PROVALIDATOR: Una herramienta para diseño y validación de *profiles*

Memoria que presenta el Licenciado en Ingeniería Electrónica Antonio Jesús Molina Henares
para aspirar al título de Doctor.

Fdo. Antonio Jesús Molina Henares

Vº Bº El Director
Juan Luis Ramos Martín
Doctor en Biología
Profesor de Investigación del CSIC

EEZ-CSIC / UGR
2010

Editor: Editorial de la Universidad de Granada
Autor: Antonio Jesús Molina Henares
D.L.: GR 3595-2010
ISBN: 978-84-693-4308-1

Esta Tesis Doctoral se ha realizado en el
Departamento de Protección Ambiental de la
Estación Experimental del Zaidín (CSIC) en Granada

Dave: Abre las compuertas del muelle de salida, Hal.
HAL: Lo siento, Dave, me temo que no puedo hacerlo.
Dave: ¿Cuál es el problema?
HAL: Creo que lo sabes tan bien como yo.
Dave: No sé de qué estás hablando.
HAL: Sé perfectamente que tú y Frank estabais
pensando en desconectarme y me temo que esto
es algo que no puedo permitir que suceda.

“2001: una odisea del espacio”
Stanley Kubrick.

Si quieres que parezca que tu vida ha durado más,
lo que tienes que hacer es perseguir cosas nuevas,
necesitas probar cosas nuevas todo el tiempo,
conducir por un camino distinto cuando vuelves a casa…
Si siempre almacenas recuerdos nuevos,
parecerá que tu vida ha sido más larga.

David Eagleman.

Índices

Índice de figuras 13
Índice de tablas 15

Introducción
Pseudomonas putida 19
Reguladores transcripcionales 21
Estado del arte 23
BacTregulators: una base de datos de reguladores transcripcionales en bacterias y arqueas 28

Herramientas
Bases de datos utilizadas 39
Provalidator: una herramienta para diseño y validación de *profiles* 42

Objetivos 49

Resultados[1]
Capítulo 1: The TetR Family of Transcriptional Repressors (Microbiology and Molecular Biology Reviews) 55
Capítulo 2.1: Prefacio: Definición de la familia de factores de transcripción IclR 89
Capítulo 2.2: The IclR family of transcriptional activators and repressors can be defined by a single profile (Protein Science) 97
Capítulo 2.3: Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors (Federation of European Microbiological Societies) 107
Capítulo 3: A general profile for the MerR family of transcriptional regulators constructed using the semi-automated Provalidator tool (Environmental Microbiology Reports) 139
Capítulo 4: Characterization of the RND family of multidrug efflux pumps: *in silico* to *in vivo* confirmation of four functionally distinct subgroups 157
Capítulo 5: BacTregulators: a database of transcriptional regulators in bacteria and archaea (Bioinformatics) 173

Discusión
BacTregulators 181
Provalidator 182

Conclusiones 193

Direcciones de sitios web utilizados 197

Bibliografía 201

[1] El material suplementario de cada artículo (si lo tuviera) está al final del capítulo.

Índice de figuras

Figura 1: Ejemplo de patrón ... 23
Figura 2: Ejemplo de *profile* ... 26
Figura 3: Ejemplo de HMM ... 27
Figura 4: Esquema de integración de secuencias de BacTregulators ... 28
Figura 5: Definición de nuevo registro en BacTregulators ... 29
Figura 6: Criterios de almacenamiento de un registro en BacTregulators ... 30
Figura 7: Estructura de la base de datos BacTregulators ... 31
Figura 8: Descripción de informe de resultados en BacTregulators ... 32
Figura 9: Formulario de filtrado en BacTregulators ... 34
Figura 10: Informe de integración de conocimiento en BacTregulators ... 35
Figura 11: Resultado de validación de la familia IclR ... 45
Figura 12: Figura de mérito para validar un *profile* ... 90
Figura 13: Comparación de alineamiento de aminoácidos y estructura secundaria ... 92
Figura 14: Distribución en genomas de la familia IclR ... 93
Figura 15: Diagrama de flujo del método de construcción de un *profile* descrito por Bucher ... 182
Figura 16: Diagrama de flujo de nuestro método de construcción de un *profile* ... 183
Figura 17: Crear sesión en Provalidator ... 184
Figura 18: Recuperación de secuencias en Expasy ... 184
Figura 19: Clusterizar secuencias para formar el conjunto semilla ... 185
Figura 20: Resultado de la clusterización ... 185
Figura 21: Construcción del *profile* ... 186
Figura 22: Modos de construcción de un *profile* ... 187

Índice de tablas

Tabla 1: Familias de reguladores de la transcripción en procariotas 22
Tabla 2: Expresiones regulares usadas 24
Tabla 3: Resultados de distintos *profiles* de IclR 89

Introducción

Pseudomonas putida

Las bacterias del género *Pseudomonas* están ampliamente distribuidas en el medio ambiente, ya que es posible encontrarlas en todo tipo de suelos y aguas de distintas latitudes. Las *Pseudomonas* son bacilos Gram negativos, rectos o ligeramente curvados (0.5-0.8 µm x 1-3 µm), que se encuadran dentro del grupo γ de las proteobacterias, quimiorganotróficos aerobios y generalmente son móviles gracias a los flagelos polares que poseen, aunque en algunas cepas se han descrito flagelos laterales. Su peculiar capacidad de crecimiento en medios de cultivo muy simples ha facilitado, principalmente en las últimas cuatro décadas, numerosos y profundos estudios de este amplio grupo taxonómico, tanto en el ámbito metabólico como en el fisiológico y el genético.

Gracias a la versatilidad metabólica de estos microorganismos y a la capacidad de formar biopelículas, no es de extrañar que se hayan aislado bacterias de este género tanto en suelos limpios como en suelos contaminados por productos biogénicos y xenobióticos. La ubicuidad de las bacterias del género *Pseudomonas* y su capacidad para explotar una amplia variedad de nutrientes refleja un sistema de adaptación al medio ambiente que no encuentra parangón en las bacterias de otros géneros. Por ello, se consideran las bacterias del género *Pseudomonas* un paradigma de versatilidad metabólica y microorganismos clave en el reciclado de materia orgánica en los compartimentos aerobios de los ecosistemas, jugando un papel esencial en la mejora y el mantenimiento de la calidad medioambiental.

Pseudomonas putida es una de las especies de mayor interés industrial entre las bacterias del género *Pseudomonas*, ya que unido a su potencial de degradación de compuestos aromáticos y xenobióticos, presenta la capacidad de colonizar el sistema radicular de plantas, formar biopelículas y ser manejable desde el punto de vista genético. Desde mediados de los años 80, distintas especies del género *Pseudomonas* están siendo aplicadas a semillas de cereales y a suelos para prevenir el crecimiento o colonización de patógenos, un proceso que se ha denominado biocontrol (Berg et al., 2001, Renault *et al*., 2007, Keane *et al*., 2008), y que representa un incremento de hasta un 25% en el rendimiento de las cosechas. Una de las cepas mejor estudiadas en este sentido es *P. putida* KT2440, derivada de la cepa *P. putida* mt-2 aislada originalmente en Japón (Nakazawa, 2002). Esta última porta el plásmido TOL, que codifica una de las rutas de degradación de tolueno y xilenos mejor caracterizados en el ámbito de la biodegradación. En 1981 el comité de ADN recombinante de los Institutos de la Salud de EEUU certificó que la cepa *P. putida* KT2440 curada del plásmido TOL no es patógena para animales o plantas, y se puede considerar como un hospedador seguro en la clonación de genes de bacterias del suelo.

El genoma completo de *P. putida* KT2440 se encuentra disponible en bases de datos de libre acceso, lo que constituye una herramienta muy valiosa para el análisis funcional de la información genética del microorganismo (Nelson *et al*., 2002). La anotación inicial del cromosoma de la cepa *P. putida* KT2440 sugería 5.420 ORFs, dentro de los cuales 80 genes aparecen truncados o alterados como consecuencia de la inserción de transposones o elementos móviles. Por otro lado, se han identificado 804 copias de un elemento específico de 35 pb conocido como REP (del inglés *Repetitive Extragenic Palindromic sequences*) (Aranda-Olmedo *et al*., 2002). El genoma de KT2440 codifica 184 proteínas relacionadas con elementos móviles, 82 genes que codifican transposasas, 8 intrones de grupo II, etc. De los 5.420 ORFs predichos, 600 aparecen anotados como proteínas hipotéticas de función desconocida. Es posible que algunos de estos casos puedan tratarse de artefactos producto de los algoritmos utilizados en la predicción de genes (dos Santos *et al*., 2004).

Una de las líneas de investigación más importantes de los últimos años llevada a cabo en especies de *Pseudomonas putida* ha sido el estudio de la regulación transcripcional de las rutas catabólicas de hidrocarburos aromáticos. Aunque en el siglo pasado se liberaron grandes cantidades de hidrocarburos aromáticos como consecuencia de la actividad industrial, la mayoría de estos compuestos se encuentran en el medio ambiente como consecuencia de la pirólisis natural de material orgánico (Dagley, 1971).

Uno de los objetivos del grupo experimental de Degradación de Tóxicos Orgánicos es contribuir a la lucha contra la contaminación medioambiental con la ayuda de herramientas biológicas. Los compuestos a combatir son hidrocarburos aromáticos (benceno, tolueno, etilbenceno y xilenos), nitroaromáticos como el TNT, haloaromáticos como los PCBs y pesticidas como el lindano y el DDT. El enfoque experimental incluye técnicas moleculares y el análisis bioquímico y genético de las rutas catabólicas, poniendo especial énfasis en las interacciones ADN/regulador, en experimentos en reactores y en la llamada rizorremediación de suelos. El grupo se ha centrado también en el estudio de las interacciones entre plantas y microorganismos en la rizosfera con el objetivo de explotar los sistemas de expresión génica en bacterias para el desarrollo de sistemas de biocontrol, biodegradación y contención biológica de microorganismos recombinantes.

Reguladores transcripcionales

Las bacterias en el medio ambiente están expuestas a variaciones de temperaturas, nutrientes, disponibilidad de agua y a la presencia de moléculas tóxicas originadas por los elementos bióticos o abióticos de su entorno (incluyendo moléculas deletéreas originadas por sus propios metabolismos). Estos cambios pueden hacerlas vivir en condiciones lejos de las óptimas. Sobrevivir en estos medios cambiantes requiere un amplio espectro de respuestas adaptativas rápidas, las cuales suelen estar controladas mediante proteínas reguladoras. Los reguladores transcripcionales pueden responder tanto a señales específicas ambientales como celulares y modular la transcripción, transducción o algunos otros eventos en la expresión genética, de manera que las respuestas fisiológicas se modifican apropiadamente (Beinlich *et al*., 2001).

Los reguladores de la transcripción de procariotas se clasifican en familias atendiendo a similitudes de secuencia, estructurales y criterios funcionales (Busenlehner et al., 2003; Ramos *et al*.,1990). La Tabla 1 incluye el listado de las familias más importantes de reguladores microbianos de la transcripción, el tipo de unión a ADN que exhiben, si sus miembros son preferentemente activadores o represores o si muestran una acción alternativa.

Los reguladores de la transcripción suelen ser proteínas de unión a ADN que presentan diferentes tipos de organización de dominios. La mayoría de los reguladores microbianos implicados en la transcripción son proteínas con dos dominios en un solo polipéptido, un dominio sensor de señal y un dominio de unión a ADN que transduce la señal. Los sistemas de un solo componente, en los que el dominio regulador de la transcripción y el dominio sensor están unidos en una sola proteína, son con diferencia los más abundantes en procariotas.

Se puede decir que el proceso de la regulación transcripcional es sumamente complicado, no sólo porque la actividad y ciclo catalítico de la ARN polimerasa es muy complejo, sino porque multitud de proteínas interfieren con la transcripción, estimulándola o reprimiéndola, pero en definitiva alterando el ciclo transcripcional. El interés del grupo por el conocimiento global de estas cuestiones se plasmó en la construcción de la base de datos AraC/XylS (Tobes y Ramos, 2000), que reunía el conocimiento disponible acerca de los reguladores pertenecientes a esta familia de activadores de la transcripción. Esto fue el inicio de una serie de investigaciones y desarrollos biotecnológicos encaminados a profundizar y mejorar la anotación del genoma de KT2440 y establecer bases para la investigación de proteínas de función desconocida.

Familia	Acción	Funciones reguladas	DBD "motif"	Posición	Referencias
LysR	Activador/Represor	Metabolismo del carbono y del nitrógeno	HTH	N-terminal	Henikoff *et al.*, 1988; Schell, 1993
AraC/XylS	Activador	Metabolismo nitrogenado, respuesta a estrés y patogénesis	HTH	C-terminal	Gallegos *et al.*, 1997; Tobes and Ramos, 2002
TetR	Represor	Biosíntesis de antibióticos, bombas de eflujo, estrés osmótico, etc.	HTH	N-terminal	Aramaki *et al.*, 1993; Aramaki *et al.*, 1995a; Aramaki *et al.*, 1995b
LuxR	Activador	"Quorum sensing", biosíntesis y metabolismo, etc.	HTH	C-terminal	Fuqua *et al.*, 1994; Pao *et al.*, 1994; Redfield, 2002
LacI	Represor	Utilización de fuente de carbono	HTH	N-terminal	Chandler, 1992, Weickert and Adhya, 1992
ArsR	Represor	Resistencia a metales	HTH	Central	Busenlehner *et al.*, 2003; Wu and Rosen, 1991
IclR	Represor/Activador	Metabolismo del carbono, bombas de eflujo	HTH	N-terminal	Neal and Chater, 1987; Reizer *et al.*, 1996; Reverchon *et al.*, 1990; Molina-Henares *et al.*, 2006
MerR	Represor	Resistencia y detoxificación	HTH	N-terminal	Helmann *et al.*, 1990; Summers, 1992
AsnC	Activador/Represor	Biosíntesis de aminoácidos	HTH	N-terminal	Friedberg *et al.*, 2001
MarR	Activador/Represor	Resistencia múltiple a antibióticos	HTH	Central	Alekshun and Levy, 1999; Ariza *et al.*, 1994; Seoane and Levy, 1995; Sulavik *et al.*, 1995
NtrC(EBP)	Activador	Asimilación de nitrógeno, síntesis de aminoácidos aromáticos, flagelos, rutas catabólicas, respuesta a fagos, etc.	HTH	C-terminal	Kustu *et al.*, 1991; Morett and Segovia, 1993
OmpR	Activador	Metales pesados y virulencia ("Response regulator of a two-component system")	Winged helix	C-terminal	Martínez-Hackert and Stock, 1997
DeoR	Represor	Metabolismo glucídico	HTH	N-terminal	Ramos *et al.*, 2002; van Rooijen and de Vos, 1990
Cold shock	Activador	Resistencia a bajas temperaturas	Dominio de unión a ARN	Variable	Brandi *et al.*, 1994; La Teana *et al.*, 1991; Schindelin *et al.*, 1993
GntR	Represor	Metabolismo general	HTH	N-terminal	Haydon and Guest, 1991; Reizer *et al.*, 1991; Rigali *et al.*, 2002
Crp	Activador/Represor	Respuesta global, represión catabólica y anaerobiosis	HTH	C-terminal	Chandler, 1992; Gambino *et al.*, 1993; Ma *et al.*, 1995

Tabla 1: Familias de reguladores de la transcripción en procariotas.

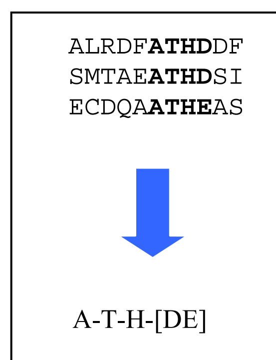
Estado del arte

La biología de los organismos vivos se basa en un conjunto de dominios funcionales "barajados" que dan lugar a enzimas, proteínas estructurales, reguladores transcripcionales, maquinaria de replicación de ADN, síntesis de ARN y otros procesos. Los dominios suelen ser regiones de tamaño variable que, debido a la conservación de secuencia, suelen presentar estructuras 3D comunes y a menudo se encuentran asociados a una función, por ej., sitios de unión a ADN, centros activos de enzimas, unión de ATP o cofactores, secuencias señal para paso a través de la membrana, etc. No obstante, estas regiones, *motivos*, se caracterizan porque su conservación no es perfecta, por lo que en algunos casos no son fácilmente detectables mediante técnicas de homología de secuencia como BLAST (Altschul *et al.*, 1997). Los motivos se conservan incluso a grandes distancias evolutivas debido a restricciones estructurales o funcionales. Luego si estos motivos están relacionados con la función, las herramientas de búsqueda de motivos nos permitirán predecir la función de las proteínas detectadas cuando ésta sea desconocida. Y, en virtud de la conservación que se produce en la secuencia primaria, también nos permitirán la detección de homólogos remotos.

Se dispone de tres herramientas principales para la detección de motivos en proteínas: *patterns* (patrones o *signatures*), *profiles* (o perfiles) y Modelos Ocultos de Markov (HMM). Los *patterns* son expresiones regulares que describen cada posición en un alineamiento múltiple. Tienen la desventaja de que no almacenan el peso de cada aminoácido en cada posición. Los *profiles* resuelven este problema al ponderar cada residuo en cada posición del alineamiento. La tercera categoría que nos encontramos son los *Hidden Markov Models*, Modelos Ocultos de Markov o HMM, en los que un modelo lineal de cadenas ocultas de Markov se corresponde con una secuencia de nodos para cada posición en un alineamiento múltiple. A continuación, se describen de manera somera los tres métodos.

Patterns:

Los *patterns* son expresiones regulares que se infieren a partir de un alineamiento múltiple de secuencias donde podemos localizar una región específica relacionada con una determinada función, como puede ser una región de unión a ADN, interacciones con membranas o un sitio activo en una enzima.



En este caso, el patrón está definido por tan pocas posiciones que seguramente obtendremos un número alto de falsos positivos (secuencias con dicho patrón que no presentan la misma función)

Figura 1: Ejemplo de patrón.

Los patrones regulares deben seguir un número de reglas previamente convenidas, como usar el código estándar de una letra para cada aminoácido, las distintas posiciones en el patrón se separan mediante guiones, o que un punto indica el final del patrón. Además, hay algunas otras reglas como las que se muestran en la siguiente tabla. Como se verá, no hay apenas diferencia con la sintaxis seguida por los motores de expresiones regulares que se pueden encontrar en cualquier editor de textos. De hecho, se podría usar uno de éstos para detectar el patrón que buscamos:

Expresiones regulares usadas	
x	Cualquier aminoácido
[A,B...]	Ambigüedad: A o B...
{A,B...}	Cualquiera menos A o B...
A(2,4)	Repetición: A-A o A-A-A o A-A-A-A
<, >	N-terminal, C-terminal

Ejemplo: [AC]-x-V-x(4)-{E,D}.
Representa: [Ala o Cys]-cq-Val-cq-cq-cq-cq-{cq menos Glu o Asp}
(cq: cualquiera)

Tabla 2: Expresiones regulares usadas.

Utilizando este código es posible extraer los patrones regulares para un alto número de sitios activos y regiones funcionales de los alineamientos de secuencias. PROSITE (Sigrist *et al.*, 2002) es la base de datos donde se encuentran almacenados todos estos patrones regulares o motivos recopilados a lo largo del tiempo. El principal criterio que debe satisfacer un patrón es el de tener la definición mínima necesaria para encontrar motivos de secuencia con la mayor eficiencia deseable, o sea, que encuentre el menor número de falsos positivos posible. En pocas palabras podríamos resumir que el patrón debe presentar alta sensibilidad y alta especificidad, conceptos sobre los que volveremos más adelante.

La creación de nuevos patrones se realiza a partir del estudio de revisiones de un grupo o familia de proteínas. De esta forma, las secuencias integradas dentro de un determinado grupo se alinean y este alineamiento se asocia al conocimiento funcional que existe sobre las mismas. Por tanto, para definir el patrón de nuevos motivos es indispensable la publicación de nuevas secuencias funcionalmente bien caracterizadas y en relación con el grupo funcional en estudio. De esta forma, es posible crear tablas de secuencias funcionalmente relacionadas que, una vez alineadas, nos permiten reconocer las regiones o residuos biológicamente importantes en el desempeño de una determinada función.

Profiles:

Los *profiles* o perfiles son matrices ponderadas sobre alineamientos múltiples de secuencias. Para la obtención de un *profile* es necesario disponer de un alineamiento múltiple de secuencias. El *profile* está definido por una matriz AxP, donde "A" corresponde al listado de los 20 aminoácidos, y "P" a las distintas posiciones del alineamiento de secuencias. De esta forma, se pondera la ocurrencia de cada uno de los 20 residuos en cada posición del alineamiento.

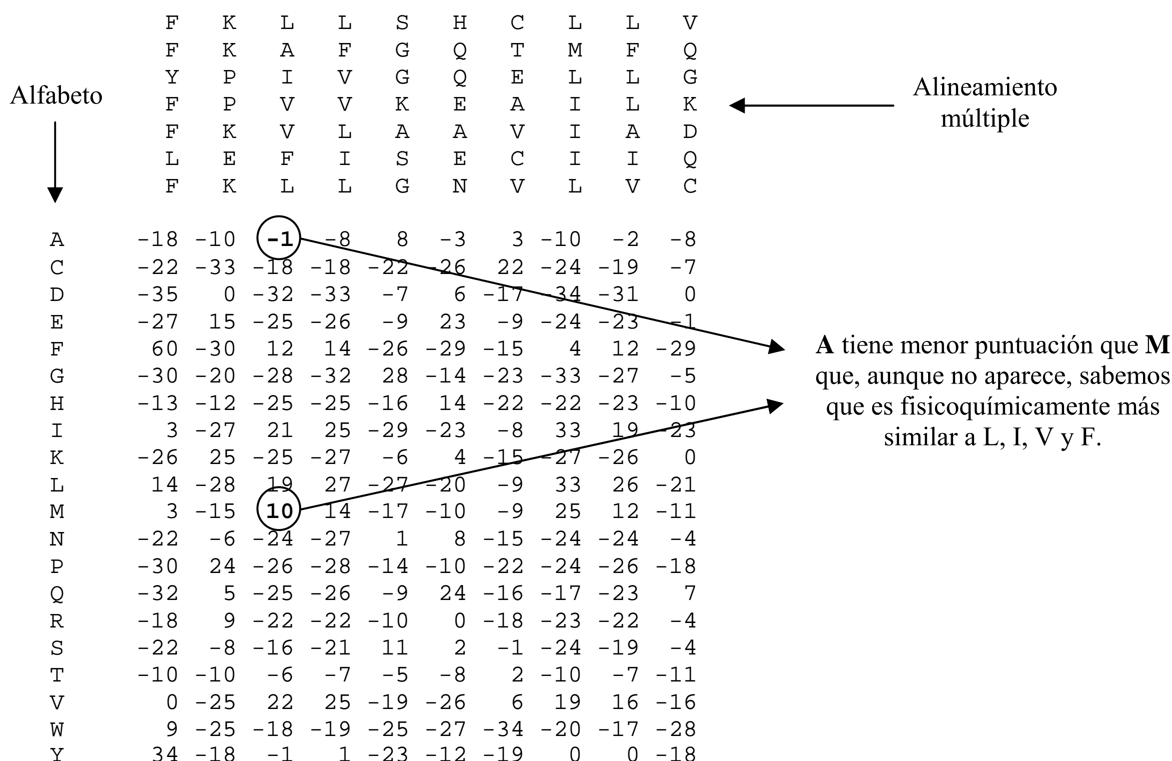
Los *profiles* se pueden construir mediante la aplicación de diferentes técnicas. El método más clásico es el descrito por Gribskov (Gribskov *et al.*, 1987; Luthy *et al.*, 1994). Dicho método requiere un alineamiento múltiple de secuencias como entrada además de una matriz de sustitución de aminoácidos (*weight matrix* o matriz de pesos) para convertir las frecuencias de los residuos en una posición en pesos (Henikoff and Henikoff, 1992). PROSITE, además de las expresiones regulares, incluye *profiles* diseñados con dicha metodología.

Una vez que disponemos del *profile* de un alineamiento, es posible iniciar búsquedas de secuencias que encajen en dicho *profile*. En este caso podemos ponderar de distinta forma un residuo en una determinada posición de la secuencia problema en función de su representatividad o valor medio en el alineamiento. Esta posibilidad incrementa la sensibilidad de la búsqueda de regiones homólogas en secuencias remotas.

Una característica que diferencia los *profiles* de las expresiones regulares es que no sólo se circunscriben a pequeñas regiones con un alto índice de similitud, sino que presentan una mayor utilidad a la hora de definir regiones o dominios más extensos que puedan caracterizar a familias de proteínas más que motivos. Debido a que el *profile* puede cubrir tanto regiones conservadas como variables del alineamiento, es menos sensible a los errores de alineamiento, pudiendo establecer una correlación general de similitud significativa con otras secuencias distantes incluso cuando ciertas partes del perfil correspondan a regiones mal alineadas. El requisito fundamentales a la hora de establecer *profiles* de calidad es que, además de presentar índices de similitud altamente significativos para los motivos que detecta y bajos para los falsos positivos, debe presentar correctamente alineados aquellos residuos con análogas funciones o propiedades estructurales de acuerdo con los datos experimentales.

Por definición, los *profiles* suelen tener una mayor sensibilidad de búsqueda y ser más robustos que los *patterns*. Esto se debe a que los *profiles* integran mayor y más completa información sobre las características conservadas y variables del dominio o región de las secuencias que definen.

A continuación se muestra el *profile* ponderado de un alineamiento sin huecos basado en el método estándar arriba mencionado:



	F	K	L	L	S	H	C	L	L	V
	F	K	A	F	G	Q	T	M	F	Q
	Y	P	I	V	G	Q	E	L	L	G
	F	P	V	V	K	E	A	I	L	K
	F	K	V	L	A	A	V	I	A	D
	L	E	F	I	S	E	C	I	I	Q
	F	K	L	L	G	N	V	L	V	C
A	-18	-10	**-1**	-8	8	-3	3	-10	-2	-8
C	-22	-33	-18	-18	-22	-26	22	-24	-19	-7
D	-35	0	-32	-33	-7	6	-17	-34	-31	0
E	-27	15	-25	-26	-9	23	-9	-24	-23	-1
F	60	-30	12	14	-26	-29	-15	4	12	-29
G	-30	-20	-28	-32	28	-14	-23	-33	-27	-5
H	-13	-12	-25	-25	-16	14	-22	-22	-23	-10
I	3	-27	21	25	-29	-23	-8	33	19	-23
K	-26	25	-25	-27	-6	4	-15	-27	-26	0
L	14	-28	19	27	-27	-20	-9	33	26	-21
M	3	-15	**10**	14	-17	-10	-9	25	12	-11
N	-22	-6	-24	-27	1	8	-15	-24	-24	-4
P	-30	24	-26	-28	-14	-10	-22	-24	-26	-18
Q	-32	5	-25	-26	-9	24	-16	-17	-23	7
R	-18	9	-22	-22	-10	0	-18	-23	-22	-4
S	-22	-8	-16	-21	11	2	-1	-24	-19	-4
T	-10	-10	-6	-7	-5	-8	2	-10	-7	-11
V	0	-25	22	25	-19	-26	6	19	16	-16
W	9	-25	-18	-19	-25	-27	-34	-20	-17	-28
Y	34	-18	-1	1	-23	-12	-19	0	0	-18

Figura 2: Ejemplo de *profile*.

La matriz usa pesos discriminatorios no sólo en los aminoácidos que aparecen, sino también en los que no aparecen, para lo cual se basa en la matriz de pesos que representa similitudes entre aminoácidos.

Hay que señalar que para ciertas posiciones, un residuo que no aparece en el alineamiento puede recibir mayores puntuaciones que otros que sí aparecen. Éste es el resultado de ponderar el resto de residuos que sí aparecen en dicha posición. Por ejemplo, el aminoácido Ala, aunque está presente en la tercera posición del alineamiento, presenta una puntuación menor (-1) que la Met (+10). Esto es debido a que los residuos más representativos en dicha columna Leu, Ile, Val y Phe están físico-químicamente más relacionados con Met que con Ala. Este procedimiento también se extrapola en el caso de las inserciones y deleciones, ponderando no sólo la posición, sino también el contexto en el alineamiento.

Modelos ocultos de Markov:

Un modelo oculto de Markov se puede ver como una máquina de estados finita de primer orden, esto es, cada estado sólo depende del estado anterior. Una máquina de estados finita se mueve a través de una serie de estados y produce una determinada salida cuando se alcanza un estado o cuando se pasa de uno a otro. Cada estado tiene una tabla de probabilidad de emisión de cada aminoácido y una tabla de probabilidad de transición para moverse de un estado a otro.

Hay tres tipos de estados diferentes (ver figura inferior): los estados de "*match*" (cuadrados en la figura de abajo), donde los aminoácidos emitidos en ellos forman la secuencia primaria conservada de la

proteína; los estados de inserción (rombos), donde se emiten aminoácidos que resultan de inserciones; y los estados de deleción o estados silenciosos (círculos), que no emiten ningún aminoácido.

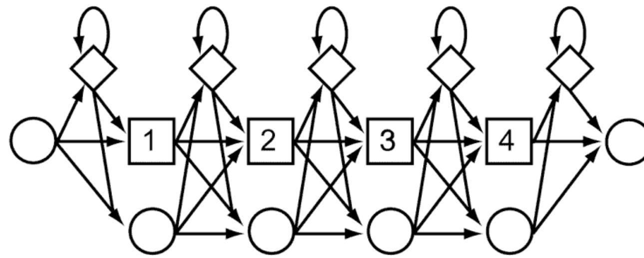


Figura 3: Ejemplo de HMM.

Una vez elegido el modelo, comienza el proceso de entrenamiento a base de ejemplos. Para esto, es una práctica común dividir el conjunto de ejemplos disponibles en dos partes: un 90% para entrenamiento y el 10% restante se reserva para muestras de ensayo, para comprobar si el modelo ha "aprendido" correctamente.

BacTregulators: una base de datos de reguladores transcripcionales en bacterias y arqueas

¿Por qué elegir familias de factores de transcripción para presentar los resultados de Provalidator? Los mecanismos más importantes que las bacterias usan para adaptar su fisiología a los cambios en las condiciones medioambientales están basados en regulación a nivel transcripcional. Los reguladores transcripcionales pueden ajustar la expresión de genes para hacer frente a cambios específicos del entorno y a condiciones de estrés. Por otra parte, reguladores transcripcionales que están presentes sólo en bacterias patógenas podrían estar involucrados en patogénesis y los reguladores transcripcionales compartidos cruciales en la fisiología de una bacteria no patógena pueden ser blanco para diseño de drogas. De ahí la importancia del estudio y diseño de *profiles* de familias de factores de transcripción.

Por otra parte, la rápida aparición de genomas de procariotas (bacterias+arqueas) recientemente secuenciados supone un crecimiento exponencial en el número de secuencias de proteínas, lo que contrasta con la limitación en la adquisición experimental de información sobre estas proteínas. Conforme se conocen más secuencias aumenta la posibilidad de que una secuencia nueva sea similar a otra ya conocida. Es importante encontrar estas similitudes porque el conocimiento de la antigua secuencia se puede usar para el análisis de la nueva. Así, una proteína cuya secuencia sea similar a otra de estructura tridimensional conocida probablemente tendrá una estructura parecida.

Otra cuestión es que los datos de genomas y bases de datos de proteínas a menudo no están integrados, por lo que hace falta un esfuerzo en esta dirección para clarificar los resultados de las búsquedas. BacTregulators se ha creado con el propósito de integrar información proveniente de los genomas secuenciados disponibles en NCBI y de la bases de datos de proteínas UniProt. Así, las secuencias de cada proteína de ambas fuentes se integran en un conjunto no redundante al que hace referencia cada proteína de BacTregulators, independientemente de la fuente de la que procede.

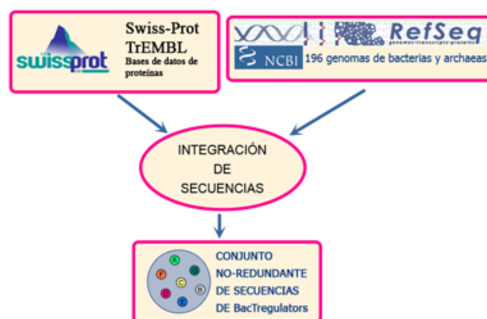


Figura 4: Esquema de integración de secuencias de BacTregulators.

BacTregulators está disponible en http://www.bacTregulators.org (Martínez-Bueno *et al*., 2004). La base de datos usa el gestor MySQL y la interfaz está programada en PHP. Actualmente, BacTregulators contiene información de tres familias de reguladores: AraC/XylS, TetR e IclR, donde los *profiles* de las dos últimas han sido definidos y validados usando Provalidator.

La extracción de información sobre reguladores es una tarea compleja y lenta. Cuando extraemos datos sobre un regulador siempre es posible asignar estos datos a una secuencia de proteína y a un microor-

ganismo. Además, si el organismo ha sido secuenciado, se puede asignar la información a una proteína codificada por un gen específico localizado en una posición precisa de un elemento de genoma. La cada vez mayor disponibilidad de genomas secuenciados acentúa la importancia de incluir datos genómicos en la definición de una proteína. Por otro lado, para asegurar la flexibilidad y precisión en la base de datos, es necesario mantener la información con el máximo grado de granularidad, considerando que es fácil combinar información al nivel que nos interese, mientras que no es posible separar información fundida.

Por tanto, la definición de un registro en BacTregulators se basa en la secuencia de la proteína, el microorganismo, el elemento de genoma (cromosoma, plásmido, profago, etc.) y la localización específica del gen en el correspondiente elemento de genoma.

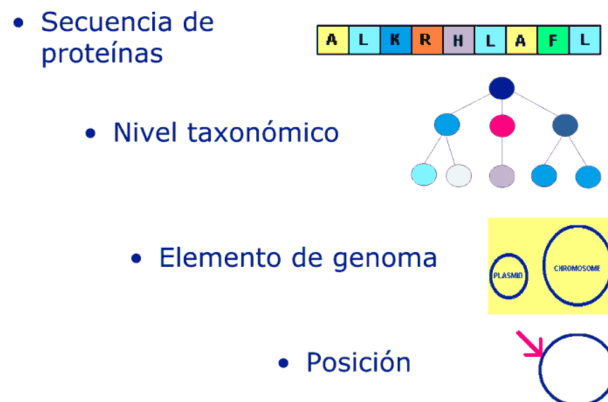


Figura 5: Definición de nuevo registro en BacTregulators.

El microorganismo se almacena siguiendo el criterio de la base de datos taxonómica de NCBI. Proteínas con diferencias en alguno de estos 4 niveles se almacenan como registros diferentes. Así, el número de proteínas en la base de datos BacTregulators es mayor que el número de secuencias porque varios registros pueden compartir la misma secuencia.

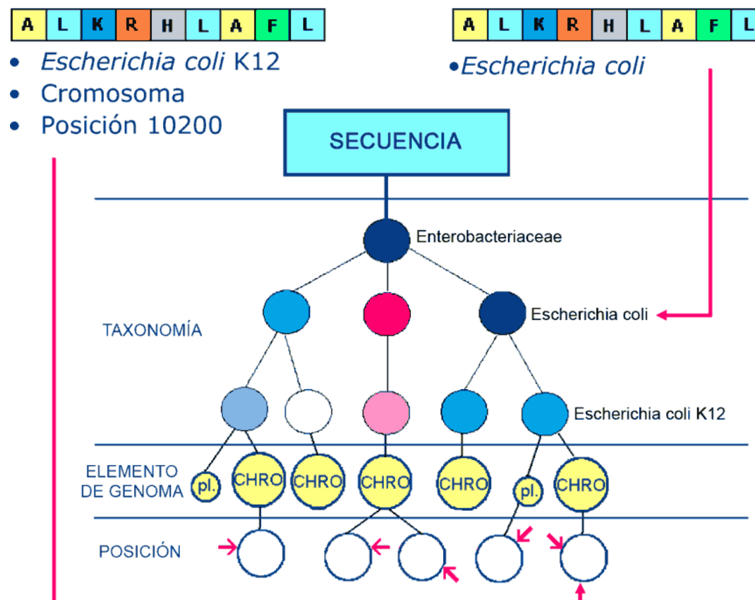


Figura 6: Criterios de almacenamiento de un registro en BacTregulators.

Durante la fase de extracción de datos de artículos publicados, se asigna información al registro apropiado, todo esto asistido por herramientas específicamente diseñadas para esta tarea.

Contenido y estructura de la base de datos BacTregulators:

BacTregulators contiene datos de tres tipos: secuencias, conocimiento y referencias. El proceso de extracción de conocimiento combina la asignación automática de datos con la manual. Los datos de asignación automática se extraen de las propias bases de datos de origen, UniProt o genomas microbianos de NCBI. El resto de conocimiento asociado a cada registro se extrae manualmente de referencias bibliográficas. Se ha usado una estructura especial para este tipo de datos, en la que se usa el párrafo como unidad elemental de información. Cada dato obtenido manualmente se organiza en párrafos de texto que se referencian individualmente. Usando esta estructura se consigue identificar el origen de cualquier dato. Durante la extracción manual de conocimiento, se pueden diseccionar datos experimentales que apoyan las cuestiones biológicas. A estos datos se les llama "evidencias experimentales" y reciben un tratamiento especial. Cuando hay alguna evidencia experimental que apoya algún dato de conocimiento expresado en un párrafo, se enlaza la evidencia al párrafo. Estos datos son importantes para evaluar la fiabilidad de un dato de conocimiento.

Cada párrafo de texto se referencia y enlaza a un artículo indexado por la base de datos de referencias Medline. También hay disponible información sobre estructuras tridimensionales de proteínas cristalizadas y tutoriales.

A continuación se muestra la estructura de BacTregulators, que le confiere las características anteriormente descritas. La línea horizontal gruesa separa físicamente lo que está separado conceptualmente. Hay dos áreas principales de conocimiento. La superior almacena los datos automáticos extraídos de las bases de datos y su tabla principal es "Proteins". La inferior almacena el conocimiento extraído manualmente y es la tabla "Paragraphs" la que ocupa el puesto principal.

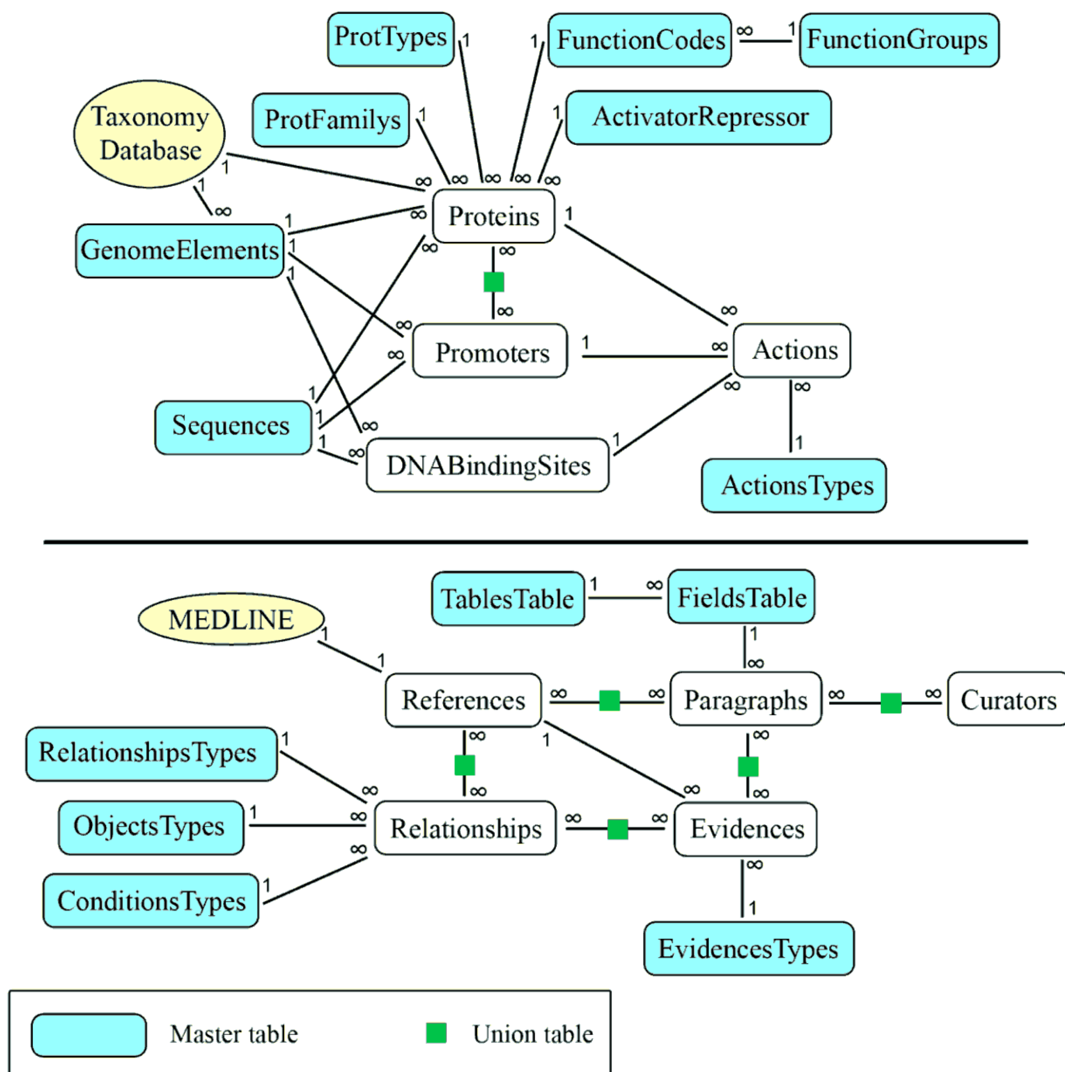


Figura 7: Estructura de la base de datos BacTregulators.

Acceso a la información de BacTregulators:

Se puede acceder a la información contenida en la base de datos BacTregulators de varias maneras. Se puede buscar por el identificador de proteína en BacTregulators o por los identificadores de UniProt o NCBI. Una búsqueda por familia, microorganismo, nombre y COG permite al usuario obtener conjuntos específicos de proteínas. Por último, una búsqueda de texto en todos los campos de la base de datos aumenta la flexibilidad. También se puede filtrar la base de datos de referencias con una búsqueda de texto.

La información sobre un registro específico se muestra en cuatro secciones. La primera sección (con una barra vertical azul oscuro) muestra los datos que definen el registro en BacTregulators: identificador de secuencia, organismo, elemento de genoma y posición en el elemento de genoma. La segunda sección

(con barra azul claro) contiene los datos obtenidos automáticamente de las bases de datos: número de acceso en UniProt o NCBI, base de datos de origen, nombre corto y nombre completo de la proteína, nombre del gen, orientación del gen, código COG, código funcional de NCBI y última fecha de actualización de este conjunto de datos. La tercera sección (barra verde) muestra el conocimiento extraído manualmente de artículos publicados. Este conocimiento está estructurado en los siguientes campos: función, genes regulados, redes de regulación, efectores, promotores de genes regulados, promotores del regulador, dependencia con factores sigma, patogenia, aplicaciones, datos relativos a mutaciones, estructura tridimensional, oligomerización, similitudes, comentarios y última fecha de actualización de este grupo de datos extraídos manualmente. La última sección (barra naranja) provee acceso a información relacionada con la secuencia de la proteína y los resultados de BLAST de comparar esta secuencia con el resto de miembros de la familia.

En la siguiente figura se muestra una descripción de cada una de estas secciones sobre una captura de pantalla de un informe real devuelto por una consulta de BacTregulators:

Figura 8: Descripción de informe de resultados en BacTregulators.

Integración de conocimiento personalizada en el paso de recuperación de información:

Cuando un usuario accede a un registro específico, el conocimiento asociado a registros relacionados puede ser útil para inferir características funcionales. Basada en esta idea, se diseñó una novedosa herramienta para visualizar un informe integrado de un conjunto de registros relacionado con un único registro inicial. Aunque la información asociada a cada registro se mantiene con el máximo grado de granularidad, la información correspondiente a un conjunto dado de registros relacionados se puede combinar en el paso de recuperación. El criterio clave para definir un grupo de registros relacionados es la similitud entre secuencias. Para seleccionar el conjunto de registros relacionados, se muestra una lista con todos los registros de BacTregulators que tienen una similitud de BLAST significativa con la secuencia inicial. En un primer paso, el usuario selecciona un valor de E como umbral para definir el conjunto de secuencias relacionadas. En un segundo paso, este conjunto puede ser acotado por criterios taxonómicos: el usuario elige el nivel taxonómico para incluir un registro en el conjunto de registros relacionados. Para esto, se muestra la clasificación taxonómica completa correspondiente al microorganismo del registro inicial.

La siguiente figura es un ejemplo del formulario que muestra BacTregulators para elegir el umbral de similitud y el filtrado taxonómico:

Figura 9: Formulario de filtrado en BacTregulators.

Una vez establecido el umbral de similitud y los límites taxonómicos, se muestra un informe integrado del conjunto de registros seleccionado (Figura 10). Los datos se muestran por campos en diferentes secciones. Cada sección incluye todos los datos de ese campo para todos los registros del conjunto. Puede accederse a un registro en particular a través de un enlace. Dentro de cada sección, los datos de conocimiento se ordenan por similitud con el registro inicial. Así, en la figura vemos dentro del campo "Function" que, obviamente, el registro 1 es el que más parecido tiene consigo mismo, después el 175 y, por último, el 192.

Cada afirmación de conocimiento que se hace está avalada por una referencia bibliográfica (enlace en verde), una evidencia experimental (en amarillo) y/o una imagen ilustrativa (en marrón). BacTregulators consta, de hecho, de una base de datos bibliográfica propia que contiene las referencias a todo el conocimiento aportado, con su correspondiente enlace a la fuente: MEDLINE, PubMed, etc.

Figura 10: Informe de integración de conocimiento en BacTregulators.

Herramientas

Bases de datos utilizadas

Las bases de datos de biología molecular se han convertido en el recurso más importante para el avance de esta disciplina. Una de las consecuencias más visibles del paso de la era genómica a la postgenómica fue el nacimiento de una verdadera comunidad de bioinformática. Actualmente son las bases de datos relativas a biología las que tienen un crecimiento más rápido y en las que se invierte más tiempo para su desarrollo. Nacen como un intento de recopilar y permitir el libre acceso a la información por parte de la comunidad de investigadores. A continuación, se describen las bases de datos más importantes usadas en esta Tesis, ya sean de secuencias de ADN o proteínas, de recursos de anotación, caracterización y clasificación de proteínas, de motivos o bibliográficas.

Bases de datos de secuencias:

UniProt (Bairoch *et al.*, 2005) es sin duda la base de datos de secuencias de proteínas por excelencia y la más utilizada debido a su alta calidad de anotación. UniProt (*Universal Protein Resource*) se creó como la unión de las bases Swiss-Prot, TrEMBL y PIR.

Genomas microbianos de NCBI (*National Center for Biotechnology Information*): Esta base de datos se concibió como lugar donde depositar las secuencias de ADN de los genomas recientemente secuenciados. Incluye un programa BLAST integrado para el que se puede personalizar el conjunto de microorganismos en el que buscar similitudes con la secuencia problema.

TIGR (Haft *et al.*, 2003): "*The Institute for Genomic Research*" es un centro dedicado al análisis e interpretación de genomas, incluye no sólo genomas completos de bacterias o arqueas, sino también de plantas como *Arabidopsis thaliana* o cultivos como el arroz. Están disponibles más de tres docenas de genomas de patógenos humanos, como las bacterias que causan neumonía, cólera, sífilis, meningitis, etc.

GOLD (Liolios *et al.*, 2008): "*Genomes Online Database*" es un recurso web para acceder a información de genomas completos o en proyecto de secuenciación, así como metagenomas y metadatos.

Bases de datos de motivos:

Cuando se tiene un número importante de motivos o dominios definidos a partir de un método determinado es posible construir una base de datos que almacene y permita disponer de dicha información. Estas bases de datos se caracterizan por poseer normalmente un sistema de consulta que nos permite determinar si nuestra secuencia problema puede ser clasificada dentro de un determinado grupo o familia de proteínas, o si posee algún motivo o dominio definido. Las bases de datos de motivos más importantes hasta el momento se describen a continuación:

PROSITE (Sigrist *et al.*, 2002): Describe actualmente más de 1500 familias de proteínas, ya sea median-

te *profiles* o motivos. Los motivos se establecen mediante patrones regulares. Como anteriormente indicamos, los patrones regulares tienen ciertos problemas de inflexibilidad a la hora de reconocer homólogos remotos o secuencias pertenecientes a una familia pero altamente divergentes. En este caso, la base de datos también incluye *profiles*, técnica que se muestra más poderosa a la hora de encontrar similitud en homólogos muy remotos con grupos o familia de proteínas. PROSITE está basado en la información existente en Swiss-Prot (Boeckmann *et al.*, 2003), pero no incluye información procedente de otras fuentes.

BLOCKS (Henikoff and Henikoff, 1994): Base de datos compuesta por pequeños segmentos de alineamientos múltiples correspondientes a entradas de PROSITE. De hecho, BLOCKS es un sistema de detección de motivos más que una base de datos propiamente.

PRINTS (Attwood, 2002): Es una base de datos integrada por una serie de motivos conservados. En este caso, el motivo no tiene por qué estar necesariamente como una secuencia contigua en el alineamiento, por lo que puede estar definido por un número de fragmentos dispersados a lo largo del gen.

Pfam (Bateman *et al.*, 2004): Es una base de datos compuesta por los perfiles HMM obtenidos para distintos dominios o regiones conservadas en las proteínas. El método HMM también se utiliza como herramienta de búsqueda de dichos dominios o regiones en las secuencias problema. Los HMM son el sistema más sensible para detectar homólogos remotos.

SMART (Letunic *et al.*, 2004): SMART (*Simple Modular Architecture Research Tool*) es una herramienta para la identificación y anotación de dominios de proteínas y proporciona una plataforma para el estudio comparativo de arquitecturas de dominios complejos en genes y proteínas.

InterPro (Mulder *et al.*, 2005): Esta base de datos proporciona una visión integrada de las bases de datos de motivos más comúnmente usadas. No usa ningún método de detección en particular sino que recopila lo que publican las otras bases de datos. Por tanto, pone a nuestro servicio un compendio de todo lo que se conoce de una determinada proteína, lo cual es extremadamente útil y cómodo ya que estas bases de datos no comparten formato ni nomenclatura y cada una tiene sus propios puntos fuertes y debilidades. InterPro analiza proteínas provenientes de la base de datos UniProt.

Bases de datos bibliográficas:

PubMed: Es un servicio de la *National Library of Medicine* (NLM) que incluye más de 15 millones de citas de MEDLINE fechados desde mediados de los años 50. Casi la totalidad de publicaciones de ámbito científico están indexadas en MEDLINE. PubMed se actualiza diariamente y contiene actualmente citas bibliográficas y *abstracts* de más de 5000 revistas biomédicas. Incluye búsquedas avanzadas.

Google Scholar: En noviembre de 2004 Google lanza una librería mundial, especialmente médica y científica, como un subconjunto del gran buscador Google, consistente en los textos completos de artículos de

revistas, informes técnicos, libros y otros documentos, incluidas las páginas webs consideradas académicas. Aunque Google Scholars cubre una amplia gama de áreas del saber, se presenta especialmente fuerte en ciencias.

Provalidator: una herramienta para diseño y validación de profiles

Provalidator es una herramienta orientada a su uso en la web que proporciona ayuda a la hora de diseñar y validar un *profile* a la manera que describe PROSITE. Esta herramienta combina una casi completa automatización en el paso de construcción del *profile* con una validación basada en la anotación de proteínas en InterPro para dar fiabilidad a los resultados obtenidos. Esta herramienta web está disponible de forma libre en:

http://www.bacTregulators.org/provalidator

La calidad de los resultados en el diseño de un *profile* viene determinada por dos características: las secuencias elegidas para formar parte de la semilla (esto es, el conjunto de secuencias del alineamiento múltiple a partir del cual se construye el *profile*) y la región seleccionada del alineamiento. Se describen estos dos procesos:

Elección de las secuencias:

A la hora de seleccionar las secuencias que formarán parte de la semilla del *profile*, Provalidator proporciona herramientas que automatizan tanto como es posible esta tarea.

Al principio partimos de todas las secuencias que están anotadas en InterPro como pertenecientes a la familia que queremos definir. El proceso de selección de secuencias se basa en tres pasos bien definidos:

1. Clusterizar las secuencias usando BLASTCLUST. Clusterizar consiste en agrupar secuencias por similitud. En nuestro caso se persigue tomar un representante de cada uno de los grupos que se forman para que ninguna secuencia pese más que otras en el alineamiento y así evitar sesgos en la construcción del *profile*. Además, un *profile* debe tener el menor número de secuencias posible para que no se especialice demasiado en el conjunto semilla y deje escapar secuencias medianamente remotas, pero no obstante miembros de la familia. Se podría decir que diseñar un *profile* consiste en abstraer lo común a un conjunto de secuencias, pero cuidando que no pierdan su idiosincrasia.
   BLASTCLUST tiene un parámetro que representa el umbral de similitud con el que se formarán los clústeres medido en tanto por ciento. Experimentalmente, se comprueba que un umbral del 60% es adecuado para descartar homólogos cercanos.

2. Los criterios para seleccionar un representante de cada clúster son (por este orden): a) Dar prioridad a las proteínas de Swiss-Prot frente a las de TrEMBL (ya que están mejor anotadas). b) Elegir proteínas cuya descripción contenga el nombre de la familia (cuando esto ocurre es porque se ha comprobado experimentalmente que ese regulador actúa como se espera de un miembro de esa familia). c) Filtrar las proteínas atendiendo a no incluir ciertas palabras clave en la descripción de la proteína (palabras como *hypothetical, probable, putative, fragment, truncated*). Si después de hacer esto no queda ninguna proteína, se descarta ese clúster por no ser suficientemente fiable.

3. Excluir secuencias cuya longitud sea muy diferente a la del resto. Esto produciría un alineamiento de peor calidad.

Elección de la región:

El siguiente paso es alinear las secuencias seleccionadas usando CLUSTALW. La salida debe estar en formato MSF (requerimientos del programa PFMAKE). A la hora de seleccionar la región del alineamiento múltiple que se usará para construir el *profile*, es crucial que la zona esté muy conservada ya que los aminoácidos esenciales para la función y/o estructura están generalmente conservados.

El motivo más recurrente de unión a ADN de los reguladores a sus correspondientes promotores es un motivo conservado llamado HTH (*helix-turn-helix*) que consiste en una α-hélice, un giro y una segunda α-hélice, llamada hélice de reconocimiento. Cuando este motivo HTH está presente (esto es, en el 95% de los reguladores transcripcionales descritos en procariotas), suele ser una región importante a tener en cuenta como posible candidato a ser incluido en el *profile*, como ocurre con las familias AraC/XylS (Gallegos *et al.*, 1997) y TetR (Ramos *et al.*, 2005). Sin embargo, no siempre es una región lo suficientemente discriminante para definir una familia de proteínas, como ocurre en IclR (Krell *et al.*, 2006).

Una vez seleccionada la región, es importante volver a clusterizar los segmentos de secuencias resultantes (o sea, los incluidos en la región seleccionada) porque la primera clusterización se hizo basándose en las secuencias completas y puede que en esta región se encuentren homologías que antes no aparecían y que son suficientes como para descartar alguna secuencia. Esto no es un proceso iterativo, basta con hacerlo una vez.

Construcción del profile usando las herramientas PFTOOLS:

Se construye el *profile* usando el programa PFMAKE, incluido en el paquete de programas PFTOOLS (Bucher and Bairoch, 1994; Bucher *et al.*, 1996) especialmente creado para diseñar *profiles* tal y como describe PROSITE. Para ello, Provalidator crea un alineamiento temporal basado en el que se le ha pasado como entrada (el alineamiento de las secuencias completas), pero que sólo incluye la región seleccionada.

El paso siguiente consiste en calibrar el *profile* (Hofmann and Bucher, 1995). Cuando enfrentamos una secuencia al *profile* usando PFSEARCH, el programa devuelve un *raw-score*, o sea, una puntuación en bruto, sin ningún significado biológico ni estadístico. Para solucionar esto se realiza una calibración, proceso que consiste en calcular los parámetros *R1* y *R2* de una función de normalización. Esto resulta en una transformación lineal de los *raw-scores* para producir *scores* normalizados o *N-scores*, que están definidos en una nueva escala con un significado estadístico concreto.

Función de normalización usada
$N\ score = R1 + R2 \times raw\ score$

Para realizar esto, se enfrenta el *profile* a una base de datos barajada y se usan los *raw-scores* obtenidos para calcular los parámetros. Esto lo hace el programa PFSCALE que, además de la lista de estos *scores*, necesita el tamaño en aminoácidos de la base de datos usada (N) y los límites superior e inferior del rango de probabilidades a los que se ajustarán los valores extremos de la distribución (P y Q, respectivamente).

Por ejemplo: si N=10.000.000 y P=0.0001, entonces se ignorarán los *scores* por debajo de 1000 (que corresponden a una probabilidad de ocurrencia >0.0001). Y si Q=0.000001, entonces se ignorarán los *scores* por encima de 10 (que corresponden a una probabilidad de ocurrencia <0.000001). El conjunto de secuencias "barajadas" se construye cogiendo ventanas no solapantes de X aminoácidos y barajando los aminoácidos dentro de cada ventana. En nuestro caso hemos usado un conjunto ya establecido como estándar: la versión 34 de la base de datos Swiss-Prot usando ventanas de tamaño 20. Los parámetros de este conjunto son: N=21.000.000, P=0.00005, Q=0.0000005. Al conjunto barajado resultante se le llama WINDOW20.SEQ (disponible en: ftp://ftp.isrec.isb-sib.ch/pub/databases/shuffled).

Este proceso es muy común en herramientas bioinformáticas y sirve para dar significación estadística a los *scores* que devuelven los programas. Así, la *E* de BLAST o el *N-score* de PFSEARCH están basados en la probabilidad de encontrar esa coincidencia de secuencias al azar.

Validación del profile usando InterPro:

A continuación, se enfrenta el *profile* calibrado contra la base de datos UniProt usando PFSEARCH, y se analizan los resultados de la validación, que presenta Provalidator. Este paso de enfrentar el *profile* contra UniProt es el más costoso en tiempo y potencia de cálculo. Hay que tener en cuenta que actualmente sólo la parte de procariotas de UniProt ocupa 1.5GB, y va creciendo de forma exponencial.

El siguiente cuadro es un ejemplo de este formulario de resultados:

Validando familia IclR (CutOff=8.5):

Falsos negativos = 18		
Q62IS0	8.337	2,3,4
Q9EWL2	8.189	1,2,3,4
Q6FBA6	7.932	2,3,4
Q82D43	7.895	1,2,4
Q5YU96	6.441	1,2,4
Q5Z0G4	6.386	1,2,4
Q7WPH1	5.558	2,4
Q762I2	5.190	1,2,4
Q5PJJ6	5.117	1,2,3,4
Q57K18	4.712	2
Q7WFW4	4.620	2,3
Q7WBG1	4.546	2,4
Q5YWX8	3.608	4
Q73W85	3.479	2
Q4NBL4	3.295	2
Q6UP88	3.258	1,4
Q9S4Y5	2.725	1,2,4
Q65WD5	2.504	1,3

Falsos positivos = 0

Métodos: 1: PF01614; 2: PS51077; 3: PS51078; 4: SM00346

Figura 11: Resultado de validación de la familia IclR.

Validar un *profile* consiste en comparar el conjunto de proteínas consideradas miembros de la familia con las anotaciones de InterPro. Como se ve, el informe consta de tres columnas para el análisis de falsos potenciales: 1) El identificador o número de acceso de las proteínas en la base de datos UniProt. 2) El *score* que han obtenido con el *profile*. 3) Los métodos con los que InterPro ha anotado esa proteína como miembro de la familia (esta columna no se incluye en la parte de potenciales falsos positivos por razones obvias). Esta información es muy útil para encontrar contradicciones en los métodos de anotación, como se verá en el análisis de resultados de la familia IclR. Los prefijos de los métodos corresponden a: PF: Pfam, PS: PROSITE, SM: SMART.

Objetivos

Generales:

- El objetivo general de esta Tesis Doctoral ha sido el desarrollo de herramientas informáticas para aplicaciones a los problemas de biología que estudia el grupo de *Pseudomonas* de la Estación Experimental del Zaidín, a la vez que se perseguía generar "conocimiento" general de aplicación en biología.

Específicos:

Los objetivos específicos se han ido estableciendo en el tiempo con objeto de definir mejor distintas familias de reguladores transcripcionales. De manera específica se ha desarrollado:

- Un *profile* que define a las familias de factores de transcripción TetR e IclR para obtener más información de las proteínas pertenecientes a estas familias.

- Desarrollar Provalidator para automatizar la construcción de *profiles*. Como parte de este objetivo, se establecen dos subobjetivos encaminados a validar el funcionamiento de esta herramienta.

    - Construcción del *profile* de MerR,
    - y el de bombas de extrusión RND.

- Crear BacTregulators, una base de datos que recoge los resultados obtenidos por Provalidator.

Resultados

Capítulo 1

The TetR Family of Transcriptional Repressors

Microbiology and Molecular Biology Reviews

MICROBIOLOGY AND MOLECULAR BIOLOGY REVIEWS, June 2005, p. 326–356 Vol. 69, No. 2
1092-2172/05/$08.00+0 doi:10.1128/MMBR.69.2.326–356.2005
Copyright © 2005, American Society for Microbiology. All Rights Reserved.

The TetR Family of Transcriptional Repressors

Juan L. Ramos,[1]* Manuel Martínez-Bueno,[1] Antonio J. Molina-Henares,[1] Wilson Terán,[1] Kazuya Watanabe,[2] Xiaodong Zhang,[3] María Trinidad Gallegos,[1] Richard Brennan,[4] and Raquel Tobes[1]†

Department of Plant Biochemistry and Molecular and Cellular Biology, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, E-18008 Granada, Spain[1]; Laboratory of Applied Microbiology, Marine Biotechnology Institute 3-75-1 Heita, Kamaishi, Iwate 026-0001, Japan[2]; Department of Biological Sciences, Imperial College London, Flowers Building, Armstrong Road, South Kensington, London, SW7 2AZ, United Kingdom[3]; and Department of Biochemistry & Molecular Biology, Oregon Health Sciences University, 3181 SW Sam Jackson Park Road, Portland, Oregon 97239-3098[4]

INTRODUCTION 326
DEFINING THE TetR FAMILY 327
TetR Family Profile 327
Identification of TetR Family Members in DNA and Protein Databases 333
PROTEINS WITH KNOWN THREE-DIMENSIONAL STRUCTURES 333
TetR Regulator 333
Tetracycline resistance and the role of the transcriptional regulator TetR 333
TetR DNA-binding domain: a symmetric TetR dimer binds a palindromic operator 337
QacR Regulator 339
Two QacR dimers bind the operator to repress the *qacA* multidrug transporter gene 339
QacR as a model for multidrug recognition 340
Three-Dimensional Structure of CprB 341
EthR Structure 341
Crystal structure of TetR family members with unknown functions 341
DNA-BINDING PREDICTIONS BASED ON TetR AND QacR CRYSTAL STRUCTURES 341
Relationship between Profile Positions and Structural Positioning 342
SOME REGULATORS ARE PART OF COMPLEX REGULATORY CIRCUITS 342
AcrR Regulator Is the Local Specific Regulator of the *acrAB* Efflux Pump 343
Mtr Circuit of *Neisseria* 343
BetI Controls the Choline-Glycine Betaine Pathway of *E. coli* 345
ArpA Regulator from *Streptomyces* 346
HapR Regulates Virulence Genes in *Vibrio cholerae* 346
Other Quorum-Sensing Circuits 347
BIOTECHNOLOGICAL APPLICATIONS AND FUTURE PROSPECTS 347
ACKNOWLEDGMENTS 347
REFERENCES 347

INTRODUCTION

Bacteria in the environment are exposed to variations in temperature and nutrient and water availability and the presence of toxic molecules that originate from their abiotic and biotic surroundings (including deleterious molecules that originate from their own metabolism). These changes can make their living conditions far from optimal. Survival in this unstable environment requires a wide range of rapid, adaptive responses which are triggered by regulatory proteins. These regulators respond to specific environmental and cellular signals that modulate transcription, translation, or some other event in gene expression, so that the physiological responses are modified appropriately (32, 52, 64, 104, 107, 145, 244, 311, 312, 326, 330, 379, 397, 409, 427).

In most cases, the adaptive responses are mediated by transcriptional regulators. Most microbial regulators involved in transcriptional control are two-domain proteins with a signal-receiving domain and a DNA-binding domain which transduces the signal (1, 18, 145, 152, 170, 207, 271, 292–294, 298, 303, 345, 369, 428, 431) (Table 1). In other cases, the sensing of signals that trigger a transcriptional process involves two proteins, as in two-component regulatory systems such as CzcR/CzcS; DcuS/DcuR; NifL/NifA; NtrB/NtrC; PhoP/PhoQ; and TodS/TodT (75, 139, 200, 206, 233, 234, 257, 307, 309, 316, 409, 423). One protein is usually a membrane-linked kinase that, upon sensing the appropriate signal, phosphorylates a DNA-binding protein that mediates transcription from its cognate promoter. Structural analyses have revealed that the helix-turn-helix (HTH) signature is the most recurrent DNA-binding motif in prokaryotic transcriptional factors, since almost 95% of all transcriptional factors described in prokaryotes use the HTH motif to bind their target DNA sequences (12, 19, 27, 41, 43, 104, 135, 136, 302, 335, 343).

Prokaryotic transcriptional regulators are classified in fami-

* Corresponding author. Mailing address: Estación Experimental del Zaidín, CSIC, Apdo. Correos 419, E-18008 Granada, Spain. Phone: 34 958 181608. Fax: 34 958 135740. E-mail: jlramos@eez.csic.es.

† Present address: Esa7, Information Technologies, Las Gabias, E-18110 Granada, Spain.

TABLE 1. Prokaryotic regulator families

Family	Action	Some regulated functions	DBD motif	Position	Reference(s)
LysR	Activator/repressor	Carbon and nitrogen metabolism	HTH	N-terminal	145, 342
AraC/XylS	Activator	Carbon metabolism, stress response and pathogenesis	HTH	C-terminal	109, 394
TetR	Repressor	Biosynthesis of antibiotics, efflux pumps, osmotic stress, etc.	HTH	C-terminal	9, 10, 11
LuxR	Activator	Quorum sensing, biosynthesis and metabolism, etc.	HTH	C-terminal	106, 298, 317
LacI	Repressor	Carbon source utilization	HTH	N-terminal	54, 420
ArsR	Repressor	Metal resistance	HTH	Central	49, 432
IcIR	Repressor/activator	Carbon metabolism, efflux pumps	HTH	N-terminal	265, 319, 321, 378
MerR	Repressor	Resistance and detoxification	HTH	N-terminal	144, 377
AsnC	Activator/repressor	Amino acid biosynthesis	HTH	N-terminal	103
MarR	Activator/repressor	Multiple antibiotic resistance	HTH	Central	4, 13, 352, 376
NtrC (EBP)	Activator	Nitrogen assimilation, aromatic amino acid synthesis, flagella, catabolic pathways, phage response, etc.	HTH	C-terminal	200, 257
OmpR	Activator	Heavy metal and virulence (response regulator of a two-component system)	Winged helix	C-terminal	237
DeoR	Repressor	Sugar metabolism	HTH	N-terminal	311, 405, 450
Cold shock	Activator	Low-temperature resistance	RNA binding domain (CSD)	Variable	42, 205, 344
GntR	Repressor	General metabolism	HTH	N-terminal	138, 318, 324
Crp	Activator/repressor	Global responses, catabolite repression and anaerobiosis	HTH	C-terminal	54, 110, 244

lies on the basis of sequence similarity and structural and functional criteria (49, 84, 106, 108, 120, 121, 138, 145, 146, 237, 274, 308, 313, 324, 342, 370, 377). Table 1 lists the most important families of microbial transcriptional regulators, the type of DNA binding motifs they exhibit, whether the members of the family are preferentially repressors or activators, and whether they show a dual action.

This review focuses on the TetR family, a family of transcriptional regulators that is well represented and widely distributed among bacteria with an HTH DNA-binding motif (210, 211, 246, 288).

Members of the TetR family of repressors are identified by a profile (see below) which can be easily used to recognize TetR family members in SWISS-PROT and TrEMBL and in all available proteins from prokaryotic genome sequences. After compiling data from protein and nucleic acid databases, the TetR family of regulators was found to include 2,353 nonredundant sequences (as of December 2004). The specific function regulated by members of the TetR family is known for only about 85 members (Table 2). These proteins control genes whose products are involved in multidrug resistance, enzymes implicated in different catabolic pathways, biosynthesis of antibiotics, osmotic stress, and pathogenicity of gram-negative and gram-positive bacteria (Table 2). The most relevant information on these proteins is collected in a database available at http://www.bactregulators.org (235). The database also supplies information for each member of the family, including identifiers, names, sequences, source, function, COG (clusters orthologous groups), position and orientation of the corresponding gene in the genome, and, when available, three-dimensional structures.

DEFINING THE TetR FAMILY

TetR Family Profile

The TetR family is named after the member of this group that has been most completely characterized genetically and biochemically, the TetR protein (141, 148, 150, 168, 288, 395). This protein controls the expression of the *tet* genes, whose products confer resistance to tetracycline (150, 183, 209, 210, 337, 434, 435). Members of the TetR family exhibit a high degree of sequence similarity at the DNA binding domain (see below). Interpro (258) assigns proteins to the TetR family based on PROSITE signature PS01081 (364), PRINTS motif PR00455 (15, 16), and Pfam Hidden Markov Model (HMM) profile PF00440 (26, 27). To establish a single criterion defining the TetR family, we decided to develop a conventional profile, because conventional profiles are easy to manage and their sensitivity is equivalent to that of HMM profiles.

To develop the TetR family profile, we first selected a set of 120 sequences as belonging to the TetR family based on two criteria: a positive score for PROSITE signature PS01081, and a high score for PF00440 HMM. The 120 sequences were clustered into 42 groups using BLAST, and a representative sequence was selected and aligned for each cluster using CLUSTAL (http://clustalw.genome.ad.jp/). This revealed that the most conserved region corresponded to the HTH domain described in the TetR and QacR crystals (120, 150, 287, 288, 289, 349, 350, 351). The initial HTH motif was progressively extended until the global score of the multialignment diminished. Figure 1 shows the final alignment of the sequences. This conserved stretch corresponded in TetR and QacR crystals to the almost complete α-helix 1, the HTH domain formed by α-helices 2 and 3, and five residues of α-helix 4 that connect the DNA-interacting region with the core of the protein (see Fig. 2 for the three-dimensional structure of TetR).

The final alignment shown in Fig. 1 was used as a seed for the construction of a conventional profile to detect TetR family members. The TetR profile was built using the pfmake program available at the Swiss Institute of Bioinformatics (http://npsa-pbil.ibcp.fr/cgi-bin/npsa__automat.pl?page=/NPSA/npsa__pfmake.html) (45, 46). The TetR profile was confronted against the 660,992 bacterial and archaeal proteins in the SWISS-PROT and TrEMBL databases (released December

TABLE 2. Specific functions regulated by members of the TetR family of repressors

No.	SPTR[a]	Name	Organism	Function	G[b]	Reference(s)
1	P34000	AcrR	*Escherichia coli*	Represses the expression of the *acrAB* operon which confers multidrug resistance and probably also controls the gene *micF*	1	102, 164, 223, 224, 314, 325, 425
2	Q53901	ActII	*Streptomyces coelicolor*	Located in the *act* cluster, which contains regulatory and antibiotic export genes	1	50, 96
3	Q9F8V9	AmeR	*Agrobacterium tumefaciens*	Negatively regulates the *ameABC* operon, which encodes proteins similar to nodulation-cell division (RND)-type efflux systems	1	301
4	Q9RG61	AmrR	*Pseudomonas aeruginosa*	Probably regulates *amrAB* genes encoding an efflux system involved in aminoglycoside impermeability phenotype in *Pseudomonas aeruginosa*	1	423
5	Q9KJC4	ArpR	*Pseudomonas putida* S12	Seems to be a repressor for the expression of the *arpABC* operon; ArpABC in *Pseudomonas putida* S12 is involved only in multidrug resistance and not in tolerance towards organic solvents.	1	178
6	Q6VV70	BpeR	*Burkholderia pseudomallel*	Controls expression of the BpeAB-OprB efflux pump that extrudes gentamycin, streptomycin erythromycin, and acryflavine	1	53
7	P31676	EnvR	*E. coli* K-12	Regulates the *acrEF* efflux pump operon, which is relevant to multidrug resistance in *E. coli*. Its substrate specificity (antibiotics, basic dyes and detergents) is similar to that of AcrAB	1	186
8	P96222	EthR	*Mycobacterium tuberculosis*	Ethionamide resistance	1	28, 90, 111
9	P72185	HemR	*Propionibacterium freudenreichii*	Probably regulates *hemX*, which appears to be involved in heme transport	1	137
10	Q93QZ7	HydR	Tn*5398* from *Clostridium difficile*	Involved in erythromycin resistance	1	93
11	O68442	IfeR	*Agrobacterium tumefaciens 1D1609*	Seems to be a repressor that controls the expression of the putative *ifeABR* isoflavonoid efflux system	1	296
12	Q9ZGB7	LanK	*Streptomyces cyanogenus*	Probably a landomycin A resistance regulator	1	315, 424
13		LfrR	*Mycobacterium segmatis*	Control of the *lfrA* gene whose end product confers resistance to fluoroquinolones, ethidium bromide, and acryflavine	1	212
14	O34619	LmrA	*Bacillus subtilis*	Probable repressor of the lincomycin-resistance operon	1	197, 198, 260
15	P39897	MtrR	*Neisseria gonorrhoeae*	A transcriptional repressor that regulates transcription of the *mtrCDE* genes, which encode a multidrug efflux pump; MtrR acts directly or indirectly as a positive regulator of *farAB* gene expression	1	72, 130, 131, 220, 221, 297, 333, 334, 339, 447, 448
16	Q9F0Y2	Pip	*Streptomyces coelicolor*	Pristinamycin I-induced regulator that controls mutidrug resistance genes	1	99
17	Q9F147	PqrA	*Streptomyces coelicolor*	Probably the repressor of *pqrB*, which encodes an efflux pump conferring resistance to paraquat	1	61
18	P23217	QacR	*Staphylococcus aureus*	Regulates the QacA multidrug efflux pump	1	119, 120, 249, 299, 300, 332, 350, 351, 390
19	O52558	RifQ	*Amycolatopsis mediterranei*	Located in the rifamycin biosynthetic gene cluster and probably related to the adjacent gene that encodes a rifamycin efflux protein	1	17
20	Q9KIH5	RmrR	*Rhizobium etli* plasmid B	Probably regulates the operon *rmrAB* related to a multidrug efflux pump involved in sensitivity to phytoalexins, flavonoids, and salicylic acids	1	113
21	Q9AMH9	SimReg 2	*Streptomyces antibioticus*	Included in the *Streptomyces antibioticus* simocyclinone biosynthetic gene cluster; probably regulates the putative export protein SimEX	1	396
22	Q8KLP4	SmeT	*Stenotrophomonas maltophilia*	A repressor of the *Stenotrophomonas maltophilia* multidrug efflux pump SmeDEF	1	340, 452
23	Q9R9T9	SrpR	*Pseudomonas putida*	Probable regulator of the solvent resistance pump SrpABC of strain S12	1	163, 179, 180, 422
24	P39885	TcmR	*Streptomyces glaucescens*	A regulator of the tetracenomycin C resistance repressing the gene *tcmA*, which encodes an export pump	1	126, 127
25	P09164	TetR	*Escherichia coli*	Controls the expression of tetracycline resistance mediated by the gene *tetA*, which encodes an efflux pump that acts as an antiporter by coupling the export of [MgTetracycline]$^+$ out of the cell with the uptake of protons	1	29, 30, 36, 38, 39, 142, 143, 148, 149, 150, 183, 189, 259, 286, 287, 288, 289, 393, 402, 430
26	Q9AIU0	TtgR	*Pseudomonas putida*	Regulates the TtgABC efflux pump mediating organic solvent tolerance and resistance to ampicillin, tetracycline, chloramphenicol, and nalidixic acid	1	81, 391
27	Q93PU7	TtgW	*Pseudomonas putida*	*ttgW* is a pseudogene	1	327, 329
28	Q9RP98	UrdK	*Streptomyces fradiae* Tu2717	Probably regulates an urdamycinA efflux pump	1	94
29	Q9AJL5	VarR	*Streptomyces virginiae*	Regulates transcription of *varS*, the virginiamycin S-specific transporter in a virginiamycin S-dependent manner	1	263
30	P96676	YdeS	*Bacillus subtilis*	Similar to a regulator of antibiotic transport complexes in *Streptomyces hygroscopicus*	1	34
31	Q54189	ArpA	*Streptomyces griseus*	Represses the expression of *adpA*; AdpA activates the expression of *strR*, and the StrR protein activates the expression of streptomycin biosynthetic genes. ArpA also controls morphogenesis	2, 5, 8	157, 278, 282, 283, 285, 375, 412, 438

Continued on facing page

TABLE 2—*Continued*

No.	SPTR[a]	Name	Organism	Function	G[b]	Reference(s)
32	Q93M20	Aur1B	*Streptomyces aureofaciens*	Included in the *Streptomyces aureofaciens* auricin polyketide biosynthesis gene cluster	2	275
33	Q9LBV6	BarA	*Streptomyces virginiae*	Probably involved in regulation of virginiamycin biosynthesis	2	175, 262
34	Q8KNI9	CalR1	*Micromonospora echinospora*	Included in the calicheamicin gene cluster	2	2
35	O66129	CprB	*Streptomyces coelicolor*	CprB is involved in the control of actinorhodin and undecylprodigiosin biosynthesis and morphogenesis	2	264, 284
36	O24741	FarA	*Streptomyces lavendulae* FRI-5	IM-2-specific receptor; plays an important role in the regulation of secondary metabolism and the biosynthesis of the antibiotics showdomycin and minimycin in *Streptomyces lavendulae*; FarA acts as a negative transcriptional regulator for the biosynthesis of nucleoside antibiotics and blue pigment, switching on their expression in the presence of IM-2; also acts as a positive transcriptional regulator for the biosynthesis of D-cycloserine, switching off its expression in the presence of IM-2	2	184, 185, 413
37	Q939Q2	JadR*	*Streptomyces venezuelae*	Included in the cluster for the biosynthesis of the dideoxysugar component of jadomycin B	2	416
38	Q56153	JadR2	*Streptomyces venezuelae*	Represses the biosynthesis of jadomycin B and seems to control cellular pigmentation	2	442, 443
39	Q9ZN97	MphB	*Escherichia coli plasmid* pTZ3721	Repressor of antibiotic biosynthesis	2	172, 272
40	Q9XDF0	NonG	*Streptomyces griseus* sbsp. *griseus*	Probably related to nonactin biosynthesis	2	414
41	Q9RF02	PhlF	*Pseudomonas fluorescens*	A repressor of the *phlABCD* operon responsible for the biosynthesis of the antifungal 2,4-diacetylphloroglucinol (PHL)	2	346
42	Q9ZHP8	TylQ	*Streptomyces fradiae*	Butyrolactone receptor TylQ is a potential regulator of production of the macrolide antibiotic tylosin	2, 8	371
43	Q8VQC6	VanT	*Vibrio anguillarum*	Positively regulates serine metalloprotease, pigment and biofilm production	2, 5	71
44	Q9RPK9	TarA	*Streptomyces tendae*	Hypothetical receptor of gamma-butyrolactone, which regulates nikkomycin synthesis	2, 8	86
45	Q9XCC7	TylP	*Streptomyces fradiae*	Regulates tylosin production and morphological differentiation, and is probably a gamma-butyrolactone receptor	2, 5, 8	25, 371, 372
46	Q59213	BM1P1	*Bacillus megaterium*	Probably acts as positive regulatory protein involved in the expression of the P450BM-1 gene by interfering with the binding of the repressor protein, Bm3R1, to the regulatory regions of P450BM-1	3	358, 361, 362
47	O68276	Bm1P1	*Bacillus megaterium* ATCC 14581	Negatively affects basal-level expression of P450BM-1, a barbiturate-inducible P450 monooxygenase; cytochromes P450BM-3 and P450BM-1 catalyze the hydroxylation of fatty acids	3	140, 214, 215, 295, 356, 358, 359, 362
48	P43506	Bm3R1	*Bacillus megaterium*	A transcriptional repressor involved in the regulation of barbiturate-inducible proteins in *Bacillus megaterium*	3	87, 88, 89, 140, 213, 214, 295, 356, 357, 358, 359, 361, 362
49	Q9AJ68	ButR	*Streptomyces cinnamonensis*	Putative transcriptional repressor of crotonyl-CoA reductase	3	218, 219
50	Q93TU7	CampR	*Rhodococcus* sp. NCIMB 9784	Probably regulates 6-oxocamphor hydrolase	3	123
51	Q51597	CamR	*Pseudomonas putida* plasmid CAM	A negative regulator of the cytochrome P-450cam hydroxylase operon	3	9, 10, 105
52	O33453	CymR	*Pseudomonas putida*	A repressor which controls expression of both the *cym* and *cmt* operons and is inducible by *p*-cumate but not *p*-cymene	3	62, 82, 83, 279
53	Q9RAJ1	DhaR	*Mycobacterium* sp. GP1	Appears to function as a repressor of *dhaA* expression, *dhaA* is an haloalkane dehalogenase gene included in the 1-clorobutane catabolic gene cluster	3	306
54	Q9RA03	KstR	*Rhodococcus erythropolis* strain SQ1	A repressor of *kstD* expression that encodes a 3-ketosteroid Δ-dehydrogenase protein involved in the degradation of steroid intermediates in phytosterol degradation	3	404
55	Q8VV87	LexA-like	*Terrabacter* sp. strain DBF63	Probably involved in degradation of dibenzofuran	3	171
56		AcnR	*Corynebacterium glutamicum*	Repressor of the *acn* gene encoding aconitrase and controlling the tricarboxylic acid cycle	3	195
57	Q9FA56	PaaR	*Azoarcus evanssi*	Probably regulates the *paa* genes, which are responsible for the aerobic phenylacetic acid catabolic pathway	3	256
58	Q9XDW2	PsbI	*Rhodopseudomonas palustris*	Included in the cluster of genes participating in aerobic biodegradation of *p*-cumate	3	310

Continued on following page

TABLE 2—*Continued*

No.	SPTR[a]	Name	Organism	Function	G[b]	Reference(s)
59	O85706	ThlR	*Clostridium acetobutylicum* DSM 792	Possibly acts as a transcriptional repressor of the *thlRBC* operon, which is involved in the biosynthesis of thiolase	3	429
60	Q59431	UidR	*E. coli*	A repressor of the *uidRABC* (*gusRABC*) operon that comprises a beta-D-glucuronidase (*uidA*), a glucuronide permease (*uidB*) and a membrane-associated protein (*uidC*)	3	40
61	P22645	YDH1	*Xanthobacter autotrophicus*	Probably regulates the *dhlA* gene involved in 1,2-dichloroethane degradation	3	162
62	P17446	BetI	*Escherichia coli*	A choline-sensing repressor of the *bet* regulon involved in osmotic stress	4	8, 201, 202, 331
63	Q8NLK1	McbR	*Corynebacterium glutamicum*	In absence of L-methionine, represses the expression of six key enzymes for the biosynthesis of the sulfur-containing amino acids L-cysteine and L-methionine including sulfonate utilization and sulfite reduction	4	322
64	Q9EVJ6	MphR	*Escherichia coli*	Represses the *mph(A)-mrx-mphR(A)* operon in the absence of erythromycin; erytromycin induces the synthesis of macrolide 2′-phosphotransferase I [Mph(A)], which inactivates erythromycin	4	273
65	Q9F9Z7	PhaD	*Pseudomonas oleovorans*	Biosynthesis of medium-chain-length (MCL) poly-3-hydroxyalkanoates (PHAs) as intracellular storage material	4	188, 451
66	Q9ZF45	Q9ZF45	*Lactococcus lactis*	Regulates the operon *purDEK*, which encodes enzymes in the de novo pathway of purine nucleotides	4	269
67	P06969	TtK	*Escherichia coli*	Co-transcribed with the *dut* (deoxyuridine triphosphatase) gene	4	85, 415
68	P32398	Yhgd or YixD	*Bacillus subtilis*	Probably related to protoheme IX biosynthesis	4	132
69	Q9F6W0	CasR	*Rhizobium etlli*	A repressor of the *casA* gene, which encodes the calmodulin-like protein calsymin involved in bacteroid development during symbiosis and in symbiotic nitrogen fixation	5	433
70	Q9RQQ0	IcaR	*Staphylococcus aureus*	A repressor of the operon *ica* which is responsible for an intercellular polysaccharide compound that acts as the slime in biofilm formation	5	70, 163
71	Q8GLC6	IcaR	*Staphylococcus epidermidis*	A repressor of the operon *ica*, which is reponsible for an intracellular polysaccharide compound that acts as the slime in biofilm formation	5	60, 66, 67, 190, 456
72	Q8KX64	LitR	*Vibrio fischeri*	Important for the normal induction of luminescence, plays a positive role in modulating the ability to colonize juvenile squid, and may control the opacity/translucent phenotype of the colony	5, 8	97
73	P21308	LuxR	*Vibrio harveyi*	Required for expression of the *luxCDABEGH* (luciferase) operon, responsible for bacterial luminescence	5	23, 24, 51, 57, 167, 232, 240, 250, 251, 253, 254, 255, 355, 365, 380, 381, 382
74	Q9ANS7	LuxT	*Vibrio harveyi*	Activates the expression of LuxO, the phosphorelay protein that regulates luminescence in *Vibrio harveyi*	5	216
75	O50285	OpaR	*Vibrio parahaemolyticus*	A transcriptional regulator that controls the opaque morphology in *Vibrio parahaemolyticus* colonies	5	240, 355
76	Q9XDV7	Orf2	*Streptomyces griseus*	Probably related to carbon-source-dependent differentiation in *Streptomyces griseus*	5	398
77	Q9L8G8	SmcR	*Vibrio vulnificus*	Appears to play an important role in starvation adaptation and in the regulation of many growth phase-regulated genes, including some virulence factors (protease, hemolysin); ScmR represses motility, fimbria production, and biofilm production	5 6 6	63, 242, 243, 355
78	O30343	HapR	*Vibrio cholerae*	A transcriptional regulator with a central role in control of the virulence of *Vibrio cholerae*, in a cell density-dependent way	6	165, 194, 196, 240, 455
79	Q8KU49	Ef0113	*Enterococcus faecalis*	Located in a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*	6	354
80	Q63B57	HlyIIR	*Bacillus cereus*	Regulates expression of *hly*II whose gene product has haemolytic activity	6	47
80	O24739	BarB	*Streptomyces virginiae*	Regulates virginiamycin biosynthesis	7	182
81	O86852	ScbR	*Streptomyces coelicolor*	Acts as the cytoplasmic receptor that specifically binds SCB1 gamma-butarylactone and negatively regulates transcription of the *scbA* gene, responsible for gamma-butyrolactone SCB1 synthesis	2, 7, 8	3, 385, 386
82	Q9JN89	MmfR	*Streptomyces coelicolor* plasmid SCP1	Putative lactone-dependent transcriptional regulator	8	437
83	Q9S3L4	AmtR	*Corynebacterium glutamicum*	Regulator of nitrogen control	9	48,161
84	Q9EX90	PsrA	*Pseudomonas putida*	Involved in the regulatory cascade controlling *rpoS* gene regulation in response to cell density	9	192
85	P36656	YjdC	*Escherichia coli*	Probably involved in copper tolerance	10	101

[a] Swiss-Prot and TrEMBL accession number.

[b] 1, regulation of efflux pumps and transporters involved in antibiotic resistance and tolerance to toxic compounds; 2, regulation of antibiotic biosynthesis; 3, regulation of catabolic pathways; 4, biosynthesis of products important for bacteria (e.g., osmoprotectants, nucleotides, amino acids, PHAs, protoheme); 5, regulation of differentiation (sporulation, mycelium formation), colony phenotype, biofilm formation; 6, regulation of genes involved in virulence; 7, regulation of butyrolactone synthesis; 8, butyrolactone or autoinducer receptors; 9, global regulation; 10, other.

```
                6  10          20         30         40         52
                |---+----+---  -+----+----+----+----+----+----+-|
TetR            RESVIDAALELLN-ETGIDGLTTRKLAQKLGIEQPTLYWHVKNKRALL
                ------- α1 -------      -- α2 --   -- α3 --   - α4 →

Profile pos.    1        10        20         30         40      47
                |---+----+---  -+----+----+----+----+----+----+-|
ENVR_ECOLI      RQELIETAIAQFA-QHGVSKTTLNDIADAANVTRGAIYWHFENKTQLF
MexZ            RDGILDAAERVFL-EKGVGTTAMADLADAAGVSRGAVYGHYKNKIEVC
BETI_RHIME      RKALVDAALRVIG-DQGTLAVTMSEIARTAGVSPALAHHYFGSKEQLL
YBIH_ECOLI      KKQLIAAALAQFG-EYG-MNATTREIAAQAGQNIAAITYYFGSKEDLY
BETI_ECOLI      RRQLIDATLEAIN-EVGMHDATIAQIARRAGVSTGIISHYFRDKNGLL
LUXR_VIBHA      KQQLMEIALEVFA-RRGIGRGGHADIAEIAQVSVATVFNYFPTREDLV
Q9KP84          RDQLIMHARELFS-VMPYDKVSTRLIASKAGVDISLIRYYFANKAGLF
TCMR_STRGA      RDQLIREALELFL-AQGYEHTTVEQIAEAVEVHPRTFFRHFASKEEVA
YCFQ_ECOLI      RDAALDKAMKLFW-QHGYEATSLADLVEATGAKAPTLYAEFTNKEGLF
YJGJ_ECOLI      PEQAISAAKVLFH-QKGFDAVSVAEVTDYLGINPPSLYAAFGSKAGLF
TETC_ECOLI      YQSLVNSARILFV-EKGYQAVSIDEISGKALVTKGAFYHHFKNKKQLL
YWCC_BACSU      FESFIDAGVNLLI-DRAYDPVSVEDISRAAGYSKGAFYVHFVSKDDFL
YERO_BACSU      KERIIKTSIRLFA-KKGFAATTIQEIASECGISKGAFYLHFKSKEALL
YPB3_LACLA      KEKILTATEQLIY-KKGYTGTSINDILDETATGKGQFYYYFDSKKEAC
YDHM_ECOLI      REHLLATGEQLCL-QRGFTGMGLSELLKTAEVPKGSFYHYFRSKEAFG
YXAF_BACSU      REKILHTASRLSQ-LQGYHATGLNQIVKESGAPKGSLYHFFPNGKEEL
BM3R_BACME      QKAIFSASLLLFA-ERGFDATTMPMIAENAKVGAGTIYRYFKNKESLV
YHGD_BACSU      KKLILEAATKSFT-QFGYKATTMDLVAKLANVGKGTIYTFFKNKEELF
YCDC_ECOLI      KKAILSAALDTFS-QFGFHGTRLEQIAELAGVSKTNLLYYFPSKEALL
UIDR_ECOLI      RTRILNAAREIFS-ENGFHSASMKAICKSCAISPGTLYHHFISKEALI
YC55_MYCTU      ADRILDAAERLFT-QRDPASIGMNEIAKAAGCSRATLYRYFDSREALR
YY05_MYCTU      AAAILQAATDLFA-ERGPAATSIRDIAARSKVNHGLVFRHFGTKDQLV
YT12_MYCTU      VARLLQASIDTII-EVGYARASAAVITKRAGVSVGALFRHFETMGDFM
TTK_ECOLI       REEILQSLALMLESSDGSQRITTAKLAASVGVSEAALYRHFPSKTRMF
Q8Y3E9          RVQILQTLASMLEQPRG-EKITTAGLAARLDVSEAALYRHFASKAQMF
Q8ZSI4          QQDVINAAIACLD-QEGEAALGVNRVARELGIKPPAIYKHLDGNLALR
Y893_HAEIN      AEQIFSATDRLMA-REGLNQLSMLKLAKEANVAAGTIYLYFKNKDELL
O33179          AERIYAAATDLIT-RYGLNALDIDKLAREVHCSRATIYRRAGGKAQIR
Y767_MYCTU      FRKVLAAAMATLR-EKSYADLTVRLVAARAKVAPATAYTYFSSKNHLI
O54171          NHRRLLQAVGELAR-EAP-DRLTMQAVASRAEIGPATAYRYYSSVDDV
YF56_MYCTU      ETELLAVTLRLLQ-EHGYDRLTVDAVAASARASKATVYRRWPSKAELV
YDH1_XANAU      TERILDVALELLE-TEGEFGLTMRQVATQADMSLSNVQYYFKSEDLLL
YI16_MYCTU      EAKIVELGRRQLL-DHGAAGLSLRAIARNLGMVSSAVYRYVSSRDELL
YD53_MYCTU      PEDIISGAFELAQ-QVSIDNLSMPLLGKHLGVGVTSIYWYFRKKDDLL
YIJC_ECOLI      RRSLVEAAFSQLSAERSFASLSLREVAREAGIAPTSFYRHFRDVDELG
YM50_MYCTU      GGRILRAAASCVV-DYGVDRVTLAEIARRAGVSRPTVYRRWPDTRSIM
Q9Z562          REDLLEGAKRCLL-EKGFARTTARDIVKESRTNLASIGYHYGSKDVLL
Q9EVJ6          DDEVLEAATVVLK-RCGPIEFTLSGVAKEVGLSRAALIQRFTNRDTLL
YXBF_BACSU      EAQVKEALLALLK-DRDIQHITMKEIAEKAKVSRGSLYLYYEDKFSII
Q99RS1          KIEIQNALVDLMA-EYPFQEISTKMICAYCNINRSTFYDYYKDKFDLL

                ------- α1 -------      -- α2 --   -- α3 --   - α4 →
QacR            KDKILGVAKELFI-KNGYNATTTGEIVKLSESSKGNLYYHFKTKENLF
                |-----+----+-  ---+----+----+----+----+----+----|
                4     10          20        30        40        50
```

FIG. 1. Alignment of 42 members of the TetR family that exemplify the TetR family profile. The blue column indicates α-helix residues involved in DNA contacts in the crystal structure of TetR and QacR. The yellow column indicates turns. The most conserved residues are shaded. Abbreviations are as follows: BACME, *Bacillus megaterium*; BACSU, *Bacillus subtilis*; ECOLI, *Escherichia coli*; HAEIN, *Haemophilus influenzae*; LACLA, *Lactobacillus lactis*; MYCTU, *Mycobacterium tuberculosis*; RHIME, *Rhizobium meliloti*; STRGA, *Streptomyces* sp.; VIBHA, *Vibrio haemophilus*; XANAU, *Xanthomonas* sp.

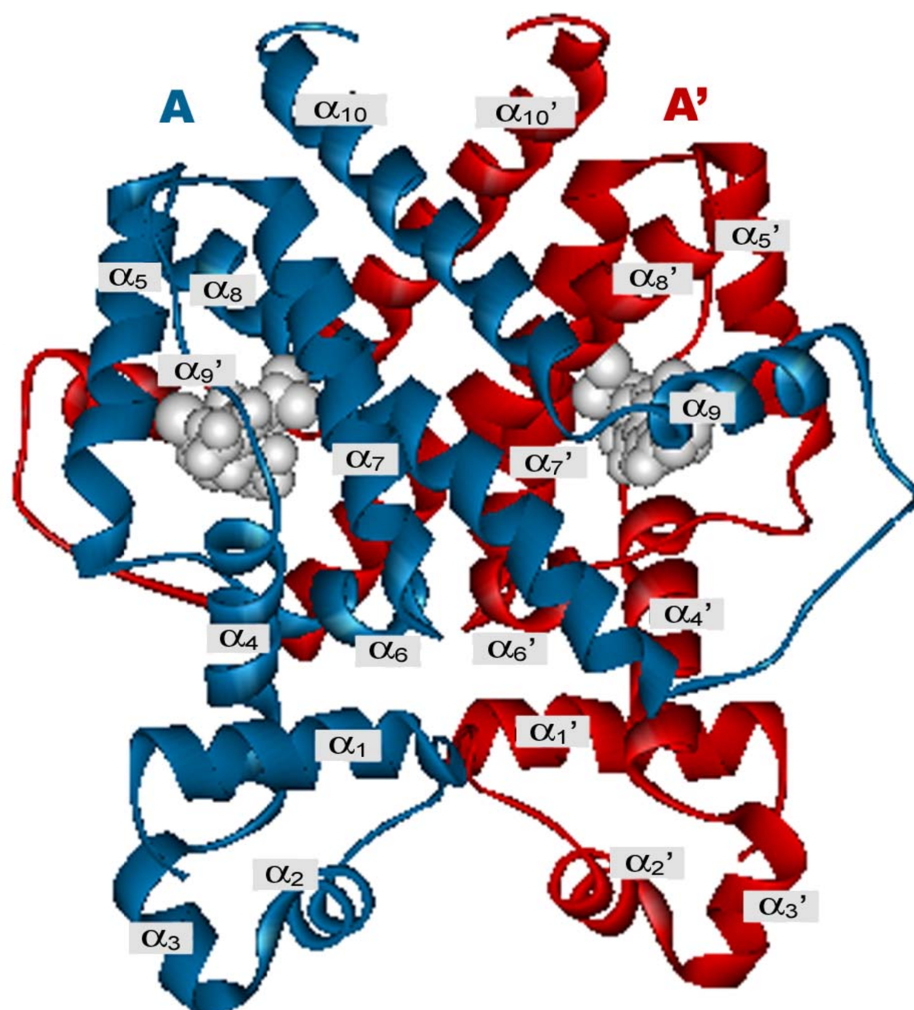


FIG. 2. Ribbon diagram of a TetR homodimer. Monomers are shown in blue or red. Two tetracycline molecules, each bound to a monomer, are shown in grey. α-Helices 2 and 3 in the blue monomer and α2′ and α3′ in the red monomer constitute the shared HTH DNA binding domain. α-Helix 1 and part of helix α-4, together with α-helices 2 and 3, comprise the sequence that best defines the TetR family profile. (Adapted from Hinrichs et al. [150] with permission of the publisher.)

2004) using the pfsearch program available at http://bioweb.pasteur.fr/seqanal/interfaces/pftools.html#pfsearch (46). The program, which proposes a tentative threshold Z-score of 8.5 to consider a protein a member of the TetR family, selected 2,357 proteins as putative members of the TetR family.

To verify the quality of this TetR profile for specificity (false positives) and sensitivity (false negatives), we implemented a new tool called Provalidator which uses Interpro, Swiss-Prot, Prodom, TIGRfam, CoGnitor, NCBI-RPS-BLAST, and PSI-BLAST resources (68, 128, 154, 323, 348, 387, 449). In the first step, we searched for false positives among the 2,353 proteins we assigned to the TetR family. Interpro assigned 2,315 proteins to the TetR family, and these 2,315 were considered true positives. The remaining 38 proteins were analyzed with other resources such as TIGRfam, Prodom, NCBI-RPS-BLAST and PSI-BLAST (128, 449). This allowed us to assign 34 proteins to the TetR family. Three of the false positives (Q89RN6, Q988I6, and Q6N8G8) that we found were protein members of the AraC/XylS family of transcription activators (109, 394). These proteins have two HTH motifs at the C-terminal end, typical of AraC/XylS family members (109, 229). These three proteins were identified as potential TetR members because

one of its HTH is highly similar to the DNA-binding domain in TetR. The fourth false positive is a transposase (Q981E7).

Provalidator detected 15 false negatives (Q742Y2, Q8CJK3, Q73ZY1, Q6D1J7, Q8KU64, Q9A917, Q880T2, Q6D2Z4, Q885G7, Q8PC90, Q9A466, Q9S6C0, Q9ZH26, Q6A626, and Q8G822), which are proteins assigned to the TetR family by INTERPRO but whose Z-score was between 6.407 and 8.487. In summary, the TetR profile with a Z-score threshold of 8.5 identified proteins that were not detected by INTERPRO, and among the 660,992 proteins analyzed, only four false positives were found. These results indicate that the new algorithm is highly effective for the detection of members of the TetR family.

Identification of TetR Family Members in DNA and Protein Databases

Using the profile defined above for the TetR family, we searched for members of this family in the Swiss-Prot and TrEMBL databases and also searched the 196 complete and incomplete microbial genomes available in NCBI (Release December 2004). We detected 73 TetR proteins in Swiss-Prot, 2,277 in TrEMBL, and 2,410 in the translated open reading frame corresponding to 196 microbial genomes. To select non-redundant sequences the set of 4,758 TetR proteins was analyzed using the SEQUNIQ program developed in our laboratory (Molina-Henares et al., unpublished results). This program integrates the set of sequences available in nucleic acid and protein databases. We found 2,353 sequences in the TetR family that surpassed the threshold Z-score of 8.5. The HTH in 2,348 members of the family was located at the N-terminal end of the proteins.

Table 3 shows that members of the TetR family were detected in 144 microbial genomes belonging to 80 genera and 113 species of gram-positive and α-, β-, and γ-proteobacteria, cyanobacteria, and archaea, indicating wide taxonomic distribution. We have found that proteins of the TetR family are encoded both in chromosomes and in plasmids, and the mobility of the latter elements could be a source of the spread of genes in this family via horizontal transfer (147, 383), as is also the case with catabolic genes (77, 160, 236, 410, 426), antimicrobial resistance determinants (20, 100, 124), and 16S rRNA genes (347).

We found that TetR family members are particularly abundant in microbes exposed to environmental changes, such as soil microorganisms (i.e., *Nocardia*, *Streptomyces*, *Bradyrhizobium*, *Mesorhizobium*, *Pseudomonas*, *Bacillus*, and *Ralstonia* spp.); plant and animal pathogens (i.e., *Agrobacterium*, *Brucella*, *Escherichia coli*, *Bordetella*, *Mycobacterium*, and *Salmonella* spp.), extremophiles (i.e., *Deinococcus*), and methanogenic bacteria such as *Methanosarcina acetivorans*. In contrast, TetR family members do not appear in intracellular pathogens such as chlamydias, mycoplasmas, and endosymbionts such as *Buchera*, in agreement with their life style in nonchanging environments (52). However, it should be noted that Dugan et al. (80) recently found that *Chlamydia suis* can acquire tetracycline resistance via horizontal gene transfer of genomic islands bearing the *tet* genes.

As a general collorarium, we can say that it seems that proteins of the TetR family are involved in the adaptation to complex and changing environments. This in turn correlates with the fact that many members of the TetR family are found among microbes with abundant extracytoplasmic function sigma factors (52, 227, 236, 277, 444).

PROTEINS WITH KNOWN THREE-DIMENSIONAL STRUCTURES

The high degree of primary sequence identity in the stretch that defines the HTH region of the TetR profile probably reflects a common three-dimensional structure in this domain in members of the family. This is supported by the almost identical three-dimensional structure of the HTH of TetR, QacR, CprB, and EthR, as deduced from the superimposition of these regions, and the high degree of sequence conservation in the alignment (79, 264, 349). As in other families of transcriptional regulators, no sequence conservation was found outside the HTH domain, which probably reflects differences in the kind of signal sensed by different regulators of the family, i.e., antibiotics with dissimilar structures, barbiturates, homoserine lactones, organic solvents, and choline (see Table 2). Nonetheless, some striking global structural conservation in the three-dimensional structure was found.

In addition, given that all members of the family whose function is known are repressors, they probably function in a similar way. Binding of an inducer molecule to the nonconserved domain of a TetR family member probably causes conformational changes in the conserved DNA-binding region that result in release of the repressor from the operator and thus allow transcription from the cognate promoter. To gain insights into the mechanisms of action of the TetR family members, we analyzed in detail the three-dimensional structure of the four members of the family, TetR, QacR, CprB, and EthR, whose crystal structures have been obtained (150, 264, 286–289, 349–351), in order to identify common and differential features of the TetR family members.

TetR Regulator

Tetracycline resistance and the role of the transcriptional regulator TetR. Tetracyclines are among the most commonly used broad-spectrum antibiotics (209, 210). They act by binding to the small ribosomal subunit, thereby interrupting polypeptide chain elongation by an unknown mechanism. Many gram-negative bacteria have developed mechanisms of resistance against this antibiotic. The most frequent mechanism involves a membrane-associated protein (TetA) that exports the antibiotic out of the bacterial cell before it inhibits polypeptide elongation (169, 211, 389, 434, 435, 453).

Adjacent to *tetA* and divergently oriented is *tetR* (112), whose gene product tightly controls expression of both *tetA* and *tetR* (148, 150). The intergenic region between the *tetR* and *tetA* genes contains two identical operators separated by 11 bp. TetR binds to these operators and thus prevents transcription from both promoters (Fig. 3) and (288). In all TetR crystal structures elucidated to date (PDB identifiers: 2TCT; 2TRT; 1A6I; 1BJO; 1BJY; 1BJZ; 1ORK; and 1RP1), this repressor appears as a homodimer (29, 30, 159, 183, 287–289, 366). The TetR homodimer binds to the operator (Fig. 3). Each 15-bp operator shows an internal palindromic symmetry with an extra

TABLE 3. Distribution of TetR proteins in microbes

Microorganism	Genome size (Mbp)	No. of members
Nocardia farcinica IFM 10152	6.21	151
Streptomyces coelicolor A3(2)	9.05	150
Streptomyces avermitilis MA-4680	9.12	116
Mycobacterium avium subsp. *paratuberculosis* k10	4.83	108
Agrobacterium tumefaciens C58	11.35	61
Bradyrhizobium japonicum USDA 110	9.11	59
Mycobacterium bovis AF2122/97	4.35	51
Mycobacterium tuberculosis CDC1551	4.40	51
Mycobacterium tuberculosis H37Rv	4.41	51
Bacillus licheniformis ATCC 14580	8.44	48
Mesorhizobium loti MAFF303099	7.60	47
Rhodopseudomonas palustris CGA009	5.46	40
Pseudomonas aeruginosa PAO1	6.26	38
Bacillus cereus ATCC 10987	5.22	36
Bacillus anthracis 'Ames Ancestor'	5.50	32
Bacillus anthracis A2012	5.37	32
Bacillus anthracis Sterne	5.23	31
Bacillus anthracis Ames	5.23	30
Bacillus cereus ZK	5.30	30
Bacillus cereus ATCC 14579	5.43	29
Bordetella bronchiseptica RB50	5.34	28
Pseudomonas syringae pv. tomato DC3000	6.40	28
Sinorhizobium meliloti 1021	6.69	28
Bacillus thuringiensis serovar konkukian 97-27	5.24	27
Clostridium acetobutylicum ATCC 824	4.13	27
Caulobacter crescentus CB15	4.02	26
Lactobacillus plantarum WCFS1	3.31	26
Pseudomonas putida KT2440	6.18	25
Burkholderia pseudomallei K96243	7.25	24
Ralstonia solanacearum GMI1000	5.81	24
Photobacterium profundum SS9	6.40	23
Oceanobacillus iheyensis HTE831	3.63	22
Bordetella parapertussis 12822	4.77	21
Burkholderia mallei ATCC 23344	5.84	21
Bacillus halodurans C-125	4.20	20
Bacillus subtilis subsp. *subtilis* 168	4.21	20
Acinetobacter sp. strain ADP1	3.60	19
Bordetella pertussis Tohama I	4.09	17
Chromobacterium violaceum ATCC 12472	4.75	17
Shewanella oneidensis MR-1	5.13	17
Vibrio vulnificus CMCP6	5.13	17
Escherichia coli CFT073	5.23	16
Gloeobacter violaceus PCC 7421	4.66	16
Methanosarcina acetivorans C2A	5.75	16
Streptococcus mutans UA159	2.03	16
Vibrio parahaemolyticus RIMD 2210633	5.17	16
Vibrio vulnificus YJ016	5.26	16
Xanthomonas axonopodis pv. citri 306	5.18	16
Corynebacterium glutamicum ATCC 13032	3.31	15
Deinococcus radiodurans R1	3.28	15
Erwinia carotovora subsp. *atroseptica* SCRI1043	5.06	15
Xanthomonas campestris pv. campestris ATCC 33913	5.08	15
Escherichia coli O157:H7	5.59	14
Corynebacterium efficiens YS-314	3.15	13
Escherichia coli O157:H7 EDL933	5.53	13
Lactococcus lactis subsp. *lactis* I11403	2.37	13
Leifsonia xyli subsp. *xyli* CTCB07	2.58	13
Listeria monocytogenes 4b F2365	2.91	13
Salmonella enterica subsp. *enterica* serovar Typhi CT18	5.13	13
Salmonella typhimurium LT2	4.95	13
Treponema denticola ATCC 35405	2.84	13
Corynebacterium diphtheriae NCTC 13129	2.49	12
Escherichia coli K12	4.64	12
Listeria innocua Clip11262	3.01	12
Listeria monocytogenes EGD-e	2.94	12
Salmonella enterica subsp. *enterica* serovar Typhi Ty2	4.79	12
Shigella flexneri 2a 301	4.61	12
Vibrio cholerae O1 biovar eltor N16961	4.03	12
Nostoc sp. strain PCC 7120	7.21	11
Enterococcus faecalis V583	3.22	10
Mycobacterium leprae TN	3.27	10
Shigella flexneri 2a 2457T	4.60	10
Geobacter sulfurreducens PCA	3.81	9
Leptospira interrogans serovar Copenhageni Fiocruz L1-130	4.63	9
Leptospira interrogans serovar Lai 56601	4.69	9
Propionibacterium acnes KPA171202	2.56	9
Staphylococcus aureus subsp. *aureus* MRSA252	2.90	9
Shigella flexneri 2a 301	4.61	12
Vibrio cholerae O1 biovar eltor N16961	4.03	12
Nostoc sp. strain PCC 7120	7.21	11
Enterococcus faecalis V583	3.22	10
Mycobacterium leprae TN	3.27	10
Shigella flexneri 2a 2457T	4.60	10
Geobacter sulfurreducens PCA	3.81	9
Leptospira interrogans serovar Copenhageni Fiocruz L1-130	4.63	9
Leptospira interrogans serovar Lai 56601	4.69	9
Propionibacterium acnes KPA171202	2.56	9
Staphylococcus aureus subsp. *aureus* MRSA252	2.90	9
Bifidobacterium longum NCC2705	2.26	8
Brucella melitensis 16M	3.29	8
Brucella suis 1330	3.32	8
Photorhabdus luminescens subsp. *laumondii* TTO1	5.69	8
Staphylococcus aureus subsp. *aureus* Mu50	2.90	8
Symbiobacterium thermophilum IAM 14863	3.57	8
Yersinia pestis biovar Medievalis 91001	4.80	8
Yersinia pseudotuberculosis IP 32953	4.84	8
Bacteroides fragilis YCH46	5.31	7
Bacteroides thetaiotaomicron VPI-5482	6.26	7
Bdellovibrio bacteriovorus HD100	3.78	7
Desulfovibrio vulgaris subsp. *vulgaris* Hildenborough	3.77	7
Fusobacterium nucleatum subsp. *nucleatum* ATCC 25586	2.17	7
Staphylococcus aureus subsp. *aureus* MSSA476	2.80	7
Staphylococcus aureus subsp. *aureus* MW2	2.82	7
Staphylococcus aureus subsp. *aureus* N315	2.84	7
Yersinia pestis CO92	4.83	7
Desulfotalea psychrophila LSv54	3.66	6
Lactobacillus johnsonii NCC 533	1.99	6
Mannheimia succiniciproducens MBEL55E	2.31	6
Rhodopirellula baltica SH 1	7.15	6
Streptococcus agalactiae NEM316	2.21	6
Yersinia pestis KIM	4.60	6
Aquifex aeolicus VF5	1.59	5
Methylococcus capsulatus Bath	3.30	5
Streptococcus agalactiae 2603V/R	2.16	5
Streptococcus pyogenes MGAS10394	1.90	5
Streptococcus pyogenes MGAS8232	1.90	5
Clostridium perfringens 13	3.09	4
Methanococcus maripaludis S2	1.66	4
Staphylococcus epidermidis ATCC 12228	2.50	4
Streptococcus pyogenes M1 GAS	1.85	4
Streptococcus pyogenes MGAS315	1.90	4
Streptococcus pyogenes SSI-1	1.89	4
Thermus thermophilus HB27	2.13	4
Clostridium tetani E88	2.80	3
Haemophilus influenzae Rd KW20	1.83	3
Methanosarcina mazei Go1	4.10	3
Nitrosomonas europaea ATCC 19718	2.81	3
Pasteurella multocida subsp. *multocida* Pm70	2.26	3
Porphyromonas gingivalis W83	2.34	3
Streptococcus pneumoniae R6	2.04	3
Streptococcus pneumoniae TIGR4	2.16	3
Synechocystis sp. strain PCC 6803	3.57	3
Thermoanaerobacter tengcongensis	2.69	3
Wolinella succinogenes DSM 1740	2.11	3
Haemophilus ducreyi 35000HP	1.70	2
Halobacterium salinarum NRC-1	2.57	2
Legionella pneumophila str. Lens	3.41	2
Legionella pneumophila str. Paris	3.64	2
Legionella pneumophila subsp. *pneumophila* Philadelphia 1	3.40	2
Methanothermobacter thermautotrophicus Delta H	1.75	2
Neisseria meningitidis MC58	2.27	2
Neisseria meningitidis Z2491	2.18	2
Thermotoga maritima MSB8	1.86	2
Xylella fastidiosa 9a5c	2.73	2
Archaeoglobus fulgidus DSM 4304	2.18	1
Campylobacter jejuni subsp. *jejuni* NCTC 11168	1.64	1
Coxiella burnetii RSA 493	2.00	1
Helicobacter hepaticus ATCC 51449	1.80	1
Mycoplasma penetrans HF-2	1.36	1
Picrophilus torridus DSM 9790	1.55	1
Pyrococcus abyssi GE5	1.77	1
Sulfolobus solfataricus P2	2.99	1
Sulfolobus tokodaii 7	2.69	1
Ureaplasma parvum serovar 3 ATCC 700970	0.75	1

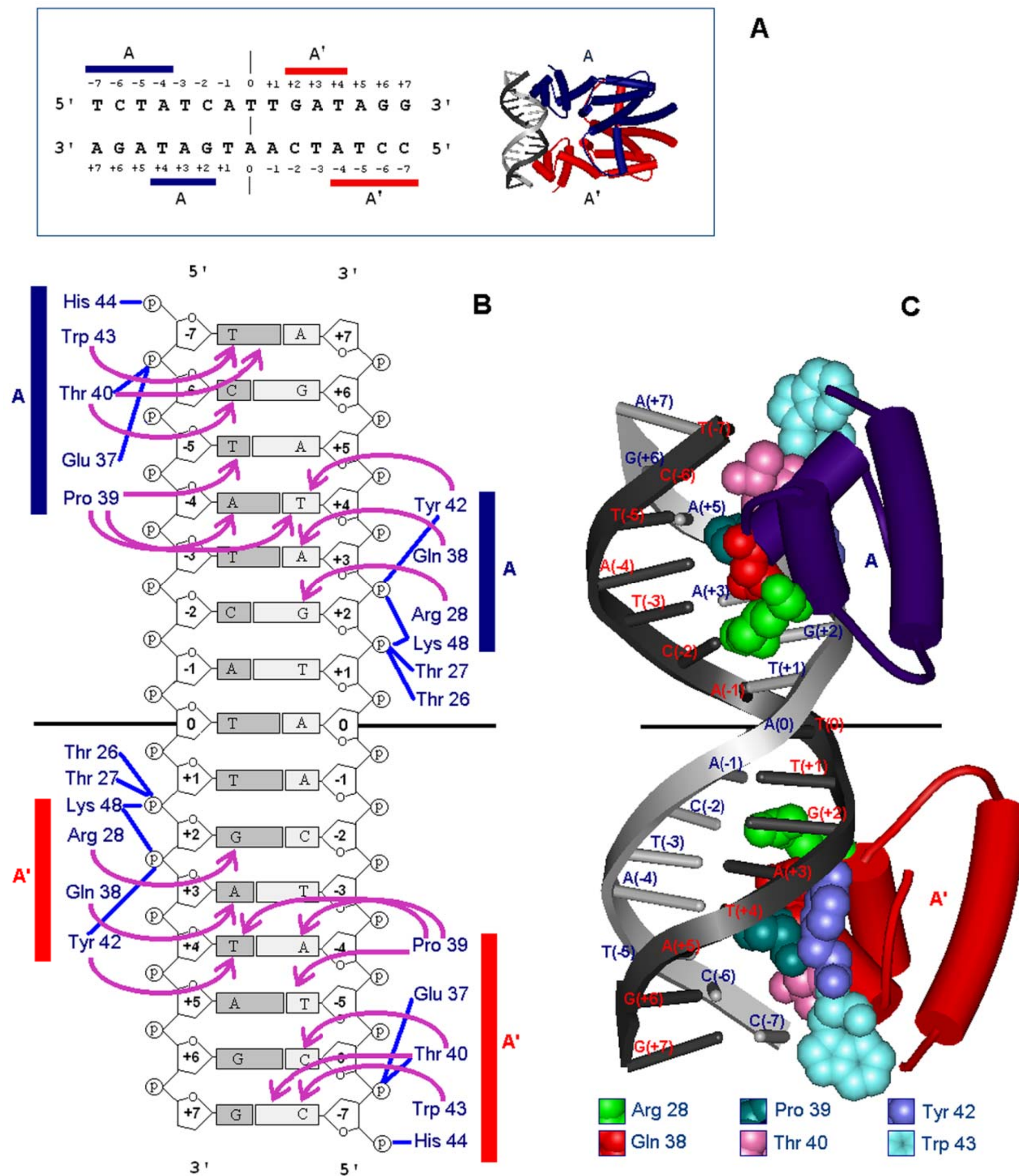


FIG. 3. Binding of TetR to its operator site. A) *tetR* operator and contact regions. The *tetR* operator is a palindromic sequence. Horizontal bars show nucleotides contacted by each monomer of the TetR dimer. B) Interaction of TetR residues with specific nucleotides (arrows) and phosphate backbone (blue lines) in the operator region. The amino acids involved in DNA binding extend from residues 27 to 48. Contacts established with the operator were confirmed by footprint assays, by analysis of TetR mutants, and by crystallographic studies (29, 30, 159, 266, 366). C) Representation of each homodimer bound to the *tet* operator in a double-helix representation. (Adapted from Orth et al. [288] with permission of the publisher.)

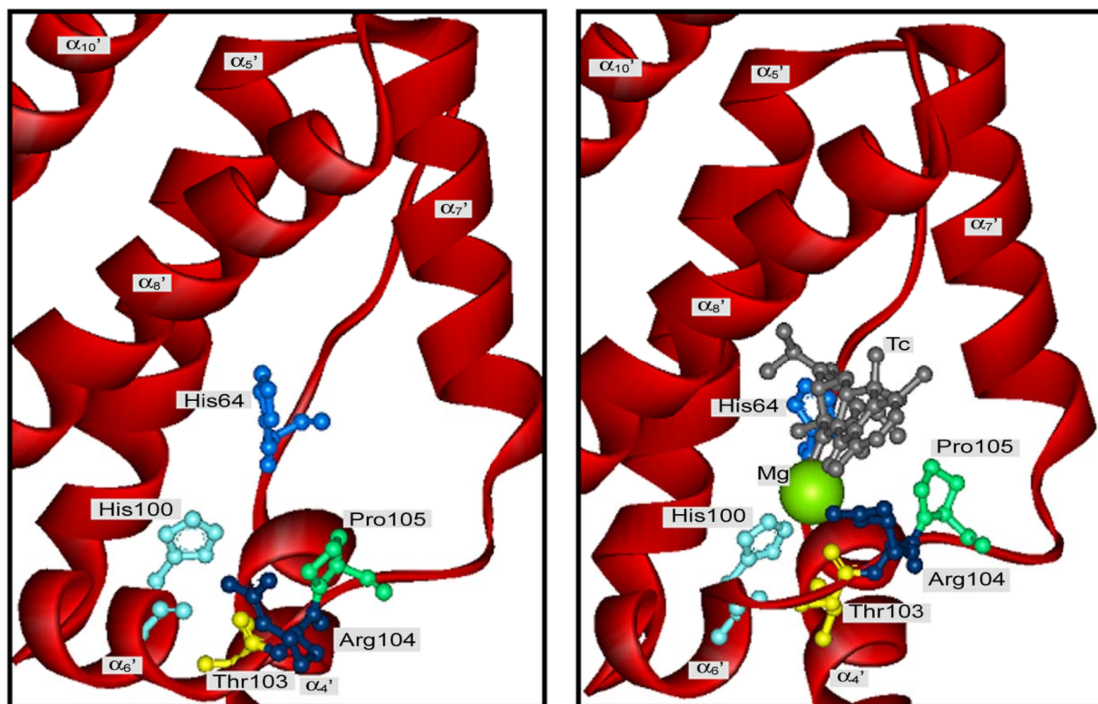


FIG. 4. Representation of the TetR cavity involved in the binding of tetracycline. Left) In the absence of tetracycline. Right) In the presence of tetracycline. The green ball represents the Mg^{2+} ion. Specific interactions are not drawn for the sake of clarity but are described in the text. (Adapted from Orth et al. [287, 289] and Kisker et al. [183] with permission of the publishers.)

central base pair (Fig. 3A). The operator sequences overlap with promoters for *tetA* and *tetR*, thereby blocking the expression of both genes. When tetracycline complexed with Mg^{2+} binds to TetR (166, 384), a conformational change takes place that renders the TetR protein unable to bind DNA. As a consequence, TetR and TetA are expressed (286).

The TetR homodimer is constituted by two identical monomers that fold into 10 α-helices with connecting turns and loops (Fig. 2). The three-dimensional structure of the TetR monomer is stabilized mainly by hydrophobic helix-to-helix contacts. The global structure of the TetR homodimer can be divided into two DNA-binding domains at the N-terminal end of each monomer, and a regulatory core domain involved in dimerization and ligand binding (150, 286–289). The DNA-binding domains are constituted by helices α1, α2, and α3 and their symmetric helices α1′, α2′, and α3′ (a prime denotes the second monomer). Helices α4 and α4′ connect these domains with the regulatory core domain composed of helices α5 to α10 and their symmetric counterparts α5′ and α10′ (150, 287, 289). The regulatory domain is responsible for dimerization and contains, for each monomer, a binding pocket that accommodates tetracycline in the presence of a divalent cation. Helices α5, α8, and α10 and their counterparts α5′, α8′, and α10′ constitute the scaffold of the core domain, and their structure is the most conserved in both TetR conformations (150, 287–289).

The tetracycline-binding pocket is identical in both monomers. The cavity to which the $[TcMg]^+$ complex binds is depicted in Fig. 4 (286, 287, 289). The entrance of this cavity is controlled by α9′ and the C-terminal end of α8′ and the loop that connects both, while the exit is closed by loop 4-5 (287–289). When $[TcMg]^+$ enters the tunnel, its A ring makes contacts with loop 4-5, and the interaction with the effector triggers a cascade of conformational changes. The contacts that His100 and Thr103, both in α6, establish with the magnesium ion of the complex displace α6, which undergoes a conformational change in its C terminus to form a β-turn (Fig. 4). The 6-7 loop is also pushed near the inducer, so that Arg104 and Pro105 interact with tetracycline. Translation of α6 forces α4 to move in the same direction due to van der Waals contacts. His64 of α4, anchored to α5 and to tetracycline, acts as a pivot point, and α4 moves like a pendulum. As a consequence of the rotation of α4 and α4′, recognition helices α3 and α3′ move further apart, and the DNA contacts are disrupted (Fig. 5) (286, 287, 289). Tetracycline is impeded from freeing the binding cavity, and TetR cannot bind its target DNA again. It should be noted that residues outside the binding cavity can influence affinity for tetracycline, as revealed by Kamionka et al. (168), who isolated a double mutant (G96E, L205S) with reduced affinity for the antibiotic.

The on/off switch mechanism used by TetR to respond to specific signals may be used similarly in other TetR family members.

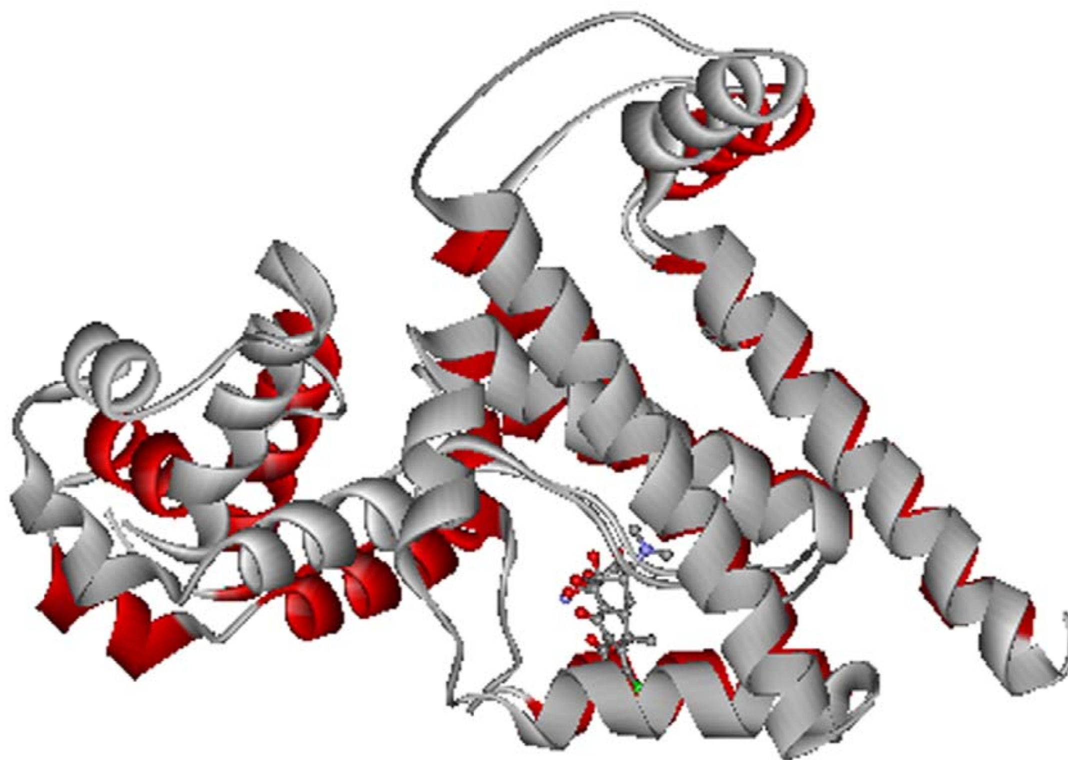


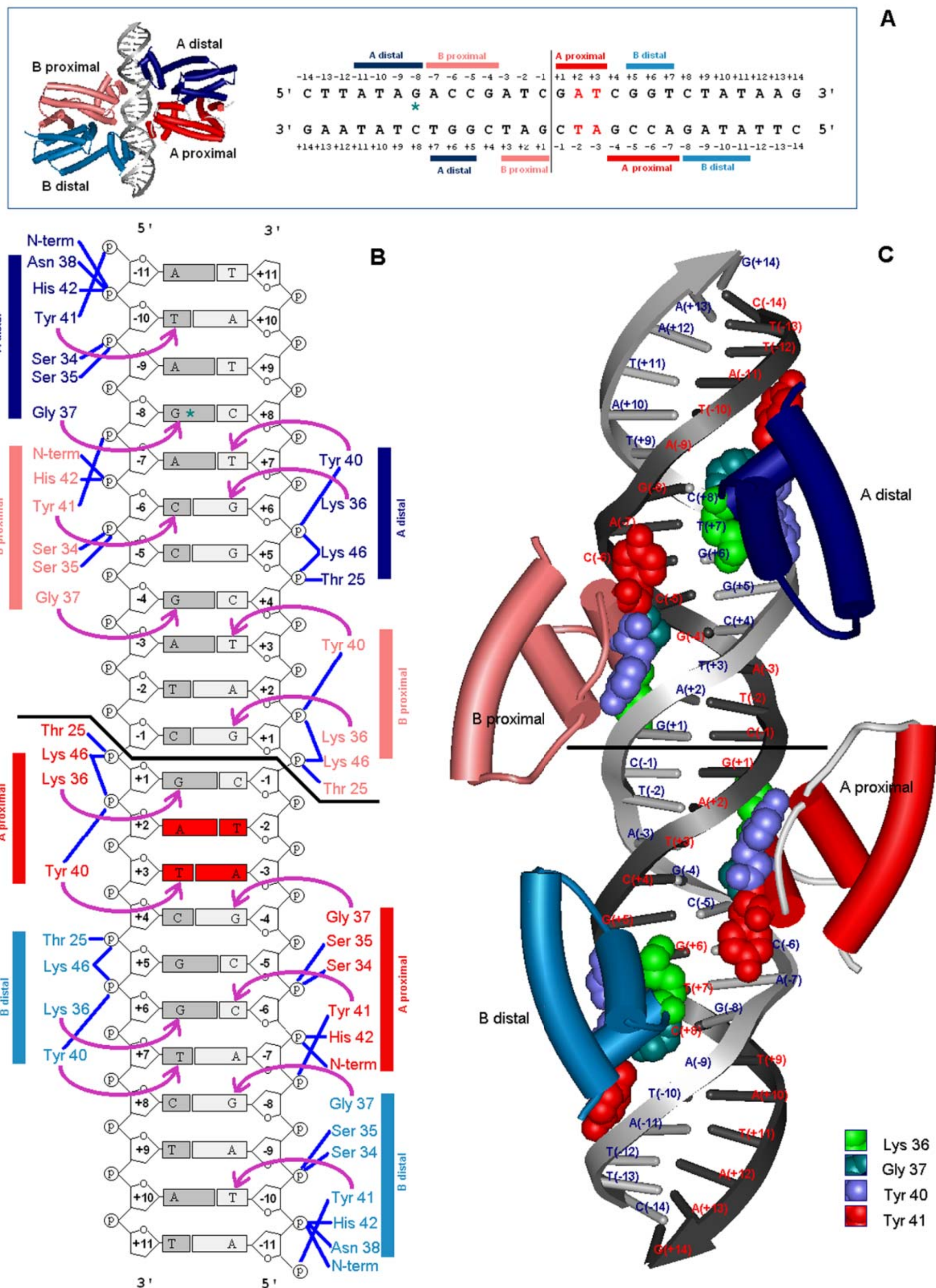
FIG. 5. Docking of a TetR monomer without (grey) and with (red) tetracycline. Note the alterations induced in the α_1-α_3 region involved in binding to the target operators. The increase in distance between α3 and α3′ with tetracycline results in the inability of TetR to maintain the specific interactions shown in Fig. 3, and therefore the repressor is released. (Adapted from Orth et al. [288] with permission of the publisher.)

TetR DNA-binding domain: a symmetric TetR dimer binds a palindromic operator. Cocrystallization of TetR with its operator DNA established that the TetR homodimer binds perpendicularly to the longitudinal DNA axis (Fig. 3A). Two adjacent DNA major groove regions covering a 6-base-pair area on both strands are involved in the almost perfect docking with the two TetR-interacting domains (Fig. 3A and 3B) (288). No water molecules were found at the TetR-DNA interface, where the crucial interactions are hydrophobic (288).

The interactions of each HTH domain with the operator DNA are summarized in Fig. 3A and 3B. The TetR monomer A binds the main strand from positions −4 to −7 while contacting the complementary strand from operator positions +4 to +2, and the symmetric monomer A' binds the main strand from positions +2 to +4 and the complementary strand from positions −4 to −7 (Fig. 3A and 3B).

Crystallographic analysis revealed that helix α3 (from Gln38 to His44) is the main element responsible for sequence-specific recognition, since all residues in this helix contribute to it, except for Leu41, which is part of the hydrophobic core stabilizing the α1, α2, and α3 helix bundle. Thr40 residue in monomer A establishes direct contacts with operator base pairs T(−7) and C(−6) in the main DNA strand (Fig. 3A and 3B). Trp43 interacts with T(−7) as well. Pro39 interacts with both strands at bases T(−5) and A(−4) of the main strand and T(+4) of the complementary strand. In the rest of the operator half site, the α3 helix of monomer A interacts with the complementary strand, Tyr42 contacting with T(+4) and Gln38 with A(+3). Helix α2 supplies an additional specific contact with the complementary strand, namely, Arg28 contacts G(+2).

Although the TetR DNA binding domain maintains its structure thanks to a hydrophobic core formed by residues from the α1, α2, and α3 bundle (288), interactions with DNA lead to changes in the TetR DNA binding domain. One such change is that α3 forms a 3_{10}–helical turn at the N-terminal end as a result of complex DNA contacts. The H-bonds between Arg28-G(+2), and Gln38-A(+3) increase the separation between base pairs 1 and 2 from 3.4 Å to 3.9 Å (288). The two phosphate groups accompanying the G at position +2 establish H-bonds with side chains of Thr26, Thr27, Tyr42, and Lys48, and with the amino groups of the main chain of Thr27 and Lys48 (Fig. 3B). These contacts draw DNA closer to TetR near G(+2). Although the DNA is kinked away from TetR at position +2 in both operator strands, bending toward TetR in


A
B
C
A distal
B proximal
A proximal
B distal
5' C T T A T A G A C C G A T C G A T C G G T C T A T A A G 3'
3' G A A T A T C T G G C T A G C T A G C C A G A T A T T C 5'
N-term
Asn 38
His 42
Tyr 41
Ser 34
Ser 35
Gly 37
Tyr 40
Lys 36
Lys 46
Thr 25
Lys 36
Gly 37
Tyr 40
Tyr 41

the area corresponding to positions +3 to +6 compensates for the DNA deviation. Crystallographic studies revealed that Lys48 located in α4, outside the HTH motif, also established contacts with the target DNA region (Fig. 3B). This lysine is relatively well conserved among TetR family members, and we are tempted to suggest that this residue plays an equivalent role in other proteins of the TetR family.

QacR Regulator

Two QacR dimers bind the operator to repress the *qacA* multidrug transporter gene. QacA confers resistance to monovalent and bivalent cationic lipophilic antiseptics and disinfectants such as *q*uaternary *a*mmonium *c*ompounds (hence the name Qac) (10, 11, 44, 239). The *qac* locus consists of the *qacA* gene and the divergently transcribed *qacR* gene, which are borne on a plasmid (119). In the absence of drug, the 188-residue QacR protein represses transcription of the *qacA* multidrug transporter gene by binding two nested palindromes located downstream from the *qacA* promoter and overlapping its transcription start site (119, 300). Therefore, QacR seems to repress transcription by hindering the transition of the RNA polymerase-promoter complex into a productively transcribing state rather than by blocking RNA polymerase binding.

The three-dimensional structure of QacR (PDB identifiers 1JTX, 1JTG, 1JTY, 1JUM, 1JUP, 1JUS, and 1JTO) revealed that it is an all-helical protein which contains a DNA-binding HTH motif embedded within an N-terminal three-helix bundle and a second domain involved in drug binding and dimerization (350, 351). It should be noted that unlike TetR, two QacR dimers, rather than one, bind the operator site (339, 340) (Fig. 6).

The monomers of each dimer have been called proximal and distal to refer to their positions with respect to the center of symmetry of the palindromic operator (Fig. 6A and 6B). It was shown that the operator to which one dimer is bound is symmetric and partially overlaps that bound by the other dimer (351) (Fig. 6 and 7). The existence within the same fragment of DNA sequence of two overlapping partial palindromes with identical symmetric bases is therefore surprising (Fig. 6). In this sense the palindromic sequences recognized by QacR are equivalent to those described for the TetR interface except for the spacer sequence length, 3 bp for TetR versus 4 bp for QacR, supporting the hypothesis that interactions of other members of the family with their target sequences may be similar, independent of the number of dimers involved.

The α3 helix of QacR A distal and B distal monomers establish the most extensive specific interactions with the operator (351). The Tyr41 residue of the A distal monomer (Fig. 6B) establishes hydrophobic contacts with base T(−10) of the DNA main strand as well as with the phosphate at position −11 in the main strand, while Tyr40 contacts T(+7) (Fig. 6B). In addition, tight docking with DNA is facilitated by specific hydrogen bonds between Lys36 and base G(+6) in the complementary strand, and between Gly37 and base G(−8) in the main strand. Gly37 is important in repression because nucleotide G(−8) is the transcription start site for the *qacA* gene. Monomers A and B proximal also establish a series of critical interactions. For instance, Tyr41 of B proximal contacts the C(−6) base in the main strand, whereas Tyr40 contacts base T(+3) and phosphate (+2) in the complementary strand (351). Gly37 in the A proximal monomer contacts G(−4) in the complementary strand, whereas Lys36 contacts G(+1) in the main strand. A number of residues in α2, loop α2-α3, α3 and the positive dipole of the α1 (N terminus) also interact with the phosphate backbone of both DNA strands (351).

Figure 6C shows how each dimer engages the DNA major groove in a face almost opposite to the other dimer, forming an angle between the two dimer axes of less than 180° (Fig. 7). Studies of QacR binding to DNA have indicated that the two dimers bind DNA cooperatively (120, 121, 351). Analysis of the three-dimensional structure suggested that such cooperativity does not arise from protein-protein interactions, as the closest approach of the dimers is 5.0 Å. Rather, binding cooperativity appears to be mediated through conversion of the DNA structure from a B-DNA conformation to the high-affinity undertwisted configuration observed in the crystal structure. Conversion of the DNA conformation is necessary because the optimal distance between each of the HTH motifs of the QacR dimer is 37 Å. This requires expansion of the 34-Å distance between successive major groove regions on one edge of the canonical B-DNA. It has been suggested that binding of the first QacR dimer forces this energetically unfavorable conformational change, which in turn produces an optimal DNA conformation for the easy binding of the second dimer (351). Experimental data reported by Grkovic et al. (121, 122) suggested that the two dimers must bind simultaneously and cooperatively to the operator in order to maintain the DNA deformation detected in the crystal.

Schumacher and Brennan (349) noticed that TetR and QacR achieve the same degree of specificity in DNA binding through different mechanisms. They noted that TetR, recruits Arg28, located outside its recognition helix, to make a base pair-specific contact (288), whereas QacR does not employ residues outside α3 to ensure DNA binding specificity. They also noted that TetR kinks its binding site and induces a 17° bend towards the protein to optimize the position of its HTH motifs for specific base interactions within each DNA half site; whereas QacR widens the major groove of the entire IR1 binding site smoothly and bends its DNA site by only 3°. These distinctions are reflected in the different HTH center-to-center distances observed in QacR (37 Å). Thus, an important lesson derived from comparisons of the QacR-DNA and TetR-DNA structures is that even structurally homologous proteins of the same family that share a similar function, i.e., repression, can utilize slightly different mechanisms of action.

FIG. 6. Binding of QacR to its operator site. A) Interaction of QacR with the *qac* operator. B) Contacts established by residues at α-helix 3 of QacR homodimers A and B with specific nucleotides (arrows) and phosphate backbone (blue lines) in the synthetic operator used for QacR-DNA cocrystal (349, 350). C) Representation of the two QacR homodimers bound to the *qac* operator in a double-helix representation. (Adapted from Schumacher et al. [351] with permission of the publisher.)

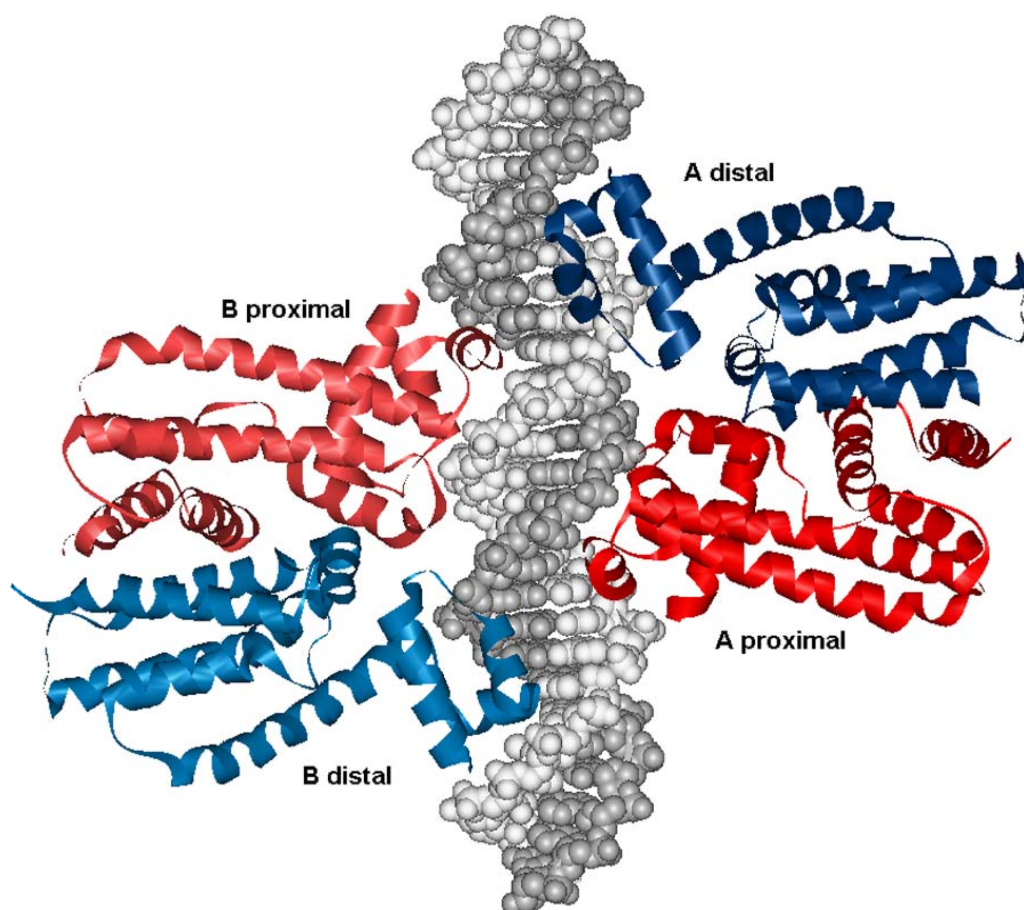


FIG. 7. Ribbon representation of the two QacR homodimers bound to target DNA in a double-helix representation (A) and details of the contacts established by α-helix 3 of monomers of different homodimers when recognizing overlapping sites (B). (Adapted from Schumacher et al. [351] with permission of the publisher.)

QacR as a model for multidrug recognition. QacR is released from the *qacA* operator by its interaction with a number of cationic lipophilic drugs such as rhodamine 6G, crystal violet, and ethidium (119). More recently, Grkovic et al. (122) showed that effector recognition of QacR can be extended to several bivalent cationic dyes and plant alkaloids. In spite of the existence of two binding pockets, only one drug molecule is bound by each homodimer, as determined by equilibrium dialysis studies and isothermal titration calorimetry for the QacR-R6G complex (350). The QacR crystal bound to different drugs revealed another remarkable finding: the presence of an expansive and multifaceted drug-binding pocket with a volume of 1,100 $Å^3$, so that different drugs partially overlap different subpockets (349, 351). A similar cavity able to bind multiple drugs was reported by Yu et al. (445, 446) for the AcrB multidrug transporter.

Crystallographic studies by Schumacher et al. (350) and Murray et al. (261) have demonstrated that multidrug recognition mediated by the QacR dimer is a rather simple process that, contrary to expectations, does not require sophisticated molecular mechanisms. Indeed, the drug binding domain of QacR consists of six α-helices (PDB identifiers: 1JTX, 1JT6, 1JTY, 1JUP, 1JUS, 1JTO, 1RKW, and 1RPW). Entry to the mostly buried drug-binding pocket is through a small opening formed by the divergence of helices α6, α7, α8, and α8′. The stoichiometry of one drug molecule for two QacR subunits led to this asymmetric induction process, in which the drug-bound monomer undergoes a major structural change. Comparison of the drug-bound structure with the DNA-bound structure reveals that drug binding triggers a coil-to-helix transition of residues 89 to 93, which extends helix α5 by a turn. This transition removes the drug surrogates Tyr92 and Tyr93 from the hydrophobic core of the protein. Expulsion of these tyrosines also leads to the relocation of nearby helix α6 and its

tethered DNA-binding domain. The result of this structural transition is a 9-Å translation and a 37° rotation of the DNA-binding domain, effectively rendering the QacR dimer unable to bind its target DNA.

Three-Dimensional Structure of CprB

The gram-positive bacterial genus *Streptomyces* uses γ-butyrolactones as autoregulators or microbial hormones, together with their specific receptors (γ-butyrolactone receptors), to control morphological differentiation, antibiotic production, or both (150, 151). The most representative of the γ-butyrolactone autoregulatory factors is 2-isocapryloyl-3R-hydroxymethyl-γ-butyrolactone, known as A-factor, which is essential for aerial mycelium formation, streptomycin production, streptomycin resistance, and yellow pigment production (133, 134, 155) in *Streptomyces griseus*. However, the A-factor receptor protein, known as ArpA, has proved to be difficult to purify. In contrast, the CprB protein from *Streptomyces coelicolor* A3(2), which is 30% identical to ArpA (284), has been purified and crystallized (264), although the ligand for CprB is still unknown. Nonetheless, CprB binds the same nucleotide sequence as does ArpA (375) and indeed CprB also serves as a negative regulator for both secondary metabolism and morphogenesis in *S. coelicolor*, as ArpA does in *S. griseus* (264, 284).

The CprB dimer is omega shaped, and the two subunits in the dimer are related by a pseudo-twofold axis. Each monomer of CprB is composed of 10 α-helices and has two domains: a DNA-binding domain (residues 1 to 52) and a regulatory domain (residues 77 to 215). The three-dimensional structure of CprB is essentially similar to that of QacR bound to DNA except for the lack of α10 (350, 351). In addition, the DNA-binding domains of the two proteins are very similar, so much so that the two DNA-binding domains can be superimposed with an rms deviation of 1.48 Å for 71 Cα atoms (264). Although no information on CprB-operator DNA is available, the high degree of sequence conservation allowed the authors to predict that the core of the DNA-binding domain is composed of Ile14, Ile15, Ala18, Phe22, Leu32, Ile35, Leu46, and Phe50.

It has been suggested that a CprB dimer binds to its target DNA as found in the TetR–DNA complex (150, 287, 288). This is because structure-based amino acid sequence alignment shows that at the amino acid sequence level the DNA-binding domains of CprB and TetR are highly identical. This suggests that there is an evolutionary relationship between the DNA-binding domains of the two proteins. The regulatory domain of CprB is composed of six α-helices (helices α5 to α10) (264), which can also be superimposed on the corresponding domain of TetR (286, 287, 289) (PDB code 1JT0).

EthR Structure

Ethionamide has been used for more than 30 years as a second-line chemotherapeutic treatment in tuberculosis patients who have developed resistance to first-line drugs such as isoniazid and rifampin. Activation of the prodrug ethionamide is regulated by the Baeyer–Villiger monooxygenase EthA and the TetR family repressor EthR, whose open reading frames are separated by 75 bp in the genome of *Mycobacterium tuberculosis*. EthR has been shown to repress transcription of the activator *ethA* gene by binding to the intergenic region and contributing to ethionamide resistance.

The expression of *ethA* is regulated by EthR in *M. tuberculosis*. Overexpression of *ethR* leads to ethionamide resistance, whereas chromosomal inactivation of *ethR* promotes ethionamide hypersensitivity (28). EthR was found to bind directly and specifically to DNA sequences corresponding to the *ethRA* intergenic region (28, 90). The large EthR operator, which comprises 55 bp in comparison with the 15-bp operators recognized by most other family members, is organized as a putative highly degenerated palindrome containing pairs of overlapping inverted and tandem repeat sequences (90). In the absence of DNA, EthR forms a homodimer in solution, and surface plasmon resonance measurements suggest that EthR octamerizes when bound to DNA (90).

The EthR monomer is an all-helical, two-domain molecule (79). The N-terminal domain comprises helices 1 to 3, with helices 2 and 3 forming the HTH DNA-binding motif seen in other TetR family protein structures. The larger C-terminal domain, which in QacR and TetR has been dubbed the drug-binding domain, consists of helices 4 to 9, and its function in EthR is unknown. The crystal structure revealed that the dimerization interface, a conserved structural feature among the TetR class of repressors, is primarily formed by helices 8 and 9 (288, 351).

One of the most striking features of the EthR structure is a narrow tunnel-like cavity formed by helices 4, 5, 7, and 8 that opens to the bottom of the molecule (79). The tunnel measures about 20 Å in length and is lined predominantly, albeit not exclusively, by aromatic residues, with helices 5 and 7 constituting the majority of side chains. The loop connecting helices 4 and 5 restricts the opening of the hydrophobic tunnel, and the electron density in this loop is only poorly defined, indicating a certain degree of structural flexibility in the loop. This cavity may serve as the binding site for an as yet unknown ligand.

Crystal structure of TetR family members with unknown functions. New genomic/proteomic approaches are leading to the crystallization of a number of proteins, many of which have no assigned function. The following proteins of the TetR family have been crystallized: Cgl2612 of *Corynebacterium glutamicum* (pdb 1V7B); YbiH of *Salmonella enterica* serovar Typhimurium (pdb 1T33); YcdC of *Escherichia coli* (pdb 1PB6); and YfiR and YsiA from *Bacillus subtilis* (pdb 1RKT and 1VIO, respectively).

DNA-BINDING PREDICTIONS BASED ON TetR AND QacR CRYSTAL STRUCTURES

There is a perfect overlap of the DNA binding domains of QacR, TetR, CprB, and EthR, and no gaps were found in the α-helices involved in contacts with DNA in the multialignment of the 2,353 members of the TetR family in this domain. Based on these findings, we hypothesized that residues at the same position in the multialignment of all family members may play equivalent roles. This prompted us to analyze each amino acid in the multialignment within the DNA binding domain.

TABLE 4. Amino acid frequency at each of the positions critical for operator recognition by TetR family members[a]

Frequency at position:															
22		33		34		35		37		38		39		43	
AA	%	AA	%	AA	%	AA	%	AA	%	AA	%	AA	%	AA	%
V	21.95	K	29.26	G	37.79	T	34.30	Y	74.16	R	22.75	H	44.50	K	77.25
L	20.60	R	20.07	A	18.93	S	20.87	F	8.05	Y	16.44	Y	33.83	R	10.07
M	18.66	P	10.27	P	12.55	A	19.6	H	4.63	H	13.29	R	4.90	L	2.82
I	17.65	Q	6.04	S	8.39	L	5.84	L	3.09	N	8.39	F	3.56	I	1.88
T	13.22	V	5.30	R	6.71	N	5.30	T	2.48	K	6.11	A	2.15	M	1.88
F	2.15	A	4.56	T	5.10	G	4.30	S	1.95	W	5.84	N	1.95	V	1.81
H	1.74	T	4.30	Q	3.56	V	2.28	N	1.88	A	5.64	Q	1.88	T	0.94
A	1.61	L	4.23	M	2.35	M	2.08	R	1.21	L	3.56	E	1.48	G	0.94
Y	1.28	E	3.36	N	1.07	Q	1.34	Q	0.6	S	3.56	L	1.34	Q	0.74
S	0.54	I	2.89	K	1.01	Y	1.28	A	0.47	F	2.89	W	1.21	A	0.47
P	0.40	H	2.89	V	0.81	I	1.14	I	0.47	Q	2.55	S	0.94	H	0.34
N	0.13	S	2.01	D	0.47	P	0.54	G	0.27	T	2.35	T	0.74	F	0.27
E	0.07	N	1.61	L	0.47	R	0.40	M	0.27	E	1.68	V	0.54	P	0.20
		G	1.07	E	0.40	E	0.20	V	0.2	V	1.61	C	0.40	S	0.13
		D	1.01	F	0.27	H	0.20	K	0.13	G	1.41	K	0.27	Y	0.07
		Y	0.67	I	0.07	K	0.13	C	0.07	D	1.07	I	0.13	E	0.07
		M	0.27	H	0.07	D	0.13	P	0.07	I	0.54	D	0.07	N	0.07
		C	0.13			C	0.07			C	0.20	G	0.07		
		F	0.07							M	0.07				
										P	0.07				

[a] The amino acid (AA) frequency is expressed as a percentage and refers to the 2,353 TetR family members.

Relationship between Profile Positions and Structural Positioning

Analysis comparison of the cocrystal of QacR and TetR with their corresponding operators revealed that residues corresponding to positions 22, 33, 34, 35, 37, 38, 39, and 43 in the family multialignment are involved in interactions with target operator DNA (Fig. 1). We analyzed the occurrence of each amino acid at these positions in the multialignment of all members of the TetR family (Table 4).

We found two types of position, one in which the residue was highly conserved and another in which the residue was poorly conserved, if at all. Positions 37, 39, and 43 were well conserved, whereas at positions 22, 33, 34, 35, and 38 the profile aligned different residues.

Tyr42 in TetR and Tyr40 in QacR corresponded to position 37 in the profile sequence displayed in Fig. 1, where a Tyr residue appeared in 74.16% of the aligned proteins (Table 4). The next most highly represented residues in this position are also aromatic amino acids: phenylalanine (8%) and histidine (4%) (Table 4). Tyr-42 in TetR and Tyr40 in QacR appear at the center of α-helix 3 and contact a thymine located at the center of the palindrome forming the operator and also contact a phosphate one position towards the center of the palindrome (Fig. 3B and 6B).

The residue at position 39 of the profile in the multialignment corresponds to His44 in TetR and His42 in QacR. In the corresponding cocrystals, these residues established contacts with the phosphate backbone (Fig. 3B and Fig. 6B). In the multiple sequence alignment of all family members, either histidine or tyrosine appears at position 39. We are tempted to propose that this residue is critical for interactions with the phosphate backbone.

A lysine-DNA phosphate interaction is shared at residues Lys48 in TetR and Lys46 in QacR, which correspond to position 43 in the multialignment and are located in the amino end of the α4 helix. A lysine residue is present in 77% of TetR proteins, and their interactions with DNA phosphates seem to be crucial to adjust the HTH domain to contact DNA (Fig. 3B and 6B). At position 22 of the profile (Thr27 in TetR and Thr25 in QacR), five residues are the most abundant (Val, Leu, Met, Ile, and Thr). Thr27 in TetR and Thr25 in QacR are involved in interactions with the phosphate backbone.

Thus, in the TetR family, the contacts established by the residue aligned at position 37 in α3 (tyrosine present in 74% of the cases) and 39 in α3 (His or Tyr present in 98% of the cases) and a residue at position 43 in α4 (Lys present in 77% of the cases) probably orient the HTH motif to interact with the DNA major groove and anchor the protein to the phosphate backbone.

Glycine at position 16, located at the end of α1, in the multialignment is highly conserved and is involved in changing the polypeptide direction in the TetR and QacR crystals to orient the HTH DNA binding domain properly.

Positions 33, 34, 35, and 38 of the profile align many different residues (Table 4). In TetR and QacR, the corresponding residues establish specific contacts with different DNA bases except Asn38 of QacR (position 35 in the multialignment), which contacts the phosphate backbone. Based on the high variability of these positions in the corresponding multiple alignment of the family, we are tempted to propose that these positions endow specificity to each protein so that it can recognize its operator through specific protein-DNA interactions.

SOME REGULATORS ARE PART OF COMPLEX REGULATORY CIRCUITS

Published data indicate that the specific function of 85 members of TetR family is known (Table 2). More infor-

mation about each TetR protein is available at http://www.bactregulators.org (235). We have clustered the functions regulated by TetR family members into 10 groups (Table 2). The most frequent function performed by TetR family proteins is the regulation of efflux pumps and transporters involved in antibiotic resistance and tolerance to toxic chemicals. We have also observed that TetR family members often regulate their own synthesis, this feedback control ensures the transcriptional repressor level within optimal concentration limits (31, 73, 231, 338, 392). In this simple regulatory scheme, synthesis of the repressor and of the regulated protein(s) is derepressed in the presence of an inducer molecule.

However, TetR family proteins also participate in other types of regulatory networks that underlie complex processes, such as homeostasis in metabolism (biosynthesis of amino acids, nucleotides, protoheme, and reserve material), synthesis of osmoprotectants, quorum sensing, drug resistance, virulence, and processes related to growth phase-dependent differentiation (sporulation and biosynthesis of antibiotics) (Table 2) (www.bactregulators.org) (235).

Figure 8 shows a series of schemes in which a TetR family member plays a role in complex circuits. Below, for the sake of brevity, we have analyzed only some representative sets of regulatory networks, including proteins involved in drug resistance (AcrR of *E. coli* and MtrR of *Neisseria gonorrhoeae*), biosynthesis of an osmoprotectant (BetI), a key protein involved in idiophase antibiotic production and differentiation in *Streptomyces* (ArpR), a protein involved in pathogenesis in *Vibrio* (HapR), and some proteins involved in quorum sensing.

AcrR Regulator Is the Local Specific Regulator of the *acrAB* Efflux Pump

Multiple antibiotic resistance in *Escherichia coli* has attracted recent attention, promoting the elucidation of a number of mechanisms that contribute to this phenomenon. One of these is the transport of diverse substrates out of the cell by the AcrAB-TolC efflux transporter, leading to a multiple antibiotic resistance (Mar) phenotype (267). The set of antibiotics to which AcrAB can confer resistance includes ampicillin, chloramphenicol, erythromycin, fluoroquinolones, β-lactams, novobiocin, tetracycline, tigecycline, and rifampin (151, 187, 223, 267, 268, 276).

AcrB is a large cytoplasmic membrane protein (224, 226, 445, 446) which associates with AcrA, a membrane fusion protein (281), and TolC, a protein that forms a channel for the extrusion of substrates into the medium (102, 193). The *acrA* and *acrB* genes form an operon (224) whose transcription is regulated by the *acrR* gene product. The *acrR* gene is divergently transcribed from the *acrAB* operon. Overexpression of AcrR represses the transcription of *acrAB*. This observation is consistent with the function of AcrR as a repressor for *acrAB* transcription. Evidence for this function has come also from gel shift mobility assays, which provided direct evidence for the binding of AcrR to the promoter region of *acrAB*. DNA sequencing (92) of certain isolates that overexpressed *acrB* mRNA revealed that the mutant strains had insertions that disrupted the *acrR* gene or point mutations that rendered a nonfunctional regulator, i.e., an amino acid substitution of cysteine for arginine at position 45 of AcrR. This biochemical and genetic evidence provides support for the regulatory role of AcrR.

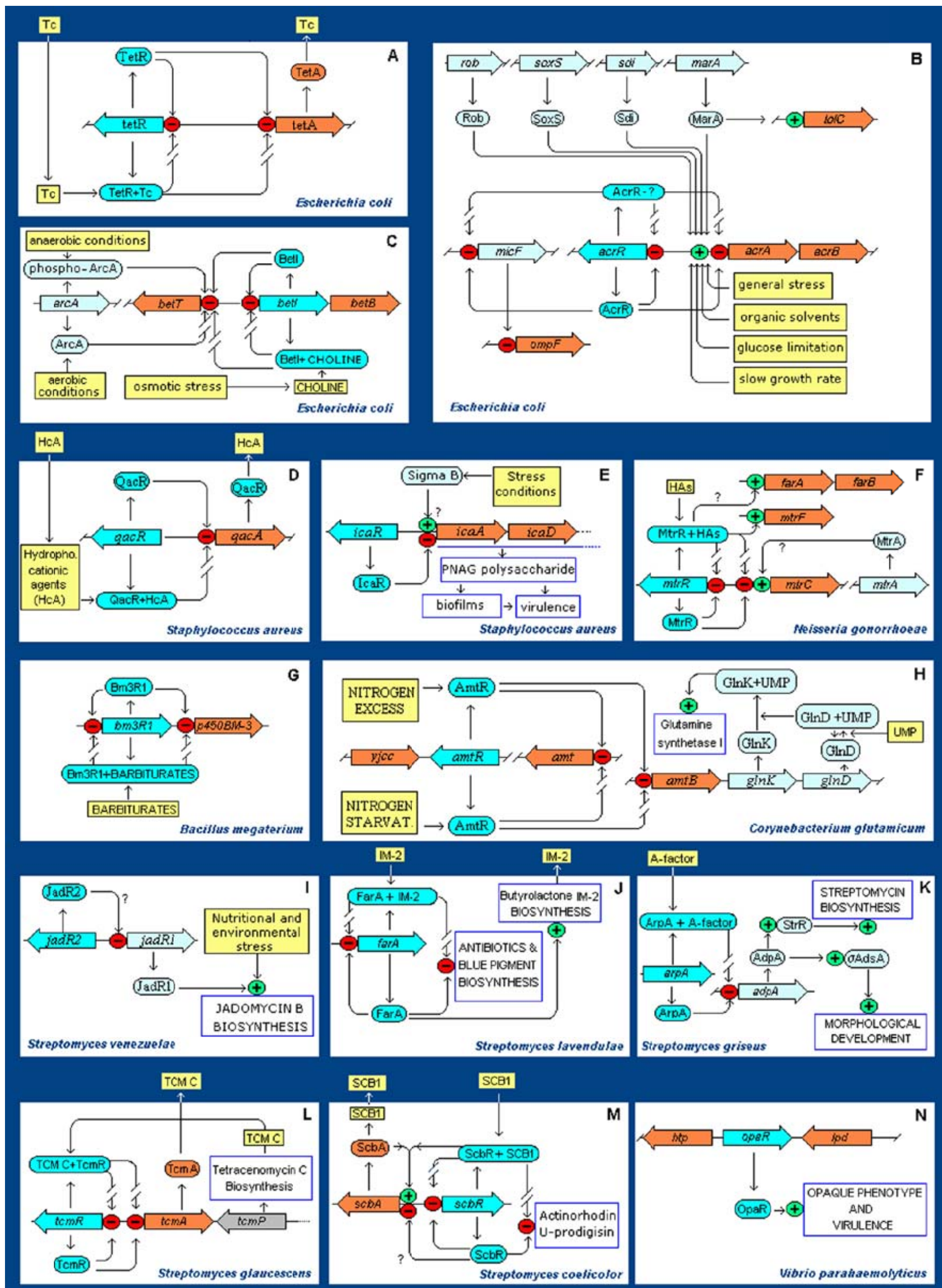
MarA, SoxS, and Rob are related transcriptional activators of the AraC/XylS family (7, 112, 367) that activate *acrAB* expression, although they are not involved in the regulation of *acrAB* in response to general stress conditions (13, 14, 21, 35, 110, 224) because the *acrAB* operon can be activated in response to these stresses in genetic backgrounds lacking *mar* and *sox* (223–225). It was also found that general stress conditions increased the transcription of *acrAB* in the absence of functional AcrR, and these conditions, surprisingly, increased the transcription of *acrR* to a greater extent than that of *acrAB*. These results suggest the existence of a *mar-sox*-independent pathway to control *acrAB* expression in response to the general stress conditions. This transcriptional control of *acrAB* is also AcrR independent. Therefore, a major role of AcrR is to function as a specific secondary modulator to fine-tune the level of *acrAB* transcription and prevent unwanted overexpression of the efflux pump. This represents a novel mechanism for regulating gene expression in *E. coli*.

Mtr Circuit of *Neisseria*

The MtrCDE efflux pump of *Neisseria gonorrhoeae* provides gonococci with a mechanism to resist structurally diverse antimicrobial hydrophobic agents and antibiotic peptides that adopt β-sheet (protegenin 1) or two-helix (PC-8 and LC37) structures (130, 228, 238, 353). Mutations that render no expression or inactivation of *mtrR*, encoding a transcriptional repressor, resulted in high expression of the *mtrCDE* operon, concomitantly increasing resistance to hydrophobic agents (69, 130, 220, 221, 297, 353, 447). It was also found that strains of *N. gonorrhoeae* that display hypersusceptibility to hydrophobic agents often contained mutations in the *mtrCDE* efflux pump genes (406).

The *mtrR* gene is divergently transcribed with respect to the *mtrCDE* operon (Fig. 8F). The promoters of *mtrR* and *mtrC* overlap in their −35 boxes, and footprinting analysis showed that MtrR binds a 40-bp region within the −10 to −35 region of the *mtrR* promoter, which contains an inverted repeat (221). MtrR bound to its target site prevented expression from the efflux pump operon and its regulator (Fig. 8F). The expression of *mtr* genes is enhanced by the AraC/XylS member MtrA, although the mechanism of activation of this protein is unknown.

On the other hand, Veal and Shafer (407) have recently identified a gene that was designated *mtrF*, located downstream of the *mtrR* gene, that is predicted to encode a 56.1-kDa cytoplasmic membrane protein containing 12 transmembrane domains. Expression of *mtrF* was enhanced in a strain deficient in MtrR production, indicating that this gene, together with the closely linked *mtrCDE* operon, is subject to MtrR-dependent transcriptional control. Genetic evidence suggests that MtrF is also important in the expression of high-level detergent resistance by gonococci, and it was proposed that MtrF acts in conjunction with the MtrC–MtrD–MtrE efflux pump to confer high-level resistance to certain hydrophobic agents in gonococci. MtrR also controls the *farAB* operon, which encodes an efflux pump involved in resistance to long-chain fatty



344

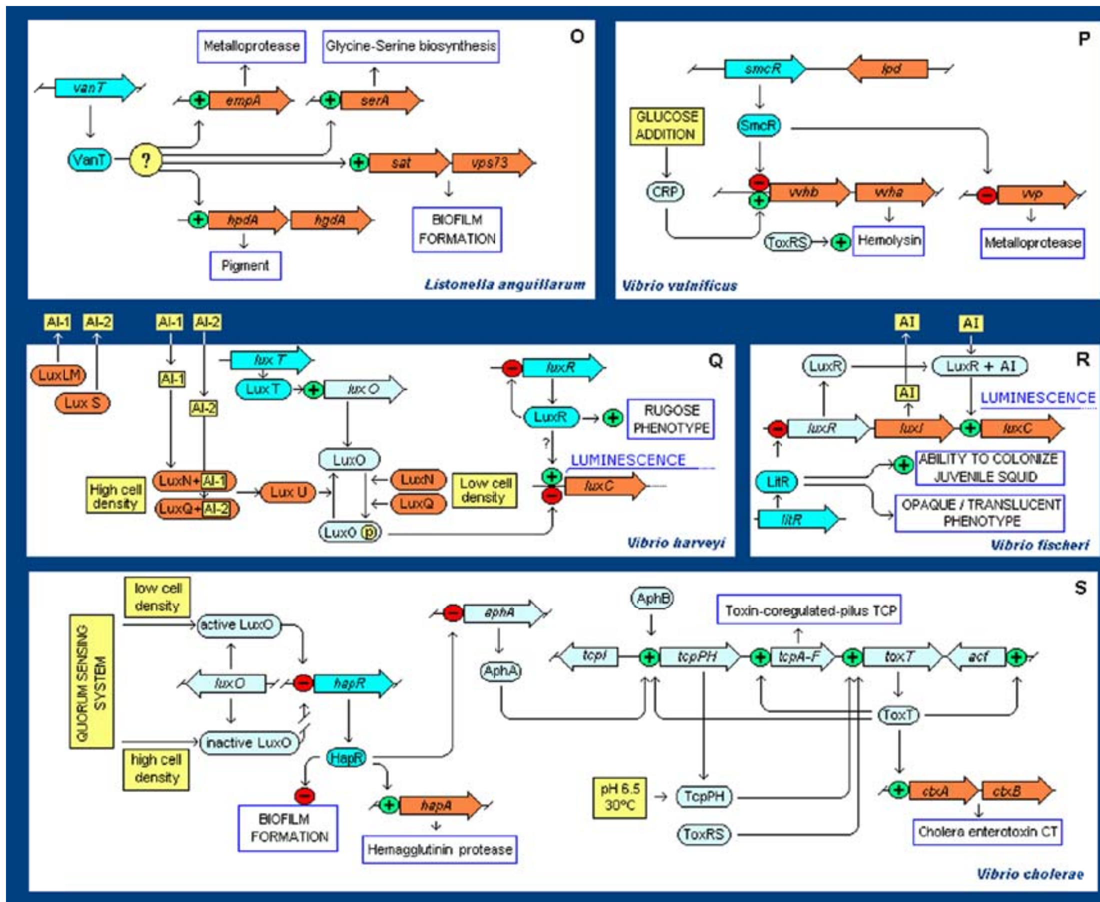


FIG. 8. Regulatory networks involving members of the TetR family. Although TetR and QacR cannot be considered part of a network, their type of control is shown because it is frequently found in members of the family. The following color code was used for complex networks: dark blue, TetR family member; orange, the gene directly regulated by the TetR family member; light blue, a regulator that modulates the expression of a TetR family member or which assists in the regulation of the gene under the control of a TetR family member; yellow boxes, signals and conditions influencing the system; open boxes, final results of the action of the system when the result is a scorable phenotype. References recommended for each circuit: panel A (29, 38, 286, 288, 388, 417, 418); panel B (4, 7, 13, 31, 35, 110, 176, 191, 230, 245, 270, 436); panel C (8, 91, 201–204, 222, 290, 331); panel D (119, 120, 122, 249, 261, 332, 350–351); panel E (5, 6, 66, 67, 116, 163); panel F (228, 238, 397, 333, 353, 408); panel G (87–89, 125, 140, 214, 215, 295, 360, 403); panel H (48, 161); panel I (78); panel J (56, 184, 185, 208, 413); panel K (133, 134, 158, 413); panel L (127); panel M (61, 266); panel N (240, 241, 345); panel O (248); panel P (59, 63, 242, 243, 355); panel Q (23, 24, 51, 107, 232, 255); panel R (97, 107, 117, 252, 341, 363); and panel S (107, 117, 181, 196, 217, 247).

acids (Fig. 8F). The efflux pump FarAB uses MtrE as the outer membrane component (208).

BetI Controls the Choline-Glycine Betaine Pathway of *E. coli*

In *Escherichia coli*, glycine betaine serves as an osmoprotector in hyperosmotically stressed cells. This osmoprotector accumulates in large amounts in the cytoplasm, which allows cells to maintain appropriate osmotic strength and thus prevents dehydration. Glycine betaine is only one of several cellular osmolytes used by *E. coli*, but its accumulation allows this microbe to achieve its highest level of osmotic tolerance (199, 373). To accumulate glycine betaine, *E. coli* needs an external supply of this compound or its precursors choline and betaine aldehyde.

The osmoregulatory choline-glycine betaine pathway is encoded by the *bet* genes. The *betA* gene encodes choline dehydrogenase; *betB* encodes betaine aldehyde dehydrogenase; *betT* encodes a transport system for choline; and *betI* encodes a 21.8-kDa repressor protein involved in choline regulation of the *bet* genes (Fig. 8C). The *bet* genes are linked, with *betT* being transcribed divergently from the *betIBA* operon (203, 374). Primer extension analysis identified two partially over-

lapping promoters which were responsible for the divergent expression of the *betT* gene and *betIBA* operon. The transcripts are initiated 61 bp apart and are induced by osmotic stress, but for full expression choline is required in the growth medium (91, 202, 204). Because the ArcA protein represses the expression of *bet* genes in *E. coli* under anaerobiosis, the *bet* genes are expressed only under aerobic conditions. An *arcA* mutation caused complete derepression of the *bet* genes. A similar pattern for derepression by ArcA has been reported previously for other genes (*sodA* and *arcA*) which are directly regulated by ArcA (65).

Results from different laboratories suggest that choline regulation but not osmotic regulation of the *bet* promoters depended on BetI, a TetR family member. This was indicated by the requirement for choline, in addition to osmotic stress, for *betT* to be expressed in a mutant strain in which *betI* was supplied in *trans*. Furthermore, this choline effect was not seen in cells lacking *betI*. These findings indicate that *betI* encodes a repressor that reduced the expression of *betT* (331).

A chimeric BetI glutathione *S*-transferase fusion protein (BetI*) was purified, and gel mobility shift assays showed that BetI* formed a complex with a 41-bp DNA fragment carrying the intergenic *betI* promoter region. Footprinting revealed the presence of two sequences of dyad symmetry which probably constitute the BetI operator.

The *Sinorhizobium meliloti bet* genes have been cloned, and their involvement in response to osmotic stress has been analyzed (304, 305, 368).

ArpA Regulator from *Streptomyces*

Streptomycetes are filamentous, soil-living, gram-positive bacteria characterized for their ability to produce a wide variety of secondary metabolites, including antibiotics and biologically active substances, and for their complex morphological differentiation culminating in the formation of chains of spores (55). Secondary metabolite synthesis is sometimes called "physiological" differentiation because it occurs during the idiophase after the main period of rapid vegetative growth and assimilative metabolism (139, 153, 291).

The ultimate regulator of the mentioned processes in *Streptomyces griseus* is a homodimeric protein called A-factor receptor protein (ArpA) (155–158), which regulates the switch for physiological and morphological differentiation. The main biologically significant target of ArpA is the *adpA* gene. The AdpA protein in turn controls the expression of other genes. These genes include *strR*, which serves as a pathway-specific transcriptional activator for streptomycin biosynthetic genes (278); an open reading frame encoding a probable pathway-specific regulator for a polyketide compound (441); *adsA*, which encodes an extracytoplasmic function sigma factor of RNA polymerase essential for aerial mycelium formation (438); *sgmA*, which encodes a metalloendopeptidase probably involved in apoptosis of substrate hyphae during aerial mycelium development (174); *ssgA*, which encodes a small acidic protein essential for spore septum formation (437); *amfR*, essential for aerial hyphae formation (398, 440); and the *sprT* and *sprU* genes, which encode trypsin-like proteases (173).

In vitro, ArpA binds its target DNA site at the −10/−35 region, which is a 22-bp palindrome (5′-GG(T/C)CGGT(A/T)(T/C)G(T/G)-3′). Addition of γ-butyrolactone effector to the ArpA-DNA complex immediately releases ArpA from the DNA. A mutant strain deficient in ArpA or producing a mutant ArpA protein unable to bind to its target DNA overproduces streptomycin and forms aerial mycelia and spores earlier than the wild-type strain (282, 283). An amino acid replacement at Val-41 to Ala in ArpA in the HTH motif at the N-terminal portion of ArpA abolished DNA binding activity but not γ-butyrolactone binding activity, suggesting the involvement of this HTH in DNA binding. On the other hand, mutation of Trp-119 (Trp 119→Ala) generated a mutant unable to bind the γ-butyrolactone, resulting in a mutant protein that did not sense the presence of A-factor. These data suggest that ArpA consists of an HTH DNA-binding at the N-terminal end and an effector binding domain at the C-terminal end of the protein.

In the streptomycin biosynthetic gene cluster in *S. griseus*, StrR is the pathway-specific regulator that serves as a transcriptional activator for the other genes in the cluster (320). Expression of the *strR* gene was controlled by the AdpA protein, which binds the region upstream of the *strR* promoter and activates its transcription (311, 312). *adpA* knockout mutants produced no streptomycin, and overexpression of *adpA* caused the wild-type *S. griseus* strain to produce streptomycin at an earlier growth stage in a larger amount. This set of events explains how A-factor triggers streptomycin biosynthesis.

Disruption of the chromosomal *adsA* gene encoding σ^{AdsA} resulted in loss of aerial hypha formation but not streptomycin production, indicating that this sigma factor is involved in morphological development (401, 438).

Several receptor proteins for γ-butyrolactone-type autoregulators have been described in other species of *Streptomyces*. For example, CprA and CprB are involved in secondary metabolism and aerial mycelium formation in *S. coelicolor* A3(2) (284). The virginiae butanolyde receptor BarA is involved in virginiamycin biosynthesis in *Streptomyces virginiae* (280), and the IM-2 receptor, FarA is involved in blue pigment production in another *Streptomyces* strain (413).

FarA in *Streptomyces lavendulae* senses the concentration of γ-butyrolactone IM-2 and also transduces this signal, thus derepressing antibiotic and blue pigment biosynthesis. In addition, FarA also seems to be necessary for IM-2 biosynthesis (Fig. 8J).

The role of ScbR protein in the quorum-sensing circuit of *Streptomyces coelicolor* is similar to that of FarA in *S. lavendulae*. ScbR is also involved in three functions: positively regulating SCB1 synthesis (the γ-butyrolactone that acts as signal), receiving the signal, and transducing this signal, thereby derepressing the production of the antibiotics actinorhodin and U-prodigiosin (Fig. 8M).

HapR Regulates Virulence Genes in *Vibrio cholerae*

The development of cholera in humans is directly related to the production of two virulence factors: toxin-coregulated pilus (TCP), which mediates colonization, and cholera toxin (CT), responsible for the severe diarrhea, characteristic of this disease. The coordinated expression of TCP and CT occurs though a complex regulatory cascade (Fig. 8S) in which the regulators AphA and AphB synergistically activate *tcpPH* tran-

scription (194). The TetR family protein HapR negatively regulates the expression of AphA and indirectly diminishes the production of TCP and CT.

HapR, in turn, is regulated by quorum-sensing signals that are sensed and transmitted by LuxO. The quorum-sensing apparatus in *Vibrio cholerae* is unusually complex and is composed of three parallel signaling systems (247). In contrast to other bacteria, in which high cell density triggers virulence gene expression, in *Vibrio cholerae* low cell density is the condition that activates the production of the pathogenic factors CT and TCP (455). At high cell density HapR positively regulates the expression of a hemagglutinin protease (Hap) that promotes detachment of *Vibrio cholerae* from the gastrointestinal epithelium (365) and exerts a negative effect on biofilm formation. Taking into account the regulatory functions of HapR and considering that some pathogenic biotypes lose HapR expression whereas others lose the *aphA* binding site, it appears that HapR expression is related to diminished toxicity and colonization capacity. These features offer potentially fruitful avenues of research to design drugs to modulate *Vibrio cholerae* pathogenicity.

Other Quorum-Sensing Circuits

Within the genus *Vibrio*, some TetR family proteins, i.e., LuxT, LuxR, and LitR, are involved in complex quorum-sensing circuits. In *Vibrio harveyi*, LuxT and LuxR participate in the circuit that regulates luminescence. This strain senses two autoinducers, AI-1 and AI-2 (58, 216). AI-1 is an acyl-homoserine lactone, as in other gram-negative bacteria. However, AI-2 is a novel autoinducer produced by many species that appears to be related to interspecies cell density detection. These two autoinducers use the same phosphorelay system to transduce the signal for bioluminescence regulation. Expression of *luxR* is in turn regulated by LitR (Fig. 8R).

BIOTECHNOLOGICAL APPLICATIONS AND FUTURE PROSPECTS

The Tet system is at present the most widely used system for conditional gene expression in eukaryotic cells (37). The system is based in the high affinity (10^{-9} M) of TetR for its operator, *tetO*, the favorable pharmacokinetics of tetracyclines (they diffuse through biological membranes), and their long record of safe clinical use. Cloning of the *tetO* elements adjacent to the TATA box of the target gene (114) was used successfully to control genes expressed by RNA polymerase I in *Leishmania donovani* and by RNA polymerase III in yeast and plant cells (76, 400). However, this system is not efficient in mammalian cells. For this reason, and based on the knowledge acquired about how TetR binds to its target operator, several chimeric versions of TetR fused to eukaryotic regulatory domains were constructed, such as the acidic activation domain (tTA) (22, 115, 403) and repression domains (tTS) (33, 336). Based on hybrid transregulators, transgenic mice able to produce diphtheria toxin or the regulated expression of Shiga toxin β to induce apoptosis in mammalian fibroblastic cells were obtained.

The *tet* system has also been used to study cancer and neurological disorders (95, 98, 419). In the future, advances to approach multifactorial biological processes like development and diseases are expected, which will be relevant for the treatment of complex diseases (37).

Solvent efflux pumps generally exhibit a broad substrate specificity, but some of them are highly specific and remove a certain number of chemicals. One such efflux pumps is the TtgABC pump of *Pseudomonas putida* DOT-T1E, which removes toluene, *m*-xylene, and propylbenzene as well as styrene and other aromatic hydrocarbons (327) in addition to several antibiotics. This efflux pump is under the control of the drug binding repressor TtgR, a member of the TetR family (391), and has applications in the safe development of solvent-resistant bacteria. Therefore, it is potentially suitable for the biotransformation of water-insoluble compounds into added-value products. One such system has been exploited to produce catechols from xylenes/toluenes in double-phase systems (328). Because these efflux systems are energy consuming, their expression in heterologous hosts has to be tightly controlled by their cognate repressor, which in turn has to be able to respond to the presence of the aromatic solvent. These types of pumps are presumed to be useful if transferred to a wide variety of bacteria that carry suitable biotransformation machineries.

As shown in Fig. 8, TetR family members are key players in multidrug resistance, virulence, and pathogenicity processes in certain bacterial pathogens. The development of drugs that bind irreversibly to the repressors and prevent their release from their cognate operators may be a strategy to fight these pathogens. Suitable screening procedures to search for these drugs can be envisaged based on current tools for gram-negative bacteria (74).

ACKNOWLEDGMENTS

Work in our laboratories was supported by EC grants MIFRIEND QLK3-CT 2000-017 and BIOCARTE QLRT-2001-01923 and projects CICYT-2003 (BIO2003-00515) and Petri-NBT (PTR1995-0653 PO) to J.L.R. and grant 0021/2002 from the Human Frontier Science Program to M.T.G., K.W., and X.Z.

We thank W. Hindrich and S. Horinouchi for comments on the article. We thank Inés Abril and Carmen Lorente for secretarial assistance and Karen Shashok for checking the use of English in the manuscript. R. Tobes acknowledges support from Eduardo Pareja.

REFERENCES

1. **Abo-Amer, A. E., J. Munn, K. Jackson, M. Aktas, P. Golby, D. J. Kelly, and S. C. Andrews.** 2004. DNA interaction and phosphotransfer of the C_4-dicarboxylate-responsive DcuS-DcuR two-component regulatory system from *Escherichia coli*. J. Bacteriol. **186:**1879–1889.
2. **Ahlert, J., E. Shepard, N. Lomovskaya, E. Zazopoulos, A. Staffa, B. O. Bachmann, K. Huang, L. Fonstein, A. Czisny, R. E. Whitwam, C. M. Farnet, and J. S. Thorson.** 2002. The calicheamicin gene cluster and its iterative type I enediyne PKS. Science **297:**1173–1176.
3. **Aigle, B., A. Wietzorrek, E. Takano, and M. J. Bibb.** 2000. A single amino acid substitution in region 1.2 of the principal sigma factor of *Streptomyces coelicolor* A3(2) results in pleiotropic loss of antibiotic production. Mol. Microbiol. **37:**995–1004.
4. **Alekshun, M. N., and S. B. Levy.** 1999. Alteration of the repressor activity of MarR, the negative regulator of the *Escherichia coli marRAB* locus, by multiple chemicals in vitro. J. Bacteriol. **181:**4669–4672.
5. **Allignet, J., S. Aubert, K. G. Dyke, and N. El Solh.** 2001. *Staphylococcus caprae* strains carry determinants known to be involved in pathogenicity: a gene encoding an autolysin-binding fibronectin and the *ica* operon involved in biofilm formation. Infect. Immun. **69:**712–718.
6. **Allmeier, H., B. Cresnar, M. Greck, and R. Schmitt.** 1992. Complete nucleotide sequence of Tn*1721*: gene organization and a novel gene product with features of a chemotaxis protein. Gene **111:**11–120.
7. **Amabile-Cuevas, C. F., and B. Demple.** 1991. Molecular characterization of the *soxRS* genes of *Escherichia coli*: two genes control a superoxide stress regulon. Nucleic Acids Res. **19:**4479–4484.

8. **Andresen, P. A., I. Kaasen, O. B. Styrvold, G. Boulnois, and A. R. Strøm.** 1988. Molecular cloning, physical mapping and expression of the *bet* genes governing the osmoregulatory choline-glycine betaine pathway of *Escherichia coli*. J. Gen. Microbiol. **134:**1737–1746.
9. **Aramaki, H., Y. Sagara, M. Hosoi, and T. Horiuchi.** 1993. Evidence for autoregulation of *camR*, which encodes a repressor for the cytochrome P-450cam hydroxylase operon on the *Pseudomonas putida* CAM plasmid. J. Bacteriol. **175:**7828–7833.
10. **Aramaki, H., Y. Sagara, H. Kabata, N. Shimamoto, and T. Horiuchi.** 1995. Purification and characterization of a *cam* repressor (CamR) for the cytochrome P-450cam hydroxylase operon on the *Pseudomonas putida* CAM plasmid. J. Bacteriol. **177:**3120–3127.
11. **Aramaki, H., N. Yagi, and M. Suzuki.** 1995. Residues important for the function of a multihelical DNA binding domain in the new transcription factor family of Cam and Tet repressors. Protein Eng. **8:**1259–1266.
12. **Argos, P., M. Hanei, J. M. Wilson, and W. N. Kelley.** 1983. A possible nucleotide-binding domain in the tertiary fold of phosphoribosyltransferases. J. Biol. Chem. **258:**6450–6457.
13. **Ariza, R. R., S. P. Cohen, N. Bachawat, S. B. Levy, and B. Demple.** 1994. Repressor mutations in the *marRAB* operon that activate oxidative stress genes and multiple antibiotic resistance in *Escherichia coli*. J. Bacteriol. **176:** 143–148.
14. **Ariza, R. R., Z. Li, N. Ringstad, and B. Demple.** 1995. Activation of multiple antibiotic resistance and binding of stress-inducible promoters by *Escherichia coli* Rob protein. J. Bacteriol. **177:**1655–1661.
15. **Attwood, T. K., P. Bradley, D. R. Flower, A. Gaulton, N. Maudling, A. L. Mitchell, G. Moulton, A. Nordle, K. Paine, P. Taylor, A. Uddin, and C. Zygouri.** 2003. PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. **31:**400–402.
16. **Attwood, T. K., M. D. Croning, D. R. Flower, A. P. Lewis, J. E. Mabey, P. Scordis, J. N. Selley, and W. Wright.** 2000. Prints-S: The database formerly known as Prints. Nucleic Acids Res. **28:**225–227.
17. **August, P. R., L. Tang, Y. J. Yoon, S. Ning, R. Muller, T. W. Yu, M. Taylor, D. Hoffmann, C. G. Kim, X. Zhang, C. R. Hutchinson, and H. G. Floss.** 1998. Biosynthesis of the ansamycin antibiotic rifamycin: deductions from the molecular analysis of the *rif* biosynthetic gene cluster of *Amycolatopsis mediterranei* S699. Chem. Biol. **5:**69–79.
18. **Babu, M. M., and S. A. Teichmann.** 2003. Evolution of transcription factors and the gene regulatory network in *Escherichia coli*. Nucleic Acids Res. **31:** 1234–1244.
19. **Bairoch, A., and K. E. Rudd.** 1995. PROSITE. Accession no. PDOC00874.
20. **Balsalobre, L., M. J. Ferrándiz, J. Liñares, F. Tubau, and A. G. de la Campa.** 2003. Viridans group Streptococci are donors in horizontal transfer of topoisomerase IV genes to *Streptococcus pneumoniae*. Antimicrob. Agents Chemother. **47:**2072–2081.
21. **Barbosa, T. M., and S. B. Levy.** 2000. Differential expression of over 60 chromosomal genes in *Escherichia coli* by constitutive expression of MarA. J. Bacteriol. **182:**3467–3474.
22. **Baron, U., M. Gossen, and H. Bujard.** 1997. Tetracycline-controlled transcription in eukaryotes: novel transactivators with graded transactivation potential. Nucleic Acids Res. **25:**2723–2729.
23. **Bassler, B. L., M. Wright, and M. R. Silverman.** 1994. Sequence and function of LuxO, a negative regulator of luminescence in *Vibrio harveyi*. Mol. Microbiol. **12:**403–412.
24. **Bassler, B. L., M. Wright, and M. R. Silverman.** 1994. Multiple signalling systems controlling expression of luminescence in *Vibrio harveyi*: sequence and function of genes encoding a second sensory pathway. Mol. Microbiol. **13:**273–286.
25. **Bate, N., A. R. Butler, A. R. Gandecha, and E. Cundliffe.** 1999. Multiple regulatory genes in the tylosin biosynthetic cluster of *Streptomyces fradiae*. Chem. Biol. **6:**617–624.
26. **Bateman, A., E. Birney, L. Cerruti, R. Durbin, L. Etwiller, S. R. Eddy, S. Griffiths-Jones, K. L. Howe, M. Marshall, and E. L. Sonnhammer.** 2002. The Pfam protein families database. Nucleic Acids Res. **30:**276–280.
27. **Bateman, A., E. Birney, R. Durbin, S. R. Eddy, K. L. Howe, and E. L. Sonnhammer.** 2000. The Pfam protein families database. Nucleic Acids Res. **28:**263–266.
28. **Baulard, A. R., J. C. Betts, J. Ndong, S. Quan, R. A. McAdam, P. J. Brennan, C. Locht, and G. S. Besra.** 2000. Activation of the pro-drug ethionamide is regulated in mycobacteria. J. Biol. Chem. **275:**28326–28331.
29. **Baumeister, R., V. Helbl, and W. Hillen.** 1992. Contacts between Tet repressor and *tet* operator revealed by new recognition specificities of single amino acid replacement mutants. J. Mol. Biol. **226:**1257–1270.
30. **Baumeister, R., G. Muller, B. Hecht, and W. Hillen.** 1992. Functional roles of amino acid residues involved in forming the alpha-helix-turn-alpha-helix operator DNA binding motif of Tet repressor from Tn10. Proteins **14:** 168–177.
31. **Becskei, A., and L. Serrano.** 2000. Engineering stability in gene networks by autoregulation. Nature **405:**590–593.
32. **Beinlich, K. L., R. Chuanchuen, and H. P. Schweizer.** 2001. Contribution of multidrug efflux pumps to multiple antibiotic resistance in veterinary clinical isolates of *Pseudomonas aeruginosa*. FEMS Microbiol. Lett. **198:**129–134.
33. **Belli, G., E. Gari, L. Piedrafita, M. Aldea, and E. Herrero.** 1998. An activator/repressor dual system allows right tetracycline regulated gene expression in budding yeast. Nucleic Acids Res. **26:**942–947.
34. **Beloin, C., S. Ayora, R. Exley, L. Hirschbein, N. Ogasawara, Y. Kasahara, J. C. Alonso, and F. Le Hegarat.** 1997. Characterization of an *lrp*-like (*lrpC*) gene from *Bacillus subtilis*. Mol. Gen. Genet. **256:**63–71.
35. **Bennik, H. J. M., P. J. Pomposiello, D. F. Thorne, and B. Demple.** 2000. Defining a *rob* regulon in *Escherichia coli* by using transposon mutagenesis. J. Bacteriol. **182:**3794–3801.
36. **Berens, C., Altschmied, L., and W. Hillen.** 1992. The role of the N terminus in Tet repressor for *tet* operator binding determined by a mutational analysis. J. Biol. Chem. **267:**1945–1952.
37. **Berens, C., and W. Hillen.** 2003. Gene regulation by tetracycline. Constraints of resistance regulation in bacteria shape TetR for application in eukaryotes. Eur. J. Biochem. **270:**3109–3121.
38. **Berens, C., D. Schnappinger, and W. Hillen.** 1997. The role of the variable region in Tet repressor for inducibility by tetracycline. J. Biol. Chem. **272:** 6936–6942.
39. **Biburger, M., C. Berens, T. Lederer, T. Krec, and W. Hillen.** 1998. Intragenic suppressors of induction-deficient TetR mutants: localization and potential mechanism of action. J. Bacteriol. **180:**737–741.
40. **Blanco, C., P. Ritzenthaler, and M. Mata-Gilsinger.** 1986. Negative dominant mutations of the *uidR* gene in *Escherichia coli*: genetic proof for a cooperative regulation of *uidA* expression. Genetics **112:**173–182.
41. **Branden, C., and J. Tooze.** 1991. Introduction to protein structure. Garland Publishing Co., New York, N.Y.
42. **Brandi, A., C. L. Pon, and C. O. Gualerzi.** 1994. Interaction of the main cold shock protein CS7.4 (CspA) of *Escherichia coli* with the promoter region of *hns*. Biochimie **76:**1090–1098.
43. **Brennan, R. G., S. L. Roderick, Y. Takeda, and B. W. Matthews.** 1990. Protein-DNA conformational changes in the crystal structure of a lambda-Cro-operator complex. Proc. Natl. Acad. Sci. USA **87:**8165–8169.
44. **Brown, M. H., and R. A. Skurray.** 2001. Staphylococcal multidrug efflux protein QacA. J. Mol. Biotechnol. **3:**163–170.
45. **Bucher, P., and A. Bairoch.** 1994. A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. Proc. Int. Conf. Intell. Syst. Mol. Biol. **2:**53–61.
46. **Bucher, P., K. Karplus, N. Moeri, and K. Hofmann.** 1996. A flexible motif search technique based on generalized profiles. Comput. Chem. **20:**3–24.
47. **Budarina, Z. I., D. V. Nikitin, N. Zenkin, M. Zakharova, E. Semenova, M. G. Shlyapnikov, E. A. Rodikova, S. Masyukova, O. Ogarkov, G. E. Baida, A. S. Solonin, and K. Severinov.** 2004. A new *Bacillus cereus*, DNA-binding protein, HlyIIR, negatively regulates expression of *B. cereus* haemolysin II. Microbiology. **150:**3691–3701.
48. **Burkovski, A.** 2003. I do it my way: Regulation of ammonium uptake and ammonium assimilation in *Corynebacterium glutamicum*. Arch. Microbiol. **179:**83–88.
49. **Busenlehner, L. S., M. A. Pennella, and D. P. Giedroc.** 2003. The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. FEMS Microbiol. Rev. **27:**131–143.
50. **Caballero, J. L., F. Malpartida, and D. A. Hopwood.** 1991. Transcriptional organization and regulation of an antibiotic export complex in the producing Streptomyces culture. Mol. Gen. Genet. **228:**372–380.
51. **Cao, J. G., and E. A. Meighen.** 1989. Purification and structural identification of an autoinducer for the luminescence system of *Vibrio harveyi*. J. Biol. Chem. **264:**21670–22166.
52. **Cases, I., V. de Lorenzo, and C. A. Ouzounis.** 2003. Transcription regulation and environmental adaptation in bacteria. Trends Microbiol. **11:**248–253.
53. **Chan, Y. Y., T. M. C. Tan, Y. M. Ong, and K. L. Chua.** 2004. BpeAB-OprB, a multidrug efflux pump in *Burkholderia pseudomallei*. Antimicrob. Agents Chemother. **48:**1128–1135.
54. **Chandler, M. S.** 1992. The gene encoding cAMP receptor protein is required for competence development in *Haemophilus influenzae* Rd. Proc. Natl. Acad. Sci. USA **89:**1626–1230.
55. **Chater, K. F.** 1993. Genetics of differentiation in *Streptomyces*. Annu. Rev. Microbiol. **47:**685–713.
56. **Chater, K. F., and C. J. Bruton.** 1985. Resistance, regulatory and production genes for the antibiotic methylenomycin are clustered. EMBO J. **4:** 1893–1897.
57. **Chatterjee, J., C. M. Miyamoto, and E. A. Meighen.** 1996. Autoregulation of *luxR*: the *Vibrio harveyi lux*-operon activator functions as a repressor. Mol. Microbiol. **20:**415–425.
58. **Chen, X., S. Schauder, N. Potier, A. Van Dorsselaer, I. Pelczer, B. L. Bassler, and F. M. Hughson.** 2002. Structural identification of a bacterial quorum-sensing signal containing boron. Nature **415:**545–549.
59. **Cheng, J. C., C. P. Shao, and L. I. Hor.** 1996. Cloning and nucleotide sequencing of the protease gene of *Vibrio vulnificus*. Gene **183:**255–257.
60. **Cho, S. H., K. Naber, J. Hacker, and W. Ziebuhr.** 2002. Detection of the *icaADBC* gene cluster and biofilm formation in *Staphylococcus epidermidis*

isolates from catheter-related urinary tract infections. Int. J. Antimicrob. Agents. **19:**570–575.
61. **Cho, Y. H., E. J. Kim, H. J. Chung, J. H. Choi, K. F. Chater, B. E. Ahn, J. H. Shin, and J. H. Roe.** 2003. The *pqrAB* operon is responsible for paraquat resistance in *Streptomyces coelicolor*. J. Bacteriol. **185:**6756–6763.
62. **Choi, E. N., M. C. Cho, Y. Kim, C. K. Kim, and K. Lee.** 2003. Expansion of growth substrate range in *Pseudomonas putida* F1 by mutations in both *cymR* and *todS*, which recruit a ring-fission hydrolase CmtE and induce the tod catabolic operon, respectively. Microbiology **149:**795–805.
63. **Choi, H. K., N. Y. Park, D. I. Kim, H. J. Chung, S. Ryu, and S. H. Choi.** 2002. Promoter analysis and regulatory characteristics of *vvhBA* encoding cytolytic hemolysin of *Vibrio vulnificus*. J. Biol. Chem. **277:**47292–47299.
64. **Cohen, S. P., H. Hachler, and S. B. Levy.** 1993. Genetic and functional analysis of the multiple antibiotic resistance (*mar*) locus in *Escherichia coli*. J. Bacteriol. **175:**1484–1492.
65. **Compan, I., and D. Touati.** 1994. Anaerobic activation of *arcA* transcription in *Escherichia coli*: roles of Fnr and ArcA. Mol. Microbiol. **11:**955–964.
66. **Conlon, K. M., H. Humphreys and J. P. O'Gara.** 2002. Regulation of *icaR* gene expression in *Staphylococcus epidermidis*. FEMS Microbiol. Lett. **216:** 171–177.
67. **Conlon, K. M., H. Humphreys, and J. P. O'Gara.** 2002. *icaR* encodes a transcriptional repressor involved in environmental regulation of *ica* operon expression and biofilm formation in *Staphylococcus epidermidis*. J. Bacteriol. **184:**4400–4408.
68. **Corpet, F., J. Gouzy, and D. Kahn.** 1999. Recent improvements of the Prodom database of protein domain families. Nucleic Acids Res. **27:**263–267.
69. **Cousin, S. L., Jr., W. L. H. Whittington, and M. C. Roberts.** 2003. Acquired macrolide resistance genes and the 1 bp deletion in the *mtrR* promoter in *Neisseria gonorrhoeae*. J. Antimicrob. Chemother. **51:**131–133.
70. **Cramton, S. E., C. Gerke, N. F. Schnell, W. W. Nichols, and F. Gotz.** 1999. The intercellular adhesion (*ica*) locus is present in *Staphylococcus aureus* and is required for biofilm formation.Infect. Immun. **67:**5427–5433.
71. **Croxatto, A., V. J. Chalker, J. Lauritz, J. Jass, A. Hardman, P. Williams, M. Camara, and D. L. Milton.** 2002. VanT, a homologue of *Vibrio harveyi* LuxR, regulates serine, metalloprotease, pigment, and biofilm production in *Vibrio anguillarum*. J. Bacteriol. **184:**1617–1629.
72. **Delahay, R. M., B. D. Robertson, J. T. Balthazar, W. M. Shafer, and C. A. Ison.** 1997. Involvement of the gonococcal MtrE protein in the resistance of *Neisseria gonorrhoeae* to toxic hydrophobic agents. Microbiology **143:**2127–2133.
73. **Delany, I., R. Ieva, C. Alaimo, R. Rappuoli, and V. Scarlato.** 2003. The iron-responsive regulator Fur is transcriptionally autoregulated and not essential in *Neisseria meningitidis*. J. Bacteriol. **185:**6032–6041.
74. **de Lorenzo, V., and H. P. Schweizer.** 2004. Molecular tools for genetic analysis of pseudomonads p. 317–350. In J. L. Ramos (ed.), *Pseudomonas*, vol. I. Kluwer, London, United Kingdom.
75. **Derzelle, S., E. Turlin, E. Duchaud, S. Pages, F. Kunst, A. Givaudan, and A. Danchin.** 2004. The PhoP-phoQ two-component regulatory system of *Photorhabdus luminescens* is essential for virulence in insects. J. Bacteriol. **186:**1270–1279.
76. **Dingermann, T., U. Frank-Stoll, H. Wener, A. Wissmann, W. Hillen, M. Jacquet, and R. Marschalek.** 1992. RNA polymerase III catalysed transcription can be regulated in *Saccharomyces cerevisiae* by the bacterial tetracycline repressor-operator system. EMBO J. **11:**1487–1492.
77. **Dogra, C., V. Raina, R. Pal, M. Suar, S. Lal, K. H. Gartemann, C. Holliger, J.R. van der Meer, and R. Lal.** 2004. Organization of *lin* genes and IS*6100* among different strains of hexachlorocyclohexane-degrading *Sphingomonas paucimobilis*: evidence for horizontal gene transfer. J. Bacteriol. **186:**2225–2235.
78. **Doull, J. L., A. K. Singh, M. Hoare, and S. W. Ayer.** 1994. Conditions for the production of jadomycin B by *Streptomyces venezuelae* ISP*5230*: effects of heat shock, ethanol treatment and phage infection. J. Ind. Microbiol. **13:** 120–125.
79. **Dover, L. G., P. E. Corsino, I. R. Daniels, S. L. Cocklin, V. Tatituri, G. S. Besra, and K. Fütterer.** 2004. Crystal structure of the TetR/CamR family repressor *Mycobacterium tuberculosis* EthR implicated in ethionamide resistance. J. Mol. Biol. **340:**1095–1105.
80. **Dugan, J., D. D. Rockey, L. Jones, and A. A. Andersen.** 2004. Tetracycline resistance in *Chlamydia suis* mediated by genomic islands inserted into the chlamydial *inv*-like gene. Antimicrob. Agents Chemother. **48:**3989–3995.
81. **Duque, E., A. Segura, G. Mosqueda, and J. L. Ramos.** 2001. Global and cognate regulators control the expression of the organic solvent efflux pumps TtgABC and TtgDEF of *Pseudomonas putida*. Mol. Microbiol. **39:** 1100–1106.
82. **Eaton, R. W.** 1996. *p*-Cumate catabolic pathway in *Pseudomonas putida* Fl: cloning and characterization of DNA carrying the *cmt* operon. J. Bacteriol. **178:**1351–1362.
83. **Eaton, R. W.** 1997. *p*-Cymene catabolic pathway in *Pseudomonas putida* F1: cloning and characterization of DNA encoding conversion of p-cymene to p-cumate. J. Bacteriol. **179:**3171–3180.
84. **Egan, S.** 2002. Growing repertoire of AraC/XylS activators. J. Bacteriol. **184:**5529–5532.
85. **El-Hajj, H. H., H. Zhang, and B. Weiss.** 1988. Lethality of a *dut* (deoxyuridine triphosphatase) mutation in *Escherichia coli*. J. Bacteriol. **170:**1069–1075.
86. **Engel, P., L. L. Scharfenstein, J. M. Dyer, and J. W. Cary.** 2001. Disruption of a gene encoding a putative gamma-butyrolactone-binding protein in *Streptomyces tendae* affects nikkomycin production. Appl. Microbiol. Biotechnol. **56:**414–419.
87. **English, N., V. Hughes, and C. R. Wolf.** 1994. Common pathways of cytochrome P450 gene regulation by peroxisome proliferators and barbiturates in *Bacillus megaterium* ATCC14581. J. Biol. Chem. **269:**26836–26841.
88. **English, N., V. Hughes, and C. R. Wolf.** 1996. Induction of cytochrome P-450 BM-3 (CYP 102) by non-steroidal anti-inflammatory drugs in *Bacillus megaterium*. Biochem. J. **316:**279–283.
89. **English, N., C. N. Palmer, W. L. Alworth, L. Kang, V. Hughes, and C. R. Wolf.** 1997. Fatty acid signals in *Bacillus megaterium* are attenuated by cytochrome P-450-mediated hydroxylation. Biochem. J. **327:**363–368.
90. **Engohang-Ndong, J., Baillat, D., Aumercier, M., Bellefontaine, F., Besra, G. S., Locht, C., and Baulard, A. R.** 2004. EthR, a repressor of the TetR/CamR family implicated in ethionamide resistance in mycobacteria, octamerizes cooperatively on its operator. Mol. Microbiol. **51:**175–188.
91. **Eshoo, M. W.** 1988. *lac* fusion analysis of the *bet* genes of *Escherichia coli*: regulation by osmolarity, temperature, oxygen, choline, and glycine betaine. J. Bacteriol. **170:**5208–5215.
92. **Everett, M., Y. F. Jin, V. Ricci, and L. J. V. Piddock.** 1996. Contribution of individual mechanisms to fluoroquinolone resistance in 36 *Escherichia coli* strains isolated from humans and animals. Antimicrob. Agents Chemother. **40:**2380–2386.
93. **Farrow, K. A., D. Lyras, and J. I. Rood.** 2001. Genomic analysis of the erythromycin resistance element Tn*5398* from *Clostridium difficile*. Microbiology **147:**2717–2728.
94. **Faust, B., D. Hoffmeister, G. Weitnauer, L. Westrich, S. Haag, P. Schneider, H. Decker, E. Kunzel, J. Rohr, and A. Bechthold.** 2000. Two new tailoring enzymes, a glycosyltransferase and an oxygenase, involved in biosynthesis of the angucycline antibiotic urdamycin A in. Streptomyces fradiae Tu2717. Microbiology. **146:**147–154.
95. **Felscher, D. W., and J. M. Bishop.** 1999. Reversible tumorogenesis by MYC in hematopoietic lineages. Mol. Cell **2:**451–461.
96. **Fernández-Moreno, M. A., J. L. Caballero, D. A. Hopwood, and F. Malpartida.** 1991. The *act* cluster contains regulatory and antibiotic export genes, direct targets for translational control by the *bldA* tRNA gene of *Streptomyces*. Cell **66:**769–780.
97. **Fidopiastis, P. M., C. M. Miyamoto, M. G. Jobling, E. A. Meighen, and E. G. Ruby.** 2002. LitR, a new transcriptional activator in *Vibrio fischeri*, regulates luminescence and symbiotic light organ colonization. Mol. Microbiol. **45:**131–143.
98. **Fischer, G. H., S.l. Wellen, D. Klimstra, J. M. Lenczowski, J. W. Tichelaar, M. J. Lizak, J. A. Whitsett, A. Korestsky, and H. E. Varmus.** 2001. Induction and apoptotic regression of lung adenocarcinomas by regulation of a K-Ras transgene in the presence and absence of tumor suppressor genes. Genes Dev. **15:**3249–3262.
99. **Folcher, M., R. P. Morris, G. Dale, K. Salah-Bey-Hocini, P. H. Viollier, and C. J. Thompson.** 2001. A transcriptional regulator of a pristinamycin resistance gene in *Streptomyces coelicolor*. J. Biol. Chem. **276:**1479–1485.
100. **Fomenko, D. E., A. Z. Metlitskaya, J. Péduzzi, C. Goulard, G. S. Katrukha, L. V. Gening, S. Rebuffat, and I. A. Khmel.** 2003. Microcin C51 plasmid genes: possible source of horizontal gene transfer. Antimicrob. Agents Chemother. **47:**2868–2874.
101. **Fong, S. T., J. Camakaris, and B. T. Lee.** 1995. Molecular genetics of a chromosomal locus involved in copper tolerance in *Escherichia coli* K-12. Mol. Microbiol. **15:**1127–1137.
102. **Fralick, J. A.** 1996. Evidence that TolC is required for functioning of the Mar/AcrAB efflux pump of *Escherichia coli*. J. Bacteriol. **178:**5803–5805.
103. **Friedberg, D., M. Midkiff, and J. M. Calvo.** 2001. Global versus local regulatory roles for Lrp-related proteins: *Haemophilus influenzae* as a case study. J. Bacteriol. **183:**4004–4011.
104. **Fromknecht, K., P. D. Vogel, and J. G. Wise.** 2003. Combinatorial redesign of the DNA binding specificity of a prokaryotic helix-turn-helix repressor. J. Bacteriol. **185:**475–481.
105. **Fujita, M., H. Aramaki, T. Horiuchi, and A. Amemura.** 1993. Transcription of the *cam* operon and *camR* genes in *Pseudomonas putida* PpG1. J. Bacteriol. **175:**6953–6958.
106. **Fuqua, W. C., S. C. Winans, and E. P. Greenberg.** 1994. Quorum sensing in bacteria: the LuxR-LuxI family of cell density-responsive transcriptional regulators. J. Bacteriol. **176:**269–275.
107. **Fuqua, C., S. C. Winans, and E. P. Greenberg.** 1996. Census and consensus in bacterial ecosystems: the LuxR-LuxI family of quorum-sensing transcriptional regulators. Annu. Rev. Microbiol. **50:**727–751.
108. **Gallegos, M. T., C. Michán, and J. L. Ramos.** 1993. The XylS/AraC family of regulators. Nucleic Acids Res. **21:**807–810.
109. **Gallegos, M. T., R. Schleif, A. Bairoch, K. Hofmann, and J. L. Ramos.** 1997.

AraC/XylS of transcriptional regulators. Microbiol. Mol. Biol. Rev. **61:** 393–410.
110. **Gambino, L., S. J. Gracheck, and P. F. Miller.** 1993. Overexpression of the MarA positive regulator is sufficient to confer multiple antibiotic resistance in *Escherichia coli*. J. Bacteriol. **175:**2888–2894.
111. **Garnier, T., K. Eiglmeier, J. C. Camus, N. Medina, H. Mansoor, M. Pryor, S. Duthoy, S. Grondin, C. Lacroix, C. Monsempe, S. Simon, B. Harris, R. Atkin, J. Doggett, R. Mayes, L. Keating, P. R. Wheeler, J. Parkhill, B. G. Barrell, S. T. Cole, S. V. Gordon, and R. G. Hewinson.** 2003. The complete genome sequence of *Mycobacterium bovis*. Proc. Natl. Acad. Sci. USA **100:** 7877–7882.
112. **George, A. M., and S. B. Levy.** 1983. Gene in the major cotransduction gap of the *Escherichia coli* K-12 linkage map required for the expression of chromosomal resistance to tetracycline and other antibiotics. J. Bacteriol. **178:**2507–2513.
113. **González-Pasayo, R., and E. Martínez-Romero.** 2000. Multiresistance genes of *Rhizobium etli* CFN42. Mol. Plant-Microbe Interact. **13:**572–577.
114. **Gossen, M., A. L. Bonin, and H. Bujard.** 1993. Control of gene activity in higher eukaryotes cells by prokaryotic regulatory elements. Trends Biochem. Sci. **18:**471–475.
115. **Gossen, M., and H. Bujard.** 1992. Tight control of gene expression in mammalian cells by tetracycline-responsive promoters. Proc. Natl. Acad. Sci. USA **89:**5547–5551.
116. **Gotz, F.** 2002. *Staphylococcus* and biofilms. Mol. Microbiol. **43:**1367–1378.
117. **Gray, K. M., and J. R. Garey.** 2001. The evolution of bacterial LuxI and LuxR quorum sensing regulators. Microbiology **147:**2379–2387.
118. **Green, J., C. Scott, and J. R. Guest.** 2001. Functional versatility in the CRP-FNR superfamily of transcription factors: FNR and FLP. Adv. Microb. Physiol. **44:**1–34.
119. **Grkovic, S., M. H. Brown, N. J. Roberts, I. T. Paulsen, and R. A. Skurray.** 1998. QacR is a repressor protein that regulates expression of the *Staphylococcus aureus* multidrug efflux pump QacA. J. Biol. Chem. **273:**18665–18673.
120. **Grkovic, S., M. H. Brown, M. A. Schumacher, R. G. Brennan, and R. A. Skurray.** 2001. Staphylococcal QacR multidrug regulator binds a correctly spaced operator as a pair of dimers. J. Bacteriol. **183:**7102–7109.
121. **Grkovic, S., M. H. Brown, and R. A. Skurray.** 2001. Transcriptional regulation of multidrug efflux pumps in bacteria. Cell Dev. Biol. **12:**225–237.
122. **Grkovic, S., K. M. Hardie, M. H. Brown, and R. A. Skurray.** 2003. Interactions of the QacR multidrug-binding protein with structurally diverse ligands: implications for the evolution of the binding pocket. Biochemistry **42:**15226–15236.
123. **Grogan, G., G. A. Roberts, D. Bougioukou, N. J. Turner, and S. L. Flitsch.** 2001. The desymmetrization of bicyclic beta-diketones by enzymatic retro-claisen reaction. A new reaction of the crotonase superfamily. J. Biol. Chem. **276:**12565–12572.
124. **Grohmann, E., G. Muth, and M. Espinosa.** 2003. Conjugative plasmid transfer in gram-positive bacteria. Microbiol. Mol. Biol. Rev. **67:**277–301.
125. **Guengerich, F. P.** 1991. Reactions and significance of cytochrome P-450 enzymes. J. Biol. Chem. **266:**10019–10022.
126. **Guilfoile, P. G., and C. R. Hutchinson.** 1992. Sequence and transcriptional analysis of the *Streptomyces glaucescens tcmAR* tetracenomycin C resistance and repressor gene loci. J. Bacteriol. **174:**3651–3658.
127. **Guilfoile, P. G., C. R. Hutchinson.** 1992. The *Streptomyces glaucescens* TcmR protein represses transcription of the divergently oriented *tcmR* and *tcmA* genes by binding to an intergenic operator region. J. Bacteriol. **174:** 3659–3666.
128. **Haft, D. H., B. J. Loftus, D. L. Richardson, F. Yang, J. A. Eisen, I. T. Paulsen, and O. White.** 2001. TIGRFAMs: a protein family resource for the functional identification of proteins. Nucleic Acids Res. **29:**41–43.
129. **Hagen, K. D., and J. C. Meeks.** 1999. Biochemical and genetic evidence for participation of DevR in a phosphorelay signal transduction pathway essential for heterocyst maturation in *Nostoc punctiforme* ATCC 29133. J. Bacteriol. **181:**4430–4434.
130. **Hagman, K. E., W. Pan, B. G. Spratt, J. T. Balthazar, R. C. Judd, and W. M. Shafer.** 1995. Resistance of *Neisseria gonorrhoeae* to antimicrobial agents is modulated by the *mtrRCDE* efflux system. Microbiology **141:** 611–622.
131. **Hagman, K. E., and W. M. Shafer.** 1995. Transcriptional control of the *mtr* efflux system of *Neisseria gonorrhoeae*. J. Bacteriol. **177:**4162–4165.
132. **Hansson, M., and L. Hederstedt.** 1992. Cloning and characterization of the *Bacillus subtilis hemEHY* gene cluster, which encodes protoheme IX biosynthetic enzymes. J. Bacteriol. **174:**8081–8093.
133. **Hara, O., and T. Beppu.** 1982. Mutants blocked in streptomycin production in *Streptomyces griseus*—the role of A-factor. J. Antibiot. **35:**349–358.
134. **Hara, O., and T. Beppu.** 1982. Induction of streptomycin-inactivating enzyme by A-factor in *Streptomyces griseus*. J. Antibiot. **35:**1208–1215.
135. **Harrison, S. C., and A. K. Aggarwal.** 1990. DNA recognition by proteins with the helix-turn-helix motif. Annu. Rev. Biochem. **59:**933–969.
136. **Harrison, S. C.** 1991. A structural taxonomy of DNA-binding domains. Nature **353:**715–719.
137. **Hashimoto, Y., M. Yamashita, and Y. Murooka.** 1997. The *Propionibacterium freudenreichii hemYHBXRL* gene cluster, which encodes enzymes and a regulator involved in the biosynthetic pathway from glutamate to protoheme. Appl. Microbiol. Biotechnol. **47:**385–392.
138. **Haydon, D. J., and J. R. Guest.** 1991. A new family of bacterial regulatory proteins. FEMS Microbiol. Lett. **63:**291–295.
139. **Hayes, A., G. Hobbs, C. P. Smith, S. G. Oliver, and P. R. Butler.** 1997. Environmental signals triggering methylenomycin production by *Streptomyces coelicolor* A3(2). J. Bacteriol. **179:**5511–5515.
140. **He, J. S., Q. Liang, and A. J. Fulco.** 1995. The molecular cloning and characterization of BM1P1 and BM1P2 proteins, putative positive transcription factors involved in barbiturate-mediated induction of the genes encoding cytochrome P450BM-1 of *Bacillus megaterium*. J. Biol. Chem. **270:** 18615–18625.
141. **Helbl, W., C. Berens, and W. Hillen.** 1995. Proximity probing of Tet repressor to *tet* operator by dimenthylsulfate reveals protected and accessible functions for each recognized base-pair in the major groove. J. Mol. Biol. **245:**538–548.
142. **Helbl, V., and W. Hillen.** 1998. Stepwise selection of TetR variants recognizing *tet* operator 4C with high affinity and specificity. J. Mol. Biol. **276:** 313–318.
143. **Helbl, V., B. Tiebel, and W. Hillen.** 1998. Stepwise selection of TetR variants recognizing *tet* operator 6C with high affinity and specificity. J. Mol. Biol. **276:**319–324.
144. **Helmann, J. D., B. T. Ballard, and C. T. Walsh.** 1990. The MerR metalo-regulatory protein binds mercuric ion as a tricoordinate, metal-bridged dimer. Science **247:**946–948.
145. **Henikoff, S., G. W. Haughn, J. M. Calvo, and J. C. Wallace.** 1988. A large family of bacterial activator proteins. Proc. Natl. Acad. Sci. USA **85:**6602–6606.
146. **Henikoff, S., J. C. Wallace, and J. P. Brown.** 1990. Finding protein similarities with nucleotide sequence databases. Methods Enzymol. **183:**111–132.
147. **Heuer, H., R. Szczepanowski, S. Schneiker, A. Pühler. E.M. Top, and A. Schlüter.** 2004. The complete sequences of plasmids pB2 and pB3 provide evidence for a recent ancestor of the IncP-1β group without any accessory genes. Microbiology **150:**3591–3599.
148. **Hillen, W., and C. Berens.** 1994. Mechanisms underlying expression of Tn*10* encoded tetracycline resistance. Annu. Rev. Microbiol. **48:**345–369.
149. **Hillen, W., K. Schollmeier, and C. Gatz.** 1984. Control of expression of the Tn*10*-encoded tetracycline resistance operon. II. Interaction of RNA polymerase and Tet repressor with the *tet* operon regulatory region. J. Mol. Biol. **172:**185–201.
150. **Hinrichs, W., C. Kisker, M. Düvel, A. Müller, K. Tovar, W. Hillen, and W. Saenger.** 1994. Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. Science **264:**418–420.
151. **Hirata, T., A. Saito, K. Nishino, N. Tamura, and A. Yamaguchi.** 2004. Effects of efflux transporter genes on susceptibility of *Escherichia coli* to tigecycline (GAR-936). Antimicrob. Agents Chemother. **48:**2179–2184.
152. **Hitomi, M., H. Nishimura, Y. Tsujimoto, H. Matsui, and K. Watanabe.** 2003. Identification of a helix-turn-helix motif of *Bacillus thermoglucosidasius* HrcA essential for binding to the CIRCE element and thermostability of the HrcA-CIRCE complex, indicating a role as a thermosensor. J. Bacteriol. **185:**381–385.
153. **Hobbs, G., A. I. Obanye, J. Petty, J. C. Mason, E. Barratt, D. C. Gardner, F. Flett, C. P. Smith, P. Broda, and S. G. Oliver.** 1992. An integrated approach to studying regulation of production of the antibiotic methylenomycin by *Streptomyces coelicolor* A3(2). J. Bacteriol. **174:**1487–1494.
154. **Hofmann, K., P. Bucher, L. Falquet, and A. Bairoch.** 1999. The Prosite database, its status in 1999. Nucleic Acids Res. **27:**215–219.
155. **Horinouchi, S., and T. Beppu.** 1990. Autoregulatory factors of secondary metabolism and morphogenesis in actinomycetes. Crit. Rev. Biotechnol. **10:** 191–204.
156. **Horinouchi, S., and T. Beppu.** 1994. A-factor as a microbial hormone that controls cellular differentiation and secondary metabolism in *Streptomyces griseus*. Mol. Microbiol. **12:**859–864.
157. **Horinouchi, S., Y. Ohnishi, and D. K. Kang.** 2001. The A-factor regulatory cascade and cAMP in the regulation of physiological and morphological development in *Streptomyces griseus*. J. Ind. Microbiol. Biotechnol. **27:**177–182.
158. **Horinouchi, S.** 2002. A microbial hormone, A-factor, as a master switch for morphological differentiation and secondary metabolism in *Streptomyces griseus*. Front. Biosci. **7:**2045–2057.
159. **Isackson, P. J., and K. P. Bertrand.** 1985. Dominant negative mutations in the Tn*10* Tet repressor: evidence for use of the conserved helix-turn-helix motif in DNA binding. Proc. Natl. Acad. Sci. USA **82:**6226–6230.
160. **Itoh, K., Y. Tashiro, K. Uobe, Y. Kamagata, K. Suyama, and H. Yamamoto.** 2004. Root nodule *Bradyrhizobium* spp. harbor *tfdA*α and *cadA*, homologous with genes encoding 2,4-dichlorophenoxyacetic acid-degrading proteins. Appl. Environ. Microbiol. **70:**2110–2118.
161. **Jakoby, M., L. Nolden, J. Meier-Wagner, R. Kramer, and A. Burkovski.** 2000. AmtR, a global repressor in the nitrogen regulation system of *Corynebacterium glutamicum*. Mol. Microbiol. **37:**964–977.
162. **Janssen, D. B., F. Pries, J. van der Ploeg, B. Kazemier, P. Terpstra, and B.**

Witholt. 1989. Cloning of 1,2-dichloroethane degradation genes of *Xanthobacter autotrophicus* GJ10 and expression and sequencing of the *dhlA* gene. J. Bacteriol. **171:**6791–6799.
163. **Jefferson, K. K., S. E. Cramton, F. Gotz, and G. B. Pier.** 2003. Identification of a 5-nucleotide sequence that controls expression of the *ica* locus in *Staphylococcus aureus* and characterization of the DNA-binding properties of IcaR. Mol. Microbiol. **48:**889–899.
164. **Jellen-Ritter, A. S., and W. V. Kern.** 2001. Enhanced expression of the multidrug efflux pumps AcrAB and AcrEF associated with insertion element transposition in *Escherichia coli* mutants selected with a fluoroquinolone. Antimicrob. Agents Chemother. **45:**1467–1472.
165. **Jobling, M. G., and R. K. Holmes.** 1997. Characterization of *hapR*, a positive regulator of the *Vibrio cholerae* HA/protease gene *hap*, and its identification as a functional homologue of the *Vibrio harveyi luxR* gene. Mol. Microbiol. **26:**1023–1034.
166. **Jogun, K. H., and J. J. Stezowski.** 1976. Coordination and conformational aspects of oxytetracycline metal ion complexation. J. Am. Chem. Soc. **98:** 6018–6026.
167. **Johnston, T. C., R. B. Thompson, and T. O. Baldwin.** 1986. Nucleotide sequence of the *luxB* gene of *Vibrio harveyi* and the complete amino acid sequence of the beta subunit of bacterial luciferase. J. Biol. Chem. **261:** 4805–4811.
168. **Kamionka, A., J. Bogdanska-Urbaniak, O. Scholz, and W. Hillen.** 2004. Two mutations in the tetracycline repressor change the inducer anhydrotetracycline to a corepressor. Nucleic Acids Res. **32:**842–847.
169. **Kaneko, M., A. Yamaguchi, and T. Sawai.** 1985. Energetics of tetracycline efflux system encoded by Tn*10* in *Escherichia coli*. FEBS Lett. **193:**893–899.
170. **Karniol, B., and R. D. Vierstra.** 2004. The HWE histidine kinases, a new family of bacterial two-component sensor kinases with potentiallly diverse roles in environmental signaling. J. Bacteriol. **186:**445–453.
171. **Kasuga, K., H. Nojiri, H. Yamane, T. Kodama, and T. Omori.** 1997. Cloning and characterization of genes involved in the degradation of dibenzofuran by *Terrabacter* sp. strain DBF63. J. Ferment. Bioeng. **84:**387–399.
172. **Katayama, J., and N. Noguchi.** 1999. Nucleotide sequence of the gene cluster containing the *mphB* gene for macrolide 2′-phosphotransferase II. Biol. Pharm. Bull. **22:**227–228.
173. **Kato, J.-Y., I. Miyahisa, M. Mashiko, Y. Ohnishi, and S. Horinouchi.** 2004. A single target is sufficient to account for biological effects of the A-factor receptor protein of *Streptomyces griseus*. J. Bacteriol. **186:**2206–2221.
174. **Kato, J., A. Suzuki, H. Yamazaki, Y. Ohnishi, and S. Horinouchi.** 2002. Control by A-factor of a metalloendopeptidase gene involved in acrial mycelium formation in *Streptomyces griseus*. J. Bacteriol. **184:**6010–6025.
175. **Kawachi, R., T. Akashi, Y. Kamitani, A. Sy, U. Wangchaisoonthorn, T. Nihira, and Y. Yamada.** 2000. Identification of an AfsA homologue (BarX) from *Streptomyces virginiae* as a pleiotropic regulator controling autoregulator biosynthesis, virginiamycin biosynthesis and virginiamycin M1 resistance. Mol. Microbiol. **36:**302–313.
176. **Kawamura-Sato, K., K. Shibayama, T. Horii, Y. Iimuma, Y. Arakawa, and M. Ohta.** 1999. Role of multiple efflux pumps in *Escherichia coli* in indole expulsion. FEMS Microbiol. Lett. **179:**345–352.
177. **Kieboom, J., R. Bruinenberg, I. Keizer-Gunnink, and J. A. de Bont.** 2001. Transposon mutations in the flagella biosynthetic pathway of the solvent-tolerant *Pseudomonas putida S12* result in a decreased expression of solvent efflux genes. FEMS Microbiol. Lett. **198:**117–122.
178. **Kieboom, J., and J. de Bont.** 2001. Identification and molecular characterization of an efflux system involved in *Pseudomonas putida* S12 multidrug resistance. Microbiology **147:**43–51.
179. **Kieboom, J., J. J. Dennis, J.A. de Bont, and G. J. Zylstra.** 1998. Identification and molecular characterization of an efflux pump involved in *Pseudomonas putida* S12 solvent tolerance. J. Biol. Chem. **273:**85–91.
180. **Kieboom, J., J. J. Dennis, G. J. Zylstra, J.A. de Bont.** 1998. Active efflux of organic solvents by *Pseudomonas putida S12* is induced by solvents. J. Bacteriol. **180:**6769–6772.
181. **Kim, Y. B., J. Okuda, C. Matsumoto, N. Takahashi, S. Hashimoto, and M. Nishibuchi.** 1999. Identification of *Vibrio parahaemolyticus* strains at the species level by PCR targeted to the *toxR* gene. J. Clin. Microbiol. **37:** 1173–1177.
182. **Kinoshita, H., H. Ipposhi, S. Okamoto, H. Nakano, T. Nihira, and Y. Yamada.** 1997. Butyrolactone autoregulator receptor protein (BarA) as a transcriptional regulator in *Streptomyces virginiae*. J. Bacteriol. **179:**6986–6993.
183. **Kisker, C., W. Hinrichs, K. Tovar, W. Hillen, and W. Saenger.** 1995. The complex formed between Tet repressor and tetracycline-Mg^{2+} reveals mechanism of antibiotic resistance. J. Mol. Biol. **247:**260–280.
184. **Kitani, S., H. Kinoshita, T. Nihira, and Y. Yamada, Y.** 1999. In vitro analysis of the butyrolactone autoregulator receptor protein (FarA) of *Streptomyces lavendulae* FRI-5 reveals that FarA acts as a DNA-binding transcriptional regulator that controls its own synthesis. J. Bacteriol. **181:** 5081–5084.
185. **Kitani, S., Y. Yamada, and T. Nihira.** 2001. Gene replacement analysis of the butyrolactone autoregulator receptor (FarA) reveals that FarA acts as a novel regulator in secondary metabolism of *Streptomyces lavendulae* FRI-5. J. Bacteriol. **183:**4357–4363.
186. **Klein, J. R., B. Henrich, and R. Plapp.** 1991. Molecular analysis and nucleotide sequence of the *envCD* operon of *Escherichia coli*. Mol. Gen. Genet. **230:**230–240.
187. **Klein, J. R., and R. Plapp.** 1992. Locations of the *envCD* genes on the physical map of the *Escherichia coli* chromosome. J. Bacteriol. **174:**3828–3829.
188. **Klinke, S., G. de Roo, B. Witholt, and B. Kessler.** 2000. Role of *phaD* in accumulation of medium-chain-length poly(3-hydroxyalkanoates) in *Pseudomonas oleovorans*. Appl. Environ. Microbiol. **66:**3705–3710.
189. **Klock, G., B. Unger, C. Gatz, W. Hillen, J. Altenbuchner, K. Schmid, and R. Schmitt.** 1985. Heterologous repressor-operator recognition among four classes of tetracycline resistance determinants. J. Bacteriol. **161:**326–332.
190. **Knobloch, J. K., K. Bartscht, A. Sabottke, H. Rohde, H. H. Feucht, and D. Mack.** 2001. Biofilm formation by *Staphylococcus epidermidis* depends on functional RsbU, an activator of the *sigB* operon: differential activation mechanisms due to ethanol and salt stress. J. Bacteriol. **183:**2624–2633.
191. **Kobayashi, K., N. Tsukagoshi, and R. Aono.** 2001. Suppression of hypersensitivity of *Escherichia coli acrB* mutant to organic solvents by integrational activation of the *acrEF* operon with the IS*1* or IS*2* element. J. Bacteriol. **183:**2646–2653.
192. **Kojic, M., and V. Venturi, V.** 2001. Regulation of *rpoS* gene expression in *Pseudomonas*: involvement of a TetR family regulator. J. Bacteriol. **183:** 3712–3720.
193. **Koronakis, V., A. Sharff, E. Koronakis, B. Luisi, and C. Hughes.** 2000. Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. Nature **405:**914–919.
194. **Kovacikova, G., and K. Skorupski.** 2002. Regulation of virulence gene expression in *Vibrio cholerae* by quorum sensing: HapR functions at the *aphA* promoter. Mol. Microbiol. **46:**1135–1147.
195. **Krug, A., V. F. Wendisch, and M. Bott.** 2005. Identification of AcnR, a TetR-type repressor of the acotinase gene *acn* in *Corynebacterium glutamicum*. J. Biol. Chem. **280:**585–595.
196. **Krukonis, E. S., R. R. Yu, and V. J. Dirita.** 2000. The *Vibrio cholerae* ToxR/TcpP/ToxT virulence cascade: distinct roles for two membrane-localized transcriptional activators on a single promoter. Mol. Microbiol. **38:** 67–84.
197. **Kumano, M., A. Tamakoshi, and K. Yamane.** 1997. A 32 kb nucleotide sequence from the region of the lincomycin-resistance gene (22 degrees-25 degrees) of the *Bacillus subtilis* chromosome and identification of the site of the *lin-2* mutation. Microbiology **143:**2775–2782.
198. **Kumano, M., M. Fujita, K. Nakamura, M. Murata, R. Ohki, and K. Yamane.** 2003. Lincomycin resistance mutations in two regions immediately downstream of the −10 region of *lmr* promoter cause overexpression of a putative multidrug efflux pump in *Bacillus subtilis* mutants. Antimicrob. Agents Chemother. **47:**432–435.
199. **Kunin, C. M., T. H. Hua, W. L. Van Arsdale, and M. Villarejo.** 1992. Growth of *Escherichia coli* in human urine: role of salt tolerance and accumulation of glycine betaine. J. Infect. Dis. **166:**1311–1315.
200. **Kustu, S., A. K. North, and D. S. Weiss.** 1991. Prokaryotic transcriptional enhancers and enhancer-binding proteins. Trends Biochem. Sci. **16:**397–402.
201. **Lamark, T., I. Kaasen, M. W. Eshoo, P. Falkenberg, J. McDougall, and A. R. Strom.** 1995. DNA sequence and analysis of the *bet* genes encoding the osmoregulatory choline-glycine betaine pathway of *Escherichia coli*. Mol. Microbiol. **5:**1049–1064.
202. **Lamark, T., T. P. Rokenes, J. McDougall, and A. R. Strom.** 1996. The complex *bet* promoters of *Escherichia coli*: regulation by oxygen (ArcA), choline (BetI), and osmotic stress. J. Bacteriol. **178(6):**1655–1662.
203. **Lamark, T., O. B. Styrvold, and A. R. Strøm.** 1992. Efflux of choline and glycine betaine from osmoregulating cells of *Escherichia coli*. FEMS Microbiol. Lett. **96:**149–154.
204. **Landfald, B., and A. R. Strom.** 1986. Choline-glycine betaine pathway a high level of osmotic tolerance in *Escherichia coli*. J. Bacteriol. **165(3):** 849–855.
205. **La Teana, A., A. Brandi, M. Falconi, R. Spurio, C. L. Pon, and C. O. Gualerzi.** 1991. Identification of a cold shock transcriptional enhancer of the *Escherichia coli* gene encoding nucleoid protein H-NS. Proc. Natl. Acad. Sci. USA **88:**10907–10911.
206. **Lau, P. C. K., Y. Wang, A. Patel, D. Labbe, H. Bergeron, R. Brousseau, Y. Konishi, and M. Rawlings.** 1997. A bacterial basic region leucine zipper histidine kinase regulating toluene degradation. Proc. Natl. Acad. Sci. USA **94:**1453–1458.
207. **Lauble, H., Y. Georgalis, and U. Heinemann.** 1989. Studies on the domain structure of the *Salmonella typhimurium* AraC protein. Eur. J. Biochem. **185:**319–325.
208. **Lee, E. H., C. Rouquette-Loughlin, J. P. Folster, and W. M. Shafer.** 2003. FarR regulates the *farAB*-encoded efflux pump of *Neisseria gonorrhoeae* via an MtrR regulatory mechanism. J. Bacteriol. **185:**7145–7152.
209. **Levy, S. B.** 1984. Resistance to tetracyclines p. 191–204. *In* L. E. Bryan (ed.), Antimicrobial drug resistance. Academic Press, New York, N.Y.

210. **Levy, S. B.** 1988. Tetracycline resistance determinants are widespread. ASM News. **54:**418–421.
211. **Levy, S. B., L. M. McMurry, T. M. Barbosa, V. Burdett, P. Courvalin, W. Hillen, M. C. Roberts, J. I. Rood, and D. E. Taylor.** 1999. Nomenclature for new tetracycline resistance determinants. Antimicrob. Agents Chemother. **43:**1523–1524.
212. **Li, X.-Z., L. Zhang, and H. Nikaido.** 2004. Efflux pump-mediated intrinsic drug resistance in *Mycobacterium smegmatis*. Antimicrob. Agents Chemother. **48:**2415–2423.
213. **Liang, Q., L. Chen, and A. J. Fulco.** 1998. *In vivo* roles of Bm3R1 repressor in the barbiturate-mediated induction of the cytochrome P450 genes (P450(BM-3) and P450(BM-)1) of *Bacillus megaterium*. Biochim. Biophys. Acta **1380:**183–197.
214. **Liang, Q., and A. J. Fulco.** 1995. Transcriptional regulation of the genes encoding cytochromes P450BM-1 and P450BM-3 in *Bacillus megaterium* by the binding of Bm3R1 repressor to Barbie box elements and operator sites. J. Biol. Chem. **270:**18606–18614.
215. **Liang, Q., J. S. He, and A. J. Fulco, A. J.** 1995. The role of Barbie box sequences as *cis*-acting elements involved in the barbiturate-mediated induction of cytochromes P450BM-1 and P450BM-3 in *Bacillus megaterium*. J. Biol. Chem. **270:**4438–4450.
216. **Lin, Y. H., C. Miyamoto, and E. A. Meighen.** 2000. Purification and characterization of a *luxO* promoter binding protein LuxT from *Vibrio harveyi*. Protein Expr. Purif. **20:**87–94.
217. **Lin, Z., K. Kumagai, K. Baba, J. J. Mekalanos, and M. Nishibuchi.** 1993. *Vibrio parahaemolyticus* has a homolog of the *Vibrio cholerae toxRS* operon that mediates environmentally induced regulation of the thermostable direct hemolysin gene. J. Bacteriol. **175:**3844–3855.
218. **Liu, H., and K. A. Reynolds.** 1999. Role of crotonyl coenzyme A reductase in determining the ratio of polyketides monensin A and monensin B produced by *Streptomyces cinnamonensis*. J. Bacteriol. **181:**6806–6813.
219. **Liu, H., and K. A. Reynolds.** 2001. Precursor supply for polyketide biosynthesis: the role of crotonyl-CoA reductase. Metab. Eng. **3:**40–48.
220. **Lucas, C. E., K. E. Hagman, J. C. Levin, D. C. Stein, and W. M. Shafer.** 1995. Importance of lipooligosaccharide structure in determining gonococcal resistance to hydrophobic antimicrobial agents resulting from the *mtr* efflux system. Mol. Microbiol. **16:**1001–1009.
221. **Lucas, C. E., J. T. Balthazar, K. E. Hagman, and W. M. Shafer.** 1997. The MtrR repressor binds the DNA sequence between the *mtrR* and *mtrC* genes of *Neisseria gonorrhoeae*. J. Bacteriol. **179:**4123–4128.
222. **Ly, A., J. Henderson, A. Lu, D. E. Culham, and J. M. Wood.** 2004. Osmoregulatory systems of *Escherichia coli*: identification of betaine-carnitine-choline transporter family member BetU and distributions of *betU* and *trkG* among pathogenic and nonpathogenic isolates. J. Bacteriol. **186:**296–306.
223. **Ma, D., M. Alberti, C. Lynch, H. Nikaido, and J. E. Hearst.** 1996. The local repressor AcrR plays a modulating role in the regulation of *acrAB* genes of *Escherichia coli* by global stress signals. Mol. Microbiol. **19:**101–112.
224. **Ma, D., M. Alberti, C. Lynch, N. G. Pon, H. Nikaido, and J. E. Hearst.** 1995. Genes *acrAB* encode a stress-induced efflux system of *Escherichia coli*. Mol. Microbiol. **16:**45–55.
225. **Ma, D., D. N. Cook, M. Alberti, N. G. Pon, H. Nikaido, and J. E. Hearst.** 1993. Molecular cloning and characterization of *acrA* and *acrE* genes of *Eschericia coli*. J. Bacteriol. **175:**6299–6313.
226. **Ma, D., D. N. Cook, J. Hearst, and H. N. Nikaido.** 1994. Efflux pumps and drug resistance in Gram-negative bacteria. Trends Microbiol. **2:**489–493.
227. **Mahren, S., and W. Braun.** 2003. The FecI extracytoplasmic-function sigma factor of *Escherichia coli* interacts with the β'subunit of RNA polymerase. J. Bacteriol. **185:**1796–1802.
228. **Maness, M. J., and P. F. Sparling.** 1973. Multiple antibiotic resistance due to a single mutation in *Neisseria gonorrhoeae*. J. Infect. Dis. **128:**321–330.
229. **Manzanera, M., S. Marqués, and J. L. Ramos.** 2000. Mutational analysis of the highly conserved C-terminal residues of the XylS protein, a member of the AraC family of transcriptional regulators. FEBS Lett. **476:**312–317.
230. **Marger, M. D., and M. H. Saier, Jr.** 1993. A major superfamily of transmembrane facilitators that catalyse uniport, symport and antiport. Trends Biochem. Sci. **18:**13–20.
231. **Marqués, S., M. T. Gallegos, M. Manzanera, A. Holtel, K. N. Timmis, and J. L. Ramos.** 1998. Activation and repression of transcription at the double tandem divergent promoters for the *xylR* and *xylS* genes of the TOL plasmid of *Pseudomonas putida*. J. Bacteriol. **180:**2889–2894.
232. **Martin, M., R. Showalter, and M. Silverman.** 1989. Identification of a locus controlling expression of luminescence genes in *Vibrio harveyi*. J. Bacteriol. **171:**2406–2414.
233. **Martínez-Argudo, I., P. Salinas, R. Maldonado, and A. Contreras.** 2002. Domain interactions on the *ntr* signal transduction pathway: two-hybrid analysis of mutant and truncated derivatives of histidine kinase NtrB. J. Bacteriol. **184:**200–206.
234. **Martínez-Argudo, I., and A. Contreras.** 2002. PII T-Loop mutations affecting signal transduction to NtrB also abolish yeast two-hybrid interactions. J. Bacteriol. **184:**3746–3748.
235. **Martinez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L., Tobes, R.** 2004. BacTregulators: a database of transcriptional regulators in bacteria and archaea. Bioinformatics **20:**2787–2791.
236. **Martínez-Bueno, M., Tobes, R., Rey, M., and Ramos, J. L.** 2002. Detection of multiple extracytoplasmic function (ECF) sigma factors in the genome of *Pseudomonas putida* KT2440 and their counterparts in *Pseudomonas aeruginosa* PA01. Environ. Microbiol. **4:**842–855.
237. **Martinez-Hackert, E., and A. M. Stock.** 1997. Structural relationships in the OmpR family of winged-helix transcription factors. J. Mol. Biol. **269:**301–312.
238. **Mavroidi, A., L. S. Tzouvelekis, K. P. Kyriakis, H. Avgerinou, M. Danilidou, and E. Tzelepi.** 2001. Multidrug-resistant strains of *Neisseria gonorrhoeae* in Greece. Antimicrob. Agents Chemother. **45:**2651–2654.
239. **Mayer, K., G. Murphy, R. Tarchini, R. Wambutt, G. Volckaert, T. Pohl, A. Duesterhoeft, W. Stiekema, K. D. Entian, N. Terryn, K. Lemcke, D. Haase, C. R. Hall, A.M. van Dodeweerd, S. V. Tingey, H. W. Mewes, M. W. Bevan, and I. Bancroft.** 2001. Conservation of microstructure between a sequenced region of the genome of rice and multiple segments of the genome of *Arabidopsis thaliana*. Genome Res. **11:**1167–1174.
240. **McCarter, L. L.** 1998. OpaR, a homolog of *Vibrio harveyi* LuxR, controls opacity of *Vibrio parahaemolyticus*. J. Bacteriol. **180:**3166–3173.
241. **McCarter, L.** 1999. The multiple identities of *Vibrio parahaemolyticus*. J. Mol. Microbiol. Biotechnol. **1:**51–57.
242. **McDougald, D., S. A. Rice, and S. Kjelleberg.** 2000. The marine pathogen *Vibrio vulnificus* encodes a putative homologue of the *Vibrio harveyi* regulatory gene, *luxR*: a genetic and phylogenetic comparison. Gene. **248:**213–221.
243. **McDougald, D., S. A. Rice and S. Kjelleberg.** 2001. SmcR-dependent regulation of adaptive phenotypes in *Vibrio vulnificus*. J. Bacteriol. **183:**758–762.
244. **McKay, D. B., and T. A. Steitz.** 1981. Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. Nature **290:**744–749.
245. **McMurry, L. M., M. Oethinger, and S. B. Levy.** 1998. Overexpression of *marA*, soxS, or *acrAB* produces resistance to triclosan in laboratory and clinical strains of *Escherichia coli*. FEMS Microbiol. Lett. **166:**305–309.
246. **Méndez, B., C. Tachibana, and S. B. Levy.** 1980. Heterogeneity of tetracycline resistance determinants. Plasmid **3:**99–108.
247. **Miller, M. B., K. Skorupski, D. H. Lenz, R. K. Taylor, and B. L. Bassler.** 2002. Parallel quorum sensing systems converge to regulate virulence in *Vibrio cholerae*. Cell **110:**303–314.
248. **Milton, D. L., A. Hardman, M. Camara, S. R. Chhabra, B. W. Bycroft, G. S. Stewart, and P. William.** 1997. Quorum sensing in *Vibrio anguillarum*: characterization of the *vanI/vanR* locus and identification of the autoinducer *N*-(3-oxodecanoyl)-L-homoserine lactone. J. Bacteriol. **179:**3004–3012.
249. **Mitchell, B. A., I. T. Paulsen, M. H. Brown, and R. A. Skurray.** 1999. Bioenergetics of the staphylococcal multidrug export protein QacA. Identification of distinct binding sites for monovalent and divalent cations. J. Biol. Chem. **274:**3541–3548.
250. **Miyamoto, C. M., M. Boylan, A. F. Graham, and E. A. Meighen.** 1988. Organization of the *lux* structural genes of *Vibrio harveyi*. Expression under the T7 bacteriophage promoter, mRNA analysis, and nucleotide sequence of the *luxD* gene. J. Biol. Chem. **263:**13393–13399.
251. **Miyamoto, C. M., J. Chatterjee, E. Swartzman, R. Szittner, and E. A. Meighen.** 1996. The role of lux autoinducer in regulating luminescence in *Vibrio harveyi*; control of *luxR* expression. Mol. Microbiol. **19:**767–775.
252. **Miyamoto, C. M., Y. H. Lin, and E. A. Meighen.** 2000. Control of bioluminescence in *Vibrio fischeri* by the LuxO signal response regulator. Mol. Microbiol. **36:**594–607.
253. **Miyamoto, C. M., E. A. Meighen, and A. F. Graham.** 1990. Transcriptional regulation of *lux* genes transferred into *Vibrio harveyi*. J. Bacteriol. **172:** 2046–2054.
254. **Miyamoto, C. M., E. E. Smith, E. Swartzman, J. G. Cao, A. F. Graham, and E. A. Meighen.** 1994. Proximal and distal sites bind LuxR independently and activate expression of the *Vibrio harveyi lux* operon. Mol. Microbiol. **14:** 255–262.
255. **Miyamoto, C. M., W. Sun, and E. A. Meighen.** 1998. The LuxR regulator protein controls synthesis of polyhydroxybutyrate in *Vibrio harveyi*. Biochim. Biophys. Acta **1384:**356–364.
256. **Mohamed, M. S., W. Ismail, J. Heider, and G. Fuchs.** 2002. Aerobic metabolism of phenylacetic acids in *Azoarcus evansii*. Arch. Microbiol. **178:** 180–192.
257. **Morett, E., and L. Segovia.** 1993. The sigma 54 bacterial enhancer-binding protein family: mechanism of action and phylogenetic relationship of their functional domains. J. Bacteriol. **175:**6067–6074.
258. **Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan, and E. M. Zdobnov.** 2003. The InterPro Data-

base, 2003 brings increased coverage and new features. Nucleic Acids Res. **31:**315–318.
259. **Muller, G., B. Hecht, V. Helbl, W. Hinrichs, W. Saenger, and W. Hillen.** 1995. Characterization of non-inducible Tet repressor mutants suggests conformational changes necessary for induction. Nat. Struct. Biol. **2:**693–703.
260. **Murata, M., S. Ohno, M. Kumano, K. Yamane, and R. Ohki.** 2003. Multidrug resistant phenotype of *Bacillus subtilis* spontaneous mutants isolated in the presence of puromycin and lincomycin. Can. J. Microbiol. **49:**71–77.
261. **Murray, D. A., M. A. Schumacher, and R. G. Brennan.** 2004. Crystal structures of QacR-diamidine complexes reveal additional multidrug-binding modes and a novel mechanism of drug charge neutralization. J. Biol. Chem. **279:**14365–14371.
262. **Nakano, H., E. Takehara, T. Nihira, and Y. Yamada.** 1998. Gene replacement analysis of the *Streptomyces virginiae barA* gene encoding the butyrolactone autoregulator receptor reveals that BarA acts as a repressor in virginiamycin biosynthesis. J. Bacteriol. **180:**3317–3322.
263. **Namwat, W., C. K. Lee, H. Kinoshita, Y. Yamada, and T. Nihira.** 2001. Identification of the *varR* gene as a transcriptional regulator of virginiamycin S resistance in Streptomyces virginiae. J. Bacteriol. **183:**2025–2031.
264. **Natsume, R., Y. Ohnishi, T. Senda, and S. Horinouchi.** 2004. Crystal structure of a γ-butyrolactone autoregulator receptor protein in *Streptomyces coelicolor* A3(2). J. Mol. Biol. **336:**409–419.
265. **Neal, R. J., and K. F. Chater.** 1987. Nucleotide sequence analysis reveals similarities between proteins determining methylenomycin A resistance in *Streptomyces* and tetracycline resistance in *eubacteria*. Gene **58:**229–241.
266. **Neal, R. J., and K. F. Chater.** 1991. Bidirectional promoter and terminator regions bracket *mmr*, a resistance gene embedded in the *Streptomyces coelicolor* A3(2) gene cluster encoding methylenomycin production. Gene **100:**75–83.
267. **Nikaido, H. N.** 1996. Multidrug efflux pumps of Gram-negative bacteria. J. Bacteriol. **178:**5853–5859.
268. **Nikaido, H. N.** 1998. Antibiotic resistance caused by Gram-negative multidrug efflux pumps. Clin. Infect. Dis. **27:**532–541.
269. **Nilsson, D., and M. Kilstrup.** 1998. Cloning and expression of the *Lactococcus lactis purDEK* genes, required for growth in milk. Appl. Environ. Microbiol. **64:**4321–4327.
270. **Nishino, K., and A. Yamaguchi.** 2001. Analysis of a complete library of putative drug transporter genes in *Escherichia coli*. J. Bacteriol. **183:**5803–5812.
271. **Nixon, B. T., C. W. Ronson, and F. M. Ausubel.** 1986. Two-component regulatory systems responsive to environmental stimuli share strongly conserved domains with the nitrogen assimilation regulatory genes *ntrB* and *ntrC*. Proc. Natl. Acad. Sci. USA **83:**7850–7854.
272. **Noguchi, N., J. Katayama, and M. Sasatsu.** 2000. A transposon carrying the gene *mphB* for macrolide 2′-phosphotransferase II. FEMS Microbiol. Lett. **192:**175–178.
273. **Noguchi, N., K. Takada, J. Katayama, A. Emura, and M. Sasatsu.** 2000. Regulation of transcription of the *mph*(A) gene for macrolide 2′-phosphotransferase I in *Escherichia coli*: characterization of the regulatory gene *mphR*(A). J. Bacteriol. **182:**5052–5058.
274. **North, A. K., K. E. Klose, K. M. Stedman, and S. Kustu.** 1993. Prokaryotic enhancer-binding proteins reflect eukaryote-like modularity: the puzzle of nitrogen regulatory protein C. J. Bacteriol. **175:**4267–4273.
275. **Novakova, R., J. Bistakova, D. Homerova, B. Rezuchova, and J. Kormanec.** 2002. Cloning and characterization of a polyketide synthase gene cluster involved in biosynthesis of a proposed angucycline-like polyketide auricin in *Streptomyces aureofaciens* CCM 3239. Gene **297:**197–208.
276. **Oethinger, M., I. Podglajen, W. V. Kern, and S. B. Levy.** 1998. Overexpression of the *marA* or *soxS* regulatory gene in clinical topoisomerase mutants of *Escherichia coli*. Antimicrob. Agents Chemother. **42:**2089–2094.
277. **Ohki, R., K. Tateno, Y. Okada, H. Okajima, K. Asai, Y. Sadaie, M. Murata, and T. Aiso.** 2003. A bacitracin-resistant *Bacillus subtilis* gene encodes a homologue of the membrane-spanning subunit of the *Bacillus licheniformis* ABC transporter. J. Bacteriol. **185:**51–59.
278. **Ohnishi, Y., S. Kameyama, H. Onaka, and S. Horinouchi.** 1999. The A-factor regulatory cascade leading to streptomycin biosynthesis in *Streptocyces griseus*: identification of a target gene of the A-factor receptor. Mol. Microbiol. **34:**102–111.
279. **Ohta, Y., M. Maeda, and T. Kudo.** 2001. *Pseudomonas putida* CE2010 can degrade biphenyl by a mosaic pathway encoded by the *tod* operon and *cmtE*, which are identical to those of *P. putida* F1 except for a single base difference in the operator-promoter region of the *cmt* operon. Microbiology **147:**31–41.
280. **Okamoto, S., K. Nakamura, T. Nihira, and Y. Yamada.** 1995. Virginiae butanolide binding protein from *Streptomyces virginiae*. Evidence that VbrA is not the virginiae butanolide binding protein and reidentification of the true binding protein. J. Biol. Chem. **270:**12319–12326.
281. **Okusu, H., D. Ma, and H. N. Nikaido.** 1996. AcrAB efflux pump plays a major role in the antibiotic resistance phenotype of *Escherichia coli* multiple-antibiotic-resistance (Mar) mutants. J. Bacteriol. **178:**306–308.
282. **Onaka, H., N. Ando, T. Nihira, Y. Yamada, T. Beppu, and S. Horinouchi.** 1995. Cloning and characterization of the A-factor receptor gene from *Streptomyces griseus*. J. Bacteriol. **177:**6083–6092.
283. **Onaka, H., and S. Horinouchi.** 1997. DNA-binding activity of the A-factor receptor protein and its recognition DNA sequences. Mol. Microbiol. **24:**991–1000.
284. **Onaka, H., T. Nakagawa, and S. Horinouchi.** 1998. Involvement of two A-factor receptor homologues in *Streptomyces coelicolor* A3(2) in the regulation of secondary metabolism and morphogenesis. Mol. Microbiol. **28:**743–753.
285. **Onaka, H., M. Sugiyama, and S. Horinouchi.** 1997. A mutation at proline-115 in the A-factor receptor protein of *Streptomyces griseus* abolishes DNA-binding ability but not ligand-binding ability. J. Bacteriol. **179:**2748–2752.
286. **Orth, P., F. Cordes, D. Schnappinger, W. Hillen, W. Saenger, and W. Hinrichs.** 1998. Conformational changes of the Tet repressor induced by tetracycline trapping. J. Mol. Biol. **279:**439–447.
287. **Orth, P., W. Saenger, and W. Hinrichs.** 1999. Tetracycline-chelated $Mg2^+$ ion initiates helix unwinding in Tet repressor induction. Biochemistry **38:**191–198.
288. **Orth, P., D. Schnappinger, W. Hillen, W. Saenger, and W. Hinrichs.** 2000. Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. Nat. Struct. Biol. **7:**215–219.
289. **Orth, P., D. Schnappinger, P.-E. Sum, G. A. Ellestad, W. Hillen, W. Saenger, and W. Hinrichs.** 1999. Crystal structure of the Tet repressor in complex with a novel tetracycline, 9-(*N*,*N*-dimethylglycylamido)-6-demethyl-6-deoxy-tetracycline. J. Mol. Biol. **285:**455–461.
290. **Osteras, M., E. Boncompagni, N. Vincent, M. C. Poggi, and D. Le Rudulier.** 1998. Presence of a gene encoding choline sulfatase in *Sinorhizobium meliloti bet* operon: choline-O-sulfate is metabolized into glycine betaine. Proc. Natl. Acad. Sci. USA **95:**11394–11399.
291. **Otten, S. L., J. Ferguson, and C. R. Hutchinson.** 1995. Regulation of daunorubicin production in *Streptomyces peucetius* by the *dnrR*2 locus. J. Bacteriol. **177:**1216–1224.
292. **Pabo, C. O., and M. Lewis.** 1992. The operator-binding domain of lambda repressor: structure and DNA recognition. Nature **298:**443–447.
293. **Pabo, C. O., and R. T. Sauer.** 1984. Protein-DNA recognition. Annu. Rev. Biochem. **53:**293–321.
294. **Pabo, C. O., and R. T. Sauer.** 1992. Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Microbiol. **61:**1053–1095.
295. **Palmer, C. N., E. Axen, V. Hughes, and C. R. Wolf.** 1998. The repressor protein, Bm3R1, mediates an adaptive response to toxic fatty acids in *Bacillus megaterium*. J. Biol. Chem. **273(29):**18109–18116.
296. **Palumbo, J. D., C. I. Kado, and D. A. Phillips.** 1998. An isoflavonoid-inducible efflux pump in *Agrobacterium tumefaciens* is involved in competitive colonization of roots. J. Bacteriol. **180:**3107–3113.
297. **Pan, W., and B. G. Spratt.** 1994. Regulation of the permeability of the gonococcal cell envelope by the *mtr* system. Mol. Microbiol. **11:**769–775.
298. **Pao, G. M., R. Tam, L. S. Lipschitz, and M. H. Saier, Jr.** 1994. Response regulators: structure, function and evolution. Res. Microbiol. **145:**356–362.
299. **Paulsen, I. T., M. H. Brown, S. J. Dunstan, and R. A. Skurray.** 1995. Molecular characterization of the staphylococcal multidrug resistance export protein QacC. J. Bacteriol. **177:**2827–2833.
300. **Paulsen, I. T., M. H. Brown, T. G. Littlejohn, B. A. Mitchell, and R. A. Skurray.** 1996. Multidrug resistance proteins QacA and QacB from *Staphylococcus aureus*: membrane topology and identification of residues involved in substrate specificity. Proc. Natl. Acad. Sci. USA **93:**3630–3635.
301. **Peng, W. T., and E. W. Nester.** 2001. Characterization of a putative RND-type efflux system in *Agrobacterium tumefaciens*. Gene **270:**245–252.
302. **Pérez-Rueda, E., and J. Collado-Vides.** 2001. Common history at the origin of the position–function correlation in transcriptional regulators in archaea and bacteria. J. Mol. Evol. **53:**172–179.
303. **Perron, K., O. Caille, C. Rossier, C. van Delden, J. L. Dumas, and T. Köhler.** 2004. CzcR-CzcS, a two-component system involved in heavy metal and carbapenem resistance in *Pseudomonas aeruginosa*. J. Biol. Chem. **279:**8761–8768.
304. **Pocard, J. A., T. Bernard, L. T. Smith, and D. Le Rudulier.** 1989. Characterization of three choline transport activities in *Rhizobium meliloti*: modulation by choline and osmotic stress. J. Bacteriol. **171:**531–537.
305. **Pocard, J. A., N. Vincent, E. Boncompagni, L. T. Smith, M. C. Poggi, and D. Le Rudulier.** 1997. Molecular characterization of the *bet* genes encoding glycine betaine synthesis in *Sinorhizobium meliloti* 102F34. Microbiology **143:**1369–1379.
306. **Poelarends, G. J., L. A. Kulakov, M. J. Larkin, J.E. van Hylckama Vlieg, and D. B. Janssen.** 2000. Roles of horizontal gene transfer and gene integration in evolution of 1,3-dichloropropene- and 1,2-dibromoethane-degradative pathways. J. Bacteriol. **182:**2191–2199.
307. **Poole, K., D. E. Heinrichs, and S. Neshat.** 1993. Cloning and sequence analysis of an EnvCD homologue in *Pseudomonas aeruginosa*: regulation by iron and possible involvement in the secretion of the siderophore pyoverdine. Mol. Microbiol. **10:**529–544.
308. **Poole, K., K. Tetro, Q. Zhao, S. Neshat, D. E. Heinrichs, and N. Bianco.** 1996. Expression of the multidrug resistance operon *mexA-mexB-oprM* in

Pseudomonas aeruginosa: *mexR* encodes a regulator of operon expression. Antimicrob. Agents Chemother. **40:**2021–2028.
309. **Pragman, A. A., J. M. Yarwood, T. J. Tripp, and P. M. Schlievert.** 2004. Characterization of virulence factor regulation by SrrAB, a two-component system in *Staphylococcus aureus*. J. Bacteriol. **186:**2430–2438.
310. **Puskas, L. G., M. Inui, Z. Kele, and H. Yukawa.** 2000. Cloning of genes participating in aerobic biodegradation of *p*-cumate from *Rhodopseudomonas palustris*. DNA Seq. **11:**9–20.
311. **Ramos, J. L., E. Duque, M. T. Gallegos, P. Godoy, M. I. Ramos-González, A. Rojas, W. Terán, and A. Segura.** 2002. Mechanisms of solvent tolerance in gram-negative bacteria. Annu. Rev. Microbiol. **56:**743–768.
312. **Ramos, J. L., M. T. Gallegos, S. Marqués, M. I. Ramos-González, M. Espinosa-Urgel, and A. Segura.** 2001. Response of gram-negative bacteria to certain environmental stresses. Curr. Opin. Microbiol. **4:**166–171.
313. **Ramos, J. L., F. Rojo, L. Zhou, and K. N. Timmis.** 1990. A family of positive regulators related to the *Pseudomonas putida* TOL plasmid XylS and the *Escherichia coli* AraC activators. Nucleic Acids Res. **18:**2149–2152.
314. **Rand, J. D., S. G. Danby, D. L. Greenway, and R. R. England.** 2002. Increased expression of the multidrug efflux genes *acrAB* occurs during slow growth of *Escherichia coli*. FEMS Microbiol. Lett. **207:**91–95.
315. **Rebets, Y., B. Ostash, A. Luzhetskyy, D. Hoffmeister, A. Brana, C. Mendez, J. A. Salas, A. Bechthold, and V. Fedorenko.** 2003. Production of landomycins in *Streptomyces globisporus* 1912 and *S. cyanogenus* S136 is regulated by genes encoding putative transcriptional activators. FEMS Microbiol. Lett. **222:**149–153.
316. **Redenbach, M., K. Ikeda, M. Yamasaki, and H. Kinashi.** 1998. Cloning and physical mapping of the EcoRI fragments of the giant linear plasmid SCP1. J. Bacteriol. **180:**2796–2799.
317. **Redfield, R. J.** 2002. Is quorum sensing a side effect of diffusion sensing? Trends Microbiol. **10:**365–370.
318. **Reizer, A., J. Deutscher, M. H. Saier, and J. Reizer.** 1991. Analysis of the gluconate (*gnt*) operon of *Bacillus subtilis*. Mol. Microbiol. **5:**1081–1089.
319. **Reizer, J., I. T. Paulsen, A. Reizer, F. Titgemeyer, and M. H. Saier, Jr.** 1996. Novel phosphotransferase system genes revealed by bacterial genome analysis: the complete complement of *pts* genes in *Mycoplasma genitalium*. Microb. Comp. Genomics **1:**151–164.
320. **Retzlaff, L., and J. Distler.** 1995. The regulator of streptomycin gene expression, StrR, of *Streptomyces griseus* is a DNA binding activator protein with multiple recognition sites. Mol. Microbiol. **18:**151–162.
321. **Reverchon, S., W. Nasser, and J. Robert-Baudouy.** 1991. Characterization of *kdgR*, a gene of *Erwinia chrysanthemi* that regulates pectin degradation. Mol. Microbiol. **5:**2203–2216.
322. **Rey, D. A., A. Puhler, and J. Kalinowski.** 2003. The putative transcriptional repressor McbR, member of the TetR-family, is involved in the regulation of the metabolic network directing the synthesis of sulfur containing amino acids in *Corynebacterium glutamicum*. J. Biotechnol. **103:**51–65.
323. **Rice, P., I. Longden, and A. Bleasby.** 2000. EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet. **16:**276–277.
324. **Rigali, S., A. Derouaux, F. Giannotta, and J. Dusart.** 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.
325. **Rodionov, D. A., M. S. Gelfand, A. A. Mironov, and A. B. Rakhmaninova.** 2001. Comparative approach to analysis of regulation in complete genomes: multidrug resistance systems in gamma-proteobacteria. J. Mol. Microbiol. Biotechnol. **3:**319–324.
326. **Roessner, C. A., K. X. Huang, M. J. Warren, E. Raux, and A. I. Scott.** 2002. Isolation and characterization of 14 additional genes specifying the anaerobic biosynthesis of cobalamin (vitamin B12) in. Propionibacterium freudenreichii (*P. shermanii*). Microbiology **148:**1845–1853.
327. **Rojas, A., E. Duque, G. Mosqueda, G. Golden, A. Hurtado, J. L. Ramos, and A. Segura.** 2001. Three efflux pumps are required to provide efficient tolerance to toluene in *Pseudomonas putida* DOT-T1E. J. Bacteriol. **183:** 3967–3973.
328. **Rojas, A., Duque, E., Schmid, A., A. Hurtado, and J. L. Ramos.** 2004. Biotransformation in double phase systems: physiological responses of *Pseudomonas putida* DOT T1E to a double phase made of aliphatic alcohols and biosynthesis of substituted catechols. Appl. Environ. Microbiol. **70:**3633–3643.
329. **Rojas, A., A. Segura, M. E. Guazzaroni, W. Terán, A. Hurtado, M. T. Gallegos, and J. L. Ramos.** 2003. In vivo and in vitro evidence that TtgV is the specific regulator of the TtgGHI multidrug and solvent efflux pump of *Pseudomonas putida*. J. Bacteriol. **185:**4755–4763.
330. **Rojo, F.** 1999. Repression of transcription initiation in bacteria. J. Bacteriol. **181:**2987–2991.
331. **Røkenes, T. P., T. Lamark, and A. R. Strøm.** 1996. DNA-binding properties of the BetI repressor protein of *Escherichia coli*: the inducer choline stimulates BetI-DNA complex formation. J. Bacteriol. **178:**1663–1670.
332. **Rouch, D. A., D. S. Cram, D. DiBerardino, T. G. Littlejohn, and R. A. Skurray.** 1990. Efflux-mediated antiseptic resistance gene *qacA* from *Staphylococcus aureus*: common ancestry with tetracycline- and sugar-transport proteins. Mol. Microbiol. **4:**2051–2062.
333. **Rouquette, C., J. B. Harmon, and W. M. Shafer.** 1999. Induction of the *mtrCDE*-encoded efflux pump system of *Neisseria gonorrhoeae* requires MtrA, an AraC-like protein. Mol. Microbiol. **33:**651–658.
334. **Rouquette-Loughlin, C., I. Stojiljkovic, T. Hrobowski, J. T. Balthazar, and W. M. Shafer.** 2002. Inducible, but not constitutive, resistance of gonococci to hydrophobic agents due to the MtrC-MtrD-MtrE efflux pump requires TonB-ExbB-ExbD proteins. Antimicrob. Agents Chemother. **46:**561–565.
335. **Rousseau, P., E. Guegen, G. Duval-Valentin, and M. Chandler.** 2004. The helix-turn-helix motif of bacterial insertion sequence IS*911* transposase is required for DNA binding. Nucleic Acids Res. **32:**1335–1344.
336. **Ryu, J. R., L. K. Olson, and D. N. Arnosti.** 2001. Cell-type specificity of short-range transcriptional repressors. Proc. Natl. Acad. Sci. USA **98:** 12960–12965.
337. **Saenger, W., P. Orth, C. Kisker, W. Hillen, and W. Hinrichs.** 2000. The tetracycline repressor-A paradigm for a biological switch. Angew. Chem. Int. ed. Engl. **39:**2042–2052.
338. **Sala, C., F. Forti, E. di Florio, F. Canneva, A. Milano, G. Riccardi, and D. Ghisotti.** 2003. *Mycobacterium tuberculosis* FurA autoregulates its own expression. J. Bacteriol. **185:**5357–5362.
339. **Salah-Bey, K., and C. J. Thompson.** 1995. Unusual regulatory mechanism for a *Streptomyces* multidrug resistance gene, *ptr*, involving three homologous protein-binding sites overlapping the promoter region. Mol. Microbiol. **17:**1109–1119.
340. **Sánchez, P., A. Alonso, and J. L. Martínez.** 2002. Cloning and characterization of SmeT, a repressor of the *Stenotrophomonas maltophilia* multidrug efflux pump SmeDEF. Antimicrob. Agents Chemother. **46:**3386–3393.
341. **Schauder, S., and B. L. Bassler.** 2001. The languages of bacteria. Genes Dev. **15:**1468–1480.
342. **Schell, M. A.** 1993. Molecular biology of the LysR family of transcriptional regulators. Annu. Rev. Microbiol. **47:**597–626.
343. **Schevitz, R. W., Z. Otwinowski, A. Joachimiak, C. L Lawson, and P. B. Singler.** 1985. The three-dimensional structure of *trp* repressor. Nature **317:** 782–786.
344. **Schindelin, H., M. A. Marahiel, and U. Heinemann.** 1993. Universal nucleic acid-binding domain revealed by crystal structure of the *B. subtilis* major cold-shock protein. Nature **364:**164–168.
345. **Schleif, R.** 1987. The L-arabinose operon, p. 1473–1481. *In* F. C. Neidhardt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed.), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology. American Society for Microbiology, Washington, D.C.
346. **Schnider-Keel, U., A. Seematter, M. Maurhofer, C. Blumer, B. Duffy, C. Gigot-Bonnefoy, C. Reimmann, R. Notz, G. Defago, D. Haas, and C. Keel.** 2000. Autoinduction of 2,4-diacetylphloroglucinol biosynthesis in the biocontrol agent *Pseudomonas fluorescens* CHA0 and repression by the bacterial metabolites salicylate and pyoluteorin. J. Bacteriol. **182:**1215–1225.
347. **Schouls, L. M., C. S. Schot, and J. A. Jacobs.** 2003. Horizontal transfer of segments of the 16S rRNA genes between species of the *Streptococcus anginosus* group. J. Bacteriol. **185:**7241–7246.
348. **Schultz, J., R. R. Copley, T. Doerks, C. P. Ponting, and P. Bork.** 2000. Smart: A Web-based tool for the study of genetically mobile domains. Nucleic Acids Res. **28:**231–234.
349. **Schumacher, M. A., and R. G. Brennan.** 2002. Structural mechanisms of multidrug recognition and regulation by bacterial multidrug transcription factors. Mol. Microbiol. **45:**885–893.
350. **Schumacher, M. A., M. C. Miller, S. Grkovic, M. H. Brown, R. A. Skurray, and R. G. Brennan.** 2001. Structural mechanisms of QacR induction and multidrug recognition. Science **294:**2158–2163.
351. **Schumacher, M. A., M. C. Miller, S. Grkovic, M. H. Brown, R. A. Skurray, and R. G. Brennan.** 2002. Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. EMBO J. **21:**1210–1218.
352. **Seoane, A. S., and S. B. Levy.** 1995. Characterization of MarR, the repressor of the multiple antibiotic resistance (*mar*) operon in *Escherichia coli*. J. Bacteriol. **177:**3414–3419.
353. **Shafer, W. M., J. T. Balthazar, K. E. Hagman, and S. A. Morse.** 1995. Missense mutations that alter the DNA-binding domain of the MtrR protein occur frequently in rectal isolates of *Neisseria gonorrhoeae* that are resistant to faecal lipids. Microbiology **141:**907–911.
354. **Shankar, N., A. S. Baghdayan, and M. S. Gilmore.** 2002. Modulation of virulence within a pathogenicity island in vancomycin-resistant *Enterococcus faecalis*. Nature **417:**746–750.
355. **Shao, C. P., and L. I. Hor.** 2001. Regulation of metalloprotease gene expression in *Vibrio vulnificus* by a *Vibrio harveyi* LuxR homologue. J. Bacteriol. **183:**1369–1375.
356. **Shaw, G. C., and A. J. Fulco.** 1992. Barbiturate-mediated regulation of expression of the cytochrome P450BM-3 gene of *Bacillus megaterium* by Bm3R1 protein. J. Biol. Chem. **267:**5515–5526.
357. **Shaw, G. C., and A. J. Fulco.** 1993. Inhibition by barbiturates of the binding of Bm3R1 repressor to its operator site on the barbiturate-inducible cytochrome P450BM-3 gene of *Bacillus megaterium*. J. Biol. Chem. **268:**2997–3004.
358. **Shaw, G. C., Y. H. Hsueh, and H. S. Kao.** 2000. The basal-level expression of the cytochrome P450(BM-1) gene is negatively affected by the bm1P1 gene of *Bacillus megaterium*. Curr. Microbiol. **40:**47–50.
359. **Shaw, G. C., Y. H. Hsueh, C. C. Sung, Y. S. Chen, and C. H. Liu.** 2000.

Negative regulation of expression of the *Bacillus megaterium bmlP1* gene by the *bmlP1* 3′ flanking region. J. Bacteriol. **182:**2654–2659.
360. **Shaw, G. C., H. S. Kao, C. C. Sung, and A. N. Chiang.** 1997. Iron and salicylate induction of cytochrome P450BM-1 in *Bacillus megaterium*. Curr. Microbiol. **35:**28–31.
361. **Shaw, G. C., C. C. Sung, C. H. Liu, and H. S. Kao.** 1997. A 53-base-pair inverted repeat negatively regulates expression of the adjacent and divergently oriented cytochrome P450(BM-1) gene and its regulatory gene, *bm1P1*, in *Bacillus megaterium*. J. Bacteriol. **179:**280–283.
362. **Shaw, G. C., C. C. Sung, C. H. Liu, and C. H. Lin.** 1998. Evidence against the Bm1P1 protein as a positive transcription factor for barbiturate-mediated induction of cytochrome P450BM-1 in *bacillus megaterium*. J. Biol. Chem. **273:**7996–8002.
363. **Showalter, R. E., M. O. Martin, and M. R. Silverman.** 1990. Cloning and nucleotide sequence of *luxR*, a regulatory gene controlling bioluminescence in *Vibrio harveyi*. J. Bacteriol. **172:**2946–2954.
364. **Sigrist, C. J., L. Cerutti, N. Hulo, A. Gattiker, L. Falquet, M. Pagni, A. Bairoch, and P. Bucher.** 2002. PROSITE: a documented database using patterns and profiles as motif descriptors. Brief Bioinform. **3:**265–274.
365. **Silva, A. J., K. Pham, and J. A. Benitez.** 2003. Haemagglutinin/protease expression and mucin gel penetration in El Tor biotype *Vibrio cholerae*. Microbiology **149:**1883–1891.
366. **Sizemore, C., A. Wissmann, U. Gülland, and W. Hillen.** 1990. Quantitative analysis of Tn*10* Tet repressor binding to a complete set of *tet* operator mutants. Nucleic Acids Res. **18:**2875–2880.
367. **Skarstad, K., B. Thony, D. S. Hwang, and A. Kornberg.** 1993. A novel binding protein of the origin of the *Escherichia coli* chromosome. J. Biol. Chem. **268:**5365–5370.
368. **Smith, L. T., J. A. Pocard, T. Bernard, and D. Le Rudulier.** 1988. Osmotic control of glycine betaine biosynthesis and degradation in *Rhizobium meliloti*. J. Bacteriol. **170:**3142–3149.
369. **Soisson, S. M., B. MacDougall-Shackleton, R. Schleif, and C. Wolberger.** 1997. Structural basis for ligand-regulated oligomerization of AraC. Science **276:**421–425.
370. **Spiro, S.** 1994. The FNR family of transcriptional regulators. Antonie van Leeuwenhoek **66:**23–36.
371. **Stratigopoulos, G., and E. Cundliffe.** 2002. Expression analysis of the tylosin-biosynthetic gene cluster: pivotal regulatory role of the *tylQ* gene product. Chem. Biol. **9:**71–78.
372. **Stratigopoulos, G., A. R. Gandecha, and E. Cundliffe.** 2002. Regulation of tylosin production and morphological differentiation in *Streptomyces fradiae* by TylP, a deduced gamma-butyrolactone receptor. Mol. Microbiol. **45:** 735–744.
373. **Strom, A. R., P. Falkenberg, and B. Landfald.** 1986. Genetics of osmoregulation in *Escherichia coli*: uptake and biosynthesis of organic osmolytes. FEMS Microbiol. Rev. **39:**79–86.
374. **Styrvold, O. B., P. Falkenberg, B. Landfald, M. W. Eshoo, T. Bjørnsen, and A. R. Strøm.** 1986. Selection, mapping, and characterization of osmoregulatory mutants of *Escherichia coli* blocked in the choline-glycine betaine pathway. J. Bacteriol. **165:**856–863.
375. **Sugiyama, M., H. Onaka, T. Nakagawa, and S. Horinouchi.** 1998. Site-directed mutagenesis of the A-factor receptor protein: Val-41 important for DNA-binding and Trp-119 important for ligand-binding. Gene **222:**133–144.
376. **Sulavik, M. C., L. F. Gambino, and P. F. Miller.** 1995. The MarR repressor of the multiple antibiotic resistance (*mar*) operon in *Escherichia coli*: prototypic member of a family of bacterial regulatory proteins involved in sensing phenolic compounds. Mol. Med. **1:**436–446.
377. **Summers, A. O.** 1992. Untwist and shout: a heavy metal-responsive transcriptional regulator. J. Bacteriol. **174:**3097–3101.
378. **Sunnarborg, A., D. Klumpp, T. Chung, and D. C. LaPorte.** 1990. Regulation of the glyoxylate bypass operon: cloning and characterization of *iclR*. J. Bacteriol. **172:**2642–1649.
379. **Surette, M. G., and B. L. Bassler.** 1998. Quorum sensing in *Escherichia coli* and *Salmonella typhimurium*. Proc. Natl. Acad. Sci. USA **95:**7046–7050.
380. **Swartzman, E., C. Miyamoto, A. Graham, and E. Meighen.** 1990. Delineation of the transcriptional boundaries of the *lux* operon of *Vibrio harveyi* demonstrates the presence of two new *lux* genes. J. Biol. Chem. **265:** 3513–3517.
381. **Swartzman, E., and E. A. Meighen, E. A.** 1993. Purification and characterization of a poly(dA-dT) *lux*-specific DNA-binding protein from *Vibrio harveyi* and identification as LuxR. J. Biol. Chem. **268:**16706–16716.
382. **Swartzman, E., M. Silverman, and E. A. Meighen.** 1992. The *luxR* gene product of *Vibrio harveyi* is a transcriptional activator of the *lux* promoter. J. Bacteriol. **174:**7490–7493.
383. **Szczepanowski, R., I. Krahn, B. Linke, A. Goesman, A. Pühler, and A. Schlüter.** 2004. Antibiotic multiresistance plasmid pRSB101 isolated from a wastewater treatment plant is related to plasmids residing in phytopathogenic bacteria and carries eight different resistance determinants including a multidrug transport system. Microbiology **150:**3613–3630.
384. **Takahashi, M., L. Altschmied, and W. Hillen.** 1986. Kinetic and equilibrium characterization of the Tet-repressor-tetracycline complex by fluorescence measurements; evidence for divalent ion requirements and energy transfer. J. Mol. Biol. **187:**341–348.
385. **Takano, E., R. Chakraburtty, T. Nihira, Y. Yamad, and M. J. Bibb.** 2001. A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in *Streptomyces coelicolor* A3(2). Mol. Microbiol. **41:**1015–1028.
386. **Takano, E., T. Nihira, Y. Hara, J. J. Jones, C. J. Gershater, Y. Yamada, and M. Bibb.** 2000. Purification and structural determination of SCB1, a gamma-butyrolactone that elicits antibiotic production in *Streptomyces coelicolor* A3(2). J. Biol. Chem. **275:**11010–11016.
387. **Tatusov, R. L., D. A. Natale, I. V. Garkavtsev, T. A. Tatusova, U. T. Shankavaram, B. S. Rao, B. Kiryutin, M. Y. Galperin, N. D. Fedorova, and E. V. Koonin.** 2001. The COG database: new developments in phylogenetic classification of proteins from complete genomes. Nucleic Acids Res. **29:** 22–28.
388. **Tauch, A., S. Gotker, A. Puhler, J. Kalinowski, and G. Thierbach.** 2002. The 27.8-kb R-plasmid pTET3 from. *Corynebacterium glutamicum* encodes the aminoglycoside adenyltransferase gene cassette aadA9 and the regulated tetracycline efflux system Tet 33 flanked by active copies of the widespread insertion sequence IS*6100*. Plasmid **48:**117–129.
389. **Tauch, A., A. Pühler, J. Kalinowski, and G. Thierbach.** 2000. TetZ, a new tetracycline resistance determinant discovered in gram-positive bacteria, shows high homology to gram-negative regulated efflux systems. Plasmid **44:** 285–291.
390. **Tennent, J. M., B. R. Lyon, M. T. Gillespie, J. W. May, and R. A. Skurray.** 1985. Cloning and expression of *Staphylococcus aureus* plasmid-mediated quaternary ammonium resistance in *Escherichia coli*. Antimicrob. Agents Chemother. **27:**79–83.
391. **Terán, W., A. Felipe, A. Segura, A. Rojas, J. L. Ramos, and M. T. Gallegos.** 2003. Antibiotic-dependent induction of *Pseudomonas putida* DOT-T1E TtgABC efflux pump is mediated by the drug binding repressor TtgR. Antimicrob. Agents Chemother. **47:**3067–3072.
392. **Thieffry, D., A. M. Huerta, E. Perez-Rueda, and J. Collado-Vides.** 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. Bioessays **20:**433–440.
393. **Tiebel, B., K. Garke, and W. Hillen, W.** 2000. Observing conformational and activity changes of *tet* repressor *in vivo*. Nat. Struct. Biol. **7:**479–481.
394. **Tobes, R., and J. L. Ramos.** 2002. AraC-XylS database: a family of positive transcriptional regulators in bacteria. Nucleic Acids Res. **30:**318–321.
395. **Tovar, K., A. Ernst, and W. Hillen.** 1988. Identification and nucleotide sequence of the class E *tet* regulatory elements and operator and inducer binding of the encoded purified Tet repressor. Mol. Gen. Genet. **215:**76–80.
396. **Trefzer, A., S. Pelzer, J. Schimana, S. Stockert, C. Bihlmaier, H. P. Fiedler, K. Welzel, A. Vente, and A. Bechthold.** 2002. Biosynthetic gene cluster of simocyclinone, a natural multihybrid antibiotic. Antimicrob. Agents Chemother. **46:**1174–1182.
397. **Tropel, D., and J. R. van der Meer.** 2004. Bacterial transcriptional regulators for degradation pathways of aromatic compounds. Microbiol. Mol. Biol. Rev. **68:**474–500.
398. **Ueda, K., K. Matsuda, H. Takano, and T. Beppu.** 1999. A putative regulatory element for carbon-source-dependent differentiation in *Streptomyces griseus*. Microbiology **145:**2265–2271.
399. **Ueda, K., H. Takano, M. Nishimoto, H. Inaba, and T. Beppu.** 2005. Dual transcriptional control of *amfTSBA*, which regulates the onset of cellular differentiation in *Streptomyces griseus*. J. Bacteriol. **187:**135–142.
400. **Ulmasou, B., J. Capone, and W. Folk.** 1997. Regulated expression of plant RNA genes by the prokaryotic *tet* and *lac* repressors. Plant Mol. Biol. **25:** 417–424.
401. **Umeyama, T., P. C. Lee, K. Ueda, and S. Horinouchi.** 1999. An AfsK/AfsR system involved in the response of aerial mycelium formation to glucose in *Streptomyces griseus*. Microbiology **145:**2281–2292.
402. **Unger, B., G. Klock, and W. Hillen.** 1984. Nucleotide sequence of the repressor gene of the RA1 tetracycline resistance determinant: structural and functional comparison with three related Tet repressor genes. Nucleic Acids Res. **12:**7693–7703.
403. **Urlinger, S., V. Helbl, J. Guthman, E. Pook, S. Grimm, and W. Hillen.** 2000. The P65 domain from NF-KB is an efficient human activator in the tetracycline-regulatable gene expression system. Gene **247:**103–110.
404. **van Der Geize, R., G. I. Hessels, R. van Gerwen, J. W. Vrijbloed, P. van Der Meijden, and L. Dijkhuizen.** 2000. Targeted disruption of the *kstD* gene encoding a 3-ketosteroid delta(1)-dehydrogenase isoenzyme of *Rhodococcus erythropolis* strain SQ1. Appl. Environ. Microbiol. **66:**2029–2036.
405. **van Rooijen, R. J., and W. M. de Vos.** 1990. Molecular cloning, transcriptional analysis and nucleotide sequence of *lacR*, a gene encoding the repressor of the lactose phosphotransferase system of *Lactococcus lactis*. J. Biol. Chem. **265:**18499–18503.
406. **Veal, W. L., R. A. Nicholas, and W. M. Shafer.** 2002. Overexpression of the MtrC-MtrD-MtrE efflux pump due to an *mtrR* mutation is required for chromosomically mediated penicillin resistance in *Neisseria gonorrhoeae*. J. Bacteriol. **184:**5619–5621.
407. **Veal, W. L., and W. M. Shafer.** 2003. Identification of a cell envelope protein (MtrF) involved in hydrophobic antimicrobial resistance in *Neisseria gonorrhoeae*. J. Antimicrob. Chemother. **51:**27–32.

408. **Veal, W. L., A. Yellen, J. T. Balthazar, W. Pan, B. G. Spratt, and W. M. Shafer.** 1998. Loss-of-function mutations in the *mtr* efflux system of *Neisseria gonorrhoeae*. Microbiology **144:**621–627.
409. **Ventre, I., A. Filloux, and A. Lazdunski.** 2004. Two-component signal transduction systems: a key to the adaptative potential of *Pseudomonas aeruginosa* p. 257–288. In J. L. Ramos (ed.), *Pseudomonas*, vol. II. Kluwer, London, England.
410. **Vimr, E. R., K. A. Kalivoda, E. L. Deszo, and S. M. Steenbergen.** 2004. Diversity of microbial sialic acid metabolism. Microbiol. Mol. Biol. Rev. **68:** 132–153.
411. **Vujaklija, D., K. Ueda, S. K. Hong, T. Beppu, and S. Horinouchi.** 1991. Identification of an A-factor-dependent promoter in the streptomycin biosynthetic gene cluster of *Streptomyces griseus*. Mol. Gen. Genet. **229:**119–128.
412. **Vujaklija, D., S. Horinouchi, and T. Beppu.** 1993. Detection of an A-factor-responsive protein that binds to the upstream activation sequence of *strR*, a regulatory gene for streptomycin biosynthesis in *Streptomyces griseus*. J. Bacteriol. **175:**2652–2661.
413. **Waki, M., T. Nihira, and Y. Yamada.** 1997. Cloning and characterization of the gene (*farA*) encoding the receptor for an extracellular regulatory factor (IM-2) from *Streptomyces* sp. strain FRI-5. J. Bacteriol. **179:**5131–5137.
414. **Walczak, R. J., A. J. Woo, W. R. Strohl, and N. D. Priestley.** 2000. Nonactin biosynthesis: the potential nonactin biosynthesis gene cluster contains type II polyketide synthase-like genes. FEMS Microbiol. Lett. **183:**171–175.
415. **Wang, L., and B. Weiss.** 1992. *dcd* (dCTP deaminase) gene of *Escherichia coli*: mapping, cloning, sequencing, and identification as a locus of suppressors of lethal *dut* (dUTPase) mutations. J. Bacteriol. **174:**5647–5653.
416. **Wang, L., R. L. White, and L. C. Vining.** 2002. Biosynthesis of the dideoxysugar component of jadomycin B: genes in the *jad* cluster of *Streptomyces venezuelae* ISP*5230* for L-digitoxose assembly and transfer to the angucycline aglycone. Microbiology **148:**1091–1103.
417. **Wasylewski, Z., P. Kaszycki, and M. Drwiega.** 1996. A fluorescence study of Tn*10*-encoded *tet* repressor. J. Protein Chem. **15:**45–58.
418. **Waters, S. H., P. Rogowsky, J. Grinsted, J. Altenbuchner, and R. Schmitt.** 1983. The tetracycline resistance determinants of RP1 and Tn*1721*: nucleotide sequence analysis. Nucleic Acids Res. **11:**6089–6105.
419. **Weber, W., R. Schoenmakers, M. Spielmann, M. D. El-Baba, M. Folcher, B. Keller, C. C. Weber, N. Link, P. van de Wetering, C. Heinzen, B. Jolivet, U. Sequin, D. Aubel, C. J. Thompson, and M. Fussenegger.** 2003. *Streptomyces*-derived quorum-sensing systems engineered for adjustable transgene expression in mammalian cells and mice. Nucleic Acids Res. **31:**71.
420. **Weickert, M. J., and S. Adhya.** 1992. A family of bacterial regulators homologous to Gal and Lac repressors. J. Biol. Chem. **267:**15869–15874.
421. **Wen, L. P., and A. J. Fulco.** 1987. Cloning of the gene encoding a catalytically self-sufficient cytochrome P-450 fatty acid monooxygenase induced by barbiturates in *Bacillus megaterium* and its functional expression and regulation in heterologous (*Escherichia coli*) and homologous (*Bacillus megaterium*) hosts. J. Biol. Chem. **262:**6676–6682.
422. **Wery, J., B. Hidayat, J. Kieboom, and J. A. de Bont.** 2001. An insertion sequence prepares *Pseudomonas putida* S12 for severe solvent stress. J. Biol. Chem. **276:**5700–5706.
423. **Westbrock-Wadman, S., D. R. Sherman, M. J. Hickey, S. N. Coulter, Y. Q. Zhu, P. Warrener, L. Y. Nguyen, R. M. Shawar, K. R. Folger, and C. K. Stover.** 1999. Characterizarion of a *Pseudomonas aeruginosa* efflux pump contributing to aminoglycoside impermeability. Antimicrob. Agents Chemother. **43:**2975–2983.
424. **Westrich, L., S. Domann, B. Faust, D. Bedford, D. A. Hopwood, and A. Bechthold.** 1999. Cloning and characterization of a gene cluster from *Streptomyces cyanogenus* S136 probably involved in landomycin biosynthesis. FEMS Microbiol. Lett. **170:**381–387.
425. **White, D. G., J. D. Goldman, B. Demple, and S. B. Levy.** 1997. Role of the *acrAB* locus in organic solvent tolerance mediated by expression of *marA*, soxS, or *rob* in *Escherichia coli*. J. Bacteriol. **179:**6122–6126.
426. **Wilson, M. S., J. B. Herrick, C. O. Jeon, D. E. Hinman, and E. L. Madsen.** 2003. Horizontal transfer of *phnAc* dioxygenase genes within one of two phenotypically and genotypically distinctive naphthalene-degrading guilds from adjacent soil environments. Appl. Environ. Microbiol. **69:**2172–2181.
427. **Winans, S. C., and B. L. Bassler.** 2002. Mob psychology. J. Bacteriol. **184:** 873–883.
428. **Wintjens, R., and M. Rooman.** 1996. Structural classification of HTH DNA-binding domains and protein-DNA interaction modes. J. Mol. Biol. **262:** 294–313.
429. **Winzer, K., K. Lorenz, B. Zickner, and P. Durre.** 2000. Differential regulation of two thiolase genes from *Clostridium acetobutylicum* DSM 792. J. Mol. Microbiol. Biotechnol. **2:**531–541.
430. **Wissmann, A., R. Baumeister, G. Muller, B. Hecht, V. Helbl, K. Pfleiderer, and W. Hillen.** 1991. Amino acids determining operator binding specificity in the helix-turn-helix motif of Tn*10* Tet repressor. EMBO J. **10:**4145–4152.
431. **Wösten, M. M. S. M., J. A. Wagenaar, and J. P. M. van Putten.** 2004. The FlgS/FlgR two-component signal transduction system regulates the *fla* regulon in *Campylobacter jejuni*. J. Biol. Chem. **279:**16214–16222.
432. **Wu, J., and B. P. Rosen.** 1991. The ArsR protein is a *trans*-acting regulatory protein. Mol. Microbiol. **5:**1331–1336.
433. **Xi, C., E. Schoeters, J. Vanderleyden, and J. Michiels.** 2000. Symbiosis-specific expression of *Rhizobium etli casA* encoding a secreted calmodulin-related protein. Proc. Natl. Acad. Sci. USA **97:**11114–11119.
434. **Yamaguchi, A., N. Ono, T. Akasaka, T. Noumi, and T. Sawai.** 1999. Metal-tetracycline/H^+ antiporter of *Escherichia coli* encoded by a transposon, Tn*10*. J. Biol. Chem. **265:**15525–15530.
435. **Yamaguchi, A., T. Udagawa, and T. Sawai.** 1999. Transport of divalent cations with tetracycline as mediated by the transposon Tn*10*-encoded tetracycline resistance protein. J. Biol. Chem. **9:**4809–4813.
436. **Yamamoto, K., K. Yata, N. Fujita, and A. Ishihama.** 2001. Novel mode of transcription regulation by SdiA, an *Escherichia coli* homologue of the quorum-sensing regulator. Mol. Microbiol. **41:**1187–1198.
437. **Yamazaki, M., Y. Ikuto, A. Ohira, K. Chater, and H. Kinashi.** 2003. Limited regions of homology between linear and circular plasmids encoding methylenomycin biosynthesis in two independently isolated streptomycetes. Microbiology **149:**1351–1356.
438. **Yamazaki, H., Y. Ohnishi, and S. Horinouchi.** 2000. An A-factor-dependent extracytoplasmic function sigma factor (σ^{AdsA}) that is essential for morphological development in *Streptomyces griseus*. J. Bacteriol. **182:**4596–4605.
439. **Yamazaki, H., Y. Ohnishi, and S. Horinouchi.** 2003. Transcriptional switch on *ssgA* by A-factor, which is essential for spore septum formation in *Streptomyces griseus*. J. Bacteriol. **185:**1273–1283.
440. **Yamazaki, H., Y. Takano, Y. Ohnishi, and S. Horinouchi.** 2003. *amfR*, an essential gene for aerial mycelium formation, is a member of the AdpA regulon in the A-factor regulatory cascade in *Streptomyces griseus*. Mol. Microbiol. **50:**1173–1187.
441. **Yamazaki, H., A. Tomono, Y. Ohnishi, and S. Horinouchi.** 2004. DNA-binding specificity of AdpA, a transcriptional activator in the A-factor regulatory cascade in *Streptomyces griseus*. Mol. Microbiol. **53:**555–572.
442. **Yang, K., L. Han, J. He, L. Wang, and L. C. Vining.** 2001. A repressor-response regulator gene pair controlling jadomycin B production in *Streptomyces venezuelae* ISP*5230*. Gene. **279:**165–173.
443. **Yang, K., L. Han, and L. C. Vining.** 1995. Regulation of jadomycin B production in *Streptomyces venezuelae* IS*P5230*: involvement of a repressor gene, jadR2. J. Bacteriol. **177:**6111–61117.
444. **Yoshimura, M., K. Asai, Y. Sadaie, and H. Yoshikawa.** 2004. Interaction of *Bacillus subtilis* extracytoplasmic function (ECF) sigma factors with the N-terminal regions of their potential anti-sigma factors. Microbiology **150:** 591–599.
445. **Yu, E. W., J. R. Aires, and H. Nikaido.** 2003. AcrB multidrug efflux pump of *Escherichia coli*: composite substrate-binding cavity of exceptional flexibility generates its extremely wide substrate specificity. J. Bacteriol. **185:** 5657–5664.
446. **Yu, E. W., G. McDermott, H. I. Zgurskaya, H. Nikaido, D. E. Koshland, Jr.** 2003. Structural basis of multiple drug-binding capacity of the AcrB multidrug efflux pump. Science **300:**976–980.
447. **Zarantonelli, L., G. Borthagaray, E. H. Lee, and W. M. Shafer.** 1999. Decreased azithromycin susceptibility of *Neisseria gonorrhoeae* due to *mtrR* mutations. Antimicrob. Agents Chemother. **43:**2468–2472.
448. **Zarantonelli, L., G. Borthagaray, E. H. Lee, E., W. Veal, and W. M. Shafer.** 2001. Decreased susceptibility to azithromycin and erythromycin mediated by a novel *mtr*(R) promoter mutation in *Neisseria gonorrhoeae*. J. Antimicrob. Chemother. **47:**651–654.
449. **Zdobnov, E. M., and R. Apweiler.** 2001. InterProScan-an integration platform for the signature-recognition methods in InterPro. Bioinformatics **17:** 847–848.
450. **Zeng, G., S. Ye, and T. J. Larson.** 1996. Repressor for the sn-glycerol 3-phosphate regulon of *Escherichia coli* K-12: primary structure and identification of the DNA-binding domain. J. Bacteriol. **178:**7080–7089.
451. **Zhang, G., X. Hang, P. Green, K. P. Ho, and G. Q. Chen.** 2001. PCR cloning of type II polyhydroxyalkanoate biosynthesis genes from two *Pseudomonas* strains. FEMS Microbiol. Lett. **198:**165–170.
452. **Zhang, L., X. Z. Li, and K. Poole.** 2001. SmeDEF multidrug efflux pump contributes to intrinsic multidrug resistance in *Stenotrophomonas maltophilia*. Antimicrob. Agents Chemother. **45:**3497–3503.
453. **Zhao, J., and T. Aoki.** 1992. Nucleotide sequence analysis of the class G tetracycline resistance determinant from *Vibrio anguillarum*. Microbiol. Immunol. **36:**1051–1060.
454. **Zhou, L., L. Xiang-He, B. R. Bochner, and B. L. Wahner.** 2003. Phenotype microarray analysis of *Escherichia coli* K-12 mutants with deletions of all two-component systems. J. Bacteriol. **185:**4956–4972.
455. **Zhu, J., M. B. Miller, R. E. Vance, M. Dziejman, B. L. Bassler, and J. J. Mekalanos.** 2002. Quorum-sensing regulators control virulence gene expression in *Vibrio cholerae*. Proc. Natl. Acad. Sci. USA **99:**3129–3134.
456. **Ziebuhr, W., V. Krimmer, S. Rachid, I. Lossner, F. Gotz, and J. Hacker.** 1999. A novel mechanism of phase variation of virulence in *Staphylococcus epidermidis*: evidence for control of the polysaccharide intercellular adhesin synthesis by alternating insertion and excision of the insertion sequence element IS*256*. Mol. Microbiol. **32:**345–356.

Capítulo 2.1

Prefacio: Definición de la familia de factores de transcripción IclR

Definición de la familia de factores de transcripción IclR

Para probar la eficacia de Provalidator, se diseñó el *profile* de la familia de factores de transcripción IclR. Se comprobó usando el predictor "PHD" de estructura secundaria (Rost *et al.*, 1994) y el predictor "HTH" de motivos HTH (Dodd and Egan, 1990) (disponibles en http://npsa-pbil.ibcp.fr) que los miembros de esta familia contienen un motivo HTH de unión a ADN. Una vez se seleccionó con la supervisión de Provalidator el conjunto de secuencias que constituyó la semilla del *profile* y se generó el alineamiento múltiple, se pasó a probar distintas regiones del alineamiento que presentaban una mayor conservación. A continuación se muestra la tabla de resultados de los intentos para definir el *profile* de esta familia:

Sesión	FN	FP	Región
1	30	1	191-274
2	34	966	HTH extend.
3	35	0	190-273
4	407	50	233-320
5	39	2	191-274
6	96	0	237-281
7	27	2	190-281
8	7	432	59-278
9	9	12	116-283 (Pfam)
10	24	3	146-281
11	28	1	168-281
12	29	3	190-288
13	27	5	190-272
14	43	0	189-280
15	39	0	189-273
16	18	0	191-274

Tabla 3: Resultados de distintos *profiles* de IclR.

Cuando se construyó el alineamiento de las secuencias elegidas por Provalidator, se pudo observar que había tres zonas de mayor conservación. Una, el HTH; otra, desde el extremo izquierdo del dominio C-terminal hasta el centro del alineamiento; y la última, desde el centro hasta el otro extremo del dominio C-terminal (en la bibliografía esta región se había definido como asociada a un dominio de unión a efector, Krell *et al.*, 2006). Las regiones elegidas como candidatas están alrededor de esas tres zonas más conservadas o bien son combinación de ellas. Como se ve, la región de HTH extendido (esto es, el HTH y aminoácidos contiguos por ambos lados) definió un *profile* de muy mala calidad. Por esto, pudimos concluir que en la familia IclR la región de HTH no es lo suficientemente discriminante como para incluirla en su *profile*. Este resultado concuerda con lo obtenido en la validación del *profile*. Los métodos PS51077 y SM00346 incluyen la región de HTH en su definición. Por tanto, se podría pensar que las proteínas estudiadas como posibles falsos negativos que InterPro considera miembros de la familia apoyándose en alguno de estos dos métodos tienen una alta homología en la región de HTH, pero no así en el dominio de unión a efector, ya que obtienen muy bajo *score* con el *profile* objeto de estudio. Toda vez que ha queda-

do demostrada la incapacidad discriminante de la región de HTH, se puede concluir que estas proteínas están erróneamente asignadas a la familia IclR.

Por fin, se decidió parar en el último intento porque parecía que ya se había alcanzado la suficiente sensibilidad en el *profile*. Existen dos características, o figuras de mérito, que evalúan la bondad de un *profile* que son la especificidad y la sensibilidad (*precision* y *recall*, respectivamente, en terminología de PROSITE). Se definen así:

$$Especificidad = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ positivos}$$

$$Sensibilidad = \frac{Verdaderos\ positivos}{Verdaderos\ positivos + Falsos\ negativos}$$

Por supuesto, un requisito indispensable para admitir un *profile* como bueno es que su especificidad sea lo más cercana al 100%, mientras que en la sensibilidad no hace falta ser tan estricto (como sabemos, en ciencia no es lo mismo afirmar algo que no negarlo), ya que decir que un determinado regulador pertenece a una familia sin serlo induce más a errores experimentales que no decir que pertenece siéndolo.

Como figura de mérito para representar las conclusiones de la validación, se generó una gráfica que muestra las cuatro categorías en que se encuadran los datos analizados:

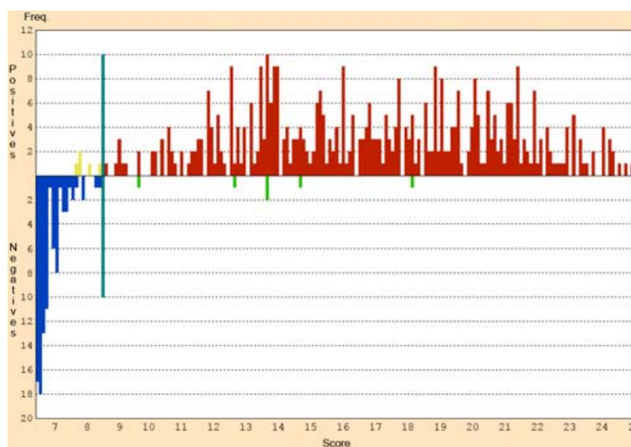


Figura 12: Figura de mérito para validar un *profile*.

En la gráfica, la barra verde marca la separación de 8,5, el umbral de *scores* que propone PFSCALE durante la calibración para pertenecer a la familia. Las barras representan la frecuencia de proteínas en ese intervalo de *scores*. Por tanto, es un histograma de frecuencias cuyo eje X está discretizado. Si la barra está hacia arriba, indica que InterPro considera esa proteína dentro de la familia, luego si esa barra está a la derecha de 8,5 será un verdadero positivo (rojas), mientras que si está a la izquierda será un falso negativo (amarillas). Por otra parte, si la barra esta hacia abajo, indica que no tiene ninguna anotación en InterPro que la vincule a la familia. Luego si esa barra está a la izquierda de 8,5 será un verdadero negativo

(azules), mientras que si está a la derecha será un falso positivo (verdes). Toda esta clasificación de verdaderos y falsos es respecto de la base de datos InterPro. La gráfica empieza en *score*=6,5, que es un segundo umbral que propone PFSCALE además del 8,5, con el que se crea una zona de incertidumbre para proteínas cuyo *score* está entre 6,5 y 8,5. De estas proteínas es difícil decir si pertenecen o no a la familia. Habría que consultar la bibliografía al respecto. Sin embargo, para proteínas bastante alejadas de este umbral (<<6,5) aumenta la probabilidad de que sean fruto de una mala anotación por parte de otros métodos más laxos o cuya construcción se ha basado en una región distinta.

A la hora de estudiar los falsos positivos representados en la gráfica, primero tenemos que descartar la posibilidad de que se produzcan como consecuencia de problemas de sincronización entre versiones de las bases de datos publicadas. Esto es, UniProt puede publicar proteínas que todavía no han sido analizadas por InterPro y que lo serán en la próxima versión (o *release*), pero hasta ese momento se están considerando potenciales falsos positivos. Esto se evita analizando a mano estas proteínas con la herramienta InterProScan (Quevillon *et al.*, 2005), que enfrenta una secuencia contra todos los métodos de anotación de que dispone InterPro y presenta un informe de resultados en el que se indica para qué métodos esa secuencia superó el umbral de pertenencia. Una vez hecho eso, se observa que todas las proteínas que eran potenciales falsos positivos pertenecían a la familia IclR por alguno de los métodos usados en InterPro. Además, para *scores* tan altos como los que se muestran (de 12 a 18), no había duda de que iban a pertenecer a la familia. Por tanto, la especificidad del *profile* es del 100%.

Para asegurar que la asignación de potenciales falsos positivos a la familia es correcta, se usó un segundo método de validación que consiste en comparar la predicción de estructuras secundarias (obtenida con el predictor "PHD"). Así, se predijo la estructura secundaria tanto para las secuencias que formaban la semilla como las secuencias de potenciales falsos positivos. Se alinearon las secuencias y se superpuso la estructura en ese alineamiento. El resultado lo podemos ver en las siguientes figuras:

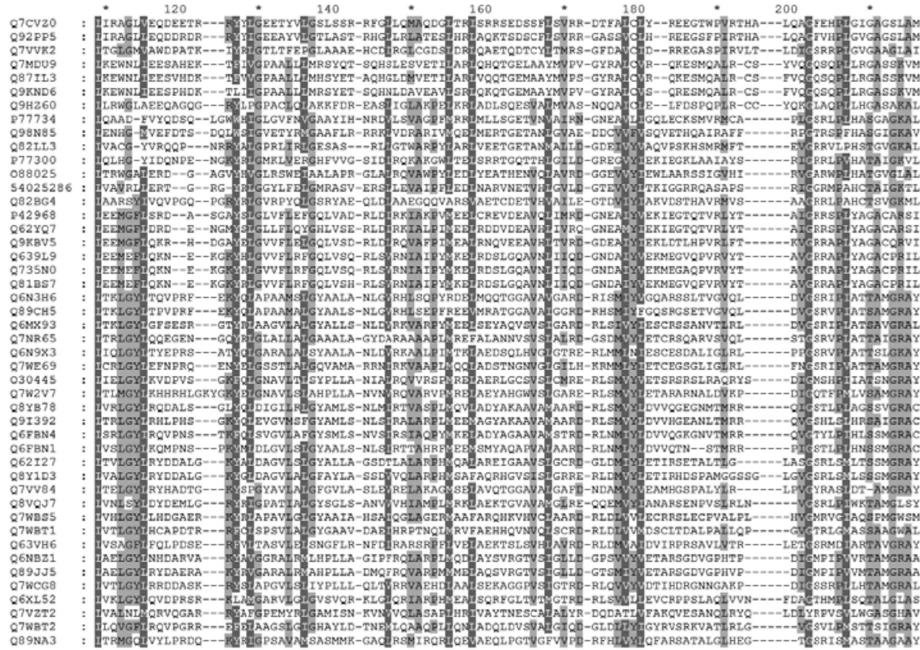


Figura 13: Comparación de alineamiento de aminoácidos y estructura secundaria.

Como se ve, esto añade un elemento de juicio a la hora de decidir si una secuencia pertenece a la familia: si además de haber obtenido un *score* significativo, comparte estructura secundaria con otros miembros de la familia (al menos en las regiones incluidas en el *profile* y en estructuras importantes como hélices o giros), podemos inferir con más seguridad que la secuencia objeto de estudio también pertenece

a la familia.

Una vez validado el *profile*, se enfrenta contra la base de datos de genomas de NCBI. De este estudio se deriva la siguiente gráfica que muestra la distribución en genomas de proteínas de IclR:

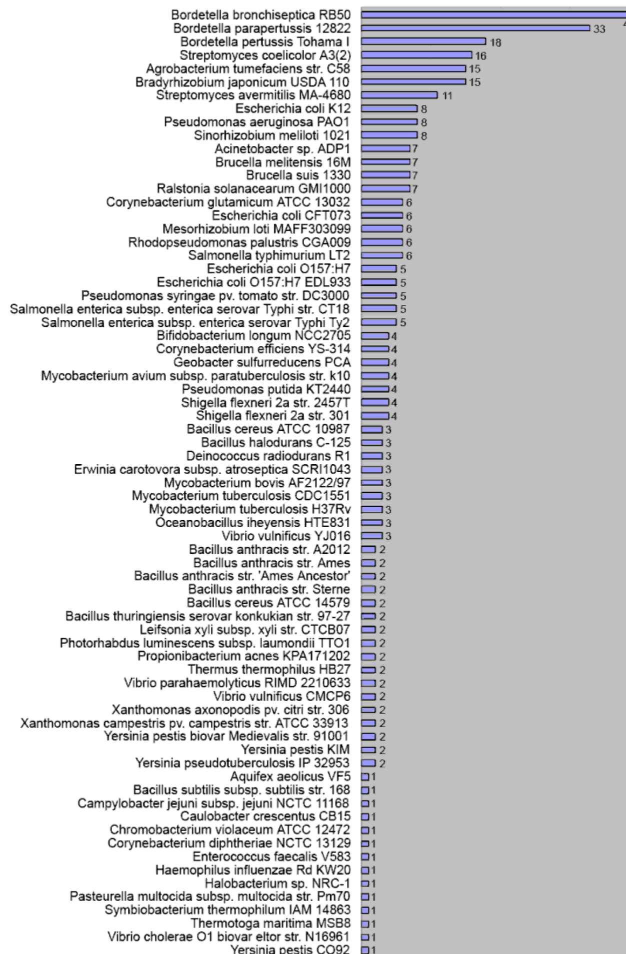


Figura 14: Distribución en genomas de la familia IclR.

Este resultado es útil para realizar estudios de genómica comparada: si una familia es muy frecuente en organismos patógenos para el humano, puede tener que ver en su patogenia; por el contrario, si es muy frecuente en no patógenos, se podría usar como blanco para diseño de drogas.

Capítulo 2.2

The IclR family of transcriptional activators and repressors can be defined by a single profile

Protein Science

FOR THE RECORD

The IclR family of transcriptional activators and repressors can be defined by a single profile

TINO KRELL, ANTONIO JESÚS MOLINA-HENARES, AND JUAN LUIS RAMOS
Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, 18008 Granada, Spain

(RECEIVED October 13, 2005; FINAL REVISION January 26, 2006; ACCEPTED January 28, 2006)

Abstract

In the last decade enormous advances in life sciences have been possible due to the information obtained from DNA sequencing projects. The optimal interpretation and analysis of genome sequence data requires the precise annotation and classification of proteins deduced from open reading frames, which is usually done with the help of family-specific signatures. Here we report a novel profile for the IclR type of transcriptional activators and repressors. In contrast to profiles for other families of transcriptional regulators, the new IclR profile is located outside the helix-turn-helix DNA-binding motif. We provide evidence that the new profile is more specific than any of the existing signatures for this family of regulators. More than 500 representatives of this family were identified with this profile. A database on bacterial regulators (http://www.bactregulators.org) was built to compile and regroup the sequences with the aid of the new profile.

Keywords: IclR; transcriptional regulator; family profile

Recent developments in functional genomics and the availability of bacterial DNA chips have revealed that microorganisms are able to alter its transcriptome pattern in response to changing environmental conditions. This involves a series of adaptive responses that are mainly triggered by regulatory proteins (Ramos et al. 2001).

The most recurrent DNA-binding motif for the binding of regulators to their corresponding promoters is a conserved DNA recognition motif that consists of an α-helix, a turn, and a second α-helix (referred to as HTH). The latter helix, termed the "recognition helix," was shown to fit into the DNA major groove (Pabo and Sauer 1992). Among HTH transcriptional regulators, families have been proposed based on common 3D structural motifs, conserved domains, and primary sequences (Nguyen and Saier 1995; Gallegos et al. 1997; Rigali et al. 2002; Ramos et al. 2005). Comparative studies have led to the determination of a specific signature for some families of bacterial regulators, and these signatures have made it possible to detect and classify new family members (Schell 1993; Gallegos et al. 1997; Rigali et al. 2002; Busenlehner et al. 2003).

One of the families of bacterial transcriptional regulators is termed IclR, which has been named after the *Escherichia coli* IclR protein. This protein controls the glyoxylate shunt and represents the best-characterized member of the family (Nègre et al. 1992; Yamamoto and Ishihama 2003). The specific functions regulated by members of the IclR family are diverse and include, for example, carbon metabolism in enterobacteriaceae (Yamamoto and Ishihama 2003), degradation of aromatic compounds by soil bacteria (Gerischer et al. 1998), solvent tolerance in *Pseudomonas* (Guazzaroni et al. 2004), inactivation of quorum sensing signals in *Agrobacterium* (Zhang et al. 2004), plant virulence by certain enterobacteriaceae (Reverchon et al. 1991), and sporulation in *Streptomyces* (Jiang and Kendrick 2000).

Reprint requests to: Juan L. Ramos, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, C/ Prof. Albareda 1, 18008 Granada, Spain; e-mail: jlramos@eez.csic.es; fax: +34-958-135740.

Article published online ahead of print. Article and publication date are at http://www.proteinscience.org/cgi/doi/10.1110/ps.051857206.

Interpro (Mulder et al. 2003) assigns proteins to the IclR family according to the PROSITE profiles PS51077 and PS51078 (Hofmann et al. 1999), the SMART domain SM00346 (Schultz et al. 2000), or the Pfam Hidden Markov Model (HMM) PF01614 (Bateman et al. 2002). The signatures used by PROSITE, SMART, and Pfam to identify IclR differ significantly and are located in different parts of the protein sequence. Pfam PF01614 HMM and PROSITE PS51078 are based on a very large segment of the proteins, the former comprising residues 82–269 and the latter comprising residues 87–272 in the IclR primary sequence, and do not consider the HTH motif. This contrasts with the SMART domain SM00346 and PROSITE profile PS51077. The former uses the HTH region of the protein and a large adjacent fragment up to residue 114 in the *E. coli* IclR primary sequence, whereas the latter is located between amino acids 24 and 86 in the primary IclR sequence. These differences between signatures do not guarantee an unequivocal identification of family members. Therefore, efforts were made to define a precise profile for the recognition of members of the IclR family of transcriptional regulators, which is reported here. This profile has allowed the identification of >500 members of the IclR family of transcriptional regulators (as of August 2005), which were found to be widely distributed in bacteria. In addition, data on IclR proteins were collected and deposited in our database of bacterial regulator proteins (http://www.bactregulators.org).

Results and Discussion

The first step in the development of the new signature for IclR family members was the selection of a seed containing 53 sequences based on the following two criteria: (1) InterPro entry IPR005471 identifies the protein unequivocally as an IclR family member; (2) the proteins were similar in size, i.e., 240–280 amino acids. BLASTCLUST analysis showed that each of the 53 proteins could be clearly distinguished from each other. The sequences were subsequently aligned with CLUSTAL (http://www.clustalw.genome.jp), which revealed three regions that were particularly well conserved in the multialignments (Fig. 1). One of the conserved regions comprises the HTH DNA binding motif located at the N terminus, a second region covers part of the N-terminal portion of the proteins toward the central region, and the third one corresponds to a segment from the central region of the protein toward the C terminus (see Fig. 1). The conserved regions were progressively extended in both directions until the global score of the multialignment diminished. The resulting alignments of these three regions were used as a seed to construct different conventional profiles, each covering a conserved region (available at http://www.bactregulators.org/docs.php). The profiles were built with the "pfmake" program available at the Swiss Institute of Bioinformatics (http://www.isrec.isb-sib.ch/ftp-server/pftools) (Bucher et al. 1996). The different profiles were confronted against all entries in the SWISS-PROT and TREMBL databases (released July 2005). We found that the profile covering the central region toward the C-terminal end (amino acids 151–229 in *E. coli* IclR) identified all IclR members recognized as such by PROSITE PS51078 and Pfam PF01614, whereas the profiles based on other segments of the protein had a reduced discriminatory capacity, and identified not only IclR family members but also regulators unequivocally ascribed to other families. A profile based on the combination of any of the conserved regions was found to be less precise than the profile that was based solely on the central region toward the C-terminal end of the multialignment. We thus considered that members of the IclR family are best identified by a profile that does not include the HTH domain of this set of proteins, and that covers a significant portion of the C terminus of the proteins. This contrasts with findings for the AraC/XylS (Gallegos et al. 1997), TetR (Orth et al. 2000; Schumacher et al. 2002; Ramos et al. 2005), and GntR (Rigali et al. 2002) families, which are best defined by a specific profile that includes the HTH DNA binding domain.

The IclR profile, available at the BacTregulators database (http://www.bactregulators.org/docs.php), was confronted against all prokaryotic proteins in the SWISS-PROT and TrEMBL (SPTR) databases (release 13-8-05) using the "pfsearch" program available at http://www.isrec.isb-sib.ch/ftp-server/pftools (Bucher et al. 1996). The program, which proposes a tentative threshold N-score of 8.5 to consider a protein as member of the IclR family, selected 546 proteins as putative members of the IclR family, of which 34 were encoded by plasmids.

To evaluate the specificity (false positives) and sensitivity (false negatives) of the new IclR profile, we used an in-house developed tool termed "Provalidator." Provalidator is a PHP-based tool that assists in the automation of profile construction and validation, and will be available free of charge at http://www.bactregulators.org. Our analysis revealed no apparent false positive proteins. A search in Interpro (Zdobnov and Apweiler 2001), a database containing all currently available classification methods for IclR proteins, assigned 587 proteins to the IclR family. The 41 proteins assigned to the IclR family by Interpro, although not identified with the new profile constructed in this study, were considered as incorrectly assigned to the family. In fact, among these 41 proteins there were three truncated polypeptides (Table 1, proteins 30, 31, and 35) and six polypeptides of reduced size (71–137 amino acids, namely proteins 23, 26, 30, 32, 34, 36, and 37 in Table 1), which made it unlikely for the latter being part of the IclR family, since our analysis revealed that these polypeptides do not possess an HTH DNA binding domain. The remaining 32 proteins assigned by

Interpro (listed in Table 1) were divided into two groups according to their score with the new profile developed here. A group of 25 proteins (Table 1, proteins 8–39, not considering the above-mentioned small or truncated proteins) yielded N-score values between 2.1 and 6.4. Alignment of these 25 proteins with IclR family members revealed substantial sequence conservation at the HTH DNA binding domain, but less sequence conservation in the C-terminal where the new profile is located (not shown). The reason why Interpro assigns these proteins to the IclR family is because PROSITE PS51077 and SMART SM00346 include the nondiscriminatory HTH region. In agreement with this observation is that, with the exception of protein Q57K18 that is exclusively recognized by SM00346, all proteins listed in Table 1 are recognized by PS51077, which is the other profile including the HTH sequence. Therefore, these proteins should be considered as incorrectly assigned to the IclR family, since a profile based on the HTH as PROSITE PS51077 lacks the necessary discriminatory potential.

The second group consisted of seven proteins with N-score values between 8.46 and 7.71 (Table 1, proteins 1–7). The alignment of these proteins to IclR family members revealed significant sequence conservation in the fragment spanning the new profile, and thus, it cannot be ruled out that these proteins are IclR family members. However, the N-score threshold of 8.5, as proposed by the "pfsearch" program, cannot be lowered in order to avoid the inclusion of non-IclR proteins. The zone between N-scores of 8.5 and 7.5 is an empirically determined buffer zone where it is recommended to consider the assignment with caution. Sequence annotation is rarely a clear-cut issue, and the purpose of this zone is to prevent the detection of false positives. We consider precision in avoiding false positives more important than the possible exclusion of any family member. Experimental characterization of these proteins will provide support for

Figure 1. (Continued on next page)

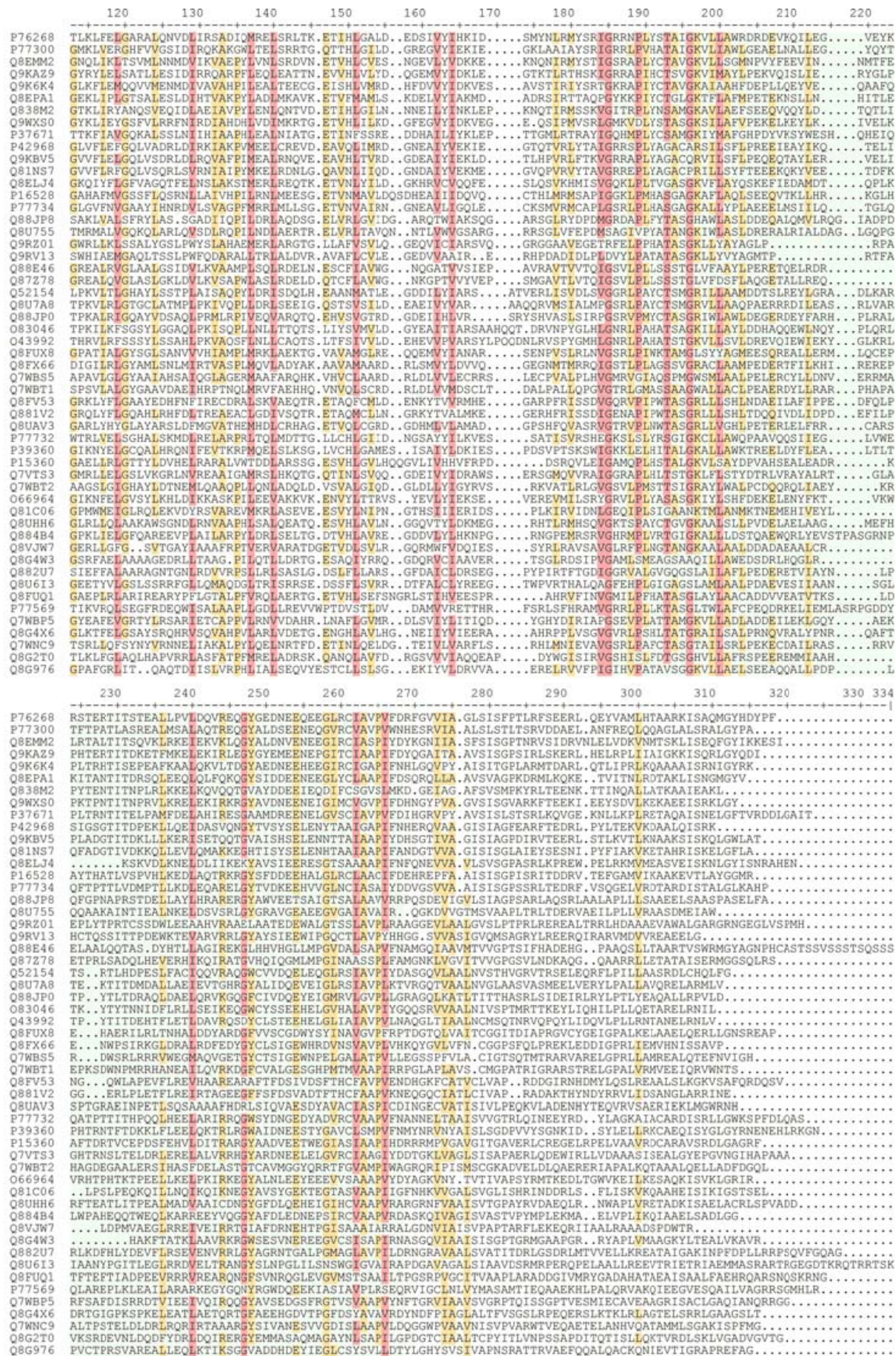


Figure 1. Multialignment of the 53 sequences used as the seed to construct the IclR family profile. Shaded in green is the conserved segment that best defined the IclR family. Bars in blue *above* the sequence indicate the HTH binding motif. Highlighted in light brown are the residues that are conserved in ≥60% of the aligned sequences, and in purple are shown the amino acids with ≥80% conservation.

Table 1. *Proteins listed as IclR family members in Interpro (Zdobnov and Apweiler 2001) but detected as non-IclR family members by the new profile*

No.	Accession no.	Description	Microorganism	Profile N-score	No. of amino acids
1	Q9AH06	Putative transcriptional regulator	*Rhodococcus erythropolis*	8.465	274
2	Q62IS0	Putative transcriptional regulator	*Burkholderia mallei*	8.337	180
3	Q9EWL2	Putative transcriptional regulator	*Streptomyces coelicolor*	8.189	255
4	Q6FBA6	Putative transcriptional regulator	*Acinetobacter* sp.	7.932	259
5	Q82D43	Putative differentiation regulon	*Streptomyces avermitilis*	7.895	213
6	Q9HW60	Probable transcriptional regulator	*Pseudomonas aeruginosa*	7.877	256
7	Q93T33	Differentiation regulon SamR	*Streptomyces ansochromogenes*	7.711	213
8	Q5YU96	Putative transcriptional regulator	*Nocardia farcinica*	6.441	299
9	Q5Z0G4	Putative transcriptional regulator	*Nocardia farcinica*	6.386	221
10	Q7W1I1	Putative transcriptional regulator	*Bordetella parapertussis*	5.558	236
11	Q7WPH1	Putative transcriptional regulator	*Bordetella bronchiseptica*	5.558	236
12	Q762I2	Hypothetical protein orf12	*Rhodococcus rhodochrous*	5.190	202
13	Q5PJJ6	Probable global regulatory protein homolog	*Salmonella paratyphi*	5.117	228
14	Q8ZM49	Putative transcriptional regulator	*Salmonella typhimurium*	4.914	228
15	Q7W4F4	Putative DNA-binding protein	*Bordetella parapertussis*	4.730	250
16	Q57K18	Putative transcriptional regulator	*Salmonella cholerae-suis*	4.712	238
17	Q7WFW4	Putative DNA-binding protein	*Bordetella bronchiseptica*	4.620	250
18	Q7WBG1	Putative transcriptional regulator	*Bordetella parapertussis*	4.546	235
19	Q7WMY2	Putative transcriptional regulator	*Bordetella bronchiseptica*	4.546	235
20	Q5YWY8	Hypothetical protein	*Nocardia farcinica*	4.197	295
21	Q938D9	Putative transcription regulator	*Mycobacterium smegmatis*	4.068	229
22	Q7VV60	Putative DNA-binding protein	*Bordetella pertussis*	3.957	250
23	Q7NW36	Hypothetical protein	*Chromobacterium violaceum*	3.810	100
24	Q5YWX8	Putative transcriptional regulator	*Nocardia farcinica*	3.608	294
25	Q73W85	Hypothetical protein	*Mycobacterium paratuberculosis*	3.479	295
26	Q4NBL4	Regulatory protein, IclR	*Arthrobacter* sp. FB24	3.295	128
27	Q6UP88	Putative enoyl-CoA hydratase	*Alcaligenes eutrophus*	3.258	448
28	Q8Z3V0	Probable global regulatory protein	*Salmonella typhi*	3.258	228
29	Q89SX5	Transcriptional regulatory protein	*Bradyrhizobium japonicum*	3.166	250
30	Q93RL8	IclR-like protein (fragment)	*Acinetobacter* sp. NCIMB9871	3.111	127
31	Q60G69	Putative regulatory protein (fragment)	*Rhodococcus* sp. DFA3	3.019	129
32	Q9AGJ8	Probable transcriptional regulator	*Corynebacterium glutamicum*	2.872	137
33	Q5YTZ3	Putative transcriptional regulator	*Nocardia farcinica*	2.798	288
34	Q6QID7	Gp12	*Burkholderia cenocepacia*	2.761	100
35	Q9S4Y5	Glyoxylate regulatory protein (fragment)	*Salmonella enteritidis*	2.725	148
36	Q65WD5	Hypothetical protein	*Mannheimia succiniciproducens*	2.504	108
37	Q5GRB7	Probable transcriptional regulator	*Alcaligenes xylosoxydans*	2.375	71
38	Q7W593	Putative transcriptional regulator	*Bordetella parapertussis*	2.338	215
39	Q7WCS4	Putative transcriptional regulator	*Bordetella bronchiseptica*	2.338	215
40	Q7VZR7	Putative transcriptional regulator	*Bordetella pertussis*	2.320	215
41	O67479	Hypothetical protein aq_1510	*Aquifex aeolicus*	2.136	219

their identification as members of the Ic1R family, but at present, such information is not available.

The IclR profile with an N-score threshold of 8.5 unequivocally identified proteins as members of the IclR family, and no false positives were found among all prokaryotic proteins that were analyzed. These results indicate that the new profile is highly effective in detecting members of the IclR family.

Using the profile defined above for the IclR family, we searched for members of this family in 228 complete microbial genomes available in NCBI (release 13-8-05). This resulted in the detection of 477 IclR members in 91 microbial genomes belonging to 60 genera of Gram-positive, α, β, and γ-proteobacteria and archaea, indicating a wide taxonomic distribution. This information can be accessed at http://www.bactregulators.org/.

The database of bacterial transcriptional regulators: BacTregulators

The profile that best defines the IclR family members, the sequences of all members of the family, their sequence alignment, as well as the available structural information together with a number of references on IclR proteins have been gathered in the BacTregulators database (http://www.bactregulators.org). This database, which can be searched with a number of different parameters such as organism, name of the regulator, accession code, or simple

text information as input information, is, in our view, a convenient tool to identify and study IclR family members.

The structural information available for IclR family proteins supports the profile as a useful tool for assigning proteins to this family

Currently, five PDB entries are available that contain structural information on IclR family members. The only full-length 3D structure of an IclR-family member is that of *Thermotoga maritima* TM0065 (PDB: 1MKM) (Zhang et al. 2002). The other four structures correspond to the effector binding domains of IclR, the glyoxylate shunt regulatory protein, YaiJ and KdgR from *E. coli* (PDB: 1TF5, 1TF1, 1YSQ, and 1YSP, respectively). All structures have in common that they were obtained in the absence of target promoter DNA or effector molecules. Structural alignments with the DALI algorithm have shown that these proteins share a similar structure, as witnessed by Z-scores >22 (see http://www.bactregulators.org/structure.php).

The TM0065 IclR protein was shown by X-ray crystallography (Zhang et al. 2002) to consist of two α/β domains: a small N-terminal DNA-binding domain with the HTH motif and a larger C-terminal effector-binding domain (Fig. 2). The latter domain consists of a five-stranded, curved β-sheet, which is flanked on both sides by several α-helices. The 79–amino acid fragment that contains the IclR profile is highlighted in yellow in Figure 2. The profile sequence forms a long loop starting at Gly151, followed by a sequence of three helices (H6–H8, of which H6 is buried and H7 and H8 are surface-exposed), and terminates with strands S5 and S6, which form the flanking part of the sheet (Fig. 2). The amino acids with the highest score (indicating that little variation is tolerated) in this new profile are shown in ball-and-stick mode. Gly151, which is labeled in Figure 2, has been proposed to play a key role in tetramerization of the protein, which is likely to occur when the protein is bound to DNA (Zhang et al. 2002). This role in tetramerization is thus likely to be responsible for the high score of Gly151 in the IclR profile. The remaining high-scoring amino acids are all located on the loop, the short buried helix, and the two strands. None of the important amino acids is located on the two long surface-exposed H7 and H8 helices. All the important amino acids are buried to a large degree and maintain multiple interactions with neighboring residues. These residues thus fulfill an important structural role, which accounts for their weight in the IclR profile.

Acknowledgments

Work at the authors' laboratory was supported by grants CICYT (2003-0515), Ministerio de Medio Ambiente (059/2004/3), and Junta de Andalucía to group CIV191. We thank M. Mar Fandila and Carmen Lorente for secretarial assistance, and Karen Shashok for improving the use of English in the manuscript.

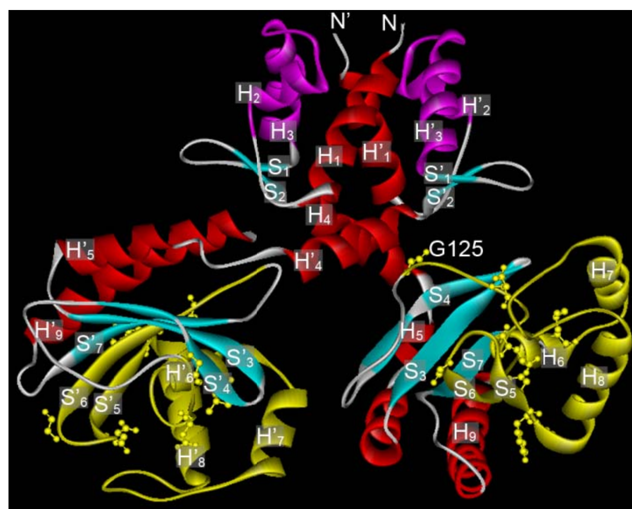


Figure 2. Schematic representation of the three-dimensional structure of the IclR dimer of *Thermotoga maritima*. Secondary structure elements are annotated, and the helix-turn-helix DNA binding domain (HTH) is shown in purple. The 79–amino acid fragment comprising the new IclR profile is highlighted in yellow. The nine amino acids with the highest score in the IclR profile are shown in ball-and stick-mode. Gly151 proposed to be involved in tetramerization is annotated.

References

Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etweiler, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M., and Sonnhammer, E.L. 2002. The Pfam protein families database. *Nucleic Acids Res.* **30:** 276–280.

Bucher, P., Karplus, K., Moeri, N., and Hoffman, K. 1996. A flexible motif search technique based on generalized profiles. *Comput. Chem.* **20:** 3–24.

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

Gallegos, M.T., Schleif, R., Bairoch, A., Hofmann, K., and Ramos, J.L. 1997. The AraC/XylS family of transcriptional regulators. *Microbiol. Mol. Biol. Rev.* **61:** 393–410.

Gerischer, U., Segura, A., and Ornston, N.L. 1998. PcaU, a transcriptional activator of genes for protocatechuate utilization in. *Acinetobacter. J. Bacteriol.* **180:** 1512–1524.

Guazzaroni, M.-E., Terán, W., Zhang, X., Gallegos, M.T., and Ramos, J.L. 2004. TtgV bound to a complex operator site represses transcription of the promoter for the multidrug and solvent extrusion TtgGHI pump. *J. Bacteriol.* **186:** 2921–2927.

Hofmann, K., Bucher, P., Falquet, L., and Bairoch, A. 1999. The Prosite database, its status in 1999. *Nucleic Acids Res.* **27:** 215–219.

Jiang, H. and Kendrick, K.E. 2000. Characterization of *ssfR* and *ssgA*, two genes involved in sporulation of *Streptomyces griseus*. *J. Bacteriol.* **182:** 5521–5529.

Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P., et al. 2003. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res.* **31:** 315–318.

Nègre, D., Cortay, J.C., Galinier, A., Sauve, P., and Cozzone, A.J. 1992. Specific interactions between the IclR repressor of the acetate operon of *Escherichia coli* and its operator. *J. Mol. Biol.* **228:** 23–29.

Nguyen, C.C. and Saier Jr., M.H. 1995. Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcription factors. *FEBS Lett.* **377:** 98–102.

Orth, P., Schnappinger, D., Hillen, W., Saenger, W., and Hinrichs, W. 2000. Structural basis of gene regulation by the tetracycline inducible Tet repressor-operator system. *Nat. Struct. Biol.* **7:** 215–219.

Pabo, C.O. and Sauer, R.T. 1992. Transcription factors: Structural families and principles of DNA recognition. *Annu. Rev. Biochem.* **61:** 1053–1095.

Ramos, J.L., Gallegos, M.T., Marqués, S., Ramos-González, M.I., Espinosa-Urgel, M., and Segura, A. 2001. Response of gram-negative bacteria to certain environmental stresses. *Curr. Opin. Microbiol.* **4:** 166–171.

Ramos, J.L., Martínez-Bueno, M., Molina-Henares, A.J., Terán, W., Watanabe, K., Zhang, X., Gallegos, M.T., Brennan, R., and Tobes, R. 2005. The TetR family of transcriptional repressors. *Microbiol. Mol. Biol. Rev.* **69:** 326–356.

Reverchon, S., Nasser, W., and Robert-Baudouy, J. 1991. Characterisation of *kdgR*, a gene of *Erwinia chrysanthemi* that regulates pectin degradation. *Mol. Microbiol.* **5:** 2203–2216.

Rigali, S., Derouaux, A., Giannotta, F., and 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.

Schell, M.A. 1993. Molecular biology of the LysR family of transcriptional regulators. *Annu. Rev. Microbiol.* **47:** 597–626.

Schultz, J., Copley, R.R., Doerks, T., Ponting, C.P., and Bork, P. 2000. SMART: A Web-based tool for the study of genetically mobile domains. *Nucleic Acids Res.* **28:** 231–234.

Schumacher, M.A., Miller, M.C., Grkovic, S., Brown, M.H., Skurray, R.A., and Brennan, R.G. 2002. Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO J.* **21:** 1210–1218.

Yamamoto, K. and Ishihama, A. 2003. Two different modes of transcription repression of the *Escherichia coli* acetate operon by IclR. *Mol. Microbiol.* **47:** 183–194.

Zdobnov, E.M. and Apweiler, R. 2001. InterProScan—An integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

Zhang, R.G., Kim, Y., Skarina, T., Beasley, S., Laskowski, R., Arrowsmith, C., Edwards, A., Joachimiak, A., and Savchenko, A. 2002. Crystal structure of *Thermotoga maritima* 0065, a member of the IclR transcriptional factor family. *J. Biol. Chem.* **277:** 19183–19190.

Zhang, H.-B., Wang, C., and Zhang, L.H. 2004. The quormone degradation system of *Agrobacterium tumefaciens* is regulated by starvation signal and stress alarmone (p)ppGpp. *Mol. Microbiol.* **52:** 1389–1401.

Capítulo 2.3

Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors

Federation of European Microbiological Societies

Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors

Antonio J. Molina-Henares, Tino Krell, Maria Eugenia Guazzaroni, Ana Segura & Juan L. Ramos

Consejo Superior de Investigaciones Científicas, Estación Experimental del Zaidín, Department of Biochemistry and Molecular and Cellular Biology of Plants, Granada, Spain

Correspondence: Juan L. Ramos, EEZ-CSIC, C/Prof. Albareda, 1, E-18008 Granada, Spain. Tel.: + 34 958 181608; fax: +34 958 135740; e-mail: jlramos@eez.csic.es

Received 22 February 2005; revised 8 June 2005; accepted 11 July 2005
First published online 13 October 2005.

doi:10.1111/j.1574-6976.2005.00008.x

Editor: Rafael Giraldo

Keywords
catabolic pathways; IclR family; Krebs cycle control; profiles; solvent efflux pumps; sporulation.

Abstract

Members of the IclR family of regulators are proteins with around 250 residues. The IclR family is best defined by a profile covering the effector binding domain. This is supported by structural data and by a number of mutants showing that effector specificity lies within a pocket in the C-terminal domain. These regulators have a helix-turn-helix DNA binding motif in the N-terminal domain and bind target promoters as dimers or as a dimer of dimers. This family comprises regulators acting as repressors, activators and proteins with a dual role. Members of the IclR family control genes whose products are involved in the glyoxylate shunt in *Enterobacteriaceae*, multidrug resistance, degradation of aromatics, inactivation of quorum-sensing signals, determinants of plant pathogenicity and sporulation. No clear consensus exists on the architecture of DNA binding sites for IclR activators: the MhpR binding site is formed by a 15-bp palindrome, but the binding sites of PcaU and PobR are three perfect 10-bp sequence repetitions forming an inverted and a direct repeat. IclR-type positive regulators bind their promoter DNA in the absence of effector. The mechanism of repression differs among IclR-type regulators. In most of them the binding sites of RNA polymerase and the repressor overlap, so that the repressor occludes RNA polymerase binding. In other cases the repressor binding site is distal to the RNA polymerase, so that the repressor destabilizes the open complex.

Introduction

Bacteria in the environment are exposed to variations in temperature, nutrients and water availability, and to toxic molecules that originate from their abiotic and biotic surroundings (including deleterious molecules that originate from their own metabolism). These changes can make their living conditions far from optimal. Survival in this unstable environment requires a wide range of rapid, adaptive responses that are triggered mainly by regulatory proteins. Most often, adaptive responses in bacteria are mediated by transcriptional regulators which, upon receiving the appropriate signal, trigger the specific transcriptional response(s) (McKay & Steitz, 1981; Henikoff *et al.*, 1988; Cohen *et al.*, 1993; Fuqua *et al.*, 1996; Surette & Bassler, 1998; Rojo, 1999; Beinlich *et al.*, 2001; Ramos *et al.*, 2001, 2002; Roessner *et al.*, 2002; Winans & Bassler, 2002; Cases *et al.*, 2003; Fromknecht *et al.*, 2003; Eggers *et al.*, 2004; Ventre *et al.*, 2004). Environmental conditions can also lead to variations in the level of RNA polymerase sigma factors, which influence promoter recognition and consequently result in altered transcriptional patterns (Karmierczak *et al.*, 2003; Rollenhagen *et al.*, 2003; Eggers *et al.*, 2004). In the case of some gram-positive microorganisms, the response to a number of environmental insults is through a cell differentiation programme that results in the formation of spores (Champness & Chater, 1994; van Wezel *et al.*, 2000; Chater, 2001).

Among transcription factors, several groups have been identified on the basis of their conserved motifs and their modes of DNA binding. Structural analyses have revealed that the helix-turn-helix (HTH) signature is the most recurrent DNA binding motif in prokaryotic transcriptional factors for the binding to their target DNA sequences (Argos *et al.*, 1983; Schevitz *et al.*, 1985; Brennan *et al.*, 1990; Harrison & Aggarwal, 1990; Branden & Tooze, 1991; Harrison, 1991; Pabo & Sauer, 1992; Bairoch & Rudd, 1995; Gallegos *et al.*, 1997; Bateman *et al.*, 2000; Pérez-Rueda & Collado-Vides,

2001; Fromknecht *et al.*, 2003; Rousseau *et al.*, 2004; Ramos *et al.*, 2005). This conserved DNA recognition motif consists of an α-helix, a turn, and a second α-helix, often called the 'recognition' helix as it is the part of the HTH motif that fits into the DNA major groove. Generally, HTH proteins bind as dimers to symmetrical DNA sequences in which each monomer recognizes a half-site (Harrison & Aggarwal, 1990; Pabo & Sauer, 1992; Orth *et al.*, 2000), although variations have been found and some proteins recognize direct repeats (Bustos & Schleif, 1993; Kessler *et al.*, 1993; Gallegos *et al.*, 1996; Harmer & Schleif, 2000; Grainger *et al.*, 2004). Among HTH transcriptional regulators, families have been identified through sequence comparisons and phylogenetic, structural, and functional analyses focused on DNA binding domains. These functional analyses have considered the HTH structure almost exclusively, as this is usually the only active motif that shows strong similarities among all members of the group (Henikoff *et al.*, 1988; Haydon & Guest, 1991; Weickert & Adhya, 1992; Gallegos *et al.*, 1993; Nguyen & Saier, 1995; Rosinski & Atchley, 1999; Rigali *et al.*, 2002). Comparative studies have led to the determination of a specific signature for each family, providing the basis for a simple method of classification and detection of new members (Henikoff *et al.*, 1988, 1990; Ramos *et al.*, 1990; Haydon & Guest, 1991; Summers, 1992; Gallegos *et al.*, 1993; North *et al.*, 1993; Schell, 1993; Fuqua *et al.*, 1994; Spiro, 1994; Poole *et al.*, 1996; Wintjens & Rooman, 1996; Martínez-Hackert & Stock, 1997; Eulberg *et al.*, 1998; Grkovic *et al.*, 2001a, b; Karmirantzou & Hamodrakas, 2001; Egan, 2002; Rigali *et al.*, 2002; Busenlehner *et al.*, 2003).

The aim of this review is to identify members of the IclR family of transcriptional regulators using a precise profile recently developed for this family, and to review the literature on the members of this family. The IclR profile can be easily used to identify IclR family members in SWISS-PROT and TrEMBL, and in all available proteins from prokaryotic genome sequences in databases.

We provide evidence that members of the IclR family of transcriptional regulators are well represented and widely distributed in bacteria. Proteins of the IclR family always have the HTH motif at their N-terminus end (Nègre *et al.*, 1991; Donald *et al.*, 1996; Pan *et al.*, 1996), and the polypeptide size of these members is typically in the range of 240–280 residues. After compiling data from protein and nucleic acid databases, we found the IclR family of regulators to include about 421 nonredundant sequences (October 2004). The specific function regulated by members of the IclR family is known for only a limited number of members (Table 1). These proteins control genes whose products are involved in the glyoxylate shunt in *Enterobacteriaceae*, multidrug resistance, degradation of aromatic compounds in soil bacteria, inactivation of quorum-sensing signals in *Agrobacterium*, determinants of plant pathogenicity by certain *Enterobacteriaceae*, and sporulation in *Streptomyces* (Table 1).

Defining the IclR family

A specific profile for the IclR family allows easy identification of IclR family members in different DNA and protein databases

The IclR family is named after the member of this group that has been most comprehensively characterized at the genetic and biochemical level – the IclR protein (Table 1). This protein controls the glyoxylate bypass in *Escherichia coli*. We found that the IclR family can be best defined by a conventional profile, which is based on sequence similarities in the region between residues 151 and 229 in the *E. coli* IclR primary sequence (T. Krell *et al.*, submitted). This profile was built using the 'pfmake' program available at the Swiss Institute of Bioinformatics (http://www.isrec.isb-sib.ch/ftp-server/pftools) (Bucher & Bairoch, 1994; Bucher *et al.*, 1996) and is available at http://www.bactregulators.org.

Using this new profile, we have searched for members of this family in 196 complete and incomplete microbial genomes available in NCBI (Release 25-10-04). Table 2 shows that members of the IclR family were detected in 61 microbial genomes belonging to 42 genera of gram-positive and *Alpha-*, *Beta-* and *Gammaproteobacteria* and *Archaea*, indicating a wide taxonomic distribution. The number of regulators in a given microorganism belonging to the IclR family was particularly high in the three strains of *Bordetella* (Table 2), which were found to contain 18–40 IclR family member proteins per genome. This observation led us to analyze whether there is a link between the pathogenicity of a microorganism and the frequency of IclR family member proteins in the genome. For these analyses we considered the complete genomes available for 135 pathogens and 57 nonpathogenic bacteria. The number of genomes for which at least one IclR protein was detected was very similar for pathogens (40%) and nonpathogens (39%). Table 2 lists the microorganisms in which IclR family members were found. The bacteria with complete genomes were classified into the following groups according to the number of IclR proteins found per genome: 0, 1–2, 3–5, 6–10, 11–20 and above 20 IclR family members per genome. The ratio of the number of pathogenic and nonpathogenic microorganisms in each of these groups was also analyzed and was found to be similar, indicating that there is no statistically relevant link between the number of IclR family members and pathogenicity.

To determine whether there is a link between IclR sequences and pathogenicity, all IclR pathogen sequences were highlighted on a phylogenetic tree showing all the IclR family sequences from completed bacterial genomes (not

Table 1. Members of the IclR family for which a function has been identified

Protein	Accession number	Microorganism	Function	References
AttJ	Q8VPD8	*Agrobacterium tumefaciens*	AttJ is the repressor of the *attMKL* operon involved in degradation of acyl-homoserine lactones that act as signals in Ti plasmid transfer	Zhang *et al.* (2004)
CatR	O33539	*Rhodococcus opacus*	Regulator of the *catABC* genes for catabolism of catechol via an *ortho*-cleavage pathway	Eulberg & Schlömann (1998)
DitR	Q9X4X3	*Pseudomonas abietaniphila*	*dit* genes encode the enzymes involved in catabolism of abietane diterpenoid. *ditR* forms a cluster with *diA3BCD* and *E* genes. Although DitR is not strictly required for growth on abietanes, a *ditR*:Km^r mutation showed reduced induction of *ditA3*.	Martin & Mohn (2000)
GclR or AllR	Q8VP30/ P77734	*Escherichia coli*	Repressor of glycolate carboligase	Rintoul *et al.* (2002)
GylR	Q828K3	*Streptomyces avermitilis*	Control of glycerol metabolism	Bolotin & Biro (1990), Hindle & Smith (1994), Smith & Chater (1988a, b)
HmgR	Q6EMJ0	*Pseudomonas putida*	The *hmgR* gene product controls the inducible expression of the divergently transcribed *hmgABC* that encodes homogentisate dioxygenase, fumarylacetoacetate hydrolase and maleylacetoacetate isomerase in the homogentisate pathway. HmgR is a repressor	Arias-Barrau *et al.* (2004)
HppR	O05148	*Rhodococcus globerulus*	The *hppR* gene forms part of the *hppCBKR* cluster involved in the metabolism of 3-(3-hydroxyphenyl)propionic acid. No functional studies on this regulator are available	Barnes *et al.* (1997)
IclR	P16528	*Escherichia coli*	Repressor of the *aceBAK* operon for glyoxylate shunt. Regulates its own synthesis	Sunnarborg *et al.* (1990)
IclR	Q8D1U3	*Yersinia pestis*	Control of acetate metabolism	Song *et al.* (2004)
KdgR	P3778/ Q6D4G5	*Erwinia* sp.	Repressor of exoenzyme production	Reverchon *et al.* (1991), Bell *et al.* (2004)
MhpR	Q9S159	*Comamonas testosteroni*	The *mhpR* gene is transcribed divergently with respect to the *mhpABDorf4orf5mhpFEC* operon. The MhpA-C proteins are involved in degradation of 3-(3-hydroxyphenyl)propionic acid. Mutants in *mhpR* fail to grow on 3-(3-hydroxyphenyl)propionic acid and 3-hydroxycinnamic acid.	Arai *et al.* (1999)
MhpR	P77569	*Escherichia coli K12*	The *mhpR* gene is transcribed divergently with respect to the *mhpABCDFE* cluster, which encodes the enzyme for catabolism of 3-(3-hydroxyphenyl)propionic acid via 2,3-dihydroxyphenyl propionic acid. MhpR induces expression of the *mhpA-E* cluster	Ferrández *et al.* (1997)
NdpR		*Rhodococcus opacus*	The NdpR protein represses expression of *ndpI* and *ndpC*, whose gene products are hydride transferases involved in the early steps of picric acid degradation. *In vivo* 2,4-dinitrophenol is an effector	Nga *et al.* (2004)
OhbR	Q9Z5W3	*Pseudomonas aeruginosa*	Putative regulator of the *ohbAB* operon for metabolism of 2-chlorobenzoate	Tsoi *et al.* (1999)
PcaR	Q52154	*Pseudomonas putida*	Positive regulator of the *pca* genes for 3,4-dihydroxybenzoate metabolism. Induces expression of *pcaIJ*, *pcaK*, *pcaF* and *pcaTBDC*	Romero-Steiner *et al.* (1994)

Table 1. Continued.

Protein	Accession number	Microorganism	Function	References
PcaR	O67983	*Rhodococcus opacus*	The *pcaR* gene is part of the *pcaHGBLRF* cluster. PcaR is the presumed regulator of the cluster.	Eulberg *et al.* (1998)
PcaR	Q6NB48	*Rhodopseudomonas palustris*	Control of protocatechuate metabolism	Larimer *et al.* (2004)
PcaU	O83046	*Acinetobacter* sp. ADP1	Positive regulator of *pcaIJFBDKCHG*, *pcaU*. The structural operon is transcribed divergently with respect to *pcaU*, and this organization influences its autoregulation	Kok *et al.* (1998)
PobR	Q43992	*Acinetobacter*	Positive regulator of *pobA*, which encodes *p*-hydroxybenzoate hydroxylase. Regulates its own synthesis	DiMarco *et al.* (1993)
RexZ	Q6CZM8	*Erwinia* sp.	Activator of exoenzyme production	Bell *et al.* (2004)
SsfR	Q9L688	*Streptomyces griseus*	Regulator involved in sporulation through activation of *ssgA*	Jiang & Kendrick (2000), Yamazaki *et al.* (2003)
SsgR		*Streptomyces coelicolor*	Regulator involved in sporulation	Zhu & Winans (1999)
TsaQ	Q6XL52	*Comamonas testosteroni*	Positive regulator needed for induction of the *tsaST* genes involved in conversion of *p*-toluenesulphonate into *p*-toluenecarboxylate	Tralau *et al.* (2003)
TtgT	Q8G976	*Pseudomonas putida*	Repressor of *ttgDEF* and *ttgGHI* operons. The TtgDEF and TtgGHI proteins form efflux pumps that extrude solvents	Duque *et al.* (2001), Mosqueda & Ramos (2000), W. Terán *et al.* (unpublished)
TtgV	Q93PU6	*Pseudomonas putida*	Repressor of *ttgDEF* and *ttgGHI* operons. The TtgDEF and TtgGHI proteins form efflux pumps that extrude solvents.	Guazzaroni *et al.* (2004), Rojas *et al.* (2003)
YiaJ	P37671	*Escherichia coli*	Positive regulator for xylulose metabolism	Schumacher *et al.* (2002)

shown). No significant clustering of sequences from pathogens was observed, demonstrating that there is no a link between IclR-family member sequences and pathogenicity, or between the number of IclR proteins per genome and pathogenicity.

Members of the IclR family are activators, repressors or proteins with dual function

A large number of studies are available on the mechanism of signal-dependent transcriptional repression and induction, and much of the information has been deduced from gene and operon structure and from genetic and mutagenesis studies. The information relevant to the IclR family shows that, in contrast to other families of regulators in which either activators or repressors are predominantly found (i.e. XylS/AraC (Gallegos *et al.*, 1997) and TetR (Ramos *et al.*, 2005), respectively), members of the IclR family include repressors, activators and proteins with a dual role: as a repressor for certain genes (including autoregulation) and as an activator for others.

The physiological role of the IclR family members has been elucidated for a limited number of these proteins using genetic or biochemical approaches. The best-characterized member of the IclR family is the *E. coli* IclR glyoxylate shunt repressor, which regulates the *aceBAK* operon encoding proteins in acetate utilization (Maloy & Nunn, 1982; Sunnarborg *et al.*, 1990). In *Streptomyces griseus* and *Streptomyces coelicolor*, IclR family members SffR and SggR are repressors that control sporulation (Nomura *et al.*, 1998; van Wezel *et al.*, 2000). In the plant pathogen *Erwinia* sp., the IclR family members KdgR (Nasser *et al.*, 1992; Liu *et al.*, 1999) and RexZ (Thomson *et al.*, 1999) regulate genes whose products are involved in the breakdown of plant cell wall polysaccharides, which are critical for the initial steps in plant tissue colonization. Whereas KgdR functions as a repressor, the RexZ protein is a positive transcriptional regulator. A good number of IclR family members regulate catabolic pathways for the degradation of aromatic compounds, such as *p*-hydroxybenzoate in *Pseudomonas putida* WCS358 (Bertani *et al.*, 2001) and *Acinetobacter* sp. ADP1 (DiMarco *et al.*, 1993), protocatechuate in *Acinetobacter* sp. ADP1 (Gerischer *et al.*, 1998; Kok *et al.*, 1998), *Rhodococcus opacus* 1CP (Eulberg & Schlömann, 1998; Eulberg *et al.*, 1998) and *Pseudomonas putida* PRS2000 (Guo & Houghton, 1999). Most of the regulators of these pathways are proteins

Table 2. Microorganisms in which IclR family members were detected

Microorganisms	No. of IclR family members	Genome size (Mbp)	Pathogenic bacteria
Acinetobacter sp. ADP1	7	3.6	–
Agrobacterium tumefaciens C58	16	11.35	+
Aquifex aeolicus VF5	1	1.59	–
Bacillus anthracis A2012	2	5.37	+
Bacillus anthracis AMES	2	5.23	+
Bacillus anthracis 'Ames Ancestor'	2	5.5	+
Bacillus anthracis Sterne	2	5.23	+
Bacillus cereus ATCC 10987	3	5.22	+
Bacillus cereus ATCC 14579	2	5.43	+
Bacillus cereus ZK	2	5.3	+
Bacillus halodurans C-125	3	4.2	–
Bacillus licheniformis ATCC 13032	6	8.44	+
Bacillus subtilis subsp. *subtilis*168	1	4.21	–
Bacillus thuringiensis 97–27	2	5.24	+
Bifidobacterium longum NCC2705	4	2.26	–
Bordetella bronchiseptica RB50	40	5.34	+
Bordetella parapertussis 12822	33	4.77	+
Bordetella pertussis Tohama I	18	4.09	+
Bradyrhizobium japonicum USDA 110	15	9.20	–
Brucella melitensis 16M	7	3.29	+
Brucella suis 1330	7	3.32	+
Burkholderia mallei ATCC 23344		–	+
Burkholderia pseudomallei K96243	13	7.25	+
Campylobacter jejuni subsp. *jejuni* NCTC 11168	1	1.64	+
Caulobacter crescentus CB15	1	4.02	–
Chromobacterium violaceum ATCC 12472	1	4.75	+
Comamonas testosteroni			–
Corynebacterium diphtheriae NCTC 13129	1	2.49	+
Corynebacterium efficiens YS-314	4	3.15	–
Corynebacterium glutamicum ATCC 13032	6	3.31	–
Deinococcus radiodurans R1	3	3.28	–
Enterococcus faecalis V583	1	3.22	+
Erwinia carotovora subsp. atroseptica SCRI1043	3	5.06	+
Erwinia sp.			–
Escherichia coli CFT073	6	5.23	+
Escherichia coli K12	8	4.64	–
Escherichia coli O157:H7	5	5.59	+
Escherichia coli O157:H7 EDL933	5	5.53	+
Geobacter sulfurreducens PCA	4	3.81	–
Haemophilus influenzae Rd KW20	1	1.83	+
Halobacterium sp NRC-1	1	2.57	–
Leifsonia xyli subsp. *xyli* CTCB07	2	2.58	+
Mannheimia succiniciproducens MBEL 55E	2	2.31	–
Mesorhizobium loti MAFF303099	6	7.6	–
Mycobacterium avium subsp. paratuberculosis k10	4	4.83	+
Mycobacterium bovis AF2122/97	3	4.35	+
Mycobacterium tuberculosis CDC1551	3	4.4	+
Mycobacterium tuberculosis H37Rv	3	4.41	+
Nocardia farcinica IFM 10152	13	6.29	+
Oceanobacillus iheyensis HTE831	3	3.63	–
Pasteurella multocida subsp. *multocida* Pm70	1	2.26	+
Photobacterium profundum SS9	4	6.4	–
Photorhabdus luminescens subsp. *laumondii* TT01	2	5.69	+
Photorhabdus luminiscens			–
Propionibacterium acnes KPA171202	2	2.56	+
Pseudomonas abietanphyla			–
Pseudomonas aeruginosa PAO1		6.26	+

Table 2. Continued.

Microorganisms	No. of IclR family members	Genome size (Mbp)	Pathogenic bacteria
Pseudomonas putida KT2440	4	6.18	–
Pseudomonas syringae DC3000	5	6.4	+
Ralstonia solanacearum GMI1000	7	5.81	+
Rhodococcus globerus			–
Rhodococcus opacus			–
Rhodopseudomonas palustris CGA009	6	5.46	+
Salmonella enterica subsp. *enterica serovar typhi* CT18	5	5.13	+
Salmonella enterica subsp. *enterica serovar typhi* Ty2	5	4.79	+
Salmonella typhimurium LT2	6	4.95	+
Shigella flexneri 2a 2457T	4	4.6	+
Shigella flexneri 2a strain 301	4	4.61	+
Sinorhizobium meliloti 1021	8	6.69	–
Streptomyces avermitilis MA-4680	11	9.12	–
Streptomyces coelicolor A3(2)	16	9.05	–
Streptomyces griseus			–
Symbiobacterium thermophilum IAM 14863	1	3.57	–
Thermotoga maritima MSB8	1	1.86	–
Thermus thermophilus HB27	2	2.13	–
Vibrio cholerae O1 *biovar eltor* N16961	1	4.03	+
Vibrio parahaemolyticus RIMD 2210633	2	5.17	+
Vibrio vulnificus CMCP6	3	5.26	+
Vibrio vulnificus YJ016	2	5.13	+
Xanthomonas axonopodis 306	2	5.18	+
Xanthomonas campestris pv. *campestris* ATCC 33913	2	5.08	+
Yersinia pestis biovar medievalis 91001	2	4.8	+
Yersinia pestis CO92	1	4.83	+
Yersinia pestis KIM	2	4.6	+
Yersinia pseudotuberculosis IP 32953	2	4.84	+

Genome size and number of IclR-like family members are only given for microorganisms for which the complete genome is known.

with dual function, since they control their own synthesis while at the same time activating transcription of the catabolic operons (see below). There are many other cases in which the gene specifying an IclR family member is adjacent to a probable or demonstrated gene or operon of unknown function, but no regulatory link between the regulator and the gene has yet been found. Thus, proteins of the IclR family appear to have diverse regulatory roles in bacteria, and an adequate understanding of their mode of action is of interest (see below).

Crystal structure of members of the IclR family

There are currently five 3-D structures of IclR proteins available, namely, the full-length protein from *Thermotoga maritima* TM0065 (Zhang RG *et al.*, 2002), the effector binding domains of two glyoxylate regulatory proteins (PDB entries 1TF1 and 1TDJ), YaiJ (PDB entry 1YSQ) and KdgR (PDB entry 1YSP). For the latter four structures there are currently no literature references available. All structures are obtained without bound DNA and effector. The full-length structure of the *Thermotoga* IclR is used below to illustrate the 3-D structure of the members of this family.

The TM-IclR protein

The *T. maritima* gene is located within a cluster of genes that are homologous to genes whose translated products are involved in xylulose metabolism. In *E. coli*, the catabolism of endogenously formed xylulose is mediated by IclR family member YiaJ (Schumacher *et al.*, 2002). The crystal structure of the TM0065 protein, also known as TM-IclR, was resolved at 2.2 Å (Zhang RG *et al.*, 2002). The protein consists of two α/β domains, a small N-terminal DNA binding domain, and a large C-terminal putative signal binding domain. These domains are linked by an α-helix and a short loop. The protein crystallized as a dimer and the dimerization interface comprises the α-helices H1 and H4 (Fig. 1). The hydrophobic interface formed by the two H1 α-helices from the N-terminal domains includes four aromatic side chains, two from each monomer. A number of additional hydrophobic interactions (such as Ala7–Leu4) and

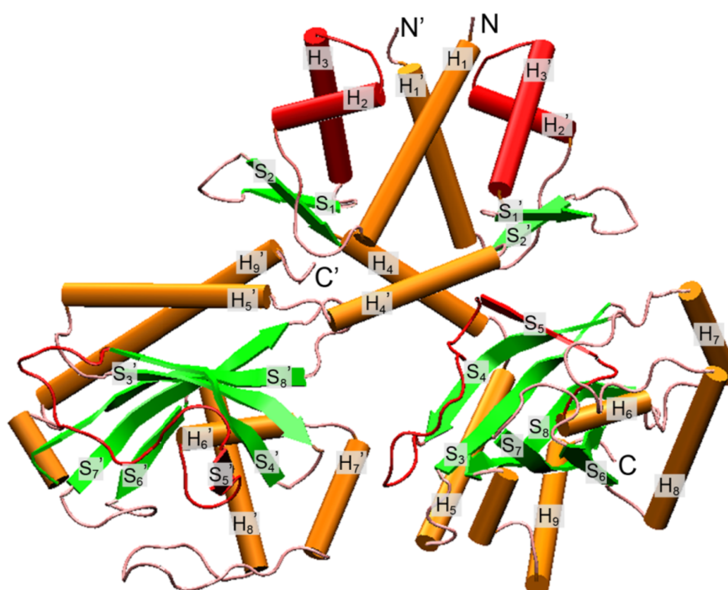


Fig. 1. *Thermotoga maritima* TM-IclR dimer. Solid orange cylinders represent the α-helixes (H) and the green arrows represent the β-sheet (S). The helix-turn-helix of each monomer is shown in red (H2 and H3) and (H_2' and H_3').

hydrogen bonds (such as Asp12–Tyr67) were also observed at the dimer interface (Zhang RG *et al.*, 2002). The dimer is asymmetrical with respect to the orientation of the signal binding domains.

The HTH DNA binding motif at the N-terminal domain is composed of two consecutive α-helices (H2 and H3 in Fig. 1 and the 4-residue turn connecting them). The H2 α-helix is aligned at almost 90° with respect to the predicted recognition α-helix (H3). The H3 of *Thermotoga* IclR has four turns, making it larger than the corresponding helix in QacR and TetR, which has two turns (Orth *et al.*, 2000; Schumacher *et al.*, 2002), and the helix in MarA and BmrR, with three turns (Rhee *et al.*, 1998; Heldwein & Brennan, 2001). The HTH motif is followed by a β-hairpin, which connects to the linking region, and the global architecture of the DNA binding domain resembles that of the winged HTH (Clubb *et al.*, 1994). The distance between HTH motifs in the dimer is 33.45 Å.

The C-terminal domain is composed of a six-stranded antiparallel β-sheet, five α-helices, and one short 3_{10} helix (Fig. 1). The strongly curved β-sheet forms a half barrel and divides the five α-helices into two subgroups: two relatively short α-helices (H4 and H5) that lie on the outside of the curved sheet, and three helices (H6, H7 and H8) on the inside, forming an α/β/α structure. The H6 α-helix is inserted into the half barrel and is parallel to the sheet surface. In the dimer, the C-terminal domains do not contact each other. The C-terminal ligand binding domains, which are not directly involved in dimer formation, are not related by two-fold symmetry, thus making the dimer asymmetrical. At this point it is not clear whether this observation is of physiological relevance or whether it is an artifact of crystallization. The linking regions (residues 62–79) in the two monomers in the asymmetric unit adopt two different conformations, which results in two different orientations of the ligand binding domains with respect to the HTH dimer.

In the TM-IclR structure, the proposed substrate binding pocket is occupied by water molecules and a divalent metal ion (assigned as Zn^{2+}) bound to Cys196. This effector binding site has been proposed to be a pocket formed by the inner surface of the curved β-sheet, helix H6 and a loop (the latter being highlighted in red in Fig. 1). Interestingly, parts of this proposed effector binding site (namely residues 214–220 and 114–118, Fig. 2) were proposed to mediate tetramer formation and the proximity of the presumed ligand binding region to the region involved in tetramerization suggests that ligand binding and tetramerization may be linked (Fig. 2). Because tetramerization and DNA binding are also connected, the proximity of the signal binding and tetramerization domains as suggested in the crystal structure may provide the link between signal binding and DNA binding, but this interpretation awaits the results of further *in vitro* studies.

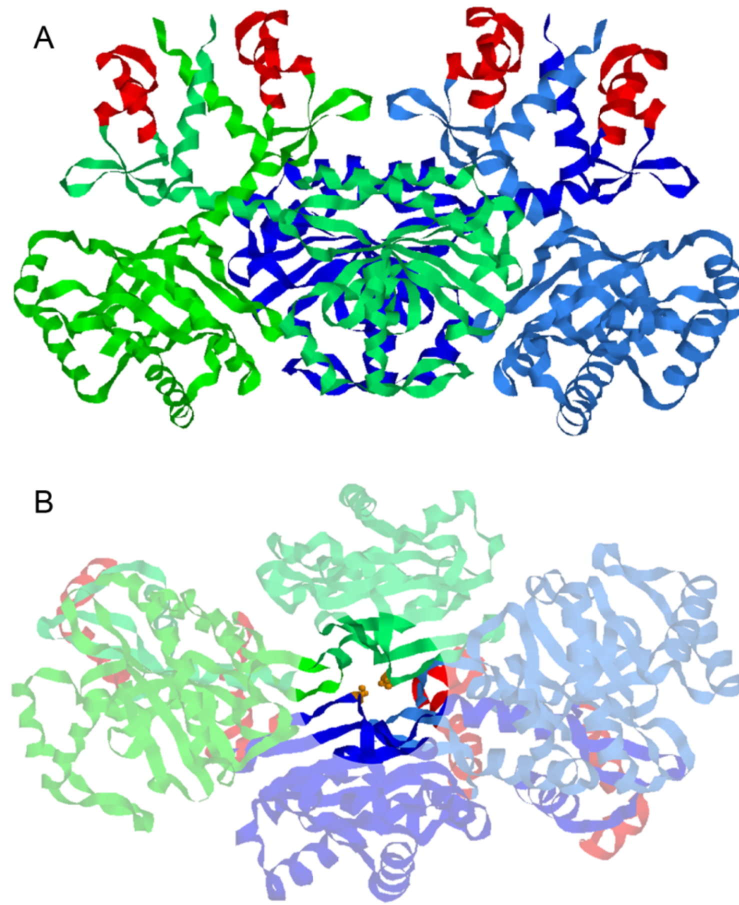


Fig. 2. (a) Tetramer of IclR. The two monomers forming a dimer are colored in blue or green and the helix-turn-helix of each monomer is highlighted in red. (b) Perpendicular view of (a) to highlight Gly151, shown as an orange ball.

IclR proteins are structurally conserved and show similarities with GAF domains

To address the question of whether IclR proteins form a separate group of structurally distinct proteins, all currently available 3-D protein structures were aligned with the effector binding domain of *E. coli* IclR using the Dali algorithm (Holm & Sander, 1996). The obtained fits were sorted using the *Z*-score (an indication of the quality of the alignment) and best structural fits are shown in Table 3. An identical alignment (*E. coli* IclR with itself) is characterized by a *Z*-score of 36.8. As mentioned above, there are a total of five IclR structures currently available in the protein data bank and very close fits were obtained for the alignment of *E. coli* IclR with the remaining four members of the IclR family (*Z*-scores > 22). These fits were global and covered the entire effector binding domain. At least 93% of amino acids could be superimposed with an Rmsd of less than 2 Å (Table 3). The sequence and length of all secondary structure elements in these five proteins were found to be very similar and it can be concluded that these five IclR proteins share a similar structure (Fig. 3).

Members of the IclR family of regulators exhibit significant *Z*-scores (9.3–6.4) with a few non-IclR proteins indicating a certain degree of structural similarities. The three best alignments with non-IclR proteins are with proteins exhibiting GAF domains (Hopper *et al.*, 1996; Korsa & Bock, 1997; Ho *et al.*, 2000; Anantharaman *et al.*, 2001). GAF domains are ubiquitous structural motifs present in c**G**MP-regulated cyclic nucleotide phosphodiesterases, certain **A**denyl cyclases, the bacterial transcription factor **F**hlA (letters in bold indicate the origin of the acronym), and hundreds of other signalling and sensory proteins from all three kingdoms of life (Beinlich *et al.*, 2001, see below). The GAF domain can be partially superimposed on the effector binding domain of *E. coli* IclR (PDB entry 1TD5), thus

Table 3. Structural alignments using the DALI algorithm of the *Escherichia coli* IclR effector domain with the equivalent region of protein with other members of the IclR family*

Protein	Microorganisms	pdb code	*Z*-score†	Rmsd (amino acids)‡	Sequence identity over equivalent region (%)
IclR	*Escherichia coli*	1TD5	36.8	0 (171)	100
Glyoxylate regulatory protein	*Escherichia coli*	1TF1	26.9	1.4 (168)	44
IclR homologue	*Thermotoga maritima*	1MKM	24.3	1.9 (166)	31
YaiJ	*Escherichia coli*	1YSQ	24.1	1.6 (167)	28
KdgR	*Erwinia*	1YSP	22.4	1.9 (159)	30

*Analyses were performed in April 2005.
†*Z*-score: strength of structural similarity in standard deviations above expected.
‡Rmsd (amino acids): positional root mean square deviation of superimposed C_A atoms in Ångstroms. The total number of equivalent residues is given in brackets.

sharing the curved and antiparallel β-sheet with the same strand order, and the three helices corresponding to H5, H6 and H9 in the *Thermotoga* TM-IclR structure (Fig. 1).

The structural similarities identified for GAF and IclR proteins may reflect a similar function, which could consist of both protein families sensing a variety of small molecules. IclR proteins were shown to bind to many structurally different small molecules, as detailed in the following sections of this review. Likewise, GAF domains do not only bind cyclic nucleotides, but have also been found to bind to formate, 2-oxoglutarate or aromatic compounds such as tetrapyrroles and photopigments (Rost & Sander, 1993; Reyes-Ramírez *et al.*, 2002; Little & Dixon, 2003).

The sequences of IclR proteins can be subdivided into five major orthologous subgroups (data not shown). However, the five 3-D structures currently available on IclR proteins all group in one of these subgroups. To determine whether members of all IclR subgroups share a similar sequence of secondary structure elements, a representative from all subgroups was selected on the basis that the function is known and that no 3-D structural information is available. The only IclR sequence from an archaebacterium has also been included in this analysis (Fig. 3).

Initial experiments were aimed at identifying the algorithm that predicts best the secondary structure for the IclR-TM of *Thermotoga* for which the 3-D structure is available. Amongst all the currently available algorithms, the prediction method using PHD (Rost & Sander, 1993) turned out to yield the prediction that coincides best with the 3-D structure. This algorithm has therefore been employed to predict the secondary structure of the representatives of each subgroup and the archaebacterial sequence. Figure 3 shows an alignment of the secondary structure found in the IclR protein of the *Thermotoga* protein with the six predictions. Considering that PHD has a precision of 70%, there is a good agreement between the different proteins. The agreement is best at the N-terminal DNA binding domain. For the overall alignment, the helical regions appear to be predicted with more precision than the strands. The region with the most pronounced differences in secondary structure was found between residues 115 and 140. In the *Thermotoga* IclR-TM protein this fragment forms loop L3 (containing a β-strand), which has been proposed to be involved in the binding of the effector molecule. The difference in secondary structure observed in this region might be due to the fact that the different proteins analyzed bind to different effector molecules. However, the overall good agreement in the secondary structure of IclR proteins might indicate that they share a very similar 3-D structure.

Potential mechanistic and structural reasons for conservation of amino acids in the IclR signature

To identify the reason for the conservation of the amino acids within the IclR signature, we analyzed four residues that are almost completely conserved in the signature defining the family, namely, G151, P155, G162 and G208 in the *E. coli* IclR primary sequence.

The asymmetrical unit of the crystal used to generate the structure of the *Thermotoga* protein (Zhang RG *et al.*, 2002) was shown to contain a dimer which is proposed to be the minimal DNA binding unit. However, the authors propose that the protein forms a tetramer when bound to DNA. This tetramer is shown in Fig. 2a, where the two monomers forming a dimer are colored blue or green. Figure 2b, a perpendicular view of Fig. 2a, shows that G151 is located on the tetramer interface. There is a distance of 2.9 Å between the two Cα atoms of the glycines, indicating a hydrophobic interaction. Furthermore, this residue is located on loop L3 (shown in red in Fig. 2b), which is proposed to form part of the effector binding site and is likely to move upon ligand binding. These features suggest that G151 is likely to play a key role in the effector-mediated dissociation of tetramers into dimers, which in turn leads to release of the protein from DNA. This important

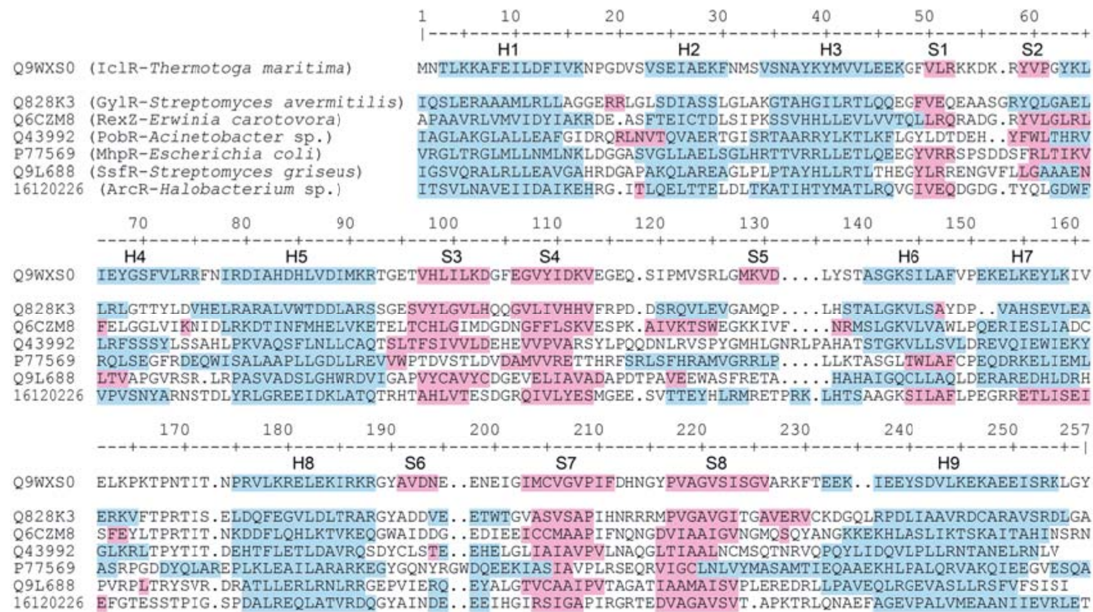


Fig. 3. Structural alignment of secondary structure elements found in the 3-D structure of the IclR protein of *Thermotoga* with secondary structure predictions of IclR sequences, representative of all five major orthologous subgroups. Sequences were selected randomly, but for all sequences no 3-D information is available and the function is known. The single sequence of an IclR protein detected in an archaebacterium (*Halobacterium* sp., see Table 2) has also been added to this alignment. Secondary structure was predicted using the PHD algorithm (Rost & Sander, 1993). Sequence numbering refers to the *Thermotoga* protein and the secondary structure elements correspond to those shown in Fig. 1. Helices are in blue and β-strands in red.

mechanistic role of G151 is thus proposed to be the reason for its conservation.

Zhang *et al.* (2002) proposed that the effector binding site is composed of the central β sheet as well as loop L3 and helix H6 (Fig. 1). Pro155 is situated on loop L3 and might contribute to the constrained conformation of this loop. Gly162 is situated in the hinge region between L3 and H6. It is likely that L3 and H6 fold over the bound effector, in analogy to the lid domain found in many ATP binding proteins. Glycine residues are frequently found in hinge regions of mobile loops, and we suggest that this is also the reason for the conservation of Gly162. In view of these findings, mechanistic roles related to effector binding are suggested to be responsible for the conservation of Pro155 and Gly162.

Helix H8 and strand S5 are linked by a sharp turn where Gly208 is located. Because of its flexible geometry, glycine residues can enable such sharp turns and thus a structural role is proposed for this residue.

Analysis of the HTH of IclR-family

Using BLAST-CLUST, Krell *et al.* (submitted) found that IclR members clustered in 189 groups based on the HTH domain. A member of each cluster was chosen and the HTH was aligned. These authors noticed that L or M, located in the turn between both helices, was highly conserved. Only two residues were highly conserved in the first helix, namely, those at positions 3 and 6 in the multialignment. Residues at positions 3 and 6 corresponded to amino acids Val23 and Ile26 in the *Thermotoga* IclR-TM protein. Both residues play a key role in maintaining the structure of this DNA binding domain by establishing numerous hydrophobic contacts. Val23 interacts with Ala37 and with Tyr38, located on helix 3, as well as with Tyr58 of strand 2 (Fig. 1). Ile 26 interacts with helix 1 by maintaining contacts with Ile10 and Phe13 and also with Val21 and Phe30 located on the N- and respectively C-terminal end of helix2. The small number of amino acids conserved in the first helix of the HTH is in contrast to findings in the XylS/AraC family and TetR families, in which the most conserved residues were in the first helix. In fact, the conservation of residues in the first helix in the HTH motif (H2) of the AraC/XylS and TetR family was considered critical for the structure and orientation of the recognition helix H3 (Gallegos *et al.*, 1997). Low conservation at the recognition helix in the AraC/XylS and TetR family was interpreted as part of variability enabling recognition (the second helix in the HTH motif) of different

promoters. An interesting feature is that the 3-D structure of IclR-TM (Fig. 1) has a longer recognition helix compared to other regulators such as QacR or TetR (Schumacher *et al.*, 2002; Berens & Hillen, 2003).

Krell *et al.* (submitted) also found that three hydrophobic residues were conserved in the recognition helix of the HTH, and suggested that they could play a certain structural role. The most conserved position in the HTH was a Leu found in 91% of the sequences. This corresponded to Leu41 in the TM-IclR protein, where it is part of the hydrophobic network at the dimer interface of the HTH domain.

Lysine and arginine residues frequently dominate the interaction of proteins with DNA. Krell *et al.* (submitted) found that there was only one highly conserved charged residue in the second helix of the HTH, which was an Arg in 75% of the cases. This residue corresponded to Arg36 in the TM-IclR protein and it is likely to interact with DNA.

PobR is the positive transcriptional regulator of the *pobA* gene involved in *p*-hydroxybenzoate metabolism in *Acinetobacter.* A number of mutants at the N-terminal of this protein with defects in regulation have been isolated (Kok *et al.*, 1998). In light of the above findings, a 3-D model of PobR was generated using the structure of the IclR-TM as a template. We found that in PobR the recognition helix extended from residue 43 to 71. Mutations within this stretch of amino acids in PobR (R56 → S, K64 → M and K67 → I) resulted in mutants with reduced DNA binding capacity.

Circuits in which IclR family members are involved

The IclR protein: genetics of the regulation of the glyoxylate cycle shunt

To grow on acetate or fatty acids, *Escherichia coli* requires the induction of the glyoxylate bypass (Fig. 4). This pathway consists of isocitrate lyase (*aceA*) and malate synthase (*aceB*), and is essential for growth on these carbon sources because it prevents the quantitative loss of the entering carbon as CO_2 in the Krebs cycle (Kornberg, 1966; Brice & Kornberg, 1968). Once induced, the flow of isocitrate through this pathway is controlled by the phosphorylation of isocitrate dehydrogenase (IDH), one of the Krebs cycle enzymes that competes with isocitrate lyase for isocitrate (Garnak & Reeves, 1979; Borthwick *et al.*, 1984; LaPorte *et al.*, 1984; LaPorte & Chung, 1985). This phosphorylation cycle is catalyzed by a bifunctional protein, IDH kinase/phosphatase (*aceK*) (LaPorte & Koshland, 1982; LaPorte *et al.*, 1989).

The *aceBAK* operon (Byrne *et al.*, 1988; Chung *et al.*, 1988; Cortay *et al.*, 1988; Klumpp *et al.*, 1988; Rieul *et al.*, 1988) is expressed from a single promoter during growth on acetate (Chung *et al.*, 1988), and is repressed if any preferred carbon source (e.g. glucose or pyruvate) is present simultaneously. The expression of this operon is regulated by a repressor protein encoded by *iclR*, as suggested by the finding that mutations in *iclR* yielded constitutive expression of the *aceBAK* operon (Maloy & Nunn, 1982; Kornberg, 1967; Sunnarborg *et al.*, 1990; Nègre *et al.*, 1991, 1992). IclR also regulates its own expression, and gel shift assays demonstrated that IclR binds to its own promoter (Gui *et al.*, 1996). IclR binds to two sites at the *aceB* promoter, one of which overlaps the −35 region (Chung *et al.*, 1988; Cortay *et al.*, 1991; Nègre *et al.*, 1991, 1992; Donald *et al.*, 1996); (see also Fig. 5). The other site is located further upstream in the −100 region and is essential for the mechanisms of action of this protein (see below). Although *aceBAK* and *iclR* are both regulated by IclR, *aceBAK* expression responds to the presence of different carbon sources, whereas the expression of *iclR* does not. The available evidence suggests that integration host factor (IHF) is responsible for this difference, since an IHF site was identified upstream from the IclR binding site of *aceBAK* (Resnik *et al.*, 1996). IHF contributes to the induction of *aceBAK* by opposing repression by IclR during growth on acetate (inducing conditions) but not on glucose (repressing conditions). It was proposed that the expression of *iclR* may be relatively insensitive to the carbon source because it does not have a binding site for IHF. Consistent with this proposal was the finding that when the IHF site upstream from *aceBAK* was inactivated, the response of *aceBAK* expression to the carbon source closely resembled the expression pattern of *iclR*.

A sequence just upstream of the *iclR* promoter bears a striking resemblance to FadR binding sites found in the fatty acid metabolic genes (DiRusso *et al.*, 1992; Henry & Cronan, 1992). Disruption of *fadR* resulted in an increase in the expression of *aceBAK* (Maloy & Nunn, 1982), as shown by measurements of the expression of IDH phosphatase during growth on glucose and acetate. Mutation in either the *fadR* gene or the FadR binding site of *iclR* yielded similar increases in the expression of IDH phosphatase. Furthermore, inactivation of the FadR binding site of *iclR* decreased the expression of an *iclR*::*lacZ* operon fusion, indicating that FadR activates the expression of *iclR*. *In vitro* gel shift and DNaseI footprint analyses confirmed that FadR binds to a site just upstream from the *iclR* promoter. This is evidence that FadR activates *iclR* expression and indirectly regulates *aceBAK* by altering the expression of IclR. The implication of these findings is that control of the glyoxylate bypass operon (*aceBAK*) in *E. coli* is mediated by two regulatory proteins, IclR and FadR.

IclR protein purified on chromatographic Sephadex G-100 columns and in sucrose density gradients behaves as

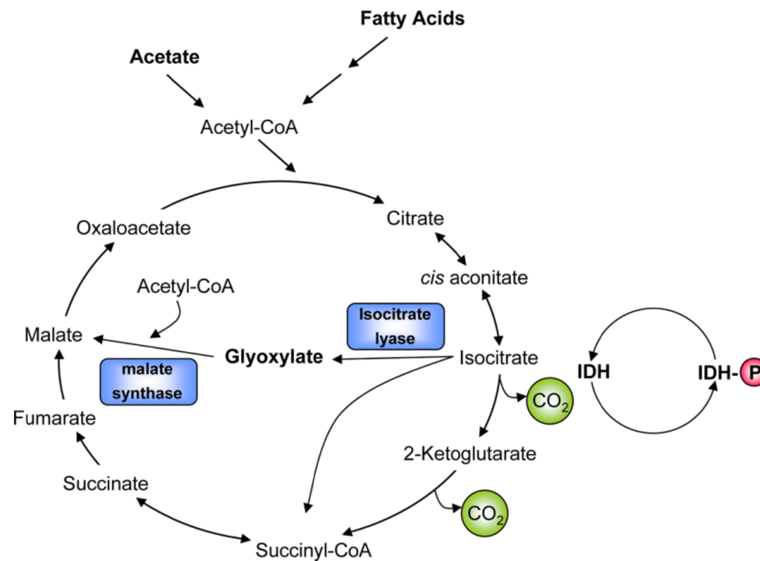


Fig. 4. Krebs cycle and the glyoxylate shunt. Isocitrate lyase and malate synthase are encoded respectively by *aceA* and *aceB* genes. The activity of isocitrate dehydrogenase is controlled by phosphorylation (shown to the right).

a tetramer (M_r~117 000) (Nègre *et al.*, 1991). IclR binding to DNA is co-operative, with polymerization of IclR subunits as they bind to the DNA. This co-operativity is consistent with evidence for an equilibrium between IclR single subunits, dimers and tetramers at high concentrations.

DNase I footprinting and *in vitro* transcription assays indicated that IclR binds to an IclR box (Chung *et al.*, 1988; Cortay *et al.*, 1991; Donald *et al.*, 1996; Pan *et al.*, 1996; Gui *et al.*, 1996). The IclR binding sequences agree with a consensus palindromic sequence, 5′-TGGAAATNATTTC-CA-3′ (Pan *et al.*, 1996). In view of the palindromic nature of the IclR box and the size of the protected region in footprint assays, the functional IclR promoter is most likely recognized by a tetramer (Nègre *et al.*, 1992; Donald *et al.*, 2001).

Yamamoto & Ishihama (2003) proposed two different mechanisms of transcription repression from the *iclR* and *aceB* promoters. The IclR protein alone protected a region between +14 and −21 in the P_{iclR} promoter (Gui *et al.*, 1996). The RNAP holoenzyme alone showed a different footprinting pattern, including the protected region between +14 and −59 (Yamamoto & Ishihama, 2003). When the RNAP holoenzyme was added after the formation of the IclR–DNA complex, the protection pattern was similar to that with IclR alone. In contrast, the addition of IclR after incubation of RNAP holoenzyme with the *iclR* promoter did not lead to significant changes in the DNase I protection pattern from that seen with the RNAP holoenzyme alone. These observations supported the competition model of repression of *iclR* transcription by IclR.

The binding sites of RNAP and IclR on the *aceB* promoter were defined by DNase I footprinting assays. The RNAP alone protected a long sequence from −62 to +17 on the coding strand, and a sequence from −61 to +22 on the noncoding strand. DNase I footprinting experiments showed two binding sites for IclR, i.e. an IclR box I between −125 and −99 and an IclR box II between −52 and −19 in the *aceB* promoter region (Fig. 7, left). All IclR contact sites were located on the same face of DNA.

Yamamoto & Ishihama (2003) also carried out mechanistic studies in the *aceB* promoter. The binding of IclR to the promoter-proximal IclR box II prevented RNAP from binding to the *aceB* promoter, as in the repression of the *iclR* promoter by IclR. In contrast, IclR appears to bind to the promoter-distal IclR box I site, even after the RNAP binds to the *aceB* promoter. Yamamoto & Ishihama (2003) measured the level of stable *aceB* open complexes in the presence and absence of IclR. The *aceB* open complex was gradually dissociated with a half-life of 16 min, after the addition of heparin, a compound known to bind to DNA. In the presence of IclR, however, the half-life of the *aceB* open complex decreased to 6 min, suggesting that IclR destabilizes the *aceB* promoter open complex. In contrast, the stability of the *iclR* promoter open complex was not affected by the addition of IclR. This series of findings led Yamamoto & Ishihama (2003) to suggest that IclR induces dissociation of the *aceB* open complex, ultimately resulting

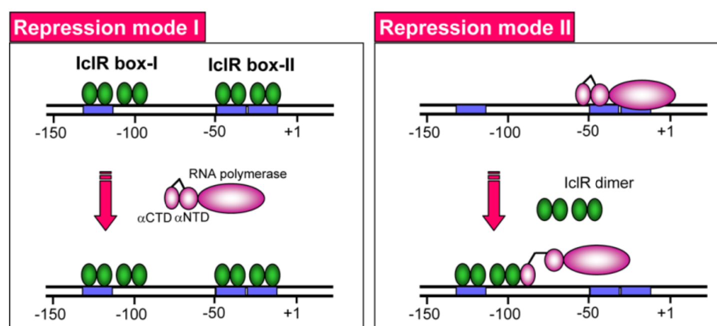


Fig. 5. Modes of repression mediated by *Escherichia coli* IclR protein. Left, repression mode I involves IclR bound to its target site overlapping RNA-polymerase binding site, so that occupancy of the site by IclR prevents the entry of the polymerase. Right, repression mode II involves IclR bound to a distal site with respect to the RNA polymerase site. The interaction of IclR with the α-subunit of the promoter-bound RNA polymerase destabilizes and disassociates the open complex (Yamamoto & Ishihama, 2003).

in the repression of *aceB* transcription even after open complex formation (Fig. 5, right).

The involvement of two different modes of repression for IclR may allow fine regulation of *aceB* transcription at different levels, in response to the intracellular concentration of IclR or the level of IclR effectors.

Control of catabolic pathways for aromatic hydrocarbons by IclR family members

A number of regulators belonging to the IclR family are involved in the control of catabolic pathways for the degradation of aromatic compounds (McFall *et al*., 1998). PcaR and PcaU in *P. putida* and *Acinetobacter* sp., respectively, are involved in the regulation of protocatechuate degradation (Romero-Steiner *et al*., 1994). CatR and PcaR in *Rhodococcus opacus* are probably involved in the regulation of catechol and protocatechuate catabolism (Eulberg *et al*., 1998; Berens & Hillen, 2003). PobR in *Acinetobacter* sp. strain ADP1 is the key regulator of *p*-hydroxybenzoate hydroxylase (Reyes-Ramírez *et al*., 2002), and MhpR controls the expression of enzymes involved in the degradation of aromatic hydroxyalkanoates in *E. coli* (Torres *et al*., 2003). Three of the proteins (PcaU, PobR and PcaR) share a common physiological function – growth with *p*-hydroxybenzoate through the β-ketoadipate pathway – although they respond to different effectors (Ornston, 1966b; Harwood & Parales, 1996; McFall *et al*., 1998; Parke *et al*., 2000). Thus, it appears that a critical stage in the evolution of control for the β-ketoadipate pathway was the selection of an ancestor of the PobR group as a transcriptional activator. Subsequent evolution probably resulted in modified transcriptional regulators that respond to different effectors. The key properties of these three regulators and those of MhpR are summarized below.

The PobR protein of *Acinetobacter*

The catabolism of *p*-hydroxybenzoate is initiated in *Acinetobacter* sp. ADP1 by *p*-hydroxybenzoate 3-hydroxylase (PobA) to yield protocatechuate. The positive regulator of the *pobA* gene is the PobR protein encoded by the *pobR* gene, which is transcribed divergently from *pobA*. A transcriptional fusion of *pobA* to *′lacZ* showed that *pobA* expression in *Acinetobacter* was enhanced up to 400-fold in response to low concentrations of *p*-hydroxybenzoate, and that *pobA* expression was prevented in a *pobR*-deficient background. On the other hand, experiments with a *pobR::lacZ* fusion showed that PobR repressed its own synthesis three- to four-fold (DiMarco *et al*., 1993; DiMarco & Ornston, 1994; Elsemore & Ornston, 1994, 1995).

The start codons of *pobA* and *pobR* in the intergenic region are separated by 134 bp. Primer extension analysis identified the *pobA* transcript start point 22 bp upstream from the structural gene, whereas that of the *pobR* transcript began 69 bp upstream from the regulatory gene. This arrangement requires superimposition of the − 10 bp and − 35 bp RNA polymerase binding sites for the respective genes (Fig. 6). Electrophoretic gel mobility shift assays (EMSA) and qualitative surface plasmon resonance revealed that, as expected, PobR binds specifically to DNA in the *pobA*–*pobR* intergenic region. However, this binding was not influenced by *p*-hydroxybenzoate: the addition of up to 1 mM *p*-hydroxybenzoate, the co-effector of *pobA* expression, did not affect the association or dissociation characteristics of wild-type PobR to DNA (DiMarco & Ornston, 1994). DNase I footprinting revealed that the *Acinetobacter* PobR binds to a 35-bp motif (Clubb *et al*., 1994) which extended from about 10 bp to about 45 bp downstream from the transcription start site of the *pobR* transcript (Fig. 6). This protected region contains an 8-bp inverted repeat that

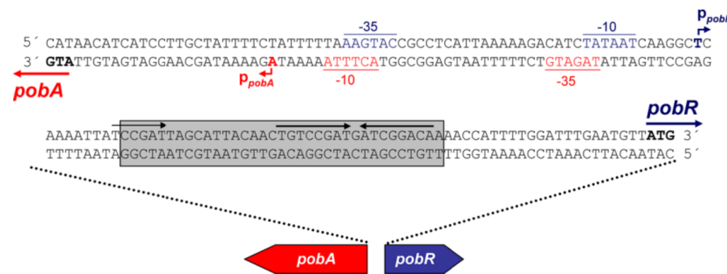


Fig. 6. The *pobApobR* intergenic region. The transcription start points are indicated by small curved arrows. The proposed translation start points are shown by long arrows. The proposed − 10/ − 35 RNA polymerase binding sites are indicated and the PobR binding site is shown in a gray box. Within this gray box the black arrows show an inverted repeat.

is probably involved in sequence-specific recognition. PobR may act as a repressor of its own transcription because it binds downstream from the *pobR* RNA polymerase binding site and most likely prevents transcription progression. In contrast, it is likely the PobR binding site is close enough to the *pobA* promoter for the α subunit of RNA polymerase to contact PobR.

Protocatechuate formed from *p*-hydroxybenzoate is cleaved by protocatechuate 3,4-dioxygenase to β-carboxy *cis,cis* muconic acid, which is further metabolized by PcaB. Ornston and colleagues designed a clever strategy to search for mutants in *pobR* based on the fact that a *pcaB* mutant of *Acinetobacter* sp. is unable to grow with succinate in the presence of *p*-hydroxybenzoate, due to the metabolic accumulation of toxic β-carboxy *cis,cis* muconate. A double mutant *pcaB pobR*, among other mutants, can grow with succinate as a carbon source on plates with *p*-hydroxybenzoate because the toxic compound is not accumulated (Hartnett & Averhoff, 1990; Clubb *et al.*, 1994; Elsemore & Ornston, 1994; Gerischer & Ornston, 1995). Kok *et al.* (1997) generated *pobR* mutants by using random PCR mutagenesis and the elegant selection procedure summarized above. Some of the mutations were clustered in and around the HTH DNA binding (residues 43–71) motif. Four mutations in the PobR recognition helix (R56 → S, R60 → Q, K64 → M, K67 → I) in the HTH motif yielded mutant proteins that completely failed to bind the operator. Other mutations adjacent to the DNA binding motif (W80 → R) showed altered operator association. In the predicted structure of PobR, tryptophan at position 80 is presumed to be located at S1 (see Fig. 1) and maintains numerous hydrophobic contacts with residues at H2 and H3; consequently, its substitution by a charged residue disturbs the structural arrangement of H2 and H3.

As mentioned above, PobR activates transcription of *pobA*, and the action of PobA on *p*-hydroxybenzoate yields protocatechuate, which in turn activates PcaU to mediate transcription from the *pca* genes. The amino acid sequences of PobR and PcaU are 54% identical and recognize very similar operators (see Fig. 7); however, each protein controls its specific set of genes. Kok *et al.* (1998) generated mutants of PobR to mimic PcaU, and found a single mutant in which Thr57 in the recognition helix was substituted by Ala. The mutant PobR (T57A) depended on *p*-hydroxybenzoate for induction of the *pca* operon.

Kok *et al.* (1998) subjected the allele that encodes PobR(T57A) to a round of PCR mutagenesis and searched for mutants that were able to activate transcription independently of *p*-hydroxybenzoate. Ten mutants which maintained the original T57 → A mutation had acquired a mutation causing an amino acid substitution in one of four residues clustered near the middle of the PobR primary sequence. This suggested that these residues formed part of a specific functional domain. In four of the six different mutations, an acidic residue (E124 or E126) was replaced by an uncharged residue (G or V). Other mutations were S118 → T, L135 → S and V120 → E. At least one of these mutants (V120 → E) was shown to bind the PobR operator *in vitro*, but the mutation prevented the response to *p*-hydroxybenzoate, thereby locking the protein in an apparently inactive conformation. In the predicted 3-D structure of PobR generated using the TM-IclR protein as a template, amino acids Ser118, Leu135 and Val120 were found on neighboring β-strands, which have been proposed to form part of the effector binding site of TM-IclR. These amino acids thus appear to be directly involved in effector binding. Furthermore, Asp124 and Asp126 are found on a short turn connecting the latter two strands. Both side chains point away from the effector binding site and appear to interact with other parts of the structure. Thus, all the mutations identified were either inside or in the vicinity of the proposed effector binding site.

DNA binding may be influenced by distal regions in the primary sequence with respect to the HTH binding domain. In this regard it should be noted that the C-terminal end of

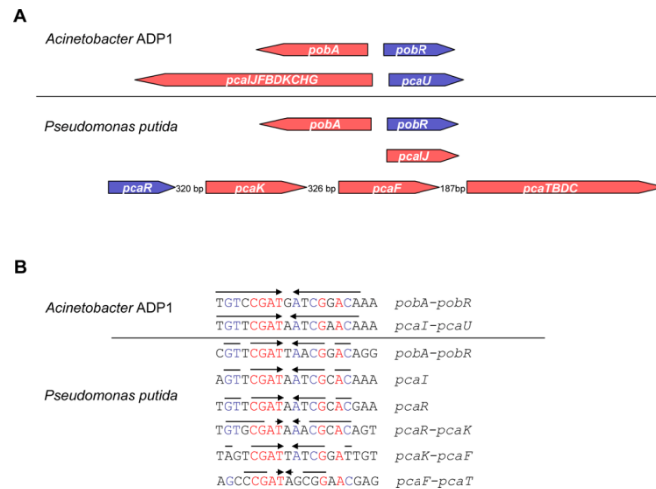


Fig. 7. Operator sequence of PobR, PcaU and PcaR. The organization of catabolic operons in *Acinetobacter* and *Pseudomonas putida* (top) and the corresponding operator sequences (bottom). The dark arrows represent the genes encoding the regulators and the clear arrows the catabolic genes. The regulator target sites in each promoter are indicated by black arrows and the corresponding operators are aligned.

the TM-IclR-TM forms a long helix which points toward the N-terminal DNA binding domain of the other monomer in the dimer (Fig. 1). In fact, there is a hydrophobic interaction between glycine 245 and tyrosine 62 in the DNA binding domain of TM-IclR. When this information is translated to PobR and considered in the light of available mutants, it is worth noting that a nonsense mutant lacking the last 15 amino acids of PobR resulted in a null mutant, and also that a PobR mutant in residue 241 (His241 → Lys) was unable to bind DNA (Kok *et al.*, 1998).

The PcaU protein of *Acinetobacter*

As mentioned above, protocatechuate generated from *p*-hydroxybenzoate is catabolized through the *pca* pathway. The promoter region of the *pcaI* gene was fused to *'lacZ*, and β-galactosidase activity was determined in a wild-type and in a Δ*pcaU* derivative background. The minimal protocatechuate concentration in the medium able to activate transcription at the *pcaI* promoter was 10 µM, whereas in the *pcaU*-deficient background, induction did not occur. In the wild type, the addition of 2 mM *p*-hydroxybenzoate also resulted in an approximately 100-fold greater induction. Thus, the process of induction at the *pcaI* promoter is completely dependent on a functional PcaU protein, and on protocatechuate supplied directly or generated from *p*-hydroxybenzoate *in vivo* (Trautwein & Gerischer, 2001; Popp *et al.*, 2002).

Divergently oriented with respect to the *pcaI* gene is the *pcaU* gene, which encodes the regulator. A *pcaU*::*lacZ* fusion was created to examine the influence of *pcaU* on its own expression. PcaU appears to decrease to one third *pcaU* expression, and this repression was relieved by protocatechuate (Gerischer *et al.*, 1998). Trautwein & Gerischer (2001) proposed that PcaU exerted a repressor function not only on its own promoter but also on the structural gene promoter in the absence of effectors, whereas it functioned as an activator of transcription only on the structural gene promoter in the presence of elevated concentrations of the effector. The PcaU protein has been shown to be a homodimer in solution. PcaU binds to the intergenic *pcaU*–*pcaI* region both in the absence and in the presence of the effector protocatechuate (100 µM). The regulator covers a 45-bp sequence which includes three perfect 10-bp repeats, two in palindromic order, and a direct repeat separated by 10 bp (Fig. 8) (Gerischer *et al.*, 1998; Guo & Houghton, 1999). The affinity of PcaU for the intergenic binding site is very high (dissociation constant [K_D], 0.16 nM). Removal of any of the two external repeats in this intergenic binding site still allowed the specific binding of PcaU, but the affinity was significantly reduced (K_D for the direct repeat was 16 nM, and for the palindrome 8 nM), suggesting an important role for all three sequence repeats in gene expression (Popp *et al.*, 2002). The described binding of PcaU to three repeats as a palindrome and as a direct repeat is unusual for regulator binding sites.

The PcaR protein of *Pseudomonas putida*

Degradation of protocatechuate in *P. putida* is accomplished by the products of the *pca* genes (*pcaHG*, *pcaTBDC*, *pcaIJ*, *pcaK* and *pcaF*) (Hughes *et al.*, 1988; Parales & Harwood, 1992; Frazee *et al.*, 1993; Harwood *et al.*, 1994; Nichols & Harwood, 1995). In *P. putida*, these genes (with the

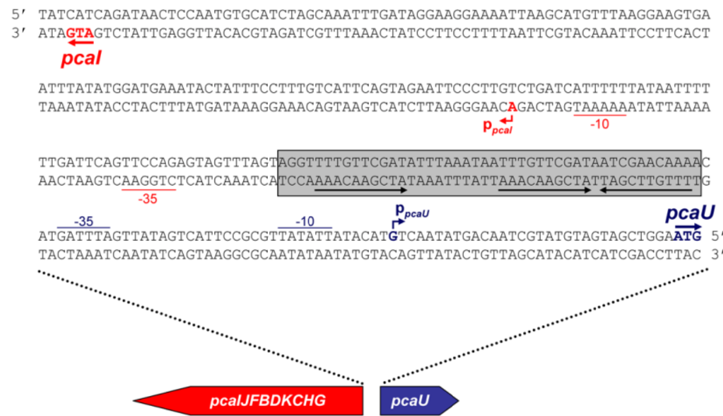


Fig. 8. The PcaU target sites at the *pcaI/pcaU* intergenic region. The gene organization of *pcaIJFBDKCH* and *pcaU* is shown in the bottom. The intergenic sequence between *pcaI* and the ATG of *pcaU* is shown. The +1, − 10/− 35 sequences of both promoters are indicated. The PcaU box is shown in gray. Within this box the PcaU motif recognized by PcaU is indicated with black arrows.

exception of *pcaHG*) are activated by the regulatory protein PcaR (Fig. 7), in association with the pathway intermediate β-ketoadipate (Ornston, 1966a; Parke & Ornston, 1976; Hughes *et al.*, 1988; Williams *et al.*, 1992; Frazee *et al.*, 1993; Parales & Harwood, 1993; Harwood *et al.*, 1994; Romero-Steiner *et al.*, 1994; Nichols & Harwood, 1995; Harwood & Parales, 1996; Orth *et al.*, 2000; Bertani *et al.*, 2001).

Pseudomonas putida cells carrying a P_{pcaF}::′*lacZ* fusion showed β-galactosidase activity, which was about 10-fold higher when the cells were grown on benzoate than when cells were grown on glucose or succinate. An about five-fold higher level of β-galactosidase activity was found in cells grown on glucose with either β-ketoadipate or its nonmetabolizable analogue, adipate. These levels of induction corresponded to those observed when activity of the pathway enzyme β-ketoadipyl coenzyme A thiolase was assayed in *P. putida* cell extracts (Harwood *et al.*, 1994). Under these conditions, β-ketoadipate, an intermediate of benzoate and *p*-hydroxybenzoate catabolism, was the physiological effector of *pcaF* expression. It was also demonstrated that β-ketoadipate was also an effector for the *pcaIJ* and *pcaK* genes (Weickert & Adhya, 1992; Harwood *et al.*, 1994).

Experiments with a fusion of the *pcaR* promoter to ′*lacZ* demonstrated that PcaR was negatively regulated (up to four-fold) by its own gene product (Hughes *et al.*, 1988). Autoregulation of PcaR was shown to be independent of any response to the co-effector molecule, β-ketoadipate, as well as to the presence of any gratuitous effector such as adipate (Parke & Ornston, 1976; Hughes *et al.*, 1988). Primer extension analyses demonstrated that *pcaR* is expressed constitutively and that the transcript is initiated 27 bp upstream from the proposed first ATG for the *pcaR* structural gene (Romero-Steiner *et al.*, 1994). Transcription was driven from a σ^{70} (*rpoD*) promoter (Domínguez & Marqués, 2004).

Purified PcaR protein was shown to form a homodimer in solution and to bind specifically to its own promoter, as well as to the promoter regions of *pcaI* and *pcaF*. Although it is likely that PcaR interacts with β-ketoadipate to activate transcription of the various target genes, the addition of β-ketoadipate to the mobility shift assay mixture did not alter the migration of the PcaR–*pcaR* promoter complex. The apparent equilibrium dissociation constants (K_D) of PcaR with each of these promoter fragments were determined by mobility shift assays and found to be 0.087 and 0.13 nM for the *pcaR* and *pcaI* promoters, respectively (Guo & Houghton, 1999). Deletion analysis of the *pcaIJ* promoter region suggested that a DNA sequence proximal to the − 10 region was sufficient for specific induction of these genes, implying that the binding site for transcriptional activation by PcaR lies very close to the transcriptional start site of the *pcaIJ* operon (Parales & Harwood, 1993). The corresponding region of *pcaIJ* revealed a similar 15-bp target (5′-GTTCGATAATCGCAC-3′) that was centered in the − 10 region. The *pcaR* binding site in *pcaK* was found to be 5′-GTTCGATAAACGGACAAT-3′ (see Fig. 7). The positioning of an operator region so close to the transcriptional start site is unusual for an inducible system, and suggests a possible mechanism for induction of the *pca* genes that is similar to that for the induction of the mercury resistance operon (Parales & Harwood, 1993). In this case the MerR protein and RNA polymerase bind to opposite faces of the DNA strand to effect open-complex formation (O'Halloran *et al.*, 1989; Ansari *et al.*, 1992). Footprint analyses demonstrated that the binding of PcaR to its own promoter occurs

within a footprint that extends from the −20 to the +4 position. This footprint can be explained assuming that a dimer binds to the *pcaR* promoter. Similar analyses with the *pcaIJ* promoter showed that the footprint of PcaR extended over a significantly larger portion of DNA, from the −38 to the +1 positions of the promoter sequence. It therefore seems that PcaR interacts with the inducible *pcaIJ* promoter as two dimers in tandem, so that a dual footprint encompassing both the −35 and the −10 regions of the promoter sequence is created.

The MhpR regulator

The 3HPP (3-(3-hydroxyphenyl)propionic acid) and 3-hydroxycinnamic acid degradation pathway of *E. coli* is encoded by the *mhp* cluster (Burlingame & Chapman, 1983; Burlingame *et al.*, 1986; Díaz & Prieto, 2000). This pathway is initiated by the MhpA monooxygenase, which transforms 3HPP into 2,3-dihydroxyphenylpropionate. This compound is then converted to succinate, pyruvate and acetyl-CoA through the action of a *meta*-cleavage hydrolytic pathway, which involves dioxygenase (MhpB), hydrolase (MhpC), hydratase (MhpD), aldolase (MhpE) and acetaldehyde dehydrogenase (MhpF) activity (Burlingame & Chapman, 1983; Díaz *et al.*, 2001). The *mhp* cluster is arranged as follows:

1. six catabolic genes encoding the initial monooxygenase (*mhpA*), the extradiol dioxygenase (*mhpB*), and the hydrolytic *meta*-cleavage enzymes (*mhpCDFE*);
2. a gene (*mhpT*) that encodes a potential transporter;
3. a regulatory gene (*mhpR*) that is adjacent to the catabolic genes but transcribed in the opposite direction to that of the *mhpABCDFE* catabolic genes.

The intergenic *mhpR–mhpA* region contains the *Pr* and *Pa* promoters which drive the expression of the *mhp* regulatory and catabolic genes, respectively.

The MhpR protein controls the expression of a *meta*-cleavage pathway for the degradation of aromatic compounds, in contrast with PobR, PcaU, CatR and PcaR, which control *ortho*-cleavage pathways (Ferrández *et al.*, 1997). The influence of the MhpR protein on the expression of the reporter *Pa:′lacZ* and *Pr:′lacZ* fusions was analyzed in cells growing in minimal medium with glycerol in the presence and in the absence of 3HPP. The expression of the *Pa:′lacZ* fusion required the presence of the *mhpR* gene and the 3HPP effector, whereas the *Pr:′lacZ* fusion was expressed constitutively both in the presence and in the absence of *Mhp*R and 3HPP. These findings indicate that MhpR is a 3HPP-dependent activator of the *Pa* promoter, and that the expression of the *mhpR* gene is constitutive and MhpR-independent. In this sense, *Mhp*R shows a distinct regulatory feature when compared with other IclR-type regulators of aromatic catabolic pathways, e.g. PcaR from *P. putida* and PobR and PcaU from *Acinetobacter* sp. ADP1. These regulators act as transcriptional activators of the cognate catabolic genes, but behave as repressors of their own expression (DiMarco & Ornston, 1994; Guo & Houghton, 1999; Trautwein & Gerischer, 2001).

Gel retardation assays with a DNA fragment bearing the Pa and Pr promoters, at subsaturating concentrations of MhpR in the presence of different concentrations of the 3HPP effector molecule, revealed that the higher the concentrations of 3HPP, the better the retardation of the DNA probe. These results indicate that this effector molecule, although dispensable, facilitates DNA binding. Furthermore, it was shown that the MhpR activator protected a region centered at position −58 with respect to the transcription start site in the *Pa* promoter, with a 17-bp imperfect palindromic motif. Torres *et al.* (2003) showed that expression of the Pa promoter was also influenced by the cAMP receptor protein (CRP), which allows expression of the *mhp* catabolic genes when the preferred carbon source (glucose) is not available and 3HPP is present in the medium. Gel retardation assays revealed that the specific activator, MhpR, is essential for binding of the second activator, CRP, to the *Pa* promoter. Such peculiar synergistic transcription activation has not yet been observed in other aromatic catabolic pathways, making the MhpR activator the first member of the IclR family of transcriptional regulators known to recruit CRP to the target promoter.

The multidrug efflux pump regulator: the TtgV repressor

Pseudomonas putida DOT-T1E is able to thrive in liquid medium with toluene (Ramos *et al.*, 1995). This strain exhibits an innate high resistance to solvents, which increases when bacteria are pre-exposed to sublethal concentrations of toluene (Ramos *et al.*, 1998; Duque *et al.*, 2001; Rojas *et al.*, 2001). Such resistance is achieved through the additive efflux of the solvent by three RND pumps called TtgABC, TtgDEF and TtgGHI (Mosqueda & Ramos, 2000; Rojas *et al.*, 2001). The expression of TtgABC is constitutive and its level does not vary significantly in the presence of toluene (Duque *et al.*, 2001). The *ttgDEF* operon is not expressed at all in the absence of solvents, and its expression is higher in the presence of aromatic hydrocarbons (Mosqueda & Ramos, 2000). The *ttgGHI* operon is expressed at a basal level in the absence of solvents, and its expression increases about four-fold in response to toluene. The TtgGHI efflux pump plays a pivotal role in the innate and induced tolerance to solvents in this strain, a role which was demonstrated on a *ttgH* knockout mutant that was extremely sensitive to solvent shocks (Rojas *et al.*, 2001).

The pattern of expression of the *ttgGHI* genes was determined by primer extension analysis and β-galactosidase assays with the *ttgGHI* promoter region fused to *lacZ*. Expression takes place from a single σ^{70}-like promoter, P_{ttgG}, and increases four-fold in the presence of toluene (Rojas *et al.*, 2001; Guazzaroni *et al.*, 2004). Rojas *et al.* (2003) reported the identification of two ORFs, *ttgV* and *ttgW*, that form an operon and that are transcribed divergently from the *ttgGHI* operon (Fig. 9). *ttgW* is a pseudogene and plays no role in the transcription of P_{ttgG}. However, a mutant lacking TtgV protein showed increased resistance towards toluene shocks under noninduced conditions when compared to the wild type. This indicated that TtgV behaved as a repressor of the *ttgGHI* operon. In fact, analysis of the expression of the *ttgGHI* operon in a *ttgV*-deficient background showed that the level of transcription also increased four-fold in the absence of toluene. These data clearly indicate that TtgV is a repressor that prevents expression of the *ttgGHI* operon.

The *ttgVW* operon was also transcribed from a single promoter regardless of the growth conditions, but its level of expression in the presence of toluene was three to four times higher than in the absence of the aromatic hydrocarbon. The pattern of expression of *ttgV* is thus similar to that of *ttgGHI*. Although this contrasts with the situation described above for the autoregulation of other IclR family promoter members, the explanation is straightforward. Sequence analysis of the promoter region showed that the −10 and −35 boxes of the P_{ttgG} overlap with the −35 and −10 boxes of the *ttgV* promoter (Fig. 9). The overlap of the genes for the promoter of the regulator and the structural efflux pump explains why they are regulated in the same way. *In vitro* gel mobility shift assays confirmed the specific binding of the regulatory protein TtgV to the intergenic region between the *ttgG* and *ttgV*, and DNAse I-footprint assays established that TtgV binds to a 40-bp region that covers the −10/−35 regions of the *ttgG* promoter and the divergently oriented *ttgV* promoter. Sequence analysis did not reveal the presence of a clear single invert repeat or direct repeat.

To gain insight into the mechanism of regulation of *ttgGHI* transcription by TtgV, *in vitro* transcription experiments were done using the purified protein and a construction of the supercoiled DNA plasmid pTE103 ligated to the *ttgV–ttgG* intergenic region. When the TtgV protein and the plasmid were incubated before RNA-polymerase was added, *ttgGHI* transcription was completely repressed. However, when TtgV was added after the formation of the RNA-polymerase–*ttgGHI* promoter open complex, repression was negligible (Guazzaroni *et al.*, 2005). These findings support the hypothesis that TtgV binding to the intergenic region blocks the entry of RNA-polymerase to transcribe these genes. This indicates that the mechanism of repression of TtgV at its target promoter is similar to the type I repression described for IclR (Fig. 5, left).

The first direct evidence of TtgV binding to ligand was obtained when EMSA assays were done with 1-hexanol (Guazzaroni *et al.*, 2004). Electrophoretic gel mobility shift assays revealed that in the presence of increasing concentrations of 1-hexanol, TtgV dissociated from its target operator. Later, isothermal titration calorimetry (ITC) and equilibrium dialysis assays suggested that TtgV can accommodate many different molecules, including mononuclear and binuclear aromatic compounds, the latter being recognized with very high affinity (1–10 μM). These assays also demonstrated that for most effectors, one molecule is bound per TtgV dimer (Guazzaroni *et al.*, 2005).

IclR proteins involved in phytopathogenicity by *Erwinia* sp.

Members of the genus *Erwinia* such as *Erwinia chrysanthemi* cause soft-rot diseases in a wide variety of host plants (Perombelon & Kelman, 1980). The phytopathogenicity of several *Erwinia* species correlates with their ability to produce and secrete plant cell wall-degrading enzymes such

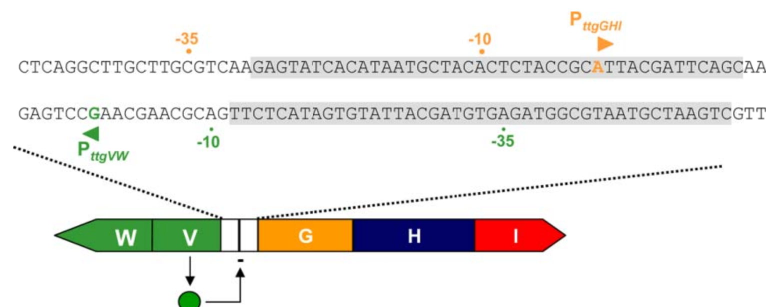


Fig. 9. *ttgG* and *ttgV* promoter fully overlap. The organization of *ttgVW* and *ttgGHI* operons is shown, as well as the intergenic region (double strand) in which the +1 of each operon is indicated by an arrowhead. The −10 and −35 positions of the promoters are indicated to highlight overlap of the promoters. Boxed in gray is the DNA sequence to which TtgV binds, as revealed by DNAse I footprinting.

as pectinases, cellulases (Cel) and proteases (Prt) (Collmer & Keen, 1986; Barras *et al*., 1994). The crucial role of these enzymes, particularly the pectate lyases (Pel), in the virulence of *Erwinia* sp. has been confirmed by the isolation of mutants that exhibit reduced virulence that are defective in the production or secretion of these enzymes (Collmer & Keen, 1986; Boccara *et al*., 1988; Hinton *et al*., 1989; Pirhonen *et al*., 1991). Exoenzyme production by soft-rot *Erwinia* species occurs in response to several environmental conditions (Hugouvieux-Cotte-Pattat *et al*., 1996), such as the presence of pectin-degradative products or plant extracts, anaerobiosis, temperature, nitrogen starvation, osmolarity, catabolite repression, iron availability and growth phase.

The degradation of pectin compounds is initiated by extracellular pectinases, including two pectin methylesterases (encoded by *pemA* and *pemB* (Laurent *et al*., 1993; Shevchik *et al*., 1996)), five major isoenzymes of pectate lyase (encoded by *pelA*, *pelB*, *pelC*, *pelD* and *pelE*) and a set of secondary pectate lyases (Hugouvieux-Cotte-Pattat *et al*., 1996), all of which generate unsaturated digalacturonates. These latter compounds are transported into the bacterium, where they are catabolized by the products of the genes *ogl*, *kduL*, *kduD*, *kdgK* and *kdgA* (Condemine *et al*., 1986; Hugouvieux-Cotte-Pattat & Robert-Baudouy, 1987; Reverchon & Robert-Baudouy, 1987). The complex regulatory network governing virulence in the erwinias involves several totally distinct, but highly conserved, members of the IclR class of DNA binding proteins: RexZ, Pir and KdgR.

The KdgR protein

The degradation of pectin in *E. chrysanthemi* is also metabolically relevant as depolymerized pectin fragments are internalized and used as a carbon source for growth.

The transcription of genes involved in pectin metabolism is tightly regulated. The key player in this control mechanism is the specific negative regulator KdgR (Nasser *et al*., 1994), which controls at least 13 operons (Hugouvieux-Cotte-Pattat *et al*., 1996, 2001; Blot *et al*., 2002). The transcription of genes involved in pectin catabolism is induced by the presence of pectin, which is converted into catabolic products by a basal level of pectinase activity. *In vitro* experiments show that the binding of the catabolic product 2-keto-3-deoxygluconate (Kdg) to KdgR releases the transcriptional repressor from its multiple binding sites in the corresponding promoters (Nasser *et al*., 1991, 1994). The KdgR-mediated induction of gene expression is thus based on a derepression mechanism, and Kdg should be considered the effector. The effector spectrum of KdgR is very narrow and a well-defined molecular moiety was identified to be essential for activity (Nasser *et al*., 1991). This finding was substantiated by lack of activity of the Kdg precursor galacturonate *in vitro* (Nasser *et al*., 1992).

An apparent K_D of 0.4 mM has been determined for the binding of Kdg to KdgR complexed with various operators. This affinity was constant for all the operators tested, indicating that different induction ratios *in vivo* are independent of the Kdg–KdgR interaction (Nasser *et al*., 1994).

KdgR was found to bind *in vitro* very tightly to different operators, with K_D values varying between 0.1 and 10 nM (Nasser *et al*., 1994). However, there was no strict correlation between affinity and the degree of repression of the corresponding gene. This suggests that in addition to the specific KdgR regulation, other signals of a global or specific nature can fine-tune the action of KdgR (see also below). The key KdgR action involves recognition of a KdgR box, which is a DNA binding motif of 17 bp (Nasser *et al*., 1994) containing two half-sites. KdgR is a dimer in solution (Thomson *et al*., 1999), and one dimer was shown to bind to a KdgR box. Although the affinity is weaker, KdgR can also bind to individual half-site sequences (Nasser *et al*., 1994).

As indicated above, the modulation of pectin metabolism is influenced markedly by several factors such as temperature, anaerobiosis, nitrogen starvation, osmolarity, iron availability, growth phase or the presence of plant extract (Hugouvieux-Cotte-Pattat *et al*., 1996). The regulatory responses to these changes might be mediated by other proteins such as PecS (MarR family regulator), PecT (LysR family regulator), for which the triggering signal molecule has yet to be identified (Reverchon *et al*., 1994; Praillet *et al*., 1996; Surgey *et al*., 1996; Castillo & Reverchon, 1997), or CRP, the global regulator of sugar metabolism (Nasser *et al*., 1997; Reverchon *et al*., 1997). Regulation of the genes involved in pectin metabolism can involve the concerted action of several of the regulators mentioned above (Rouanet *et al*., 1999; Hugouvieux-Cotte-Pattat *et al*., 2002; Shevchik & Hugouvieux-Cotte-Pattat, 2003). For instance, regulation of the *pelDE* operon, which encodes two of the secreted pectate lyases, involves PecS, KdgR and CRP, all of which to bind to the *pelD* promoter (Rouanet *et al*., 1999). DNaseI footprinting assays demonstrated that PecS and KdgR binding sites are close together. Although neither co-operativity nor competition is observed *in vitro* for the binding of these proteins, the presence of bound KdgR stabilized the interaction of adjacent-bound PecS with DNA. The presence of both proteins at the promoter gave rise to more efficient repression *in vivo* as compared to the binding of either protein to the promoter. Furthermore, PecS partially inhibited binding of the CRP activator. The molecular basis for the mechanisms of action of the repressors needs to be studied in further detail.

KdgR is highly conserved among different bacterial species. This is exemplified by the fact that sequences with 88–93% identity to the *E. chrysanthemi* protein are found in

Erwinia carotovora, *E. coli*, *Shigella flexneri*, *Yersinia pestis* and *Salmonella typhimurium*.

The RexZ protein

Erwinia carotovora contains a homologue of the *E. chrysanthemi* KdgR and in addition has another regulator that belongs to the IclR family, called RexZ, which is also involved in virulence. The *E. carotovora rexZ* gene encodes a 262-amino acid protein, which unlike KdgR acts as an activator of exoenzyme production and therefore virulence. An *E. carotovora rexZ* mutant generated by reverse genetics failed to synthesize depolymerizing enzymes such as pectate lyases, cellulases and proteases. These results suggest that the *E. carotovora* RexZ protein acts as an activator of exoenzyme synthesis.

As described above, CRP and KdgR are major regulators of genes involved in pectinolysis in *E. chrysanthemi* (Nasser *et al.*, 1997; Reverchon *et al.*, 1997). Computer searches revealed the presence of potential binding sites for CRP and KdgR in the promoter region of the *rexZ* gene. In EMSA assays the addition of subsaturating amounts of KdgR and CRP to a solution containing the *rexZ* promoter fragment resulted in three protein–DNA complexes, two corresponding to the KdgR–DNA and CRP–DNA individual complexes and one corresponding to a CRP–KdgR–DNA complex. At a saturating concentration of CRP and KdgR proteins, only one ternary complex was observed.

The precise location of the CRP and KdgR binding sites in the regulatory region of the *rexZ* gene were mapped using DNaseI protection experiments. A single protected region with an average length of approximately 36 and 47 bp was obtained in the presence of CRP and KdgR at subsaturating and saturating concentrations, respectively. These protected regions encompassed the predicted CRP binding site and the KdgR box. The CRP binding site is centered at position − 41.5 with respect to the putative transcription start site of *rexZ*, which is typical for class II CRP-dependent promoters. This suggests that the CRP complex may act as a direct activator of the *rexZ* promoter. However, the KdgR-protected region covers the nucleotides between − 67 and − 110; this makes it unlikely that KdgR binding would interfere with the expression of the *rexZ* promoter, and suggests that the *rexZ* promoter is independent of KdgR repressor control. In the absence of a direct effect on *rexZ* expression, it is formally possible that, because the binding sites of KdgR and CRP are close together, KdgR may affect *rexZ* expression by interfering with the ability of CRP to bind DNA and thus activate gene expression.

The Pir protein

Nomura *et al.* (1998) isolated a plant-inducible regulatory (Pir) protein and cloned its structural gene (*pir*) from *E. chrysanthemi* EC16. Around 50% of the Pir amino acid sequence was either identical or highly similar to *E. chrysanthemi* KdgR and *E. carotovora* RexZ proteins. The experimental approach used was based on protein enrichment through binding of a crude cell extract of cells grown in the presence of polypectate-Na (NaPP) NaPP and glycerol with or without potato extract to a DNA fragment containing the *pel* promoter. The protein that accounted for binding of a fragment that contained the promoter region of *pelE* was called Pir (Nomura *et al.*, 1998, 1999). Its deduced molecular mass is around 30 kDa and gel filtration assays suggested that Pir is a dimer.

Mutation of *pir* resulted in the loss of pectate lyase hyperinduction in response to plant signals, and reduced bacterial virulence on plant tissues, but did not affect the regulation of other extracellular enzymes such as cellulases or proteases.

The promoter region of *pelE* (from − 150 to + 110) was fused to the *lux* genes and the construction was inserted into the chromosome of the wild-type bacteria and a *pir*-deficient mutant. In the wild type, expression of the *pelE:'lux* fusion was inducible in the presence of NaPP and glycerol, and was hyperinducible by the further addition of potato extract. In the mutant strain, hyperinduction in the presence of potato extract with NaPP and glycerol was not observed. Nomura *et al.* (1998) showed that the virulence of *pir*-mutant bacteria was markedly reduced in potato, celery and Chinese cabbage tissues when low concentrations of cells ($\approx 10^6$ cells mL^{-1}) were inoculated, demonstrating the role of this protein in pathogenesis.

Sporulation in *Streptomyces* involves IclR family members

Streptomycetes are aerobic gram-positive soil bacteria that grow as multinucleoidal, multicellular, branched filaments and undergo morphological and physiological differentiation in response to environmental factors (Kendrick & Ensign, 1983; Champness & Chater, 1994; Chater, 1998; Keijser *et al.*, 2002). The first sign of morphological differentiation is the formation of aerial sporogenic hyphae, which give a fuzzy, white appearance to the colonies. When these specialized hyphae have stopped growing, DNA segregates and sporulation septa form to generate the uninucleoidal compartments that become the spores. Sporulation septation occurs relatively synchronously throughout a single sporogenic hypha (Kwak & Kendrick, 1996; Schwedock *et al.*, 1997).

The characterization of nonsporulating mutants of *Streptomyces* has been critical to understand this development process. These mutants have been divided into bald (*bld*) strains, which appear not to form aerial hyphae, and white

(*whi*) strains, which form aerial hyphae but which are blocked at various stages of sporulation.

Chater & Horinouchi (2003) compared the developmental regulatory cascades in *S. coelicolor* and *S. griseus*, and described an important difference between them: *S. griseus* sporulates in submerged culture, and the signal for onset, although poorly understood, seems to be mediated by the γ-butyrolactone A-factor. *S. griseus* mutants unable to produce A-factor are defective in development and antibiotic production (Horinouchi, 2002; Chater & Horinouchi, 2003), whereas these processes seem barely affected in γ-butyrolactone-deficient mutants of *S. coelicolor* (Takano *et al.*, 2001).

SsgA plays a key role in the production of sporulation septa, as deduced from the finding that *ssgA* mutants of *S. coelicolor* and *S. griseus* are defective in sporulation even though they form apparently normal vegetative septa (Jiang & Kendrick, 2000; van Wezel *et al.*, 2000; Traag *et al.*, 2004). Traag *et al.* (2004) identified significant differences in the regulation of *ssgA* between *S. coelicolor* and *S. griseus*, and proposed that this is one of the determinants of the morphological and developmental divergence between the two microorganisms. The IclR type regulators SsfR and SsgR were shown to be associated with sporulation in *S. griseus* and *S. coelicolor*, respectively. These proteins are homologues and share 76% sequence identity.

The SsfR transcriptional regulator of *Streptomyces griseus*

Streptomyces griseus is a species that can be induced to sporulate in submerged cultures in response to phosphate or nutritional downshift (Kendrick & Ensign, 1983; Kroening & Kendrick, 1987). In the wild-type strain, sporogenic hyphae emerge after about 4 h of phosphate starvation, and after about 10 h, centripetal growth of the sporulation septa is visible. The formation of sporulation septa is complete by 12 h (Kwak & Kendrick, 1996; Hao & Kendrick, 1998). Strain SKK2663 is a mutant in which the formation of the sporulation septum is impaired. This mutant carries a point mutation in *ssfR* that introduces an early stop codon. The SsfR protein is required for sporulation in rich medium (SpM) or minimal medium with glucose, but not in minimal medium with mannitol. In experiments with a Δ*ssfR* mutant, Jiang & Kendrick (2000) showed that the mutation was complemented by the *ssfR* gene provided *in trans*. Although the *ssfR* gene is located upstream of *ssgA*, Jiang & Kendrick (2000) demonstrated that transcription of *ssgA* is not under the control of the *ssfR* gene product. An interesting feature of SsfR as well as SsgR is that both proteins are predicted to have a transmembrane domain in the C-terminal part of the protein. Both predictions favor an inside-out model, indicating that the large part of the protein including the HTH domain is inside the cell.

The SsgR protein of *Streptomyces coelicolor*

In *S. coelicolor*, transcription of *ssgA* and the upstream-located *ssgR* is developmentally regulated and activated around the time of the onset of sporulation. The expression of *ssgR* is very low, but its expression is upregulated at the onset of sporulation.

The *ssgR* gene encodes a 241-amino acid protein. An in-frame deletion mutant of *S. coelicolor ssgR* was generated to create mutant GSR1. The mutant had a phenotype similar to that of the *ssgA* mutant strain and formed aerial hyphae, but failed to produce spores on rich media such as R2YE or minimal medium with glucose as the sole carbon source. In this mutant the *ssgA* gene was not transcribed, and this was probably the sole cause of its sporulation deficiency, as wild-type levels of sporulation could be restored by the SsgR-independent expression of *ssgA* from a constitutive promoter.

Failure to detect *ssgA* transcripts in the *ssgR* mutant indicated that the growth phase-dependent induction of *ssgA* transcription is dependent on SsgR. Transcription of *ssgA* was induced after *ssgR* and at a time when sporulation had already commenced. The *ssgA* gene was found to be transcribed from two promoters (*ssgA* P1 and *ssgA* P2) separated by approximately 15 nucleotides, suggesting that they overlap (van Wezel *et al.*, 2000).

Binding of a truncated version of SsgR to the *ssgA* promoter region showed that *ssgA* transcription is directly activated by SsgR; the dependence of *ssgA* on SsgR in *S. coelicolor* is in clear contrast to the situation in *S. griseus*, where *ssgA* transcription is activated by A-factor, and where its control by the SsgR orthologue, SsfR, is far less clear. The *ssgR* mutant of *S. coelicolor* was not complemented by the *S. griseus ssfR*, which suggests functional differences between the genes.

Degradation of quorum-sensing signals

Quorum sensing is a bacterial community behavior to sense the change in bacterial population density and coordinate different biological functions. Bacterial cells in a prequorum population produce a basal level of quorum-sensing signal (quormone), a good example of which is the *N*-acyl-homoserine lactones (AHL) identified in many gram-negative *Proteobacteria* (Eberhard *et al.*, 1981; Cao & Meighen, 1989; Jones *et al.*, 1993; Zhang *et al.*, 1993; Pearson *et al.*, 1994). The quormone AHL accumulates as bacterial cells proliferate. Once a threshold concentration is reached, AHL interacts with its cognate transcription factor, usually a member of the LuxR-type proteins, to form a complex that triggers the expression of target genes (Zhu & Winans, 1999; Qin *et al.*, 2000; Welch *et al.*, 2000). Evidence is now accumulating that bacterial cells can switch off the quorum-sensing machinery by degradation of quormone signals in response to a change in growth (Zhang HB *et al.*, 2002).

Quormone *N*-3-oxo-octanoyl homoserine lactone (3OC8HSL), originally known as conjugation factor, regulates Ti-plasmid conjugal transfer in *Agrobacterium tumefaciens* (Zhang & Kerr, 1991; Piper *et al.*, 1993; Zhang *et al.*, 1993). In combination with the cognate transcription factor TraR, 3OC8HSL positively regulates the expression of genes required for Ti-plasmid transfer (*tra* genes) (Piper *et al.*, 1993; Hwang *et al.*, 1994; Zhu & Winans, 1999; Qin *et al.*, 2000). However, the efficiency of Ti-plasmid conjugal transfer decreases rapidly after bacterial cells enter the stationary phase (Hwang *et al.*, 1994). Coincidentally, 3OC8HSL concentration also declines rapidly at the stationary phase (Zhang HB *et al.*, 2002). Molecular and biochemical analyses demonstrated that the AHL-lactonase encoded by *attM* is expressed and degrades 3OC8HSL in the stationary phase. The *attM* gene is negatively regulated by the IclR-type transcription factor AttJ. When *attJ* was knocked out in *A. tumefaciens* strain A6 by transposon mutagenesis, AHL-lactonase became constitutively expressed, which resulted in phenotypes that were completely deficient in 3OC8HSL accumulation and Ti-plasmid conjugal transfer (Zhang HB *et al.*, 2002). *In vitro* analysis showed that AttJ specifically binds to the P_{attKLM} promoter that directs AHL-lactonase expression (Zhang HB *et al.*, 2002).

To identify the signalling pathway and regulatory mechanisms that mediate quormone degradation, *lacZ* was transcriptionally fused to the *attKLM* promoter. Transposon mutagenesis of strain A6(*attKLM*::*lacZ*) led to identification of the *relA* gene, which is involved in the synthesis of the stress alarmone (p)ppGpp. Tn*5* knock-out of *relA* abolished the stationary phase-dependent expression of *attM*. It was concluded that the *A. tumefaciens* quormone degradation system is coupled to and regulated by the generic (p)ppGpp stress response machinery (Zhang *et al.*, 2004).

Future perspectives and concluding remarks

In prokaryotic microorganisms, transcriptional regulators play an essential role in the response to environmental changes. Based on the structural and functional characteristics of the regulators they can be grouped into families. Members of a given family are preferentially repressors or activators (Méndez *et al.*, 1980; Levy, 1988; Brennan & Matthews, 1989; Jourdan & Stauffer, 1998; Orth *et al.*, 2000). Members of the IclR family are either activators or repressors, and some of them have dual transcriptional activity, because they regulate their own synthesis and activate the transcription of other(s) gene(s).

The signature that best identified members of this family does not involve the HTH DNA binding domain but about 80 amino acids that cover part of the central domain and the C-terminal end of the primary structure of the proteins. Conservation of such a large segment may imply a large degree of similarity in the 3-D structure between members of this family. This is probably the case as demonstrated from the five available 3-D structures for members of the IclR protein family. All of these structures belong to IclR-like proteins that are members of only one of the five orthologous groups of the IclR family. Therefore, additional protein structures for members of the family are needed to establish generalities. However, a number of mutants among members of the IclR family suggest that the conserved domain may be involved in signal sensing and signal transmission to the DNA binding domain. Although the HTH fails to define the family, in contrast with the TetR and AraC/XylS families (Gallegos *et al.*, 1997; Ramos *et al.*, 2005), a certain degree of conservation within this region was found, in particular for a few noncharged residues. These residues seem to play a role in the maintenance of the HTH structure and in establishing nonspecific contacts with DNA. This conservation at the protein level of the HTH domain contrasts with the relative diversity of promoters recognized by members of this family. Some promoters are relatively simple in structure (e.g. a classical inverted repeat), whereas others are more complex and include inverted repeats plus a half repeat or a series of potential direct repeats. The lack of information about the 3-D structures of members of the family complexed with their target DNA prevents us from proposing a general model for interactions with target promoters.

With regard to the mechanism of repression, the IclR members can exert repression by occluding the RNA polymerase binding site at the corresponding cognate promoter(s) or by actively inhibiting RNA polymerase progression. An intriguing issue is that there is no preferred distance for the IclR members that act as a positive transcriptional regulator with respect to the binding site of RNA polymerase. We therefore presume that a number of mechanistic approaches will be needed to elucidate the intimate mechanisms of action of these regulators in the near future.

Members of the IclR family regulate a variety of metabolic processes in a diverse range of microorganisms. It will be interesting to see whether these regulators are associated with a specific physiological process in the prokaryotic world. For example, is IclR a universal regulator of the glyoxylate shunt in aerobes? Are IclR family members involved in the control of catechol pathways in soil microorganisms (Eulberg & Schlömann, 1998)? The answer to these and other related questions could assist bioinformaticians in the functional annotation of new genomes.

Acknowledgements

Work in the authors' laboratory was supported by grants QLRT-2001-00435 and QLK3-CT-2002-01923 from the

European Commission, grant BIO2003-00515 from the Spanish Comisión Interministerial de Ciencia y Tecnología, and PETRI project 2001261-P from the Spanish Ministerio de Ciencia y Tecnología. We thank Carmen Lorente and María del Mar Fandila for secretarial assistance and Karen Shashok for checking the English.

References

Anantharaman V, Koonin EV & Aravind L (2001) Regulatory potential, phyletic distribution and evolution of ancient, intracellular small-molecule-binding domains. *J Mol Biol* **307**: 1271–1292.

Ansari A, Chael ML & O'Halloran TV (1992) Allosteric underwinding of DNA is a critical step in positive control of transcription by Hg-MerR. *Nature* **355**: 87–89.

Arai H, Yamamoto T, Ohishi T, Shimizu T, Nakata T & Wudo T (1999) Genetic organization and characterization of the 3-(3-hydroxyphenyl) propionic and degradation pathway of *Comamonas testosteroni* TA441. *Microbiology* **145**: 2813–2820.

Argos P, Hanei M, Wilson JM & Kelley WN (1983) A possible nucleotide-binding domain in the tertiary fold of phosphoribosyltransferases. *J Biol Chem* **258**: 6450–6457.

Arias-Barrau E, Olivera ER, Luengo JM, Fernández C, Galán B, García JL, Díaz E & Miñambres B (2004) The homogentisate pathway: a central catabolic pathway involved in the degradation of L-phenylalanine, L-tyrosine, 3-hydroxyphenylacetate in *Pseudomonas putida*. *J Bacteriol* **186**: 5062–5077.

Bairoch A & Rudd KE (1995) PROSITE. Accession no. PDOC00874.

Barnes MR, Duetz WA & Williams PA (1997) A 3-(3-hydroxyphenyl) propionic acid catabolic pathway in *Rhodococcus globerulus* PWD1: cloning and characterization of the hpp operon. *J Bacteriol* **179**: 6145–6153.

Barras F, Vangijsegem F & Chatterjee AK (1994) Extracellular enzymes and pathogenesis of soft-rot *Erwinia*. *Ann Rev Phytopathol* **32**: 201–234.

Bateman A, Birney E, Durbin R, Eddy SR, Howe KL & Sonnhammer ELL (2000) The Pfam protein families database. *Nucleic Acids Res* **28**: 263–266.

Beinlich KL, Chuanchuen R & Schweizer HP (2001) Contribution of multidrug efflux pumps to multiple antibiotic resistance in veterinary clinical isolates of *Pseudomonas aeruginosa*. *FEMS Microbiol Lett* **198**: 129–134.

Bell KS, Sebaihia M, Pritchard L, *et al.* (2004) Genome sequence of the enterobacterial phytopathogen *Erwinia carotovora* sp. atroseptica and characterization of virulence factors. *Proc Natl Acad Sci USA* **101**: 11105–11110.

Berens C & Hillen W (2003) Gene regulation by tetracyclines. Constraints of resistance regulation in bacteria shape TetR for application in eurkaryotes. *Eur J Biochem* **270**: 3109–3121.

Bertani I, Kojic M & Venturi V (2001) Regulation of the *p*-hydroxybenzoic acid hydroxylase gene (*pobA*) in plant-growth-promoting *Pseudomonas putida* WCS358. *Microbiology* **147**: 1611–1620.

Blot N, Berrier C, Hugouvieux-Cotte-Pattat N, Ghazi A & Condemine G (2002) The oligogalacturonate-specific porin KdgM of *Erwinia chrysanthemi* belongs to a new porin family. *J Biol Chem* **277**: 7936–7944.

Boccara M, Diolez A, Rouve M & Kotoujansky A (1988) The role of individual pectate lyases of *Erwinia chrysanthemi* strain 3937 in pathogenicity of *Saintpaulia* plants. *Physiol Mol Plant Pathol* **33**: 95–104.

Bolotin A & Biro S (1990) Nucleotide sequence of the putative regulatory gene and major promoter region of the *Streptomyces griseus* glycerol operon. *Gene* **87**: 151–152.

Borthwick AC, Holms WH & Nimmo HG (1984) The phosphorylation of *Escherichia coli* isocitrate dehydrogenase in intact cells. *Biochem J* **222**: 797–804.

Branden C & Tooze J (1991) *Introduction to Protein Structure*. Garland, New York, NY.

Brennan RG & Matthews BW (1989) The helix-turn-helix DNA binding motif. *J Biol Chem* **264**: 1903–1906.

Brennan RG, Roderick SL, Takeda Y & Matthews BW (1990) Protein-DNA conformational changes in the crystal structure of a lambda-Cro-operator complex. *Proc Natl Acad Sci USA* **87**: 8165–8169.

Brice CD & Kornberg HL (1968) Genetic control of isocitrate lyase activity in *Escherichia coli*. *J Bacteriol* **96**: 2185–2186.

Bucher P & Bairoch A (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol* **2**: 53–61.

Bucher P, Karplus K, Moeri N & Hofmann K (1996) A flexible motif search technique based on generalized profiles. *Comput Chem* **20**: 3–24.

Burlingame R & Chapman PJ (1983) Stereospecificity in *meta*-fission catabolic pathways. *J Bacteriol* **155**: 424–426.

Burlingame RP, Wyman L & Chapman PJ (1986) Isolation and characterization of *Escherichia coli* mutants defective for phenylpropionate degradation. *J Bacteriol* **168**: 55–64.

Busenlehner LS, Pennella MA & Giedroc DP (2003) The SmtB/ArsR family of metalloregulatory transcriptional repressors: structural insights into prokaryotic metal resistance. *FEMS Microbiol Rev* **27**: 131–143.

Bustos SA & Schleif RF (1993) Functional domains of the AraC protein. *Proc Natl Assoc Sci USA* **90**: 5638–5642.

Byrne C, Stokes HW & Ward KA (1988) Nucleotide sequence of the *aceB* gene encoding malate synthase A in *Escherichia coli*. *Nucleic Acids Res* **16**: 9342–9342.

Cao JG & Meighen EA (1989) Purification and structural identification of an autoinducer for the luminescence system in *Vibrio harveyi*. *J Biol Chem* **264**: 21670–21676.

Cases I, de Lorenzo V & Ouzounis CA (2003) Transcription regulation and environmental adaptation in bacteria. *Trends Microbiol* **11**: 248–243.

Castillo A & Reverchon S (1997) Characterization of the *pecT* control region from *Erwinia chrysanthemi* strain 3937. *J Bacteriol* **179**: 4909–4918.
Champness WC & Chater KF (1994) Regulation and integration of antibiotic production and morphological differentiation in *Streptomyces* spp. *Regulation of Bacterial Differentiation* (Piggot Jr. P Jr., Moran CP & Youngman P, eds), pp. 61–94. American Society for Microbiology, Washington, DC.
Chater KF (1998) Taking a genetic scalpel to the *Streptomyces* colony. *Microbiology* **144**: 1456–1478.
Chater KF (2001) Regulation of sporulation in *Streptomyces coelicolor* A3 (2): a checkpoint multiplex? *Curr Opin Microbiol* **4**: 667–673.
Chater KF & Horinouchi S (2003) Signalling early developmental events in two highly diverged *Streptomyces* species. *Mol Microbiol* **48**: 9–15.
Chung T, Klumpp DJ & LaPorte DD (1988) Glyoxylate bypass operon of *Escherichia coli*: cloning and determination of the functional map. *J Bacteriol* **170**: 386–392.
Clubb RT, Omichinski JG, Savilahti H, Mizuuchi K, Gronenborg AM & Clore G (1994) A novel class of winged helix-turn-helix protein: the DNA-binding domain of Mu transposase. *Structure* **2**: 1041–1048.
Cohen SP, Hachler H & Levy SB (1993) Genetic and functional analysis of the multiple antibiotic resistance (*mar*) locus in *Escherichia coli*. *J Bacteriol* **175**: 1484–1492.
Collmer A & Keen NT (1986) The role of pectic enzymes in plant pathogenesis. *Ann Rev Phytopathol* **24**: 383–409.
Condemine G, Hugouvieux-Cotte-Pattat N & Robert-Baudouy J (1986) Isolation of *Erwinia chrysanthemi kduD* mutants altered in pectin degradation. *J Bacteriol* **165**: 937–941.
Cortay JC, Bleicher F, Rieul C, Reeves HC & Cozzone AJ (1988) Nucleotide sequence and expression of the *aceK* gene coding for isocitrate dehydrogenase kinase/phosphatase in *Escherichia coli*. *J Bacteriol* **170**: 89–97.
Cortay JC, Nègre D, Galinier A, Duclos B, Perrière G & Cozzone AJ (1991) Regulation of the acetate operon in *Escherichia coli*: purification and functional characterization of the IclR repressor. *EMBO J* **10**: 675–679.
DiMarco AA & Ornston LN (1994) Regulation of *p*-hydroxybenzoate hydroxylase synthesis by PobR bound to an operator in *Acinetobacter calcoaceticus*. *J Bacteriol* **176**: 4277–4284.
DiMarco AA, Averhoff B & Ornston LN (1993) Identification of the transcriptional activator *pobR* and characterization of its role in the expression of *pobA*, the structural gene for *p*-hydroxybenzoate hydroxylase in *Acinetobacter calcoaceticus*. *J Bacteriol* **175**: 4499–4506.
DiRusso C, Heimert TL & Metzger AK (1992) Characterization of FadR, a global transcriptional regulatory of fatty acid metabolism: interaction with the *fadB* promoter is prevented by long chain fatty acyl coenzyme A. *J Biol Chem* **267**: 8685–8691.
Domínguez P & Marqués S (2004) Compiling sigma-70 dependent promoters. *Pseudomonas. Vol. II* (Ramos JL, ed), pp. 319–343. Kluwer, London.
Donald LJ, Chernushevich IV, Zhou J, Verentchikov A, Poppe-Schriemer N, Hosfield DJ, Westmore JB, Ens W, Duckworth HW & Standing KG (1996) Preparation and properties of pure, full-length IclR protein of *Escherichia coli*. Use of time-of-flight mass spectrometry to investigate the problems encountered. *Protein Sci* **5**: 1613–1624.
Donald LJ, Hosfield DJ, Cuvelier SL, Ens W, Standig KG & Duckworth HW (2001) Mass spectrometric study of the *Escherichia coli* repressor protein, IclR and GalR, their complexes with DNA. *Protein Sci* **10**: 1370–1380.
Duque E, Segura A, Mosqueda G & Ramos JL (2001) Global and cognate regulators control the expression of the organic solvent efflux pumps TtgABC and TtgDEF of *Pseudomonas putida*. *Mol Microbiol* **39**: 1100–1106.
Díaz E & Prieto MA (2000) Bacterial promoter triggering biodegradation of aromatic pollutants. *Curr Opin Biotechnol* **11**: 467–475.
Díaz E, Ferrández A, Prieto MA & García JL (2001) Biodegradation of aromatic compounds by *Escherichia coli*. *Microbiol Mol Biol Rev* **65**: 523–569.
Eberhard A, Burlingame AL, Eberhard C, Kenyon GL, Nealson KH & Oppenheimer NJ (1981) Structural identification of autoinducer of *Photobacterium fischeri*. *Biochemistry* **20**: 2444–2449.
Egan S (2002) Growing repertoire of AraC/XylS activators. *J Bacteriol* **184**: 5529–5532.
Eggers CH, Caimano MJ & Radolf JD (2004) Analysis of promoter elements involved in the transcriptional initiation of RpoS-dependent *Borrelia burgdorferi* genes. *J Bacteriol* **186**: 7390–7402.
Elsemore DA & Ornston LN (1994) The *pca-pob* supraoperonic cluster of *Acinetobacter calcoaceticus* contains *quiA*, the structural gene for quinate-shikimate dehydrogenase. *J Bacteriol* **176**: 7659–7666.
Elsemore DA & Ornston LN (1995) Unusual ancestry of dehydratases associated with quinate catabolism in *Acinetobacter calcoaceticus*. *J Bacteriol* **177**: 5971–5978.
Eulberg D & Schlömann M (1998) The putative regulator of catechol catabolism in *Rhodococcus opacus* 1CP, is an IclR-type, not a LysR-type transcriptional regulator. *Antonie van Leeuwenhoek* **74**: 71–82.
Eulberg D, Lakner S, Golovleva LA & Schlömann M (1998) Characterization of a protocatechuate catabolic gene cluster from *Rhodococcus opacus* 1CP: evidence for a merged enzyme with 4-carboxymuconolactone-decarboxylating and 3-oxoadipate enol-lactone-hydrolyzing activity. *J Bacteriol* **180**: 1072–1081.
Ferrández A, García JL & Díaz E (1997) Genetic characterization and expression in heterologous hosts of the 3-3-(hydroxyphenyl)-propionate catabolic pathway of *Escherichia coli* K-12. *J Bacteriol* **170**: 2573–2581.

Frazee RW, Livingston DM, LaPorte DC & Lipscomb JD (1993) Cloning, sequencing, expression of the *Pseudomonas putida* protocatechuate 3,4-dioxygenase genes. *J Bacteriol* **175**: 6194–6202.

Fromknecht K, Vogel PD & Wise JG (2003) Combinatorial redesign of the DNA binding specificity of a prokaryotic helix-turn-helix repressor. *J Bacteriol* **185**: 475–481.

Fuqua WC, Winans SC & Greenberg EP (1994) Quorum sensing in bacteria: the LuxR–LuxI family of cell density-responsive transcriptional regulators. *J Bacteriol* **176**: 269–275.

Fuqua C, Winans SC & Greenberg EP (1996) Census and consensus in bacterial ecosystems: the LuxR–LuxI family of quorum-sensing transcriptional regulators. *Annu Rev Microbiol* **50**: 727–751.

Gallegos MT, Michán C & Ramos JL (1993) The XylS/AraC family of regulators. *Nucleic Acids Res* **21**: 807–810.

Gallegos MT, Marqués S & Ramos JL (1996) The TACAN$_4$TGCA motif upstream from the −35 region in the σ^{70}/σ^{S}-dependent Pm promoter of the TOL plasmid is the minimum DNA segment required for transcription stimulation by XylS regulators. *J Bacteriol* **178**: 6427–6434.

Gallegos MT, Schleif R, Bairoch A, Hofmann K & Ramos JL (1997) AraC/XylS family of transcriptional regulators. *Microbiol Mol Biol Rev* **61**: 393–410.

Garnak M & Reeves HC (1979) Phosphorylation of isocitrate dehydrogenase of *Escherichia coli*. *Science* **203**: 1111–1112.

Gerischer U & Ornston LN (1995) Spontaneous mutations in *pcaH* and -*G*, structural genes for protocatechuate 3,4-dioxygenase in *Acinetobacter calcoaceticus*. *J Bacteriol* **177**: 1336–1347.

Gerischer U, Segura A & Ornston LN (1998) PcaU, a transcriptional activator of genes for protocatechuate utilization in *Acinetobacter*. *J Bacteriol* **180**: 1512–1524.

Grainger DC, Overton TW, Reppas N, Wade JT, Tamai E, Hobman JL, Constantinidou C, Struhl K, Church G & Busby SJW (2004) Genomic studies with *Escherichia coli* MelR protein: applications of chromatin immunoprecipitation and microarrays. *J Bacteriol* **186**: 6938–6943.

Grkovic S, Brown MH, Schumacher MA, Brennan RG & Skurray RA (2001a) The Staphylococcal QacR multidrug regulator binds a correctly spaced operator as a pair of dimers. *J Bacteriol* **183**: 7102–7109.

Grkovic S, Brown MH & Skurray RA (2001b) Transcriptional regulation of multidrug efflux pumps in bacteria. *Cell Develop Biol* **12**: 225–237.

Guazzaroni M-E, Terán W, Zhang X, Gallegos M-T & Ramos JL (2004) TtgV Bound to a complex operator site represses transcription of the promoter for the multidrug and solvent extrusion TtgGHI pump. *J Bacteriol* **186**: 2921–2927.

Guazzaroni M-E, Krell T, Felipe A, Ruíz R, Meng C, Zhang X, Gallegos M-T & Ramos JL (2005) The multidrug efflux regulator TtgV recognizes a wide range of ligands in solution and complexed with target DNA. Evidence from isothermal titration calorimetry. *J Biol Chem* **208**: 20887–20893.

Gui L, Sunnarborg A, Pan B & LaPorte DC (1996) Autoregulation of *iclR*, the gene encoding the repressor of the glyoxylate bypass operon. *J Bacteriol* **178**: 321–324.

Guo Z & Houghton JE (1999) PcaR-mediated activation and repression of *pca* genes from *Pseudomonas putida* are propagated by its binding to both the −35 and the −10 promoter elements. *Mol Microbiol* **32**: 253–263.

Hao J & Kendrick KE (1998) Visualization of penicillin-binding proteins during sporulation of *Streptomyces griseus*. *J Bacteriol* **180**: 2125–2132.

Harmer TL & Schleif R (2000) The C-terminal end of AraC tightly binds to the rest of its domain. *J Biol Chem* **276**: 4886–4888.

Harrison SC (1991) A structural taxonomy of DNA-binding domains. *Nature* **353**: 715–719.

Harrison SC & Aggarwal AK (1990) DNA recognition by proteins with the helix-turn-helix motif. *Annu Rev Biochem* **59**: 933–969.

Hartnett GB & Averhoff B (1990) Selection of *Acinetobacter calcoaceticus* mutants deficient in the *p*-hydroxybenzoate hydroxylase gene (*pobA*), a member of a suprooperonic cluster. *J Bacteriol* **172**: 6160–6161.

Harwood CS & Parales RE (1996) The β-ketoadipate pathway and the biology of self-identity. *Annu Rev Microbiol* **50**: 553–590.

Harwood CS, Nichols NN, Kim MK, Ditty JL & Parales RE (1994) Identification of the *pcaRKF* gene cluster from *Pseudomonas putida*: involvement in chemotaxis, biodegradation, transport of 4-hydroxybenzoate. *J Bacteriol* **176**: 6479–6488.

Haydon DJ & Guest JR (1991) A new family of bacterial regulatory proteins. *FEMS Microbiol Lett* **63**: 291–295.

Heldwein EE & Brennan RG (2001) Crystal structure of the transcription factor BmrR bound to DNA and a drug. *Nature* **409**: 378–382.

Henikoff S, Haughn GW, Calvo JM & Wallace JC (1988) A large family of bacterial activator proteins. *Proc Natl Acad Sci USA* **85**: 6602–6606.

Henikoff S, Wallace JC & Brown JP (1990) Finding protein similarities with nucleotide sequence databases. *Meth Enzymol* **183**: 111–132.

Henry MF & Cronan JE (1992) A new mechanism of transcriptional regulation: release of a activator triggered by small molecule binding. *Cell* **70**: 671–679.

Hindle Z & Smith CP (1994) Substrate induction and catabolite repression of the *Streptomyces coelicolor* glycerol operon are mediated through the GylR protein. *Mol Microbiol* **12**: 737–745.

Hinton JCD, Sidebotham JM, Hyman LJ, Perombelon MCM & Salmond GPC (1989) Isolation and characterisation of transposon-induced mutants of *Erwinia carotovora* subsp. *atroseptica* exhibiting reduced virulence. *Mol Gen Genet* **217**: 141–148.

Ho YS, Burden LM & Hurley JH (2000) Structure of the GAF domain, a ubiquitous signaling motif and a new class of cyclic GMP receptor. *EMBO J* **19**: 5288–5299.

Holm L & Sander C (1996) Mapping the protein universe. *Science* **273**: 595–602.
Hopper S, Korsa I & Bock A (1996) The nucleotide concentration determines the specificity of in vitro transcription activation by the sigma 54-dependent activator FhlA. *J Bacteriol* **178**: 199–203.
Horinouchi S (2002) A microbial hormone, A-factor, as a master switch for morphological differentiation and secondary metabolism in *Streptomyces griseus*. *Front Biosci* **7**: 2045–2057.
Hughes EJ, Shapiro MJ, Houghton JE & Ornston LN (1988) Cloning and expression of *pca* genes from *Pseudomonas putida* in *Escherichia coli*. *J Gen Microbiol* **134**: 2877–2887.
Hugouvieux-Cotte-Pattat N & Robert-Baudouy J (1987) Hexuronate catabolism in *Erwinia chrysanthemi*. *J Bacteriol* **169**: 1223–1231.
Hugouvieux-Cotte-Pattat N, Condemine G, Nasser W & Reverchon S (1996) Regulation of pectinolysis in *Erwinia chrysanthemi*. *Ann Rev Microbiol* **50**: 213–257.
Hugouvieux-Cotte-Pattat N, Blot N & Reverchon S (2001) Identification of TogMNAB, an ABC transporter which mediates the uptake of pectic oligomers in *Erwinia chrysanthemi* 3937. *Mol Microbiol* **41**: 1113–1123.
Hugouvieux-Cotte-Pattat N, Shevchik VE & Nasser W (2002) PehN, a polygalacturonase homologue with a low hydrolase activity, is coregulated with the other *Erwinia chrysanthemi* polygalacturonases. *J Bacteriol* **184**: 2664–2673.
Hwang I, Li PL, Zhang LH, Piper KR, Cook DM, Tate ME & Farrand SK (1994) TraI, a LuxI homologue, is responsible for production of conjugation factor, the Ti plasmid *N*-acylhomoserine lactone autoinducer. *Proc Natl Acad Sci USA* **91**: 4639–4643.
Jiang H & Kendrick KE (2000) Characterization of *ssfR* and *ssgA*, two genes involved in sporulation of *Streptomyces griseus*. *J Bacteriol* **182**: 5521–5529.
Jones S, Yu B, Bainton NJ, Birdsall M, Bycroft BW & Chhabra SR (1993) The *lux* autoinducer regulates the production of exoenzyme virulence determinants in *Erwinia carotovora* and *Pseudomonas aeruginosa*. *EMBO J* **12**: 2477–2482.
Jourdan AD & Stauffer GV (1998) Mutational analysis of the transcriptional regulator GcvA: amino acids important for activation, repression, DNA binding. *J Bacteriol* **180**: 4865–4871.
Karmierczak MJ, Mithoe SC, Boor KJ & Wiedmann M (2003) *Listeria monocytogenes* σ^B regulates stress response and virulence functions. *J Bacteriol* **185**: 5722–5734.
Karmirantzou M & Hamodrakas SJ (2001) A Web-based classification system of DNA-binding protein families. *Protein Eng* **14**: 465–472.
Keijser BJ, van Wezel GP, Canters GW & Vijgenboom E (2002) Developmental regulation of the *Streptomyces lividans ram* genes: involvement of RamR in regulation of the *ramCSAB* operon. *J Bacteriol* **184**: 4420–4429.
Kendrick KE & Ensign JC (1983) Sporulation of *Streptomyces griseus* in submerged culture. *J Bacteriol* **155**: 357–366.
Kessler B, de Lorenzo V & Timmis KN (1993) Identification of a *cis*-acting sequence within the Pm promoter of the TOL plasmid which confers XylS-mediated responsiveness to substituted benzoates. *J Mol Biol* **5**: 699–703.
Klumpp DJ, Plank DW, Bowdin LJ, Stueland CS, Chung T & LaPorte DC (1988) Nucleotide sequence of *aceK*, the gene encoding isocitrate dehydrogenase kinase/phosphatase. *J Bacteriol* **170**: 2763–2769.
Kok RG, D'Argenio DA & Ornston LN (1997) Combining localized PCR mutagenesis and natural transformation in direct genetic analysis of a transcriptional regulator gene, *pobR*. *J Bacteriol* **179**: 4270–4276.
Kok RG, D'Argenio DA & Ornston LN (1998) Mutation analysis of PobR and PcaU, closely related transcriptional activators in *Acinetobacter*. *J Bacteriol* **180**: 5058–5069.
Kornberg HL (1966) The role and control of the glyoxylate cycle in *Escherichia coli*. *Biochem J* **99**: 1–11.
Kornberg HL (1967) The regulation of anaplerotic enzymes in *E. coli*. *Bull Biochem Soc Chem Biol* **49**: 1479–1490.
Korsa I & Bock A (1997) Characterization of *fhlA* mutations resulting in ligand-independent transcriptional activation and ATP hydrolysis. *J Bacteriol* **179**: 41–45.
Kroening TA & Kendrick KE (1987) *In vivo* regulation of histidine ammonia-lyase activity from *Streptomyces griseus*. *J Bacteriol* **169**: 823–829.
Kwak J & Kendrick KE (1996) Bald mutants of *Streptomyces griseus* that prematurely undergo key events of sporulation. *J Bacteriol* **178**: 4643–4650.
LaPorte DC & Chung T (1985) A single gene codes for the kinase and phosphatase which regulate isocitrate dehydrogenase. *J Biol Chem* **260**: 1591–1597.
LaPorte DC & Koshland DE Jr (1982) A protein with kinase and phosphatase activities involved in the regulation of the tricarboxylic acid cycle. *Nature* **300**: 458–460.
LaPorte DC, Walsh K & Koshland DE Jr (1984) The branch point effect: ultrasensitivity and subsensitivity to metabolic control. *J Biol Chem* **259**: 14068–14075.
LaPorte DC, Stueland CS & Ikeda T (1989) Isocitrate dehydrogenase kinase/phosphatase. *Biochimie* **71**: 1051–1057.
Larimer FW, Chain P, Hauser L, *et al.* (2004) Complete genome sequence of the metabolically versatile photosynthetic bacterium *Rhodospirillum palustris*. *Nat Biotech* **22**: 55–61.
Laurent F, Kotoujansky A, Labesse G & Bertheau Y (1993) Characterisation and overexpression of the *pem* gene encoding pectin methylesterase of *Erwinia chrysanthemi* strain 3937. *Gene* **131**: 17–25.
Levy SB (1988) Tetracycline resistance determinants are widespread. *ASM News* **54**: 418–421.
Little R & Dixon R (2003) The amino-terminal GAF domain of *Azotobacter vinelandii* NifA binds 2-oxoglutarate to resist inhibition by NifL under nitrogen-limiting conditions. *J Biol Chem* **278**: 28711–28718.
Liu Y, Jiang G, Cui Y, Mukherjee A, Ma WL & Chatterjee AK (1999) *kdgR*$_{Ecc}$ negatively regulates genes for pectinases, cellulase, protease, Harpin$_{Ecc}$, a global RNA regulator in

Erwinia carotovora subsp. *carotovora*. *J Bacteriol* **181**: 2411–2421.
Maloy SR & Nunn WD (1982) Genetic regulation of the glyoxylate shunt in *Escherichia coli* K-12. *J Bacteriol* **149**: 173–180.
Martin VJJ & Mohn WW (2000) Genetic investigation of the catabolic pathway for degradation of abietane diterpenoids by *Pseudomonas abietaniphila* BKME-9. *J Bacteriol* **182**: 3784–3793.
Martínez-Hackert E & Stock AM (1997) Structural relationships in the OmpR family of winged-helix transcription factors. *J Mol Biol* **269**: 301–312.
McFall SM, Chugani SA & Chakrabarty AM (1998) Transcriptional activation of the catechol operons: variations on a theme. *Gene* **223**: 257–267.
McKay DB & Steitz TA (1981) Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. *Nature* **290**: 744–749.
Mosqueda G & Ramos JL (2000) A set of genes encoding a second toluene efflux system in *Pseudomonas putida* DOT-T1E is linked to the *tod* genes for toluene metabolism. *J Bacteriol* **182**: 937–943.
Méndez B, Tachibana C & Levy SB (1980) Heterogeneity of tetracycline resistance determinants. *Plasmid* **3**: 99–108.
Nasser W, Condemine G, Plantier R, Anker D & Robert-Baudouy J (1991) Inducing properties of analogues of 2-keto-3-deoxygluconate on the expression of pectinase genes of *Erwinia chrysanthemi*. *FEMS Microbiol Lett* **81**: 73–78.
Nasser W, Reverchon S & Robert-Baudouy J (1992) Purification and functional characterisation of the KdgR protein, a major repressor of pectinolysis genes of *Erwinia chrysanthemi*. *Mol Microbiol* **6**: 257–265.
Nasser W, Reverchon C, Condemine G & Robert-Baudouy J (1994) Specific interactions of *Erwinia chrysanthemi* KdgR repressor with different operators of genes involved in pectinolysis. *J Mol Biol* **236**: 427–440.
Nasser W, Robert-Baudouy J & Reverchon S (1997) Antagonistic effect of CRP and KdgR in the transcription control of the *Erwinia chrysanthemi* pectinolysis genes. *Mol Microbiol* **26**: 1071–1082.
Nga DP, Altenbuchner J & Heiss GS (2004) NpdR, a repressor involved in 2,4,6-Trinitrophenol degradation in *Rhodococcus opacus* HL PM-1. *J Bacteriol* **186**: 98–103.
Nguyen CC & Saier MH Jr (1995) Phylogenetic, structural and functional analyses of the LacI–GalR family of bacterial transcription factors. *FEBS Lett* **377**: 98–102.
Nichols NN & Harwood CS (1995) Repression of 4-hydroxybenzoate transport and degradation by benzoate: a new layer of regulatory control in the *Pseudomonas putida* β-ketoadipate pathway. *J Bacteriol* **177**: 7033–7040.
Nomura K, Nasser W, Kawagishi H & Tsuyumu S (1998) The *pir* gene of *Erwinia chrysanthemi* EC16 regulates hyperinduction of pectate lyase virulence genes in response to plant signals. *Proc Natl Acad Sci USA* **95**: 14034–14039.
Nomura K, Nasser W & Tsuyumu S (1999) Self-regulation of *pir*, a regulatory protein responsible for hyperinduction of pectate lyase in *Erwinia chrysanthemi* EC16. *Mol Plant Microbe Interact* **12**: 385–390.
North AK, Klose KE, Stedman KM & Kustu S (1993) Prokaryotic enhancer-binding proteins reflect eukaryote-like modularity: the puzzle of nitrogen regulatory protein C. *J Bacteriol* **175**: 4267–4273.
Nègre D, Cortay JC, Old IG, Galinier A, Richaud C, Saint Girons I & Cozzone AJ (1991) Overproduction and characterization of the *iclR* gene product of *Escherichia coli* K-12 and comparison with that of *Salmonella typhimurium* LT2. *Gene* **97**: 29–37.
Nègre D, Cortay JC, Galinier A, Sauve P & Cozzone AJ (1992) Specific interactions between the IclR repressor of the acetate operon of *Escherichia coli* and its operator. *J Mol Biol* **228**: 23–29.
O'Halloran TV, Frantz B, Shin K, Ralston DM & Wright JG (1989) The MerR heavy metal receptor mediates positive activation in a topologically novel transcription complex. *Cell* **56**: 119–129.
Ornston LN (1966a) The conversion of catechol and protocatechuate to β-ketoadipate by *Pseudomonas putida*. II. Enzymes of the protocatechuate pathway. *J Biol Chem* **241**: 3787–3794.
Ornston LN (1966b) The conversion of catechol and protocatechuate to β-ketoadipate by *Pseudomonas putida*. IV. Regulation. *J Biol Chem* **241**: 3800–3810.
Orth P, Schnappinger D, Hillen W, Saenger W & Hinrichs W (2000) Structural basis of gene regulation by the tetracycline inducible Tet repressor–operator system. *Nat Struct Biol* **7**: 215–219.
Pabo CO & Sauer RT (1992) Transcription factors: structural families and principles of DNA recognition. *Annu Rev Bichem* **61**: 1053–1095.
Pan B, Unnikrishnan I & LaPorte DC (1996) The binding site of the IclR repressor protein overlaps the promoter of *aceBAK*. *J Bacteriol* **178**: 3982–3984.
Parales RE & Harwood CS (1992) Characterization of the genes encoding β-ketoadipate: uccinyl-coenzyme A transferase in *Pseudomonas putida*. *J Bacteriol* **174**: 4657–4666.
Parales RE & Harwood CS (1993) Regulation of the *pcaIJ* genes for aromatic acid degradation in *Pseudomonas putida*. *J Bacteriol* **175**: 5829–5838.
Parke D & Ornston LN (1976) Constitutive synthesis of the enzymes of the protocatechuate and of the β-ketoadipate uptake system in mutant strains of *Pseudomonas putida*. *J Bacteriol* **126**: 271–281.
Parke D, D'Argenio DA & Ornston LN (2000) Bacteria are not what they eat: that is why they are so diverse. *J Bacteriol* **182**: 257–263.
Pearson JP, Gray KM, Passador L, Tucker KD, Eberhard A, Iglewski BH & Greenberg EP (1994) Structure of the autoinducer required for expression of *Pseudomonas aeruginosa* virulence genes. *Proc Natl Acad Sci USA* **91**: 197–201.

Perombelon MCM & Kelman A (1980) Ecology of the soft rot erwinias. *Annu Rev Phytopathol* **18**: 361–387.
Piper KR, von Bodman SB & Farrand SK (1993) Conjugation factor of *Agrobacterium tumefaciens* regulates Ti plasmid transfer by autoinduction. *Nature* **362**: 448–450.
Pirhonen M, Saarilahti H, Karlsson MB & Palva ET (1991) Identification of pathogenicity determinants of *Erwinia carotovora* subsp. *carotovora* by transposon mutagenesis. *Mol Plant–Microbe Interact* **4**: 276–283.
Poole K, Tetro K, Zhao Q, Neshat S, Heinrichs DE & Bianco N (1996) Expression of the multidrug resistance operon *mexA–mexB–oprM* in *Pseudomonas aeruginosa*: *mexR* encodes a regulator of operon expression. *Antimicrob Agents Chemother* **40**: 2021–2028.
Popp R, Kohl T, Patz P, Trautwein G & Gerischer U (2002) Differential DNA binding of transcriptional regulator PcaU from *Acinetobacter* sp. strain ADP1. *J Bacteriol* **184**: 1988–1997.
Praillet T, Nasser W, Robert-Baudouy J & Reverchon S (1996) Purification and functional characterisation of PecS: a regulator of virulence factor synthesis in *Erwinia chrysanthemi*. *Mol Microbiol* **20**: 391–402.
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.
Qin Y, Luo ZQ, Smyth AJ, Gao P, von Bodman SB & Farrand SK (2000) Quorum-sensing signal binding results in dimerization of TraR and its release from membranes into the cytoplasm. *EMBO J* **19**: 5212–5221.
Ramos JL, Rojo F, Zhou L & Timmis KN (1990) A family of positive regulators related to the *Pseudomonas putida* TOL plasmid XylS and the *Escherichia coli* AraC activators. *Nucleic Acids Res* **18**: 2149–2152.
Ramos JL, Duque E, Huertas MJ & Haïdour A (1995) Isolation and expansion of the catabolic potential of a *Pseudomonas putida* strain able to grow in the presence of high concentrations of aromatic hydrocarbons. *J Bacteriol* **177**: 3911–3916.
Ramos JL, Duque E, Godoy P & Segura A (1998) Efflux pumps involved in toluene tolerance in *Pseudomonas putida* DOT-T1E. *J Bacteriol* **180**: 3323–3329.
Ramos JL, Gallegos MT, Marqués S, Ramos-González MI, Espinosa-Urgel M & Segura A (2001) Response of gram-negative bacteria to certain environmental stresses. *Curr Opin Microbiol* **4**: 166–171.
Ramos JL, Duque E, Gallegos MT, Godoy P, Ramos-González MI, Rojas A, Terán W & Segura A (2002) Mechanisms of solvent tolerance in gram-negative bacteria. *Ann Rev Microbiol* **56**: 743–768.
Ramos JL, Martínez-Bueno M, Molina-Henares AJ, Terán W, Watanabe K, Zhang X, Gallegos MT, Brennan R & Tobes R (2005) The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* **69**: 326–356.
Resnik E, Pan B, Ramani N, Freundlich M & LaPorte DC (1996) Integration host factor amplifies the induction of the *aceBAK* operon of *Escherichia coli* by relieving IclR repression. *J Bacteriol* **178**: 2715–2717.
Reverchon S & Robert-Baudouy J (1987) Molecular cloning of *Erwinia chrysanthemi* oligogalacturonate lyase gene involved in pectin degradation. *Gene* **55**: 125–133.
Reverchon S, Nasser W & Robert-Baudouy J (1991) Characterisation of *kdgR*, a gene of *Erwinia chrysanthemi* that regulates pectin degradation. *Mol Microbiol* **5**: 2203–2216.
Reverchon S, Nasser W & Robert-Baudouy J (1994) *pecS*: a locus controlling pectinase, cellulase and blue pigment production in *Erwinia chrysanthemi*. *Mol Microbiol* **11**: 1127–1139.
Reverchon S, Expert D, Robert-Baudouy J & Nasser W (1997) The cyclic AMP receptor protein is the main activator of the pectinolysis genes in *Erwinia chysanthemi*. *J Bacteriol* **179**: 3500–3508.
Reyes-Ramírez F, Little R & Dixon R (2002) Mutant forms of the *Azotobacter vinelandii* transcriptional activator NifA resistant to inhibition by the NifL regulatory protein. *J Bacteriol* **184**: 6777–6785.
Rhee S, Martin RG, Rosner JL & Davies DR (1998) A novel DNA-binding motif in MarA: the first structure for an AraC family transcriptional activator. *Proc Natl Acad Sci USA* **95**: 10413–10418.
Rieul C, Bleicher F, Duclos B, Cortay JC & Cozzone AJ (1988) Nucleotide sequence of the *aceA* gene coding for isocitrate lyase in *Escherichia coli*. *Nucleic Acids Res* **16**: 5689.
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, YtrA subfamilies. *J Biol Chem* **277**: 12507–12515.
Rintoul MR, Cusa E, Baldomà L, Badia J, Reitzer L & Aguilar J (2002) Regulation of the *Escherichia coli* allantoin regulon: coordinated function of the repressor AllR and the activator AllS. *J Mol Biol* **324**: 599–610.
Roessner CA, Huang KX, Warren MJ, Raux E & Scott AI (2002) Isolation and characterization of 14 additional genes specifying the anaerobic biosynthesis of cobalamin (vitamin B12) in *Propionibacterium freudenreichii* (*P. shermanii*). *Microbiology* **148**: 1845–1853.
Rojas A, Duque E, Mosqueda G, Golden G, Hurtado A, Ramos JL & Segura A (2001) Three efflux pumps are required to provide efficient tolerance to toluene in *Pseudomonas putida* DOT-T1E. *J Bacteriol* **183**: 3967–3973.
Rojas A, Segura A, Guazzaroni ME, Teran W, Hurtado A, Gallegos MT & Ramos JL (2003) *In vivo* and *in vitro* evidence that TtgV is the specific regulator of the TtgGHI multidrug and solvent efflux pump of *Pseudomonas putida*. *J Bacteriol* **185**: 4755–4763.
Rojo F (1999) Repression of transcription initiation in bacteria. *J Bacteriol* **181**: 2987–2991.
Rollenhagen C, Antelmann H, Kirstein J, Delumeau O, Hecker M & Yudkin MD (2003) Binding of A and B to core RNA polymerase after environmental stress in *Bacillus subtilis*. *J Bacteriol* **185**: 35–40.

Romero-Steiner S, Parales RE, Harwood CS & Houghton JE (1994) Characterization of the *pcaR* regulatory gene from *Pseudomonas putida*, which is required for the complete degradation of *p*-hydroxybenzoate. *J Bacteriol* **176**: 5771–5779.

Rosinski JA & Atchley WR (1999) Molecular evolution of helix-turn-helix proteins. *J Mol Evol* **49**: 301–309.

Rost B & Sander C (1993) Prediction of protein secondary structure at better than 70% accuracy. *J Mol Biol* **232**: 584–99.

Rouanet C, Nomura K, Tsuyumu S & Nasser W (1999) Regulation of *pelD* and *pelE*, encoding major alkaline pectate lyases in *Erwinia chrysanthemi*: involvement of the main transcriptional factors. *J Bacteriol* **181**: 5948–5957.

Rousseau P, Guegen E, Duval-Valentin G & Chandler M (2004) The helix-turn-helix motif of bacterial insertion sequence IS*911* transposase is required for DNA binding. *Nucleic Acids Res* **32**: 1335–1344.

Schell MA (1993) Molecular biology of the LysR family of transcriptional regulators. *Annu Rev Microbiol* **47**: 597–626.

Schevitz RW, Otwinowski Z, Joachimiak A, Lawson CL & Singler PB (1985) The three-dimensional structure of *trp* repressor. *Nature* **317**: 782–786.

Schumacher MA, Miller MC, Grkovic S, Brown MH, Skurray RA & Brennan R (2002) Structural basis for cooperative DNA binding by two dimers of the multidrug-binding protein QacR. *EMBO J* **21**: 1210–1218.

Schwedock J, McCormick JR, Angert ER, Nodwell JR & Losick R (1997) Assembly of the cell division protein FtsZ into ladder-like structures in the aerial hyphae of *Streptomyces coelicolor*. *Mol Microbiol* **26**: 847–858.

Shevchik VE, Condemine G, Hugouvieux-Cotte-Pattat N & Robert-Baudouy J (1996) Characterization of pectin methylesterase B, an outer membrane lipoprotein of *Erwinia chrysanthemi* 3937. *Mol Microbiol* **19**: 455–466.

Shevchik VE & Hugouvieux-Cotte-Pattat N (2003) PaeX, a second pectin acetylesterase of *Erwinia chrysanthemi* 3937. *J Bacteriol* **185**: 3091–3100.

Smith CP & Chater KF (1988a) Structure and regulation of controlling sequences for the *Streptomyces coelicolor* glycerol operon. *J Mol Biol* **204**: 569–580.

Smith CP & Chater KF (1988b) Cloning and transcription analysis of the entire glycerol utilization (*gylABX*) operon of *Streptomyces coelicolor*A3(2) and identification of a closely associated transcription unit. *Mol Gen Genet* **211**: 129–137.

Song Y, Tong Z, Wang J, *et al.* (2004) Complete genome sequence of *Yersinia pestis* strain 91001, an isolate avirulent to humans. *DNA Res* **11**: 179–197.

Spiro S (1994) The FNR family of transcriptional regulators. *Antonie van Leeuwenhoek* **66**: 23–36.

Summers AO (1992) Untwist and shout: a heavy metal-responsive transcriptional regulator. *J Bacteriol* **174**: 3097–3101.

Sunnarborg A, Klumpp D, Chung T & LaPorte DC (1990) Regulation of the glyoxylate bypass operon: cloning and characterization of IclR. *J Bacteriol* **172**: 2642–2649.

Surette MG & Bassler BL (1998) Quorum sensing in *Escherichia coli* and *Salmonella typhimurium*. *Proc Natl Acad Sci USA* **95**: 7046–7050.

Surgey N, Robert-Baudouy J & Condemine G (1996) The *Erwinia chrysanthemi pecT* gene regulates pectinase gene expression. *J Bacteriol* **178**: 1593–1599.

Takano E, Chakraburtty R, Nihira T, Yamada Y & Bibb MJ (2001) A complex role for the gamma-butyrolactone SCB1 in regulating antibiotic production in *Streptomyces coelicolor* A3 (2). *Mol Microbiol* **41**: 1015–1028.

Thomson NR, Nasser W, McGowan S, Sebaihia M & Salmond GP (1999) *Erwinia carotovora* has two KdgR-like proteins belonging to the IclR family of transcriptional regulator: identification and characterization of the RexZ activator and the KdgR repressor of pathogenesis. *Microbiology* **145**: 1531–1545.

Torres B, Porras G, García JL & Díaz E (2003) Regulation of the *mhp* cluster responsible for 3-(3-hydroxyphenyl)propionic acid degradation in *Escherichia coli*. *J Biol Chem* **278**: 27575–27585.

Traag BA, Kelemen GH & van Wezel GP (2004) Transcription of the sporulation gene *ssgA* is activated by the IclR-type regulator SsgR in a *whi*-independent manner in *Streptomyces coelicolor* A3(2). *Mol Microbiol* **53**: 985–1000.

Tralau T, Cook AM & Ruff J (2003) An additional regulator, TsaQ, is involved with TsaR in regulation of transport during the degradation of *p*-toluenesulfonate in *Comamonas testosteroni* T-2. *Arch Microbiol* **180**: 319–326.

Trautwein G & Gerischer U (2001) Effects exerted by transcriptional regulator PcaU from *Acinetobacter* sp. strain ADP1. *J Bacteriol* **183**: 873–881.

Tsoi TV, Plotnikova EG, Cole JR, Guerin WF, Bagdasarian M & Tiedje JM (1999) Cloning, expression, nucleotide sequence of the *Pseudomonas aeruginosa* 142 *ohb* genes coding for oxygenolytic ortho dehalogenation of halobenzoates. *Appl Environ Microbiol* **65**: 2151–2162.

Ventre I, Filloux A & Lazdunski A (2004) Two-component signal transduction systems: a key to the adaptative potential of *Pseudomonas aeruginosa*. *Pseudomonas. Vol. II* (Ramos JL, ed), pp. 257–288. Kluwer, London.

Weickert MJ & Adhya S (1992) A family of bacterial regulators homologous to Gal and Lac repressors. *J Biol Chem* **267**: 15869–15874.

Welch M, Todd DE, Whitehead NA, McGowan SJ, Bycroft BW & Salmond GPC (2000) N-acyl homoserine lactone binding to the CarR receptor determines quorum-sensing specificity in *Erwinia*. *EMBO J* **19**: 631–641.

van Wezel GP, van der Meulen J, Kawamoto S, Luiten RG, Koerten HK & Kraal B (2000) *ssgA* is essential for sporulation of *Streptomyces coelicolor* A3(2) and affects hyphal development by stimulating septum formation. *J Bacteriol* **182**: 5653–5662.

Williams SE, Woolridge EM, Ransom SC, Landro JA, Babbit PC & Kozarich JW (1992) 3-Carboxy-*cis*, *cis*-muconate lactonizing enzyme from *Pseudomonas putida* is homologous to the class II

fumarase family: a new reaction in the evolution of a mechanistic motif. *Biochemistry* **31**: 9768–9776.

Winans SC & Bassler BL (2002) Mob psychology. *J Bacteriol* **184**: 873–883.

Wintjens R & Rooman M (1996) Structural classification of HTH DNA-binding domains and protein–DNA interaction modes. *J Mol Biol* **262**: 294–313.

Yamamoto K & Ishihama A (2003) Two different modes of transcription repression of the *Escherichia coli* acetate operon by IclR. *Mol Microbiol* **47**: 183–194.

Yamazaki H, Ohnishi Y & Horinouchi S (2003) Transcriptional switch on of *ssgA* by A-factor, which is essential for spore septum formation in *Streptomyces griseus*. *J Bacteriol* **185**: 1273–1283.

Zhang HB, Wang LH & Zhang LH (2002) Genetic control of quorum-sensing signal turnover in *Agrobacterium tumefaciens*. *Proc Natl Acad Sci USA* **99**: 4638–4643.

Zhang H-B, Wang C & Zhang L-H (2004) The quormone degradation system of *Agrobacterium tumefaciens* is regulated by starvation signal and stress alarmone (p)ppGpp. *Mol Microbiol* **52**: 1389–1401.

Zhang LH & Kerr A (1991) A diffusible compound can enhance conjugal transfer of the Ti plasmid in *Agrobacterium tumefaciens*. *J Bacteriol* **173**: 1867–1872.

Zhang LH, Murphy PJ, Kerr A & Tate ME (1993) *Agrobacterium* conjugation and gene regulation by *N*-acyl-L-homoserine lactones. *Nature* **362**: 446–448.

Zhang RG, Kim Y, Skarina T, Beasley S, Laskowski R & Arrowsmith C (2002) Crystal structure of *Thermotoga maritima* 0065, a member of the IclR transcriptional factor family. *J Biol Chem* **277**: 19183–19190.

Zhu J & Winans SC (1999) Autoinducer binding by the quorum-sensing regulator TraR increases affinity for target promoters *in vitro* and decreases TraR turnover rates in whole cells. *Proc Natl Acad Sci USA* **96**: 4832–4837.

Capítulo 3

A general profile for the MerR family of transcriptional regulators constructed using the semi-automated Provalidator tool

Environmental Microbiology Reports

Environmental Microbiology Reports (2009) doi:10.1111/j.1758-2229.2009.00067.x

A general profile for the MerR family of transcriptional regulators constructed using the semi-automated Provalidator tool

Antonio J. Molina-Henares, Patricia Godoy, Estrella Duque and Juan L. Ramos*
Consejo Superior de Investigaciones Científicas, Estación Experimental del Zaidín, Department of Environmental Protection, E-18008 Granada, Spain.

Summary

Provalidator is a web-based tool that facilitates the design and validation of generalized profiles of protein families in prokaryotes. This tool combines the nearly full automation of profile building with a search for family members in all available databases. The tool is useful for assigning a given protein to a specific family, and is also useful for genome mining in annotated prokaryotic genomes. The tool is freely available at http://www.bactregulators.org. As proof of concept we constructed a profile that best defines the MerR family of transcriptional regulators. The profile created includes functional residues that are part of the helix–turn–helix DNA binding domain and accessory elements defined as wings 1 and 2, suggesting that members of the MerR family of regulators may exhibit conserved 3D structure in the region that defines the family profile. The profile defined for MerR was used to search for members of this family in the Swiss-Prot and TrEMBL databases, and also to identify members of the family in the genome of *Pseudomonas putida*. One of these identified regulators was found to be involved in zinc tolerance, showing the usefulness of identifying family members and assigning phenotypes.

Introduction

The analysis of protein sequences grouped according to function makes it possible to identify conserved domains in proteins, such as enzyme catalytic sites, cofactor binding sites, small molecule ligand domains, DNA binding domains and many others (Corpet *et al.*, 1999; Bradley *et al.*, 2008). Proteins or protein domains belonging to a particular family often share functional attributes, and therefore the grouping of proteins has been used in turn to characterize them at the functional level (Holm and Sander, 1996; Tatusov *et al.*, 2001).

Received 27 May, 2009; accepted 8 July, 2009. *For correspondence. E-mail: jlramos@eez.csic.es; Tel. (−34) 958 181608; Fax (+34) 958 135740.

Several approaches are available to define domains and protein families. The most frequently used of these approaches are Hidden Markov Models (HMM) and conventional profiles (Bateman *et al.*, 2004; Hulo *et al.*, 2006). Conventional profiles have the advantage of being easier to manage (Bucher *et al.*, 1996; Tobes and Ramos, 2005) than HMM and although the latter can detect some subtle patterns, this advantage is not relevant to define the families. Profiles are not necessarily based on small regions with high sequence similarity, but rather they attempt to characterize a protein family (or domain) through the consideration of its entire length (Hulo *et al.*, 2006). Our research group has constructed conventional profiles to define several families of transcriptional regulators, e.g. the AraC/XylS family of positive regulators (Gallegos *et al.*, 1997), the TetR family of repressors (Ramos *et al.*, 2005), and the IclR family, which includes activators and repressors (Krell *et al.*, 2006; Molina-Henares *et al.*, 2006). Other researchers have constructed profiles that define the GntR family of regulators (Rigali *et al.*, 2002), while a number of other profiles are available at the PROSITE database (Hulo *et al.*, 2006 and http://www.expasy.ch/prosite/). Information about the AraC/XylS, TetR and IclR families is available through the BacTregulators database, which aims to collect and integrate information about proteins belonging to defined families of transcriptional regulators in prokaryotes (Martínez-Bueno *et al.*, 2002).

An added value of the use of profiles for defining protein families is that the domain that best identifies the family often corresponds to highly conserved 3D structures. For example, the stretch that best defines the profile of the TetR family is made up of 47 amino acid residues that correspond to the helix–turn–helix DNA binding motif and adjacent regions. Analysis of a set of 2353 non-redundant proteins belonging to the TetR family by screening genome and protein databases with the TetR profile identified two groups of residues (Ramos *et al.*, 2005). One of these groups includes highly conserved residues that are

involved in the proper orientation of the helix–turn–helix motif and hence seem to play a structural role. The other set of residues is less conserved, and are involved in establishing contacts with the phosphate backbone and with target bases in the operator. In the case of the AraC/XylS protein family, the profile defining the family was used to propose the existence of two helix–turn–helix DNA biding domains in these proteins, which was later confirmed when AraC/XylS family members, MarA and Rob, were co-crystallized with target DNA sequences, and also when they were analysed by nuclear magnetic resonance (Kwon *et al.*, 2000; Dangi *et al.*, 2001). Therefore, general profiles are not only useful tools to identify new members of protein families, but also to define putative functional domains.

Although there are several ways to construct profiles, a general limitation of all methods is that they require a dedicated manual process. In light of the rapidly increasing volume of raw genomics data, automated or semi-automated programmes are needed to classify proteins into families, which will in turn facilitate their further biochemical and genetic characterization. Below we describe the construction of a general profile for the MerR family of transcriptional regulator using Provalidator, a tool that combines nearly full automation of profile construction and validation.

Provalidator: a tool for semi-automatic construction of profiles

The construction of a general profile begins with the selection of the sequences that are part of the profile seed, which can be chosen based on literature searches, or can be selected through the retrieval of sequences from databases, for which Provalidator provides the user with tools that automate this task (http://www.bactregulators.com). The selection of sequences retrieved from databases for the profile seed involves three well-defined steps: (i) clustering of the sequences using BLASTCLUST; (ii) filtering the sequences to avoid the inclusion of proteins whose description includes words such as *hypothetical, probable, putative, fragment*, and to give priority to Swiss-Prot sequences versus TrEMBL; and (iii) excluding sequences with extreme differences in length compared with the rest.

The N sequences are then clustered into groups using BLASTCLUST, and a representative sequence of each group is selected and aligned using CLUSTALW (http://align.genome.jp). Alignment of the entire sequence of this set of proteins is used to identify the most conserved regions, which often correspond to functional domains such as motifs involved in DNA binding, catalytic sites, metal binding sites, etc (see also *Supporting information*).

The final alignment is then used as a seed for the construction of a conventional profile using the pfmake program, which is part of the PFTOOLS package of programs, available at the Swiss Institute of Bioinformatics (http://www.isrec.isb-sib.ch/ftp-server/pftools) (Bucher *et al.*, 1996). Thereafter, the profile is calibrated and validated (see *Supporting information*). To verify the quality of the profile in regard to specificity (false positives) and sensitivity (false negatives), Provalidator uses the InterPro database because it contains a large amount of information and allows automatic comparisons that facilitate the validation of the constructed profile (Zbobnov and Apweiler, 2001). Proteins identified by Provalidator that match those assigned by InterPro to the family of interest are considered true positives. False positives are proteins assigned to the family by Provalidator but not assigned to the family by InterPro. Potential false positive proteins are analysed in detail manually by curators in order to ensure that they belong to the family. Provalidator considers as potential false negatives those proteins assigned to the family by InterPro but whose score is below 8.5, a previously established value that is commonly used for the construction of profiles.

The profile defined for a given family can then be used to search for members of this family in the Swiss-Prot and TrEMBL databases, and also to screen the set of complete and incomplete microbial genomes available in NCBI and GOLD (Siezen and Wilson, 2008). This is an added-value of the profile because it allows the researcher to learn about the set of sequences that belong to the group defined by the profiles.

A profile for the MerR family of transcriptional regulators

As a proof of concept we generated a general profile for MerR family, a set of transcriptional regulators typically involved in the control of genes that confer cells with tolerance to toxic metals, and some members seem to be also involved in responses to oxidative stress and certain toxic drugs (Brown *et al.*, 2003; Watanabe *et al.*, 2007). Heavy metals are toxic for all living organisms, and among these compounds mercury is one of the most noxious (Van der Lelie *et al.*, 1997; Grosse *et al.*, 1999). Microorganisms have developed several systems to remove mercury, the most frequent being reduction of Hg^{2+} to Hg^{0}, which volatilizes. The best-studied regulator that controls the expression of mercury reductase is the archetypal MerR protein (Barkay *et al.*, 2003), which regulates both its own synthesis and expression of the polycistronic *mer* operon (reviewed in Brown *et al.*, 2003; Hobman *et al.*, 2005; Pennella and Giedroc, 2005). MerR is a dimer in solution that binds to a symmetric operator DNA sequence within the promoter region. Normally transcription is repressed because the promoter DNA is bent

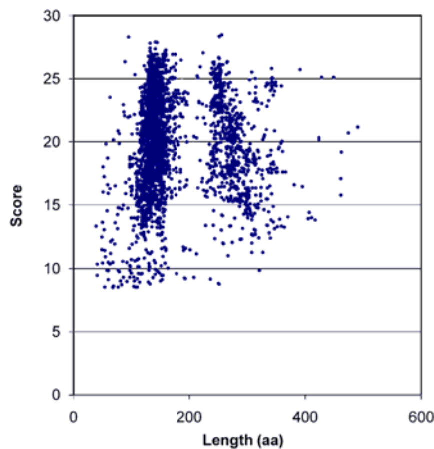


Fig. 1. Two groups of MerR family members based on length. The MerR profile was used to screen all available databases. Members of the MerR family received a Z-score value between 8.5 and 29. Two groups were defined based on length.

by the apo-MerR homodimer bound to this motif. When the MerR dimer binds a mercuric ion, the Hg^{2+}-MerR dimer complex causes allosteric underwinding of the DNA at the operator/promoter site, thereby realigning the –35 and –10 sequences of the promoter. This in turn allows the binding of σ^{70} RNA polymerase to the promoter sequences and the initiation of transcription (reviewed by Ansari *et al*., 1995; Hobman *et al*., 2005; Hobman, 2007).

Using Provalidator, we generated a profile to identify proteins that belong to the MerR family of transcriptional regulators. To construct the MerR profile we scanned SwissProt and retrieved all 1022 sequences which contained the word MerR. These sequences were aligned and clusters were established, which reduced the number of sequences to 309 non-redundant sequences. For seed generation, the programme excluded proteins annotated as truncated, fragment, hypothetical, probable and putative. This reduced the number of sequences for the seed to 164, which were re-aligned with the Multialignment programme.

Figure S1 shows the alignment of the seed sequences. Figure 1 shows that the proteins selected for seed construction fell into two groups; one in which the set of sequences was around 140 amino acids long, and another comprising longer proteins of around 270 amino acid residues. Alignment of all proteins revealed that regardless of their size, the set of conserved amino acids was a stretch of around 70 residues at the N-terminal end of the protein (see Fig. 2 for a limited number of sequences). This region included the helix–turn–helix DNA binding domain (positions 36 to 57 in the archetypal Tn*501* MerR protein), and an adjacent coiled-coil region. The functional role of this segment and the corresponding

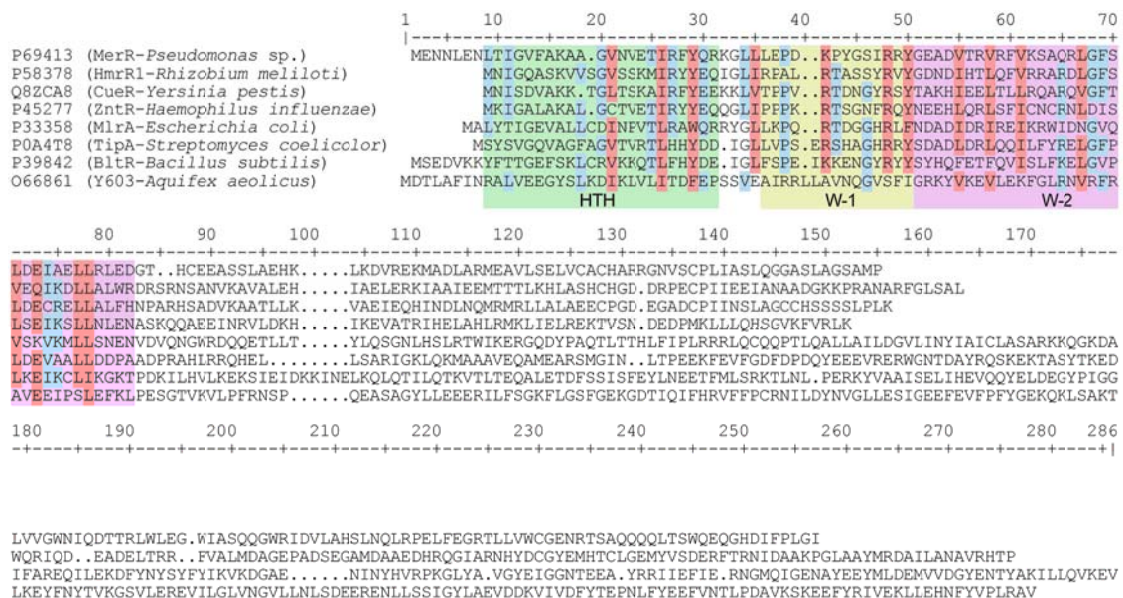


Fig. 2. Multialignment of the conserved region of the MerR family members used to create the generalized MerR family profile. The top four sequences are proteins that belong to group 1 (140 amino acids long, on average), while the other four are proteins that belong to the group of long (around 270 residues) MerR family members. Red residues were present in more than 80% of the seed proteins, and blue residues were present in more than 60% of the seed proteins. The green background indicates residues within the helix–turn–helix DNA binding domain, while the yellow and orange backgrounds indicate residues within the accessory elements wing 1 and wing 2 respectively.

3D structure is available for BmrR, CueR, ZntR and SoxR (Outten *et al.*, 1999; Heldwein and Brennan, 2001; Changela *et al.*, 2003; Watanabe *et al.*, 2007). Given the availability of experimental and structural data supporting the functional role of the selected motif, we decided to construct the profile based on this conserved region. On the other hand, in CoaR, a member of MerR, the extended C-terminal end has been shown to act as a 'sensing' region (Rutherford *et al.*, 1999).

To validate the profile we ran the profile against all entries in InterPro to search for potential false positives (Table S1) and potential false negatives (Table S2). Proteins identified by Provalidator and what matched those assigned by InterPro to the family were considered true positives. False positives were proteins assigned to the family by Provalidator, but not assigned to the family by InterPro. Potential false positive proteins were then manually analysed in detail in order to ensure that they did in fact belong to the family. Provalidator considers potential false negatives those proteins assigned to the family by InterPro, but whose score is below 8.5. We found 164 potential false positive and 264 potential false negative proteins, which we analysed manually. Alignment of each of the potential false positives against the proteins in the profile seed revealed that all 164 of these proteins exhibited a high degree of identity within the 70-residue conserved stretch, which was considered confirmation that they were in fact MerR family members. As such, the potential false positives seem to have resulted from a delay in the updating of InterPro. Regarding the 264 potential false negatives, we found that most of them (~ 90%) belong in the MerR family based on HMM profiles available in the Pfam database (http://pfam.sanger.ac.uk/), and among this set, a few proteins are putative enzymes, such as GlcNAc transferase or citrate synthase (Table S2). The remaining potential false negatives were, in general, uncharacterized proteins whose annotation included words such as 'Replication protein', 'Plasmid partition protein', or 'Integrase/Resolvase' (Table S2). When each of the false negative proteins were aligned with the seed proteins used to define MerR family members, we found that they exhibited very little or no similarity. We therefore considered the original assignment of a number of proteins to the MerR family based on the HMM profile at PFAM to be incorrect.

Once the MerR family profile had been validated we used it to identify MerR family members in SwissProt/TrEMBL databases, which included complete and uncomplete genomes. Close to 4500 (31 August 2008) MerR family proteins were identified by the profile. We found no correlation between the genome size and the number of MerR family members in a given genome; however, we found that MerR family members were well represented in soil and water microbes living in niches exposed to environmental fluctuations. For instance, MerR family members were well represented in genomes of bacteria of the genus *Streptomyces*, *Bacillus*, *Salinispora*, *Burkholderia*, *Frankia* and *Pseudomonas* among others (see Table S3). In contrast, microorganisms living in environments less susceptible to environmental changes, such as *Helicobacter*, *Xylella*, *Neisseria*, *Rickettsia* exhibited a number of MerR family members. It is worth noting differences in the number of MerR family members between species of the genera *Clostridium* and *Synechococcus*. Each of the genomes of *C. phytofermentants*, *C. beijerinckii* and *C. difficile* contained more than 10 MerR family members, while the genome of *C. perfringens* only encoded a single MerR family member. In the genus *Synechococcus* we found a single gene encoding a MerR family member in *S. elongatus* PCC7942, whereas 22 MerR family members were found in the genome of *Synechococcus* sp. JA-3-3Ab. Therefore, MerR family members are well distributed in microorganisms living under changing environmental conditions, but no correlation between genome size and number of encoded MerR family members was found.

It is known that many members of the MerR family are involved in resistance to a number of toxic metals (Grosse *et al.*, 2004). The wild-type *Pseudomonas putida* KT2440 strain grows in the presence of a number of metals, namely Ag (5 μM), Cd (1.25 mM), Co (0.625 mM), Cu (6.25 mM), Hg (3 μM), Ni (2.5 mM) and Zn (1.25 mM). However, no information on potential regulators responding to these metals is available. In the genome of *P. putida* 10 open reading frames (ORFs) have been annotated as potential members of the MerR family. Among these regulators is PP4273 (SwissProt Q88F26) for which a mini-Tn*5* knockout mutant was available in the *P. putida* mutant collection (Duque *et al.*, 2007), and whose potential involvement in metal resistance was unknown. PP4273 was assigned to the MerR family by PFAM (PF00376) too. We carried out Minimal Inhibitory Concentration (MIC) assays using the mutant strain against the metals cited above to which the parental strain was resistant. We found that the MIC90 of the mutant was only different from the parental strain in that the mutant tolerated up to 5 mM Zn, i.e. fourfold more than the parental strain, suggesting that PP4273 may sense Zn in *P. putida*. Next we examined the genomic regions adjacent to PP4273 and found that the PP4272 ORF, which encodes a protein of unknown function, was transcribed divergently with respect to PP4273 and that only a short intergenic region exists between the genes suggesting that their promoters may overlap. However, while PP4273 seems to be involved in Zn tolerance, whether or not the adjacent PP4272 gene is involved in tolerance to Zn is currently under investigation.

In summary, our work shows that automation is feasible for the construction of profiles that identify members of a specific family of proteins with high sensitivity. Provalidator software shows promise for use as a tool for the classification of proteins within families, and for the facilitation of new insights into the biological functions of such proteins. In addition, Provalidator can be used to efficiently derive a consensus sequence in the region that defines the family protein profile, and to consolidate existing knowledge in database, while providing another way in which to detect and thereby correct erroneously annotated proteins. As proof of concept we developed a profile for the MerR family of regulators. The derived profile for this family included residues present in functional domains such as the helix–turn–helix DNA binding domain and the accessory elements defined as wings 1 and 2. We previously showed that members of the TetR and AraC/XylS families exhibited conserved 3D structure in the region that defines the family profile. In the case of TetR, these regions included the helix–turn–helix DNA binding domain and an extended alpha helix that connects this domain to the dimerization and effector binding domain (Hinrichs *et al.*, 1994; Schumacher *et al.*, 2001). We therefore suggest that members of the MerR family of regulators will also exhibit conserved 3D structure in the region that defines the family profile. We have also shown that *in silico* information and putative functions for proteins acquired by Provalidator analysis can be quickly confirmed using wet testing, as we have carried out for the *P. putida* gene product of ORF PP4273, which has been shown to be involved in tolerance to zinc.

Acknowledgements

Work in Ramos' laboratory was supported by PSYSMO (Ref. GEN2006-27750-C5-5-E/SYS) and Consolider-Ingenio-2010 (Ref. CSD2007-00005). We thank M.M. Fandila and C. Lorente for secretarial assistance, and K. Shashok and B. Pakuts for improving the English in the manuscript.

References

Ansari, A.Z., Bradner, J.E., and O'Halloran, T.V. (1995) DNA bend modulation in a repressor-to-activator switching mechanism. *Nature* **374:** 371–375.

Barkay, T., Miller, S.M., and Summers, A.O. (2003) Bacterial mercury resistance from atoms to ecosystems. *FEMS Microbiol Rev* **27:** 355–384.

Bateman, A., Coin, L.R., Durbin, R.D., Finn, V., Hollich, S., Griffith-Jones, A., *et al.* (2004) The Pfam protein families database. *Nucleic Acids Res* **32:** D138–D141.

Bradley, M.J., Chivers, P.T., and Baker, N.A. (2008) Molecular dynamics simulation of the Escherichia coli, NikR: equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J Mol Biol* **378:** 1155–1173.

Brown, N.L., Stoyanov, J.V., Kidd, S.P., and Hobman, J.L. (2003) The MerR family of transcriptional regulators. *FEMS Microbiol Rev* **27:** 145–163.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem* **20:** 3–24.

Changela, A., Chen, K., Xue, Y., Holschen, J., Outten, C.E., O'Halloran, T.V., and Mondragón, A. (2003) Molecular basis of metal-ion selectivity and zeptomolar sensitivity by CueR. *Science* **301:** 1383–1387.

Corpet, F., Gouzy, J., and Kahn, D. (1999) Recent improvements of the PRODOM database of protein domain families. *Nucleic Acids Res* **27:** 263–267.

Dangi, B., Delupessey, P., Martin, R.G., Rosner, J.L., Louis, J.M., and Gronenborn, A.M. (2001) Structure and dynamics of MarA-DNA complexes: an NMR investigation. *J Mol Biol* **314:** 113–127.

Duque, E., Molina-Henares, A.J., de la Torre, J., Molina-Henares, M.A., del Castillo, T., Lam, J., and Ramos, J.L. (2007) Towards a genome-wide mutant library of *Pseudomonas putida* strains kt2440. In *Pseudomonas*, Vol. V. Ramos, J.L. (ed.). London, UK: Kluwer Academic Press, Chap 8, pp. 227–254.

Gallegos, M.T., Scheleif, R., Bairoch, A., Hofmann, K., and Ramos, J.L. (1997) AraC/XylS family of transcriptional regulators. *Microbiol Mol Biol Rev* **61:** 393–410.

Grosse, C., Grass, G., Anton, A., Franke, S., Navarrete-Santos, A., Lawley, B., *et al.* (1999) Transcriptional organization of the *czc* heavy metal homeostasis determinant from *Alcaligenes eutrophus*. *J Bacteriol* **181:** 2385–2393.

Grosse, C., Anton, A., Hoffmann, T., Franke, S., Schleuder, G., and Nies, D.H. (2004) Identification of a regulatory pathway that controls the heavy-metal resistance system Czc via promoter *czcNP* in *Ralstonia metallidurans*. *Arch Microbiol* **182:** 109–118.

Heldwein, E.E., and Brennan, R.G. (2001) Crystal structure of the transcription activator BmrR bound to DNA and a drug. *Nature* **409:** 378–382.

Hinrichs, W., Kisker, C., Düvel, M., Müller, A., Tovar, K., Hillen, W., and Saenger, W. (1994) Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. *Science* **264:** 418–420.

Hobman, J.L. (2007) MerR family transcription activators: similar designs, different specificities. *Mol Microbiol* **63:** 1275–1278.

Hobman, J.L., Wilkie, J., and Brown, N.L. (2005) A design for life: prokaryotic metal-binding MerR family regulators. *Biometals* **18:** 429–436.

Holm, L., and Sander, C. (1996) Mapping the protein universe. *Science* **273:** 595–602.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., de Castro, E., Lagendijk-Genevana, P.S., *et al.* (2006) The PROSITE database. *Nucleic Acids Res* **34:** D227–D230.

Krell, T., Molina-Henares, A.J., and Ramos, J.L. (2006) The IclR family of transcriptional activators and repressors can be defined by a single profile. *Protein Sci* **15:** 1207–1213.

Kwon, H.I.M., Bennik, H.J., Demple, B., and Ellenberger, T. (2000) Crystal structure of the *Escherichia coli* Rob transcription factor in complex with DNA. *Nat Struct Biol* **7:** 424–430.

Martínez-Bueno, M., Tobes, R., Rey, M., and Ramos, J.L. (2002) Detection of multiple extractoplasmatic function (ECF) sigma factors in the genome of *Pseudomonas putida* KT2440 and their counterparts in *Pseudomonas aeruginosa* PAO1. *Environ Microbiol* **4:** 842–855.

Molina-Henares, A.J., Krell, T., Guazzaroni, M.E., Segura, A., and Ramos, J.L. (2006) Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors. *FEMS Microbiol Rev* **30:** 157–186.

Outten, C.E., Outten, W.F., and O'Halloran, T.V. (1999) DNA Distortion mechanism for transcriptional activation by ZntR, a Zn(II)-responsive MerR homologue in *Escherichia coli*. *J Biol Chem* **274:** 37517–37524.

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

Ramos, J.L., Martínez-Bueno, M., Molina-Henares, A.J., Terán, W., Watanabe, K., Zhang, X., *et al.* (2005) The TetR family of transcriptional repressors. *Microbiol Mol Bio Rev* **69:** 326–356.

Rigali, S., Derouaux, A., Giannotta, F., and Dusart, J. (2002) Subdivision of the Helix-turn-Helix GntR family of bacterial regulators in the FadR, HutC, and YtrA subfamilies. *J Biol Chem* **277:** 12507–12545.

Rutherford, J.C., Cavet, J.S., and Robinson, N.J. (1999) Cobalt-dependent transcriptional switching by a dual-effector MerR-like protein regulates a cobalt-exporting variant CPx-type ATPase. *J Biol Chem* **274:** 25827–25832.

Schumacher, M.A., Miller, M.C., Grkovic, S., Brown, M.H., Skurray, R.A., and Brennan, R.G. (2001) Structural mechanisms of QacR induction and multidrug recognition. *Science* **294:** 2158–2163.

Siezen, R.J., and Wilson, G. (2008) Unpublished but public microbial genomes with biotechnological relevance. *Microb Biotechnol* **1:** 202–207.

Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., *et al.* (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* **29:** 22–28.

Tobes, R., and Ramos, J.L. (2005) REP code: defining bacterial identity in extragenic space. *Environ Microbiol* **7:** 225–228.

Van der Lelie, D., Schwuchow, T., Schwidetzky, U., Wuertz, S., Baeyens, W., Mergeay, M., and Nies, D.H. (1997) Two-component regulatory system involved in transcriptional control of heavy-metal homoeostasis in *Alcaligenes eutrophus*. *Mol Microbiol* **23:** 493–503.

Watanabe, S., Kita, A., Kobayashi, K., and Miki, K. (2007) Crystal structure of the [2Fe-2S] oxidative-stress sensor SoxR bound to DNA. *Proc Natl Acad Sci USA* **105:** 4121–4126.

Zbobnov, E.M., and Apweiler, R. (2001) InterProScan an integration platform for the signature-recognition methods in InterPro. *Bioinformatics* **17:** 847–848.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. The multialignment is intended to show the different length of the two length groups of MerR family members and the conserved residues in the common N-terminal part of the protein that are highlighted in blue.

Table S1. Potential false positives. Sequences were ordered by score value. The Swiss Prot accession number and the given name are indicated.

Table S2. Potential false negatives. Sequences are ordered by score value. The Swiss Prot accession number and the given name are indicated.

Table S3. MerR family members in different available full and partial genomes. The size of the full genome or part of it is provided together with the number of MerR family members identified by Provalidator.

Appendix S1. Seed sequences for profile construction.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

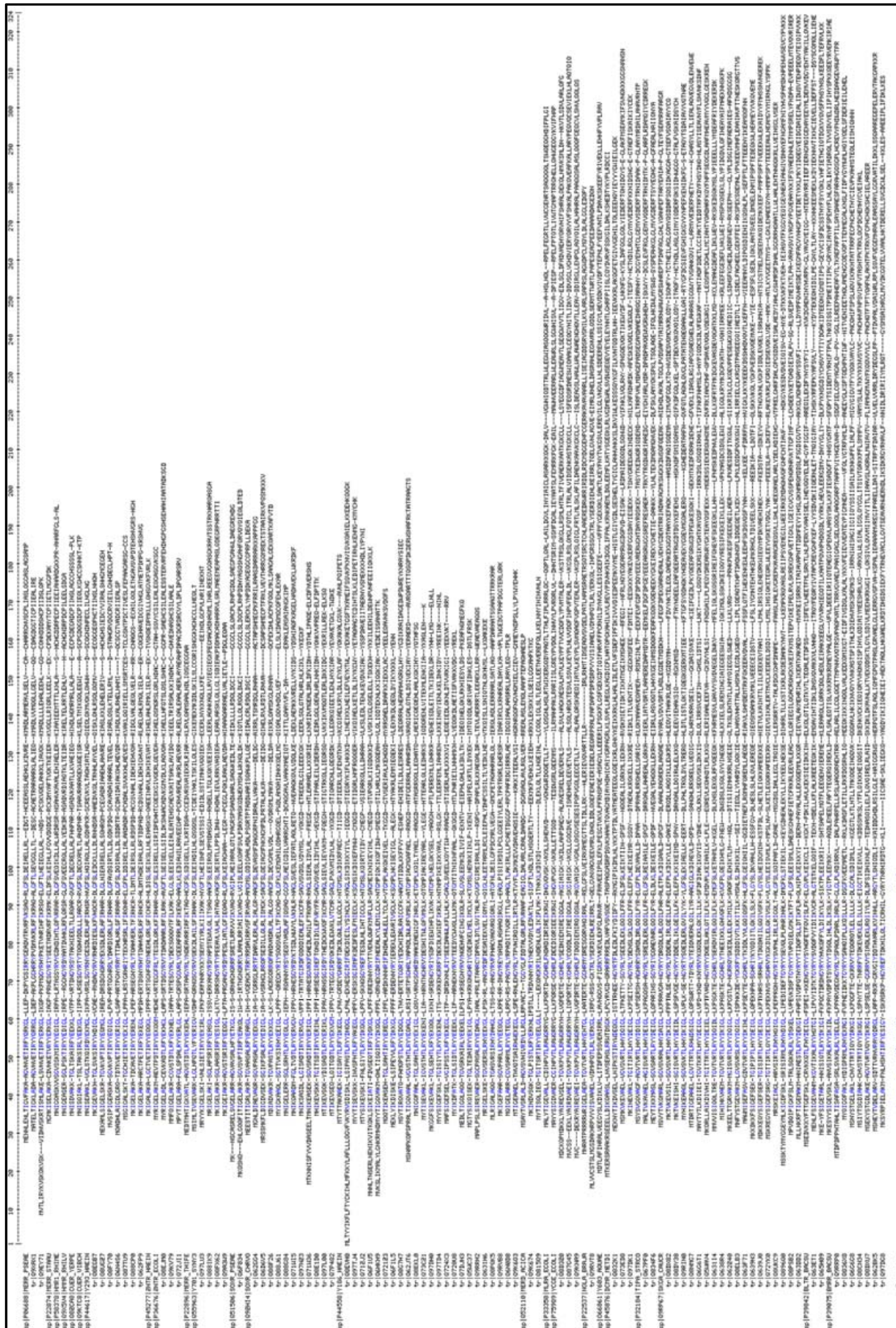


Fig. S1. The multialignment is intended to show the different length of the two length groups of MerR family members and the conserved residues in the common N-terminal part of the protein that are highlighted in blue.

Table S1. Potential false positives. Sequences were ordered by score value. The Swiss Prot accession number and the given name are indicated.

Score	AC	Name
26.107	A3HL45	Putative transcriptional regulator, MerR family
24.094	A3HF85	Putative transcriptional regulator, MerR family
23.017	Q3CHE9	Regulatory protein, MerR
20.882	A3HBL8	Putative transcriptional regulator, MerR family
20.312	A3HC76	Putative transcriptional regulator, MerR family
19.031	A3HQ97	Putative transcriptional regulator, MerR family
18.950	A3HIJ2	Putative transcriptional regulator, MerR family
17.242	A3HHF5	Putative transcriptional regulator, MerR family
17.018	A3HG31	Putative transcriptional regulator, MerR family
14.700	A9W4E6	Regulatory protein MerR
13.867	Q1VGQ7	Transcription regulator
11.854	A7M7D8	Putative uncharacterized protein
11.772	A0RLV4	Putative uncharacterized protein
11.732	Q4V0Z6	Putative uncharacterized protein
11.691	Q45444	DNA-binding protein
11.691	Q7BUV5	Putative uncharacterized protein
11.691	Q84E49	Orf6
11.589	Q3YFI4	Putative DNA-binding protein
11.589	Q45449	Putative DNA-binding protein
11.589	Q45453	Putative DNA-binding protein
11.589	Q9X3Z8	DNA-binding protein Ptr
11.549	A7HA58	Putative transcriptional regulator, MerR family
11.528	A4INV5	Transcriptional regulator MerR family
11.447	Q5DUS4	Putative uncharacterized protein
11.447	Q5MCL1	Putative DNA-binding protein
11.447	Q4MIS3	PXO2-68
11.427	A9VVU7	Putative uncharacterized protein
11.366	Q3CGJ5	Regulatory protein, MerR:Resolvase, N-terminal
11.366	Q0ESX3	Resolvase-like
11.305	A9VVC4	Putative DNA-binding protein
11.264	A8VNV6	Putative transcriptional regulator, MerR family
11.264	Q2IKR4	Putative transcriptional regulator, MerR family
11.183	Q5LK76	Putative uncharacterized protein
11.183	Q3YN23	Putative uncharacterized protein
11.101	A9GVK2	Putative RepA replication protein
10.817	Q08RL5	Putative uncharacterized protein
10.715	Q7MS23	Putative uncharacterized protein
10.695	Q4V0T1	Possible DNA-binding protein
10.430	A9G305	Transcriptional regulator, MerR family
10.288	A8YWP7	Putative uncharacterized protein
10.146	Q1D8A0	Putative uncharacterized protein
10.085	Q5FMN5	Putative uncharacterized protein
10.044	Q70JB3	Putative DNA binding protein
10.044	Q70JC4	Putative DNA binding protein
9.942	A9EE06	RepA partitioning protein/ATPase, ParA type
9.942	A3T2E5	RepA partitioning protein/ATPase, ParA type
9.922	A9EER8	RepA partitioning protein/ATPase, ParA type
9.902	A8FID2	Possible chromosome-anchoring protein RacA
9.881	Q8KVZ6	RC194
9.861	A6KYA9	Putative uncharacterized protein
9.861	A7V4U2	Putative uncharacterized protein
9.861	A4ITZ1	Putative DNA binding protein
9.861	A7GQL4	Putative uncharacterized protein
9.841	A4C8K0	Putative uncharacterized protein
9.800	Q5SGM0	Putative uncharacterized protein TTHC014
9.780	A9JAE1	Putative transcriptional regulator, MerR family
9.759	A8UD55	Putative uncharacterized protein
9.658	A3XEI7	RepA partitioning protein/ATPase, ParA type
9.658	A3T1V2	RepA partitioning protein/ATPase, ParA type
9.637	Q67T11	Putative uncharacterized protein
9.597	A9E9N7	Putative uncharacterized protein
9.556	A7W9E6	Binding-protein-dependent transport systems inner membrane component precursor
9.556	A6NU69	Putative uncharacterized protein
9.536	A8EWB5	Transcriptional regulator, AraC family
9.475	Q8YR38	Alr3614 protein
9.434	O67508	Putative uncharacterized protein
9.434	A9FIW2	Citrate synthase (EC 2.3.3.1)
9.414	A4XEU0	Cobyrinic acid a,c-diamide synthase
9.373	Q3ETK6	Putative uncharacterized protein
9.170	A5V0H1	Putative transcriptional regulator, MerR family
9.170	Q04CI1	Predicted transcriptional regulator
9.129	A6G005	Putative uncharacterized protein
9.109	A0FJX8	Putative DNA-binding protein
9.088	A3W5S0	ISPsy20, transposase IstA
9.088	P45870	Chromosome-anchoring protein racA
9.088	A3W5F1	ISPsy20, transposase IstA
9.027	A9WI72	DNA binding domain protein, excisionase family
9.027	A6FSG1	Putative uncharacterized protein
9.027	A3SWL5	Putative uncharacterized protein
9.027	A3SZF7	Putative uncharacterized protein
9.007	A9DW06	Putative uncharacterized protein
8.987	Q82CK9	Putative MarR-family transcriptional regulator
8.987	Q74LS4	Putative uncharacterized protein
8.966	Q8TRW1	Transposase
8.966	Q1N8A9	Probable replication protein A
8.946	A3IT96	Putative uncharacterized protein
8.946	A8GNW4	Fic
8.926	Q5DUS2	Putative uncharacterized protein
8.926	A9HGM8	Putative transposase
8.926	A7Z9S2	YwkC
8.905	A9GSN4	Putative uncharacterized protein
8.905	Q13H49	Putative integrase
8.905	Q13IZ1	Putative integrase
8.885	A4FHU3	Transcriptional regulator, GntR family
8.885	Q8PTM1	Transposase
8.865	Q0KJA0	Putative partitioning protein ParA
8.844	Q8TH46	Transposase
8.844	Q8TH42	Transposase
8.804	Q58AT4	Partitioning protein (ATPase, ParA family)
8.804	A5VH06	Cobyrinic acid a,c-diamide synthase
8.804	Q046V2	Predicted transcriptional regulator
8.783	Q9L1I9	Putative uncharacterized protein SCO2728
8.763	Q8TIG5	Transposase
8.763	O85859	RepA replication protein
8.763	A4XE00	Cobyrinic acid a,c-diamide synthase
8.743	A7JXA6	Possible DNA-binding protein
8.722	A9QNZ2	SopA
8.682	A9G9Y0	Integrase, catalytic region
8.682	A9GF91	Integrase, catalytic region
8.682	P45916	Uncharacterized protein yqaT (ORF50)
8.682	A9GMZ1	Integrase, catalytic region
8.682	A9FRI4	Integrase, catalytic region (Molybdate ABC transporter permease protein)
8.682	A9G302	Integrase, catalytic region
8.682	A5G6S1	Putative uncharacterized protein
8.682	A9G4W6	Integrase, catalytic region
8.682	A9G863	Integrase, catalytic region
8.641	A7NQ80	Putative transcriptional regulator, MerR family
8.641	A5N5P4	Putative uncharacterized protein
8.621	A3WSR1	Putative uncharacterized protein
8.621	Q317G9	Putative transcriptional regulator, MerR family
8.600	A7WBH6	Putative uncharacterized protein
8.600	A9VYU7	Putative uncharacterized protein
8.539	Q8A8N7	Putative DNA binding protein
8.539	A7AB94	Putative uncharacterized protein
8.539	Q4EBH5	Transcriptional regulator, merR family (Fragment)
8.539	A7ALF0	Putative uncharacterized protein
8.539	A7ALY4	Putative uncharacterized protein

Table S2. Potential false negatives. Sequences are ordered by score value. The Swiss Prot accession number and the given name are indicated.

Score	AC	Name
8.478	Q8KY11	RepA - Rhizobium etli.
8.478	Q2K288	Plasmid partitioning protein RepAb - Rhizobium etli (strain CFN 42 / ATCC 51251).
8.478	A5WY09	RepA' - Agrobacterium tumefaciens.
8.478	A4XFW4	Regulatory protein, MerR - Caldicellulosiruptor saccharolyticus (strain ATCC 43494 / DSM 8903).
8.458	Q8XLF2	Putative uncharacterized protein CPE1090 - Clostridium perfringens.
8.458	Q47L88	Excisionase/Xis, DNA-binding - Thermobifida fusca (strain YX).
8.458	Q2JZZ2	Plasmid partitioning protein RepAe - Rhizobium etli (strain CFN 42 / ATCC 51251).
8.458	Q0TRE8	Transcriptional regulator, MerR family - Clostridium perfringens (strain ATCC 13124 / NCTC 8237 / Type A).
8.458	A0LWK3	DNA binding domain, excisionase family - Acidothermus cellulolyticus (strain ATCC 43068 / 11B).
8.438	Q0SUE2	Resolvase, N terminal domain family - Clostridium perfringens (strain SM101 / Type A).
8.438	Q0RM36	Putative site-specific integrase-resolvase - Frankia alni (strain ACN14a).
8.417	A3ZLC0	Probable O-linked GlcNAc transferase - Blastopirellula marina DSM 3645.
8.377	Q4V235	Putative uncharacterized protein - Bacillus cereus (strain ZK / E33L).
8.356	Q0FMK7	ATPase, ParA type - Roseovarius sp. HTCC2601.
8.275	Q5KVB2	Chromosome-anchoring protein racA - Geobacillus kaustophilus.
8.275	A6X7L5	Cobyrinic acid ac-diamide synthase - Ochrobactrum anthropi (strain ATCC 49188 / DSM 6882 / NCTC 12168).
8.255	Q47LE0	Excisionase/Xis, DNA-binding - Thermobifida fusca (strain YX).
8.255	Q46R44	Excisionase/Xis, DNA-binding - Ralstonia eutropha (strain JMP134) (Alcaligenes eutrophus).
8.255	Q2JY77	Plasmid partitioning protein RepAf1 - Rhizobium etli (strain CFN 42 / ATCC 51251).
8.255	A7GBR3	Conserved domain protein - Clostridium botulinum (strain Langeland / NCTC 10281 / Type F).
8.234	A4KVA6	Replication protein A - Rhizobium meliloti (Sinorhizobium meliloti).
8.234	A0K1F4	TOBE domain protein - Arthrobacter sp. (strain FB24).
8.214	Q82C29	Putative MerR-family transcriptional regulator - Streptomyces avermitilis.
8.214	Q1M9T7	Putative replication protein - Rhizobium leguminosarum bv. viciae (strain 3841).
8.214	A8YJK4	Genome sequencing data, contig C321 - Microcystis aeruginosa PCC 7806.
8.194	B0J8W5	Cobyrinic acid ac-diamide synthase - Rhizobium leguminosarum bv. trifolii WSM1325.
8.194	A7A2S5	Putative uncharacterized protein - Bifidobacterium adolescentis L2-32.
8.173	Q7UXZ0	Probable O-linked GlcNAc transferase - Rhodopirellula baltica.
8.173	A7C1K0	ISSoc2, resolvase - Beggiatoa sp. PS.
8.153	Q7D3E2	Related to citrate synthase - Agrobacterium tumefaciens (strain C58 / ATCC 33970).
8.153	Q7BLR4	RepA - Agrobacterium tumefaciens.
8.153	Q52225	RepA protein - Plasmid pTiB6S3.
8.153	A6UH66	Cobyrinic acid ac-diamide synthase - Sinorhizobium medicae (strain WSM419) (Ensifer medicae).
8.112	Q4MUT1	DNA binding domain, excisionase family protein - Bacillus cereus G9241.
8.112	Q2J5V8	TOBE - Frankia sp. (strain CcI3).
8.112	Q1M736	Putative replication protein A - Rhizobium leguminosarum bv. viciae (strain 3841).
8.112	A5EU56	Putative replication protein A - Bradyrhizobium sp. (strain BTAi1 / ATCC BAA-1182).
8.072	A4IST0	Putative uncharacterized protein - Geobacillus thermodenitrificans (strain NG80-2).
8.031	Q0AK02	Transcriptional regulator, MerR family protein - Nitrosomonas eutropha (strain C71).
8.011	Q9HKN4	Putative uncharacterized protein Ta0563 - Thermoplasma acidophilum.
8.011	A9GJ31	Cobyrinic acid ac-diamide synthase - Methylobacterium nodulans ORS 2060.
8.011	A4FPW4	Excisionase/Xis, DNA-binding - Saccharopolyspora erythraea (strain NRRL 23338).
7.97	Q981J0	Replication protein A - Rhizobium loti (Mesorhizobium loti).
7.97	A8M6R6	DNA binding domain, excisionase family - Salinispora arenicola (strain CNS-205).
7.848	A0QI31	TobE protein - Mycobacterium avium (strain 104).
7.828	A1RAG7	Putative molybdenum-pterin binding domain protein - Arthrobacter aurescens (strain TC1).
7.787	Q0SAJ4	Probable ABC molybdate transporter, substrate binding protein - Rhodococcus sp. (strain RHA1).
7.767	Q7WZH4	MerR (Fragment) - Vibrio cholerae.
7.746	Q98P11	Replication protein A - Rhizobium loti (Mesorhizobium loti).
7.665	Q0RN63	Putative citrate synthase-like protein - Frankia alni (strain ACN14a).
7.462	A2SFT7	Putative transposase - Methylibium petroleiphilum (strain PM1).
7.441	Q4BZD9	Cytosine-specific methyltransferase - Crocosphaera watsonii.
7.441	A5WY06	RepA - Agrobacterium tumefaciens.
7.441	A3WFM2	Putative uncharacterized protein - Erythrobacter sp. NAP1.
7.34	A4LY71	Regulatory protein, MerR - Geobacter bemidjiensis Bem.
7.197	A9AWL8	Citrate synthase - Herpetosiphon aurantiacus ATCC 23779.
7.197	A3T2U3	ATPase, ParA type - Sulfitobacter sp. NAS-14.1.
7.177	Q5YRL5	Putative transporter - Nocardia farcinica.
7.096	Q5L6D6	PTS system, IIa component - Chlamydophila abortus.
7.075	Q1A0H8	Gp39 - Mycobacterium phage Che12.
7.055	Q89XX3	Citrate synthase - Bradyrhizobium japonicum.
7.055	A1HB31	Citrate synthase - Ralstonia pickettii 12J.
7.035	Q1AXE1	Excisionase/Xis, DNA-binding protein - Rubrobacter xylanophilus (strain DSM 9941 / NBRC 16129).
7.035	A7HGM8	DNA binding domain, excisionase family - Anaeromyxobacter sp. (strain Fw109-5).
7.014	Q1M8W1	Putative replication partitioning protein - Rhizobium leguminosarum bv. viciae (strain 3841).
7.014	A0HI38	Response regulator receiver protein - Comamonas testosteroni KF-1.
6.994	Q1BG16	Molybdenum-pterin binding domain protein - Mycobacterium sp. (strain MCS).
6.994	A7CCW7	DNA binding domain, excisionase family - Ralstonia pickettii 12D.
6.994	A5CTJ1	Putative transcriptional regulator involved in molybdate uptake - Clavibacter michiganensis subsp. michiganensis (strain NCPPB 382).
6.852	Q0RUC7	Putative DNA-binding protein - Frankia alni (strain ACN14a).
6.852	A7IPX9	DNA binding domain, excisionase family - Xanthobacter sp. (strain Py2).
6.831	Q1IYS7	Excisionase/Xis related HTH transcription regulator - Deinococcus geothermalis (strain DSM 11300).

6.831	A8YWD8	Transposase ORF_A - Lactobacillus helveticus (strain DPC 4571).
6.831	A8YVY1	Transposase ORF_A - Lactobacillus helveticus (strain DPC 4571).
6.831	A8LA68	DNA binding domain, excisionase family - Frankia sp. EAN1pec.
6.811	Q36689	RepA protein - Rhizobium leguminosarum.
6.811	Q215A6	Putative GAF sensor protein - Rhodopseudomonas palustris (strain BisB18).
6.811	A3PSH2	DNA binding domain, excisionase family - Mycobacterium sp. (strain JLS).
6.709	A6DLD7	DNA-binding response regulator, excisionase family protein - Lentisphaera araneosa HTCC2155.
6.689	Q140J1	Putative citrate synthase - Burkholderia xenovorans (strain LB400).
6.689	A8VQ61	Glucan endo-1,3-beta-D-glucosidase - Bacillus selenitireducens MLS10.
6.669	Q856T0	Gp14 - Mycobacterium phage Corndog.
6.669	Q1JYF0	Regulatory protein, MerR - Desulfuromonas acetoxidans DSM 684.
6.669	Q05246	Gene 36 protein - Mycobacteriophage L5.
6.608	A6C2K5	Type II adenine specific methyltransferase - Planctomyces maris DSM 8797.
6.526	Q1N036	Response regulator - Oceanobacter sp. RED65.
6.506	Q0F233	Bacterial regulatory proteins, MerR family - Mariprofundus ferrooxydans PV-1.
6.506	Q07NE3	Putative GAF sensor protein - Rhodopseudomonas palustris (strain BisA53).
6.486	A6MN77	Putative uncharacterized protein - Thermus sp. 4C.
6.465	Q2RM21	Excisionase/Xis, DNA-binding - Moorella thermoacetica (strain ATCC 39073).
6.445	Q3SUF2	Putative uncharacterized protein - Nitrobacter winogradskyi (strain Nb-255 / ATCC 25391).
6.445	Q1DF95	Citrate synthase - Myxococcus xanthus (strain DK 1622).
6.201	A9BMJ2	Citrate synthase - Delftia acidovorans SPH-1.
6.16	A6M9A5	Putative regulatory protein - Geobacillus virus E2.
6.16	A6C2F2	Putative uncharacterized protein - Planctomyces maris DSM 8797.
6.14	Q92XG0	RepA2 replication protein - Rhizobium meliloti (Sinorhizobium meliloti).
6.099	A6L6L9	Putative transposase - Bacteroides vulgatus (strain ATCC 8482 / DSM 1447 / NCTC 11154).
6.059	A6PLL8	Phage transcriptional regulator, AlpA - Victivallis vadensis ATCC BAA-548.
5.876	Q11FP9	Citrate synthase - Mesorhizobium sp. (strain BNC1).
5.855	Q1WLM8	RepA - Rhizobium meliloti (Sinorhizobium meliloti).
5.855	A3TQ77	Putative uncharacterized protein - Janibacter sp. HTCC2649.
5.794	A8V8E1	TPR repeat protein - Anaeromyxobacter sp. K.
5.774	Q7UXD8	Similar to DNA-binding protein - Rhodopirellula baltica.
5.774	Q48567	Orf protein - Lactobacillus helveticus.
5.774	Q1N8F9	Transcriptional regulator, MerR family protein - Sphingomonas sp. SKA58.
5.591	Q9LCU5	Putative excisionase - Arthrobacter sp. TM1.
5.591	Q746H1	Soj protein - Thermus thermophilus (strain HB27 / ATCC BAA-163 / DSM 7039).
5.53	Q4J7B3	ISC1913-like resolvase - Sulfolobus acidocaldarius.
5.53	A6MN84	Putative uncharacterized protein - Thermus sp. 4C.
5.53	A4ALM8	Putative uncharacterized protein - marine actinobacterium PHSC20C1.
5.164	A0K263	DNA binding domain, excisionase family - Arthrobacter sp. (strain FB24).
5.001	Q8FRD4	Putative uncharacterized protein - Corynebacterium efficiens.
4.981	Q6W4X5	Putative uncharacterized protein - Streptococcus thermophilus.
4.94	Q9V1M0	Resolvase related protein - Pyrococcus abyssi.
4.92	A1RBC8	Putative DNA binding domain, excisionase family protein - Arthrobacter aurescens (strain TC1).
4.9	Q6MBC0	Putative uncharacterized protein - Protochlamydia amoebophila (strain UWE25).
4.859	A3DHM0	Regulatory protein, MerR - Clostridium thermocellum (strain ATCC 27405 / DSM 1237).
4.859	A0FRL6	Regulatory protein, MerR - Burkholderia phymatum STM815.
4.839	Q2JB82	Recombinase - Frankia sp. (strain CcI3).
4.798	Q6NJZ3	Putative excisionase - Corynebacterium diphtheriae.
4.798	Q124F4	Response regulator receiver domain protein - Polaromonas sp. (strain JS666 / ATCC BAA-500).
4.188	Q62P29	Sporulation inhibition like protein - Bacillus licheniformis (strain DSM 13 / ATCC 14580).
4.147	B0K4R8	DNA binding domain, excisionase family - Thermoanaerobacter sp. (strain X514).
4.147	A2SC00	Transcriptional regulator, LysR family - Methylibium petroleiphilum (strain PM1).
3.944	Q96Z87	Putative uncharacterized protein ST1944 - Sulfolobus tokodaii.
3.944	A6DNV8	Putative uncharacterized protein - Lentisphaera araneosa HTCC2155.
3.619	Q96Z93	Putative uncharacterized protein ST1939 - Sulfolobus tokodaii.
3.619	Q96XM3	Putative uncharacterized protein ST2495 - Sulfolobus tokodaii.
3.619	Q72JH4	Putative excisionase - Thermus thermophilus (strain HB27 / ATCC BAA-163 / DSM 7039).
3.619	A8M3Q6	Excision promoter, Xis - Salinispora arenicola (strain CNS-205).
3.558	A7VSL4	Putative uncharacterized protein - Clostridium leptum DSM 753.
3.537	Q74F74	PTS system IIA component, fructose subfamily - Geobacter sulfurreducens.
3.253	Q5YZS3	Putative uncharacterized protein - Nocardia farcinica.
3.232	A3HT73	Mobilizable transposon, tnpA protein - Algoriphagus sp. PR1.
3.212	Q2YZL9	Aspartate carbamoyltransferase, catalytic chain - uncultured epsilon proteobacterium.
3.131	Q0REM3	Excisionase - Frankia alni (strain ACN14a).
3.131	A5U977	Putative excisionase - Mycobacterium tuberculosis (strain ATCC 25177 / H37Ra).
3.131	A2VMC9	Putative uncharacterized protein - Mycobacterium tuberculosis C.
3.11	A3J9D6	MerR family regulatory protein - Marinobacter sp. ELB17.
2.846	Q5Z143	Putative uncharacterized protein - Nocardia farcinica.
2.744	Q8XUE9	Putative excisionase protein - Ralstonia solanacearum (Pseudomonas solanacearum).
2.602	Q9K9Q4	BH2591 protein - Bacillus halodurans.
2.602	A3YB66	Putative uncharacterized protein - Marinomonas sp. MED121.
2.48	Q604A7	Putative uncharacterized protein - Methylococcus capsulatus.
2.317	Q4AJK9	Acetyl-CoA biotin carboxyl carrier - Chlorobium phaeobacteroides BS1.
2.277	A4LD75	Excisionase domain protein - Burkholderia pseudomallei 305.
2.256	O85865	Putative excisionase - Sphingomonas aromaticivorans.
2.216	A6W2E5	DNA binding domain, excisionase family - Marinomonas sp. (strain MWYL1).
2.155	Q4C9A6	Excisionase/Xis, DNA-binding - Crocosphaera watsonii.
1.687	Q20Y39	Excisionase/Xis, DNA-binding - Rhodopseudomonas palustris (strain BisB18).
1.545	Q7NJV9	Glr1722 protein - Gloeobacter violaceus.

Table S3. MerR family members in different available full and partial genomes. The size of the full genome or part of it is provided together with the number of MerR family members identified by Provalidator.

Genome	Size (M)
34	
Streptomyces coelicolor A3(2)	8.66
31	
Streptomyces avermitilis MA-4680	9.02
24	
Bacillus licheniformis ATCC 14580	4.22
Salinispora tropica CNB-440	5.18
23	
Saccharopolyspora erythraea NRRL 2338	8.21
22	
Nocardia farcinica IFM 10152	6.02
Synechococcus sp. JA-3-3Ab	2.93
20	
Burkholderia xenovorans LB400	4.89
19	
Salinispora arenicola CNS-205	5.78
Rhodococcus sp. RHA1	7.80
18	
Frankia sp. EAN1pec	8.98
Clostridium beijerinckii NCIMB 8052	6.00
Bacillus cereus E33L	5.30
17	
Bacillus weihenstephanensis KBAB4	0.07
16	
Bacillus cereus ATCC 10987	5.22
Ralstonia metallidurans CH34	0.23
Kineococcus radiotolerans SRS30216	4.76
Burkholderia sp. 383	1.39
15	
Frankia alni ACN14a	7.49
Burkholderia cenocepacia HI2424	3.48
Bacillus cereus ATCC 14579	5.41
Bacillus clausii KSM-K16	4.30
Burkholderia multivorans ATCC 17616	3.44
Desulfitobacterium hafniense Y51	5.72
14	
Delftia acidovorans SPH-1	6.76
Burkholderia vietnamiensis G4	0.26
Sphingopyxis alaskensis RB2256	3.34
Clostridium phytofermentans ISDg	4.84
Burkholderia cenocepacia AU 1054	3.29
Pseudomonas aeruginosa PA7	6.58
13	
Bacillus thuringiensis serovar konkukian str. 97-27	5.23
Bacillus anthracis str. 'Ames Ancestor'	0.09
Frankia sp. CcI3	5.43
Burkholderia ambifaria AMMD	3.55
Acidovorax sp. JS42	4.44
Shewanella sp. ANA-3	0.27
Clostridium difficile 630	4.29
Bacillus thuringiensis str. Al Hakam	0.05
Marinobacter aquaeolei VT8	0.21
12	
Bacillus anthracis str. Sterne	5.22
Mycobacterium vanbaalenii PYR-1	6.49
Bacillus subtilis subsp. subtilis str. 168	4.21
Pseudomonas putida KT2440	6.18
Lactobacillus plantarum WCFS1	3.30
Bacillus anthracis str. Ames	5.22
Pseudomonas stutzeri A1501	4.56
Clostridium botulinum F str. Langeland	3.99
Sorangium cellulosum 'So ce 56'	13.03
Pseudomonas putida F1	5.96
Pseudomonas aeruginosa UCBPP-PA14	6.53
Ochrobactrum anthropi ATCC 49188	2.88
11	
Mycobacterium smegmatis str. MC2 155	6.98
Clostridium kluyveri DSM 555	3.96
Herpetosiphon aurantiacus ATCC 23779	6.34
Lactobacillus brevis ATCC 367	2.29
Shewanella sp. W3-18-1	4.70
Rhizobium leguminosarum bv. viciae 3841	5.05
Listeria monocytogenes EGD-e	2.94
Mycobacterium sp. JLS	6.04
Pseudomonas fluorescens Pf-5	7.07
Polaromonas sp. JS666	5.20
Sulfolobus solfataricus P2	2.99
Listeria monocytogenes str. 4b F2365	2.90
Mycobacterium sp. KMS	0.30
Alkaliphilus oremlandii OhILAs	3.12
10	
Rhizobium etli CFN 42	4.38
Agrobacterium tumefaciens str. C58	2.84
Corynebacterium glutamicum ATCC 13032	3.30
Clostridium botulinum A str. ATCC 19397	3.86
Burkholderia thailandensis E264	2.91
Listeria innocua Clip11262	3.01
Burkholderia pseudomallei 668	3.91
Pseudomonas mendocina ymp	5.07
Clostridium botulinum A str. ATCC 3502	3.88
Alkaliphilus metalliredigens QYMF	4.92
Hahella chejuensis KCTC 2396	7.21
Nocardioides sp. JS614	0.30
Bacillus pumilus SAFR-032	3.70
Mycobacterium sp. MCS	5.70
Burkholderia pseudomallei K96243	4.07
Mesorhizobium sp. BNC1	0.34
Shewanella putrefaciens CN-32	4.65
Clostridium botulinum A str. Hall	3.76
Dinoroseobacter shibae DFL 12	3.78
Arthrobacter sp. FB24	0.15
Salmonella enterica subsp. enterica serovar Typhi str. CT18	4.80
Pseudomonas aeruginosa PAO1	6.26
Burkholderia pseudomallei 1106a	3.98
Pseudomonas putida GB-1	6.07
9	
Burkholderia mallei ATCC 23344	3.51
Pseudomonas fluorescens PfO-1	6.43
Pseudomonas entomophila L48	5.88
Maricaulis maris MCS10	3.36
Salmonella enterica subsp. enterica serovar Choleraesuis str. SC-B67	0.13
Novosphingobium aromaticivorans DSM 12444	3.56
Vibrio parahaemolyticus RIMD 2210633	3.28
Aeromonas salmonicida subsp. salmonicida A449	4.70
Clavibacter michiganensis subsp. michiganensis NCPPB 382	3.29
Burkholderia pseudomallei 1710b	4.12
Mycobacterium avium 104	5.47
Sinorhizobium meliloti 1021	1.35
Rhodobacter sphaeroides 2.4.1	0.11
Clostridium acetobutylicum ATCC 824	0.19
Silicibacter pomeroyi DSS-3	4.10
Rhodoferax ferrireducens T118	4.71
Thermobifida fusca YX	3.64
Mycobacterium gilvum PYR-GCK	5.61
Streptococcus mutans UA159	2.03
Bacillus amyloliquefaciens FZB42	3.91
Enterobacter sakazakii ATCC BAA-894	4.36
8	
Mycobacterium avium subsp. paratuberculosis K-10	4.82
Sinorhizobium medicae WSM419	1.57
Caulobacter sp. K31	5.47
Salmonella typhimurium LT2	4.85
Aeromonas hydrophila subsp. hydrophila ATCC 7966	4.74
Xanthobacter autotrophicus Py2	5.30
Acidobacteria bacterium Ellin345	5.65
Lactococcus lactis subsp. cremoris SK11	2.43
Burkholderia mallei NCTC 10247	2.35
Arthrobacter aurescens TC1	4.59
Methylibium petroleiphilum PM1	4.04
Lactococcus lactis subsp. cremoris MG1363	2.52
Oceanobacillus iheyensis HTE831	3.63
Photobacterium profundum SS9	4.08
Ralstonia eutropha JMP134	0.63
Bordetella parapertussis 12822	4.77
Ralstonia eutropha H16	4.05
Lactobacillus sakei subsp. sakei 23K	1.88

Bacillus cereus subsp. cytotoxis NVH 391-98	4.08
Shewanella frigidimarina NCIMB 400	4.84
Methanosarcina acetivorans C2A	5.75
Chromobacterium violaceum ATCC 12472	4.75
Silicibacter sp. TM1040	0.82
Geobacillus thermodenitrificans NG80-2	3.55
Klebsiella pneumoniae subsp. pneumoniae MGH 78578	5.31
Paracoccus denitrificans PD1222	2.85
Azoarcus sp. BH72	4.37
Burkholderia mallei SAVP1	1.73
7	
Mesorhizobium loti MAFF303099	7.03
Burkholderia mallei NCTC 10229	2.28
Herminiimonas arsenicoxydans	3.42
Streptococcus agalactiae 2603V/R	2.16
Solibacter usitatus Ellin6076	9.96
Bordetella pertussis Tohama I	4.08
Shewanella sp. MR-4	4.70
Ralstonia solanacearum GMI1000	3.71
Nitrosomonas eutropha C91	0.06
Thermoanaerobacter tengcongensis MB4	2.68
Alcanivorax borkumensis SK2	3.12
Rhodobacter sphaeroides ATCC 17029	0.12
Pseudomonas syringae pv. tomato str. DC3000	6.39
Janthinobacterium sp. Marseille	4.11
Parvibaculum lavamentivorans DS-1	3.91
Bacillus halodurans C-125	4.20
Acidiphilium cryptum JF-5	0.20
Salmonella enterica subsp. enterica serovar Paratyphi B str. SPB7	4.85
Deinococcus radiodurans R1	2.64
Bordetella bronchiseptica RB50	5.33
Polaromonas naphthalenivorans CJ2	4.41
Salmonella enterica subsp. arizonae serovar 62:z4,z23:--	4.60
Colwellia psychrerythraea 34H	5.37
Enterobacter sp. 638	0.15
6	
Desulfococcus oleovorans Hxd3	3.94
Legionella pneumophila str. Paris	0.13
Brucella melitensis biovar Abortus 2308	2.12
Brucella ovis ATCC 25840	1.16
Shewanella denitrificans OS217	4.54
Geobacillus kaustophilus HTA426	3.54
Myxococcus xanthus DK 1622	9.13
Shewanella amazonensis SB2B	4.30
Lactobacillus helveticus DPC 4571	2.08
Corynebacterium jeikeium K411	2.46
Jannaschia sp. CCS1	4.31
Brucella canis ATCC 23365	2.10
Pseudoalteromonas atlantica T6c	5.18
Bordetella petrii DSM 12804	5.28
Vibrio vulnificus YJ016	3.35
Escherichia coli APEC O1	5.08

Corynebacterium glutamicum R	3.31
Salmonella enterica subsp. enterica serovar Typhi str. Ty2	4.79
Idiomarina loihiensis L2TR	2.83
Lactobacillus casei ATCC 334	2.89
Geobacter uraniireducens Rf4	5.13
Escherichia coli E24377A	0.07
Hyphomonas neptunium ATCC 15444	3.70
Mycobacterium ulcerans Agy99	5.63
Acidovorax avenae subsp. citrulli AAC00-1	5.35
Sphingomonas wittichii RW1	0.22
Renibacterium salmoninarum ATCC 33209	3.15
Brucella suis 1330	2.10
Brucella abortus biovar 1 str. 9-941	2.12
Listeria welshimeri serovar 6b str. SLCC5334	2.81
Salmonella enterica subsp. enterica serovar Paratyphi A str. ATCC 9150	4.58
Staphylococcus haemolyticus JCSC1435	2.68
Erythrobacter litoralis HTCC2594	3.05
Shewanella sp. MR-7	4.79
Brucella melitensis 16M	2.11
5	
Brucella suis ATCC 23445	1.40
Rhodospirillum rubrum ATCC 11170	4.35
Escherichia coli O157:H7 str. Sakai	0.09
Actinobacillus succinogenes 130Z	2.31
Methylobacterium extorquens PA1	5.47
Shewanella oneidensis MR-1	4.96
Dehalococcoides sp. CBDB1	1.39
Escherichia coli W3110	4.64
Vibrio cholerae O1 biovar eltor str. N16961	2.96
Vibrio cholerae O395	1.10
Corynebacterium efficiens YS-314	3.14
Rubrobacter xylanophilus DSM 9941	3.22
Escherichia coli O157:H7 EDL933	5.52
Shewanella baltica OS185	5.22
Roseiflexus castenholzii DSM 13941	5.72
Vibrio vulnificus CMCP6	3.28
Saccharophagus degradans 2-40	5.05
Legionella pneumophila subsp. pneumophila str. Philadelphia 1	3.39
Shewanella loihica PV-4	4.60
Vibrio harveyi ATCC BAA-1116	3.76
Dechloromonas aromatica RCB	4.50
Escherichia coli UTI89	5.06
Escherichia coli 536	4.93
Citrobacter koseri ATCC BAA-895	4.72
Clostridium tetani E88	2.79
Bifidobacterium adolescentis ATCC 15703	2.08
Marinomonas sp. MWYL1	5.10
Shewanella sediminis HAW-EB3	5.51
Pseudomonas syringae pv. syringae B728a	6.09
Nitrobacter hamburgensis X14	0.29
Streptococcus agalactiae NEM316	2.21
Nostoc sp. PCC 7120	0.18

Mannheimia succiniciproducens MBEL55E	2.31
Roseobacter denitrificans OCh 114	4.13
Rhodobacter sphaeroides ATCC 17025	3.21
Shewanella halifaxensis HAW-EB4	5.22
Lactobacillus salivarius UCC118	1.82
Shewanella baltica OS195	5.34
Shewanella baltica OS155	5.12
Roseiflexus sp. RS-1	5.80
Shigella sonnei Ss046	4.82
Pediococcus pentosaceus ATCC 25745	1.83
Escherichia coli CFT073	5.23
Enterococcus faecalis V583	3.21
Pseudomonas syringae pv. phaseolicola 1448A	5.92
Deinococcus geothermalis DSM 11300	0.57
Oenococcus oeni PSU-1	1.78
Escherichia coli K12	4.63
Lactococcus lactis subsp. lactis Il1403	2.36
4	
Staphylococcus epidermidis ATCC 12228	2.49
Shigella dysenteriae Sd197	4.36
Staphylococcus aureus subsp. aureus JH9	2.90
Yersinia pestis Angola	0.06
Cytophaga hutchinsonii ATCC 33406	4.43
Staphylococcus aureus subsp. aureus N315	2.81
Staphylococcus aureus subsp. aureus str. Newman	2.87
Escherichia coli HS	4.64
Gloeobacter violaceus PCC 7421	4.65
Xanthomonas campestris pv. vesicatoria str. 85-10	5.17
Staphylococcus aureus subsp. aureus Mu50	2.87
Mycobacterium bovis BCG str. Pasteur 1173P2	4.37
Mycobacterium tuberculosis F11	4.42
Bdellovibrio bacteriovorus HD100	3.78
Dehalococcoides ethenogenes 195	1.46
Lactobacillus gasseri ATCC 33323	1.89
Azorhizobium caulinodans ORS 571	5.36
Mycobacterium tuberculosis H37Rv	4.41
Vibrio fischeri ES114	2.90
Staphylococcus aureus subsp. aureus USA300	2.87
Mycobacterium tuberculosis CDC1551	4.40
Staphylococcus aureus subsp. aureus COL	2.80
Thermus thermophilus HB8	1.84
Dehalococcoides sp. BAV1	1.34
Mycobacterium tuberculosis H37Ra	4.42
Staphylococcus aureus subsp. aureus USA300_TCH1516	2.87
Lactobacillus johnsonii NCC 533	1.99
Symbiobacterium thermophilum IAM 14863	3.56
Haemophilus influenzae 86-028NP	1.91

Corynebacterium diphtheriae NCTC 13129	2.48
Haemophilus influenzae PittGG	1.88
Shewanella pealeana ATCC 700345	5.17
Staphylococcus aureus subsp. aureus MW2	2.82
Yersinia pestis biovar Microtus str. 91001	4.59
Streptococcus agalactiae A909	2.12
Yersinia enterocolitica subsp. enterocolitica 8081	0.06
Staphylococcus aureus subsp. aureus Mu3	2.88
Acaryochloris marina MBIC11017	6.50
Lactobacillus acidophilus NCFM	1.99
Caulobacter crescentus CB15	4.01
Mycobacterium bovis AF2122/97	4.34
Staphylococcus aureus subsp. aureus JH1	2.90
Pelobacter carbinolicus DSM 2380	3.66
Staphylococcus aureus RF122	2.74
Yersinia pestis CO92	0.07
Staphylococcus aureus subsp. aureus NCTC 8325	2.82
Rhodopseudomonas palustris BisB5	4.89
Staphylococcus aureus subsp. aureus MSSA476	2.79
Yersinia pseudotuberculosis IP 32953	0.06
Psychrobacter cryohalolentis K5	0.04
Yersinia pestis Antiqua	0.07
Yersinia pestis Pestoides F	0.07
Microcystis aeruginosa NIES-843	5.84
Acinetobacter baumannii ATCC 17978	3.97
3	
Nitrosomonas europaea ATCC 19718	2.81
Gluconobacter oxydans 621H	0.16
Pseudoalteromonas haloplanktis TAC125	3.21
Leuconostoc mesenteroides subsp. mesenteroides ATCC 8293	2.03
Streptococcus gordonii str. Challis substr. CH1	2.19
Tropheryma whipplei TW08/27	0.92
Rhodopirellula baltica SH 1	7.14
Pelobacter propionicus DSM 2379	4.00
Streptococcus sanguinis SK36	2.38
Flavobacterium johnsoniae UW101	6.09
Bifidobacterium longum NCC2705	2.25
Propionibacterium acnes KPA171202	2.56
Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305	2.51
Clostridium perfringens ATCC 13124	3.25
Staphylococcus aureus subsp. aureus MRSA252	2.90
Serratia proteamaculans 568	5.44
Desulfotalea psychrophila LSv54	3.52
Chromohalobacter salexigens DSM 3043	3.69
Arcobacter butzleri RM4018	2.34
Bacteroides thetaiotaomicron VPI-5482	6.26
Thermus thermophilus HB27	1.89
Mycobacterium leprae TN	3.26

Anaeromyxobacter sp. Fw109-5	5.27
Gluconacetobacter diazotrophicus PAl 5	3.94
Synechococcus sp. JA-2-3B'a(2-13)	3.04
Magnetospirillum magneticum AMB-1	4.96
Streptococcus suis 98HAH33	2.09
Yersinia pseudotuberculosis IP 31758	4.72
Leifsonia xyli subsp. xyli str. CTCB07	2.58
Thiobacillus denitrificans ATCC 25259	2.90
Desulfovibrio desulfuricans G20	3.73
Xanthomonas oryzae pv. oryzae MAFF 311018	4.94
Staphylococcus epidermidis RP62A	2.61
Verminephrobacter eiseniae EF01-2	5.56
Lactobacillus delbrueckii subsp. bulgaricus ATCC BAA-365	1.85
Streptococcus pneumoniae TIGR4	2.16
Clostridium perfringens str. 13	3.03
Psychromonas ingrahamii 37	4.55
Yersinia pestis KIM	4.60
Bradyrhizobium japonicum USDA 110	9.10
Nitrosococcus oceani ATCC 19707	3.48
Shigella boydii Sb227	4.51
Streptococcus pneumoniae R6	2.03
Legionella pneumophila str. Corby	3.57
Acinetobacter sp. ADP1	3.59
Legionella pneumophila str. Lens	3.34
Xanthomonas oryzae pv. oryzae KACC10331	4.94
Tropheryma whipplei str. Twist	0.92
Photorhabdus luminescens subsp. laumondii TTO1	5.68
Campylobacter curvus 525.92	1.97
Rhodopseudomonas palustris CGA009	5.45
Trichodesmium erythraeum IMS101	7.75
Haemophilus influenzae Rd KW20	1.83
Haemophilus influenzae PittEE	1.81
Sulfurovum sp. NBC37-1	2.56
Xanthomonas axonopodis pv. citri str. 306	5.17
Anabaena variabilis ATCC 29413	6.36
Yersinia pestis Nepal516	4.53
Acidothermus cellulolyticus 11B	2.44
Bacteroides vulgatus ATCC 8482	5.16
2	
Streptococcus pyogenes MGAS315	1.90
Streptococcus pyogenes MGAS6180	1.89
Streptococcus pyogenes MGAS10270	1.92
Granulibacter bethesdensis CGDNIH1	2.70
Aquifex aeolicus VF5	1.55
Caldicellulosiruptor saccharolyticus DSM 8903	2.97
Methylobacillus flagellatus KT	2.97
Desulfotomaculum reducens MI-1	3.60
Streptococcus thermophilus LMG 18311	1.79
Lactobacillus delbrueckii subsp. bulgaricus ATCC 11842	1.86
Bradyrhizobium sp. ORS278	7.45

Ehrlichia ruminantium str. Welgevonden	1.51
Xanthomonas campestris pv. campestris str. 8004	5.14
Actinobacillus pleuropneumoniae serovar 3 str. JL03	2.24
Azoarcus sp. EbN1	4.29
Nitrobacter winogradskyi Nb-255	3.40
Thiomicrospira crunogena XCL-2	2.42
Magnetococcus sp. MC-1	4.71
Synechocystis sp. PCC 6803	3.57
Campylobacter jejuni subsp. jejuni 81-176	1.61
Flavobacterium psychrophilum JIP02/86	2.86
Thermoanaerobacter pseudethanolicus ATCC 33223	2.36
Shigella flexneri 2a str. 2457T	4.59
Anaeromyxobacter dehalogenans 2CP-C	5.01
Streptococcus pyogenes MGAS9429	1.83
Pyrococcus furiosus DSM 3638	1.90
Bradyrhizobium sp. BTAi1	8.26
Heliobacterium modesticaldum Ice1	3.07
Streptococcus pyogenes MGAS8232	1.89
Erwinia carotovora subsp. atroseptica SCRI1043	5.06
Helicobacter hepaticus ATCC 51449	1.79
Bartonella henselae str. Houston-1	1.93
Streptococcus pyogenes MGAS10750	1.93
Syntrophomonas wolfei subsp. wolfei str. Goettingen	2.93
Streptococcus pneumoniae D39	2.04
Streptococcus pyogenes M1 GAS	1.85
Wolinella succinogenes DSM 1740	2.11
Mesoplasma florum L1	0.79
Carboxydothermus hydrogenoformans Z-2901	2.40
Streptococcus pyogenes str. Manfredo	1.84
Xanthomonas campestris pv. campestris str. ATCC 33913	5.07
Shigella flexneri 2a str. 301	4.60
Actinobacillus pleuropneumoniae L20	2.27
Polynucleobacter sp. QLW-P1DMWA-1	2.15
Alkalilimnicola ehrlichei MLHE-1	3.27
Clostridium novyi NT	2.54
Geobacter sulfurreducens PCA	3.81
Parabacteroides distasonis ATCC 8503	4.81
Streptococcus pyogenes MGAS5005	1.83
Streptococcus pyogenes MGAS2096	1.86
Streptococcus thermophilus LMD-9	1.85
Bartonella tribocorum CIP 105476	2.61
Rhodopseudomonas palustris HaA2	5.33
Shigella flexneri 5 str. 8401	4.57
Clostridium thermocellum ATCC 27405	3.84
Streptococcus suis 05ZYH33	2.09
Methanosarcina mazei Go1	4.09
Streptococcus pyogenes MGAS10394	1.89
Streptococcus thermophilus CNRZ1066	1.79
Moorella thermoacetica ATCC 39073	2.62

Campylobacter jejuni subsp. jejuni NCTC 11168	1.64
Campylobacter jejuni subsp. jejuni 81116	1.62
Chloroflexus aurantiacus J-10-fl	5.25
Streptococcus pyogenes SSI-1	1.89
Geobacter metallireducens GS-15	3.99
Gramella forsetii KT0803	3.79
Zymomonas mobilis subsp. mobilis ZM4	2.05
Sulfurimonas denitrificans DSM 1251	2.20
Halorhodospira halophila SL1	2.67
1	
Rhodopseudomonas palustris BisB18	5.51
Helicobacter acinonychis str. Sheeba	1.55
Methanopyrus kandleri AV19	1.69
Wolbachia endosymbiont strain TRS of Brugia malayi	1.08
Neisseria meningitidis Z2491	2.18
Desulfovibrio vulgaris subsp. vulgaris DP4	3.46
Syntrophobacter fumaroxidans MPOB	4.99
Chlorobium phaeobacteroides DSM 266	3.13
Methanoculleus marisnigri JR1	2.47
Prosthecochloris vibrioformis DSM 265	1.96
Lactobacillus reuteri F275	1.99
Campylobacter hominis ATCC BAA-381	1.71
Chlorobium chlorochromatii CaD3	2.57
Anaplasma phagocytophilum HZ	1.47
Coxiella burnetii RSA 493	1.99
Bartonella bacilliformis KC583	1.44
Nitratiruptor sp. SB155-2	1.87
Campylobacter jejuni RM1221	1.77

Sodalis glossinidius str. 'morsitans'	4.17
Methylococcus capsulatus str. Bath	3.30
Bacteroides fragilis YCH46	5.27
Thermococcus kodakarensis KOD1	2.08
Candidatus Pelagibacter ubique HTCC1062	1.30
Synechococcus elongatus PCC 7942	2.69
Archaeoglobus fulgidus DSM 4304	2.17
Thermosynechococcus elongatus BP-1	2.59
Ehrlichia ruminantium str. Gardel	1.49
Nitrosospira multiformis ATCC 25196	3.18
Campylobacter fetus subsp. fetus 82-40	1.77
Neisseria gonorrhoeae FA 1090	2.15
Neisseria meningitidis MC58	2.27
Neisseria meningitidis FAM18	2.19
Methanosarcina barkeri str. Fusaro	4.83
Pelodictyon luteolum DSM 273	2.36
Methanospirillum hungatei JF-1	3.54
Helicobacter pylori HPAG1	1.59
Xylella fastidiosa 9a5c	2.67
Treponema denticola ATCC 35405	2.84
Ehrlichia chaffeensis str. Arkansas	1.17
Methanocaldococcus jannaschii DSM 2661	1.66
Coxiella burnetii Dugway 5J108-111	2.15
Thermoanaerobacter sp. X514	2.45
Mycoplasma penetrans HF-2	1.35
Psychrobacter arcticus 273-4	2.65
Syntrophus aciditrophicus SB	3.17
Lawsonia intracellularis PHE/MN1-00	1.45
Helicobacter pylori 26695	1.66

Pasteurella multocida subsp. multocida str. Pm70	2.25
Neisseria meningitidis 053442	2.15
Anaplasma marginale str. St. Maries	1.19
Bartonella quintana str. Toulouse	1.58
Helicobacter pylori J99	1.64
Fusobacterium nucleatum subsp. nucleatum ATCC 25586	2.17
Methanocorpusculum labreanum Z	1.80
Xylella fastidiosa Temecula1	2.51
Porphyromonas gingivalis W83	2.34
Campylobacter jejuni subsp. doylei 269.97	1.84
Campylobacter concisus 13826	2.05
Ehrlichia canis str. Jake	1.31
Clostridium perfringens SM101	2.89
Rhodopseudomonas palustris BisA53	5.50
Pelotomaculum thermopropionicum SI	3.02
Rickettsia felis URRWXCal2	1.48
Salinibacter ruber DSM 13855	3.55
Neorickettsia sennetsu str. Miyayama	0.85
Chlorobium tepidum TLS	2.15
Bacteroides fragilis NCTC 9343	5.20
uncultured methanogenic archaeon RC-I	3.17
Rickettsia akari str. Hartford	1.23
Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough	3.57
Wolbachia endosymbiont of Drosophila melanogaster	1.26
Synechococcus sp. RCC307	2.22
Coxiella burnetii RSA 331	2.01

Appendix S1. Seed sequences for profile construction.

Name	Genome	NCBI Id	SPTR Id	Target substrate
ArpB	Pseudomonas putida S12	8163736	Q9KJC2	Antibiotics: carbenicillin, chloramphenicol, erythromycin, novobiocin, streptomycin and tetracyline. It is not involved in organic solvent efflux
SmeB	Xanthomonas maltophilia	5764625	Q9RBY8	Multidrug
EefB	Enterobacter aerogenes		Q8GC83	Multidrug
IfeB	Agrobacterium tumefaciens		O68441	Isoflavonoid-inducible
MexD	Pseudomonas aeruginosa	1399758	Q51396	
AdeB	Acinetobacter baumannii	16118478	Q93E19	Aminoglycoside
MtrD	Neisseria gonorrhoeae	1408202	Q51073	
MexY	Pseudomonas aeruginosa	3868984	Q9ZNG8	Multidrug
MdtB	Escherichia coli		P76398	Novobiocin and deoxycholate
CnrA	Ralstonia eutropha	584962	P37972	Nickel and cobalt
CzcA	Ralstonia eutropha	2507003	P13511	Divalent cation
SilA	Salmonella typhimurium	13633956	Q9ZHC9	Silver
NolG	Sinorhizobium meliloti	16262933	P25197	Involved in the production of Medicago-specific nodulation signal molecule

Capítulo 4

Characterization of the RND family of multidrug efflux pumps: in silico to in vivo confirmation of four functionally distinct subgroups

Characterization of the RND family of multidrug efflux pumps: *in silico* to *in vivo* confirmation of four functionally distinct subgroups

Patricia Godoy, Antonio J. Molina-Henares, Jesús de la Torre, Estrella Duque, and Juan L. Ramos

Department of Environmental Protection, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, C/ Profesor Albareda, 1, 18008 Granada, Spain

We have developed a generalized profile that identifies members of the root-nodulation-cell-division (RND) family of efflux pumps and classifies them into four functional subfamilies. According to Z-score values, efflux pumps can be grouped by their metabolic function, thus making it possible to distinguish pumps involved in antibiotic-resistance (group 1) from those involved in metalresistance (group 3). In silico data regarding efflux pumps in group 1 were validated after identification of RND efflux pumps in a number of environmental microbes that were isolated as resistant to ethidium bromide. Analysis of Pseudomonas putida KT2440 identified efflux pumps in all groups. A collection of mutants in efflux pumps and a screening platform consisting of 50 drugs were created to assign a function to the efflux pumps. We validated in silico data regarding efflux pumps in groups 1 and 3 using 9 different mutants. Four mutants belonging to group 2 were found to be more sensitive than the wild type to oxidative stressinducing agents such as bipyridyl and methyl viologen. The two remaining mutants belonging to group 4 were found to be more sensitive than the parental to tetracycline and one of them was particularly sensitive to rubidium and chromate. By effectively combining in vivo data with generalized profiles and gene annotation data, this approach allowed the assignment, according to metabolic function, of both known and uncharacterized RND efflux pumps into subgroups, thereby providing important new insight into the functions of proteins within this family.

efflux pumps | RND family | antibiotics resistance | multidrugs

The analysis of protein sequences makes it possible to identify conserved domains in proteins, such as enzyme catalytic sites, cofactor binding sites, small molecule ligand domains, DNA binding domains, and many others (1,2). Proteins or protein domains belonging to a particular family often share functional attributes, and therefore the grouping of proteins has been used in turn to characterize them at the functional level (3,4). Several approaches are available to define domains and protein families, and the availability of semi-automatic methods for profile construction, as well as their high sensitivity have improved the efficiency and eased the process involved in the definition of protein families (5-10). Profiles are not necessarily confined to small regions with high sequence similarity, but rather they attempt to characterize a protein family (or domain) based on full length sequences (6,11,12).

Microorganisms in the environment are exposed to a large number of drugs of natural and xenobiotic origin, and, as such, have developed strategies to cope with toxic compounds. The extensive use of some drugs in medicine, such as biocides and antibiotics, have lead to a major therapeutic problem as bacteria have developed resistance to multiple antibiotics (13-18). Resistance-Nodulation-cell Division (RND) efflux pumps are common elements in multidrug resistance, and their wide substrate specificity explains cross-resistance between antibiotics, biocides, dyes, and solvents in laboratory strains (13,19-23). A number of RND efflux pumps have also been described that extrude heavy metals, and represent a major determinant in the proliferation of microorganisms at sites polluted with zinc, lead, mercury, cobalt and other metals (24). Although the entire suite of physiological functions of RND pumps has not yet been well established, a number of recent studies have shown that these pumps may be involved in the extrusion of intracellularly generated toxic compounds. These RND pumpexcluded compounds may include, for example, formaldehyde produced from the metabolism of histidine and methoxylated chemicals (25); amino acids to maintain amino acid homeostasis (26); and quorum sensing molecules (27,28). As well, RND pumps may also be important for protecting cells against toxic compounds in the cells environment (23,29).

A typical RND efflux pump consists of three components, with one of these components being the inner membrane protein, which acts as the extrusion element, and which is often more than 1000 amino acids long, consisting of 12 transmembrane helices (23,30-32). A second component of RND efflux pumps is the outer membrane protein that penetrates into the periplasmic space to form a channel (22,33,34). The third component is a lipoprotein that is linked to the inner membrane, and which plays a role in stabilizing the interactions between the two other elements (35-37). The best studied antibiotic-extruding RND pump in structural and functional terms is the AcrAB-TolC system in *Escherichia coli* (23). This pump was initially described as a transporter for the topical antiseptic acriflavin (hence the name Acr), but it was later shown to transport a large variety of other substrates (14,21). Another well characterised RND pump is the CzcABC metal-extruding pump, which is involved in the extrusion of heavy metals such as zinc, cadmium and cobalt (38,39).

In this study we have constructed a stringent profile for RND efflux pumps and have used it to search for and identify members of this family within annotated genomes. With the protein sequences that we found we were able to group them, using Z-score values, into four groups, which represents a novel method regarding the application of profiles. Based on gene annotation data, those within groups 1 and 3 appeared to be involved in extrusion of antibiotic and metals, respectively. Using a series of *P. putida* KT24440 mutants, we were able to validate that the efflux pumps found in group 1 do in fact extrude a wide range of antibiotic compounds, and that those in group 3 extrude heavy metals. The paucity of information available for proteins in groups 2 and 4 made it necessary to carry out screens, using mutants, in order to assign functional activities to these groups. Our results show that the efflux pumps found in group 2 are involved in the extrusion of oxidative stress-causing agents, while the pumps in group 4 appear to be involved in the extrusion of organic and inorganic chemicals.

Results and Discussion

Construction of the RND profile. We have constructed a stringent profile for RND efflux pumps using Provalidator, a tool that combines nearlyfull automation of profile construction and validation (12), and have used it to identify RND efflux pumps in annotated genomes. The first step in the construction of a profile was to select sequences to produce the seed profile (12). We chose a set of 16 sequences of well characterized RND

efflux pumps that extrude antibiotics, solvents and heavy metals (Suppl. Table 1). We generated a multialignment of these proteins using CLUSTALW (http://align.genome.jp) (Suppl. Figure 1), which revealed conserved sequence identity throughout the full length of the proteins. We therefore opted to use the full length of the proteins to construct the profile using Provalidator (www.bactregulators.org). This is the first report of a profile being generated based on the whole protein sequence rather than on a fragment of conserved sequences. The generalized profile for RND efflux pumps was run against all entries in the SwissProt and TrEMBL databases (40). Around 2000 proteins ranging in length from 950 to 1200 amino acids were identified as members of the RND efflux pump family (Suppl. Table 2). The profile yielded a Z-score of >50 for all of these proteins. Z-scores were used to classify the proteins into 4 well-defined groups based on intervals of about 50 arbitrary units on the Z-score scale (Figure 1).

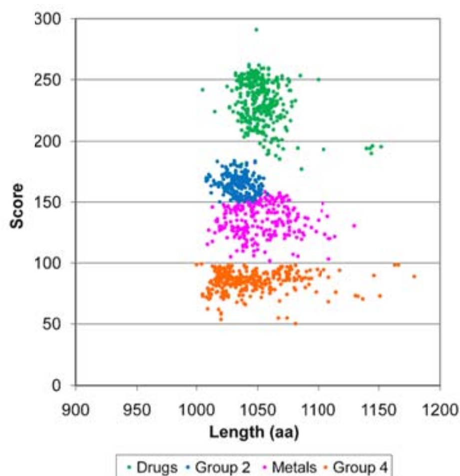


Fig. 1. Grouping of efflux pumps based on Z-score. All proteins identified by the RND efflux pump profile were at least 950 residues long. Based on Z-scores we established four groups of proteins. Each group corresponds to a set of 50-unit intervals.

A phylogenetic tree was also constructed, and the sequences of each Z-score group tended to cluster together (Figure 2). When proteins in these clusters were analysed with regard to their function, we found that those in group 1 were antibiotic/drug efflux pumps such as the Mex/Ttg pumps of *Pseudomonas* (41-48), the Acr pumps of enteric bacteria (35,49,50), and the Cme pump of *Acinetobacter* spp (51,52). Group 3 included metal resistance efflux pumps such as Czc/Cnr from *Ralstonia* (39,53), and Sil from *Salmonella* (54-56), among others. No known function was found to be previously assigned to members of the proteins in groups 2 and 4.

Wet testing of the efflux pump profile groups. We then hypothesized that it would be possible to validate the Z-score groups of the RND pumps by identifying efflux pumps in environmental microorganisms that exhibited tolerance to drugs. To test this hypothesis we isolated environmental microorganisms tolerant to ethidium bromide (EtBr). We choose EtBr because it has been reported that a number of pathogenic and nonpathogenic microorganisms, such as *Acinetobacter braumannii*, *Chromohalobacter* sp., *Lactococcus lactis*, and several species of the genus *Pseudomonas* (41,44,46,51,52,57-59) are tolerant to the compound thanks to efflux pumps. Subsequently we planned to identify (by PCR amplification and DNA sequencing) genes that encoded efflux pumps in these microorganisms related to the phenomenon under scrutiny.

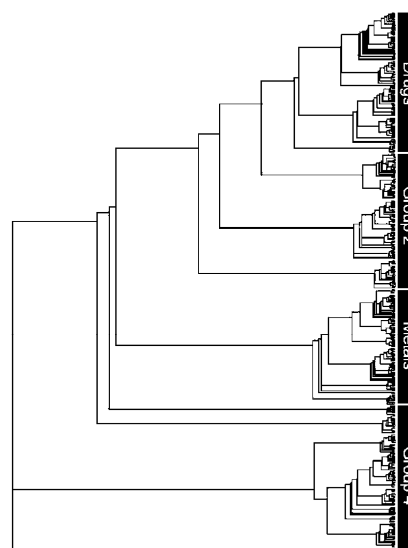


Fig. 2. Phylogenetic tree of efflux pumps and assigned function. The phylogenetic tree was constructed using the TREE program. The function of the characterized pumps in each of the four groups is indicated to the right of the cluster.

To isolate environmental microorganisms tolerant to EtBr, 20 ml of waste water samples from the city of Granada's wastewater treatment plant were mixed with 80 ml M9 minimal medium (60) containing glucose (0.5% w/v) as a carbon source and EtBr at a final concentration of 10 mg/ml and the samples were used for enrichment, as described in Materials and Methods. Based on differences in colony morphology, size and color we retained 11 clones from these enrichments (Suppl. Table 3). The clones were taxonomically identified based on Gram staining, the API growth test, metabolic profiles, and sequencing of the 16S rRNA gene. Three clones belonged to the genus *Klebsiella*, two to the genus *Kluyvera*, and one of each belonged to the genera *Shewanella*, *Empedobacter* and *Pseudomonas*. Two others exhibited high similarity to uncultured microorganisms and one clone exhibited significant homology to Bacterium G3 Greenlake (Suppl. Table 3). The predominance of enteric bacteria among the isolates and their ability to use a wide variety of sugars (Suppl. Table 4) was not surprising, considering that the microorganisms were isolated from wastewater. We characterized all these strains with regard to EtBr and antibiotic resistance using Minimal Inhibitory Concentration (MIC) assays (61). The results showed that all EtBr-resistant clones exhibited high resistance to certain antibiotics, but we did not find a regular pattern of antibiotic resistance, except that clones resistant to EtBr were often highly resistant to penicillin-derived antibiotics (Suppl. Table 5).

We next tested whether these environmental isolates exhibited genes that encoded efflux pumps. To this end, and based on the alignment in Supplementary Figure 1 and previous studies by Meguro *et al*. (62), we used two oligoprimers to amplify by PCR a region of about 500 nucleotides of the RND efflux pumps within the isolates. As a control for amplification we included *P. putida* DOT-T1E, a strain that is known to be resistant to EtBr due to its extrusion through *ttgABC*-encoded gene products (63-64). As expected, the *ttgB* gene was amplified from the DOT-T1E strain using the above primers. We also found that a related gene was also present in eight of the eleven strains (including two *Kluivera* strains, the three *Klebsiella* strains, one of the *Pseudomonas* strains, *Bacterium* G3, and one of the strains exhibiting 16S rRNA similar to the EV821 uncultured bacterium). These amplified fragments were cloned, sequenced,

and clustered within the sequences used to construct the phylogenetic tree in Figure 2. Six of the sequences clustered with the RND efflux pumps assigned to group 1, which includes extrusion pumps for acriflavine, acridine and other drugs. One of the new sequences, which belongs to fragments amplified from *Pseudomonas* sp., clustered with pumps in Group 4, suggesting that this phylogenetically separate group of RND pumps may also include antibiotic/dyes efflux pumps (Suppl. Figure 2).

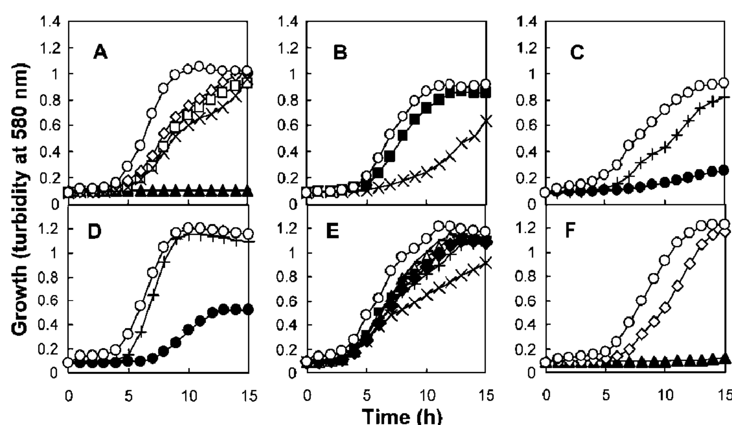


Fig. 3. Growth characteristics of a set of knock-out mutants in efflux pumps in Pseudomonas putida. Wild-type Pseudomonas putida and mutants strains were grown overnight in M9 minimal medium with glucose as a carbon source. Cultures were diluted 100-fold in the same medium and 180 µL of resulting culture was placed in 100-well polystyrene plates and incubated at 30 °C in a Bio-screen-C incubator that agitated the plates while measuring growth every 30 min, in the presence of either 20 µg/ml of tetracycline (panel A), 0.25 mM Cd^{2+} (panel B), 0.25 M Rb^{2+} (panel C), 0.5 mM Zn^{2+} (panel D), 1 mM Cr_2O_7 (panel E) or 30 µg/ml chloramphenicol (panel F). Symbols: ○, wild-type; ●, mutant PP0043; △, mutant PP0906; ▲, mutant PP1385; □, mutant PP1517; ■, mutant PP2065; ◆, mutant PP3583; ◖, mutant PP3584; ×, mutant PP5173; +, mutant PP5397.

Identification of extrusion pumps in the genome of *P. putida* KT2440. Using the RND family profile that we constructed, we screened the genome of *P. putida* KT2440 (65) and identified 14 RND efflux pumps corresponding to genes PP0043, PP0906, PP1385, PP1517, PP2065, PP2411, PP2818, PP3302, PP3426, PP3456, PP3583, PP3584, PP5173, and PP5387 (Suppl. Table 6). Three of these efflux pumps were grouped as being potentially involved in metal extrusion in group 3 (PP0043, PP2411 and PP5387); four of these efflux pumps were placed within group 1 (PP1385, PP3456, PP2818 and PP3426), while another four pumps fell within group 2 (PP2065, PP0906, PP3584, PP3583). The three remaining pumps were clustered within group 4 (PP1517, PP5173, PP3302) (See Suppl. Figure 3). Mutants for two of the pumps in group 1 (TtgABC [44] and MexEF/OprN [25]) have been isolated before and partially characterized. As such, it is known that the TtgABC pump (PP1384-PP1386) is an antibiotic extrusion pump, and that it is able to extrude ampicillin, chloramphenicol, tetracycline, and flavonoids (25,63,66). The MexEF/OprN (PP3425-PP3427) has been shown to be involved in formaldehyde detoxification and in phenylalanine homeostasis (25,26). No information was available for the pumps annotated as potential Metal Resistance pumps, nor was information available for the other pumps. In order to characterise the function of these putative efflux pumps, we used a number of mini-Tn*5*, including mutants for four pumps within group 2 (PP2065, PP0906, PP3583 and PP3584), two mutants in pumps within group 3 (PP0043 and PP5387) and two pumps within group 4 (PP1517 and PP5173). This set of mutants was subjected to a systematic phenomic screen that allowed us to test for growth under 190 different conditions (see Material and Methods). The platform allows for growth under different carbon, nitrogen, sulphur, and phosphate sources to be tested, as well as growth in the presence of 47 different stressors. The results showed that none of the mutants exhibited significant differences with respect to growth under different carbon, nitrogen, sulphur, or phosphate sources, as expected. However, significant differences in growth curves were observed for some of the mutants when compared to the parental strain in the presence of certain stressors (Figure 3). The mutants were found to exhibit two different growth patterns. One of these growth patterns corresponded to a deep inhibition of growth so that turbidity of the cultures did not increase with time. This indicated to us that the knocked-out efflux pump was an essential element for tolerance to the stressor. This is clearly exemplified in Figure 3A, which shows that novobiocin strongly inhibits growth of the *ttgB* mutant. The second observed growth pattern was different in that growth was not fully inhibited, but rather that it was significantly decreased. This indicated to us that more than one efflux pump was required in order to extrude the chemical. Table 1 summarises the results organised according to placement of the pumps within the four groups identified by the RND profile. As expected efflux pumps in group 1 were found to be involved in the extrusion of antibiotics. TtgB was the most critical efflux pump for antibiotic resistance as its deficiency lead to growth inhibition in the presence of tetracycline, novobiocin, gentamycin, ampicillin, chloramphenicol, cefotaxime, erythromicin, and the detergent deoxycholate. These results raise the profile of the TtgB efflux pump, which has previously only been shown to be able to extrude EtBr, ampicillin, chloramphenicol and a number of flavonoids. We also found that MexEF/OprN plays a role in extrusion of tetracycline and chloramphenicol, as shown by delayed growth of the corresponding mutant in the presence of these compounds (Figure 3F). Since deficiency in TtgABC fully prevents growth (Figure 3A), we propose that the MexEFOprN efflux pump is of less relevance to the removal of these antibiotics. In group 3 the tested mutants were found to exhibit inhibited growth in the presence of certain metals. As such, mutants deficient in PP0043 showed in increased sensitivity to Cd^{2+} and Zn^{2+}, while the mutant deficient in PP5387 was more sensitive to Cd^{2+}, Ni^{2+} and chromate. These results show that metal efflux pumps within *P. putida* KT2440 have overlapping specificity regarding removal of Cd^{2+}, but that they also show specificity regarding other metals, such as Ni^{2+} and Zn^{2+} (Figure 3D and 3E). The presence of multiple efflux pumps with similar specificity is advantageous to cells in that, although some redundancy exists, they are provided with a wider total spread of chemicals that can be extruded.

Table 1. Potential 1 substrates for efflux pumps of *Pseudomonas putida* KT2440 as deduced from growth inhibition of mutants in the presence of the indicated drugs.

Number	Group	Proposed name	Substrates
PP1385	1	TTgB	Tc, Nov, Gm, Sm, Ap, Cm, Ctx, Ery, Bip, DOC
PP3426	1	MexF	Tc, Cm
PP0906	2	OapB	Bip, Mv
PP2065	2	OapE	Bip, Mv
PP3584	2	OapH	Bip, Mv
PP0043	3	CzpB	Cd, Zn
PP5387	3	CneB	Cd, Ni, Cr_2O_7, MV
PP1517	4	TepB	Tc
PP5173	4	TrcB	Tc, Rb, Cr_2O_7

Abbreviations: Tc, Nov, Gm, Sm, Ap, Cm, Ctx, Ery, Bip, DOC, MV stand for resistance to 5 tetracycline, novobioxin, gentamycin, streptomycin, ampicillin, chloramphenicol, ceto, 6 bipyridyl, deoxycholate and methyl viologen, respectively.

The four mutants within group 2 have a common phenotypic background, and appear to be involved in the extrusion of organic compounds that generate oxidative stress. In fact, all four mutants exhibited retarded growth in the presence of bipyridyl and methyl viologen. Additionally, the mutant lacking PP3583 may also be involved in the extrusion of Ni^{2+} and chromate (Table 1). Mutants in the pumps within group 4 exhibited certain sensitivity to tetraycycline, and the mutant deficient in PP5173 showed striking inhibition of growth in the presence of rubidium (0.25 mM) and chromate (Table 1). This group, then, may include a number of efflux pumps able to remove heavy metals and antibiotics.

Our results support that, based on a limited number of sequences derived from a literature search, it is possible to construct a semiautomated profile for RND efflux pumps, which not only serves to identify members of the family in databases but that can also be used to group them into functional subfamilies based on Z-score values. Most importantly, the subfamilies exhibit differential specificity with regards to substrate specificity. As a proof of concept we isolated environmental microbes tolerant to EtBr and found that, in many of these microorganisms, it was possible to amplify genes that encode pumps that are specifically assigned to the expected Z-score groups, as confirmed with *in vivo* and *in vitro* assays against EtBr. The profile also allowed us to identify new efflux pumps within *P. putida* KT2440. These results were combined with *in vivo* experiments using mutants in these genes, which consisted of stringent phenomics screens that allowed us to assign functions and reveal the metabolic context within which these previously uncharacterised efflux pumps operate. These results provide new foundations for further studies in the area of substrate specificity of RND efflux pumps in environmental isolates and metagenomes.

Materials and Methods

Seed Sequences for Profile Construction. The 16 sequences that were used to create the seed profile (Supplementary Table 1) were chosen based on literature searches of well characterized RND efflux pumps. The sequences were aligned using CLUSTALW (http://align.genome.jp) revealing that there exists a high degree of sequence conservation along the whole sequence of the proteins (Suppl. Fig. 1). The final alignment was then used as seed for the construction of a conventional profile using the pfmake program, which is part of the PFTOOLS package of programs, available from the Swiss Institute of Bioinformatics. The RND profile was subsequently calibrated by running it against shuffled sequences in the database (http://www.isrec.isbsib.ch/pub/databases/shuffled). PFTOOLS proposes a tentative threshold score of 8.5 for a protein to be considered a member of the family of interest. The constructed RND profile then was used to search for members of this family in the Swiss-Prot and TrEMBL databases, and also to screen the set of complete and incomplete microbial genomes available at NCBI and GOLD. A set of almost 2000 members of the RND family was identified (Suppl. Table 2).

Bacterial Strains and growth conditions. *P. putida* KT2440 (65) and its derived Tn*5*-Km mutant strains were obtained from the *Pseudomonas* Reference Culture Collection established at Estación Experimental del Zaidín in Granada (Spain). Strains were grown at 30 ºC with shaking at 200 rpm in an orbital platform in LB medium supplemented with rifampicin (30 µg/ml) or kanamycin (50 µg/ml), respectively. When required, M9 minimal medium (60) was used and supplemented with the appropriate carbon source.

Isolation of Ethidium Bromide Resistant Microorganisms. 20 ml of waste water samples from the wastewater treatment plant of city of Granada was mixed with 80 ml M9 minimal medium (60), with glucose (0.5% w/v) as a carbon source, and Ethidium bromide (EtBr) was added at a final concentration of 10 mg/ml. Flasks were shaken (200 rpm) at 30 ºC for 24 h, at which point the samples were diluted 50-fold in the same medium. After incubating 24 h more, serial dilutions were spread on solid M9 minimal medium with glucose (0.5% w/v) as the C-source and 10 mg/ml EtBr as a counter selective agent. The clones were retained based on colony morphology, size and color.

Ethidium bromide-resistant strains were grown according to routine procedures in LB medium. When required, M9 minimal medium (60) was used and supplemented with different carbon sources at 10 mM.

DNA Techniques. Preparation of chromosomal DNA was carried out using standard methods (67). To amplify HAE-1 RND efflux pumps, we used the A24f2 (5'-CCSRTITTYGCITGGGT-3') and A577r2 (5'-SAICCARAIRCGCATSGC-3') primers, as described in Meguro *et al.* (62) to PCR amplify a 500-bp fragment. The PCR reaction procedure was as follows: an initial step at 94 ºC for 10', was followed by 30 cycles at 94 ºC for 1', then 50 ºC for 1', and a final step at 72 ºC for 1' was run. Amplification products were visualized on a 1.5% (w/v) agarose gel stained with ethidium bromide, and the corresponding band was extracted with the QUIAEX II Gel Extraction kit (Quiagen). The PCR products were sequenced using an ABI-PRISM Dye Terminator Cycle Sequencing Ready Reaction kit with Amplitaq DNA polymerase in an automatic DNA sequencer (model ABI-PRISM 3100; Applied Biosystems, USA).

Phenotypic Characterization. *P. putida* strains were grown on solid M9 minimal medium (60) supplemented with 0.1% (w/v)

glucose. The following day cells were recovered with a loop and resuspended in M8 minimal medium (67). The wild-type and the mutant strains were inoculated in microwell plates in M9 minimal medium with different carbon (5 mM), nitrogen (5 mM) and sulphur (5 mM) sources, and grown at 30 ºC with continuous shaking while at 60 minute intervals the turbidity was measured by a Bioscreen C (ThermoFisher Scientific) at 420-580 nm for 24 hours. Minimal M9 medium was prepared as in Abril *et al*. (60) but when appropriate $MgSO_4$ or NH_4Cl was replaced with other sulphur or nitrogen sources. All data recordings were performed using a Bioscreen C MBR analyzer type system FP-1100-C (OY Growth Curves Ab Ltd., Raisio, Finland). At least three independent experiments were performed for each condition. Growth curves were drawn using the average values of the three experiments (standard deviations were always less than 0.1 units of the O.D. value).

Carbon sources: D-glucose, D-fructose, D-glucuronic acid, glycerol, sodium acetate, trisodium citrate, fumaric acid, sodium succinate, sodium lactate, malic acid, sodium pyruvate, methyl pyruvate, propionic acid, sodium benzoate, sodium 4-hydroxybenzoate, quinic acid, aminobutyric acid, 5-amino-n-valeric acid, 2,4-dihydroxyphenylacetic acid, sodium decanoate, Tween 20, 2-phenyl ethanolamine, L-Leu, LLys, L-His, L-Gln, L-Glu, L-Phe, L-Arg, L-Asn, L-Ala, L-Pro, L-Tyr, L-Ile, L-Cys, L-Met, L-Val, glutaric acid, and xylose. *Nitrogen sources*: D- and L-Arg, D- and L-Lys, D- and L-Pro, D- and L-Val, D- and L-Ala, D- and L-Asn, D-Met, D-Leu, L-Asp, L-Cys, L-Phe, L-Glu, L-Gln, L-Gly, L-His, L-homoserine, L-Ile, L-trans hydroxyproline, L-Ser, L-Tyr, Ala-Glu, Ala- Gly, Ala-His, Ala-Leu, Ala-Phe, Gly-Gln, Gly-Gly, Gly-Leu, Gly-Ser, Gly- Val, Tyr-Ala, adenine, agmatine sulphate, hypoxanthine, phenylethanolamine, ethanolamine, putrescine, DL-ornithine, NH4Cl and NH4NO3 and urea. *Sulphur sources*: L-Cys, D- and L-Met, Lcystine, cysteamine, D,L-ethionine, D,L-homocysteine, taurine, thiourea, 2-thiouracil, N-acetyl-cysteamine, 2-thiohidanthoin, sodium taurocholate, agmatine sulphate, Na2SO4, Na2SO3.

Toxic Compound Resistance Assays. Individual colonies of *P. putida* KT2440 and mutant strains were picked from freshly cultured LB plates, and streaked onto LB medium plates supplemented with the suitable antibiotic and grown overnight at 30 ºC. The biomass of this overnight plate was recovered from the plate surface and resuspended in 15 ml of LB liquid medium to an OD_{660nm} of 0.1. Microwell plate wells were filled with 180 µl of the above cell suspension and 20 µl of the different 10x concentrated solutions of stressors. Microwell plates were incubated and data recordings were processed using the Bioscreen C MBR analyzer type system as described above. *Toxic compounds*: $AgNO_3$ (3 µM), $CdCl_2$ (0.156 mM), $CoCl_2$ (0.156 mM), $CuSO_4$ (1 mM), $HgCl_2$ (2.5 µg/ml), LiCl (0.25 M), $MnSO_4$ (1 mM), $NiCl_2$ (1 mM), RbCl (0.5 M), K_2TeO_3 (0.9 µg/ml), $ZnCl_2$ (0.5 mM), Non detergent sulphobetaine (NDSB-201) (1%), Cetyl trimethylammonium bromide (CTAB) (0.002%), N-lauryl sarcosine (0.3%), SDS (0.06%), deoxycholate (DOC) (1%), Triton X-100 (1%), ethylenediaminetetraacetic acid (EDTA) (0.125 mM), 2,2'-Bipyridine (1 mM), NaCl (0.5 M), $K_2Cr_2O_7$ (12.5 µg/ml), H_2O_2 (0.004%), NH_2OH (5 y 2.5 mM), Methyl viologen (100 µM), Tert-butyl hydroperoxide (0.00078%), ethidium bromide (0.1 mg/ml), KSCN (100 mM), KCN (0.325 mg/ml), K_2HAsO_4 (0.9 mg/ml), NaBr (0.25 M), ampicillin (20 µg/ml), carbenicillin (160 µg/ml), chloramphenicol (30 µg/ml), cefotaxime (0.375 µg/ml), erythromycin (30 µg/ml), gentamicin (4 µg/ml), kanamycin (0.195 µg/ml), nalidixic acid (0.0125 mg/ml), neomycin (1 µg/ml), norfloxacin (0.05 µg/ml), novomycin (20 µg/ml), piperacillin (10 µg/ml), streptomycin (2 µg/ml), spectomycin (0.1 mg/ml), tellurite (0.5 µg/ml), tetracycline (1 µg/ml).

Work in Ramos laboratory was supported by funding from PSYSMO (Ref. GEN2006-27750-C5-5-E/SYS), Excelencia Junta de Andalucía (CVI-3010), Consolider-C (BIO2006-05668), and Consolider-Ingenio-2010 (Ref. CSD2007-00005) grants. We thank M. Trini Gallegos for advice on the choice of RND efflux pump sequences, M.M. Fandila and C. Lorente for secretarial assistance, and Ben Pakuts for improving the English in the manuscript.

1. Bradley, M.J., Chivers, P.T., and Baker, N.A. (2008) Molecular dynamics simulation of the *Escherichia coli*, NikR: Equilibrium conformational fluctuations reveal interdomain allosteric communication pathways. *J Mol Biol* 378: 1155-1173.
2. Corpet, F., Gouzy, J., and Kahn, D. (1999) Recent improvements of the PRODOM database of protein domain families. *Nucleic Acids Res* 27: 263267.
3. Holm, L., and Sander, C. (1996) Mapping the protein universe. *Science* 273: 595-602.
4. Tatusov, R.L., Natale, D.A., Garkavtsev, I.V., Tatusova, T.A., Shankavaram, U.T., Rao, B.S., Kiryutin, B., Galperin, M.Y., Fedorova, N.D., and Koonin, E.V. (2001) The COG database: new developments in phylogenetic classification of proteins from complete genomes. *Nucleic Acids Res* 29: 2228.
5. Bateman, A., Coin, L.R., Durbin, R.D., Finn, V., Hollich, S., Griffith-Jones, A., Khanna, M., Marshall, S., Moxon, E.L.L., Sonnhammer, D.J., Studholme, C., Yeats, C., and Eddy, S.R. (2004) The Pfam protein families database. *Nucl Acids Res* 32: D138-141.
6. Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., de Castro, E., Langendijk-Genevana, P.S., Pagni, M., and Sigrist, C.J.A. (2006) The PROSITE database. *Nucl Acids Res* 34: D227-D230.
7. Bucher, P., and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc Int Conf Intell Syst Mol Biol* 2: 5361.
8. Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem* 20: 324.
9. Tobes, R., and Ramos, J.L. (2005) REP code: defining bacterial identity in extragenic space. *Environ Microbiol* 7: 225-228.
10. Ramos, J.L., Martínez-Bueno, M., Molina-Henares, A.J., Terán, W., Watanabe, K., Zhang, X., Gallegos, M.T., Brennan, R., Tobes, R. (2005) The TetR family of transcriptional repressors. *Microbiol Mol Biol Rev* 69: 326-356.
11. Rigali, S., Derouaux, A., Giannotta, F., and Dusart, J. (2002) Subdivision of the helix-turnhelix GntR family of bacterial regulators in the FadR, HutC, and YtrA subfamilies. *J Biol Chem* 277: 12507-12545.
12. Molina-Henares, A.J., Godoy, P., Duque, E., and Ramos, J.L. (2009) A general profile for the MerR family of transcriptional regulators constructed using the semi-automated Provalidator tool. *Environ Microbiol Reports* 1-6: 518-523.
13. Daniels, C., and Ramos, J.L. (2009) Microbial Biotechnology from medicine to bacterial population dynamics. *Microbial Biotechnol* 2: 304-307.
14. Fábrega, A., Madurga, S., Giralt, E., and Vila, J. (2009) Mechanism of action of and resistance to quinolones. *Microbial Biotechnol* 2: 40-61.
15. Aminov, R. (2009) The role of antibiotics and antibiotic resistance in nature. *Environ. Microbiol. Reports*, in press.
16. Baquero, F., Álvarez-Ortega, C., and Martínez, J.L. (2009) Ecology and evolution of antibiotic resistance. *Environ. Microbiol. Reports*, in press.
17. Grill, S., Busendender, S., Pfeiffer, M., Khler, M., and Mark, M. (2008) The bifunctional flavokinase/flavin adenine dinucleotide synthetase from *Streptomyces davawensis* produces inactive flavin cofactors and is not involved in resistance to the antibiotic roseoflavin. *J Bacteriol.* 190:1546-1553.
18. Zhang, L., and Mah, T.-F. (2008) Involvement of a novel efflux system in biofilm-specific resistance to antibiotics. *J Bacteriol* 190: 4447-4452.
19. Poole, K. (2004) Efflux-mediated multiresistance in Gram-negative bacteria. *Clin Microbiol Infect* 10: 12-26.
20. Nikaido, H. (1996) Multidrug efflux pumps of gram-negative bacteria. *J Bacteriol* 178: 5853-5859.
21. Nikaido, H. (1998) Multiple antibiotic resistance and efflux. *Curr Op Microbiol* 1: 516-523.
22. Nikaido, H. (2000) How do exported proteins and antibiotics bypass the periplasm in gramnegative cells? *Trends Microbiol* 8: 841-483.
23. Nikaido, H., and Takatsuka, Y. (2009) Mechanism of RND multidru efflux pumps. *Biochim. Biophys – Acta Proteins and Proteomics*

1794: 769-781.
24. Checa, S.K., Espariz, M., Pérez Audero, M.E., Botta, P.E., Spinelli, S.V., and Soncini, F.C. (2007) Bacterial sensing and detoxification of gold salts. *Mol Microbiol* 63: 1307-1318.
25. Roca, A., Rodríguez-Herva, J.J., Duque, E., and Ramos, J.L. (2008) Physiological responses of *Pseudomonas putida* to formaldehyde during detoxification. *Microbial Biotechnol* 1: 158-169.
26. Herrera, M.C., Duque, E., Rodríguez-Herva, J.J., and Ramos, J.L. (2009) PhhR Controls a regulon including two phenylalanine degradation pathways and the MexEF/OprN system in *Pseudomonas putida. Environ. Microbiol.* Submitted.
27. Pearson, J.P., van Delden, C., Iglewski, B.H. (1999) Active efflux and diffusion are involved in transport of *Pseudomonas aeruginosa* cell-to-cell signals. *J Bacteriol* 191: 1203-1210.
28. Ueda, A., Attila, C., Whiteley, M., and Wood, T.K. (2009) Uracil influences quorum sensing and biofilm formation in *Pseudomonas aeruginosa* and fluorouracil is an antagonist. *Microbial Biotechnol* 2: 62-74.
29. Levy, S. (1998) Multidrug resistance – A sign of the times. *New England J Med* 338:1376- 1378.
30. Murakami, S., Nakashima, R., Yamashita, E., and Yamaguchi, A. (2002) Crystal structure of bacterial multidrug transporter AcrB. *Nature* 419: 587-593.
31. Yu, E.W., Aires, J.R., McDermott, G., and Nikaido, H. (2005) A periplasmic drug-binding site of the AcrB multidrug efflux pump: a crystallographic and site-directed mutagenesis study. *J Bacteriol* 187: 6804-6815.
32. Törnroth-Horsefield, S., Gourdon, P., Horsefield, R., Brive, L., Yamamoto, N., Mori, H., Snijder, A., and Neutze, R. (2007) *Structure* 15: 1663-1673.
33. Koronakis, V., Sharft, A., Koronakis, E., Luisi, B., and Hughes, C. (2000) Crystal structure of the bacterial membrane protein TolC central to multidrug efflux and protein export. *Nature* 405: 914-919.
34. Wong, K.K.Y., Brinkman, F.S.L., Benz, R.S., and Hancock, R.E.W. (2001) Evaluation of a structural model of *Pseudomonas aeruginosa* outer membrane protein OprM, and efflux component involved in intrinsic antibiotic resistance. *J Bacteriol* 183: 367-374.
35. Zgurskaya, H.I., and Nikaido, H. (1999) Bypassing the periplasm: Reconstitution of the AcrAB multidrug efflux pump of *Escherichia coli. Proc Nat Acad Sci USA* 96: 7190-7195.
36. Mikolosko, J., Bobyk, K., Zgurskaya, H., and Ghosh, P. (2006) Conformational flexibility in the multidrug efflux system protein AcrA. *Structure* 14: 577-587.
37. Takatsuka, Y., and Nikaido, H. (2009) Covalently linked trimer of the AcrAB multidrug efflux pump provides support for the functioning rotating mechanism. *J Bacteriol* 190: 1729-1737.
38. Nies, D.H. (2003) Efflux-mediated heavy metal resistance in prokaryotes. *FEMS Microbiol Rev* 27: 313-339.
39. Nies, D.H. (1999) Microbial heavy-metal resistance. *Appl Microbiol Biotechnol* 51: 730-750.
40. Siezen, R.J., and Wilson, G. (2008) Unpublished but public microbial genomes with biotechnological relevance. *Microb Biotechnol* 1: 202-207.
41. Li, X.Z., Zhang, L., and Poole, K. (1998) Role of the multidrug efflux systems of *Pseudomonas aeruginosa* in organic solvent tolerance. *J Bacteriol* 180: 2987-2991.
42. Poole, K. (2004) Efflux pumps. In *Pseudomonas* vol. 1 pp. 635-674. Edited by Juan L. Ramos. Kluwer Acad. New York.
43. Köhler, T., Michea-Hamzehpour, M., Henze, U., Gotoh, N., Curty, L.K., and Pechere, J.C. (1997) Characterization of MexE-MexF-OprN, a positively regulated multidrug efflux system of *Pseudomonas aeruginosa. Mol Microbiol* 23: 345-354.
44. Ramos, J.L., Duque, E., Godoy, P., and Segura, A. (1998) Efflux pumps involved in toluene tolerance in *Pseudomonas putida* DOT-T1E. *J Bacteriol* 180: 3323-3329.
45. Poole, K., Gotoh, N., Tsujimoto, H., Zhao, Q., Wada, A., Yamasaki, T., et al., (1996) Overexpression of the *mexC-mexD-oprJ* efflux operon in *nfxB*-type multidrug-resistant strains of *Pseudomonas aeruginosa. Mol Microbiol* 21: 713-724.
46. Rojas, A., Duque, E., Mosqueda, G., Golden, G., Hurtado, A., Ramos, J.L., and Segura, A. (2001) Three efflux pumps are required to provide efficient tolerance to toluene in *Pseudomonas putida* DOT-T1E. *J Bacteriol* 183: 3967-3973.
47. Fernández, M., E. Duque, P. Pizarro-Tobías, P. van Dillewijn, R. Wittich, and J.L. Ramos. 2009. Microbial responses to xenobiotic compounds. Identification of genes that allow *Pseudomonas putida* KT2440 to cope with 2,4,6-trinitrotoluene. *Microbial Biotechnol* 2:287-294.
48. Jeannot, K., Elsen, S., Khler, T., Attree, I., van Delden, C., Plsiat, P. (2008) Resistance and virulence of *Pseudomonas aeruginosa* clinical strains overproducing the MexCD-OprJ efflux pump. *Antimicrobial Agents Chemother* 52: 2455-2462.
49. Kobayashi, K., Tsukagoshi, N., and Aono, R. (2001) suppression of hypersensitivity of *Escherichi coli* acrB mutant to organic solvents by integrational activation of the *acrEF* operon with the IS/1 or IS2 element. *J Bacteriol* 183: 2646-2653.
50. Ma, D., Cook, D.N., Alberti, M., Pon, N.G., Nikaido, H., and Hearst, J.E. (1993) Molecular cloning and characterization of *acrA* and *acrE* genes of *Escherichi coli. J Bacteriol* 175: 6299-6313.
51. Magnet, S., Courvalin, P., and Lambert, T. (2001) Resistance-nodulation-cell division type efflux pump involved in aminoglycoside resistance in *Acinetobacter baumannii* strain BM4454. *Agents Chemother* 45: 3375-3380.
52. Chau,S.L., Chu, Y.W., and Houang, E.T.S. (2004) Novel resistance-nodulation-cell division efflux system AdeDE in *Acinetobacter* genomic DNA group 3. *Antimicrob Agents Chemother* 48: 4054-4055.
53. Sthler, F.N., Odenbreit, S., Haas, R., Wilrich, J., van Vliet, A.H.H., Kusters, J.G., Kist, M., and (2006) *Infect and Imm* 74:3845-3852.
54. McHugh, G.L., Moellering, R.C., Hopkins, C.C., and Swartz, M.N. (1975) *Salmonella typhimurium* resistant to silver nitrate, chloramphenicol, and ampicillin. *Lancet* 1: 235-240.
55. Gupta, A., Matsui, K., Lo, J.F, and Silver, S. (1999) Molecular basis for resistance to silver cations in *Salmonella. Nat Med* 5: 183-188.
56. Silver, S. (2003) Bacterial silver resistance: molecular biology and uses and misuses of silver compounds. *FEMS Microbiol Rev* 27: 341-353.
57. Bolhius, H., Molenaar, D., Poelarends, G., van Veen, H.W., Poolman, B., Driesen, A.J., and Konings, W.N. (1994) Proton motive force-driven and ATP-dependent drug extrusion systems in multi-drug-resistant *Lactococcus lactis. J Bacteriol* 176: 6957-6964.
58. Li, X.Z., Nikaido, H., and Poole, K. (1995) Role of mexA-mexB-oprM in antibiotic efflux in *Pseudomonas aeruginosa. Antimicrob Agents Chemother* 39: 1948-1953.
59. Tokunaga, H., Mitsuo, K., Ichinose, S., Omori, A., Ventosa, A., Nakae, T., and Tokunaga, M. (2004) Salt-inducible multidrug efflux pump protein in the moderately halophilic bacterium *Chromohalobacter* sp. *Appl Environ Microbiol* 70: 4424-4431.
60. Abril, M.A., Michn, C., Timmis, K.N., and Ramos, J.L. (1989) Regulator and enzyme specificities of the TOL plasmid-encoded upper pathway for degradation of aromatic hydrocarbons and expansion of the substrate range of the pathway. *J Bacteriol* 171: 6782-6790.
61. Amsterdam, D. (1991) Antibiotics in laboratory medicine, In Lorian (ed.)p. 72-78. The Williams & Wilkins Co., Baltimore, Md. Meguro, N., Kodama, Y., Gallegos, M.T., and Watanabe, K. (2005) Molecular characterization of resistance-nodulation-division transporters from solvent- and drugresistance bacteria in petroleum-contaminated soil. *Microbiology* 71: 580-586.
62. Meguro, N., Kodama, Y., Gallegos, M.T., and Watanabe, K. (2005) Molecular characterization of resistance-nodulation-division transporters from solvent- and drug resistance bacteria in petroleum-contaminated soil. *Microbiology* 71: 580-586.
63. Terán, W., Felipe, A., Segura, A., Rojas, A., Ramos, J.L., and Gallegos, M.T. (2003) Antibioticdependent induction of *Pseudomonas putida* DOT-T1E TtgABC efflux pump is mediated by the drug binding repressor TtgR. *Antimicrob Agents Chemother* 47: 3067-3072.
64. Bernal, P., Muñoz-Rojas, J., Hurtado, A., Ramos, J.L., and Segura, A. (2007) A *Pseudomonas putida* cardiolipin synthesis mutant exhibits increased sensitivity to drugs related to transport functionality. *Environ Microbiol* 9: 1135-1145.
65. Nelson et al., (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ Microbiol* 4:799-808.
66. Duque, E., Segura, A., Mosqueda, G., and Ramos, J.L. (2001) Global and cognate regulators control the expression of the organic solvent efflux pumps TtgABC and TtgDEF of *Pseudomonas putida. Mol Microbiol* 39: 1100-1106.
67. Ausubel, F.M., Brent, R., Kingston, R.E., Moore, D.D., Seidman, J.G., Smith, J.A., and Struhl, K. (1991) Current protocols in molecular Biology. Wiley, New York, NY.

Suppl. Table 1. RND efflux pumps chosen for seed construction. All these pumps had been characterized through *in vivo* or *in vitro* assays.

RND protein	Protein description	Organism	Length (aa)	Accesion number
IfeB	Transmembrane efflux pump protein	*Agrobacterium tumefaciens*	1046	O68441
CzcA	Cobalt-zinc-cadmium resistance protein	*Ralstonia metallidurans* CH34	1063	P13511
NolG	Nodulation protein	*Sinorhizobium meliloti*	1065	P25197
CnrA	Nickel and cobalt resistance protein	*Ralstonia metallidurans* CH34	1076	P37972
MdtB	Multidrug resistance protein	*Escherichia coli* K-12	1040	P76398
MdtC	Multidrug resistance protein	*Escherichia coli* K-12	1025	P76399
MtrD	RND protein	*Neisseria gonorrhoeae*	1067	Q51073
MexD	RND family exporter	*Pseudomonas aeruginosa* PAO1	1043	Q51396
CmeB	RND efflux system, inner membrane transporter	*Campylobacter jejuni* RM1221	1040	Q5HWA1
EefB	Inner membrane protein	*Enterobacter aerogenes*	1035	Q8GC83
AdeB	RND protein	*Acinetobacter baumannii*	1035	Q93E19
MexF	RND multidrug efflux transporter	*Pseudomonas aeruginosa* PAO1	1062	Q9I0Y8
ArpB	Antibiotic efflux pump membrane transporter	*Pseudomonas putida* S12	1050	Q9KJC2
SmeB	Multidrug efflux pump	*Stenotrophomonas maltophilia*	1049	Q9RBY8
SilA	Putative cation efflux system protein	*Salmonella typhimurium*	1048	Q9ZHC9
MexY	Multidrug efflux pump	*Pseudomonas aeruginosa* PAO1	1046	Q9ZNG8

Suppl. Table 2. RND efflux pumps identified in databanks with the corresponding identification number, score and length. Data obtained on July 14th, 2009. (Extract. Complete table has 32 pages)

Q9F7M0_PRB01	Predicted cation efflux system (AcrB/AcrD/AcrF family)	53.704	1020
Q5LB43_BACFN	Putative ABC transport system, membrane protein	54.885	1074
Q64RI7_BACFR	Drug efflux protein	54.885	1067
Q6IVS4_9GAMM	Predicted cation efflux system	58.071	1020
Q6Q918_9GAMM	Predicted cation efflux system	59.217	1020
Q3NYY4_9GAMM	Similar to Cation/multidrug efflux pump precursor	61.934	1018
Q3P6H4_9GAMM	Similar to Cation/multidrug efflux pump precursor	62.472	1009
Q82WW7_NITEU	Acriflavin resistance protein	65.899	1040
Q30TY0_THIDN	Acriflavin resistance protein	66.862	1039
Q442D7_SOLUS	Acriflavin resistance protein precursor	67.871	1033
Q7UGN1_RHOBA	Cation/multidrug efflux pump	68.238	1109
Q3CRH6_ALTAT	Acriflavin resistance protein	68.879	1068
Q3H8V6_TRIER	Acriflavin resistance protein precursor	69.636	1037
Q5LUY1_SILPO	Transporter, AcrB/AcrD/AcrF family	70.358	1010
Q8RG07_FUSNN	Acriflavin resistance protein D	70.496	1022
Q3QS85_9RHOB	Acriflavin resistance protein precursor	70.507	1137
Q64R25_BACFR	AcrB/D/F family transporter	71.321	1011
Q5LAM7_BACFN	Putative AcrB/AcrD/AcrF family membrane protein	71.413	1011
Q3NSZ3_SHEFR	Acriflavin resistance protein	71.894	1053
Q35ZX0_9GAMM	AcrB/AcrD/AcrF family protein precursor	72.1	1026
Q36NZ0_MARHY	AcrB/AcrD/AcrF family protein precursor	72.112	1008
Q36U61_MARHY	Cation efflux system protein XF2083 precursor	72.169	1011
Q87J90_VIBPA	Putative efflux protein	72.226	1020
Q8EC65_SHEON	AcrB/AcrD/AcrF family protein	72.467	1026
Q30SF3_THIDN	Acriflavin resistance protein	72.616	1044
Q2ZCB0_9GAMM	AcrB/AcrD/AcrF family protein precursor	72.673	1026
Q5LX87_SILPO	Transporter, AcrB/AcrD/AcrF family	72.742	1133
Q33W13_9GAMM	Acriflavin resistance protein precursor	72.742	1023
Q92Y52_RHIME	Probable cation efflux protein	72.765	1005
Q3FNR0_9BURK	Acriflavin resistance protein precursor	72.937	1151
Q40HQ2_9RHOB	Acriflavin resistance protein	73.281	1131
Q36B16_9GAMM	AcrB/AcrD/AcrF family protein precursor	73.338	1026
Q35W96_9GAMM	Putative efflux protein	73.407	1035
Q36AT2_9GAMM	Putative efflux protein	73.51	1035
Q6W1F3_RHISN	Acriflavin resistance plasma membrane protein	73.522	1038
Q2Z6D8_9GAMM	Putative efflux protein	73.602	1035
Q3QVF2_9RHOB	Acriflavin resistance protein	73.9	1267
Q36J71_MARHY	Cation efflux system protein XF2083 precursor	73.98	1012
Q47VP1_COLP3	AcrB/AcrD/AcrF family protein	74.255	1074
Q2WZQ1_9GAMM	AcrB/AcrD/AcrF family protein precursor	74.404	1025
Q30QF5_THIDN	Acriflavin resistance protein	74.461	1005
Q8A731_BACTN	Transporter, AcrB/D/F family	74.633	1012
Q3Q996_9GAMM	Acriflavin resistance protein precursor	74.645	1022
Q5LR00_SILPO	Transporter, AcrB/AcrD/AcrF family	74.828	1259
Q8A899_BACTN	Transporter, AcrB/D/F family	75.057	1010
Q33NZ0_9GAMM	Acriflavin resistance protein	75.195	1060
Q3NUC7_SHEFR	Acriflavin resistance protein precursor	75.436	1021
Q31EX8_THICR	Acriflavin resistance protein	75.86	1021
Q3QWF5_9RHOB	Acriflavin resistance protein	75.883	1016
Q484D7_COLP3	AcrB/AcrD/AcrF family protein	75.997	1054
Q40D90_9RHOB	Acriflavin resistance protein	75.997	1335
Q315C5_DESDG	Transporter, AcrB/D/F family	76.043	1010
Q7UJY1_RHOBA	Probable RND efflux transporter	76.077	1115
Q2Y7Y4_NITMU	Acriflavin resistance protein	76.192	1098
Q8U880_AGRT5	Cation efflux system protein (AGR_L_1281p)	76.444	1036

Suppl. Table 3. Isolated 1 ethidium bromide resistant bacteria strains.
The names of the bacteria most closely related to isolates based on 16S rDNA nucleotide sequence identity are listed in column 2. Column 3 shows the percentage of identity between the two strains and column 4 lists the number of nucleotides different from those of the corresponding sequence deposited in Genebank.

Strain	Closest strain	Nucleotide identity (%)	Number of nucleotides different
EB(1)	*Shewanella putrefaciens*	99	4 out of 1501
EB(2)	*Kluyvera cryocrescens* WAB1904	99	3 out of 1476
EB(3)	*Klebsiella oxytoca* GR6	99	4 out of 1498
EB(4)	Uncultured bacterium clone EV821FW101601SAC11	99	4 out of 1497
EB(5)	Uncultured g-proteobacterium BioIuz K33	99	8 out of 1511
EB(7)	*Klebsiella oxytoca* GR6	99	4 out of 1498
EB(8)	*Bacterium G3 GreenLake*	100	0 out of 1495
EB9	*Kluyvera cryocrescens*	99	3 out of 1476
EB(10)	*Empedobacter brevis* LMG4011T	99	5 out of 1487
EB(11)	*Klebsiella* sp. LB-2	99	2 out of 1528
EB(12)	*Pseudomonas* sp. BWDY-9 *Pseudomonas mendocina* PC10	99 99	6 out of 1498 7 out of 1481

Suppl. Table 4. Growth of the 1 isolated EtBr resistant strains in the presence of different carbon sources.
Cells were grown overnight in LB medium. The following day, cells were washed with 50 mM phosphate buffer, and resuspended in M9 medium at a cell density corresponding to an OD_{660} of 0.1. Aliquots were supplemented with 10 mM of each of the different C sources, and cells were incubated at 30º C at 200 rpm, and growth was measured as turbidity after 24 h incubation. (+++) $OD_{660} > 1$; (++) $1 < OD_{660} < 0.5$; (+) $0.5 < OD_{660} < 0.2$; (+/-) $0.2 < OD_{660} < 0.1$; (-) $OD_{660} < 0.1$

	EB-1	EB-2	EB-3	EB-4	EB-5	EB-7	EB-8	EB-9	EB-10	EB-11	EB-12
LB	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++	+++
Glucose	-	+++	+++	-	+++	+++	+++	+++	-	+++	+++
Sucrose	-	+++	+++	-	+++	+++	+++	+++	-	+++	+++
Maltose	-	+++	+++	-	+++	+++	+++	+++	-	+++	+++
Xylose	-	+++	+++	-	-	+++	+++	+++	-	+++	+++
Fructose	-	+++	+++	-	++	+++	+++	+++	-	+++	+++
Glycerol	-	+++	+++	-	++	+++	+++	+++	+	++	+++
Benzoate	-	-	-	+++	-	-	-	-	-	++	+/-
Succinate	+	++	++	+++	++	++	+	++	-	++	++
Citrate	-	+++	+++	-	+	+++	+++	+++	-	+++	+++
Salicilate	-	-	-	-	-	-	-	-	-	-	-
p-Hydroxybenzoate	-	++	++	-	-	++	++	++	-	+++	+++
Quinate	-	+++	+++	+/-	-	+++	+++	+++	-	+++	+++
Vanillate	-	-	-	-	-	-	-	-	-	-	-
Glutamate	+/-	+/-	+	+++	+++	+	+++	+/-	+/-	++	+++
Glutamine	-	-	-	++	-	-	+/-	-	-	+/-	++
Lysine	-	-	-	-	-	-	-	-	-	-	-

Suppl. Table 5. Susceptibility of EtBr resistant 1 clones and *P. putida* DOT-T1E strains to different antimicrobial agents.

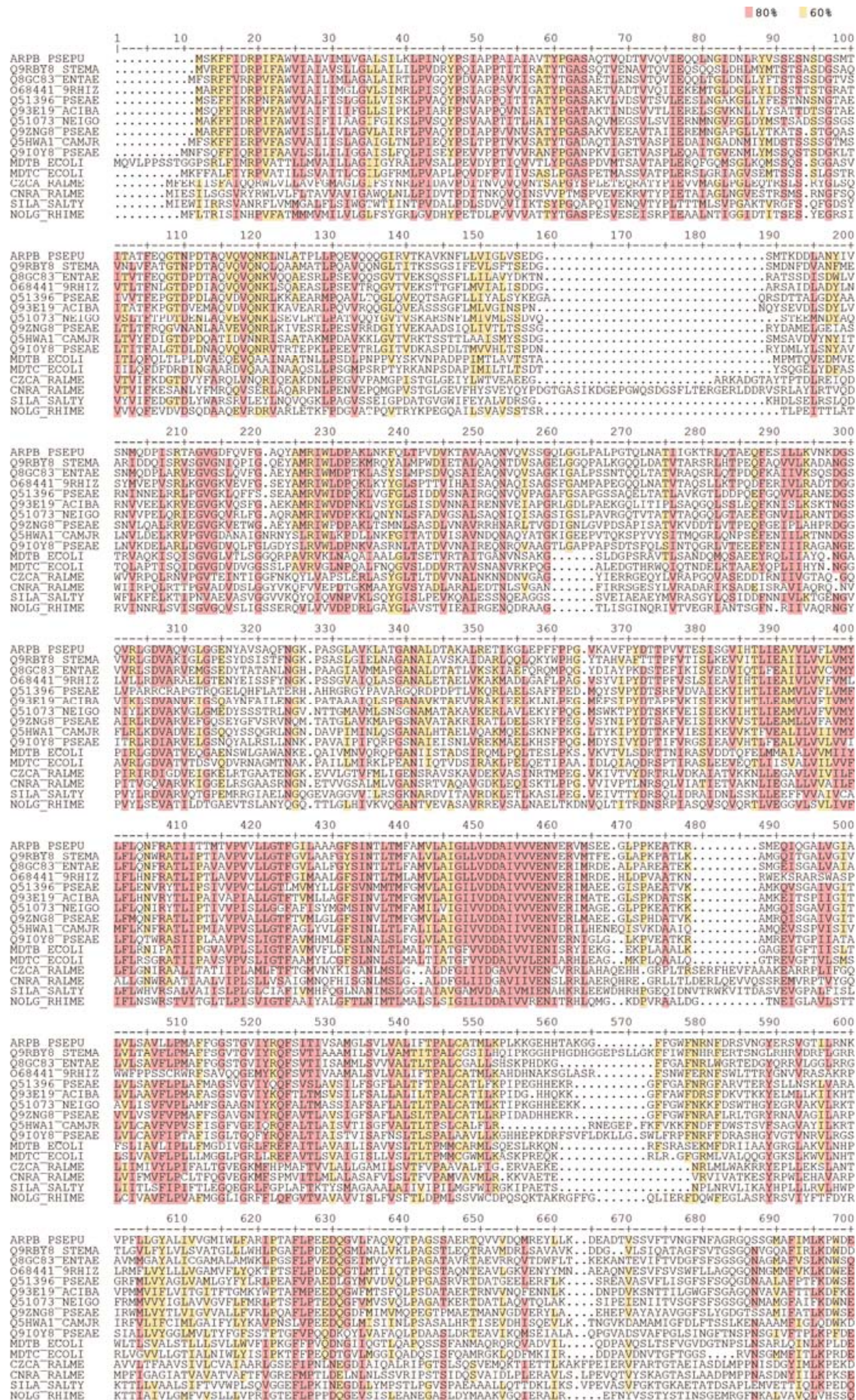
Strain	MIC$_{90}$ (ug/ml)													
	Ap	**Cb**	**Cm**	**Ctx**	**Gm**	**Km**	**Nal**	**Neo**	**Nor**	**Nov**	**Sm**	**Tc**	**Tcl**	**EtBr**
DOT-T1E	1000	1500	1000	3.125	25	3.125	125	2.5	3.125	2000	18.75	15	>1000	>1000
EB(1)	62.5	23.43	1.563	0.078	50	15.63	0.3125	25	0.78	250	37.5	0.937	12.5	500
EB(2)	250	187.5	6.25	0.3125	50	15.63	500	25	1.562	62.5	750	1.875	0.078	>1000
EB(3)	500	187.5	3.125	0.039	100	12.5	3.9	10	0.313	62.5	18.75	0.937	0.078	>1000
EB(4)	15.62	30	31.25	12.5	25	6.25	1.25	10	3.125	6.125	9.375	1.875	0.00976	>1000
EB(5)	>2000	3000	3.9	7.81	1000	>2000	125	>100	1.562	25	1500	0.3125	6.25	500
EB(7)	500	375	3.125	0.039	50	31.25	1.562	25	0.097	31.25	23.43	1.875	0.039	>1000
EB(8)	250	93.75	31.25	0.313	50	500	125	50	0.78	62.5	187.5	3.75	12.5	>1000
EB(9)	500	375	7.81	0.78	50	15.63	3.9	25	0.1563	125	23.43	1.875	0.039	>1000
EB(10)	125	187.5	31.25	50	>1000	2000	3.9	>100	12.5	15.63	750	15	1.563	>1000
EB(11)	1000	750	12.5	0.0195	50	15.63	1.95	12.5	0.097	62.5	11.72	0.937	0.078	>1000
EB(12)	500	93.75	125	6.25	12.5	3.125	7.81	2.5	0.156	31.25	9.375	1.875	0.313	>1000

Abbreviations: Ap, ampicillin; Cb, carbenicillin; Cm, chloramphenicol; Ctx, cefotaxime; Gm, gentamycin; Km, kanamycin; Nal, nalidixic acid; Neo, neomycin; Nor, norfloxacin; Nov, novobiocin; Sm, streptomycin; Tc, tetracycline; Tcl, triclosan; EtBr, ethidium bromide.

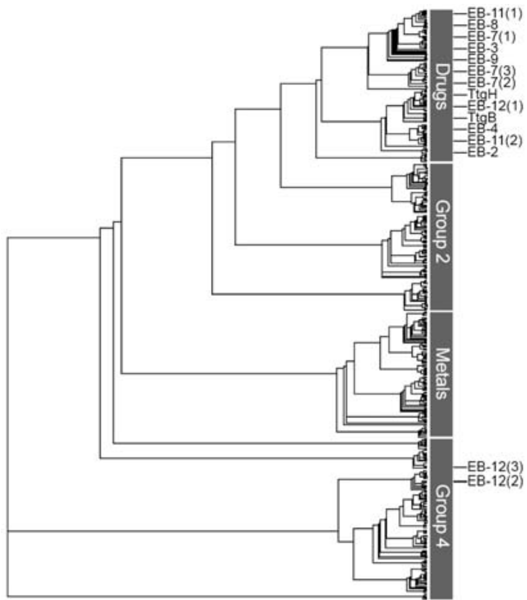
Suppl. Table 6. *Pseudomonas putida* KT2440 RND efflux pumps identified in the genome.

Locus name	RND efflux pump	Annotation	Length (aa)	UniProt AC
PP0043	CzcA	Putative cobalt/zinc/cadmium efflux RND transporter, permease protein, CzcA family	1053	Q88RT6
PP0906		Multidrug efflux RND transporter, putative	1014	Q88PE4
PP1385	TtgB	Multidrug/solvent RND transporter TtgB	1050	Q88N31
PP1517		RND efflux transporter, hydrophobe/amphiphile efflux-1 (HAE1) family	1021	Q88MQ3
PP2065		Multidrug efflux RND transporter	1027	Q88L70
PP2411		Putative cobalt/zinc/cadmium resistance protein CzcA	1048	Q88J31
PP2818	MexD	Multidrug efflux RND transporter MexD	1042	Q88K81
PP3302		RND efflux transporter, hydrophobe/amphiphile efflux-1 (HAE1) family	1030	Q88HQ1
PP3426	MexF	Multidrug efflux RND transporter MexF	1059	Q88HD4
PP3456		Multidrug efflux RND transporter	1042	Q88HA4
PP3583	MdtC	Multidrug RND efflux transporter, permease protein MdtC	1035	Q88GY2
PP3584	MdtB	RND efflux transporter, transmembrane protein MdtB, authentic frameshift	-	-
PP5173		RND efflux transporter, hydrophobe/amphiphile efflux-1 (HAE1) family	1017	Q88CK7
PP5387	CzcA	Heavy metal RND efflux transporter, CzcA family	1052	Q88BZ6

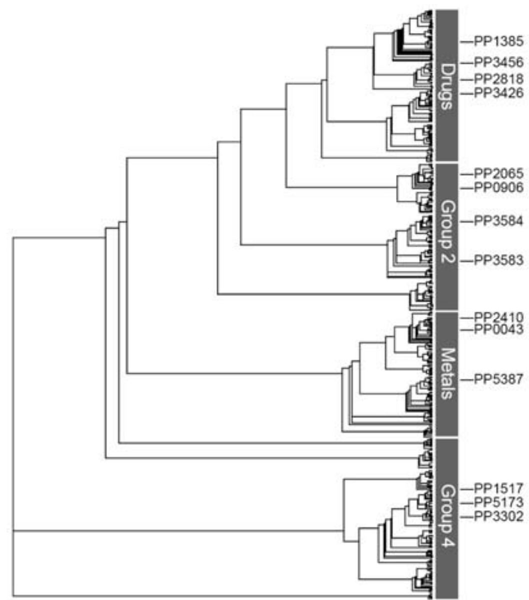
Suppl. Figure 1. Multialignment of the 16 RND efflux pumps initially chosen for the construction of an RND efflux pump profile.



```
                    710       720       730       740       750       760       770       780       790       800
ARPB_PSEPU   RS.KENSVFALAQRAQ.QHFFTFRDAMVFAFAPPAVLELGNATGFDVFLQDRGGVGHAKLMEARNQFLAKAAQSKILSAVRPNG..LNDEPQYQLTIDDE
Q9RBY8_STEMA R...KDDADTIAARLT.RAMASVPDAQVFITSPPAILGLGDAGGFTLELQDEGGAGHAAAVAARNTLLKEAAKDPKLVNVRYAS..LEDAPVYAVKVDDA
Q8GC83_ENTAE RKGDENTAQAIALRAN.KELGAIRDATLFAMTPPSVDGLGQSNGFTFELMASGGTDRDSLMKLRSQLLAAANQSGELQSVRAND..LPQMPQLQVDIDNN
O68441_9RHIZ RKGKENSAQAIAARAFGPLMGGIKEAIVVPLVPPAVTELGNSNGFTAFLQARSGQSHEQLLEARNMLLGLAAQSPKLMAVRPNG..VEDASQFELNIDWG
Q51396_PSEAE RGAEQSSAAEIA..ALNEHFALPDDGTVMAVSPPPINGLGNSGGFALRLMDRSGVGREALLQARDTLLGEIQTNPKFLYAMMEG..LAEAPQLRLLIDRE
Q93E19_ACIBA R..TSSASKMTS..DVNSSMANSTEGETMAVLPPAIDELGTFSGFSLRLQDRANLGMPALLAAQDELMAMAAKNKKFYMVWNEG..LPQGDNISLKIDRE
Q51073_NEIGO RTASGSDAVAVAGKLTGMMMGTLKDGFGIAVVPPPILELGNGSGLSINLQDRNNTGHTGIAGEGNELIQKMRASGLFDPSTVRAGGLEDSPQLKIDINRA
Q9ZNG8_PSEAE RR.EASQHVGAIVERINQRFAGLPNRTVYAMNSPPLPDLGSTSGFDFRLQDRGGVGYEALVKARDQLLARAAEDPRLANVMFAG..QGEAPQIRLDIDRR
Q5HWA1_CAMJR R...NVSADQIIAELN.KKFAFDRNASSVFIGLPPIPGLSITGGFEMYVQNKSGKSYDEIQKDVNKLVAVANQRKELSR..VRTTLDTTFPQYKLIIDRD
Q9I0Y8_PSEAE RKDPSQSAGAIAAALN.AKYADIQDAYIAIFPPPPVQGLGTIGGFRLQIEDRGNQGYEELFKQTQNIITKARALPELEPSSVFSSYQVNVPQIDADIDRE
MDTB_ECOLI   RD.....DRVQKVIARLQTAVDKVPGVDLFLQPTQDLTIDTQVSRTQYQFTLQATSLDALSTWVPQLMEKLQQLPQLSDVSSDW..QDKGLVAYVNVDRD
MDTC_ECOLI   RS.....ETAQQIIDRLRVKLAKEPGANLFLMAVQDIRVGGRQSNASYQYTLLSDDLAALREWEPKIRKKLATLPELADVNSDQ..QDNGAEMNLVYDRD
CZCA_RALME   WPEPKKTHAELLSAIQEEAGKIPGNNYEFSQPIQLRFNELISGVRSDVAVKIFGDDNNVLSETAKKVSAVLQGIPGAQEVKVEQ..TTGLPMLTVKIDRE
CNRA_RALME   WPEGVTTKEQVIERIREKTAPMVGNNYDVTQPIEMRFNELIGGVRSDVAVKVYGENLDELAATAQRIAAVLKKTPGATDVRVPL..TSGFPTFDIVFDRA
SILA_SALTY   WRPG..MTIDKIIDELDRTVRLPGLANLWVPPIRNRIDMLSTGIKSPIGIKVSGTVLSDIDATAQSIEAVAKTVPGVVSVLAER..LEGGRYIDIDINRE
NOLG_RHIME   RR.........LKTAETLGPIRRRLSRIAGLEISVGQRSEVVGSIKPLQLSILGDGDEELRRISDHITSVLAAIPGATEIESSI..EKLRPTLAVRVRRE

                    810       820       830       840       850       860       870       880       890       900
ARPB_PSEPU   RASALGVTIADINNTLSIALGASYVNDFIDR.GRVKKVYIQGEPSARMSPEDLQKWYVRNGAG......EMVPFSSFAKGEWTYGSPKLSRYNGVEAMEI
Q9RBY8_STEMA KAQAMGVNPQDVNDTLNAALGGDFVNNFIYK.GRIKKVFIQGTAEARMQPQDIERWSVRNQAG......QMVPLSSLISTHWTSAPAAVQRYNGISAMEI
Q8GC83_ENTAE KAVSLGLSLSDVTDTLSSAWGGTYVNDFIDR.GRVKKVYIQGESDARAVPSDLGKWFVRASDN......SMTPFTAFATTHWEYGPESLVRYNGSAAFEI
O68441_9RHIZ KAGAVGLSAADVGSFLTTVWSSSYVNDFLYE.GRMKRVYVQGEPLARTGPEDLALWRVPNANG......DFVDLSTIASQNWVFGPQQVIRYDALPAMSI
Q51396_PSEAE KARALGVSFETISGTLSAAFGSEVINDFTNA.GRQQRVVIQAEQGNRMTPESVLELYVPNAAG......NLVPLSAFVSVKWEEGPVQLVRYNGYPSIRI
Q93E19_ACIBA KLSAFGVKFSDVSDIISTSMGSMYINDFPNQ.GRMQQVIVQVEAKSRMQLKDILNLKVMGSSG......QLVSLSEVVTPQWNKAPQQYNRYNGRPSLSI
Q51073_NEIGO AAAAQGISFADIRTALASALSSSYVSDFPNQ.GRLQRVMVQADEDAKMQPADILNLTVPNKSG......VAVPLSTIATVSWENGTEQSVRFNGYPSMKL
Q9ZNG8_PSEAE KAETLGVSMDEINTTLAVMFGSDYIGDFMHG.SQVRKVVVQADRRKRLGIDDIGRLHVRNEQG......EMGAAGDVRQGRLDPRPAATDPLQRLSLVQP
Q5HWA1_CAMJR KLKHYNLNMQDVFNTMNATIGTYYVNDFSML.GKNFQVNIRAKGDFRNTQDALKNIFVRSNDG......KMIPLDSFLTLQRSSGPDDVKRFNLFPAAQV
Q9I0Y8_PSEAE KAKTHGVAISDIFDTLQVYLGSLYANDFNRF.GRTYQVNVQAEQQFRLEPEQIGQLKVRNNLG......EMVPLASFIKVSDTSGPDRVMHYNGFITAEL
MDTB_ECOLI   SASRLGISMADVDNALYNAFGQRLISTIYTQ.ANQYRVVLEHNTENTPGLAALDTIRLTSSDG......GVVPLSSIAKIEQRFAPLSINHLDQFPVTTI
MDTC_ECOLI   TMARLGIDVQAANSLLNNAFGQRQISTIYQP.MNQYKVVMEVDPRYTQDISALEKMFVINNEG......KAIPLSYFAKWQPANAPLSVNHQGLSAASTI
CZCA_RALME   KAARYGLNMSDVQDAVATGVGGRDSGTFFQG.DRRFDIVVRLPEAVRGEVEALRRLPIPLPKGVDART.TFIPLSEVATLEMAPGPNQISRENGKRRIVI
CNRA_RALME   AIARYGLTVKEVADTISTAMAGRPAGQIFDG.DRRFDIVIRLPGEQRENLDVLGALPVMLPLSEGQAR.ASVPLRQLVQFRFTQGLNEVSRDNGKRRVYV
SILA_SALTY   KASRYGMTVGDVQLFVSSAIGGAMVGETVEG.VARYPINIRYPQDYRNSPQALKQMPILTPMK......QQITLGDVADINVVSGPTMLKTENARPASWI
NOLG_RHIME   AASDLGVSIATIGDTLRSLVAGDAISVWNSPDGETHDVVVRLPAAGRENAAQLRNLPIATARMDDNGKPIMVLLDQVADVVESTAPAQITRKDLSRDIRI

                    910       920       930       940       950       960       970       980       990      1000
ARPB_PSEPU   LGAPAPGYSTGEAMAEVERIAGELP.SGIGFSWTGMSYEEK.LSGSQMPALFALSVLFVFLCLAALYESWSIPIAVVLVVPLGIIGALIATSLR......
Q9RBY8_STEMA TGQPAPGVSSGEAMAEIARLADTLP.EGFSHAWSDMAYQEQ.LSGNQAPMLYAISLLFVFLCLAALYESWAVPFAVMLAVPVGMFGAVLMMNLR......
Q8GC83_ENTAE QGENAAGFSSGAAMDKMEKLADSLP.AGSTWAWSGISLQEK.LASGQAMSLYAISILVVFLCLAALYESWSVPFSVIMVIPLGLLGAALAATLR......
O68441_9RHIZ EGSAAPGFSSGEALAEIEALAAKLP.PGFSLQWTGMSLEEK.DAGAGALPLYGLALATMFLCLAALYESWTIPVAVLLAMPVGILGALLGAWIG......
Q51396_PSEAE VGDAAPGFSTGEAMAEMERLASQLP.AGIGYEWTGLSYQEK.VSAGQATSLFALAILVVFLLLVALYESWSIPLSVMLIVPIGAIGAVLAVMVS......
Q93E19_ACIBA AGIPNFDTSSGEAMREMEQLIAKLP.KGIGYEWTGISLQEK.QSESQMAFLLGLSMLVVFLVLAALYESWAIPLSVMLVVPLGIFGAIIAIMSR......
Q51073_NEIGO SASPATGVSTGQAMEAVQKMVDELG.GGYSFEWGGQSSEEA.KGGSQTLILYGLAVAAVFLVLAALYESWSIPLAVILVIPLGLIGAAAGVTGRNLFEGL
Q9ZNG8_PSEAE RGPGRAGLQQREAMQAMEQLMQGTA.RGIRPRVVRPVLRRTPVAGAQAPALFALSVLIVFLALAALYESWSIPLAVILVVPLGVLGALLGVSLR......
Q5HWA1_CAMJR QGQPAPGYTSGQAIEAIAQVAKETLGDDYSIAWSGSAYQEV.SSKGTASYAFALGMVFVFLILAAQYERWLIPLAVVTAVPFAVFGSFLLVYLR......
Q9I0Y8_PSEAE NGAPAAGYSSGQAQAAIEKLLKEELPNGMTYEWTELTYQQI.LAGNTALFVFPLCVLLAFLVLAAQYESWSLPLAVILIVPMTLLSAITGVILA......
MDTB_ECOLI   SFNVPDNYSLGDAVQAIMDTEKTLNLPVDITTQFQGSTLAFQSALGSTVWLIVAAVVAMYIVLGILYESFIHPITILSTLPTAGVGALLALLIAG.....
MDTC_ECOLI   SFNLPTGKSLSDASAAIDRAMTQLGVPSTVRGSFAGTAQVFQETMNSQVILIIAAIATVYIVLGILYESYVHPLTILSTLPSAGVGALLALELFN.....
CZCA_RALME   SANVRGRDIGSFVPEAEAAIQSQVKIPAGYWMTWGGTFEQLQSATTRLQVVVPVALLLVFVLLFAMFNNIKDGLLVFTGIPFALTGGILALWIRG.....
CNRA_RALME   EANVGGRDLGSFVDDAAARIAKEVKLPPGMYIEWGGQFQNLQAATKRLAIIVPLCFILIAATLYMAIGSAALTATVLTAVPLALAGGVFALLLRG.....
SILA_SALTY   YVDARGRDMVSVVNDIKTAISEKVKLRPGTSVAFSGQFELLEHANKKLKLMVPMTVMIIFILLYLAFRRVDEALLILMSLPFALVGGIWFLYWQG.....
NOLG_RHIME   SSN.IEGRTLGDVVADLKAAMTKMDIPVGFRISFGGDAENLTESTAYALQSLAMAVIFIYIILASQFGSFIQPIAIIMTMPLSLMGVLLGLLFTG.....

                   1010      1020      1030      1040      1050      1060      1070      1080      1090      1100
ARPB_PSEPU   .....GLSNDVYFLVGLLTTIGLAAKNAILIVEFAKELHEQG..........RSLYDAAIEACRMRLRPIIMTSLAFILGVVPLTIASGAGAGSQHAIG
Q9RBY8_STEMA .....GLNNDVYFQVGLLTTIGLAAKNGILIVEFARILEQQG..........KSTREAILQAVYLRLRPIVMTSLAFLMGVLPLVFATGAGSAARRSLG
Q8GC83_ENTAE .....GLSNDVYFQVALLTTIGLSSKNAILIVEFAESAVDEG..........YSLSRAAIRAAQTRLRPIVMTSLAFIAGVLPLAIATGAGANSRVAIG
O68441_9RHIZ .....GQSNGVYFQVGLLTVVGLTGKNGIMIVEFARERMAQLG.........ESAFEAVCQAAKLRFRPILMTSLAFGLGVIPLVISTGAGAGARQAIG
Q51396_PSEAE .....GMSNDVYFKVGLITIIGLSAKNAILIVEFAKELWEQG..........HSLRDAAIEAARLRFRPIIMTSMAFILGVIPLALASGAGAASQRAIG
Q93E19_ACIBA .....GLMNDVFFKIGLITIIGLSAKNAILIVEFAKMLKEEG..........MSLIEATVAAAKLRLRPILMTSLAFTCGVIPLVIASGASSETQHALG
Q51073_NEIGO LGSVPSFANDIYFQVGFVTVMGLSAKNAILIIEFAKDLQAQG..........KSAVEAALEAARLRFRPIIMTSFAFILGVVPLYIAAGASSASQRAIG
Q9ZNG8_PSEAE .....GLPNDIYFKVGLITIIGLSAKNAILIIEVAKDHYQEG..........MSLLQATLEAARLRLRPIVMTSLAFGFGVVPLALSSGAGIRAQVAIG
Q5HWA1_CAMJR .....GFSNDIYFQTGLLLLIGLSAKNAILIVEFAMEERFKKG.........KGVFEAAVAAAKLRFRPIIMTSLAFTFGVLPMIFATGAGSASRHSLG
Q9I0Y8_PSEAE .....GSDNNIFTQIGLIVLVGLACKNAILIVEFAKDKQEE.G.........MDRVAAVLEACRLRLRPILMTSIAFIMGVVPLVISTGAGAEMRHAMG
MDTB_ECOLI   ......SELDVIAIIGIILLIGIVKKNAIMMIDFALAAEREQG.........MSPREAIYQACLLRFRPILMTTLAALLGALPLMLSTGVGAELRRPLG
MDTC_ECOLI   ......APFSLIALIGIMLLIGIVKKNAIMMVDFALEAQRHGN.........LTPQEAIFQACLLRFRPIMMTTLAALFGALPLVLSGGDGSELRQPLG
CZCA_RALME   ......IPMSITAAVGFIALCGVAVLNGLVMLSFIRSLREEG..........HSLDSAVRVGALTRLRPVLMTALVASLGFVPMAIATGTGAEVQRPLA
CNRA_RALME   ......IPFSISAAVGFIAVSGVAVLNGLVLISAIRKRLDDG..........MAPDAAVIEGAMERVRPVLMTALVASLGFVPMAIATGTGAEVQKPLA
SILA_SALTY   ......FHMSVATGTGFIALAGVAAEFGVVMLMYLRHAIEAHPELSRKETFTPEGLDEALYHGAVLRVRPKAMTVAVIIAGLLPILWGTGAGSEVMSRIA
NOLG_RHIME   ......STLNMFSMIGIMMLMGLVTKNAILLVDYSNLGVREGK.........SLRQSLADAGAVRLRPIVMTTLAMIFGMLPTALGLGEGGAQRAPMA

                   1110      1120      1130      1140      1150      1160      1172
ARPB_PSEPU   TGVIGGMISATVLAIFWVPLFFVAVSSLFGSKEPEKDVTPENPRYEAGQ.......................
Q9RBY8_STEMA TGVAGGTVASMVLGMFFVPLFYLLVRRLFPGRAPADATVPETSP............................
Q8GC83_ENTAE TGIIGGTLTATLLAVFFVPLFFVLVKRLFTRPRPSQE...................................
O68441_9RHIZ YATFFGTVTGTLLAIVFVPVFFVLISGIFSRRSVKAAEPKTA..............................
Q51396_PSEAE TGVIGGMLSATFLGVLFVPICFVWLLSLLRSKPAPIEQAASAGE............................
Q93E19_ACIBA TGVFGGMISATILAIFFVPVFFIFILGAVEKLFSSKKKISS...............................
Q51073_NEIGO TTVFWGMLVGTLLSVFLVPLFYVVVRKFFKETAHEHEMAVRHASKAGITGSDDKQY................
Q9ZNG8_PSEAE TGVLGGIVTATVLAVFLVPLFFLVVGRLFRLRKAPRTGNSPQIPTEQA........................
Q5HWA1_CAMJR TGLIGGMIAASTLAIFFVPLFFYLLENFNEWLDKKRGKVHE...............................
Q9I0Y8_PSEAE VAVFSGMIGVTFFGLLLTPVFYVLIRRFVENREARRAANDKGLPEVHA........................
MDTB_ECOLI   IGMVGGLIVSQVLTLFTTPVIYLLFDRLALWTKSRFARHEEEA.............................
MDTC_ECOLI   ITIVGGLVMSQLLTLYTTPVVYLFFDRLRLRFSRKPKQTVTE..............................
CZCA_RALME   TVVIGGILSSTALTLLVLPVLYRLAHRKDEDAEDTREPVTQTHQPDQGRQPA....................
CNRA_RALME   TVVIGGLVTATVLTLFVLPALCGIVLKRRTAGRPEAQAALEA..............................
SILA_SALTY   APMIGGMITAPLLSLFIIPAAYKLIWLRRHKKSVS.....................................
NOLG_RHIME   HAIIGGLISSTLLSLVFVPVVLTYLDAFAGRVRRWVPSPTGSNASAQHDGSDKTKTPACALTSQDQLGTTIM
```



Suppl. Figure 2. Location of RND efflux pumps amplified from EtBr resistant clones within the general phylogenetic tree of efflux pumps. The pumps identified were included in the phylogenetic tree according to their sequence identity. They are named on the basis of the microorganism from which they were amplified. TtgB and TtgH are RND efflux pumps of the *P. putida* DOT-T1E.



Suppl. Figure 3. Location of RND efflux pumps present in the genome of *P. putida* KT2440 within the general tree of efflux pumps. Pumps are named based on the PP number.

Capítulo 5

BacTregulators: a database of transcriptional regulators in bacteria and archaea

Bioinformatics

BIOINFORMATICS

Vol. 20 no. 16 2004, pages 2787–2791
doi:10.1093/bioinformatics/bth330

BacTregulators: a database of transcriptional regulators in bacteria and archaea

Manuel Martínez-Bueno[1,†], Antonio J. Molina-Henares[1,†], Eduardo Pareja[2], Juan L. Ramos[1] and Raquel Tobes[1,]*

[1]Department of Plant Biochemistry and Molecular and Cellular Biology, Estación Experimental del Zaidín, Consejo Superior de Investigaciones Científicas, E-18008 Granada, Spain and [2]Bioinformatics, Hospital Universitario Virgen de las Nieves, E-18014 Granada, Spain

Received on February 23, 2004; revised on April 27, 2004; accepted on May 10, 2004
Advance Access publication May 27, 2004

ABSTRACT

Motivation: The BacTregulators database is intended to collect and to integrate information on proteins belonging to defined families of transcriptional regulators in prokaryotes.

Results: The BacTregulators database currently contains data on two families of transcriptional regulators: AraC-XylS and TetR. The proteins included in the BacTregulators database have been identified by screening 123 genomes from archaea and bacteria and the SWISS-PROT and TrEMBL databases with profiles defining each family. As the result of an integration process, we have included 1326 different protein sequences from the AraC-XylS family and 1487 different protein sequences from the TetR family. The definition of an entry in BacTregulators is based on protein sequence, source organism, genome element and position in this genome element. The BacTregulators site allows the user to retrieve protein sequences, functional features and experimental evidence supporting the functions, references and the three-dimensional structure of the regulator when available. BacTregulators supplies an innovative tool that allows the researcher to obtain an integrated report that shows the data corresponding to other entries which are related by sequence similarity to the query entry. BacTregulators detects and classifies the regulators belonging to AraC-XylS and TetR families present in prokaryotic genomes, and thus contributes to a more accurate annotation of regulators in genomes. The information collected on each protein in the family can be useful to characterize a new regulator or compile information on the biological properties of a known regulator.

Availability: The BacTregulators is available at www.bactregulators.org

Contact: rtobes@bioinformatica.org

Supplementary information: www.bactregulators.org/supplementary

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

INTRODUCTION

The most important mechanisms that bacteria use to adapt their physiology to changing environmental conditions are based on regulation at the transcriptional level (Ramos *et al.*, 1997, 2001; Ishihama, 2000). Transcriptional regulators can fine-tune gene expression to face the specific environmental changes and stress conditions. This control of gene expression is achieved through the delicate interplay of sigma factors that confer promoter selectivity, and transcriptional regulators that modulate RNA polymerase activity (Ramos *et al.*, 1997; Gallegos *et al.*, 1997). Since 1990, we have been updating information on the AraC-XylS family of transcriptional regulators (Gallegos *et al.*, 1997), and we have compiled information on proteins of this family in the AraC-XylS database (Tobes and Ramos, 2002). In an effort to understand the regulatory networks of complex processes that involve regulators of different families, we decided to extend the database to incorporate new families of regulators. This new approach can be useful to define the regulatory networks that control microbial responses to different environmental challenges. In addition, the rapid appearance of newly sequenced prokaryotic genomes generates exponential growth in the number of protein sequences, but in contrast, the experimental acquisition of information about these proteins is relatively limited.

Another issue is that data from sequenced genomes and protein databases are often not integrated, so further efforts are needed in that direction. BacTregulators (www.bactregulators.org) has been created with the aim of integrating information about regulators from the NCBI microbial genome resource and from the SWISS-PROT and TrEMBL protein databases.

The BacTregulators database currently compiles data on two families of regulators: AraC-XylS and TetR. Information

on AraC-XylS family members has been updated and at present covers 1326 sequences. These sequences have been obtained after screening the bacterial and archaeal genomes included in the NCBI Refseq database (June 11, 2003) and the SWISS-PROT and TrEMBL (SPTR) protein databases (June 16, 2003). The sequences have been integrated in a non-redundant set.

The new family we have analyzed and incorporated into this new database is the TetR family. We have developed a TetR profile, which has been validated (Tobes *et al.*, 2004) and used to detect TetR proteins both in genomes and in the SPTR database. In addition, we have compiled a non-redundant set of 1487 protein sequences belonging to the TetR family.

SYSTEMS AND METHODS

Selection of a non-redundant set of protein sequences of each family by analyzing the SPTR database and NCBI proteomes

The first step in the development of the BacTregulators database was to select the sets of transcriptional regulators with profiles defining each family (Gallegos *et al.*, 1997; Tobes *et al.*, 2004). The screening included searches in SPTR and in all the proteins from whole genome sequencing projects (available at the NCBI microbial genomes resource). The HAMAP project (High-quality Automated and Manual Annotation of microbial Proteomes) (Gattiker *et al.*, 2003) is oriented to automatically annotate in SWISS-PROT a significant percentage of proteins originating from bacterial and archaeal genome sequencing projects. This project has initiated a process of incorporation of transcriptional regulators from genome sequencing projects to SPTR. At present, the overlap between the sets of regulatory protein sequences detected in SPTR and in genome sequencing projects is only partial. We have developed a tool that integrates protein sequences from SPTR with protein sequences from genomes available at NCBI in a non-redundant set for each family (Supplementary Figure 1). This information is available at the BacTregulators website (www.bactregulators.org).

BacTregulators entry

The extraction of information about regulators is a complex and time-consuming task. When we extract data about a regulator it is always possible to assign these data to a protein sequence and to a microorganism. Furthermore, if the microorganism has been sequenced we can assign the information to a protein encoded by a specific gene located in a precise position of a genome element. The expanding availability of sequenced genomes emphasizes the importance of genomic data in the definition of a protein. On the other hand, to ensure flexibility and precision in a database it is necessary to maintain the information at the maximum level of granularity, considering that it is easy to combine information by common features but it is impossible to separate merged information.

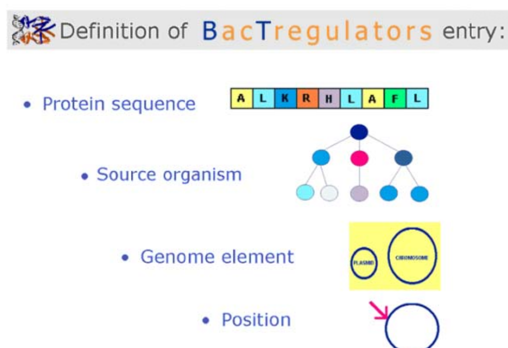


Fig. 1. Each BTR entry is defined considering its protein sequence, the taxonomic classification of the source organism, the genome element and the position in which is located its corresponding gene.

Thus, the definition of an entry in BacTregulators is based on the protein sequence, the source microorganism, the genome element (chromosome, plasmid, phage, etc.) and the specific location of the gene in the corresponding genome element (Fig. 1). The source microorganism is registered according to the taxonomic criteria of the NCBI taxonomy database (Wheeler *et al.*, 2000) (ftp://ftp.ncbi.nih.gov/pub/taxonomy/). In addition, the BacTregulators database stores the taxonomic tree structure,thus making it possible to access the lineage of each source microorganism. Proteins with differences at any of the four levels are recorded as different entries (Fig. 2). Thus, the number of entries in the BacTregulators database (5119) is higher than the number of sequences (2813) because several entries can share the same sequence. However, all entries with identical sequences are interconnected.

Most of the entries included in BacTregulators have been automatically generated after screening the SWISS-PROT, TrEMBL and TrEMBL-new databases (Boeckmann *et al.*, 2003) and the complete proteomes available at the Refseq NCBI microbial genomes resource (Pruitt and Maglott, 2001) for the TetR and XylS/AraC family profiles. The automatic generation of entries is different depending on the source database. When the protein comes from the NCBI Microbial Genome database the process is simple: each regulator detected in a proteome corresponding to a complete genome generates an entry with information for all four levels of definition (sequence, source microorganism, genome element and position). However, if the source database is SPTR the process is more complex. The first difference is that the definition of an entry in the SPTR database is based solely on the protein sequence (Boeckmann *et al.*, 2003), consequently the same SPTR entry can correspond to more than one organism and to more than one genome element. Another difference to

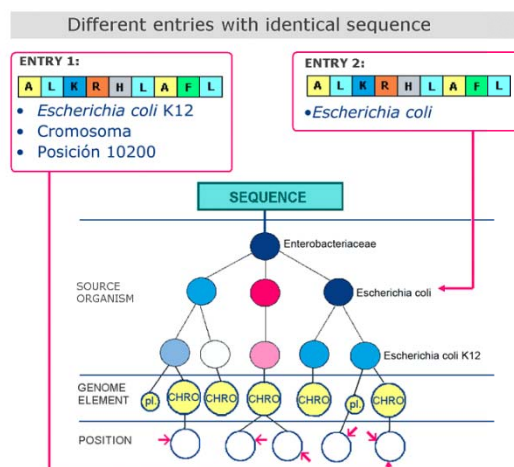


Fig. 2. Two identical sequences can have different entries depending on differences at any of the four levels that define a BTR entry: protein sequence, source organism, genome element and position. In this figure, for example, entry 1 and entry 2 share the same sequence but differ in the source organism, in the genome element and in the position in which the gene encoding the protein is located.

consider is the possibility of redundancy, especially between TrEMBL and TrEMBL-new entries (Gattiker *et al.*, 2003). Thus, for the automatic generation of BacTregulators entries from SPTR database we have used the following strategy: (1) If the SPTR entry corresponds to only one organism and to only one genome element, only one BacTregulators entry is generated. (2) If the SPTR entry corresponds to more than one organism or to more than one genome element, several BacTregulators entries are generated. To avoid entry redundancy, the presence of each entry generated in the database is checked before its inclusion as a new BacTregulators entry.

During the phase of data extraction from published papers, the curators assign information to the appropriate entry, assisted by tools specifically designed for this task. If occasionally the appropriate entry is not included in the set of the automatically generated entries, the curators can generate additional entries to assign the information in the most specific way. Curators are continuously updating BacTregulators database but new data about knowledge and sequences will be available by releases to ensure high quality and integrity of the data.

RESULTS

Information content and structure of the BacTregulators database

The data structure of BacTregulators database is especially designed to include data from genomes. The information is stored in a highly structured way, organized in 158 fields grouped into 33 tables in a relational database managed by the MySQL database management system and the PHP scripting language.

BacTregulators contains data of three types: sequences, knowledge and references.

(1) Information about sequences has been specially treated by integrating, filtering and checking each protein sequence. We supply three sets of non-redundant complete protein sequences for each family of regulators:

 (i) The whole set of protein sequences obtained as the result of the integration of sequences from SPTR and NCBI prokaryotic genomes.

 (ii) The set of protein sequences that result from the integration of sequences from the SWISS-PROT, TrEMBL and TrEMBL-new databases.

 (iii) The set of protein sequences from NCBI prokaryotic genomes.

 As in the AraC-XylS database described previously (Tobes and Ramos, 2002), we also supply these three sets of sequences by domains (DNA-binding domain, N-terminal domain and C-terminal domain) in 9 additional sets of sequences for each family of regulators.

(2) In BacTregulators, the knowledge extraction process combines automatic data extraction with manual extraction. Some data assigned to each entry have been obtained automatically from NCBI Microbial resources and from the SPTR knowledgebase (see below). The rest of the knowledge associated with each entry has been extracted manually from bibliographic references. We have adopted a special structure for this type of data, using the text paragraph as the information unit. All manually obtained data are organized in text paragraphs that are always individually referenced. This structure identifies the source of each piece of data. During the manual extraction of knowledge, we have dissected experimental data that support biological features. These data are termed 'experimental evidence' and receive a special treatment in BacTregulators. When there is experimental evidence to support the knowledge expressed in a paragraph, a link to the experimental data is displayed. These data are useful to evaluate the reliability of the knowledge data.

(3) Each text paragraph is referenced and linked to the Medline abstract. Currently the reference database contains 457 references corresponding to AraC-XylS and TetR families.

Graphical information about the three-dimensional (3D)-structures of crystallized proteins and tutorials are also available at the BacTregulators website.

Access to BacTregulators information

The information contained in the BacTregulators database can be accessed in various ways. Data can be searched by BacTregulators identifier or by SWISS-PROT-TrEMBL or NCBI identifiers. A search by family, microorganism, name and COG allows the user to retrieve specific sets of proteins. Finally, a text search in all the fields of the database allows more flexible browsing. The reference database can also be independently browsed by text searching.

Information about a specific entry is displayed in four sections (Supplementary Figure 2). The first section (blue vertical bar) shows the data that define the BacTregulators entry: sequence identifier, source organism, genome element and position in the genome element. The second section (blue-green vertical bar) contains data automatically obtained from databases: SPTR or NCBI accession number, source database, the protein's full name and short name, gene name, gene orientation, COG code, NCBI functional code and last update of this set of data. The third section (green vertical bar) displays knowledge manually extracted from published research articles. This knowledge is structured in the following fields: function, genes regulated, regulatory networks, effectors, genomic allocation, promoters of regulated genes, promoters of the regulator, sigma dependence, pathogenicity, applications, mutational data, 3D-structure, oligomerization, similarities, comments and last update of this group of manually extracted data (Tobes and Ramos, 2002). For each of these knowledge fields, we also provide downloadable text files containing all the information corresponding to each field. The last section (orange vertical bar) provides access to information related to protein sequences. The subsection on protein sequences provides access to the complete protein sequence and to the sequence of the different domains. Finally, the subsection of BLAST similarities allows the user to access the file with the BLAST results as compared to the rest of the members of the family, and a list of entries with a similar sequence ranked by significance (BLAST E-value).

DISCUSSION

Integration of knowledge customizable at the retrieval step

When a user is accessing a specific entry, the knowledge associated with related entries can be useful to infer functional features. This was our rationale for designing an innovative tool to display an integrated report of a set of entries related to a single initial entry. Although the information is maintained at the maximum level of granularity associated with each entry, the information corresponding to a given set of related entries can be combined at the retrieval step. We have considered similarity between sequences as the key criterion to define a group of related entries. This first selection of related entries can be restricted by taxonomic requirements. In this way the level of integration of knowledge can be tuned by the user, who sets similarity parameters and taxonomic level.

To select the set of related entries, a list is displayed with all BacTregulators entries that have significant BLAST similarity with the initial sequence. All entries that share a sequence identical to the initial entry are displayed against a yellow background. In a first step, the user selects an E-value as the threshold to define the set of related sequences. In a second step, this set can be limited by taxonomic criteria: the user chooses the taxonomic level for including an entry in the set of related entries. The lineage corresponding to the source microorganism of the initial entry is displayed to allow the user to fix taxonomic limits (Supplementary Figure 3). Once the similarity threshold and taxonomic limits are set, an integrated report of the selected set of entries is displayed. The data are displayed by field in different sections. Each section includes all the data for a field from all entries in the set. The corresponding BacTregulators entry identifier is indicated at the beginning of each paragraph. A list of integrated entries is also accessible at the integrated report page. (Supplementary Figure 4).

It is unquestionable that orthologous proteins frequently share many functional features. Hence, inference of function from sequence similarity (Gattiker *et al.*, 2003) makes particular sense for orthologous proteins. However, in practise, it is extremely difficult to define orthologous proteins. We have avoided this problem by allowing the user to choose the set of entries from which function may be inferred. This tool can help orient the user to the possible function of regulators with no defined function. Thus, a selected group of entries with a similar sequence can include entries with functional features and entries with no known function. If the function is similar in all the characterized proteins in the group, it is likely that the uncharacterized proteins of the group have a similar function. The tool described here thus facilitates the selection of functional clusters of entries for which knowledge can be propagated to entries with no published function.

New families of transcriptional regulators will be added to BacTregulators database to extend the understanding of the complex circuitry of gene regulatory networks in bacteria and archaea.

BacTregulators is especially intended to detect the regulators in prokaryotic genomes and classify then into families. As designed, BacTregulators facilitates the accurate annotation of regulators in genomes. In addition, the information collected on each protein of the family can be useful for researchers who wish to characterize a new regulator or investigate the biological properties of a known regulator.

ACKNOWLEDGEMENTS

We thank Carmen Lorente and Karen Shashok for checking the language of the manuscript and Wilson Teran for the information about TtgR. Work in J.L.R.' laboratory was

financed by European Commission grants (BIO-CT3-00435 and BIO-CT2000-00170) and by grants from the CICYT (GEN2001-4698-C05-3 and BIO2003-00515).

REFERENCES

Boeckmann,B., Bairoch,A., Apweiler,R., Blatter,M.-C., Estreicher,A., Gasteiger,E., Martin,M.J., Michoud,K., O'Donovan,C., Phan,I., Pilbout,S. and Schneider,M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365–370.

Gallegos,M.T., Schleif,R., Bairoch,A., Hofmann,K. and Ramos,J.L. (1997) AraC/XylS family of transcriptional regulators. *Microbiol. Mol. Biol. Rev.*, **61**, 393–410.

Gattiker,A., Michoud,K., Rivoire,C., Auchincloss,A.H., Coudert,E., Lima,T., Kersey,P., Pagni,M., Sigrist,C.J., Lachaize,C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput. Biol. Chem.*, **27**, 49–58.

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

Pruitt,K.D. and Maglott,D.R. (2001) RefSeq and LocusLink: NCBI gene-centered resources. *Nucleic Acids Res.*, **29**, 137–140.

Ramos,J.L., Marques,S. and Timmis,K.N. (1997) Transcriptional control of the *Pseudomonas* TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. *Annu. Rev. Microbiol.*, **51**, 341–373.

Ramos,J.L., Gallegos,M.T., Marqués,S., Ramos-González,M.I., Espinosa-Urgel,M. and Segura,A. (2001) Responses of Gram-negative bacteria to certain environmental stressors. *Curr. Opin. Microbiol.*, **4**, 166–171.

Tobes,R. and Ramos,J.L. (2002) AraC-XylS database: a family of positive transcriptional regulators in bacteria. *Nucleic Acids Res.*, **30**, 318–321.

Tobes,R., Martínez-Bueno,M., Molina-Henares,A.J., Pareja,E. and Ramos,J.L. (2004) The TetR family of transcriptional repressors, in preparation.

Wheeler,D.L., Chappey,C., Lash,A.E., Leipe,D.D., Madden,T.L. and Schuler,G.D., Tatusova,T.A. and Rapp,B.A. (2000) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **28**, 10–14.

Material suplementario del capítulo 5. Equivalencias:

Suppl. Figure 1 – Figura 4 de esta Tesis.
Suppl. Figure 2 – Figura 8 de esta Tesis.
Suppl. Figure 3 – Figura 9 de esta Tesis.
Suppl. Figure 4 – Figura 10 de esta Tesis.

Discusión

BacTregulators

Es incuestionable que proteínas ortólogas comparten frecuentemente características funcionales. Por tanto, la inferencia de función por similitud de secuencia tiene especial sentido en el caso de proteínas ortólogas. Sin embargo, en la práctica es extremadamente difícil definir proteínas ortólogas.

En BacTregulators se ha resuelto este problema permitiendo al usuario elegir el conjunto de registros del que se puede inferir la función. Esta herramienta puede orientar al usuario sobre la posible función de reguladores cuya función se desconoce. Así, el grupo de registros seleccionado con una secuencia similar puede incluir registros con características funcionales definidas y registros con función no conocida. Si la función es similar en todas las proteínas caracterizadas del grupo, es probable que las proteínas no caracterizadas tengan una función similar. Por tanto, se facilita la selección de clústeres funcionales de registros para los que la función se puede extrapolar a registros cuya función no está publicada.

Se planea añadir a BacTregulators nuevas familias de reguladores transcripcionales para comprender mejor los complejos circuitos de redes de regulación génica en bacterias y arqueas. Un caso es el de la familia MerR, que se ha definido en esta Tesis Doctoral.

BacTregulators está especialmente diseñada para detectar y anotar con precisión reguladores en genomas de procariotas y clasificarlos en familias. Además, la información reunida sobre cada proteína de la familia puede ser de utilidad a quien desee caracterizar un nuevo regulador o investigar las propiedades biológicas de uno conocido.

Provalidator

¿Por qué investigar en diseño de *profiles*? La visualización, comprensión y manipulación de *profiles* es posible gracias a que la información que contiene se almacena de manera legible, mientras que un HMM, como su misma palabra indica, oculta la información almacenada, por lo que es imposible un ajuste a mano de los parámetros. Así, si por razones estructurales o basando nuestra decisión en cualquier dato bibliográfico, queremos dar más o menos peso a un determinado residuo de una posición (o incluso hacerlo obligatorio), con un *profile* podremos hacerlo fácilmente.

La verdadera novedad de Provalidator radica en que ofrece la posibilidad de enfrentar a la base de datos UniProt (o cualquier conjunto de secuencias en formato FASTA) un *profile* construido por el usuario. Hasta ahora, ningún sitio web ofrecía esa alternativa. Lo que PROSITE brinda a través de su aplicación ScanProsite o InterPro a través de InterProScan es la posibilidad de detectar para qué *profiles* almacenados en su base de datos las secuencias pasadas por el usuario dan positivo, o bien el *score* obtenido para un determinado *profile*. Pero no permite pasar también el *profile* que puntuará las secuencias.

El problema al que nos enfrentamos a la hora de construir un *profile* a partir de un conjunto de secuencias semilla es que tenemos 2 variables: las secuencias que forman parte del conjunto y la región del alineamiento elegida. Kay Hofmann y Philipp Bucher resolvieron esta dualidad fijando una de las variables y moviendo la otra. En su caso, fijaron la región y movieron las secuencias pertenecientes a la semilla. Esto nos da un diagrama de flujo del proceso tal y como se representa en la siguiente figura:

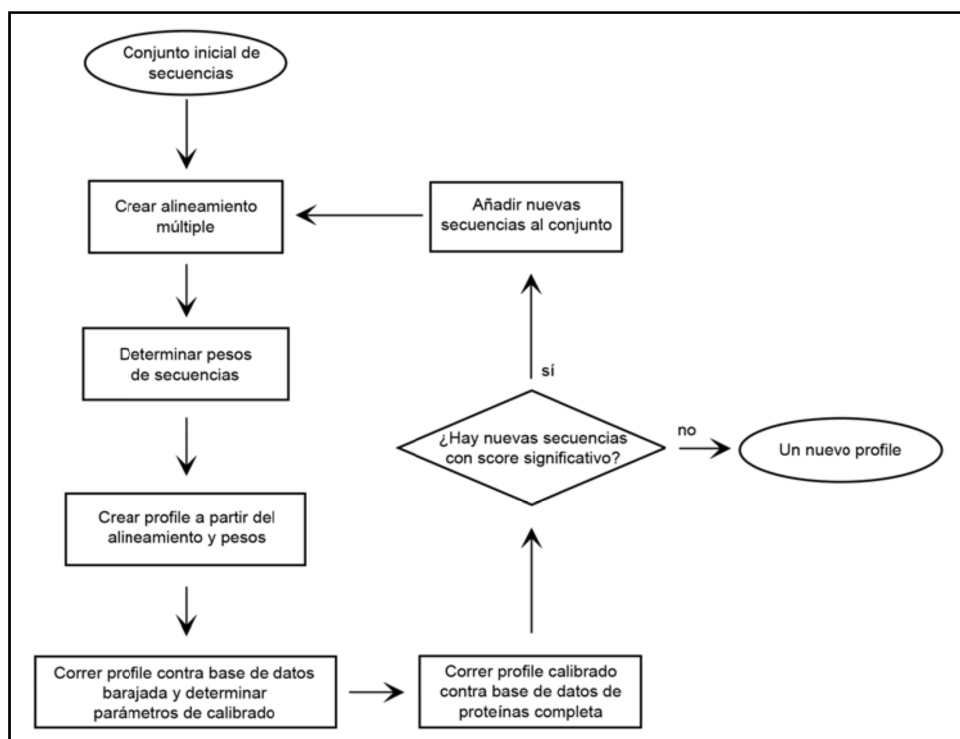


Figura 15: Diagrama de flujo del método de construcción de un *profile* descrito por Bucher.

Nuestra experiencia nos dice que este método iterativo no converge, esto es, cada vez que añadimos nuevas secuencias al conjunto inicial que antes no estaban a pesar de obtener un *score* significativamente alto (y por tanto definir bien al grupo en cuestión), del conjunto de secuencias que captura el nuevo *profile* salen otras que antes pertenecían de pleno derecho, con un *score* alto. A esto nos referimos con que este método es no convergente. El proceso se puede repetir constantemente, que no llegaremos a un conjunto inicial que genere un *profile* que capture todas las secuencias que en iteraciones anteriores han obtenido un *score* por encima del umbral de pertenencia. Por eso, nosotros hemos decidido fijar el conjunto de secuencias que forman la semilla (basando esta elección en motivos bibliográficos, estructurales o en una anotación inequívoca) y jugar con la región elegida del alineamiento. Con esto, podemos sacar conclusiones más interesantes, como averiguar qué región es discriminante para definir a una familia de proteínas. Esto se representa en el siguiente diagrama de flujo:

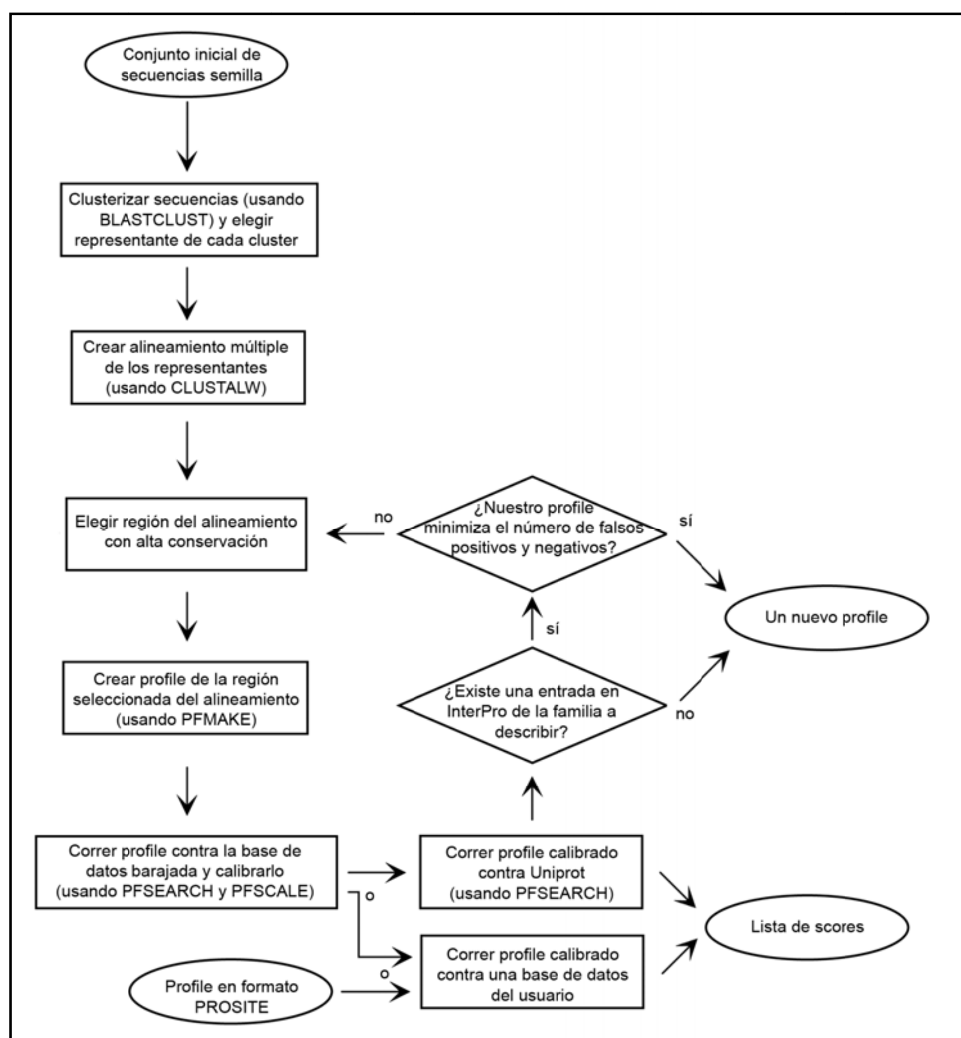


Figura 16: Diagrama de flujo de nuestro método de construcción de un *profile*.

A la hora de variar la región del alineamiento para comparar qué *profile* define mejor a la familia objeto de estudio, hay que tener en cuenta dos conceptos de las técnicas heurísticas que se pueden usar en este contexto: diversidad y exhaustividad. La diversidad consiste en abarcar diferentes zonas del alineamiento en las que haya una alta conservación. Exhaustividad consiste en hacer ligeros ajustes en los límites de la región que mejor resultado nos esté dando. O sea, hacer un ajuste fino una vez que se ha barrido todo el espacio de soluciones.

Ejemplo de uso:

Paso 1: Lo primero que se debe hacer para comenzar a trabajar con Provalidator es crear una nueva sesión.

Haciendo click en "New session" se generará un nuevo número de sesión aleatorio que la identificará unívocamente y que servirá para poder entrar en la sesión en cualquier momento y seguir trabajando en ella o ver resultados anteriores. Este número se deberá recordar. La sesión (y todos los datos que en ella se generen) durará en el servidor una semana desde el momento de la última operación realizada. Cuando se haya terminado de trabajar, se puede borrar (junto con los datos) pulsando "Close".

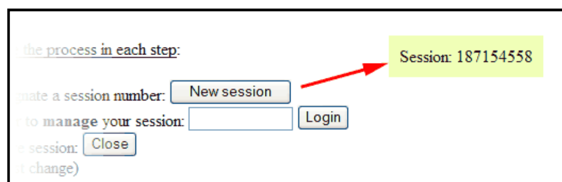


Figura 17: Crear sesión en Provalidator.

Paso 2: Se selecciona el conjunto inicial de secuencias que forman parte del grupo que se quiere definir.

Ejemplo: supongamos que queremos crear un *profile* para la familia de reguladores transcripcionales IclR. Buscaremos en Swiss-Prot/TrEMBL (www.expasy.org) secuencias pertenecientes a este grupo.

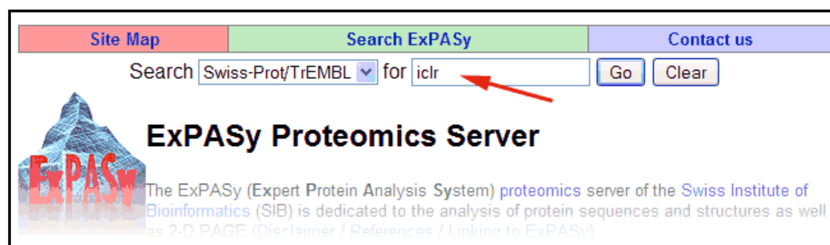


Figura 18: Recuperación de secuencias en Expasy.

Este conjunto puede proceder de cualquier fuente, no necesariamente UniProt. Se podrá usar cualquier conjunto de secuencias en formato FASTA. La ventaja de usar UniProt como fuente para las secuencias es que Provalidator puede dar prioridad a las secuencias de Swiss-Prot (las cuales están anotadas a mano por expertos) frente a TrEMBL a la hora de elegir representante del clúster, lo cual da una mayor solvencia al conjunto inicial elegido.

Una vez tengamos ese conjunto, se pega en la ventana "Sequences". Se selecciona un umbral de simi-

litud, una "lista negra" (esto es, palabras que no deben aparecer en la cabecera de la secuencia para que ésta sea elegida) y una palabra clave (esto es, una palabra que debe aparecer en la cabecera de la secuencia para ser elegida. Es opcional), y se pulsa "Clusterize".

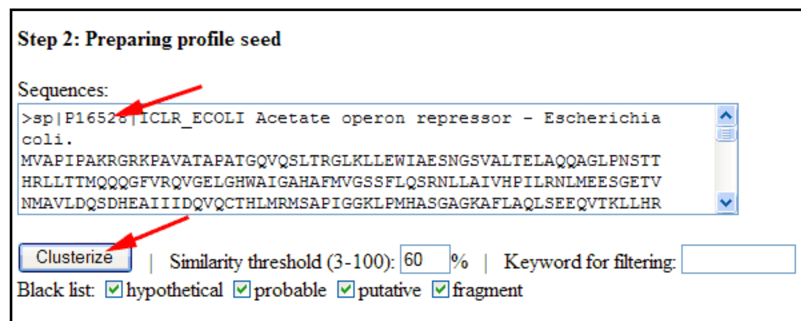


Figura 19: Clusterizar secuencias para formar el conjunto semilla.

Después de unos minutos, se obtiene el resultado tal y como se presenta en la siguiente tabla:

Click here for a st[...]

Análisis estadístico, marcando secuencias con longitud muy diferente.

Result of the [...] et: 88 sequences. Similarity=60%):

```
Original output from blastclust:

#1: sp|P37728|KDGR_ERWCH tr|Q87K22 sp|P76268|KDGR_ECOLI tr|Q83RJ0 tr|Q8FGT5 tr|Q8KM23 tr|Q8XC
#2: sp|P37671|YIAJ_ECOLI tr|Q83PQ7 tr|Q8FCD7 tr|O87190 tr|Q93KA0 tr|Q93Q65 tr|Q9CLH4 sp|P4499
#3: tr|Q8FB60 tr|Q8X5Z9 sp|P16528|ICLR_ECOLI sp|P17430|ICLR_SALTY tr|Q8Z1V7 tr|Q83MI8 tr|Q8D
#4: sp|Q52154|PCAR_PSEPU tr|Q88N41 tr|Q9R9T0 tr|Q87X78 tr|Q9I6X7
#5: tr|Q8Z8R2 tr|Q8ZR86 sp|P77734|ALLR_ECOLI tr|Q83SD3 tr|Q9S
#6: sp|P77732|YFAX_ECOLI tr|Q8FFM7 tr|Q8XE06 tr|Q8Z547 tr|Q8Z
#7: tr|Q8U7A8 tr|Q8FW26 tr|Q8YC90 tr|Q92TL3
#8: tr|Q8G2T0 tr|Q8YF10 tr|Q986Y9 tr|Q92VL9
#9: tr|Q849Q8 tr|Q8G976 tr|Q93PU6 tr|Q9R9U0
#10: tr|Q9L688 tr|Q82FI4 tr|Q9X9U3
#11: tr|Q8YDF7 tr|Q8VQJ7 tr|Q8FUX8
#12: tr|Q9I392 tr|Q8FX66 tr|Q8YB78
#13: tr|Q9RDB9 tr|Q82CW2 tr|Q54411
#14: tr|Q8NSW1

-------------------------------------------------

Header of the 18 filtered sequences:

#1: >sp|P37728|KDGR_ERWCH Pectin degradation repressor protein kdgR - Erwinia chrysanthemi.
#2: >tr|O87190 Pir protein - Erwinia chrysanthemi.
#3: >sp|P16528|ICLR_ECOLI Acetate operon repressor - Escherichia coli.
#4: >sp|Q52154|PCAR_PSEPU Pca regulon regulatory pr
#5: >sp|P77734|ALLR_ECOLI Negative regulator of all                tion operons
#7: >tr|Q8U7A8 Transcriptional regulator, IclR fami                s (strain C58
#8: >tr|Q8G2T0 Transcriptional regulator, IclR fami
#9: >tr|Q849Q8 Sep repressor SepR - Pseudomonas put
#11: >tr|Q8YDF7 Transcriptional regulator - Brucell
#12: >tr|Q8FX66 Transcriptional regulator, IclR fami
#14: >tr|Q8NSW1 Transcriptional regulator - Corynebacterium glutamicum (Brevibacterium flavum).
#15: >tr|Q8FV53 Transcriptional regulator, IclR family - Brucella suis.

-------------------------------------------------
```

Cada línea representa un clúster, con todos sus elementos listados. Cuando la línea sólo tiene un *id*, éste es el único miembro del clúster.

Cabeceras de la secuencia representante en cada clúster, dando prioridad a las anotadas en Swiss-Prot.

cont.

```
-----------------------------------------------

Header of sequences that are in lines with no choosen sequence:
  -for including some of the selected words in the black list.
  -for not including the keyword (if exists).
  -for having a length very dissimilar.

#6:
  >sp|P77732|YFAX_ECOLI Putative HTH-type transcriptional regulator
  >tr|Q8FFM7 Hypothetical transcriptional regulator yfaX - Escheric
  >tr|Q8XE06 Putative regulator - Escherichia coli O157:H7.
  >tr|Q8Z547 Putative transcriptional regulator - Salmonella typhi.
  >tr|Q8ZNF8 Putative transcriptional regulator - Salmonella typhim
#10:
  >tr|Q9L688 Sporulation protein SsfR (Putative transcriptional regulator) - Streptomyces griseus.
  >tr|Q82FI4 Putative IclR-family transcriptional regulator - Streptomyces avermitilis.
  >tr|Q9X9U3 Putative transcriptional regulator - Streptomyces coelicolor.
#13:
  >tr|Q9RDB9 Putative IclR-family transcriptional regulator - Streptomyces coelicolor.
  >tr|Q82CW2 Putative IclR-family transcriptional regulator - Streptomyces avermitilis.
  >tr|Q54411 66 lysT-alpha transfer RNA-Lys(AAG) (Fragment) - Streptomyces lividans.
#20:
  >tr|O87768 AcoR protein - Pseudomonas putida. LENGTH!
  >tr|Q88QD7 Acetoin catabolism regulatory protein - Pseudomonas putida (strain KT2440). LENGTH!
#21:
  >tr|Q8UJL2 Transcriptional regulator, IclR family/regucalcin - Agrobacterium tumefaciens
  >tr|Q92ZR8 Putative calcium binding transcriptional regulatory protein - Rhizobium meliloti

-----------------------------------------------

Choosen sequences:

>sp|P37728|KDGR_ERWCH Pectin degradation repressor protei                    santhemi.
MIFNRSVTYSNARLPYSKRSLYTKTRVLFFLKQKILSRVTTKMAIADLDKQPDSVSS
VFGILQALGEEREIGITELSQRVMMSKSTVYRFLQTMKSLGYVAQEGESEKYSLTLK
LGAKALQNVDLIRSADIQMRELSALTRETIHLGALDEDSIVYIHKIDSMYNLRMYSRIGR
RNPLHSTAIGKVLLAWRDREEVKEILSQVEFKRTTVHTIGSTEELLPQLDLVRQQGYGED
NEEQEEGLRCIAVPVFDRFGVVIAGLSISFPTIRFSEDNKHEYVAMLHTAARNISDQMGY
HDYPF
>tr|O87190 Pir protein - Erwinia chrysanthemi.
MINEHSIDDEVKAEKPLGSQSLFRGLQLIEILSDYPNGCPLARIAELSGMNKSTVHRLLQ
GLAGCGYVVPARQPGSYRLTTKFIAIGQKALSSLNVLMVAVPHLEALNLTVGETVNFSTR
EDDHAILIYKLEPTTGMMRTRAYIGQHMPLHSSAMGKIFMAYGTEDYPAEYWRTHQNSIC
PLTDNTIVELSAMRQELAEIRRQGLAMDREENELGVSCISAPVFDIQHRVPYAVSISLSK
ARLQQIGIDTLVAPLRETARNISLELGFTPPT
```

Clústeres con ninguna secuencia elegida por una de estas razones: lista negra, palabra clave o longitud.

Secuencias finalmente elegidas por el programa.

Figura 20: Resultado de la clusterización.

Ahora tenemos que alinear el conjunto de secuencias final que formará la semilla del *profile*. Para esto, recomendamos el sitio web: align.genome.jp, basado en el algoritmo CLUSTALW.

Paso 3: Una vez el conjunto de secuencias clusterizado se ha alineado, se selecciona una región conservada a fin de construir un *profile* con esta región. Para esto, se establecen los límites del alineamiento.

Step 3: Constructing profile

Alignment (in MSF format):

```
PileUp

   MSF:  220  Type: P    Check:  6971   ..

 Name: Q97NZ3|Q97NZ3_STRPN_Transcript  Len:  220  Check:  5910
```

Limits: 15 31

Mode: ○ Global ○ Domain global ○ Semiglobal ⦿ Local

☑ Calibrate profile (calibrating can take 10-15 min.)

Construct

Figura 21: Construcción del *profile*.

Hay 4 modos de construcción:

- Global: El alineamiento debe empezar al principio del *profile* y al principio de la secuencia, y debe terminar al final del *profile* y al final de la secuencia.
- Domain global: El alineamiento debe incluir el *profile* entero; éste puede empezar y terminar en cualquier posición dentro de la secuencia.
- Semi-global: El alineamiento debe empezar o al principio del *profile* o al principio de la secuencia, y debe terminar o al final del *profile* o al final de la secuencia.
- Local: El alineamiento es posible en cualquier parte.

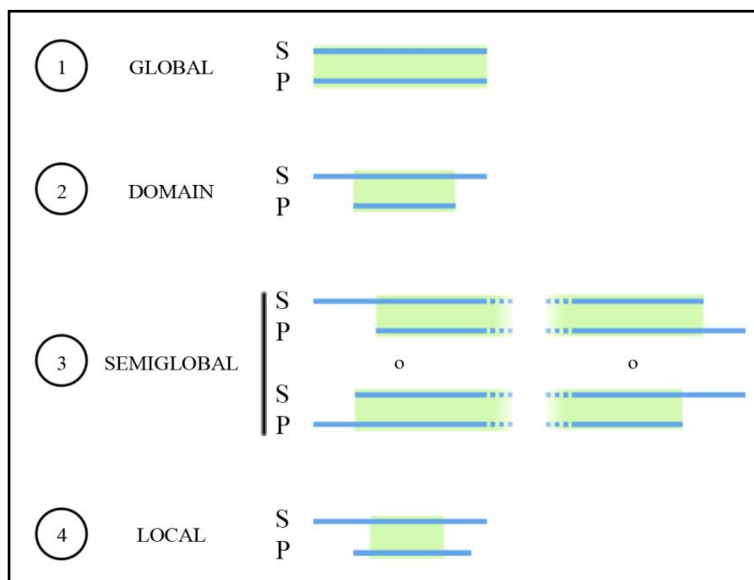


Figura 22: Modos de construcción de un *profile*.

Como ya hemos explicado, calibrar el *profile* consiste en enfrentarlo contra una base de datos barajada y calcular los parámetros de normalización que fijan el umbral de pertenencia en 8.5. Este proceso puede durar varios minutos, dependiendo de la longitud del *profile*.

Paso 4: Enfrentar las secuencias a un *profile* con objeto de calcular el *score* obtenido por cada secuencia.

Se puede elegir entre el último *profile* creado o uno pasado en la ventana "Profile", y la base de datos puede ser la versión de UniProt almacenada en el servidor o un conjunto de secuencias en formato FASTA pasado en la ventana "Database". Si se elige UniProt como base de datos, ésta es la parte más costosa en tiempo de todo el proceso, y puede durar de 5 a 10 horas, dependiendo de la longitud del *profile*. Éste precisamente es el principal problema que nos encontramos a la hora de automatizar este paso: no pudimos usar alguna de las técnicas heurísticas existentes para toma de decisiones a la hora de hacer evolucionar el sistema hasta la solución óptima debido al gran tiempo empleado para generar una solución en cada iteración. En nuestro caso, una solución es el número de potenciales falsos positivos y negativos, y la meta consiste en minimizarlos.

Paso 5: Validación de resultados obtenidos.

Siempre que hayamos elegido UniProt como conjunto de secuencias al que enfrentar el *profile* y que exista una entrada en InterPro de la familia o grupo de proteínas que se pretende definir, podremos validar el resultado obtenido usando la información recogida en InterPro. Esto nos da una idea de los diferentes conjuntos de secuencias que capturan los demás métodos usados en InterPro y el *profile* que acabamos de definir.

Órdenes ejecutadas por el código:

A continuación, se enumeran las órdenes que el código ejecuta en cada paso, explicando los argumentos de cada programa y los valores que toman.

Clusterización:

```
blastclust -i archivo_entrada -o archivo_salida -S umbral -b F -e F
```

-i: Archivo de secuencias en formato FASTA.
-o: Archivo de resultados.
-S: Umbral de similitud expresado en % (valores posibles: 3-100).
-b: requiere cumplir la condición del parámetro S en ambas (T) o sólo en una (F) secuencia del par.
-e: Debe valer F si múltiples secuencias pueden tener el mismo nombre (para evitar errores).

Construcción del *profile*:

```
pfmake -0123 alineamiento weight_matrix >salida
```

-0123: Selecciona el modo de construcción.
-alineamiento: alineamiento de las secuencias semilla.
-weight-matrix: Matriz de pesos usada.
-salida: Archivo de salida.

Calibración del *profile*:

```
pfsearch -fr -C 150 profile_no_calibrado window20.seq | sort -r | pfscale -
profile_no_calibrado -N 21000000 -P 0.00005 -Q 0.0000005 >profile_calibrado
```

Esta orden encadena tres programas:
- pfsearch: calcula los *scores* del *profile* contra el archivo de secuencias barajadas.
- sort: ordena descendentemente los *scores* obtenidos en el paso anterior.
- pfscale: usa estos *scores* para calcular los parámetros de la función de normalización y devuelve un *profile* idéntico pero con estos parámetros almacenados.

-N, -P y -Q son los parámetros que definen el conjunto de secuencias barajadas window20.seq.

Mejoras previstas en Provalidator:

Está previsto incluir alguna funcionalidad más en Provalidator que añada automatización al proceso. Aquí se enumeran algunas de esas mejoras:

- Reconocimiento automático de la región de HTH (si existe).
- Cálculo automático de los límites de las distintas regiones conservadas que se detecten en el alineamiento.

Conclusiones

1. Se ha desarrollado un programa semiautomático para diseñar y validar *profiles* que definen familias de proteínas. El programa Provalidator está libremente accesible en Internet.

2. La validación de Provalidator se consiguió definiendo el *profile* de la familia de MerR. En esta familia, el *profile* ha permitido diferenciar dos grupos de secuencias atendiendo a su longitud.

3. En el diseño del *profile* de la familia IclR, se encontró que la región de HTH no era representativa para determinar la pertenencia de una proteína a esta familia. En cambio, la región de multimerización y unión a efectores permite asignar proteínas a la familia.

4. Asimismo, y como vía de validación de la herramienta Provalidator, se construyó un *profile* para definir la familia de bombas de extrusión RND. Provalidator no sólo permitió asignar proteínas a la familia, sino que permitió clasificarlas en grupos basándose en los valores normalizados de *score*. Se ha determinado que cada grupo de Z-*score* coincide con una subfamilia. Estas subfamilias se diferencian en el sustrato que expulsan, habiéndose encontrado bombas que expulsan antibióticos, metales pesados y otras dos de sustratos desconocidos. Entre éstas, una podría eliminar agentes oxidantes, y la cuarta podría representar un grupo de bombas de amplio perfil de sustratos que expulsa metales y compuestos orgánicos.

5. El *profile* de la familia TetR reveló que la región que comprende el HTH y la hélice adyacente α_1 es la región más representativa de los miembros de la familia. Analizados los modelos tridimensionales de las proteínas de la familia TetR que están cristalizadas, la región usada para definir el *profile* corresponde a una estructura altamente conservada en los miembros de la familia.

6. Se ha construido BacTregulators, una base de datos y una interfaz web para presentar los resultados obtenidos a partir de los *profiles* definidos con Provalidator. Esta herramienta permite extrapolar información de proteínas de función conocida a otras con secuencias similares.

Direcciones de sitios web utilizados

Programas:

- BLAST: ftp://ftp.ncbi.nih.gov/blast
- PFTOOLS: http://www.isrec.isb-sib.ch/ftp-server/pftools

Bases de datos:

- BacTregulators: http://www.bacTregulators.org
- Secuencias de UniProt de procariotas: ftp://ftp.expasy.org/databases/uniprot/current_release/knowledgebase/taxonomic_divisions
- Base de datos barajada: ftp://ftp.isrec.isb-sib.ch/pub/databases/shuffled
- Base de datos taxonómica de NCBI: ftp://ftp.ncbi.nih.gov/pub/taxonomy
- Base de datos InterPro: ftp://ftp.ebi.ac.uk/pub/databases/interpro
- Secuencias de genomas bacterianos de NCBI: ftp://ftp.ncbi.nih.gov/genomes/Bacteria
- Código de ortólogos de NCBI: ftp://ftp.ncbi.nih.gov/pub/COG/COG
- PubMed: http://www.ncbi.nlm.nih.gov/entrez

Sitios web:

- UniProt, Swiss-Prot y TrEMBL: http://www.expasy.org
- PROSITE: http://www.expasy.org/prosite
- InterPro: http://www.ebi.ac.uk/interpro
- InterProScan: http://www.ebi.ac.uk/InterProScan
- NCBI: http://www.ncbi.nlm.nih.gov
- TIGR: http://www.tigr.org
- CLUSTALW: http://align.genome.jp
- NPS@ (Network Protein Sequence Analysis): aquí encontramos el software de predicción de estructura “PHD” y “HTH”. http://npsa-pbil.ibcp.fr

Bibliografía

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389-3402.

Attwood, T. K. (2002) The PRINTS database: a resource for identification of protein families. *Brief.Bioinform.*, **3**, 252-263.

Aranda-Olmedo, I., Tobes, R., Manzanera, M., Ramos, J.L. and Marqués, S. (2002) Species-specific repetitive extragenic palindromic (REP) sequences in *Pseudomonas putida*. *Nucleic Acids Res.*, **30**, 1826-1833.

Bairoch, A., Apweiler, R., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2005) The Universal Protein Resource (UniProt). *Nucleic Acids Res.*, **33**, D154-D159.

Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004) The Pfam protein families database. *Nucleic Acids Res.*, **32**, D138-D141.

Beinlich, K.L., R. Chuanchuen, and H.P. Schweizer (2001) Contribution of multidrug efflux pumps to multiple resistance in veterinary clinical isolates of *Pseudomonas aeruginosa*. *FEMS Microbiol Lett.*, **198**, 129-134.

Berg, G., A. Fritze, N. Roskot, and K. Smalla (2001) Evaluation of potential biocontrol rhizobacteria from different host plants of *Verticillium dahliae* Kleb. *J. Appl Microbiol*, **91**, 963-971.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M. C., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., Pilbout, S., and Schneider, M. (2003) The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res.*, **31**, 365-370.

Bucher, P. and Bairoch, A. (1994) A generalized profile syntax for biomolecular sequence motifs and its function in automatic sequence interpretation. *Proc.Int.Conf.Intell.Syst.Mol.Biol.*, **2**, 53-61.

Bucher, P., Karplus, K., Moeri, N., and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput.Chem.*, **20**, 3-23.

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

Dagley, S. (1971) Catabolism of aromatic compounds by microorganisms. *Adv Microb Physiol*, **6**, 1-46.

Dodd, I. B. and Egan, J. B. (1990) Improved detection of helix-turn-helix DNA-binding motifs in protein sequences. *Nucleic Acids Res.*, **18**, 5019-5026.

Dos Santos, V.A., Heim, S., Moore, E.R., Stratz, M. and Timmis, K.N. (2004) Insights into the genomic basis of niche specificity of *Pseudomonas putida* KT2440. *Environ Microbiol*, **6**, 1264-86.

Gallegos, M. T., Schleif, R., Bairoch, A., Hofmann, K., and Ramos, J. L. (1997) Arac/XylS family of transcriptional regulators. *Microbiol.Mol.Biol.Rev.*, **61**, 393-410.

Gribskov, M., McLachlan, A. D., and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc.Natl.Acad.Sci.U.S.A*, **84**, 4355-4358.

Haft, D. H., Selengut, J. D., and White, O. (2003) The TIGRFAMs database of protein families. *Nucleic Acids Res.*, **31**, 371-373.

Henikoff, S. and Henikoff, J. G. (1992) Amino acid substitution matrices from protein blocks. *Proc.Natl.Acad.Sci.U.S.A*, **89**, 10915-10919.

Henikoff, S. and Henikoff, J. G. (1994) Protein family classification based on searching a database of blocks. *Genomics*, **19**, 97-107.

Hofmann, K. and Bucher, P. (1995) The FHA domain: a putative nuclear signalling domain found in protein kinases and transcription factors. *Trends Biochem.Sci.*, **20**, 347-349.

Keane, A., P. C. Lau, and S. Ghoshal. (2008) Use of a whole-cell biosensor to assess the bioavailability enhancement of aromatic hydrocarbon compounds by nonionic surfactants. *Biotechnol Bioeng*, **99**, 86-98.

Krell, T., Molina-Henares, A. J., and Ramos, J. L. (2006) The IclR family of transcriptional activators and repressors can be defined by a single profile. *Protein Science*, **15**, 1207-1213.

Letunic, I., Copley, R. R., Schmidt, S., Ciccarelli, F. D., Doerks, T., Schultz, J., Ponting, C. P., and Bork, P. (2004) SMART 4.0: towards genomic data integration. *Nucleic Acids Res.*, **32**, D142-D144.

Luthy, R., Xenarios, I., and Bucher, P. (1994) Improving the sensitivity of the sequence profile method. *Protein Sci.*, **3**, 139-146.

Martínez-Bueno, M., Molina-Henares, A. J., Pareja, E., Ramos, J. L., and Tobes, R. (2004) BacTregulators: a database of transcriptional regulators in bacteria and archaea. *Bioinformatics.*, **20**, 2787-2791.

Molina-Henares, A. J., Krell, T., Guazzaroni, M. E., Segura, A., and Ramos, J. L. (2006) Members of the IclR family of bacterial transcriptional regulators function as activators and/or repressors. *FEMS Microbiol.Rev*. **30**, 157-186.

Mulder, N. J., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bradley, P., Bork, P., Bucher, P., Cerutti, L., Copley, R., Courcelle, E., Das, U., Durbin, R., Fleischmann, W., Gough, J., Haft, D., Harte, N., Hulo, N., Kahn, D., Kanapin, A., Krestyaninova, M., Lonsdale, D., Lopez, R., Letunic, I., Madera, M., Maslen, J., McDowall, J., Mitchell, A., Nikolskaya, A. N., Orchard, S., Pagni, M., Ponting, C. P., Quevillon, E., Selengut, J., Sigrist, C. J., Silventoinen, V., Studholme, D. J., Vaughan, R., and Wu, C. H. (2005) InterPro, progress and status in 2005. *Nucleic Acids Res.*, **33**, D201-D205.

Nakazawa, T. (2002) Travels of a *Pseudomonas*, from Japan around the world. *Environ Microbiol*, **4**, 782-786.

Nelson, K. E., C. Weinel, I. T. Paulsen, R. J. Dodson, H. Hilbert, d. S. Martins, V, D. E. Fouts, S. R. Gill, M. Pop, M. Holmes, L. Brinkac, M. Beanan, R. T. DeBoy, S. Daugherty, J. Kolonay, R. Madupu, W. Nelson, O. White, J. Peterson, H. Khouri, I. Hance, L. P. Chris, E. Holtzapple, D. Scanlan, K. Tran, A. Moazzez, T. Utterback, M. Rizzo, K. Lee, D. Kosack, D. Moestl, H. Wedler, J. Lauber, D. Stjepandic, J. Hoheisel, M. Straetz, S. Heim, C. Kiewitz, J. A. Eisen, K. N. Timmis, A. Dusterhoft, B. Tummler, and C. M. Fraser (2002) Complete genome sequence and comparative analysis of the metabolically versatile *Pseudomonas putida* KT2440. *Environ. Microbiol.*, **4**, 799-808.

Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R., and Lopez, R. (2005) InterProScan: protein domains identifier. *Nucleic Acids Res.*, **33**, W116-W120.

Ramos, J.L., F. Rojo, L. Zhou, and K.N. Timmis (1990) A family of positive regulators related to the *Pseudomonas putida* TOL plasmid XylS and the *Escherichia coli* AraC activators. *Nucleic Acids Res.*, **18**, 2149-2152.

Ramos, J. L., Martínez-Bueno, M., Molina-Henares, A. J., Terán, W., Watanabe, K., Zhang, X., Gallegos, M. T., Brennan, R., and Tobes, R. (2005) The TetR family of transcriptional repressors. *Microbiol.Mol.Biol.Rev.*, **69**, 326-356.

Renault, D., F. Déniel, E. Benizri, D. Sohier, G. Barbier, and P. Rey (2007) Characterization of *Bacillus* and *Pseudomonas* strains with suppressive traits isolated from tomato hydroponic-slow filtration unit. *Can J. Microbiol*, **53**, 784-797.

Rost, B., Sander, C., and Schneider, R. (1994) PHD--an automatic mail server for protein secondary structure prediction. *Comput.Appl.Biosci.*, **10**, 53-60.

Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief.Bioinform.*, **3**, 265-274.

Tobes, R., and J.L. Ramos (2002) AraC/XylS database: a family of positive transcriptional regulators in bacteria. *Nucl Acids Res.*, **30**, 318-321.