UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

CENTRO DE INVESTIGACION SOBRE FIJACION DE NITROGENO

Reconocimiento de Motivos Estructurales en Reguladores Transcripcionales

T E S I S

QUE PARA OBTENER EL GRADO DE:

DOCTOR EN CIENCIAS

P R E S E N T A:

ERNESTO PEREZ RUEDA

CUERNAVACA, MORELOS 1999

A la memoria de Fidel Rueda

A Silvia... la compañera de mis ilusiones...

A Mis padres Carmen y Daniel, a mis hermanos Esperanza, Ricardo, Daniel y Víctor...por todo lo que nos falta por hacer y vivir juntos.

A Julio Collado Vides por su amistad, confianza y los años trabajando juntos.

A mis grandes y queridos amigos... Brenda, Ramón, Agustino, Juan, Humberto, Cano y Josué... mil gracias.

A los Vaqueros de la aguerrida Facultad de Ciencias.

Mil gracias a todos los del laboratorio por su profesionalismo, pero principalmente a Conchita y Hely por los seis años que hemos pasado juntos.

A Alex, Lies, Ilse, Standra, SanToon, Lasse y Vasco... por toda su ayuda y confianza.

A Shoshana Wodak, Jean Richelle, Denis Thieffry y Jacques Van Helden... por sus paciencia al enseñar a aprender!!

A la Universidad Nacional Autónoma de México y al Centro de Investigación Sobre Fijación de Nitrógeno por permitirme continuar con mi formación académica y por enseñarme a ser Universitario.

Un reconocimiento especial al Jurado que me sugirió muchas cosas y se entusiasmo otras tantas.

Dr. Julio Collado Vides

Dr. Guillermo Dávila Ramos

Dr. Lorenzo Segovia Forcella

Dr. Antonio Lazcano Araujo

Dra. Carmen Gómez Eichelmann

Dra. Alicia González Manjarrez

Dr. Pedro Miramontes Vidal

Agradezco la valiosa ayuda de Dolores Cuéllar para la realización en los trámites administrativos... que paciencia de mujer!!

Este trabajo se realizó bajo la asesoria del Dr. Julio Collado Vides en el Programa de Biología Molecular Computacional del Centro de Investigación Sobre Fijación de Nitrógeno. Cuernavaca, Morelos México.

Durante mis estudios de Doctorado fuí apoyado por una beca de CONACYT y DGEP. Además del apoyo de PAEP No. 201365.

INDICE

RESUMEN

Un total de 314 reguladores transcripcionales fueron detectados en el genoma completo de *E. coli* K12. De ellos, 151 están descritos experimentalmente, mientras que 163 proteínas se predicen por métodos computacionales. La información posicional del motivo de unión al DNA o HTH fue utilizada para asignar la función a las proteínas reguladoras predichas. En el total de reguladores hay una proporción similar de activadores, represores y duales. Todos los reguladores se agrupan en alrededor de 20 familias evolutivas. Pensamos que una sobrerepresentacion de los reguladores de algunas familias nos mostraría procesos de duplicación-divergencia dentro de las familias de *E. coli.* Asimismo, la cantidad de reguladores se explicaría con base en la capacidad de *E. coli* de habitar y explotar diferentes ambientes. Adicionalmente, proponemos la presencia de un supergrupo funcional y evolutivo en términos de comparación de secuencias. Este supergrupo relacionaría aquellas proteínas que tienen el HTH en el N-terminal y que son represores de la transcripción. En conclusión, este es el primer trabajo donde se presenta la compilación de los factores transcripcionales de *E. coli* K12 y representa una importante aportación al área de la regulación transcripcional y del análisis de genomas completos.

ABSTRACT

314 transcriptional factors characterized and predicted were detected in the genome of *E. coli* K12. Protein predictions were performed by several computational approaches. The positional information of the helix-turn-helix DNA-binding motif was used to assign regulatory roles in putative protein factors. In our set, activator, repressor and dual proteins are distributed in similar proportion. All regulatory proteins are grouped in around 20 families. The overrepresentation of members in some families might be explained in terms of duplication-divergent processes inside the families and within of *E. coli*, and it explains the capacity of *E. coli* to live and exploit several environments. Additionally, we proposed one supergroup in terms of sequence analysis. This supergroup relates proteins with the classic HTH in the N-terminal position, and with repression roles. In conclusion, this work presents the first compilation and analysis of all regulatory proteins in *E. coli* K12 and represents one important contribution to the transcription regulation and genome analysis.

INTRODUCCION

El genoma completo de *Escherichia coli* cepa K12 consta de 4.6 Mb y se le han identificado alrededor de 4280 regiones codificantes (Blattner et al. 1997). Muchos de los genes actuales del genoma son derivados de duplicación génica y divergencia (Labedan et al.1995), aunque también se han descrito eventos de transferencia horizontal (Lawrence et al.1998). De hecho, se ha sugerido que aproximadamente el 25% de todas las proteínas de *E. coli* pertenecen a cuatro grandes supergrupos: Permeasas; proteínas reguladoras con el motivo hélice-vuelta-hélice (HTH); ATPasas y GTPasas con el motivo *Walker-type* conservado; y proteínas que pegan NAD (FAD) (Koonin et al.1995).
E. coli representa sin lugar a dudas el organismo mejor conocido en biología y en donde muchos de los procesos celulares se han dilucidado. Uno de los procesos mejor descritos en esta bacteria, es el del mecanismo de la transcripción y su regulación a nivel inicio (Collado-Vides et al.1991; Gralla, 1991).

Mecanismo de la transcripción en procariotes

El evento central en la transcripción es el copiado de la secuencia templado de DNA en una cadena complementaria de RNA. La RNA polimerasa (RNApol) cataliza una cadena de RNA que puede funcionar como un adaptador en la traducción a una proteína o bien puede formar un RNA estructural (RNA de transferencia o ribosomal). En otras palabras, el mecanismo de la transcripción es uno de los procesos centrales en el desarrollo y crecimiento celular, en el cual los genes codificados en el DNA son selectivamente localizados, reconocidos y transcritos en RNA mensajeros y RNA estructurales (von Hippel et al.1984. von Hippel, 1998). En *E. coli* la frecuencia del inicio de la transcripción es determinada por la fuerza inherente de los promotores y como resultado de los mecanismos regulatorios (Neidhardt et al.1990).
El inicio de la transcripción involucra varios elementos proteicos que forman complejos funcionales. Uno de dichos complejos es la RNA polimerasa (RNApol) que en *E. coli* y en procariotes es la enzima encargada de sintetizar los diferentes tipos de RNA (McClure, 1985). La enzima básica o

core mide aproximadamente 90x95x160 Å, pesa 480 kD y consiste en un complejo multimérico con cuatro subunidades: dos subunidades α de 35kD cada una (codificadas por los genes *rpoA*); una subunidad ß de 155kD (codificada por *rpoB*) y una subunidad ß´ de 165kD (*rpoC*) (Record et al.1996). La subunidad α, interactúa con algunos activadores, tales como Crp y OxyR (Ishihama, 1993), y es parte importante en el ensamblaje de la enzima. La subunidad ß, se encarga del pegado de los nucleótidos, mientras que la subunidad ß' se encuentra unida a la cadena templado del DNA. En condiciones normales de crecimiento, el factor σ70 se une a la enzima básica formando la holoenzima (RNApol-factor sigma). El factor σ70 sólo se une al DNA cuando está en contacto con la enzima básica, requiriéndose para el reconocimiento específico del promotor en el inicio de la transcripción (Gralla, 1990). La afinidad de la RNApol aumenta por el DNA (promotor) en un orden de 10^7 veces cuando el factor σ esta asociado al *core* de la polimerasa. La holoenzima cataliza la transferencia de un ribonucleósido monofosfato al 3´-OH terminal de la cadena del RNA en crecimiento, usando ribonucleósidos trifosfatos como sustrato (von Hippel et al.1984). Figura 1.

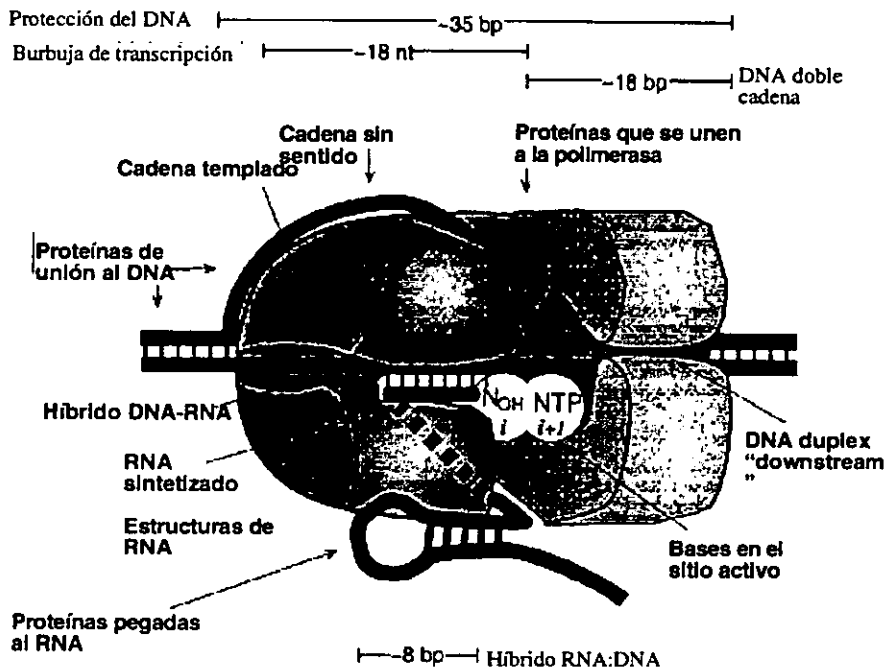


Figura 1 . Mecanismo de la transcripción. Se señalan los sitios activos de la RNApol y la región del DNA que es protegido por la enzima. (Mooney et al.1998).

E. coli expresa varios tipos de factores sigma adicionales al del tipo σ70 (Gross et al.1992) los cuales se han clasificado en dos familias con base en la comparación de secuencias. La primer familia está relacionada al factor σ 70 primario de *E. coli* (Lonetto et al.1992); mientras que el segundo grupo es similar al factor σ 54 alternativo (Gross et al.1992). Los promotores son reconocidos por las holoenzimas conteniendo algún factor σ y presentan secuencias conservadas para cada factor asociado. Tabla 1.En general, las holoenzimas con factores σ70 forman complejos abiertos en ausencia de factores auxiliares, mientras que las que contienen al factor σ54 requieren proteínas activadoras para la formación del complejo abierto (Gross et al.1992).

Sigma	Organismo	Región -35	Región -10	Función
σ70	*E. coli*	**TTGACA**	**TATAAT**	Sigma primario. *Housekeeping*
σ32	*E. coli*	**TCTC-CCCTTGAA**	**CCCCAT-TA**	Respuesta a *heat-shock*
σ38	*E. coli*	**G-T-AGC**	**-C-TCC**	Fase estacionaria y resistencia a peróxido
σ24	*E. coli*	**AACT**	**AAAAA---TCTGA**	Respuesta a *heat-shock*
σ28	*E. coli*	**CTAAA**	**GCCGATAA**	Síntesis de flagelos
σA	*B. subtilis*	**TTGACA**	**TATAAT**	Sigma primario. *Housekeeping*
σB	*B. subtilis*	**AGGTTTAA**	**GGGTAT**	Desconocido
σD	*B. subtilis*	**CTAAA**	**CCGATAT**	Biosíntesis de flagelos. Fragmentación celular
σE	*B. subtilis*	**ATATT**	**ATACA**	Esporulación. Expresión en la célula madre
σG	*B. subtilis*	**TGAATA**	**CATACTA**	Esporulación. Forespora.
σK	*B. subtilis*	**AC**	**CATA--T**	Esporulación. Expresión en la célula madre
σH	*B. subtilis*	**CAGGA**	**GAATT--T**	Esporulación. Septación. Competencia
σgp28	*Fago SPO1*	**AGGAGA**	**TTT-TTTa**	Expresión media
σgp55	*Fago T4*	**Ninguna**	**TATAAATA**	Expresión tardía
Sigma	Organismo	Región -24	Región -12	
σ54	*E. coli, S.typhimurium, K.pneumoniae, Rhizobium sp.*	**CTGGA-A**	**TTGCA**	Asimilación de Nitrógeno

Tabla 1. Promotores reconocidos por la RNA polimerasa conteniendo varios factores sigma. *Ta* de SPO1 representa hidroximetiluracil que reemplaza a la Timina (Modificado de Gross et al.1992).

Localización del promotor

Un elemento importante para la interacción DNA-RNApol es la arquitectura del promotor. El promotor es un segmento de DNA que contiene las señales que dirigen el pegado específico de la holoenzima y su subsecuente activación a una forma capaz de iniciar la transcripción. La comparación de secuencias en regiones reconocidas por el factor σ70 muestran la presencia de dos hexanucleótidos conservados (Lisser et al.1993). La caja de Pribnow (TATAAT), se localiza aproximadamente 10 pares de bases *upstream* o corriente arriba con respecto al inicio de la transcripción (+1). El segundo hexanucleótido (TTGACA) se localiza 35 pb *upstream* respecto al +1. Ambas cajas tienen una región espaciadora que varía de los 15 a los 21 pb de longitud; mientras que una segunda secuencia espaciadora se presenta entre el -10 y el +1 con una longitud de entre 5 a 9 pb (Lewin, 1994). Tabla 1. En algunos promotores se ha descrito el elemento *UP-like* rico en AT y que se localiza aproximadamente 20 pb *upstream* del -35 (Ross et al.1993). Esta región está implicada en la formación de contactos adicionales con la región del C-terminal de la subunidad α (αCTD) de la RNApol y en aumentar la tasa del inicio de la transcripción.

En el caso de los promotores reconocidos por la RNApol-σ54 se han descrito dos dinucleótidos invariables en la posiciones -24 (GG) y –12 (GC) (Tabla 1). Aunque la RNApol-σ54 se une al promotor y forma el complejo cerrado, la transición hacia el complejo abierto requiere la presencia de un activador que se une a sitios localizados de 100 a 150 pb *upstream* con respecto al +1 (Collado-Vides et al.1991; Gralla, 1996). Los activadores transcripcionales en este tipo de promotores ayudan a la isomerización del complejo cerrado al abierto involucrando la hidrólisis de ATP.

La formación del complejo de iniciación

La formación de un complejo estable promotor-RNApol, depende de la fuerza del promotor. La fuerza del promotor está definida por dos componentes, el de equilibrio que relaciona el grado de ocupación del promotor (complejo cerrado) con una concentración dada de la holoenzima y el

componente cinético que relaciona la velocidad a la cual se isomeriza el complejo cerrrado a complejo abierto (McClure, 1980).

El complejo cerrado se forma por la interacción específica de la holoenzima y su promotor. Dicho complejo se isomeriza espontáneamente a una burbuja de transcripción en el cual alrededor el DNA se abre en una longitud de 12 pares de bases –incluido el +1- (Gralla, 1990). La formación del complejo abierto puede ir acompañado de cambios conformacionales en la RNApol (Mooney et al.1998). Además de los cambios conformacionales, se han descrito múltiples complejos abiertos y cerrados que se forman como intermediarios durante el proceso de reconocimiento del promotor.

En el complejo abierto, la doble cadena del DNA se desnaturaliza a través de una reacción irreversible en la zona de contacto entre el DNA y la RNApol (alrededor del -10 y el +1). La burbuja de transcripción se produce por la desestabilización de los puentes de hidrógeno que mantienen unida la doble hélice de l DNA. Adicionalmente se produce un desenrollamiento local que inicia en el sitio de unión de la holoenzima y expone la cadena templado para su complementación con los ribonucleótidos a una tasa de 40nt/seg (von Hippel et al.1984). La RNApol-σ70 protege alrededor de 75 pb al formar el complejo abierto y 60 pb cuando se libera el factor sigma, sintetizando y elongando el RNAm a lo partir de la cadena templado (Mooney et al.1998, von Hippel et al.1984). Figura 2.

Regulación del inicio de la transcripción

Los genes se organizan en operones y/o regulones que responden a estímulos ambientales externos e internos, por ejemplo, el operon lactosa que agrupa tres genes (*lacZYA*) para la asimilación de la lactosa (Beckwith 1987. Lewin, 1994) o el regulon para la biosíntesis de arginina que incluye alrededor de 11 operones y/o genes conocidos (Charlier et al.1992). Una de las ventajas que ofrece la organización de los genes en sistemas coordinados es que pueden ser regulados para una respuesta o función conjunta o sincronizada (Neidhardt et al. 1996). La regulación, es por lo tanto, común a ese conjunto de genes y se da principalmente en el inicio de la transcripción, aunque también se le ha descrito en niveles posteriores.

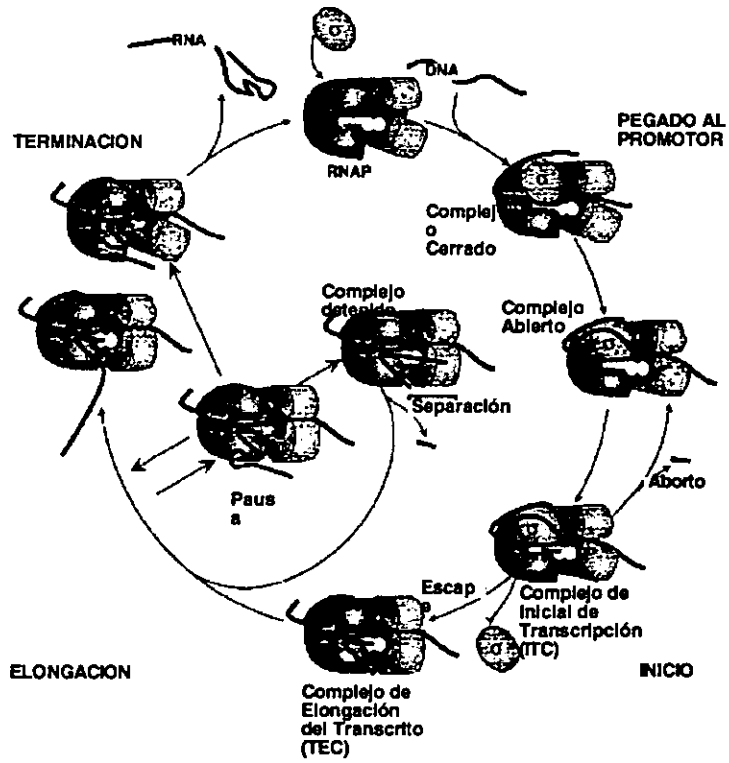


Figura 2. El ciclo de la transcripción. Se presentan las etapas de la transcripción: Reconocimiento del promotor, inicio del transcrito, elongación del RNA mensajero y la terminación. (Tomado de Mooney et al.1998)

En la fase del inicio de la transcripción, las regiones intergénicas compiten por la cantidad limitada de la RNApol. Cada gene para ser competitivo, es activado por múltiples factores que incrementan la afinidad de su promotor por la polimerasa y facilitan la apertura del DNA, la síntesis inicial del transcrito, su elongación o el inicio de otro ciclo (von Hippel, 1998).

El inicio de la transcripción está algunas veces acoplado a eventos de represión y que involucran el pegado de proteínas al promotor o bien a sitios más distantes que actúan por medio de *loops* (Matthews, 1992). En resumen, la regulación transcripcional implica la modulación de la especificidad y afinidad de la holoenzima por el promotor, y está mediada por proteínas que interactúan en sitios cercanos o sobrelapados al promotor activando o reprimiendo la transcripción del gene (Collado-Vides et al.1991).

En el mecanismo de la represión, las proteínas reguladoras bloquean al promotor (eliminándolo del grupo de promotores que compiten por la Eσ70), ya sea bloqueando el pegado de la RNApol al promotor, impidiendo la formación del complejo cerrado o evitando la isomerización del complejo cerrado a complejo abierto (von Hippel, 1998). La región preferencial de contacto de estas proteínas al DNA va del -50 al +20 con respecto al +1 (Collado-Vides et al.1991).

Para los genes bajo control positivo, la expresión únicamente es posible cuando la proteína reguladora está presente (Gralla, 1996). Las proteínas activadoras incrementan la afinidad de la holoenzima por el promotor, la tasa de formación del complejo abierto o bien aumentan la velocidad con que la holoenzima abandona el promotor (von Hippel, 1998). Los activadores se unen al DNA en regiones *upstream* del promotor y que van del -80 al -30 (Collado-Vides et al.1991). Figura 3.

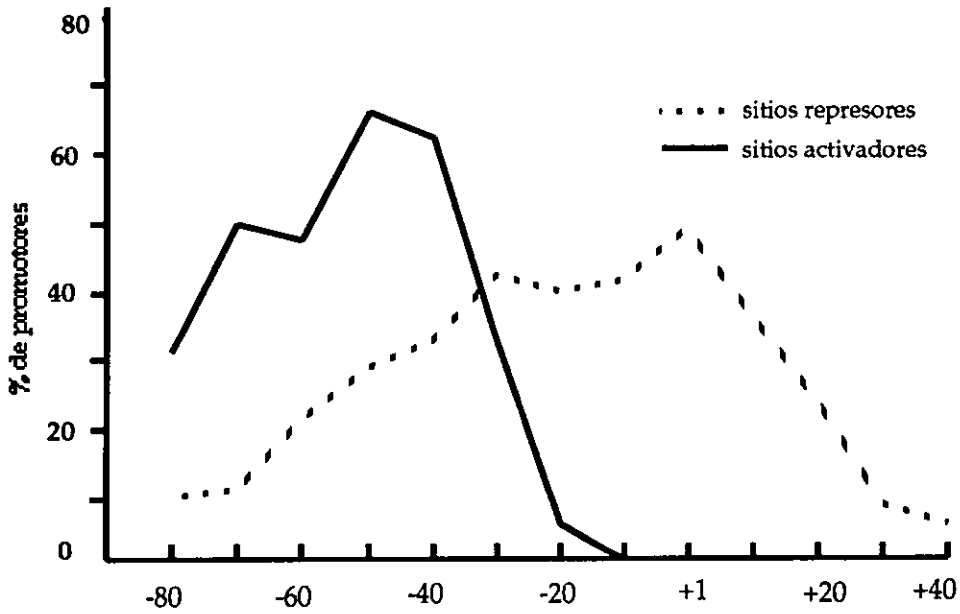


Figura 3. Sitios de pegado al DNA de las proteínas reguladoras asociadas al factor sigma70 de *E. coli*. (Tomado de Collado-Vides et al.1991)

Proteínas Reguladoras

La regulación de la expresión genética es uno de los temas centrales para entender los mecanismos de diferenciación y/o expresión celular. Para que una célula se diferencie o responda a una función

definida es necesario que se presenten diferentes condiciones ambientales. La presencia de estos componentes en el medio extra e intracelular está ligado a la presencia de proteínas que regulan la expresión de los genes para la utilización de los compuestos disponibles. Los factores proteicos van desde los represores del inicio de la transcripción hasta las proteínas.

Los reguladores transcripcionales presentan una organización con dominios para diversas funciones: multimerización, pegado al DNA y unión a inductores, entre otros (Kaptein, R, 1993). Estos dominios pueden conservarse entre las diferentes familias y ha revelado –en el caso particular del dominio de pegado al DNA- una amplia variedad de diseños: pueden plegarse y presentarse como protuberancias en la superficie molecular o bien como estructuras flexibles que se extienden a través de la proteína para contactar con las bases del DNA. Estos contactos ocurren generalmente en el surco mayor del DNA e incluyen puentes de hidrógeno, contactor mediados por moléculas de agua (represor del operón triptófano) o interacciones de Van der Waals (Harrison, S.C. et al.1990).

Una característica en este tipo de proteínas reguladoras en procariotes es la estructura de pegado al DNA del tipo hélice-vuelta-hélice (HTH). Está estructura es importante para clasificar las proteínas en grupos. El HTH se describió originalmente en la el regulador CRO y el represor del fago lambda siendo una de las estructuras más diversificadas en los reguladores transcripcionales tanto de procariotes como de eucariotes. También se le ha descrito en algunas enzimas que no se unen al DNA (Suzuki, M. et al.1995), así como en proteínas que estabilizan el RNA ribosomal y el RNA de transferencia (Yonath, A. et al. 1997. Xing, Y. et al.1997). El HTH consiste de alrededor de 20 aminoácidos de longitud dividido en dos α-hélices que adoptan un ángulo de 120°. La primer α-hélice, esta constituída de aminoácidos básicos que interaccionan con los grupos fosfato del DNA. La segunda α-hélice reconoce específicamente los nucleótidos expuestos en el surco mayor(Wharton, R.P, et al, 1984). Ambas hélices están separadas por una vuelta que contiene predominantemente Glicina. Muchas proteínas tienen tres α-hélices que pueden formar una estructura globular más pequeña y ayudan a estabilizar la hélice de reconocimiento. Tabla 2 y Figura 4.

Recientemente se ha descrito un HTH diferente del tipo clásico. A este tipo de HTH se le denominó *winged* o alado. En OmpR se le ha descrito como una estructura con tres α-hélices flanqueadas por

hojas-ß antiparalelas y un anillo C-terminal que interactúa con una cadena corta de hojas-ß y que conecta las α-hélices 1 y 2 formando una hoja-ß de tres cadenas (Brennan, R.G. 1993). Tabla 2.

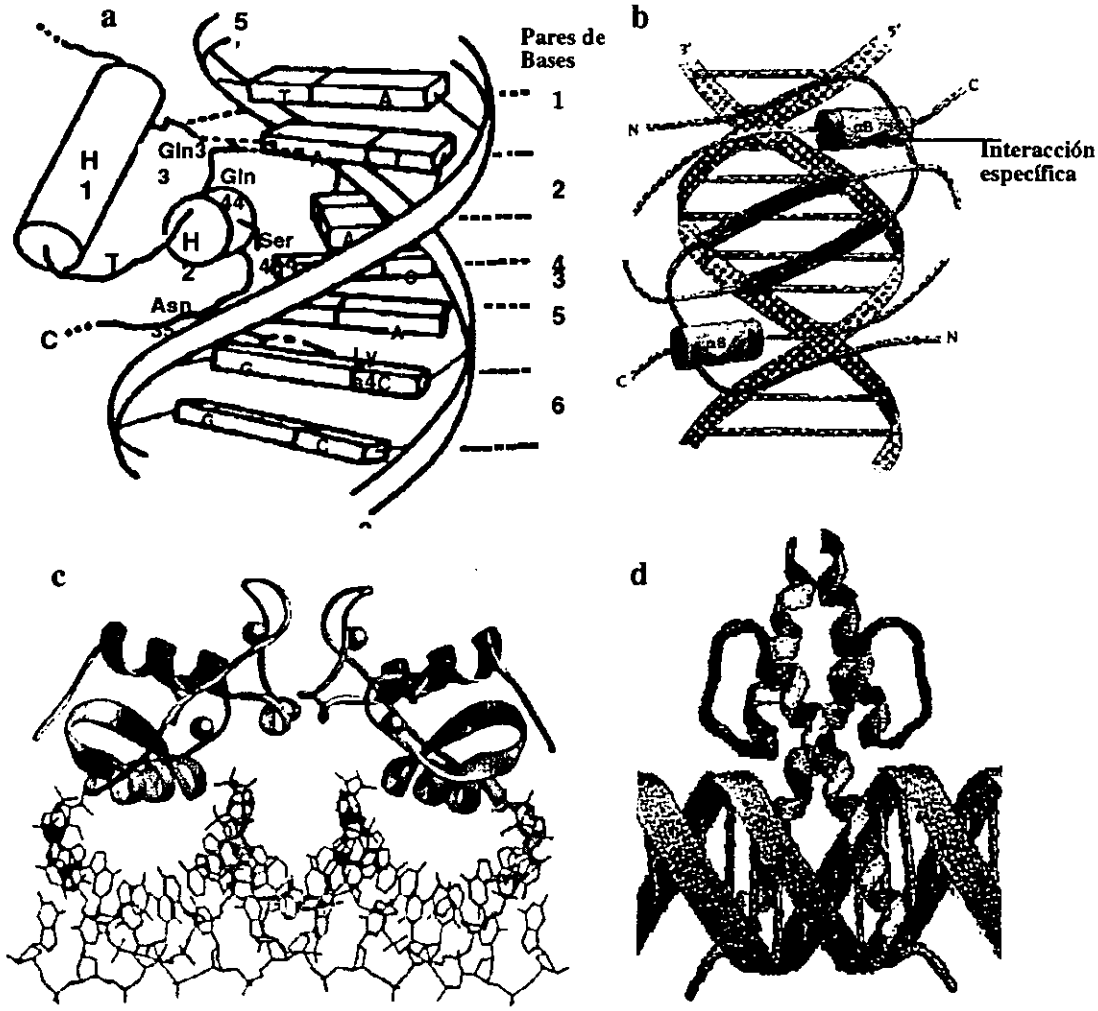


Figura 4. Estructuras de pegado al DNA descritas en reguladores de *E. coli*: a) Hélice-vuelta-hélice, b) ß-plegada antiparalela, c) Dedos de zinc y d) Hélice-loop-hélice. (Tomado de Branden et al.1991)

La estructura ß-plegada antiparalela también ha sido descrita en los procariotes. El motivo está constituido por 3 α-hélices y una ß-plegada que ocupan aproximadamente el 50% de la secuencia en los reguladores MetJ y ArcA. Al asociarse las proteínas en dímeros, las hojas ß-plegadas forman la estructura funcional *ß-plegada antiparalela* que se expone en la superficie de la proteína a manera de

una protuberancia (Branden, C. et al.1991), reconociendo los nucleótidos del surco menor del DNA (Stragier, P. et al. 1996). Ver tabla 2.

Motivo	Sitio de Interacción	Ejemplos
Hélice-vuelta-Hélice	Surco Mayor. La orientación dentro del surco mayor es variable	CRP, LacI, Fur, BirA, PurR*
Hélice-Loop-Hélice	El reconocimiento se realiza por la α-hélice	DnaA
Dedos de Zinc	Contactan bases en el surco Mayor	HypF
Hojas ß	Contactan bases en el surco Menor	MetJ, ArcA,

Tabla 2. Estructuras de unión al DNA reportadas en *E. coli*. *PurR contacta el surco mayor y el surco menor.

Familias de proteínas reguladoras

Las proteínas reguladoras se han clasificado con base en la comparación de secuencias de aminoácidos. Dicha comparación ha mostrado la existencia de alrededor 20 familias en procariotes.
En SwissProt y Prosite (Bairoch.1992) se describen muchas de las familias de reguladores transcripcionales de procariotes, tales como ArsR (Bairoch 1993), AraC/XylS (Gallegos et al.1997), LysR (Henikoff et al.1988, Schell.1993), Crp/Fnr (Spiro.1993), GalR/LacI (Vartak et al. 1991, Nguyen et al. 1995, Weickert et al.1992), LuxR (Pao et al. 1995), EBP (Morett et al.1993) y de OmpR (Pao et al. 1995). Muchos de los miembros de las familias LuxR, EBP y OmpR forman parte del sistema de los dos componentes (Stock et al.1989).
Las funciones reguladas por las proteínas que forman parte de una misma familia tienden a ser homogéneas, por ejemplo, la mayoría de los reguladores de la familia GalR/LacI regulan la expresión de los genes para el catabolismo de fuentes de carbono (Weickert et al.1992), mientras que las proteínas de la familia LysR regulan preferentemente a genes para la biosíntesis de aminoácidos (Henikoff et al.1988). Las proteínas de la familia EBP regulan los genes para la asimilación de nitrógeno o algunos procesos que no son totalmente requeridos para que la sobrevivencia de la célula

(Morett et al.1993). Finalmente, en la familia AraC/XylS se agrupan reguladores involucrados en el metabolismo de fuentes de carbono y en mecanismos de patogénesis (Gallegos et al.1997).

ANTECEDENTES

El genoma completo de *E. coli* cepa K12 (Blattner et al.1997) ofrece la oportunidad de dilucidar el panorama completo de la regulación transcripcional en esta bacteria. Hasta el momento existe una base de datos en donde se describen los elementos de la regulación en este organismo: RegulonDB (Huerta et al.1998. Salgado et al.1999). En dicha base de datos, se ha almacenado información respecto a la organización en operones, sitios de pegado al DNA para proteínas reguladoras, mecanismos de regulación (activación, represión y actividad dual), y promotores, entre otros. Sin embargo, solo una pequeña fracción de los reguladores transcripcionales de *E. coli* se han almacenado en RegulonDB debido a la falta de evidencias experimentales (Thieffry et al.1998), por lo que la detección de todos los reguladores en *E. coli* nos ayudará a completar el panorama general de la regulación transcripcional en dicha bacteria.

Un análisis de la posición del HTH en la secuencia de 227 reguladores transcripcionales de procariotes, mostró una distribución uniforme de grupos funcionales asociados a la posición del motivo de unión al DNA. Esta distribución muestra que la mayoría de los represores presentan el HTH en posición N-terminal, mientras que los activadores lo presentan en el C-terminal. Las proteínas duales están constituidas principalmente por miembros de la familia LysR y presentan el HTH en el N-terminal (Pérez-Rueda et al.1998).

Los grupos obtenidos a partir de la distribución del HTH y la asociación a la actividad regulatoria nos mostró que hay regiones preferenciales de localización del motivo de unión al DNA. Asimismo, hay grupos en donde el HTH se localiza en un intervalo más estrecho y grupos donde la distribución del HTH es más amplia. Dichos grupos corresponden a familias evolutivas. La relación nos hace pensar en la hipótesis de un HTH ancestral al que se le fueron adicionando fragmentos proteicos que determinaron posteriormente su función. Ver a Pérez-Rueda and Collado-Vides,1998 (Artículo anexo).

Artículo "Pérez-Rueda, E., Gralla, J.D. and Collado-Vides, J. 1998. Genomic Position Analyses and the Transcription Machinery. J. Mol. Biol. 275:165-170"

J. Mol. Biol. (1998) 275, 165–170

JMB

COMMUNICATION

Genomic Position Analyses and the Transcription Machinery

Ernesto Pérez-Rueda[1], Jay D. Gralla[2] and Julio Collado-Vides[1]*

[1]*Centro de Investigación Sobre Fijación de Nitrógeno Universidad Nacional Autónoma de México Cuernavaca A.P.565-A Morelos 62110, México*

[2]*Department of Chemistry and Biochemistry and the Molecular Biology Institute, University of California, Los Angeles CA 90095, USA*

**Corresponding author*

Position analyses have been devised to extract additional transcriptional information from rapidly expanding genomic data bases. The locations of promoter regulatory sites and also the locations of transcription factor DNA-binding domains are analyzed. Strongly preferred positions of activator binding sites occur in both *Escherichia coli* and eukaryotes, suggesting specific common features of transcription in the two systems. In both systems, regulatory proteins are found to have their DNA-binding domains near termini and the data suggest an evolutionary analysis that complements a phylogenetic analysis based on sequence alignments. The results indicate that positional information can be an important adjunct to sequence comparisons in analyzing genomic information.

Keywords: positional analysis; transcriptional machinery; protein domains; operator binding sites

As genome sequences are analyzed what new are we learning about transcriptional control? For the most part the mass of accumulating data has been used to identify genes for transcription factors. Recently, analysis of bacterial systems raised the possibility that there might be an important additional level of information. The location of transcription elements along the genome has a clear influence on how regulation through those elements occurs (Gralla & Collado-Vides, 1996). The influence of such preferred positions is known for selected aspects of the *Escherichia coli* transcription machinery. In this paper we extend the location analysis to extract information regarding both extragenic promoter elements and intragenic transcription factor domains. The results lead to interesting conclusions concerning the organization and evolution of the transcription apparatus. Initial extension of the analysis to mammalian genome sequences suggests that it has the potential to add a new level of information useful in understanding mammalian transcriptional control.

We begin by extending the analysis to analyze locations of proximal activator sites in *E. coli* and mammalian promoters. We chose to use two common and well characterized activators, CRP in *E. coli* and Sp1 in mammalian cells.

Figure 1(a) shows that CRP protein has strongly preferred locations for binding as an activator. The source of these preferences is clearly functional. CRP functions *via* contacts to the polymerase using two primary activating regions (Busby & Ebright, 1994; Niu *et al.*, 1996). Such contacts can be made optimally only from certain locations (Gaston *et al.*, 1990; Niu *et al.*, 1996), which show up as preferred binding sites in Figure 1(a). The optimal contacts are in fact intimately related to function because moving the CRP sites just a few base-pairs away from these few preferred positions leads to strong reductions in transcription (Gaston *et al.*, 1990; Niu *et al.*, 1996).

In an attempt to learn if such principles might apply to mammalian activation we analyzed the locations of the binding sites for the Sp1 activator that are collected in the TRANSFAC data base (Wingender *et al.*, 1996). Sp1 sites were analyzed seven years ago and at that time it was found that they occurred preferentially near position −50 (Bucher, 1990). Figure 1(b) shows the distribution of sites that now exist in the TRANSFAC data

Abbreviations used: CRP, catabolite receptor protein; Sp1, transcription factor Sp1; TRANSFAC, the transcription factor data base; TAF, transcriptional activator factor; HTH, helix-turn-helix DNA-binding motif; LuxR, LuxR regulatory protein family; AraC, arabinose regulatory protein; DeoR, DeoR regulatory protein family; GalR, GalR regulatory protein family; AsnC, AsnC regulatory protein family; GntR, GntR regulatory protein family.

0022-2836/98/020165-06 $25.00/0/mb971465

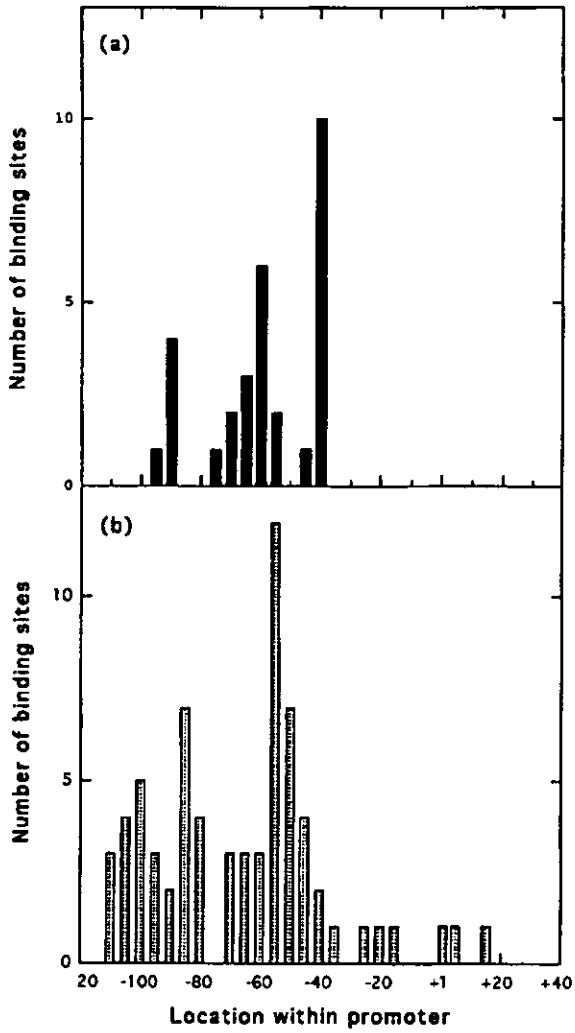


Figure 1. Distribution of CRP and Sp1 binding sites in promoters. (a) The 43 binding sites for CRP protein in *E. coli*-activated promoters (Gralla & Collado-Vides, 1996) are plotted in 5 bp intervals, upstream from the position indicated, with +1 being the start of transcription. (b) Sp1 sites were obtained from TRANSFAC (Wingender *et al.*, 1996). Redundant entries were eliminated, as well as sites longer than 45 bp, or sites beyond position −100, leaving 113 sites. When multiple sites were present each was counted separately. The center of the consensus DNA sequence was identified and assigned to a 5 bp window between −100 and +20, with +1 being the major start of transcription. Multiple transcriptional starts can occur at these promoters and the positions of Sp1 sites with respect to these minor sites were not analyzed.

base, over the same region displayed for CRP, from −100 to +20.

The results show that Sp1, like CRP, has several closely spaced preferred locations for binding in this region. Binding just upstream from −50 is by far the preferred situation for Sp1; CRP has a strongly preferred site just upstream from −40. Secondary preferred binding sites, with somewhat lower statistical significance, occur slightly further upstream for both activators, near −80 for Sp1 and near −60 for CRP. In both cases tertiary preferred sites appear to occur slightly further upstream.

The common property of preferred site locations within the promoter proximal region raises possibilities for analogous mechanisms of transcriptional activation in prokaryotes and eukaryotes. CRP activates differently from different locations. For example, when it binds at −62 it recruits polymerase primarily *via* a single contact region whereas at −42 CRP can use a different activating surface (Busby & Ebright, 1994; Gaston *et al.*, 1990; Niu *et al.*, 1996). If prokaryotic activators are bound simultaneously to two different closely spaced locations activation can be synergistic, presumably because two different polymerase contacts are made (see Busby *et al.*, 1994; Joung *et al.*, 1994). A large set of bacterial promoters with multiple sites for the same protein were analyzed, showing that the most common arrangement involves phasing at 11-base-pair intervals (data not shown; the method was that of Haykinson & Johnson, 1993). This organization supports the importance of conservation of position, although the interval suggests that in prokaryotes the site synergy may be primarily *via* cooperative interactions along a helix face.

These issues, synergy and recruitment, are central to mammalian activation but have not been discussed in the context of site location (Chi *et al.*, 1995; Roberts *et al.*, 1995; Sauer *et al.*, 1995; Ptashne & Gann, 1997). The locations of closely spaced preferred Sp1 sites suggest the new possibility that Sp1, and perhaps other mammalian activators, might work differently from different locations. As one example, Sp1 bound near −50 could contact a TAF (Hoey *et al.*, 1993) and work synergistically with another Sp1 or other activator that binds in a different location and thus is poised to make a different contact. The new idea that is raised is not synergy from multiple contacts but rather the possibility that a single activator may have multiple activation mechanisms by virtue of making different contacts to the general transcription factors from different locations. This has not yet been tested by moving Sp1 sites and probing for the consequences, but the results of this location analysis suggest that such experiments would be informative.

Another strong similarity in the two systems (Figure 1) is the remarkably small number of activation sites downstream of −35. In the prokaryotic case this phenomenon applies to all activators. Indeed, in the rare cases when activator sites appear in this location the proteins are converted to repressor function (Gralla & Collado-Vides, 1996). The small number of Sp1 sites in this region raises similar possibilities in this eukaryotic case, that is, that mammalian activators may be converted to repressor function by placing their binding sites downstream of −35. This idea, raised by

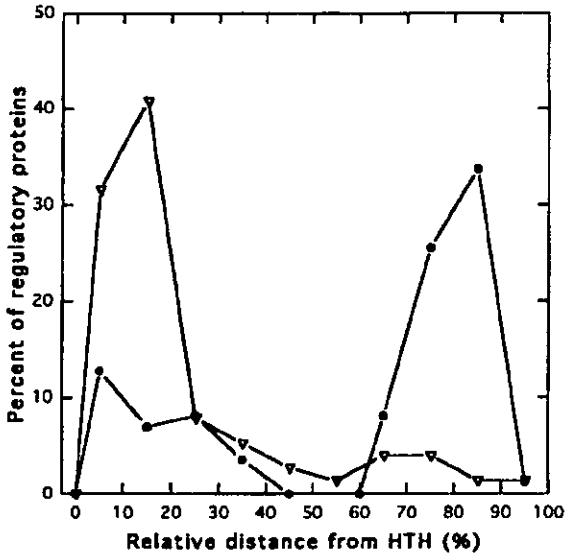


Figure 2. Location of HTH domains in bacterial transcription factors. 232 bacterial transcriptional regulators containing a HTH motif were obtained from Swiss-Prot release 31.0 (Bairoch & Apweiler, 1996; Dodd & Egan, 1987). The center of the each motif was found and attributed to a window with respect to the normalized length of each protein. The number of proteins that have the domain centered within each 10% window were counted and plotted as shown. 0% refers to the N terminus and 100% refers to the C terminus. The number of proteins is normalized separately for 86 activators (filled circles) and 76 repressors (open triangles). The distribution shown was not changed when orthologous proteins (analogous proteins from different organisms) were considered as one entry, leaving a minimum of 70 proteins.

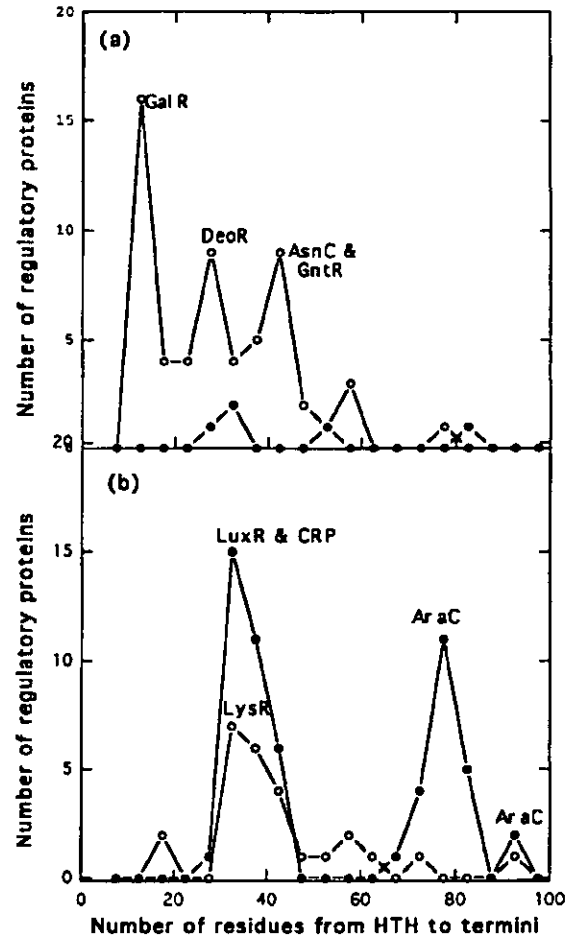


Figure 3. Absolute position of HTH domains in bacterial transcription factors. (a) For the repressor proteins, the number of amino acids from the center of the HTH motif to the terminus was calculated. Open circles show the distribution of distances between the HTH and the N terminus for HTH domains within the first 100 amino acids. Filled circles show the distribution of distances between the HTH and the C terminus. The major protein families present in each peak are indicated: DeoR, GalR, AsnC, and GntR. (b) As in (a), activator proteins are analyzed. Filled circles refer to distances from the C terminus and open circles refer to distances from the N terminus. The major protein families present in each peak are indicated: CRP, LuxR, and AraC. AraC is the only family with the HTH proteins defining more than one peak; a minority of AraC members have an approximately 15 amino acid C-terminal insertion.

comparative location analysis of prokaryotic and eukaryotic genomes, has not yet been tested.

The above analysis applies to the locations of elements within promoters. Genome analysis also gives information about the proteins that bind these sequences (Crowley *et al.*, 1997). We attempted to extract additional information about the transcription apparatus by analyzing intragenic positions of domains. The most common regulator domain in prokaryotes is the DNA-binding helix-turn-helix (HTH) motif. A total of 232 bacterial regulatory protein sequences in the Swiss-Prot data base were analyzed with regard to the location of their HTH (Bairoch & Apweiler, 1996; Dodd & Egan, 1987). These include activators, repressors and dual function regulators.

The results show that functional groups of proteins largely segregate on the basis of the location of their DNA-binding domains. Figure 2 shows that most repressors (open triangles) have such domains located near their N terminus, whereas most activators (filled circles) have their domains located near their C terminus. The large LysR family of dual regulators forms a distinct group with a repressor-like location (not shown). There is no apparent functional reason why repressor proteins should have N-terminal DNA-binding domains, whereas activators should bind C-terminally, suggesting that the groupings are related to evolutionary issues.

The idea that each functional type of protein forms an evolutionary grouping in which the location of the DNA binding domain has been strongly preserved is supported by the analyses shown in Figure 3. This Figure plots the absolute

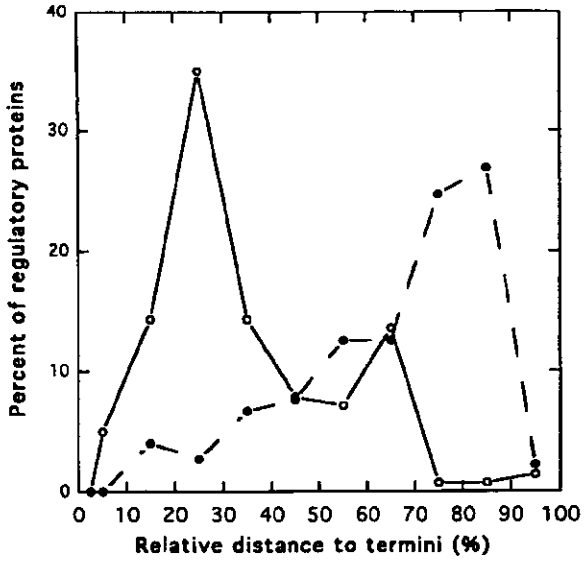


Figure 4. Location of DNA-binding domains in homeodomain and zinc finger eukaryotic transcription factors. 223 homeodomain proteins (filled circles) and 140 proteins with two zinc fingers (open circles) were analyzed for the distribution of the centers of their DNA-binding domains as described for Figure 2. 0% refers to an N-terminal location and 100% to a C-terminal location. These proteins were obtained from Swiss-Prot release 31.0.

number of amino acids between the termini and the HTH domain. Sharp sub-peaks are observed where large groups of proteins have retained a precise position of the DNA-binding domain. This occurs within a heterogeneous group of proteins, including large proteins like the 360 amino acid Lac repressor and proteins less than a quarter of this size such as the phage Cro dimer. The large majority of repressors have their HTH domain located between 15 and 50 amino acids from the N terminus (Figure 3(a), open circles). It is interesting to note that the largest group of activators have their HTH at a similar distance from the C terminus, 30 to 45 amino acids distant (Fig. 3(b), filled circles). For the minority of activators with an HTH in an N-terminal location this same distance applies; Figure 3(b) (open circles) shows that the HTH domains are typically 30 to 45 amino acids distant from the terminus, in this case the N terminus. All families analyzed showed groupings, indicating conservation of both absolute and relative positions (data not shown).

The overall conservation of HTH location within 50 amino acids of a protein terminus is clearly non-random and requires explanation. The 232 regulatory proteins represent more than 70 non-orthologous proteins. The conservation of location suggests that there were a very small number of ancestors for this large group of diverse proteins. The data even raise the possibility that there was a single small ancestral HTH protein of less than 50 amino acids. This ancestor could have been joined on one terminus, ultimately yielding most activators, and independently joined on its other terminus, ultimately yielding most repressors and dual function proteins. This would lead to the existence of the two superfamilies observed today.

Further analysis suggests that positional subgroups exist that share more detailed function. The HTH activators (Figure 3(b)) display two major peaks, the greater one at position 35 and the lesser one at position 75. The greater peak includes the known sequence-related proteins of the CRP and LuxR types and the lesser peak includes the AraC family of proteins. Each sub-peak also contains proteins that were not identified previously as being related. Thus, the position analysis suggests additional possibilities for close relatedness among the proteins in the genomic data base. In this manner this position analysis complements the known phylogenetic family analysis based on simple sequence comparisons (Schleif, 1996; Nguyen & Saier, 1995; Pao & Saier, 1995; Ninfa, 1996).

The results suggest that it may be possible to begin to apply the analysis to the vast number of eukaryotic regulatory protein sequences that are piling up in eukaryotic genomic data bases. As an initial test we analyzed three common types of eukaryotic DNA-binding regulators, homeobox, zinc finger proteins, and helix-loop-helix proteins. Figure 4 (filled circles) shows the location of the homeobox domains in 223 proteins obtained from the Swiss-Prot data base (Bairoch & Apweiler, 1996). The result shows that a C-terminal location is strongly preserved in this large group of proteins. A more detailed analysis shows that approximately one-third of proteins in this very diverse group have their HTH domain located in a narrow range within ten amino acids of position 50 from the C terminus (data not shown). The C-terminal position is similar to the HTH position of bacterial activators, but there is not yet enough functional information available to decide if the correlation is meaningful.

This intragenic position analysis was also applied to the 140 double zinc finger proteins in the data base (Bairoch & Apweiler, 1996). These form a very diverse group with regard to function. Nonetheless the result (Figure 4, open circles) shows that the location of the double zinc fingers is N-terminal and decidedly non-random. Moreover the location is quite different from the DNA-binding homeodomains, those typically being in the C-terminal region. There is no apparent functional reason that homeoproteins should have C-terminal domains, whereas zinc finger proteins should have N-terminal domains. Thus, as argued for the analogous prokaryotic results, it appears that DNA-binding domains appeared very early in the evolution of these many regulatory proteins and diversity was built onto this early base.

The analysis of a third family of eukaryotic proteins, the helix-loop-helix regulators, shows two prominent preferred positions for the HLH domain

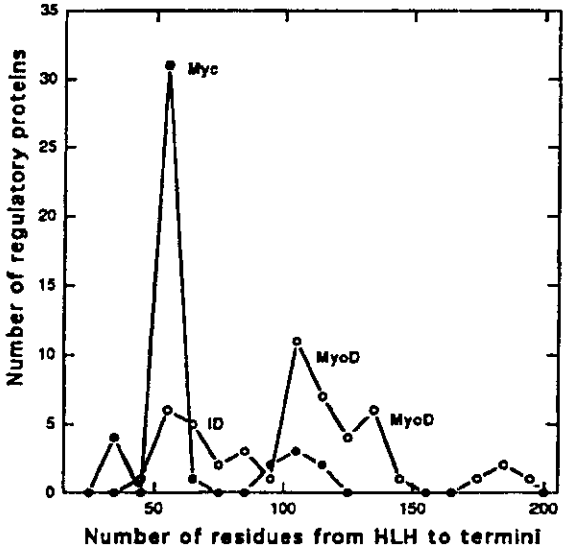


Figure 5. Absolute position of HLH domains in eukaryotic transcription factors. 95 proteins were obtained from Swiss-Prot. The number of amino acids from the center of the helix-loop-helix motif to the terminus was calculated and plotted as described for the other regulatory proteins. Filled circles refer to distances from the C terminus and open circles refer to distances from the N terminus. The same distribution was observed when orthologous proteins were considered as one entry.

(Figure 5). The terminal position includes mostly Myc and ID proteins, whereas the more distal peak includes mostly MyoD proteins. Therefore, a terminal location for DNA-binding domains is probably very common but not necessarily unique. It is also interesting to note that the separate peaks identified by this HLH position analysis correspond to evolutionary groups defined in terms of a tree analysis (Atchley & Fitch, 1997).

The above application of several new types of position analyses leads to results that bear on the evolution and function of the current transcription apparatus. One type of position analysis showed that activators bind at strongly preferred proximal promoter positions in both prokaryotes and eukaryotes. As discussed above, application of concepts derived in the former case suggest possibilities for synergistic activation in the latter, even using a single activator.

A different type of position analysis showed a strikingly non-random distribution of DNA-binding domains in transcription regulators, again in both prokaryotes and eukaryotes. The domain was almost always near a protein terminus, but which terminus depended on the class of protein analyzed. This classification of proteins based on domain position has the potential to complement simple sequence alignments and greatly expand the information output from sequencing programs. As further information and programs become available one should be able to apply a position analysis to identify sub-groups of eukaryotic regulators with regard to function, as occurred in the bacterial analysis where repressors and activators could be distinguished.

In addition, the position analysis contains potential evolutionary information that again complements that obtained from phylogenetic analysis of genomic sequences. In the prokaryotic case the repressors and activators have their HTH domains located near different termini. Each group responds to a great variety of ligands. Even small DNA-binding proteins without ligand response domains, like lambda cro, can act as repressors when they bind promoter sites downstream from −35 (see Gralla & Collado-Vides, 1996). It is possible that such a simple regulator evolved first, as an ancestor to both repressors and activators. Ligand response domains, which are generally larger and uninterruptible as they involve complex protein folds, would be added later. After the joining of the large and small domains during evolution the center of the smaller DNA-binding domain would inevitably be near to an end. The above data suggest that such ancestral protein with a DNA domain C-terminal to the ligand domain could have been adapted to serve as an activator by evolving sequences that could contact polymerase from nearby upstream sites. A different ancestral protein with a DNA domain N-terminal to the ligand domain could have given rise to repressors. The ultimate result of this process could be the modern-day prokaryotic transcription apparatus that is described by the above position analysis. The provocative conservation of DNA-binding domain location in classes of eukaryotic regulators suggests that similar analyses of sequences from eukaryotic genome projects would be a useful adjunct to standard phlyogenetic analysis.

Human genomic sequences are being collected at a rapid rate and the current methods of analysis typically do not include comparisons of the positions of protein domains and DNA elements. As more transcription factors are sequenced from mammalian genomes the grouping of locations of their DNA-binding domains (or indeed any domains) may be analyzed. The development of such position analyses should maximize the information output and hasten our understanding of the circuitry of mammalian transcriptional regulation (Thanos & Maniatis, 1995).

Acknowledgments

We acknowledge useful discussions with Enrique Morett about the AraC regulators.

This research was supported by grants GM35754 and GM49048 from the NIH to J.D.G. and by grants from DGAPA-UNAM and CONACYT to J.C.V. E.P.R. has been supported by a graduate fellowship from CONACYT.

References

Atchley, W. R. & Fitch, W. M. (1997). A natural classification of the basic helix-loop-helix class of transcription factors. *Proc. Natl Acad. Sci. USA*, **94**, 5172–5176.

Bairoch, A. & Apweiler, R. (1996). The Swiss-Prot protein sequence databank and its new supplement TREMBL. *Nucl. Acids Res.* **24**, 21–25.

Bucher, P. (1990). Weight matrix descriptions of four eukaryotic RNA polymerase II promoter elements derived from 502 unrelated promoter sequences. *J. Mol. Biol.* **212**, 563–578.

Busby, S. & Ebright, R. (1994). Promoter structure, promoter recognition, and transcription activation in prokaryotes. *Cell*, **79**, 743–746.

Busby, S., West, D., Lawes, M., Webster, C., Ishihama, A. & Kolb, A. (1994). Transcription activation by the *Escherichia coli* cyclic AMP receptor protein. Receptors bound in tandem at promoters can interact synergistically. *J. Mol. Biol.* **241**, 341–352.

Chi, T., Lieberman, P., Ellwood, K. & Carey, M. (1995). A general mechanism for transcriptional synergy by eukaryotic activators. *Nature*, **377**, 254–257.

Crowley, E. M., Roeder, K. & Bina, M. (1997). A statistical model for locating regulatory regions in genomic DNA. *J. Mol. Biol.* **268**, 8–14.

Dodd, I. B. & Egan, J. B. (1987). Systematic method for the detection of potential cro-like DNA-binding regions in proteins. *J. Mol. Biol.* **194**, 557–564.

Gaston, K., Bell, A. I., Kolb, A., Buc, H. & Busby, S. J. (1990). Stringent spacing requirements for transcription activation by CRP. *Cell*, **62**, 733–743.

Gralla, J. D. & Collado-Vides, J. (1996). Organization and function of transcriptional regulatory elements. In Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology* (Neidhart, F. C., ed.), 2nd edit., pp. 1232–1246, ASM Press, Washington, DC.

Haykinson, M. J. & Johnson, R. C. (1993). DNA looping and the helical repeat *in vitro* and *in vivo*: effect of HU protein and the enhancer location on Hin invertasome assembly. *EMBO J.* **12**, 2503–2512.

Hoey, T., Weinzierl, R. O., Gill, G., Chen, J. L., Dynlacht, B. D. & Tjian, R. (1993). Molecular cloning and functional analysis of *Drosophila* TAF110 reveal properties expected of coactivators. *Cell*, **72**, 247–260.

Joung, J. K., Koepp, D. M. & Hochschild, A. (1994). Synergistic activation of transcription by bacteriophage I cI protein and *E. coli* cAMP receptor protein. *Science*, **265**, 1863–1866.

Nguyen, C. C. & Saier, M. H., Jr (1995). Phylogenetic, structural and functional analyses of the LacI-GalR family of bacterial transcriptional factors. *FEBS Letters*, **377**, 98–102.

Ninfa, A. J. (1996). Regulation of gene transcription by extracellular stimuli. In Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology* (Neidhart, F. C., ed.), 2nd edit., pp. 1246–1262, ASM Press, Washington, DC.

Niu, W., Kim, Y., Tau, G., Heyduk, T. & Ebright, R. H. (1996). Transcription activation at class II CAP-dependent promoters: two interactions between CAP and RNA polymerase. *Cell*, **87**, 1123–1134.

Pao, G. M. & Saier, M. H., Jr (1995). Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. *J. Mol. Evol.* **40**, 136–154.

Ptashne, M. & Gann, A. A. (1997). Transcriptional activation by recruitment. *Nature*, **386**, 569–577.

Roberts, S. G., Choy, B., Walker, S. S., Lin, Y. S. & Green, M. R. (1995). A role for activator-mediated TFIIB recruitment in diverse aspects of transcriptional regulation. *Curr. Biol.* **5**, 508–516.

Sauer, F., Hansen, S. K. & Tjian, R. (1995). Multiple TAFIIs directing synergistic activation of transcription. *Science*, **270**, 1783–1788.

Schleif, R. (1996). Two positively regulated systems, *ara* and *mal*. In Escherichia coli *and* Salmonella typhimurium: *Cellular and Molecular Biology* (Neidhart, F. C., ed.), 2nd edit., pp. 1300–1309, ASM Press, Washington, DC.

Thanos, D. & Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, **83**, 1091–1100.

Wingender, E., Dietze, P., Karas, H. & Knuppel, R. (1996). TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucl. Acids Res.* **24**, 238–241.

Edited by M. Yaniv

(Received 15 July 1997; received in revised form 24 September 1997; accepted 24 September 1997)

HIPOTESIS

A) Esperamos encontrar alrededor de 400 reguladores transcripcionales en el genoma completo de *E. coli* K12. Este número de genes estaría cercano al máximo de reguladores que *E. coli* necesitaría para regular a todos sus genes.

B) Podemos pensar que los reguladores se agrupan en al menos tres supergrupos de acuerdo a la posición del HTH: El primer grupo conformado por proteínas con el HTH en el N-terminal, el segundo grupo con proteínas con el HTH en el C-terminal. La familia LysR podría formar un tercer supergrupo, tanto por la posición del HTH como por la conservación de sus funciones regulatorias. De acuerdo a esta distribución, cada uno de estos grupos podría presentar una historia evolutiva diferente.

OBJETIVOS

De acuerdo a los datos descritos anteriormente, hemos decidido:

A) Detectar y/o predecir todos los reguladores transcripcionales de *E. coli.*, mediante la utilización de diferentes enfoques: análisis de secuencias, búsqueda de patrones conservados y búsqueda de homología distante.

B) Describir en términos funcionales a todos los reguladores de *E. coli.*

C) Utilizar la información posicional para asignar la función a los reguladores predichos.

D) Tratar de explicar la relación de la posición del HTH y la actividad regulatoria, con base en los reguladores de *E. coli.*

RESULTADOS

Los resultados obtenidos los hemos dividido en dos partes para mayor facilidad del lector y las presentamos a continuación.

PRIMERA PARTE

Presentamos una descripción funcional de *todos* los reguladores transcripcionales de *E.coli*. Este conjunto agrupa tanto a proteínas reguladoras conocidas como predichas. Para las predicciones se utilizaron diversos métodos computacionales: Comparación de secuencias, búsqueda de patrones comunes tanto en el genoma completo de *E. coli* versión m53, como en la base de datos SwissProt, y finalmente, una revisión bibliográfica.

Artículo: "Pérez-Rueda, E. and Collado-Vides, J. The Repertoire of DNA-Binding Transcriptional Regulators in Escherichia coli.

The Repertoire of DNA-Binding Transcriptional Regulators in *Escherichia coli*

Ernesto Pérez-Rueda and J. Collado-Vides
ernesto@cifn.unam.mx, collado@cifn.unam.mx
Programa de Biología Molecular Computacional. Centro de Investigación sobre Fijación de Nitrógeno.
Universidad Nacional Autónoma de México. Cuernavaca, Morelos. A. P. 565-A. 62110.

Abstract

Most of the transcriptional regulatory proteins have a modular architecture with distinct domains for various functions such as DNA-binding, dimerization, and ligand binding. The most thoroughly studied DNA-binding motif contains a helix-turn-helix (HTH) structure formed by two α-helices linked by a turn. Using a combination of several approaches we estimated and characterized the total set of regulatory DNA-binding proteins in *E. coli*. This could be the minimal set of transcriptional regulators in *E. coli*. Activation, repression, and dual regulation are widely diversified in this bacterium. Protein families are homogeneous in terms of regulatory roles, regulated physiological function, length size, and genome position, showing that these families have evolved homogeneously both in prokaryotes and in *E. coli*. This is a first step towards a full characterization of the repertoire of regulatory interactions in a free living cell, and it shall play a central role in the interpretation of global expression profiles of *E. coli* in different experimental conditions.

Introduction

As more bacterial genomes continue to be sequenced, the central role of *Escherichia coli* as a privileged bacterium that synthesizes a large amount of the experimental legacy of previous decades in molecular biology will become more evident (Neidhardt, et al. 1987, 1996). Our laboratory has been engaged in gathering information from the literature on transcriptional regulation and operon organization, and organizing it into a database, RegulonDB (Salgado, H. et al. 1999). This database has enabled us to initiate global studies of transcriptional regulation, such as a comprehensive prediction of promoters, regulatory sites, and operon organization in the complete genome of *E. coli* (Blattner, F. et al., 1997). It also enabled us to identify preliminary observations on the architecture and connectivity of this characterized fraction of the regulatory network in *E. coli* (Thieffry, D. et al. 1998). We have shown that in addition to sequence analysis, other sources of information are relevant to understand the regulation of gene expression, such as the relative position of motifs both in the upstream DNA regions of transcriptional regulation (Collado-Vides, J. 1991. Gralla J. and J. Collado-Vides, 1996), as well as the relative position of the HTH motif within the protein sequence (Pérez-Rueda, E. et al.1998).

Methods of sequence and positional analysis of motifs can now be applied to the full genome sequence of *E. coli*, in order to characterize and extensively define the set of regulatory DNA-binding proteins in the cell. Combining two informative sources we provide here an estimate of the complete set of transcriptional regulators in this bacterium. First, we have exhaustively collected information from the literature. Second, we used methods to search for similar motifs in the complete set of *E. coli* proteins, and completed the putative set. Finally, we present a set of regulatory proteins, and analyze them in terms of their structural and functional properties.

The collection we present here will contribute to the organized knowledge available in *E. coli*, and shall represent a relevant reference collection to the analysis of the global expression profiles in *E. coli*.

Methods

Detection and prediction of transcriptional factors in *E. coli*

Prediction of protein function using computational tools are becoming more and more important as the gap between the increasing amount of sequences and the experimental characterization of the respective proteins widens (Bork, P., and E. Koonin, 1998). One problem in current predictive methods is that they fail to detect all the members with one particular function. For this reason, in this work, different methods were used to search for transcriptional factors, as discussed in the following. The general strategy to detect all transcriptional factors, schematized in Figure 1, implied several methods, from regular expressions to profile algorithms. SwissProt database version 34.0 plus weekly updates (Bairoch, A. and R. Apweiler. 1996), and references were consulted. We used the keywords "Escherichia and coli", "transcription", and "regulation". Subsequently, we added the keywords "DNA-binding and helix-turn-helix" to exclude proteins that do not bind to DNA. At this stage, we removed proteins involved in DNA-binding like transposons, and sensor proteins that belong to the two component systems (i. e. BarA sensory protein). In this first search, 232 proteins were considered as potential transcriptional factors.

It is known that most of the prokaryotic regulatory proteins often recognize DNA-operator sequences using the helix-turn-helix (HTH) motif (Harrison, S. C. 1991). The program called HTH 1.05 that predicts whether a protein contains a helix-turn-helix was implemented to scan all proteins of the *E. coli* genome (Dodd, I. B. and J. B. Egan. 1987). This method, in its original version, is based in a master set of 37 aligned sequences from where a frequency matrix is built. This matrix is used as a weighting factor to determine in a candidate sequence whether the amino acid that actually occurs favours an HTH structure. It evaluates how well the amino acids correspond to those typically found at similar positions in the canonical HTH sequences. Favored amino acids have a positive weight, disfavored amino acids a negative weight (Yudkin, M. D. 1987). At certain positions in the DNA-binding region, the variability of residues is crucial to the function of the protein. Amino acids in specific positions in the structure are severely penalized by the method. For instance, at position 9 there is always a glycine implied in the turn conformation, and the positions 5 and 15 usually contain small hydrophobic residues, because they form van der Waals contacts and maintain the correct inter-helix angle. One disadvantage in this search for is that the method is unable to find an HTH motif in the TrpR repressor, because the protein has an isoleucine at position 12 (heavily penalized by the method). Moreover, the algorithm detects one putative HTH in the ß-galactosidase enzyme (Yudkin, M. D. 1987). A total of 4283 Open Reading Frames (ORFs) were scanned detecting 283 motifs in 276 proteins, 7 of them with two potential HTH regions. See Table 1. All the predicted HTH motifs are located either in the N-terminal or in the C-terminal, in similar positions as previously shown in the distribution of known similar DNA-binding motifs in eukaryotic and prokaryotic transcriptional factors (Pérez-Rueda, E., et al. 1998).

Detection of the HTH motifs has been reported using pattern searches in unaligned protein sequences. The Gibbs sampler method (Lawrence, C., et al. 1993) can detect shared motifs in either protein or nucleic unaligned sequences, with or without gaps. The program is a heuristic method, and not exhaustive, however it converges towards an optimal matrix. Most sub-optimum alignments found by the Gibbs sampler are often closely related to the optimum alignment. In order to calibrate the Gibbs sampler we first made several essays with a set of fifteen known 3D structures from *E. coli* and phage DNA-binding proteins. (See table 9). The best pattern found in the training set was running 1000 independent searches 100 times, with no gaps allowed. The best pattern detected corresponds to the HTH and reflects the array of it secondary structure obtained from the Dictionary of Secondary Structure Protein database or DSSP (Kabsch, W, and C. Sander. 1983). Under these conditions, we searched in the complete protein set of *E. coli* (derived from the search in SwissProt) and detected a helix-turn-helix DNA-binding motif in 146 proteins. Matrices derived from Gibbs were used to search for more proteins in the *E. coli* genome. A cross-reference between SwissProt annotations, the

Dodd and Egan method, and the Gibbs sampler method, shows that around 90% DNA-binding motifs are described by at least three of these methods. Figure 2.

Many of the known protein families have been collected in the PROSITE database (Bairoch, A.1997), where diagnostic patterns are described. PROSITE can be used to get an idea of the function of uncharacterized proteins translated from genomic or cDNA sequences. For some families, the pattern given is not the best to make a functional diagnostic; in fact, we observed that it may fail to match some of the sequences within the family, and may also match some sequences outside of the family (Brazma, A. et al 1998). For example, in the IclR family the most conserved region is the C-terminal, and not the DNA-binding domain (located in the N-terminus). For this reason, we have to be careful in determinate whether a new candidate is a transcriptional factor. Prosite patterns for 17 families were used to detect putative transcriptional factors in the *E. coli* genome, as summarized in Table 2. In this way we described around 171 proteins as putative transcriptional factors.

Finally, a sequence comparison was performed in order to detect more regulatory proteins. Identity values below 25% in fragments of at least 100 amino acid residues were not considered significant. This empirical calibration is widely used, suggesting that sequences with scores higher than this threshold share structural and functional similarity (Holm, L. 1998). In the case of the AraC/XylS family, a significant homology was found extending over a 100-residue stretch constituting the DNA binding domain (Gallegos, M-T. et al, 1997), whereas the GalR/LacI family (Weickert, M. J. and S. Adhya, 1992) shows the longest conservation involving almost all the protein sequence in the definition of the family

In addition, during all this work, we were monitoring the literature and adding new experimentally described regulatory proteins to the dataset. Furthermore, we compared our dataset with the annotations of open reading frames (ORFs) generated by Fred Blattner's group (Blattner, F. et al 1997), and in this way the set of transcriptional factors was finally defined.

A Blast search (using the default parameters, and Blosum 62 matrix) was found not to be sensitive enough to detect all the known regulatory proteins (Altschul, S. F. et al. 1997). The sequential order of the complete search is the following:

1. Regulatory proteins were initially searched in SwissProt. A set of 232 were detected, out of which we excluded 55 since they did not show an HTH motif. Some excluded members were part of the two-component class of regulators that do not bind to the DNA, and some were transposons, among others. This cleaning left a set of 177 out of which only 7 were predictions. This search gave us also literature information that we used later to find additional members, and also to complete different properties we gathered about all the regulators.
2. The weight matrix designed for HTH recognition by Dodd and Egan (1987) was then used. We obtained 276 proteins after searching in the genome, out of which, we manually excluded 85 for several reasons (transposons, insertion sequences, and the ß-galactosidase), keeping finally 191 proteins. Of these, 162 are detected by another method, and are 29 new predicted regulatory proteins. In parallel, we ran the Gibbs search using a weight matrix built with the known HTH proteins. A total of 133 proteins were detected, of which 124 are known, generating only 9 new candidates.
3. From the set of 171 described in PROSITE, we eliminated 13 proteins that are architecturally similar to regulators but lack a DNA-binding domain (DBD). This leaves 158 proteins, out of which 156 are either part of the known set, or are recognized by a previous method, leaving only 2 new candidates.
4. We took the 252 regulators annotated in the genome by Blattner (Blattner, F. et al. 1997), and excluded 57 proteins that either lack a DBD, or are enzymes, or the literature indicates that they are not DNA-binding regulatory proteins, such as for instance, the D-xylose-binding periplasmic protein XylF that is known not to be a transcriptional regulator.
5. We re-analyzed the initial blast searches and added few more proteins with sequence identity higher than 25%. This enabled us to add some few proteins with a different type of DNA-binding domain to the HTH,

such as the Cold shock domain (CSD), ß-strand, Helix-loop-helix (HLH), or Zinc-finger domains. It also helped to eliminate proteins with a similar architecture as regulators (i.e. MgbL, XylF, and RbsB that are sugar transporters).

6. During all this process we kept searching the literature and completed a total of 163 transcriptional regulators -supported by either 3D structure, mutation in the regulatory gene or in the DNA-binding site. A Venn diagram in Figure 2 summarizes all these results.

In brief, the criteria to accept a regulatory protein were:

1. Presence of a known DNA-binding domain, HTH preferably.
2. Proteins that are part of the known dataset and that are recognized by any of the methods used.
3. Proteins detected by any two methods out of: i) PROSITE, ii) Blattner annotations, iii) Dodd and Egan weight matrix, or iv) Gibbs-sampler.
4. Proteins detected by only one method, and that by sequence comparison are similar to known regulatory proteins.
5. Proteins with a 25% or higher homology to a known regulator, preferably around the HTH region.

The final set after searching, filtering and selecting by these automated methods and by manual inspection, groups 314 proteins, of which 163 have an experimental support and 151 are predicted regulatory proteins. See Table 3. Table 4 shows a more detailed comparison of the predicted 151 proteins and indicate the methods associated to each prediction.

In order to get an estimate of the sensitivity and accuracy of the predictions, we evaluated the output of the predictive method with the known dataset of 163 experimentally supported proteins. Based on that dataset we estimated the fraction of true and false positives of the different methods used. See Table 5. From this table we can see that overall the dataset of predictions is based on 65 to 67% true positives, whereas we must be adding between 9 to 15% false positives to the set.

A fold prediction was performed for all transcriptional DNA-binding proteins, and around of 83% of all characterized proteins and 77% of the predicted regulators have a DBD fold. This analysis based on structural comparisons (Jones, D.T. 1999) provides additional support to the dataset of predicted regulators and therefore to the complete repertoire here presented.

All DNA-binding transcriptional factors were grouped in families based on information available in the literature, as well as based on sequence comparisons (Clustalw, using parameters the default. Thompson, J. D. et al. 1994). Several families have already been proposed previously, such as the Enhancer Binding Protein (EBP) family, the LuxR/UhpA family, and the OmpR family (Pao, G. M. and M. H. Saier, 1995) (many of them are involved in two component systems); as well as the GalR/LacI (Weickert, M. J. and S. Adhya, 1992), the LysR (Henikoff, S. et al. 1988. Schell, M. A. 1993), AraC/XylS (Gallegos, M-T. et al, 1997, the ArsR (Bairoch, A. 1993); and the CRP (Spiro, S. 1994) families. The criterion to define a family is based on sequence comparison. If a protein shares at least 30% of identity in its complete sequence, or within the DNA-binding domain (DBD) with any family member of a family, then it is considered part of that family. Sometimes, several domains are detected in a protein. For this reason, the DBD and its similarity to the family plays a defining role to discriminate among different families of transcriptional factors.

Results

We will first present an overview of the dataset, emphasizing the structural and functional properties of all transcriptional factors, such as the DNA-binding motif and regulatory roles. Then we describe the distribution of the regulators into their evolutionary families. Furthermore, we analyze the position of the regulators in the genome, as well as some correlations with functional properties of these families.

The Repertoire of DNA-Binding Transcriptional Regulators

To fully understand the genetic regulatory mechanisms it is necessary to study their properties in the context of the cellular processes they regulate: cellular division, differentiation, responses to several

environmental changes, etc. The purpose of this work is to analyze the organization of 314 transcriptional factors of the *E. coli* genome that influences the gene expression in terms of both structural and physiological properties. It should be clear that our predictions are preliminary in the sense that they await further experimental testing.

An important question to answer is if our dataset is the total of regulatory proteins in *E. coli*. Based on the estimated percentage of false positives of the methods used, the minimum regulatory predicted proteins would be on the order of 100 proteins (65% of 151), adding to the 163 already characterized. This eliminates around 20% of the estimated 314 proteins. On the other hand, we know that several methods fail to detect true positives. In the case of the matrix-based methods, 30% can be missed (based on comparing with pattern search or sequence comparisons). This gives the upper limit of around 350 regulatory proteins, adding 8% to the set. See Table 5.

Previous rough estimates pointed to around 400 regulatory genes in the all *E. coli* genome (Thieffry, D. et al., 1998). They assume a 1:10 ratio of regulatory to regulated genes, and a 10% of constitutive genes, there would be around 400 regulators, since *E. coli* contains around 4400 genes. Another source of information is the collection of operons and regulatory interactions described in RegulonDB (Salgado et al., 1999). Based on previous versions of such dataset, the same estimate of around 400 regulators was made. Note that it is hard to know or guess what percentage of genes in *E. coli* are constitutive. In an updated analysis of this approximation, we can re-calculate an estimated minimal number of transcriptional regulators in *E. coli* as follows. We know that 933 genes are grouped in 361 operons, and that 78 regulatory proteins are associated to this dataset. This gives a 1:12 ratio of regulatory to regulated genes. If we assume that this set corresponds 25% of all regulatory genes, since they regulate a fourth of the total set of genes in *E. coli*, then we get surprisingly the number of 78 times 4, or 312 total transcriptional DNA-binding regulators in *E. coli*, which corresponds very much to our estimate here of 314. Of course the precision of these numbers may be mere coincidence. Overall, based on this information and the analysis performed, we consider that the universe of DNA-binding regulatory proteins in *E.coli* contains on the order of 300 to 350 proteins.

Table 6 describes the collection of individual regulatory proteins in the chromosome of *E. coli*. 163 transcriptional factors have experimental evidence (mutation, footprints, structure analysis, etc,), while 151 are putative factor proteins. The table presents all regulators organized in families, their protein name, the HTH relative position in the protein sequence, their regulatory function known or predicted based on family membership, the physiological function, and the genes regulated by each regulator. See legend of Table 4 for additional explanations.

The set of characterized regulatory proteins has a diversity of functions. Some proteins regulate the bacterial housekeeping $\sigma 70$ promoters, the $\sigma 54$ promoters (some EBP proteins), or both promoters (NtrC, a dual protein). Some regulators are affecting a particular pathway, like the L-cysteine biosynthesis (CysB regulator). Many of the regulatory proteins control operons with one or more promoters (Gralla, J. and J. Collado-Vides, 1997); others are involved in catabolic regulons (Crp regulator), or have structural and regulatory roles (i.e., ArgR and Fis). In few cases regulatory proteins affect the expression of other regulators, but there is no single closed loop of interactions involving several regulatory proteins, at least within the known dataset of regulatory interactions (Thieffry, D. et al 1998).

This diversity of regulatory role and physiology of the regulated genes is better summarized when described at the level of each family. We have observed that families of regulatory proteins tend to participate in certain physiological processes. Table 10 shows the most frequent physiological participation *per* family.

As mentioned before, regulatory proteins are grouped into evolutionary families based on their sequence similarity. The total set of *E. coli* K-12 chromosomal regulators fall within 25 protein families, with 17 families containing 50% of all members of the collection. The helix-turn-helix DNA-binding motif is detected in 234 known and predicted transcriptional regulators. The rest of the predictions are based on homology to known transcriptional regulators. Also,

motifs such as zinc-fingers, ß-sheet antiparallel, RNA-binding like, and helix-loop-helix, have been described in transcriptional factor proteins in *E. coli*, although the fraction of regulators described with these motifs is much smaller than the HTH proteins.

Regulatory Activity and HTH Relative Position

We have previously observed an interesting correlation between the relative position of the HTH motif in the protein sequence, and its role as a negative or positive activator. Repressor proteins have their HTH usually in the N-terminal of the protein, whereas activator proteins tend to have their HTH close to the C-terminal end of the protein. Furthermore, this position is conserved within the different evolutionary families of regulatory proteins. A preferred position was also observed within the HLH family of eukaryotic transcription factors (Pérez-Rueda, E. et al.1998). We wanted to re-estimate whether this behavior is also followed by the predicted set of 151 regulators. Figure 3.a shows that a similar distribution of the HTH motif is observed with the 125 known and with the 123 predicted regulators, with in fact, a smaller peak in the middle positions for the predicted set. Figure 3.b shows the distribution of all predicted regulators with a putative HTH motif, and the separated distributions of the known activators, repressor and dual proteins. It is clear that activators (in red) have a strong preference for their HTH to be located in the C-terminal of the protein (78% of activators), with only 22% of them in the N-terminal, whereas repressor proteins have almost all of them their HTH in the N-terminal (96%) with only 4% of them in the C-terminal. There are very few proteins with their HTH in central position, mostly from the ArsR family. The regulatory role in proteins using the HTH position as an indicator of function predicts around 70% of cases correctly, with 15% of false positives. In addition family membership criteria is used in conflicting cases, such as in the EBP or IclR families.

Dual proteins, which exert an activator effect in some promoters and a negative effect in other promoters, show mostly a similar behavior to the repressor proteins. This structural observation supports the unexpected suggestion that dual proteins might have initially been repressor proteins that acquired the function to activate. As discussed before, given the mechanistic requirements to repress, we conceived most likely the opposite situation where activators become repressors by a simple displacement of their DNA-binding site in relation to promoter initiation (Gralla, J. D. and J. Collado-Vides, 1996).

This distribution of positions was used to predict if a putative transcriptional regulator is expected to be a repressor or an activator. If a protein has the HTH in N-terminal, it will be predicted as a repressor protein, unless it belongs to the LysR family. As discussed in earlier work, the peak of dual proteins observed at the N-terminal (blue in Figure 3) is mostly contributed by members of the LysR family (Pérez-Rueda, E. et al.1998). A protein with the HTH in the C-terminal is assumed to be an activator. Using this simple rule, we assigned the putative function to several HTH proteins (Tables 6 and 7). We considered all proteins that have their HTH in the first 40% of the protein (relative position) as N-terminal, and C-terminal when its HTH is present in the last 40% of the protein.

Table 7 groups all proteins by their regulatory roles. We defined regulatory function based on the experimental evidence and the prediction rule explained above. We considered dual proteins those whose sequence is similar to any member of the LysR family. Dual proteins are either activators of several genes and repressors on their own expression (proteins of the LysR family) (Schell, M. A.1993), or activators and repressors of different sets of genes, such as the case of Crp and FruR (Kolb, A. et al. 1993). Around 67 proteins do not have any function described, because we do not have sufficient information to assign them a regulatory role. Remember that 80 regulators were added to the collection even if they do not have an identified HTH motif.

Grouping known and predicted proteins gives around 85 activators, 104 repressor proteins, and 59 dual proteins, corresponding to 34.3%, 42% and 24%, respectively. A previous evaluation with a much smaller database gave some years ago 10%, 55% and

38% respectively (Gralla, J. D. and J. Collado-Vides, 1996). The current numbers show a much even distribution of repressors, activators and dual proteins. It is quite surprising that the contribution of one single family, LysR, can account for almost a quarter of all dual regulatory proteins in *E. coli*. This may be a selected family that has been used in evolution with a wide number of members in *E. coli* to control the regulation of amino acid biosynthesis where, depending on the conditions of growth, the same pathways have to be either activated or repressed (Newman, E. B. et al. 1996). This dual property is not limited to this family. Proteins from different families also add to this dual set of regulators, such as CRP and FNR, two important global regulators.

Negative autoregulation is predominant in transcriptional factors of *E. coli* (Thieffry, D. et al. 1998). The updated numbers in RegulonDB show that only 6 regulators are positively autoregulated - PhoB, GutM, TdcA, CadC, RhaS, and RhaR- while 3 regulators are exerting both positive and negative regulation on their own expression, Ada, Crp, and NtrC (Magasanik, B. and F. C. Neidhardt. 1987). The LysR family that provides almost 25% of proteins with a known function, accounts also for 25% of the negatively autoregulated proteins (10 out of 40). LysR is a unique family with negative autoregulation present in most of its members.

DNA-binding Motifs Described in *E. coli*

Structural evidence have shown the heterogeneity of the DNA-binding motifs in transcriptional factors both in prokaryotes and eukaryotes. The three-dimensional structure of around 17 different DNA-binding structures of transcriptional factors in *E. coli* have been determined by X-ray crystallography, nuclear magnetic resonance (NMR) and spectroscopy methods. See table 8.

In *E. coli*, the HTH is the most diversified structure. Two types of them have been described, the classical (cHTH) and the winged (wHTH) helix-turn-helix motifs. The cHTH in its simplest form consists of two alpha helices separated by a ß-turn. The whole structure contains 20 amino acid residues, some of which interact specifically with the DNA (Harrison, S. C. 1991). The first helix (residues 1-8) can lie across the major groove of DNA, locking the second helix (residues 12-20) into position in the major groove, where the side chains of crucial amino acid residues interact specifically with certain bases in the DNA (Pabo, C. O. and R. T. Sauer, 1984). At position 9 there is almost always a glycine, which easily adopts the configuration required at the turn, connecting the two helices (See Branden, C. and J. Tooze, 1991). This DNA-binding motif is conserved due to the structural constraints imposed by the fold. Such residues can be used to define a fingerprint for the motif and then applied to other protein sequences to predict whether a similar fold is found. The winged helix is a DNA-binding motif composed of an α/ß structure. This structure contains 3 N-terminal α-helices and a 3-stranded antiparallel ß -sheet. The folding of the ß -sheet region about the α-helices gives the appearance of wings on the helices, hence the term "winged-helix". This motif was first identified in the transcription factor HNF-3g (HNF-3g is a member of a large family of transcription factors related to the *Drosophila* gene fork head, hence the gene family is termed the fork head (FKH) family) (Brennan, R. G. 1993). It was later described in OmpR, Crp, and other transcription factors.

The HTH DNA-binding motif has been identified in many prokaryotic DNA-binding regulatory proteins, and it may constitute a structure almost exclusive of this class of proteins. However, this motif, also has been described in three proteins that are not transcriptional regulators: the aspartyl tRNA-synthetase (AspRS) (the HTH is a domain implicated in the stabilization of the complex with tRNA) (Delarue, M. et al, 1994); the C-terminal domain of L7/L12 protein that may constitute a HTH binding structure for interaction with nucleic acids, most probably RNA (Rice, P. A. and T. A. Steitz, 1989); and the A/G mismatch-specific glycosylase MutY that may form a HTH structure (Tsai-Wu, J. J. et al . 1991).

ß -sheet antiparallel

This motif has been characterized in a DNA-binding role only in two proteins in *E. coli*, MetJ and HU. The repressor of methionine biosynthesis, MetJ, forms a

dimer with a core composed of four α-helices, two helices from each subunit. An antiparallel ß ribbon, formed in the N-terminal segment contributed by each monomer protrudes from the core (Harrison, S. C. 1991). This protein binds to the major groove of DNA (Suzuki, M. 1995). HU and its relatives are believed to contact the minor groove of DNA. HU has a core composed of two α-helices from each monomer, and two projecting ß-ribbons, each formed by the C-terminal third in the polypeptide chains. The tips of these ribbons are disordered, so that presumably they can wrap around the DNA helix (Harrison, S. C. 1991).

Helix-loop-helix

This motif has only been suggested in one transcriptional regulator in *E. coli*. DnaA is a protein that promotes the initiation of replication of the bacterial chromosome, and of several plasmids. It contains four functional domains. The first domain included the P-loop or Walker A motif (involved in ATP binding). The second domain maps to a region near the C-terminal and is involved in DNA binding. The third domain maps near the N-terminal and might be involved in the ability of DnaA to oligomerize. Finally, the fourth domain is essential for replication from *oriC*. (Sutton, M. D. and J. M. Kaguni, 1997). The DNA-binding domain contains two amphipathic α-helices, and a third α-helix. Secondary structure predictions suggest the presence of three potential α–helices; two of them are potential amphipathic helices (helix A and B) which might form a HLH DNA binding motif (Roth, A. and W. Messer, 1995).

Zinc finger

HypF is the regulator of the micro aerobically inducible region of the hydrogenase operon (Olson, J. W. and R. J. Maier, 1997). The *E. coli* gene products HypA, HypB, HypF, and HypD organized in an operon, contain the CX2C motifs characteristic of metal-binding proteins. In addition, HypB has a long histidine-rich stretch of amino acids near the N-terminal, suggesting a possible role of this protein in nickel binding. The gene product HypF, which is translationally coupled to HypB, exhibits two cysteine motifs (CX2CX18CX2C) with a capacity to form zinc finger-like structures in the N-terminal of the protein. A role in nickel metabolism in relation to hydrogenase synthesis is postulated for proteins HypB and HypF (Rey, L. et al 1993). HypF proteins have also been described in *Bradyrizobium japonicum*, Rhodobacter *capsulatus* B10 (Colbeau, A. et al. 1993), and *Rhizobium leguminosarum biovar viciae*. They contain two zinc-finger motifs (CX2CX18CX2C) characteristic of other HypF proteins (Rey, L. et al. 1993).

Cold shock (CSD)

The CSD is a nucleic acid-binding protein domain found both in bacteria and eukaryotes, that has been shown to bind double-stranded DNA (dsDNA), single-stranded DNA (ssDNA) and/or RNA. Members of the cold-shock family contain a conserved RBD (RNA binding domain) on similar single-stranded nucleic acid-binding surfaces. Although CSD and RNP motifs show little similarity in topology or amino acid sequence, they are an example of convergent evolution (Graumman, P. and M. A. Marahiel, 1996). Both X-ray crystallography and nuclear magnetic resonance (NMR) spectroscopy have been used to determine the three-dimensional structure of CspA. It consists of five anti-parallel ß-strands, ß1 to ß5, forming a ß-barrel structure with two ß-sheets. One ß-sheet consisting of ß1-ß2-ß3, contains seven out of eight of the aromatic residues (W11, F12, F18, F31, and F42), all of which are highly solvent exposed, with in addition, two lysine residues (K10, and K16) in the surface. On this surface, there are two RNA-binding motifs, RN-1 (KGFGFI) on ß2 and RNP2 (VFVHF) on ß3 (Yamanaka, K. et al. 1998).

Out of the around 95% of the regulatory proteins with a known and predicted motif have an HTH motif. Only MetJ, DnaA, HypF, and the cold-shock family have a different DNA-binding motif.

Organization of Regulators in Families

The group of 314 regulatory proteins (known and putative) can be clustered into subsets based on their sequence similarity. This generates 20 clusters, which correspond to evolutionary related proteins, with a

similarity of at least 25% within members of the same family. In addition, these groups match with the 20 families that have been described when studying regulatory proteins across all prokaryotes (Pao, G. M. and M. H. Saier, 1995. Weickert,, M. J. and S. Adhya, 1992. Henikoff, S. et al. 1988. Schell, M. A. 1993. Gallegos, M-T., et al, 1997. Bairoch, A. 1993. Spiro, S. 1994). In other words, we found that all evolutionary families of putative transcriptional regulators have representatives in *E. coli*.

These groups vary considerably in their number of members, as summarized in Table 9 and figure 4. The average number of members *per* family is on the order of 10, ranging from LysR, the largest family in *E. coli* with 45 members, to families with one or two members, such as the CRP, and Fur families. Families with very few members have homologue members in different bacteria, giving evidence that they form separate identifiable groups.

The largest families are those that affect genes involved in amino acid biosynthesis or carbon uptake. The large *csp* gene family in *E. coli* evolved by a process of repeated duplication and diversification of genes, and as an adaptive response to different environmental stresses (Yamanaka, K. et al. 1998).

All protein families described in prokaryotes have been detected in *E. coli*, see Table 9. It certainly makes sense for *E. coli* to have a large diversity of regulatory proteins given its ability to grow and respond to different environmental conditions, and its capacity to live in the human gut as well as a free-living organism. *E. coli* must survive under much more diverse environments including wide ranges of temperature (8 °C to 50 °C), osmolarity, and pH, inside the human intestine and as a free-living cell (Death, A. and T. Ferenci, 1994).

Table 9 shows the percentage of *E. coli* proteins *per* family in relation to all known bacterial members of that family. *E. coli* provides 30% of all prokaryotic proteins within the GalR/LacI family, which may also reflect the fact that many other bacteria do not need sugar metabolism because all carbon compounds are imported from the host (Tomb et. al., 1997). Note however, that this fraction is going to diminish as more bacterial genomes continue to be sequenced in the future. For instance, in the genomes of *E. coli* (Blattner, F. et al. 1997), *H. influenzae* (Fleischmann, R. D. et al. 1995.), and *B. Subtilis* (Kunst, F. et al. 1997), *the* DeoR family has been predicted to be formed by 12, 4, and 6 members respectively. At present there at least 30 members of the DeoR family found in species ranging from Gram-positive cocci such as *Lactococcus lactis* (van Rooijen, R. T. J. et al. 1993) to *Streptococcus mutants* (Rosey E. L. and G. C. Stewart, 1992), and Gram-negative bacteria such as *A. tumefaciens* (Kim, H. and S. K. Farrand. 1997), and *P. aeruginosa* (Schweizer, H. P. and C. Po. 1996).

The LysR family is mostly abundant in Proteobacteria (purple bacteria) and α and γ subgroups. Few proteins of the LysR family have been found in the ß subgroup, and none within the δ subgroup, also members of the family have been described in gram-positive bacteria (Schell, M. A. 1993). However, because many prokaryotic genera have not been subjected to extensive genetic characterization, the observed distribution of LysR family may not yet be conclusive. The large genetic distances between prokaryotes with members of this family, and vast differences in G+C content suggests a progenitor LysR that arose early in prokaryotic evolution (Schell, M. A. 1993). On the contrary, members of the AraC/XylS family are widely distributed in diverse prokaryote genera. Most of the members of this family occurs in the gamma subdivision of the Proteobacteria (purple bacteria). A few have been found in low G+C and high G+C gram positive bacteria and in cyanobacteria (Olsen, G. J. et al. 1994). The AraC/XylS and LysR families have few members in archaeabacteria or in eukaryotes. Members of the GntR family have been described in *E. coli*, *B. subtilis*, *P. putida* and *K. aerogenes* (Haydon, D. and J. Guest. 1991). An interesting observation related to a still earlier evolutive scenario is that some regulatory protein families such as GalR/LacI, or GntR, are only present within prokaryotes, while some other families, like the AsnC family, show a wider distribution both in prokaryotes and archeobacterial organisms (*Methanococcus jannaschi* and *Pyroccocus furiosus*) (Kyrpides, N. C. and C. A. Ouzounis, 1995).

Our knowledge on the distribution of regulatory families in different bacteria will change as more

genomes are finished. However, the fraction of regulatory proteins within genomes should not vary that much in the near future. For instance, *H. pylori* is an example of a bacterium living in a quite stable environment, and its annotated genome indicates 13 regulatory proteins, accounting for 0.82% of the total 1590 genes it contains (Tomb, J. F. et al., 1997). It may be that some proteins are not required in microorganisms that are living exclusively in relatively stable environments. Nonetheless, the cell has alternative coccoid and bacillus forms which shall involve regulatory interactions for their differentiation,. Such interactions may involve DNA-binding motifs other than the HTH.

Families of transcriptional regulators share a type of DNA-binding motif, plus inducer binding or oligomerization motifs. These similarities define a signature sequence shared by the family members. In most of the cases the region that characterizes a regulatory family contains the HTH motif. Therefore, the HTH motif is not only the general main signature in bacteria characteristic of DNA-binding transcriptional regulators, but it also provides the internal distinction among different families. Remember that one of the helix motif is common to many proteins, and is considered to provide for a non-specific binding ability, whereas the second helix motif is the one conferring the DNA binding specificity. The domain organization has been experimentally identified in most transcriptional regulatory proteins. For instance, the DBD from GalR/LacI family is around 59 amino acids (Weickert, M. J. and S. Adhya. 1992). The family contains a second domain of around 270 residues implied in multimerization and induction function. A second additional domain is also found in proteins that bind sugars, such as RbsB (ribose transport). The DBD of AraC/XylS is around 60 residues. The LysR family shares a highly conserved N-terminal DBD (consisting of an HTH motif and flanking sequences). For most members of the LysR family, the less conserved C-terminal domain of the proteins has a sensory function (Ninfa, A. J. 1996). In the case of Crp protein, the C-terminal domain contains an HTH motif able to bind to DNA, while the N-terminal domain is a large structure of around 170 residues containing the nucleotide-binding site that shows homology with cAMP-dependent protein kinases (Kolb, A. et al. 1993). The main feature in the set of transcriptional factors is the presence of the HTH DBD. In fact, this is the domain that characterizes all DNA-binding regulatory proteins. In most protein families the domain covalently linked to the DBD is less well conserved and is implied in several responses (for instance, in the AraC protein the N-terminal domain is involved in allosteric regulation by the coinducer and dimerization) (Ninfa, A. J. 1996). The IclR and EBP families are the only families with a signature that involves other domains in addition to the DBD. In IclR the DBD is located in the N-terminal, while the most conserved domain is located in the C-terminal. In the EBP family the ATPase and DBD domains are strongly conserved, as signatures of the family.

Conservation of Protein Sizes Within Families.

Given the high degree of similarity of proteins within a family, it is not suprising that their size in amino acid residues is also rather conserved. Observing the size distribution within families, it is natural to group them in two groups. Families of one group have members with a rather homogeneous size, whereas in another group of families their members show a more heterogeneous size distribution.

The homogeneous families can be defined by having at least 70% of its members with a conserved length size, shown by an overall standard deviation (SD) smaller than 20. We show a calculus of the mean and standard deviation in the most conserved regulatory proteins. See table 10. This analysis shows homogeneous groups with a small SD, such as Cold shock, GalR/LacI, TetR/AcrR, and IclR families. The IciR family has one protein with 315 residues (MhpR) and 7 proteins with a mean length size of 272.4 residues (SD of 10.01). DeoR is a family where except for the two subunits of GatR, the rest of family members have a conserved size (See table 10).

The GntR family has two small proteins (b3694 and DgoR) with 98 and 177 residues, and two much larger ones, b1439 and YjiR with 468 and 470 residues. Otherwise the remaining 16 proteins have a mean

length of 246.0 residues and SD of 1.41, showing a highly homologous group. A sequence comparison between the biggest proteins in the family (b1439 and YijR) shows an identity of 35%. This result might reflect a genetic duplication in the *E. coli* genome.

Similarly, 70% of the LuxR/UhpA members fall within a mean size of 219.16 residues (SD of 13.21). The maltose activator, MalT is included in this group. It represents one of the biggest proteins in the family and in the collection (with 901 amino acids. Pao and Saier, 1995). Many proteins of LuxR/UhpA family belong to receiver-response regulators. MalT is a protein that lacks receiver modules and is not a response regulator, but possesses homology in the DNA-binding domain. In general, the organization of MalT, shows three domains, the DBD is located in the C-terminal and contains the HTH; in the N-terminal there are two domains of around 400 and 200 residues which are not shared with the other members of the family.

The OmpR protein group belongs to a very homogeneous family in size. 93% of the members fall within a mean size of 230.66 (SD of 7.33) residues. Smaller proteins tend to have a higher percentage of identity than bigger proteins. This is a consequence of the definition of families, mostly in terms of the DBD that is the most conserved region. Smaller proteins have only the DBD, while bigger proteins have additional domains probably as a result of events of acquisition of novel domains.

The homogenous distribution observed in *E. coli* families suggests that the members of these families are derived from a common ancestor through pathways in evolution involving individual amino acid substitutions and small insertions/deletions that could lead to incremental changes and sizes (Savageau, M. 1986). This observation is consistent with the homogeneity of functional properties. Certainly, GalR/LacI, Cold Shock, OmpR, and LuxR/UhpA families are very homogeneous in terms of HTH location, regulatory roles, and physiology of the genes regulated. Therefore, this homogeneity in length is probably correlated with conserved functions that are shared in the domains of all regulators within a family.

As mentioned before, heterogeneous families in terms of the distribution in size of their protein members form a second group. For instance, the LysR family shows a bimodal distribution with one subset of 15 proteins with a size between 215 to 299 amino acids (mean of 284.8, and SD=24.66), and a second subset with 29 transcriptional factors whose length falls between 300 to 354 amino acids (mean of 312.9, and SD=11.62). Similarly, the NagR/XylS family also shows two subsets, one with a mean of 304 and SD of 3.36, and a second one with a mean of 403.6, and SD of 4.03. These subsets might reflect two different evolutionary events within members of the same family.

AraC/XylS, an activator protein family that shows heterogeneity in terms of the location of its HTH motifs (Pérez-Rueda, E. et al. 1998) also shows a high variability in the size of its members. The family can be decomposed into three subsets. A subset of two small proteins size 107 and 129, the larger set of 16 proteins within the range of 239 to 292 residues (mean of 267.8, SD of 18.1), and the group of 9 larger proteins with sized from 300 to 400 residues (mean of 332.8, SD of 35.3). The first subset has its HTH in the N-terminal, the second one has the HTH in the last third of the sequence, while the third has the HTH in central position. Similarly, the EBP family shows a high variability with proteins ranging from 98 to 668 residues. The larger subsets are four proteins with a mean of 454 residues (SD 13.21), a subset of 5 proteins with a mean of 538 residues (SD of 30.63), and three larger proteins with around 668 residues, SD of 25.05. This family is very heterogeneous in terms of sequence length, probably because the proteins have a multidomain organization.

In general within a family, proteins of the same subset or similar length show a higher degree of sequence similarity, as would be expected. In some cases the high sequence similarity and length conservation suggests the notion of gene duplication. Such is the case of Cold shock, GntR, IclR and GalS/GalR families, and the two AraC/XylS members of 107 and 129 residues. This degree of conservation could reflect duplication events and homogeneity in regulatory functions.

The families with a conserved size might have all come within one family by gene duplication events, as has already been suggested (Nguyen, C. C. and M. H. Saier. 1995). Size conservation, presence of the HTH motif in a conserved relative position, as well as homogeneity in the motif that defines the family, all these factors contribute to the notion of a homogeneous family of regulatory proteins. On the other hand, there are families with more heterogeneous behavior suggesting a more diverse evolutionary history involving multi-domain shuffling and acquisition of novel domains. Protein families with more variable length such as EBP and the AraC/XylS show multi-domain organization.

Regulatory Families and The Physiology of Genes they Regulate

Regulatory proteins within a family share structural properties as discussed before. They also tend to affect genes and operons involved in related metabolic functions. In Table 10 we describe the most common physiological classes of genes regulated by the different regulatory families. We calculated the percentage of members for each family dedicated to the regulation of one physiological function. Some families regulate genes that can clearly be associated to a particular physiological class. This is clear for 7 families that group a total of 47% of all potential transcriptional regulators. The LysR, GalR/LacI, GntR, DeoR, TetR/AcrR, IclR, OmpR, EBP and Cold Shock families are included in this group. For instance, the GalR/LacI family regulators are involved in regulating different carbon sources (from all regulators associated to carbon uptake, 20% belongs to the GalR/LacI family); while members of the LysR family are devoted to regulation of amino acid biosynthesis (40% of this biosynthetic activity is controlled by proteins from this family). Members of different families, such as the Cold Shock family, the ArsR (arsenical resistance), and the TetR/AcrR (antibiotic resistance) families control resistance responses. *E. coli* can grow in many different carbon sources (galactose, melibiose, lactose, rhamnose, etc.). At least 50 transcriptional factors are devoted to degradation of carbon compounds, which belong to 5 regulatory families. Some of these involve two regulators involved in the regulation of carbon uptake (GalR/S or RhaR/S), See Table 10.

Table 10 shows the most common physiological class that is subject to regulation by each regulatory family, in decreasing order of dedicated physiology. Thus, the CRP, GalR/LacI and Cold Shock proteins devote either all or 90% of their regulatory proteins to one single class of cellular activity.

One regulatory family is not always involved in regulating the same metabolic response as can be observed in Table 10. This is the case, for instance, of the growth-phase-dependent expression of CspD protein, a member of the CspA family. All other regulatory proteins of this family have been suggested to be inducers under cold shock (Yamanaka, K. et al 1998). The *cspD* expression is induced by stationary-phase growth, and it does not dependent on the stationary-phase sigma factor σS. Moreover, the expression of *cspD* is inversely dependent on growth rates and induced upon glucose starvation. The (p)ppGpp is one of the positive factors for the regulation of *cspD* expression. CspD is then the only protein of its family involved in a different cellular function. Recall that a more precise description for each protein is found in Table 6.

Regulatory Genes and Their Distribution in the Genome.

Regulatory genes within one family have structural, and functional properties related as discussed before. As mentioned there are several cases suggesting a history of gene duplications. Genes that originated by duplication are more likely to be contiguous neighbors or somehow clustered in the chromosome (Huynen, M. A. and E. van Nimwegen, 1998). Figure 5 shows a graph of the position of all regulator genes and the position *per* family in the chromosome of *E. coli*, indicating the transcription/replication direction (table 11). We expect that overall, transcriptional regulators should reflect the distribution gene duplications in the chromosome.

The distribution of all regulators to detect duplications is difficult to see in a single overview showing all

regulators in the genome (figure 5), for this reason, we analyzed further family *per* family. In this way we find at least, three families whose genome position could reflect duplication events. For instance, the location of *cspA* homologous genes in the *E. coli* chromosome has been discussed suggesting that the Cold shock family probably evolved from a number of gene duplications and, after subsequent adaptations, resulted in specific groups of genes that respond to different environmental stresses (Yamanaka, K. et al. 1998). The *cspG-cspH* and *cspB-cspF* are two highly homologous gene clusters including the intergenic region, and are located at symmetric distances from one of the DNA replication termination sites, *treA*, centered around 29 min. in the *E. coli* chromosome. In addition, *cspC* and *cspE* are located downstream of *cspF* and *cspH*, respectively, forming a mirror image (Figure 5b). CspF and CspH are highly homologous to one another while CspB and CspG are both cold-shock inducible genes (Yamanaka, K. et al. 1998).

Other regulatory families, such as the LysR and GntR families show similar noticeable properties. As already mentioned, the LysR family contains most dual proteins described both in prokaryotes and in *E. coli*, while the GntR family of repressor proteins is associated to carbon sources uptake. Figure 5c shows the distribution of the members of the LysR family in the *E. coli* genome. It is remarkable to see, that most of the LysR family regulators (72% of its members) have a transcriptional direction that is anti-parallel with the direction of replication of the chromosome, while overall in *E. coli*, around 55% of all genes have the same transcription/replication direction. The GntR family shows a similar distribution to that of the Cold-shock family. In this family, around 70% of gene regulators have a mirror distribution in the *E. coli* genome and it could reflect a probable vestige of duplication events. Figure 5d.

Recalling the analysis of direction of transcription/replication in *E. coli* and the suggestion that weakly transcribed genes should preferentially be located near the terminus of replication, and/or with transcription and replication in opposite orientations (Brewer, 1988), one would expect that overall, transcriptional regulators should reflect the distribution of weakly expressed genes. The transcription/replication direction was determined. In many cases the families have a similar distribution to all *E. coli* genes (around 50% with a parallel and 50% with anti-parallel directions), such as, Cold Shock, EBP, LuxR/UhpA, NagC/XylR, TetR/AcrR, and Crp families. Additionally, there are families where the direction of transcription is different to replication, for instance, AraC/XylS, and LysR, OmpR, DeoR, GalR/LacI, IclR families. Four families have a predominance of genes with the same transcription/replication direction, such as, GntR, MerR, and AsnC. Additionally, 45% of all transcriptional genes have the same transcription/replication direction, while 55% have different direction. In figure 5 we graphed the position of all regulator genes in the chromosome of *E. coli.*

In the table 11 we divided the genome of *E. coli* in replichores (Blattner, F. et al. 1997). For each protein families we calculated the parallel and antiparallel genes percentage.

Duplication events can be proposed for proteins that have the following properties: A similar position in the *E. coli* chromosome (particularly, when the family genes have a mirror position), when the length size is homogeneous, and physiological and regulatory function are conserved. In this sense, Cold Shock, GntR could be proposed as protein families originated from gene duplication in the *E. coli* genome.

Although the bias of anti-parallel orientation of transcription and replication is not strong, overall, genes for regulatory proteins tend in effect to be in anti-parallel orientation, with some few proteins having this tendency much more marked, such as LysR, IcIR and GalR/LacI. One may wonder if there is a slight correlation with marked anti-parallel orientation and size of the family, with LysR the largest family, followed by AraC/XylS. It would be interesting to test the hypothesis that large the families should have their members at low levels of expression at a given time, to avoid cross-recognition of relatively similar binding sites.

Immediate Neighbor Genes

Gene order has been observed within several genomes. Functionally related genes tend to be neighbors more often than do unrelated genes. In this sense, we analyzed immediate neighbor genes of the set of 314 transcriptional regulators, concentrating our analysis to neighbors that are also transcriptional regulators. Thus, proteins that belong to two component proteins, as kinases were not included.

We have defined the following functional groups:

1) Regulatory gene-products that interact physically, and that belong to the same transcriptional unit. For instance, *fhlCD*, and *hipAB*.

2) Regulatory genes whose final products are influencing the response of each other, in cascades of regulation. For instance, positive cascades of RhaR/RhaS (Egan, S. M. and R. F. Schleif, 1993), and SoxR/SoxS gene pairs; negative cascades of GutM/GutR, and MarR/MarA. The two component systems belong to this category.

3) Regulatory genes for which there is no evidence of physical interaction of their products, such as the *lacI-mhpR*, and *ycdC-putA* gene pairs. (PutA is divergent to PutP, and LacI has the same transcription direction to *lacZYA* operon).

Gene interactions have been identified in other systems, for instance, the operon *purEK* (Tiedeman, A. A. et al. 1989) whose products are forming a complex implied in the purine ribonucleotide biosynthesis. The correlation between neighbor conservation and genes whose product interact has already been suggested before (Dandekar, T. et al.1998). This kind of organization is implicated in horizontal transfer events.

The analysis of regulatory neighbor genes suggests 5 cases of plausible gene duplication where pairs of contiguos genes belong to the same family. For instance, two gene pairs of the Cold Shock family have been found (*cspB-cspF*, and *cspH-cspG*). Another couple of pairs that might have resulted from gene duplication are the *yhiW-yhiX* and *rhaR-rhaS* genes, belonging to the AraC/XylS family, and finally, the *cbl-nac* gene regulatory pair of the LysR family.

Additionally, around 20 gene regulators are organized in neighbor pairs with one gene in between. For instance, MelR-AdiY and EnvY-AppY genes that belong to AraC/XylS family.

The existence of these neighbor regulatory genes and duplicated pairs should not obscure the fact that the dominant organization of regulatory genes is that of single units of transcription. Out of the 314 known and predicted regulatory genes, 33.4% or 105 regulatory genes are organized as single transcriptional units, and 13% are organized in operons with two or more genes. Finally, 16% of gene regulators have another regulator as neighbor. It is not easy to know if regulatory proteins tend to have a higher tendency to be single units of transcription. Taking a dataset of operons and transcription units grouping 933 genes, 124 genes corresponding to 13% are single genes (Salgado, H. et. al. 1999). However, when we take the complete set of known and predicted transcription units in the complete chromosome, we predict 39% of all genes to be transcribed as single units. We know there are reasons to over-generate isolated genes. Given that there are 480 genes surrounded by genes in different orientation, the minimum number of genes transcribed as single genes in the chromosome correspond to 13%. Given these numbers, it is difficult to assess if the set of 324 transcriptional regulator genes tend to be organized as single-transcriptions units in higher proportion than the total set of genes in *E.coli*.

Four families have a strong tendency to have genes as transcription units isolated. These are LysR (dual regulators), EBP (mostly activators) GalR/LacI (repressor regulators), and LuxR/UhpA (activators). When we analyzed the groups of activator, repressor, and dual proteins in terms of family organization, we obtained that 62.5% of dual proteins belong to the LysR family (remember that proteins of this family are gene activators and repress their own transcription). Members of the GalR/LacI and GntR families contribute with 30% and 12% of all repressor proteins. And members of the AraC/XylS (20%), EBP (16%), and LuxR/UhpA (20%) families integrate activator proteins. AraC/XylS is evenly distributed. Given the small number of regulators in several families -see Table 9- their absence may just be due to the small number of total genes (105 or 33% of all regulators) in this single-transcription set.

Although there is not a clear tendency of regulators to be organized as single-transcription units, its important to emphasize that many gene-regulator clusters can be identified in the chromosome of *E. coli*. Different clusters are identified in the chromosome: regulator-regulator (around 60 genes), and regulator-any gene-regulator (around 20 genes are organized in this form). The second class of cases groups proteins of the OmpR protein family or some regulators whose organization is clearly into operons.

In brief, four regulatory families tend to have their genes as single-transcriptional units in *E. coli*: around the half of the LysR and EBP family genes; 71% and 41% of the GalR/LacI and LuxR/UhpA families. Table 13. This result has at least two alternative explanations: either, there is an over-representation of these families in the collection, or, this genomic organization has been selected for some reason. Isolated genes allow in principle the regulator to be uncoupled on its regulation from the regulated genes. Uncoupled regulation has been analyzed in association with the levels of expression and the gain in Neidhardt and Savageau (1996). It is not clear, currently whether this organization that dominates in some few regulatory proteins is a result of evolution and whether it is being exploited in the functioning of the regulatory circuits in these genes.

On other hand, having two or more genes in the same operon provides the option for a tight equimolar expression in some conditions, whereas at the same time, additional regulatory elements in the same operon can offer for these same genes the option of different expression levels in a different condition. Having regulatory genes within operons provides the genomic architectural basis for coordinate regulation, and cascades of interactions. See table of operon organization.

Additional functions of transcriptional regulators

Some transcriptional factors have been described as proteins with more than one function. Multimodular organization in these proteins could reflect several. Table 14. In this sense, regulatory proteins that affect the DNA organization (chromatin) have been suggested, for instance the housekeeping gene *crl* (Arnqvist, A. 1992), or histone-like protein H1 (Ueguchi, C. et al. 1997). Additionally, *crl* could interact with specific regulatory protein(s) controlling transcription of genes required for curli formation and fibronectin binding (Arnqvist, A, 1992). Similarly H-NS, functions as a global regulator for expression of a wide variety of genes and is at the same time a nucleoid protein (Ueguchi, C. et al. 1997). NadR is a bifunctional protein, serving in both regulatory and transport capacities (NAD transport) (Foster, J. W. 1990). The DnaA protein (52 kD) is essential of the initiation of replication of the *E. coli* chromosome, and it is a repressor for various genes (Shaefer, C. and W. Messer. 1991). DnaA contains four functionally domains: an N-terminal domain implicated in the oligomerization; a second domain containing a P-loop involved in nucleotide binding; a third domain proposed to interact with pSC101 RepA protein; and the fourth domain implied in the DNA-binding (Sutton, M. D. and J. M. Kaguni, 1997). DnaA binds to the origin of replication specifically at a 9 bp consensus (DnaA box): 5'-TAATC(C/A)A(C/A)A-3'. DnaA binds to ATP and to acidic phospholipids and it can inhibit its own gene expression. The binding to a *DnaA* box located within its own promoter gene leads to repression of transcription of the *dnaA* gene (Sutton, M. D. and J. M. Kaguni, 1997). Additional similar cases are described in the following:

Ada is a dual regulator that also functions as an enzyme. *ada* gene encodes a 39kD protein with two ATase functional domains, one acting on O6-alkylguanine and O4-alkylthimine (Gonzaga, P. E. et al. 1990), and the other acting on alkylphosphotriesters. These activities are conserved in 19kDa and 20kD proteolytic fragments respectively. Ada repairs one alkylated guanine in DNA by stoichiometrically transferring the alkyl group at the O-6 position to a cysteine residue in the enzyme. The enzyme is irreversibly inactivated. *Ada* is expressed constitutively at a very low level, but is induced by exposure to low doses of mehylating agents such as N-methyl-N'-nitro-N-nitrosoguanidine (adaptive response) (Gonzaga, P. E. et al. 1990). The transcription-activating function of the Ada protein resides in its N-terminal domain. The methylated Ada

protein activates its own synthesis (Gallegos, M-T. et al. 1997).

The repressor of the biotin biosynthetic operon, BirA (35 kD), is a protein that also catalyzes the formation of biotinyl-5'-adenylate from biotin and ATP and transfers the biotin moiety to other proteins. The level of biotin biosynthetic enzymes in the cell is controlled by the amount of biotinyl-5'-adenylate, which is the BirA correpresor (Streaker, E. D. and D. Beckett, 1998). BirA consists of three structural domains. The N-terminal domain contains a HTH DNA-binding motif, and is loosely connected to the remainder of the molecule. The central domain consists of seven-stranded mixed ß-sheet with α-helices covering one face; while the other face contains the active site. The C-terminal domain comprises a six-stranded, antiparallel ß-sandwich (Wilson, K. P. et al. 1992).

PutA oxidizes proline to glutamate for use as a carbon and nitrogen source and also function as a transcriptional repressor of the *put* operon. PutA has been described as a multifunctional protein: flavin adenine dinucleotide-dependent dehydrogenase activity, NAD-dependent dehydrogenase activity, membrane-binding site, and DNA-binding site. This property of multifunctionality may have evolved to prevent futile cycling of endogenously synthesized proline (Maloy, S. 1987).

The arginine biosynthesis repressor, ArgR, controlled around 20 genes organized in 9 operons. ArgR (17 kD as monomer) is unusual protein repressor because it binds DNA as a hexamer (98 kD) and it is implicated in more than one function (Sunnerhagen, M. et al. 1997). ArgR was found to function as an accessory factor in the resolution of plasmid ColE1 multimers by intramolecular recombination at *cer* sites, where it is implicated both in synapsis and as activator of the XerCD recombines (Chen, S. H. et al. 1997). ArgR is organized in two domains. The C-terminal domain houses the hexamerization and arginine-binding functions, and an N-terminal domain, which contains the DNA-binding function (with a HTH motif).

Fis (Factor for Inversion Stimulation) is a 12kD protein, identified biochemically as a host factor required in vitro for the phage Mu *gin* and Salmonella *hin* site specific DNA inversion reactions (Ross, W. et al. 1990). Fis also participate in a related recombination system, *cin* from phage P1 and to stimulate the excision reaction of the lambda site-specific recombination system. Fis protein is a positive regulator of the tRNA and rRNA promoter *rrnB* B1. Also, Fis prevents initiation of DNA replication from *oriC* (Harrison, S. C. 1991) and regulates its own expression (Kostrewa, D. et al. 1992). Fis binds to DNA as a homodimeric molecule. This protein has three structural domains. The N-terminal domain is disordered; the second domain stabilizes the homodimer; and the third domain presents the HTH DNA-binding motif (Kostrewa, D. et al. 1992).

Hns binds tightly to DNA and increases its thermal stability and inhibits transcription. It also binds to ss-DNA and RNA but with a much lower affinity. Hns has possible histone-like function. Hns plays a role in the thermal control of pili production and it is subject to transcriptional auto-repression. This regulator binds preferentially to the upstream region of its own gene recognizing two segments of DNA on both sides of a bend centered on -150 (Free, A, et al. 1998).

Sunnerhagen, M. et al. (1997) propose that these kind of DNA organization roles are reminiscent of the functions of integration host factor (IHF) in facilitating chromosomal integration and excision of bacteriophage lambda (Segall, A. M. et al. 1994). ArgR and Lrp are members of a growing class of proteins that share features of regulatory proteins, global regulators, and nonspecific gene organizers (Oshima, T. et al. 1995).

Discussion

Based on different sequence similarity search strategies we have defined a comprehensive set of 314 DNA-binding transcriptional regulators in *E. coli* K12. The definition of this set was facilitated by the predominant occurrence of the helix-turn-helix DNA-binding motif in regulatory proteins in *E. coli*, and in fact also in the prokaryotic kingdom. Other DNA-binding motifs are also present in *E. coli*, but in very few regulatory proteins. Based on the specificity of

recognition we estimate around 300 to 350 transcriptional regulators in *E. coli*.

Regulatory proteins share a significant amount of protein similarity, enabling also a clustering of them into families of plausible common evolutionary origin. The diagnostic region shared within a family imposes an identity of around 25% to members within one family. We find in this way that all the twenty families of the HTH regulatory proteins of the bacterial kingdom have representatives in *E. coli*. Most of these families appear to be quite homogeneous groups whose members share several properties. These are families with proteins with a rather similar length, and with their HTH domain localized in the same relative position either in the C-terminal or in the N-terminal. Regulators within a family tend to be mostly activators, or mostly repressors in others, with the dual regulators concentrated in the most abundant family -LysR- with 45 members in *E. coli*. These families group regulators that tend to affect genes involved in related biological functions. All these common structural and functional properties support the notion of a family even if some of them have only one member in *E. coli*. The additional correlation between the HTH in the C-terminal for activators and the HTH positioning for known repressors in the N-terminal was used, in combination with family membership, to generate a predicted functional role for most of the 314 regulatory proteins in *E. coli*. This produces a picture of a quite even distribution of 34%, 42%, and 24% of activators, repressors, and dual regulators respectively.

Evolution is however more flexible as illustrated by multi-domain regulatory proteins that are more difficult to group. Thus, some families group a less homogeneous set of proteins. The existence of multiple domain makes their clustering more difficult to achieve, grouping proteins with a more elaborate evolutionary history with some motifs appearing only in some members of the family. In some cases, multiple domains reflect the existence of several evolutionary events, such as the shuffling process in prokaryotic proteins of the two component systems.

An interesting question related to these observations is that of the different size or abundance of regulatory proteins within the different families. Huynen and van Nimwegen (1998) have shown that genes within one family have similar functions, but as the requirements of this function vary over time so does the presence of the gene family in the genome. Activation of transcription by different types of metabolites (e.g. aromatics, ions, amino acid derivatives, nucleic acids, sugars, etc) suggests an ancient divergence of signal recognition within regulatory families. Remember that the HTH region contributes in an important way to the sequence similarity within members of the families. The evolutionary flexibility within regulatory families can be appreciated when observing the structural diversity of the different co-inducers that stimulate various transcriptional factors that belong to the same family, as opposed to the highly conserved HTH domain. An additional source of diversity is the presence of self-transmissible plasmids, which probably move freely throughout the prokaryotic community, and may have promoted a more recent and rapid dissemination and evolution (Schell, M. A. 1993).

A good number of well documented regulatory families show a tendency to members with similar size. We hypothesize that the largest proteins (such as the maltose activator, MalT) could have other functions in addition to transcriptional regulation (like the proline-dehydrogenase of PutA), while smaller or medium proteins, usually acting as multimers could rarely have additional functions.

The amount of information of known binding sites is limited to around 1/6 (50 out of 300) of all transcriptional potential regulators of *E. coli*. It is interesting to address the question of whether the grouping of proteins into families would limit or structure their DNA-binding available space. It is possible to imagine one scenario where proteins of the same family could be recognizing the same DNA-binding sites, or at least, similar sites. Whether a mechanism of coevolution between one family and one set of DNA-binding sites exist, and their functional implications, should be investigated.

We analyzed as another potential trait that could help to characterize the different regulatory families their gene location into operons and transcription-isolated

genes, as well as their neighboring genes specially if these are also regulatory proteins. One salient feature is that 30% of all transcriptional regulators occur as isolated transcriptional units. This organization may be relevant for their independent transcriptional regulation as uncoupled from the set of genes they regulate.

We considered equally interesting to study the orientation of transcription and replication in this set of regulatory proteins, given the rational hypothesis presented years ago about such distribution (Brewer, 1988). This predicted tendency is in fact confirmed, although not very strongly. It will be interesting to test with global studies if this set of genes is expressed in low amounts as assumed. It may also be that transcriptional regulators have a pattern in genome localization that does not differ particularly from the complete set of genes of *E. coli*. This same conclusion seems justified when studying their organization into operons or into single transcriptional units.

In brief, we have learned that *E. coli* transcriptional regulators are grouped into families reflecting their common evolutionary origin. Also these families, share functional and structural conserved properties. Less than 10% of all genes in *E. coli* would participate as transcriptional regulators. Other proteins are certainly involved in regulation of transcription, making the fraction of regulator genes. Whether the properties of this collection in *E. coli* are also found to occur in other bacteria is an interesting question.

1. References

2. **Aldea, M., T. Garrido, C. Hernandez-Chico, M. Vicente, and S. R. Kushner.** 1989. Induction of a growth-phase-dependent promoter triggers transcription of *bolA*, an *Escherichia coli* morphogene. EMBO J. **8**:3923-3931.
3. **Altschul, S. F., T. L. Madden, A. A. Schäffer, J. Zhang, Z. Zhang, W. Miller, and D. Lipman.** 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. **25**:3389-3402.
4. **Ariza, R. R., S. P. Cohen, N. Bachhawat, 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.
5. **Arnqvist, A., A. Olsen, J. Pfeifer, D. G. Russell, and S. Normark.** 1992. The Crl protein activates cryptic genes for curli formation and fibronectin binding in *Escherichia coli* HB101. Mol. Microbiol. **6**:2443-2452.
6. **Autexier, C., and M. S. DuBow.** 1992. The *Escherichia coli* Mu/D108 phage *ner* homologue gene (*nlp*) is transcribed and evolutionarily conserved among the Enterobacteriaceae. Gene. **114**:13-18.
7. **Baikalov, I., I. Schroeder, M. Kaczor-Grzeskowiak, D. Cascio, R. P. Gunsalus, and R. E. Dickerson.** 1998. NarL dimerization? Suggestive evidence from a new crystal form. Biochemistry. **37**:3665-3676.
8. **Bailey, M.J., C. Hughes, and V. Koronakis.** 1997. RfaH and the *ops* element, components of a novel system controlling bacterial transcription elongation. Mol. Microbiol. **26**:845-851.
9. **Bairoch, A.** 1993. A possible mechanism form metal-ion induced DNA-protein dissociation in a family of a prokaryotic transcriptional regulators. Nucleic Acids Res. **21**:2515.
10. **Bairoch A. and R. Apweiler. 1996.** The Swiss-Prot protein sequence databank and its new supplement TREMBL. Nucleic Acids Res. **24**:21-25.
11. **Bairoch, A., and B. Boeckmann. 1992.** The SWISS-PROT protein sequence data bank. Nucleic Acids Res. **20**:2019-2022.
12. **Bairoch, A. , Bucher, P. and K. Hoffman.** 1997. The PROSITE database, its status and progress. Nucleic Acids Res. **25**:217-221.
13. **Bejar, S., K. Cam, and J. P. Bouche.** 1986. Control of cell division in *Escherichia coli*. DNA sequence of *dicA* and of a second gene complementing mutation *dicA1*, *dicC*. Nucleic Acids Res. **14**:6821-6833.
14. **Black, D.S., A. J. Kelly, M. J. Mardis, and H. S. Moyed.** 1991. Structure and organization of *hip*,

an operon that affects lethality due to inhibition of peptidoglycan or DNA synthesis. J. Bacteriol. **173**:5732-5739.

15. **Blattner, F. R., G. 3rd. Plunkett, C. A. Bloch, N. T. Perna, V. Burland, M. Riley, J. Collado-Vides, J. D.Glasner, C. K. Rode, G.F. Mayhew, J. Gregor, N. W. Davis, H. A. Kirkpatrick, M. A. Goeden, D. J. Rose, B. Mau, and Y. Shao.** 1997. The complete genome sequence of *Escherichia coli* K-12. Science. **277**:1453-1462.
16. **Bongaerts, J., S. Zoske, U. Weidner, and G. Unden.** 1995. Transcriptional regulation of the proton translocating NADH dehydrogenase genes (*nuoA-N*) of *Escherichia coli* by electron acceptors, electron donors and gene regulators. Mol. Microbiol. **16**:521-34
17. **Bork, P., and E. V. Koonin.** 1998. Predicting functions from protein sequences--where are the bottlenecks?. Nat. Genet. **18**:313-8
18. **Branden, C. and J. Tooze.** 1991. Introduction to Protein Structure. Garland Publishing, Inc. New York and London.
19. **Brazma, A., I. Jonassen, I. Eidhammer, and D. Gilbert.** 1998. Approaches to the automatic discovery of patterns in biosequences. J. Comput. Biol. 5:279-305.
20. **Brennan, R. G.** 1993. The winged-helix DNA-binding motif: another helix-turn-helix takeoff. Cell. **74**:773-776.
21. **Brewer, B. J.** 1988. When polymerases collide: replication and the transcriptional organization of the *E. coli* chromosome. Cell. **53**:679-686.
22. **Bsat, N., A. Herbig, L. Casillas-Martinez, P. Setlow, and J. D. Helmann.**1998. *Bacillus subtilis* contains multiple Fur homologues: identification of the iron uptake (Fur) and peroxide regulon (PerR) repressors. Mol. Microbiol. **29**:189-198.
23. **Canellakis, E. S., A. A. Paterakis, S. C. Huang, C. A. Panagiotidis, and D. A. Kyriakidis.** 1993. Identification, cloning, and nucleotide sequencing of the ornithine decarboxylase antizyme gene of *Escherichia coli*. Proc. Natl. Acad. Sci. **90**:7129-7133.
24. **Castillo, I., J. E. Gonzalez-Pastor, J. L. San Millan, and F. Moreno.** 1991. Nucleotide sequence of the *Escherichia coli* regulatory gene *mprA* and construction and characterization of *mprA*-deficient mutants. J Bacteriol. **173**:3924-3929.
25. **Chen, S. H., A. F. Merican, and D. J. Sherratt.** 1997. DNA binding of *Escherichia coli* arginine repressor mutants altered in oligomeric state. Mol. Microbiol. **24**:1143-56.
26. **Chen, Y.M., Y. Zhu, and E. C. Lin.** 1987. The organization of the *fuc* regulon specifying L-fucose dissimilation in *Escherichia coli* K12 as determined by gene cloning. Mol. Gen. Genet. **210**:331-337.
27. **Chothia, C. 1992.** One thousand families for the molecular biologist. Nature. **357**: 543-544.
28. **Chu, S., J. DeRisi, M. Eisen, J. Mulholland, D. Botstein, P. O. Brown, and I. Herkowitz.** 1998. The transcriptional program of sporulation in budding yeast. *Science* **282**:699-705.
29. **Colbeau, A., P. Richaud, B. Toussaint, F. J. Caballero, C. Elster, C. Delphin, R. L. Smith, J. Chabert, and P. Mau Vignais.** 1993. Organization of the genes necessary for hydrogenase expression in *Rhodobacter capsulatus*. Sequence analysis and identification of two *hyp* regulatory mutants. Mol. Microbiol. **8**:15-29.
30. **Collado-Vides, J., B. Magasanik and J. D. Gralla. 1991.** Control site location and transcriptional regulation in *Escherichia coli*. Microbiol. Reviews. **55**:371-394.
31. **Coppola, G., F. Huang, J. Riley, J. L. Cox, P. Hantzopoulos, L. B. Zhou, and D. H. Calhoun.** 1991. Sequence and transcriptional activity of the *Escherichia coli* K-12 chromosome region between *rrnC* and *ilvGMEDA*. Gene. **97**:21-27.
32. **Christie, G. E., T. J. White, and T. S. Goodwin.** 1994. A *merR* homologue at 74 minutes on the *Escherichia coli* genome. Gene. **146**:131-132.
33. **Dandekar, T., B. Snel, M. Huynen, and P. Bork.** 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem Sci. **23**:324-328.
34. **Death, A., and T. Ferenci.** 1994. Between feast and famine: endogenous inducer synthesis in the

adaptation of *Escherichia coli* to growth with limiting carbohydrates. J Bacteriol. **176**:5101-7.

35. **Delarue, M., A. Poterszman, and S. Nikonov.** 1994. Crystal structure of a prokaryotic aspartyl tRNA-synthase. EMBO J. **13**:3219-3229.
36. **deRisi, J. L., V. R. Iyer, and P. O. Brown.** 1997. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**: 680-686.
37. **Dodd, I. B., and J. B. Egan**, 1987. Systematic method for the detection of potential λ Cro-like DNA-binding regions in proteins. J. Mol. Biol. **194**:557-564.
38. **Dong, J. M., J. S. Taylor, D. J. Latour, S. Iuchi, and E. C. Lin.** 1993. Three overlapping *lct* genes involved in L-lactate utilization by *Escherichia coli*. J. Bacteriol. **175**:6671-8
39. **Eichler, K., A. Buchet, R. Lemke, H. P. Kleber, and M. A. Mandrand-Berthelot.** 1996. Identification and characterization of the *caiF* gene encoding a potential transcriptional activator of carnitine metabolism in *Escherichia coli*. J. Bacteriol. **178**:1248-1257.
40. **Egan, S. M., and R. F. Schleif.** 1993. A regulator cascade in the induction of *rhaBAD*. J. Mol. Biol. **234**:87-98.
41. **Escolar, L., J. Pérez-Martin, and V. de Lorenzo.** 1998. Binding of the FUR (Ferric Uptake Regulator) repressor of *Escherichia coli* to arrays of the GATAAT sequence. J. Mol. Biol. 283:537-547.
42. **Feng, W., Tejero, R., Zimmerman, D. E., Inouye, M., and Montelione, G. T.** 1998. Resonance assignments, solution NMR structure, and backbone dynamics of the major Cold-Shock protein C from *Escherichia coli*: evidence for conformational dynamics in the proposed ssRNA-binding site. Biochemistry. **37**:10881.
43. **Ferrandez, A., J. L. Garcia, and E. Diaz.** 1997. Genetic characterization and expression in heterologous hosts of the 3-(3-hydroxyphenyl)propionate catabolic pathway of *Escherichia coli* K-12. J. Bacteriol. **179**:2573-81.
44. **Figge, R. M., T. M. Ramseier, and M. H. Saier Jr.** 1994. The mannitol repressor (MtlR) of *Escherichia coli*. J. Bacteriol. **176**:840-847.
45. **Fleischmann, R. D., M. D. Adam, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick, et al.** 1995. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. Science. **269**:496-512.
46. **Fogh R. H., G. Ottleben, H. Rueterjans, M. Schnarr, R. Boelens, and R. Kaptein.** 1994. Solution structure of the LexA repressor DNA binding domain determined by 1H NMR spectroscopy. EMBO J. **13**:3936-3944.
47. **Foster, J. W., Y. K. Park, T. Penfound, T. Fenger, and M. P. Spector.** 1990. Regulation of NAD metabolism in *Salmonella typhimurium*: molecular sequence analysis of the bifunctional *nadR* regulator and the *nadA-pnuC* operon. J. Bacteriol. **172**:4187-4196.
48. **Free, A., R. M. Williams, and C. J. Dorman.** 1998.The StpA protein functions as a molecular adapter to mediate repression of the *bgl* operon by truncated H-NS in *Escherichia coli*. J. Bacteriol. **180**:994-997.
49. **Gallegos, M-T., R. Schelif, A. Bairoch, K. Hoffman, and J. L. Ramos.** 1997. AraC/XylS family of transcriptional regulators. Microb. and Mol. Biol. Rev. **61**:393-410.
50. **Ghrist, A. C., and G. V. Stauffer.** 1998. Promoter characterization and constitutive expression of the *Escherichia coli gcvR* gene.J. Bacteriol. **180**:1803-1807.
51. **Ghrist, A. C., and G. V. Stauffer.** 1995. Characterization of the *Escherichia coli gcvR* gene encoding a negative regulator of *gcv* expression. J. Bacteriol. **177**:4980-4984.
52. **Gonzaga, P. E., L. Harris, G. P. Margison, and T. P. Brent.** 1990. Evidence that covalent complex formation between BCNU-treated oligonucleotides and *E. coli* alkyltransferases requires the O6-alkylgunanine function. Nucleic Acids Res. **18**: 3961-3966.
53. **Gralla J. D. and J. Collado-Vides.** 1996. Organization and function of transcriptional

regulatory elements. p. 1232-1246. In F. C. Neidhart, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, Vol.1. American Society for Microbiology, Washington, D. C.

54. **Graumman, P. and M. A. Marahiel.** 1996. A case of convergent evolution of nucleic acid binding modules. BioEssays. **18**:309-315.
55. **Gui, L., A. Sunnarborg, B. Pan, and D. C. LaPorte.** 1996.Autoregulation of *iclR*, the gene encoding the repressor of the glyoxylate bypass operon. J. Bacteriol. **178**:321-324
56. **Gutierrez, C., S. Gordia, and S. Bonnassie.** 1995. Characterization of the osmotically inducible gene *osmE* of *Escherichia coli* K-12. Mol. Microbiol. **16**:553-563.
57. **Hall, B. G., and L. Xu.** 1992. Nucleotide sequence, function, activation, and evolution of the cryptic *asc* operon of *Escherichia coli* K12. Mol. Biol. Evol. **9**:688-706.
58. **Hammar, M., A. Arnqvist, Z. Bian, A. Olsen, and S. Normark.** 1995. Expression of two *csg* operons is required for production of fibronectin- and congo red-binding curli polymers in *Escherichia coli* K-12.Mol. Microbiol. **18**:661-670.
59. **Harrison, S. C.** 1991. A structural taxonomy of DNA-binding domains. Nature. **353**: 715-719.
60. **Haydon, D. and J. Guest.** 1991. A new family of bacterial regulatory proteins. FEMS Microbiol. Lett. **79**: 291-296.
61. **Henikoff, S., G. W. Haughn, J. M. Calvo, and J. C. Wallace.** 1988. A large family of bacterial activator proteins. Proc. Natl. Acad. Sci. **85**:6602-6606.
62. **Henikoff, S., J. C. Wallace, and J. P. Brown.** 1990. Finding protein similarities with nucleotide sequence databases. Meth. Enzym. **183**:111-132.
63. **Hinrichs W., C. Kisker, C. Duevel, A. Mueller, K. Tovar, W. Hillen, and W. Saenger.** 1994. Structure of the Tet repressor-tetracycline complex and regulation of antibiotic resistance. Science **264**:418-420.
64. **Holm, L., C. Sander, H. Rüterjans, M. Schnarr, R. Fogh, R. Boelens, and R. Kaptein.** 1994. LexA repressor and iron uptake regulator from *Escherichia coli*: new members of the CAP-like DNA binding domain superfamily. Prot. Engin. **7**:1449-1453.
65. **Holm, L.** 1998. Unification of protein families. Current Opinion in Structural Biology. **8**:372-379.
66. **Horswill, A. R., and J. C. Escalante-Semerena.** 1997. Propionate catabolism in *Salmonella typhimurium* LT2: two divergently transcribed units comprise the *prp* locus at 8.5 centisomes, *prpR* encodes a member of the sigma-54 family of activators, and the *prpBCDE* genes constitute an operon. J. Bacteriol.**179**:928-940.
67. **Huerta, A.M., H. Salgado, D. Thieffry, and J. Collado-Vides.** 1998. RegulonDB: a database on transcriptional regulation in *Escherichia coli*. Nucleic Acids Res. **26**:55-59.
68. **Huynen, M. A., and E. van Nimwegen.** 1998. The frequency distribution of gene family sizes in complete genomes. Mol. Biol. Evol. **15**:583-589.
69. **Iwanicka-Nowicka, R., and M. M. Hryniewicz** .1995. A new gene, *cbl*, encoding a member of the LysR family of transcriptional regulators belongs to *Escherichia coli* cys regulon. Gene. **166**:11-17.
70. **Jones, D.T.** 1999. GenTHREADER: an efficient and reliable protein fold recognition method for genomic sequences. J Mol Biol. **287**:797-815
71. **Jourlin, C., M. Ansaldi, and V. Mejean.** 1997. Transphosphorylation of the TorR response regulator requires the three phosphorylation sites of the TorS unorthodox sensor in *Escherichia coli*. J. Mol. Biol. **267**:770-777.
72. **Jovanovic, G., J. Rakonjac, and P. Model.** 1999. In vivo and in vitro activities of the *Escherichia coli* sigma54 transcription activator, PspF, and its DNA-binding mutant, PspFDeltaHTH. J Mol Biol. **285**:469-483.
73. **Kabsch, W., and C. Sander.** 1983. Definition of secondary structure of proteins given a set of 3D coordinates. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers **22**: 2577-2637.

74. **Kawamukai, M., R. Utsumi, K. Takeda, A. Higashi, H. Matsuda, Y.L. Choi, and T. Komano.** 1991. Nucleotide sequence and characterization of the *sfs1* gene: *sfs1* is involved in CRP*-dependent mal gene expression in *Escherichia coli*. J. Bacteriol. **173**:2644-2648.

75. **Kim, H. and S. K. Farrand.** 1997. Characterization of the *acc* operon from the nopaline-type Ti plasmid pTiC58 which encodes utilization of agrocinopines A and B and suceptibility to agrocin 84. J. Bacteriol. **179**:7559-7572.

76. **Kimata, K., T. Inada, H. Tagami, and H. Aiba.** 1998. A global repressor (Mlc) is involved in glucose induction of the *ptsG* gene encoding major glucose transporter in *Escherichia coli*. Mol Microbiol. **29**:1509-1519.

77. **Kisker, C., W. Hinrichs, K. Tovar, W. Hillen, and W. Saenger.** 1995. The complex formed between Ter repressor and tetracycline-Mg2+ reveals mechanism of antibiotic resistence. J. Mol. Biol. **247**:260-280.

78. **Klein, W., R. Horlacher, and W. Boos.** 1995. Molecular analysis of *treB* encoding the *Escherichia coli* enzyme II specific for trehalose. J. Bacteriol. **177**:4043-4052.

79. **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.

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

81. **Koonin, E.V., R. L. Tatusov, and K. E. Rudd.** 1995. Sequence similarity analysis of *Escherichia coli* proteins: Functional and evolutionary implications.Proc. Natl. Acad. Sci. **92**: 11921-11925.

82. **Koonin, E. V., R. L. Tatusov, and K. E. Rudd.** 1997. *Escherichia coli* protein sequences: Functional and Evolutionay implications. In p. 2203-2217. In F. C. Neidhart, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, Vol.2. American Society for Microbiology, Washington, D. C

83. **Kostrewa, D., J. Granzin, D. Stock, H-W. Choe, J. Labahn, and W. Saenger.** 1992. Crystal structure of the factor for inversion stimulation FIS at 2.0Å resolution. J. Mol. Biol. **226**:209-226.

84. **Kunst, F., N. Ogasawara, I. Moszer, A. M. Albertini, G. Alloni, V. Azevedo, M. G. Bertero, P. Bessieres, A. Bolotin, S. Borchert, R. Borriss, L. Boursier, A. Brans, M. Braun, S. C. Brignell, S. Bron, S. C. V. Brouillet, B. Caldwell, V. Capuano, N. M. Carter, S. K. Choi, J. J. Codani, I. F. Connerton, A. Danchin. et al.** 1997. The complete genome of the gram-positive bacterium *Bacillus subtilis*. Nature. **390**:249-256.40.

85. **Kyrpides, N. C., and C. A. Ouzounis.** 1995. The eubacterial transcriptional activator Lrp is present in the archaeon *Pyrococcus furiosus*. Trends Biochem Sci. **20**:140-1.

86. **Labedan, B., and M. Riley.** 1995. Gene products of *Escherichia coli*: Sequence comparisons and common ancestries. Mol. Biol. Evol. **12**: 980-987.

87. **Lawrence, C. E., S. F. Altschul, M.S. Boguski, J.S.Liu, A.F. Neuwald, and J.C. Wooton.** 1993. Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. Science. **262**:208-214.

88. **Lawson, C. L., and P. B. Sigler.** 1988. The structure of *trp* pseudorepressor at 1.65A shows why indole propionate acts as a *trp* 'inducer'. Nature **333**:869-871.

89. **Le, S. J., A. Xie, W. Jiang, J. P. Etchegaray, P. G. Jones, and M. Inouye.** 1994. Family of the major cold-shock protein, CspA (CS7.4), of *Escherichia coli*, whose members show a high sequence similarity with the eukaryotic Y-box binding proteins. Mol. Microbiol. 11:833-839.

90. **Lewis, M., G. Chang, N. C. Horton, M. A. Kercher, H. C. Pace, M. A. Schumacher, R. R. Brennan, and P. Lu.** 1996. Crystal structure of the lactose operon repressor and its complexes with DNA and inducer. Science **271**:1247-1254.

91. **Liu, X., and P. Matsumura.** 1996. Differential regulation of multiple overlapping promoters in flagellar class II operons in *Escherichia coli*. Mol Microbiol. **21**:613-620

92. **Liu, X., and P. Matsumura.** 1994. The FlhD/FlhC complex, a transcriptional activator of the *Escherichia coli* flagellar class II operons. J. Bacteriol. **176**:7345-7351

93. **Lomovskaya, O., K. Lewis, and A. Matin.** 1995. EmrR is a negative regulator of the *Escherichia coli* multidrug resistance pump EmrAB. J. Bacteriol. **177**:2328-2334

94. **Lundberg, L.G., H.O. Thoresson, O.H. Karlstrom, and P.O. Nyman.** 1983. Nucleotide sequence of the structural gene for dUTPase of *Escherichia coli* K-12. EMBO J. **2**:967-971.

95. **Ludwig, A., C. Tengel, S. Bauer, A. Bubert, R. Benz, H.J. Mollenkopf, and W. Goebel.** 1995. SlyA, a regulatory protein from *Salmonella typhimurium*, induces a haemolytic and pore-forming protein in *Escherichia coli*. Mol Gen Genet. **249**:474-486.

96. **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.

97. **Maier, T., U. Binder, and A. Bock.** 1996. Analysis of the *hydA* locus of *Escherichia coli*: two genes (*hydN* and *hypF*) involved in formate and hydrogen metabolism. Arch. Microbiol. **165**:333-41.

98. **Magasanik, B. and F. C. Neidhardt.** 1987. Regulation of carbon and nitrogen utilization. 1318-1325. In F. C. Neidhadrt, J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, Vol. II. American Society for Microbiology, Washington, D. C.

99. **Maloy, S.** 1987. The proline utilization operon. p. 1513-1519. 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, Vol. II. American Society for Microbiology, Washington, D. C.

100. **Martinez-Hackert, E, and M. Stock.** 1997. Structural Relationships in the OmpR family of Winged-Helix transcription factors. J. Mol. Biol. **269**:301-312.

101. **Masuda, Y., K. Miyakawa, Y. Nishimura, and E. Ohtsubo.** 1993. *chpA* and *chpB*, *Escherichia coli* chromosomal homologs of the *pem* locus responsible for stable maintenance of plasmid R100. J. Bacteriol. **175**:6850-6856.

102. **Mauzy, C.A., and M.A. Hermodson.** 1992. Structural and functional analyses of the repressor, RbsR, of the ribose operon of *Escherichia coli*. Protein Sci. **1**:831-842.

103. **Meyer, M., P. Dimroth, and M. Bott.** 1997. In vitro binding of the response regulator CitB and of its carboxy-terminal domain to A + T-rich DNA target sequences in the control region of the divergent *citC* and *citS* operons of *Klebsiella pneumoniae*. J. Mol. Biol. **269**:719-31.

104. **Missiakas, D., C. Georgopoulos, and S. Raina.** 1993. The *Escherichia coli* heat shock gene *htpY* mutational analysis, cloning, sequencing, and transcriptional regulation. J. Bacteriol. **175**:2613-24. Published erratum appears in J. Bacteriol. 1993. **175**:7124.

105. **Morett, E. and L. Segovia.** 1993. The σ54 bacterial enhancer-binding protein family: Mechanism of action and phylogenetic relationship of their functional domains. J. Bacteriol. **175**:6067-6074.

106. **Moyed, H. S., and S. H. Broderick.** 1986. Molecular cloning and expression of *hipA*, a gene of *Escherichia coli* K-12 that affects frequency of persistence after inhibition of murein synthesis. J. Bacteriol. **166**:399-403.

107. **Murooka, Y., H. Azakami, and M. Yamashita.** 1996. The monoamine regulon including syntheses of arylsulfatase and monoamine oxidase in bacteria. Biosci. Biotechnol. Biochem. **60**:935-941.

108. **Muse, W. B., and R. A. Bender.** 1999. The amino-terminal 100 residues of the nitrogen assimilation control protein (NAC) encode all known properties of NAC from *Klebsiella*

aerogenes and *Escherichia coli*. J. Bacteriol. **181**:934-940

109. **Neidhardt, F. C., J. L. Ingraham, K. B. Low, B. Magasanik, M. Schaechter, and H. E. Umbarger (ed).** 1987.*Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, Vol. II. American Society for Microbiology, Washington, D. C.

110. Neidhardt, F. C., R. Curtiss III., J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaerchter, and H. E. Umbarger (ed.). (1996). *Escherichia coli and Salmonella typhimurium. celular and Molecular Biology* (2nd ed.). American Society for Microbiology, Washington D.C.

111. **Neidhardt, F. C. and M. Savageau.** 1996. Regulation beyond the operon. p. 1310-1324. In F. C. Neidhardt , R. Curtiss III., J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaerchter, and H. E. Umbarger (ed.). (1996). *Escherichia coli and Salmonella typhimurium. cellular and Molecular Biology* (2nd ed.). American Society for Microbiology, Washington D.C

112. **Newman, E. B., R. T. Lin, and R. D'Ari.** 1996. The leucine/Lrp regulon. 1513-1526. Neidhardt, F. C., R. Curtiss III., J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. S. Reznikoff, M. Riley, M. Schaerchter, and H. E. Umbarger (ed.). Vol. 1. American Society for Microbiology, Washington, D. C.

113. **Ninfa, A.** J. 1996. Regulation of gene transcription by extracellular stimuli. p. 1246-1262. In F. C. Neidhart, R. Curtiss III, J. L. Ingraham, E. C. C. Lin, K. B. Low, B. Magasanik, W. Reznikoff, M. Riley, M. Schaechter, and H. E. Umbarger (ed), *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology, Vol.1. American Society for Microbiology, Washington, D. C.

114. **Nobelmann, B., and J. W. Lengeler.** 1996. Molecular analysis of the *gat* genes from *Escherichia coli* and of their roles in galactitol transport and metabolism. J. Bacteriol. **178**:6790-6795.

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

116. **Olsen, G. J., C. R. Woese, and R. Overbeek.** 1994. The winds of (evolutionary) change: breathing new life into microbiology. J. Bacteriol. **176**:1-6.

117. **Olson, J. W., and R. J. Maier.** 1997.The sequences of *hypF*, *hypC* and *hypD* complete the hyp gene cluster required for hydrogenase activity in *Bradyrhizobium japonicum*. Gene. 199:93-9.

118. **Oshima, T., K. Ito, H. Kabayama, and Y. Nakamura.** 1995. Regulation of *lrp* gene expression by H-NS and Lrp proteins in *Escherichia coli*: dominant negative mutations in *lrp*. Mol Gen Genet. **247**:521-8.

119. **Pao G. M. and M. H. Saier Jr.**1995. Response regulators of bacterial signal transduction systems: Selective domain shuffling during evolution. J. Mol. Evol. **40**: 136-154.

120. **Pabo, C. O., and R.T. Sauer.** 1984. Protein-DNA recognition. Annu. Rev. Biochem. **53**:293-321.

121. **Peekhaus, N., and T. Conway.** 1998. What's for dinner?: Entner-Doudoroff Metabolism in *Escherichia coli*. J. Bacteriol. **180**:3495-3502.

122. **Pellicer, M. T., J. Badia, J. Aguilar, and L. Baldoma.** 1996. *glc* locus of *Escherichia coli*: characterization of genes encoding the subunits of glycolate oxidase and the glc regulator protein. J. Bacteriol. **178**:2051-2059.

123. **Penin F., C. Georjon, R. Montserret, A. Bockmann, A. Lesage, Y. Yang, C. Bonod-Bidaud, J. C. Cortay, D. Negre, A. J. Cozzone, and G. Deleage.** 1997. Three-dimensional structure of the DNA-binding domain of the fructose repressor from *Escherichia coli* by 1H and 15N NMR. J. Mol. Biol. **270**:496-510.

124. **Pérez-Rueda, E., J. Gralla, and J. Collado-Vides.** 1998. Genomic position analyses and the transcription machinery. J. Mol. Biol. **275**:165-170.

125. **Pujic, P., R. Dervyn, A. Sorokin, and S. D. Ehrlich.** 1998. The *kdgRKAT* operon of *Bacillus*

subtilis: detection of the transcript and regulation by the *kdgR* and *ccpA* genes. Microbiology. **144**:3111-3118.

126. **Rabin, R. S., and V. Stewart**. 1993. Dual response regulators (NarL and NarP) interact with dual sensors (NarX and NarQ) to control nitrate- and nitrite-regulated gene expression in *Escherichia coli* K-12. J. Bacteriol. **175**:3259-3268.

127. **Reizer, J., A. Charbit, A. Reizer, M.H. Saier Jr.** 1996. Novel phosphotransferases system genes revealed by bacterial genome analysis: operons encoding homologues of sugar-specific permease domains of the phosphotransferase system and pentose catabolic enzymes. Genome Sci. Technol. **1**:52-75.

128. **Reizer, J., T.M. Ramseier, A. Reizer, A. Charbit, and M. H. Saier Jr. 1996**. Novel phosphotransferase genes revealed by bacterial genome sequencing: a gene cluster encoding a putative N-acetylgalactosamine metabolic pathway in *Escherichia coli*. Microbiology. **142**:231-250.

129. **Reizer, J., V. Michotey, A. Reizer, and M. H. Saier Jr. 1994.** Novel phosphotransferase system genes revealed by bacterial genome analysis: unique, putative fructose- and glucoside-specific systems. Protein Sci. **3**:440-450.

130. **Rey, L., J. Murillo, Y. Hernando, E. Hidalgo, E. Cabrera, J. Imperial and T. Ruiz-Argueso**. 1993. Molecular analysis of the microaerobically induced operon required for hydrogenase synthesis in *Rhizobium leguminosarum* biovar viciae. Mol. Microbiol. **8**:471-481.

131. **Rice, P. A., and T. A. Steitz**. 1989. Ribosomal protein L7/L12 has a helix-turn-helix motif similar to that found in DNA-binding regulatory proteins. Nucleic Acids Res. **17**:3757-62.

132. **Richmond, C.S., J. D. Glasner, R. Mau, H. Jin, and F. R. Blattner**. 1999. Genome-wide expression profiling in *Escherichia coli* K12. Nucleic Acids Res. **27**:3821-3835

133. **Robison, K. A. M. MGuire, and G.M.Church.** 1998. A comprehensive library of DNA-binding Site Matrices for 55 proteins applied to the complete *Escherichia coli* K-12 Genome. J. Mol. Biol. 284:241-254.

134. **Rosey, E. L. and G.C. Stewart**. 1992. Nucleotide and deduced amino acid sequences of the *lacR*, *lacABCD*, and *lacFE* genes encoding the repressor, tagatose 6-phosphate gene cluster, and sugar-specific phosphotransferase system components of the lactose operon of *Streptococcus mutants*. J. Bacteriol. **174**:6159-6170.

135. **Ross, W., J. F. Thompson, J. T. Newlands, and R. L. Gourse.** 1990. *E. coli* Fis protein activates ribosomal RNA transcription *in vitro* and *in vivo*. EMBO J. **9**:3733-3742.

136. **Roth, A. and W. Messer**. 1995. The DNA binding domain of the initiator protein DnaA. EMBO J. **14**:2106-2111.

137. **Roth, F. P., J. D. Hughes, P. W. Estep, and G. M. Church**. 1998. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat. Biotechnol. **16**:939-45.

138. **Salgado, H., A. Santos, U. Garza-Ramos, J. van Helden, E. Diaz, and J. Collado-Vides.** 1999. RegulonDB (version 2.0): a database on transcriptional regulation in *Escherichia coli*. Nucleic Acids Res. **27**:59-60.

139. **Savageau, M.** 1986. Proteins of *Escherichia coli* come in sizes that are multiples of 14kDa: Domain concepts and evolutionary implications. Proc. Natl. Acad. Sci. **83**:1198-1202.

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

141. **Schlensog, V., S. Lutz, and A. Bock.** 1994. Purification and DNA-binding properties of FhlA, the transcriptional activator of the formate hydrogenlyase system from *Escherichia coli*. J Biol Chem **269**:19590-19596.

142. **Schumacher, M. A., A. Glasfeld, H. Zalkin, and R. G. Brennan**. 1997. The X-ray structure of the PurR-guanine-*purF* operator complex reveals the contributions of complementary electrostatic surfaces and a water-mediated hydrogen bond to

corepressor specificity and binding affinity. J. Biol. Chem. **272**:22648-22653.

143. **Schweizer, H. P., and C. Po.** 1996. Regulation of glycerol metabolism in *Pseudomonas aeruginosa*: characterization of the *glpR* repressor gene. J. Bacteriol. **178**:5215-5221.

144. **Segall, A. M., S. D. Goodman, and H. A. Nash.** 1994. Architectural elements in nucleoprotein complexes: interchangeability of specific and non-specific DNA binding proteins. EMBO J. **13**:4536-4548.

145. **Shaefer, C. and W. Messer.** 1991. DnaA protein/DNA interaction. Modulation of the recognition sequence. Mol. Gen. Genet. **226**:34-40.

146. **Sharma, S., T. F. Stark, W. G. Beattie, and R. E. Moses.** 1986. Multiple control elements for the *uvrC* gene unit of *Escherichia coli*. Nucleic Acids Res. **14**:2301-2318..

147. **Slettan, A., K. Gebhardt, E. Kristiansen, N. K. Birkeland, and B. H. Lindqvist.** 1992. *Escherichia coli* K-12 and B contain functional bacteriophage P2 *ogr* genes. J. Bacteriol. **174**:4094-4100.

148. **Somers, W. S., and S. E. Phillips.** 1992. Crystal structure of the *met* repressor-operator complex at 2.8 A resolution reveals DNA recognition by beta-strands. Nature. **359**:387-93.

149. **Sorensen, K. I., and B. Hove-Jensen.** 1996. Ribose catabolism of *Escherichia coli*: characterization of the *rpiB* gene encoding ribose phosphate isomerase B and of the *rpiR* gene, which is involved in regulation of *rpiB* expression. J. Bacteriol. **178**:1003-1011.

150. **Spiro, S.** 1994. The FNR family of transcriptional regulators. Antonie van Leewenhoek. **66**:23-36

151. **Spiro, S. and J. R. Guest.** 1991. Adaptive responses to oxygen limitation in *Escherichia coli*. Trends Biochem Sci. 310-314.

152. **Streaker, E. D., and D. Beckett.** 1998. A map of the biotin repressor-biotin operator interface: binding of a winged helix-turn-helix protein dimer to a forty base-pair site. J. Mol. Biol. **278**:787-800.

153. **Sunnerhagen, M., M. Nilges, G. Otting, and J. Carey.** 1997. Solution structure of the DNA-binding domain and model for the complex of multifunctional hexameric arginine repressor with DNA. Nat. Struct. Biol. **4**:819-825.

154. **Sutton. M. D. and J. M. Kaguni. 1997.** The *Escherichia coli dnaA* gene: Four functional domains. J. Mol. Biol. **274**:546-561.

155. **Suzuki, M.** 1995. DNA recognition by a ß-sheet. Protein Engineering. **8**: 1-4.

156. **Tanabe, H., K. Yamasaki, A. Katoh, S. Yoshioka, and R. Utsumi.** 1998. Identification of the promoter region and the transcriptional regulatory sequence of the *evgAS* operon of *Escherichia coli*. Biosci Biotechnol Biochem. **62**:286-290.

157. **Tiedeman, A. A., J. Keyhani, J. Kamholz, H. A. Daum, J. S. Gots, and J. M. Smith.** 1989. Nucleotide sequence analysis of the *purEK* operon encoding 5'-phosphoribosyl-5-aminoimidazole carboxylase of *Escherichia coli* K-12. J. Bacteriol. **171**:205-212.

158. **Titgemeyer, F., J. Reizer, A. Reizer, and M. H. Saier, Jr.** 1994. Evolutionary relationships between sugar kinases and transcriptional repressors in bacteria. Microbiology. **140**:2349-2354

159. **Thieffry, D., A. Huerta, E. Pérez-Rueda, and J. Collado-Vides.** 1998. From specific gene regulation to genomic networks: a global analysis of transcriptional regulation in *Escherichia coli*. BioEssays. **20**:433-440.24.

160. **Thlivieris, A. T., and D. W. Mount.** 1992. Genetic identification of the DNA binding domain of *Escherichia coli* LexA protein. Proc. Natl. Acad. Sci. 1992. **89**:4500-4504.

161. **Thompson, J. D., D. G. Higgins, and T. J. Gibson.** 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, positions-specific gap penalties and weight matrix choice. Nucleic Acids Res. **22**:4673-4680.

162. **Tomb, J. F., O. White, A. R. Kerlavage, R. A. Clayton, G. G. Sutton, R. D. Fleischmann RD, K. A. Ketchum, H. P. Klenk, S. Gill, B. A. Dougherty, K. Nelson, J. Quackenbush, L. Zhou, E. F. Kirkness, S. Peterson, B. Loftus, D.**

Richardson, R. Dodson, H. G. Khalak, A. Glodek, K. McKenney, L. M. Fitzegerald, N. Lee, M. D. Adams, and J. C. Venter. 1997. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. Nature. **388**:539-547.

163. **Trempy, J. E., J. E. Kirby, and S. Gottesman.** 1994. Alp suppression of Lon: dependence on the *slpA* gene. J. Bacteriol. **176**:2061-2067.

164. **Tsai-Wu, J. J., Radicella, J. P., and A. L. Lu.** 1991. Nucleotide sequence of the *Escherichia coli micA* gene required for A/G-specific mismatch repair: identity of *micA* and *mutY*. J. Bacteriol. **173**:1902-10.

165. **Ueguchi, C., C. Seto, T. Suzuki, and T. Mizuno.** 1997. Clarification of the dimerization domain and its functional significance for the *Escherichia coli* nucleoid protein H-NS. J. Mol. Biol. **274**:145-51.

166. **Van Rooijen, R. T. J., K. J. Dechering, C. Niek, J. Wilmink, and W. M. de Vos.** 1993. Lysines 72, 80 and 213 and aspartic acid 210 of the *Lactococcus lactis* LacR repressor are involved in the response to the inducer tagatose-6-phosphate leading to induction of *lac* operon expression. Protein Engineering. **6**:201-206.

167. **Walkenhorst, H. M., S. K. Hemschemeier, and R. Eichenlaub.** 1995. Molecular analysis of the molybdate uptake operon, *modABCD*, of *Escherichia coli* and *modR*, a regulatory gene. Microbiol Res. 150:347-361.

168. **Wanner, B. L.** 1986. Novel regulatory mutants of the phosphate regulon in *Escherichia coli* K-12. J Mol Biol. **191**:39-58.

169. **Weber, I. T., and T. A. Steitz.** 1987. Structure of a complex of catabolite gene activator protein and cyclic AMP refined at 2.5-A resolution. J. Mol. Biol. **198**:311-326.

170. **Weickert, M. J., and S. Adhya**, 1992. A family of bacterial regulators homologous to Gal and Lac represssors. J. Biol. Chem. **267**:15869-15874.

171. **Weiner, L., J. L. Brissette, and P. Model.** 1991. Stress-induced expression of the *Escherichia coli* phage shock protein operon is dependent on sigma 54 and modulated by positive and negative feedback mechanisms. Genes Dev. **5**:1912-1923.

172. **Wilson, K. P., L. M. Shewchuk, R. G. Brennan, A. J. Otsuka, and B. W. Matthews.** 1992. *Escherichia coli* biotin holoenzyme synthetase/*bio* repressor crystal structure delineates the biotin- and DNA-binding domains. Proc. Natl. Acad. Sci. **89**:9257-9261.

173. **Wu, H., K. L. Tyson, J. A. Cole, and S. J. Busby.** 1998. Regulation of transcription initiation at the *Escherichia coli nir* operon promoter: a new mechanism to account for co-dependence on two transcription factors. Mol. Microbiol. **27**:493-505.

174. **Yamanaka, K., L. Fang, and M. Inouye.** 1998. The CspA family in *Escherichia coli*: multiple gene duplication for stress adaptation. Mol. Microbiol. **27**:247-255

175. **Yamanaka, K., and M. Inouye.** 1997.Growth-phase-dependent expression of *cspD*, encoding a member of the CspA family in *Escherichia coli*. J. Bacteriol. **179**:5126-30

176. **Yamanaka, K., T. Mitani, T. Ogura, H. Niki, and S. Hiraga.** 1994. Cloning, sequencing, and characterization of multicopy suppressors of a *mukB* mutation in *Escherichia coli*. Mol Microbiol. **13**:301-12

177. **Yamashita, M., H. Azakami, N. Yokoro, J. H. Roh, H. Suzuki, H. Kumagai, and Y. Murooka.** 1996. *maoB*, a gene that encodes a positive regulator of the monoamine oxidase gene (*maoA*) in *Escherichia coli*. J. Bacteriol. **178**:2941-2947

178. **Yang, B., and T. J. Larson.** 1998. Multiple promoters are responsible for transcription of the *glpEGR* operon of *Escherichia coli* K-12.Biochim Biophys Acta. **1396**:114-126.

179. **Yudkin, M. D.** 1987. The prediction of helix-turn-helix DNA-binding regions in proteins. Prot. Enginnering. **1**:371-372.

180. **Zhang, R-G., A. Joachimiak, C. L. Lawson, R. W. Schevitz, Z. Otwinowski, and P. B. Sigler.** 1987. The crystal structure of *trp* aporepressor at 1.8Å shows how binding tryptophan enhances DNA affinity. Nature. **327**:591-597.

Tables and Figures

Score Interval	Probability	Proteins
4.51-7.76	1.0	102
4.01-4.48	0.9	27
3.52-3.99	0.71	36
3.01-3.48	0.5	47
2.51-3.0	0.25	71

Table 1. Proteins with HTH by Dodd and Egan method.

Proteins with HTH predicted by the method of Dodd and Egan. HTH proteins are grouped in 5 score intervals. Scores below 2.51 are not reported. High probability values close to 1 are the most likely HTH. The LacI repressor, for instance, has a HTH with score of 6.64 and a probability of 1.

Family	Pattern	PROSITE-ID	Proteins
AraC/XylS	[KRQ]-[LIVMA]-x(2)-[GSTALIV]-{FYWPGDN}-x(2)-[LIVMSA]-x(4,9)-[LIVMF]-x(2)-[LIVMSTA]-{GSTA CIL]-x(3)-[GANQRF]-[LIVMFY]-x(4,5)-[LFY]-x(3)-[FYIVA]-{FYWHCM}-x(3)-[GSADENQKR]-x-[NSTAPKL]-[PARL]	PS00041 PS01124	22
ArsR	C-x(2)-D-[LIVM]-x(6)-[ST]-x(4)-S-[HYR]-[HQ]	PS00846	1
Cold-shock	[FY]-G-F-I-x(6,7)-[DER]-[LIVM]-F-x-H-x-[STKR]-x-[LIVMFY]	PDOC00304 PS00352	9
AsnC	[GSTAP]-x(2)-[DNEA]-[LIVM]-[GSA]-x(2)-[LIVMFY]-[GN]-[LIVMST]-[ST]-x(6)-R-[LVT]-x(2)-[LIVM]-x(3)-G	PS00519	3
CRP	[LIVM]-[STAG]-[RHNW]-x(2)-[LIM]-[GA]-x-[LIVMFYA]-[LIVSC]-[GA]-x-[STACN]-x(2)-[MST]-x-[GSTN]-R-x-[LIVMF]-x(2)-[LIVMF]	PS00042	2
DeoR	R-X(3)-[LIVM]-X(3)-[LIVM]-X(16,17)-[STA]-X(2)-T-[LIVMA]-[RH]-[KRNA]-D-[LIVMF]	PS00894	12
DnaA	I-[GA]-x(2)-[LIVMF]-[SGDNK]-x(0,1)-[KR]-x-H-[STP]-[STV]-[LIVM](2)-x-[SA]-x(2)-[KRE]-[LIVM]	PDOC00771 PS01008	1
EBP	[LIVMFY](3)-x-G-[DEQ]-[STE]-G-[STAV]-G-K-x(2)-[LIVMFY]	PS00675	13
GntR	[LIVAPKR]-[PILV]-x-[EQTIVMR]-x(2)-[LIVM]-x(3)-[LIVMFYK]x-[LIVFT]-[DNGSTK]-[RGTLV]-x-[STAIVP]-[LIVA]-x(2)-[STAGV]-[LIVMFYH]-x(2)-[LMA]	PS00356	15
IclR	[GA]-X(3)-[DS]-X(2)-E-X(6)-[CSA]-[LIVM]-[GSA]-X(2)-[LIVM]-[FYH]	PS01051	6
GalR/LacI	[LIVM]-x-[DE]-[LIVM]-A-x(2)-[STAGV]-x-V-[GSTP]-x(2)-[STAG]-[LIVMA]-x(2)-[LIVMFYAN]-[LIVMC]	PS00356	13
LuxR/UhpA	[GDC]-x(2)-[NSTAVY]-x(2)-[IV]-[GSTA]-x(2)-[LIVMFYWCT]-x-[LIVMFYWCR]-x(3)-[NST]-[LIVM]-x(5)-[NRHSA]-[LIVMSTA]-x(2)-[KR]	PS00622	17
LysR	[NQKRHSTAG]-[LIVMFYTA]-x(2)-[STAGLV]-[STAG]-x(4)-[LIV39MYCTQR]-[PSTANLVER]-x-[PSTAGQV]-[PSTAGNVMF]-[LIVMFA]-[STAGH]-x(2)-[LIVMF]-x(2)-[LIVMFW]-[RKEAV]-x(2)-[LIVMFYNTAE]-x(3)-[LIMVT]	PS00044	39
MarR	[STNA]-[LIA]-x-[RNGS]-x(4)-[LM]-[EIV]-x(2)-[GES]-[LFYW]-[LIVC]-x(7)-[DN]-[RKQG]-[RK]-x(6)-T-x(2)-[GA]	PS01117	3
MerR	[GSA]-x-[LIVMFA]-[ASM]-x(2)-[STACLIV]-[GSDENQR]-[LIVC]-[STANHK]-x(3)-[LIVM]-[RHF]-x-[YW]-[DEQ]-x(2,3)-[GHDNQ]-[LIVMF]	PS00552	4
NagC/XylR	[LIVM]-x(2)-G-[LIVMFCT]-G-x-[GA]-[LIVMFA]-x(8)-G-x(3,5)-[GATP]-x(2)-G-[RKH]	PDOC00866 PS01125	4
TetR/AcrR	G-[LIVMFYS]-x(2,3)-[TS]-[LIVMT]-x(2)-[LIVM]-x(5)-[LIVQS]-[STAGENQH]-x-[GPAR]-x-[LIVMF]-[FYST]-x-[HFY]-[FV]-x-[DNST]-K-x(2)-[LIVM]	PS01081	7

Table 2. Prosite patterns used to detect regulatory proteins in the *E. coli* genome.

Proteins with some experimental evidence:	159	
Total set of predicted regulators:		155
Set with homologous to regulators with experimental evidence :		71
Proteins detected by one or more of the four methods (See Table 4)	80	

Table 3. The dataset of 314 known and predicted regulatory proteins.

Protein	Dodd-Egan	Blattner	Collection	Gibbs	Prosite
29	+	-	-	-	-
23	-	+	-	-	-
13	+	+	-	-	-
5	-	-	+	-	-
39	-	+	+	-	-
13	+	+	+	-	-
2	-	-	-	-	+
15	-	+	-	-	+
7	+	+	-	-	+
6	-	-	+	-	+
12	-	+	+	-	+
17	+	+	+	-	+
9	-	-	-	+	-
9	+	-	-	+	-
2	-	+	-	+	-
10	+	+	-	+	-
2	-	+	+	+	-
4	+	+	+	+	-
2	+	-	-	+	+
4	-	+	-	+	+
28	+	+	-	+	+
2	+	-	+	+	+
14	-	+	+	+	+
49	+	+	+	+	+
314	191	252	**163**	133	158

Table 4. Comparison of transcriptional factor predictions with different approaches.

314 protein factors were detected using several approaches. Sets of proteins found by a given method are indicated by as (+) and not found by a (-). Dodd and Egan and Gibbs sampler methods, generate roughly similar sets of predictions. On the other hand, Prosite and Blattner annotations are mutually consistent. Annotations by Blattner were based initially on a Smith-Waterman overall analysis and ORF assignment in the genome, and the functional annotation was a mixture of literature search and the method of remote similarities by Labedan and Riley (1995).

Method	% true	% false positives
Dodd and Egan	65.0	15.0
Gibbs	50.7	30.0
Prosite	67.9	9.3
Blattner annotations	97.6	21.0

Table 5. Evaluation of the dataset.

We used the dataset of 163 experimentally supported proteins to evaluate the performance of the methods. The evaluations of Dodd and Egan, and the Gibbs methods were done comparing with the set of 128 proteins that have reported a HTH motif.

Table 6...

Protein	B-number	Role	Family	Motif	Start	End	Position	Length	Description	Function	Condition	Operon	References
AcrR	b0464	-	TetR/AcrR	H-T-H	33	52	19.7	215	Known	Multidrug efflux pump	General stress	*acrAB*	95
Ada	b2213	+/-	AraC/XylS	H-T-H	100	121	31.2	354	Known	DNA repair	Alkylated guanine in DNA	*Ada*	48
AdiY	b4116	+	AraC/XylS	H-T-H	165	184	68.9	253	Known				3, 15
AgaR	b3131	-	DeoR	H-T-H	32	51	15.4	269	Known	N-acetyl galactosamine metabolism		*Aga*	126
AlpA	b2624	+	Unclass	H-T-H	12	31	30.7	70	Known	Prophage CP4-57 regulatory protein		*slpA*	162
AppY	b0564	+/-	AraC/XylS	H-T-H	149	168	65.2	243	Known	Acid phosphatase synthesis	Deceleration phase of growth	*appA/appBC*	48
AraC	b0064	+/-	AraC/XylS	H-T-H	197	216	70.7	292	Known	Arabinose catabolism		*araBAD and FGH,araJ ,araE*	48
ArcA	b4401	+/-	Two	H-T-H	181	223	84.8	238	Known	Aerobic respiration control		*pfl/aceBAK*	99
ArgR	b3237	-	ArgR	H-T-H				156	Known	Arginine synthesis	L-arg concentration	*carAB/argCBH(and 3 more)*	152
ArsR	b3501	-	ArsR	H-T-H	33	52	36.3	117	Known	Arsenical resistance		*arsEFG*	8
AscG	b2714	-	GalR/LacI	H-T-H	4	23	4.1	334	Known	Carbon catabolism		*ascBF*	57
AslB	b3800	+	Unclass					411	Known	Regulation of arylsulfatase			106
AsnC	b3743	+/-	AsnC	H-T-H	25	44	22.6	152	Known	Asparagine biosynthesis		*asnA/gidA/ asnC*	158
AtoC	b2220	+	EBP	H-T-H	433	452	95.9	461	Known	Acetoacetate metabolism		*atoDAB*	22
b0298	b0298	?	Unclass	H-T-H	33	54	42.6	102	Putative				
b0373	b0373	?	Unclass	H-T-H	33	54	42.6	102	Putative				
b0502	b0502	?	Unclass	H-T-H	35	56	33.7	135	Putative				
b0540	b0540	?	Unclass	H-T-H	33	54	42.6	102	Putative				
b0845	b0845	?	Unclass					402	Putative				
b0846	b0846	?	TetR/AcrR					178	Putative				
b1027	b1027	?	Unclass	H-T-H	33	54	42.6	102	Putative				
b1146	b1146	+	Unclass	H-T-H	112	133	73.3	167	Putative				
b1201	b1201	+	EBP	H-T-H	608	629	96.3	642	Putative				
b1284	b1284	-	DeoR	H-T-H	18	39	11.4	249	Putative				
b1422	b1422	+/-	LysR	H-T-H	73	94	23.5	354	Putative				
b1438	b1438	+	Unclass	H-T-H	98	119	74.8	145	Putative				
b1439	b1439	?	GntR					468	Putative				
b1450	b1450	?	GntR					240	Putative				
b1499	b1499	-	AraC/XylS	H-T-H	42	63	20.7	253	Putative				
b1649	b1649	?	TetR/AcrR					199	Putative				
b1696	b1696	?	AraC/XylS					303	Putative				
b1747	b1747	?	Unclass					344	Putative				
b1770	b1770	?	DeoR					252	Putative				
b1978	b1978	-	Unclass	H-T-H	102	123	4.7	2383	Putative				
b2088	b2088	?	Unclass	H-T-H	33	54	42.6	102	Putative				
b2248	b2248	?	IclR					260	Putative				
b2382	b2382	+	AraC/XylS	H-T-H	198	219	73.1	285	Putative				
b2531	b2531	-	Unclass	H-T-H	26	47	22.5	162	Putative				
b2667	b2667	?	Unclass	H-T-H	39	60	50	99	Putative				
b2981	b2981	-	Unclass	H-T-H	69	90	20.3	390	Putative				
b3021	b3021	+	Unclass	H-T-H	83	104	71.3	131	Putative				
b3694	b3694	?	GntR					98	Putative				
BaeR	b2079	+	Two	H-T-H	181	213	82.1	240	Known	RNA synthesis, modification, DNA transcription			99
BasR	b4113	+	Two	H-T-H	170	202	83.7	222	Known	RNA synthesis, modification, DNA transcription			99
BetI	b0313	-	TetR/AcrR	H-T-H	31	50	20.7	195	Known	Osmoregulatory choline-glycine betaine		*betIBA*	158

Table 6...

									pathway			
BirA	b3973	-	BirA	H-T-H	22	41	9.8	321	Known	Biotin synthesis	*bio*	151
BolA	b0435	+	BolA	H-T-H	37	46	35.7	116	Known	Involved in morphogenetic pathway	*murein genes*	1
CadC	b4133	+	Two	H-T-H	48	80	12.5	512	Known	pH-regulated	*cadAB*	99
CaiF	b0034	+	Unclass					166	Known		*cai operon*	38
Cbl	b1987	+/-	LysR	H-T-H	19	39	9.1	316	Known	Cysteine biosynthesis	*cys regulon*	68
CelD	b1735	-	AraC/XylS	H-T-H	230	280	91.1	280	Known	Carbon uptake	*celABCF*	48
ChpA	b2782	-	PemK					111	Known	Regulation of cell growth	*chpRA*	100
ChpB	b4225	-	PemK					116	Known	Regulation of cell growth	*chpSB*	100
CpxR	b3912	+	Two	H-T-H	177	209	83.1	232	Known			99
CreB	b4398	+	Two	H-T-H	175	207	83.4	229	Known	Catabolic regulation	*creC/creB*	99
CriR	b0620	+	Two	H-T-H	180	199	83.8	226	Known	Citrate fermentation	*citCDEFG & citS-oadGAB-citAB*	99
Crl	b0240	+	Unclass					132	Known	Cryptic genes for curli formation and fibronectin binding	Temperature-regulated (26°C)	*csgA*	4
Crp	b3357	+/-	CRP	H-T-H	170	189	85.4	210	Known	Catabolite-sensitives	Catabolite repression		79, 149
CsgD	b1040	+	LuxR/UhpA	H-T-H	173	192	84.4	216	Known	Curli formation and fibronectin binding		*csgBA*	57
CspA	b3556	+	Cold	CSD				69	Known	Cold-shock	Low temperatures	*hns*	53, 88, 173
CspB	b1557	+	Cold	CSD				71	Known	Cold-shock	Low temperatures		54, 173
CspC	b1823	+	Cold	CSD				68	Known	Cold-shock	Low temperatures		88, 173
CspD	b0880	+	Cold	CSD				74	Known		Nutritional deprivation		88, 173
CspE	b0623	+	Cold	CSD				68	Known	Cold-shock	Low temperatures		173
CspF	b1558	+	Cold	CSD				70	Known	Cold-shock	Low temperatures		88
CspG	b0990	+	Cold	CSD				70	Known	Cold-shock	Low temperatures		174
CspH	b0989	+	Cold	CSD				70	Putative	Cold-shock	Low temperatures		174
CspI	b1552	+	Cold	CSD				70	Known	Cold-shock	Low temperatures		173
CynR	b0338	+/-	LysR	H-T-H	18	37	9.1	299	Known	Cyanate		*cynTSX*	60
CysB	b1275	+/-	LysR	H-T-H	19	38	8.7	324	Known	Biosynthesis of L-cysteine		*cys regulon*	60, 139
CytR	b3934	-	GalR/LacI	H-T-H	12	31	6.3	341	Known			*deoCABD,udp,cdd and nupC, nupG, and tsx*	169
DeoR	b0840	-	DeoR	H-T-H	22	41	12.5	252	Known	Nucleotide and deoxyribonucleotide catabolism		*Deoxyribose operon*	158
DgoR	b3695	-	GntR	H-T-H	31	50	22.8	177	Putative	D-galactonate catabolism		*dgoRKAT*	
DicA	b1570	-	Unclass	H-T-H	23	43	24.4	135	Known	Division cell inhibition	Temperature-sensitive	*dicC*	12
DicC	b1569	-	Unclass	H-T-H	13	33	30.2	76	Known	Division inhibition			12
DnaA	b3702	+/-	DnaA	H-L-H	373	427	85.6	467	Known	Chromosomal replication initiator		*dam,rpoH,ftsA,mioC ,nrdA*	158
DniR	b0211	+	Unclass	H-T-H	349	370	79.5	452	Putative	Nitrite reductase (cytochrome c552) regulation			
DsdC	b2364	+/-	LysR	H-T-H	32	51	16.9	245	Known	D-serine deaminase		*dsdXA*	158
EbgR	b3075	-	GalR/LacI	H-T-H	4	23	4.1	327	Known	Repressor of β galactosidase		*Eb gA and ebgC*	169
EnvR	b3264	-	TetR/AcrR	H-T-H	33	52	19.3	220	Known			*acrEF/envCD*	54
EnvY	b0566	+	AraC/XylS	H-T-H	165	184	68.9	253	Known	Porin thermoregulatory	Temperature-dependent expression	*ompF & ompC*	48
EvgA	b2369	+	LuxR/UhpA	H-T-H	161	180	83.5	204	Known			*evgAS*	155
ExuR	b3094	-	GntR	H-T-H	40	59	18.8	263	Known	Carbon degradation		*exu regulon (exuT, uxaAC, & uxuB)*	120
FadR	b1187	+/-	GntR	H-T-H	34	53	18.2	239	Known	Fatty acid metabolism		*fad regulon & fabA*	59
FarR	b0730	-	GntR	H-T-H	32	51	17.2	240	Known	TCA cycle regulator		*succinylCoA synthetase operon*	59
FeaR	b1384	+	AraC/XylS	H-T-H	216	235	74.9	301	Known	2-phenylethylamine catabolism	Tyramine	*maoA*	48
FhlA	b2731	+	EBP	H-T-H	663	682	97.1	692	Known	Induction of expression of the formate		*fdhF, hyc & Putative*	104

Table 6...

										dehydrogenase H		*operons*	
FimZ	b0535	+	LuxR/UhpA	H-T-H	188	207	85.4	231	Known	Fimbrial Z protein; signal transducer			118
Fis	b3261	+/-	EBP	H-T-H	74	93	85.2	98	Known	Factor-for-inversion stimulation		*rRNA promoters fis*	58
FlhC	b1891	+	σ70					192	Known	Flagellar class II synthesis		*fliA &fliL*	90, 91
FlhD	b1892	+	σ70					119	Known	Flagellar class II synthesis		*fliA &fliL*	90, 91
Fnr	b1334	+/-	CRP	H-T-H	197	216	82.6	250	Known	Aerobic, anaerobic respiration, Osmotic balance		*ansBp2*	149, 150
FruR	b0080	+/-	GalR/LacI	H-T-H	3	22	3.7	334	Known	Operons encoding enzymes which Comprise central		*aceBAK*	169
FrvR	b3897	-	Unclass	H-T-H	20	39	5.1	582	Known	Sugar phosphotransferase system		*frv*	127
FucR	b2805	+	DeoR	H-T-H	19	38	11.7	243	Known	L-Fucose utilization		*(regulon) fucO-fucA-fucPIK-fucR*	25
Fur	b0683	-	Fur	H-T-H	114	129	82.1	148	Known	Iron transport		*cir*	40
GalR	b2837	-	GalR/LacI	H-T-H	4	23	3.9	343	Known	Galactose utilization		*galETKM*	169
GalS	b2151	-	GalR/LacI	H-T-H	4	23	3.9	346	Known	Galactose utilization	Galactose and fucose inducer	*mgl*	169
GatR	b2087 b2090	-	DeoR	H-T-H	22	41	12.1	259	Known	Galactitol metabolism		*gatYZABCDR*	113
GcvA	b2808	+/-	LysR	H-T-H	23	42	10.6	305	Known	Cleavage of glycine		*gcvA*	158
GcvR	b2479	-	Unclass					77	Known	Cleavage of glycine	Induced by glycine and repressed by purines		48, 49
GlcC	b2980	+	GntR	H-T-H	34	53	17.1	254	Known	Glycolate utilization		*glcDEFG.glcGB*	121
GlpR	b3423	-	DeoR	H-T-H	20	39	11.7	252	Known	Glycerol-3-phosphate		*glpEGR/glpACB*	177
GntR	b3438	-	GalR/LacI	H-T-H	8	27	5.5	313	Known	Gluconate utilization		*gnt-I. gntUKR.*	120
GutM	b2706	+	Unclass	H-T-H	23	29	21.8	119	Known	Glucitol utilization		*srlAABD-gutM-srlR-gutQ*	158
GutR	b2707	-	DeoR	H-T-H	20	39	11.4	257	Known	Glucitol utilization		*srlAABD-gutM-srlR-gutQ*	158
HipA	b1507	-	Unclass					440	Known	Frequency of persistence		*hip*	13
HipB	b1508	-	Unclass	H-T-H	28	47	42.6	88	Known				13
Hns	b1237	-	Histone-like					136	Known	Increases DNA thermal stability and inhibits transcription		*hns*	47
HtgA	b0012	+	Unclass					196	Known	Heat shock gene			103
HybF	b2991	-	Unclass					113	Known	Modulate levels of hydrogenease-2			96
HycA	b2725	+	Unclass					153	Known	Formate hydrogenylase system		*hyc and Putative*	139
HydG	b4004	+	EBP	H-T-H	421	440	97.6	441	Known	hydrogenase 3 activity regulator			104
HyfR	b2491	+	EBP	H-T-H	641	660	97.1	670	Known	Induction of expression of the hydrogenase-4 genes.			104
HypF	b2712	-	HypF	Zn-finger	109	184	19.5	750	Known	Hydrogenase expression			129
IciA	b2916	+/-	LysR	H-T-H	21	40	10.2	297	Known	Specific inhibitor of chromosomal initiation			60
IclR	b4018	-	IclR	H-T-H	34	54	16.1	274	Known	Glyoxylate bypass	Carbon source	*aceBAK*	54
IhfA/B	b0912	+/-	HI-HN-S					94	Known	Host factor, lysogenic life		*carAB*	158
IhfA/B	b1712	+/-	HI-HN-S					99	Known	Host factor, lysogenic life		*carAB*	158
IlvY	b3773	+/-	LysR	H-T-H	18	37	9.2	297	Known	Isoleucine, Valine synthesis		*ilvc*	60, 139
KdgR	b1827	-	IclR	H-T-H	32	53	16.1	263	Putative	Galacturonic acid		*kdgRKAT*	124
KdpE	b0694	+	Two	H-T-H	170	204	83.1	225	Known	Kdp transcriptional regulatory protein *kdpE.*		*kdpABC*	99
LacI	b0345	-	GalR/LacI	H-T-H	6	25	4.3	360	Known	Lactose catabolism		*lacZYA*	168, 58
LeuO	b0076	+	LysR	H-T-H	39	58	16.7	290	Known	Leucine biosynthesis		*leuABCD*	60, 139
LexA	b4043	-	LexA	H-T-H	28	48	18.8	202	Known	DNA damage (SOS response)	DNA damage	*sos(lexA) regulon/cloDF13*	159
LldR	b3604	-	GntR	H-T-H	34	53	16.8	258	Known	L-Lactate utilization		*lctDRP*	37
LrhA	b2289	-	LysR	H-T-H	28	47	12.2	305	Known	Proton translocating nadh dehydrogenase genes		*lrhA*	15
Lrp	b0889	+/-	AsnC	H-T-H	31	50	24.6	164	Known	Global response to leucine and high-affinity		*lrp regulon*	158

Table 6...

										branched-chain amino acid transport system		
LysR	b2839	+/-	LysR	H-T-H	21	40	9.8	311	Known	Diaminopimelate decarboxylase; Lysine biosynthesis	*lysA*	60, 139
MalI	b1620	-	GalR/LacI	H-T-H	9	28	5.6	325	Known	Maltose regulon	*malX/malY*	169
MalT	b3418	+	LuxR/UhpA	H-T-H	853	872	95.7	901	Known	Maltose utilization	*maltose regulon*	118
MarA	b1531	+	AraC/XylS	H-T-H	30	49	30.6	129	Known	Multiple antibiotic resistance	*sodA, zwf,micF,fpr*	48
MarR	b1530	-	MarR					125	Known	Antibiotic resistance and oxidative stress	*marRAB*	158
MelR	b4118	+	AraC/XylS	H-T-H	210	229	72.6	302	Known	Melibiose regulatory protein	*melAB*	48
MetJ	b3938	-	MetJ	β-sheet	21	67	42.3	104	Known	Methionine biosynthesis	*metJ,methionine regulon*	58, 147
MetR	b3828	+/-	LysR	H-T-H	19	38	8.9	317	Known	Methionine biosynthesis	*metE and metH*	60, 139
MhpR	b0346	+	IclR	H-T-H	72	91	25.8	315	Known	3-Hydroxyphenylpropionate degradation	*mhpRABCDEF*	42
Mlc	b1594	-	NagC/XylR	H-T-H	33	42	9.2	406	Known	Glucose uptake or glycolisis	*ptsG*	75
ModE	b0761	-	Unclass					262	Known	Molybdate uptake	*modABCD*	166
MprA	b2684	-	MarR					176	Known	Multidrug resistance pump	*emrAB/mcbABCDEFG*	92
MtlR	b3601	-	Unclass					195	Known	Mannitol	*mtlA-mtlD,mtlR*	43
MviN	b1069	?	Unclass	H-T-H	229	250	46.8	511	Putative	Putative transcriptional regulator		
Nac	b1988	+/-	LysR	H-T-H	18	37	9.0	305	Known	Histidine utilization/nitrogen assimilation	Nitrogen limitation	*hut,gdh,ure,put,glnB*	107
NadR	b4390	-	Unclass	H-T-H	23	44	8.1	410	Known	Novo synthesis of NAD	NAD levels	*nadA-B-pncB*	46
NagC	b0676	-	NagC/XylR	H-T-H	35	44	9.7	406	Known	N-acetylglucosamine operon		*nagEBACD; glmSU*	157
NarL	b1221	+/-	LuxR/UhpA	H-T-H	173	192	84.4	216	Known	Reductase (*narGHJI*) and formate dehydrogenase-n (*fdnGHI*)s. Anaerobic respiration	Nitrate/Nitrite induction	*frdABCD/adhE*	118
NarP	b2193	+/-	LuxR/UhpA	H-T-H	171	190	83.9	215	Known	Anaerobic respiratory	Nitrate and nitrite regulated	*frdABCD/narGHJIf dnGHI/aeg*	125
NhaR	b0020	+	LysR	H-T-H	23	42	12.4	262	Known	Transport of cations		*nhaA*	60
Nlp	b3188	+	Unclass	H-T-H	50	69	64.6	92	Known	Transposable coliphages			5
NtrC	b3868	+/-	EBP	H-T-H	445	464	96.9	469	Known	Nitrogen assimilation		*argT*	104
OgrK	b2082	+	Unclass					72	Known	Cryptic		*ogr*	146
OmpR	b3405	+/-	Two	H-T-H	181	223	84.5	239	Known	Outer membrane protein synthesis		*ompF/ompC fadL*	99
OsmE	b1739	?	Unclass	H-T-H	43	64	47.7	112	Putative	Osmotically inducible		*ntrL gene*	55
OxyR	b3961	+/-	LysR	H-T-H	18	37	9.0	305	Known	Hydrogen peroxide-inducible genes		*oxyR*	60
PdhR	b0113	-	GntR	H-T-H	37	56	18.3	254	Known	Pyruvate dehydrogenase complex		*aceE*	59
PerR	b0254	-	LysR	H-T-H	24	44	11.4	297	Known	Peroxide resistance		*peroxide regulon*	21
PhnF	b4102	-	GntR	H-T-H	38	57	19.7	241	Known	Alkylphosphonate			60
PhoB	b0399	+	Two	H-T-H	175	207	83.4	229	Known	Phosphate	Phosphate limited	*pho regulon /phnC*	99
PhoP	b1130	+	Two	H-T-H	170	202	83.4	223	Known				99
PhoU	b3724	-	Unclass					241	Known		Phosphate limited	*phosphate regulon*	167
PrpD	b0334	?	Unclass	H-T-H	178	199	39.0	483	Putative				
PrpR	b0330	+	EBP	H-T-H	508	527	98.0	528	Known	Propionate catabolism		*prpBCDE*	65
PspC	b1306	+	Unclass					119	Known	Phage-shock protein	Stresses	*pspABCDE*	170
PspF	b1303	+	EBP	H-T-H	302	321	95.8	325	Known	Phage-shock protein	Heat,ethanol,osmotic shock,infection by filamentous bacteriophages	*pspABCDE*	71
PssR	b3763	+/-	LysR	H-T-H	18	37	20.6	133	Known	Phosphatidylserine synthetase			30
PurR	b1658	-	GalR/LacI	H-T-H	4	23	3.9	341	Known	De novo purine nucleotide synthesis		*purB,purC,purEK,purHD,purL,purMN,guaBA,glyA,glnB,prsA,speA,codBA*	169
PutA	b1014	-	PutA					1320	Known	Proline synthesis		*putAB*	158
RbsR	b3753	-	GalR/LacI	H-T-H	3	22	3.7	329	Known	Ribose metabolism	Ribose inducer	*rbsDACBK*	101

Table 6...

RcsA	b1951	+	LuxR/UhpA	H-T-H	155	174	79.4	207	Known	Capsular polysaccharide synthesis		61
RcsB	b2217	+	LuxR/UhpA	H-T-H	168	187	82.1	216	Known	Capsular polysaccharide synthesis	*ftsZ,cpsGB*	118
RfaH	b3842	+	Unclass					162	Known	Synthesis, assembly and export of the lipopolysaccharide core, Exopolysaccharide, f conjugation pilus and haemolysin toxin	*rfaD/htrM,htrL,hlyC ADB*	7
RhaR	b3906	+	AraC/XylS	H-T-H	225	244	75.1	312	Known	L-Rhamnose catabolism	*rhaRS*	48
RhaS	b3905	+	AraC/XylS	H-T-H	190	209	71.7	278	Known	L-Rhamnose catabolism	*rhaBAD*	48
Rob	b4396	+	AraC/XylS	H-T-H	24	43	11.5	289	Known	Replication	*inaA*	48
RpiR	b4089	-	RpiR/YebK/YfhH	H-T-H	50	69	20.1	296	Known	Ribose catabolism	*rpiB*	148
RstA	b1608	+	Two	H-T-H	185	217	83.1	242	Known	Member of the two-component regulatory system rstb/rsta.		99
SdiA	b1916	+	LuxR/UhpA	H-T-H	198	217	86.1	241	Known	Cell division	*ftsQAZ*	118
SfsA	b0146	?	Unclass					234	Putative	Maltose metabolism		73
SgcR	b4300	-	DeoR	H-T-H	22	41	12.1	260	Known	Sugar-specific permease domains of the phosphotransferase system and Pentose catabolic enzymes	*sgcREAQCX*	126
SlyA	b1642	+	MarR	H-T-H	49	70	40.7	146	Known	Haemolytic phenotype	*slyAB*	94
SoxR	b4063	+	MerR	H-T-H	14	33	15.2	154	Known	Superoxide response	Oxidative stress	*soxS*	158
SoxS	b4062	+	AraC/XylS	H-T-H	24	43	31.3	107	Known	Superoxide response	*sodA,nfo,zwf, micF,fpr*	48
TdcA	b3118	+	LysR	H-T-H	24	43	10.7	312	Known	Amino acids degradation	*tdcABC*	158
TdcR	b3119	+	Unclass					99	Known	Threonine dehydratase	*tdcABC*	158
TorR	b0995	+	Two	H-T-H	178	211	84.5	230	Known	Trimethylamine n-oxide reductase respiratory system	*torCAD*	70
TreR	b4241	-	GalR/LacI	H-T-H	7	26	5.2	315	Known	Uptake of trehalose	*treABC*	77
TrpR	b4393	-	TrpR	H-T-H	67	90	73.3	107	Known	Tryptophan biosynthesis	*trp & amH*	58
Tdk	b3641	+/-	TetR/AcrR	H-T-H	47	66	26.6	212	Known	dUTPase subunit	*dut gene*	93
TyrR	b1323	+/-	EBP	H-T-H	483	502	96.0	513	Known	Aromatic amino acid synthesis	*aroF, aroG, tyrA,and at least 8 operons else*	158
UhpA	b3669	+	LuxR/UhpA	H-T-H	155	174	83.9	196	Known	Carbon sources transport	*uhpT*	61
UidR	b1618	-	Unclass		33	52	21.6	196	Known	Carbon uptake	*uidRABC*	120
UvrY	b1914	+	LuxR/UhpA	H-T-H	167	186	80.9	218	Known		*uvrC*	145
UxuR	b4324	-	GntR	H-T-H	36	55	17.7	257	Known	Carbon uptake	*uidA,uxuBA*	120
WrbA	b1004	-	Unclass					198	Known	Tryptophan biosynthesis	*trpR*	158
XapR	b2405	+	LysR	H-T-H	24	43	11.3	294	Known	Nucleotide interconversions	*xapAB*	60
XylR	b3569	+/-	AraC/XylS	H-T-H	304	323	79.9	392	Known	Xylose utilization	*xylBAFGHR*	48
YadD	b0132	?	Unclass	H-T-H	277	298	95.8	300	Putative			
YafC	b0208	+/-	LysR	H-T-H	20	39	9.7	304	Putative			
YafY	b0251	-	DeoR	H-T-H	28	47	13.1	285	Putative			
YagA	b0267	?	Unclass	H-T-H	28	49	10.0	384	Putative			
YagI	b0272	-	IclR	H-T-H	25	45	13.8	252	Putative			
YahA	b0315	-	LuxR/UhpA	H-T-H	41	70	15.3	362	Putative			
YahB	b0316	+/-	LysR	H-T-H	22	42	10.3	310	Putative			
YaiV	b0375	?	Unclass					222	Putative			
YaiW	b0378	?	Unclass	H-T-H	277	298	78.9	364	Putative			
YajF	b0394	?	NagC/XylR					302	Putative			
YbaO	b0447	+/-	AsnC	H-T-H	21	40	20.1	152	Putative			
YbaQ	b0483	-	Unclass	H-T-H	43	64	40.8	131	Putative			
YbhI	b0487	-	MerR	H-T-H	4	23	10.0	135	Putative			
YbhS	b0504	+/-	LysR	H-T-H	19	38	9.2	308	Putative			

Table 6...

YbbU	b0506	-	IclR	H-T-H	43	62	19.3	271	Putative
YbcM	b0546	+	AraC/XylS	H-T-H	181	200	71.8	265	Putative
YbdO	b0603	?	LysR	H-T-H	25	76	16.8	300	Putative
YbeF	b0629	+/-	LysR	H-T-H	46	65	17.5	317	Putative
YbhD	b0768	+/-	LysR	H-T-H	41	61	15.1	338	Putative
YbhN	b0788	-	Unclass	H-T-H	107	128	36.9	318	Putative
YbiH	b0796	-	TetR/AcrR	H-T-H	33	52	19.0	223	Putative
YbjN	b0853	+	Unclass	H-T-H	111	132	76.8	158	Putative
YcaL	b0909	?	Unclass					268	Putative
YcaN	b0900	+/-	LysR	H-T-H	20	40	9.9	302	Putative
YcdC	b1013	-	TetR/AcrR	H-T-H	39	58	22.8	212	Putative
YcfQ	b1111	?	TetR/AcrR					236	Putative
YcfX	b1119	?	NagC/XylR					303	Putative
YcgE	b1162	-	MerR	H-T-H	6	25	6.3	243	Putative
YcjC	b1299	-	Unclass	H-T-H	21	42	17.0	185	Putative
YcjW	b1320	-	GalR/LacI	H-T-H	3	24	4.1	332	Putative
YcjZ	b1328	+/-	LysR	H-T-H	21	41	10.3	299	Putative
YdaK	b1339	+/-	LysR	H-T-H	20	41	10.1	302	Putative
YdaR	b1356	-	Unclass	H-T-H	25	46	22.4	158	Putative
YdaS	b1357	-	Unclass	H-T-H	16	37	27.1	98	Putative
YdcE	b1461	?	Unclass					77	Putative
YdcN	b1434	-	Unclass	H-T-H	21	42	17.6	178	Putative
YddM	b1477	?	Two	H-T-H	48	69	48.7	120	Putative
YdeW	b1512	-	Unclass	H-T-H	31	52	13.1	317	Putative
YdfH	b1540	-	GntR	H-T-H	36	57	20.3	228	Putative
YdfK	b1544	?	Unclass					88	Putative
YdhB	b1659	+/-	LysR	H-T-H	19	38	9.1	310	Putative
YeaM	b1790	?	AraC/XylS					273	Putative
YeaT	b1799	+/-	LysR	H-T-H	27	48	11.9	314	Putative
YebK	b1853	?	Unclass	H-T-H	35	56	15.7	289	Putative
YedW	b1969	?	Two	H-T-H	184	216	83.6	239	Putative
YeeY	b2015	+/-	LysR	H-T-H	25	76	15.9	316	Putative
YegW	b2101	-	GntR	H-T-H	50	69	23.9	248	Putative
YehV	b2127	-	MerR	H-T-H	6	25	6.3	243	Putative
YeiE	b2157	+/-	LysR	H-T-H	20	39	10.1	293	Putative
YeiL	b2163	+	Unclass	H-T-H	157	178	76.4	219	Putative
YfeC	b2398	-	Unclass	H-T-H	6	27	13.8	119	Putative
YfeD	b2399	-	Unclass	H-T-H	8	29	14.2	130	Putative
YfeG	b2437	+	AraC/XylS	H-T-H	259	278	76.7	350	Putative
YfeR	b2409	+/-	LysR	H-T-H	18	39	9.2	308	Putative
YfeT	b2427	-	Unclass	H-T-H	35	56	15.9	285	Putative
YfhA	b2554	+	EBP	H-T-H	414	433	95.1	445	Putative
YfhH	b2561	-	Unclass	H-T-H	35	56	16.1	282	Putative
YfhT	b2537	+/-	LysR	H-T-H	18	38	9.4	296	Putative
YfiE	b2577	+/-	LysR	H-T-H	18	37	12.7	215	Putative
YfjR	b2634	-	DeoR	H-T-H	28	47	16.1	233	Putative
YgaA	b2709	+	EBP	H-T-H	504	523	97.1	529	Putative
YgaE	b2664	-	GntR	H-T-H	29	48	17.5	220	Putative
Ygbl	b2735	-	DeoR	H-T-H	30	49	14.9	265	Putative
YgeK	b2855	+	LuxR/UhpA	H-T-H	103	122	76.1	148	Putative
YgeV	b2869	+	EBP	H-T-H	567	585	97.2	592	Putative
Ygfl	b2921	+/-	LysR	H-T-H	26	45	11.7	303	Putative

Table 6...

YggD	b2929	?	Unclass					169	Putative			
YggG	b2936	?	Unclass					294	Putative		*speB*	
YgiP	b3060	+/-	LysR	H-T-H	23	42	10.4	310	Putative			
YgiX	b3025	?	Two	H-T-H	170	202	84.9	219	Putative			
YhaJ	b3105	+/-	LysR	H-T-H	24	43	11.2	298	Putative			
YhcI	b3222	?	NagC/XylR					302	Putative			
YhcK	b3226	-	GntR	H-T-H	58	77	25.6	263	Putative			
YhcS	b3243	+/-	LysR	H-T-H	19	38	9.2	309	Putative			
YhdM	b3292	-	MerR	H-T-H	4	23	9.5	141	Putative	Detoxication-putative		31
YheN	b3345	?	Unclass					128	Putative			
YheO	b3346	+	Unclass	H-T-H	211	232	90.7	244	Putative			
YhfR	b3375	-	GntR	H-T-H	60	79	26.2	265	Putative			
YhgB	b3422	?	EBP					532	Putative			
YhiF	b3507	+	LuxR/UhpA	H-T-H	133	152	80.9	176	Putative			
YhiW	b3515	+	AraC/XylS	H-T-H	155	174	67.9	242	Putative			48
YhiX	b3516	+	AraC/XylS	H-T-H	161	180	62.2	274	Putative			48
YhjB	b3520	+	LuxR/UhpA	H-T-H	159	178	84.2	200	Putative			
YhjC	b3521	+/-	LysR	H-T-H	43	62	16.2	323	Putative			
YiaJ	b3574	-	IclR	H-T-H	45	64	19.3	282	Putative			
YiaU	b3585	+/-	LysR	H-T-H	23	42	10.0	324	Putative			
YidF	b3674	?	Unclass					165	Putative			
YidL	b3680	+	AraC/XylS	H-T-H	213	232	72.4	307	Putative			48
YidP	b3684	-	GntR	H-T-H	28	47	15.7	238	Putative			
YidZ	b3711	+/-	LysR	H-T-H	25	44	10.8	319	Putative			
YihL	b3872	-	GntR	H-T-H	33	52	18.0	236	Putative			
YihW	b3884	-	DeoR	H-T-H	33	52	15.7	269	Putative			
YijC	b3963	-	Unclass	H-T-H	51	72	26.2	234	Putative			
YijO	b3954	+	AraC/XylS	H-T-H	188	207	69.7	283	Putative			48
YjaE	b3995	?	Unclass					158	Putative			
YjbK	b4046	-	Fur	H-T-H	127	146	86.9	157	Putative			64
YjcT	b4084	?	NagC/XylR					309	Putative			
YjcB	b4178	-	Unclass	H-T-H	26	47	25.8	141	Putative			
YjfQ	b4191	-	DeoR	H-T-H	20	39	11.7	251	Putative			
YjgS	b4264	-	GalR/LacI	H-T-H	8	27	5.2	332	Putative			
YjhI	b4299	-	IclR	H-T-H	27	46	13.9	262	Putative			
YjhU	b4295	-	Unclass	H-T-H	10	29	7.3	266	Putative			
YjiE	b4327	+/-	LysR	H-T-H	28	47	12.3	303	Putative			
YjiR	b4340	-	GntR	H-T-H	27	48	7.9	470	Putative			
YjjJ	b4385	-	Unclass	H-T-H	13	34	5.3	443	Putative			
YjjM	b4357	?	Unclass	H-T-H	74	95	31.5	268	Putative			
YjjQ	b4365	?	LuxR/UhpA					241	Putative			
YjjR	b4366	+	LuxR/UhpA	H-T-H	168	189	79.3	225	Putative			
YkgA	b0300	-	AraC/XylS	H-T-H	35	54	18.6	239	Putative			
YkgD	b0305	+	AraC/XylS	H-T-H	196	215	72.3	284	Putative			
YlcA	b0571	?	Two	H-T-H	171	203	82.4	227	Putative			
YmfL	b1147	-	Unclass	H-T-H	32	53	22.4	189	Putative			
YmfN	b1149	-	Unclass	H-T-H	41	62	11.3	455	Putative			
YnaE	b1375	?	Unclass					88	Putative			
YneJ	b1526	+/-	LysR	H-T-H	18	38	9.5	293	Putative			
YnfL	b1595	+/-	LysR	H-T-H	20	40	10.1	297	Putative			
YphH	b2550	-	NagC/XylR	H-T-H	30	51	10.1	399	Putative			
YqhC	b3010	+	AraC/XylS	H-T-H	284	305	78.5	375	Putative			

YrbA	b3190	?	BolA	84	Putative

Table 6. *E. coli* K-12 MG1655 transcriptional factors.

Only chromosomal proteins of *E. coli* were considered. The notation is as follows: protein name; bnumber, regulatory role (activator, repressor or dual), protein family, type of DNA-binding motif (HTH, zinc-finger, H-L-H, cold shock, and ß-sheet antiparallel), position of the motif in the sequence (start, end, and relative position), sequence length; type (know or hypothetical), regulatory function; physiological condition, operon regulated, and references associated. Families nomenclature: TetR/AcrR tetracicline repressor; AraC/XylS Arabinose/Xylose regulator; DeoR Deoxyribose regulator; EBP, EnhancerBinding Proteins; Cold, Cold-shock proteins; Crp, Crp-like proteins; GalR/LacI, galactose/lactose-regulators; GntR, gluconate repressor; LysR activator protein family; LuxR/UhpA; OmpR-like, OmpR-like proteins; ArgR, arginine repressor; ArsR, arsenical resistance regulator; AsnC, asparagine-like regulator; Fur, Ferric uptake repressor; BirA, Biotin repressor; DnaA, DNA activator; BolA, morphogenetic pathway family; HI-HN-S histone like proteins; HypF, hydrogenase regulator proteins; IclR, acetate operon repressor; LexA, SOS response; MarR, antibiotic resistance protein; MerR, mercuric resistance protein; MetJ, methionine biosynthesis repressor; NagR/XylR, sugar kinases repressor family; PemK; PutA, proline dehydrogenase repressor; RpiR; TrpR, tryptophan repressor;σ70, σ70-like proteins; and Unclass, proteins not classified so far. Two component proteins are included in the EBP, LuxR/UhpA and OmpR-like families.

A)

	Proteins	Classified	Unclassified	DBD	No-DBD
Characterized	159	135	25	134	26
Predicted	155	100	55	125	30
Total	314	235	80	259	56

B)

Function	Activator	Repressor	Dual	Unknown
Characterized	67	60	32	-
Predicted	25	53	27	50
Total	92	113	59	50

C)

	Positive	Negative	Dual
Autoregulation	6	40	3

Table 7. Functional description of all regulatory proteins.

A) Characterization of all proteins in terms of function and DBD. Hypothetical functions have been assigned based on the HTH relative position in the protein sequence. The columns are as follows: Proteins are the total of proteins *per* category; the second and third columns describe the classified and unclassified proteins; DBD is the DNA-binding domain described, and no-DBD are proteins with no DBD described so far. B) Proteins grouped by their regulatory roles. Activator, repressor and dual functions both in characterized and predicted proteins are shown. "Unknown" indicates proteins with no function assigned, and C) Proteins with autoregulation.

Protein	Family	ID	Resolution (Å)	Motif	References
ArgR*	ArgR	1aoy	NMR	wHTH	152
BirA*	BirA	1bia	2.3	wHTH	171
Crp*	Crp	1ruo	2.7	wHTH	168
CspA	Cold shock	3emf	NMR	RBD	41
Fis*	EBP	1fia	2.0	HTH	82
FruR*	GalR/LacI	1uxc	NMR	cHTH	122
LacI*	GalR/LacI	1lcc	4.8Å	cHTH	89
LexA*	LexA	1lea	NMR	wHTH	45
MetJ	MetJ	1cma	2.8	ß-sheet	147
NarL*	LuxR/UhpA	1rnl	2.2	wHTH	6
OmpR*	OmpR	1opc	1.95	wHTH	99
PurR*	GalR/LacI	1wet	2.6	cHTH	141
TetR*	TetR/AcrR	2trt, 2tct	2.5	HTH	62, 76
TrpR*	TrpR	1tro, 1trr	1.9, 2.4	CHTH	179, 87

Table 8. DNA-binding structures described in transcriptional factors of *E. coli*.

DNA-binding motifs described in transcriptional factors of *E. coli*. Regulatory families are named as in Table 6. ID is the identifier of the protein in the PDB database. The resolution of the structure is given in Amstrongs. The motifs are: HLH: helix-loop-helix, c-HTH for classic HTH, wHTH for winged HTH, and RBD: RNA Binding Domain. *Proteins used in the Gibbs sampler calibration.

Family	*E. coli*		Total		Prokaryotes	
	Know	Predicted	#	%	#	%
AraC/XylS	14	13	27	8. 6	80	33.75
ArgR	1	0	1	0.3	2	50
ArsR	1	0	1	0.3	12	8.3
AsnC	2	1	3	0.9	16	18.75
BirA	1	0	1	0.3	5	20
BolA	1	1	2	0.6	nd	nd
Cold	9	0	9	2.8	50	18
CRP	2	0	2	0.6	28	7.1
DeoR	7	7	14	4.4	24	58.3
DnaA	1	0	1	0.3	24	4.1
EBP	9	5	14	4.4	56	25
Fur	1	1	2	0.6	14	14.3
GalR/LacI	12	2	14	4.4	49	28.5
GntR	8	12	20	6.4	38	52.6
HN-S	2	0	2	0.6	nd	nd
HYPF	1	0	1	0.3	nd	nd
IclR	3	5	8	2.5	12	66.6
LexA	1	0	1	0.3	7	14.3
LuxR/UhpA	11	6	17	5.4	59	28.8
LysR	18	27	45	14.3	173	26
MarR	3	0	3	0.9	13	23.1
MerR	1	4	5	1.5	26	19.2
MetJ	1	0	1	0.3	2	50
NagC/XylR	2	5	7	2.2	19	36.8
PemK	2	0	2	0.6	nd	nd
PutA	1	0	1	0.3	3	33.3
RpiR	1	0	1	0.3	nd	nd
σ70	2	0	2	0.6	nd	nd
TetR/AcrR	4	5	9	2.8	22	40.9
TrpR	1	0	1	0.3	5	20
OmpR	13	3	16	5.1	61	26.2
Unclass	30	52	82	26.1	nd	nd
TOTAL	165	149	314	100	701	44.79

Table 9. Regulatory Protein families: *E. coli vs.* Prokaryotes

All regulatory protein families in prokaryotes (*E. coli* included) were obtained by an exhaustively search in SwissProt and in the *E. coli* genome. The first column is family name, second column represents the *E. coli* known and predicted regulatory proteins, and the third and fourth are the total and percent of proteins *per* family (The percent for each one family represents the proportion of the family in the collection of 314 proteins). Five and six columns describe the total of eubacterial transcriptional factors *per* family, and the percentage that represents the *E. coli* family in all prokaryotes, respectively.

Family	Function	Mean	Physiological function	%	Members
CRP	Dual	230 ± 20	Global responses	100%	2
Cold	Activator	70.0 ± 1.8	Low temperatures: Cold Shock	90%	9
GalR/LacI	Repressor	333.7 ± 12.3	Carbon sources uptake	90%	14
EBP	Activator	Heterogeneous	Nitrogen assimilation, aromatic amino acid synthesis, and several functions	70%	14
TetR/AcrR	Repressor	210.0 ± 17.1	Tetracycline resistance	66%	9
AsnC	Dual	156 ± 6,9	Amino acid biosynthesis	66%	3
IclR	Repressor	272.4 ± 19.5	Carbon sources uptake	60%	8
GntR	Repressor	246.0 ± 1.4	Carbon metabolism	55%	20
DeoR	Repressor	256.8 ± 12.7	Sugar metabolism	50%	14
LysR	Dual	Heterogeneous	Amino acid biosynthesis	50%	45
OmpR#	Activator	230.6 ± 7.3	Adaptative response	45%	16
LuxR/UhpA	Activator	219.2 ± 13.2	Biosynthesis, and glycerol metabolism	40%	17
AraC/XylS	Activator	Heterogeneous	Virulence, transposition, sugar metabolism	22%	27
Others	Several	233.9 ± 258.6	Several functions	**	116

Table 10. Functional conservation of transcriptional factor families.
We whow the most conserved families in terms of length size. Only families with more than 8 members in *E. coli* were considered in this analysis. The first column indicates the family name; the second column contains the mean of the length sequence size for each family, and the standard deviation (±). Additionally, the main physiological function, the number of proteins *per* family, and the regulatory role associated are indicated in the last three columns. Although Crp and AsnC families are rarely represented in *E. coli* to consider a main physiological function, in the most prokaryotes the function is conserved (data not shown). ** Not calculated.

Family	Replichore 1	Replichore 2	Parallel	Antiparallel
AraC/XylS	60.0	40.0	40.0	60.0
Cold	75.0	25.0	50.0	50.0
DeoR	43.7	56.2	31.2	68.7
EBP	38.4	61.5	53.8	46.1
GalR/LacI	42.8	57.1	35.7	64.3
GntR	40.0	60.0	54.5	45.4
IclR	71.4	28.5	14.3	85.7
Luxr/UhpA	35.3	64.7	52.9	47.1
LysR	43.2	56.8	29.5	70.4
NagR/XylR	66.6	33.3	50.0	50.0
TetR/AcrR	57.1	42.8	42.8	57.1
OmpR	66.6	33.3	40.0	60.0
Others	60.1	39.8	50.0	50.0
TOTAL	45.6	54.3	44.7	55.2

Table11. Direction of transcription/replication in all regulators of *E. coli*.

The first column indicates the family name. The second and third columns desribe the replichore (defined in Blattner, F. et al. 1997. Fourth and fifth columns contain the percentage of proteins that have the same direction of transcription/replication (Parallel); and percentage of members with different direction of transcription/replication (antiparallel). 45% of the regulators are in the replichore_1 and 55% in the replichore_2.

Protein1	Protein2	Family	Address		Regulatory function	
Parallel transcription						
LacI	MhpR	GalR/LacI-IclR	R	R	-	+
HipA	HipB	Unclass- Unclass	R	R	?	-
Mlc	YnfL	NagC/XylR-LysR	R	R	-	+/-
FlhC&	FhlD&	Unclass- Unclass	R	R	+	+
Cbl*	Nac*	LysR-LysR	R	R	+/-	+/-
YjhI	SgcR	IclR-DeoR	R	R	-	-
YhiW*	YhiX*	AraC/XylS - AraC/XylS	R	R	+	?
YheN	YheO	Unclass - Unclass	R	R	?	+
b1146	YmfL	Unclass- Unclass	F	F	+	-
b1438	b1439	Unclass- GntR	F	F	+	?
MarR#	MarA#	MarR- AraC/XylS	F	F	-	+
YfeC	YfeD	Homol-Unclass	F	F	-	-
GutM#	SrlR#	Unclass -DeoR	F	F	+	-
RhaS*#	RhaR*#	AraC/XylS - AraC/XylS	F	F	+	+
Divergent transcription						
CspB*	CspF*	Cold-Cold	R	F	+	+
CspH*	CspG*	Cold-Cold	D	P	+	+
b0845	b0846	Unclass- TetR/AcrR	R	F	?	?
YdaR	YdaS	Unclass- Unclass	R	F	-	-
DicC	DicA	Unclass- Unclass	R	F	-	-
GatR_1	b2088	DeoR- Unclass	R	F	-	?
TdcA	TdcR	LysR- Unclass	R	F	+	+
YhjB	YhjC	LuxR-LysR	R	F	+	+/-
SoxS#	SoxR#	AraC/XylS-MerR	R	F	+	+
YahA	YahB	Unclass -LysR	F	R	-	+/-
YhgB	GlpR	EBP-DeoR	F	R	?	-
YcdC	PutA	TetR/AcrR-PutA	F	R	-	-
PurR	YdhB	GalR/LacI -LysR	F	R	-	+/-
GlcC	b2981	GntR-Unclass	F	R	+	-

Table 12. Neighbor transcriptional factor genes.

The first and second columns describe the names of the neighboring genes. The family of the regulator is indicated in the third column. The direction of transcription (forward or reverse) is indicated in the fourth column. The fifth and sixth columns indicate the regulatory roles of the proteins. Plus: activator, Minus: Repressor, Plus/Minor: dual, and unknown. Genetic duplication are marked with stars (sequence comparison). In the Cold shock family there are evidences of genetic duplication: Genome position and sequence conservation (see text). #Cascades of regulation. & Protein complex. Parallel refers to the same direction of transcription and replication for both genes, and divergent refers to genes in opposite direction of transcription/replication.

Family	% in total of 314	% in 105 single-direction
LysR	14	21
EBP	4.5	7.6
GalR/LacI	4.5	9.6
LuxR/UhpA	5.4	6.7
AraC/XylS	8.6	8.6
GntR	6.4	5.7
Other*	45	40.8
Total	88.4	100%

Table 13. 105 regulatory genes are organized as single-transcription units.

* DeoR, Crp, Fur, IclR, OmpR-like, TetR, Cold, and unclass proteins.

Regulator	Family	Regulatory function	Additional Function
Ada*	ArsR/XylS	Dual regulator	Methylated DNA protein cysteine methyltransferase
ArgR*	ArgR	Arginine repressor	Site specific recombination in the plasmid ColE1
BirA*	BirA	Biotin repressor	Biotin-Acetyl-CoA-Carboxylase synthetase
DnaA	DnaA	Dual regulator	Involved in the chromosomal replication initiation
Fis*	EBP	Dual regulator	Factor for invertion stimulation protein
HNS	Hns	Global repressor	Histone like protein
PutA	PutA	Proline repressor	Proline dehydrogenase
Crl	Unclass	Activator	DNA organization, Curli formation and fibronectin binding
NadR*	Unclass	Repressor	NAD transport

Table 14. Additional functions described in regulatory proteins.

Proteins with a HTH described as marked with an asterisk. DnaA present a HLH domain. HNS has a ß-sheet, and PutA does not have an apparent DNA-binding motif

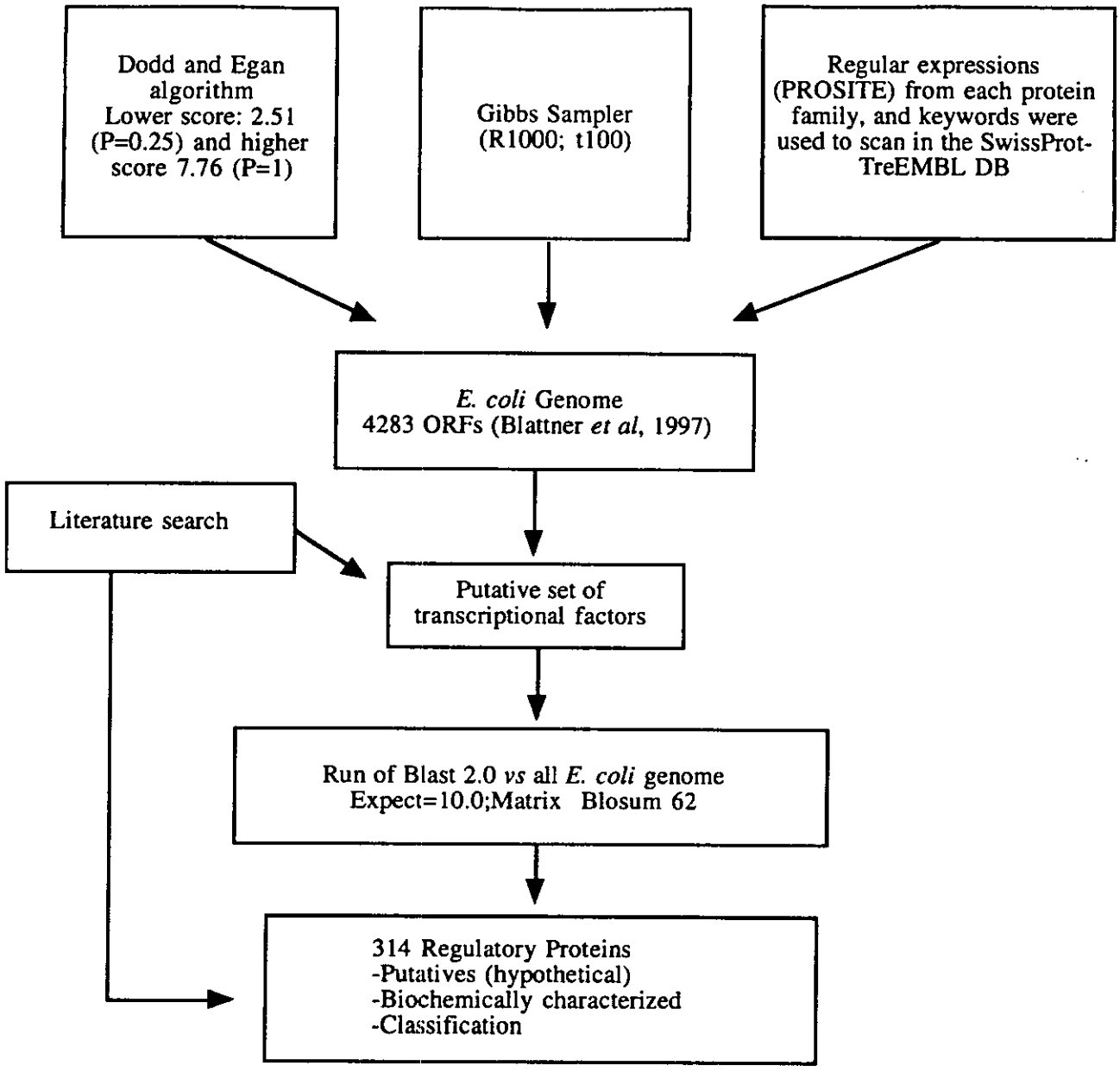


Figure 1. Prediction of regulatory proteins in the *E. coli* genome.
The method is divided in two stages. In the first stage, a set of putative transcriptional factors is described. Search in SwissProt DB (keywords, and regular expressions were used), and scanning all ORFs (4283) of *E. coli* genome using Dodd and Egan, Prosite patterns, and Gibbs sampler algorithms were performed. Parallel to scanning the genome and the SwissProt search, a literature search was performed for evidences about our transcriptional factors. We filtered proteins that are not transcriptional regulators (a transcriptional factor has a DNA-bind functional motif). In the second part, we made a sequence comparison using blast to detect additional proteins that in previous stages were not detected. Additionally, we used, the annotations of the *E. coli* genome provided by Blattner, and the collection of RegulonDB (Huerta et al 1998).

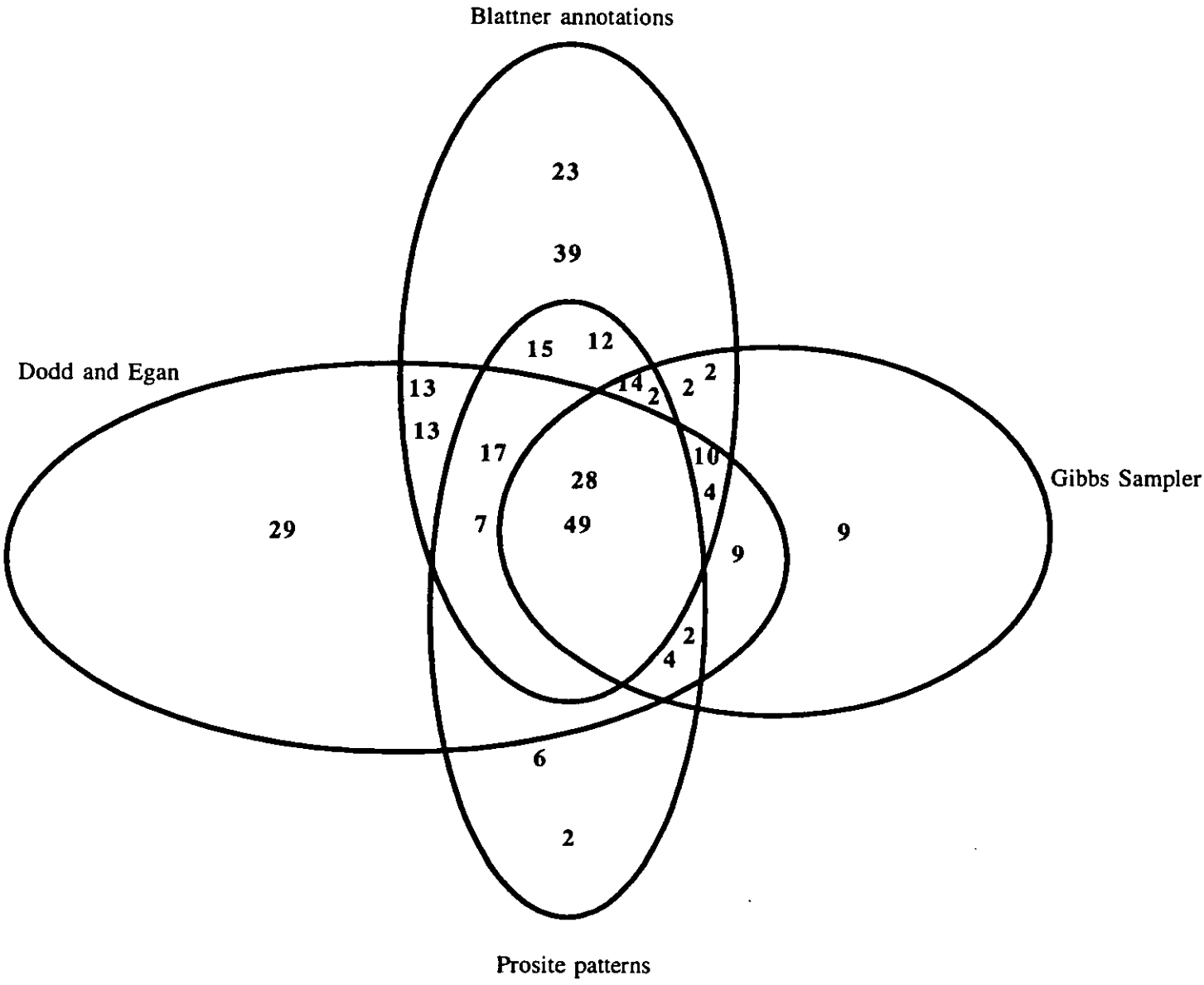


Figure 2. Venn diagram of all predictions.
Venn diagram showing all proteins and the methods used. In blue are characterized proteins and in black are predicted proteins.

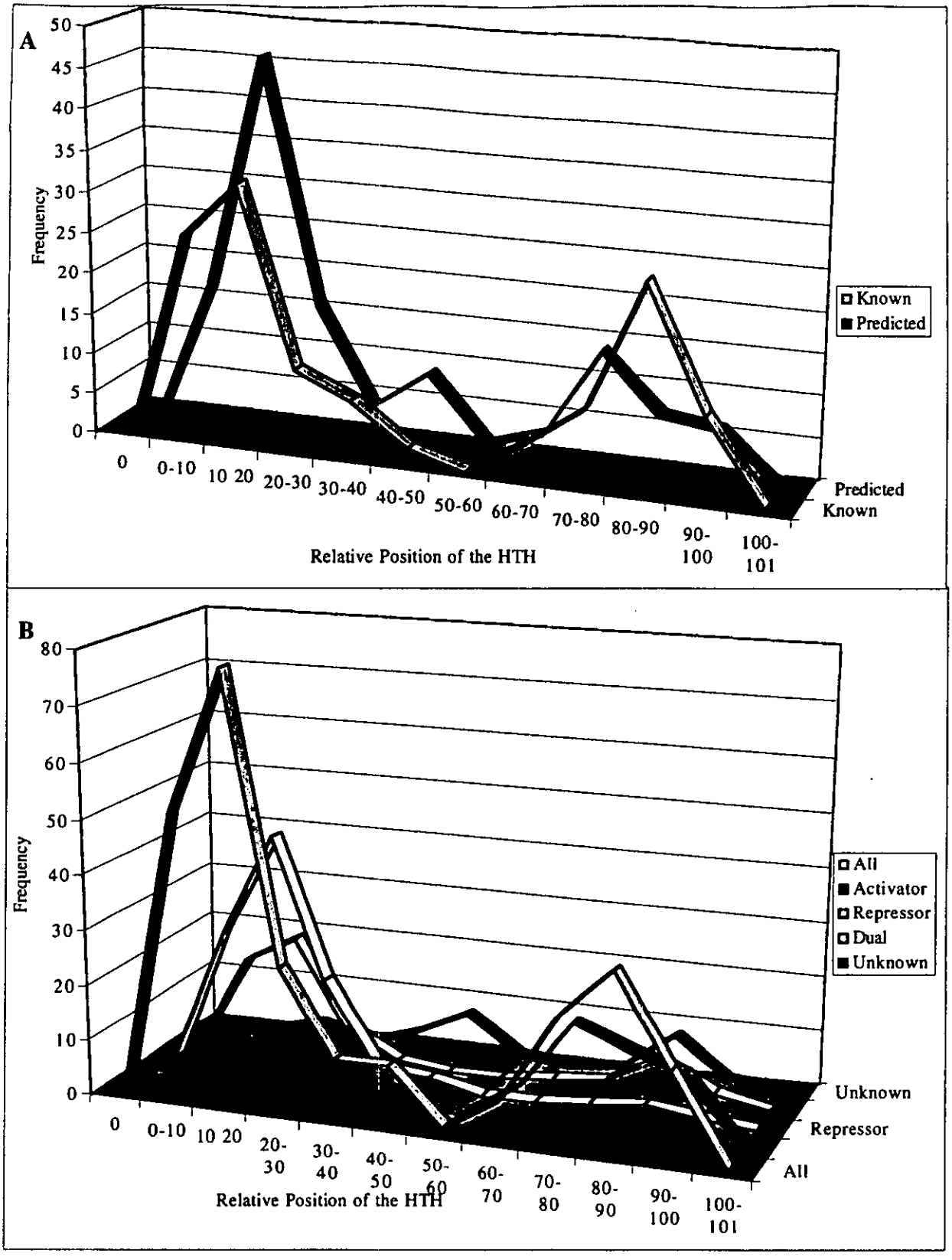


Figure 3. HTH distribution in the complete set of *E. coli* proteins.
In the X-axis, 0% represents the N-terminal and 100% the C-terminal end of the protein. The Y-axis shows the frequency of proteins. 234 transcriptional factors with the HTH and regulatory role were taken into account for this analysis. Known and hypothetical transcriptional factors have a similar distribution as the one observed for all prokaryotic proteins. In panel A, the relative position of the HTH was calculated into characterized (blue line) and predicted (red line) transcriptional factors. In panel B) all proteins were plotted respect to regulatory role.

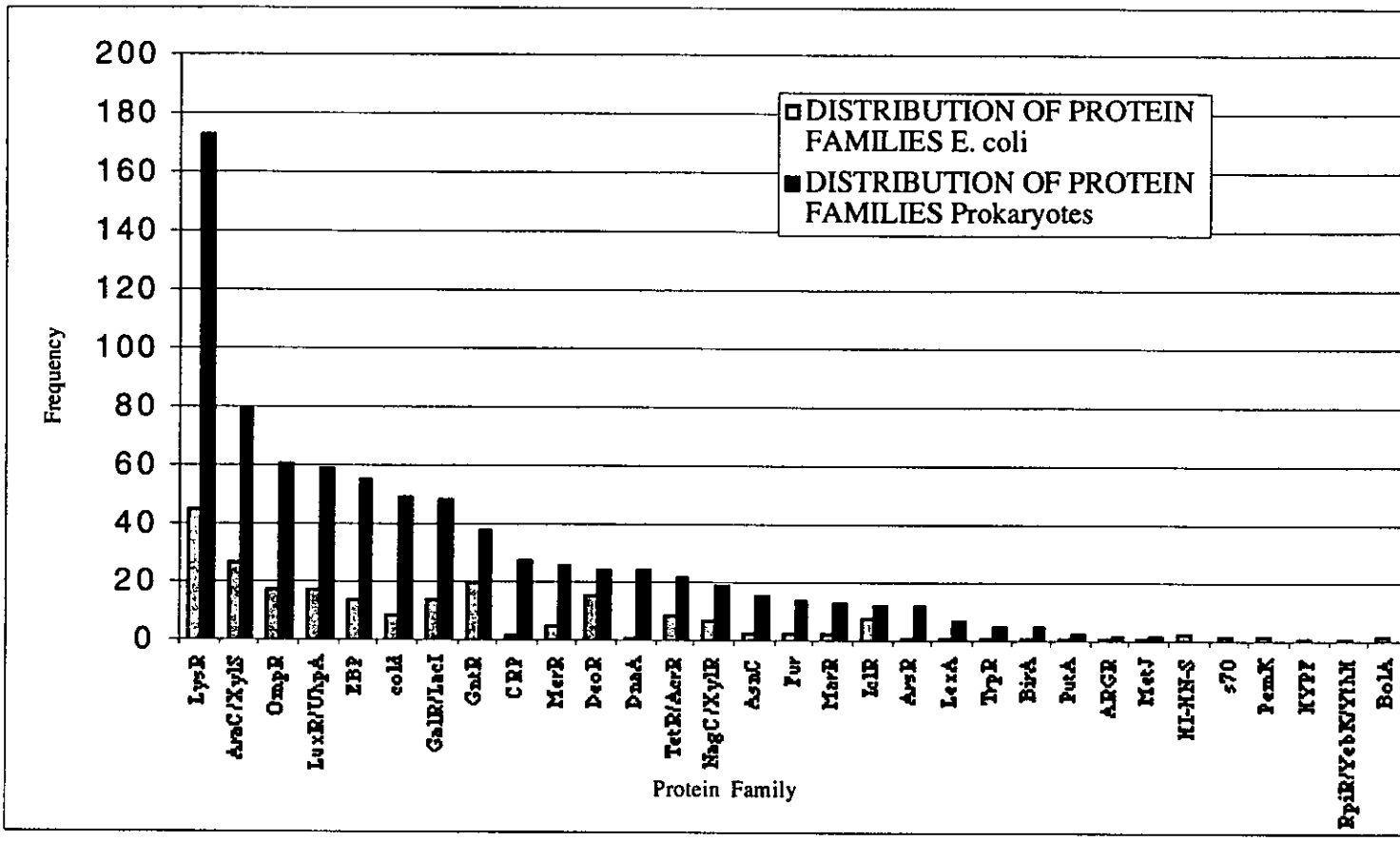


Figure 4. Protein families distribution in Prokaryotes and *E. coli*.
All protein families reported in prokaryotes were obtained by an exhaustively search in SwissProt and all ORFs of *E. coli*, using several approaches. All protein families are present in the same proportion both in *E. coli* (blue) and prokaryotes (red). In the "X" axis the protein families are represented while in the "Y" axis the frequency.

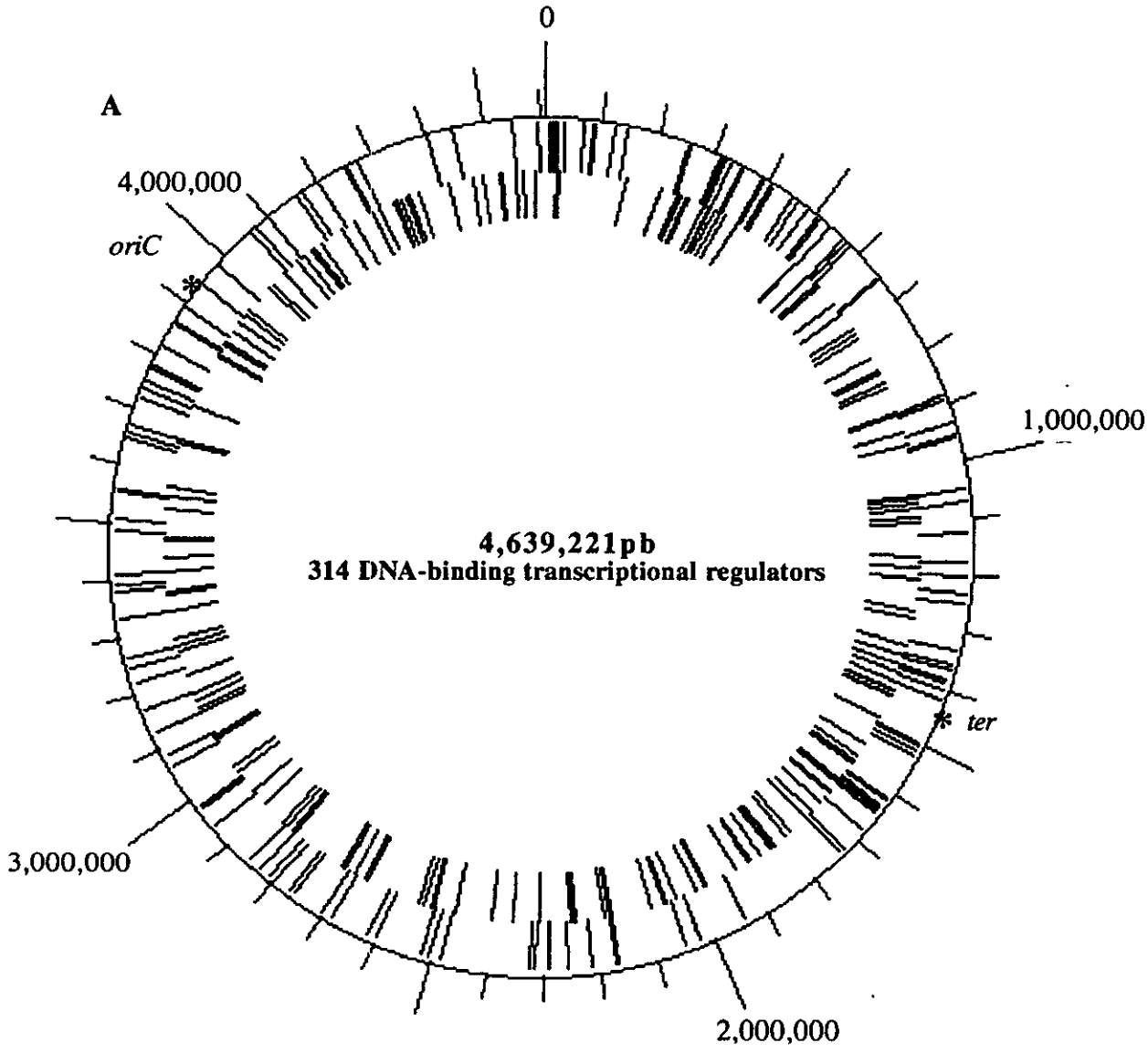
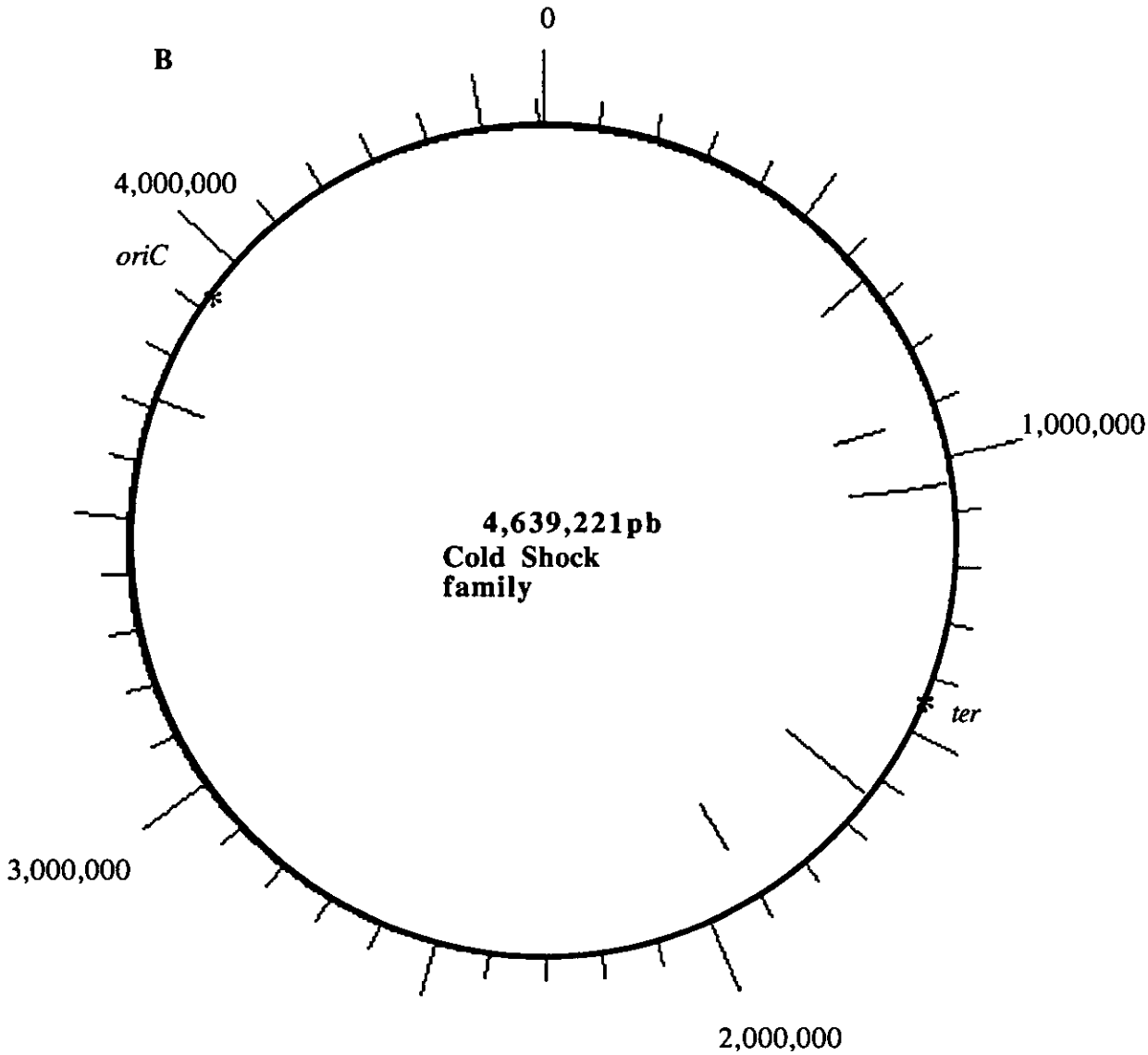
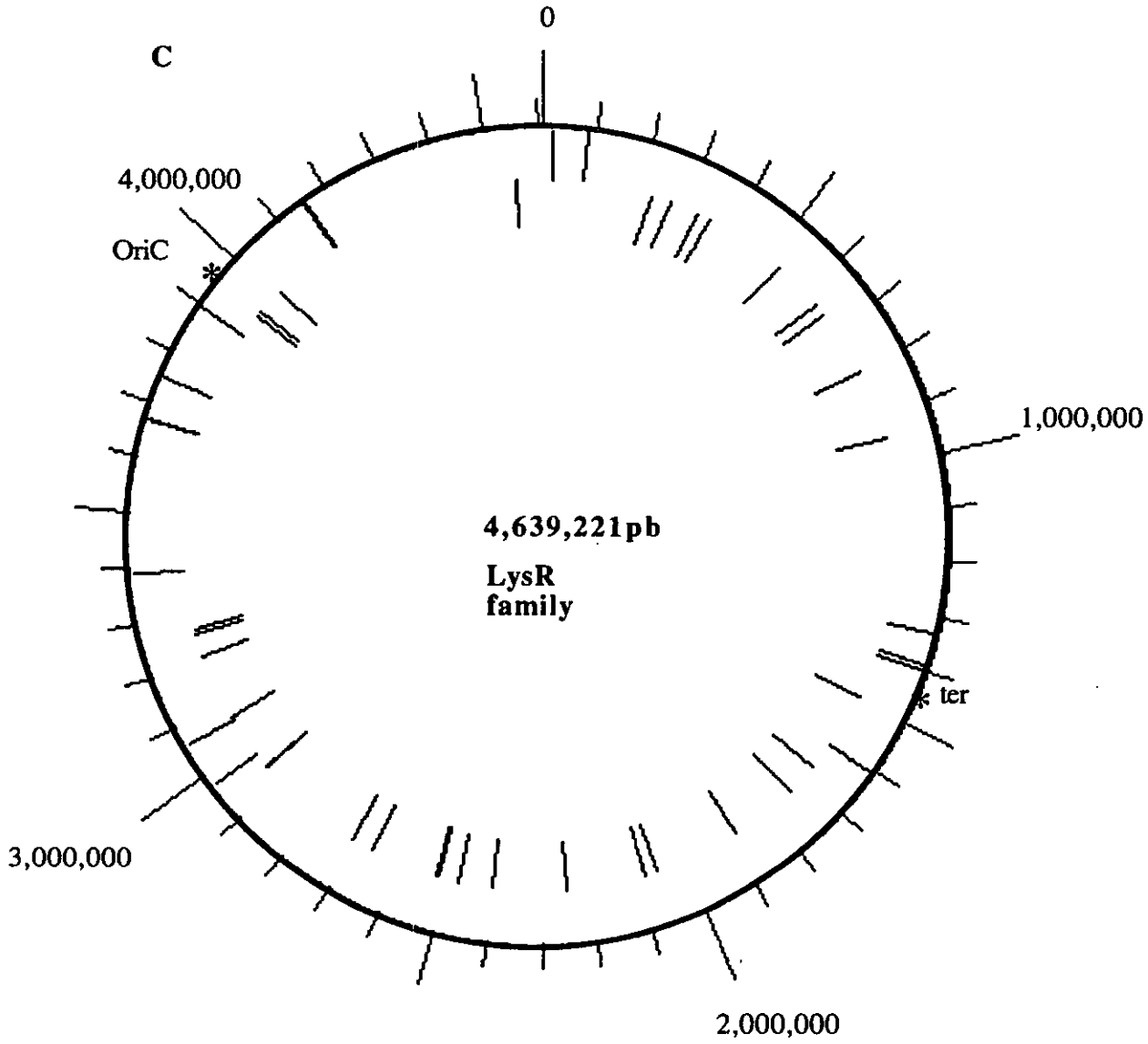
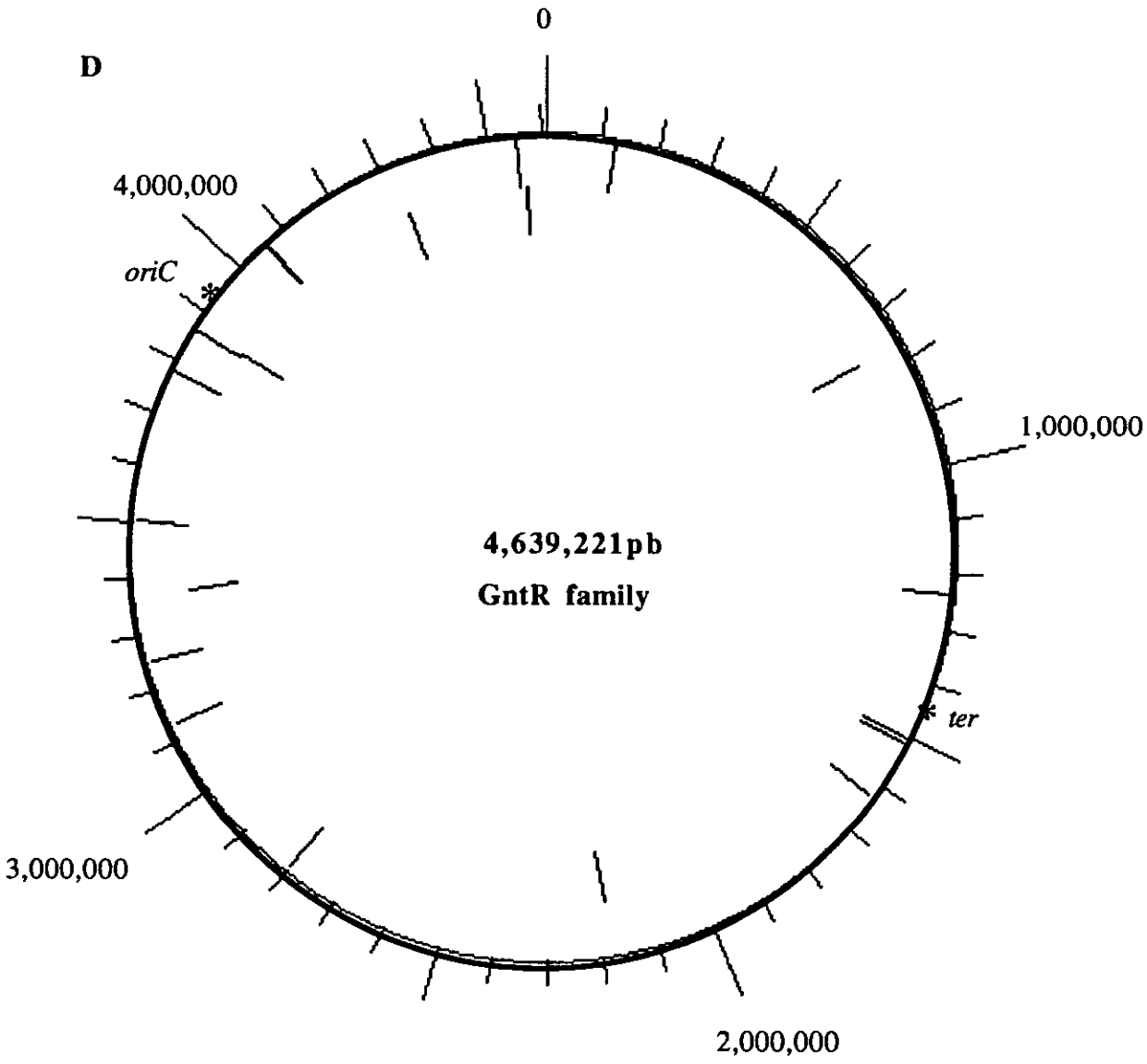


Figure 5. Regulatory gene location in the *E. coli* chromosome
In A, All gene regulators were plotted in a circular map of *E. coli*. *oriC* (replication origin) and *ter* (termination site) are represented by a star. Genes are represented in blue lines. Genes that touch the line have the same direction of transcription and replication, while genes that do not touch the genome are transcribed opposite direction to thereplication, B) the Cold Shock family, C) the LysR family and d) the GntR family.


B
0
4,000,000
oriC
1,000,000
4,639,221pb
Cold Shock
family
ter
3,000,000
2,000,000


C
0
4,000,000
OriC
1,000,000
4,639,221pb
LysR
family
ter
3,000,000
2,000,000


D
0
4,000,000
oriC
1,000,000
4,639,221pb
GntR family
ter
3,000,000
2,000,000

SEGUNDA PARTE: Propuesta del supergrupo HTH N-terminal

Un tema importante en la regulación transcripcional es la evolución de los reguladores, y en particular aquellas proteínas que presentan el motivo de unión al DNA del tipo Hélice-vuelta-hélice (HTH). Aunque la estructura primaria de estos motivos no esta totalmente conservada en todos los reguladores que la presentan, la estructura base parece no estar alterada, por ejemplo, comparaciones estructurales del HTH y regiones adyacentes han evidenciado la existencia de al menos siete familias. Únicamente la región que puede ayudar a diferenciar entre un HTH y otro -a nivel de estructura- es la orientación de una tercer α-hélice que bordea al motivo funcional (Wintjens et al.1997). Esto significa que el motivo requiere un número limitado de residuos para adoptar un plegamiento correcto (Sauer et al.1982. Brennan et al.1989) y que conlleva a que la secuencia adicional permita un número de sustituciones que no afectarían la estructura central. En esta segunda parte del trabajo, mostramos la existencia de un supergrupo de proteínas reguladoras con el HTH en el N-terminal originado posiblemente de un ancestro común. Esta especulación esta basada en la existencia de patrones comunes, pero cuyo tiempo de divergencia es tal, que difícilmente pueden detectarse por métodos de análisis de secuencia tradicionales.

Material y Método

El genoma completo de *E. coli* cepa K12 versión m53 que contiene alrededor de 4400 *ORFs* fue utilizado para este trabajo (Blattner et al.1997). Por otra parte, todos los dominios de pegado al DNA (DBD) se definieron por medio de una consulta en las bases de datos de ProDom (Corpet et al.1999; Corpet et al.1998; Gouzy et al.1996), y en el *Protein DataBank* (Sussman et al.1998). Posteriormente, estas delimitaciones fueron corroboradas por consultas bibliográficas.

En general, podemos afirmar que los DBD presentan una longitud de alrededor de 60 residuos de aminoácidos, por ejemplo, el DBD en miembros de la familia GalR/LacI, como LacI (Lewis et al.1996), PurR (Schumacher et al.1994), y FruR (Penin et al.1997) es de alrededor 59 residuos y se localiza cerca del N-terminal (Weickert et al.1992). El DBD en los miembros de la familia

AraC/XylS consta de alrededor 60 residuos de longitud (Gallegos et al.1997), y presenta una posición variable dentro de la secuencia. En los activadores Rob (Ariza et al.1995), y MarA (Jair et al.1995) el DBD se localiza en el N-terminal, mientras que en AraC (Bustos et al.1993), RhaR (Tobin et al.1990), y RhaS (Tobin et al.1990) se localiza en el último tercio del C-terminal.

En este trabajo consideramos 81 reguladores transcripcionales agrupados en 13 familias (ver tabla 3) para detectar relaciones evolutivas distantes. El DBD en estas proteínas se localizado en el N-terminal o en el C-terminal. La estrategia para comparar y detectar posibles regiones comunes entre las secuencias proteicas ya ha sido descrito (Neuwald et al.1997). Cada uno de los DBD fueron utilizados como secuencias *queries* para buscar y detectar homólogos distantes en todos los *ORFs* del genoma de *E. coli* utilizando el programa denominado Probe (Neuwald et al.1997). Probe construye un alineamiento utilizando diferentes herramientas de comparación de secuencias. En el paso inicial, se compara por Blast2.0 una secuencia *query* (los DBD ya mencionados) contra todo el genoma de *E. coli*, esta comparación genera un conjunto de secuencias relacionadas que constituyen a su vez nuevas secuencias *queries*. Aquí se introduce el criterio de homología indirecta, es decir, si A es homologa de B y B es homologo de C, entonces se asume que A y C son *homólogos*. Una vez que se converge a un *set* que no puede crecer más (ya que se recorrió todo el espacio de búsqueda), entonces se utiliza el método de Gibbs Sampler para detectar un patrón común en todas estas secuencias. Adicionalmente se generan matrices a partir de estos patrones y se vuelve a realizar otra serie de búsquedas para detectar posibles secuencias relacionadas que no se detectaron con anterioridad. Al detectarse el patrón común se realinea varias veces con algoritmos genéticos para determinar el mejor alineamiento y presentarlo como el óptimo. (Neuwald et al.1997). Cabe señalar que el proceso de búsqueda es *iterativo* y termina cuando ya no se detectan más secuencias relacionadas o bien cuando el conjunto de proteínas detectado ya no crece más.

En el caso de las proteínas reguladoras utilizadas como secuencias *query*, el programa convirgió varias veces en un patrón común que es compartido por proteínas de ocho familias reguladoras y que corresponde al DBD (ver tabla 3 y figura 5).

Familia	DBD	# de Proteínas	Familias detectadas
AraC/XylS	N y C- terminal	7	AraC/XylS
AsnC	N- terminal	1	AsnC
Crp	C- terminal	2	-
DeoR	N- terminal	7	DeoR, GntR, IclR, GalR/LacI, TetR, Crp, MarR
EBP	C- terminal	5	EBP
IciR	N- terminal	2	GntR, DeoR, IclR, MarR
GalR/LacI	N- terminal	12	GalR/LacI, TetR/AcrR, DeoR
LexA	N- terminal	1	LexA
GntR	N- terminal	9	GntR,IciR, DeoR. Fur
LuxR/UhpA	C-terminal	6	-
LysR	N- terminal	19	LysR
MerR	N- terminal	1	MerR
TetR/AcrR	N- terminal	5	GalR/LacI

Tabla 3.Familias empleadas para detectas relaciones evolutivas distantes. 81 secuencias de diferentes familias fueron utilizadas como *queries* para la detección de patrones comunes, 4 de las cuales no están clasificadas. Nomenclatura: Familia; DBD: Región donde se localiza el dominio de pegado al DNA en la secuencia; Número de proteínas utilzadas como queries; y, familias que comparten un patrón.

Resultados

Con base en la comparación de secuencias y pensando que el HTH es un motivo relativamente conservado (Ohlendorf et al.1983), nosotros realizamos una búsqueda *iterativa* en todo el genoma de *E. coli* K12. En este sentido, la convergencia de los patrones detectados en diversas proteínas agrupadas en ocho familias deben ser interpretados como las secuencias más relacionadas en todo el universo de proteínas de *E.coli* y que corresponden a varias proteínas reguladoras agrupadas en al menos ocho familias evolutivas

Alineamiento mostrando las relaciones entre las familias proteicas

La búsqueda independiente de patrones con las 81 secuencias *query* (ver tabla 3), convirgió hacia un patrón que relaciona a proteínas de ocho familias reguladoras: GalR/LacI (Weickert et al.1992), DeoR (Valentin-Hansen et al.1985), IclR (Ferrandez et al. 1997), GntR (Haydon et al.1991), Crp (Spiro.1993), MarR (Sulavik et al.1995) MerR (Zeng et al.1998), y AsnC (De Mot et al. 1996), además de cinco proteínas no clasificadas (P77484, ~DtxR, YbaE, YmfN y YabN). En todas estas

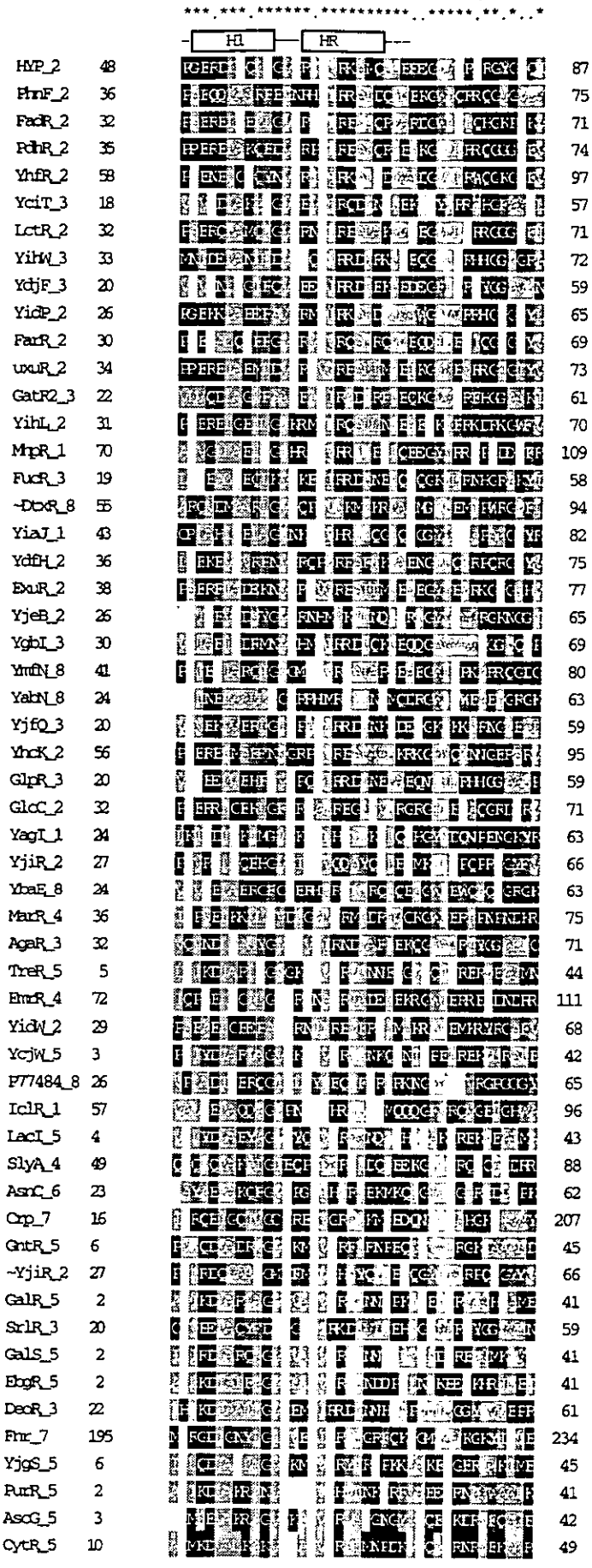
proteínas el patrón detectado se localiza cerca del N-terminal, con excepción de tres proteínas que lo presentan en el C-terminal (Crp, Fnr y YjdG). Dicho patrón esta constituido por 40 residuos de aminoácidos, y corresponde al DBD que contiene tanto el HTH como una tercer α-hélice en varias de las proteínas (Wintjens et al.1997; Suzuki et al.1995). Cabe mencionar que independientemente del patrón utilizado para cada una de las búsquedas en el genoma completo de *E. coli*, la convergencia fue hacia un mismo conjunto de proteínas reguladoras, por lo que podemos señalar que las proteínas relacionadas son aquellas que son recuperadas por diferentes secuencias pregunta en búsquedas independientes. De hecho, alrededor del 80% del total de las proteínas de estas ocho familias fueron recuperadas en las diferentes búsquedas.

Los primeros 10 residuos del patrón incluyen la hélice de reconocimiento inespecífico y son principalmente de naturaleza hidrofóbica, excepto en la posición 5 donde un aminoácido con carga negativa está presente. Las características de esta región son consistentes con que la función de la primer α-hélice, ya que contacta inespecificamente los grupos fosfatos del DNA (Brennan et al.1989). La segunda α-hélice es importante dentro del proceso de reconocimiento, tiene tres sitios hidrofóbicos que han sido descritos en miembros de la familia Crp que estarían formando un *pocket* hidrofóbico que estabilizaría el HTH, mientras que las posiciones variables dentro de la estructura serían importantes para el reconocimiento específico del DNA (Brennan et al. 1989). Los residuos Ala8 y Gly9 son importantes para formar la vuelta entre las dos hélices (Steitz et al.1982. Weber et al.1982).

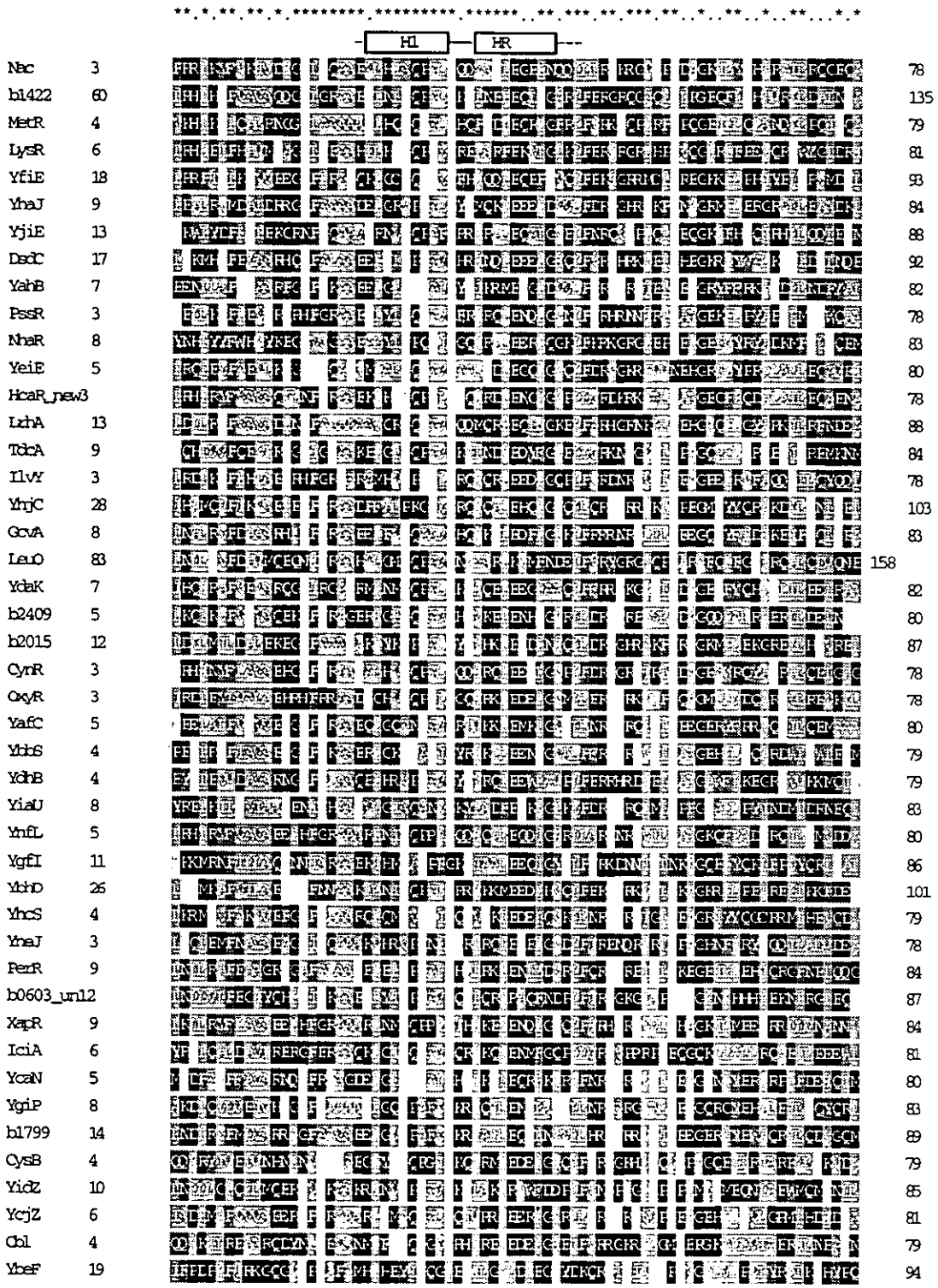
Debemos mencionar, que detectamos patrones adicionales a la región donde se localiza el HTH. Dichos patrones son específicos para algunas familias de reguladores y corresponden a sitios de interacción con inductores o sitios para multimerización de las proteínas. Por ejemplo, en las familias GalR/LacI, DeoR y TetR/AcrR, se localiza un segundo patrón en posición C-terminal y que estaría implicado en el pegado al inductor (Weickert et al.1992). Figura 5.

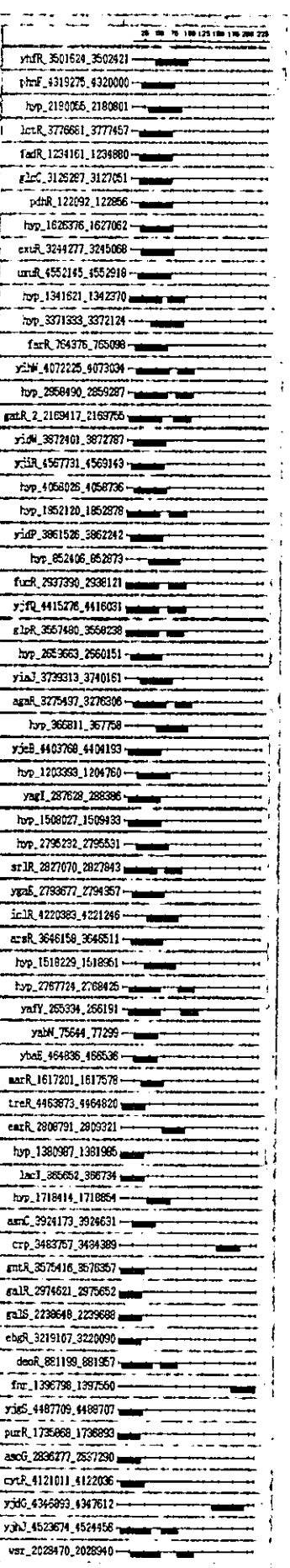
Un segundo grupo de proteínas reguladoras claramente delimitado fue detectado utilizando como secuencias pregunta a miembros de la familia LysR. El segundo grupo corresponde a miembros de la familia LysR (Henikoff et al.1988. Schell.1993).

A) Supergrupo



B) Familia LysR




yhfR_3501624_3502421
phnF_4319275_4320000
hyp_2180055_2180801
lctR_3776661_3777457
fadR_1234161_1234880
glcC_3126267_3127051
pdhR_122092_122856
hyp_1626376_1627062
exuR_3244277_3245068
uxuR_4552145_4552918
hyp_1341621_1342370
hyp_3371333_3372124
fsrR_764376_765098
ydhW_4072225_4073034
hyp_2858490_2859287
gntR_2_2169417_2169755
ydiM_3872401_3872787
ydiR_4567731_4569143
hyp_4056026_4056736
hyp_1852120_1852878
ydiP_3861526_3862242
hyp_852406_852873
fucR_2937390_2938121
yjfQ_4415276_4416031
glpR_3557480_3558238
hyp_2559663_2560151
yiaJ_3739313_3740151
agaR_3275437_3276306
hyp_366811_367758
yjeB_4403768_4404193
hyp_1203393_1204760
yagI_287628_288386
hyp_1508027_1509433
hyp_2795232_2795531
srlR_2827070_2827843
ygaE_2793677_2794357
iclR_4220383_4221246
srsR_3646158_3646511
hyp_1518229_1518961
hyp_2767724_2768425
yafY_255334_256191
yahN_75644_77299
ybaS_464836_466536
marR_1617201_1617578
treR_4463873_4464820
carR_2806791_2809321
hyp_1380987_1381986
lacI_365652_366734
hyp_1718414_1718854
asnC_3924173_3924631
crp_3483767_3484389
gntR_3575416_3576357
galR_2974621_2975652
galS_2238648_2239688
ebgR_3219107_3220090
deoR_881199_881957
fnr_1396798_1397560
ydeS_4487709_4488707
purR_1735868_1736893
ascG_2836277_2837230
cytR_4121011_4122036
ydiG_4346893_4347612
yghJ_4523674_4524456
vsr_2028470_2028940

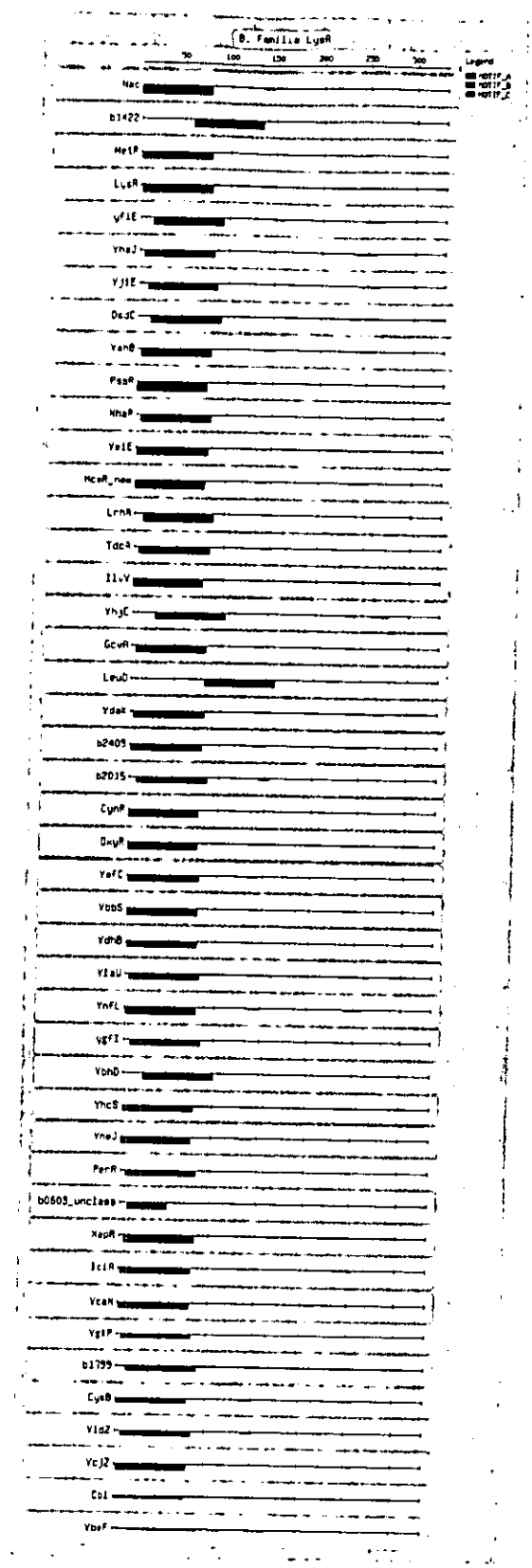

Nac
b1422
MetR
LysR
yfiE
YhaJ
YjiE
DsdC
YahB
PsaR
NhaR
YeiE
HcaR_new
LrhA
TdcA
IlvY
YhjC
GcvA
LeuO
YdeK
b2409
b2015
CynR
OxyR
YafC
YbbS
YdhB
YiaU
YnfL
YgfI
YbhD
YhcS
YneJ
PerR
b0603_unclass
XapR
IciA
YcaN
YgiP
b1799
CysB
YidZ
YcjZ
Cbl
YbaF

Figura 5. Patrones detectados y alineamiento de las proteínas relacionadas en *E. coli*. A) Supergrupo HTH-N terminal y B) familia LysR. En azul intenso se señala el DBD (motif A) y en otros colores (motifs B, C) sitios probables de contacto con el inductor.
En la primer columna se muestra el nombre de la proteína; el número representa a las siguientes familias: 1 IclR; 2 GntR; 3 DeoR; 4 MarR, 5 GalR/lacI; 7 Crp/Fnr; y 8 no clasificadas. En el caso de la familia LysR se presenta solo el nombre de los miembros de la misma. Los números bordeando la secuencias indican el inicio y el final del patrón en la secuencia. Los aminoácidos son representados de acuerdo al siguiente código: rojo, residuos ácidos (D y E); rosa, residuos amídicos (N y Q); azul fuerte, residuos básicos (K, R y H); azul claro, residuos alifáticos (A, V, L, e I); verde, residuos aromáticos (F, Y y W); amarillo, residuos hidroxilos (S,T); negro: G, P, M y C. En la parte superior se muestra el HTH descrito en Crp, PurR, FruR y LacI. Mientras que en la familia LysR se indica el HTH con respecto a las predicciones reportadas para LysR, OxyR y CysB, así como por las predicciones hechas en este trabajo. Se muestran las posiciones conservadas (*) dentro del H-T-H y alrededor del mismo. Los puntos indican menor conservación por posición. Los patrones adicionales a la región donde se localiza el HTH corresponden a sitios de interacción con inductores, descritos en miembros de la familia LysR, así como en proteínas de la familia DeoR.

En la familia LysR detectamos tres motivos o patrones: El primer patrón incluye al dominio de pegado al DNA (localizado en el N-terminal), mientras que los dos restantes (localizados en el C-terminal) estarían implicados en el reconocimiento al inductor (Schell, 1993). Figura 5. A continuación daremos una descripción de los grupos relacionados en términos funcionales (tabla 4):

Supergrupo

A) De todas las proteínas con el HTH en el N-terminal, el 43.3% forma parte del supergrupo.

B) El 47.4% de las proteínas con el HTH en el intervalo del 10 al 20% del N-terminal se localizan en el supergrupo. Alrededor del 80% de todos los miembros de las familias relacionadas fueron detectados en el proceso de búsqueda,

C) Crp y Fnr son los únicos reguladores del supergrupo que presentan el HTH en el C-terminal. Crp ha sido propuesto como miembro de la superfamilia Crp donde se incluyen proteínas como LexA, BirA (con el HTH en el N-terminal), así como el represor de la toxina de la difteria (DtxR) (Holm et al.1994). En el supergrupo, detectamos un homólogo de DtxR.

C) El supergrupo incluye a más de la mitad de las proteínas represoras (alrededor del 55%) reportadas en *E. coli*. Sólo unas pocas proteínas con actividad dual son incluídas.

D) La mayoría de los miembros del supergrupo estarían implicados en la regulación de los genes para la asimilación de fuentes de carbono.

En conclusión, las proteínas incluidas en el supergrupo contienen a la mayoría de los represores reportados en el N-terminal y contiene a varias familias con propiedades funcionales homogéneas.

Familia LysR

A) Esta familia agrupa alrededor del 65% del total de los reguladores duales. Es decir, los reguladores de esta familia activan diversos genes y se autoregulan así mismos.

B) El HTH se localiza hacia el primer 10% N-terminal de la secuencia.

C) La mayoría de los miembros de la familia regulan genes para la biosíntesis de aminoácidos.

D) Su dirección de transcripción va en contra de la dirección de replicación en el 71% de sus genesEn el genoma completo de *E. coli* hay una proporción del 50% para cada grupo de genes (divergentes y paralelos a la replicación).

Características	Supergrupo	LysR	Colección
Represor	55.1%	1.8%	43.0%
Activador	16.4%	4.4%	80.0%
Dual	14.0%	65.5%	20.1%
Paralelo	44.0%	29.0%	48.5%
Divergente	56.0%	71.0%	51.5%

Tabla 4. Comparación funcional entre los grupos detectados. 55% de las proteínas activadoras, 16% de los represores y 14% de los reguladores duales de la colección están incluídos en el supergrupo. 44% de los miembros del supergrupo presentan una dirección paralela de transcripción/replicación, mientras que el 56% es antiparalela. En la familia LysR, se agrupa el 1.8% de los represores, 4.4% de los activadores y el 65.5% de los reguladores duales de la colección. Alrededor del 70% de sus genes tienen una dirección antiparalela de replicación/transcripción.

Discusión

El HTH es una estructura de pegado a nucleótidos muy antigua, descrita en arqueobacterias, tales como en *Methanococcus jannaschi* y *Pyrococcus furiosus* (Kyrpides et al.1995), así como en proteínas que estabilizan el RNA ribosomal (Yonath et al. 1997. Xing et al.1997). Este motivo esta ampliamente diversificado entre los reguladores transcripcionales tanto de procariotes y eucariotes (Treisman et al.1992).

Una de las preguntas a responder es si las proteínas reguladoras con HTH presentan un solo origen o han sido creados en diferentes momentos de la evolución. Las evidencias que hemos colectado en este trabajo apoyan la idea de un ancestro común para un gran supergrupo de reguladores que comparten la posición del HTH, la función reguladora y las funciones metabólicas reguladas. La existencia de un segundo grupo –representado por la familia LysR- con el HTH en la misma región que el supergrupo puede indicarnos la posibilidad de que el tiempo de divergencia ente un grupo y otro es tan antigua que no podemos detectarlo por los métodos existentes o bien, existe la posibilidad de diferentes orígenes evolutivos.

En el caso de la familia LysR, posiciones cruciales en el plegamiento del HTH no están conservados, a diferencia del supergrupo que si las presenta (Wintjens et al.1997). Por ejemplo, en la familia LysR no está presente la glicina de la vuelta, el residuo 5 de la primer α-hélice tiene una cadena lateral ramificada (en la estructura del HTH clásico no se presentan ningún aminoácido ramificado) y los residuos 8 y 10 están cargados. En resumen, la familia LysR no presenta las restricciones estereoquímicas impuestas en el HTH, por lo que cabe la posibilidad de que esta familia se haya derivado paralelamente al HTH de los represores.

En este contexto, cada grupo funcional de proteínas reguladoras (mencionados anteriormente) formarían un grupo evolutivo donde la posición del dominio de pegado al DNA se conserva. Nuestra hipótesis, es que un regulador evolucionó como un ancestro de las proteínas represoras y activadoras. Posteriormente, los dominios que responderían a señales inductoras se adicionarían para formar una sola entidad funcional que respondería a las diferentes condiciones ambientales. El HTH, por lo tanto, quedaría englobado en un dominio pequeño que se localizaría cerca de la región terminal de la proteína. Propuestas similares han sido analizadas para la construcción del HTH (Yura et al.1993), y en algunas familias de reguladores, tales como en los dos componentes (Pao et al. 1995), GalR/LacI (Vartak et al. 1991) y NagR/XylR (Titgemeyer et al.1994).

Cabe señalar que para ambos grupos de reguladores se realizó un análisis filogénetico, utilizando los métodos de máxima parsimonia y de distancia genética. Sin embargo, para el caso del supergrupo HTH-N-terminal, el tamaño del patrón no es suficiente para inferir una historia evolutiva, por lo cual para la propuesta del grupo no fue tomado en cuenta dicho análisis.

DISCUSION GENERAL

El genoma completo de *E. coli* cepa K12 ofrece una de las mejores oportunidades para dilucidar el panorama completo del mecanismo de la regulación transcripcional. En este trabajo nos hemos enfocado en describir desde dos perspectivas a las proteínas reguladoras de la transcripción. El primer enfoque hace énfasis en la compilación de los factores transcripcionales caracterizados experimentalmente y predichos por análisis computacionales. El segundo enfoque hace referencia a

la descripción de dos grupos de proteínas reguladoras que posiblemente comparten diferentes orígenes. Pero ¿qué hemos aportado con la compilación de los reguladores transcripcionales?.

A) La colección de los 314 reguladores transcripcionales de *E. coli* K12 refleja -posiblemente- el total de factores que *E. coli* necesita para regular todos sus genes, si consideramos que hay una relación de 1 gene regulador por cada 10 o 12 genes regulados. Esto significa, que alrededor del 10% de los productos génicos de *E. coli* se destinan a regular el inicio de la transcripción. Aproximadamente la mitad de los reguladores descritos en la colección han sido descritos experimentalmente, mientras que la otra mitad son propuestos en este trabajo.

B) El motivo de unión al DNA o HTH es uno de los mejores indicadores para describir a un regulador transcripcional y para asignarle una función. Es decir, si describimos un HTH en una proteína y además se localiza en un extremo proteico es muy probable que sea un regulador transcripcional al cual se le puede asignar la función de represor o activador.

C) Las proteínas reguladoras de *E. coli* se han agrupado en familias, lo que ha evidenciado que algunas familias presenten una gran cantidad de miembros (LysR con 45 miembros), mientras que hay familias donde sólo se presentan uno o dos miembros (Crp). Es posible que algunas familias han duplicado su número de genes al interior de *E. coli* y por lo tanto es posible predecir que estás familias serán homogéneas en términos de secuencia y función. Ver el artículo "The Repertoire of DNA-Binding Transcriptional Regulators in *Escherichia coli*".

D) En términos funcionales, qué significa este número de reguladores y la gran cantidad de miembros de una misma familia?. Una posible explicación es que *E. coli* es una bacteria que tiene la capacidad de habitar y explotar diferentes ambientes. Esta característica de *E. coli* implicaría también que la bacteria está sometida a diferentes presiones de selección y que provocaría un aumento en el número de genes reguladores para contender con dichas condiciones.

En conclusión, este trabajo aporta tanto la descripción funcional de los reguladores y además intenta dar una explicación al por qué de la presencia de estos genes reguladores.

En relación a la segunda parte del trabajo, ¿qué significa la presencia de grupos evolutivamente relacionados con el HTH?.

Pensando en que los reguladores transcripcionales han sido posiblemente originados a partir de un ancestro común, hemos intentado describirlos en términos de regiones o patrones compartidos. Estos patrones reflejan la existencia de al menos dos grupos claramente delimitados, tanto por secuencia como por función. En consecuencia, podemos definir un grupo de represores y un grupo de proteínas duales en el extremo N-terminal. La existencia de diferencias en aminoácidos que ocupan posiciones fundamentales en el motivo de unión al DNA nos hace pensar que ambos grupos se han originado a partir de dos eventos evolutivos diferentes o bien que el tiempo de divergencia es tal que aún somos incapaces de detectar relaciones evolutivas entre ellos. En resumen, la presencia de grupos funcionales al interior de todos los reguladores transcripcionales no sólo de *E. coli* sino también de procariotes, nos hace pensar en que la posición del HTH es una de las mejores huellas evolutivas que valdría la pena rastrear entre todas las bacterias.

PERSPECTIVAS GENERALES

Con la descripción de los reguladores transcripcionales y la propuesta del supergrupo HTH N-terminal, se abren varias líneas de investigación que podrían ser abordadas y que mencionamos a continuación

- Con el fin de determinar la antigüedad de los mecanismos regulatorios, hemos empezado la detección y descripción de proteínas homólogas a los reguladores de *E. coli* en otros genomas, tales como en la arqueobacteria *Methanococcus jannaschi*. En esta bacteria detectamos alrededor de 17 posibles reguladores transcripcionales utilizando al regulador DeoR de *E. coli* como secuencia *query*. De los 17 posibles reguladores, describimos a 12 como represores y 5 como activadores (utilizando la posición del HTH como indicador de función). De las predicciones realizadas, cuatro proteínas se describen en el genoma como reguladoras y una de ellas como un posible represor. Análisis similares deberán de ser realizados en todos los genomas disponibles. Con la detección de los reguladores transcripcionales en otros genomas, se pueden responder varias preguntas: ¿Cuál es el panorama de regulación transcripcional en otros organismos?, ¿Cómo han evolucionado dichos sistemas de regulación? y ¿La conservación de funciones

reguladas por los reguladores en *E. coli* es similar en otros organismos?. En conclusión, la detección de los reguladores transcripcionales (utilizando un enfoque de búsqueda similar al de *E. coli*) en otros genomas representa la oportunidad para trazar la historia evolutiva de la transcripción en otros sistemas biológicos con un enfoque más integrado gracias a la existencia de genomas completos.

- La detección de los sitios de regulación en *E. coli* para todos los reguladores transcripcionales será de gran importancia para complementar desde la perspectiva del DNA, los sitios de interacción DNA-proteína. En RegulonDB se han almacenado los sitios de pegado con evidencia experimental, para alrededor de 50 proteínas reguladoras. Creemos que si utilizamos los sitios que corresponden a miembros de una misma familia y generamos matrices de búsqueda genérica para varias proteínas, se localizarán los sitios de pegado al DNA para las proteínas tanto conocidas como predichas.
- La realización de experimentos donde se hagan mutantes de los reguladores predichos será importante para determinar y/o solidificar las predicciones realizadas por métodos computacionales. Un primer candidato para realizar el experimento es YeiL_B2163 una proteína homologa –en *fold*- a Crp/Fnr y que bien podría ser otro regulador global. Un primer análisis podría ser probando la mutante en medio mínimo y con diferentes fuentes de carbono, describiendo el fenotipo de la mutante.
- Buscar relaciones entre las proteínas que presentan el HTH en el C-terminal, para determinar si presentan uno o más posibles orígenes evolutivos.
- El análisis posicional de los motivos proteicos que ha sido utilizado en este trabajo será relavante para realizar predicciones en proteínas con posible función de regulador transcripcional. En el caso de las proteínas que se unen al DNA con el motivo HTH, la posición de esta estructura dentro de la secuencia nos ha ayudado a determinar si una proteína es reguladora y cuál es su función más probable, utilizando el mismo enfoque en otros genomas completos se puede dilucidar la función de cada uno de los posibles reguladores transcripcionales que se detecten.
- ¿Qué otro tipo de motivos estructurales pueden ser utilizados como indicadores evolutivos en las proteínas?. En un análisis preliminar en proteínas de la familia de las EBP (Enhancer Binding

Proteins) se calculó la posición de los diferentes motivos funcionales. Con base en este análisis encontramos que el HTH está localizado en los últimos 35 residuos de aminoácidos del extremo C-terminal. La distancia de este motivo con respecto al sitio de interacción con el factor sigma es de alrededor 160 residuos y la distancia entre el sitio de interacción de la proteína-factor sigma con respecto al posible sitio de interacción con el ATP es de alrededor 80 residuos. Por lo que la organización de las proteínas de esta familia sería de la siguiente forma: extremo N-terminal---ATP *binding site*-**80 aa**--Factor sigma—**160 aa**--HTH---C-terminal.

- Finalmente, estamos diseñando una herramienta que utilice la predicción del *fold* y la detección de patrones en secuencias proteicas, esto con el fin de determinar con un enfoque estructural la presencia de reguladores y/o proteínas que se asocian al DNA en otros genomas.

REFERENCIAS

Ariza, R.R., Li, Z., Ringstad, N. & Demple, B. (1995) J Bacteriol 177, 1655-1661

Bairoch A. (1992) Nucleic Acids Res 20, 2013-3018

Bairoch A. (1993) Nucleic Acids Res 21, 2515

Beckwith, J. (1987) in *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology (Neidhardt, F.C., Ingraham, J.L., Low, B., Magasanik, B., Schaechter M. & Umbarger HM, eds.), pp. 1439-1443, American Society for Microbiology, Washington, D. C.

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. (1997) Science 277, 1453-1474

Blattner, F.R., Plunkett, G., Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F., Gregor, J., Davis, N.W., Kirkpatrick, H.A., Goeden, M.A., Rose, D.J., Mau, B. & Shao, Y. (1997) Science 277, 1453-1474

Brennan, R.G. & Matthews, B.W. (1989) J Biol Chem 264, 1903-1906

Branden, C. & Tooze, J. (1991) In Introduction to Protein Structure. pp. 85-110. NY and London. Garland Publishing, Inc..

Bustos, S.A. & Schleif, R.F. (1993) Proc Natl Acad Sci U S A 90, 5638-5642

Charlier, D., Roovers, M., Van, V.F., Boyen, A., Cunin, R., Nakamura, Y., Glansdorff, N. & Pierard, A. (1992) Journal of Molecular Biology 226, 367-386

Collado-Vides, J., Magasanik, B. & Gralla, J.D. (1991) Microbiol Rev 55, 371-394

Corpet, F., Gouzy, J. & Kahn, D. (1998) Nucleic Acids Res 26, 323-326

Corpet, F., Gouzy, J. & Kahn, D. (1999) Nucleic Acids Res 27, 263-267

De Mot, R., Nagy, I., Schoofs, G., & Vanderleyden J.(1996) Curr Microbiol 33:26-30

Ferrandez, A., Garcia, J.L. & Diaz, E. (1997) J Bacteriol 179, 2573-81

Gallegos, M.T., Schleif, R., Bairoch, A., Hofmann, K. & Ramos, J.L. (1997) Microbiol Mol Biol Rev 61, 393-410

Gouzy, J., Corpet, F. & Kahn, D. (1996) Trends Biochem Sci 21, 493

Gralla, J.D. (1990) Methods Enzymol 185, 37-54

Gralla, J.D. (1991) Cell 66, 415-418

Gralla, J.D. (1996) Genetics and Development 6, 526-530

Gross, C,A., Lonetto, M. & Losick, R. (1992) in Transcriptional Regulation (McKnight, S.L. & Yamamoto, K.R., eds.), pp. 129-176, Cold Spring Harbor Laboratory Press, New York

Harrison, S.C. & Aggarwal, A.K. (1990) Annu Rev Biochem 59, 933-969

Haydon, D.J. & Guest, J.R. (1991) FEMS Microbiol Lett 63, 291-295

Henikoff, S., Haughn, G.W., Calvo, J.M. & Wallace, J.C. (1988) Proceedings of the National Academy Of Sciences Of The United States Of America 85, 6602-6606

Holm, L., Sander, C., Rüterjans, H., Schnarr, M., Fogh, R., Boelens, R. & Kaptein. R. (1994) Prot. Engin. 7,1449-1453.

Huerta, A.M., Salgado, H., Thieffry, D. & Collado-Vides, J. (1998) Nucleic Acids Res 26, 55-59

Ishihama, A. (1993) J Bacteriol 175, 2483-2489

Jair, K.W., Martin, R.G., Rosner, J.L., Fujita, N., Ishihama, A. & Wolf, R.E. Jr. (1995) J Bacteriol 177, 7100-7104

Koonin, E.V., Tatusov, R.L. & Rudd, K.E. (1995) Proc Natl Acad Sci U S A 92, 11921-11925

Kyrpides, N.C. & Ouzounis, C.A. (1995) Trends Biochem Sci 20, 140-1

Labedan, B. & Riley, M. (1995) J Bacteriol 177, 1585-1588

Lawrence, J.G. & Ochman, H. (1998) Proc Natl Acad Sci U S A 95, 9413-9417

Lewin, B. (1994) Genes V, Oxford University Press,

Lewis, M., Chang, G., Horton, N.C., Kercher, M.A., Pace, H.C., Schumacher, M.A., Brennan, R.G. & Lu, P. (1996) Science 271, 1247-1254

Lisser, S. & Margalit, H. (1993) Nucleic Acids Res 21, 1507-1516

Lonetto, M., Gribskov, M. & Gross, C.A. (1992) J Bacteriol 174, 3843-3849

Matthews, K.S. (1992) Microbiol Rev 56,123-36

McClure, W.R. (1985) Annu Rev Biochem 54, 171-204

McClure, W.R. (1980) Proc Natl Acad Sci U S A 77, 5634-5638

Mooney, R.A., Artsimovitch, I. & Landick, R. (1998) J Bacteriol 180, 3265-3275

Morett, E. & Segovia, L. (1993) J Bacteriol 175, 6067-6074

Neidhardt, F. & Savageau, M. (1996) in Escherichia coli and Salmonella typhimurium: cellular and molecular biology (Neidhardt, F.C., Curtiss, R.III, Ingraham, J.L., Lin, E.C.C., Low, B., Magasanik, B., Reznikoff, W., Riley M, , Schaechter M, & Umbarger HM, , eds.), pp. 792-821, American Society for Microbiology, Washington, D. C.

Neidhardt, F.C., Ingraham, J.L. & Schaech, M. (1990) Physiology of the bacterial cell. A Molecular Approach,

Neuwald, A.F., Liu, J.S., Lipman, D.J. & Lawrence, C.E. (1997) Nucleic Acids Res 25, 1665-1677

Nguyen, C.C. & Saier, M.H. Jr. (1995) FEBS lett 377:98-102

Ohlendorf, D.H., Anderson, W.F. & Matthews, B.W. (1983) J Mol Evol 19, 109-114

Pao, G.M. & Saier, M.J.Jr. (1995) J Mol Evol 40, 136-154

Penin F., Georjon, C., Montserret, R., Bockmann, A., Lesage, A., Yang, Y., Bonod-Bidaud, C., Cortay, J.C., Negre, D., Cozzone, A.J. & Deleage, G. (1997) J. Mol. Biol. 270,496-510.

Record, M.T., Reznikoff, W.S., Craig, M.L., McQuade, K.L. & Schlax, P.J. (1996) in *Escherichia coli* and *Salmonella typhimurium*: cellular and molecular biology (Neidhardt, F.C., Curtiss, R.III, Ingraham, J.L., Lin, E.C.C., Low, B., Magasanik, B., Reznikoff, W., Riley M, , Schaechter M, & Umbarger HM, , eds.), pp. 792-821, American Society for Microbiology, Washington, D. C.

Ross, W., Gosink, K.K., Salomon, J., Igarashi, K., Zou, C., Ishihama, A., Severinov, K. & Gourse, R.L. (1993) Science 262, 1407-1413

Salgado, H., Santos, A., Garza-Ramos, U., van Helden, J., Diaz, E. & Collado-Vides, J. (1999) Nucleic Acids Res 27, 59-60

Sauer, R.T., Yocum, R.R., Doolittle, R. F., Lewis, M. & Pabo, C. O. (1982) Nature. 298,447-451

Schell, M.A.(1993). Ann. Rev. Microbiol. 597-626.

Schumacher, M.A., Choi, K.Y., Zalkin, H. & Brennan, R.G. (1994) J Mol Biol 242, 302-305

Spiro S. (1993) Antonie Van Leeuwenhoek. 66, 23-36

Steitz, T.A., Ohlendorf, D.H., McKay, D.B., Anderson, W.F. & Matthews, B.W. (1982) Proc Natl Acad Sci U S A 79, 3097-100

Stock, J.B., Ninfa, A.J. & Stock, A.M. (1989) Microbiol. Rev. 53, 450-490

Sulavik, M.C., Gambino, L.F. & Miller, P.F. (1995) Mol Med 1:436-46

Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., Abola, E.E.(1998) Acta Crystallogr D Biol Crystallogr 54, 1078-84

Suzuki, M. & Brenner S.E. (1995) FEBS Lett. 372, 215-21

Suzuki, M., Yagi, N. & Gerstein, M. (1995) Protein Eng 8, 329-338

Thieffry, D., Huerta, A.M., Pérez-Rueda, E. & Collado-Vides, J. (1998) Bioessays 20, 433-440

Titgemeyer, F., J. Reizer, A. Reizer &Saier, M. H. Jr. (1994) Microbiology. 140,2349-2354

Tobin, J.F. & Schleif, R.F. (1990) Journal of Molecular Biology 211, 75-90

Treisman, J., Harris, E., Wilson, D. & Desplan, C. (1992) Bioessays 14, 145-150

Valentin-Hansen, P., Hojrup, P. & Short, S. (1985) Nucleic Acids Res 13, 5927-36

Vartak, N.B., Reizer, J., Reizer, A., Gripp, J.T., Groisman, E.A., Wu, L.F., Tomich, J.M. & Saier, M.H.Jr. (1991) Res. Microb. 142, 951-963

von Hippel, P.H. (1998) Science 281, 660-665

von Hippel, P.H., Bear, D.G., Morgan, W.D. & McSwiggen, J.A. (1984) Annu Rev Biochem 53, 389-446

Weber, I.T., McKay, D.B. & Steitz, T.A. (1982) Nucleic Acids Res 10, 5085-102

Weickert, M.J. & Adhya, S. (1992) Journal of Biological Chemistry 267, 15869-15874

Wharton, R.P., Brown, E.L. & Ptashne, M. (1984) Cell. 38, 361-9

Wintjens, R. & Rooman, M. (1997) J. Mol. Biol. 262, 294-313

Xing, Y. Guha Thakurta D. & Draper D.E. Nat Struct Biol 4, 24-7

Yonath A. Franceschi F. (1997) Nat Struct Biol 4, 3-5

Yura, K., Tomoda, S. & Go, M. (1993) protein eng 6, 6

Zeng, Q., Stalhandske, C,, Anderson, M.C., Scott, R.A. & Summers, A.O. (1998) Biochemistry 37, 15885-95