INSTITUTO DE BIOTECNOLOGÍA

“Análisis de operones divergentes conservados en procariontes para la identificación de potenciales sitios de unión a factores transcripcionales”

T E S I S

QUE PARA OBTENER EL GRADO DE MAESTRÍA EN CIENCIAS BIOQUÍMICAS

Patricia María Rufina Oliver Ocaño

DIRECTOR DE TESIS:
Dr. Enrique Merino Pérez

Cuernavaca, Morelos **Enero del 2008**

Tutor Principal:

Dr. Enrique Merino Pérez
Instituto de Biotecnología (IBT), UNAM.

Comité tutoral:

Dr. Lorenzo Segovia Forcella
Instituto de Biotecnología (IBT), UNAM.

Dra. Alicia González Manjárrez
Instituto de Fisiología Celular (IFC), UNAM.

Miembros del Jurado:

Dr. Martín Peralta Gil
CCG–UNAM

Dr. Guillermo Gosset Lagarda
IBT–UNAM

Dr. Enrique Merino Pérez
IBT–UNAM

Dr. Víctor Bustamante Santillán
IBT–UNAM

Dr. Ernesto Pérez Rueda
IBT–UNAM

Agradecimientos

A Arturo Medrano Soto, por tu adecuada presencia siempre que tengo cara de What? Pero sobre todo, por tu profundo amor, te quiero muchísimo.

A mis padres, Ana y Francisco.

A Elany,Miguelito,Naomi, Diego, Millyam, Claudia Iraíz, Paco, y bebé, mis sobrinitos.

A J, por tu presencia y tu cariño.

Un agradecimiento especial a mis amigos lejanos porque siempre están en mi corazón: Pablo; Félix, Mandis, Viole; Chamín; Vladimir. Y a otros más cercanos por acompañar mis días, Vero, Irma Vichido; Freddy; Héctor; Josué; Nocturno; Monik; Adán; Santiago, Sarita.

Un agradecimiento muy especial al Dr. Martín Peralta, por su amistad y su apoyo substancial en la realización de esta tesis, gracias por regresarme la confianza.

A las discusiones y valiosos comentarios de mi amigo Sarath Chandra Janga que del mismo modo fueron fundamentales para el proyecto.

A Kike Merino, que más que un jefe ha sido un amigo, gracias por tu paciencia, tu apoyo, tu nobleza, tu carisma y tu alegría, siempre encuentro algo que aprender de ti.

Gracias a mis amigos y compañeros del Instituto, les agradezco su compañía, su alegría, sus palabras de ánimo y sus aventones en las noches frías :P.

Por brindarme su tiempo, sus comentarios y sugerencias: gracias a Rosa María Gutiérrez; Cei Abreu; Ruy Jáuregui; Santiago Ramírez; Heladia Salgado; Raúl Noguez; María Luisa Tabche; Alejandro García Rubio; Javier Díaz; Zuemy Rodríguez; Mario Núñez; Lorenzo Segovia.

A mis compañeros del laboratorio Merino-Spin, porque los días son más alegres con la armonía que se percibe aquí.

Este proyecto se llevó a cabo bajo Fuentes de Financiamiento CONACYT: 44213-Q y DGAPA: IN203705, IX210204.

A las personas que más amo y que nunca faltan.

"¡No corras, ve despacio,

que a donde tienes que ir es a ti solo!

¡Ve despacio, no corras,

que el niño de tu yo, recién nacido

eterno,

no te puede seguir!"

Juan Ramón Jiménez

(premio novel de literatura 1956, obra: Eternidades, 1918)

CONTENIDO

RESUMEN

La conservación de operones estructuralmente equivalentes en diferentes genomas constituye una evidencia de la importancia de la vecindad de los genes en el cromosoma para permitir su regulación coordinada. De manera similar, se ha identificado a la cercanía cromosomal de los genes divergentes como parte de otro mecanismo de corregulación. Particularmente se han estudiado aquellos casos donde el factor transcripcional del sistema se autorregula y regula al operón divergente, pero a la fecha aún existe una gran cantidad de factores transcripcionales cuyos sitios de unión todavía se desconocen. Por esta razón, nuestro interés se orientó al estudio de los sitios de autorregulación de factores transcripcionales divergentes (FTDs) empleando un enfoque comparativo.

En esta tesis se presenta un análisis multigenómico sobre los sitios de unión que reconocen los FTDs. La estrategia se enfocó especialmente a la identificación de los sitios de unión de los FTDs conocidos haciendo un análisis comparativo del conjunto de unidades transcripcionales divergentes ortólogas (UTDOs) entre 176 genomas eubacterianos no redundantes (nr). Se utilizó a *Escherichia coli K12* como organismo modelo para definir los pares de unidades transcripcionales divergentes (UTD) referencia que tuvieran al menos un factor transcripcional. Empleando una metodología para la identificación de huellas filogenéticas, nuestro objetivo fue evaluar si las predicciones de los sitios de unión de los FTDs corresponden a lo descrito en la literatura. Argumentamos que las matrices obtenidas en nuestro estudio, a partir de regiones multigenómicas, conservan más los sitios de unión en comparación con los reportados por métodos que utilizan secuencias que pertenecen a un solo genoma.

De los 176 genomas analizados se trabajó con 48 grupos de UTDOs, de éstos, sólo 18 grupos tienen sitios de unión al DNA reportados y nuestro método logró recuperar el 83% de los sitios. Posteriormente se realizó la predicción de sitios de unión para los 16 grupos de FTDs con actividad conocida (pero sin sitios de unión caracterizados) y para 14 grupos de FTDs hipotéticos. Se logró proponer sitios de unión probables para el 50% y 64% respectivamente, de dichos grupos de FTDs.

1 INTRODUCCION

El mecanismo por el cual los microorganismos pueden adaptarse y sobrevivir ante cambios internos y condiciones medioambientales, involucra una serie de modificaciones genéticas que permite alterar el metabolismo de las bacterias (71). La mayoría de los estímulos externos son percibidos por la bacteria a través de sistemas de transducción de señales denominados sistemas de dos componentes (64). Por otra parte, la mayoría de los estímulos internos son provocados por componentes solubles en el citoplasma, ya sea por que el agente externo logró penetrar en la célula, o bien por simple acumulación de metabolitos internos que pueden afectar directamente el comportamiento de las proteínas que regulan la transcripción, logrando mediar con ello una gran parte de la respuesta celular y su modificación en el metabolismo (45). Por tal razón es fundamental que los microorganismos puedan percibir inmediatamente dichos cambios para activar o reprimir en su momento a los genes relacionados con el estímulo en cuestión.

El genoma de la bacteria está compuesto por miles de genes, que desde el punto de vista de la expresión genética, pueden ser constitutivos, cuando se transcriben permanente, e independientemente de las condiciones medioambientales; o genes de expresión regulada que resultan muy interesantes dado que se expresan en función de las condiciones del medio y que pueden ser inducibles o reprimibles. Para controlar la activación y represión de los genes, la célula utiliza mecanismos de regulación genética muy variados y estrategias rápidas y eficientes que le permiten economizar energía al momento de regular un grupo de genes simultáneamente.

La expresión coordinada de los genes involucra toda una red de sistemas de control en el que están representados los distintos niveles de regulación genética. Durante la regulación **a nivel de la transcripción**, los genes pueden ser activados o reprimidos para alterar la síntesis de mRNA. El procesamiento o degradación del mRNA corresponde a la regulación **traduccional**. En la regulación **a nivel postraduccional**, los productos o proteínas pueden ser modificados o

degradados. Y por último, la regulación **a nivel postranscripcional** se refiere a las modificaciones que sufren los productos generados durante la transcripción (mRNA) y la traducción (proteínas). En bacterias, el control crítico y de regulación primaria se da fundamentalmente al inicio de la transcripción, en donde generalmente están implicadas proteínas

reguladoras, comúnmente denominadas factores de transcripción (FT), que se unen a secuencias específicas de DNA con el objetivo de controlar la expresión de un operón de genes estructurales (5).

1.1 Mecanismo de la regulación transcripcional

En 1961, Francois Jacob y Jaques Monod formularon un mecanismo básico sobre el control de la transcripción en bacterias (21), en el cual proponen que la expresión de los genes está regulada por las interacciones específicas que ocurren entre secuencias cercanas (*cis*) localizadas en el DNA con los productos (FTs) de secuencias codificadas en regiones a distancia (*trans)*. Ahora sabemos que la transcripción de los genes puede además regularse de otras formas, como por ejemplo, a través de las modificaciones en la RNA polimerasa (RNApol) (36), o rearreglos en la estructura del DNA que favorecen la activación o represión de promotores particulares, debido a la acción de las proteínas reguladoras que modifican la topología del DNA (25, 41) o bien, a través de pequeñas moléculas de RNA que activan o inhiben indirectamente la transcripción (32).

Uno de los mecanismos de regulación transcripcional que emplea la célula para controlar la cantidad y los tiempos de aparición del producto de un gen, es mediante la utilización de diferentes subunidades de la RNApol (24). La RNApol es la enzima central de la transcripción, que cataliza la síntesis de RNA y esta constituida por múltiples subunidades. La parte principal o corazón de la enzima se compone de dos subunidades α, dos subunidades β (β,β'), una subunidad ω, y adicionalmente la subunidad σ, que resulta ser el elemento principal para el reconocimiento de promotores (12). A su vez, cada subunidad α consiste de un dominio amino terminal (NTD) y un dominio carboxilo terminal (CTD). Los dos CTDs se unen a regiones río arriba ricas en A/T

(5). Las subunidades β y β' comprenden el sitio activo de la enzima y la subunidad ω actúa como una chaperona para ayudar al correcto plegamiento de la subunidad β'(**Figura 1**). La subunidad σ esta formada por cuatro dominios que reconocen los sitios de las cajas -10, -35 y el -10 extendido (12; 24).

En general, la mayoría de las bacterias cuentan con un factor σ estándar que está presente en condiciones normales de crecimiento y reconoce a la mayoría de los promotores del genoma. Este factor σ puede ser sustituido por otro tipo de σ alternativo bajo condiciones específicas de crecimiento, con lo que la RNApol adquirirá la capacidad de reconocer e iniciar la transcripción a partir de un tipo distinto de promotor (29), permitiéndole transcribir operones que permanecían inactivos durante el crecimiento vegetativo. Un ejemplo es la respuesta al choque por calor que experimenta *E. coli K12*, que al someterse a una agresión de altas temperaturas, induce la expresión de más de 30 genes cuyos productos tienden a evitar ciertos daños ocasionados por este agente. Esta respuesta está mediada por la subunidad σ (σ^{32}) que desplaza a la σ^{70} (16; 72). La nueva holoenzima reconoce los genes de la respuesta al calor, los cuales poseen un promotor con una región -10 totalmente diferente a la estándar (56).

Otro de los mecanismos es la regulación del inicio de la transcripción mediado por FTs. En el esquema de la **Figura 1**, se representan los elementos necesarios para el mecanismo mediado por FTs: inicio de transcripción, promotor, sitios de unión de los FTs y regiones de unión a las subunidades α. El extremo 5´ del mRNA (+1) esta representando el sitio de inicio de la transcripción. Río arriba del sitio de inicio (+1) se localiza el promotor, el cual esta compuesto por dos hexámeros que se ubican en las posiciones -10 y -35. Las secuencias consenso de σ70, de las cajas -10 y –35, están representadas por los hexámeros TATAAT y TTGACA respectivamente, y su conservación permite altas tasas de transcripción. El promotor es reconocido por el factor σ y facilita la unión de la RNApol para formar el complejo abierto e iniciar la transcripción de los genes. Río arriba del promotor se encuentran secuencias específicas que son reconocidas por FTs, que al unirse al DNA reclutan a la RNApol facilitando su unión, y del mismo modo ayudan a la formación del complejo abierto de la transcripción.

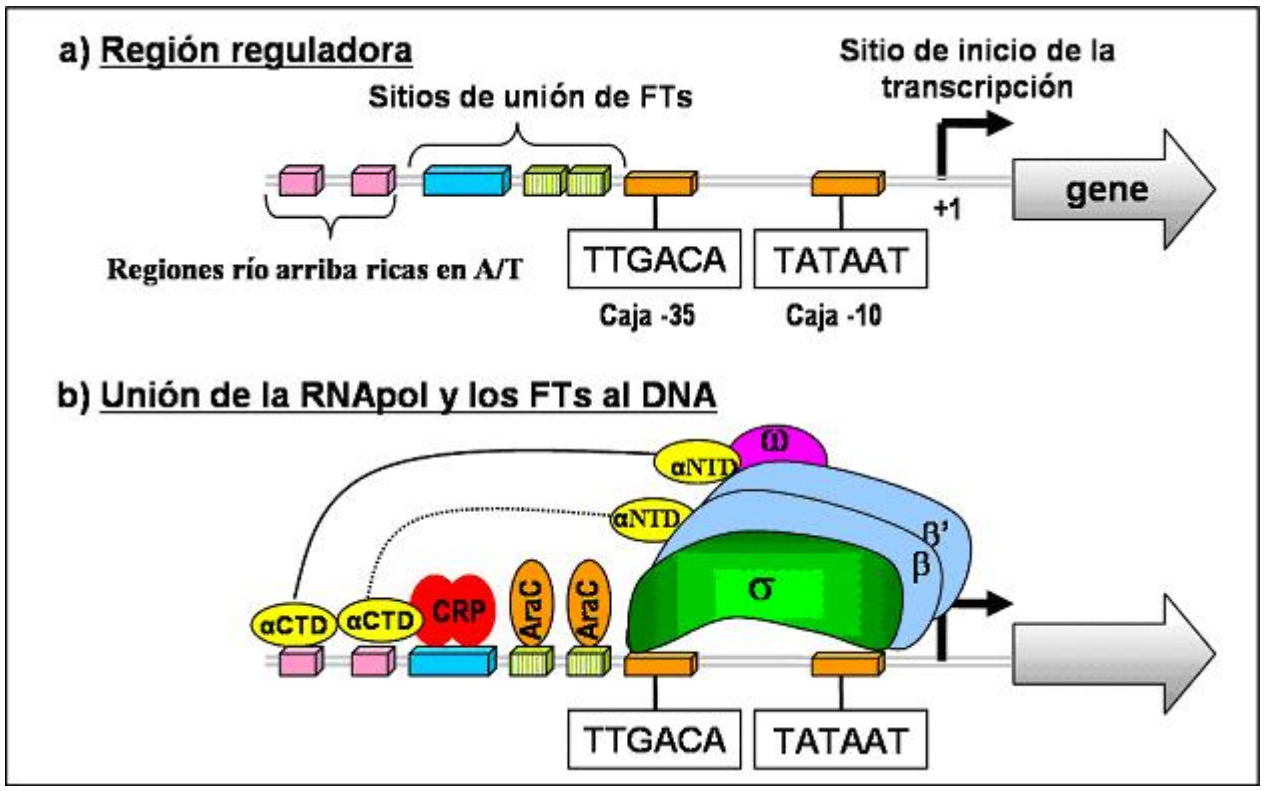


Figura 1. Mecanismo de la regulación del inicio de la transcripción en procariontes.

1.2 Factores de transcripción (FTs)

En general, los FTs pueden ser activadores o represores (23). Los activadores transcripcionales son FTs que pueden unirse a secuencias específicas de DNA cercanas al promotor y promueven la unión de la RNApol, para dar inicio a la transcripción (**Figura 1**); este reclutamiento se debe mayoritariamente a un contacto directo entre el activador y la RNApol (5;

58). En otros casos, la activación sucede indirectamente por interrupción de la represión, donde el activador puede unirse a secuencias que se ubican lejanas al promotor (58). Los represores actúan pegándose a sitios específicos de DNA, ubicados cerca de los promotores, una vez unidos bloquean la unión de la RNApol para impedir el inicio de la transcripción (17). Un FT puede ser también bifuncional y actuar como activador o represor dependiendo del contexto.

La regulación genética mediada por FTs se lleva a cabo mediante fenómenos de inducción y represión genética, los cuales a su vez, pueden realizarse por mecanismos de control positivo y negativo (**Figura 2, Figura 3**). Estos mecanismos se disparan en el interior de la célula debido a estímulos ambientales químicos relacionados con el metabolismo, que informan a la bacteria sobre un determinado cambio, lo que desencadena la respuesta celular (47).

El mecanismo de control negativo por efectos de inducción (*e.g.* el operón *lac*), ocurre cuando interviene un inductor, un metabolito que funciona como sustrato de una ruta metabólica, que desencadena la reacción. A la izquierda de la **Figura 2**, el FT con actividad represora *per se* es activo, a menos que el inductor se le una y lo inactive, induciendo la expresión de los genes que codifican las enzimas que a su vez, pueden metabolizarlos. Para el mecanismo de control negativo por efectos represores (*e.g.* el operón *trp*), ocurre que el represor, se encuentra del mismo modo inactivo (aporrepresor), pero en presencia del correpresor (así le llamaríamos en esta reacción al inductor), se activa y reprime la expresión del operón estructural. En ausencia del correpresor se promueve la transcripción dado que el represor se mantiene inactivo (derecha de la **Figura 2**).

Por el contrario, para el mecanismo de control positivo por efectos de inducción, hablamos de un FT activador que se encuentra inactivo, pero una vez que se le une el inductor, lo activa y promueve la transcripción. En ausencia del inductor, el FT permanecería inactivo y la transcripción no se realiza (izquierda de la **Figura 3**). El mecanismo de control positivo por efectos represores, se refiere a un FT activador que *per se* es activo, pero cuando se le une el correpresor, lo inactiva, y evita que se lleve a cabo la transcripción, en ausencia del correpresor el activador promueve la transcripción (derecha de la **Figura 3**).

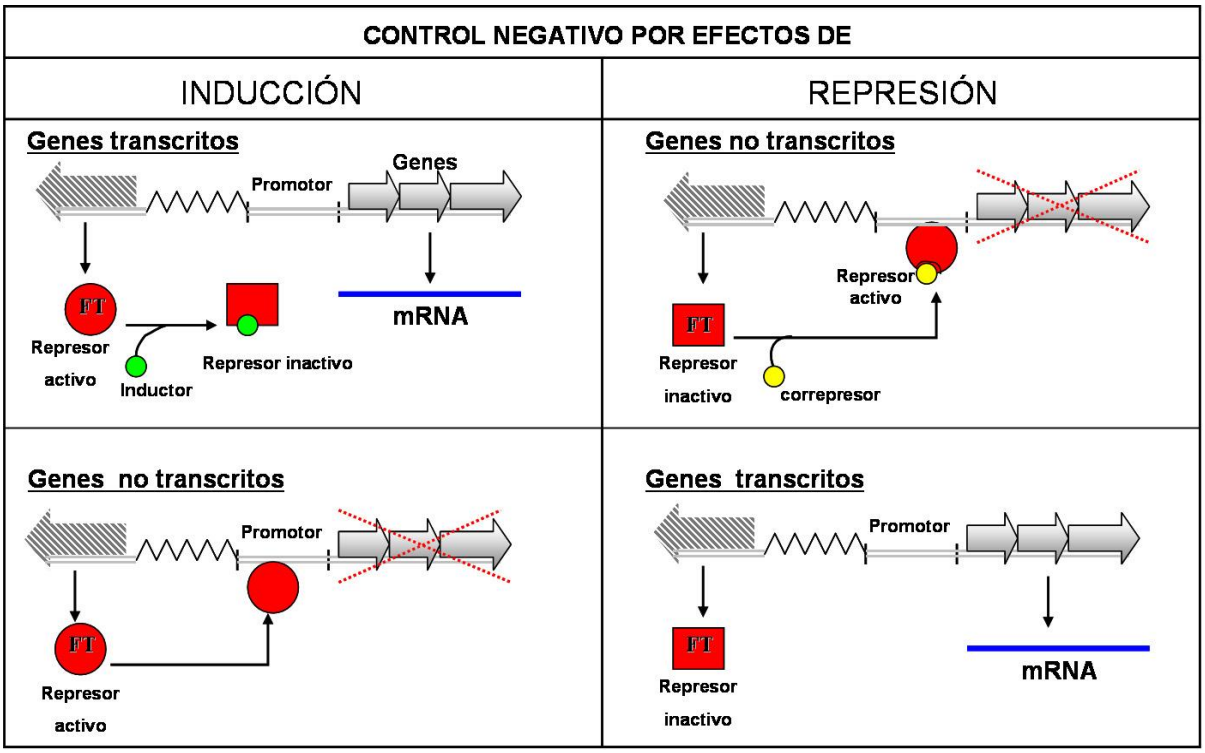


Figura 2. Para el control negativo hablamos siempre de un FT represor, que puede ser afectado por un inductor que mantiene al FT activo (izq.), o por un correpresor, cuando el FT se inactiva (der).

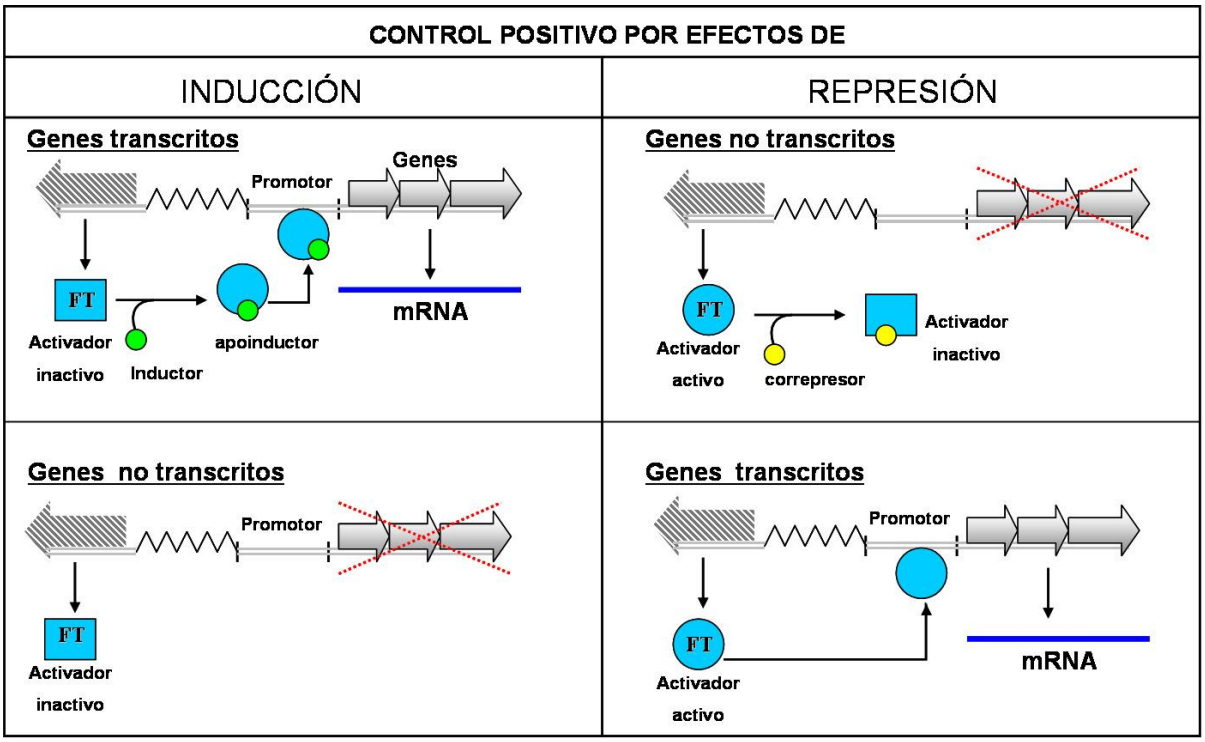


Figura 3. En el control positivo siempre nos referimos a un FT activador, que puede inactivarse o activarse al ser afectado por un inductor o un correpresor respectivamente.

1.2.1 Clasificación de los FTs

Los FTs pueden ser **reguladores globales**, cuando controlan coordinadamente la activación o represión de gran cantidad de genes y operones (*e.g.* CRP, FNR, FIS, H-NS, IHF, Lrp y ArcA) (46); **locales** cuando están encargados de regular sistemas de genes con funciones particulares (reparación de DNA por LexA (15); biosíntesis de biotina por BirA (75), degradación de arabinosa por AraC) (22). Entre los FT se encuentran los **reguladores de respuesta (RR)** que son parte de los sistemas de dos componentes (un sensor y un regulador de respuesta) (64). En este sistema la señal ambiental no entra a la célula, sino que es detectada por un sensor a nivel de membrana, el cual reemite el estímulo hacia una proteína citoplásmica, que a su vez interacciona con secuencias determinadas al comienzo del operón para regularlo, generando así la respuesta adaptativa correspondiente a la señal ambiental (64). Las proteínas sensoras pertenecen a la familia de histidín-proteín-quinasas (HPK), y son capaces de autofosforilarse cuando detectan algún estímulo procedente del ambiente y cambian de conformación (36), pero su función

principal es la de fosforilar al correspondiente RR, que tras ser forsforilado ejerce algún efecto regulatorio, normalmente como activadores de la transcripción de ciertos operones (*e.g*. regulón NarL; regulón OmpR) (57; 69).

1.2.2 Sitios de unión de FTs

Los FT son de suma importancia en los procesos biológicos de los microorganismos y realizan su función uniéndose a regiones específicas del DNA denominados sitios de unión (5). En términos generales, los sitios de unión de los FT son definidos como regiones regulatorias en el DNA (regiones-*cis*) conformadas por secuencias específicas, con tamaños y direcciones definidas, separados por nucleótidos (nt) poco conservados. Muchos de esos sitios están representados por pequeños motivos de secuencias arregladas en forma de invertidos repetidos (**IR**), directos repetidos (**DR**) o convergentes (**CV**), que pueden o no tener espacios variables entre los motivos (**Figura 4**). Los sitios de unión pequeños tienen longitudes de 5 a 7 pares de bases (pb), mientras que los sitios grandes tienen longitudes entre 17 a 26 pb.

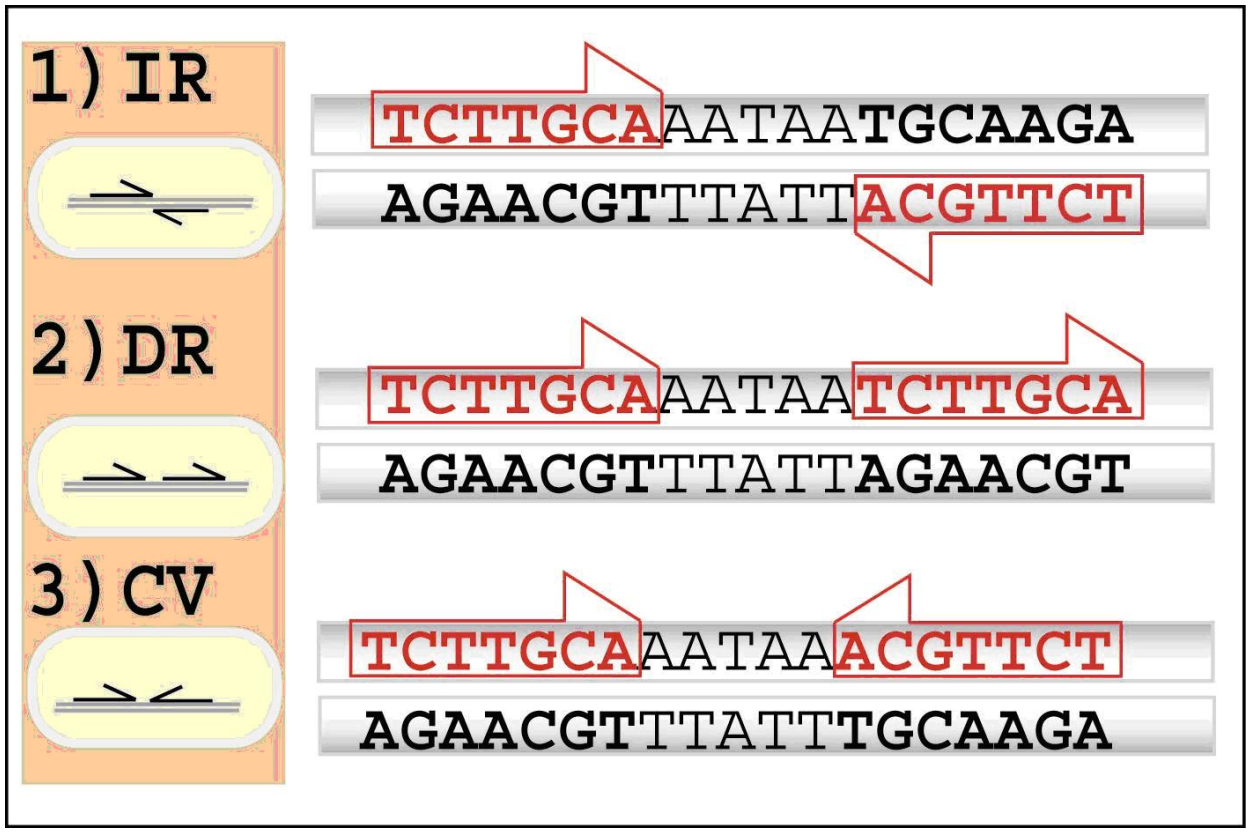


Figura 4. Ejemplos de la dirección de los sitios de unión encontrados en el DNA. 1) invertidos repetidos (IR), 2) directos repetidos (DR) y 3) convergentes (CV).

Cuando un FT regula varios genes, su sitio de unión puede encontrarse repetidas veces en el genoma. El alineamiento de los diferentes sitios de unión de un FT en particular, permite la obtención de una secuencia general que muestra un sitio idealizado o secuencia consenso que puede ser representada por un LOGO. En Bioinformática, dicho consenso se conoce como patrón o motivo, y al conjunto de técnicas y métodos estadístico-computacionales empleados para su identificación se le llama generalmente "reconocimiento o búsqueda de patrones". La secuencia consenso representa un posible sitio en el DNA con alta afinidad a su FT, mientras que los sitios de unión que difieren mucho a la secuencia consenso tendrán baja afinidad. Por ejemplo, la secuencia consenso de CRP es: AAATGTGAnnnnnnTCACATTT, donde la región espaciadora de 6 nt ubicada a la mitad del sitio puede ser representada por cualquier nt (78).

Los FTs que funcionan regulando de manera global se unen a múltiples sitios para controlar la expresión coordinada de diversos genes. Las regiones de DNA donde este tipo de

reguladores se unen varían mucho en su composición si se les compara, es decir, no son muy conservadas y pueden tener una gran gama de afinidades (78). Por otro lado, los FTs modulares o locales, se unen a muchos menos sitios y su papel en la regulación está restringido a unos cuantos genes involucrados en una función común (*e.g.* biosíntesis de leucina). Otra diferencia importante respecto a los FTs globales es que los sitios de unión al DNA para las FTs modulares están muy conservados y por lo tanto son sitios con alta afinidad (22).

1.2.3 Identificación de los sitios de unión en el DNA

Se han desarrollado varias metodologías para identificar, en el DNA, los sitios de unión de los FTs. Experimentalmente se ha usado la técnica de retardamiento en gel (EMSA), el cual es un método, *in vitro*, para el estudio de interacciones proteína-DNA. Los complejos proteína-DNA se moverán lentamente a través de un gel de electroforesis, respecto a la sonda radiactiva de DNA que no tiene la proteína unida. Las bandas radiactivas se pueden visualizar discretamente en geles de acrilamida. Sin embargo, este ensayo no es suficiente para localizar con exactitud la posición del sitio de contacto.

El ensayo que puede complementarlo es el que permite observar la huella de la posición del sitio de unión. La técnica de impresión de la huella o "footprinting" permite estudiar interacciones proteína-DNA e identificar la secuencia de unión en el DNA, a la cual se une la proteína. La molécula de DNA que lleva la proteína unida estará protegida de cualquier modificación o degradación. De esta forma, al tratar el complejo proteína-DNA con una nucleasa, para romper todos los enlaces fosfodiéster, se protegerán solamente aquellos nt protegidos por la proteína. Posteriormente se remueve la proteína del complejo y el DNA se resuelve en geles de poliacrilamida. Los espacios corresponderán a la huella de la posición del sitio de unión y a partir de ellos se obtienen las secuencias de unión al FT.

3 Organización genética de los FTs en el genoma

En las bacterias se han desarrollado gran variedad de arreglos genéticos que permiten coordinar grupos de genes con funciones relacionadas. A un conjunto de genes adyacentes que se cotranscriben en un mismo mensajero policistrónico se le denomina **operón** (21). Asimismo, al total de genes y operones regulados por el mismo FT se le conoce como **regulón** (30).

Además, hemos notado que la disposición de los genes codificadores de FTs en el genoma, es un punto importante que la célula ha desarrollado a su favor, con el fin de economizar energía y facilitar la regulación de sus genes en respuesta a las necesidades del medio donde crece. En el caso de los genes que codifican **FTs globales,** éstos se han identificado posicionados de manera individual en el DNA, sin observar alguna relación funcional con sus genes adyacentes (**Figura 5a**). Los genes que codifican **FTs del tipo RR**, implicadas en los sistemas de dos componentes, se organizan comúnmente en operones, lo que permite su regulación coordinada (**Figura 5b**). A su vez, los genes codificadores de **FTs locales** se organizan de diversas formas (**Figura 5c**). La gran ventaja de colocar al FT junto a los genes que regula es que se facilita el trabajo a la célula en sus mecanismos de regulación.

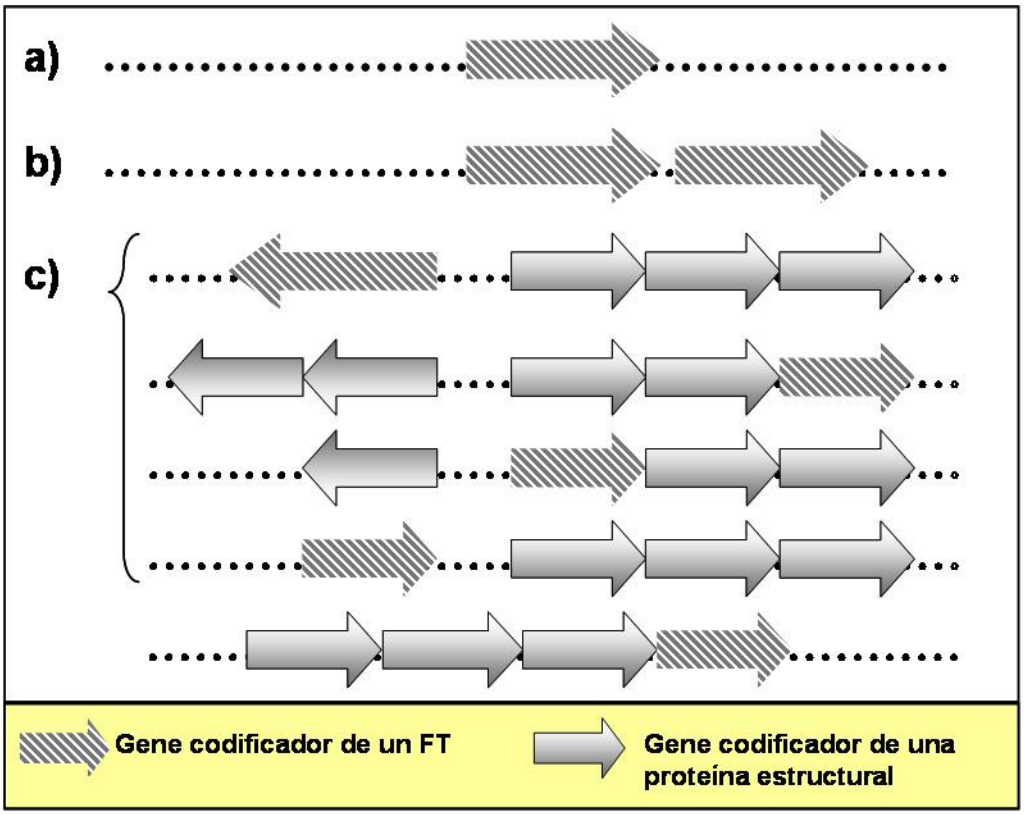


Figura 5. Disposición de los genes que codifican para FTs en el genoma. a) Reguladores globales, b) Reguladores de respuesta (sistema de dos componentes), y c) Los reguladores locales, que pueden encontrarse en disposiciones genéticas muy variadas respecto a los genes que regulan.

4 Corregulación en unidades de transcripción divergentes

Siempre que hablemos de dos unidades transcripcionales adyacentes cuyas direcciones de transcripción son contrarias (una UT ubicada en una cadena de DNA inmediatamente seguida de una UT en la cadena opuesta) nos referiremos a ellas como **U**nidades **T**ranscripcionales **D**ivergentes (UTDs). Cada UTD comparte una región de DNA no codificador (región intergénica) en la cual se pueden encontrar elementos implicados en su regulación, como regiones promotoras y potenciales sitios de unión a FTs (51).

A la izquierda de la **Figura 6**, se ilustra la clasificación de los UTDs en tres categorías de acuerdo a la función de los productos génicos involucrados: a) P-P, cuando ambos productos son proteínas estructurales (usualmente enzimas o transportadores) (10) b) R-R, cuando ambos genes codifican proteínas reguladoras denominadas Factores Transcripcionales Divergentes (FTDs) (73). c) R-P, cuando uno de los genes es un FTD y el otro un gen estructural (55). Asimismo, la **Figura 6** esquematiza los tres tipos de arreglos de promotores divergentes que pueden identificarse en las regiones intergénicas de las UTDs (6; 76): 1) Con los promotores opuestos y un fragmento de DNA entre ellos que puede contener sitios de reconocimiento para FTs, 2) promotores opuestos traslapados (*e.g*. reconocidos por diferentes factores σ, o dos promotores para un mismo factor sigma pero desfasados por uno o varios nt) y 3) con los promotores posicionados de frente (traslapados y/o contenidos en la región codificadora de los genes) (13).

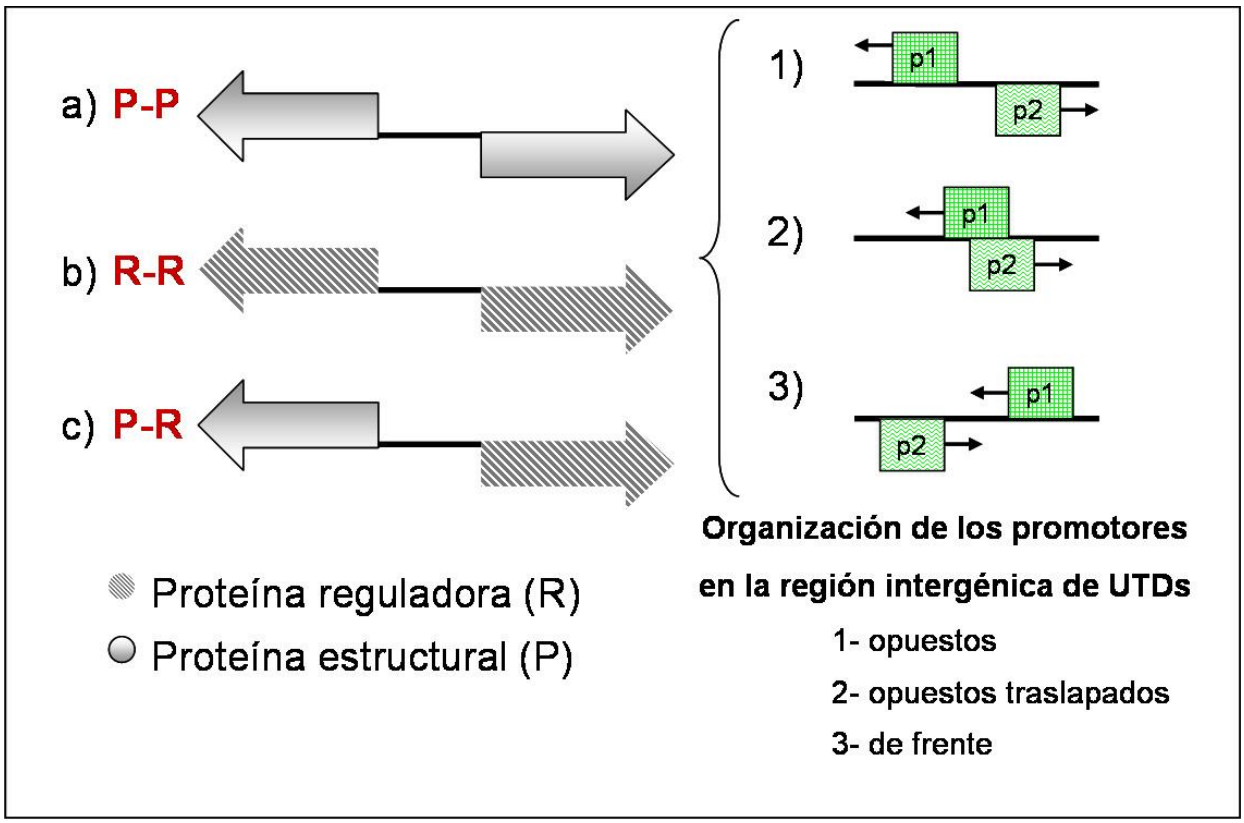


Figura 6. Clasificación de los UTDs de acuerdo a la función de los genes divergentes (*izq*) y según el tipo de promotores posibles de encontrar en su región intergénica (*der*). Los promotores se representan por cajas verdes y las flechas indican el sentido del promotor.

1.4.1 Ejemplos de UTDs

Los mecanismos de regulación divergente se empezaron a estudiar desde hace tres décadas y se han realizado tanto en fagos como en bacterias. Como se ha mencionado anteriormente, uno de los arreglos divergentes se constituye por genes que codifican proteínas estructurales y pueden ser regulados simultáneamente por uno o varios reguladores o ser constitutivo (caso P-P **Figura 6**). Para los otros dos casos (P-R y R-R) se ha logrado demostrar que los FTDs juegan un papel muy importante en la autorregulación de UTDs, la cual puede ser positiva o negativa (**Figura 2, Figura 3**).

En general, los FTDs reconocen sitios específicos de unión con orientaciones bien definidas (**Figura 4**), que pueden identificarse en ambas cadenas del DNA de la región intergénica compartida entre el par de UTDs.

Para esquematizar los ejemplos de organización divergente con base en los productos génicos se han seleccionado tres ejemplos que se describen a continuación, éstos son el caso del sistema de reguladores involucrados en el estrés oxidativo (55), la degradación de arabinosa (74) y la degradación de maltosa (7) en la bacteria *E. coli.*

1.4.1.1 El sistema de estrés oxidativo soxRS

El sistema *soxRS* (**Figura 7**) presenta la topología R-R. SoxR es un regulador transcripcional dual que pertenece a la familia de MerR. Esta proteína se transcribe constitutivamente y se encarga de percibir el estado redox de la célula activándose por exposición a agentes externos como el óxido nítrico (40). En la región intergénica de la UTD *soxRS*, se reporta un sitio de unión de SoxR con orientación de IR, la posición central del sitio está ubicada a dos pb de distancia del sitio de inicio de la transcripción y traslapa completamente la caja -10 del promotor, lo que evita la unión de la RNApol y la formación del complejo abierto. De esta manera, en su forma oxidada SoxR se autorreprime y activa de manera indirecta la expresión de *soxS* (31).

A su vez, existe evidencia experimental de que el regulador SoxS puede autorreprimirse y activar la transcripción de otras UTs involucradas en incrementar la resistencia a oxidantes y antibióticos. Aunque la ubicación de su sitio de unión aún no ha sido reportado para esta UTD, ya se ha caracterizado su secuencia consenso como IR en el regulón (55).

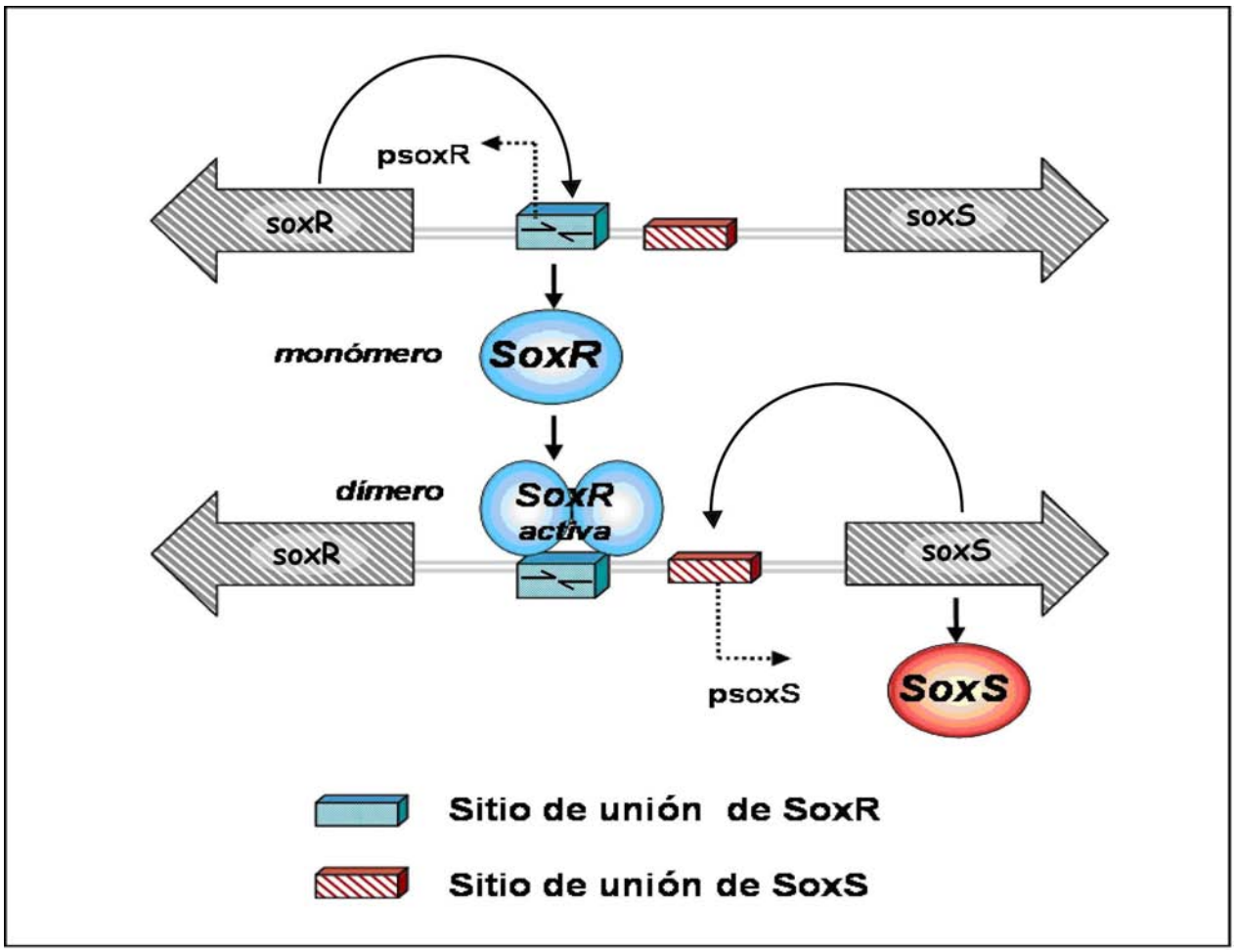


Figura 7. Mecanismo de regulación *soxR soxS* (55). Ver discusión en el texto.

2 Sistema de degradación de arabinosa

El par de UTDs implicados en la degradación de arabinosa se conforma por el gen regulador *araC* y el operón *araBAD,* que codifica para las proteínas catabólicas AraA, AraB y AraD (topología R-P). AraC puede actuar como represor o activador dependiendo de la presencia o ausencia de arabinosa en el medio (74) . En la región intergénica entre *araC* y *araBAD* existe un sitio de reconocimiento para CRP (requerido para la activación del promotor *araBAD*) y cinco sitios de unión para AraC: las regiones *araO2* y *araO1* están representados por dos sitios de unión cada uno, mientras que en *araI* se encuentra el quinto sitio de reconocimiento de AraC (**Figura 8**). En ausencia de arabinosa y presencia de glucosa, dos monómeros de AraC se unen a los sitios $araI_1$ y $araO_2$ respectivamente, lo cual origina la formación de un bucle en el DNA debido a la estabilización que se logra entre los sitios $araI_1$ y $araO_2$ (22). La presencia de este bucle impide la activación del promotor *araBAD* por la RNApol y es así como se reprime la expresión de ambas unidades de transcripción (39). En presencia del sustrato y cuando los niveles de glucosa son limitados, se forman dímeros de la proteína AraC que se unen a la región $araO_2$ y activan la expresión del operón *araBAD* (63).

Los cinco sitios de unión de este FT son regiones de 17 pb que se representan como DR codireccionales (en tándem) en la misma cadena de DNA, su secuencia consenso esta representada por dos submotivos separados por 7nt: nAGCnnnnnnnTCCATA.

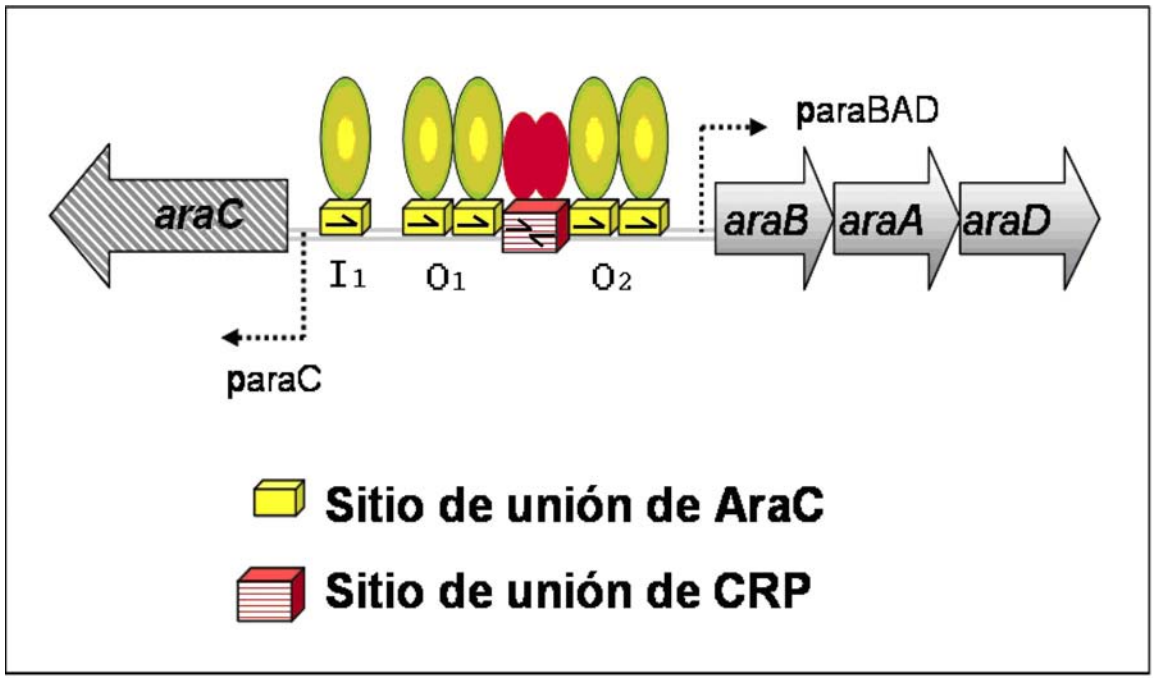


Figura 8. Corregulación divergente de *araC* y el operón *araBAD* (44; 62; 63).

1.4.1.3 Sistema de degradación de maltosa

Este sistema es un ejemplo del arreglo estructural P-P que involucra a los operones *malEFG* y *malK-lamB-malM,* los cuales se encuentran separados por una región intergénica de 271 pb (10). En la **Figura 9** se muestra cómo la UTD es regulada por el FT externo MalT (10). Este regulador se une a cinco sitios DR en tándem, en la región divergente, que flanquean a cuatro sitios de alta afinidad para CRP (11). La acción sinérgica de CRP y MalT forma una estructura que es responsable de la activación de ambos promotores y, por consecuencia, de la transcripción bidireccional de la UTD (7; 59). Los sitios de unión de MalT son pequeños con una longitud de 10 pb, y su secuencia consenso es: GGGGAGGAGG (11).

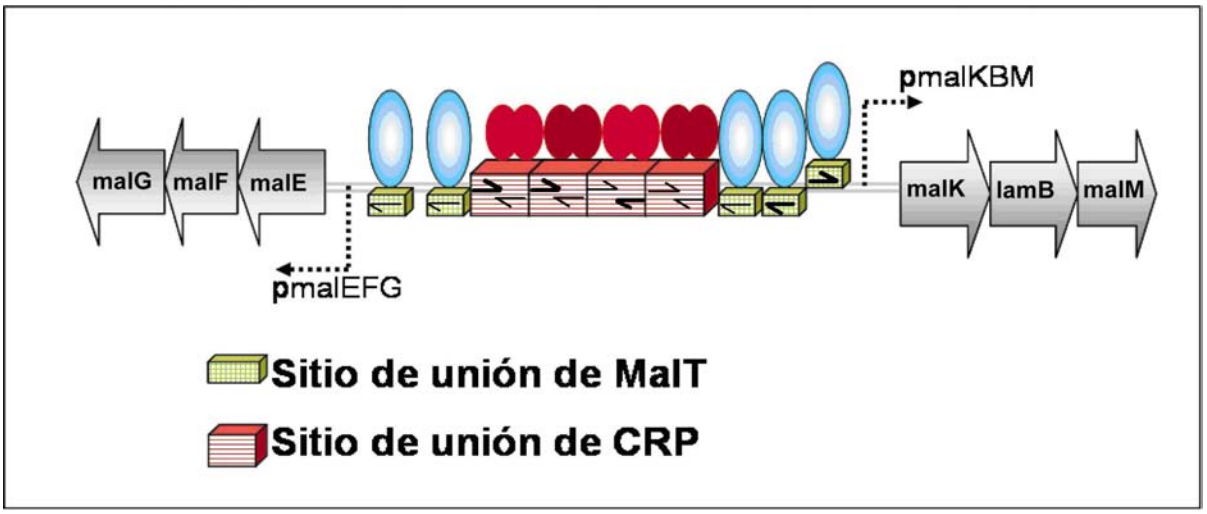


Figura 9. Sistema de regulación de genes involucrados en la degradación de maltosa por la acción cooperativa de los reguladores CRP y MalT.

Además de los ejemplos aquí descritos, existen otros casos que ayudan a entender la complejidad del encendido y apagado de genes divergentes, como por ejemplo la biosíntesis de arginina (26), la expresión de los genes *tetA* y *tetR* relacionados con resistencia a tetraciclina (35), entre otros. Una compilación de sesenta casos caracterizados experimentalmente se puede encontrar en la referencia (13).

2 ANTECEDENTES

Opel *et al*. describe que aproximadamente el 40% de las unidades de transcripción en este microorganismo se transcriben en forma divergente (51), y aproximadamente el 60% de las UTDs contienen uno o más genes que codifican a un regulador específico. Por tal razón, a la fecha se han estudiado numerosos mecanismos de regulación que involucran UTDs, como los mencionados en la introducción (sección 1.4.1). Este caudal de conocimiento ha permitido entender otros procesos biológicos muy complejos. Por ejemplo, la incidencia de genes conservados en operones ha sido ampliamente utilizada para predecir relaciones funcionales entre genes de organismos procariontes (48). Además de la estructura del operón (33), la conservación de las secuencias regulatorias en las regiones intergénicas de genes divergentes podría ser evidencia de corregulación genética. En particular, los genes adyacentes transcritos a partir de promotores divergentes constituyen una forma de organización común en procariontes (13). El mismo fenómeno se ha observado en organismos superiores como *Saccharomyces cerevisiae* (8; 42) y *Homo sapiens* (1; 38; 68).

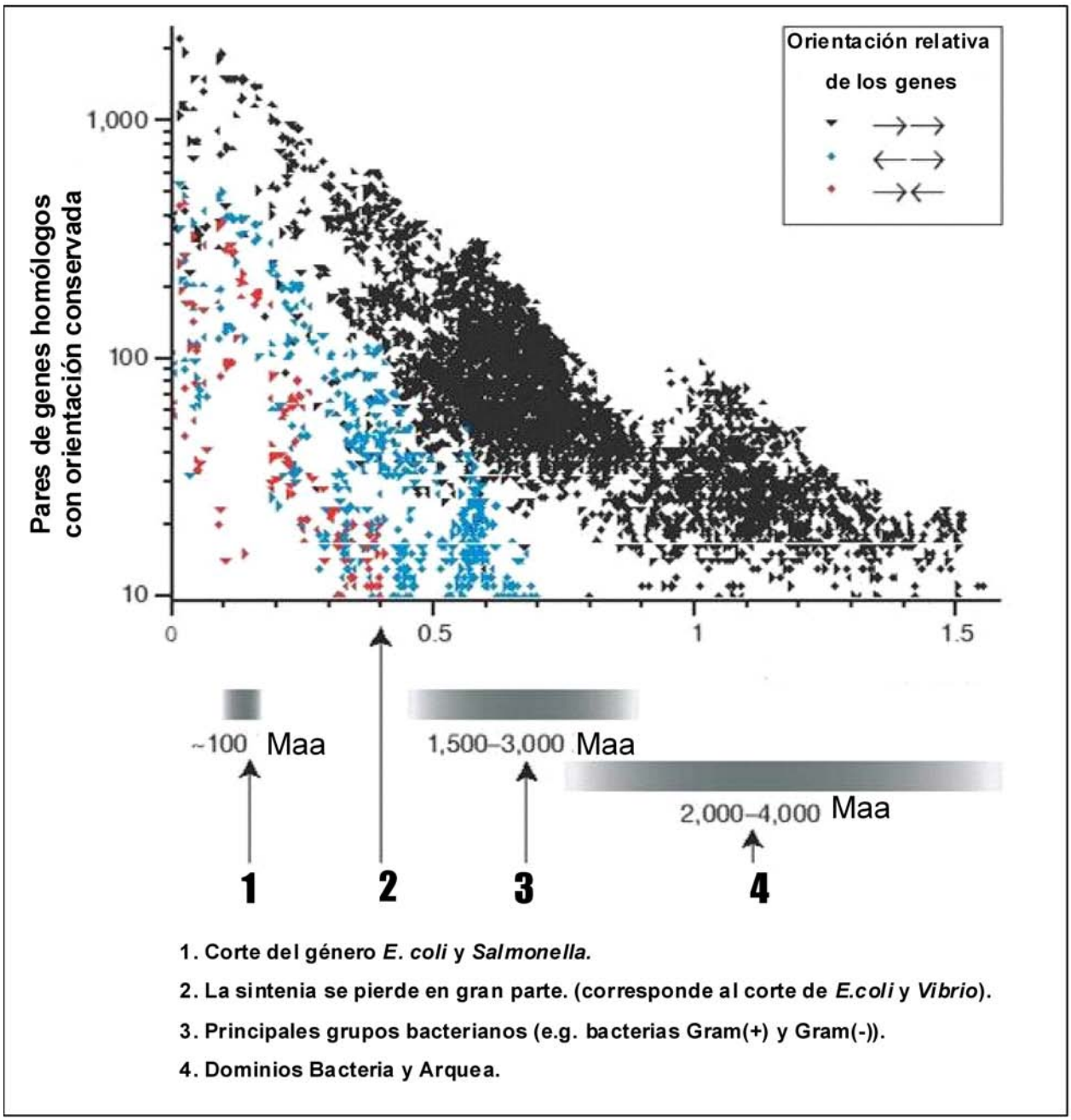
Por otro lado, el estudio de las relaciones funcionales entre genes requiere de diversas estrategias como la fusión de genes o piedra *Rosetta* (45), la vecindad genética (18; 52) y la ocurrencia de genes entre genomas (perfiles filogenéticos) (19; 54). Recientemente, un trabajo apoyado en experimentos de microarreglos, evaluó la conservación evolutiva de unidades transcripcionales divergentes (UTDs), convergentes y en tándem en organismos procariontes (37). Aunque los autores reportaron que la organización de los genes en tándem es razonablemente más frecuente, también observaron que la orientación divergente se conserva más que la convergente especialmente en eubacterias (**Figura 10**), y propusieron una nueva estrategia que explota la conservación de los pares de genes transcritos divergentemente (bidireccionales) para predecir asociaciones funcionales (37). Estos hallazgos abren la posibilidad de que en las regiones intergénicas comunes a genes divergentes *homólogos*, se conserven también las secuencias de reconocimiento de proteínas encargadas de regular la transcripción de los genes. Hasta donde nosotros sabemos, no existe a la fecha un estudio a nivel genómico enfocado a los mecanismos de corregulación observados en UTDs y sobre la

conservación de sitios de regulación en sus regiones intergénicas. Esta tesis está inspirada en esa observación.

La identificación de motivos puede hacerse tanto a nivel de secuencias de DNA, como, en el caso de proteínas, a nivel de secuencias de aminoácidos (2). En este trabajo nos concentramos en la identificación de motivos en las secuencias de DNA.

Dado que la tasa de variación es substancialmente diferente entre organismos filogenéticamente distantes, la conservación de un motivo (un probable sitio de regulación) entre ellos podría indicar la existencia de una presión selectiva por mantener una misma secuencia que es reconocida por factores transcripcionales *ortólogos*. No obstante, en organismos filogenéticamente muy lejanos, es posible que los FT no se conserven o que la secuencia de reconocimiento para un mismo tipo de FT haya variado tanto que sea imposible de identificar mediante análisis comparativo de secuencias. Por este motivo, en la presente tesis sólo son considerados organismos filogenéticamente cercanos, en este caso al grupo de las eubacterias, pero al mismo tiempo, se eliminaron los genomas más parecidos entre si.

Además de las técnicas experimentales, mencionadas en la introducción, también existen herramientas computacionales disponibles para la identificación de sitios de regulación y el estudio sistemático del contexto genómico. Algunos de los programas más comunes para el análisis de sitios de regulación son: AlignACE (32), Consensus/Patser (30), MotifSampler (66), Oligo/Dyad-analysis (70) y Weeder (53), entre otros. En el presente trabajo utilizamos los programas MEME-MAST (4) para identificar motivos sobre representados. Su descripción se encuentra en la Sección de Metodología (5.1.4).



1. Corte del género *E. coli* y *Salmonella.*
2. La sintenia se pierde en gran parte. (corresponde al corte de *E.coli* y *Vibrio*).
3. Principales grupos bacterianos (e.g. bacterias Gram(+) y Gram(-)).
4. Dominios Bacteria y Arquea.

Figura 10. Conservación evolutiva de la orientación de genes vecinos entre linajes procariontes. En la gráfica se aprecia que la orientación de genes divergentes se conserva más que la convergente, y se mantiene principalmente en la rama de las eubacterias (Maa = Millones de años atrás). Figura modificada de (37).

3 JUSTIFICACIÓN

A la fecha existen varios métodos de localización de huellas filogenéticas para la identificación de sitios de unión de factores transcripcionales, sin embargo, la utilidad de estas predicciones solamente se puede conocer si se realiza un análisis biológico exhaustivo individual de cada uno de los casos predichos. Nuestra estrategia es un análisis multigenómico que se enfoca en identificar sitios conservados en regiones intergénicas comunes a Unidades Transcripcionales Divergentes Ortólogas (UTDO), donde es probable que existan sitios reales de unión a Factores Transcripcionales Divergentes (FTDs). Si nuestro enfoque nos permite localizar los sitios de autorregulación caracterizados de los FTDs conocidos, puede ser de mucha utilidad extender el análisis y proponer sitios de unión para aquellos FTDs conocidos que no tienen sitio de unión identificado o predecir los sitios de unión a FTDs hipotéticos.

4 OBJETIVO GENERAL

Desarrollar una estrategia computacional que permita la identificación de los sitios de unión de Factores Transcripcionales Divergentes, con base en la identificación de huellas filogenéticas conservadas río arriba de Unidades Transcripcionales Divergentes Ortólogas.

1 Objetivos Particulares

i) Identificar al conjunto de UTDs ortólogos entre genomas bacterianos no redundantes.

ii) Desarrollar la estrategia computacional para identificar sitios de unión conservados en regiones intergénicas de UTDOs conocidas.

iii) Evaluar la capacidad predictiva de nuestra estrategia, revisando la congruencia de los resultados, con lo reportado en la literatura.

5 MÉTODOS

Dado el razonamiento de nuestra justificación de trabajo, el empleo de metodologías para identificar huellas filogenéticas de sitios de unión, comúnmente denominadas “Phylogenetic footprinting”(9), es útil en el diseño de una estrategia para someter a prueba la hipótesis del proyecto. Si un número significativo de sitios conocidos son encontrados en UTDs ortólogos (no muy cercanos filogenéticamente), se podría argumentar que adicionalmente a la conservación de la organización transcripcional divergente, se mantienen también los sitios de unión de FTDs. En l a

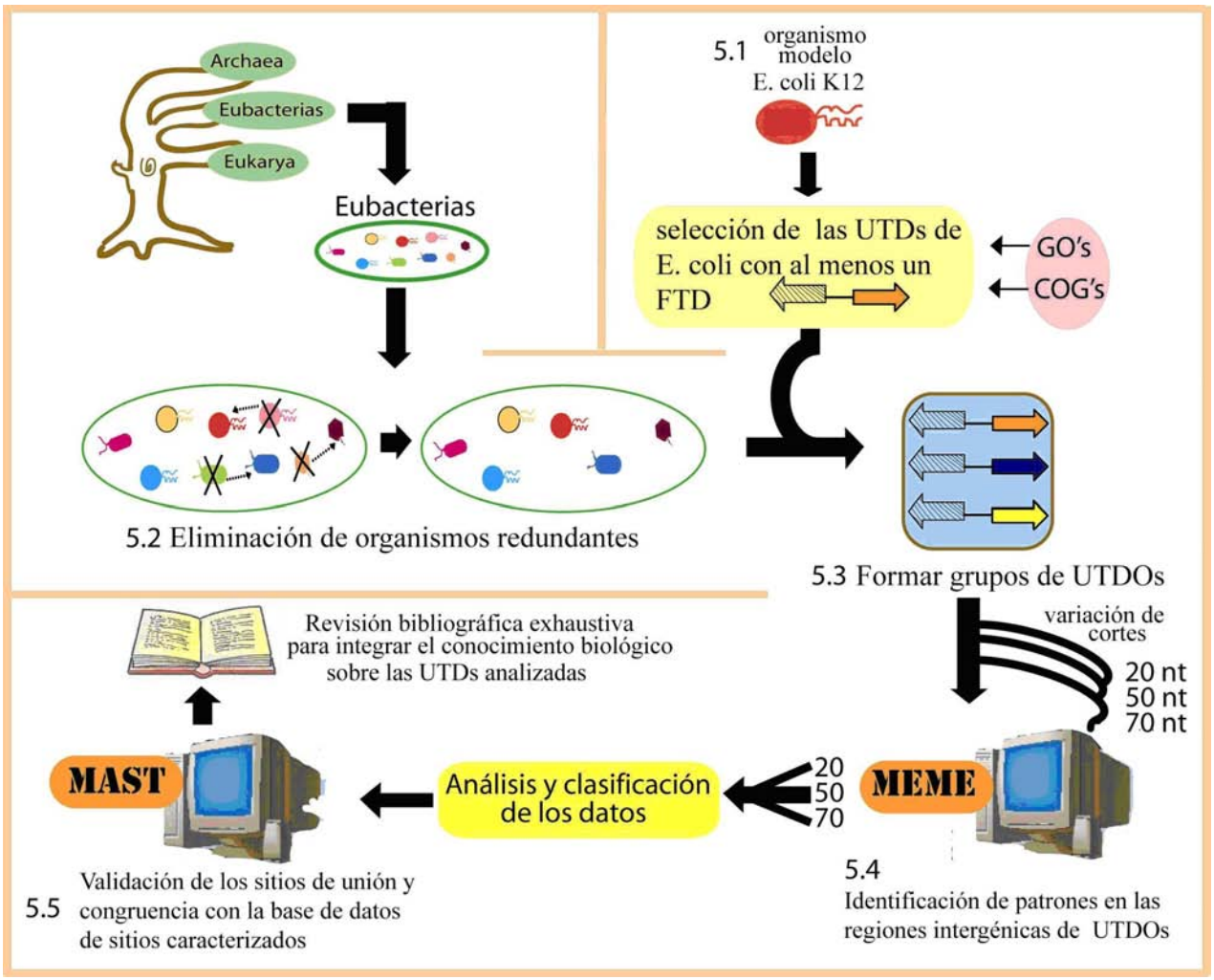


Figura 11 se resume la metodología utilizada en este proyecto seguida de una breve descripción de cada paso.

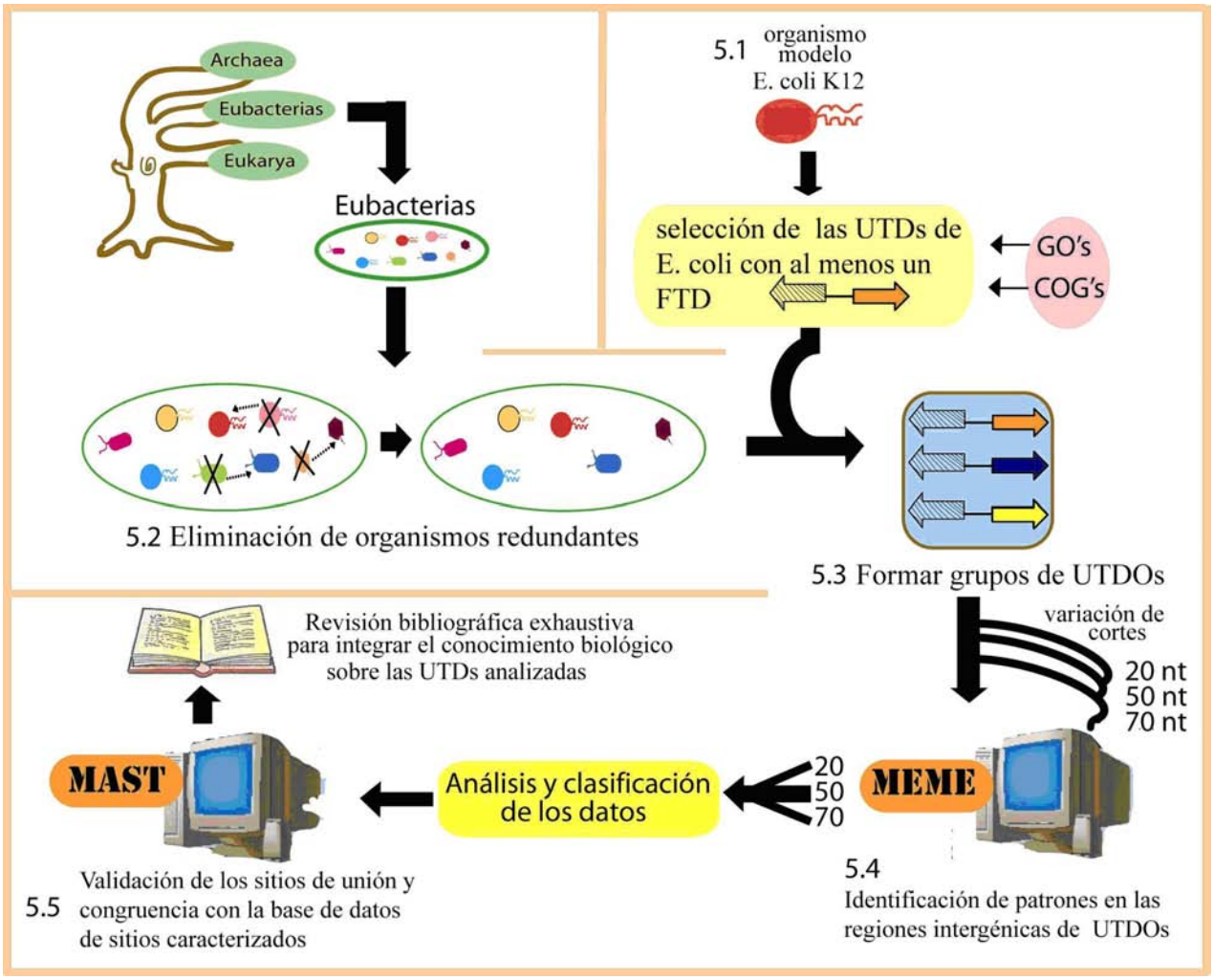


Figura 11. Descripción gráfica de la estrategia computacional desarrollada.

5.1 Obtención de los UTDs referencia de *Escherichia coli K12*

Como organismo modelo se utilizó la bacteria *Escherichia coli K12*, tomando en consideración el cúmulo de información experimental bien documentada sobre los sitios de unión de FTs depositados en la base de datos de RegulonDB (60). De referencia, se utilizaron las UTDs de *E. coli* conformadas al menos, por un FT y un gen estructural (topología R-P y R-R ver **Figura 6**) (61). En esta etapa, se decidió no considerar el análisis de los casos P-P dado que se intentaba estudiar el mecanismo de corregulación bidireccional más caracterizado. Este criterio obedece al hecho de que las UTDs más conservadas en eubacterias cuentan con un FT que se transcribe divergentemente a uno o más genes codificadores de proteínas (37). Para llevar a cabo este primer filtro, nos basamos en las asociaciones y clasificaciones funcionales de la base de datos COG (por sus siglas en inglés, Cluster of Orthologous Groups) y GO (Gene Ontology) (3; 65). Si al menos uno de los genes del par divergente era FT (activador o represor), se consideraba para la lista del conjunto referencia. De este modo, se trabajó únicamente con aquellos pares de UTDs que aseguraban tener al menos un FT, independientemente de que fuera conocido o hipotético (dicho FT será referido como FTD en el resto de la tesis).

5.2 Selección del conjunto de genomas bacterianos no redundantes (nr)

Se entiende por "genomas bacterianos redundantes" aquellos genomas cuya cercanía filogenética los hacen ser extremadamente similares. Por ejemplo, las diferentes cepas de *Escherichia coli: E.coli 536; E.coli APEC O1; E.coli CFT073; E.coli O157H7 EDL933; E.coli O157H7; E.coli UTI89; E.coli W3110, E.coli K12,* son consideradas cepas redundantes entre sí. El eliminar genomas redundantes de nuestro estudio evitará identificar sitios altamente conservados por azar debido a la cercanía filogenética entre dos especies que no han tenido tiempo para divergir. Se decidió trabajar únicamente con organismos eubacterianos porque se ha observado que la conservación de genes divergentes es más común en este grupo (37). La identificación de motivos de regulación biológicamente funcionales conservados puede resultar poco confiable, al comparar regiones de regulación de genomas filogenéticamente cercanos (*e.g.*

cinco cepas de *Salmonella* y cuatro de *Escherichia*). En estos casos las secuencias se parecen mucho porque el evento de especiación pudo ser reciente y no porque haya una presión selectiva por preservar la función de una secuencia particular. El número de genomas bacterianos secuenciados actualmente (noviembre del 2007) es más de 500, y la existencia de técnicas para identificar redundancia genética permite seleccionar un número razonablemente grande de genomas. En breve, el método aquí empleado para identificar y eliminar genomas redundantes consta de dos pasos (2). Primero, obtener la calificación promedio de BlastP de un genoma contra sí mismo, es decir, la autocalificación promedio de todas sus proteínas al compararse contra ellas mismas. Segundo, obtener la calificación promedio de BlastP de todos los *BDBH (*por sus siglas en inglés *BiDirectional Best Hit)* entre un genoma referencia y cualquier otro (*e.g. E. coli* K12 vs *Vibrio cholerae*) (**Figura 12**), es decir, la calificación promedio comparada. Finalmente, dos genomas son considerados redundantes si el cociente de la calificación promedio comparada y la autocalificación promedio es mayor o igual a 0.9 (corte considerado para este trabajo).

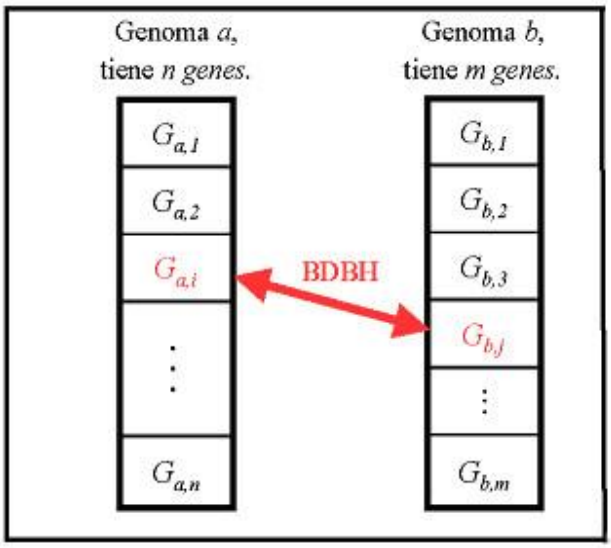


Figura 12. La estrategia utilizada para detectar ortología involucra a los pares de genes que son más parecidos en ambas direcciones (48; 49).

5.3 Identificación de UTDs ortólogos (UTDOs) entre el conjunto de genomas nr

La identificación de genes ortólogos se realizó aplicando la definición operativa de ortología basada en los genes recíprocamente más parecidos (*BDBH*) entre pares de genomas (**Figura 12**). Se identificaron los pares de UTDOs, es decir, UTDs que tienen un par de UTs ortólogos orientados de la misma manera en uno o varios genomas. La organización de genoma puede variar tomando en cuenta que los genes de interés pueden estar organizados con otros genes distintos, u ordenados de manera diferente de genoma a genoma. En este sentido se tomaron sólo los casos representados a la izquierda de la **Figura 13**, en todas sus variaciones.

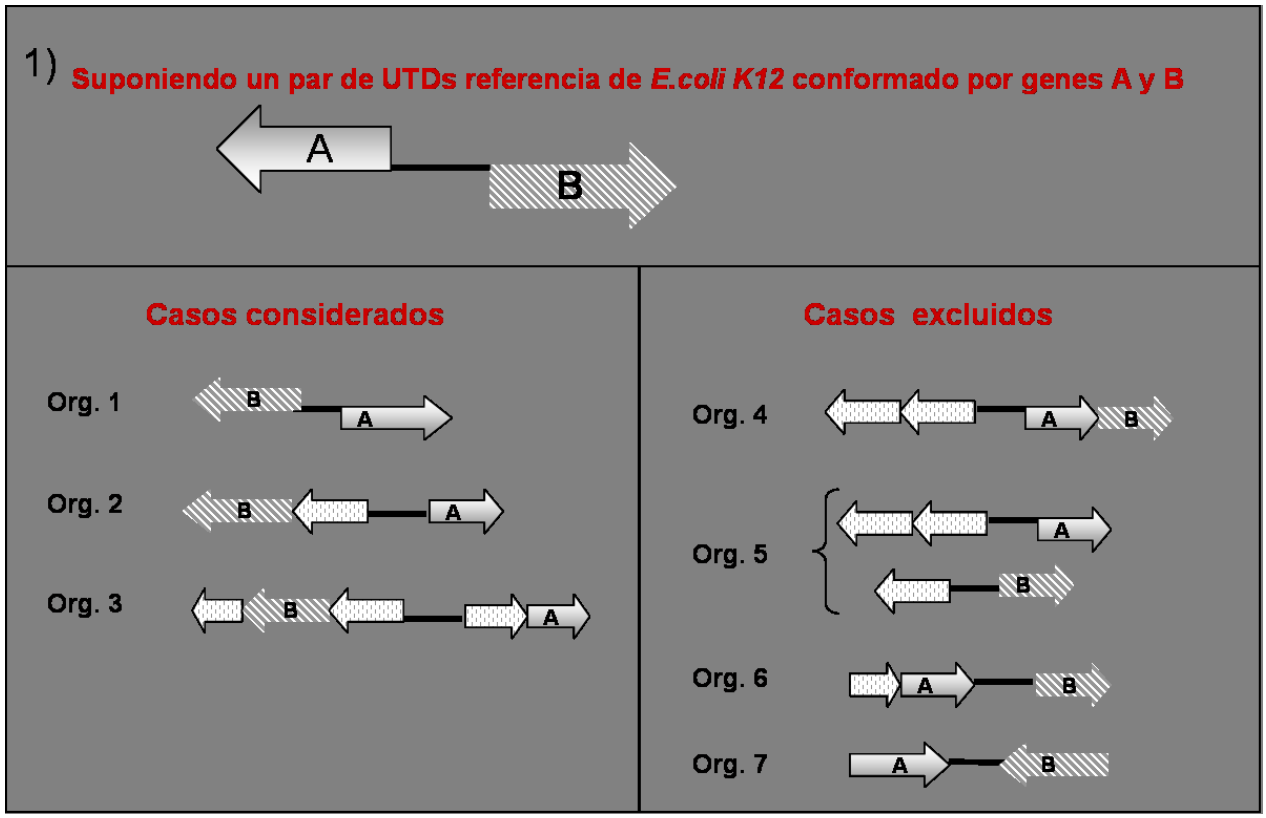


Figura 13. A partir de un par de UTDs referencia, se hizo la búsqueda de los genes que conservan la misma organización en el resto de los organismos no redundantes considerados.

5.4 Búsqueda y clasificación de motivos en regiones intergénicas de UTDOs

Se tomó la región *intergénica* de uno de los genes del par de UTDOs, considerando una longitud de al menos 46 pb para cada uno de los grupos de pares de UTDOs. Tomando en cuenta que 46 pb es la distancia mínima reportada que permite la existencia de dos promotores divergentes para el operón IlvY-IlvC (51), considerándolo como un parámetro de corte muy exigente.

Para descubrir motivos conservados en los grupos de regiones intergénicas, se utilizó un programa del dominio público MEME (por sus siglas en inglés ***M**ultiple **E**M for **M**otif **E**licitation*), y donde EM se refiere al algoritmo de ***E**xpectation **M**aximization*) (4). MEME es considerado uno de los mejores métodos de acuerdo a un análisis realizado sobre una variedad de herramientas para predicción de motivos (67). MEME requiere como entrada un conjunto de secuencias en formato FASTAy puede buscar varios patrones sobre representados

simultáneamente. El método consiste en ajustar una mezcla de modelos a la serie de secuencias que se le proporcionan.

Considerando que los tamaños más pequeños de los sitios de unión de los FTs tienen longitudes entre 7 pb y 10 pb, y a su vez, las longitudes más grandes oscilan entre 18 y 25 pb (27; 28; 41), seleccionamos tres rangos de búsqueda cambiando el parámetro de corte del tamaño de los patrones a identificar por MEME: 20 pb, 50 pb y 70 pb. Se consideraron únicamente patrones conservados cuya significancia estadística (Valor Esperado o "E-value") fuera menor de 10^{-1} y la opción zoops ("zero or one *per* sequence") que no considera necesario que los patrones conservados estén presentes en todas las secuencias de entrada. En general el valor esperado empleado permite encontrar elementos de regulación presentes en solamente unas pocas de las secuencias de un grupo de UTDOs.

Las matrices obtenidas con MEME no necesariamente representan los sitios de unión, pero si contienen los segmentos de secuencia (entre 20 pb y 70 pb) más conservados. El análisis y clasificación de los sitios de unión dentro de las matrices requiere un complejo análisis en donde se identifican sus diversas orientaciones (IR, DR, CV), posiciones y distancias respecto al gen.

Una UTD puede ser regulada por diversos FTs, incluyendo el propio FTD, por lo tanto se pueden identificar diversos sitios de unión. En el presente trabajo nos enfocaremos en analizar los sitios de unión de los FTDs, ya que es común que las UTDs sean autorreguladas.

Una vez determinados los sitios de unión y sus orientaciones, procedimos a clasificar las UTDOs de acuerdo a la presencia o ausencia de sitios a sus FTD, en su región intergénica: a) UTDOs con FTD conocido y que poseen sitios caracterizados, detectados por nuestro método; b) UTDOs con FTD caracterizado, sin sitios reportados y nuestro método posiblemente lo esta detectando; c) UTDOs con FTD hipotético, sin sitios reportados y nuestro método potencialmente los predice; d) UTDOs que aparentemente no presentan sitios de unión pero conservan su organización divergente y poseen un FTD hipotético.

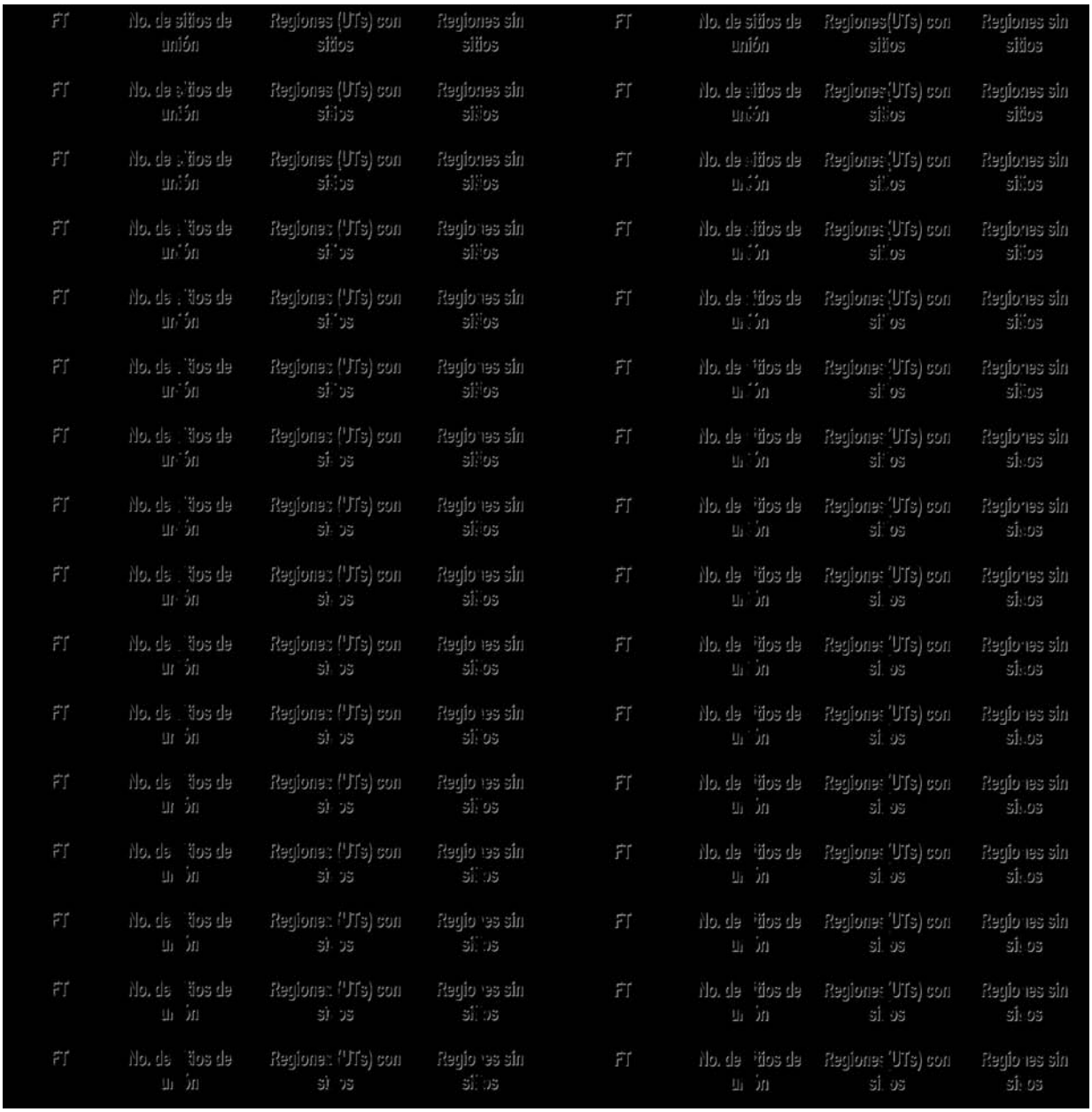
5.5 Validación de los motivos predichos.

Como primer paso, se verificó si nuestro método era capaz de detectar los grupos de UTDOs que poseen sitios caracterizados experimentalmente para su FTD. Este grupo serviría como control positivo ya que es una evidencia que apoya el potencial de predicción de nuestro método.

Para ello utilizamos la herramienta MAST (por sus siglas en inglés *Motif Alignment and Search Tool*) (4; 18), que funciona comparando las matrices (salida de MEME) contra una base de datos determinada; dado el propósito de nuestro estudio, nuestra base de datos se construyó con la información de los sitios de unión de FTs recopilados en la base de datos RegulonDB para la bacteria *E. coli K12* (**Tabla 1**), esta base de datos contiene información de 88 regulones, de los cuales sólo 75 tienen al menos 2 sitios de unión para sus correspondientes FTs. Debido a que los sitios registrados en RegulonDB son de longitudes pequeñas (~5 a 25 nt), se decidió extender la longitud de las secuencias reportadas a secuencias de 100 pb antes y 100 pb posteriores a los sitios reportados en la base de datos.

Por otro lado, el control negativo se realizó formando nuevos grupos de UTDOs al azar. Se permitió mantener tanto el par referencia de *E.coli K12*, como la lista de organismos originales que integran el grupo. Lo único que varía y se elige al azar son las nuevas UTDs de los organismos nr, exigiendo que la nueva UTD incorporada no sea ortólogo del par de UTD referencia. Se espera que con los nuevos grupos, al hacer el análisis no se logren identificar patrones conservados, ya que se trata de UTDs no relacionadas y no existe una justificación biológica para conservar los sitios de unión. Lo anterior a su vez nos corroboraría que el estar identificando los patrones conservados en las UTDOs no es un evento azaroso.

Tabla 1. Resumen del conjunto de sitios de reconocimiento de factores transcripcionales en la base de datos RegulonDB. Los FTs se listan en la primera columna, la segunda columna señala el número total de sitios de unión identificados para cada FT, distribuidos en las regiones río arriba de las UTs numeradas en la tercera columna. La cuarta columna indica el número probables sitios río arriba, para las cuales aún no hay una prueba directa de la unión del regulador a este sitio, pero existe evidencia experimental que sugiere regulación. Por ejemplo, para el caso de ArcA (tercer caso) hay 74 sitios de unión identificados, localizados en 31 regiones río arriba de UTs distintas (en promedio 2 sitios por región), más 9 regiones río arriba de genes distintos para los cuales existe evidencia de regulación por ArcA pero cuyo sitio de unión no está definido aún (60).



RESULTADOS Y DISCUSIÓN

De 652 pares de UTDs en *Escherichia coli* K12, sólo 158 tienen en su organización al menos un gen regulador (caracterizado o hipotético) (**Figura 14** filtro A). Al momento de realizar el análisis se contaba con 253 genomas bacterianos completamente secuenciados que se redujeron a 177 con la eliminación de genomas redundantes (Apéndice 1). Partiendo de los 158 UTDs

referencia, se tomaron los genes ortólogos de los genomas nr y se definieron los grupos de UTDOs.

Tomando en cuenta que realizaríamos análisis comparativos entre las secuencias, decidimos considerar aquellos grupos que tuvieran un mínimo de organismos, el corte lo realizamos tomando en consideración el valor de la moda de la distribución del tamaño de los grupos de UTDOs (**Figura 15**), se decidió trabajar únicamente con aquellos que constaran al menos de 5 UTDs de organismos distintos. En este segundo filtro nuestro conjunto de datos se redujo a 75 grupos (**Figura 14** filtro B). Como criterio adicional, se consideraron únicamente aquellos grupos que compartieran una región intergénica común de al menos 46 pb. Finalmente se obtuvieron 48 grupos de UTDOs (**Figura 14** filtro C), listados en el Apéndice 2.

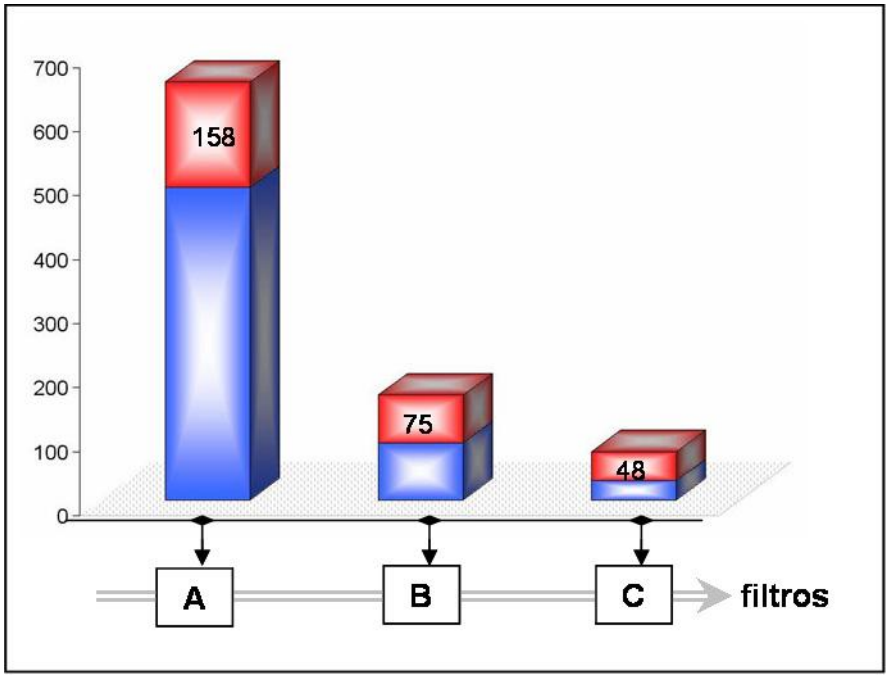


Figura 14. Discriminación de UTDs del organismo modelo *E.coli K12*. Filtro A) de un total de 652 UTDs únicamente 158 pares tienen en su composición al menos un gene regulador, entran aquí los casos R-R y R-E. Filtro B) 75 grupos de UTDOs contienen al menos cinco organismos. Filtro C) 48 grupos de UTDOs que además de las consideraciones anteriores poseen regiones intergénicas mayores de 46 nt.

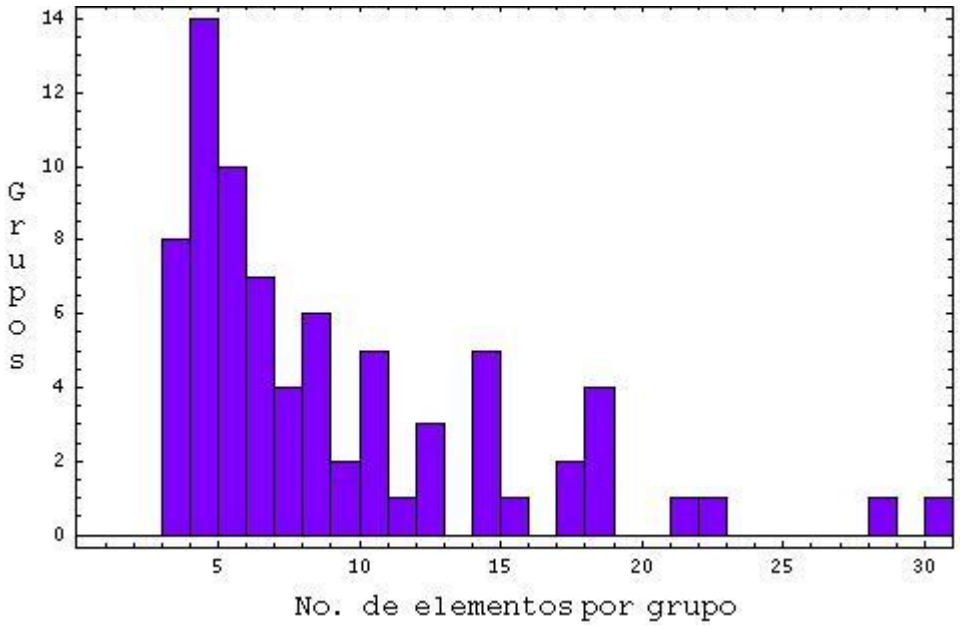


Figura 15. Gráfica de distribución del número de organismos que forman los grupos de UTDOs.

Una vez definidos los 48 grupos de trabajo (Apéndice 2), se analizaron sus regiones intergénicas para identificar secuencias conservadas. La herramienta para identificación de patrones MEME construye una matriz por cada motivo encontrado en cada una de las secuencias intergénicas ortólogas, y a través de ella el programa construye una secuencia consenso que representa los residuos más conservados entre el conjunto de secuencias comparadas. Ya obtenidas las matrices, se verificó si en las secuencias identificadas estaban representados los sitios de unión de los FTDs caracterizados en *E. coli K12*. En teoría, los resultados de MAST nos permitirían apreciar la congruencia entre los sitios identificados y lo que se tiene reportado en la literatura, pero es importante mencionar que se requiere un análisis con más escrutinio para la verificación de verdaderos positivos, es decir, el hecho de acertar a un sitio de unión en la base de datos de

RegulonDB no siempre significa que se trata del sitio de unión *per se* ya que puede tratarse de una secuencia que corresponde a otro sitio distinto cercano al supuesto sitio con el que nuestra matriz tuvo un "match". Con el resultado de MAST tenemos también el potencial de identificar regiones en otros genes que pueden pertenecer al mismo regulón.

El objetivo de variar el parámetro de corte de la longitud de las secuencias, en la búsqueda de patrones con MEME, fue localizar los sitios de longitudes pequeñas en el rango de 20 pb y los más grandes en el rango de 50 pb, el último corte (70 pb) lo usamos para verificar que no se quedara ningún sitio fuera.

Al comparar los resultados de los tres rangos (20pb, 50pb y 70pb) se observó que conforme abríamos el márgen de corte, se introducía más variación en los datos obtenidos. Apoyándonos en los casos bien caracterizados que disponen de información sobre sus sitios de unión, promotores y secuencias consenso, logramos obtener mayor información sobre las regiones de los sitios de unión menos conservados en nuestras matrices.

En la mayoría de los casos, las matrices con una mejor representación de la secuencia consenso las obtuvimos usando como parámetro de corte una longitud de 20 pb. Con éste primer corte se observaron los sitios más conservados al analizar las secuencias. En menor proporción identificamos DR en tándem con longitudes de 7 a 10 pb, y mayoritariamente secuencias muy conservadas que gracias al análisis adicional de los parámetros de corte extendidos, nos llevó a concluir que constituyen la mitad del fragmento más conservado de los sitios de IRs. Al ampliar los rangos de corte (50 pb y 70 pb), encontramos que además de incluirse el fragmento más conservado del sitio de unión (localizado previamente con el rango de 20 pb), se incorporan también los fragmentos menos conservados que no se lograron ver en el rango de 20 pb, lo que nos permitió identificar los IR completos.

La ventaja de estas matrices es que a diferencia de las que reportan actualmente con otras estrategias, éstas se construyen basándose en una región de DNA conservada en organismos distintos, en cambio las matrices que se han venido manejando en RegulonDB, son el resultado de un conjunto de secuencias de sitios de unión localizados en varias regiones intergénicas del mismo organismo, que por ende tienen más variación. Por lo tanto, se observó que las secuencias

o matrices obtenidas a través del análisis de identificación de huellas filogenéticas son muy conservadas.

6.1 Clasificación de los grupos

Después de realizar la búsqueda y el análisis de los sitios de unión, los 48 grupos de UTDOs seleccionados, se clasificaron en 4 bloques de acuerdo a la información disponible en la literatura. En el bloque I se agruparon 18 UTDOs que tienen evidencias experimentales de ser reguladas por sus propios FTDs y que cuentan con las secuencias de los sitios de unión para dichos reguladores. El bloque II agrupó 16 UTDOs que también presentan evidencias experimentales de ser reguladas por sus FTDs, pero a diferencia del bloque I, para estos casos no se han identificado los sitios de unión, en este bloque únicamente la mitad de los grupos presentaron sitios de unión, y los restantes se decidió agruparlos en bloque IV. El bloque III se conforma por 9 UTDOs con FTDs hipotéticos sobre los que no se conoce nada sobre su estructura y forma de unirse al DNA, sin embargo conservan su organización divergente entre los genomas y presentan motivos conservados en sus regiones intergénicas. Para algunos casos se observan claramente algunas estructuras de IR que podrían proponerse como sitios de unión. Por último en el bloque IV se organizó el resto de los UTDOs que contienen tanto FTDs hipotéticos como conocidos (incluidos aquí 8 grupos del bloque II), pero que en las matrices obtenidas fue difícil identificar posibles sitios de unión.

1 Análisis del bloque I: UTDs autorreguladas con sitio de unión caracterizado

El bloque I contiene 18 grupos de UTDOs, de los cuales, basándonos en la información sobre su UTD referencia, contenida en las bases de datos de RegulonDB y EcoCyc (34), se observó que cada uno conserva los sitios de unión de sus FTDs respectivos. En general 15 de las 18 UTDOs incluidas en este bloque son autorreguladas por sus FTDs (**Figura 16**), sin embargo, en las tres UTDOs restantes fue posible identificar los sitos de unión con estructuras de IR sobre

las regiones de la UTD referencia, y son los casos que se presentan en los recuadros verdes de la **Figura 16**.

Es importante resaltar que el análisis de los sitios de unión se enfocó únicamente en los sitios blanco de los FTD, de tal forma que los primeros 10 grupos del bloque I (gpos. 1-10) presentan sitios de unión IR en tándem y sus longitudes oscilan entre 11 a 26 pb. En estas secuencias IR se une un dímero del FTD correspondiente a la UTD. Los siguientes 4 grupos (gpos. 11-14) a su vez, presentan sitios de unión DR en tándem con longitudes entre 5 y 10 pb. Las UTDOs representadas por *nagBAD-nagE* (gpo. 15) y *cynR-cynTSX* (gpo. 16), presentan dos sitios de unión CV repetidos. Los grupos 17 y 18 contienen una mezcla de sitios de unión IR en tándem y DR, es el caso de los reguladores MetJ y NarL, que son un ejemplo de FTs que pueden unirse como dímeros o como monómeros tomando en consideración los distintos arreglos de sus sitios de unión.

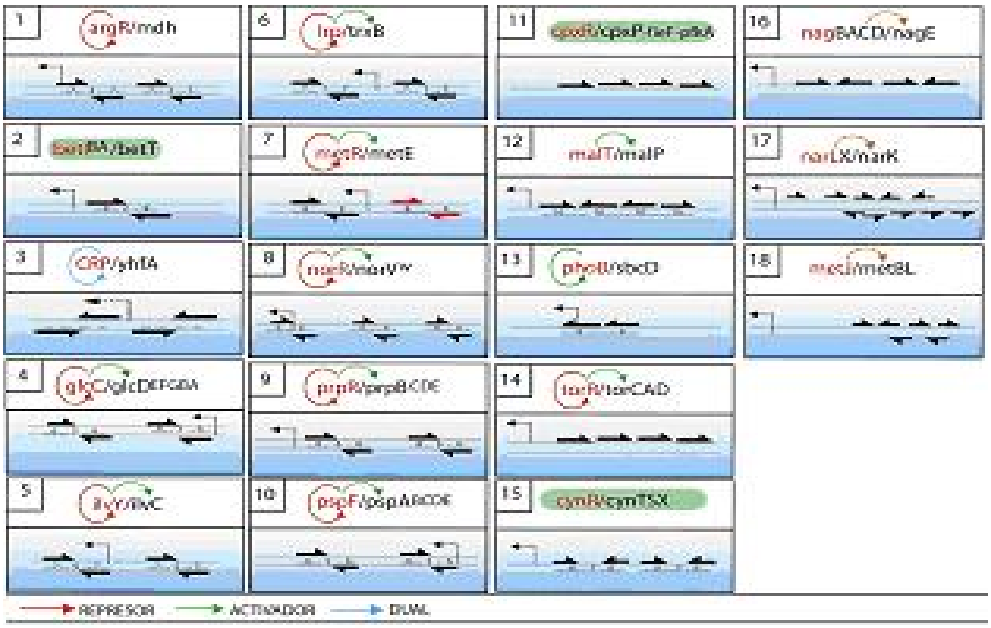


Figura 16. Los casos del bloque I representados aquí son UTDs que poseen sitios de unión a su FTD.

1 Discusión de casos particulares en el bloque I

Análisis de los sitios de unión del FT GlcC en la UTDO *glcC-glcDEFGBA*.

El sitio de unión que se identificó con nuestro método para la proteína GlcC es un IR perfecto de 15nt (S1), mientras que la secuencia reportada en la literatura (RegulonDB), es una secuencia continua (no mencionan nada sobre su arreglo de IR) de 16 nt, que al compararse con S1 se observa desplazada en 2pb (**Figura 17**). Tomando como referencia el sitio S1 de GlcC, la matriz con el rango de corte 50pb obtenida con MEME, pone en evidencia que a una distancia de 14 nt se encuentra un segundo sitio de unión menos conservado para GlcC que podría estar traslapando el promotor de E.coli (S2) lo que justifica el efecto represor del FT. La secuencia correspondiente

a S2 no ha sido publicada y por lo tanto no esta registrada en las bases de datos de RegulonDB y EcoCyc. Por lo tanto, con base en estas observaciones se podrían diseñar experimentos para verificar lo que proponemos sobre la corrección del sitio de unión de GlcC y la identificación de un segundo sitio de unión.

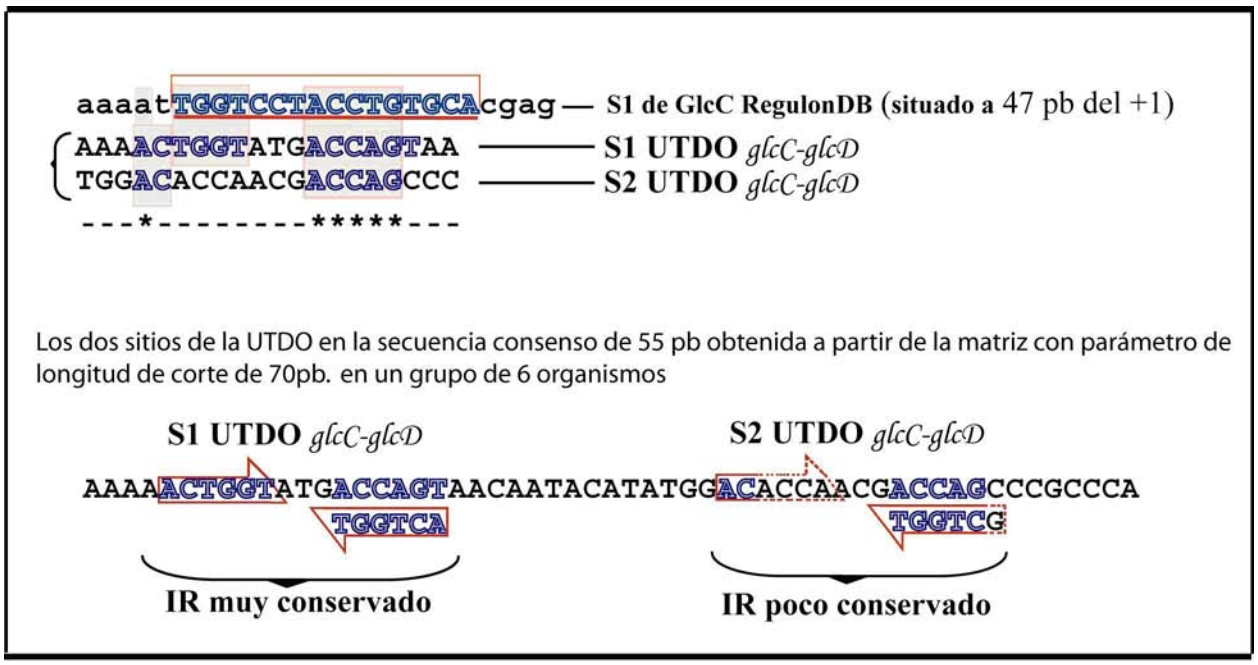


Figura 17. Alineamiento de los sitios de GlcC, S1 y S2, obtenidos en nuestro análisis en la región glcC-glcD *versus* el sitio de unión reportado en la base de datos de RegulonDB (S1 RDB).

Análisis de los sitios de unión del FT Lrp en la UTDO *lrp-trxB*

En este caso, la secuencia consenso publicada para la proteína Lrp corresponde a un IR poco conservado de 15pb (Y**AG**H**A**WA**TT**W**T**D**CT**R) (77). Los resultados obtenidos por nuestro método muestran dos sitios de unión de Lrp (S1 y S2). En la **Figura 18** se alinean las secuencias

identificadas en este trabajo, S1 y S2, y se comparan contra los sitios de unión reconocidos por Lrp, mostrados en la base de datos de RegulonDB, en azul se resaltan los nt del sitio de unión. El alineamiento entre los cuatro sitios de unión muestra un motivo conservado (gcaTGT) que correspondería al sitio de unión de un monómero, la otra mitad del sitio de unión del invertido es menos conservada y se encuentra representado por la secuencias CATgt. Estas evidencias muestran que la secuencia consenso publicada en la literatura esta incompleta y carece por completo de la parte menos conservada del posible IR (CATgt). Además los sitios publicados y reportados en la literatura, S1 y S2 de RegulonDB (en cajas rojas), estarían mal anotados. En este caso en particular aunque las secuencias de los sitios S1 y S2, obtenidas por nuestra metodología, no son muy conservadas, nos permiten definir mejor la secuencia consenso del sitio de unión de la proteína Lrp. Otra posibilidad que existe es que los sitios están representados por dos DR de 6 pb (GCATGT) separados por 5 nt.

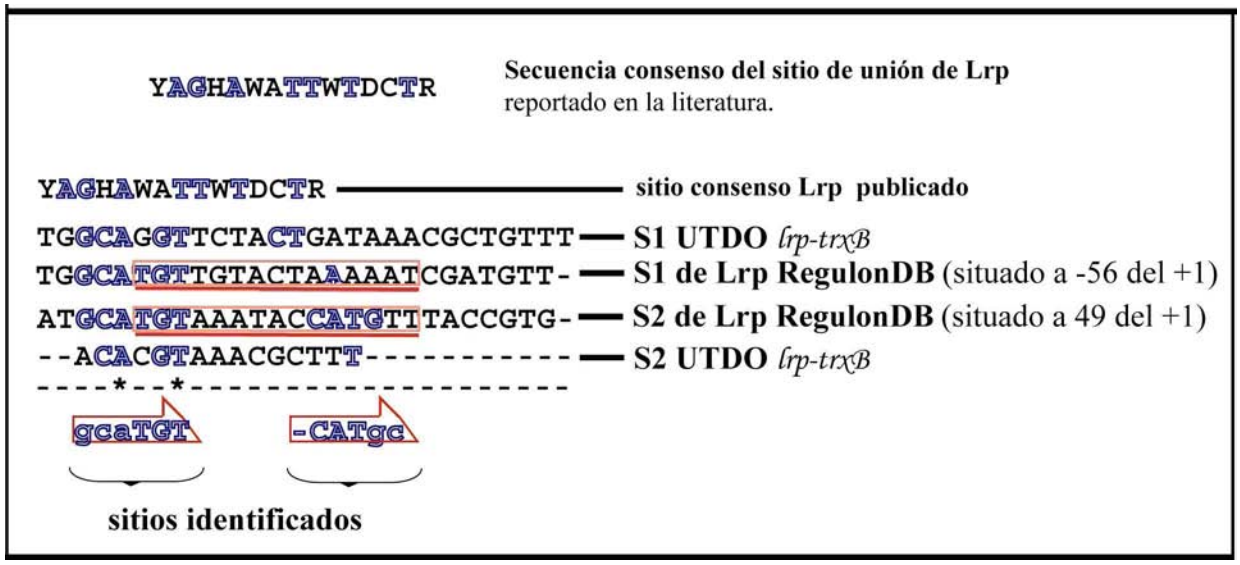


Figura 18. Alineamiento de los sitios identificados en la UTDO *lrp-trxB versus* los sitios conocidos en E.coli K12.

Análisis de los sitios de unión del FT MetR en la UTDO *metR-metE*.

La secuencia consenso de la proteína MetR es un motivo IR de 16pb (AGACGTCTAGACGTCT). En la **Figura 19**, se alinean los sitios de unión identificados y se observa que tanto la secuencia del sitio uno (S1) obtenida de la UTDO (compuesta por una matriz de sitios en distintos organismos) y su correspondiente en RegulonDB (compuesta por una matriz de sitios del mismo organismo, *E. coli*) son poco conservadas. Sin embargo, en nuestro análisis también logramos identificar un segundo sitio (S2) más conservado, que no ha sido reportado para MetR pero que se parece más a la secuencia consenso, e inclusive se localiza en la región intergénica de *E.coli K12*, pero este segundo sitio no está reportado.

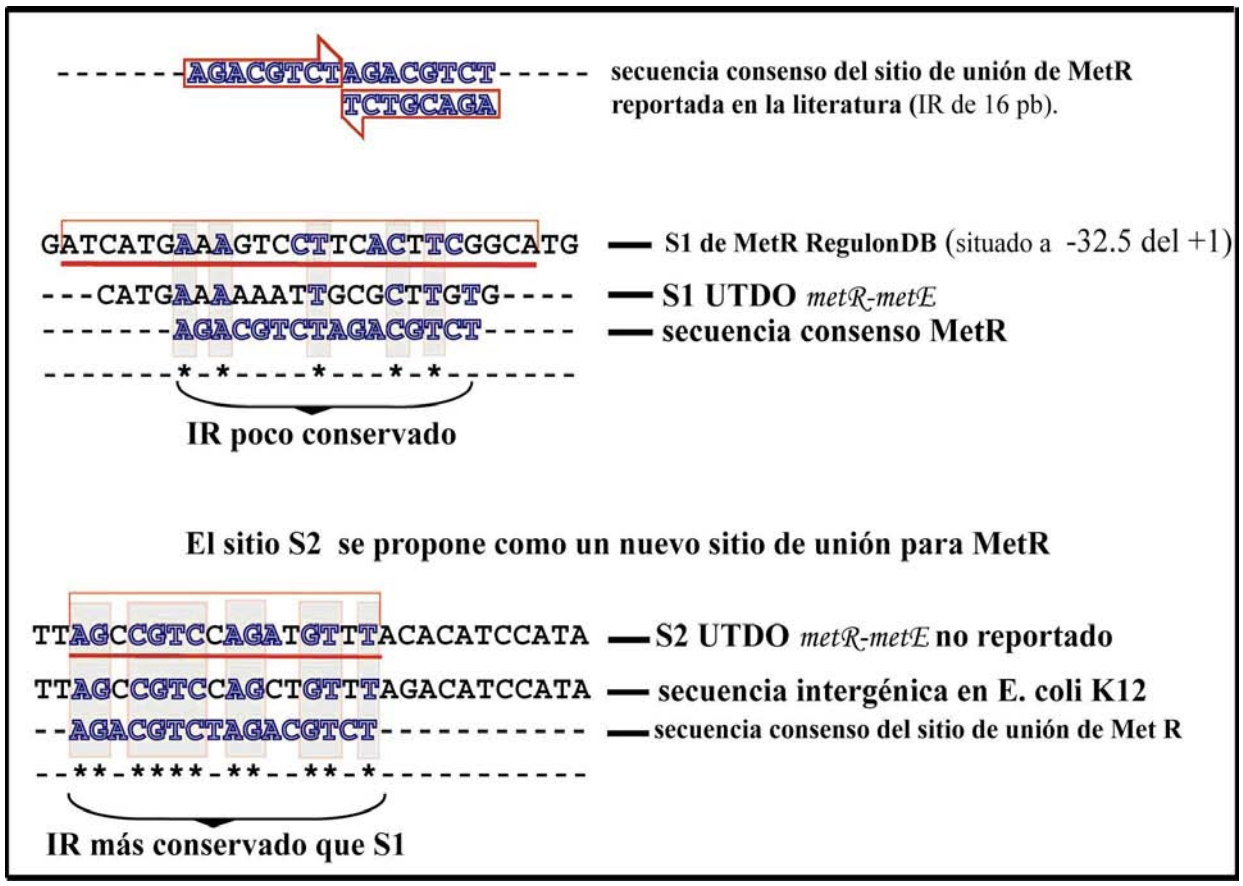


Figura 19. Alineamiento y análisis de los sitios de unión de MetR encontrados en la UTDO *metR-metE*.

En general en este bloque logramos demostrar que los resultados computacionales apoyados con la información biológica disponible es buena estrategia, ya que en cada UTDO se lograron identificar los sitios de unión del FTD. Además de conservar el número de sitios de unión, también se conservan las orientaciones y en conjunto estas características nos reflejan la confiabilidad que podemos tener sobre el análisis de nuestros resultados. Este primer bloque aporta las evidencias necesarias para predecir sitios de unión a FTs que tienen evidencia experimental de su regulación pero que no se han identificado los sitios de reconocimiento en el DNA (bloque II) o posibles sitios de unión a FTs hipotéticos (bloque III).

2 ANÁLISIS DEL BLOQUE II : UTDOs con FTs caracterizados y sitio de unión no identificado

En el bloque II están representados 16 grupos de UTDO, cuyos FTDs son caracterizados pero que a diferencia del bloque I, no se han identificado experimentalmente sus sitios de unión, con excepción de FadR. A su vez, este bloque se subdivide en 3 partes: i) Los FTDs que tienen evidencias experimentales de autorregulación, ii) Los FTDs que no tienen evidencias experimentales de autorregulación, y iii) Los FTDs que regulan otros genes.

En este segundo bloque se encontraron posibles sitios de unión que no alinean con ningún sitio conocido en la base de datos de RegulonDB y EcoCyc (**Figura 20**), con excepción de FadR, que se decidió agrupar aquí porque se ha reportado que controla la expresión de múltiples genes y operones, pero no se sabe si se autorregula. La secuencia consenso determinada de los sitios de unión de FadR en su regulón es: AacTGGTcngACCAGTt.

En la primera parte del bloque II (grupos 1-4) se muestran 4 de las 6 UTDOs, con evidencia experimental de autorregulación, que presentan posibles sitios de unión: *acrR-acrA*, *dsdC-dsdX*, *lsrR-lsrACDBFG*, *putA-putP*. La segunda sección de la **Figura 20** (grupos 5 y 6), corresponde a FTDs que carecen de evidencias de autorregulación, y se muestran dos UTDOs con posibles sitios IR (*rob-creABCD*, *uvrYC-yecF*). Por último en la tercera parte (grupos 7 y 8) se

ubican UTDOs cuyos FTDs regulan a otros genes pero sin evidencia de autorregulación. En esta última parte solamente se identificaron posibles sitios de unión IR en 2 UTDOs (*fadR-rhaB*, *yeiE-yeiH*). Aunque inicialmente mencionamos que este bloque consta de 16 grupos, únicamente reportamos los 8 UTDOs en los cuales se logró identificar un posible sitio de unión. Las 8 UTDOs restantes se consideran como parte del bloque IV.

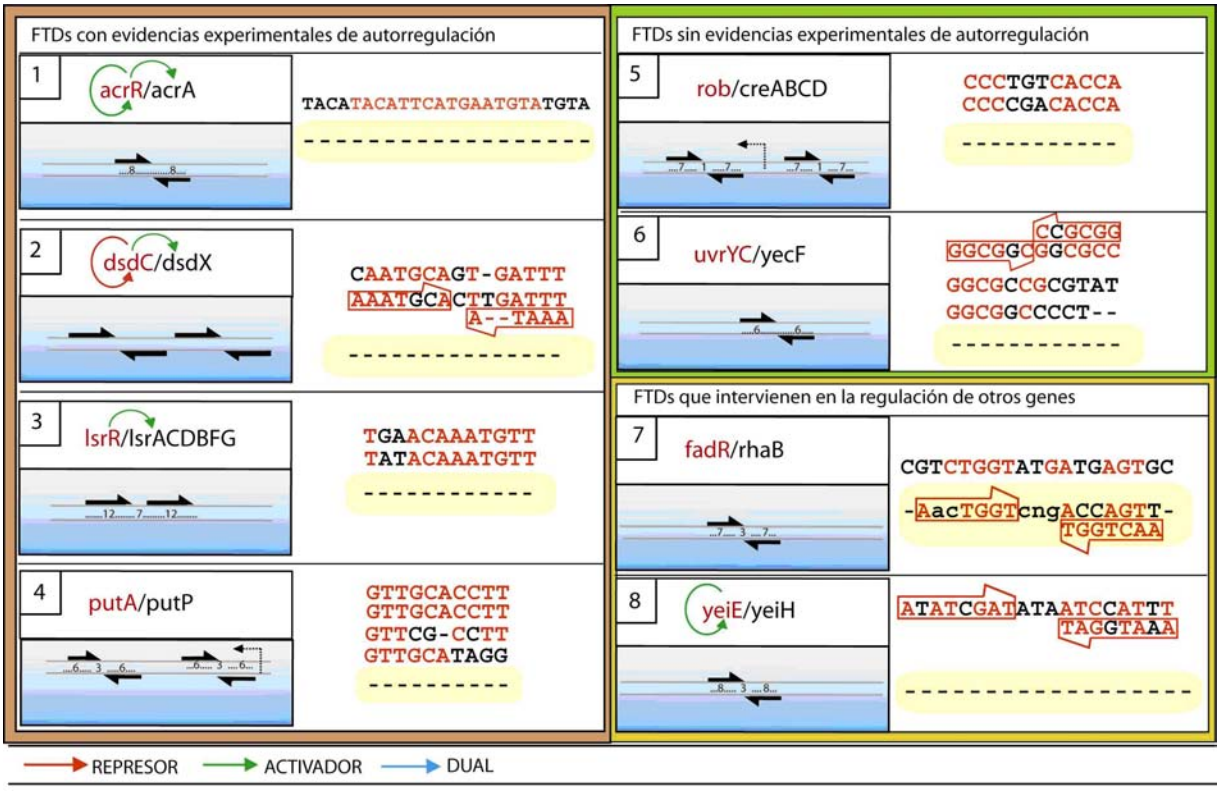


Figura 20 Esquema de los sitios identificados en grupos del bloque dos, conformado por UTDOs con FTDs caracterizados pero sin sitios de unión reportados.

2 Discusión de casos particulares en el Bloque II:

Análisis de los sitios de unión del FT AcrR en la UTDO *acrR- acrA*

Existe evidencia experimental de la represión transcripcional de *acrA* por parte de AcrR, pero el sitio de unión de este FT no se ha identificado. El resultado de nuestro análisis para esta UTDO muestra una secuencia IR de 16pb (TACATTCATGAATGTATGTA) que al compararla contra su secuencia correspondiente en la región intergénica del gen *acrA* de *E. coli K12*, muestra que traslapa por completo al promotor *acrA*p

(

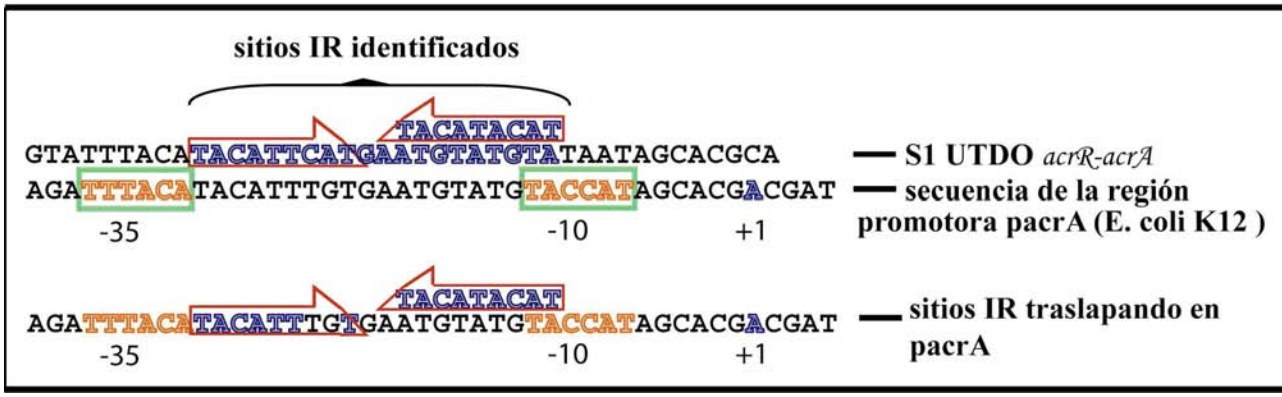


Figura 21). AcrR es un FT que se ha reportado como represor, y su regulación sobre el gene *acrA* se sustenta por las siguientes evidencias reportadas en literatura: unión por extracto celular y análisis de expresión del gen mediante una fusión transcripcional (43). El hecho de que el sitio IR que proponemos este ubicado sobre el promotor puede justificar su efecto represor.

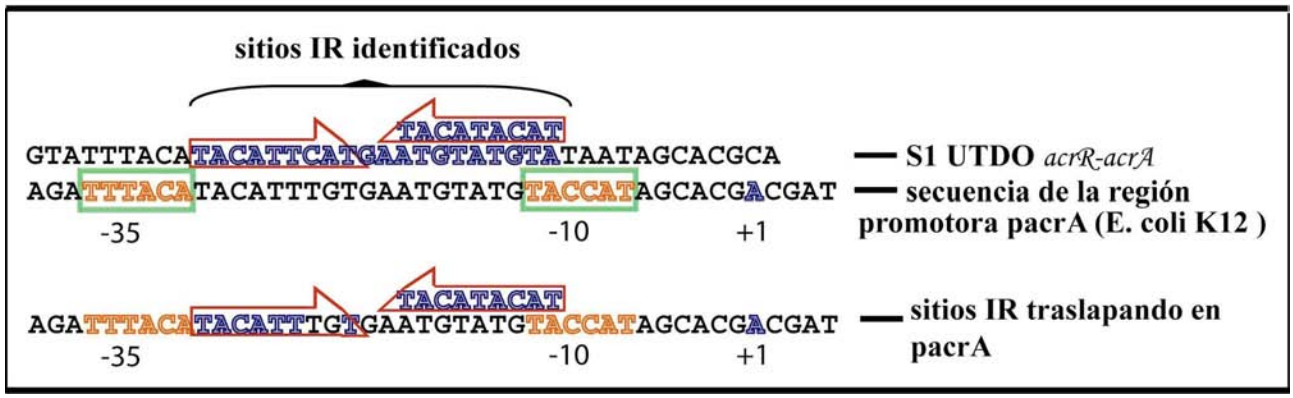


Figura 21. Representación del traslape del sitio identificado en la UTDO acrR-acrA sobre el promotor de AcrA.

Análisis de los sitios de unión del FT DsdC en la UTDO *dsdC- dsdXA*

Se ha reportado que DsdC es un activador transcripcional del operón *dsdXA,* y en presencia de CRP, incrementa alrededor de 200 veces la transcripción de dicho operón (50). Norregaard *et al.* ha reportado que CRP se une a dos sitios de unión localizados a -87.5pb y

–117.5pb, pero por si solo es incapaz de activar la transcripción. Adicionalmente agrega que DsdC y CRP actúan sinérgicamente en la activación del operón *dsdXA*. Para esta UTDO se lograron identificar dos sitios DR que traslapan los sitios de unión de CRP reportados, como se aprecia en la **Figura 22**. De acuerdo a la información anterior existe la posibilidad de que DsdC este formando un complejo con CRP.

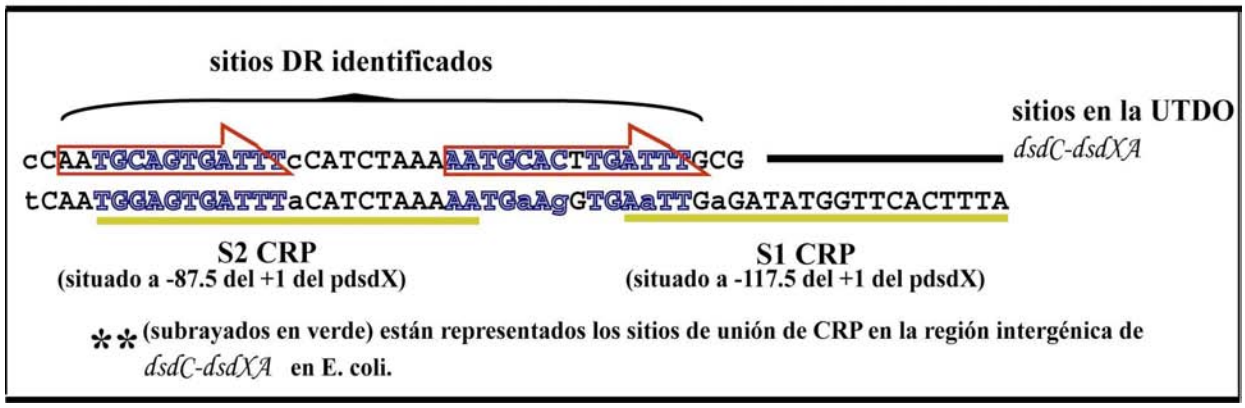


Figura 22. Esquema de los sitios identificados en la UTDO *dsdC-dsdXA.*

Análisis del caso de la UTDO *lsrR-lsrACDBFG*

En esta UTDO logramos identificar dos posibles sitios de unión (S1 y S2), repetidos en tándem, separados por 7 pb. Como se observa en el alineamiento de la **Figura 23** los dos posibles sitios de unión son muy conservados a diferencia se sus ortólgos en *E. coli*. A pesar de que existe evidencia de la represión del operón *lsrACDBFG* por parte LsrR, aún no se ha logrado identificar los posibles sitios de unión de este regulador. La falta de conservación de los sitios de unión puede ser un factor importante que impide la identificación de los sitios de unión en general.

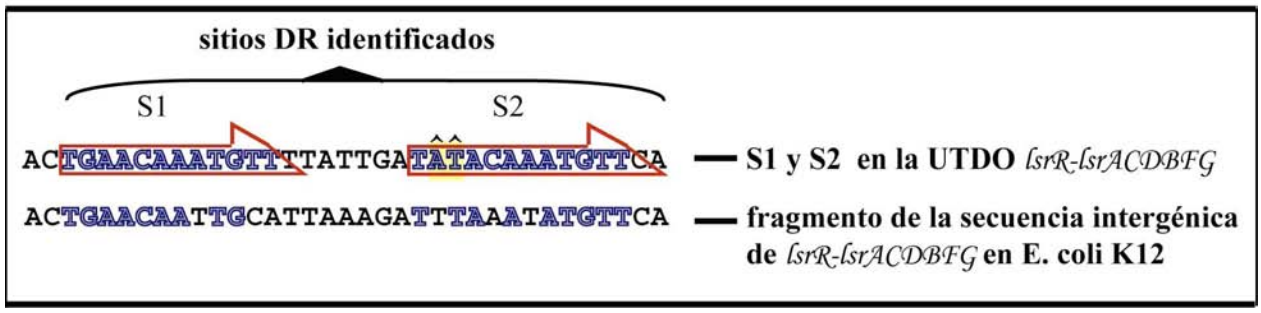


Figura 23. Esquema de los sitios identificados en la UTDO *lsrR-lsrA* comparado contra su secuencia correspondiente en *Escherichia coli K12*.

Análisis de la UTDO *fadR-rhaB*.

FadR es un FT que pertenece a la familia GntR, regula la transcripción de varios genes involucrados en el metabolismo de ácidos grasos y transporte (20; 21). La secuencia consenso para los sitios de unión de FadR esta reportada y es un IR de 17pb (14; 25). Se trata de un FT que controla la transcripción de múltiples genes, pero aun no se sabe si se autorregula. En los resultados obtenidos logramos identificar un posible sitio de unión en la UTDO *fadR-rhaB* que se parece a la secuencia consenso de FadR. Ver alineamiento en la **Figura 24**.

Figura 24. Como se observa en el alineamiento, el sitio encontrado se parece bastante a la secuencia consenso de FadR, lo que puede indicar que este regulador se autorregule o regule al gen divergente.

.

1 ANÁLISIS DEL BLOQUE III y BLOQUE IV

El bloque III esta conformado por 9 UTDOs cuyos FTDs son hipotéticos, no se tiene evidencia de autorregulación y tampoco se conocen sus sitios de unión. Siete UTDOs presentan regiones IRs y dos contienen secuencias repetidas en tándem (DRs). Cabe mencionar que ninguno de los sitios propuestos corresponde a un FT conocido, todos son motivos conservados, a considerar, como posibles sitios de unión. En la figura se esquematizan siete de los casos con IR conservados.

En el bloque IV, además de incluir a los 6 UTDOs excluidos del bloque II, se agruparon otros 5 UTDOs con FTDs predichos, de los cuales no se conocen sitios de unión y no se tiene evidencia de autorregulación. En las matrices obtenidas de estas UTDOs no se logró detectar ningún sitio de unión tentativo, posiblemente por que son regiones más complicadas de encontrar o menos conservadas o definitivamente no son reguladas por ningún FT.

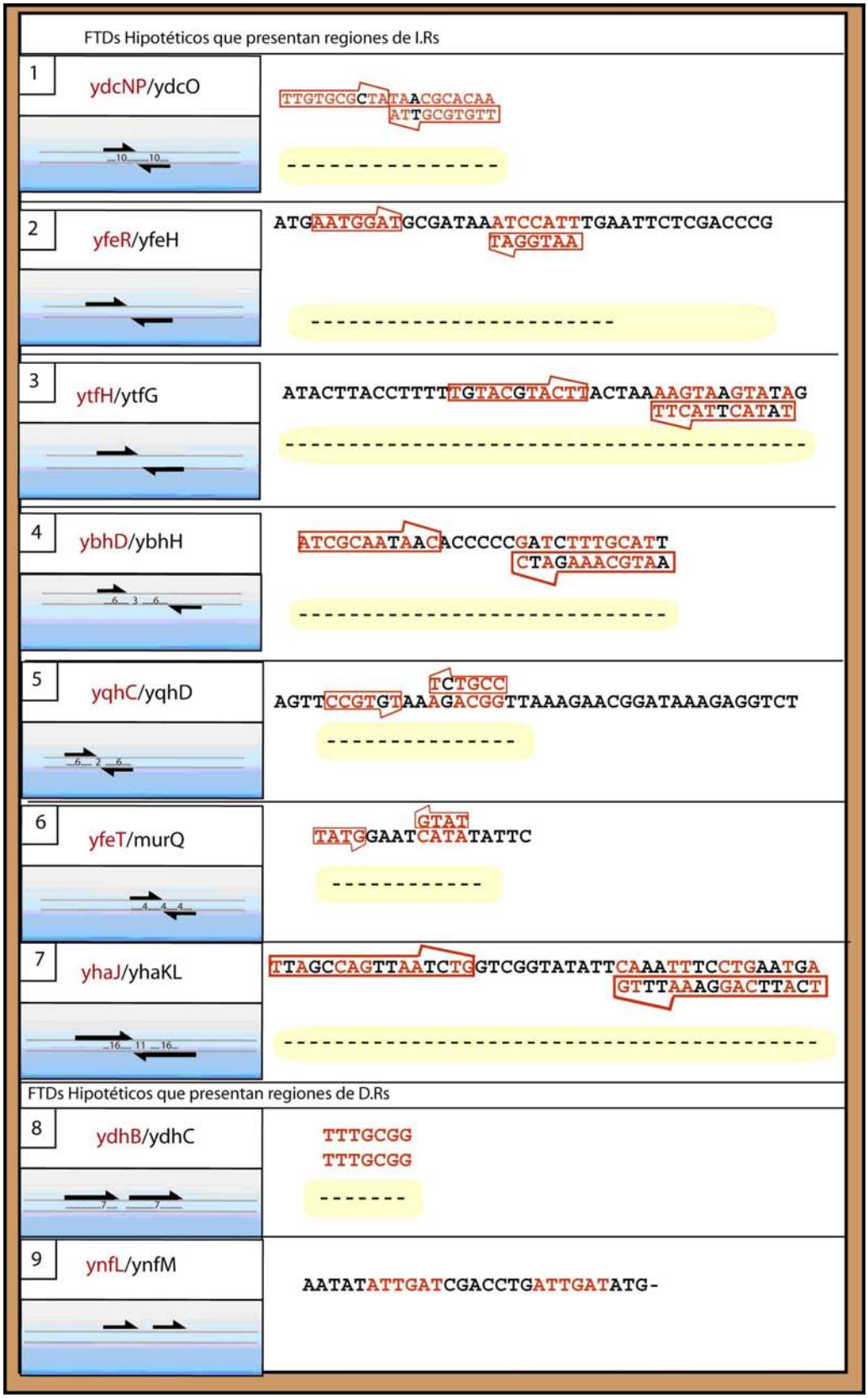

FTDs Hipotéticos que presentan regiones de I.Rs
1
ydcNP/ydcO
TTGTGCGCTATAACGCACAA
ATTGCGTGTT
2
yfeR/yfeH
ATGAATGGATGCGATAAATCCATTTGAATTCTCGACCCG
TAGGTAA
3
ytfH/ytfG
ATACTTACCTTTTTGTACGTACTTACTAAAAGTAAGTATAG
TTCATTCATAT
4
ybhD/ybhH
ATCGCAATAACACCCCCGATCTTTGCATT
CTAGAAACGTAA
5
yqhC/yqhD
TCTGCC
AGTTCCGTGTAAAGACGGTTAAAGAACGGATAAAGAGGTCT
6
yfeT/murQ
GTAT
TATGGAATCATATATTC
7
yhaJ/yhaKL
TTAGCCAGTTAATCTGGTCGGTATATTCAAATTTCCTGAATGA
GTTTAAAGGACTTACT
FTDs Hipotéticos que presentan regiones de D.Rs
8
ydhB/ydhC
TTTGCGG
TTTGCGG
9
ynfL/ynfM
AATATATTGATCGACCTGATTGATATG-

Figura 25. Esquema de los sitios invertidos repetidos predichos para los FT hipotéticos.

En resumen, para la mayoría de los casos las mejores matrices se obtuvieron usando el rango de corte de 20 pb. Al analizar las secuencias, se observaron los sitios más conservados y de mejor afinidad. Las matrices obtenidas con este parámetro permitieron la identificación de sitios pequeños, menores de 10 pb, y RD en tándem de 7 a 10 pb. También se logró observar la parte más conservada de sitios IR mayores a 10 pb. El ampliar los rangos de corte (50 y 70 pb) nos permitió recuperar en su totalidad los sitios IR grandes (mayores a 10 pb), conformados tanto por regiones muy conservadas como por sitios menos conservados que no se lograron ver en el rango de 20 pb. De modo que el trabajo conjunto tanto de la información que nos proveían las distintas matrices, como la integración de la información biológica de cada UTD nos permitió hacer mejores inferencias de nuestros resultados.

La ventaja de construir matrices basadas en huellas filogenéticas, a diferencia de otras estrategias, es que se construyen basándose en las secuencias conservadas de una misma región de DNA en organismos distintos que por ende están sujetas a menor variación respecto a matrices que son el resultado de un conjunto de secuencias (de sitios de unión) tomadas de distintas regiones intergénicas del mismo organismo. En general se observó que las matrices nos reflejan muy bien la conservación de los sitios. Nuestra estrategia permite proponer corrección de sitios de unión y corrección de matrices basadas en estrategias monogenómicas.

Las diferencias de conservación entre los sitios de unión de una misma proteína ocurren porque las afinidades de unión son distintas, de esta forma podemos encontrar sitios de alta afinidad con secuencias muy conservadas o sitios de baja afinidad con secuencias poco conservadas. Cuando los sitios son muy conservados puede tratarse de sitios de unión de proteínas locales, cuya especificidad propicia la conservación del sitio.

8.

CONCLUSIÓN

En general, la estrategia desarrollada presenta resultados congruentes entre el análisis computacional y la información biológica sobre los grupos estudiados. Para la mayoría de las UTDOs del primer bloque, conformado por FTDs con sitio de unión caracterizado, se lograron identificar los sitios, lo cual nos refleja un control positivo substancialmente bueno que nos da confianza para utilizar la estrategia en el análisis extendido al resto de los grupos. El análisis de las matrices de los sitios de unión conocidos, nos permitió hacer propuestas de corrección de sitios e identificación de nuevos sitios para FTD ya caracterizados. Asimismo, se identificaron sitios de unión que corresponden en gran medida con el efecto de FTDs que carecen de sitios de unión. Estas UTDO son candidatos perfectos para ser evaluados experimentalmente. La estrategia informática, predictiva, puede considerarse buena, pero requiere el sustento de un gran conocimiento biológico para interpretar los resultados, ya que las propuestas de nuevos sitios o corrección de sitios surgen a partir de observaciones basadas en conocimiento a priori de la UTD, del FTD, así como de la ubicación de los sitios de unión y sus posiciones respecto al promotor entre muchas otras observaciones.

Se observó que las matrices obtenidas con un parámetro de corte de 20 pb permiten identificar las regiones más conservadas en sitios a 10 pb. Ya que el análisis de estos patrones mostró los sitios de unión más conservados con secuencias consenso que reflejaban mejor las distintas organizaciones de los sitios de unión respecto a las reportadas en la literatura actual. Por otra parte las matrices con márgenes de corte más amplio (50 pb y 70 pb) permitieron recuperar los sitios de unión, grandes, que incluyen regiones tanto muy conservadas como regiones poco conservados, mayores a 10 pb.

9. PERSPECTIVAS

Una vez hecho el análisis *in silico*, para los casos interesantes de conservación divergente se pueden realizar diseños de trabajo experimental para comprobar nuestras predicciones. La mutación de genes que codifican FT y su sobre expresión es un método muy utilizado para analizar la expresión de los genes regulados, así como el enfoque que combina los datos de expresión de genes (microarreglos) con los análisis del sitio de pegado de un FT en la secuencia completa de un genoma. Adicionalmente, una vez que la secuencia del sitio de unión se ha inferido, es factible realizar mutagénesis dirigida para verificar su función. Por otro lado, también es posible usar la conservación divergente de un par de UTDs para extender y/o caracterizar nuevos regulones.

Para trabajo futuro, una vez que se tenga disponibilidad a bases de datos más robustas y completas, se propone el mismo enfoque que se realizó en esta tesis pero basado en otros organismos modelo como *Bacillus subtilis*, *Saccharomyces cerevisiae,* etc, con el fin de ampliar nuestra búsqueda. Inclusive se podría considerar realizar el análisis formando grupos de FTDs ortólogos pero sin tomar inicialmente organismos referencia para formar los grupos, ya que existen casos muy conservados en un conjunto de organismos, pero que no están presentes en el organismo modelo. Del mismo modo se podrían analizar los casos excluidos en esta tesis, especialmente las disposiciones genéticas en donde el FT forma parte del mismo operón que regula.

Y finalmente, considerando que un algoritmo de predicción de patrones no siempre puede ser el que arroje los mejores resultados, otra propuesta sería plantear una estrategia donde se conjunten diversos métodos de búsqueda de firmas con el objetivo de identificar el mismo patrón. De esta forma, al estar los métodos respaldados por diferentes bases estadísticas, las firmas identificadas en común serían objetos de análisis más confiables.

APÉNDICES

Apéndice 1 Relación de organismos no redundantes utilizados en el análisis, a la izquierda se presenta la abreviación utilizada para facilitar su manejo en el texto.

Aae	Aquifex aeolicus VF5
Aci	Acinetobacter sp. ADP1
Ama	Anaplasma marginale str. S t. Maries
Atu	Agrobacterium tumefaciens str. C58
Azo	Azoarcus sp. EbN1
Bap	Buchnera aphidicola str. APS (Acyrthosiphon pisum)
Bas	Bacillus subtilis subsp. subtilis str. 168
Bba	Bdellovibrio bacteriovorus HD100
Bbp	Buchnera aphidicola str. Bp (Baizong ia pistaciae)
Bbr	Bordetella bronchiseptica RB50
Bce	Bacillus cereus E33L
Bcl	Bacillus clausii KSM -K16
Bfr	Bacteroides fragilis YCH46
Bha	Bacillus halodurans C -125
Bhe	Bartonella henselae str. Houston -1
Bif	Bifidobacterium longum NCC2705
Bja	Bradyr hizobium japonicum USDA 110
Bli	Bacillus licheniformis ATCC 14580 (DSM 13)
Bor	Borrelia burgdorferi B31
Bps	Burkholderia pseudomallei K96243
Bqt	Bartonella quintana str. Toulouse
Bru	Brucella melitensis 16M
Bsg	Buchnera aphidicola str. Sg (Schizaphis graminum)
Bth	Bacteroides thetaiotaomicron VPI -5482
Bur	Burkholderia sp. 383
Cah	Corynebacterium glutamicum ATCC 13032
Caj	Campylobacter jejuni RM1221
Can	Candidatus Protochlamydia amoebophila UWE25
Cbf	Candidatus Blochmannia floridanus
Cbp	Candida tus Blochmannia pennsylvanicus str. BPEN
Cbu	Coxiella burnetii RSA 493
Cce	Clostridium acetobutylicum ATCC 824
Chc	Chlamydophila caviae GPIC
Chl	Chlamydophila abortus S26/3
Cho	Chlorobium chlorochromatii CaD3
Chy	Carboxydothermus hydrogenoformans Z -2901
Cje	Corynebacterium jeikeium K411
Clt	Clostridium tetani E88
Cmn	Chlamydia muridarum Nigg
Coe	Corynebacterium efficiens YS -314
Cor	Corynebacterium diphtheriae NCTC 13129
Cpe	Clostridium perfringens str. 13
Cpn	Chlamydophila pneumoniae TW -183
Cps	Colwellia psychrerythraea 34H
Cpu	Candidatus Pelagibacter ubique HTCC1062
Cre	Caulobacter crescentus CB15
Cte	Chlorobium tepidum TLS
Ctr	Chlamydia trachomatis A/HAR -13
Cvi	Chromobacterium violaceum ATCC 12472
Dar	Dechloromonas aromatica RCB
Dde	Desulfovibrio desulfuricans G20
Det	Dehalococcoides ethenogenes 195
Dps	Desulfotalea psychrophila LSv54
Dra	Deinococcus radiodurans R1
Dvu	Desulfovibrio vulgaris subsp. vulgaris str. Hildenborough
Eco	Escherichia coli K12
Efa	Enterococcus faecalis V583
Ehc	Ehrlichia canis str. Jake
Erc	Erwinia carotovora subsp. atroseptica SCRI1043
Eru	Ehrlichia ruminantium str. Welgevonden
Fnu	Fusobacterium nucleatum subsp. nucleatum ATCC 25586
Ftu	Francisella tularensis subsp. tularensis SCHU S4
Geo	Geobacillus kaustophilus HTA426
Ges	Geobacter sulfurreducens PCA
Glo	Gluconobacter oxydans 621H
Glv	Gloeobacter violaceus PCC 7421
Gme	Geobacter metallireducens GS -15
Hdu	Haemophilus ducreyi 35000HP
Hhe	Helicobacter hepaticus ATCC 51449
Hin	Haemophilus influenz ae 86 -028NP
Hpy	Helicobacter pylori 26695
Idi	Idiomarina loihiensis L2TR
Lac	Lactobacillus acidophilus NCFM
Laj	Lactobacillus johnsonii NCC 533
Lcl	Lactococcus lactis subsp. lactis II1403
Lep	Legionella pneumophila str. Paris
Lin	Listeria innocua Cl ip11262
Lis	Leptospira interrogans serovar Lai str. 56601
Lls	Lactobacillus plantarum WCFS1
Lxs	Leifsonia xyli subsp. xyli str. CTCB07
Mav	Mycobacterium avium subsp. paratuberculosis K -10
Mca	Methylococcus capsulatus str. Bath
Mel	Mesorhizobium loti MAFF303099
Mes	Mesoplasma florum L1
Mne	Mycoplasma pneumoniae M129
Mnn	Mannheimia succiniciproducens MBEL55E
Mpp	Mycoplasma penetrans HF -2
Mpu	Mycoplasma pulmonis UAB CTIP
Msy	Mycoplasma synoviae 53
Mtr	Mycobacterium tuberculosis H37Rv
Myc	Mycoplas ma gallisepticum R
Myg	Mycoplasma genitalium G37
Myh	Mycoplasma hyopneumoniae 7448
Myl	Mycobacterium leprae TN
Mym	Mycoplasma mobile 163K
Myp	Mycoplasma mycoides subsp. mycoides SC str. PG1
Nei	Neisseria meningitidis MC58
Nio	Nitrosococcus oceani AT CC 19707
Nit	Nitrosomonas europaea ATCC 19718
Nof	Nocardia farcinica IFM 10152
Nos	Nostoc sp. PCC 7120
Nwi	Nitrobacter winogradskyi Nb -255
Oih	Oceanobacillus iheyensis HTE831
Ony	Onion yellows phytoplasma OY -M
Pac	Propionibacterium acnes KPA171202
Pae	Pseudomonas aeruginosa PAO1
Pel	Pelobacter carbinolicus DSM 2380
Pfl	Pseudomonas fluorescens Pf -5
Pfo	Pseudomonas fluorescens PfO -1
Pgi	Porphyromonas gingivalis W83
Pha	Pseudoalteromonas haloplanktis TAC125
Pho	Photorhabdus luminescens subsp. lau mondii TTO1
Plu	Pelodictyon luteolum DSM 273
Pma	Prochlorococcus marinus str. MIT 9313
Pna	Prochlorococcus marinus str. NATL2A
Ppr	Photobacterium profundum SS9

Ppt	Pseudomonas putida KT2440
Prm	Prochlorococcus marinus subsp. marin CCMP1375
Prp	Prochlorococcus marinus subsp. pastc CCMP1986
Psm	Pasteurella multocida subsp. multocida
Psy	Pseudomonas syringae pv. tomato str. DC3000
Psy	Psychrobacter arcticus 273
Ral	Ralstonia solanacearum GMI1000
Rev	Ralstonia eutropha JMP134
Rfe	Rickettsia felis URRWXCal2
Rhi	Rhizobium etli CFN 42
Rho	Rhodopirellula baltica SH 1
Ric	Rickettsia conorii str. Malish 7
Rpa	Rhodopseudomonas palustris CGA009
Rph	Rhodobacter sphaeroides 2.4.1
Rpr	Rickettsia prowazekii str. Madrid E
Scc	Synechococcus sp. CC9902
Sch	Synechococcus sp. CC9605
Sco	Streptomyces coelicolor A3(2)
Sec	Synechocystis sp. PCC 6803
Sha	Staphylococcus haemolyticus JCSC1435
Sho	Shewanella oneidensis MR-1
Sil	Silicibacter pomeroyi DSS-3
Sin	Sinorhizobium meliloti 1021
Smu	Streptococcus mutans UA159
Sne	Streptococcus pneumoniae R6
Spy	Streptococcus pyogenes MGAS5005
Ssa	Staphylococcus saprophyticus subsp. saprophyticus ATCC 15305
Sta	Staphylococcus aureus subsp. aureus Mu
Ste	Staphylococcus epidermidis RP62A
Sth	Streptococcus thermophilus LMG 1831
Stm	Salmonella typhimurium LT2
Str	Streptococcus agalactiae 2603V/R
Sve	Streptomyces avermitilis MA-4680
Syc	Synechococcus sp. WH 8102
Sym	Symbiobacterium thermophilum IAM 14
Syn	Synechococcus elongatus PCC 6301
Tcr	Thiomicrospira crunogena XCL-2
Tde	Thiobacillus denitrificans ATCC 25525
Tel	Thermosynechococcus elongatus BP-1
Tfu	Thermobifida fusca YX
The	Thermoanaerobacter tengcongensis MB4
Thh	Thermus thermophilus HB27
Tpl	Treponema pallidum subsp. pallidum str. Nichols
Tre	Treponema denticola ATCC 35405
Tri	Thermotoga maritima MSB8
Tro	Tropheryma whipplei str. Twist
Ure	Ureaplasma parvum serovar 3 str. ATCC 700970
Vch	Vibrio cholerae O1 biovar eltor str. N16961
Vfi	Vibrio fischeri ES114
Vpa	Vibrio parahaemolyticus RIMD 2210633
Vvu	Vibrio vulnificus YJ016
Wdr	Wolbachia endosymbiont of Drosophila melanogaster
Wen	Wolbachia endosymbiont strain TRS of Brugia malayi
Wig	Wigglesworthia glossinidia endosymbiont of Glossina brevipalpis
Wol	Wolinella succinogenes DSM 1740
Xan	Xanthomonas campestris pv. campestris str. ATCC 3391
Xcv	Xanthomonas campestris pv. vesicatoria str. 85
Xor	Xanthomonas oryzae pv. oryzae KACC10331
Xyl	Xylella fastidiosa 9a5c
Ype	Yersinia pestis biovar Microtus str. 91001
Zym	Zymomonas mobilis subsp. mobilis ZM4

Apéndice 2. LISTA DE UTDs ANALIZADAS (Gpo = grupo de UTDO; R= gen regulador; E= operón estructural; D.I. = Distancia Intergénica del par divergente en *E. coli* K12; C.C. = Coeficiente de Correlación de expresión de cada par de UTDs calculado sobre 78 experimentos de microarreglos; Org = Organismos del grupo de UTDO; Regulador) A/R = Activación/Represión del gen regulador; (Estructural) A/R = Activación/Represión del gen Estructural; Familia reguladora del gen regulador).

Gpo	R	E	D.I	C.C	Org	Organismos	(Regulador) A/R	(Estructural) A/R	Familia reguladora
1	argR	mdh	434	0.2704	18	Cps;Hdu;Idi;Pha;Psm;Sho;Vch;Vpa;Ype;Erc;Hin;Mnn;Pho;Ppr;Stm;Vfi;Vvu;Eco.	-/-	CRP/ArcA, FlhDC	-
2	betI betB berA	betT	128	0.8752	5	Cor;Cje;Erc;Ype;Eco.	-		-
3	crp,yhfK	yhfA	301		14	Aci;Nio;Pfl;Pho;Ppt;Stm;Ype;Erc;Pae;Pfl;Ppr;Psy;Vfi;Eco.	CRP/CRP	CRP/CRP	CRP
4	glcC	glcDEFGBA	250	0.6966	6	Bbr;Pae;Ppt;Bps;Pfl;Eco.	CRP/GlcC	CLCC,IHF/ArcA	GntR
5	ilvY	ilvC	149	0.5344	18	Cps;Erc;Mnn;Pha;Psm;Sho;Vch;Vpa;Ype;Dps;Hin;Pel;Pho;Ppr;Stm;Vfi;Vvu;Eco.	-/IlvY	IlvY/-	LysR
6	lrp	trxB	544		8	Erc;Pho;Sho;Ype;Nwi;Rpa;Stm;Eco.	GadE/Lrp	-/-	AsnC
7	metR	metE	236	0.4419	22	Bap;Bps;Erc;Nit;Pho;Ppr;Rev;Sho;Tcr;Vfi;Ype;Bbr;Cvi;Mca;Pfl;Psm;Psy;Ral;Stm;Tde;Vvu;Eco.	-/MetJ, MerR	MetR/MetJ	LysR
8	norR	norV,norW	111	-	5	Erc;Stm;Vfi;Vvu;Eco.	-/NorR	IHF,NorR/FNR,IHF,NarL, NarP	EBP
9	prpR	prpB	239	0.6590	10	Bps;Pho;Stm;Xan;Xor;Bps;Rev;Xcv;Eco.	CRP/PrpR	CRP,PrpR/-	EBP
10	pspF	pspA, pspB, pspC,	152	0.7791 0.7617	17	Dde;Erc;Idi;Pho;Sho;Vch;Vpa;Ype;Eco;Dvu;Glo;Pha;Ppr;Stm;Vfi;Vvu;Zym.	-/PspF	IHF.PspF/-	EBP **RR**

		pspD		0.8143 0.8012					
11	cpxR	cpxP fieF pfkA	149	-	7	Eco;Stm;Erc;Ype;Pho;Vfi;Ppr			
12	malT	malQP	611	0.5071	7	Mnn;Ppr;Vfi;Eco;Pho;Stm;Ype.	CRP, Lrp/ DgsA	FNR,MalT/-	LuxR/UhpA
13	phoB	sbcD sbcC	189	0.7758	5	Erc;Stm;Eco;Pho;Ype.	PhoB/-	Por atenuación transcripcional	Two
14	torR	torC,torA, torD	129	0.6180 0.76011 0.6178	7	Ppr;Vch;Vpa;Eco;Stm;Vfi;Vvu.	-/TorR	TorR/NarL	Two
15	nagBACD	nagE	333		10	Bha;Psm;Vch;Vpa;Ype;Erc;Stm;Vfi;Vvu;Eco	CRP/CRP, NagC	CRP/CRP,NagC	NagC/XylR
16	cynR	cynT,cynS,cynX	108	0.0460 0.1959 -0.08625	8	Bps;Cvi;Pfl;Pho;Bps;Pae;Pfl;Eco	-/CynR	CynR/-	LysR
17	metJ	metB,metL	277	0.9031	14	Cps;Idi;Pho;Sho;Vch;Vpa;Ype;Erc;Pha;Ppr;Stm;Vfi;Vvu;Eco;	-/Fur	PhoP/MetJ	MetJ
18	narL narX	narK	338	0.55	6	Erc;Pae;Tde;Idi;Stm;Eco.	ModE/FNR		LuxR/UhpA
						Bloque II			
1	acrR	acrAB	141	-	15	Bru;Bps;Nei;Bur;Dar;Erc;Pfl;Pfo;Pho;Ppt ;Psy;Ral;Stm;Ype;Eco.	-/AcrR	MarA,Rob, SoxS/AcrR, PhoP	TetR/AcrR

2	dsdC	dsdX, dsdA	217	-	5	Cps;Stm;Eco;Pho;Vch.	-/DsdC	CRP,DNDC/-	LysR
3	lsrR	lsrACD BFG	248	-	7	Bce;Psm;Sin;Eco;Pho;Rph;Stm.	-	CRP/LsrR	-
4	putA	putP	422	0.6042	8	Erc;Pfl;Pfo;Pho;Ppt;Psy;Ype;Eco.	MarA/putA	-/CRP	-
5	rob	creA creB	210	0.7480	5	Erc;Pho;Stm;Ype;Eco.	-		Two components
6	uvrY	yecF	458	0.7399	5	Erc;Pho;Stm;Ype;Eco.	-		Predicted protein
7	fadR	nhaB	220	0.90201	14	Erc;Hdu;Hin;Mnn;Pho;Psm;Ppr;Sho;Stm;Vch;Vfi;Vpa;Ype;Eco.			
8	yeiE	yeiH	98	0.84	30	Atu;Bbr;Bce;Bru;Cah ;Caj ;Cps;Cvi;Erc;Gme;Ges;Idi;Lin;Lac;Lls;Nof;Oih;Pae;Par;Pel;Pho;Rev;Ral;Smu;Stm;Wol;Xcv;Xor;Ype;Eco.			
9	asnC	asnA	151	-	10	Erc;Idi;Pha;Psm;Ype;Hin;Mnn;Pho;Stm;Eco.	-/AsnC,Nac	AsnC/-	AsnC
10	lysR	lysA	121	0.7724	11	Bur;Cvi;Dar;Erc;Pfl ;Pho ;Ppt;Par ;Psy ;Stm;Ype;Eco.		C	Biosíntesis de lisina
11	bolA	yajG	169	0.7485	12	Cps;Erc;Mnn;Psm;Stm;Vfi;Ype;Eco;Hdu;Pho;Vch;Vpa;		Predicted lipoprotein	H-NS OmpR
12	yeaT	yeaU	81	0.82495	6	Bbr;Bps;Bur;Psy;Ype ;Eco.	-		Predicted Dehidrogenase
13	barA	rumA	56	-	14	Aci;Psy;Vfi;Pha;Sho;Erc;Pho;Stm;Vvu;Cps;Ppr;Vch;Vpa;Eco.	-	-	Putative methyltransferase
14	rob	creA creB	210	0.7480	5	Eco;Stm;Pho;Ype;Erc.	-*Rob es regulado por marA, es un FT		Two components

							conocido regulado por otro FT pero del cual no se sabe si se autorregula. Se identificaron dos DR de 11 pb, separados por 5 pb.		
						FTs que se sabe son reprimidos por otros FTs pero no se identificó el sitio de unión. Estos dos grupos se ubicarían también en el bloque IV			
15	rpoE rseA rseB rseC	nadB	407	-	21	Azo;Cps;Dar;Erc;Idi;Mca;Nio;Pae;Pfl;Pha;Pho;Ppt;Psy;Sho;Stm;Tcr;Tde;Vfi;Vvu;Ype;Eco.	/lexA		L-aspartato oxidasa
16	ompR	greB, yhgF,	191	0.5255	6	Eco;Stm;Ppr;Pha;Sho;Pho.	*ompR se regula por CRP e IHF pero no se identificó sitio de unión.		
						Bloque III			
Gpo	**R**	**E**	**D.I**	**C.C**	**Org**	**Organismos**	**(Regulador) A/R**	**(Estructural) A/R**	**Familia reguladora/ o función**
1	ydcN ydcP	ydcO	91	-	7	Cps;Stm;Vch;Eco;Erc;Vch;Eco.	-	-	-
2	yfeR	yfeH		0.8197	5	Cvi;Rev;Ral;Stm;Eco.			/Proteína de

									membrana
3	ytfH	ytfG	-		17	Atu;Cre;Cah;Erc;Lcl ;Mnn ;Nos;Psy;Rpa;Ral;Sin;Stm;Xcv;Xan;Xyl;Xor;Eco.			Quinona oxidorreductasa
4	ybhD	ybhH		0.6980	6	Bja;Atu;Ppt;Mel;Pfl ;Eco.			-
5	yqhD	yqhC		0.7153	8	Erc;Pha;Stm;Vfi;Vpa;Vvu;Ype;Eco.			Probable alcohol deshidrogenasa
6	yfeT	murQ murP yfeW	342	-	10	Cps;Pho;Stm;Vfi;Ype;Erc;Ppr;Vch;Vpa;Eco	-	-	
7	yhaJ	yhaK,yhaL	104		5	Erc;Stm;Eco;Pho;Ype.	-	-	-
8	ydhB	ydhC	113	-	10	Pho;Ppr;Sho;Stm;Vch;Vfi;Vpa;Vvu;Ype;Eco.	-	-	-
9	ynfL	ynfM		-	19	Aci;Azo;Bbr;Bru;Cvi;Erc;Pae;Pfl;Pfo;Pho;Ppt;Psy;Ssa;Stm;Xcv;Xor;Eco.			Probable proteína de transporte
						Bloque IV			
1	yneJ	yneI	77	-	6	Ppt;Pfo;Pfl;Psy;Stm;Eco.	-	-	LysR
2	ycjC	ycjL, ycjK	121	-	10	Cps;Pfl;Sve;Xcv;Xor;Pae;Pfo;Vpa;Xan;Eco.	-	-	-
3	yebK	zwf	337	-	12	Cps;Pae;Pfl;Ppt;Sho;Ype;Erc;Pfo;Pho;Psy;Stm; Eco.	-/-	MarA, Rob, SoxS/-	-
4	hdfR	yifE	119	-	11	Erc;Pha;Pho;Ppr;Sho;Stm;Vch;Vfi;Vpa;Vvu;Eco	-	-	LysR

BIBLIOGRAFÍA

1. **Adachi N and Lieber MR**. Bidirectional gene organization: A common architectural feature of the human genome. *Cell* 109: 807-809, 2002.

2. **Altschul SF, Madden TL, Schaffer AA, Zhang JH, Zhang Z, Miller W and Lipman DJ**. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25: 3389-3402, 1997.

3. **Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT, Harris MA, Hill DP, Issel-Tarver L, Kasarskis A, Lewis S, Matese JC, Richardson JE, Ringwald M, Rubin GM and Sherlock G**. Gene Ontology: tool for the unification of biology. *Nature Genetics* 25: 25-29, 2000.

4. **Bailey TL and y Elkan C**. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* 2: 28-36, 1994.

5. **Barnard A, Wolfe A and Busby S**. Regulation at complex bacterial promoters: how bacteria use different promoter organizations to produce different regulatory outcomes. *Current Opinion in Microbiology* 7: 102-108, 2004.

6. **Beck CF WRA**. Divergent promoters, a common form of gene organization. *Microbiol Rev* 52: 318-326, 1988.

7. **Bedouelle H, Schmeissner U, Hofnung M and Rosenberg M**. Promoters of the malEFG and malK-lamB operons in Escherichia coli K12. *J Mol Biol* 161: 519-531, 1982.

8. **Bell PJL, Bissinger PH, Evans RJ and Dawes IW**. A 2 Reporter Gene System for the Analysis of Bidirectional Transcription from the Divergent Mal6T Mal6S Promoter in Saccharomyces-Cerevisiae. *Current Genetics* 28: 441-446, 1995.

9. **Blanchette M and Tompa M**. Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Research* 12: 739-748, 2002.

10. **Boos W and Bohm A**. Learning new tricks from an old dog - MaIT of the Escherichia coli maltose system is part of a complex regulatory network. *Trends in Genetics* 16: 404-409, 2000.

11. **Boos W and Shuman H**. Maltose/maltodextrin system of Escherichia coli: Transport, metabolism, and regulation. *Microbiology and Molecular Biology Reviews* 62: 204-+, 1998.

12. **Borukhov S and Severinov K**. Role of the RNA polymerase sigma subunit in transcription initiation. *Research in Microbiology* 153: 557-562, 2002.

13. **C.F.Beck and R.A.J.Warren**. Divergent Promoters, a Common Form of Gene Organization. *Microbiological Reviews* 52: 318-326, 1988.

14. **Campbell JW and Cronan JE**. Escherichia coli FadR positively regulates transcription of the fabB fatty acid biosynthetic gene. *Journal of Bacteriology* 183: 5982-5990, 2001.

15. **Campoy S, Mazon G, Fernandez de Henestrosa AR, Llagostera M, Monteiro PB and Barbe J**. A new regulatory DNA motif of the gamma subclass proteobacteria: identification of the LexA protein binding site of the plant pathogen *Xylella fastidiosa*. *Microbiology* 148: 3583-3597, 2002.

16. **Charpentier B BC**. The Escherichia coli gapA gene is transcribed by the vegetative RNA polymerase holoenzyme E sigma 70 and by the heat shock RNA polymerase E sigma 32. *J Bacteriol* 176: 830-839, 1994.

17. **Collado-Vides J MBGJD**. Control site location and transcriptional regulation in Escherichia coli. *Microbiol Rev* 55: 371-394, 1991.

18. **Dandekar T, Snel B, Huynen M and Bork P**. Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences* 23: 324-328, 1998.

19. **Date SV and Marcotte EM**. Discovery of uncharacterized cellular systems by genome-wide analysis of functional linkages. *Nature Biotechnology* 21: 1055-1062, 2003.

20. **DiRusso CC HTMAK**. Characterization of FadR, a global transcriptional regulator of fatty acid metabolism in Escherichia coli. Interaction with the fadB promoter is prevented by long chain fatty acyl coenzyme A. *J Biol Chem* 267: 8685-8691, 1992.

21. **DiRusso CC MAHTL**. Regulation of transcription of genes required for fatty acid transport and unsaturated fatty acid biosynthesis in Escherichia coli by FadR. *Molecular Microbiology* 7: 311-322, 1993.

22. **Gallegos MT, Schleif R, Bairoch A, Hofmann K and Ramos JL**. AraC/XylS family of transcriptional regulators. *Microbiology and Molecular Biology Reviews* 61: 393-&, 1997.

23. **Gralla JD**. Activation and repression of E-coli promoters. *Current Opinion in Genetics & Development* 6: 526-530, 1996.

24. **Gruber TM and Gross CA**. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annual Review of Microbiology* 57: 441-466, 2003.

25. **Gui LZ, Sunnarborg A and LaPorte DC**. Regulated expression of a repressor protein: FadR activates iclR. *Journal of Bacteriology* 178: 4704-4709, 1996.

26. **Guido Beny, Raymond Cunin, Nicolas Glansdorff, Anne Boyen, Josée Charlier and Norman Kelker**. Transcription of Regions Within the Divergent argECBH Operon of Escherichia coli: Evidence for Lack of an Attenuation Mechanism. *Journal of Bacteriology* 151: 58-61, 1982.

27. **Harley C.B and Reynolds R.P**. Analysis of E.coli promoter sequences. *Nucleic Acids Research* 15: 2343-2361, 1987.

28. **Hawley D.K and and McClure W.R**. Compilation and analysis of Escherichia coli promoter DNA sequences. *Nucleic Acids Research* 11: 2237-2255, 1983.

29. **Helmann JD CMJ**. Structure and function of bacterial sigma factors. *Annu Rev Biochem* 57: 839-872, 1988.

30. **Hertz GZ and Stormo GD**. Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15: 563-577, 1999.

31. **Hidalgo E DB**. An iron-sulfur center essential for transcriptional activation by the redox-sensing SoxR protein. *EMBO J* 13: 138-146, 1994.

32. **Hughes JD, Estep PW, Tavazoie S and Church GM**. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *Journal of Molecular Biology* 296: 1205-1214, 2000.

33. **Jacob F, Perrin D, Sanchez C and Monod J**. Operon: a group of genes with the expression coordinanted by an operator. *C R Hebd Seances Acad Sci* 250: 1727-1729, 1960.

34. **Keseler IM, Collado-Vides J, Gama-Castro S, Ingraham J, Paley S, Paulsen IT, Peralta-Gill M and Karp PD**. EcoCyc: a comprehensive database resource for Escherichia coli. *Nucleic Acids Research* 33: D334-D337, 2005.

35. **Kevin P.Bertrand, Kathleen Postle, Lewis V.Wray Jr and William S.Reznikoff**. Overlapping divergent promoters control expression of Tn10 tetracycline resistance. *Elsevier Science Publishers B V* 23: 149-156, 1983.

36. **Khorchid A and Ikura M**. Bacterial histidine kinase as signal sensor and transducer. *International Journal of Biochemistry & Cell Biology* 38: 307-312, 2006.

37. **Korbel JO, Jensen LJ, von Mering C and Bork P**. Analysis of genomic context: prediction of functional associations from conserved bidirectionally transcribed gene pairs. *Nature Biotechnology* 22: 911-917, 2004.

38. **Koyanagi KO, Hagiwara M, Itoh T, Gojobori T and Imanishi T**. Comparative genomics of bidirectional gene pairs and its implications for the evolution of a transcriptional regulation system. *Gene* 353: 169-176, 2005.

39. **Lee DH HLSR**. Repression of the araBAD promoter from araO1. *J Mol Biol* 224: 335-341, 1992.

40. **Li ZY and Demple B**. Sequence specificity for DNA binding by Escherichia coli SoxS and Rob proteins. *Molecular Microbiology* 20: 937-945, 1996.

41. **Lisser S and Margalit H**. Compilation of E. coli mRNA promoter sequences. *Nucleic Acids Research* 21: 1507-1516, 1993.

42. **Liu YL and Xiao W**. Bidirectional regulation of two DNA-damage-inducible genes, MAG1 and DDI1, from Saccharomyces cerevisiae. *Molecular Microbiology* 23: 777-789, 1997.

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

44. **Malcolm J.Casadaban**. Regulation of the Regulatory Gene for the Arabinose Pathway, araC. *Journal of Molecular Biology* 104: 557-566, 1976.

45. **Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO and Eisenberg D**. Detecting protein function and protein-protein interactions from genome sequences. *Science* 285: 751-753, 1999.

46. **Martinez-Antonio A and Collado-Vides J**. Identifying global regulators in transcriptional regulatory networks in bacteria. *Current Opinion in Microbiology* 6: 482-489, 2003.

47. **Martinez-Antonio A, Janga SC, Salgado H and Collado-Vides J**. Internal-sensing machinery directs the activity of the regulatory network in Escherichia coli. *Trends in Microbiology* 14: 22-27, 2006.

48. **Moreno-Hagelsieb G and Collado-Vides J**. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* 18: 329-336, 2002.

49. **Moreno-Hagelsieb GyC-VJ**. Operon conservation from the point of view of *Escherichia coli*, and inference of functional interdependence of gene products from genome context. *In Silico Biol* 2: 87-95, 2002.

50. **Norregaardmadsen M, Mcfall E and Valentinhansen P**. Organization and Transcriptional Regulation of the Escherichia-Coli K-12 D-Serine Tolerance Locus. *Journal of Bacteriology* 177: 6456-6461, 1995.

51. **Opel ML, Arfin SM and Hatfield GW**. The effects of DNA supercoiling on the expression of operons of the ilv regulon of Escherichia coli suggest a physiological rationale for divergently transcribed operons. *Molecular Microbiology* 39: 1109-1115, 2001.

52. **Overbeek R, Fonstein M, D'Souza M, Pusch GD and Maltsev N**. The use of gene clusters to infer functional coupling. *Proceedings of the National Academy of Sciences of the United States of America* 96: 2896-2901, 1999.

53. **Pavesi G, Mereghetti P, Mauri G and Pesole G**. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Research* 32: W199-W203, 2004.

54. **Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D and Yeates TO**. Assigning protein functions by comparative genome analysis: Protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America* 96: 4285-4288, 1999.

55. **Pomposiello PJ and Demple B**. Redox-operated genetic switches: the SoxR and OxyR transcription factors. *Trends in Biotechnology* 19: 109-114, 2001.

56. **Ramírez Santos J, Solís Guzmán G and Gómez Eichelmann MC**. Genetic regulation of the heat-shock response in Escherichia coli. *Rev Latinoam Microbiol* 43: 51-63, 2001.

57. **Reitzer L and Schneider BL**. Metabolic context and possible physiological themes of sigma(54)-dependent genes in Escherichia coli. *Microbiology and Molecular Biology Reviews* 65: 422-+, 2001.

58. **Rhodius VA BSJ**. Positive activation of gene expression. *Curr Opin Microbiol* 1: 152-159, 1998.

59. **Richet E**. On the role of the multiple regulatory elements involved in the activation of the Escherichia coli malEp promoter. *J Mol Biol* 264: 852-862, 1996.

60. **Salgado H, Gama-Castro S, Peralta-Gil M, Diaz-Peredo E, Sanchez-Solano F, Santos-Zavaleta A, Martinez-Flores I, Jimenez-Jacinto V, Bonavides-Martinez C, Segura-Salazar J, Martinez-Antonio A and Collado-Vides J**. RegulonDB (version 5.0): Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Research* 34: D394-D397, 2006.

61. **Salgado H, Moreno-Hagelsieb G, Smith TF and Collado-Vides J**. Operons in Escherichia coli: Genomic analyses and predictions. *Proceedings of the National Academy of Sciences of the United States of America* 97: 6652-6657, 2000.

62. **Schleif R**. *L-arabinose operon. In Escherichia coli and Salmonella typhimurium*. **Washington, DC**: **American Society for Microbiology**, 1987.

63. **Schleif R**. AraC protein: a love-hate relationship. *Bioessays* 25: 274-282, 2003.

64. **Stock AM, Robinson VL and Goudreau PN**. Two-component signal transduction. *Annual Review of Biochemistry* 69: 183-215, 2000.

65. **Tatusov RL, Koonin EV and Lipman DJ**. A genomic perspective on protein families. *Science* 278: 631-637, 1997.

66. **Thijs G, Lescot M, Marchal K, Rombauts S, De Moor B, Rouze P and Moreau Y**. A higher-order background model improves the detection of promoter regulatory elements by Gibbs sampling. *Bioinformatics* 17: 1113-1122, 2001.

67. **Tompa M, Li N, Bailey TL, Church GM, De Moor B, Eskin E, Favorov AV, Frith MC, Fu YT, Kent WJ, Makeev VJ, Mironov AA, Noble WS, Pavesi G, Pesole G, Regnier M, Simonis N, Sinha S, Thijs G, van Helden J, Vandenbogaert M, Weng ZP, Workman C, Ye C and Zhu Z**. Assessing computational tools for the discovery of transcription factor binding sites. *Nature Biotechnology* 23: 137-144, 2005.

68. **Trinklein ND, Aldred SF, Hartman SJ, Schroeder DI, Otillar RP and Myers RM**. An abundance of bidirectional promoters in the human genome. *Genome Research* 14: 62-66, 2004.

69. **Tsung K BRIM**. Identification of the DNA-binding domain of the OmpR protein required for transcriptional activation of the ompF and ompC genes of Escherichia coli by in vivo DNA footprinting. *J Biol Chem* 264: 10104-10109, 1989.

70. **van Helden J, Andre B and Collado-Vides J**. Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *Journal of Molecular Biology* 281: 827-842, 1998.

71. **Vicente M, Chater KF and de Lorenzo V**. Bacterial transcription factors involved in global regulation. *Molecular Microbiology* 33: 8-17, 1999.

72. **Wade JT, Roa DC, Grainger DC, Hurd D, Busby SJW, Struhl K and Nudler E**. Extensive functional overlap between sigma factors in Escherichia coli. *Nature Structural & Molecular Biology* 13: 806-814, 2006.

73. **Wickstrum JR, Skredenske JM, Kolin A, Jin DJ, Fang J and Egan SM**. Transcription activation by the DNA-binding domain of the AraC family protein RhaS in the absence of its effector-binding domain. *Journal of Bacteriology* 189: 4984-4993, 2007.

74. **Wilcox G, Meuris P, Bass R and Englesberg E**. Regulation of the L-arabinose operon BAD in vitro. *J Biol Chem* 249: 2946-2952, 1974.

75. **Wilson KP, Shewchuk LM, Brennan RG, Otsuka AJ and Matthews BW**. *Escherichia coli* biotin holoenzyme synthetase/bio repressor crystal structure delineates the biotin- and DNA binding domains. *Proc Nat Acad Sci* 89: 9257-9261, 1992.

76. **Yamada M, Kabir MS and Tsunedomi R**. Divergent promoter organization may be a preferred structure for gene control in Escherichia coli. *Journal of Molecular Microbiology and Biotechnology* 6: 206-210, 2003.

77. **Yuhai Cui, Michael A.Midkiff, Qing Wang and Joseph M.Calvo**. **The Leucine-responsive Regulatory Protein (Lrp) from *Escherichia coli***. *American Society for Biochemistry and Molecular Biology* 271: 6611-6617, 1996.

78. **Zheng DL, Constantinidou C, Hobman JL and Minchin SD**. Identification of the CRP regulon using in vitro and in vivo transcriptional profiling. *Nucleic Acids Research* 32: 5874-5893, 2004.