11281

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

INSTITUTO DE CIENCIAS BIOMÉDICAS
CENTRO DE INVESTIGACIÓN SOBRE LA FIJACIÓN DEL NITRÓGENO

DOCTORADO EN CIENCIAS BIOMEDICAS

T E S I S

ANÁLISIS Y PREDICCIÓN DE SECUENCIAS DE PROMOTORES BACTERIANOS DEPENDIENDO DE LOS FACTORES SIGMA Y ALPHA Y DE OTROS FACTORES REGULADORES DIFERENTES A LA RNAP.

PRESENTADA POR

Araceli Huerta Moreno

MARZO 2004

ANÁLISIS Y PREDICCIÓN DE SECUENCIAS DE PROMOTORES BACTERIANOS DEPENDIENDO DE LOS FACTORES SIGMA Y ALPHA Y DE OTROS FACTORES REGULADORES DIFERENTES A LA RNAP.

POR

ARACELI HUERTA MORENO

TESIS

Realizada en el Programa de Genómica Computacional del Centro de Investigación sobre la Fijación del Nitrógeno bajo la supervisión del Dr. Pedro Julio Collado Vides.

Y

Presentada a la División de Estudios de Posgrado.
Este trabajo es Requisito Parcial para obtener el Título de Doctorado en Ciencias del programa en Ciencias Biomédicas.

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO.

MARZO 2004

ANÁLISIS Y PREDICCIÓN DE SECUENCIAS DE PROMOTORES BACTERIANOS DEPENDIENDO DE LOS FACTORES SIGMA Y ALPHA Y DE OTROS FACTORES REGULADORES DIFERENTES A LA RNAP.

Tesis de doctorado en Ciencias Biomédicas.

por

Araceli Huerta Moreno

Resumen.

En este trabajo se presenta un análisis computacional que muestra que los promotores reconocidos por la sigma endógena (σ70) de *Escherichia coli* se encuentran inmersos en regiones con altas densidades de señales tipo promotor. Fenómeno que encontramos presente en una gran variedad de genomas bacterianos. Utilizando 599 promotores de *E. coli* mapeados experimentalmente se construyeron y evaluaron mas de 200 matrices de peso. Las matrices que generaron los mejores modelos estadísticos no correspondieron al modelo canónico del promotor sigma70 reportado en la literatura. Sin embargo, las matrices correspondientes a tal modelo se desempeñaron mejor como herramientas de reconocimiento de promotores funcionales. Al utilizar estas matrices para buscar señales tipo promotor en regiones intergenicas de 250 bp se encontraron en promedio 38 señales promotoras. En mas del 50% de las regiones analizadas, el promotor mapeado no tuvo la mejor calificación de homología al consenso de sigma70. Lo que se observó es que los promotores funcionales existen dentro de regiones con una alta densidad de señales tipo promotor superpuestas. Se evaluaron diferentes estrategias para identificar al promotor funcional y se fue capaz de reconocer correctamente el 86% de los promotores mapeados generando un promedio de 4.7 promotores putativos por región, de los cuales el 3.7 en promedio existen in clusters formando una serie de sitios de pegado potencialmente competentes para el pegado de la RNAP. Esta densidad de señales se encuentra presente solamente en las regiones intergenicas entre genes divergentes. Estos resultados son consistentes con evidencia experimental que muestra la existencia de múltiples promotores superpuestos que se vuelven funcionales bajo ciertas condiciones experimentales. Esta densidad es formada probablemente por vestigios de promotores que quedaron como un resultado del proceso de la evolución. Nosotros sugerimos que los reguladores transcripcionales juegan un papel muy importante para mantener esas señales latentes suprimidas.

Palabras clave: promotores sigma70; *Escherichia coli*; predicciones computacionales; señales sobrepuestas o sobrelapadas.

Director de Tesis: Dr. Pedro Julio Collado Vides.
Título: Director del Programa de Genómica Computacional en el Centro de Investigación sobre la Fijación del Nitrógeno, UNAM Campus Morelos, México.

ANALYSIS AND PREDICTION OF BACTERIAL PROMOTER SEQUENCES DEPENDING ON SIGMA AND ALPHA FACTORS AND OTHERS REGULATORY FACTORS THAN RNAP

PhD Thesis in Biomedical Science.

by

Araceli Huerta Moreno

Abstract.

We present here a computational analysis showing that sigma70 housekeeping promoters are located within zones with high densities of promoter-like signals in *Escherichia coli* and we introduce strategies that allow for the correct computer prediction of sigma70 promoters. Based on 599 experimentally verified promoters of *E. coli* K12, we generated and evaluated more than 200 weight matrices optimizing different criteria to obtain the best recognition matrices. The alignments generating the best statistical models did not fully correspond with the canonical sigma70 model. However, matrices that correspond to such canonical model performed better as tools for prediction. We tested the predictive capacity of these matrices on 250bps long regions upstream of gene starts, where 90% of the known promoters occur. The computational matrix models generated an average of 38 promoter-like signals within each 250 bps region. In more than 50% of the cases, the true promoter does not have the best score within the region. We observed, in fact, that real promoters occur mostly within regions with high densities of overlapping putative promoters. We evaluated several strategies to identify promoters and we were able to correctly identify 86% true promoters generating an average of 4.7 putative promoters per region as output, of which 3.7 on average exist in clusters, as a series of overlapping potentially competing RNAP binding sites. This high signal density is mainly found within regions upstream of genes, contrasting with coding regions and regions located between convergently transcribed genes. These results are consistent with experimental evidence that show the existence of multiple overlapping promoter sites that become functional under particular conditions. This density is probably the consequence of a rich number of vestiges of promoters in evolution. We suggest that transcriptional regulators as well as other functional promoters play an important role in keeping these latent signals suppressed.

Keywords: sigma70 promoters; *Escherichia coli*; computational predictions; overlapping signals.

Research Supervisor: Dr. Pedro Julio Collado Vides.
Title: Head of the Computational Genomics Program at the Nitrogen Fixation Research Center, UNAM Campus Morelos, México.

AGRADECIMIENTOS.

A mis padres, Meche y Benito:

Por todas las cosas lindas que me han dado, por los buenos y los malos momentos, ahora sé que siempre me dieron lo mejor de si mismos. Gracias por su paciencia, confianza, y comprensión a mis fallas.

A mis hermanos, sobrinos:

Lupita, Merce, Hugo, Erika, y Omar: gracias por red de cariño y unión que forman conmigo y que ha sobrevivido a pesar de las tormentas. A ellos les agradezco esas luces, con las que han alegrado a nuestra familia: Jazar, Isaac, Offo, Jetzemani, Quetzalli, Erick y Grace. Todo mi amor para ustedes

A mi segunda familia:

Lucila, Mariela, Sadot, Jesús Adrian, Eduardo, y Erick con su apoyo este logro ha sido posible, gracias por sus cuidados y preocupaciones.

A mis amigos:

Por compartir conmigo esta aventura llamada Nitróge no, en especial quiero agradecer a Laura Esmirna, Magda, Ilenys, Victoria, Ernesto, Espe, Rosa María, Gabo, Vero, Romualdo, Delfino, Fabis y Edgar por su cariño, Heli por su paciencia y cariño, Victor por su inmensa *macrotimia* y apoyo, en especial agradezco a Conchita todo el cariño y apoyo, por ser un angelito de la guarda. I thank Rob Rohde-Szudy for his his daily conversation which made more enjoyable this time.

En especial quiero agradecer a las personas que marcaron un cambio en mi vida con su ministerio del perdón y la sanidad del corazón: Yalu, Hortensia Saldaña, Alma, Hortensia Arnaud, Tere, Tete, Alice (en memoria). Siempre las llevaré en mi corazón.

A mis maestro y tutores:

Dr. Carlos Gómez-Mont Avalos y Dr. Jaime Rangel Mondragón quienes materialmente me empujaron dentro del quehacer académico al ver en mi cualidades propicias para esta tarea.

Dr. Julio Collado Vides por las lecciones de vida que me enseñaron que el ser humano es una mezcla de cualidades y defectos en la que estos últimos no pueden opacar a los primeros. Gracias a su actitud visionaria y apoyo pude incursionar exitosamente en un área de la ciencia, ajena a mi formación previa, la Biología. El reto y la satisfacción que significa el entender y abstraer los mecanismos de la biología molecular y genómica hace que la vida en su laboratorio sea una aventura difícil de olvidar.

Agradezco a todos los catedráticos e investigadores de la UNAM el conocimiento sembrado en mi persona y la paciencia con la que enseñaron a una persona educada en las ciencias computacionales

Finalmente a Ti Padre por que has cumplido Tu promesa en mi vida:

cuando las montañas se han movido y algunos de los cielos han caído, Tu lealtad y amor han permanecido conmigo como la roca inamovible en la que me sustentas.

A los bastiones de la casa:
Meche, Lucila, Maye, Omar, Erika y Sadot.
Con todo mi cariño.

ANÁLISIS Y PREDICCIÓN DE SECUENCIAS DE PROMOTORES BACTERIANOS DEPENDIENDO DE LOS FACTORES SIGMA Y ALPHA Y DE OTROS FACTORES REGULADORES DIFERENTES A LA RNAP.

Llámame, y yo te responderé, y te enseñaré cosas grandes y secretas que tú no conoces.

- Jer. 33:3

ÍNDICE

CAPÍTULO I. INTRODUCCIÓN.

I.1 QUÉ ES UN PROMOTOR.

Los mecanismos que regulan la expresión genética en organismos bacterianos les permiten adaptarse fácilmente a los cambios en su medio ambiente. Los elementos que desempeñan un papel en estos mecanismos son el DNA, y productos difusibles como el RNA y las proteínas; estas últimas se unen a sitios específicos en el DNA [Raibaud_1984, Walker_1987, Gralla_1990, Snyder_1997].

La expresión genética en organismos procariontes está regulada básicamente al nivel de la transcripción, que es el proceso de síntesis de RNA usando el DNA como templado y que es llevado a cabo por la holoenzima RNA-polimerasa (RNApol, o RNAP), ver Figura I.1.

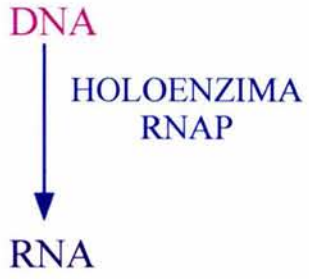


Figura I.1. El DNA es copiado a RNA como resultado del proceso de la transcripción.

La etapa más importante de la regulación de la expresión genética es el *inicio de la transcripción.* Este proceso se realiza por la asociación o pegado específico de la RNApol a un segmento de secuencia particular que está presente en el DNA -el *promotor*. Los *promotores* están localizados en las regiones previas (río arriba, o upstream) a la de los genes transcritos [Snyder_1997]. La frecuencia con la que se inicia la transcripción depende de la afinidad de la RNAP por el promotor; se sabe que esta frecuencia puede incrementarse hasta 100 veces dependiendo de las pares de bases (bp por sus siglas en Inglés), o secuencia de nucleótidos, que conforman al promotor [Mulligan_1984, Weller_1994].

Los promotores de *Escherichia coli* son de las señales reguladoras en el DNA mejor estudiadas. La RNAP de *E. coli* es un complejo multimérico compuesto por las subunidades: α_2,β,β', ω y σ [Chan_1994, Murakami_2002, Murakami_2002b]. Se ha descrito que la subunidad σ (factor) es quien le confiere la especificidad a la RNAP por el promotor en el DNA [Burgess_1969]. Al factor sigma que participa en la transcripción de la mayoría de los genes que componen un genoma, se le conoce como sigma endógeno, primario o "house-keeping". El factor σ^{70} es el sigma primario responsable del reconocimiento de la mayoría de los promotores mRNA, rRNA, y tRNA en muchos organismos bacterianos [Dombroski_1992, Hertz_1996]. *Escherichia coli* puede hacer uso de otros 6 factores sigmas secundarios para el reconocimiento de promotores en el DNA. La mayoría de los factores sigma bacterianos forman una familia conocida como la familia σ70 [Paget_2003]. Los factores sigmas pertenecientes a esta familia reconocen promotores en el DNA con una estructura muy similar a la reconocida por el factor σ70 de *E. coli*, la cual se describe a continuación.

Cuando la RNAP reconoce un promotor se pega a una de las caras del DNA de doble cadena de la región promotora protegiendo una región de 75 a 80 bp aproximadamente [Mooney_1998].

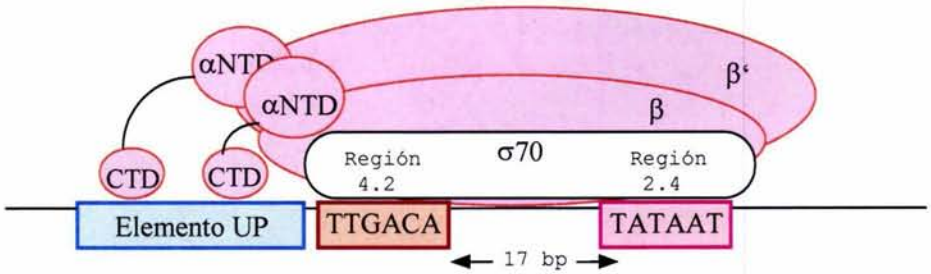


Figura I.2. Modelo del reconocimiento de los promotores σ70 de *E. coli* [Gourse_2000]. La región correspondiente al motivo TTGACA es conocida como la **caja -35**, y la región de TATAAT se le llama la **caja -10**. Además de estas cajas existe un tercer elemento de reconocimiento, el **elemento UP**, caracterizado hasta ahora en los promotores ribosomales. Las regiones 4.2 y 2.4 del factor σ70 reconocen respectivamente las cajas –35 y –10 del promotor.

Los análisis funcionales y estadísticos sobre las secuencias promotoras reconocidas por la RNAP-σ^{70} de *E. coli* han hecho posible la identificación de dos hexámeros consenso, **TTGACA** y **TATAAT**, localizados respectivamente a ~35 y 10 bp arriba del punto de inicio de la transcripción, y también han identificado un espacio entre ambos hexámeros que puede ser de 15 a 21 bp, con una preferencia a ser de 17 bp [Hawley_1983, Harley_1987, Lisser_1993], ver Figuras I.2 y I.3.

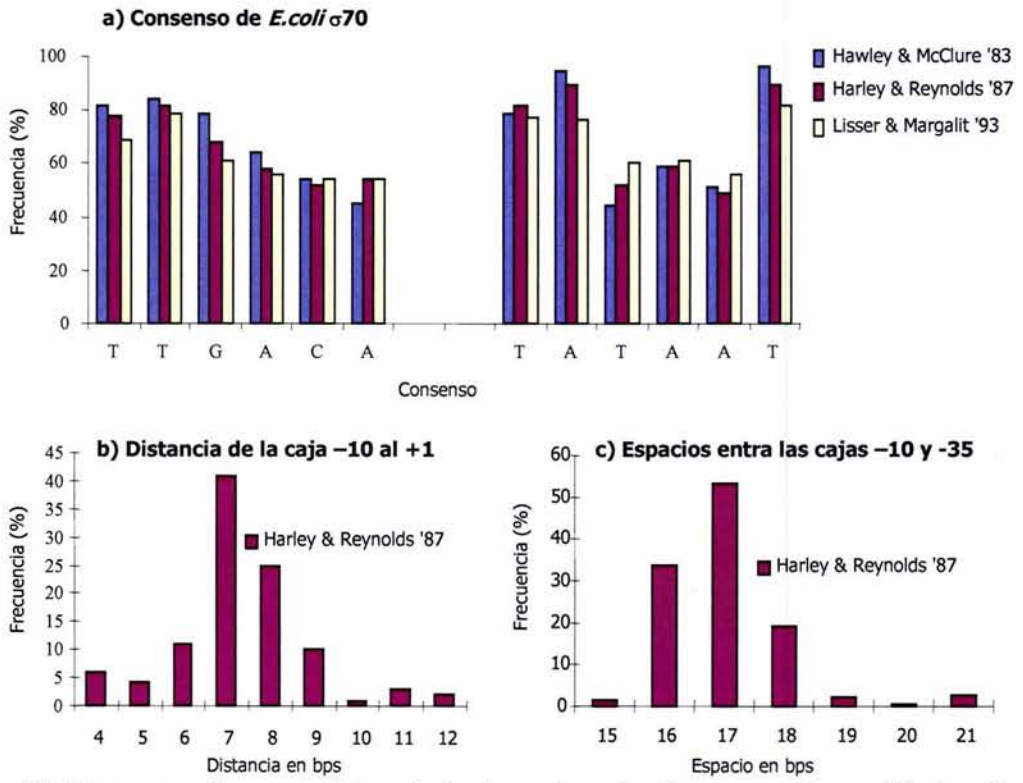


Figura I.3. Consensos más representativos de las 3 grandes colecciones que se han publicado sobre los promotores reconocidos por el factor σ70. (a) Frecuencias de las bases más conservadas de las cajas -10 y -35 basadas en 112 promotores en color azul, en 263 promotores en color púrpura, y en 298 promotores en

color amarillo [Hawley 1983, Harley 1987, Lisser_1993]. (b) Distribución de 263 promotores respecto al inicio de la transcripción [Harley_1987]. (c) Distribución de promotores respecto al espacio, en bp, separando a las cajas -10 y -35 [Harley_1987].

Estos tres elementos se convirtieron en los cánones que caracterizan a un promotor σ^{70}. Cada par de base de ambos hexámeros está localizada en una posición específica, sin embargo son secuencias que tiene mucha variabilidad. Tan solo en la colección de los primeros 6 promotores de *Escherichia coli*, las cajas -10 tuvieron solo 2 posiciones conservadas de las 6 del hexámero [Stormo_2000, Pribnow_1975], ver figura I.4. Esta variabilidad se conservó en las subsecuentes colecciones ampliadas de promotores que se publicaron en la literatura [Hawley_1983, Harley_1987, Lisser_1993]. Para ambas regiones el promedio de conservación es de 7.9 nt por promotor y aproximadamente sólo el 10% de los promotores eficientes son iguales al consenso en 5 de las seis posiciones [Lisser_1993, Ozoline_1997].

TACGAT	
TATAAT	
TATAAT	
GATACT	
TATGAT	
TATGTT	
TATAAT	**Secuencia Consenso**
TATRNT	Secuencia Consenso alterada

Figura I.4. La región -10 de los seis promotores de Pribnow [Pribnow_1975], y dos posibles representaciones del consenso de las secuencias [Stormo_2000].

I.2 ELEMENTOS NO CANÓNICOS QUE SON PARTE DEL PROMOTOR.

Aproximadamente 25 años después del descubrimiento del factor σ fue encontrado un tercer elemento de reconocimiento para promotores bacterianos, el cual está reconocido por la subunidad α de la RNAP [Ross_1993, Rao_1994, Gaal_1996, Estrem_1998]. Este tercer elemento, una secuencia de nucleótidos rica en (A+T), localizado en la región río arriba de los hexámeros -10 y -35 del promotor ribosomal *rrnBp*1, fue llamado el **elemento UP** (por "**up**stream"), ver figura I.5. Dicho UP, en ausencia de otros factores adicionales a la RNAP, estimula la transcripción por un factor de 30 veces *in vivo* como *in vitro*. Las secuencias UP han mostrado incrementar las actividades de varios promotores *E. coli* y *Bacillus subtilis, in vitro*, bajo la misma condición anterior [Ross_1993]. Asimismo estudios estadísticos y computacionales han detectado estas regiones ricas en (A+T) río arriba de muchos promotores [Ozoline_1997, Deuschle_1986, Galas_1985, Plaskon_1987].

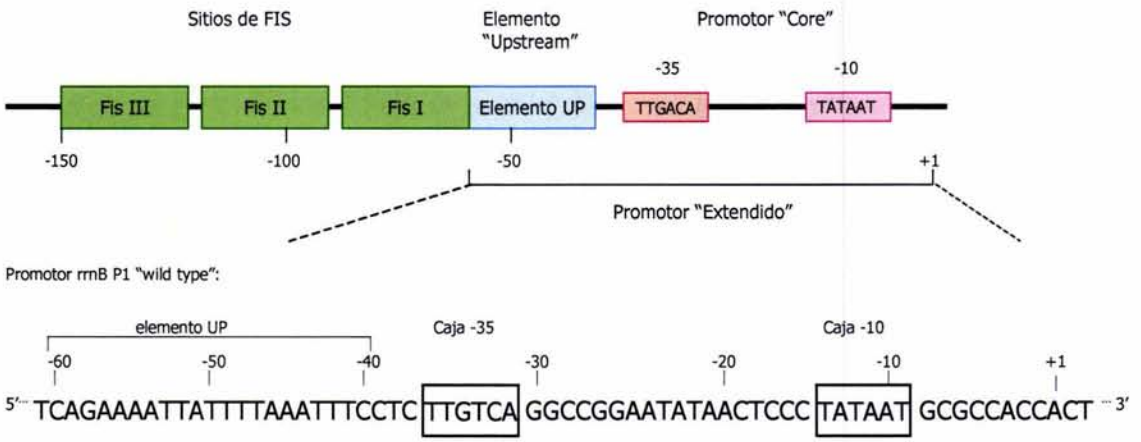


Figura I.5. Estructura de la región del promotor ribosomal *rrnBp*1. El "core" del promotor incluye las cajas -10 y -35 (rectángulos naranja y rosa). El elemento UP (la región de -40 a -60 bp; rectángulo azul) incrementa a 30 veces (fold) la actividad del promotor en ausencia de otros factores diferentes a la RNA polimerasa, y es considerado parte de un promotor "extendido" [Rao_1994]. Los sitios de pegado para la proteína activadora FIS se muestran en rectángulos verdes.

Varios experimentos han mostrado también la existencia de promotores cuya transcripción es dependiente de una secuencia conservada 5'-TG-3' localizada una base río arriba de la caja -10 [Gross_1998, Burr_2000, Mitchell_2003]. Lo interesante de este motivo, conocido como la "**caja -10 extendida**", es que hace innecesario el requerimiento de una caja -35 para que se de el proceso de la transcripción. Este motivo ha sido reportado en el 30% de los promotores en *E.coli*, pero en algunas bacterias gram-positivas está presente en el 50% de los promotores.

Por otro lado en promotores dependientes de un activador, la caja -35 puede ser sustituida por sitios de pegado para proteínas reguladoras de la transcripción las cuales tienen la misma función de una caja -35. El compromiso de la región -35 por acomodar estos sitios de pegado y su dispensabilidad debido al motivo TG hace que las secuencias correspondientes a esta región muestren una amplia heterogeneidad.

Toda la amplia gama de posibles cajas -35 y -10 que pueden formar un promotor funcional hace que un algoritmo computacional no pueda diferenciar con certeza cuando una región puede ser un promotor o no. Esta limitante provoca que no se puedan generar algoritmos eficientes para la búsqueda de sitios de promotores, ya sea usando calificaciones estadísticas, o usando reconocimiento por identificación de patrones ("pattern matching") de las bases canónicas del promotor.

I.3 Planteamiento del Problema.

La RNAP se pega de manera reversible a una cara del DNA de doble cadena cuando reconoce al promotor. Experimentos de "footprinting" del complejo RNAP-DNA indican que la RNAP cubre una región de 70 a 80 bp que se extiende de -55 a aproximadamente +20 con respecto al sitio de inicio de la transcripción [Mooney_1998]. La identificación de promotores en organismos ya sean procariontes o eucariontes es un problema extremadamente difícil dada la alta heterogeneidad de las secuencias que conforman al promotor. Una gran variedad de estrategias ha sido usada para dar solución al

problema de caracterizar las secuencias consensos de promotores y reconocerlos en una secuencia de DNA dada [Vanet_1999]. Con el advenimiento de la era genómica, y por ende con la disponibilidad de la secuencia completa del genoma de un organismo, el problema de predecir promotores en el DNA puede contextualizarse a las regiones no codificadoras entre genes no convergentes las cuales presumiblemente pueden tener un promotor. Esto se traduce en buscar promotores en regiones de aproximadamente 250 bp ya que la mayoría de los promotores descritos en la literatura se encuentra en ese rango de distancia desde el inicio de la traducción, como se mostrara mas adelante.

En este proyecto se definió y desarrolló algorítmicamente un método analítico para la detección de patrones canónicos y no-canónicos, que están presentes imperfectamente en conjuntos de secuencias promotoras. Este método está basado en la metodología de Teoría de la Información y de Conjuntos, y toma en consideración toda la información biológica disponible acerca del factor σ que da la especificidad de pegado a la RNA polimerasa. En este trabajo hemos explorado los conceptos de Teoría de Conjuntos y Relaciones pertenecientes al área de las Matemáticas Discretas para trabajar con este tipo de información biológica. El método que se ha desarrollado es capaz de detectar señales promotoras usando los argumentos canónicos y no-canónicos del promotor. El elemento no-canónico que es usado aquí se define por primera vez en este trabajo. Específicamente nuestro algoritmo emplea la siguiente información biológica:

Canónica:

a) Los consensos de la caja -10 y la caja -35.
b) Las distancias entre las cajas -10 y -35.

No canónica:

c) La posición del más uno con respecto al inicio del gen.

En el estudio de las bases biológicas y teóricas necesarias para el desarrollo de este método se contempló la obtención de las siguientes metas específicas:

A. Basándose en una colección de mas de 560 promotores con inicio de la transcripción mapeado; re-evaluar la variabilidad de las secuencias versus el consenso canónico.

B. Evaluar la frecuencia de los otros elementos que conforman al promotor para analizar su utilidad en la detección de predicciones positivas de promotores σ^{70}. Este análisis intentara verificar la sugerencia, hecha en [Estrem_1998, Estrem_1999], de que los elementos UP son un componente común de los promotores en *E. coli*. Hasta ahora, el elemento UP más extensamente caracterizado es aquel que fue localizado en uno de los siete *E.coli* promotores ribosomales de *E. coli*, *rrnBp*1, el cual tiene una fuerza de activación de la transcripción extraordinaria [Ross_1993, Rao_1994, Gaal_1996].

C. Proponer métodos de predicción de promotores reconocidos por el factor σ^{70} por medio de la detección de patrones canónicos, y no-canónicos.

D. Establecer algunas de las características propias de las secuencias promotoras que las distinguen de otro tipo de secuencias a nivel genómico. ¿Cómo la RNAP reconoce específicamente en el DNA las cajas –35 y –10 las cuales están separadas por longitudes variables de pares de bases? Es una pregunta que permanece aun sin ser completamente contestada [Murakami_2002b]. Los resultados teóricos expuestos aquí pretenden dar algunas sugerencias que podrían ayudar a contestar esta pregunta.

E. Aplicar algunos de los métodos de predicción de promotores desarrollados en este trabajo para el reconocimiento de promotores de otras sigmas alternativas. Ya que 5 de las 6 sigmas alternativas de *E. coli* pertenecen a la familia σ70, la aplicación de las metodologías desarrolladas para σ70 se extiende de manera automática para estos otros tipos de sigmas. En particular se intentara aplicar la metodología en el reconocimiento de promotores σ32.

F. Asumiendo que la mayoría de los factores sigma bacterianos pertenecen a la familia σ70 y asumiendo que los sigmas "house-keeping" reconocerán secuencias que presenten las mismas características de las secuencias promotoras reconocidas por el sigma "house-keeping" de *E. coli*, se propondrá expandir este análisis a otros genomas bacterianos.

I.4 CONTENIDO DEL TRABAJO.

En el capítulo 1 se dio una introducción al área en la que se circunscribe este proyecto, presentando al lector el concepto del promotor bacteriano, los elementos que lo componen en el DNA, y una descripción del problema de identificarlos inequívocamente en secuencias de DNA no-codificantes.

En el capítulo 2 presenta un breve resumen de los diferentes esquemas que se han utilizado para resolver el problema de identificar señales de control, como la de los promotores, en secuencia de DNA, presenta una descripción de la manera clásica que se usó para detectar estas señales, y finalmente presenta las bases teóricas en las que se sustenta la metodología que se introduce aquí para la detección de promotores.

En el capítulo 3 se dan los resultados del análisis realizados con las metodologías presentadas en este trabajo para el reconocimiento de promotores σ70. Aquí se replantea brevemente los elementos estructurales que forman parte de un promotor, y se anexa el artículo en el que nuestros resultados fueron publicados. En este artículo presentamos 2 modelos alternativos para la caracterización de promotores: ***Cover Function*** y ***DRHES Method***. Como un añadido se explica la metodología general que implementamos a través de un ejemplo en la región promotora del gen *tyrR*.

El capítulo 4 presenta los resultados colaterales de nuestro análisis e incluye los resultados de aplicar los algoritmos presentados en el capítulo 3 a la detección de promotores tipo σ32 en el genoma de *E. coli*, así mismo se presenta en este capítulo los resultados de un análisis comparativo entre las predicciones de σ32 y los resultados de dos experimentos de micro-arreglos en los cuales la condición ensayada fue un choque de calor a 50° ("heat-shock"). Los resultados del análisis comparativo de las predicciones de σ32 con los experimentos de micro-arreglos mostraron que el desempeño del algoritmo de

predicción de promotores para esta sigma es muy baja, además de que se encontró que no existe ninguna correlación entre el valor de la expresión del gen, y la calificación (score) de homología del promotor σ32 predicho. Una explicación posible de esta disparidad puede ser el ruido presente en mucho de los experimentos de micro-arreglos y a que la literatura ha reportado sobre promotores en choques de calor de 37 a 47° condiciones muy diferentes a las de los experimentos analizados aquí. En ese capítulo se propone un criterio de congruencia para evaluar la bondad de un experimento de transcriptoma, y se muestran los resultados de este análisis de la congruencia en 18 diferentes experimentos de micro-arreglos. También se presenta al lector la herramienta de web que se desarrolló para análisis de los datos que vienen de experimentos de perfiles globales de expresión en *Escherichia coli*. Se incluyen resultados preliminares del uso de la metodología presentada aquí para el estudio de promotores en 2 genomas, uno cercano a E. coli, *Salmonella typhimurium LT2*, y uno alejado, el plásmido simbiótico de *Rizhobium etli CFN42*.

En el capítulo 5 presentamos la discusión y las conclusiones de este trabajo así como las perspectivas de trabajos futuros inmediatos a desarrollar.

El capítulo 6, Materiales y Métodos, presenta los datos que se utilizaron en el desarrollo de este proyecto de doctorado. La eficiencia del método para predecir promotores depende en gran medida de la elección de las matrices consensos que se usaran para el reconocimiento de los promotores; en este capítulo se describe el procedimiento que se siguió para las construcciones de las matrices y los criterios de selección de las mejores matrices a usarse como sensores de promotores en el DNA. Mas de 200 matrices fueron analizadas y se encontró que los patrones canónicos se comportan mejor como herramientas predictivas aun cuando no resultaros ser los mejores modelos estadísticos. La presencia de otros elementos parece ser importante para distinguir lo que es una región reguladora de una que no lo es, problema que aun en la literatura no se encuentra resuelto. El anexo A presenta un resumen somero de la Teoría de la Información aplicada al descubrimiento de patrones de control en un conjunto no alineado de secuencias biológicamente relacionadas.

Finalmente el capítulo 7 contiene la bibliografía.

CAPÍTULO II. ANTECEDENTES.

Este proyecto involucra dos áreas del conocimiento, Biología y Ciencias Computacionales. En ambos campos el problema de detectar o identificar patrones de control en cadenas de símbolos, como el DNA, ha evolucionado dentro de una área activa de investigación. Este problema puede dividirse en dos sub-problemas, el primer problema es el de extraer un consenso (el motivo común) de un conjunto de secuencias; y el segundo problema es el de la búsqueda en un genoma por un consenso (un motivo) previamente definido [Vanet_1999]. Esto quiere decir que, primero, dado un conjunto de secuencias biológicamente relacionadas encontrar el patrón de control común y la posición de éste en cada una de las secuencias esas secuencias, dando como resultado una representación de la especificidad del factor que las regula; y segundo, usar la representación de esos sitios para buscar secuencias nuevas, prediciendo de manera confiable el lugar donde un sitio adicional puede ocurrir en el DNA [Stormo_2000, Waterman_1984a].

Los métodos que se han propuesto para resolver este tipo de problemas vienen de paradigmas determinísticos como: "String-Pattern Matching" [Cormen_1990], Programación Dinámica (comentados en [Waterman_1984a, Landau_1986]), y métodos de Teoría Lingüística como las Gramáticas Generativas [Collado_Vides_1992, Dong_1994, Lafebvre_1996]; y de paradigmas heurísticos: como los modelos de Teoría de la Información [Schneider_1986, Hertz_1990, Hertz_1999], de Teoría de Mecánica Estadística [Berg_1987, O'Neill_1989], métodos probabilísticos como los de Funciones de densidad de Probabilidad [Waterman_1984b, Mulligan_1984, Staden_1984, van Helden_1998, Bockhorst_2003], modelos de optimización de similaridad local [Altschul_1990, Pearson_2000], modelos de Redes Neuronales y cadenas de Markov [Stormo_1982, Lukashin_1989, O'Neill_1992, Pedersen_1996], modelos estocásticos como los basados en los métodos de Montecarlo [Lawrence_1993]; y de modelos híbridos que mezclan ambos paradigmas [Roth_1998]. Cabe notar que se han mencionado solo los modelos más representativas de cada paradigma.

Todos estos enfoques han sido empleados para caracterizar secuencias de control y/o para descubrir nuevos patrones de control en secuencias de aminoácidos, DNA y RNA. Algunos de los métodos toman una cantidad de tiempo prohibitiva para su ejecución, son de naturaleza no-polinomial, como lo son todos los algoritmos determinísticos que crecen exponencialmente en tiempo y espacio de acuerdo al número y tamaño de las secuencias a analizar; otros necesitan secuencias ya alineadas con respecto a algún punto ya conocido, como puede ser el punto de inicio de la transcripción. Otros han sido usados para detectar sitios de pegado de proteínas u otro tipo de señales por entrenamiento, auto-aprendizaje, o procesos estocásticos. Estos últimos han demostrado una eficiencia buena en el reconocimiento de señales en contextos definidos, sin embargo ellos no pueden ser usados para mostrar cómo la predicción fue hecha (cajas negras) contribuyendo poco al conocimiento biológico, o carecen de una evaluación cuantitativa que refleje una correlación con la actividad del proceso biológico real, y en algunos casos es difícil incorporarles conocimiento biológico nuevo.

La bondad de un algoritmo de reconocimiento de sitios de pegado de proteínas o de cualquier motivo especial en secuencias de letras es cuantificada de varias maneras. Dos de las más importantes medidas son la **sensibilidad** y la **especificidad**, las cuales son

respectivamente, la capacidad del método para reconocer correctamente las señales de interés y la capacidad de discriminar las secuencias que no son una señal deseada.

II.1 MÉTODO CLÁSICO PARA CARACTERIZAR SECUENCIAS DE PROMOTORES.

El método usado inicialmente para analizar secuencias de promotores, y en general para analizar los sitios de pegado de proteínas en el DNA o el RNA ha sido el alineamiento de las secuencias reconocidas por una misma molécula, e. g. proteínas reguladoras, RNAP's, etc. [Schneider_1986, Vanet_1999]. Una vez alineadas se seleccionan las bases más comunes en cada posición para crear una secuencia consenso, o expresión regular, ver el tope de la figura II.1. Es difícil trabajar con un esquema como éste, además de que no es confiable a la hora de buscar por sitios nuevos, ésto se debe a que mucha información se pierde cuando la frecuencia relativa de las bases presentes en cada posición es ignorada. Para evitar este problema el siguiente esquema empleado para representar la información de la alineación (**alineamiento**) fue el histograma o matrices que registran la frecuencia de cada base en cada posición de las secuencias alineadas obteniéndose entonces la frecuencia relativa de cada base por posición [e. g. Staden_1984, Schneider_1986], ver la parte inferior de la figura II.1.

A partir de entonces se han empleado diferentes esquemas estadísticos para hacer alineaciones y caracterizar numéricamente las secuencias alineadas, generando matrices de peso, que sirvan primero, para distinguirlas de otros conjuntos de secuencias, y segundo, para evaluar la pertenencia o no de nuevos elementos al conjunto de sitios reconocidos por el mismo reconocedor molecular. Una lista por orden de aparición de estos métodos es resumida en [Stormo_2000]. Uno de los métodos más ampliamente usado para la generación de matrices de peso por posición (PWM) fue el propuesto por [Schneider_1986, O'Neill_1998] el cual aplica el índice de Contenido de Información de Shannon para evaluar la significancia de cada posición dentro del consenso de una familia de secuencias alineadas cuando son comparadas con su frecuencia en el genoma [Schneider_1986, Berg_1987, O'Neill_1989]. Más adelante se propuso una estadística de normalización (*log-likelihood ratio*) del índice de contenido de información que puede ser usada para estimar la significancia de la matriz consenso completa [Hertz_1999, Stormo_2000, Benos_2002].

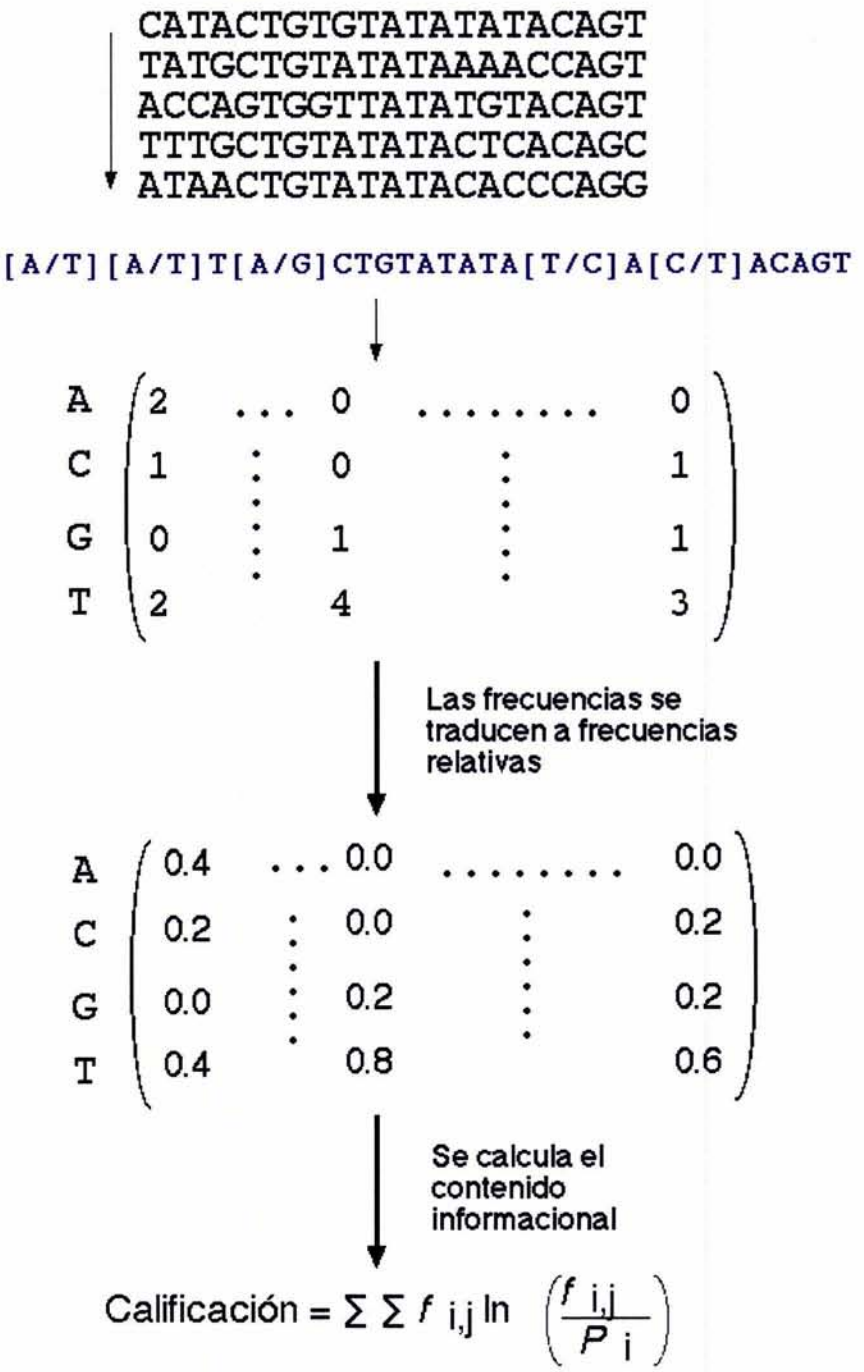


Figura II.1. Método clásico para la predicción de promotores y sitios de pegado. En la fórmula $f_{i,j}$ es la frecuencia relativa de la letra i-esima en la columna j-esima de la matriz de frecuencias relativas. P_i es la probabilidad *a priori* de encontrar la i-esima letra al azar.

Ver en la figura II.2 un ejemplo de como calificar una secuencia blanco usando una matriz de peso, calculando la calificación con la metodología de índice de C0ontenido de

Información. Una matriz de peso asigna un peso a cada posible nucleótido en cada posición de un sitio de pegado putativo y da como la calificación final del sitio la suma de esos pesos. Es obvio que una matriz de peso da una descripción más informativa de la especificidad del sitio de pegado que una secuencia consenso [Schneider_1986, Hertz_1990, Fickett_1997, Stormo_1998, Hertz_1999].

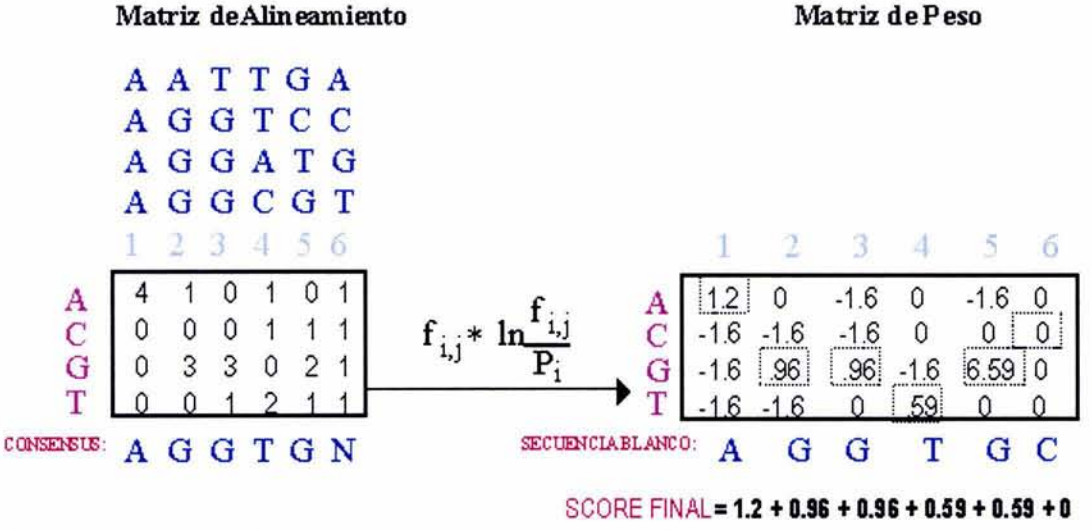


Figura II.2. Ejemplo del modelo de matrices para medir cuantitativamente la relación de las secuencias en una alineación y para calificar la cercanía de una secuencia blanco al alineamiento. En la formula $f_{i,j}$ es la frecuencia relativa de la base *i* en la posición *j*, y P_i es la probabilidad *a priori* de la letra *i* (0.25 para este ejemplo) [Hertz_1999].

La selección del algoritmo, o de la metodología, más eficiente para extraer la mejor matriz consenso de las secuencias promotoras que pueda ser usada como sensor por algoritmos de reconocimiento de promotores en secuencias reguladoras, es un paso critico en cualquier método de detección de promotores. Entre mejor represente una matriz al elemento del promotor que se desea buscar, mejor será la capacidad de los algoritmos para identificar de manera especifica a una región promotora en el DNA.

La ventaja de utilizar los modelos estadísticos de matrices de peso, como el de la figura II.2, es que son simples, y permiten establecer correlaciones biológicas entre la calificación de una secuencia particular y su actividad como promotor, u obtener un estimado de la contribución de energía de pegado de cada base en cada posición [Mulligan_1984, Berg_1987, Stormo_1998]. El estudio de la proteína Mnt es un buen ejemplo que demuestra que la matriz de peso generada con la familia de sitios de Mnt generados por SELEX es un estimador bueno de la matriz de afinidad calculada por las mutaciones únicas de todas posibles bases en cada posición del sitio silvestre [Stormo_1998]. Las limitantes de este tipo de modelos son que asume que las bases en cada posición son mutuamente independientes, lo cual no es el caso de algunas señales en el DNA donde la contribución de los nucleótidos a la fuerza o especificidad del sitio no es individual [Fickett_1997, Benos_2002]. En esos casos de no siempre existe una correlación entre la calificación de un sitio y la energía de pegado [Man_2001, Bulyk_2002]. Otra limitante es que no se toma en cuenta información del contexto de las secuencias, como puede ser la conformación local del DNA, o la presencia en fase de ciertos elementos del DNA [Vanet_1999]. A pesar de estas limitantes se ha demostrado que el uso de Teoría de Información, para la representación de los sitios de pegado de una

proteína, es un buen modelo para medir las interacciones proteína-DNA, y un método conveniente para la búsqueda de nuevos sitios [Stormo_1990, Benos_2002].

El problema de encontrar patrones desconocidos que ocurren imperfectamente a lo largo de una secuencia genómica, y/o el de buscar posiciones nuevas de motivos conocidos en un genoma que también ocurren imperfectamente en secuencias geonómicas, es un problema no trivial. Nosotros pensamos que herramientas de las Matemáticas Discretas pueden ofrecernos alternativas para su solución.

II.2 ANTECEDENTES TEÓRICOS.

Conjunto.

Un *conjunto* es una colección de objetos de cualquier tipo. Un objeto a que pertenece a un conjunto **A** es llamado un *elemento* de ese conjunto. Y se denota por $a \in \mathbf{A}$.

Inclusión de un conjunto.

Sí **A** y **B** son dos conjuntos cualesquiera y si todo elemento de **A** es un elemento de **B**, entonces se dice que **A** es un *subconjunto* de **B**, o que **A** está *incluido* en **B**, o que **B** *incluye* a **A**. Simbólicamente esta relación se denota por $\mathbf{A} \subseteq \mathbf{B}$.

Conjunto Potencia.

Para cualquier conjunto **A**, el conjunto vacío $\varnothing$ y el conjunto **A** son subconjuntos de **A**. También para cualquier elemento $a \in \mathbf{A}$, el conjunto $\{a\}$ es un subconjunto de **A**. De manera similar se pueden estimar otros subconjuntos de **A**. El conjunto de todos los posibles subconjuntos de **A** es llamado el *conjunto potencia* de **A**. Para cualquier conjunto **A**, su conjunto potencia se denota como $p(\boldsymbol{A})$, o $2^{|\boldsymbol{A}|}$, donde $|\boldsymbol{A}|$ es el número de elementos en **A**.

Relación.

Asociado a la idea de una relación está el acto de comparar objetos los cuales están relacionados a otro. Dos elementos, a y b forman una pareja ordenada si ellos pueden relacionarse o ser comparables bajo una relación dada **R**. Entonces se dice que $a, b \in \mathbf{R}$ donde **R** es la relación, y se denota por $a\ \mathbf{R}\ b$, lo que se lee como, *a está en relación R con b.*

Las siguientes son las propiedades que puede tener una la relación **R** sobre un conjunto **A** cualquiera.

- **R** es reflexiva, sí para cualquier $a \in \mathbf{A}$, $a\ \mathbf{R}\ a$.
- **R** es simétrica, sí para cualquier a y b en **A**, siempre que $a\ \mathbf{R}\ b$, $b\ \mathbf{R}\ a$.
- **R** es transitiva, sí para cualquier a, b, y c en **A**, siempre que $a\ \mathbf{R}\ b$ y $b\ \mathbf{R}\ c$, $a\ \mathbf{R}\ c$.
- **R** es irreflexiva, sí para cualquier $a \in \mathbf{A}$, $(a, a) \notin \mathbf{R}$.
- **R** es asimétrico, sí para cualquier a y b en **A**, siempre que $a\ \mathbf{R}\ b$ entonces $(a,b) \notin \mathbf{R}$.

II.2.1 ORDENES PARCIALES.

Se dice que un conjunto está ordenado si existe una *relación de orden* sobre ese conjunto. Un orden siempre implica una jerarquía entre los elementos ordenados. Una relación **R** en un conjunto **P** es llamada una *relación de orden parcial* o un *semi-orden* en **P** si y sólo si **R** es reflexiva, asimétrica, y transitiva [Tremblay_1987]. Es una convención denotar al orden parcial por el símbolo "≤". Este símbolo no necesariamente significa "menor o igual que", como se usa para los números reales. Ya que la relación de semi-orden es reflexiva, la llamaremos de aquí en adelante una relación sobre un conjunto digamos **P**. Si ≤ es un semi-orden sobre **P**, entonces el par ordenado (**P**, ≤) es llamado un *conjunto parcialmente ordenado*.

Obsérvese que no es necesario tener $a \leq b$ o $b \leq a$ para cualquier a y b en un conjunto parcialmente ordenado **P**. De hecho, a podría no estar relacionado con b, en cuyo caso diremos que a y b son incomparables. Dos relaciones de orden parcial usadas frecuentemente son, la relación *Menor o Igual que*:

- (**Z**, ≤) donde ≤ es la relación conocida comúnmente "*menor o igual que*" definida sobre el conjunto de los números enteros.

y la relación *Inclusión*:

- (**p(X)**, ⊆) donde **p(X)** es el conjunto potencia del conjunto **X** y la relación de semi-orden está dada por ⊆, la relación comúnmente conocida como inclusión.

II.2.2 LA COBERTURA DE UN CONJUNTO.

Dejemos que **S** sea un conjunto dado y $\mathbf{A} = \{ \mathbf{A_1}, \mathbf{A_2}, ..., \mathbf{A_m} \}$ donde cada $\mathbf{A_i}$, $i = 1,..., m$, es un subconjunto de **S** y la unión de todos los subconjuntos de **A** es **S**, $\bigcup_{i=1}^{m} \mathbf{A}_i = \mathbf{S}$, entonces el conjunto **A** es llamado la *cobertura* de **S**, y se dice que los subconjuntos $\mathbf{A_1}$, $\mathbf{A_2}$, ..., $\mathbf{A_m}$ *cubren* a **S** [Tremblay_1987].

Cualquier relación de orden parcial aplicada sobre un conjunto define una cobertura del conjunto. Para cada subconjunto de la cobertura de **S** se puede definir un *borde superior* y un *borde inferior*.

Sea (**P**, ≤) un conjunto parcialmente ordenado y **A** un subconjunto de **P**, $\mathbf{A} \subseteq \mathbf{P}$. Cualquier elemento $x \in \mathbf{P}$ es un *borde superior* de **A** si para toda $a \in \mathbf{A}$, $a \leq x$. De manera similar, cualquier elemento $x \in \mathbf{P}$ es un *borde inferior* de **A** si para toda $a \in \mathbf{A}$, $x \leq a$.

CAPÍTULO III. RESULTADOS.

III.1 RECONOCIMIENTO DE PROMOTORES σ70.

El factor σ70 es el sigma de mantenimiento endógeno del genoma *E. coli* y por si mismo da forma a toda una familia de factores sigmas que comparten las mismas características y dominios conservados que ejercen la misma función biológica [Arthur_1998]. Características que se han conservado a través de la evolución [Helmann_1988, Conaway_1990, Lonetto_1992]. Resolver el problema de detección de promotores σ70 sienta las bases para el reconocimiento de todo factor perteneciente a la misma familia de factores σ70 en bacterias.

Como ya se mencionó en el capítulo I, los elementos estructurales en el DNA que componen a un promotor tipo σ70 pueden ser: una región -10; una región -35 que puede estar presente o no dependiendo de la presencia de una proteína activadora que supla la función de la caja -35; los espacios permitidos a los que una caja -35 puede estar de una posible caja -10; una subregión (**TG**) 1 base río arriba de la caja -10 que da forma a lo que se le llama la "caja -10 extendida", que puede estar presente o no, y cuya presencia hace, al igual que algunas proteínas activadoras, que la región -35 no sea funcionalmente necesaria [Gross_1998, Burr_2000, Mitchell_2003]; un elemento UP que se encuentra localizado aproximadamente 4 bases río arriba de la caja -35 y que le confiere una fuerza de pegado extraordinaria al promotor [Ross_1993]. Podemos decir que todos estos elementos y la mezcla de alguno de ellos pueden hacer que una región en el DNA sea reconocida por una RNApol y sea factible de tener una función de promotor. Las combinaciones generales que pueden definir a un promotor pueden ser aquellas determinadas por las reglas de la figura III.1. Esto no implica que otras combinaciones no puedan darse en la naturaleza.

a) PromotorFuerte → (ElementoUP, Caja-35, [15...21], Caja-10)

b) PromotorCanónico → (Caja-35, [15...21], Caja-10)

c) Promotor-10Extendida → (ElementoTGn, Caja-10)

d) PromotorClaseII → (SitioActivador, Caja-10)

e) Promotor~ClaseI → (SitioActivador, ElementoUP, Caja –35, Caja-10)

Figura III.1. Reglas generales de los elementos que pueden componer a un promotor σ70. Estas reglas asumen que cada elemento se encuentra en las distancias convenientes a las diferentes topologías que aseguran que los contactos de la RNAP y el promotor y/o proteína activadora se mantendrán. La regla (a) modela los a promotores ribosomales [Gourse_1998]. La regla (b) da cuenta por la mayoría de los promotores en *E. coli* y no sugiere algún tipo de regulación particular [Youderian _1982, Lisser_1993]. La regla (c) muestra como la caja –35 es dispensable, pudiendo ser ignorada o permitiendo que su localización esté mas allá de las 21 bases como espaciador [Mitchell_2003]. La regla (d) asume la presencia de una proteína activadora aun cuando este motivo no es parte del promotor *per se*. Esta regla da cuenta por los promotores dependientes de activación de la clase II, es decir aquellos donde la interacción del activador con la RNAP es la única base de la activación y el sitio del activador se superpone a la caja -35 [Dhiman_2000, Benoff_2002]. La regla (e) permite la descripción de promotores en los cuales la

interacción de la RNAP con el activador es una de muchas interacciones involucradas en la activación; da cuenta por promotores de clase I y III [Busby_1999, Ishihama_2000]. Esta regla contempla los casos en que las otras interacciones involucradas son con uno de los elementos que componen al promotor, en los casos reportados este otro elemento puede ser parte del elemento UP ya que se asume que el activador esta río arriba del promotor y RNAP esta haciendo contacto con este y con el elemento UP para mediar la transcripción.

E. coli es, hasta donde se sabe, el primer genoma bacteriano completamente secuenciado con caracterización experimental y computacional de una gran variedad de elementos tales como operones, genes, promotores y sitios de pegado de proteínas reguladoras. Esta anotación exhaustiva y esfuerzo predictivo está soportado en la riqueza de la información disponible para *Escherichia coli*. Para la implementación del método de predicción de promotores usamos una colección promotores contenida en **RegulonDB**, (http://www.cifn.unam.mx/Computational_Biology/regulondb), una base de datos relacional de regulación transcripcional y organización de operones [Salgado_2001]. RegulonDB tiene un conjunto de 599 inicios de transcripción mapeados en el genoma de *E. coli*. Como parte de esta colección se encuentra el conjunto de promotores reportados en [Hawley_1983, Harley_1987, Lisser_1993], así como información detallada sobre la regulación de la transcripción de algunos de los promotores ahí colectados.

El análisis del genoma completo de *E. coli* ha representado un reto dada la gran cantidad de conocimiento biológico que se genera sobre esta bacteria [Neidhardt_1996]. El grado de integridad del análisis en *E. coli* proveerá una referencia para la investigación de un gran número de nuevos genomas en donde la Biología no está aún ricamente documentada. El siguiente artículo muestra los resultados del análisis y de las estrategias implementadas en este trabajo para resolver el problema de caracterizar y reconocer promotores $\sigma 70$ en el genoma de *E. coli*.

III.2 La función "cover".

El problema de seleccionar un promotor entre un conglomerado de señales que parecen promotor fue resuelto usando la función cobertura ("Cover"), la cual fue introducida en el artículo anterior [Huerta_2003]. El aplicar la función cobertura en un conjunto de señales tipo promotor implica la detección de esta serie de posibles señales en el DNA.

El método general de reconocimiento de promotores presentado aquí fue implementado para que como primer paso, usando las matrices de peso, realice una búsqueda de todas las posibles señales promotoras en una región y después compare los valores relativos de los promotores putativos encontrados en la región, y seleccione el mejor candidato. Los valores a tomar en cuenta para cada señal tipo promotor son: la suma de las calificaciones de las cajas –10 y –35 y la posición de la caja –10 de la señal tipo promotor respecto del inicio del gen. Este paso es el único que hace que este método sea claramente dependiente sobre el contexto de búsqueda de secuencias y el único que claramente distingue a nuestro método de otras herramientas estándares. La figura III.2 muestra cómo este método trabaja usando la función cobertura aplicada a cada subconjunto de posibles promotores candidatos para una región dada; estos subconjuntos fueron generados al clasificar los promotores por los espaciadores conocidos que hay entre las cajas -10 y -35.

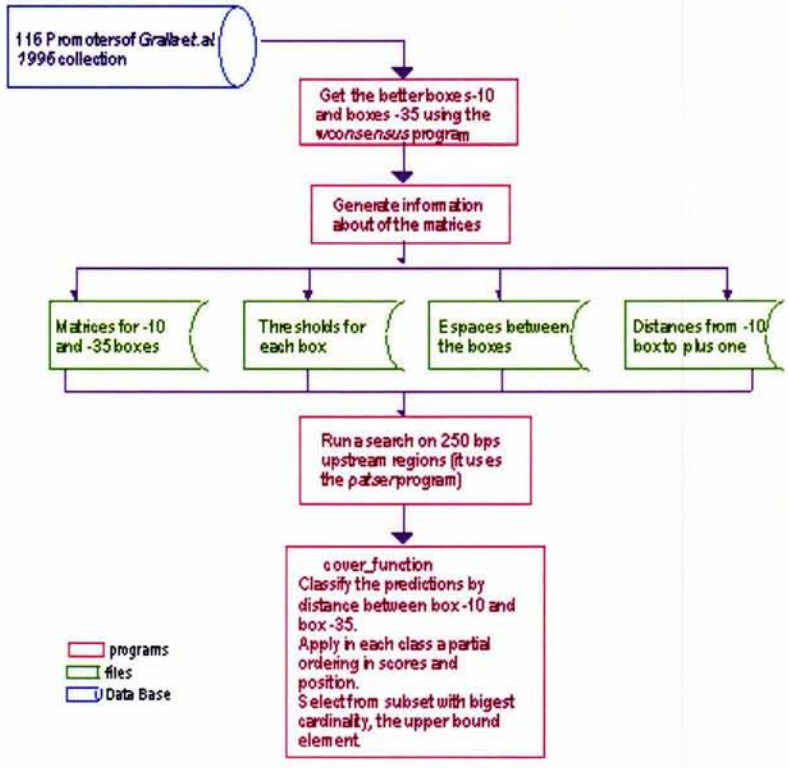


Figura III.2. Método para predecir promotores. Los cilindros azules corresponden a los datos que viene de la base de datos RegulonDB, los cuadros rojos corresponden a programas que ejecutan las instrucciones que se pueden leer al interior de estos cuadrados. Los rectángulos incompletos de color verde representan las salidas de los programas (datos) y que sirven como entrada a otros programas.

La esencia del reconocimiento de promotores está basada en la idea de divide y vencerás, muy socorrida en Programación Dinámica, y en el uso de un ordenamiento parcial - la función cobertura - y fue implementado en 2 pasos:

1. La selección del mejor par -10 y -35 para cada región reguladora fue hecha en dos pasos. En el primer paso clasificamos a todos los posibles promotores por el espacio entre las cajas -10 y -35, espacios de 15 a 21 pares de bases fueron permitidos, obteniendo subconjuntos los cuales llamaremos *subconjuntos-espacio*.

2. Una relación de orden parcial, llamada la *relación de inclusión,* fue aplicada a cada una de las 7 clases o *subconjuntos-espacio*. Esta *relación de inclusión* está basada en la comparación de los tres valores para cada promotor. Un promotor dado A se dice que esta incluido por otro promotor B sí y solamente si los valores del promotor B son calificaciones mejores que las del promotor A, decimos que $A \subseteq B$. La relación $\subseteq$ define una *cobertura* para cada una de los *subconjuntos-espacio*. Las matrices de peso para las cajas -10 y -35 fueron usadas para generar las calificaciones. Las posiciones absolutas fueron modificadas para reflejar la distribución de las posiciones conocidas con relación al inicio del gen. Ver Materiales y Métodos.

El mejor candidato es seleccionado desde los subconjuntos que forman la *cobertura* de cada uno de los *subconjuntos-espacio*. La selección es hecha de la siguiente manera, fueron elegidos aquellos subconjuntos cuya cardinalidad fuera la mayor; para cada uno de estos subconjuntos con cardinalidad mayor se calcularon los elementos *borde superior*, estos son los mejores candidatos para cada *subconjunto-espacio* disponible.

III.2.1 UN EJEMPLO.

Se corrió una búsqueda usando las matrices descritas en la sección VI.2 de Materiales y Métodos en la región promotora del gen *tyrR*, el umbral fue establecido a ser $\mu - 2.5\sigma$. Los resultados son mostrados en la figura III.3a. Se encontraron 22 promotores putativos para esta región. Para localizar al promotor funcional, se agrupan los 22 promotores por clase de espaciador entre las cajas -10 y -35, la figura III.3b nos muestra la renglón para cada *subconjunto-espacio*.

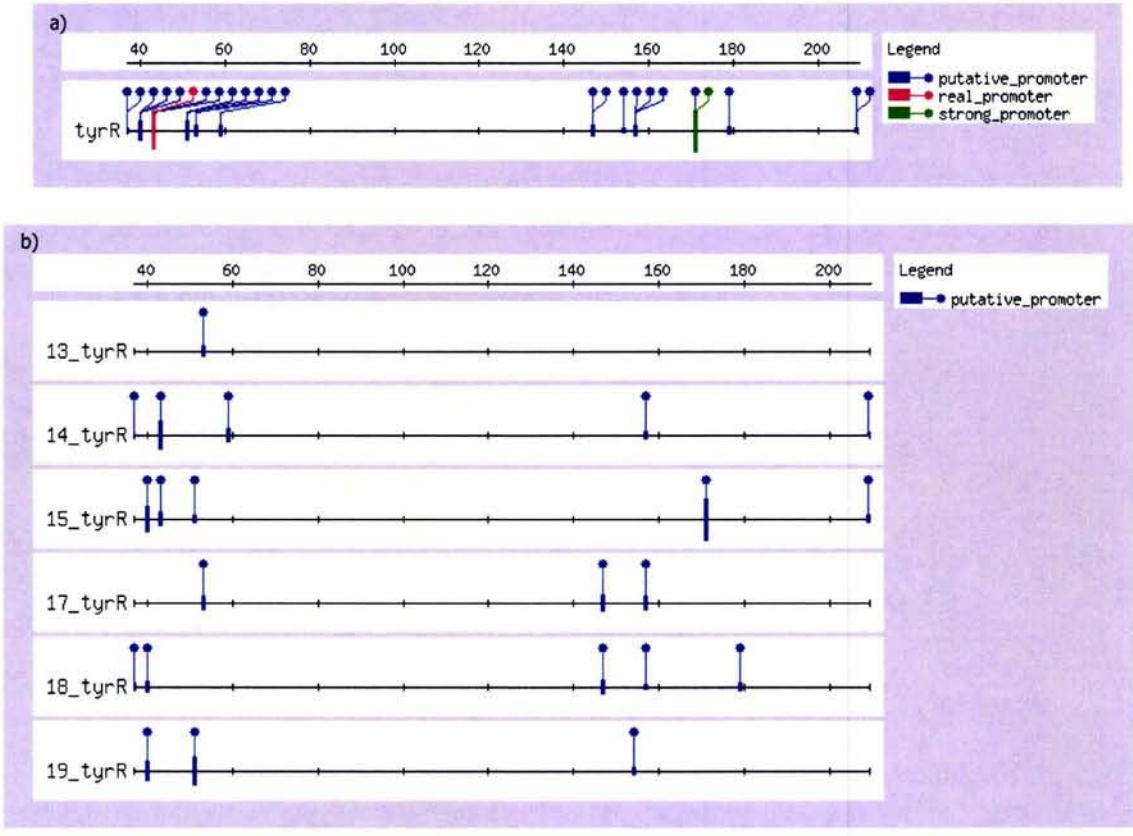


Figura III.3. a) Resultados de correr una búsqueda de promotores usando matrices de peso, y permitiendo un umbral de μ - 2.5σ. En rojo se puede ver al promotor reportado experimentalmente y en azul a las predicciones de promotor. La altura de los rectángulos en la base de cada símbolo-promotor es proporcional a su calificación, por ejemplo, en el renglón *15_tyrR*, el cuarto símbolo-promotor es el de mayor calificación. b) El problema de encontrar el promotor(es) funcional(es) tyrR entre 22 posibles predicciones de promotor se divide en este caso en encontrar la mejor predicción de promotor en cada clase. Esta figura muestra el resultado de dividir el problema usando los espaciadores como criterio de clasificación. La idea de separar a los promotores por su clase también fue propuesta por [Oneill_1989b] aduciendo que existen 3 principales clases de espaciadores, 16, 17, 18 bp, en las cuales los promotores divergen en la secuencia.

En la tabla III.1 se muestra los resultados de esta búsqueda. El mas-uno del promotor funcional reportado en RegulonDB está a 28 bp arriba del inicio del gen. El promotor funcional puede ser entonces cualquier promotor predicho cuya caja -10 este en la posición 37, 40, 43, e incluso el de 50, arriba del inicio del gen.

Espacio en bp	Posición de la caja -10	Calificación -10	Calificación -35	Caja -35	Espacio	Caja -10
14	37	1.34	0.87	TGTCATATC	aatagtgtcatatc	ATCATATTA
18	37	1.34	1.68	TGTCATATC	tccaaatagtgtcatatc	ATCATATTA
19	40	2.02	1.83	TAGTGTCAT	gctatccaaatagtgtcat	ATCATCATA
18	40	2.02	-0.57	TAGTGTCAT	ctatccaaatagtgtcat	ATCATCATA
15	40	2.02	1.68	TAGTGTCAT	tccaaatagtgtcat	ATCATCATA
16	43	3.09	1.83	AAATAGTGT	gctatccaaatagtgt	CATATCATC
15	43	3.09	-0.57	AAATAGTGT	ctatccaaatagtgt	CATATCATC
15	51	1.43	0.48	TGCTATCCA	ccttcctgctatcca	AATAGTGTC
19	51	1.43	3.24	TGCTATCCA	gaaaccttcctgctatcca	AATAGTGTC
17	53	0.58	3.24	CCTGCTATC	gaaaccttcctgctatc	CAAATAGTG
13	53	0.58	0.48	CCTGCTATC	ccttcctgctatc	CAAATAGTG
14	59	1.89	1.41	AACCTTCCT	acagaaaccttcct	GCTATCCAA

Espacio en bp	Posición de la caja -10	Calificación -10	Calificación -35	Caja -35	Espacio	Caja -10
18	147	1.35	-0.36	GAAAAATAA	acgcccagcgaaaaataa	TGCAATATC
17	147	1.35	-0.25	GAAAAATAA	cgcccagcgaaaaataa	TGCAATATC
19	154	0.37	0.97	GCCCAGCGA	aaggcaaaacgcccagcga	AAAATAATG
17	157	0.54	0.75	AACGCCCAG	aaaggcaaaacgcccag	CGAAAAATA
16	157	0.54	0.97	AACGCCCAG	aaggcaaaacgcccag	CGAAAAATA
18	157	0.54	1.38	AACGCCCAG	aaaaggcaaaacgcccag	CGAAAAATA
15	171	1.21	4.10	CTGCAAAAA	gaaacgctgcaaaaa	GGCAAAACG
18	179	0.48	1.08	AAGAAACGC	gtcaggtgaaagaaacgc	TGCAAAAAG
15	209	1.84	0.13	ATTTCCGTC	tcggggatttccgtc	GTCAGCTTA
14	209	1.84	0.27	ATTTCCGTC	cggggatttccgtc	GTCAGCTTA

Tabla III.1. Las señales que calificaron a ser un promotor en la región promotora del gen *tyrR*. La primera columna muestra los espacios que hay entre las cajas -10 y -35, estos coinciden con aquellos de la literatura, ya que la caja -10 encontrada es de 9 bp y en el extremo izquierdo hay 2 bp que el hexámero -10 canónico no tiene, esto quiere decir que, por ejemplo, el espacio de 13 bp es el equivalente al espacio de 15 bp reportado en la literatura, y así sucesivamente para el resto de los espacios, ver la figura VI.6 de la sección de Materiales y Métodos.

El proceso para seleccionar al mejor candidato empieza por clasificar a los promotores por su espaciador entre las cajas esto nos deja los siguientes *subconjuntos-espacio*:

subconjunto_espacio 13 = { (53,0.58,0.48) }
subconjunto_espacio 14 = { (209,1.84,0.27), (37,1.34,0.87), (59,1.89,1.41) }
subconjunto_espacio 15 = { (171,1.21,4.10), (209,1.84,0.13), (40,2.02,1.68), (43,3.09,-0.57), (51,1.43,0.48) }
subconjunto_espacio 16 = { (157,0.54,0.97), (43,3.09,1.83) }
subconjunto_espacio 17 = { (147,1.35,-0.25), (157,0.54,0.75), (53,0.58,3.24) }
subconjunto_espacio 18 = { (147,1.35,-0.36), (157,0.54,1.38), (179,0.48,1.08), (37,1.34,1.68), (40,2.02,-0.57) }
subconjunto_espacio 19 = { (154,0.37,0.97), (40,2.02,1.83), (51,1.43,3.24) }

donde cada elemento en los *subconjuntos-espacio* queda definido por los siguientes atributos, posición de la caja -10, la calificación de la caja -10, y la calificación de la caja -35.

La relación de orden parcial es una competencia entre los promotores que pertenecen a cada *subconjunto-espacio*, llamémosla una competencia local. Decimos que un elemento P_j contiene a otro P_i, cuando P_j es mejor en sus atributos, entonces decimos que $P_i \subseteq P_j$. Para efectos de la comparación, las calificaciones -10 y -35 de cada promotor se suman ya que son eventos independientes, ver la figura VI.8a de la sección de Materiales y Métodos, y las posiciones de las cajas -10 se codifican de acuerdo a la formula (h) descrita en la misma sección. Para todo *subconjunto_espacio* y para todo elemento P_i =(a,b,c) y P_j = (x,y,w) $\in$ *subconjunto_espacio*, se dice que $P_i \subseteq P_j$ si:

$$\mathbf{(a+b) <= (x+y)\ \ \&\ (c <= w)} \qquad \text{(e)}$$

esta relación de orden parcial nos divide a los *subconjuntos_espacio* en subconjuntos como se muestra en la figura III.4.

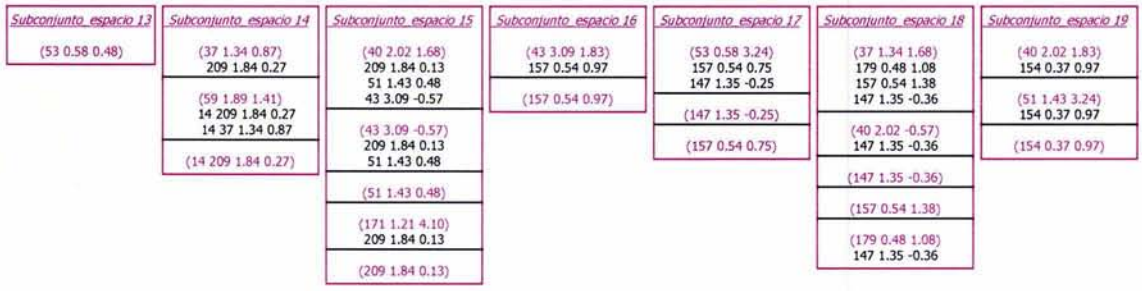


Figura III.4. Después de aplicar la relación de orden parcial a cada *subconjunto_espacio* estos quedan a su vez divididos en subconjuntos, la unión de los cuales forma la cobertura de cada *subconjunto_espacio*. Los promotores en color uva son los elementos *borde-superior* de cada subset.

Se selecciona para cada subconjunto-espacio el subconjunto con mayor cardinalidad, en el caso del *subconjunto-espacio 15*, el subconjunto con mayor cardinalidad es el que aparece, primero en la tercera caja de la figura III.4, con 4 elementos. De cada subconjunto seleccionado como el de mayor cardinalidad se obtiene el *borde-superior* y este es el que se considera el candidato optimo a ser el promotor de la región. Aquí se muestran los resultados de esta selección en cada uno de los *subconjuntos_espacio*:

subconjunto_espacio 13 = { (53,0.58,0.48) }
subconjunto_espacio 14 = { (59,1.89,1.41) }
subconjunto_espacio 15 = { (40,2.02,1.68) }
subconjunto_espacio 16 = { (43,3.09,1.83) }
subconjunto_espacio 17 = { (53,0.58,3.24) }
subconjunto_espacio 18 = { (37,1.34,1.68) }
subconjunto_espacio 19 = { (40,2.02,1.83), (51,1.43,3.24) }

Al final quedan 6 promotores con posiciones 37, 40, 43, 51, 53, y 59, de entre los cuales elegir cual es el mejor. La distancia entre 59 y 37 es de 22 bp, y hemos dicho que la RNAP cubre aproximadamente 60 bp. Si se desea un resultado más fino y preciso, los criterios para filtrar este último conjunto de resultados pueden ser varios.

Por ejemplo por posición repetida, en este caso es obvio que los promotores en las posiciones 37,40 y 43 pueden definir a un mismo promotor, se puede elegir a 40 como el candidato. También podemos pensar que 51, 53 y 59 definen también a un solo promotor, se puede elegir a 53 como el promotor candidato, de esta manera para la región de *tyrR* se tendrán dos candidatos posibles a ser un promotor, siendo la posición 37 el verdadero.

También puede elegirse como candidato a ser promotor aquel que es *borde-superior* del subconjunto con mayor cardinalidad, podemos decir el que "cubrió" a más promotores en proporción al tamaño del *subconjunto-espacio*. Los resultados para cada *borde-superior* son los mostrados en la tabla III.2. En este caso se seleccionaría a los promotores con posición 43 y 59, de esta manera se tendrán dos candidatos posibles a ser un promotor

subconjuntos_espacio	***borde_superior***	***cobertura***
subconjunto_espacio 13	(53,0.58,0.48)	-

subconjunto_espacio 14	(59,1.89,1.41)	100% (3/3)
subconjunto_espacio 15	(40,2.02,1.68)	80% (4/5)
subconjunto_espacio 16	(43,3.09,1.83)	100% (2/2)
subconjunto_espacio 17	(53,0.58,3.24)	3/3% (3/3)
subconjunto_espacio 18	(37,1.34,1.68)	80% (4/5)
subconjunto_espacio 19	(40,2.02,1.83) (51,1.43,3.24)	67% (2/3) 67% (2/3)

Tabla III.2. Para cada borde superior del *subconjunto-espacio* se calculo un porcentaje de cobertura que es igual al número de elementos los cuales son menores que el *borde_superior* entre el tamaño del *subconjuto_espacio* al que pertenece.

Otra opción para el filtrado podría ser eligiendo el borde_superior más fuerte, usando la formula descrita en la figura 3 del artículo [Huerta_2003]. En este caso el promotor en la posición 43 sería el ganador. En la tabla III.3 se ven los valores para cada *borde_superior* en cada *subconjunto_espacio*.

Espacio (bp)	Posición caja -10	Calificación -10	Calificación – 35	Posición codificada	Suma -10 y -35	Total
13	53	0.58	0.48	0.0948	1.06	1.24103
14	59	1.89	1.41	0.1379	3.3	3.48103
15	40	2.02	1.68	0.2758	3.7	3.91552
16	43	3.09	1.83	0.1896	4.92	5.13552
17	53	0.58	3.24	0.0775	3.82	4.00103
18	37	1.34	1.68	0.0775	3.02	3.09759
19	40	2.02	1.83	0.1465	3.85	4.06552

Tabla III.3. Para cada borde_superior se calcula su calificación final con la formula descrita en la figura 3 del artículo [Huerta_2003]. Se suman las calificaciones -10 y -35, y se codifica el espaciador basándose en la formula (i), ver Materiales y Métodos, y se suma a las calificaciones de las cajas.

La última fase de filtrado es opcional, y debe hacerse sobre todo cuando se requiere de predicciones más precisas. Como se ha mostrado hay varios criterios para filtrar el mejor promotor putativo. Puede hacerse mezclas como el de mayor cobertura y el más fuerte, en ese caso, 43 y 59 serían los posibles promotores putativos de *tyrR*.

CAPÍTULO IV. RESULTADOS ADICIONALES.

IV.1 VARIABILIDAD DE LAS SECUENCIAS PROMOTORAS VS. EL CONSENSO CANÓNICO.

RegulonDB es al momento la base de datos con la mayor cantidad de promotores σ70 con mas uno mapeado. Esta enorme nueva cantidad de promotores justifica un análisis sobre la variabilidad en los consensos generados por el alineamiento de estos promotores versus el consenso canónico generado a partir de conjunto de promotores más pequeños. Se alienaron 584 promotores σ70 utilizando la estrategia mencionada en la figura VI.3 de la sección de Materiales y Métodos. Las matrices seleccionadas por los criterios mencionados en la estrategia, Mínima Frecuencia Esperada, Máximo Contenido de Información, Distribución más compacta de las calificaciones de los promotores, y homología al consenso canónico, resultaron no muy ser diferentes de aquellas generadas para los 116 promotores del conjunto de entrenamiento. Las matrices representando los consensos canónicos fueron encontradas usando este conjunto de 584 promotores. Sin embargo al igual que con las matrices del conjunto de entrenamiento los mejores modelos estadísticos, aquellas matrices con la menor frecuencia esperada o con el máximo contenido de información no arrojaron los consensos canónicos. En estos casos, patrones como A-tracks (polyAAA), el consenso del elemento UP generado por SELEX y el consenso UP como el encontrado en el promotor *rrnBp1* aparecieron para el alineamiento correspondiente a la región –35. Para la región –10 las matrices contienen el motivo TG perteneciente a una caja –10 extendida. La figura IV.1 muestra la conservación de cada posición de los hexámeros canónicos en el conjunto alineado de los 584 promotores σ70.

Figura IV.1. Porcentaje de conservación de cada una de las bases del consenso canónico en un conjunto de 584 promotores σ70.

IV.2 RECONOCIENDO PROMOTORES σ32 EN *E. COLI* K12.

Cuando la célula se expone a un cambio brusco a altas temperaturas necesita generar una respuesta que le permita ajustarse a su nueva condición de crecimiento. Esta respuesta celular es clásicamente conocida como *heat-shock* (hs). La respuesta hs es transciente y por medio de cambios mantiene homeostasis. Durante el heat-shock se presenta un aumento en la cantidad de un conjunto de proteínas, las *proteínas hs* (hsp), se ha observado que en la mayoría de los casos esta inducción ocurre a nivel del proceso de la transcripción de los genes que las codifican [Gross_1996].

La mayoría de los genes que codifican para hsp son transcritos por la holoenzima RNApol que contiene al factor σ32. Hasta ahora se han descrito 22 promotores que se inducen durante el heat-shock y los cuales son dependientes de σ32. Este factor sigma pertenece a la familia de factores σ70.

La metodología presentada para detectar promotores σ70 se probó en la detección de promotores σ32 en el genoma de *E. coli* K12. Se compararon estos resultados con experimentos globales de expresión de genes (micro arreglos), hechos en el laboratorio del Dr. Blattner [Richmond_1999], que midieron la respuesta celular a hs.

IV.2.1 Perfil de expresión global en *E. coli* K12 a un choque de calor (Heat-Shock).

El experimento realizado para medir la respuesta celular a un choque de calor se hizo con una sola colonia de *E. coli* K12 para inocular 60 ml de preparado de Luria-Bertini (LB) en matraces de Elenmeyer de 250 ml y se dejó crecer a un ODA_{600} entre 0.8 y 0.9 a 37°C con una constante ventilación. La inducción del heat shock se realizó dividiendo un cultivo mid-log* en 2 muestras de 30 ml, manteniendo el cultivo de control a 37° durante 7 minutos mientras que el cultivo experimental se expuso a un choque de calor de 50°C agitándolo en agua por 7 minutos. Para el monitoreo en el tiempo, las células se dejaron crecer en medio rico (rich defined medium) a un OD600 = 0.8 y se tomaron muestras de 30 ml a 0, 5, 10, 15 y 20 minutos después de la transferencia de temperatura de 37 a 50°C. Las muestras se marcaron, se hibridizaron en un array y se midió la intensidad de las señales para los 4290 genes de la versión m54 de *E. coli* K12.

Con el experimento se pudo detectar el conjunto de genes que fueron afectados durante el choque de calor, se tomó como criterio arbitrario de afectación una razón de 5 veces entre la condición de heat-shock y el control. Es decir, todos aquellos genes que presentaron una inducción o represión de 5 veces (fold), respecto de la muestra control, fueron dichos a ser genes que participan en la respuesta celular que se dispara en *E. coli* al recibir el choque de calor. La tabla IV.1 muestra el número de genes que fueron afectados.

Experimento	Total de Genes $> \pm 5$ fold	Inducidos	Reprimidos
Heat-Shock	119	77	42

Tabla IV.1. Número de genes afectados por un fold ≥ 5.

IV.2.2 Comparando predicción de promotores σ32 vs. experimentos globales de expresión.

Se recolectó de la literatura toda la información disponible acerca del regulón de heat-shock, en total se obtuvieron 22 promotores σ32, ver tabla VI.5 de la sección de Materiales y Métodos. Debido a que se disponía de un conjunto muy pequeño de

promotores σ32 se decidió usar la colección de 17 promotores σ32 reportadas en [Gross_1996] como el conjunto de entrenamiento, ver tabla VI.5a. El conjunto total de 22 promotores que incluye los 17 del conjunto de entrenamiento mas 5 reportados en [Zhou_1988, Lesley_1990, Raina_1990, Dartigalongue_1998, Decker_1998] fue seleccionado como el conjunto de prueba, ver tabla VI.5b.

Con los 17 promotores de la tabla VI.5a se generó una alineación usando el programa de *wconsensus* [Hertz_1990] para obtener las cajas -10 y -35, las cuales en su mayoría coincidieron con aquellas reportadas en [Gross_1996], ver la columna 5 de la tabla VI.5 de Materiales y Métodos. Se calcularon los estadísticos de la media y la desviación estándar (μ,σ) y se encontró que el 90% de los promotores del conjunto de entrenamiento está contenido en la $\mu \pm 1.5 * \sigma$. Esta desviación estándar fue establecida como umbral de corte para decir cuando una secuencia blanco puede ser un promotor σ32. Con las matrices generadas se corrió la búsqueda de promotores σ32 en el conjunto de entrenamiento de los 21 promotores. El programa detectó 18 de esos 22 promotores. Se obtiene una sensibilidad del 81% con una especificidad del 100%.

Comparando los datos reportados en la literatura con los resultados del experimento con el micro arreglo pudimos observar que solo 16 de los 22 promotores fueron transcritos. Los 6 promotores que no parecen ser afectados en el experimento son mostrados en la tabla IV.2. 3 de estos 6 promotores no fueron detectados por el programa usando un umbral de corte de $1.5 * \sigma$ debajo de la media de los promotores del conjunto de entrenamiento.

Genes	Micro arreglos	Programa
HtgA	1.1	nodetectado
GapA	-1.3	detectado
rfaDFCL (htrM)	-1.7	detectado
HtrC	1.2	nodetectado
YbaU	-1.3	detectado
Mlc	2	nodetectado

Tabla IV.2. Genes hs reportados en la literatura y que no fueron afectados en el experimento global micro arreglos. La tercera columna muestra si el promotor fue detectado por el programa usando un umbral de corte de 1.5 desviaciones estándar por debajo de la media de las calificaciones de los 18 promotores que componen el conjunto de entrenamiento.

Desde las condiciones del experimento podemos explicar porque algunos de los genes, reportados en la literatura a ser expresado durante un heat-shock, no fueron afectados. *htgA*, y *gapA* son genes parálogos entonces su no-afectación puede deberse a una mala hibridizacion de los mRNAs. *ybaU* fue marcado para exclusión por mal producto de PCR. *mlc* tienen un promotor σ70 conocido, entonces su poca inducción puede deberse a un intercambio de sigma, es decir ya era producido el mRNA antes del hs. Algunos genes tienen promotores para σ70 y σ32.

Se hizo una búsqueda de promotores σ32 en el genoma de *E. coli* K12. Se encontraron 66 predicciones de promotores σ32. Nota que estas 66 no incluyen los 22 promotores conocidos.

Se verificaron estos resultados con aquellos 119 genes que se detectaron en el experimento de heat-shock usando el micro arreglo [Richmond_1999]. 13 de esos 119 corresponden a genes regulados por un σ32 conocido. Solo 4 de las predicciones hechas con nuestros programas coincidieron con genes afectados con al menos ± 5 fold en el micro arreglo. Este quiere decir que 102 genes no tienen ninguna predicción de ser regulados por σ32. Los genes con un promotor σ32 predicho son mostrados en la tabla IV.3.

Bnumber	Posición +1	Calificación -10	Calificación -35	Gen	MicroArray
b3498	-71	7.57	4.22	prlC	16.7
b0209	-65	5.79	5.44	yafD	6.7
b0032	-121	3.58	3.26	carA	-23.27
b0660	-61	7.57	3.14	b0660	5.94

Tabla IV.3. Genes afectados en el experimento de hs usando micro arreglos a los cuales el programa les detecto un promotor σ32.

Tomando en cuenta la especificidad del método, se asume que las predicciones de σ32 son en su mayoría correctas. Es decir, los 62 promotores predichos en genes cuyos fold de afectación no están arriba del corte de ±5 usado para el micro arreglo, presentan alguna de las siguientes características mostradas en la tabla IV.4, las cuales podrían dar evidencia de porque no se trata de una mala predicción computacional si no condiciones "especiales" del experimento:

Características de los genes	No. Genes
Parálogos	9
Excluidos (mal producto de PCR)	7
Función que coincide con las de los genes HS*	22
Posible regulación por cambio de sigma[§]	25

Tabla IV.4. Características que presentaron los genes en los cuales se detecto una predicción de promotor σ32. Excluidos son aquellos genes cuyo producto de PCR no fue bueno, o no hibridizó en el micro arreglo, ver referencia [Richmond_1999]. * Debiesen haber sido detectados en el experimento. [§] Evidencia de regulación por σ70.

Se analizaron las predicciones de promotores en términos de su calificación y posición dentro de los operones, no existe correlación entre la calificación del promotor vs. su posición dentro del operon, es decir, se quería ver si las calificaciones eran bajas en predicciones inter-operon, datos no presentados aquí muestran que no existe dicha relación,

de hecho la mayoría de las calificaciones "bajas" caen en la posición 1ª del operon, es decir en la región de regulación del primer gen del operon. Y el 62% de las predicciones cayeron en la posición 1ª del operon, el resto fueron predicciones intergenicas las cuales estarían evidenciando nuevas unidades de transcripción dentro de estos operones.

Se graficó las calificaciones de las predicciones de promotores versus los índices de afectación de los genes a los que transcriben, y no se vio una correlación que indicara que en folds altos, es decir, que en genes con niveles de expresión altos tienen promotores predichos con una alta calificación. La gráfica de la figura IV.2 muestra los resultados de dos repeticiones del experimento de heat-shock, y en los cuales se predijeron promotores σ32, la gráfica muestra que no existe una relación entre folds de afectación bajos y calificación de la predicción del promotor.

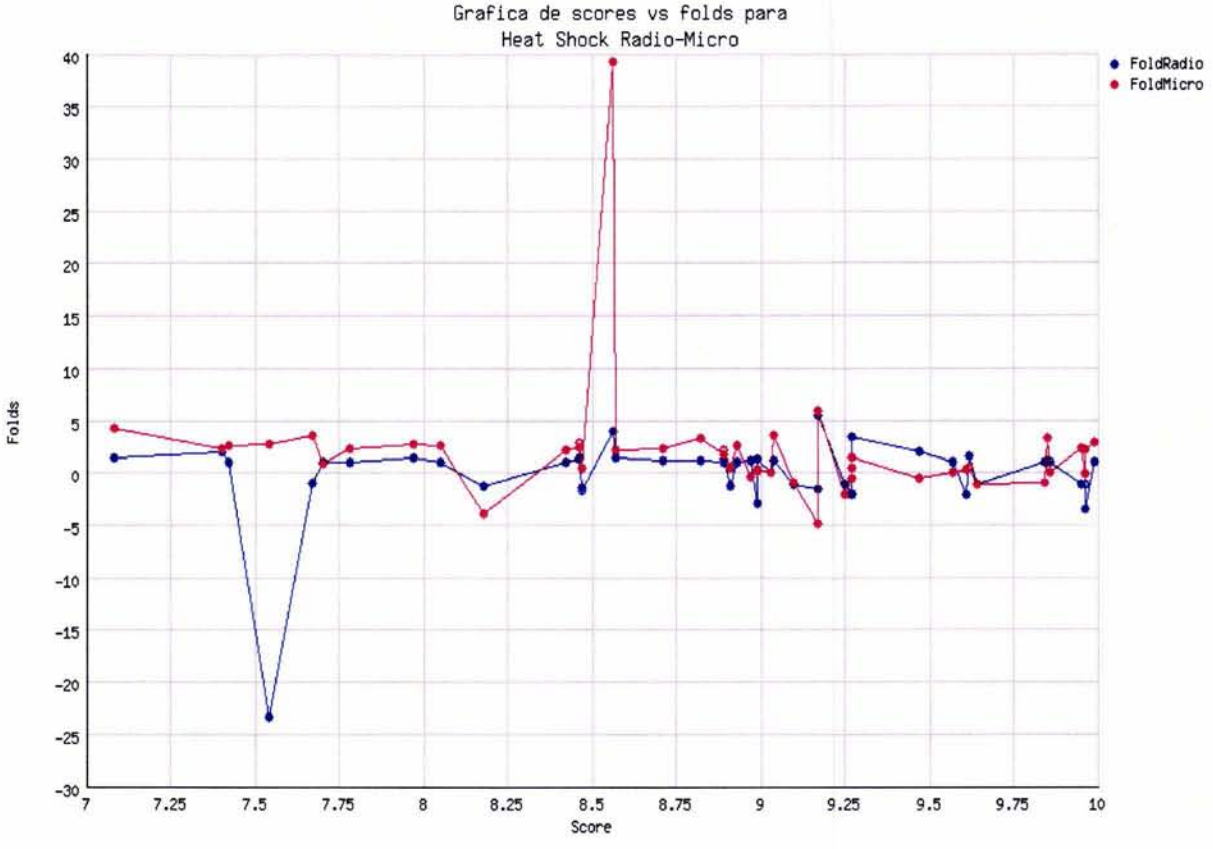


Figura IV.2. Genes inducidos en el experimento de hs usando micro arreglos vs. las calificaciones de los promotores σ32 predichos. Cada punto corresponde a un promotor. La distribución en rojo corresponde a los resultados del experimento de expresión global hibridizado en membrana de nylon. La distribución azul corresponde al experimento hibridizado en el microarreglo de vidrio.

IV.2.3 Congruencia en la expresión de genes en operones y en clases funcionales.

Cuando estuvimos analizando el conjunto de predicciones de promotores σ32 que fueron localizados por nuestros programas en regiones inter-operon, es decir promotores al interior de operones que pudieran estar definiendo unidades de transcripción internas, y su correlación con su calificación de homología a los consensi de σ32, nos encontramos con

un comportamiento en la expresión de los genes, agrupados en operones, que fueron afectados con folds altos. La figura IV.3 muestra el conjunto de operones que contienen al menos un gen afectado arriba o igual a ±5 fold.

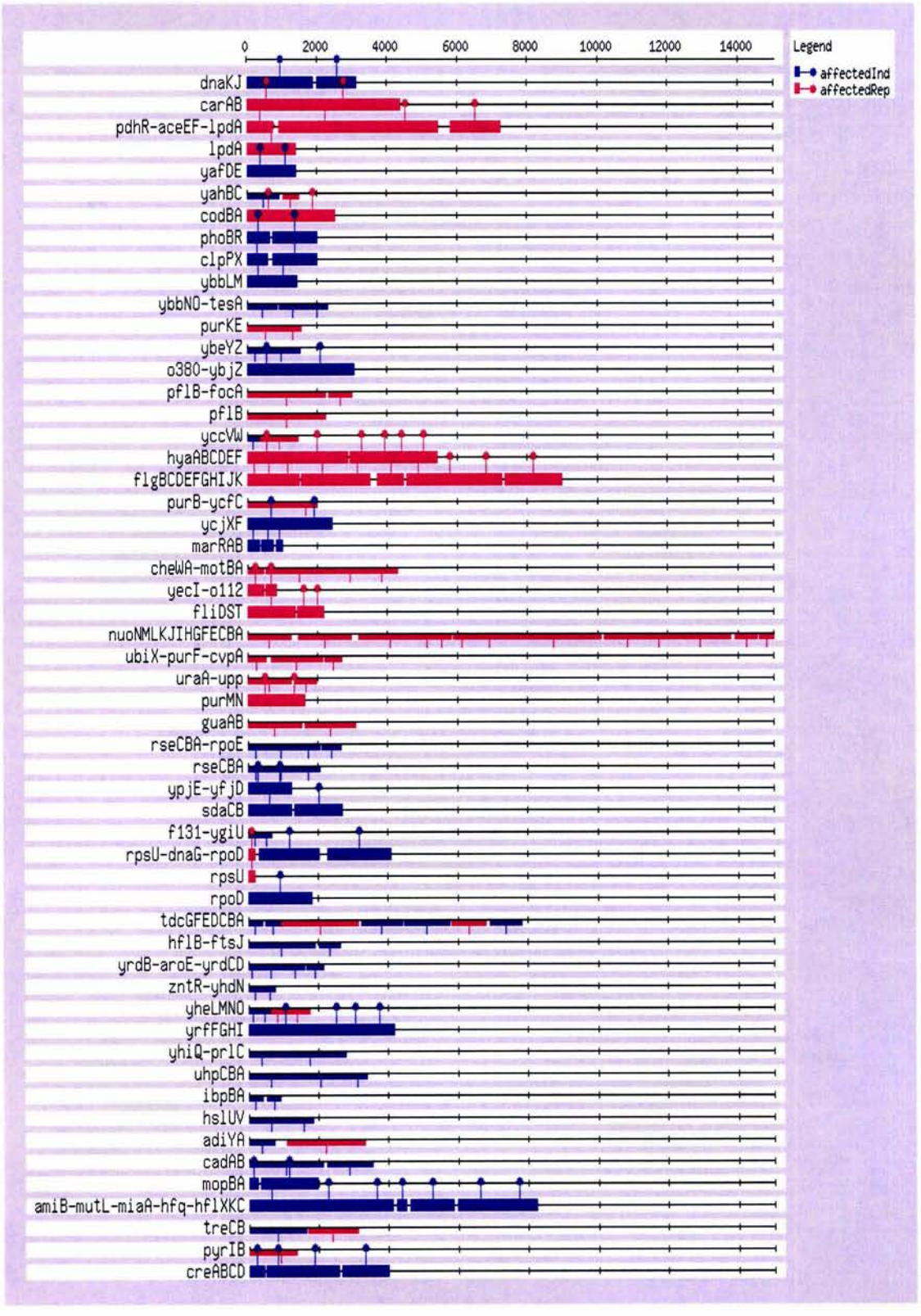


Figura IV.3. Gráfica de los operones afectados durante heat-shock. La gráfica fue hecha usando el programa *feature-map* [van Helden_2003] (http://embnet.cifn.unam.mx/rsa-tools).

Después de un heat-shock de 50°, 119 genes se detectaron como afectados con al menos ±5 fold [Richmond_1999]. Estos 119 genes dan cuenta por 100 unidades de transcripción de las cuales 52 son unidades de transcripción de mas de un solo gen, estas 52 agrupan a un conjunto de 153 genes. En la figura IV.3 se marca con color azul los genes que fueron inducidos y en rojo los genes reprimidos, es claro que la mayoría de las líneas que marcan las 52 unidades polisincrónicas de transcripción o son todas azules o son todas rojas, solo 6 líneas presentan una mezcla de colores indicando que esos operones contienen genes que están siendo reprimidos e inducidos al mismo tiempo, es decir operones cuya afectación no es *congruente*. Para cuestiones de la gráfica todo gen con fold negativo es considerado reprimido (rojo) y todo gen con fold positivo es considerado inducido (azul). Sobre esta premisa nosotros establecemos que usando un umbral de afectación de ±5 fold el experimento muestra un 88% de operones siendo *congruentemente* afectados.

Esta observación nos llevo a hacernos las siguientes preguntas:

a) ¿Que tan bajo puede ser el umbral de corte tal que aun podamos encontrar una alta congruencia?
b) Si agrupamos los genes por su clasificación de función fisiológica [Riley_1996], ¿veremos también congruencia como lo hicimos al agrupar a los genes en sus operones?'
c) ¿Estas observaciones son propias de la respuesta de la célula a heat-shock, o podríamos encontrar congruencia en otras respuestas celulares?

Las figuras IV.4 a y b muestran las gráficas de la congruencia en operones y clases fisiológicas para el experimento de heat-shock.

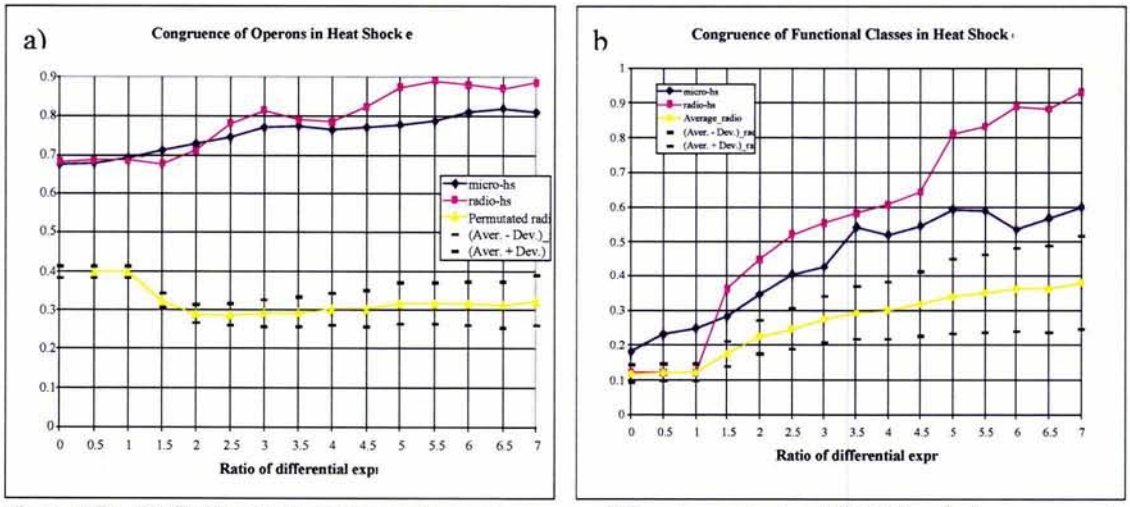


Figura IV.4. (a) Gráfica de la congruencia en operones en diferentes umbrales. (b) Gráfica de la congruencia en clases funcionales.

Las gráficas anteriores se generaron con las siguientes definiciones de medidas de congruencia.

MEDIDAS DE CONGRUENCIA.

Definimos dos medidas de congruencia que nos permitieran medir el comportamiento de los experimentos a lo largo de diferentes folds de expresión:

$$CongruenteOperon(fold_i) = \frac{\#\text{ de Operones completamente inducidos o reprimidos en } fold_i}{\text{Total de operones afectados en } fold_i}$$

y, (g)

$$CongruenteClase(fold_i) = \frac{\#\text{ de Clases completamente inducidas o reprimidas en } fold_i}{\text{Total de clases afectadas en } fold_i}$$

usando estas formulas analizamos 13 diferentes condiciones o respuestas celulares y el comportamiento de congruencia fue parecido que para heat-shock. Solo un pequeño número de clases fisiológicas u operones permanecen incongruentes cuando seleccionamos umbrales de folds altos. Los resultados son mostrados es las figuras IV.5 a y b.

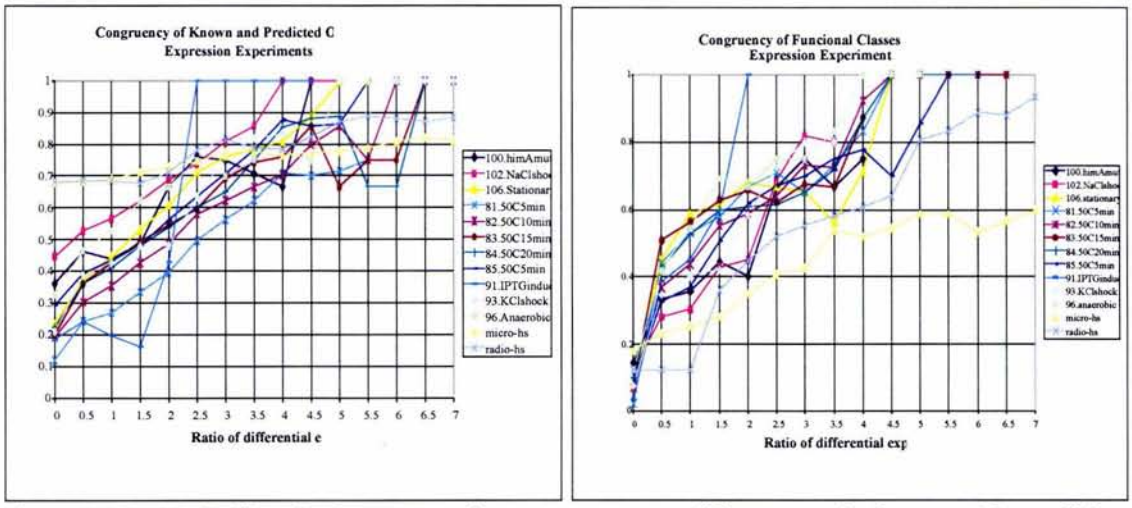


Figura IV.5. (a) Gráfica de la congruencia en operones en diferentes umbrales para 13 condiciones experimentales diferentes. (b) Gráfica de la congruencia en clases funcionales en diferentes umbrales para 13 condiciones experimentales.

Como un resultado de este análisis se diseñó una herramienta de web que permite la comparación de experimentos de micro arreglos con información reportada en la literatura. Esta herramienta hace uso de la información contenida en la base de datos RegulonDB [Salgado_2001]. El artículo anexo explica esta herramienta.

IV.3 USANDO LA FUNCIÓN "COVER" EN OTROS GENOMAS.

El alineamientos hechos para las sigmas endógenas de las mayoría de los organismos bacterianos secuenciados al momentos nos muestran que las regiones de contacto con las regiones consenso del promotor es ampliamente conservado. Esto indica que la estructura de los promotores se conserva permitiéndonos migrar nuestras metodologías de análisis a otros genomas bacterianos.

Para probar la función Cover en la predicción de promotores σ70 en otras bacterias seleccionamos dos genomas, uno muy cercano a *Escherichia coli*: *Salmonella typhimurium* LT2, y el otro alejado: el plásmido simbiótico de *Rhizobium etli* CFN42. La figura IV.6 muestra la posición de estos genomas en el árbol taxonómico generado de acuerdo a la clasificación de la base de datos de taxonomias de NCBI a Octubre del 2003 (http://www.ncbi.nlm.nih.gov:80/Taxonomy/CommonTree/wwwcmt.cgi)

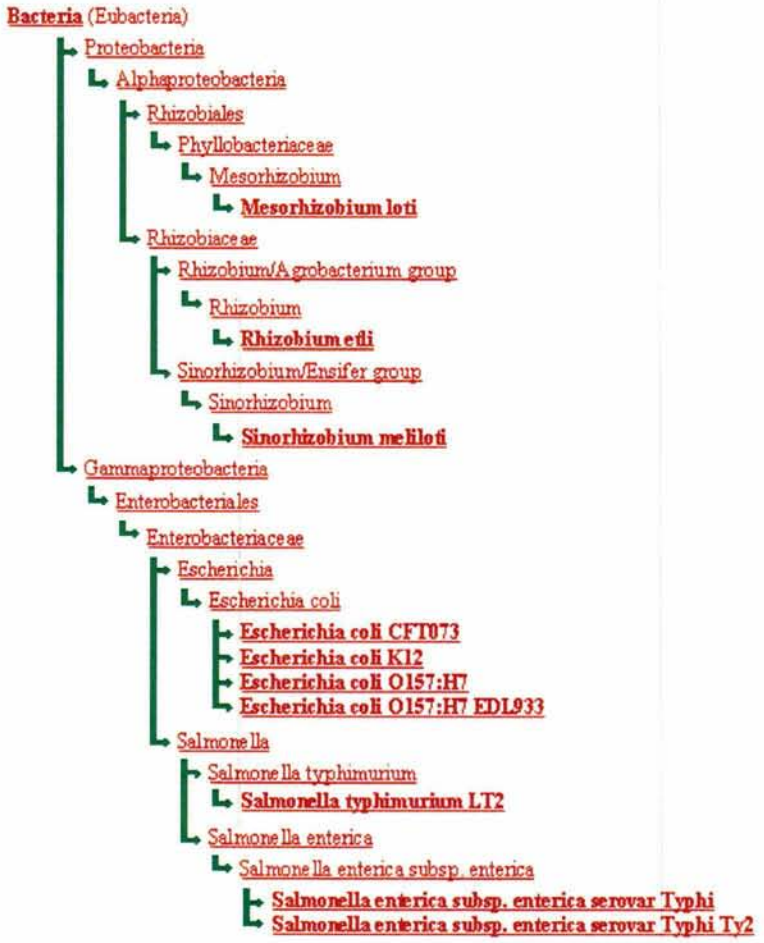


Figura IV.6. Árbol taxonómico que muestra la ubicación de *E. coli* K12 respecto a *S. typhimurium LT2* y al plásmido *R. etli* CFN42.

IV.3.1 *SALMONELLA TYPHIMURIUM LT2.*

La obtención y análisis del genoma de *Salmonella typhimurium* son relevantes ya que se trata de una bacteria causante de la gastroenteritis humana, así como es un modelo en ratones de la fiebre tifoidea que afecta a los seres humanos. *S. typhimurium* fue secuenciada, analizada y publicada en [McClelland_2001]. Hicimos la predicción de promotores para este genoma usando la función Cover, los promotores anotados están contenidos en [GenBank accession number AE006468]. Y presentamos aquí un breve resumen de los resultados.

Para las búsquedas de señales tipo promotor usamos el juego de matrices "Matrix_18_15_13_2_1.5" descrito en [Huerta_2003]. Se recalcularon los umbrales de corte calibrando las matrices con las frecuencias de las bases de las regiones estrictamente no-codificantes del genoma *S. typhimurium*. Se analizó si con las matrices calibradas el fenómeno de las densidades localizadas en regiones estrictamente no-codificantes visto en *E. coli* también se presentaba en *S. typhimurium*. La figura IV.7 muestra la densidad promedio de señales tipo promotor en regiones codificadoras y en regiones no codificantes.

De la figura se puede observar que *S. typhimurium* muestra el mismo comportamiento visto para *E. coli* (figura 5 de Huerta_2003). *S. typhimurium* tiene 4450 ORFs, 28% de ellos no están en *E. coli* [McClelland_2001]. De los 3240 restantes, el 77% de los genes tiene predicciones de promotor por Cover en ambos genomas. 2476 predicciones están en la misma posición para 1234 genes es decir el 40% de los genes. La distancia entre *E. coli* y *S. typhimurium* es del 87%, esta distancia está basada en el parecido de todos los genes ortólogos [Moreno_Hagelsieb_2002].

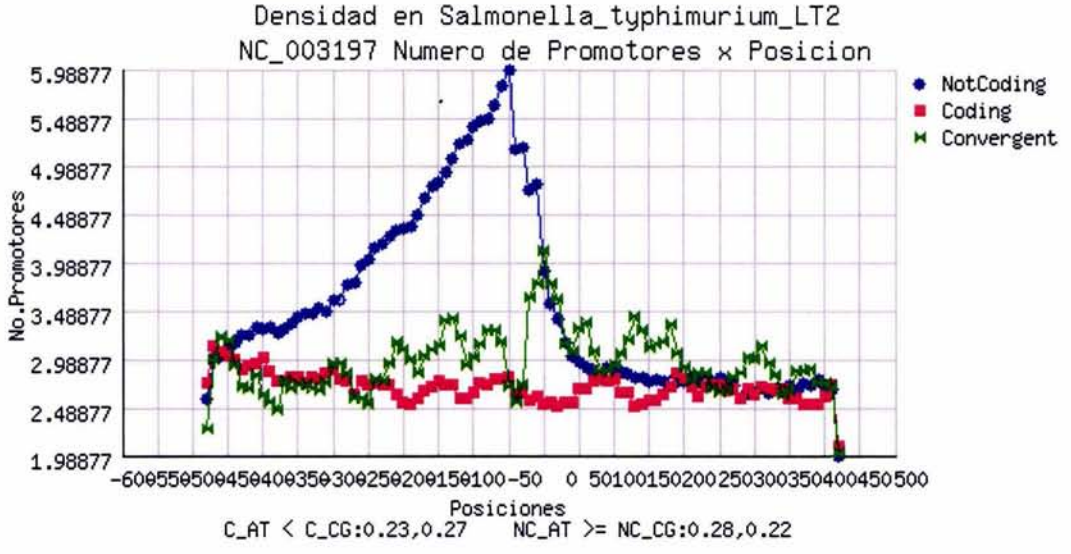


Figura IV.7. Densidad promedio de *señales_promotor* en regiones codificadoras y en regiones no codificantes de *Salmonella typhimurium*. Después de hacer la búsqueda usando las matrices de peso y umbral de la $\mu - 3\sigma$, se alinearon las regiones no-codificantes respecto al inicio de la traducción y se obtuvo el promedio de las señales en intervalos de 10 bp (curva azul). Para la distribución de las regiones codificadoras, en color rojo, se tomaron genes mayores de 1000 bp, el 0 hace referencia a la posición media adentro de cada gen, tal que la gráfica representa las primeras 1000 bases centrales de los genes. Las regiones convergentes se usaron como control ya que son regiones estrictamente no-codificantes en las que no debería haber un promotor. El pico al final de los genes es explicado por interacciones de la RNAP con el final de los mismos.

Se buscó en la literatura análisis comparativos de regiones promotoras entre *S. typhimurium* y *E. coli* para comparar estos resultados con los arrojados por la función Cover. En un estudio comparativo del gen *crp* en diferentes genomas se demostró que este gen codifica para una de las proteínas mas conservadas que se han observado hasta el momento [Cossart_1986]. Hay un aminoácido de diferencia entre la secuencia de aminoácidos del *crp* de *S. typhimurium* y la de *E. coli*, y 77 cambios en la secuencia de nucleótidos. Aquí presentamos un análisis comparativo de la región corriente arriba del gen *crp* en ambos genomas usando los resultados de la función Cover. La figura IV.8 muestra que las predicciones encontradas son casi las mismas para ambas regiones sugiriendo que la región de regulación es también bastante conservada.

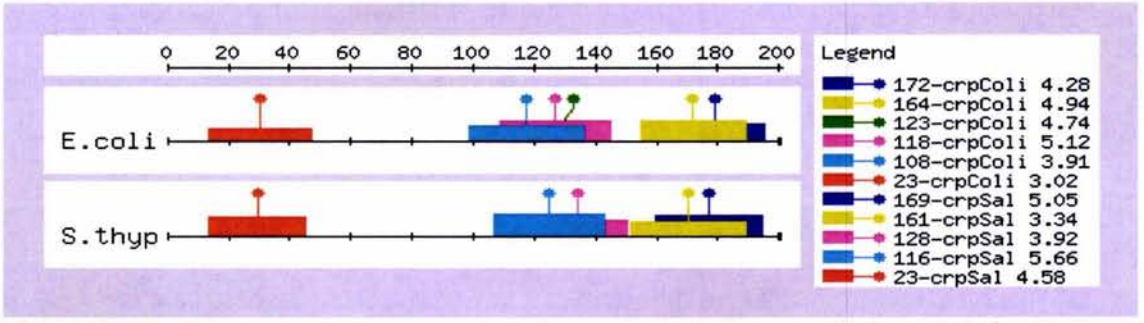


Figure IV.8. Región promotora del gen *crp* en *S. typhimurium* y *E. coli*. Las predicciones de la función Cover son casi iguales en ambas regiones. *E. coli* presenta un promotor mas que *S. typhimurium*, con una caja -10 a -123 bp del codón de inicio.

Un análisis comparativo de la región promotora del gen *fliA*, el cual codifica para el factor σ28 (*flagellum*), en *Salmonella typhimurium* mostró que el operon *fliAYZ* es transcrito desde un promotor σ70 y un promotor σ28 [Ikebe_1999]. El inicio de la transcripción del promotor σ70 fue reportado a -29 bp arriba del codón de inicio del gen *fliA*. Ese estudio también comprobó que, como en *E. coli*, en presencia de las proteínas FlhD y FlhC el promotor es transcrito, lo que sugiere que funcionalmente la región promotora se conserva en ambos genomas. La figura IV.9 muestra los resultados de la función Cover en las regiones promotoras de 200 bp arriba del gen *fliA en S. typhimurium* y *E. coli*.

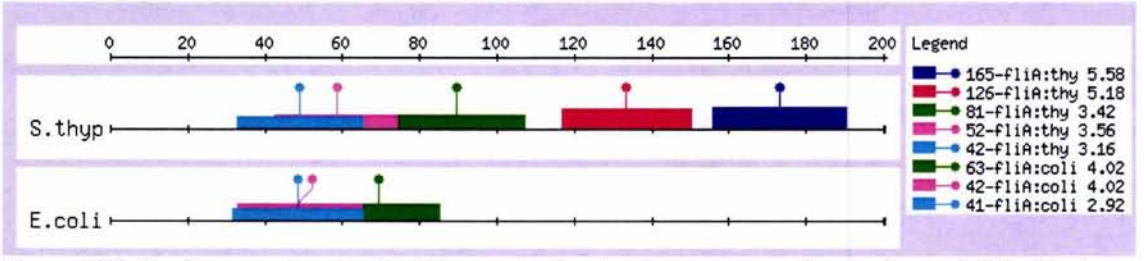


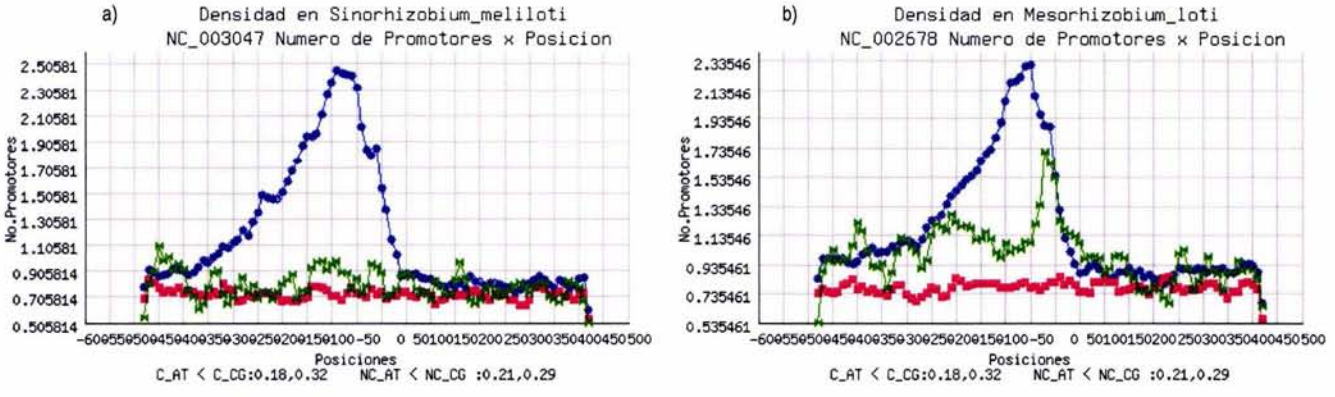
Figura IV.9. Región promotora de *fliA*. El promotor funcional es representado en color azul cielo. El cluster de 3 promotores superpuestos en la región del promotor descrito experimentalmente coincide con los resultados observados para los promotores σ70 de *Escherichia coli* [Huerta_2003]. El promotor funcional en *S. typhimurium* tiene la caja -10 localizada a -41 bp del codón de inicio de *fliA*, y en posición -42 para *E. coli*. Dos promotores adicionales son encontrados en el gen *fliA* de *S. typhimurium* uno con la caja -10 en posición -126 y otro con una caja -10 a -126 bp arriba del inicio del gen.

De las anteriores figuras podemos observar que para las regiones de regulación conservadas en ambos genomas, la función Cover encuentra predicciones similares. Esto nos puede ayudar para anotar los promotores reportados en de *E. coli* en *S. typhimurium* con un grado de certeza. Del árbol de la figura IV.6 sabemos que *S. typhimurium* y *E. coli* son genomas cercanos y que la homología entre ellos es alta. Nos interesa ahora saber como se comportara la función Cover en un genoma no tan cercano a *E. coli*.

IV.3.2 EL PLÁSMIDO SIMBIÓTICO DE *RHIZOBIUM ETLI* CFN42.

Rhizobia es un término que congrega las especies *Rhizobium*, *Sinorhizobium*, *Mesorhizobium* y *Bradyrhizobium*. Estas son bacterias simbióticas que interactúan con las raíces de las legumbres e inducen la formación de nódulos que fijan nitrógeno, el cual es necesario para el crecimiento de las plantas y su desarrollo. Se sabe que el nitrógeno es un nutriente limitante del crecimiento de la planta del fríjol, el alimento básico en muchos de los países latinoamericanos, entonces el entendimiento del proceso biológico de la fijación del nitrógeno es importante por razones económicas y humanitarias [CIFN_2001].

En los Rhizobia, los genes esenciales para la simbiosis están contenidos en los plásmidos simbióticos o en islas simbióticas dentro del cromosoma. La secuencia completa del plásmido simbiótico de *Rhizobium etli* CFN42 ha sido obtenida [González_2003]. Predicciones de promotores fueron hechas en este plásmido con nuestros programas y aquí presentamos brevemente los resultados de la función Cover. Del árbol taxonómico de la figura IV.6 podemos observar que *R. etli* no es tan cercano a *E. coli* como lo es *S. typhimurium*. La distancia en términos de homología entre ortólogos entre *Bradyrhizobium japonicum*, *Sinorhizobium meliloti*, *Mesorhizobium loti,* y *Escherichia coli* es del 32%, 33%, y 32% respectivamente. Podría esperarse que el fenómeno de las densidades localizado en las regiones estrictamente no-codificadoras del genoma de *E. coli* no se presentase en genomas alejados a él. Usando el juego de matrices "Matrix_18_15_13_2_1.5", descrito en [Huerta_2003], se corrieron búsquedas en las distintas regiones de los genomas de *Sinorhizobium meliloti* y *Mesorhizobium loti*. La figura IV.10 (a y b) muestra los resultados de estas búsquedas.



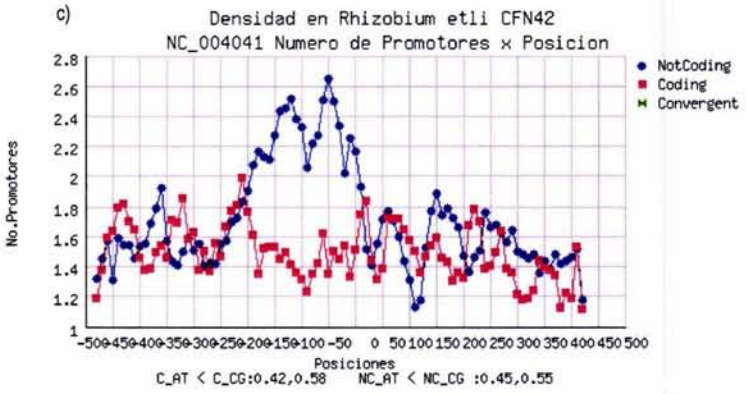


Figura IV.10. (a,b). Densidad promedio de *señales_promotor* en regiones codificadoras y en regiones no codificadoras de *S. meliloti* y *M. loti*. Como con *S. typhimurium*, las matrices "Matrix_18_15_13_2_1.5" fueron calibradas a las frecuencias de las bases presentes en las regiones codificadoras de los dos genomas analizados. La diferencia entre las densidades en regiones codificadoras y no-codificantes es evidente. (c) Densidad promedio de *señales_promotor* en regiones codificadoras y en regiones no-codificantes del plásmido simbiótico CFN42. Las búsquedas fueron hechas utilizando matrices generadas con 34 promotores de Rhizobias. La diferencia entre las regiones codificadoras y no-codificantes en términos de densidad de señales no es muy limpia mas allá de las 200 bp arriba de los genes. La distribución de las señales en las regiones divergentes no se muestra ya que solo 16 regiones presentaron la característica de ser regiones divergentes de mas de 50 bp.

Se obtuvo de la literatura secuencias de promotores en rhizobias que evidenciaran la existencia promotores tipo σ70 en estos organismos. SigA ha sido identificado en varias rhizobias como el factor de mantenimiento endógeno que reconoce promotores del tipo –35/-10 de *E. coli* [Beck_1997]. La figura IV.11 muestra los consensi de 34 promotores de rizhobias, el cual, como ya reportado en la literatura, resulta muy parecido al de *E. coli*. SigA es considerado un factor de la familia σ70 con un porcentaje de identidad de 50% entre el σ70 y sigA.

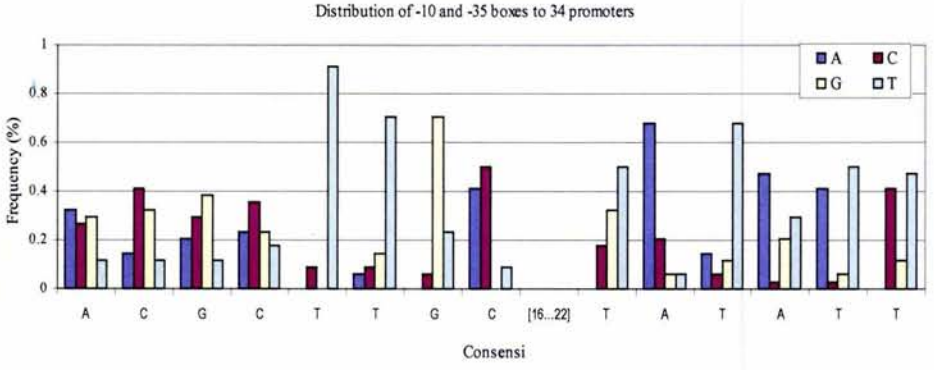


Figure IV.11. Consensi de 34 promotores tipo σ70 encontrados en diferentes Rhizobias.

Búsquedas en las regiones codificadoras y estrictamente no-codificantes del plásmido CFN42 fueron hechas con las matrices obtenidas de los 34 promotores para verificar el fenómeno de las densidades en el plásmido. La figura IV.10 (c) muestra distribuciones más ruidosas que las presentadas en los cromosomas, (a) y (b). Las

secuencias promotoras de los genes simbióticos carecen de similitud con los promotores σ70 de *Escherichia coli*. Los genes de fijación de nitrógeno *nif* son reconocidos por la RNApol requiriendo el factor σ54 [Beck_1997]. Ya que muchos de los genes involucrados en la simbiosis se encuentran en los plásmidos parecería lógico que no existan muchas señales reconocidas por el "house-heeping" sigA en el plásmido CFN42.

CAPÍTULO V. DISCUSIÓN, CONCLUSIÓN Y PERSPECTIVAS.

V.1 DISCUSIÓN.

Hasta ahora se ha pensado que el promotor es una región bien definida de aproximadamente 60 bp la cual contiene dos hexámeros bien definidos: la caja -35 y la caja -10 separadas por secuencias variables de 15 a 21 bp de tamaño. Pero lo que hemos visto aquí es que los promotores existen en clusters de señales tipo promotor, como una serie de sitios de pegado de la RNAP potencialmente competentes. También observamos que una alta densidad de señales tipo promotor está localizada en las regiones no-codificantes del genoma de *E. coli* vs. aquella vista en las regiones codificantes, indicando que estas señales tienen un papel funcional y evolutivo en el contexto genómico.

Esta nueva visión de las regiones de regulación de un genoma replantea el problema de reconocer por medios computacionales una señal promotor funcional. El problema se convirtió en uno que plantea la pregunta de cómo elegir una señal promotora funcional entre un conglomerado de señales tipo promotor. Nosotros no sabemos realmente como se da este reconocimiento específico del promotor por la RNAP en medio de este cúmulo de señales, sabemos que existen muchos factores que afectan este proceso de reconocimiento, como son la presencia de proteínas activadoras, el supercoiling del DNA, la curvatura del DNA producida por la presencia de A-tracks (polyAAAA), concentraciones de salinidad en el medio ambiente, etc. La solución aquí presentada es una estrategia que no resuelve el problema al 100%, pero es una herramienta que nos permitió estudio *in silico* más detallado de las secuencias promotoras.

El primer punto interesante con el que nos encontramos en el análisis de los promotores en *E. coli* fue que no siempre el promotor mejor calificado en una región es el promotor descrito experimentalmente. Posibles razones pueden deberse a la naturaleza de los promotores analizados y a al contexto en el que se encuentran, es decir, la secuencia del promotor puede estar influida por el tipo de regulación a la que este está sometido y por la presencia de otros elementos adicionales como el motivo TG de la caja -10 extendida o el elemento UP. Sin embargo, los elementos adicionales como el UP o el TG de la caja extendida no mostraron ser muy útiles en *E. coli* para discriminar entre un conjunto de posibles candidatos al promotor funcional. Creemos que son las proteínas reguladoras las que determinan cuál es el promotor que deberá "usarse" mediante la inhibición de los otros sitios de pegado del promotor al dirigir a la RNAP al promotor apropiado. La región promotora del gen *lac* es el ejemplo idóneo que evidencia esta conclusión [Reznikoff_1992]. El promotor *lac* (P1) coexiste con 5 señales tipo-promotor que se superponen, y es la proteína activadora CRP la que le permite a la RNAP el transcribirlo. En ausencia de CRP los promotores *lac* P2 y P3 son transcritos, y los 3 restantes se vuelven activos por una simple mutación en cualquiera de ellos, representando lo que nosotros hemos llamado vestigios de promotores producto de la evolución. Lo que hemos demostrado aquí es que este fenómeno de señales promotoras superpuestas no es exclusivo del promotor *lac*, si no que es un fenómeno presente en los promotores bacterianos.

El segundo punto interesante es que el espectro de las calificaciones de homología al consenso mostrado por los promotores funcionales es muy amplio, variando hasta 7 veces. Esto genera un problema muy serio a la hora de determinar los umbrales de corte a

partir de los cuales podemos decir que una secuencia blanco es o no una señal promotor. Entre mejores sean las matrices consensi que se construyan para censar regiones en busca de promotores putativos, menos variabilidad en las calificaciones de los promotores puede encontrarse. La construcción de las matrices de peso, o sensores, es clave en cualquier método de predicción de promotores.

Y tercero, se piensa que promotores regulados positivamente tiene señales -35 muy débiles, o que promotores con señal -10 extendida pueden incluso no tener una caja -35. También se piensa que un promotor regulado negativamente tiene señales -10 y -35 fuertes. En la figura V.1 se muestra las calificaciones promedio de las cajas -10 y -35 por tipo de regulación. El promedio de las calificaciones de los 116 promotores del conjunto de entrenamiento nos muestra que las cajas -35 no presentan ninguna variabilidad con respecto al tipo de regulación. Sin embargo, las cajas -10 de los reprimidos (repressor) muestran una calificación promedio más alta que las de los duales y activados, en una razón aproximada de 1.25 veces. Hace sentido pensar que un promotor que es estrictamente reprimido debería tener señales -35 y -10 buenas tal que, sin el represor presente se comporte como un promotor constitutivo. En general podría decirse que si un promotor regulado negativamente no tiene calificaciones -10 y/o -35 buenos, entonces debe haber ahí algún otro mecanismo no reportado que de cuenta por la transcripción en ausencia del represor.

Figura V.1. El promedio de las calificaciones de los 116 promotores del conjunto de entrenamiento visto por elemento del promotor. Los 116 promotores fueron agrupados por el tipo de regulación a la que son sometidos de acuerdo a la información reportada en RegulonDB. El conjunto de los promotores reprimidos es mostrado en color rojo en la figura, e incluye solamente aquellos promotores que son estrictamente regulados negativamente, es decir, no existe reportado en la literatura otro mecanismo de regulación que no sea el de represión. El conjunto de promotores regulados únicamente por mecanismo de inducción se presenta en color azul. El conjunto de promotores que son regulados positiva y negativamente se presenta en color amarillo, los mecanismos de inducción y represión se activan dependiendo de la condición celular.

En cualquier modo de regulación las cajas -10 muestran una calificación promedio más alta que el de las cajas -35. De la figura VI.7b de la sección de Materiales y Métodos, se puede ver que la distribución de las calificaciones de la caja -10 se encuentra sesgada a la derecha indicando que no hay muchas calificaciones débiles y que las cajas -10 se encuentran muy conservadas.

La distribución de las calificaciones de las cajas -35 indica la presencia de varias calificaciones bajas. De acuerdo a la teoría de Contenido de Información, calificaciones negativas representan secuencias que no pertenecen al conjunto de patrones que definen al consenso. Es decir una caja -35 negativa representaría la ausencia de esa señal, esto quiere decir que para esos promotores debe haber un elemento extra que ayude a la RNAP a

anclarse en el promotor una vez que ya reconoció la caja -10, estos elementos pueden ser sitios de pegado de proteínas activadoras, y en muy pocos casos la -10 extendida, o quizá un elemento UP. La tabla V.1 muestra los promotores del conjunto de entrenamiento con cajas -35 negativas o muy alejadas del consenso.

Promotor	Calificación -35	Caja -35	Calificación -10	Caja -10	Calificación -10 + -35
phoHp1	0.93	CACTGTCAT	0.61	AGTAGAAAC	1.54
glnL	0.90	TGCTTCGCG	6.24	GCTATAATG	7.14
glpD	0.70	AATGTTACC	5.10	GCTAATATG	5.8
glnBp2	0.65	GGGGCGCAA	4.45	TATAAACTG	5.1
argE	0.63	CGGCTGGTG	4.28	GTTAGTGTA	4.91
purFp1	0.52	ATCCCTACG	5.92	GTTAGAATG	6.44
glnAp1	0.49	AAGGTCATT	4.17	CTTAATGTT	4.66
focAp2	0.38	ATGTAGGCT	5.86	GTTATATTA	6.24
dmsA	0.05	TAATTACTC	1.66	ACTACTTTC	1.71
narKp1	-0.03	TTACATCAA	2.66	ACTAAGGTG	2.63
tsxp2	-0.13	AACGAAACA	5.96	TTTATAATA	5.83
malPQ	-0.13	GATGAGGAA	3.14	GGCAAACTA	3.01
ccd	-0.19	AGATCACAT	4.38	CGTAAACTG	4.19
purH	-0.30	CGAGCGTTG	5.73	GTTACAATG	5.43
nirB	-0.31	CAATAAGCG	4.13	GTTAAGGTA	3.82
ycfCp	-0.33	AACGGTGAA	5.60	GGTATTATT	5.27
ilvY	-0.45	TGACAGGAA	4.57	tGATATATTC	4.12
rpoHp5	-0.68	CACGGTCTG	3.51	tGATAACCTC	2.83

Tabla V.1. 18 de los 116 promotores del conjunto de entrenamiento tiene calificaciones bajas para las cajas -35, incluso 10 de ellos tienen calificaciones negativas.

De esta tabla puede verse que, por ejemplo, el promotor *rpoHp5* tiene una caja -10 extendida, los promotores con -10 extendida pueden prescindir de una caja -35 sin ningún problema ya que el motivo "TG" en -15 y -14 respectivamente anclan a la polimerasa como lo haría una caja -35 [Gross_1998], *rpoHp5* tiene reportado un mecanismo negativo de regulación dado por los complejo de proteínas formado por cAMP-CRP y CytR, y un mecanismo de activación por CRP en -48. El promotor *ilvY* tiene también una caja -10 extendida y sufre auto-regulación negativa teniendo un sitio para el represor IlvY en -18. El promotor *ycfCp* tiene reportado un mecanismo negativo ejercido por PurR a 837.5 bp del promotor, este promotor tiene una caja -10 muy conservada, solo 1 substitución desde el consenso, se sugiere aquí que este promotor debe tener asociado algún mecanismo de regulación positiva que aun no se ha reportado.

El problema de reconocer promotores endógenos es un problema complejo y representa aun un reto para las ciencias computacionales el resolverlo. En este trabajo se presentaron algoritmos que para el reconocimiento de promotores de la familia $\sigma 70$. Estos algoritmos están contextualizados a secuencias arriba de un inicio de gen y usa la información de la distancia del promotor al codón de inicio como elemento de análisis. Este recurso es ficticio ya que no existe un mecanismo biológico mediante el cual la RNAP cense esta distancia. Cuando enfrentados con una secuencia intergenica los algoritmos determinan cual es la mejor(es) subsecuencia a ser promotor, pero no son capaces de decir cuando una secuencia por sus características no puede tener un promotor. Esfuerzos en este sentido mejorarían la especificidad de los algoritmos aquí presentados. Pero mejor aun, el

diseño de programas que emulen el proceso biológico del reconocimiento del promotor por la RNAP ayudaran a entender mejor como este se lleva a cabo.

V.2 Conclusión y perspectivas.

Pensar en los promotores como un cúmulo de señales del mismo tipo compitiendo con el promotor "funcional" por el pegado de la RNAP nos puede dejar un mal sabor de boca si nuestra tarea es la de estudiar las propiedades bioquímicas de un promotor de manera experimental, o si nuestra tarea es la de reconocer de manera inequívoca al promotor funcional por medios computacionales y/o analíticos. Las complicaciones que esta idea de promotores sobrepuestos supone a la biología experimental son [Reznikoff_1987]: mutaciones las cuales aumentan o no alteran la expresión del gene estudiado podrían estar desenmascarando un promotor latente; y pruebas bioquímicas que miden la actividad de un promotor se complican en la presencia de reacciones de competencia inesperadas entre la RNAP con otras señales promotoras funcionales. Las complicaciones prácticas a la que los métodos computacionales se enfrentan son: generalmente el promotor funcional reportado en la literatura califica mas bajo que algunas de las señales tipo promotor con las que se superpone; y si una simple mutación en un nucleótido puede dar vida a un promotor que no era funcional, cómo puede cualquier metodología ser capaz de medir esto para decrementar la calificación asignada a esa señal tipo promotor tal que pueda ser discriminada como un falso promotor.

Por otro lado esta nueva perspectiva de lo que un promotor es nos abre la puerta a nuevas líneas del pensamiento que expliquen el porqué los promotores existen dentro de clusters de potencialmente funcionales y completamente funcionales promotores. Quizás este cúmulo de señales promotoras latentes y funcionales nos este hablando de la robustez del proceso de la transcripción a mutaciones que el DNA puede sufrir al azar. Este conjunto de señales disponibles capaces de ejecutar la misma función ofrece una vasta gama de recursos que la célula puede usar ante cambios que puedan afectar el mecanismo de la transcripción en una región promotora particular.

También podemos pensar de los promotores latentes, esas señales potencialmente funcionales que con una mutación cobran vida, como vestigios evolutivos de promotores que nos estarían hablando de la existencia de mecanismos de transcripción que por alguna razón dejaron de funcionar. Pudiera ser que la transcripción en ese promotor del gen corriente abajo, estando en una célula de un organismo ancestro, se diera por la presencia de factores de la transcripción o debido a condiciones particulares que no existieron en la célula a la que el gen emigró, forzando así una mutación que permitiera el desenmascaramiento de un nuevo promotor que diera cuenta por la transcripción de ese gen.

Las señales completamente funcionales, esas que compiten con el promotor mapeado experimentalmente pero que no se han reportado en la literatura, pueden estar presentes para responder a diferentes condiciones ambientales en las cuales el promotor "conocido" no puede realizar su función. En estos casos factores de la transcripción como las proteínas reguladoras o las proteínas que cambian la estructura del DNA deben estar jugando un papel clave para la selección del promotor a transcribir.

La bondad en términos computacionales de la presencia de ese cúmulo señales es que nos permitirá distinguir inequívocamente en donde se encuentran las regiones inicio de unidades de transcripción en un genoma. La densidad de señales tipo promotor solo se da en ese tipo de regiones. La detección de esas regiones ayudara al biólogo experimental a seleccionar las regiones en donde vale la pena hacer experimentos de mapeo de promotores.

Trabajos futuros encaminados a detectar solo a los promotores funcionales en estos conglomerados de señales putativas, ofrecen un reto tanto experimental como teórico. Análisis del comportamiento de los promotores en otros genomas bacterianos en donde no haya datos experimentales ni métodos de predicción de promotores implementados representan un trabajo futuro que ayudará a la anotación de esos genomas. Inferir el papel de la regulación mediada por las proteínas en los diferentes genomas dependiendo del arreglo que formen los promotores en el DNA, si existen densidades de señales o no, es también un trabajo que aportará nuevo conocimiento a las ciencias genómicas.

En *E. coli* seria interesante conocer si existen diferencias al nivel de las matrices consensi de las cajas –10 y –35 entre promotores regulados positiva, dual y negativamente así como verificar la aseveración que dice que promotores constitutivos no existen. Este análisis iría encaminado a sentar las bases de un modelo de predicción del tipo regulación que un promotor puede sufrir dependiendo solo del análisis de su secuencia.

CAPÍTULO VI. MATERIALES Y MÉTODOS.

Se realizó una recopilación bibliográfica referente al mecanismo del inicio de la transcripción mediada por la RNAP, y los factores σ70 y σ32 que permitieran un análisis del proceso del reconocimiento del promotor y el diseño de algoritmos para el reconocimiento de estos inicios.

Para el desarrollo de los algoritmos se utilizaron dos diferentes conjuntos de secuencias promotoras reguladas por σ70, un conjunto de entrenamiento de 116 promotores de la colección publicada en [Gralla_1996], y un conjunto de prueba de 392 promotores, ambos colectados en la base de datos RegulonDB (http://www.cifn.unam.mx/Computational_Biology/regulondb). Para σ32 se tomo de la literatura un conjunto de 22 promotores heat-shock, esta colección también se encuentra en RegulonDB. Se extrajo de RegulonDB la información de las posiciones absolutas en el genoma de *E. coli* K12 del inicio de la transcripción de cada uno de los promotores mencionados. La tabla VI.1 muestra los tipos de sigmas y el número de promotores transcritos por éstas y que están colectados en RegulonDB.

Tipo de Sigma	Número de promotores regulados
sigma19	1
sigma24	4
sigma28	3
sigma32	26
sigma38	46
sigma54	12
sigma70	580
total	672

Tabla VI.1. Factores sigma recolectados en RegulonDB, y número de promotores regulados por cada sigma.

Las evidencias que RegulonDB tiene para los promotores utilizados en este trabajo pueden verse en la tabla VI.2.

a)

Tipo de Evidencia	Número de Promotores
Identificación de las cajas -35 y/o -10	68 (58)
Mapeo del inicio de la transcripción & "footprinting" con la RNA polimerasa	4 (3)
Mapeo del inicio de la transcripción & Identificación de las cajas -35 y/o -10	164 (127)
Mapeo del inicio de la transcripción	165 (86)
Total	401

b)

Análisis del conjunto de promotores σ70	#Promotores con evidencia
Conjunto de entrenamiento (116 promotores)	114
Conjunto de prueba (392)	160
Total	274

Tabla VI.2. (a) Evidencias de los promotores en colectados en RegulonDB. (b) Número de evidencias que se tienen en RegulonDB para cada una de las colecciones usadas en este análisis. RegulonDB incluye entre sus promotores las colecciones publicadas en [Hawley_1983, Harley_1987, Lisser_1993].

En RegulonDB solo se tiene reportada la base del inicio de la transcripción para los 584 promotores tipo σ70. Se hizo una análisis de las distancias de estos inicios de la transcripción al inicio de la traducción, es decir al codón de inicio del gen. La figura VI.1 muestra esta distribución para la colección completa de promotores sin considerar su tipo de sigma asociada y la distribución de solo los promotores reconocidos por el factor σ70.

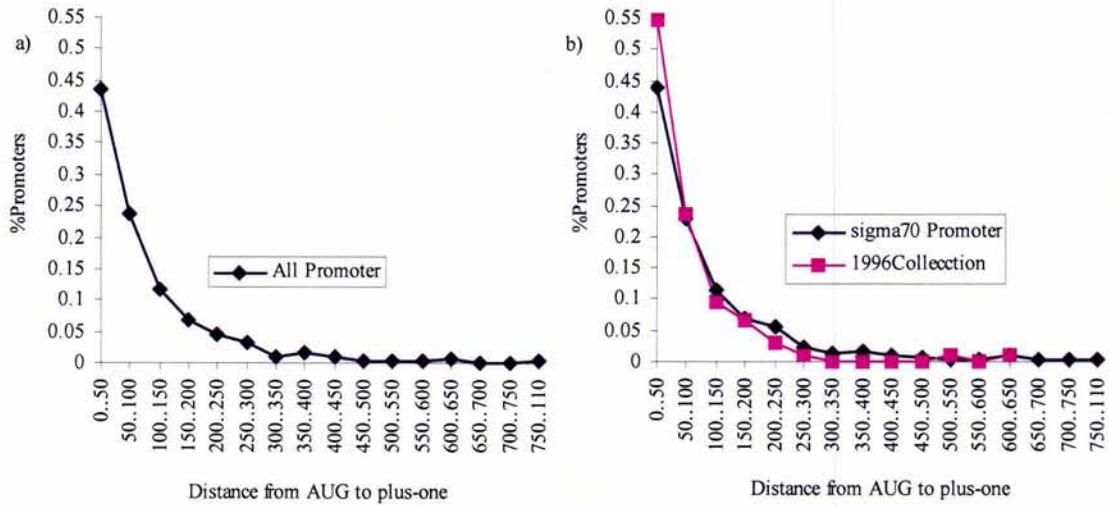


Figura VI.1. Distribución de las distancias entre el mas-uno de los promotores y el codón de inicio de los genes transcritos. En (a) la distribución de las distancias de los 678 promotores reportados en RegulonDB. En (b) la distribución de los 584 promotores tipo σ70 en azul, y la distribución de un subconjunto de estos (116 promotores) que conforman el conjunto de entrenamiento - curva rosa -.

Puede verse que existe una marcada preferencia de los promotores a estar a -50 bp del inicio del gen. El 90% de los promotores se encuentra ubicado en las primeras 200 bases arriba de un gen. Tabulamos las distancias del +1 al inicio del gen en intervalos de 10 pares de bases. Hicimos 20 grupos desde el intervalo [0-10] hasta el intervalo [240-250] y fue calculada la probabilidad relativa de cada uno, ver formula (h),

$$Score\_Position(i) = \frac{\text{número de promotores en el intervalo } i}{\text{número total de promotores analizados}} \qquad \text{(h)}$$

Por lo tanto, las regiones favorecidas son aquellas cuya cercanía es de 100 pares de bases al inicio del gen.

También en *RegulonDB* se consulto el modo de regulación de los 584 promotores σ70. La gráfica de la figura VI.2 muestra que casi la mitad de los promotores reportados en la literatura no tiene un modo regulación asociado.

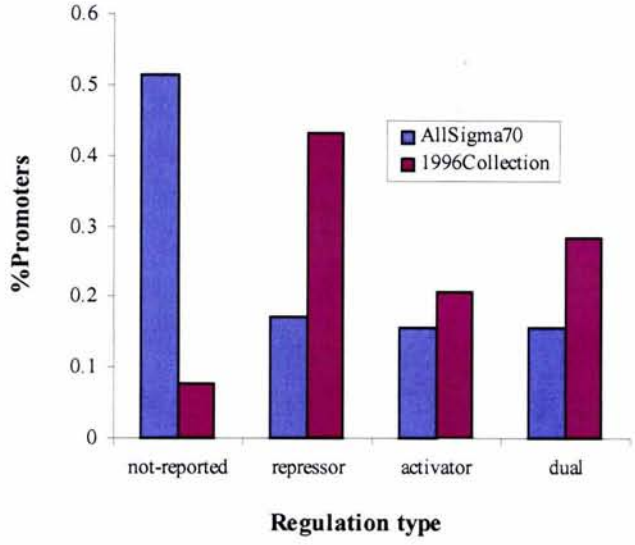


Figura VI.2. Distribución de los modos de regulación de los promotores σ70 colectados en RegulonDB. En rojo se presenta la información para la colección de entrenamiento de 116 promotores.

En la tabla VI.3 puede verse las 116 secuencias del conjunto de entrenamiento - publicada en [Gralla_1996] - y el modo de regulación de cada uno de estos promotores. Este conjunto de promotores fue elegido arbitrariamente para ser el conjunto de análisis que determina los elementos y argumentos que nuestro algoritmo de reconocimiento de promotores usa.

Promotor	RegulonDB referencia	Tipo de regulación	Mas-uno	Cadena	Secuencia del Promotor (alineación hecha con el programa *wconsesus*)			
pyrD	ECK120009425	repressor	1003955	+	CCGCAGGTCAATTCCCT	tttggtccgaactc	GCACATAATA	c
fabA	ECK120009421	activator	1015721	-	CTGATCGGACTTGTTCA	gcgtacacgtgtt	AGCTATCCTG	cgtg
ompA	ECK120009446	repressor	1019414	-	CGGAGTTCACACTTGTA	agttttcaactacg	TTGTAGACTT	t
sulA	ECK120009441	repressor	1020168	-	GAAAATAGGGTTGATCT	ttgttgtcactgg	ATGTACTGTA	catccatac
putA	ECK120009724	unknown	1078148	-	ATTTAACATGGTTGCAC	aaagttgcaacatc	ATGGATATTT	cac
putPp1	ECK120009726	unknown	1078391	+	AAATAGTGTCGCTGAGC	actaaaatttaat	GTAAATGGTG	tgt
phoHp1	ECK120009263	activator	1084088	+	CATCACTGTCATCACTC	tgtcatctttcca	GTAGAAACTA	atgtc
pyrC	ECK120009434	repressor	1121868	-	TTTATTTTTCGTGCAAA	ggaaaacgtttcc	GCTTATCCTT	tgtgt
ycfCp	ECK120009100	repressor	1191909	-	CTTCTATAACGGTGAAG	tgtgcctcggtgg	CGGTATTATT	gagcagc
aroP	ECK120009483	repressor	121671	-	GCGGGTGCATTCGCTGC	cgcataccattat	TCTTGATCTG	acgg
narKp1	ECK120009467	activator	1277155	+	GATTTA CATCAAATTGC	ctttagctacaga	CACTAAGGTG	gcagac
narG	ECK120009466	activator	1279037	+	ATTCCCACGCTGTTTCA	gagcgttaccttgc	CCTTAAACAT	t
trp	ECK120009442	repressor	1321132	-	AAATGAGCTGTTGACAA	ttaatcatcgaact	AGTTAACTAG	tacgc
cysB	ECK120009602	repressor	1331812	+	AAATGATATAGTGGTTA	tagttagcacctt	TTTTATTATT	aaatc
tyrR	ECK120009444	repressor	1384716	+	GACAGAAACCTTCCTGC	tatccaaatagtg	TCATATCATC	atatta
fnr	ECK120009085	repressor	1397576	-	TTGCTTAGACTTACTTG	ctccctaaaaaga	TGTTAAAATT	gacaa
purR	ECK120009457	repressor	1735713	+	TTTTTACCACTTCCCCT	tttcgtcaagatc	GGCCAAAATT	ccacgct
aroH	ECK120009094	repressor	1786342	+	ACTAGAGAACTAGTGCA	ttagcttatttttt	TGTTATCATG	cta
tyrP	ECK120009378	dual	1987666	+	TAACGTCGGTTTGACGA	agcagccgttatg	CCTTAACCTG	cgccgc
alkA	ECK120009088	activator	2145581	-	ATATGAAAGCAAAGCGC	agcgtctgaataac	GTTTATGCTG	aaagc
ccd	ECK120012682	unknown	2229839	+	GATTCAGATCACATATA	aagccacaacgggt	TCGTAAACTG	ttatc
galS	ECK120009470	dual	2239730	-	GACTCGATTCACGAAGT	cctgtattcagtgc	TGACAAAATA	gccgcc
cirp1	ECK120009461	dual	2244962	-	GATATAAAATTTAACAT	ttggattgataat	TGTTATCGTT	tgcatt
cirp2	ECK120009462	dual	2244949	-	TAACATTTGGATTGATA	attgttatcgtttg	CATTATCGTT	acgccgc
ada	ECK120009405	dual	2308447	-	CAAGATTGTTGGTTTTT	gcgtgatggtgacc	GGGCAGCCTA	aaggct
ompCp1	ECK120009426	dual	2310850	-	CATATTCGTGTTGGATT	attctgcattttt	GGGGAGAATG	gac
nrd	ECK120009469	activator	2342775	+	CCACAAAGTTATGCACT	tgcaagagggtcat	TTTCACACTA	tcttgca

Promotor	RegulonDB referencia	Tipo de regulación	Mas-uno	Cadena	Secuencia del Promotor (alineación hecha con el programa *wconsesus*)			
glpTQ	ECK120009452	dual	2350471	-	GCGGCAATTCACATTTA	atttatgaatgttt	TCTTAACATC	gcggc
glpACB	ECK120009450	dual	2350603	+	GCATGAAATCACGTTTC	actttcgaattatg	AGCGAATATG	cgcg
purFp1	ECK120009605	repressor	2428820	-	GGAAATCCCTACGCAAA	cgttttcttttc	TGTTAGAATG	cgccc
fadL	ECK120009431	repressor	2459225	+	AAACATACTTGCAACAT	tccagctggtccg	ACCTATACTC	tcgcc
cysKp1	ECK120009410	activator	2530397	+	CTAAATCCTTACTTCCG	catattctctgagc	GGGTATGCTA	cctgtt
ptsp0	ECK120009190	activator	2531521	+	CCCTCCTGGCATTGATT	cagcctgtcggaac	TGGTATTTAA	ccagact
phoE	ECK120009400	activator	259382	-	AATCTGTAATATATCTT	taacaatctcagg	TTAAAAACTT	tcct
purC	ECK120009453	repressor	2595801	-	TACAGGGCTGGAATCAT	ccggccctttttc	TGATATGATA	cgcaa
purM	ECK120009456	repressor	2619173	+	TAAAGCAGTCTCGCAAA	cgtttgctttccc	TGTTAGAATT	gcgcc
glnBp2	ECK120009379	repressor	2685523	-	CGAGAATGGCGGGGCGC	aaccggacagaatt	TTATAAACTG	ctttccc
purL	ECK120009455	repressor	2693609	-	CACGCAAACGGTTTCGT	cagcgcatcagatt	CTTTATAATG	acgccc
aroF	ECK120009092	repressor	2739221	-	CGAAATATGGATTGAAA	actttactttatg	TGTTATCGTT	acgtc
recA	ECK120009414	repressor	2821841	-	TCTACAAAACACTTGAT	actgtatgagcata	CAGTATAATT	gcttc
argF	ECK120009474	repressor	289564	-	GTGAAATGGGGTTGCAA	atgaataattacac	ATATAAAGTG	aatttt
cysJIH	ECK120009409	activator	2889987	-	CTATCCCGTCTTTAATC	cacaccgtttgccc	CGTTAACCTT	acct
carAp2	ECK120009460	repressor	29619	+	CAGATTTGCATTGATTT	acgtcatcattgtg	AATTAATATG	caa
araE	ECK120009408	dual	2980232	-	GTTTCCGACCTGACACC	tgcgtgagttgttc	ACGTATTTTT	tcact
ansBp2	ECK120009407	activator	3098770	-	GTTTAACGTCAAATTTC	ccatacagagcta	AGGGATAATG	cgtagc
mtr	ECK120009422	dual	3303501	-	GCTTTTTTTCTGTCTTT	tgtactcgtgtac	TGGTACAGTG	caatgc
argRp1	ECK120009476	repressor	3382312	+	CTGACTGTTTGCATAAA	aattcatctgtatg	CACAATAATG	tt
fis	ECK120009084	repressor	3407883	+	AAGTTTGGCCTTTCATC	tcgtgcaaaaaat	GCGTAATATA	cgcc
nirB	ECK120009468	dual	3491624	+	GATTTACATCAATAAGC	ggggttgctgaat	CGTTAAGGTA	ggcggt
malPQ	ECK120009435	dual	3550143	-	CCGCAGGATGAGGAAGG	tcaacatcgagcc	TGGCAAACTA	gcgat
malT	ECK120009420	dual	3550657	+	ACGTCATCGCTTGCATT	agaaaggtttctg	GCCGACCTTA	taacc
glpD	ECK120009451	dual	3559604	+	TTTTTCAATGTTACCTA	aagcgcgattctt	TGCTAATATG	ttcgat
ugpp1	ECK120009181	unknown	3590009	-	CAAAAAAGTTATTTTTC	tgtaattcgagca	TGTCATGTTA	cccc
rpoHp4	ECK120009433	repressor	3598494	-	CATTGAACTTGTGGATA	aaatcacggtctga	TAAAACAGTG	aat
rpoHp5	ECK120009416	dual	3598480	-	GGATAAAATCACGGTCT	gataaaacagtgaa	TGATAACCTC	gttgc
lac	ECK120009481	dual	365567	-	CACCCCAGGCTTTACAC	tttatgcttccggc	TCGTATGTTG	tgtgg
tdh	ECK120009398	dual	3790204	-	ACGCGTTATCTCGTCGC	gacctataagttt	GGGTAATATG	tgctg
tnaA	ECK120009399	activator	3886040	+	AACAATTTCAGAATAGA	caaaaactctgag	TGTAATAATG	tagcctc
pstS	ECK120009298	activator	3909196	-	TATTCCTTACATATAAC	tgtcacctgtttg	TCCTATTTTG	cttctc
ilvY	ECK120009478	repressor	3955488	-	ACTATATGACAGGAAAT	ttattgcggaaat	TGATATATTC	ac
ilvC	ECK120009419	activator	3955533	+	AATATATCAATTTCCGC	aataaatttcctg	TCATATAGTG	a
cya	ECK120009463	repressor	3988612	+	CGCATCTTTCTTTACGG	tcaatcagcaagg	TGTTAAATTG	atcac
uvrD	ECK120009404	repressor	3995520	+	GAAATTTCCCGGTTGGC	atctctgacctcgc	TGATATAATC	agcaa
phoA	ECK120009596	activator	400931	+	AACAGCTGTCATAAAGT	tgtcacggccgag	ACTTATAGTC	gcttt
fadBA	ECK120009430	repressor	4028603	-	AATCTTTTGTTTGCATA	tttttaacacaaa	ATACACACTT	cgactc
spf	ECK120009146	unknown	4047479	+	CAAAAAGTGCTTTCTGA	actgaacaaaaaa	GAGTAAAGTT	agtcgc
glnL	ECK120009449	repressor	4053951	-	CTCTGATGCTTCGCGCT	ttttatccgtaaaa	AGCTATAATG	cact
glnAp1	ECK120009448	dual	4055802	-	GGTTATCCAAAGGTCAT	tgcaccaacatggt	GCTTAATGTT	tccattgaa
aroL	ECK120009095	repressor	405504	+	TATTATTTACACTTCAT	tcttgaatatttat	TGGTATAGTA	aggggtgt
rhaS	ECK120009415	activator	4095293	+	TCGAAAATTTAAGGTAA	gaacctgacctcgt	GATTACTATT	tcgccg
sodA	ECK120009417	dual	4098338	+	ATTGATAATCATTTTCA	atatcatttaatt	AACTATAATG	aac
cytR	ECK120009149	unknown	4122089	-	GAAAATTCTGTAACCGT	tttcacgcgctatc	TGCTAAAAAT	gttgccg
metJp1	ECK120009129	unknown	4126138	-	TCAATACATCTGGACAT	ctaaacttctttgc	GTATAGATTG	agca
metB	ECK120009110	unknown	4126216	+	TCCAGATGTATTGACGT	ccattaacacaat	GTTTACTCTG	gtgcct
metF	ECK120009438	repressor	4130126	+	CATTTTTCGGTTGACGC	ccttcggcttttc	CTTCATCTTT	a
katG	ECK120009479	activator	4131392	+	ATTAATTCAATTATAAC	ttctctctaacgct	GTGTATCGTA	acggta
argE	ECK120009473	repressor	4152476	-	GGCTGGTGGGTTTTATT	acgctcaacgtta	GTGTATTTTT	attc
argCBH	ECK120009472	repressor	4152464	+	TTGTTTTTCATTGTTGA	cacacctctggtca	TGATAGTATC	aatattc
oxyR	ECK120009428	repressor	4156036	+	GTCAGAATGCTTGATAG	ggataatcgttcat	TGCTATTCTA	cctatc
phoB	ECK120009429	activator	416325	+	CGAGCTTTTCATAAATC	tgtcataaatctg	ACGCATAATG	acgtc
purH	ECK120009454	repressor	4205205	-	AAAAATTCGCGAGCGTT	gcgcaaacgtttt	CGTTACAATG	cgggc
metAp1	ECK120009437	dual	4211812	+	ACATGCAGGCTCGACAT	tggcaaattttct	GGTTATCTTC	agct
aceBAK	ECK120009087	dual	4212981	+	GGAAATTGTTTTTGATT	ttgcattttaaat	GAGTAGTCTT	agtt
queA	ECK120009285	unknown	424209	+	GGGGTAAAGGTTGCAGG	gagagcgccccgg	CACTAGACTA	cccgc
malE	ECK120009458	activator	4244043	-	GGAGGATGGAAAGAGGT	tgccgtataaaga	AACTAGAGTC	c
malK	ECK120009459	activator	4244315	+	GGGGGTGGAGGATTTAA	gccatctcctgatg	ACGCATAGTC	agccc
lexA	ECK120009482	repressor	4254666	+	GCAGTTTATGGTTCCAA	aatcgcctttgct	GTATATACTC	acagc
tyrB	ECK120009389	repressor	4264661	+	GAACATCCACTCGATCT	tcgccttcttccg	GTTTATTGTG	tttta

Promotor	RegulonDB referencia	Tipo de regulación	Mas-uno	Cadena	Secuencia del Promotor (alineación hecha con el programa *wconsesus*)			
uvrA	ECK120009445	repressor	4271512	-	CCAATACTGTATATTCA	ttcaggtcaatttg	TGTCATAATT	aacc
ssb	ECK120009418	repressor	4271590	+	CAGTATTGGAATGCATT	acccggagtgttg	TGTAACAATG	t
tsxp1	ECK120009287	repressor	431469	-	TGCGTCCCGCAACATCT	ttccccgtcattt	TGTTACTCTG	cttac
tsxp2	ECK120009286	dual	431315	-	GAAACGAAACATATTTT	tgtgagcaatgat	TTTTATAATA	ggctc
melR	ECK120009436	dual	4339228	-	GCTCCCACTCGCAGTCA	tcctccctcactcc	TGCCATAATT	ctgat
frd	ECK120009086	dual	4379990	-	TCTCGTCAAATTTCAGA	cttatccatcaga	CTATACTGTT	gtacctat
argI	ECK120009475	repressor	4475907	-	TGCTTTAGACTTGCAAA	tgaataatcatcc	ATATAAATTG	aatttt
deop1	ECK120009464	repressor	4614248	+	AACGTTTTATTCGAACA	tcgatctcgtcttg	TGTTAGAATT	ctaac
deop2	ECK120009081	dual	4614847	+	GATGTGTATCGAAGTGT	gttgcggagtaga	TGTTAGAATA	ctaac
trpR	ECK120009443	repressor	4630273	+	ACGTCGTTACTGATCCG	cacgtttatgata	TGCTATCGTA	ctcttt
purE	ECK120009432	repressor	552365	-	CCACGCAACCGTTTTCC	ttgctctctttccg	TGCTATTCTC	tgtgcc
ahpC	ECK120009406	activator	638144	+	GGAAACGCATTAGCCGA	atcggcaaaaatt	GGTTACCTTA	catctc
araBAD	ECK120009089	dual	70075	-	GGATCCTACCTGACGCT	ttttatcgcaact	CTCTACTGTT	tctccat
nagB	ECK120009440	dual	702848	-	GTTACGCTTAAAGATGC	ctaatccgccaacg	GCTTACATTT	tact
nagE	ECK120009465	dual	703078	+	GATACGAATTAAATTTT	cacacactctgta	GCAGATGATC	taacaat
araC	ECK120009090	dual	70241	+	TGCCGTGATTATAGACA	cttttgttacgcgt	TTTTGTCATG	gctt
kdpABC	ECK120009480	activator	728072	-	GCAGATTTTTGCGAAAT	ctttgcagccagaa	TTCTACCCTT	ccggt
aroG	ECK120009093	repressor	784815	+	GTAAAACCCCGTTTACA	cattctgacggaag	ATATAGATTG	gaagt
galp1	ECK120009471	dual	791304	-	ATTCCATGTCACACTTT	tcgcatctttgtta	TGCTATGGTT	atttc
galp2	ECK120009447	repressor	791309	-	AATTTATTCCATGTCAC	acttttcgcatctt	TGTTATGCTA	tggtt
bioA	ECK120009484	repressor	808515	-	TCTCCAAAACGTGTTTT	ttgttgttaattcg	GTGTAGACTT	gtaa
bioB	ECK120009485	repressor	808525	+	CATAATCGACTTGTAAA	ccaaattgaaaag	ATTTAGGTTT	acaagtc
uvrBp2	ECK120009403	repressor	812685	+	AAATATTATGGTGATGA	actgttttttatc	CAGTATAATT	tgttg
ilvlHp1	ECK120009597	activator	85597	+	CTGGCTGCCAATTGCTT	aagcaagatcggac	GGTTAATGTG	tttt
dmsA	ECK120009082	dual	940053	+	GAACAATAATTACTCCT	cacttacacgtaa	TACTACTTTC	gagtgaa
focAp1	ECK120009114	dual	953715	-	GATCTATATCAATTTCT	catctataatgctt	TGTTAGTATC	tcgtc
focAp2	ECK120009255	activator	954037	-	CTACGCAATGTAGGCTT	aatgattagtctg	AGTTATATTA	cggggc
ompF	ECK120009427	dual	986315	-	CACTTTCACGGTAGCGA	aacgttagtttgaa	TGGAAAGATG	cctgc

Tabla VI.3. La columna 4 corresponde a la posición absoluta del mas-uno del promotor en el genoma completo de *E. coli* K12. F o R en la columna 5 indican si la secuencia no-templado del promotor esta en la cadena mas (5'-3'), o en la cadena menos (3'-5') del genoma. Las secuencias en mayúsculas corresponden a las secuencias más significativas en términos de contenido de información y las secuencias en minúsculas indican regiones del promotor que no fueron alineadas; el algoritmo que se utilizó para generar el alineamiento es *wconsensus* v5c de [Hertz_1999]. La tabla muestra 116 promotores σ70 pertenecientes al conjunto de entrenamiento reportado en [Gralla_1996] y que se encuentra colectada en la base de datos RegulonDB (http://www.cifn.unam.mx/Computational_Biology/regulondb).

Del trabajo que se ha realizado con los elementos UP en el laboratorio del Dr. Gourse [Estrem_1998] recibimos las secuencias de los mejores elementos UP generados por SELEX, la tabla VI.4 muestra una la colección de 32 secuencias UP.

Nombre de la secuencia UP	Actividad Relativa	Secuencia UP (UpStream)
4192	326	GGAAAATTTTTTTTCAAAAGTA
4181	320	AGAAATTTTTTTTCGAAAAACA
4176	316	TAAAAATTTTTTTTGAAAAGGG
4173	297	CAAAAATATTTTTGAAAAAAGA
4209*	293	GGAAATATTTTTTCATAAACCC
4206	274	AGAAAAATATTTTCGAAAACTA
4202	269	AAAAATATTTTTTCGAAAAGTA
4196	268	TAAATTTTTTTTTGCAAAAGTA
4193	265	ACAAAAATATTTTTCAAAACCC
4179	265	TTAAATTTTTTTTCGTAAACCC
4203	262	TTAAATTTTTTTTTCATAAACCC
4191	262	TCAAATTTTTTTTTGCAAACCC
4204	257	CAAATTTTTTTTTGCTAAACCC

Nombre de la secuencia UP	Actividad Relativa	Secuencia UP (UpStream)
4190	248	AAAAATATTTTTTGAAAAGTA
4219*	245	TAAAAATATTTTTCGTTTACCC
4198	240	ACAAAAATATTTTTCGAAACCC
4200	239	TCAAAATTTTTTTTGCAAAGTA
4218*	238	TGAATTTTTTTTTCGTCTACCC
4171	228	AGAAAAATATTTTTGAAAACTA
4177	222	GCAAAATAATTGTAAAAAAGTA
4220*	221	AGAAATTTATTTTAAAAAAGGG
4205	215	TGAAAAATATTTTTGAAAACTA
4199	213	TAAACTATTTTTTCAAAAAGGA
4174	210	TGAAATTTATTTTGCGAAAGGG
4197	206	TAAACTTTTTTTTTCGAAAGTG
4207	199	TGAAATATTTTTTTGAAAACCC
4194	194	AGATTTTTTTTTTGTAAAAGTG
4168	193	GCAAAAATATTTCGTCAAACCC
4201	185	GAAAAATATTTTTGATAAAGTA
4185	178	GCAAAATTATTTTGCTAAAGTA
4195	136	GAAAATATATTTTTCAAAAGTA
WT-rrnB p1	69	AGAAAATTATTTTAAATTTCCT
core rrnB p1	1	GACTGCAGTGGTACCTAGGAAT

Tabla VI.4. 32 secuencias Upstream, y las actividades relativas en promotores, seleccionadas *in vitro*.

Se tomaron 22 promotores regulados por σ32 en *E. coli* reportados en la literatura y colectados en RegulonDB para hacer el análisis de los elementos usados por el programa para reconocer secuencias de promotores de tipo σ32. La tabla VI.5 muestra las secuencias de estos 22 promotores.

(a)

Promotor	RegulonDB referencia	Mas-uno	Cadena	Secuencia del Promotor (alineación hecha con el programa *wconsesus*)			
htgAp1	ECK120009520	10644	+	TTGAGGGGAA	aatgaaaattttc	CCCGGT	ttccggtat
dnaKp1	ECK120009279	12047	+	CCCCCTTGAT	gacgtggtttacga	CCCCAT	ttagtagt
dnaKp2	ECK120009269	12121	+	GGCAGTTGAA	accagacgtttcgc	CCCTAT	tacagac
clpPp1	ECK120009288	455801	+	TTAGCGTAAC	aacaaaagattgttatg	CTTGAA	at
lon	ECK120009096	458039	+	CGGCGTTGAA	tgtgggggaaacat	CCCCAT	atactgacgtac
htpGp1	ECK120009266	494299	+	CTCGCTTGAA	attattctcccttgt	CCCCAT	ctctccca
htpGp2	ECK120009267	494308	+	CTCCCTTGTC	Cccatctctc	CACAT	cctgtt
gapAp1	ECK120009231	1860642	+	TGCCCTTTAA	aattcggggcgccga	CCCCAT	gtggtctcaa
htpX	ECK120009516	1910642	-	CAGACTTGAA	aatagtcgcgta	ACCCAT	acgatgtgggt
clpB	ECK120009187	2732225	-	TAACCTTGAA	taattgagggatga	CCTCAT	ttaatctcc
grpE	ECK120009177	2748768	-	TTCCCTTGAA	accctgaaactgat	CCCCAT	aataagcga
rpoDp3	ECK120009552	3210329	+	CACCCTTGAA	aaactgtcgatgtggga	CGATAT	agcagat
ftsJp2	ECK120009174	3325372	-	TGGGATTGAA	aacgggtcattctac	CGCCAT	ctcccatat
htrM	ECK120009098	3791492	+	CCGCCATGAA	ggactagctaaa	ACCCAA	actagt
ibpA	ECK120009097	3865145	-	AAGGCTTGAA	aagttcatttccag	ACCCAT	ttttacatc
hslV	ECK120009148	4119934	-	GGGGGTTGAA	accctcaaaatccc	CCCCAT	ctat
mopB	ECK120009126	4368194	+	CCCCCTTGAA	ggggcgaagcctcat	CCCCAT	ttctctggtc

(b)

Promotor	RegulonDB referencia	Mas-uno	Cadena
hnaK	ECK120009531	12144	+
topAp1	ECK120009509	1329004	+
htrC	ECK120009515	4187331	+
ybaUp1	ECK120009584	461062	+

Promotor	RegulonDB referencia	Mas-uno	Cadena
mlcp2	ECK120012510	1666615	-

Tabla VI.5. La tabla muestra 22 promotores σ32 con el mas-uno mapeado que se encuentra colectada en la base de datos RegulonDB. La columna 3 corresponde a la posición absoluta del mas-uno del promotor en el genoma completo de *E. coli* K12. F o R en la columna 4 indican si la secuencia no-templado del promotor esta en la cadena mas (5'-3'), o en la cadena menos (3'-5') del genoma. (a) Las secuencias de promotores en mayúsculas corresponden a las secuencias más significativas en términos de contenido informacional y las secuencias en minúsculas indican regiones del promotor que no fueron alineadas; el algoritmo que se utilizó para generar el alineamiento es *wconsensus* v5c de [Hertz_1999]. (b) Este conjunto formó parte del conjunto de prueba.

VI.1 ALINEAMIENTO DE LAS SECUENCIAS Y SELECCIÓN DE LA MATRIZ CONSENSO.

Como se mencionó anteriormente el problema de reconocer secuencias promotoras en una cadena de DNA es un problema sujeto a muchas variables, principalmente la poca homogeneidad entre las secuencias del promotor y las regiones canónicas -10 y -35 que definen un promotor es una limitante. Por lo tanto la elección del alineamiento que produzca la matriz óptima para la predicción de la ocurrencia de nuevos sitios promotores es de critica importancia [Stormo_2000].

Para el alineamiento de secuencias, y la obtención de las matrices de peso que serán utilizadas para asignar una calificación a una secuencia blanco y saber si esta es una caja -10 o una caja -35, se utilizaron los programas de *wconsensus* y *patser* [Hertz_1990, Hertz_1999]. Estos programas están basados sobre la metodología de Teoría de la Información para encontrar el mejor alineamiento en un conjunto de secuencias biológicamente relacionadas, y generan una matriz consenso que representa este alineamiento. El programa *patser* permite usar esta matriz consenso para calificar secuencias blanco, asignándoles una calificación que es estadísticamente significativa y que mide la cercanía de la secuencia blanco al consenso definido por la matriz.

Los alineamientos para encontrar las cajas -10 y -35 se llevaron a cabo de la siguiente manera:

a) Se obtuvo de E. *coli* K12 las secuencias de los 116 promotores σ70 de la colección publicada en [Gralla_1996], se alinearon estas 116 secuencias con respecto al inicio de la transcripción.
b) Se cortó un número determinado *X* de bases desde cada inicio de la transcripción obteniéndose subsecuencias de tamaño *X*. Se sabe que la distancia del hexámero -10 al mas-uno del promotor es de 4 a 12 bp [Harley_1987], esta consideración fue tomada en cuenta en la selección de ese número X.
c) Este nuevo subconjunto de secuencia se usó como entrada al programa *wconsensus* [Hertz_1999] para encontrar la mejor matriz consenso que representara a la caja -10. Se hicieron 6 diferentes corridas del programa con subsecuencias de longitudes de 15 a 20 bp, y con desviaciones estándares de 0.5, 1.0, 1.5, y 2.0.
d) Se alineó el conjunto original de secuencias ahora con respecto a la primera letra de la caja -10 encontrada.
e) Se tomó un número determinado *(Y + W)* de bases desde la primera letra de la caja -10, y se cortaron las primeras *Y* bases, obteniéndose subsecuencias de tamaño *Y*. Se sabe que la RNAP se puede pegar con cierta flexibilidad al DNA, las distancias

permitidas entre los dominios de la RNAP que contactan al DNA son de 15 a 21 bp en medidas de DNA, esta consideración fue tomada en cuenta en la selección del número *W*. En este trabajo se eligió arbitrariamente una $W = 13$.

f) Este nuevo subconjunto de secuencias se usó como entrada al programa *wconsensus* para encontrar la mejor matriz que representara a la caja -35.

Una vez que se obtuvieron todas las pajeras de matrices -10 y sus correspondientes -35 se calcularon los otros atributos de los promotores de la siguiente manera:

g) Para cada pareja de alineamientos se calcularon los espaciadores entre las cajas -10 y -35 de cada promotor. Se calculó la probabilidad relativa de cada espaciador para cada pareja de matrices o alineamientos como:

$$Score\_Espaciador\ (s) = \frac{\text{número de promotores con espaciador } s}{\text{número total de promotores analizados}} \qquad \text{(i)}$$

h) Para cada alineamiento se calculó la distancia de la primera letra corriente arriba de la caja -10 al codón de inicio del gen transcrito y se calculó la probabilidad relativa como se describe para la formula (h). Se usó el logaritmo natural de esa calificación para cualquier predicción de caja -10.

i) Para cada alineación se calculó la distancia de la caja -10 al inicio de transcripción, el +1, la distancia mínima y la máxima fueron utilizadas para determinar si una predicción de caja –10 era una predicción positiva verdadera.

La figura VI.3 muestra un diagrama de la estrategia que se siguió en este trabajo para la búsqueda de la matriz más óptima para reconocer promotores tipo σ70. Los primeros consensos reportados en la literatura para las cajas -35 y -10 fueron identificados por comparación de muy pocos promotores, por ejemplo, se usaron 6 para identificar la caja -10 [Pribnow_1975, Rosenberg_1979]. Análisis en colecciones grandes de promotores fueron hechos por [Hawley_1983, Harley_1987, Lisser_1993] utilizando métodos estadísticos que tomaron como punto referencia los consensos TTGACA y TATAAT, alineando de tal manera que se maximizara la homología a las 12 bp. En este trabajo no se sembró ninguna semilla inicial o de arranque que guiara el alineamiento. Se permitió que el programa de *wconsensus* determinara la matriz consenso del patrón mas conservado en el conjunto de secuencias a analizar, y por lo tanto la longitud de éste.

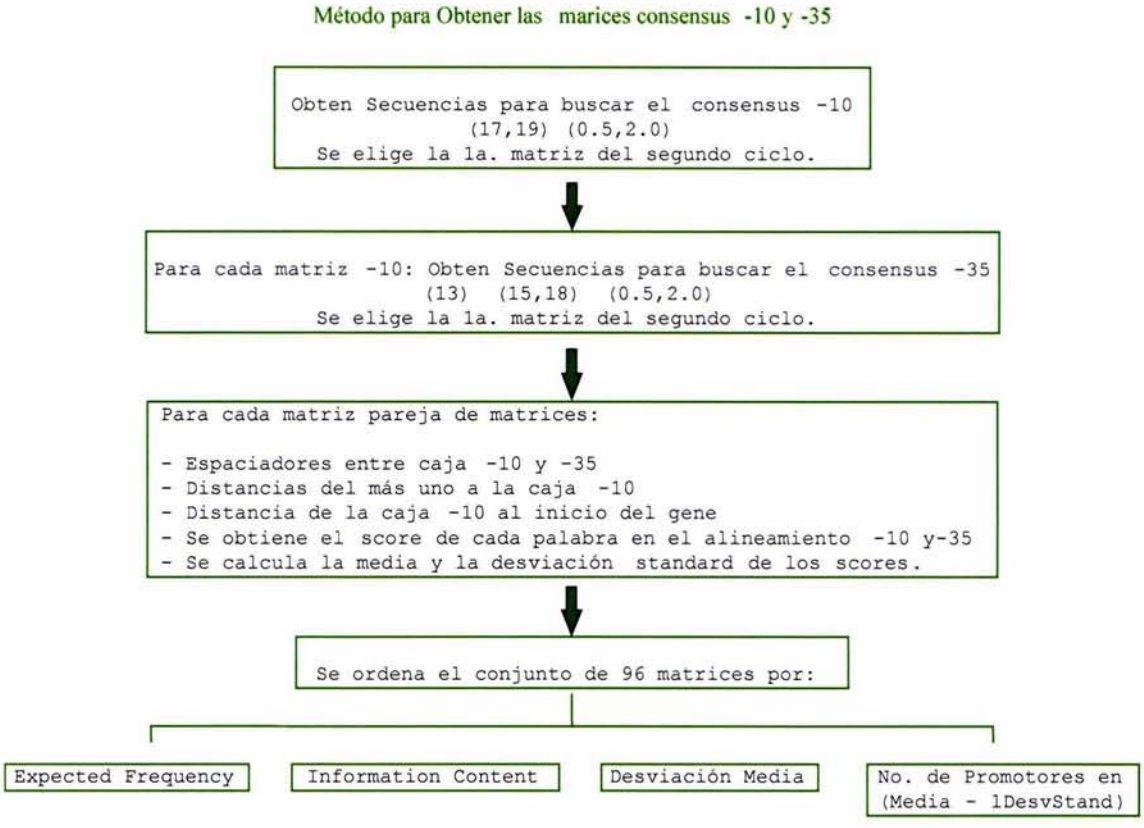


Figura VI.3. Proceso seguido para obtener las matrices consenso -10 y -35.

El programa de *wconsensus* selecciona la mejor matriz maximizando el índice de contenido de información de las secuencias y minimizando la frecuencia esperada de encontrar la matriz al azar. La selección de las probabilidades *a priori* de cada nucleótido pueden ser dadas al programa o dejar que este use las frecuencias observadas de las letras en las secuencias a analizar. La corrección del contenido de información de cada matriz esta dado por un valor de desviación estándar esperado de un alineamiento hecho al azar. Se recomiendan valores de desviación estándar de 0.5, 1, 1.5, y 2. Siguiendo el procedimiento de la figura VI.3 se generó una librería de 96 diferentes matrices para representar a las dos señales del promotor. La elección de la pareja de matrices -35 y -10, que sea la mejor en la tarea del reconocimiento de promotores, puede seguir 4 diferentes criterios:

1. Aquellas con el mayor contenido de información.
2. Aquellas con la mínima frecuencia esperada
3. Aquellas que contienen la mayor cantidad de promotores agrupados en μ de las calificaciones menos una desviación estándar (σ).
4. Aquellas cuyo consenso homologó a los patrones canónicos -35, -10 y espaciadores entre las cajas de 15 a 21 bp.

Las matrices fueron usadas para calificar a cada una de las cajas alineadas de los 116 promotores, y se obtuvo así el conjunto de calificaciones de cada promotor. Se

calcularon los estadísticos de la media y la desviación estándar de estas calificaciones. Estos estadísticos fueron usados en la determinación del umbral de corte a partir del cual se dice si una secuencia blanco es una caja -10 o -35.

VI.2 Eligiendo las matrices que cumplen con los cánones de un promotor.

Secuencias arriba del +1 de longitud de 18 bp para la región -10 y de longitudes de 15 y 16 para la región -35 con desviaciones estándares de 1.5 y 2.0 produjeron las matrices consenso cuyas posiciones más conservadas son aquellas de los elementos canónicos que definen a un promotor. Para efectos de ejemplificación, a este juego de matrices le llamaremos "Matriz.18.16.13.2.2". La figura VI.4 muestra el consenso obtenido y las frecuencias de cada base en cada posición. La figura VI.5 muestra las matrices consensi correspondientes a estos alineamientos.

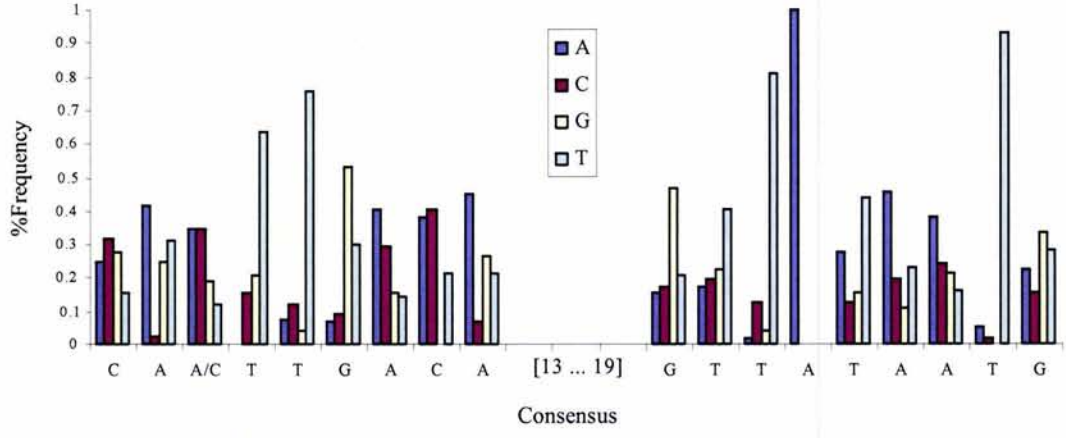


Figura VI.4. Distribución de las bases en las secuencias alienadas con *wconsensus*.

a)

A |	18	20	2	116	32	53	44	6	26
C |	20	23	15	0	15	23	28	2	18
G |	54	26	5	0	18	13	25	0	39
T |	24	47	94	0	51	27	19	108	33
	G	T	T	A	T	A	A	T	G
					|				
					-10				

b)

A |	29	48	40	0	9	8	47	44	52
C |	37	3	40	18	14	11	34	47	8
G |	32	29	22	24	5	62	18	0	31
T |	18	36	14	74	88	35	17	25	25
	C	A	A/C	T	T	G	A	C/A	A
				|					
				-35					

Figura VI.5. (a) Matriz consenso -10 generada con el programa *wconsensus*. Cabe notar que a diferencia con el patrón reportado en la literatura esta matriz contiene 9 bp en lugar de 6. (b) La matriz -35 también resulta en un patrón de 9 bp a diferencia del consenso reportado en la literatura de 6 bp.

Una vez ubicadas ambas cajas en las 116 secuencias promotoras, se calcularon los espaciadores que hay entre cada una las cajas -35 y -10. Se calculó la probabilidad relativa para cada una de los espacios que se encontraron. Más adelante, nuestro algoritmo usará estos espacios entre las cajas -10 y -35 como un argumento de clasificación [Mulligan_1985, Harley_1987, Oneill_1992]. La figura VI.6 muestra la distribución de los espaciadores en conjunto de entrenamiento.

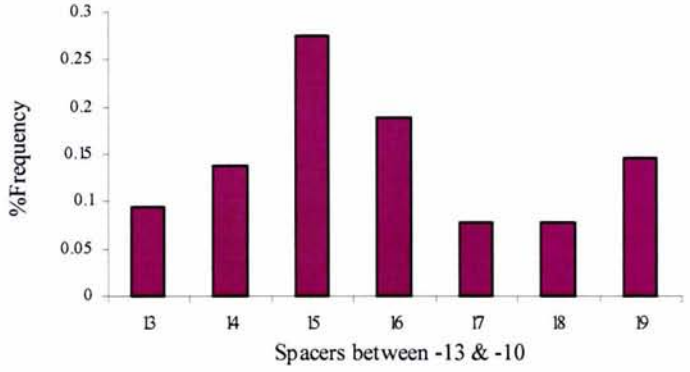


Figura VI.6. Distribución de los espacios encontrados en los 116 promotores del conjunto de entrenamiento después de haber obtenido las matrices para las señales -10 y -35. Debe observarse que los espacios corresponden a aquellos de la literatura, ya que la caja -10 encontrada es de 9 bp y en el extremo izquierdo hay 2 bp que el hexámero -10 canónico no tiene, esto quiere decir que, por ejemplo, el espacio de 13 bp es el equivalente al espacio de 15 bp reportado en la literatura.

Usando las matrices de la figura VI.5 se asigno una calificación a cada una de las secuencias resultantes del alineamiento. Se obtuvo la media y la desviación estándar de las calificaciones de las cajas -35 y de las calificaciones de la -10, así como los de la suma de las calificaciones -35 + -10. Se calculo la distribución de frecuencia de las calificaciones, lo que resulto en lo que ya se sabía, que hay gran variabilidad para cada caja. La figura VI.7 muestra la distribución de frecuencias de las 116 calificaciones de las cajas -35 y -10. Se buscaron correlaciones entre las cajas -35 y -10, y entre la suma de las calificaciones y el espaciador entre las cajas. La figura VI.8 muestra que no existe correlación entre las calificaciones de estos elementos del promotor.

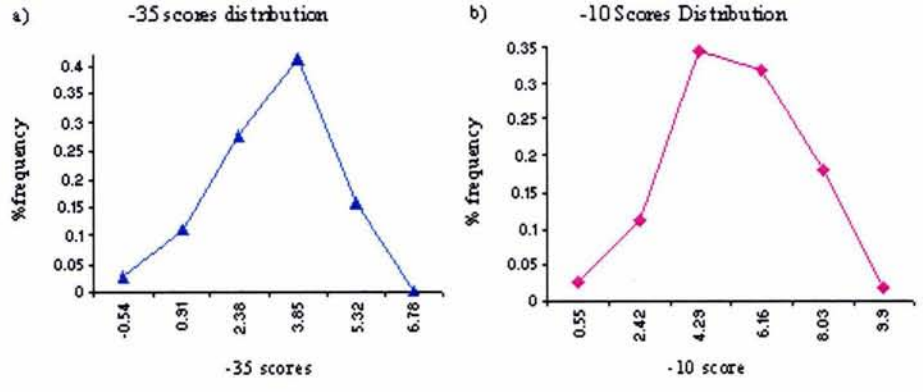


Figura VI.7. Distribución de las calificaciones de las cajas -35 y -10. Las medias de las cajas -35 y -10 son 2.31 y 3.85 respectivamente.

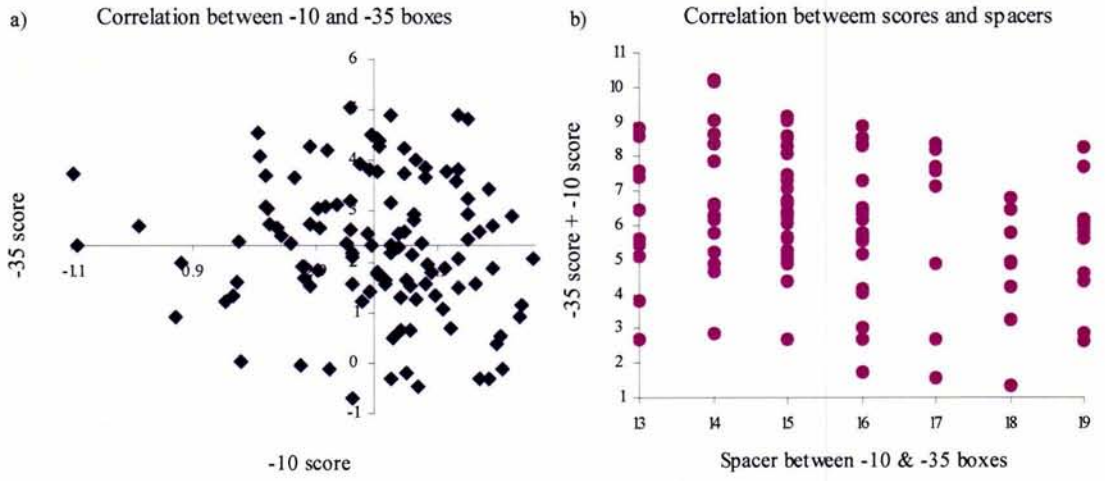


Figura VI.8. (a) La correlación entre las calificaciones de las cajas -10 y las calificaciones de las cajas -35 indica que una caja -10 muy cercana al consenso no tiene necesariamente una caja -35 débil. (b) Correlación entre la calificación completa, el valor en -10 más el valor en -35, y los diferentes espacios en bp que están permitidos por la RNAP para pegarse al DNA.

Finalmente, generamos una búsqueda usando estas matrices y el programa *patser* [Hertz_1999] en el conjunto potencial de regiones reguladoras del genoma de *E. coli*, que suman 4290 regiones, permitiendo un espacio de 15 a 21 pares de bases entre cada caja potencial -10 y -35, y guardando aquellas posibles señales cuyos cajas tuvieron calificaciones más altas que el respectivo umbral, y aplicamos nuestra función Cover para obtener los promotores "funcionales".

Todo este proceso fue realizado para detectar las cajas -10 y -35 de los promotores σ32 usando los promotores mostrados en la tabla VI.5 y promotores σ54, datos no presentados. Los programas para la predicción de promotores fueron escritos en Perl [Wall_1991].

VI.3 ANEXO A.

Diferentes métodos han sido usados para buscar señales de regulación en el DNA. Búsquedas con matrices de peso por posición han sido usadas en numerosos estudios para caracterizar la distribución de bases en cada poción in la secuencia reconocida (Vanet_1999, Stormo_2000).

Una matriz de peso por posición es generada mediante el alineamiento de un conjunto secuencias que maximiza la conservación de la secuencia. En estas matrices un valor específico es asignado a cada nucleótido para cada posición en la secuencia indicando la contribución de esa base particular a la especificidad de la secuencia-señal o sitio de pegado. La calificación total de una secuencia de blanco de igual longitud a la secuencia-señal o sitio de pegado es la suma de todos los valores para cada posición. La calificación entonces refleja la afinidad de pegado o especificidad de la secuencia por la molécula o proteína que se pega a DNA, y permite que diferentes secuencias sean cuantitativamente comparadas como secuencias-señal o sitios de pegado potenciales. Esto

es cierto cuando no existe dependencia entre las bases del sitio de pegado, característica que no todas los sitios de pegado de proteínas tienen.

Una matriz de peso por posición puede ser generada desde una tabla de frecuencias u ocurrencias usando diferentes enfoques matemáticos. Cada enfoque intenta dar una calificación, c_{bp}, para cada nucleótido base (b) in la posición (p) de las secuencias-señal alineadas, o sitios de pegado. Para cualquier secuencia a calificar de la misma longitud que el sitio de pegado, una calificación total (C) puede ser calculada como una combinación estrictamente linear de las calificaciones asignadas a cada base en la secuencia analizada de la siguiente manera:

$$C = \sum_{1}^{L} s_{bp}$$

donde L es la longitud de la secuencia.

La teoría de la información ofrece una manera de interpretar matrices de peso por posición. Este enfoque intenta determinar que tan restringido, u obligado, está la selección de una base en cada posición dentro de la secuencia-señal o sitio de pegado, y entonces estima la cantidad de información local que está codificada en una secuencia como la sumatoria lineal de las calificaciones en cada posición. La transformación logarítmica estándar usada para calcular las calificaciones de las posiciones individuales es la siguiente:

$$s_{bp} = \log_2 \frac{f_{bp}}{q_b}$$

donde f_{bp} es igual a la frecuencia observada del nucleótido b en la posición p en la serie de secuencias alineadas, y q_b es la frecuencia de encontrar el nucleótido b dentro de la secuencia o fuente genómica del sitio de pegado o secuencia-señal (determinada por la composición de nucleótidos en el genoma). Ya que solamente un numero limitado de sitios son usados para generar la matriz, la frecuencia en la cual algunos nucleótidos son observados en una posición particular puede ser igual a cero. Esto crea una calificación indefinida en la ecuación anterior. Una simple solución a este problema es introducir un parámetro que es llamado un "seudo-contador" (s) para asegurar que la fracción de la ecuación anterior nunca será cero para el caso de un nucleótido no observado. En lugar de calcular el número de ocurrencias observadas (o_{bp}) dividido por el numero total de sitios (N), la frecuencia f_{bp} puede ser calculada usando la siguiente formula:

$$f_{bp} = \frac{o_{bp} + sq_b}{N + s}$$

Puede verse que el contenido total de información de una secuencia será generalmente positivo si hay mas información en la secuencia de lo que podría ser esperado al azar. Además, entre mas alto es la calificación total (C), mas grande se vuelve la similitud entre la secuencia particular y la lista de sitios usada para ejecutar la búsqueda.

VII. BIBLIOGRAFÍA.

Altschul S.F., Gish W., Miller W., Myers E.W., and Lipman D.J. (1990). Basic local alignment search tool. *J.Mol. Biol.* 215, 403-410.

Arthur T.M., and Burgess R.R. (1998) Localization of a 70 binding site on the N-terminus of the *Escherichia coli* RNA polymerase ß' subunit. *J. Biol. Chem.* 273, 31381–31387.

Beck C., Marty R., Klausli S., Hennecke H., and Gottfert M. (1997) Dissection of the transcription machinery for housekeeping genes of *Bradyrhizobium japonicum*. *J Bacteriol.* 179(2), 364-369.

Benoff B., Yang H., Lawson C.L., Parkinson G., Liu J., Blatter E., Ebright Y.W., Berman H.M., and Ebright, R.H. (2002) Structural Basis of Transcription Activation: The CAP-αCTD-DNA Complex. *Science*. **297**, 1562-1566.

Benos P.V., Bulyk M.L., and Stormo G.D. (2002) Additivity in protein-DNA interactions: how good an approximation is it? *Nucl. Acids. Res.* **30**(20), 4442-4451.

Berg O.G., and von Hippel P.H. (1987) Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and aplication to operators and promoters. *J. Mol. Biol.* 193, 723-750.

Bockhorst J., Qiu Y., Glasner J., Liu M., Blattner F., and Craven M. (2003) Predicting bacterial transcription units using sequence and expression data. *Bioinformatics*. 19(1), 34–43.

Bulyk M., Johnson P., and Church G. (2002) Nucleotides of transcription factor binding sites exert inter-dependent effects on the binding affinities of transcription factors. *Nucleic Acids Res.*, **30**, 1255-1261.

Burgess R.R., Traves A.A., Dunn J.J., and Bautz E.K.F. (1969) Factor stimulating transcription by RNA polimerase. *Nature.* 221(175), 43-46.

Burr T., Mitchell J., Kolb A., Minchin S., and Busby S. (2000). DNA sequence elements located immediately upstream of the -10 hexamer in *Escherichia coli* promoters: a systematic study. *Nucl. Acids. Res.* 28, 1864-1870.

Busby S., and Ebright R.H. (1999). Transcription activation by catabolite activator protein (CAP). *J Mol Biol.* **293**(2), 199-213. Review.

Chan C.L., and Landick R. (1994) In Conaway (eds), *Transcription mechanism and regulation.* 297.

CIFN (2001) Web page: http://www.cifn.unam.mx/iii_research/research.htm#3.1.

Collado-Vides J. (1992). Grammatical model of the regulation of gene expression. *Proc. Natl. Acad. Sci USA.* 89, 9405-9409.

Conaway J.W., and Conaway R.C. (1990) An RNA polymerase II transcription factor shares functional properties with *Escherichia coli* sigma70. *Science.* 248(4962), 1550-1553.

Cormen T.H., Leiserson C.E., and Rivest R.L. (1990) Introduction to Algorithms. *The MIT Press* , ISBN 0-262-03141-8.

Cossart P., Groisman E.A., Serre M.C., Casadaban M.J., and Gicquel-Sanzey B. (1986) crp genes of Shigella flexneri, Salmonella typhimurium, and Escherichia coli. J Bacteriol. 1986 167(2), 639-646.

Dartigalongue C., and Raina S. (1998) A new heat-shock gene, ppiD, encodes a peptidyl-prolyl isomerase required for folding of outer membrane proteins in *Escherichia coli*. *EMBO.* 17(14), 3968-3980.

Decker K., Plumbridge J., and Boos W. (1998) Negative transcriptional regulation of a positive regulator: the expression of malT, encoding the transcriptional activator of the

maltose regulon of *Escherichia coli*, is negatively controlled by Mlc. *Mol Microbiol.* 27(2), 381-390.
Deuschle U., Kammerer W., Gentz R., and Bujard H. (1986) Promoters of *Escherichia coli*: a hierarchy of in vivo strength indicates alternate structures. *EMBO J.* (11), 2987-2994.
Dhiman A., and Schleif R R. (2000). Recognition of overlapping nucleotides by AraC and the sigma subunit of RNApolymerase. *J Bacteriol.* 182(18), 5076-5081.
Dombroski A.J., Walter W.A., Record M.T. Jr, Siegele D.A., and Gross C.A. (1992) Polypeptides containing highly conserved regions of transcription initiation factor sigma70 exhibit specificity of binding to promoter DNA. *Cell.* 70(3), 501-512.
Dong S., and Searls D.B. (1994). Gene structure prediction by Linguistic methods. *Genomics.* 23, 540-551.
Estrem S.T., Gaal T., Ross W., and Gourse R.L. (1998) Identification of an UP Element Consensus Sequence for Bacterial Promoters. *Proc. Natl. Acad. Sci USA.* 95, 9761-9766.
Estrem S.T., Ross W., Gaal T., Chen Z.W., Niu W., Ebright R.H., and Gourse R.L. (1999) Bacterial promoter architecture: subsite structure of UP elements and interactions with the carboxy-terminal domain of the RNA polymerase alpha subunit. *Genes Dev.* 13(16), 2134-2147.
Fickett J.W., and Hatzigeorgiou A.G. (1997). Eukaryotic promoter recognition. *Genome Res.* 7(9), 861-878. Review.
Gaal T., Ross W., Blatter E.E., Tang H., Jia X., Krishnan V.V., Assa-Munt N., Ebright R.H., and Gourse R.L. (1996) DNA-binding determinants of the alpha subunit of RNA polymerase: novel DNA-binding domain architecture. *Genes Dev.* 10(1), 16-26.
Gaal T., Ross W., Estrem S.T., Nguyen L.H., Burgess R.R., and Gourse R.L. (2001) Promoter recognition and discrimination by $E\sigma^S$ RNA polymerase. *Mol Microbiol.* 42(4), 939-954.
Galas D.J., Eggert M., and Waterman M.S. (1985) Rigorous pattern-recognition methods for DNA sequences. Analysis of promoter sequences from *Escherichia coli. J Mol Biol.* 186(1), 117-128.
González V., Bustos P., Ramírez-Romero M.A., Medrano-Soto A., Salgado H., Hernández-González I., Hernández-Celis J.C., Quintero V., Moreno-Hagelsieb G., Girard L., Rodríguez O., Flores M., Cevallos M.A., Collado-Vides J., Romero D., and Dávila G. (2003) The mosaic structure of the symbiotic plasmid of *Rhizobium etli* CFN42 and its relation to other symbiotic genome compartments. *Genome Biol.* 4, R36 (http://genomebiology.com/2003/4/6/R36).
Gourse R.L., Ross W., and Gaal T. (2000). UPs and Downs in bacterial transcription initation: the role of alpha subunit of RNA polymerase in promoter recognition. *Molecular Microbiology.* 37(4), 687-695.
Gourse R.L., Gaal T., Aiyar S.E., Barker M.M., Estrem S.T., Hirvonen C.A., and Ross W. (1998) Strength and regulation without transcription factors: lessons from bacterial rRNA promoters. *Cold Spring Harb Symp Quant Biol.* **63**, 131-139. Review.
Gralla J.D. (1990) Promoter recognition and mRNA initiation by *Escherichia coli* E sigma70. in *Methods in Enzymology.* 185, 37-54.
Gralla J., and Collado-Vides J. (1996) In Neidhart F.C., Curtiss R., Ingraham J.L., Lin E.C.C., Low K.B., Magasanik B., Reznikoff W., Riley M., Schaechter M. and Umbarger H.E. (eds), *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC, 1232–1246.
Gross C.A., Chan C., Dombroski A., Gruber T., Sharp M., Tupy J., and Young B. (1998). The functional and regulatory roles of sigma factors in transcription. *Cold Spring Harb Symp Quant Biol.* 63, 141-155.

Gross C. (1996). In Neidhart,F.C., Curtiss,R., Ingraham,J.L., Lin,E.C.C., Low,K.B., Magasanik B., Reznikoff W., Riley M., Schaechter M. and Umbarger H.E. (eds), *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology*. American Society for Microbiology, Washington, DC, 1382–1399.

Harley C.B, and Reynolds R.P. (1987). Analysis of *E.coli* promoter sequences. *Nucleic Acids Res.* 15(5), 2343-2361.

Hawley D.K., and McClure W.R. (1983). Compilation and analysis of *Escherichia coli* promoter DNA sequences. *Nucleic Acids Res.* 11(8), 2237-2255.

Helmann J.D., and Chamberlin M.J. (1988) Structure and function of bacterial sigma factors. *Annu Rev Biochem.* 57, 839-872. Review.

Hertz G.Z., and Stormo G.D. (1999). Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics*. 15, 563-577.

Hertz G.Z., and G.D. Stormo G.D. (1996). *Escherichia coli* promoter sequences: analysis and prediction. *Methods Enzymol.* 273, 30-42. Review.

Hertz G.Z., Hartzell G.W. 3rd, and Stormo G.D. (1990). Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Comput Appl Biosci.* 6(2), 81-92.

Huerta A.M., Collado-Vides J. (2003) Sigma70 Promoters in *Escherichia coli*: Specific Transcription in Dense Regions of Overlapping Promoter-like Signals. *J Mol Biol.* 333(2), 261-278.

Ikebe T., Iyoda S., and Kutsukake K. (1999) Structure and expression of the fliA operon of *Salmonella typhimurium. Microbiology*. 145 (Pt 6), 1389-1396.

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

Lafebvre F. (1996) A grammar-based unification of several alignment and folding algorithms. *Proc Int Conf Intell Syst Mol Biol.* 4, 143-154.

Landau G.M., Vishkin U., and Nussinov R. (1986) An efficient algorithm with k differences for nucleotide and amino acid sequences. *Nucleic Acids Res.* 14(1), 31-46.

Lawrence C.E., Altschul S.F., Boguski M.S., Liu J.S., Neuwald A.F., and Wootton J.C. (1993). *Science*. 262, 208-214.

Lesley S.A., Jovanovich S.B., Tse-Dinh Y.C., and Burgess R.R. (1990) Identification of a heat shock promoter in the topA gene of *Escherichia coli. J Bacteriol.* 172(12), 6871-6874.

Lisser S., and Margalit H. (1993) Compilation of *E. coli* mRNA promoter sequences. *Nucleic Acids Res.* **21**, 1507-1516.

Lonetto M., Gribskov M., and Gross C.A. (1992) The sigma70 family: sequence conservation and evolutionary relationships. *J Bacteriol.* 174(12), 3843-3849. Review.

Lukashin A.V., Anshelevich V.V., Amirikyan B.R., Gragerov A.I., and Frank-Kamenetskii M.D. (1989). Neural network models for promoter recognition. *J.Biomolecular Struc & Dynamics.* 6(6), 1123-1133.

Man T.K., and Stormo G. (2001) Non-independence of Mnt repressoroperator interaction determined by a new quantitative multiple fuorescence relative affinity (QuMFRA) assay. *Nucleic Acids Res.* **15**, 2471-2478.

McClelland M., Sanderson K.E., Spieth J., Clifton S.W., Latreille P., Courtney L., Porwollik S., Ali J., Dante M., Du F., Hou S., Layman D., Leonard S., Nguyen C., Scott K., Holmes A., Grewal N., Mulvaney E., Ryan E., Sun H., Florea L., Miller W., Stoneking T., Nhan M., Waterston R., and Wilson R.K. (2001) Complete genome sequence of *Salmonella enterica serovar Typhimurium* LT2. *Nature*. 413(6858), 852-856.

Mitchell J.E., Zheng D., Busby S.J., and Minchin S.D. (2003) Identification and analysis of 'extended -10' promoters in *Escherichia coli. Nucleic Acids Res.* **31**(16), 4689-4695.

Mooney R.A., Artsimovitch I., and Landick R. (1998). Information processing by RNA polymerase: recognition of regulatory signals during RNA chain elongation. *J Bacteriol.* 180(13), 3265-3275. Review.
Moreno-Hagelsieb G., and Collado-Vides J. (2002). A powerful non-homology method for the prediction of operons in prokaryotes. Bioinformatics 18 Suppl 1, S329-S336.
Moyle H., Waldburger C., and Susskind M.M. (1991) Hierarchies of base pair preferences in P22 *ant* promoter. *J Bacteriol.* 173(6), 1944-1950.
Mulligan M.E., Hawley D.K., Entriken R., and McClure W. (1984). *Escherichia coli* promoters sequences predict *in vitro* RNA polymerase selectivity. *Nucleic Acids Res.* 12, 789-800.
Mulligan M. E., Brosius J., McClure and W. R. (1985). Characterization in vitro of the effect of spacer length on the activity of *Escherichia coli* RNA polymerase at the TAC promoter. *The J. of Biological Chem.* 260(6), 3529-3538.
Murakami K.S., Masuda S., and Darst S.A. (2002) Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution. *Science.* **296**, 1280-1284.
Murakami K.S., Masuda S., Campbell E.A., Muzzin O., and Darst S.A. (2002b) Structural Basis of Transcription Initiation: An RNA Polymerase Holoenzyme-DNA Complex. *Science.* **296**, 1285-1290.
Neidhardt F.C., Curtiss R., Jingraham L., Lin E.C.C., Low K.B., Magasanik B., Reznikoff W.S., Riley M., Schaerchter M., and Umbarger H.E. (eds) (1996) *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology.* American Society for Microbiology, Washington, DC.
O'Neill M.C. (1992). Escherichia coli promoters: neural networks develop distinct descriptions in learning to search for promoters of different spacing classes. *Nucleic Acids Res.* 20(13), 3471-3477.
O'Neill M.C. (1989). Consensus method for finding and ranking DNA binding sites. Aplication to *Escherichia coli* promoters. *J. Mol. Biol.* 207, 301-310.
O'Neill M.C. (1989b). Escherichia coli promoters. I. Consensus as it relates to spacing class, specificity, repeat substructure, and three-dimensional organization. *J Biol Chem.* 264(10), 5522-5530.
O'Neill M.C. (1998) A general procedure for locating and analyzing protein-binding sequence motifs in nucleic acids. *Proc Natl Acad Sci U S A.* **95**(18), 10710-10715.
Ozoline O.N., Deev A.A., and Arkhipova M.V. (1997). Non-canonical sequence elements in the promoter structure. Cluster analysis of promoters recognized by *Escherichia coli* RNA polymerase. *Nucleic Acids Res.* 25(23), 4703-4709.
Ozoline O.N., Deev A.A., Arkhipova M.V., Chasov V.V, and Travers A. (1999). Proximal transcribed regions of bacterial promoters have a non-random distribution of A/T tracts. *Nucleic Acids Res.* 27(24), 4768-4774.
Paget M.S., and Helmann J.D. (2003) The sigma70 family of sigma factors. *Genome Biol.* **4**(1), 203. Review.
Pearson W.R. (2000). Flexible sequence similarity searching with the FASTA3 program package. *Methods Mol Biol.* 132, 185-219.
Pedersen A.G., Baldi P., Brunak S., and Chauvin Y. (1996) Characterization of Prokaryotic and Eukaryotic Promoters Using Hidden Markov Models. *ISMB.* 182-191.
Plaskon R.R., and Wartell R.M. (1987) Sequence distributions associated with DNA curvature are found upstream of strong *E. coli* promoters. *Nucleic Acids Res.* 15(2), 785-796.
Pribnow D. (1975). Nucleotide sequence of an RNA polymerase binding site at an early T7 promoter. *Proc. Natl. Acad. Sci USA.* 72, 784-788.

Raibaud O., and M. Schwartz M. (1984). Positive control of transcription initiation in bacteria. *Annu. Rev. Genet.* 18,173-206. Review.
Raina S., and Georgopoulos C. (1990) A new *Escherichia coli* heat shock gene, htrC, whose product is essential for viability only at high temperatures. *J Bacteriol.* 172(6), 3417-3426.
Rao L., Ross W., Appleman J.A., Gaal T., Leirmo S., Schlax P.J., Record M.T. Jr., and Gourse R.L. (1994) Factor independent activation of *rrnB* P1. An "extended" promoter with an upstream element that dramatically increases promoter strength. *J Mol Biol.* 235(5), 1421-1435.
Reznikoff W.S., Bertrand K., Donnelly C., Krebs M., Maquat L.E., Peterson M., Wray L., Yin J., and Yu X-M. (1987). Complex promoters. In Reznikoff W.S., Burgess R.R., Dahlberg J.E., Gross C.A., Record M.T. Jr, and Wickens M.P. (eds). *RNA Polymerase and the Regulation of Transcription.* Elsevier, New York. 105-113.
Reznikoff W.S. (1992) The lactose operon-controlling elements: a complex paradigm. *Mol. Microbiol.* **6**(17), 2419-2422. Review.
Richmond C.S., Glasner J.D., Mau R., Jin H., Blattner F.R. (1999) Genome-wide expression profiling in Escherichia coli K-12. *Nucleic Acids Res.* 27(19), 3821-3835.
Riley M., and Labedan. B. (1996) *E.coli* gene products: Physiological functions and common ancestries" In *Escherichia coli and Salmonella: Cellular and Molecular Biology* (ed. Neidhardt, F. *et. al*), 2118-2202. ASM Press, Washington, D.C. 2nd Ed.
Rosenberg M.,and D. Court (1979). Regulatory sequences involved in the promotion and termination of RNA transcription. *Annu Rev Genet.* 13, 319-353. Review.
Ross W., Gosink K.K., Salomon J., Igarashi K., Zou C., Ishihama A., Severinov K., and Gourse R.L. (1993) A third recognition element in bacterial promoters: DNA binding by the alpha subunit of RNA polymerase. *Science.* 262(5138), 1407-1413.
Roth F.P., Hughes J.D., Estep P.W., and Church G.M. (1998). Finding DNA regulatory mitifs within unaligned noncoding sequences clustered by whole-genome quantitation. *Nat Biotechnol.* 16, 939-945.
Salgado H., Santos-Zavaleta A., Gama-Castro S., Millan-Zarate D., Díaz-Peredo E., Sánchez-Solano F., Pérez-Rueda E., Bonavides-Martinez C., and Collado-Vides J. (2001) RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12. *Nucleic Acids Res.* **29**(1), 72-74.
Schneider T.D., Stormo G.D., Gold L., and Ehrenfeucht A. (1986) Information content of binding sites on nucleotide sequences. *J Mol Biol.* **188**(3), 415-431.
Snyder L., and Champness W. (1997) Molecular Genetics of Bacteria. *ASM Press Washington, D.C.* ISBN 1-55581-102-7.
Staden R. (1984). Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res.* 12(1), 505-519.
Stormo G.D., Schneider T.D., Gold L., and Ehrenfeucht A. (1982). Use of the 'Perceptron' algorithm to distinguish traslational initiation sites in *E.coli. Nucleic Acids Res.* 10, 2997-3011.
Stormo G.D. (2000). DNA binding sites: representation and discovery. *Bioinformatics,* Special History Issue. 16(1), 16-23.
Stormo G.D., and Fields D.S. (1998). Specificity, free energy and information content in protein-DNA interactions. *Trends Biochem. Sci.* 23, 109-113.
Tremblay J.P., and Manohar R. (1987) Discrete Mathematical Structures with Applications to Computer Science. *McGraw-Hill Book Company*, ISBN 0-07-100322-3.
van Helden J., André B., and Collado-Vides J. (1998). Extracting Regulatory Sites from the Upstream Region of *Yeast* Genes by Computational Analysis of Oligonucleotide Frecuencies. *J. Mol. Biol.* 00, 1-16.

Vanet A., Marsan L., and Sagot M.F. (1999) Promoter sequences and algorithmical methods for identifying them. *Res Microbiol.* 150(9-10), 779-799.
Walker G.C. (1996) In Neidhart F.C., Curtiss R., Ingraham J.L., Lin E.C.C., Low K.B., Magasanik B., Reznikoff W., Riley M., Schaechter M., and Umbarger H.E. (eds), *Escherichia coli and Salmonella typhimurium, Cellular and Molecular Biology*. American Society for Microbiology, Washington, DC. 1334.
Wall L., and Schwartz R.L. (1991). Programming Perl. *O'Reilly and Associates, Inc., Sebastol, CA*. ISBN 0-937175-64-1.
Waterman M.S., Galas D., and Arratia R. (1984b). Pattern Recognition in Several Sequences: Consensus and Aligment. *Bull. Math. Biol.* 46(4), 515-527.
Waterman M.S. (1984a). General Methods of Sequence Comparison. *Bull. Math. Biol.* 46(4), 473-500.
Weller K., and Recknagel R.D. (1994). Promoter strength prediction based on occurrence frequencies of consensus patterns., *J. theor. Biol.* 171(4), 355-359.
Youderian P., Bouvier S., and Susskind M.M. (1982). Sequence determinants of promoter activity. *Cell.* **30**, 843-853.
Zhou Y.N., Kusukawa N., Erickson J.W., Gross C.A., and Yura T. (1988) Isolation and characterization of *Escherichia coli* mutants that lack the heat shock sigma factor sigma 32. *J Bacteriol.* 170(8), 3640-3649.