTNF002-00	1581	80	13	7	7	3
		Francisella tularensis ssp. holarctica OSU18	1555	60	14	7	7	3
		Francisella tularensis ssp. mediasiatica FSC147	1406	20	12	6	6	2
		Francisella tularensis ssp. tularensis FSC198	1605	71	16	6	11	2
		Francisella tularensis ssp. tularensis SCHU S4	1604	71	16	6	11	2
		Francisella tularensis ssp. tularensis WY96-3418	1634	81	20	7	13	3
		Haemophilus ducreyi 35000HP	1717	37	16	11	13	8
		Haemophilus influenzae 86-028NP	1792	51	21	10	18	8

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Gamma)	Haemophilus influenzae PittEE	1613	47	21	10	17	8
		Haemophilus influenzae PittGG	1661	47	22	10	17	9
		Haemophilus influenzae Rd KW20	1657	46	22	10	19	9
		Haemophilus somnus 129PT	1792	52	27	9	22	8
		Haemophilus somnus 2336	1980	58	26	10	21	8
		Hahella chejuensis KCTC 2396	6778	322	130	17	84	9
		Halorhodospira halophila SL1	2407	55	26	13	14	8
		Idiomarina loihiensis L2TR	2628	94	43	13	28	10
		Klebsiella pneumoniae ssp. pneumoniae MGH 78578	4776	380	185	18	145	15
		Legionella pneumophila Corby	3204	65	28	11	20	6
		Legionella pneumophila Lens	2878	71	27	13	18	6
		Legionella pneumophila Paris	3027	66	28	12	20	6
		Legionella pneumophila ssp. pneumophila Philadelphia 1	2942	67	25	9	18	5
		Mannheimia succiniciproducens MBEL55E	2369	95	48	13	41	9
		Marinobacter aquaeolei VT8	3858	168	59	14	37	9
		Marinomonas sp. MWYL1	4439	343	176	16	134	12
		Methylococcus capsulatus Bath	2956	61	24	14	10	5
		Mycobacterium sp. JLS	5739	386	131	20	41	15
		Nitrosococcus oceani ATCC 19707	2974	76	20	12	9	4
		Pasteurella multocida ssp. multocida Pm70	2015	50	25	11	19	10
		Pectobacterium atrosepticum SCRI1043	4472	268	130	17	92	15
		Photobacterium profundum SS9	5491	316	142	14	113	16
		Photorhabdus luminescens ssp. laumondii TTO1	4683	299	66	13	49	13
		Pseudoalteromonas atlantica T6c	4281	211	98	17	66	12
		Pseudoalteromonas haloplanktis TAC125	3486	139	70	14	54	12
		Pseudomonas aeruginosa PA7	6286	426	189	15	144	9
		Pseudomonas aeruginosa PAO1	5571	429	201	18	154	11
		Pseudomonas aeruginosa UCBPP-PA14	5892	444	203	19	154	11
		Pseudomonas entomophila L48	5134	340	168	15	133	12
		Pseudomonas fluorescens Pf-5	6138	479	233	17	191	13
		Pseudomonas fluorescens Pf0-1	5722	405	204	15	160	12
		Pseudomonas mendocina ymp	4594	292	121	14	91	12
		Pseudomonas putida F1	5250	385	201	15	158	11

(Proteobacteria)	(Gamma)	Pseudomonas putida GB-1	5408	404	209	18	164	12
		Pseudomonas putida KT2440	5350	371	191	16	150	11
		Pseudomonas putida W619	5182	375	191	18	143	11
		Pseudomonas stutzeri A1501	4128	182	70	12	55	10
		Pseudomonas syringae pv. phaseolicola 1448A	4985	285	141	15	99	10
		Pseudomonas syringae pv. tomato DC3000	5481	298	141	17	103	11
		Psychrobacter arcticus 273-4	2120	56	24	8	17	3
		Psychrobacter cryohalolentis K5	2467	84	36	11	26	4
		Psychrobacter sp. PRwf-1	2370	63	30	10	20	2
		Psychromonas ingrahamii 37	3545	130	61	14	38	11
		Saccharophagus degradans 2-40	4007	146	52	18	29	11
		Salmonella enterica ssp. arizonae ser. 62:z4,z23:--	4498	237	88	16	67	16
		Salmonella enterica ssp. enterica ser. Choleraesuis str. SC-B67	4406	250	106	15	77	14
		Salmonella enterica ssp. enterica ser. Heidelberg SL476	4651	256	118	15	85	14
		Salmonella enterica ssp. enterica ser. Newport SL254	4613	254	119	16	90	17
		Salmonella enterica ssp. enterica ser. Paratyphi A ATCC 9150	4091	249	109	16	81	16
		Salmonella enterica ssp. enterica ser. Paratyphi B SPB7	5592	271	121	16	89	17
		Salmonella enterica ssp. enterica ser. Schwarzengrund str. CVM19633	4503	250	111	15	82	14
		Salmonella enterica ssp. enterica ser. Typhi CT18	4391	253	110	16	80	16
		Salmonella enterica ssp. enterica ser. Typhi Ty2	4314	252	112	16	82	17
		Salmonella typhimurium LT2	4423	272	117	15	85	14
		Serratia proteamaculans 568	4891	400	193	19	153	16
		Shewanella amazonensis SB2B	3645	160	77	12	59	12
		Shewanella baltica OS155	4307	204	93	16	71	13
		Shewanella baltica OS185	4323	216	95	17	71	13
		Shewanella baltica OS195	4499	225	99	15	74	13
		Shewanella denitrificans OS217	3754	149	57	16	38	11
		Shewanella frigidimarina NCIMB 400	4029	186	92	14	66	12
		Shewanella halifaxensis HAW-EB4	4278	203	96	14	79	14
		Shewanella loihica PV-4	3859	185	91	14	67	12

(Proteobacteria)	(Gamma)	**Shewanella oneidensis MR-1**	4318	220	88	16	68	13
		Shewanella pealeana ATCC 700345	4241	201	103	16	82	14
		Shewanella putrefaciens CN-32	3972	172	88	15	60	12
		Shewanella sediminis HAW-EB3	4497	247	123	14	99	13
		Shewanella sp. ANA-3	4111	195	94	14	68	14
		Shewanella sp. MR-4	3924	169	82	13	60	13
		Shewanella sp. MR-7	4006	175	85	13	61	13
		Shewanella sp. W3-18-1	4044	176	84	15	57	12
		Shigella boydii CDC 3083-94	4246	185	81	15	56	14
		Shigella boydii Sb227	4133	191	83	17	62	17
		Shigella dysenteriae Sd197	4274	176	77	15	56	14
		Shigella flexneri 2a 2457T	4060	200	89	18	61	14
		Shigella flexneri 2a 301	4650	199	85	18	61	14
		Shigella flexneri 5 8401	4114	208	89	19	60	14
		Shigella sonnei Ss046	4218	223	80	16	57	15
		Sodalis glossinidius morsitans	2432	122	50	15	30	13
		Stenotrophomonas maltophilia K279a	4386	236	105	14	75	12
		Stenotrophomonas maltophilia R551-3	4039	247	108	15	77	13
		Thiomicrospira crunogena XCL-2	2196	43	22	12	16	8
		Vibrio cholerae O1 biov. eltor N16961	3834	178	76	15	60	13
		Vibrio cholerae O395	3875	189	73	14	58	13
		Vibrio fischeri ES114	3760	200	94	13	79	11
		Vibrio parahaemolyticus RIMD 2210633	4832	273	128	17	100	13
		Vibrio vulnificus YJ016	4955	265	113	15	87	13
		Wigglesworthia glossinidia	611	7	2	1	0	0
		Xanthomonas axonopodis pv. citri 306	4312	171	72	16	47	11
		Xanthomonas campestris pv. campestris 8004	4271	170	78	15	49	10
		Xanthomonas campestris pv. campestris ATCC 33913	4179	170	79	16	50	11
		Xanthomonas campestris pv. vesicatoria 85-10	4487	182	84	16	53	11
		Xanthomonas oryzae pv. oryzae KACC10331	4064	139	47	15	28	9
		Xanthomonas oryzae pv. oryzae MAFF 311018	4372	141	47	15	28	9
		Xylella fastidiosa 9a5c	2766	55	21	10	13	8
		Xylella fastidiosa M12	2104	42	17	10	10	8
		Xylella fastidiosa M23	2161	42	17	10	10	8

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

(Proteobacteria)	(Gamma)	Xylella fastidiosa Temecula1	2034	40	17	10	10	8
		Yersinia enterocolitica ssp. enterocolitica 8081	3978	224	94	15	71	13
		Yersinia pestis Angola	3831	149	67	15	49	15
		Yersinia pestis Antiqua	4164	178	74	15	55	15
		Yersinia pestis biov. Microtus 91001	3890	171	72	15	54	15
		Yersinia pestis CO92	3885	170	71	15	53	15
		Yersinia pestis KIM	3978	170	72	15	52	14
		Yersinia pestis Nepal516	3981	172	71	15	51	15
		Yersinia pestis Pestoides F	3849	176	74	15	55	15
		Yersinia pseudotuberculosis IP 31758	4124	184	75	15	56	15
		Yersinia pseudotuberculosis PB1/+	4150	178	73	15	54	15
		Yersinia pseudotuberculosis YPIII	4192	188	74	15	53	14
Spirochaetes	Spirochaetes							
		Borrelia afzelii PKo	850	3	1	1	0	0
		Borrelia burgdorferi B31	851	3	1	1	0	0
		Borrelia garinii PBi	832	3	1	1	0	0
		Borrelia turicatae 91E135	818	4	1	1	0	0
		Leptospira biflexa ser. Patoc Patoc 1 (Ames)	3543	101	32	10	8	6
		Leptospira biflexa ser. Patoc Patoc 1 (Paris)	3667	97	30	10	7	5
		Leptospira borgpetersenii ser. Hardjo-bovis JB197	2880	51	19	8	5	5
		Leptospira borgpetersenii ser. Hardjo-bovis L550	2945	57	20	8	5	5
		Leptospira interrogans ser. Copenhageni Fiocruz L1-130	3667	55	22	10	6	4
		Leptospira interrogans ser. Lai 56601	3702	56	21	10	5	4
		Treponema denticola ATCC 35405	2767	67	25	13	9	5
		Treponema pallidum ssp. pallidum Nichols	1028	5	1	1	1	1
		Treponema pallidum ssp. pallidum SS14	1028	5	1	1	1	1
Tenericutes	Mollicutes							
		Acholeplasma laidlawii PG-8A	1380	38	18	10	6	5
		Aster yellows witches-broom phytoplasma AYWB	671	2	1	1	1	1
		Candidatus Phytoplasma mali	479	2	1	1	1	1
		Mesoplasma florum L1	682	9	6	5	2	2
		Mycoplasma agalactiae PG2	742	2	1	1	1	1

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

		Mycoplasma arthritidis 158L3-1	631	2	1	1	1	1
		Mycoplasma capricolum ssp. capricolum ATCC 27343	812	6	5	5	3	2
		Mycoplasma gallisepticum R	723	3	3	3	1	1
		Mycoplasma genitalium G37	475	3	3	3	1	1
		Mycoplasma hyopneumoniae 232	691	1	1	1	1	1
		Mycoplasma hyopneumoniae 7448	657	1	1	1	1	1
		Mycoplasma hyopneumoniae J	657	1	1	1	1	1
		Mycoplasma mobile 163K	633	5	3	3	2	2
		Mycoplasma mycoides ssp. mycoides SC PG1	1017	7	5	5	3	2
		Mycoplasma penetrans HF-2	1037	9	6	6	2	2
		Mycoplasma pneumoniae M129	689	3	3	3	1	1
		Mycoplasma pulmonis UAB CTIP	782	3	2	2	1	1
		Mycoplasma synoviae 53	659	2	2	2	1	1
		Onion yellows phytoplasma OY-M	750	2	1	1	1	1
		Ureaplasma parvum ser. 3 ATCC 27815	609	3	2	2	1	1
		Ureaplasma parvum ser. 3 ATCC 700970	614	3	2	2	1	1
Thermotogae	Thermotogae							
		Fervidobacterium nodosum Rt17-B1	1750	36	21	13	9	7
		Petrotoga mobilis SJ95	1898	48	29	12	9	8
		Thermosipho melanesiensis BI429	1879	38	27	16	11	9
		Thermotoga lettingae TMO	2040	66	43	15	24	11
		Thermotoga maritima MSB8	1858	46	31	17	24	11
		Thermotoga petrophila RKU-1	1785	42	30	16	22	11
		Thermotoga sp. RQ2	1819	43	31	16	25	12
Verrucomicrobia	Verrucomicrobiae							
		Akkermansia muciniphila ATCC BAA-835	2138	27	18	9	8	6
	Opitutae							
		Opitutus terrae PB90-1	4612	165	46	14	25	10
		Methylacidiphilum infernorum V4	2472	33	11	9	6	4

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

Supplementary Table S2. Families belonging to the winged helix–turn–helix superfamily identified in this work

Abbreviation	Family name	Number of proteins	Proportion
GntR	GntR-like transcriptional regulators	4039	0.131465026
Lrp	Lrp/AsnC-like transcriptional regulator N-terminal domain	2548	0.082934609
Biotin	Biotin repressor-like	1020	0.033199883
IclR	Transcriptional regulator IclR, N-terminal domain	984	0.032028122
CAP	CAP C-terminal domain-like	1186	0.038603001
LysR	LysR-like transcriptional regulators	9736	0.316896136
MarR	MarR-like transcriptional regulators	3852	0.125378381
ArsR	ArsR-like transcriptional regulators	2080	0.067701722
FUR	FUR-like	967	0.031474791
LexA	LexA repressor, N-terminal DNA-binding domain	478	0.015558376
Rrf2	Transcriptional regulator Rrf2 (Pfam 02082)	449	0.014614458
HrcA	Heat-inducible transcription repressor HrcA, N-terminal domain	358	0.011652508
ArgR	Arginine repressor (ArgR), N-terminal DNA-binding domain	343	0.011164274
Iron	Iron-dependent repressor protein	423	0.013768187
Rex	Transcriptional repressor Rex, N-terminal domain	138	0.004491749
Peni	Penicillinase repressor	339	0.011034079
Mj22	DNA-binding protein Mj223	275	0.008950949
Z-DNA	Z-DNA-binding domain	329	0.01070859
AF2	Hypothetical protein AF2008	143	0.004654493
E-II	Transcription factor E/IIe-alpha, N-terminal domain	169	0.005500765
Heli	Helicase DNA-binding domain	70	0.002278423
ModE	N-terminal domain of molybdate-dependent transcriptional regulator ModE	210	0.00683527
F93	Hypothetical protein F93	193	0.006281939
RTP	Replication terminator protein (RTP)	46	0.00149725
Arch	Archaeal DNA-binding protein	32	0.001041565
PurR	N-terminal domain of Bacillus PurR	67	0.002180777
SCF	SCF ubiquitin ligase complex WHB domain	10	0.000325489
H1/H5	Linker histone H1/H5	7	0.000227842
DsvD	Dissimilatory sulfite reductase DsvD	7	0.000227842
SelB	C-terminal fragment of elongation factor SelB	64	0.00208313
MotA	Transcription factor MotA, activation domain	19	0.000618429
RPA32	C-terminal domain of RPA32	12	0.000390587
Plan	Plant O-methyltransferase, N-terminal domain	81	0.002636461
P4	P4 origin-binding domain-like	12	0.000390587
FokI	Restriction endonuclease FokI, N-terminal (recognition) domain	27	0.00087882
Ets	ets domain	1	3.25489E-05
Meth	Methionine aminopeptidase, insert domain	1	3.25489E-05
Vacu	Vacuolar sorting protein domain	7	0.000227842
Rio2	Rio2 serine protein kinase N-terminal domain	1	3.25489E-05

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

Supplementary Table S3. Partner domains (PaDos) identified and analysed in this work

Abbreviation	PaDo Name	No. of PaDos	Proportion
PBP II	Periplasmic binding protein-like II	9799	0.54276061
GAF	GAF domain-like	1503	0.08325025
cAMP	cAMP-binding domain-like	1169	0.06475019
Dimeric	Dimeric alpha+beta barrel	1029	0.05699568
PLP	PLP-dependent transferases	914	0.0506259
NagB	NagB/RpiA/CoA transferase-like	763	0.0422621
LexA	LexA/Signal peptidase	332	0.01838928
Arginine	C-terminal domain of arginine repressor	238	0.01318267
Actin	Actin-like ATPase domain	235	0.01301651
Trans-repress	C-terminal domain of transcriptional repressors	233	0.01290573
Iron	Iron-dependent repressor protein, dimerization domain	212	0.01174255
MOP	MOP-like	186	0.01030243
ClassII	Class II aaRS and biotin synthetases	173	0.00958236
NAD(P)	NAD(P)-binding Rossmann-fold domains	161	0.00891769
PTS	PTS-regulatory domain, PRD	144	0.00797607
CBS	CBS-domain	102	0.00564972
Phos/Ani t.p.	Phoshotransferase/anion transport protein	73	0.00404343
S-adenosyl	*S*-adenosyl-L-methionine-dependent methyltransferases	72	0.00398804
Acyl-CoA	Acyl-CoA *N*-acyltransferases (Nat)	71	0.00393265
PTS-sis	PTS system, Lactose/Cellobiose specific IIB subunit (Pfam 02302)	69	0.00382187
PRTase-like	PRTase-like	65	0.00360031
FadR-C-T	Fatty acid responsive transcription factor FadR, C-terminal domain	57	0.0031572
PBP I	Periplasmic binding protein-like I	56	0.00310181
TM1602 C-T	Putative transcriptional regulator TM1602, C-terminal domain	51	0.00282486
Rhodanese	Rhodanese/Cell cycle control phosphatase	37	0.00204941
Ribokinase-like	Ribokinase-like	36	0.00199402
P-loop	P-loop containing nucleoside triphosphate hydrolases	35	0.00193863
SIS domain	SIS domain	33	0.00182785
Phosphotyrosine	Phosphotyrosine protein phosphatases I	25	0.00138473
2-methylcitrate PrpD	2-methylcitrate dehydratase PrpD	1	5.5389E-05
4Fe-4S	4Fe-4S ferredoxins	3	0.00016617
Acetyl-CoA	Acetyl-CoA synthetase-like	2	0.00011078
Alpha/Beta knot	Alpha/Beta knot	1	5.5389E-05
Alpha/Beta-Hydrolases	Alpha/Beta-Hydrolases	1	5.5389E-05
Amidase signature	Amidase signature (AS) enzymes	1	5.5389E-05
ATPase domain of HSP90	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase	1	5.5389E-05
Bet v1-like	Bet v1-like	19	0.0010524
CheY-like	CheY-like	20	0.00110779
C-T bipartite	C-terminal effector domain of the bipartite response regulators	1	5.5389E-05
DNA primase core	DNA primase core	1	5.5389E-05
DNA-glycosylase	DNA-glycosylase	2	0.00011078

DsbC/DsbG N-T	DsbC/DsbG N-terminal domain-like	1	5.5389E-05
FMN-binding	FMN-binding split barrel	9	0.0004985
FMN-dependent	FMN-dependent nitroreductase-like	1	5.5389E-05
GIY-YIG	GIY-YIG endonuclease	2	0.00011078
GST	Glutathione S-transferase (GST), C-terminal domain	1	5.5389E-05
Gly-3-p d-like C-T	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain	1	5.5389E-05
Gly-3-p-(1)-at	Glycerol-3-phosphate (1)-acyltransferase	1	5.5389E-05
Glyo/B rp/D d	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase	1	5.5389E-05
Haem-dependent catalases	Haem-dependent catalases	1	5.5389E-05
Homeodomain-like	Homeodomain-like	8	0.00044312
HPr kN-T	HPr kinase/phoshatase HprK N-terminal domain	8	0.00044312
IIA-Man	IIA domain of mannose transporter, IIA-Man	2	0.00011078
JAB1/MPN domain	JAB1/MPN domain	1	5.5389E-05
L30e-like	L30e-like	1	5.5389E-05
LDH C-terminal domain-like	LDH C-terminal domain-like	1	5.5389E-05
MFS	MFS general substrate transporter	2	0.00011078
Nitrogenase accessory factor-like	Nitrogenase accessory factor-like	1	5.5389E-05
NTF2-like	NTF2-like	1	5.5389E-05
Nucleoside p/pl c d	Nucleoside phosphorylase/phosphoribosyltransferase catalytic domain	2	0.00011078
Nucleoside p/prt N-T	Nucleoside phosphorylase/phosphoribosyltransferase N-terminal domain	2	0.00011078
Nucleotide-d-s t	Nucleotide-diphospho-sugar transferases	5	0.00027695
Nucleotidyltransferase	Nucleotidyltransferase	5	0.00027695
Nudix	Nudix	1	5.5389E-05
PIN domain-like	PIN domain-like	1	5.5389E-05
Prim-pol domain	Prim-pol domain	1	5.5389E-05
PK-like	Protein kinase-like (PK-like)	3	0.00016617
PTS IIb component	PTS IIb component	1	5.5389E-05
P. lysine decarboxylase	Putative lysine decarboxylase	4	0.00022156
PAS domain	PYP-like sensor domain (PAS domain)	9	0.0004985
Radical SAM enzymes	Radical SAM enzymes	1	5.5389E-05
Sensory d two-component	Sensory domain of two-component sensor kinase	1	5.5389E-05
SirA-like	SirA-like	4	0.00022156
SCP	Sterol carrier protein, SCP	20	0.00110779
Tetracyclin r C-T	Tetracyclin repressor-like, C-terminal domain	2	0.00011078
Thio/Thiol	Thioesterase/Thiol ester dehydrase-isomerase	16	0.00088623
Thioredoxin-like	Thioredoxin-like	1	5.5389E-05
Urocanase	Urocanase (Pfam 01175)	1	5.5389E-05
Zinc beta-ribbon	Zinc beta-ribbon	5	0.00027695

Rivera-Gómez, N., Segovia, L. & Perez-Rueda, E. (2011). Diversity and distribution of transcription factors: their partner domains play an important role in regulatory plasticity in bacteria. *Microbiology* **157**, 2308–2318.

14. REFERENCIAS

1. Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K., & Watson, J. (1994). *Molecular biology of the cell.* CSHL, E.U: Garland Science.
2. Apic, G., Gough, J., & Teichmann, S. (2001). An insight into domain combination. *Bioinformatics*, S83-S89.
3. Apic, G., Gough, J., & Teichmann, S. (2001). Domain combinations in Archeal, Eubacterial and Eukaryotic proteomes. *310*, 311-325.
4. Aravind, L., Anantharaman, V., Balaji, S., Babu, M., & Iyer, L. (2005). The many faces of the heliz-turn-helix domain: transcription regulation and beyond. *29*, 231-262.
5. Babu, M., & Teichmann, S. (2003). Evolution of transription factors and the gene regulatory network in Escherichia coli. *31*(4), 1234-1244.
6. Baumbach, J., & Apeltsin, L. (2008). Linking Cytoscape and the corynebacteria reference database CoryneRegNet. *9*(184), 2164-2169.
7. Beker, C. H., Tomlinson, S. R., García, A. E., & Harman, J. G. (2001). Amino acid substitution at position 99 affects the rate of CRP subunit Exchange. *40*, 12329-12338.
8. Bernard, A., & Busby, S. (2004). Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *7*, 102-108.
9. Borukhov, S., & Severinov, K. (2002). Role of the RNA polymerase sigma subunit in transcription initiation. *153*, 557-562.
10. Braun, V., Mahren, S., & Sauter, A. (2005). Gene regulation by transmembrane signaling. *19*, 103-113.
11. Brennan, R., & Matthews, B. (1989). Structural basis of DNA-protein recognition. *14*(7), 286-290.
12. Brinkrolf, K., Brune, I., & Tauch, A. (2007). The transcriptional regulatory network of the amino acid procer Corynebacterium glutamicum. *219*(2), 191-211.
13. Browning, D. F., & Busby, S. J. (2004). The regulation of bacterial Transcription initiation. *Nature*, 1-9.
14. Busby, S., & Savery, N. (2002). Transcription Activation at bacterial promoters. *Nature*, 1-7.
15. Chothia, C., & Gough, J. (2009). Genomic and structural aspects of protein evolution. *419*, 18-28.
16. Collado-Vides, J., Pérez-Rueda, E., & Segovia, L. (2004). Phylogenetic distribution of DNA-binding trancription factors in bacteria and archea. *28*, 341-350.
17. Crick, F. (1970). Central dogma of molecular biology. *227*, 561-563.
18. Doi, R. H., & Wang, L. (1986). Multiple procaryotic ribonuclei acid polymerase sigma factors. *50*(3), 227-243.
19. Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., y

otros. (2009). The Pfam protein families database. *38*, D211-D222.

20. gama, c. (2008). *http://regulondb.ccg.unam.mx/*.
21. Gama-Castro, S., Jiménez-Jacinto, V., Peralta-Gil, M., Santos-Zabaleta, A., Contreras-Moreira, Contreras-Moreira, B., y otros. (2008). RegulonDB (version 6.0): gene regulation model of Escheichia coli K-12 beyond transcription, active (experimental) annotated promoters and texpresso navegation. *36*, D120-D124.
22. Gough, J., Karplus, K., Hughey, R., & Chothia, C. (2001). Assignment of hology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure. *313*(4), 903-919.
23. Gourse, R. L., Ross, W., & Gaal, T. (2000). UPs and Downs in bacterial transcriptiona initiation: the role of the alpha subunit of RNA polymerase in promoter recognition. *37*(4), 687-695.
24. Haldenwang, W. G. (1995). The sigma factors of bacillus subtilis. *59*(1), 1-30.
25. Hengge , R. (2009). Proteolysis of sigmaS (RpoS) and the geneal stress response in EScherichia coli. *160*(9), 667-676.
26. Hinton, D. M., Pande, S., Wais, N., Johnson, X. B., Vuthoori, M., Makela, A., y otros. (2005). Transcriptional takeover by sigma appropriation: remodelling of the sigma70 subunit of Escherichia coli RNA polymerase by the bacteriophage T4 activator MotA and co-activator AsiA. *151*, 1729-1740.
27. Hsu, L. M. (2002). Promoter clearence and escape in prokaryotes. *1577*, 191-207.
28. Huffman, J., & Brennan, R. (2002). Prokaryotic transcription regulators: more than just the helix-turn-motif. *12*, 98-106.
29. Janga, S., & Collado-Vides, J. (2007). Structure and evolution of gene regulatory networks in microbial genomes. *Research in Microbiology* , 787-794.
30. Kumar, A., Malloch, R. A., Fujita, N., Smillie, D. A., Ishihama, A., & Hayward, R. S. (1993). The minus 35-recognition region of eschrichia coli sigma 70 is inessetial for initiation of transcription at an "extended minus 10" promoter. *232*, 406-418.
31. Lewin, B. (2001). *Genes VII.* Madrid, España: Marbán.
32. Marmorstein, R., & Fitzgerald, M. X. (2003). Modulation of DNA.-binding domains for sequece-specific DNA recognition. *Gene*, 1-12.
33. Martínez-Antonio, A., & Collado-Vides, J. (2003). Identifying blobal regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology*, 482-489.
34. Martínez-Antonio, A., Janga, S., Saldado, H., & Collado-Vides, J. (2006). Internal-sensing machinary directs the activity of the regulatory network in Eschericia coli. *TRENDS in Mircrobiology*, 22-27.
35. Moore, A., Björklund, A., Ekman, D., Bornger-Bauer, E., & Elofsson, A. (2008). Arrengements in the modular evolution of proteins. *33*(9), 444-451.
36. Moreno-Campuzano, S., Janga, S. C., & Pérez-Rueda, E. (2006). Identification and analysis of DNA-binding transcriptional factors in Bacillus

subitlis and other firmicutes- a genome approach. *7*(147).

37. Murzin, A., Brenner, S., Hubbard, T., & Chothia, C. (1995). SCOP: A structural classification of proteins datavase for the investigation of sequences and estrucures. *247*, 536-540.

38. Orengo, C., & Thorton, J. (s.f.). Protein Families and their evolution -a setructural prespective. *74*, 867-900.

39. Orengo, C., Michie, A., Jones, S., Jones, D., Swindells, M., & Thornton, J. (1997). CHAT: a hierarchic classification of protein structure. *5*, 1093-1098.

40. Paget, M., & Helmann, J. (2003). The sigma70 family of sigma factors. *4*.

41. Paul, B., Ross, W., Gaal, T., & Gourse, R. L. (2004). rRNA transcription in Escherichia coli. *38*, 749-770.

42. Pérez-Rueda, E., & Collado-Vides, J. (2000). The repertorie of DNA-binding transcriptional regulatores in Escherichi coli K-12. *Nucleic Acid Research*, 1838-1847.

43. Pérez-Rueda, E., & Collado-Vides, J. (2001). Common history at the origin of the position–function correlation in transcriptional regulators in archaea and bacteria. *Journal of Molecular Evolucion* , 172-179.

44. Ranea, J., Buchan, D., Thornton, J., & Orengo, C. (2004). Evolution of protein superfamilies and bacterial genome size. *336*, 871-887.

45. Rigali, S., Derouaux, A., Giannotta, F., & Dusart, J. (2002). Subdivision of the helix-turn-helix GntR family of bacterial regulators in the FadR, HutC, MocR, and YtrA subfamilies. *227*(15), 12507-12515.

46. Rigali, S., Schlicht, M., Hoskisson, P., Nothaft, H., Merzbacher, M., Joris, B., y otros. (2004). Extending the classification of bacterial transcriptional factors beyon the helix-turn-helix motif as an alternative approach to discover new cis/trans relationships. *32*(11), 3418-3426.

47. Rodriguez, S., Provvedi, R., Jacques, P., Gaudreau, L., & Manganelli, R. (2006). The sigma factors o Mycobacterium tuberculosis. *Federation of European Microbiological Societies*, 926-941.

48. Rosinski, J., & Atchley, W. (1999). Molecular evolution of helix-turn-helix proteins. *49*, 301-309.

49. Saeed, A. I., Sharov, V., White, J., Liang, W., Bhagabati, N., Braisted, J., y otros. (2003). TM4: a free, open source system for microarray data management and analysis. *34*(2), 374-378.

50. Salgado, H., Matínez-Antonio, A., & Janga, S. (2007). Conservation of transcriptional sensing systems in prokaryotes: A perspective fromo Escherichia coli. *581*, 3499-3506.

51. schell, M. A. (1993). Molecular biology of the LysR family of transcriptional regulators. *47*, 597-626.

52. Schreiter, E., & Drennan, C. (2007). Ribbon-helix-helix transcription factors: variations on a theme. *5*, 710-720.

53. Sierro, N., Makita, Y., de Hoon, M. J., & Nakai, K. (2008). DBTBS: a database of transcriptional regulation in Bacillus subtilis containing upstream intergenic consevation information. *36*, D93-D96.

54. Steitz, T. A. (1990). Structural studies od protin-nucleic acid interaction:the sources of sequence-specific binding. *Quarterly reviews*, 205-280.

55. Thorton, J., Orengo, C., Todd, A., & Pearl, F. (1999). Protein folds, function and evolution. *293*, 333-342.

56. Vogel, C., Bashton, M., Kerrison, N., Chothia, C., & Teichmann, S. (2004). Structure, funcion and evolution of multidomain proteins. *14*, 208-216.

57. Wang, M., & Caetano-Anollés, G. (2009). The evolutionay mechanics of domain organizationin proteomes and the rise of modularity in the protein world. *17*, 66-78.

58. Wilson, D., Charoensawan, V., Kummerfeld, S., & Teichmann, S. (2008). DBD-taxonomically broad transcription factor predictions: new content and functionality. *36*, D88-D92.

59. Wintjens, R., & Rooman, M. (1996). Structural Classification of HTH DNA-binding domains and protein-DNA interaction modes. *Journal Molecular Biology, 262*, 294-313.

60. Zhou, D., & Yang, R. (2006). Global analysis of gene transcription regulation in prokaryotes. *Cellular and Molecular Life Sciences*, 2260-2290.