UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

DOCTORADO EN
CIENCIAS BIOMÉDICAS

CENTRO DE CIENCIAS GENÓMICAS

"Desarrollo de una nueva estrategia para la predicción de sitios de unión de factores transcripcionales en bacterias, basado en el análisis de huellas filogenéticas y su caracterización biológica: la familia LysR como modelo de estudio"

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

PRESENTA:
PATRICIA MARÍA RUFINA
OLIVER OCAÑO

TUTOR PRINCIPAL:
DR. ENRIQUE MERINO PÉREZ ..INSITUTO DE BIOTECNOLOGÍA

CIUDAD DE MÉXICO A 24 DE FEBRERO DEL 2016

Agradecimiento al programa de Posgrado en Ciencias Biomédicas de la Universidad Nacional Autónoma de México (UNAM). Este fué un proyecto realizado gracias al donativo 23556 y a la beca para posgrado 45230 otorgados por CONACYT.

Tutor Principal:

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

Comité tutoral:

Dr. Ernesto Pérez Rueda
Instituto de Biotecnología (IBT), UNAM.

Dr. Sergio Encarnación Guevara
Centro de Ciencias Genómicas (CCG), UNAM.

Miembros del Jurado:

Presidente:

Dr. Miguel Ángel Cevallos Gaos
CCG-UNAM

Vocal:

Dr. Juan Miranda Ríos
IIBO–UNAM

Secretario:

Dr. Enrique Merino Pérez
IBT–UNAM

Vocal:

Dra. María de Lourdes Girard Cuesy
CCG–UNAM

Vocal:

Dra. Bertha María Josefina

González Pedrajo
IFC–UNAM

CONTENIDO

RESUMEN

En bioinformática, la mayoría de los algoritmos desarrollados para identificar sitios de unión de factores transcripcionales (TFs, por ***T**ranscriptional **F**actors*) se basan en identificar patrones estadísticamente significativos, sobre-representados en un conjunto de secuencias, entre las que se espera que el TF reconozca y se una a sus correspondientes sitios de unión en el DNA (TFBSs, por ***T**ranscriptional **F**actor **B**inding **S**ites*). A pesar de usarse con frecuencia, estas estrategias aún son poco precisas para identificar a los verdaderos TFBSs, en especial si son sitios de baja afinidad (degenerados) que pueden ser de importancia crucial para el sistema de regulación.

En esta tesis se presenta el protocolo “Perfil Filogenético de Motivos Consenso” (PProCoM, por ***P**hylogenetic **Pro**file of **Co**nsensus **M**otifs*), una estrategia bioinformática para identificar TFBSs mediante el análisis de huellas filogenéticas y la integración de información relativa a los aspectos biológicos implicados en el proceso de unión TF-DNA, como son: *i*) la conformación homodimérica activa de algunos TFs, la cual impone la formación de una estructura simétrica en los TFBSs, *ii*) la cooperatividad de unión de algunos TFs, *iii*) el efecto de la presencia/ausencia de metabolitos coinductores, *iv*) la distancia entre dos TFBSs o bien, entre promotores y TFBS, *v*) secuencias ricas en A/T como parte integral del TFBS, y *vi*) el orden de participación dinámico de los diferentes eventos de unión, para determinar una respuesta reguladora.

En este trabajo se analizaron las regiones intergénicas de seis FTs de tipo LysR en Gammaproteobacterias, con base en el organismo modelo *E. coli K12*. Como resultado

se presenta la arquitectura de los TFBSs identificados para cada sistema, la propuesta de una secuencia consenso (5'-CTATAtcattatgaTATAG-3') para los miembros de la familia LysR y un nuevo modelo de regulación congruente con las evidencias experimentales reportadas.

ABSTRACT

The goal of most Bioinformatic programs developed to find transcription factor binding sites (TFBSs) is the identification of discrete sequence motifs that are significantly over-represented in a given set of sequences where a transcription factor (TF) is expected to bind. Despite their extensive use, the accuracies reached with these programs remain low. In many cases, true TFBSs are excluded from the identification process, especially when they correspond to low-affinity but important binding sites of regulatory systems.

This thesis presents a computational protocol named PProCoM (Phylogenetic Profile of Consensus Motifs), this is a bioinformatic pipeline based on molecular and structural criteria to perform biologically meaningful and accurate phylogenetic footprinting analyses. PProCoM protocol considers fundamental aspects of the TF-DNA binding process, such as: i) the active homodimeric conformations of TFs that impose symmetric structures on the TFBSs, ii) the cooperative binding of TFs, iii) the effects of the presence or absence of co-inducers, iv) the proximity between two TFBSs or one TFBS and a promoter that leads to very long spurious motifs, v) the presence of AT-rich sequences not recognized by the TF but that are required for DNA flexibility, and vi) the dynamic order in which the different binding events take place to determine a regulatory response (i.e., activation or repression).

In this job, the abovementioned criteria were used to analyze a profile of consensus motifs generated from canonical Phylogenetic Footprinting Analyses using a set of analysis windows of incremental sizes. To evaluate the performance of our protocol, we

analyzed six members of the LysR-type TF family in Gammaproteobacteria. As a result we present the architecture of TFBSs identified for each system, the proposal of a new consensus sequence (5'-CTATAtcattatgaTATAG-3 ') for members of the LysR family and a new regulation model, congruent with experimental evidence.

DEDICATORIA

Dedico este trabajo al risueño de mi padre, de espíritu presente y de cuerpo ausente, te extrañaré todos los días de mi vida. Y a ti madre, por tu resistencia, inteligencia y amor, gracias hermosa mujer.

AGRADECIMIENTO

Quiero expresar mi profunda gratitud a mis supervisores y amigos el Dr. Enrique Merino Pérez y el Dr. Martín Peralta Gil, ya que su experiencia profesional, sus consejos y orientación durante mis estudios de doctorado fueron determinantes para mi formación, hubiera sido imposible terminar mi tesis sin su dirección académica y apoyo personal. Un agradecimiento especial al Ing. Ricardo Ciria Mercé por su asesoría constante en cuestiones técnicas computacionales y por poner a mi disposición la base de datos que utilicé en la realización de este trabajo, asimismo a la técnica Maria Luisa Tabche por su participación en el análisis experimental y su amistad.

Durante mis estudios de posgrado aprendí cosas muy valiosas del equipo de trabajo del Dr. Merino, de la Dra. Espín y del Dr. Collado, les agradezco a todos los colegas y amigos del Instituto de Biotecnología y el Centro de Ciencias Genómicas porque hicieron más agradable mi estancia ahí con su amistad, gracias Pablo Loera y Arturo Medrano por todas sus enseñanzas técnicas, a Santiago Castillo, Irma Vichido, Vero Jiménez, Rosa María, Raúl Noguez, y muchos amigos que hice en el día a día, omito nombrarlos a todos por cuestión de espacio, pero cada quien sabe lo mucho que los aprecio y les agradezco los momentos compartidos.

Agradezco a mis padres Francisco Oliver y Ana Alicia Ocaño que siempre me guiaron con cariño y respeto, esto es resultado del esfuerzo de taquería Los ASES, de donde vienen mis mejores memorias. A mis herman@s Claudia, Miguel, Lorenia, Luis y Natalie por su cariño y presencia constante en mis días, a mi esposo Carlos Pineda por su amor, compañía y tolerancia, porque trabajar en familia y para una familia es un

proceso difícil pero satisfactorio y alentador, y a ti Juliana Marie, porque caminar a tu lado me ha permitido ser la expectadora en primera fila de esta aventura del desarrollo humano, te amo hijita.

ABREVIATURAS

DNA	Ácido desoxirribonucléico (*Desoxyribonucleic acid*)
LTTR	Factores de Transcripción de tipo LysR (*LysR-Type transcriptional regulator family proteins*)
mRNA	RNA mensajero
nt	nucleótidos
PProCoM	Perfil filogenético de motivos consenso (*Phylogenetic Profile of Consensus Motifs*)
PWM	Matrices de posición peso-específicas (*Position-Weight Matrices)*
RNA	Ácido ribonucleico (Ribonucleic acid)
RNApol	RNA polimerasa
TF	Factor Transcripcional
TFBS	Sitio de unión a factores transcripcionales (*Transcription Factor Binding Site*)
TSS	Sitio de inicio de la transcripción (Transcription Start Site)
TG	Gene Blanco o *Target Gene*

1 INTRODUCCIÓN

Los seres vivos se caracterizan por su capacidad de responder a los estímulos de su exterior. Desde el organismo unicelular más primitivo, hasta los animales más complejos, son capaces de percibir cambios en su medio y reaccionar de una manera apropiada. Sus respuestas, en un momento dado requerirán de cambios en el estado de expresión del genoma, es decir, genes que estaban prendidos necesitarán ser apagados y genes latentes necesitarán ser activados. En los organismos bacterianos, en última instancia, la regulación de la expresión genética ocurre primordialmente al inicio de la transcripción y está mediada comúnmente por proteínas reguladoras, también llamadas factores transcripcionales (TF, por sus siglas en inglés *Transcription Factors*). Los TFs inhiben o favorecen la transcripción de acuerdo a las necesidades metabólicas de la célula.

Se ha observado que el proceso de control de la transcripción de un gen procariota, en general depende de la frecuencia con la que la RNA polimerasa (RNApol) reconoce a sus promotores (“fuerza del promotor”) y de la presencia de proteínas activadoras y represoras que actúan modulando dicho reconocimiento al unirse a las secuencias de DNA específicas cercanas al promotor. Los TFs son las proteínas que directamente modulan la transcripción de los genes, mientras que los sitios en el DNA a los cuáles se unen se les conoce como sitios de unión de los factores transcripcionales o simplemente TFBS (por sus siglas en inglés *Transcriptional Factor Binding Sites*). La capacidad de los TFs para incrementar o decrecer la transcripción de los genes (activar/reprimir), depende de las necesidades metabólicas de la célula y juegan papeles

muy importantes en la regulación transcripcional de los genes. Dado lo anterior, la identificación precisa de los TFBSs de cada uno de los Factores transcripcionales de un organismo resulta crítica para comprender la regulación biológica en la célula. A pesar de que hoy en día existen repositorios públicos que hacen disponibles muchas secuencias genómicas, los elementos funcionales codificados en el DNA, tales como los TFBSs, no se han caracterizado completamente. Esto se debe en parte, a la variada complejidad que presenta la actividad de unión de los TFs y a la degeneración de la secuencia del sitio de unión que reconocen. En este sentido se han desarrollado una variedad de técnicas experimentales y de programas de cómputo especializados para identificar los diferentes TFBSs en un genoma. Las predicciones de TFBSs *in silico* se realizan mediante la técnica llamada *huellas filogenéticas* (6-9) de los motivos de DNA. Este enfoque utiliza la comparación de secuencias nucleotídicas entre genomas ortólogos, y se basa tanto en la identificación de secuencias consenso, como en la determinación de matrices representativas del sitio de unión. Aunque con este método se pretenden identificar las secuencias conservadas, tanto los promotores como los TFBSs, son elementos reguladores que poseen variabilidad en sus secuencias nucleotídicas, por lo tanto, su identificación resulta muy compleja. Para la RNApol, se sabe que su afinidad de unión por el promotor será mayor si su secuencia nucleotídica a dicho promotor es más conservada respecto al consenso reconocido por los diferentes factores sigma. Para el caso de un determinado TF, se ha observado que su afinidad de unión por sus TFBSs será mayor o menor según el grado de conservación de la secuencia nucleotídica del TFBS respecto a una secuencia consenso específica, descrita para el TF. Dicha conservación se origina por la fuerza selectiva que permite el reconocimiento molecular de una región de DNA de manera específica. La variabilidad de las secuencias

nucleotídicas en estos elementos de regulación es muy relevante ya que determina con qué afinidad el TF reconocerá a sus TFBSs. En otras palabras, sitios cercanos al consenso serán de mayor afinidad, mientras que alejados del consenso serán reconocidos con menor afinidad.

Según análisis independientes, se ha determinado que el número relativo de moléculas de TFs en genomas bacterianos típicamente se incrementa cuadráticamente con el número total de genes en el genoma (2). Para el organismo modelo *Escherichia coli K12,* que posee un total de 4405 genes, a la fecha se ha estimado que aproximadamente el 8% de sus genes codifican para TFs conocidos o predichos (3) de los cuales, el 35% corresponde a activadores, el 43% a represores y el 22% con actividad dual (4).

La identificación *in silico* de TFBSs representa un tema clave para muchos estudios de Biología Molecular, cuyos objetivos son caracterizar los elementos reguladores en secuencias genómicas. Este tipo de análisis se ha realizado a través de varios enfoques, considerando tanto a diferentes genes coregulados en un genoma (5), como a un conjunto de secuencias de DNA localizadas río arriba de genes ortólogos (6-9) asumiendo que la conservación de nucleótidos de una región específica de DNA, en un conjunto de secuencias, corresponde a los TFBS que reconocen los TFs. Lo anterior es el principio en el cual se basan muchos algoritmos computacionales, y han sido desarrollados para identificar las secuencias más sobre-representadas (motivos) en un conjunto de secuencias dado, donde se espera que un TF una. Se considera que esos motivos son parte de los TFBS y comúnmente son representados como matrices específicas de peso por posición (PWM, por sus siglas en inglés *Position-Weight Matrices*). Los TFBSs y sus correspondientes PWMs se han compilado en diferentes

bases de datos tales como RegulonDB (5), EcoCyc (10), RegPrecise (11), Prodoric (12) y Tractor DB (13). Para evaluar la significancia estadística de la predicción de TFBSs, se han desarrollado diferentes enfoques basados en los modelos teóricos tales como valores *log-odds* o *entropy-weighted* (14) o la combinación de valores de distribución teóricos y empíricos (15). Con estas estrategias, los verdaderos TFBSs en muchos casos se excluyen o se identifican con poca precisión, especialmente cuando corresponden a sitios de baja afinidad que son de gran relevancia en el sistema de regulación. Dicho de otra forma, el hecho de que la significancia estadística de un patrón conservado esté dada por su sobre-representación en un conjunto de secuencias de genes coregulados, no es necesariamente la mejor vía para identificar la totalidad del conjunto de TFBSs de un regulón.

2 ANTECEDENTES

En este capítulo se presentan los conceptos biológicos necesarios para comprender el trabajo de esta tesis, y se incluye una explicación muy extensa sobre los motivos de unión a TFs y la forma en la que se representa su identificación mediante enfoques experimentales y computacionales.

2.1 Regulación de la expresión genética

Las células responden a estímulos internos o externos cambiando la expresión de sus genes, a través de procesos en los que cada gen es transcrito para transmitir información a los ribosomas que a su vez sintetizarán las proteínas. Desde el punto de vista de la expresión genética, los genes pueden ser: 1) de expresión constitutiva ("*housekeeping*") es decir, que se transcriben permanentemente independientemente de las condiciones ambientales (*i.e.* operones para las DNApol y RNApol; las proteínas ribosómicas, *etc.*), y 2) genes de expresión regulada en función de las condiciones ambientales.

En la célula, los procesos de síntesis proteica requieren un costo energético, por lo que ésta tiende a optimizar el uso de la energía disponible. En ese sentido, para el caso de genes cuya expresión es regulable, son esenciales los mecanismos de control de la regulación, ya que representan un componente crítico en la regulación del metabolismo celular y en el mantenimiento de las diferencias estructurales y funcionales que existen en las células durante el desarrollo. En los procesos celulares, se han descrito diversas etapas donde la cantidad de proteína puede regularse: 1) síntesis del transcrito primario de RNA,

2) procesamiento post-transcripcional del mRNA, 3) degradación del mRNA, 4) traducción, 5) modificación de la proteína, y 6) degradación de proteínas. La participación de los RNAs no codificantes (ncRNAs) se asocia también a la regulación del flujo de información de DNA a proteínas. Los ncRNAs incluyen a los RNAs de transferencia (tRNAs), RNAs ribosomales y otra variedad tales como los snoRNAs, microRNAs, siRNAs, exRNAs piRNAs, scaRNAs y long ncRNAs.

2.2 Mecanismos de regulación genética

Los diversos niveles de los que depende la expresión de la información genética (*i.e.* nivel transcripcional, traduccional y postraduccional) pueden estar sometidos a algún tipo de regulación. La regulación en principio puede llevarse a cabo por procesos de inducción o represión. La inducción ocurre cuando se sintetizan ciertas enzimas debido a la presencia en el medio de sustratos o estímulos ambientales (*i.e.* la producción de la enzima β -galactosidasa se induce en determinadas bacterias cuando en el medio aparece un azúcar de tipo β -galactósido (lactosa)). La represión puede ocurrir cuando hay una desconexión rápida de la ruta biosintética de un determinado compuesto, cuando éste aparece aportado en el medio de la bacteria (*i.e.* si *E. coli.* crece en ausencia de triptófano (Trp), la ruta para su biosíntesis funciona hasta que ese aa. aparece en el medio). Además de atender un estímulo nutricional, la represión de genes también ocurre para evitar que su expresión interfiera con otros procesos que ya están en curso en la célula.

A nivel transcripcional, la regulación puede darse: a) al inicio de la transcripción, b) por terminación prematura de la transcripción (atenuación de la transcripción), y c) por

procesamiento de RNA (casos raros en procariotas). A nivel traduccional, la regulación ocurre por la síntesis de proteínas ribosómicas ó por regulación de RNA antiparalelo que interfiere con la traducción del RNAm. A nivel postraduccional, los mecanismos no son puramente de regulación genética, y pueden ocurrir por degradación de proteínas, modificación covalente de proteínas (*i.e.* fosforilación) y regulación alostérica por retroalimentación (*feed-back*) de la actividad de las proteínas enzimáticas.

Tanto la inducción como la represión pueden ser de tipo positivo y negativo:

1. Control negativo: el sistema se expresa a menos que sea reprimido por la acción de una proteína reguladora denominada represor. Dentro de este control se distinguen a su vez:
    a. Control negativo con efectos inductores (*i.e*: el operón *lac*): el represor, *per se* es activo, pero se inactiva en presencia del inductor.
    b. Control negativo con efectos represores (*i.e*: en el operón *trp*): el represor, *per se* es inactivo, pero en presencia del correpresor se activa (adquiere su capacidad funcional), y es entonces cuando reprime al operón estructural.
2. Control positivo: cuando los genes estructurales no se transcriben (o en todo caso lo hacen a un nivel basal bajo) a no ser que exista una proteína reguladora activa llamada activadora. También aquí puede haber dos categorías:
    a. Control positivo por inducción: la proteína activadora, por sí misma es inactiva, pero queda activada cuando se le une el inductor.
    b. Control positivo por represión: la proteína activadora, *per se*, es activa, pero se inactiva cuando se le une el correpresor.

2.2.1 Regulación al inicio de la transcripción

La transcripción es un proceso en el que la información genética del DNA se copia a mRNA. En bacterias, controlar el inicio de este proceso constituye el mecanismo básico para regular la expresión de los genes, y su regulación puede ocurrir por dos vías *i*) sustitución del factor σ de la RNApol ó *ii*) por interacción de proteínas reguladoras sobre secuencias de DNA cercanas al promotor (fenómenos de inducción y represión génica).

En *E. coli*, existen diversas familias de factores σ que interactúan distinto dependiendo del gen y de las diferentes condiciones medioambientales. Un ejemplo de desplazamiento de subunidades σ lo representa la respuesta al choque por calor en *E. coli*, cuando la bacteria se somete a una agresión por altas temperaturas, produce la inducción de 17 genes de respuesta al calor, cuyos productos tienden a evitar los daños ocasionados por esta condición. La respuesta al daño por calor está mediada por la nueva subunidad σ^{32} que desplaza a la σ^{70} de la célula normal. La nueva holoenzima reconoce a los genes de la respuesta al calor, los cuales poseen un promotor con una región -10 (CCATNT) totalmente diferente a la caja TATAAT descrita como consenso. Otro ejemplo lo constituyen las enterobacterias que requieren utilizar otras fuentes de Nitrógeno en condiciones donde no exista el NH3, para este caso la subunidad σ^{54} desplaza a la σ^{70} de la RNApol para que la holoenzima reconozca promotores distintos correspondientes a operones cuyos productos permiten utilizar fuentes de N más raras. Existen otros casos como el factor de citrato férrico σ^{19} (FecI) que regula el transporte de hierro, el factor σ^{24} (RpoE) que participa en situaciones de estrés por calor, el factor sigma flagelar

σ^{28} (RpoF), el factor σ^{38} (RpoS) de fase estacionaria/inanición y el factor σ^{54} (RpoN), sin embargo el factor primario o "house keeping" que transcribe la mayoría de los genes esenciales es el σ^{70} (RpoD) (138), el cual pertenece a una familia de proteínas que interactúan con los promotores de los genes permitiendo la unión específica de la RNApol (1), y facilita la formación del complejo de cadena abierto que da inicio a la síntesis de RNA (138)

Por otro lado, para el caso de proteínas reguladoras que controlan el inicio de la transcripción (TFs), los desencadenantes de las respuestas de inducción/represión se llaman efectores (inductor/correpresor), suelen ser moléculas de tamaño pequeño que informan sobre el ambiente exterior, y su función depende de su interacción con los TFs específicos para cada sistema, las cuales interactúan con una zona de la región reguladora cercana al promotor.

2.2.2 Función de los TFs

La función de los TFs depende generalmente de la posición del sitio donde reconocen y se unen, de esta forma se ha descrito que pueden funcionar como:

1. Activadores; cuando su función es reconocer sitios específicos de DNA, adyacentes al promotor, con el fin de aumentar las posibilidades de interacción de la RNApol y dar inicio a la transcripción; este proceso se debe principalmente a un contacto directo entre el activador y la RNApol (17,18). Un ejemplo es el activador CRP (cAMP *receptor protein*) que activa la transcripción del operón *lac* en *E. coli* (140) al unir cAMP producida por ausencia de glucosa en el medio. También existen casos

indirectos donde la activación puede suceder por interrupción de la represión, permitiendo que el activador se una a secuencias que se ubican lejanas al promotor (18).

2. Represores; cuando su función es reconocer sitios específicos de DNA que traslapan o están muy cercanos a los promotores y, por lo tanto, al ser reconocidos por el TF bloquean el acceso de la RNApol para reconocer los promotores, impidiendo el inicio de la transcripción (19). Ejemplos clásicos de este tipo de reguladores son el represor TrpR, que regula negativamente la transcripción del operón de biosíntesis de triptófano (137). Otros ejemplos de reguladores con actividad represora son MetJ y LacI (141).

3. Actividad dual o bifuncional; cuando el TF puede funcionar como activador o represor (20). Un ejemplo de este tipo también lo representa CRP o CAP (140)

2.2.4 Descripción de elementos adicionales para el mecanismo de regulación en procariotas

Además de los TFs y sus correspondientes TFBSs otros elementos importantes para llevar a cabo el mecanismo mediado por TFs en organismos procariotas son: inicio de transcripción (TSS por sus siglas en inglés *Transcription Start Site*), los promotores, y las regiones de contacto con las subunidades de la RNApol.

El TSS es el primer nucleótido en ser transcrito y se localiza en la posición +1 de la secuencia de un gen. Esta coordenada marca la división entre la región codificante y la no

codificante. Las posiciones de los promotores y los TFBSs son definidas respecto a la ubicación del TSS.

Los promotores, son secuencias hexaméricas de nucleótidos (nt) bien definidos, ubicados en las posiciones -10 y -35 pb río arriba respecto al TSS del gen. Se representan por las secuencias consenso TATAAT y TTGACA respectivamente, sin embargo, a pesar de tener un consenso bien definido, éste no siempre se encuentra 100% conservado. Con frecuencia los mRNAs bacterianos se organizan en operones y se transcriben en unidades policistrónicas que sólo requieren un promotor único que regula la transcripción y expresión coordinada del grupo de genes. Otros dos elementos comúnmente presentes en la región de regulación son la región -10 extendida que contacta también a la RNApol y el elemento UP (UPstream del +1) que contacta a las subunidades α de la RNApol.

La RNApol es el elemento central en el mecanismo de transcripción, una holoenzima formada por cinco subunidades proteicas que utiliza la célula, para controlar la cantidad y los tiempos de aparición del mRNA que se genera como producto de un gen (21). Estructuralmente, la RNApol se compone de dos subunidades β (β,β'), una subunidad ω, y 2 subunidades α. Adicionalmente requiere de la subunidad σ para aumentar la especificidad de la RNApol en el reconocimiento de los promotores (**Figura 1**) (22). Cada subunidad α a su vez, 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 (17). Por otro lado, la subunidad β contiene parte del centro activo y la subunidad β' participa en la unión del DNA, mientras la subunidad ω actúa como una chaperona para ayudar al correcto plegamiento de la subunidad β'. Finalmente, la subunidad σ está formada por cuatro

dominios que reconocen las cajas -10, -35 y una vez iniciada la transcripción se disocia del resto de la enzima (21,22).

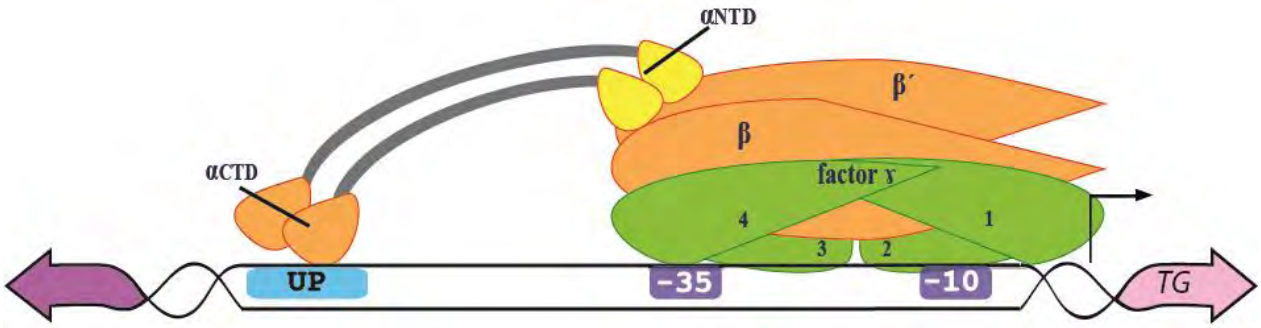

Figura 1 Maquinaria de transcripción en bacterias. En verde se representa el factor σ haciendo contacto con los promotores mediante sus dominios 2 y 4, las subunidades α a su vez hacen contacto con el elemento UP.

2.3 Identificación de Sitios de Unión a Factores Transcripcionales (TFBS)

Para los diferentes estudios de Biología Molecular, ha sido muy relevante lograr identificar los TFBSs en posiciones río arriba de los TGs (genes blanco o "Target Gene"), debido a que caracterizarlos permite determinar en gran medida la respuesta transcripcional del organismo frente a un estímulo específico. Sin embargo, identificar en conjunto los diferentes TFBSs en un genoma, es fundamental para estudios de genómica funcional, ya que permitirían establecer las complejas redes de regulación genética.

Dada la importancia de la identificación de los TFBSs, desde hace más de cuatro décadas se han desarrollado diferentes aproximaciones metodológicas, tanto experimentales como *in silico*.

2.3.1 Estado del arte en la identificación de TFBS

Los enfoques computacionales actuales para la identificación de TFBSs se clasifican en dos categorías generales: la identificación de secuencias con motivos conocidos "*known*" (23,24) y el descubrimiento o identificación de nuevos motivos "*unknown*" (25,26). La primera categoría tiene disponibles, de antemano, algunos TFBSs que ya han sido anotados para el TF de interés, y se "conocen" a *priori*, mientras para la categoría de descubrir motivos de *novo*, únicamente se cuenta con un conjunto de secuencias de DNA en las cuales se cree que pueden encontrarse los TFBSs de un TF determinado. El protocolo que presentamos en esta tesis, nos ha permitido identificar ambos tipos de motivos e inclusive, hacer planteamientos sobre el modelo dinámico del mecanismo de regulación, considerando todos los elementos disponibles de cada sistema.

La mayoría de los enfoques sobre el descubrimiento de motivos se basan en determinar la sobre-representación de las secuencias de interés. En ellos se realiza una búsqueda exhaustiva a través del muestreo de todas las posibles "palabras de DNA" de cierta longitud. Posteriormente, se evalúan esas palabras basándose en calificaciones (*scores*) de significancia estadística asociados a cada palabra. Tales *scores* indican la sobre-representación de dichas palabras en secuencias que son potencialmente reconocidas por el TF de interés. Otros métodos evitan la búsqueda exhaustiva, aplicando muestreos (27,28). Mediante el uso de cualquiera de las técnicas antes mencionadas, el espacio de búsqueda se puede reducir únicamente a aquellas regiones que son más susceptibles de contener TFBSs.

Adicionalmente, gracias a los avances biotecnológicos en los laboratorios de todo el mundo, se ha derivado una creciente cantidad de datos biológicos que aún faltan por completarse, no obstante, trabajar con conjuntos de datos de secuencias intergénicas de genes coregulados en los cuales ya han sido anotados algunos TFBSs, promotores o TSS es muy frecuente. El uso de información existente sobre las regiones de regulación estudiadas nos permite mejorar la identificación de nuevos TFBSs.

Los motivos de TFBSs anotados en la literatura representados por matrices PWMs (29,30), nos presentan información sobre la probabilidad, por nucleótido, de ocurrir en cada posición de los sitios de unión alineados. TRANSFAC (31) y JASPAR (32) son dos bases de datos públicas que almacenan PWMs verificadas experimentalmente para una variedad de TFs. Muchos enfoques en la literatura (33,34) han utilizado las PWMs para intentar resolver el problema. Dada la matriz PWM de un TF, se calcula un "*matching score*" entre la PWM y la secuencia de DNA de la misma longitud representada por la matriz (subsecuencia de DNA de cierta longitud). Si la puntuación es mayor a la especificada por el usuario, la subsecuencia de DNA se considera como un nuevo motivo. Sin embargo, la pequeña cantidad de sitios de unión conocidos en ambas bases de datos provocan un sobreajuste de datos de PWMs, lo que trae como resultado que tales métodos generen gran cantidad de falsos positivos.

Cuando se trata de motivos degenerados o sitios de menor afinidad "débiles" que no se encuentran sobre-representados respecto al ruido de fondo, la identificación de TFBSs con estas metodologías es todavía más complicada. Algunos métodos actuales (35,36) han utilizado información adicional para hacer frente a la identificación de motivos degenerados, haciendo uso de características variables de las secuencias de

DNA, las cuales reflejan las propiedades genómicas, celulares y fisiológicas, tales como las características basadas en la estructura del DNA, temperatura de congelación, contenido de GC, *etc*.

La identificación computacional actual de TFBSs se centra en utilizar algún algoritmo específico y asignar un valor (*score*) de salida a una secuencia de entrada dada, pero adicionalmente existen los algoritmos de clasificación utilizados en la estrategia aprendizaje automático conocida como "*machine learning"*. En esta metodología se requieren datos de entrenamiento tanto positivos como negativos. Los datos de entrenamiento positivos son el conjunto de sitios de unión conocidos, mientras que los negativos son las secuencias de DNA que no son reconocidas por el TF de interés. Tanto los ejemplos positivos como los negativos, constituyen los datos de entrenamiento para los algoritmos de clasificación que derivan en modelos para la predicción de sitios de unión.

2.3.2 El problema de identificación de TFBS

En el presente apartado se discuten las razones más importantes por las cuales la identificación de TFBSs se considera un desafío. 1) Para considerar que la identificación de un TFBS representa un reto, el factor más significativo es la degeneración de su secuencia, la cual se le atribuye a la naturaleza biológica de la unión de los TFs sobre sus sitios de reconocimiento. Como ejemplo hablemos del caso de los sitios de unión de una proteína hipotética X para la cual se han descrito 4 secuencias nucleotídicas (Figura 2). Evidentemente al comparar entre sí las secuencias, se pueden apreciar las diferencias y similitudes entre los 4 sitios de unión, sin embargo, es posible

que la causa de esta degeneración se deba a que la afinidad requerida no siempre es la máxima, razón por la cual no todos los nucleótidos son igualmente importantes para el reconocimiento. Lo que aún no logramos predecir es ¿Cuáles serán todos los posibles sitios blanco a partir de la secuencia o estructura de un TF?. Para el caso de organismos ortólogos la variabilidad de los sitios se debe también a que los TFs ortólogos no son 100% idénticos en su secuencia y esos cambios hacen que sus secuencias blanco también varíen.

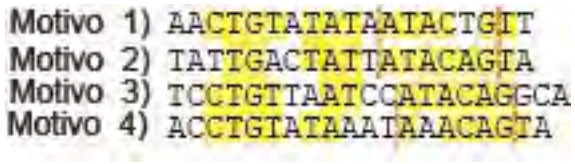

Figura 2 Alineamiento de cuatro motivos de unión que forman parte del regulón de una misma proteína X. Se resaltan en amarillo los nucleótidos conservados y se marca con dos verticales punteadas la representación de un posible desplazamiento de la región conservada del motivo.

2) Otro factor relevante es que los TFBSs son secuencias de tamaños relativamente cortas, con longitudes que van de 7 a 20 nt, por lo que secuencias similares no relacionadas pueden encontrarse en cualquier lugar del cromosoma, incluyendo en las regiones 5´ UTRs de los genes. El hecho de que los TFBSs sean secuencias cortas y degeneradas hace que los enfoques computacionales actuales sean vulnerables a reportar falsos positivos. Un falso positivo corresponde a secuencias que son similares a un TFBS, encontradas por azar en las regiones reguladoras, pero sin ser biológicamente funcionales. 3) El tercer factor, es la posibilidad de que existan múltiples sitios de unión para un solo TF en la misma región de regulación, o bien que un gen esté regulado por

más de un TF y en consecuencia existan diferentes TFBSs en una misma región de regulación. 4) Otro factor que constituye un impedimento para la identificación de los TFBSs en regiones de regulación, de genes que se asumen estar coregulados, es la poca confiabilidad de los datos de expresión producidos por las tecnologías de alto rendimiento ("*high throughput*"), ya que existe la posibilidad de que los datos contengan ruido debido a la variabilidad de los experimentos biológicos o bien, a que los algoritmos de agrupación o "*clustering"* utilizados para obtener conjuntos de genes coregulados no sean perfectos en términos de precisión.

Además de la degeneración, los TFBSs asociados con un TF, hasta cierto punto también están conservados en algunas posiciones a lo largo de la evolución de las especies. Si uno alinea los TFBSs, normalmente puede observarse un patrón o motivo entre los TFBSs alineados. Sin embargo, la identificación de TFBSs no puede considerarse como un asunto de reconocimiento de patrones tradicional, dado que identificar un patrón o "*match*" de un TFBS, no implicará necesariamente que corresponde a un verdadero sitio de unión. Regularmente un motivo se puede obtener fácilmente mediante alineamiento de TFBSs, sin embargo, en este trabajo proponemos que la identificación de patrones maneja otras métricas que hasta ahora no habían sido descritas, como la distancia del TFBS dada por vueltas de DNA respecto al TSS, la presencia de secuencias ricas en A/T en medio de los TFBS, para dar flexibilidad a los sitios, la posición central del TFBS respecto a las cajas de los promotores o al TSS, *etc*. En esta tesis nos referiremos a un "motivo" como una representación de un conjunto de sitios de unión de TFs.

2.3.3 Metodologías experimentales para identificación de TFBS

Identificar la localización exacta de los TFBSs y sus especificidades de unión resulta un paso importante para avanzar en la comprensión de los mecanismos de expresión de los genes. Dentro de las metodologías experimentales comúnmente empleadas, se han utilizado los ensayos *in vitro* y los ensayos *in vivo* que se describirán brevemente a continuación.

Dentro de los ensayos *in vitro*, para estudiar las interacciones de la proteína unida a DNA se ha usado la técnica de retardo en gel EMSA (*electrophoretic mobility shift assay*). El fundamento de esta técnica se basa en que los complejos proteína-DNA se moverán más lentamente a través de un gel de electroforesis, respecto a la sonda radiactiva de DNA que no tiene la proteína unida, y para observar la migración, las bandas radiactivas se pueden visualizar discretamente en geles de acrilamida. Sin embargo, este ensayo no es lo suficientemente sensible para localizar con exactitud la posición del sitio de contacto.

El ensayo que puede complementar la movilidad electroforética es la técnica que permite observar e imprimir la huella de la posición del sitio de unión, la técnica de *footprinting* con DNAsa I permite estudiar interacciones TF-DNA e identificar la secuencia de unión en el DNA a la cual se une el TF. El fundamento de esta técnica se basa en que la molécula de DNA que lleva el TF unido estará protegida de cualquier modificación o degradación. De esta forma, al tratar el complejo TF-DNA con una nucleasa (DNasa I), para romper todos los enlaces fosfodiéster, se protegerán solamente aquellos nt que mantienen a la proteína (TF) unida. Posteriormente se remueve la proteína del complejo y el DNA se resuelve en geles de poliacrilamida. El área en blanco que se observe en el

gel corresponderá a la huella de la posición del sitio de unión (donde estaba unido el TF) y a partir de ahí se obtienen las secuencias consenso. 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". Otras técnicas utilizadas *in vitro* son los arreglos de unión a DNA (37).

Dentro de los ensayos *in vivo*, los TFBS pueden identificarse mediante técnicas tales como los ensayos de ChIP-chip (38) o arreglos de expresión de genes (39,40) donde es posible reconocer los TFBS con alta precisión. La tecnología ChIP-chip combina la inmunoprecipitación de la cromatina (***Ch****romatin* ***I****mmuno****p****recipitation*) con microarreglos de DNA (***chip***) y permite identificar las interacciones de las proteínas con el DNA *in vivo* en genomas completos, dando como resultado el conocido "cistroma" o suma de los sitios de unión en el genoma, lo cual es relevante para identificar sus elementos funcionales. Con los avances en estas últimas tecnologías y el surgimiento de las técnicas de secuenciación masiva, se han determinado las secuencias de miles de genomas de los tres reinos de la vida, Eubacteria, Arqueobacteria y Eucariota, por lo que este gran volumen de información ha hecho necesaria la identificación masiva de TFBSs mediante análisis comparativo de secuencias, a través del desarrollo de algoritmos bioinformáticos.

El principio fundamental para el reconocimiento computacional de TFBS a partir de estos datos, es que los genes coexpresados pueden organizarse en un conjunto o "*cluster"* y a su vez, la coexpresión en cada *cluster* puede ser considerada como un resultado de coregulación transcripcional (41). De esta forma, se asume que los genes coexpresados deben ser regulados por un TF común y uno puede buscar los sitios de unión en sus regiones promotoras. Hasta este punto el problema de identificación de

TFBS puede describirse como la simple identificación de TFBSs “escondidos” en un conjunto de secuencias reguladoras de genes coexpresados. Para este problema se han desarrollado varios programas y herramientas, sin embargo, distan mucho de ser perfectos y el problema sigue sin resolverse (42).

2.3.4 Metodologías computacionales para identificación de TFBS

Además de las aproximaciones experimentales, los enfoques computacionales han recibido un creciente interés y han demostrado un buen potencial para resolver el problema, sobre todo por la posibilidad de ahorrar tiempo y costos en el diseño de experimentos.

Con los enfoques actuales de identificación de motivos, al buscar patrones sobrerepresentados se tiende a discriminar los segmentos degenerados de los TFBS, llegando a aproximaciones poco satisfactorias o bien, a la incapacidad de identificar elementos de regulación adicionales por separado como los promotores y sitios de inicio, ya que su identificación también se basa en la conservación de la secuencia y suelen reportarse erróneamente. En la sección de apéndices se provee una lista de las bases de datos y herramientas computacionales más utilizados para detección de TFBS (**Apéndice 1 y 2**).

2.3.5 Enfoque canónico de identificación de huellas filogenéticas

Como se ha mencionado, una de las características más ampliamente utilizadas para detectar TFBSs es la conservación filogenética. El sustento de este enfoque es que

las regiones funcionales de los genes (tales como TFBSs) están propensas a una presión que hace que las mutaciones a las que puedan ser sometidas se acumulen más lentamente que el resto de la secuencia (segmentos no funcionales) (43). Es posible distinguir a los TFBSs de la secuencia de fondo (*background*) mediante el alineamiento de regiones de regulación ortólogas. Los genes ortólogos son aquellos que comparten un gen ancestral común, que han evolucionado y después de un evento de especiación se encuentran en organismos distintos. Normalmente, esos genes retienen la misma función que la del gen del cual descienden. Para poder realizar el alineamiento, las secuencias ortólogas deben ser de especies disímiles pero relacionados, es decir organismos que tengan una distancia filogenética no mayor de 50 millones de años respecto a su ancestro común más reciente (36,142). Una vez alineadas las regiones intergénicas, probablemente una gran porción corresponda a secuencia no funcional y se descarte, dejando los segmentos potenciales para una posterior identificación. Los enfoques basados en conservación filogenética, por lo tanto, requieren algoritmos de alineamiento de alta precisión (44).

La premisa del enfoque de "*Huellas filogenéticas"* es que las secuencias discretas (motivos) que pueden encontrarse en las regiones río arriba de genes que son coregulados por el mismo TF en un genoma, o bien en la región reguladora de un conjunto de genes ortólogos, de organismos filogenéticamente cercanos, tienden a presentar una mayor similitud entre sus secuencias, debido a la restricción que les impone la interacción específica TF-DNA.

El método de *Huellas filogenéticas* se reportó por primera vez en 1988, por Tagle *et al.* en un trabajo donde identificaban los elementos de regulación codificados en *cis,* responsables de la expresión de los genes de las globinas embrionarias *epsilon* y *gama*

en primates (45). Posterior a ello, se han realizado diversas implementaciones del método y a grandes rasgos se pueden dividir en tres: determinista, estocástico y enumerativo (46). Como ejemplos del método determinista se encuentran MEME *(Multiple Expectation Maximization Estimation*) (46-48) basado en algoritmos que maximizan la expectativa de encontrar motivos sobre-representados en un grupo de secuencias de estudio (48) o aquellos basados en el contenido informacional del conjunto de posibles alineamientos de las secuencias de estudio (49). Por otro lado, el método de Gibbs-sampler (50) y su implementación PhyloGibbs (27) para el análisis comparativo en varios genomas, son ejemplos de aplicaciones del método estocástico que utilizan técnicas Monte Carlo basadas en cadenas de Markov y estadística bayesiana, éstas utilizan valores de las muestras anteriores para generar aleatoriamente la muestra siguiente en donde se espera que se encuentren los motivos mayormente representados. Dado que la búsqueda se realiza de manera aleatoria, los resultados obtenidos por los métodos estocásticos no siempre son los mismos. Finalmente, los métodos enumerativos permiten representar los sitios de unión identificados por los métodos deterministas y estocásticos mediante expresiones regulares y matrices de frecuencias específicas de las bases que constituyen los motivos. Sin embargo, con estas metodologías el porcentaje de falsos positivos obtenidos continúa siendo elevado.

A pesar de las diferentes aproximaciones metodológicas de los algoritmos antes mencionados, todos tienen como objetivo identificar aquellos motivos de secuencia mayormente representados asumiendo que éstos corresponden o son potenciales TFBSs. Dado lo anterior, un motivo es comúnmente considerado como un verdadero TFBS, cuando su valor estadístico p-valor (*P*-value), asociado a la probabilidad de encontrarlo sobre-representado en un conjunto de secuencias determinadas, es

significativamente pequeño. Como consecuencia de lo anterior, los motivos de secuencias con valores de *p*-value poco significativos nunca son considerados como verdaderos TFBS o no son considerados, a pesar de que pudieran ser relevantes en la regulación de los genes de estudio. Estos TFBSs de baja conservación pudieran corresponder a los sitios de baja afinidad, y suelen ser reconocidos por los TFs en sistemas de regulación en donde los TFs realizan mecanismos de unión cooperativa. La cooperatividad de unión sucede cuando el TF reconoce un primer TFBS de mayor afinidad (muy conservado) que promueve la unión del TF sobre un segundo sitio de menor afinidad (menos conservado o sitio “degenerado”). El sitio degenerado únicamente es reconocido después de que el primer sitio ha sido unido. Las diferencias entre los conceptos de P-valor y E-valor según el algoritmo utilizado en este análisis, es que MEME reporta un estimado de la significancia estadística de cada motivo que encuentra (E-valor del motivo) así como la probabilidad de qué tan buen balance (*match*) hace cada ocurrencia con el motivo (P-valor del sitio). El E-valor del motivo es un estimado muy conservador sobre la probabilidad de que el motivo encontrado no sea sólo un artefacto estadístico. El caso del P-valor no es conservativo y se usa como un indicador relativo de lo bien que cada sitio coincide con el motivo, pero no puede usarse para decir si el motivo es real o no. Convencionalmente, en un análisis de huellas filogenéticas se considera como "motivo" al resultado que posea el E-valor más pequeño (el más significativo).

2.4 Los motivos de DNA y su representación

Para iniciar la transcripción, es necesario que los TFs reconozcan segmentos de DNA conocidos como sitios de unión, que son secuencias de longitud corta (7-20 nt),

frecuentemente degeneradas en su composición nucleotídica. Las variaciones de los TFBS son toleradas por el TF que se une a ellos distinguiendo entre los de mayor a menor afinidad (más conservados o muy degenerados), lo que determina la modulación de los mecanismos de transcripción celular. A pesar de la variabilidad de los sitios, al mismo tiempo sucede que entre los TFBS de un TF debe existir un conjunto común de nt lo suficientemente conservado, para que el TF pueda reconocerlos. Esta ocurrencia de nt nos permite “dibujar” un patrón común, o “motivo”, a partir de un conjunto de sitios de unión de un TF. En pocas palabras, un motivo es otra forma de referirnos a un conjunto de sitios de unión.

Para representar la caracterización de los TFBS, tanto de variación como de conservación, computacionalmente se han desarrollado formas para presentar los motivos, ya sea por medio de secuencias consenso, matrices de posición-peso (PWM) o secuencias llamadas “logos” (inciso 4 de la **Figura 3**). A continuación, se discute detalladamente cada una de ellas.

2.4.1 Secuencia consenso

Las secuencias consenso son patrones construidos por un alfabeto de 15 letras, las cuales consisten en cuatro bases nucleotídicas (A,C,T,G) y once letras específicas que hacen referencia a once posibles subconjuntos de cuatro nucleótidos, que son: R=[AG], Y=[CT], W=[AT], S=[GC], M=[AC], K=[GT], B=[CGT], D=[AGT], H=[ACT], V=[ACG], N=[ACGT]. Dado lo anterior, si suponemos una secuencia consenso YRCAMC, esta cadena de caracteres representa el motivo que captura todos los sitios que empiezan con C o T seguidos de A o G, CA, A o C y terminados con C.

La representación de una secuencia consenso es simple, y la búsqueda de secuencias genómicas que representen nuevos motivos con base en la secuencia consenso también es fácil y sencilla, sin embargo, cuando se trata de buscar nuevos sitios, las secuencias consenso no son muy confiables (51,52), en parte porque las secuencias consenso ignoran la frecuencia relativa de cada substitución en una posición dada. Es posible que muchos sitios funcionales se pierdan, y muchos otros sitios no funcionales se recuperen como falsos positivos (29).

Tampoco queda claro cómo construir una secuencia consenso de un conjunto de sitios de unión, cuando en alguna posición es necesario usar las letras específicas de dos o tres sustituciones de base, lo cual se vuelve un poco arbitrario. Para superar esta deficiencia se han indicado las matrices PWM como un método alternativo.

2.4.2 Position Weight Matrix (PWM)

La representación de un sitio con una matriz de peso por posición “PWM”, mejora el poder descriptivo de las secuencias consenso, ya que la matriz nos ofrece las probabilidades de peso (“*weight*”) para cada uno de los cuatro nucleótidos, de aparecer en una posición dada del motivo. A una matriz PWM también se le conoce como Position Specific Weight Matrix (PSWM), y consta filas que se asocian a las cuatro posibles bases (A,C,G,T) y columnas que representan posiciones consecutivas del motivo. En caso de proteínas, las filas se asocian a los 20 símbolos del alfabeto de aminoácidos. El primer paso para construir una PWM es crear una matriz de frecuencias contando las ocurrencias de aparición de cada nucleótido por posición. A partir de esta matriz de frecuencias inicial se puede crear la matriz de probabilidad, dividiendo el número anterior de nucleótidos en cada posición por el número de secuencias, normalizando así los

valores. Las ventajas que poseen las PWMs sobre las secuencias consenso las han hecho un método popular para representar patrones de secuencias biológicas y son un componente principal en algoritmos modernos para descubrimiento de motivos.

2.4.3 Representación en logos

Además de las secuencias consenso y las PWMs, los motivos pueden ser representados gráficamente por logos (53), estas imágenes muestran la predominancia de las substituciones de nt en cada posición del motivo, y se representan por un apilamiento de nt (A,T,G o C) que varían de tamaño en relación a su frecuencia de aparición por posición (**Figura 3**). Para cada posición del motivo, las substituciones se apilan en orden descendente, de modo que en la parte superior se encuentra aquella con una frecuencia relativa mayor. La altura total de todos los nt apilados es proporcional al contenido de información de esa posición.

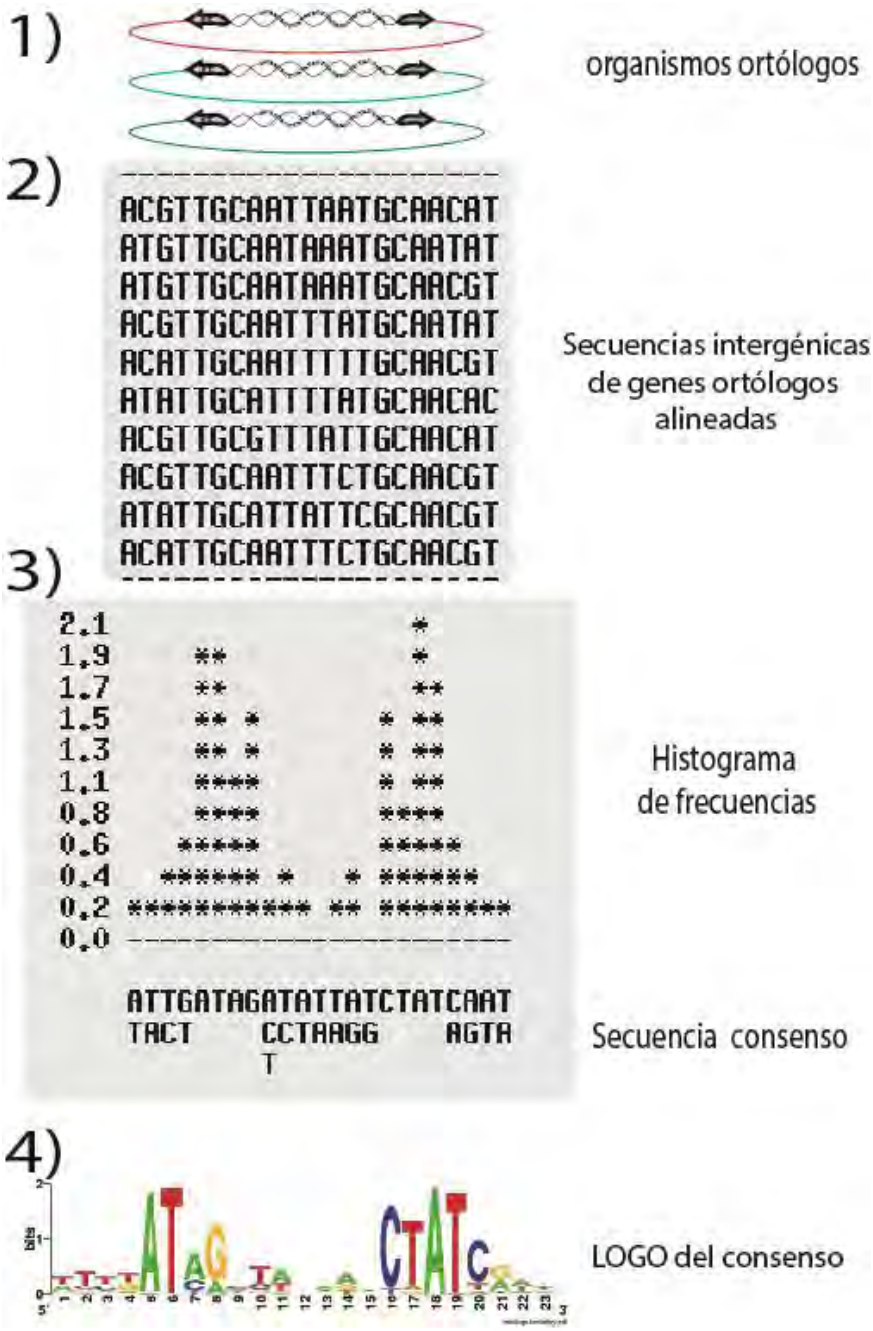

Figura 3 Diversas representaciones de un motivo. 1) conjunto de secuencias ortólogas, 2) alineamiento de secuencias intergénicas, 3) histograma de frecuencia y generación de la secuencia consenso, 4) representación gráfica de un logo del motivo de unión de IlvY creada por el software weblogo (132)

2.5 Propiedades biológicas complementarias en la identificación de TFBS

Identificar los TFBSs *bona fide* en un conjunto de motivos, provee una herramienta muy importante a los biólogos para descifrar la compleja red de regulación de la célula. Aunque la mayoría de los enfoques existentes para el reconocimiento de motivos explota únicamente el contenido nucleotídico de una secuencia de DNA, existen también evidencias biológicas provenientes de experimentos de laboratorio, dirigidas a caracterizar la unión del TF a sus sitios de reconocimiento. En este trabajo demostramos que el proceso de identificación de TFBSs también involucra otros factores, tales como la distancia de los TFBSs respecto al sitio de inicio de la transcripción del gen codificante, relativa a la posición de los promotores, así como propiedades conformacionales de las secuencias de DNA, estructura tridimensional activa del TF, participación de metabolitos coinductores, características específicas de los TFBSs como son: la simetría de sus secuencias, presencia de nt A/T intermedias al sitio, posiciones centrales determinadas de los TFBSs respecto a los promotores, *etc*. Integrar esta diversidad de información mejora en gran medida el rendimiento de la búsqueda de motivos. Dados los diferentes requerimientos de interacción específica que tienen los TFs con el DNA, proteínas y cofactores, éstos comparten ciertas características estructurales que son importantes de considerar en el análisis integrativo de los sistemas que hemos analizado. A continuación, se mencionarán brevemente las más relevantes.

2.5.1 Motivo proteico de unión al DNA

Las interacciones de las proteínas con el DNA son de gran importancia para llevar a cabo los mecanismos básicos de la célula, como la regulación de la expresión génica, la división celular y la diferenciación. La información codificada en el DNA se organiza, se replica y se lee por una variedad de proteínas que se unen al DNA, de las cuales algunas reconocen determinadas secuencias de DNA con poca o nula afinidad (*i.e.* histonas o DNA polimerasas), mientras que otras como los represores, activadores transcripcionales y endonucleasas de restricción poseen una especificidad muy alta para secuencias de DNA y son capaces de distinguir secuencias pequeñas de 10 a 20 nt de longitud, en un fondo de 10^6 o más pb.

Las bases moleculares del reconocimiento de un sitio específico de unión a DNA requieren de la caracterización de las conformaciones tridimensionales de las proteínas, del complejo proteína-DNA, de la secuencia del TFBS en el DNA, y de los cambios estructurales que suceden como consecuencia de la interacción.

Las proteínas que se unen a DNA pueden clasificarse estructuralmente según el motivo de unión al DNA en las siguientes grandes "familias": 1.- Hélice-Vuelta-Hélice (*helix-turn-helix* o HTH), 2.- Dedos de Zinc (*Zinc fingers*), 3.- Cierre de Leucina (leucine zipper o ZIP) y 4.-Hojas Beta (beta sheets). En procariotes, la estructura del motivo de unión que más se ha caracterizado en los dominios de proteínas de unión a DNA es la HTH. Este motivo consta de dos hélices alfa unidas por una secuencia corta en forma de asa, ejemplos de estas proteínas son lambda cro, CAP y la proteína represora lambda de *E coli.* La arquitectura HTH presente en los TFs hace que los TFBSs como

consecuencia, presenten a su vez simetría en su secuencia, ya sea como Invertidos Repetidos (IR), Directos Repetidos (DR), Reversos Repetidos Directos (ER, por sus siglas en inglés*, Everted Repeated*) y Asimétricos (A).

2.5.2 Estructura funcional de los TFs

La estructura cuaternaria de una proteína es la característica que generalmente le confiere la función, sin embargo, también la unión a un sustrato (o ligando) puede influir para cambiar la función de la proteína al activarla o desactivarla. Los TFs se caracterizan por la diversidad de su estructura cuaternaria, la cual puede comprender : *i*) TFs monoméricos, como algunos miembros de la familia AraC/XylS (54), MalT (55), PutA (56) o el activador transcripcional MotA del bacteriófago T4 (57); *ii*) TFs homodiméricos, como los reguladores lambda y CRO del bacteriófago Lambda (58), los reguladores bacterianos CRP (59), MarR (60), NtrC (61), FadR (62), los represores TrpR (63) y LacI (64), o miembros de la familia LysR (65); y *iii*) TFs heterodiméricos: como por ejemplo dímeros de la nucleoproteína H-NS con otras nucleoproteínas, Hha o StpA (66), o el caso del regulador flagelar maestro FlhD4C2 (66) y otros reguladores como: RcsB- GadE (67), RcsB-BglJ (68). Las proteínas que se activan o desactivan mediante la unión por ligando consisten mayoritariamente en conjuntos simétricos de subunidades idénticas. Con esta disposición, la union de una molécula de ligando a un sitio único en una subunidad puede desencadenar cambios conformacionales alostéricos que pueden transmitirse a las subunidades vecinas ayudándolas a unirse al mismo ligando. Como resultado, la transición produce cambios en la estructura de la proteína que pueden activarla o inactivarla. Para las proteínas de la familia de reguladores transcripcionales LysR se ha

descrito que la formación de un dímero de dímeros es la estructura functional activa para actuar como un activador de la transcripción en presencia de un metabolito inductor (143).

2.5.3 Propiedades generales de los TFBSs y la relación con sus TFs

A partir del análisis de múltiples TFBSs caracterizados experimentalmente, actualmente se sabe que los TFBSs bacterianos suelen ser pequeñas regiones de DNA, con longitudes de 7 a 20 pb. Su ubicación con respecto al promotor define la función que tendrán sus correspondientes TFs sobre la transcripción de los genes regulados (activación/represión). Considerando la información de los TFBSs de *E. coli,* depositada en la base de datos de RegulonDB, se puede cuantificar que el 22% de los TFs se unen como monómeros a regiones asimétricas; el 65% de los TFs pueden unirse como homodímeros a regiones simétricas DR; y sólo el 8%, de los TFs se unen como homodímeros a regiones simétricas IR; por último, el 5% de los TFs de *E. coli* se unen como heterodímeros a regiones asimétricas (69). Además de la tendencia de los reguladores a formar mayoritariamente homodímeros, la conservación de simetrías en sus TFBSs (IR, DR, ER y A) se relaciona directamente. Se cree que ambas características forman parte de una estrategia de regulación que consiste en permitir que la proteína reguladora amplíe su zona de contacto con el DNA e incremente su especificidad de unión. Adicionalmente se sabe que las posiciones de los TFBSs están estrechamente ligadas con las distancias respecto a las subunidades de reconocimiento de la RNApol dentro de la región de regulación, de tal manera que se favorezca el contacto entre ellas. Lo anterior determina la distancia entre los TFBSs respecto a la RNApol, con distancias de longitudes en múltiplos de 10 pb, equivalentes en promedio a

una vuelta de la hélice del DNA. En este sentido existe una clasificación de los promotores regulados por CRP que toma en cuenta la distancia a la que se une con respecto al TSS (+1) (70). Esta clasificación divide a los promotores en tres: a) promotores Clase I, cuando el sitio de unión de CRP se encuentra ubicado a distancias de -60.5, -70.5, -80.5 y -90.5 pb respecto al TSS; b) promotores Clase II, cuando el sitio de unión de CRP, ubicado a -41.5 pb, traslapa la caja -35 del promotor (71,72); y c) promotores Clase III, que representan la unión cooperativa de dos dímeros de CRP a diferentes distancias (73). Dada la naturaleza de los TFs que intervienen como activadores de la transcripción, se ha visto que sus posiciones preferenciales, respecto al TSS (+1), pueden localizarse a -60.5, -70.5, -80.5 y -90.5 pb (**Figura 4**). La mayoría de los activadores transcripcionales se unen a una distancia de -41.5 pb respecto al inicio de la transcripción, quedando sobrepuesta la caja -35 de su promotor. Debido a lo anterior, es común encontrar que la mayoría de los algoritmos de identificación de motivos de regulación, predicen erróneamente que el sitio de unión del TF y la caja -35 corresponde a un único motivo de longitud muy grande.

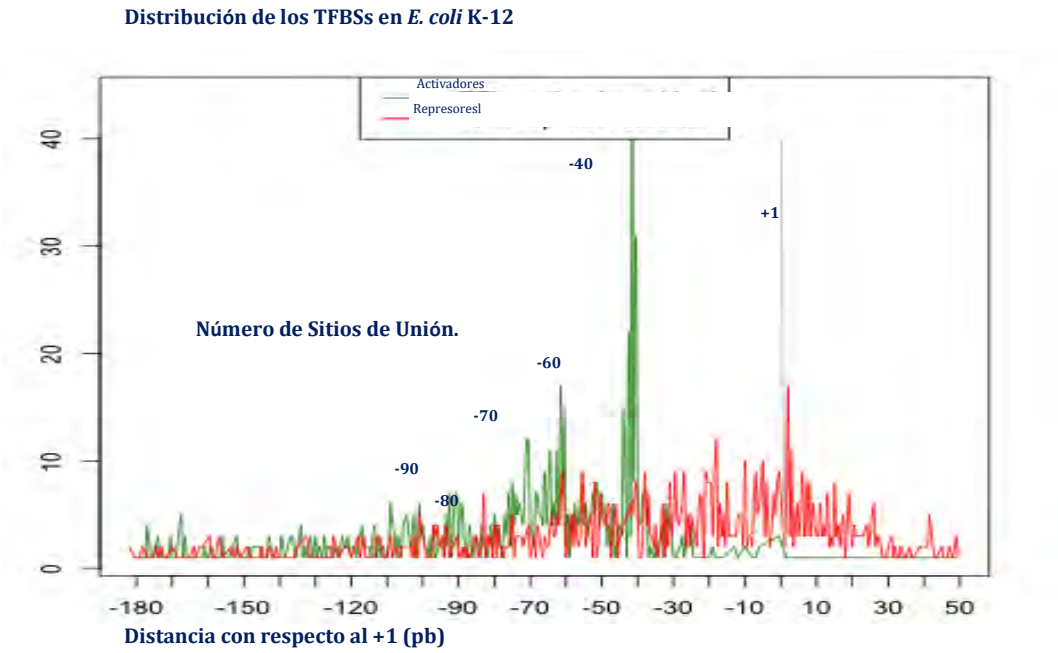

Figura 4 Posiciones preferenciales de unión de los TFs sobre sus TFBSs, en relación al TSS. Los activadores (verde) se unen preferencialmente en posiciones a -40 y -60, mientras que la frecuencia de TFBSs cuya actividad es negativa (rojo) tiene una cobertura más amplia que abarca regiones sobre las cajas promotoras -10 y -35.

2.5.4 Unión cooperativa de ciertos TFs

Una de las propiedades fundamentales en los sistemas de regulación basados en TFs, es su capacidad de unión cooperativa al DNA. Durante este proceso, los sitios de baja afinidad, con secuencias poco conservadas, son reconocidos por un TF únicamente cuando otro TF se ha unido previamente a un sitio contiguo de alta afinidad (con secuencia más conservada). Los sitios de mayor afinidad, son los primeros en ser ocupados, posteriormente, al incrementar las concentraciones celulares de los TFs, se

ocupan también los sitios de menor afinidad. Este tipo de comportamientos se presentan en TFs de *E. coli*, tales como: OmpR, NtrC, ArcA, entre otros. Cabe resaltar que la unión del segundo TF sucede exclusivamente cuando el primer sitio ha sido ocupado. Otro ejemplo relativo se puede observar en el fino mecanismo de regulación de los reguladores maestros CI y Cro para establecer los procesos de lisis o lisogenia del fago lambda (58).

Los TFs que se unen de manera cooperativa suelen ser homodímeros que comúnmente requieren de un coinductor para favorecer su interacción. Esta cinética de unión cooperativa resulta muy relevante en la dinámica de la respuesta transcripcional de los genes regulados, sin embargo, es común observar que los sitios de baja afinidad suelen no ser identificados por los métodos convencionales de identificación de patrones sobrerepresentados, debido precisamente a su baja conservación. En resumen, la mayoría de los activadores transcripcionales dependen de inductores y de concentraciones del regulador que permitan la formación de dímeros que, a su vez, favorecen la unión a TFBSs poco conservados. Hasta donde sabemos, ningún programa o algoritmo computacional para la identificación de TFBSs ha considerado la naturaleza de la unión cooperativa de los TFs.

2.6 Reguladores transcripcionales de la familia LysR

La familia LysR representa el tipo más abundante de reguladores transcripcionales en los organismos bacterianos ya que cuenta con más de 50 elementos. Los miembros de esta familia se caracterizan por tener una estructura conservada que incluye a un motivo de unión al DNA del HTH en su extremo N-terminal, un dominio central implicado

en la unión de coinductores y un dominio en su extremo C-terminal requerido tanto para la unión al DNA como de respuesta al coinductor (74).

A pesar de la alta conservación tanto estructural como funcional entre los miembros de la familia LysR, los TG que regulan participan en funciones muy diversas tales como la virulencia, el metabolismo, la detección de la densidad celular (*quorum sensing*), motilidad, fijación de nitrógeno, respuestas al estrés oxidativo, producción de toxinas, y secreción, entre otros (75).

Sobre su actividad reguladora, se sabe que pueden actuar como activadores o represores de la transcripción, o inclusive tener una función dual dependiendo de la posición donde se localizan sus TFBSs respecto a los promotores de sus TG.

Adicionalmente, los miembros de esta familia en su mayoría, pueden unir metabolitos que funcionan como coinductores, lo cual modifica sustancialmente su actividad reguladora, siendo común que dicha modificación resulta relevante para la formación de un circuito de autorregulación (75).

En términos generales, los TFBS de los reguladores de esta familia están muy poco conservados, por lo que identificarlos mediante aproximaciones computacionales, ha sido una estrategia poco exitosa y consideramos que esto ha representado el principal problema en el estudio d

3 HIPÓTESIS

La identificación de TFBSs en secuencias de genomas bacterianos puede realizarse de manera precisa, si además de los valores estadísticos de la sobre-representación de motivos calculados mediante algoritmos computacionales, se consideran también las propiedades biológicas relevantes del complejo de interacción TF-DNA, tales como: a) la capacidad de unión cooperativa de los TFs, y b) la naturaleza multimérica de los TFs, lo cual origina diferentes simetrías en las secuencias de sus correspondientes TFBSs (IR, DR, ER, A). Adicionalmente, si se tomara en cuenta la conservación relativa de los TFBSs entre sí y en relación a la posición de los promotores adyacentes, es posible proponer modelos de regulación que consideren la dinámica de unión de los TFs a sus correspondientes TFBSs.

4 JUSTIFICACIÓN

La regulación de la expresión genética en bacterias ocurre principalmente a nivel del inicio de la transcripción y es mediada primordialmente por TFs. La identificación de los TFBSs que éstos reconocen es muy importante para diversos tipos de estudios de Biología Molecular, así como para la reconstrucción de redes de regulación metabólicas. Con este fin, se han desarrollado diferentes algoritmos computacionales y han sido ampliamente utilizados. En la mayoría de los casos, tales algoritmos utilizan métodos estadísticos que tienen como objetivo el identificar motivos de secuencias que se encuentren sobre-representados de una manera estadísticamente significativa. A pesar de que esta aproximación estadística pareciera ser adecuada, su precisión sigue siendo limitada.

En este proyecto presentamos un nuevo protocolo de análisis computacional que integra el análisis estadístico de elementos de secuencia sobre-representados, con las propiedades biológicamente relevantes, tanto de los diferentes TFs como de los componentes con los que interactúa, el DNA y la RNApol. Lo anterior es de suma importancia en aquellos sistemas de regulación en donde los TFs pueden unirse de manera cooperativa a TFBSs de baja afinidad y que no suelen reconocerse con la mayoría de los algoritmos convencionales. Nuestro protocolo de análisis no solo considera a los motivos “estadísticamente significativos”, sino que también incluye a todos los motivos cuya longitud se hace variar de manera creciente y consecutiva. La integración de todos los motivos (de alta y baja afinidad/conservación) permite mapearlos contra las coordenadas de una secuencia de regulación en un organismo referencia.

Nuestro protocolo de análisis PProCoM permitirá mejorar las predicciones de TFBSs, mismas que posteriormente estarán sujetas a comprobación experimental.

Consideramos que nuestra nueva aproximación conceptual y metodológica aplicada a la identificación de TFBSs nos permitirá obtener modelos de regulación transcripcional consistentes con el conocimiento biológico-molecular de los mismos. En este sentido hemos obtenido resultados relevantes en el análisis de los reguladores GcvA, MetR, OxyR, IlvY, LysR y CynR, de la familia LysR. Mediante un estudio detallado del grado de conservación de los TFBSs y sus posiciones relativas, hemos propuesto modelos que proponen la dinámica de unión de los TFs, mismas que son consistentes con su naturaleza dimérica y propiedades de unión cooperativa en presencia de sus inductores.

5 OBJETIVO GENERAL

Desarrollar un nuevo protocolo para la identificación de TFBSs, con base en una nueva estrategia computacional, que incorpore la significancia estadística de la sobre-representación de motivos de secuencia nucleotídica y considere el conocimiento biológico-molecular adicional, referente a los diversos procesos involucrados en la unión TF-DNA para cada caso de estudio.

5.1 Objetivos particulares

1) Implementación de un *pipeline* computacional para la detección de motivos de secuencias, correspondientes a TFBSs en organismos bacterianos, con base en la metodología conocida como Perfiles Filogenéticos y que considere de manera simultánea: *i*) el grado de sobre-representación (conservación) de los motivos dentro de un conjunto de secuencias determinadas, *ii*) la posición relativa de los motivos entre sí, *iii*) la posición relativa de los motivos referente a los promotores transcripcionales, *iv*) la posición relativa de los motivos en relación a su grado de conservación y *v*) la posible unión cooperativa de los TFs en presencia de sus correspondientes coinductores.

2) Predicción computacional de los TFBSs en los seis sistemas de regulación de los miembros de la familia LysR, del phyla Gammaproteobacterias, para los cuales existe algún tipo de evidencia experimental que demuestre que se organizan de forma divergente y contigua respecto a sus correspondientes genes regulados. Se utilizará el

pipeline desarrollado en el objetivo particular #1 y el conjunto de secuencias intergénicas ortólogas no redundantes.

3) Elaboración de modelos dinámicos de regulación de los sistemas de estudio, con base en nuestras predicciones computacionales y en los elementos adicionales.

6 MÉTODOS

En términos generales, nuestra aproximación computacional se basa en la técnica de huellas filogenéticas (45) con modificaciones metodológicas que permiten incluir, no solo a los motivos de secuencia mayormente representados (con valores estadísticos significativos), sino también a los motivos de menor conservación. La identificación de todos los motivos se obtiene al realizar múltiples análisis individuales de sobre-representación, en donde el tamaño de búsqueda se va incrementando paulatinamente en cada análisis efectuado.

Los pasos fundamentales del protocolo computacional se ejemplifican en el análisis del sistema de regulación *ilvY-ilvC* de *E. coli* W3110 y se mencionan a continuación: *i*) identificación de organismos no-redundantes considerados en el estudio; *ii*) identificación de genes ortólogos; *iii*) predicción de operones bacterianos; *iv*) obtención de regiones intergénicas; *v*) obtención de motivos estadísticamente significativos; *vi*) mapeo de motivos significativos sobre una región de regulación de un organismo modelo. *vii*) alineamiento múltiple del conjunto de motivos (**Figura 5**). Se mencionan a continuación los aspectos fundamentales de los pasos antes mencionados.

i) Identificación de organismos no-redundantes considerados en el estudio.

Con el objetivo de evitar un sesgo introducido por la secuenciación preferencial de ciertos organismos modelo (por ejemplo, se tiene la secuencia de más de 30 cepas de *E. coli*), nuestro análisis contempla como primer paso una selección de genomas no-redundantes del grupo de las Gammaproteobacterias. Dicha selección se efectúa con base en el valor de distancias filogenéticas evaluadas con el programa PROTDIST del

paquete de inferencia filogenética PHYLIP (76), a partir de un alineamiento múltiple de la secuencia concatenada de un conjunto de 31 genes considerados como constitutivos y definidos en Ciccarelli, F.D., *et al* (77).

ii) Identificación de genes ortólogos.

En nuestro grupo de trabajo, la asignación de grupos de ortología se realiza utilizando el programa HMMERsearch (139) y los modelos de Markov escondidos para cada una de las de 4,873 familias de genes ortólogos definidas en la base de datos COG (*Clusters of Orthologous Groups of genes*) (78).

iii) Predicción de operones bacterianos.

En nuestro estudio se utilizó un algoritmo computacional de predicción de operones altamente preciso desarrollado en nuestro grupo de investigación (79). Éste se basa en la construcción de una red neuronal cuyas variables de entrada son las distancias intergénicas entre los genes divergentes contiguos (codificados en direcciones y cadenas opuestas contiguas en el DNA) y la relación funcional de sus correspondientes productos polipeptídicos. Esta relación funcional se define para el conjunto de genes en la base de datos STRING (80).

iv) Obtención de regiones intergénicas.

Las regiones intergénicas serán aquellas correspondientes a las secuencias río arriba de cada operón regulado por miembros de la familia LysR, considerando el conjunto de organismos no-redundantes definidos en las primeras etapas de nuestro protocolo de análisis.

v) Obtención de motivos estadísticamente significativos.

El análisis de motivos de regulación, sobrerepresentados en cada conjunto de regiones ortólogas, se llevará a cabo utilizando el programa MEME (46,81,82). A pesar de que este programa puede definir automáticamente el tamaño del motivo con el cual se obtiene el valor de sobrerepresentación estadísticamente más significativo, con nuestro protocolo realizaremos alrededor de 50 análisis por cada uno de los reguladores de estudio. En dichos análisis se hará variar la longitud del ancho de motivo, desde el más pequeño de 10 nt, hasta el más grande 100 nt, con incrementos de dos pares de bases en cada iteración. Solo se considera el motivo estadísticamente más significativo por cada corrida.

vi) Mapeo de motivos significativos sobre una región de regulación de un organismo modelo.

Con el propósito de identificar las posiciones de los diferentes motivos al variar la longitud de cada corrida de MEME, cada uno de ellos se mapea en la región intergénica de los correspondientes operones regulados de un organismo modelo. En nuestro caso, el organismo modelo corresponde a la cepa W3110 de *E. coli.* Todos los nucléotidos que no pertenecen al motivo en cuestión, se enmascaran en la secuencia con la letra “n”.

vii) Alineamiento múltiple del conjunto de motivos.

El alineamiento múltiple de todos los motivos obtenidos se obtiene sencillamente apilando las secuencias obtenidas durante el mapeo de motivos significativos tomando en consideración las coordenadas de una región de regulación de un organismo modelo.

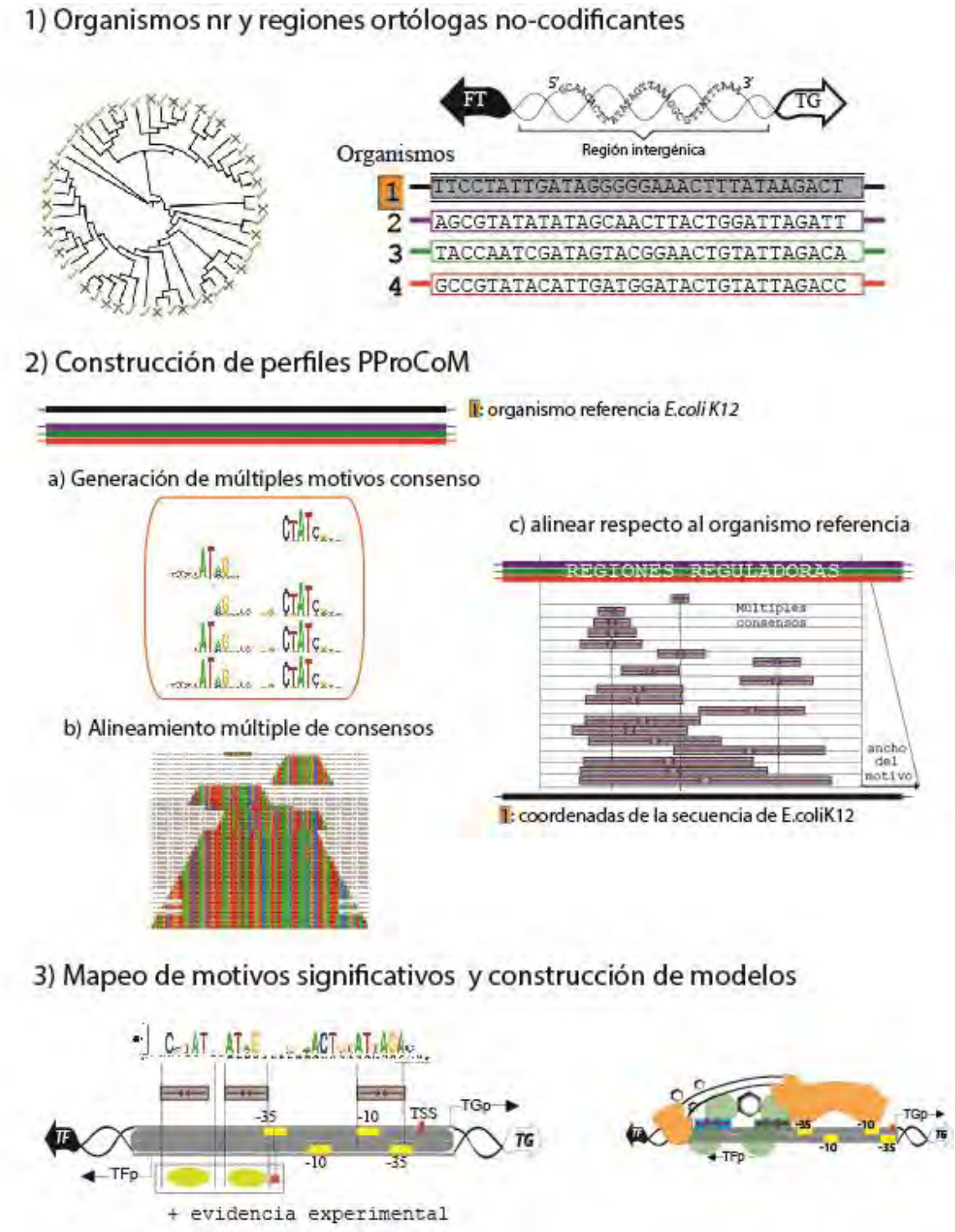

Figura 5 Diagrama general de la metodología desarrollada. 1) Selección de organismos no-redundantes (nr) y utilización de las regiones intergénicas. 2) Construcción de perfiles filogenéticos de secuencias consenso de múltiples longitudes. 3) Mapeo de los motivos significativos e interpretación de resultados mediante la construcción de mecanismos de regulación.

7 RESULTADOS

En el presente trabajo se evaluó el desempeño del protocolo PProCoM mediante la identificación *in silico* de los TFBSs para seis TFs miembros de la familia de reguladores LysR. La elección de este conjunto de proteínas se hizo con base en sus características comunes. Se trata de TFs codificados de manera divergente a su TG regulado, con propiedades de unión cooperativa, estructura funcional de dímero de dímeros, y corresponden a sistemas en donde participa un solo TF, por lo cual se analizaron regiones donde participan reguladores locales. El análisis genómico se realizó para el conjunto de organismos no redundantes (nr) del *phyla* Gammaproteobacteria y se utilizó a *Escherichia coli* como organismo modelo debido a la gran cantidad de información disponible sobre su regulación. Como resultado del análisis comparativo, los sistemas de regulación se dividieron en tres grupos distintos al realizar la interpretación del análisis, según la estructura cuaternaria de los TFs y las propiedades moleculares de sus correspondientes sitios de unión. Las características generales de los tres grupos se resumen en la **Tabla 1**.

GRUPO	TF	TG	DI	Función
1	GcvA	*gcvB*	128	Degradación de glicina
	MetR	*metJ*	236	Biosíntesis de metionina
2	OxyR	*oxyS*	95	Estrés oxidativo
	IlvY	*ilvC*	149	Biosíntesis de acetolactato
	CynR	*cynT*	108	Detoxificación de cianato
3	LysR	*LysA*	122	Biosíntesis de lisina

Tabla 1 Descripción general de los TFs miembros de la familia LysR considerados en el estudio clasificados por grupo de acuerdo a las características de sus TFBSs. Factor transcripcional (TF), gen blanco (TF), longitud de la región intergénica común entre los genes divergentes (DI) y función del regulador.

7.1 Grupo uno: GcvA y MetR

Para el análisis PProCoM de los TFs GcvA y MetR, se identificaron dos TFBSs con simetrías de Invertidos Repetidos (IR1 e IR2 de **Figura 6** y **Figura 7)**. Ambos reguladores comparten las siguientes características: i) los genes codificantes del TF y su TG se transcriben en direcciones opuestas, ii) la activación transcripcional del TG ocurre cuando un dímero del TF se une al TFBS localizado junto a la caja -35 del promotor TG, e interactúa con la RNApol (IR2 de **Figura 6** y **Figura 7)**. Simultáneamente, la auto-represión del TF sucede cuando el mismo TF se une al TFBS que traslapa su propio promotor ubicado en la cadena opuesta del DNA (IR1 de **Figura 6** y **Figura 7)**

7.1.1 Sistema de regulación GcvA

La proteína GcvA (*Glycine Cleavage A*), es un TF que regula la transcripción de genes involucrados en la vía serina-glicina de *E. coli* (86,87). Se codifica por el operón divergente *gcvA-gcvB* con promotores traslapados, y en su mecanismo de regulación se ha descrito que GcvA funciona como represor en presencia de glicina, autorregulándose negativamente al hacer contacto directo con las subunidades beta de la RNApol (88), al mismo tiempo incrementa coordinadamente la transcripción del gen divergente *gcvB*, el cual codifica para GcvB, un sRNA (*small* RNA) que regula la expresión de componentes de transporte periplásmico DppA y OppA entre otros (90). Adicionalmente, GcvA regula la transcripción del operón *gcvTHP* (86).

Mediante un análisis de *footprinting* con DNasa I, Wilson *et al*. reportaron que en la secuencia de la región intergénica de los genes *gcvA-gcvB* del organismo *E. coli*, la proteína GcvA protege una región de DNA de 48 pb de longitud mientras que en la región intergénica del operón *gcvTHP* protege otras dos secuencias de 35 y 57 pb de largo (86). El alineamiento de esas secuencias reveló un motivo conservado de 5'-CTAAT-3', y se determinó experimentalmente, por mutagénesis de sitio dirigida, que el motivo era importante para la unión de GcvA, la cual regula negativamente la transcripción de *gcvA y gcvTHP* (86,89,90). En general, los sitios de GcvA no presentan una clara conservación de la secuencia, excepto por un motivo corto de 5'-CTAAT-3'. Además, las regiones protegidas de GcvA contienen la secuencia IR 5'-ATTA-n7-TAAT-3' (86), la cual coincide con los sitios de unión de GcvA reportados en la base de datos de RegPrecise (91).

Los resultados obtenidos con PProCoM para el análisis de las regiones intergénicas de los genes *gcvB-gcvA*, permiten identificar la presencia de dos secuencias

IR con longitudes de 15 pb (5'-ATTAG-n5-CTAAT-3', ver **Figura 6**), mismas que están incluidas en los motivos reportados por Wilson *et al*., 5'-CTAAT-3' y 5'-ATTA-n7-TAAT-3' (86). Asimismo, las posiciones centrales de los motivos IR1 e IR2 predichos, se localizan a -65 y -43 pb del TSS, respectivamente (**Figura 6**).

Es importante resaltar que las secuencias mostradas en la **Figura 6** no representan el resultado estándar de un alineamiento de secuencias, ya que se obtienen de acuerdo con la localización de múltiples motivos conservados de diferentes tamaños en la región de regulación de los genes *gcvB-gcvA* de *E. coli* (ver sección de Métodos).

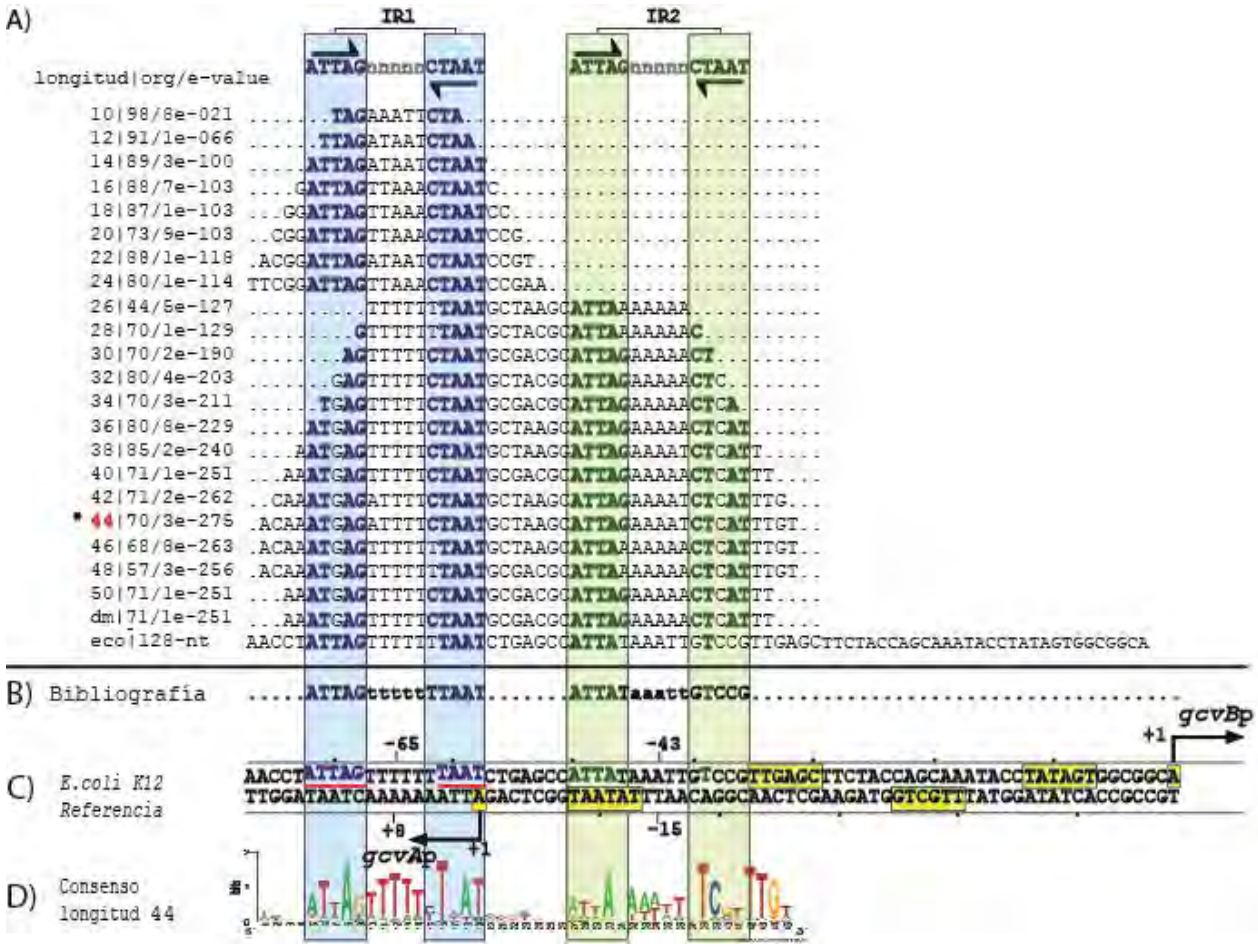

Figura 6 Análisis PProCoM de la región intergénica gcvA-gcvB en Gammaproteobacteria. A) perfil de múltiples secuencias consenso de longitud creciente posicionadas de acuerdo a la región intergénica correspondiente en *E. coli* K12. A la izquierda de la columna separada por un pipe, el ancho de la ventana usado en cada análisis de MEME, el E-valor obtenido para cada motivo y se indica el número de organismos que presentan el motivo identificado (150 organismos utilizados en nuestro análisis). La secuencia indicada como "dm" es la correspondiente al "default motif" y corresponde al motivo por default sin forzar el tamaño de la ventana de análisis (sección de métodos). Los motivos consenso de las secuencias IR (IR1 e IR2) se despliegan en la parte superior de la figura y son representados con flechas invertidas negras. B) los TFBSs reportados con evidencia experimental citadas en Wilson RL y Jourdan AD *et al.(*86,89). Los TSS se reportan con flechas negras sólidas y han sido previamente identificados o bien se reportan con nuestro análisis. Las cajas de los promotores -10 y -35 se indican por líneas sólidas si esos elementos se han reportado previamente y con líneas punteadas si los elementos fueron identificados basados en nuestros análisis PProCoM. Se indican las posiciones centrales de los motivos IR con relación al TSS de los genes codificantes para el TF o el TG. Los nt de la secuencia IR1 en *E. coli,* que hacen *match* con el consenso se delinean con una raya roja. D) Se seleccionó un LOGO correspondiente a una secuencia representativa del perfil de consensos (marcado con un asterisco) y éste incluye todos los motivos reguladores de la región intergénica de estudio.

7.1.2 Sistema de regulación MetR

La proteína MetR es un TF que regula la expresión de genes involucrados en la biosíntesis del aminoácido metionina y protección contra el óxido nítrico (92-97). La actividad transcripcional de MetR está modulada por homocisteína, un precursor metabólico de la metionina. En presencia de homocisteína, la proteína MetR activa la transcripción de genes tales como *metE y glyA* y autoreprime su propia transcripción, pero también la de otros genes, como el caso de *metH, meta y hmp* (92-100).

En los organismos *E. coli* y *Salmonella typhimurium*, los genes de *metE* y *metR* comparten la región de regulación y se transcriben divergentemente de promotores traslapados (92-98). Mediante análisis mutacionales y *footprinting* con DNasa I, se probó experimentalmente que en *S.typhimurium*, MetR se une a dos secuencias IR organizadas codireccionalmente, las cuales presentan diferente conservación de la secuencia respecto a una secuencia consenso TGAAnnTnnTTCA-3' (98) , por lo tanto se cree que son sitios que varían su afinidad de unión. En *E. coli*, se reportaron dos sitios de unión con las mismas características para la región de regulación de genes *hmp-glyA*, los cuales son regulados por MetR y se encuentran transcritos divergentemente (99,100). Asimismo, se ha reportado que la presencia de homocisteína aumenta la afinidad de unión de MetR para reconocer esos sitios de unión a DNA contiguos y activar la transcripción de *metE* y la represión de *metR* (98). Actualmente no existe evidencia experimental que sustente la existencia de dos sitios de unión en la región intergénica *metR-metE*, sin embargo, mediante los análisis realizados con PProCoM, logramos identificar dos posibles secuencias IR de 15 pb, con un consenso 5'-ATGAA-n5-TTCAT-3', el cual equivale al tamaño reportado para los TFBSs de la familia LysR (101).

Basándonos en la secuencia de referencia de *E. coli* logramos identificar un motivo distal IR1 localizado a -63 pb del TSS de *metE* y un motivo proximal IR2, menos conservado localizado a -41 pb del TSS de *metE* (**Figura 7**). Las ubicaciones centrales de estos posibles TFBS representan posiciones preferenciales de los activadores transcripcionales en *E. coli* (102,103*)*. En la misma figura, se muestra que la secuencia inter-motivo compartida entre los motivos IR1-IR2 de *E. coli K12* es 1 pb más corta que la secuencia inter-motivo IR1-IR2 mostrada en el alineamiento del perfil PProCoM (líneas punteadas de la **Figura 7**), posiblemente este efecto se deba a una base nucleotídica que se perdió en el espacio inter-motivo del organismo *E. coli K12*, al respecto se ha demostrado que las variaciones largas en el espacio inter-motivo, de alrededor de 6 pb (media vuelta de la hélice de DNA) producen un efecto negativo en la transcripción de *metE* en *S. typhimurium* (104). Adicionalmente, respecto a las mutaciones puntuales en cualquiera de los dos TFBSs propuestos, también se ha reportado que disminuyen la transcripción de *metE*, indicando que para la completa activación de *metE* se requieren ambos TFBSs (98). La secuencia consenso de 15 pb obtenida por nuestro análisis PProCoM coincide con aquella reportada para MetR en la base de datos RegPrecise (*i.e*., 5'-ATGAAAATTTTTCAT-3') (91).

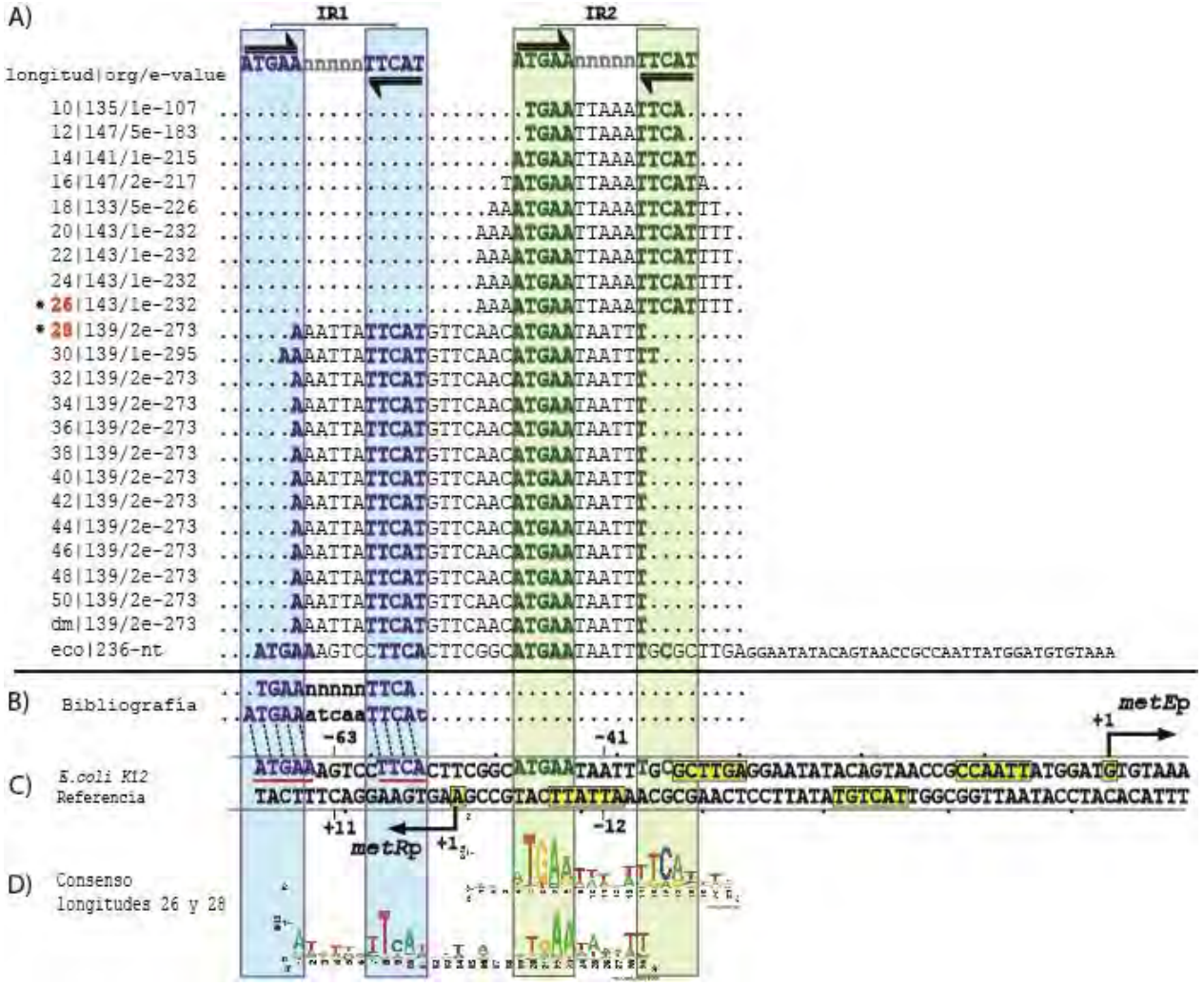

Figura 7 Análisis PProCoM para la región reguladora *metR-metS.* A) Alineamiento de las secuencias consenso de múltiples longitudes. B) Secuencias reportadas en la literatura de acuerdo con evidencias experimentales (11,98), útiles para mapear mejor los sitios predichos. C) Secuencia de doble cadena del organismo modelo *E. coli K12*, con la representación de sus cajas de promotores, TSS y posición central de los TFBS identificados (desplazado 1 pb respecto al análisis PProCoM). D) Se incluye la representación gráfica del LOGO de dos secuencias consenso para una mayor cobertura de la región de regulación (26 y 28 pb). En la parte superior se enmarcan los nt conservados en dos sitios con simetría de IR y se distinguen en dos colores. IR1= azul, IR2 = verde.

7.2 Grupo dos: **OxyR, IlvY y CynR**

El grupo dos se integra por las proteínas OxyR, IlvY y CynR, a las cuales se les realizó un análisis de sus regiones reguladoras mediante el protocolo PProCoM. Se identificaron tres TFBSs en sus regiones intergénicas respectivas (IR1, IR2 e IR3 **Figura 8).** De acuerdo a nuestras interpretaciones, para los tres casos la activación del TG ocurre gracias a la unión cooperativa de dos dímeros del TF que reconocen los TFBS IR1 e IR2, en presencia de los inductores respectivos. El sitio IR2 se localiza adyacente a la caja -35 del promotor del TG, y es reconocido por el dímero del TF para promover la transcripción del TG mediante su interacción con la RNApol. Adicionalmente, ocurre la auto-represión simultánea del TF gracias a que IR1 se traslapa con el promotor del mismo TF localizado en la cadena opuesta del DNA (IR1 de la **Figura 8**). La principal diferencia respecto a los sistemas reguladores del grupo uno, es que además se identificó la presencia de un tercer TFBS que se traslapa con la caja -35 del promotor del TG (IR3; **Figura 8**). Una característica notable de este tercer TFBS (IR3, usado para la represión del TG) es que traslapa parcialmente el segundo TFBS (IR2, usado para activación del TG), de manera que la unión del TF sobre algunos de esos dos sitios es mutuamente excluyente y determina la actividad reguladora (*i.e,* activación o represión) del TF sobre el TG.

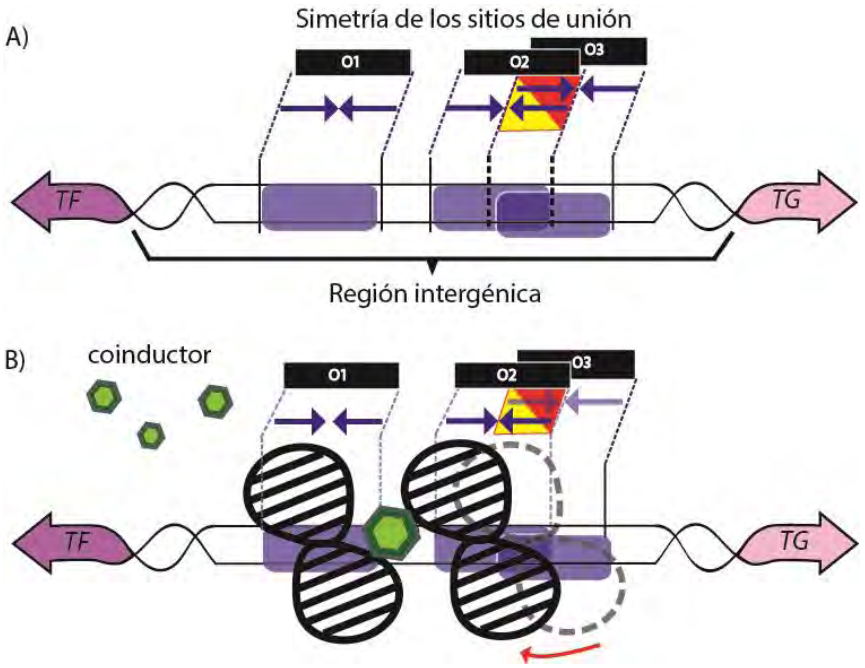

Figura 8 A) Esquema general de los TFBSs identificados para el grupo 2 (proteínas OxyR, IlvY, CynR), B) se muestra el desplazamiento del sitio en presencia de inductor.

7.2.1 El sistema de regulación OxyR

La proteína OxyR es un factor de transcripción sensible a los cambios oxidación-reducción que sufre la célula. OxyR participa en la regulación de la expresión de genes implicados en diversos mecanismos moleculares, tales como la protección al estrés oxidativo, el balance óxido-reducción y la ingesta/absorción de manganeso (105-108). La actividad de OxyR depende de su estado oxidado, el cual determina la formación de los enlaces disulfuro reversibles de un par de cisteínas presentes en su secuencia de aminoácidos (109). La forma de OxyR en su estado oxidado, permite activar la transcripción del gen divergente *oxyS*, el cual codifica un RNA pequeño. Además, OxyR

reprime su propia expresión bajo condiciones tanto de oxidación como de reducción (110).

Mediante análisis basados en experimentos de *footprinting* con DNasa I, el grupo de Tartaglia *et al* demostró que OxyR se une a una región inusualmente larga de DNA que cubre 45 pb, con dos probables sitios de unión a OxyR que no presentan una similitud evidente entre sus secuencias (111). Posteriormente, utilizando un ensayo de unión *in vitro* de OxyR sobre los oligonucleótidos al azar y el análisis de *footprinting* con DNasa I, Toledano *et al.* demostraron que el reconocimiento del DNA por OxyR depende de ambos estados, oxidado y reducido. En su forma oxidada, OxyR reconoce una región de DNA que incluye cuatro repeticiones de la secuencia 5'-ATAGnt-3', localizada en una región que cubre cuatro zurcos mayores contiguos de una de las caras de la hebra de DNA. En su forma reducida, OxyR se une a una secuencia más corta de dos repeticiones de la secuencia 5'-ATAGnt-3' localizada en una región que cubre dos pares de zurcos mayores y separadas por una vuelta hélice (110).

Con el análisis PProCoM realizado para la región intergénica *oxyR-oxyS* de múltiples organismos ortólogos, se identificó la presencia de tres secuencias IR con una longitud de 15-pb (5'-ATAG-n7-CTAT-3'). Al comparar el perfil con la región intergénica de *oxyR-oxyS* contra las secuencias homólogas en el organismo referencia *E. coli*, se observó que los motivos predichos IR1, IR2 e IR3 se localizan en las posiciones centrales respectivas a -66, -44 y -35 pb del TSS de *oxyS* (ver **Figura 9**). El análisis e interpretación de nuestros resultados nos permitieron dilucidar y proponer un mecanismo de regulación transcripcional (**Figura 10**).

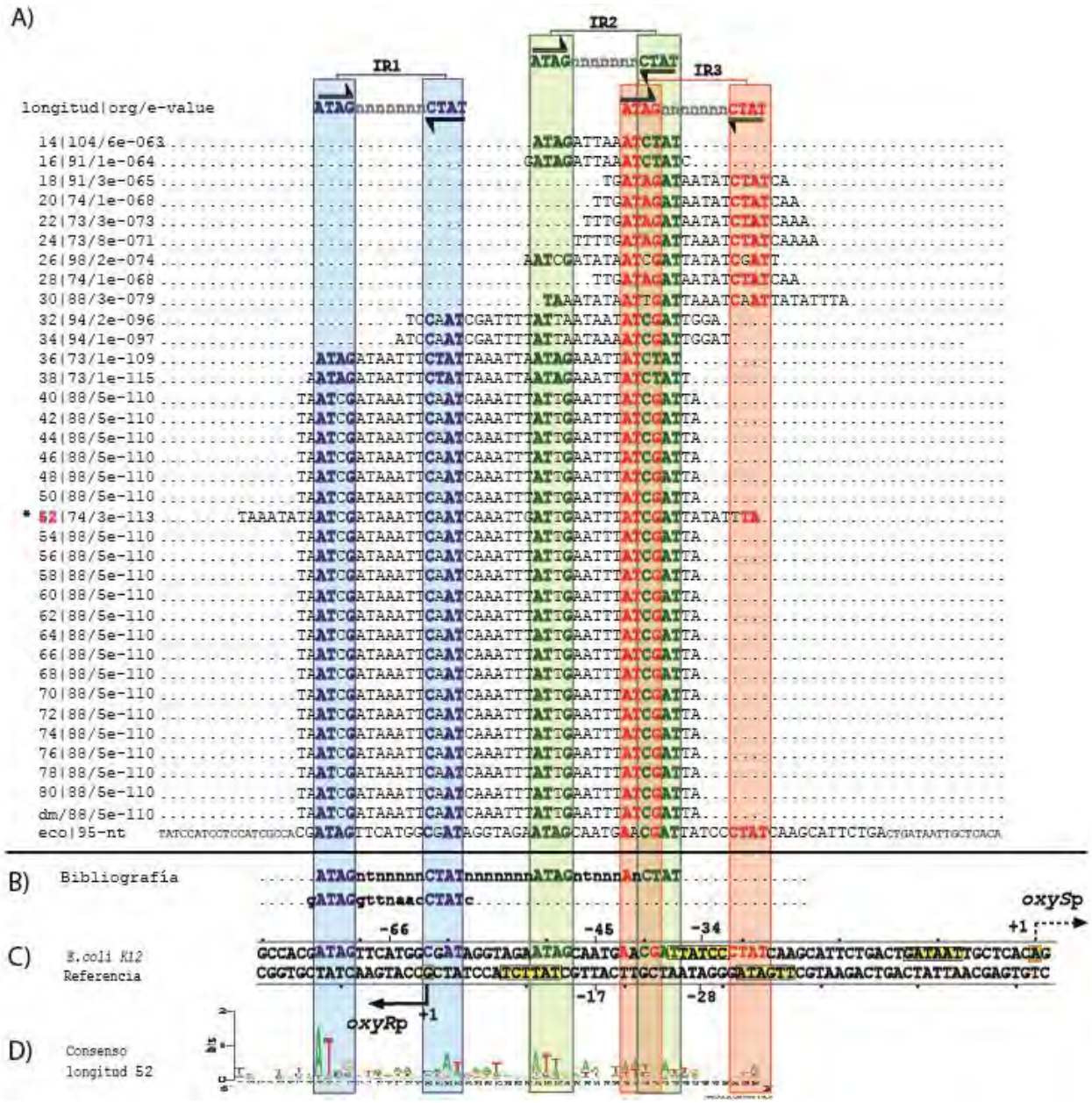

Figura 9 Análisis PProCoM para la región reguladora *oxyR-oxyS*. A) Alineamiento de las secuencias consenso de múltiples longitudes. B) Secuencias reportadas en la literatura de acuerdo con evidencias experimentales (110,69), útiles para mapear mejor los sitios predichos. C) Secuencia de doble cadena del organismo modelo *E. coli K12*, con la representación de sus cajas de promotores, TSS y posición central de los TFBS identificados. D) Representación gráfica del LOGO de la secuencia consenso de tamaño 52 pb. En la parte superior se enmarcan los nt conservados en tres sitios con simetría de IR y se distinguen en tres colores. IR1= azul, IR2 = verde, IR3= rojo. Se observa un traslape en la mitad del motivo IR2-IR3.

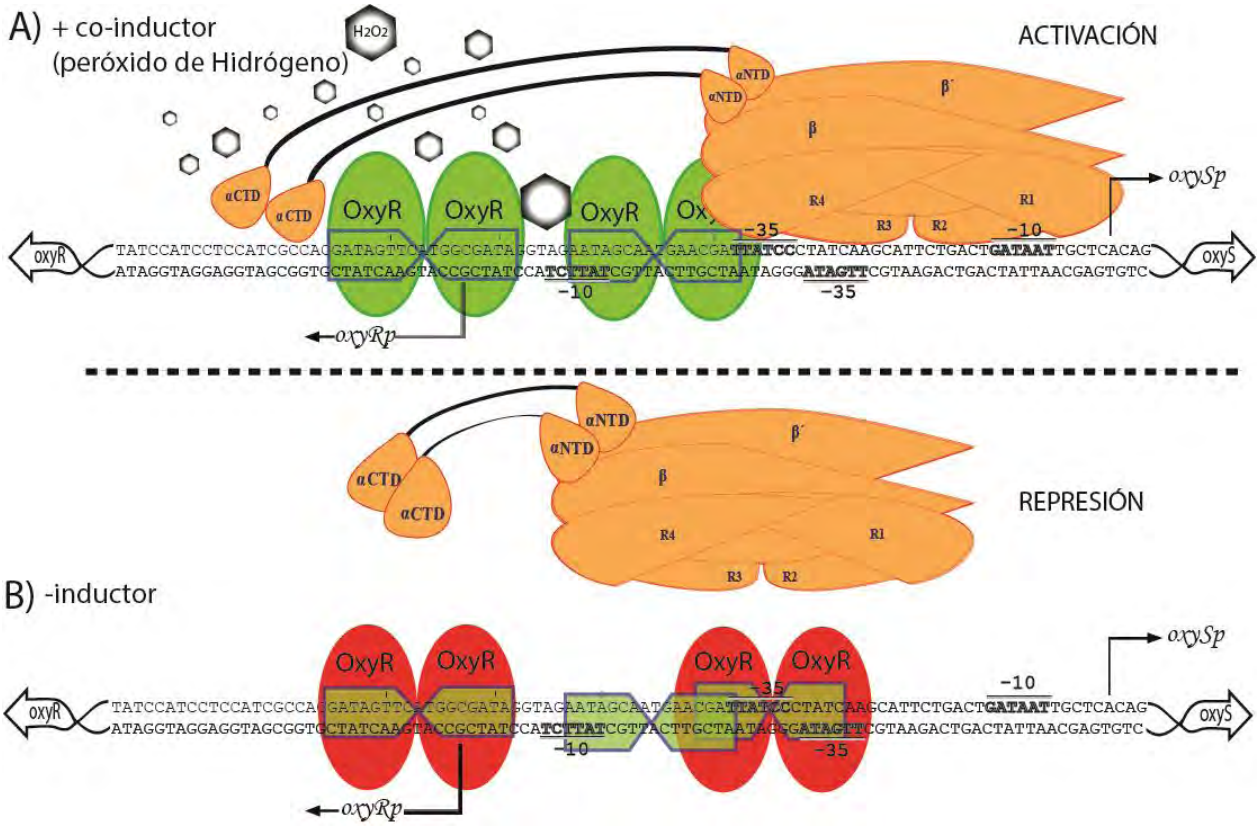

Figura 10 Modelo de regulación dual propuesto para el TF OxyR en presencia (A) y ausencia (B) del inductor (detalles en la sección de Modelo Dinámico de Regulación).

7.2.2 El sistema de regulación IlvY

La proteína IlvY regula positivamente la transcripción de *ilvC*, un gen involucrado en la biosíntesis de los aminoácidos valina e isoleucina. La activación transcripcional de *ilvC* por IlvY depende de la presencia del inductor acetolactato o acetohidroxibutirato. Al mismo tiempo, IlvY autoregula negativamente su propia transcripción, independientemente de la presencia del inductor (112,113).

Los genes *ilvY* e *ilvC* se transcriben divergentemente de promotores traslapados. Utilizando el análisis de *footprinting* con DNasa I, Wek y Hatfield propusieron que IlvY se

une a dos secuencias de 27 pb localizadas en la región promotora de *ilvY-ilvC*, los nombró operadores O1 y O2 (113). Esas regiones se organizan codireccionalmente y poseen dos motivos imperfectos con simetría de invertidos repetidos, de 21 pb de longitud. La secuencia del operador O1, 5'-ACgTTGCAAaaaTTGCAAtGT-3' (centrado en la posición +17 relativo al TSS de *ilvY*), y la secuencia O2, 5'-aTATatCaatttccGcaATAa-3' (misma que traslapa las cajas de los promotores a -10 y -35 pb propuestos para *ilvY* y la caja -35 del promotor propuesto para *ilvC*). El motivo de unión consenso propuesto para IlvY es 5'-A[C/T]ATTGCAA-3' (113) y resulta común para los operadores O1 y O2. Esos autores proponen que IlvY reprime su propia transcripción al unirse a O1 independientemente del inductor, y activa la transcripción de *ilvC* en presencia de los inductores del sistema, al unirse a los operadores O1 y O2 de manera cooperativa. Los mismos autores también proponen que la activación transcripcional de *ilvC* ocurre gracias a las interacciones de la RNApol y el TF (IlvY) cuando éste se une a O2 en presencia del metabolito inductor, o por un cambio en la conformación de la caja -35 del promotor de *ilvC*. En este sentido, Rhee *et al*. propusieron que la transcripción de los genes divergentes *ilvY* e *ilvC* está acoplada a un superenrollamiento del DNA, que incrementa la unión de la RNApol a su promotor, alrededor de 100 plegamientos (foldings) (112). Mediante el análisis PProCoM de la región intergénica *ilvY-ilvC* se identificó la presencia de tres secuencias IR de 15 pb (5'-TTGCA-n5-TGCAA-3'; ver **Figura 11**). Considerando la secuencia referencia de la región intergénica de *ilvY-ilvC* en *E. coli*, las posiciones centrales de los motivos IR1, IR2 e IR3 se localizan a -65, -43 y -34 pb del TSS de *ilvC*, respectivamente (**Figura 11**).

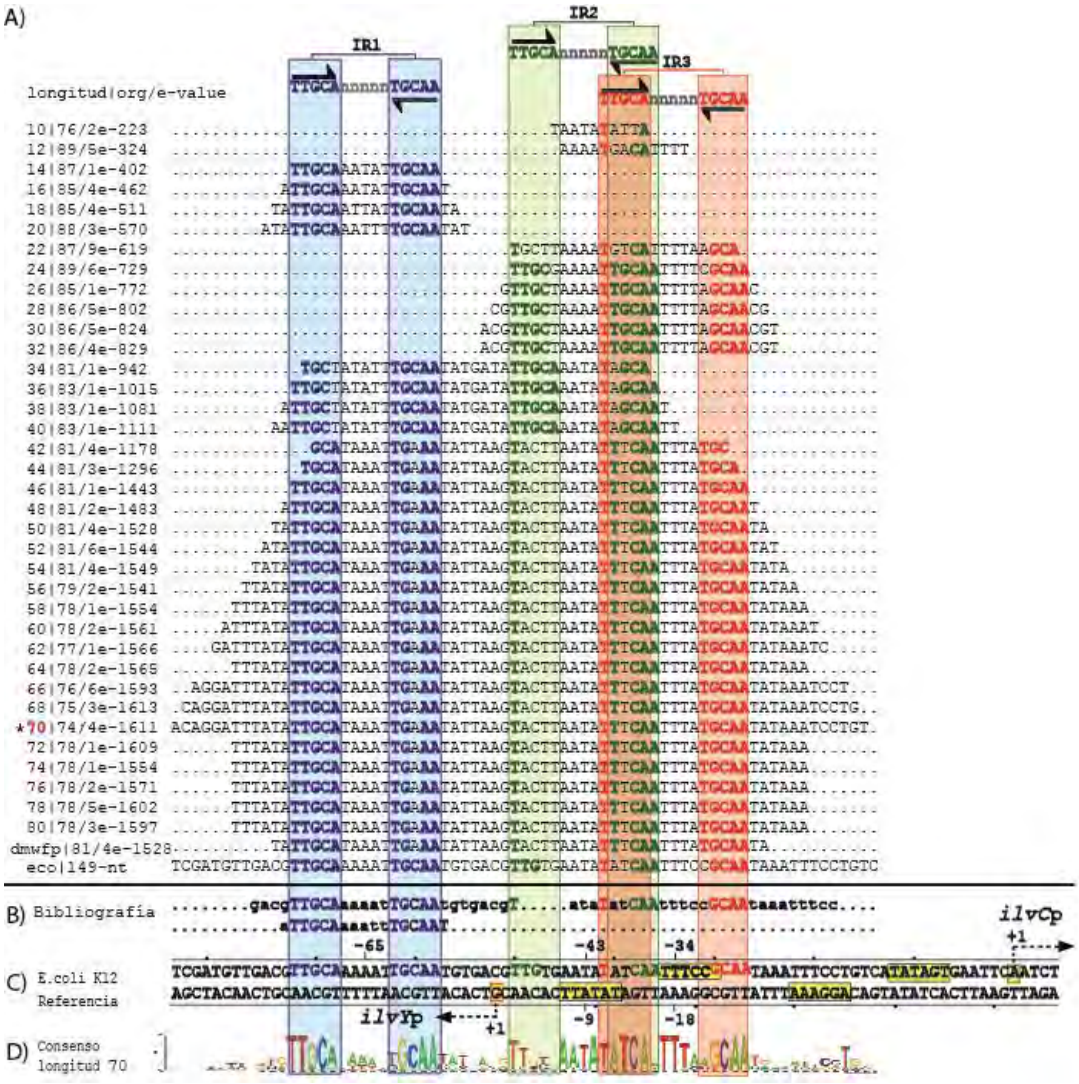

Figura 11 Análisis PProCoM para la región reguladora *ilvY-ilvC.* A) Alineamiento de las secuencias consenso de múltiples longitudes. B) Secuencias reportadas en la literatura de acuerdo con evidencias experimentales (113,98), útiles para mapear mejor los sitios predichos. C) Secuencia de doble cadena del organismo modelo *E. coli K12*, con la representación de sus cajas de promotores, TSS y posición central de los TFBS identificados. D) Representación gráfica del LOGO de la secuencia consenso de tamaño 70 pb. En la parte superior se enmarcan los nt conservados en tres sitios con simetría de IR y se distinguen en tres colores. IR1= azul, IR2 = verde, IR3= rojo. Se observa un traslape en la mitad del motivo IR2-IR3.

7.2.3 El sistema de regulación CynR

La proteína CynR es un TF que regula la transcripción del operón *cynTSX*, el cual está involucrado en la detoxificación de cianato. El cianato se usa como fuente de nitrógeno debido a su hidrólisis, con la consecuente producción de amonio y bicarbonato (114). La activación del operón *cynTSX* depende de la presencia de cianato en el medio. Asimismo, CynR regula negativamente su propia transcripción independientemente de la existencia de cianato en el medio (114). Como es el caso de los sistemas reguladores de tipo LysR antes mencionados, el gen codificante para el TF (*cynR*) y sus genes blanco regulados (*cynTSX*) se transcriben en direcciones opuestas, y sus promotores correspondientes se traslapan (115,116). Utilizando un análisis de digestión con DNasa I, Lamblin y Fuchs demostraron que CynR se une a una secuencia de 60 pb en la región intergénica de los genes *cynR-cynTSX*, y propusieron que esta región contiene dos probables sitios de unión con diferentes afinidades (116). Se dijo que la primera de esas regiones, denominada “R1” (de secuencia 5'-ATAAGTAAA-3´), tiene la mayor afinidad de unión, mientras que la segunda denominada “R2” (de secuencia 5'-ATAAGGTAA-3´), se traslapa completamente sobre la región promotora de cynR, así como la región del promotor -35 del operón divergente *cynTSX* (115,117).

Los autores sugieren como primera instancia, que un dímero de CynR podría unirse a R1 (*i.e.* la región más conservada), y simultáneamente como una segunda instancia, otro dímero de CynR podría unir fuertemente a R2 de manera cooperativa. Esos autores también propusieron que la activación transcripcional del operón *cynTSX*

ocurre en presencia de cianato, el cual se cree que desencadena un cambio conformacional en CynR, modificando su interacción con el DNA (116).

Con nuestro análisis PProCoM de la región intergénica *cynR-cynTSX*, se identificó la presencia de tres secuencias IR de 15 pb de longitud (5'-ATAA-n7-TTAT-3'), incluyendo las secuencias propuestas por Lamblin y Fuchs (**Figura 12**). Considerando la región intergénica *cynR-cynTSX*, las posiciones centrales de los motivos IR1, IR2 e IR3 predichos se localizaron a -66, -44 y -34 pb del TSS de *cynTSX*, respectivamente (**Figura 12**).

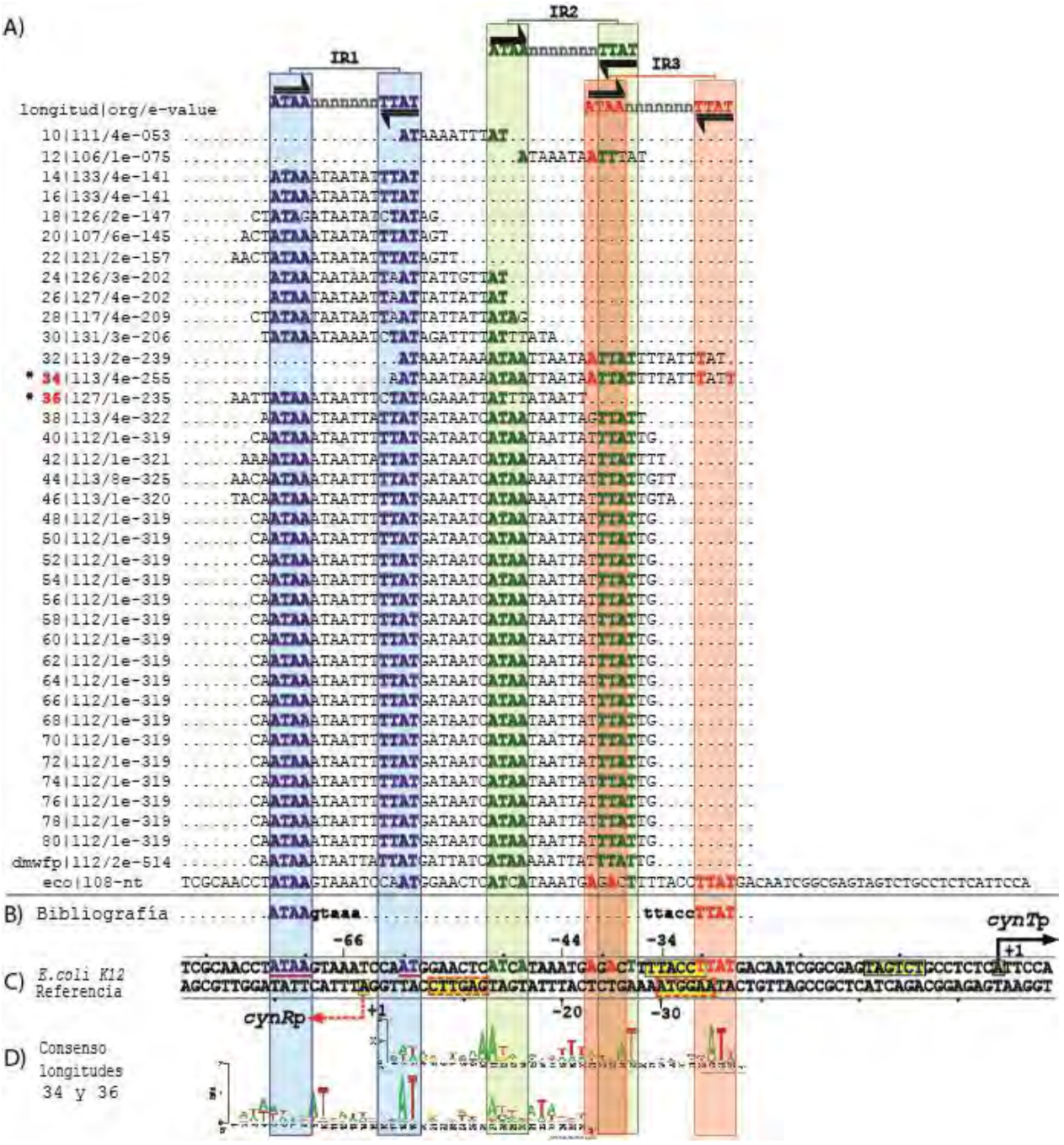

Figura 12 Análisis PProCoM para la región reguladora *cynR-cynT.* A) Alineamiento de las secuencias consenso de múltiples longitudes. B) Secuencias reportadas en la literatura de acuerdo con evidencias experimentales (114,115), útiles para mapear mejor los sitios predichos. C) Secuencia de doble cadena del organismo modelo *E. coli K12*, con la representación de sus cajas de promotores, TSS y posición central de los TFBS identificados. D) Se incluye la representación gráfica del LOGO de dos secuencias consenso para una mayor cobertura de la región de regulación (34 y 36 pb). En la parte superior se enmarcan los nt conservados en tres sitios con simetría de IR y se distinguen en tres colores. IR1= azul, IR2 = verde, IR3= rojo. Se observa un traslape en la mitad del motivo IR2-IR3.

7.3 Grupo tres: LysR

El grupo tres se compone por un solo sistema regulador, el TF LysR. En este caso se lograron identificar tres TFBSs en la región reguladora intergénica (IR1, IR2 e IR3, **Figura 13**). De la misma forma como se aprecia en los grupos uno y dos de los miembros de la familia LysR, la activación transcripcional del TG (*lysA*) ocurre gracias a la unión cooperativa de dos dímeros, los cuales reconocen los sitios IR1 e IR2 en presencia del inductor (*i.e.* ácido diaminopimélico). La auto-represión del TF (*lysR*) ocurre simultáneamente porque el sitio IR1 superpone el promotor del TF localizado en la cadena opuesta del DNA (IR1 de la **Figura 13**). La principal diferencia respecto al sistema regulador del grupo dos es que el segundo y tercer sitio no están traslapados (IR2 y IR3; **Figura 13**).

El sistema de regulación LysR

La proteína LysR es un TF que regula la transcripción del gen *lysA*, el cual codifica para una enzima que cataliza la vía final del metabolismo de biosíntesis de lisina. LysR se autoregula negativamente y regula positivamente la transcripción de *lysA* en presencia de su inductor (el ácido diaminopimélico) (118-120). Como se describió con los casos anteriores, los genes codificantes del TF (*lysR*) y su TG regulado (*lysA*) se transcriben en direcciones opuestas. Los TFBSs de LysR y su mecanismo de regulación no han sido identificados. Sin embargo, se ha determinado que los sitios de unión de LysR se encuentran en fragmentos de DNA de 73 pb, localizado a 48 pb río arriba del gen estructural *lysA* (119). La concentración intracelular de LysR en su forma activa podría

ser limitante ya que su papel como regulador disminuye cuando el fragmento de DNA antes mencionado es clonado en un plásmido. Basado en análisis experimentales, se predijo que el TSS se localiza 26 pb río arriba de su gen estructural (121). Sin embargo, se ha predicho un probable promotor de *lysA*, con una caja -35 (TTGcat) y una caja -10 (TATTTT), localizado a 52 pb de la región codificante de *lysA (122)*. Tomando en cuenta lo anterior, se ha propuesto que el correspondiente TSS se localiza a 3 pb río abajo de la caja -10 de los promotores predichos (122).

Con nuestro análisis PProCoM de la región intergénica *lysR-lysA*, se identificó la presencia de tres secuencias IR de 15 pb (5'-ATATC-n5-GATAT-3', ver **Figura 13**), IR1, IR2 e IR3, mismas que al compararse con la secuencia de referencia en *E. coli*, se localizan respectivamente con posiciones centrales localizadas a -64, -43 y -9 pb del TSS de *lysA* (ver **Figura 13**). Basándonos en las posiciones de los tres TFBSs, postulamos que el TSS de *lysA* se localiza a 22 pb río arriba de su gen estructural.

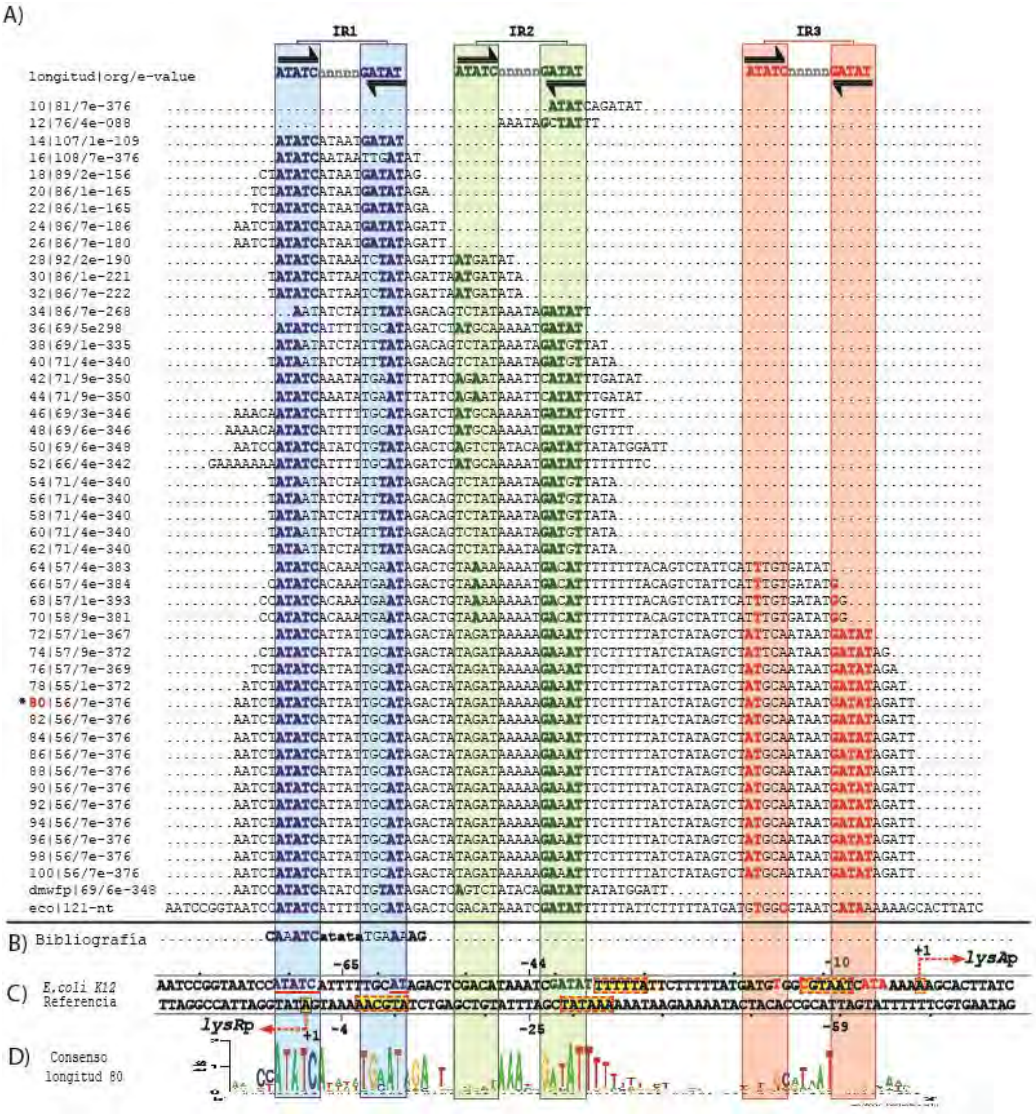

Figura 13 Análisis PProCoM para la región reguladora *lysR-lysA*. A) Alineamiento de las secuencias consenso de múltiples longitudes. B) Secuencia reportada en la literatura de acuerdo con evidencias experimentales (98) útiles para mapear mejor los sitios predichos. C) Secuencia de doble cadena del organismo modelo *E. coli K12*, con la representación de sus cajas de promotores, TSS y posición central de los TFBS predichos. D) Se incluye la representación gráfica del LOGO de la secuencia consenso de 80 pb. En la parte superior se enmarcan los nt conservados en tres sitios con simetría de IR y se distinguen en tres colores. IR1= azul, IR2 = verde, IR3= rojo.

8 DISCUSIÓN

Motivos comunes entre los TFBSs de los miembros de la familia LysR

La utilización del protocolo PProCoM para realizar un análisis sistemático en seis miembros representativos de la familia LysR, en organismos del *phyla* Gammaproteobacteria, nos permitió identificar con precisión los TFBSs y sus características comunes, mismas que resumimos en la **Tabla 2**. Dentro de las características generales del conjunto de datos, se observa que los genes que codifican para los TFs y sus TGs se transcriben en orientaciones divergentes, y sus regiones intergénicas presentan al menos dos de los tres motivos invertidos repetidos que se lograron identificar (IR1, IR2 e IR3). Con base en la información que resume la **Tabla 2,** se observa claramente la conservación de las secuencias en términos de sus longitudes y distancias intra e inter-motivo, así como similitudes en sus mecanismos de regulación molecular. Además de las secuencias específicas conservadas, identificar correctamente los TFBSs nos permitió hacer análisis comparativos para identificar similitudes entre las secuencias conservadas, las cuales se representan por una secuencia consenso 5'-CTATA-n9-TATAG-3', como se observa en la **Figura 14**. La secuencia consenso obtenida para los miembros de la familia LysR puede considerarse como una versión extendida del motivo “T-n11-A“ postulado originalmente por Goethals *et al*, como un consenso de la familia LysR basado en el análisis de los TFBSs de NodD en *Azorhizobium*, y otros miembros de la familia de tipo LysR (123). La conservación de la secuencia consenso en miembros de la familia LysR es muy relevante y puede explicarse, si consideramos que los nuevos genomas adquieren esos TFs con gran frecuencia, vía transferencia horizontal

de genes. Además de que los TFs compartieron un ancestro común, al evolucionar conservaron la similitud de las secuencias de sus motivos de unión y los mecanismos moleculares que regulan las respuestas a la transcripción de una variedad de estímulos y funciones, incluyendo entre ellos el metabolismo, detección de *quorum*, motilidad y virulencia, entre otros (75). De la **Figura 14** también se observa que el espacio común (intra-motivo) entre los monómeros que forman parte del motivo 5'-CTATA-3' y 5'-TATAG-3' es de 9 nt y hay variaciones de longitud mínimas; la variación más larga se observó para IlvY, con un espacio intra-motivo de 11 nt. En los análisis PProCoM se demostró que la secuencia intra-motivo es rica en nt A/T, estos nucleótidos le proveen flexibilidad al DNA, requerida para un adecuado reconocimiento TF-DNA. Existen pocas variaciones de la secuencia consenso 5'-CTATA-n9-TATAG-3' porque su conservación es requerida para el reconocimiento específico de un TF a su correspondiente TFBS (ver ***Figura 6*** y ***Figura* 13**). En la **Figura 14** también se incluyen ejemplos representativos de otros miembros de tipo LysR con TFBSs caracterizados experimentalmente. Esos TFBSs son consistentes con nuestro motivo consenso extendido para LysR. Por ejemplo en la región intergénica compartida entre los genes *catR-catBC* de *Pseudomonas putida*, el TFBS distal de *catBC* también conocido como sitio de unión del represor, tiene una secuencia palidrómica imperfecta 5'-tcA**g**A-n9-TATgG-3' (note el nucleótido g en negritas) que se asemeja a nuestro motivo LysR extendido 5'-CTATA-n9-TATAG-3'. Al hacer mutagénesis sitio dirigida cambiando la G por la T en el cuarto nt de este motivo, se crea una secuencia más parecida al consenso y como resultado un incremento de la unión de CatR y un incremento del nivel de transcripción del operón *catBC*. Sin embargo, las substituciones de la primer A por T en el mismo TFBS 5'-tc**A**gA-n9-TATgG-3´ (note la A negrita) hacen que esta secuencia sea menos parecida a la secuencia consenso, lo cual provoca un

decremento en el reconocimiento de CatR y disminuye la transcripción del operón *catBC* *(124)*.

A	B	C	D	E	F	G	H	I
FT	TG	longitud IR1	longitud intermotivo	longitud IR2	longitud intermotivo	longitud IR3	Posiciones Centrales IR1/IR2/IR3	Distancia entre motivos IR1-IR2/IR2-IR3
MetR	*metE*	(5-5-5)=15	7	(5-5-5)=15	–	NP	-63/-41/NP	21/–
GcvA	*gcvB*	(5-5-5)=15	7	(5-5-5)=15	–	NP	-65/-43/NP	21/–
OxyR	*oxyS*	(4-7-4)=15	7	(4-7-4)=15	Traslapado	(4-7-4)=15	-66/-45/-34	20/10
IlvY	*ilvC*	(5-5-5)=15	7	(5-5-5)=15	Traslapado	(5-5-5)=15	-65/-43/-34	21/8
CynR	*cynT*	(4-7-4)=15	7	(4-7-4)=15	Traslapado	(4-7-4)=15	-66/-44/-34	21/9
LysR	*lysA*	(5-5-5)=15	6	(5-5-5)=15	19	(5-5-5)=15	-65/-44/-10	20/33

Tabla 2 Relaciones reveladas con los análisis PProCoM, de acuerdo con las arquitecturas de los TFBSs de la familia de TFs de tipo LysR. Dentro de las características comunes es que los genes codificantes del TF y su correspondiente gene blanco se transcriben en orientaciones divergentes y sus regiones intergénicas presentan dos o tres motivos invertidos repetidos (IR1, IR2, e IR3). En esta tabla se resumen regiones intergénicas de los seis TFs analizados en nuestro estudio. Se observa una clara conservación de la longitud de los motivos y las distancias inter-motivo sugieren que existen similitudes en sus mecanismos regulatorios.

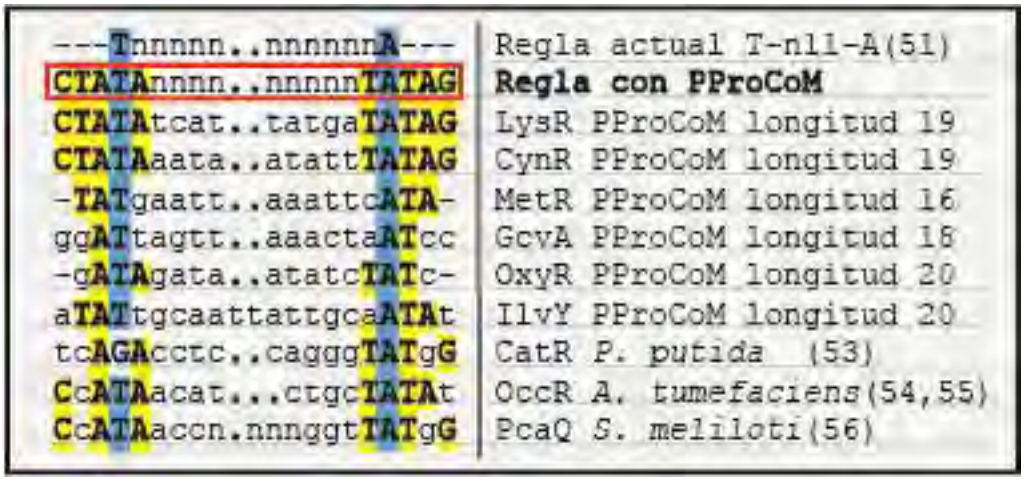

Figura 14 Secuencias consenso de los TFBSs para TFs miembros de la familia LysR. El motivo T-n11-A se propuso originalmente por Goethals *et al*, como la secuencia consenso reconocida por miembros de la familia de tipo LysR. Considerando nuestros análisis PProCoM de seis miembros representativos de esta familia en Gammaproteobacteria, definimos una nueva y extendida versión de este motivo: 5´-CTATA-n9-TATAG-3´. Adicionalmente, también se muestran ejemplos de la secuencia consenso de los TFBSs de otros miembros de la familia de tipo LysR que han sido verificados experimentalmente, y se incluyen los TFBSs distales de CatR de la bacteria *Pseudomonas Putida (124),* OccR de *Agrobacterium tumefaciens (125,126)* y PcaQ de *Sinorhizobium meliloti* (127). Los puntos dentro de las secuencias inter-motivos se utilizaron para alinear los nt conservados de las secuencias consenso.

Nuestro segundo ejemplo corresponde al sistema regulador OccR en *Agrobacterium tumefaciens*, donde se clonaron y caracterizaron regiones discretas de la región intergénica de *occR-occQ* utilizando *footprinting* con DNasa I y ensayos de movilidad "*gel shift*". En el trabajo de Wang *et al*. definieron cinco sitios de unión de OccR y sus afinidades relativas (125,126). Los sitios 1 y 2 formaron un IR localizado a -33 pb del TSS *occQ* y según nuestros análisis corresponde al sitio IR3 de los IRs identificados en este trabajo. Los sitios 4 y 5 formaron otro IR localizado a -64 pb del TSS de *occQ*, que corresponden también con el sitio IR1 (de mayor afinidad para OccR). Los sitios 3 y 2 corresponden con IR2 (*i.e.* el sitio con menor afinidad de los tres IRs del sistema) (125). Reemplazar IR3 con IR1 (el IR con mayor afinidad), trae como consecuencia una unión más fuerte de OccR y aumento de la represión del TG *occQ* (126). OccR se une

únicamente a IR2 (IR con menor afinidad) de modo cooperativo, y exclusivamente en presencia de octopina, el inductor del sistema (126).

A pesar de que ocurre el reemplazo de IR2 por IR1, (*i.e*. el IR con mayor afinidad) la unión de OccR a este sitio resulta ser parcialmente independiente del inductor octopina (126). Finalmente, nuestro tercer ejemplo corresponde a la región intergénica *pcaQ-pcaMNVWX* en *Sinorhizobium meliloti*. McLean *et al.* se basaron en experimentos de mutagénesis sitio dirigida, para proponer la unión de la proteína PcaQ sobre el sitio con secuencia 5'-ATAaccgggggatTAT-3' cuya posición central se localiza a -65.5 pb río arriba del gen estructural (**Figura 14**). Los cambios relevantes en los nucleótidos de la secuencia consenso, provocaron un decremento en la activación transcripcional del operón blanco *pcaMNVWX*, en la presencia de su inductor, debido a la alteración del reconocimiento del TF (127). Esas mutaciones involucraron cambios A-G en los nucleótidos subrayados de la secuencia 5'-ATA-n10-TAT-3', generando las secuencias 5'-GTA-n10-TAT-3', 5'-ATG- n10-TAT-3' y 5'-ATA- n10-TGT-3' (127).

8.1 Modelos Dinámicos de Regulación

Además de la descripción estática de los TFBSs identificados por nuestros análisis PProCoM, se pudieron elucidar modelos de regulación dinámicos para cada uno de los sistemas, basándonos en las características de los elementos del sistema regulatorio, los cuales se describen a continuación:

1) Las secuencias intergénicas de los sistemas reguladores del grupo uno (*metR-metE* y *gcvA-gcvB*) contenían dos motivos IR, mientras que los sistemas regulatorios del grupo dos (*oxyR-oxyS*, *ilvY-ilvC*, y *cynR-cynT*) y el grupo tres (*lysR-lysA*) contenían tres motivos

IR. En los tres grupos, los motivos IR muestran diferente conservación de la secuencia, y de esta forma, diferente afinidad. En el grupo uno, IR1 es el motivo más conservado, mientras que IR2 es el menos conservado. En los grupos dos y tres, IR1 y IR3 son los más conservados, mientras IR2 es el menos conservado.

2) Todos los TFs analizados (GcvA, MetR, OxyR, IlvY, CynR y LysR), adoptan dos conformaciones diferentes dependiendo de la presencia o ausencia de su inductor correspondiente: glicina, homocisteína, especies reactivas al oxígeno, acetolactato, cianato y ácido diaminopimélico, respectivamente.

3) Sin el sistema de inductores, para el caso del grupo uno, los TFs se unen como dímeros preferentemente al sitio IR1, mientras que el grupo dos y tres se unen a IR1 e IR3. De acuerdo con este reconocimiento, los ensayos de *footprinting* con miembros de la familia LysR muestran una región hipersensible de 50 pb río arriba del TSS de IlvY (112,113), CynR (117), OccR (125,128). Resultados similares se han observado en estudios con otros TFs reguladores de la familia LysR tales como ClcR (129), CatR (129) y PcaQ (127). En el caso de CynR, esta región hipersensible corresponde a la región donde el DNA se curva con la unión de CynR (117).

4) En presencia del sistema de inductores, el TF se une a DNA como dímero de dímeros de manera cooperativa. Sólo a través de esta unión cooperativa es que los TFs pueden reconocer al IR2 (el TFBS menos conservado). Este tipo de unión para TFs miembros de la familia LysR ha sido demostrada por ensayos de *footprinting* con DNasa I (86,98,112,113,125,128,130,131) y ensayos de mutagénesis de sitio dirigida (89,98,124-128,130,131). Como consecuencia de esta unión, las regiones hipersensibles de DNA localizadas alrededor de -50 pb río arriba del TSS decrecen significativamente.

Adicionalmente, se ha demostrado que alterar la distancia entre IR1 e IR2 reduce la unión cooperativa de los TFs (125-128,131).

5) Los TFs actúan como activadores o represores de la transcripción del propio TF o de los genes TG, dependiendo de la posición del IR al cual se unen.

6) Los motivos IR1 superponen la caja -10 río abajo de los promotores, esto suscita la auto-represión de la transcripción, cuando los TFs se unen a los sitios IR1.

7) Los motivos IR2 traslapan los promotores de los TFs, los cuales también se localizan río abajo de la caja -35 de los promotores de los TGs, por lo tanto, un TF unido a un IR2 reprime la transcripción del TF y activa la transcripción del TG.

8) En el caso del grupo dos, el motivo IR3 traslapa los promotores del TF y del TG, por lo que un TF se une a este sitio simultáneamente bloqueando la transcripción de los genes del TF y del TG. En el caso del grupo tres, el motivo IR3 solo traslapa el promotor TG, de acuerdo a esta unión, el TF que se une a este sitio bloquea exclusivamente la transcripción del TF.

9) Además de los resultados de regulación mencionados anteriormente, vale la pena mencionar que en el caso del grupo dos, los sitios IR2 e IR3 se superponen, por lo tanto, la unión de TFs a estos sitios es mutuamente excluyente.

En ausencia de los inductores del sistema, los TFs se unen preferentemente a IR3 dado que este sitio tiene una mayor conservación de la secuencia que IR2; sin embargo, en presencia de los inductores del sistema, los TFs se unirían de forma cooperativa como un dímero de dímeros a IR1 e IR2. En este caso, la unión de los TFs a IR2 tendría dos efectos positivos en la transcripción del TG; un efecto directo por su interacción con la RNA polimerasa, y un efecto indirecto por el bloqueo de la unión de los TFs a IR3, un evento que de otro modo reprimiría la transcripción del TF.

Gracias a los análisis con PProCoM y a la identificación de los motivos de secuencias de TFBSs, fue posible revelar modelos de regulación representativos de la familia de TFs de tipo LysR en Gammaproteobacteria, nuestro modelo de regulación también incluye los efectos de la unión del TF y la curvatura del DNA.

Estos efectos también se han reportado para varias TFs, como GcvA (86), MetR (99), OxyR (110), IlvY (112), CynR (116), LysR (133), CysB (134), CatR (124), ClcR (129), y OccR (128). En los sistemas de regulación mencionados anteriormente, basados en los análisis de *footprinting* con DNasa I se ha dicho, que en ausencia de inductores del sistema los TFs se unen a largas regiones de DNA. Por el contrario, en presencia de inductores, el área protegida del DNA en el análisis de la huella disminuye significativamente. Por ejemplo, en ausencia del inductor, OccR protege una región de aproximadamente 60 pares de bases, lo que resulta en el DNA con un ángulo de curvatura de 62 grados, que muestra las regiones hipersensibles alrededor de la posición -50 (135).

En la presencia del inductor, el ángulo se reduce a 46 grados, el acortamiento de la longitud del DNA protegido a 50 pb en el ensayo de *footprinting* con DNasa I, disminuye la región hipersensible (128,135). Toledano ***et al.*** propuso que esta reducción de la longitud en el DNA protegido es causada por el reordenamiento de los dímeros de dímeros de los TFs. En ausencia de inductores, los dímeros se unen a sitios distales, por ejemplo, IR1 e IR3. Una sola vuelta de la separación entre los dos dímeros provoca una curva en el DNA y como consecuencia la inhibición de la transcripción de las unidades de transcripción divergentes (110).

Un modelo regulador similar fue propuesto por Wang y Winans en el sistema regulador *occR* - *occQ* (128). Los resultados obtenidos con el protocolo PProCoM, que se resumen como nuestro 5'-CTATA- n9 -TATAG- 3' extendieron el motivo consenso para el TFBS de tipo LysR (**Figura 14**) y se esquematizan en nuestro modelo (**Figura 15**), son consistentes con las observaciones en curvatura de DNA disponible en la literatura. El uso potencial de PProCoM para identificar TFBSs puede inclusive utilizarse para otros sistemas reguladores diferentes a los de la familia LysR. Nuestro protocolo PProCoM se puede utilizar para identificar TFBSs de casi cualquier sistema de regulación bacteriana si se consideran las características de los TFs. Por ejemplo, además de las del sistema regulador LysR, actualmente se lleva a cabo un estudio para identificar los sitios de unión de los miembros de la familia AraC / XylS (136).

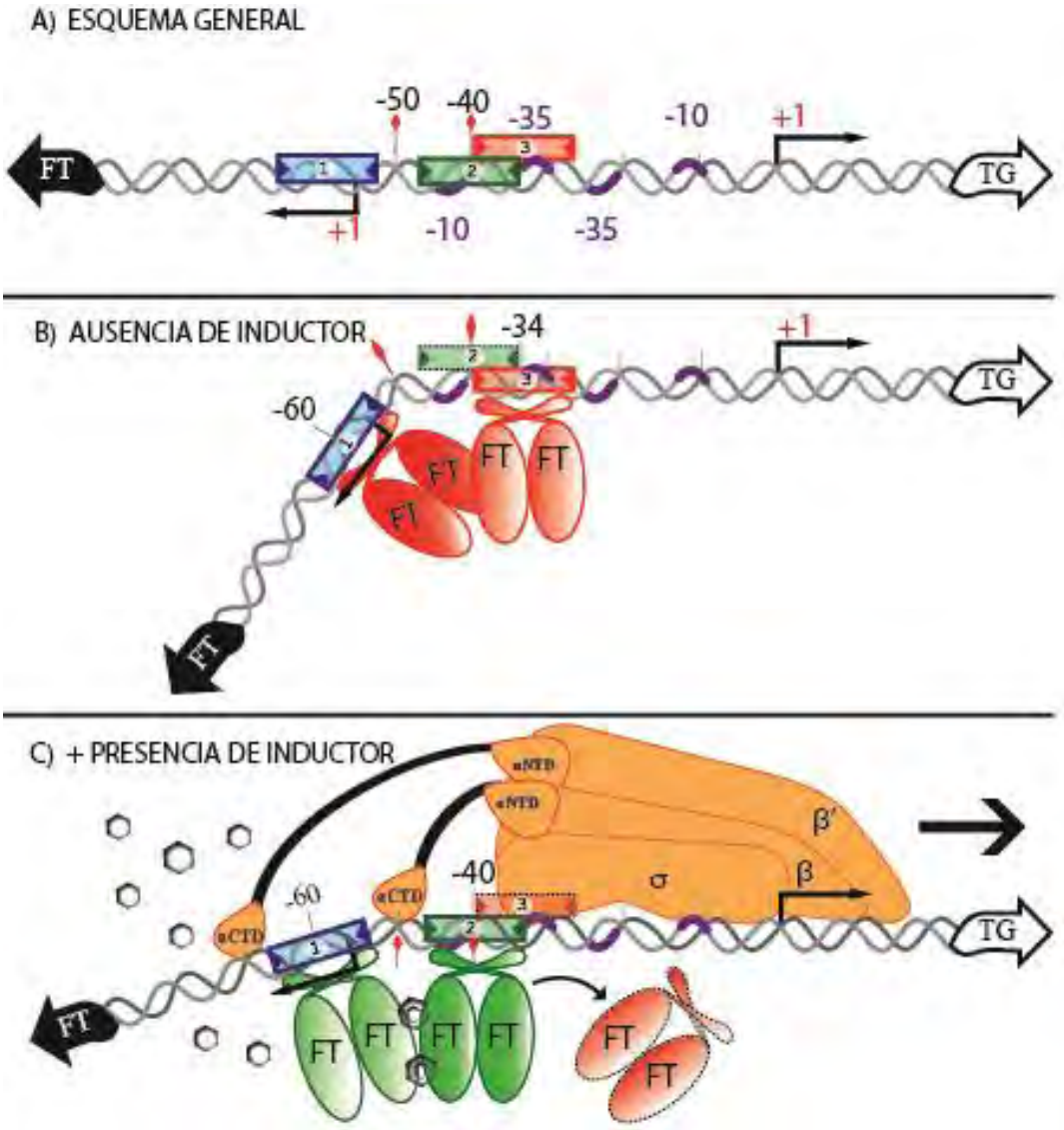

Figura 15 Modelo representativo general para TFs miembros de la familia LysR en Gammaproteobacterias, revelado por los análisis PProCoM. A) Arquitectura típica de las regiones regulatorias de los TFs en presencia de tres secuencias IR representadas en azul (IR1), verde (IR2) y rojo (IR3). Algunos sistemas reguladores, tales como aquellos de nuestro primer grupo de análisis, GcvA y MetR, pierden el tercer elemento IR. B) Dadas las afinidades de las secuencias IR1 e IR3 (observadas según la conservación de secuencia de los motivos) son mayores que aquella observada para el motivo IR2, en ausencia del inductor, el TF del sistema únicamente se une a los sitios IR1 e IR3. Las posiciones de IR1 e IR3 son críticas para la represión transcripcional de los sistemas divergentes. El motivo IR1 traslapa el promotor del TF, mientras IR3 traslapa los promotores de TF y TG. C) En presencia de un inductor en el sistema, el dímero del TF puede unir cooperativamente a un motivo IR menos conservado y menos afín en el sistema, *i.e.* IR2. Una característica muy significativa de varios sistemas reguladores en esta familia es que el IR2 traslapa parcialmente el IR3; por lo tanto, una primera consecuencia de la unión de del TF a IR2 es el desplazamiento estérico del TF que se unió a IR3, que dió como resultado la represión de la transcripción del TG. Adicionalmente a este efecto de des-represión, un segundo efecto de la unión del TF a IR2 es la activación transcripcional directa del TG debida a la posición del IR2 localizado inmediatamente arriba de la caja del promotor -35 del TG, donde el TF interactúa con la RNApol. Figura modificada de Wang L, et al.(128).

9 CONCLUSIONES

El protocolo PProCoM representa el análisis de un alineamiento de motivos consenso múltiples no convencional, con secuencias de longitud creciente, dispuestas de acuerdo con las coordenadas de los nucleótidos de referencia en la región intergénica del organismo referencia *E. coli* (*i.e* inciso C **Figura 6)**. Esta estrategia permite fusionar los motivos más representados (de E-valor significativo) con los motivos menos conservados, que desempeñan papeles importantes en los sistemas dinámicos de regulación de la transcripción. En general, los motivos menos conservados no han sido identificados o incluidos en los estudios anteriores, inclusive en los casos de análisis experimentales, como el análisis de *footprinting* con DNasa I. Nuestro análisis PProCoM de seis miembros de la familia de TFs de tipo LysR ha puesto en evidencia el gran interés de los motivos menos conservados en las regiones intergénicas de sus secuencias reguladoras. Este enfoque permite comprender la naturaleza homodimérica de los TFBS y proporciona una imagen más integrada y completa de sus procesos de regulación.

Consideramos que del mismo modo que existe una distancia genética en la que están ubicados los promotores (-10 y -35) respecto al sitio de inicio de la transcripción, también se podrían identificar las posibles métricas mediante las cuales operan los TFs para su posicionamiento sobre los TFBSs que reconocen.

Enfoques Computacionales para Identificación de TFBS

El presente trabajo tiene como objetivo desarrollar un protocolo computacional para predecir las ubicaciones exactas de los TFBSs en regiones promotoras de DNA, considerando también las zonas de mayor degeneración de los sitios, a través de interpretaciones que suman tanto las propiedades biológicas como las evidencias experimentales.

Esta tesis contribuye a resolver el problema de identificación de los sitios en los siguientes aspectos. 1) se propone una métrica de reconocimiento de los sitios, considerando las posiciones centrales de los TFBS. 2) incluye tanto la parte del análisis predictivo como la curación de la información biológica de cada sistema biológico a analizar.

El marco que proponemos para trabajar, demuestra que el protocolo tiene un buen potencial para identificar los elementos de regulación, y demuestra que la predicción de patrones sobre-representados no es suficiente, y que es necesario integrar el conocimiento biológico en conjunto para mejorar la predicción.

10 APÉNDICES

Apéndice 1. BASES DE DATOS PARA IDENTIFICACIÓN DE MOTIVOS

Nombre de la Base de Datos	Descripción
TRANSFAC	Una base de datos de matrices de peso por posición verificadas para TFs, disponible en: (http://www.gene-regulation.com)
JASPAR	Base de datos pública, de perfiles de matrices que describen una variedad de sitios de unión de TFs de vertebrados, plantas, insectos, nemátodos, hongos y urochordata. (http://jaspar.genereg.net)
ABS	Base de datos pública, de TFBS verificados experimentalmente. También se proporcionan las secuencias de promotores con las entradas originales de GenBank o RefSeq. Para cada sitio, están disponibles la posición, el motivo y la secuencia en la que el sitio está presente. (http://big.crg.cat/)
ECRBase	Una base de datos de las regiones conservadas evolutivamente, promotores, y TFBSs anotados, en genomas de vertebrados. La base de datos actualmente incluye los genomas de humano, macaco rhesus, perro, zarigüeya, rata, ratón, pollo, rana, pez cebra y Fugu. (http://ecrbase.dcode.org)

UniPROBE	Base de datos de datos de unión de proteínas a DNA (PWMs, secuencias, logos) para 574 proteínas no redundantes de organismos variables, incluyendo el procariota *Vibrio harveyi,* el parásito eucariótico de la malaria *Plasmodium falciparum*, el parásito *Apicomplexan Cryptosporidium parvum*, la levadura *Sacharomyces cerevisiae*, el gusano *Caenorhabditis elegans,* y humanos. (http://thebrain.bwh.harvard.edu/uniprobe/)
YEASTRACT	(**Yea**st **S**earch for **T**ranscriptional **R**egulators **A**nd **C**onsensus **T**racking) es un repositorio curado de más de 206000 asociaciones regulatorias entre TFs y genes blanco en el genoma de *Saccharomyces cerevisiae.* Incluye también la descripción de 326 sitios de unión a DNA específicos compartidos entre 113 FTs caracterizados. (http://www.yeastract.com/)
SwissRegulon	Base de datos de anotaciones de genoma completo sobre sitios de regulación en las regiones intergénicas de los genomas. Actualmente incluye anotaciones para 17 procariotes y 3 eucariotes (http://swissregulon.unibas.ch/fcgi/sr)
GenBank	Base de datos de secuencias de nucleótidos disponible públicamente actualmente alberga más de 363 millones de secuencias WGS. Disponible en https://www.ncbi.nlm.nih.gov/genbank/
RefSeq	Una colección no redundante de secuencias que representan genomas, transcritos y proteínas. La base de datos proporciona información para secuencias en regiones codificantes, dominios conservados, tRNAs, sitios marcados con secuencia (STS), variación, referencias, nombres de productos de genes y proteínas y referencias cruzadas de bases de datos. Actualmente, la base de datos contiene secuencias para 64,227 organismos que abarcan procariotas, eucariotas y virus. (https://www.ncbi.nlm.nih.gov/refseq/)

EPD	Base de Datos de Promotores Eucariotes (http://epd.vital-it.ch/)
CSHLmpd	Base de datos de promotores de mamíferos del laboratorio de Cold Spring Harbor. Alberga promotores de genomas humanos, ratones, y rata. (http://rulai.cshl.edu/cshlmpd/)
Repbase	Una base de datos de elementos eucarióticos repetitivos. http://www.girinst.org/repbase/

Apéndice 2. HERRAMIENTAS PARA IDENTIFICACIÓN DE MOTIVOS

Nombre de la herramienta	Descripción
PHAST	Es un paquete de software gratuito para genómica comparativa y evolutiva. Se utiliza principalmente para la identificación de nuevos elementos funcionales, incluyendo exones codificadores de proteínas y secuencias evolutivamente conservadas. (http://compgen.bscb.cornell.edu/phast/)
RepeatMasker	RepeatMasker es un programa que detecta secuencias de DNA para repeticiones intercaladas y secuencias de DNA de baja complejidad. (http://www.repeatmasker.org/)
CLUSTALW	Herramienta para múltiples alineamientos de secuencias. Disponible en http://www.genome.jp/tools/clustalw/
P-Match	Una herramienta para buscar probables sitios de unión a factor de transcripción en secuencias de DNA usando matrices de peso por posición disponibles.
MEME	Herramienta de identificación de motivos conocidos y desconocidos (http://meme.nbcr.net)
RSAT	Sitio Web que provee una serie de programas de cómputo modulares diseñados específicamente para la detección de señales reguladoras en secuencias no codificantes. (http://embnet.ccg.unam.mx/rsa-tools/)

11 BIBLIOGRAFÍA

1. Busby, S; Ebright R. Promoter Structure, Promoter Recognition, and Transcription Activation in Prokaryotes. *Cell*. 1994;79.

2. Molina N, van Nimwegen E. Scaling laws in functional genome content across prokaryotic clades and lifestyles. *Trends Genet.* 2009;25:243–247.

3. Martínez-Antonio A, Collado-Vides J. Identifying global regulators in transcriptional regulatory networks in bacteria. *Curr. Opin. Microbiol.* 2003;6:482–489.

4. Pérez-Rueda E, Collado-Vides J. The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12. *Nucleic Acids Res.* 2000;28:1838–1847.

5. Huerta AM, Salgado H, Thieffry D, Collado-Vides J. RegulonDB: a database on transcriptional regulation in Escherichia coli. *Nucleic Acids Res*. 1998;26:55–59. Available at: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=9399800.

6. Janky R, van Helden J. Evaluation of phylogenetic footprint discovery for predicting bacterial cis-regulatory elements and revealing their evolution. *BMC Bioinformatics*. 2008;9:37.

7. Tan K, McCue LA, Stormo GD. Making connections between novel transcription factors and their DNA motifs. *Genome Res.* 2005;15:312–320.

8. Tan K, Moreno-Hagelsieb G, Collado-Vides J, Stormo GD. A comparative genomics approach to prediction of new members of regulons. *Genome Res*. 2001;11:566–84. Available at: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.genome.org/cgi/content/full/11/4/566.

9. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 1988;203:439–455.

10. Keseler IM, Mackie A, Peralta-Gil M, et al. EcoCyc: Fusing model organism databases with systems biology. *Nucleic Acids Res.* 2013;41.

11. Novichkov PS, Brettin TS, Novichkova ES, et al. RegPrecise web services interface: Programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. *Nucleic Acids Res.* 2012;40.

12. Grote A, Klein J, Retter I, et al. PRODORIC (release 2009): A database and tool platform for the analysis of gene regulation in prokaryotes. *Nucleic Acids Res.* 2009;37.

13. Pérez AG, Angarica VE, Vasconcelos ATR, Collado-Vides J. Tractor_DB (version 2.0): A database of regulatory interactions in gamma-proteobacterial genomes. *Nucleic Acids Res.* 2007;35.

14. Oberto J. FITBAR: a web tool for the robust prediction of prokaryotic regulons. *BMC Bioinformatics*. 2010;11:554.

15. Medina-Rivera A, Abreu-Goodger C, Thomas-Chollier M, Salgado H, Collado-Vides J, Van Helden J. Theoretical and empirical quality assessment of transcription factor-binding motifs. *Nucleic Acids Res.* 2011;39:808–824.

16. DiRusso CC, Metzger AK, Heimert TL. Regulation of transcription of genes required for fatty acid transport and unsaturated fatty acid biosynthesis in Escherichia coli by FadR. *Mol. Microbiol.* 1993;7:311–322.

17. Barnard A, Wolfe A, Busby S. Regulation at complex bacterial promoters: How bacteria use different promoter organizations to produce different regulatory outcomes. *Curr. Opin. Microbiol.* 2004;7:102–108.

18. Rhodius VA, Busby SJ. Positive activation of gene expression. *Curr. Opin. Microbiol.* 1998;1:152–159.

19. Collado-Vides J, Magasanik B, Gralla JD. Control site location and transcriptional regulation in Escherichia coli. *Microbiol. Rev.* 1991;55:371–394.

20. Gralla JD. Activation and repression of E. coli promoters. *Curr. Opin. Genet. Dev.* 1996;6:526–530.

21. Gruber TM, Gross CA. Multiple sigma subunits and the partitioning of bacterial transcription space. *Annu. Rev. Microbiol.* 2003;57:441–466.

22. Borukhov S, Severinov K. Role of the RNA polymerase sigma subunit in transcription initiation. *Res. Microbiol.* 2002;153:557–562.

23. Hannenhalli S. Eukaryotic transcription factor binding sites--modeling and integrative search methods. *Bioinformatics*. 2008;24:1325–31. Available at: http://bioinformatics.oxfordjournals.org/content/24/11/1325.

24. Osada R, Zaslavsky E, Singh M. Comparative analysis of methods for representing and searching for transcription factor binding sites. *Bioinformatics*. 2004;20:3516–25. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15297295.

25. Li N, Tompa M. Analysis of computational approaches for motif discovery. *Algorithms Mol. Biol.* 2006;1:8. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1540429&tool=pmcentrez&rendertype=abstract.

26. Sandve GK, Drabløs F. A survey of motif discovery methods in an integrated framework. *Biol. Direct*. 2006;1:11.

27. Siddharthan R, Siggia ED, Van Nsmwegea E. PhyloGibbs: A gibbs sampling motif finder that incorporates phylogeny that incorporates phylogeny. *PLoS Comput. Biol.* 2005;1:0534–0556.

28. Dellaert F. The Expectation Maximization Algorithm. *Technology*. 2002;2:1–7. Available at: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.9.9735&rep=rep1&type=pdf.

29. Stormo GD. DNA binding sites: representation and discovery. *Bioinformatics*. 2000;16:16–23. Available at: http://www.ncbi.nlm.nih.gov/pubmed/10812473.

30. Hughes JD, Estep PW, Tavazoie S, Church GM. Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae. *J Mol Biol*. 2000;296:1205–1214. Available at: http://www.ncbi.nlm.nih.gov/cgi-bin/Entrez/referer?http://www.idealibrary.com/links/citation/0022-2836/296/1205.

31. Matys V, Fricke E, Geffers R, et al. TRANSFAC®: Transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.* 2003;31:374–378.

32. Sandelin A, Alkema W, Engström P, Wasserman WW, Lenhard B. JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.* 2004;32:91D–94. Available at: http://nar.oxfordjournals.org/content/32/suppl_1/D91.

33. Kel AE, Gößling E, Reuter I, Cheremushkin E, Kel-Margoulis O V., Wingender E. MATCH^{TM}: A tool for searching transcription factor binding sites in DNA sequences. *Nucleic Acids Res.* 2003;31:3576–3579.

34. Quandt K, Frech K, Karas H, Wingender E, Werner T. MatInd and MatInspector: new fast and versatile tools for detection of consensus matches in nucleotide sequence data. *Nucleic Acids Res.* 1995;23:4878–4884.

35. Pudimat R, Schukat-Talamazzini E-G, Backofen R. A multiple-feature framework for modelling and predicting transcription factor binding sites. *Bioinformatics*. 2005;21:3082–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15905283.

36. Fu W, Ray P, Xing EP. DISCOVER: A feature-based discriminative method for motif search in complex genomes. In: *Bioinformatics*.Vol 25.; 2009.

37. Mukherjee S, Berger MF, Jona G, et al. Rapid analysis of the DNA-binding specificities of transcription factors with DNA microarrays. *Nat. Genet.* 2004;36:1331–9. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15543148.

38. Harbison CT, Gordon DB, Lee TI, et al. Transcriptional regulatory code of a eukaryotic genome. *Nature*. 2004;431:99–104. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15343339.

39. Spellman PT, Sherlock G, Zhang MQ, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*. 1998;9:3273–97. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=25624&tool=pmcentrez&rendertype=abstract.

40. Kim SK, Lund J, Kiraly M, et al. A Gene Expression Map for Caenorhabditis elegans. *Science (80-.).* 2001;293:2087–2092. Available at: http://www.sciencemag.org/cgi/doi/10.1126/science.1061603.

41. Narlikar L, Gord??n R, Ohler U, Hartemink AJ. Informative priors based on transcription factor structural class improve de novo motif discovery. In: *Bioinformatics*.Vol 22.; 2006.

42. Tompa M, Li N, Bailey TL, et al. Assessing computational tools for the discovery of transcription factor binding sites. *Nat. Biotechnol.* 2005;23:137–44. Available at: http://www.ncbi.nlm.nih.gov/pubmed/15637633.

43. Wasserman WW, Sandelin A. Applied bioinformatics for the identification of regulatory elements. *Nat. Rev. Genet.* 2004;5:276–287.

44. Satija R, Pachter L, Hein J. Combining statistical alignment and phylogenetic footprinting to detect regulatory elements. *Bioinformatics*. 2008;24:1236–1242.

45. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT. Embryonic epsilon and gamma globin genes of a prosimian primate (Galago crassicaudatus). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J. Mol. Biol.* 1988;203:439–455.

46. Bailey TL. Discovering sequence motifs. *Methods Mol. Biol.* 2008;452:231–251.

47. Bailey TL, Bodén M, Whitington T, Machanick P. The value of position-specific priors in motif discovery using MEME. *BMC Bioinformatics*. 2010;11:179.

48. Bailey TL. Discovering novel sequence motifs with MEME. *Curr. Protoc. Bioinformatics*. 2002;Chapter 2:Unit 2.4.

49. Stormo GD, Hartzell GW. Identifying protein-binding sites from unaligned DNA fragments. *Proc. Natl. Acad. Sci. U. S. A.* 1989;86:1183–1187.

50. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF, Wootton JC. Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment. *Science*. 1993;262:208–214.

51. J. R. Sadler, M. S. Waterman TFS. Regulatory pattern identification in nucleic acid sequences. *Nucleic Acids Res.* 1983;11(7):2221–2232.

52. Schneider TD, Stormo GD, Gold L, Ehrenfeucht A. Information content of binding sites on nucleotide sequences. *J. Mol. Biol.* 1986;188:415–431.

53. Schneider TD, Stephens RM. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.* 1990;18:6097–6100. Available at: http://nar.oxfordjournals.org/cgi/reprint/18/20/6097\npapers2://publication/uuid/C799DFF0-112F-46C2-8077-BD70AEE1E3BD.

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

55. Boos W, Shuman H. Maltose/maltodextrin system of Escherichia coli: transport, metabolism, and regulation. *Microbiol. Mol. Biol. Rev.* 1998;62:204–229.

56. Zhou Y, Larson JD, Bottoms CA, et al. Structural Basis of the Transcriptional Regulation of the Proline Utilization Regulon by Multifunctional PutA. *J. Mol. Biol.* 2008;381:174–188.

57. Bonocora RP, Caignan G, Woodrell C, Werner MH, Hinton DM. A basic/hydrophobic cleft of the T4 activator MotA interacts with the C-terminus of E. coli??70 to activate middle gene transcription. *Mol. Microbiol.* 2008;69:331–343.

58. Ptashne M, Jeffrey A, Johnson AD, Maurer R, Meyer BJ, Pabo CO, Roberts TM SR. How the lambda repressor and cro work. *Cell*. 1980;19(1):1–11.

59. Spronk CAEM, Bonvin AMJJ, Radha PK, Melacini G, Boelens R, Kaptein R. The solution structure of Lac repressor headpiece 62 complexed to a symmetrical lac operator. *Structure*. 1999;7:1483–1492.

60. Alekshun MN, Levy SB, Mealy TR, Seaton BA, Head JF. The crystal structure of MarR, a regulator of multiple antibiotic resistance, at 2.3 A resolution. *Nat. Struct. Biol.* 2001;8:710–714.

61. Pelton JG, Kustu S, Wemmer DE. Solution structure of the DNA-binding domain of NtrC with three alanine substitutions. *J. Mol. Biol.* 1999;292:1095–1110.

62. Xu Y, Heath RJ, Li Z, Rock CO, White SW. The FadR.DNA complex. Transcriptional control of fatty acid metabolism in Escherichia coli. *J. Biol. Chem.* 2001;276:17373–17379.

63. Joachimiak A, Kelley RL, Gunsalus RP, Yanofsky C, Sigler PB. Purification and characterization of trp aporepressor. *Proc. Natl. Acad. Sci. U. S. A.* 1983;80:668–672.

64. Bell CE, Lewis M. The Lac repressor: A second generation of structural and functional studies. *Curr. Opin. Struct. Biol.* 2001;11:19–25.

65. Brinkman AB, Ettema TJG, De Vos WM, Van Der Oost J. The Lrp family of transcriptional regulators. *Mol. Microbiol.* 2003;48:287–294.

66. Lee YY, Barker CS, Matsumura P, Belas R. Refining the binding of the Escherichia coli flagellar master regulator, FlhD 4C 2, on a base-specific level. *J. Bacteriol.* 2011;193:4057–4068.

67. Castanié-Cornet MP, Cam K, Bastiat B, Cros A, Bordes P, Gutierrez C. Acid stress response in Escherichia coli: Mechanism of regulation of gadA transcription by RcsB and GadE. *Nucleic Acids Res.* 2010;38:3546–3554.

68. Stratmann T, Pul Ü, Wurm R, Wagner R, Schnetz K. RcsB-BglJ activates the Escherichia coli leuO gene, encoding an H-NS antagonist and pleiotropic regulator of virulence determinants. *Mol. Microbiol.* 2012;83:1109–1123.

69. Salgado H, Peralta-Gil M, Gama-Castro S, et al. RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013;41.

70. Busby S, Ebright RH. Transcription activation by catabolite activator protein (CAP). *J. Mol. Biol.* 1999;293:199–213.

71. Lawson CL, Swigon D, Murakami KS, Darst SA, Berman HM, Ebright RH. Catabolite activator protein: DNA binding and transcription activation. *Curr. Opin. Struct. Biol.* 2004;14:10–20.

72. Savery NJ, Lloyd GS, Kainz M, et al. Transcription activation at class II CRP-dependent promoters: Identification of determinants in the C-terminal domain of the RNA polymerase ?? subunit. *EMBO J.* 1998;17:3439–3447.

73. Browning DF, Busby SJ. The regulation of bacterial transcription initiation. *Nat. Rev. Microbiol.* 2004;2:57–65.

74. Tam R, Saier MH. Structural, functional, and evolutionary relationships among extracellular solute-binding receptors of bacteria. *Microbiol. Rev.* 1993;57:320–346.

75. Maddocks SE, Oyston PCF. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. *Microbiology*. 2008;154:3609–3623.

76. Felsenstein J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* 1981;17:368–376.

77. Ciccarelli F, Doerks T, Mering C Von. Toward automatic reconstruction of a highly resolved tree of life. *Science (80-.).* 2006;(May).

78. Tatusov RL, Fedorova ND, Jackson JD, et al. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*. 2003;4:41.

79. Taboada B, Verde C, Merino E. High accuracy operon prediction method based on STRING database scores. *Nucleic Acids Res.* 2010;38.

80. Jensen LJ, Kuhn M, Stark M, et al. STRING 8 - A global view on proteins and their functional interactions in 630 organisms. *Nucleic Acids Res.* 2009;37.

81. Bailey TL, Boden M, Buske FA, et al. MEME Suite: Tools for motif discovery and searching. *Nucleic Acids Res.* 2009;37.

82. Bailey TL. Discovering sequence motifs. *Methods Mol. Biol.* 2008;452:231–251.

83. Zhang M, Leong HW. Bidirectional best hit r-window gene clusters. *BMC Bioinformatics*. 2010;11 Suppl 1:S63.

84. Altschul SF, Madden TL, Schäffer a a, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 1997;25(17):3389–402.

85. Bailey TL. Discovering novel sequence motifs with MEME. *Curr. Protoc. Bioinformatics*. 2002;Chapter 2:Unit 2.4.

86. Wilson RL, Urbanowski ML, Stauffer G V. DNA binding sites of the LysR-type regulator GcvA in the gcv and gcvA control regions of Escherichia coli. *J. Bacteriol.* 1995;177:4940–4946.

87. Wilson RL, Steiert PS, Stauffer G V. Positive regulation of the Escherichia coli glycine cleavage enzyme system. *J. Bacteriol.* 1993;175:902–904.

88. Stauffer LT, Stauffer G V. GcvA interacts with both the ?? and ?? subunits of RNA polymerase to activate the Escherichia coli gcvB gene and the gcvTHP operon. *FEMS Microbiol. Lett.* 2005;242:333–338.

89. Jourdan AD, Stauffer G V. Genetic analysis of the GcvA binding site in the gcvA control region. *Microbiology*. 1999;145:2153–2162.

90. Wonderling LD, Urbanowski ML, Stauffer G V. GcvA binding site 1 in the gcvTHP promoter of Escherichia coli is required for GcvA-mediated repression but not for GcvA-mediated activation. *Microbiology*. 2000;146:2909–2918.

91. Novichkov PS, Brettin TS, Novichkova ES, et al. RegPrecise web services interface: Programmatic access to the transcriptional regulatory interactions in bacteria reconstructed by comparative genomics. *Nucleic Acids Res.* 2012;40.

92. Cai X-Y, Maxon ME, Redfield B, Glasst R, Brot N, Weissbach H. Methionine synthesis in Escherichia coli: Effect of the MetR protein on metE and metH expression. *Biochemistry*. 1989;86:4407–4411.

93. Weissbach H, Brot N. Regulation of methionine synthesis in Escherichia coli. *Mol. Microbiol.* 1991;5:1593–1597.

94. Flatley J, Barrett J, Pullan ST, Hughes MN, Green J, Poolet RK. Transcriptional responses of Escherichia coli to S-nitrosoglutathione under defined chemostat conditions reveal major changes in methionine biosynthesis. *J. Biol. Chem.* 2005;280:10065–10072.

95. Membrillo-Hernández J, Coopamah MD, Channa A, Hughes MN PR. A novel mechanism for upregulation of the Escherichia coli K-12 hmp (flavohaemoglobin) gene by the “NO releaser”, S-nitrosoglutathione: nitrosation of homocysteine and modulation of MetR binding to the glyA-hmp intergenic region. *Mol. Microbiol.* 1998;29:1101–1112.

96. Maxon ME, Redfield B, Cai XY, et al. Regulation of methionine synthesis in Escherichia coli: effect of the MetR protein on the expression of the metE and metR genes. *Proc. Natl. Acad. Sci. U. S. A.* 1989;86:85–89.

97. Jafri S, Urbanowski ML, Stauffer G V. A mutation in the rpoA gene encoding the ?? subunit of RNA polymerase that affects metE-metR transcription in Escherichia coli. *J. Bacteriol.* 1995;177:524–529.

98. Wu WF, Urbanowski ML, Stauffer G V. Characterization of a second MetR-binding site in the metE metR regulatory region of Salmonella typhimurium. *J. Bacteriol.* 1995;177:1834–1839.

99. Lorenz E, Stauffer G V. Characterization of the MetR binding sites for the glyA gene of Escherichia coli. *J. Bacteriol.* 1995;177:4113–4120.

100. Lorenz E, Stauffer G V. Cooperative MetR binding in the Escherichia coli glyA control region. *FEMS Microbiol. Lett.* 1996;137:147–152.

101. MA S. Molecular biology of the LysR family of transcriptional regulators. Annu Rev Microbiol. *Annu Rev Microbiol*. 1993;47:597–626.

102. Harari O, del Val C, Romero-Zaliz R, et al. Identifying promoter features of co-regulated genes with similar network motifs. *BMC Bioinformatics*. 2009;10 Suppl 4:S1.

103. Collado-Vides J, Salgado H, Morett E, Gama-Castro S, Jiménez-Jacinto V M-FI. Bioinformatics resources for the study of gene regulation in bacteria. *J Bacteriol*. 2009;191:23–31.

104. Anjem A, Varghese S, Imlay JA. Manganese import is a key element of the OxyR response to hydrogen peroxide in Escherichia coli. *Mol. Microbiol.* 2009;72:844–858.

105. Anjem A, Varghese S, Imlay JA. Manganese import is a key element of the OxyR response to hydrogen peroxide in Escherichia coli. *Mol. Microbiol.* 2009;72:844–858.

106. Storz G, Tartaglia LA, Ames BN. The OxyR regulon. In: *Antonie van Leeuwenhoek, International Journal of General and Molecular Microbiology*.Vol 58.; 1990:157–161.

107. Mongkolsuk S, Helmann JD. Regulation of inducible peroxide stress responses. *Mol. Microbiol.* 2002;45:9–15.

108. Zheng M, Wang X, Templeton LJ, Smulski DR, LaRossa RA, Storz G. DNA microarray-mediated transcriptional profiling of the Escherichia coli response to hydrogen peroxide. *J. Bacteriol.* 2001;183:4562–4570.

109. Zheng M, Aslund F, Storz G. Activation of the OxyR transcription factor by reversible disulfide bond formation. *Science*. 1998;279:1718–1721.

110. Toledano MB, Kullik I, Trinh F, Baird PT, Schneider TD, Storz G. Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: A mechanism for differential promoter selection. *Cell*. 1994;78:897–909.

111. Tartaglia LA, Gimeno CJ, Storz G, Ames BN. Multidegenerate DNA recognition by the OxyR transcriptional regulator. *J. Biol. Chem.* 1992;267:2038–2045.

112. Rhee KY, Senear DF, Hatfield GW. Activation of gene expression by a ligand-induced conformational change of a protein-DNA complex. *J. Biol. Chem.* 1998;273:11257–11266.

113. Wek RC, Hatfield GW. Transcriptional activation at adjacent operators in the divergent-overlapping ilvY and ilvC promoters of Escherichia coli. *J. Mol. Biol.* 1988;203:643–663.

114. Sung YC, Fuchs JA. Characterization of the cyn operon in Escherichia coli K12. *J. Biol. Chem.* 1988;263:14769–14775.

115. Lamblin AFJ, Fuchs JA. Expression and purification of the CynR regulatory gene product: CynR is a DNA-binding protein. *J. Bacteriol.* 1993;175:7990–7999.

116. Lamblin AF, Fuchs JA. Functional analysis of the Escherichia coli K-12 cyn operon transcriptional regulation. *J. Bacteriol.* 1994;176:6613–6622.

117. Lamblin AF, Fuchs JA. Functional analysis of the Escherichia coli K-12 cyn operon transcriptional regulation. *J. Bacteriol.* 1994;176:6613–6622.

118. Stragier P, Richaud F, Borne F, Patte JC. Regulation of diaminopimelate decarboxylase synthesis in Escherichia coli. I. Identification of a lysR gene encoding an activator of the lysA gene. *J. Mol. Biol.* 1983;168:307–320.

119. Stragier P, Danos O, Patte JC. Regulation of diaminopimelate decarboxylase synthesis in Escherichia coli. II. Nucleotide sequence of the lysA gene and its regulatory region. *J. Mol. Biol.* 1983;168:321–331.

120. Stragier P, Patte JC. Regulation of diaminopimelate decarboxylase synthesis in Escherichia coli. III. nucleotide sequence and regulation of the lysR gene. *J. Mol. Biol.* 1983;168:333–350.

121. Salgado H, Peralta-Gil M, Gama-Castro S, et al. RegulonDB v8.0: Omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res.* 2013;41.

122. Huerta AM, Collado-Vides J. Sigma70 promoters in Escherichia coli: Specific transcription in dense regions of overlapping promoter-like signals. *J. Mol. Biol.* 2003;333:261–278.

123. Goethals K, Van Montagu M, Holsters M. Conserved motifs in a divergent nod box of Azorhizobium caulinodans ORS571 reveal a common structure in promoters regulated by LysR-type proteins. *Proc. Natl. Acad. Sci. U. S. A.* 1992;89:1646–1650.

124. Parsek MR, Ye RW, Pun P, Chakrabarty AM. Critical nucleotides in the interaction of a LysR-type regulator with its target promoter region: catBC promoter activation by CatR. *J. Biol. Chem.* 1994;269:11279–11284.

125. Wang L, Winans SC. The sixty nucleotide OccR operator contains a subsite essential and sufficient for OccR binding and a second subsite required for ligand-responsive DNA bending. *J. Mol. Biol.* 1995;253:691–702.

126. Akakura R, Winans SC. Mutations in the occQ operator that decrease OccR-induced DNA bending do not cause constitutive promoter activity. *J. Biol. Chem.* 2002;277:15773–15780.

127. MacLean AM, Haerty W, Brian Golding G, Finan TM. The LysR-type PcaQ protein regulates expression of a protocatechuate-inducible ABC-type transport system in Sinorhizobium meliloti. *Microbiology*. 2011;157:2522–2533.

128. Wang L, Winans SC. High angle and ligand-induced low angle DNA bends incited by OccR lie in the same plane with OccR bound to the interior angle. *J. Mol. Biol.* 1995;253:32–8. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7473714.

129. Parsek MR, McFall SM, Shinabarger DL, Chakrabarty a M. Interaction of two LysR-type regulatory proteins CatR and ClcR with heterologous promoters: functional and evolutionary implications. *Proc. Natl. Acad. Sci. U. S. A.* 1994;91:12393–7. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=45444&tool=pmcentrez&rendertype=abstract.

130. Urbanowski ML, Stauffer G V. Genetic and biochemical analysis of the MetR activator-binding site in the metE metR control region of Salmonella typhimurium. *J. Bacteriol.* 1989;171:5620–5629.

131. Toledano MB, Kullik I, Trinh F, Baird PT, Schneider TD, Storz G. Redox-dependent shift of OxyR-DNA contacts along an extended DNA-binding site: A mechanism for differential promoter selection. *Cell*. 1994;78:897–909.

132. Crooks GE, Hon G, Chandonia JM, Brenner SE WebLogo: A sequence logo generator, *Genome Research*, 14:1188-1190, (2004)

133. Maddocks SE, Oyston PC. Structure and function of the LysR-type transcriptional regulator (LTTR) family proteins. Microbiology. 2008;154:3609–23.

134. Hryniewicz MM, Kredich NM. Stoichiometry of binding of CysB to the cysJIH, cysK, and cysP promoter regions of salmonella typhimurium. J Bacteriol. 1994;176:3673–82.

135. McFall SM, Klem TJ, Fujita N, Ishihama A, Chakrabarty AM. DNase I footprinting, DNA bending and in vitro transcription analyses of ClcR and CatR interactions with the clcABD promoter: evidence of a conserved transcriptional activation mechanism. Mol Microbiol. 1997;24:965–76.

136. Gallegos MT, Schleif R, Bairoch A, Hofmann K, Ramos JL. Arac/XylS family of transcriptional regulators. Microbiol Mol Biol Rev. 1997;61:393–410.

137. Kumamoto87: Kumamoto AA, Miller WG, Gunsalus RP (1987). “Escherichia coli tryptophan repressor binds multiple sites within the aroH and trp operators”, Genes Dev 1(6);556-64. PMID: 3315853

138. Gruber, T. M.; Gross, C. A. (2003). "Multiple Sigma Subunits and the Partitioning of Bacterial Transcription Space". *Annual Review of Microbiology*. **57**: 441–466

139. Durbin, Richard; Sean R. Eddy; Graeme Mitchison (1998). *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*.

140 *Busby S., Ebright RH. (2001). "Transcription activation by catabolite activator protein (CAP)". J. Mol. Biol. **293**: 199–213.*

141. Slonczewski, Joan, and John Watkins. Foster. Microbiology: An Evolving Science. New York: W.W. Norton &, 2009.

142. Ray P, et al. Csmet: comparative genomic motif detection via multi-resolution phylogenetic shadowing. PLoS Comput. Biol. 2008.

143. Symmetric Protein Assemblies Produce Cooperative Allosteric Transitions Molecular Biology of the Cell. 4th edition Alberts B, Johnson A, Lewis J, et al. New York: Garland Science; 2002.