03091

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

UNIDAD ACADÉMICA DE LOS CICLOS PROFESIONAL Y DE POSGRADO DEL COLEGIO DE CIENCIAS Y HUMANIDADES

EVALUACIÓN Y DETERMINACIÓN DE CRITERIOS PARA LA DETECCIÓN DE SITIOS DE UNIÓN EN EL DNA PARA LOS REGULADORES TRANSCRIPCIONALES DE *E. coli* K12

T E S I S

PRESENTADA POR

ESPERANZA BENÍTEZ BELLÓN

PARA OBTENER EL TÍTULO DE

DOCTORADO EN CIENCIAS

INSTITUTO DE BIOTECNOLOGÍA

CUERNAVACA, MOR.

MAYO DE 2002

El presente trabajo se realizó en el Programa de Biología Molecular Computacional del Centro de Investigación sobre Fijación de Nitrógeno de la Universidad Nacional Autónoma de México, con la asesoría y bajo la supervisión del Dr. Julio Collado Vides y la asesoría del Dr. Gabriel Moreno Hagelsieb.

A mi corazón, Rodrigo

"Evaluación y determinación de criterios para la detección de sitios de unión en el DNA para los reguladores transcripcionales en *E. coli* K12"

INDICE DE FIGURAS

INDICE DE TABLAS

RESUMEN

La detección computacional de las secuencias de unión a DNA (descubrimiento de patrón = pattern discovery) de una colección de genes considerados co-regulados, como resultado de datos de transcriptomas o por comparación de genomas, o por el uso de secuencias de sitios de unión conocidas para encontrar otros genes regulados por la misma proteína (búsqueda de patrón = pattern search), son estrategias importantes en el descubrimiento y enriquecimiento de las redes de regulación. El experimentalista puede necesitar una predicción precisa de los sitios, o sin mucho detalle, la identificación de otros genes dentro del mismo regulón. Como han sido muchos los métodos de identificación de sitos de unión publicados, una minuciosa evaluación de éstos es necesaria. En este trabajo nos enfocamos al estudio de dos algoritmos computacionales publicados con ese fin: Dyad-analysis (van Helden *et al.*, 2000) y Consensus (Hertz and Stormo, 1999). Para ello, contamos en el laboratorio con la información referente a las familias, genes, proteínas de interés y secuencias que forman la base de datos, RegulonDB (Salgado *et al.*, 2001). De igual forma, se utilizó la información proveniente de las secuencias del genoma completo de *E. coli* que se encuentra en el genbank (Blattner *et al.*, 1997). Colaboramos con el desarrollo de una nueva estrategia que, mediante el barrido base por base de las secuencias analizadas, se utiliza para determinar las secuencias de los sitios de unión de los reguladores transcripcionales. La elección de un umbral adecuado para la admisión de nuevos miembros a una familia de regulación es de gran importancia debido a la necesidad de evitar un alto número de falsos positivos y asegurar el mayor número posible de verdaderos positivos. En este trabajo se analizó la eficacia de los resultados obtenidos con los dos algoritmos en el descubrimiento de patrones, y el nivel de confianza en las predicciones propuestas y se estableció que la combinación de los dos algoritmos de búsqueda es la mejor estrategia para la detección de sitios de unión al DNA de los reguladores transcripcionales de *E. coli* K12.

ABSTRACT

Sites in DNA that bind regulatory proteins can be detected computationally in various ways. The computational detection of DNA-binding sites from a collection of genes suspected to be co-regulated (pattern discovery) as a result, for instance, of clustering of transcriptome data, or the use of sequences with known binding sites to find other genes regulated by a given protein (pattern search) are important strategies in the discovery and enrichment of regulatory networks. The experimentalist may need a precise prediction of a site, or to identify a gene within a regulon. As more variations of methods are published, thorough evaluation is necessary. The performance of the methods may differ depending on the conditions of their usage. A detailed evaluation also helps to improve and understand the behavior of the different methods and computational strategies. In this paper, we search for regulatory motifs using two reported methods: CONSENSUS and Dyad detector (first time to prokaryote oligonucleotide analysis). To define optimize parameters to prediction and the efficiency in transcriptional protein binding-sites determination, is essential a good evaluation of computer algorithms to computational identification of *cis*-regulatory elements. For each regulatory family, we want to know its regulon, that is, all the genes each regulatory protein regulates. At the same time, we have adopted a combined approach, the evaluation of the algorithm *per se*, with the optention of optimized method to find unknown binding sites in some new coregulated genes cluster. As more variations of methods are published, thorough evaluation is necessary. The performance of the methods may differ depending on the conditions of their usage. A detailed evaluation also helps to improve and understand the behavior of the different methods and computational strategies.

ABREVIATURAS

E. coli	*Escherichia coli*
RNApol	RNA-polimerasa
Pb	Pares de bases
DNA	Acido-Desoxirribo-Nucleico
HTH	Helix-turn-helix
LOO	Leave-one-out
VP	Verdaderos positivos
FP	Falsos positivos
VN	Verdaderos negativos
FN	Falsos negativos
PPV	Positive predictive value
OP	Overall performance
ROMs	Regions of Overlapping Matches

Introducción

Escherichia coli es sin lugar a dudas uno de los organismos mejor estudiados de todos los tiempos. Es un procariote muy simple, bacteria Gram negativa, de grandes ventajas experimentales, que gracias a las técnicas de biología molecular y de bioinformática, ha ayudado a dilucidar muchos de los procesos celulares de los organismos vivos.

En el genoma completo de la cepa K12 de *E.coli* se han identificado aproximadamente unas 4300 regiones codificantes. Uno de los procesos mejor descritos y al que enfocamos nuestro interés, así como muchos otros investigadores en estos momentos de la era genómica, es la regulación de la transcripción [1-3]. El proceso de la transcripción consiste en el copiado de la secuencia molde de DNA en una secuencia complementaria de RNA. La regulación de este proceso es de vital importancia para las células ya que los genes codificados en el DNA son específicamente seleccionados para su expresión. Cada situación ambiental o momento de desarrollo por el que atraviesa una célula necesita la expresión de determinados genes, o para ser más precisos, en un organismo las células expresan un grupo de genes idóneos a los necesitados para un equilibrio óptimo.

La regulación de la transcripción es por lo tanto uno de los pasos relevantes en la vida celular y a la vez, es uno de los pasos para nuestro entendimiento del complicado proceso de expresión de la información genética. Para la regulación a nivel de transcripción, nos enfocaremos a la regulación en su inicio. Diferentes proteínas se coordinan para una primera localización del promotor, posteriormente se lleva a cabo la formación del complejo que permite el inicio, y por cambios sucesivos prosiguen la elongación y la terminación de la transcripción. Todas estas proteínas tienen que encontrarse en las concentraciones necesarias para su óptimo funcionamiento, teniendo en cuenta que la competencia entre secuencias es también relevante. Las diferentes

subunidades que formarán la holoenzima de la RNA polimerasa, mencionadas a continuación, junto con todos los reguladores que determinan el pegado de este complejo al DNA, además de los factores que regularán su permanencia en el lugar o el que se despegue de dicha secuencia, son posibles sujetos de regulación. El promotor es una región de DNA que posee señales estructurales que facilitan su interacción con la holoenzima RNA-polimerasa (RNApol). La enzima básica es un complejo multimérico con cuatro subunidades: dos subunidades α, una β y otra β'. Los promotores son reconocidos por el factor σ, que aparece en cada condición de crecimiento de las células, y la unión de este factor con la enzima básica es lo que forma la holoenzima. El factor sigma primario es el factor σ70, sin embargo, existen otros muchos factores σ que reconocen diferentes secuencias en el promotor y su unión está determinada por diferentes condiciones "heat-shock" (σ32 y σ24), fase estacionaria (σ38), resistencia a compuestos (σ38), asimilación de nitrógeno (σ54). Para el reconocimiento específico del promotor, acción llevada a cabo por la holoenzima, dos secuencias de hexanucleótidos se mantienen conservadas, posibles responsables de la arquitectura del promotor, la caja Pribnow (TATAAT) aproximadamente 10pb arriba del inicio de la transcripción y la caja que se encuentra a 35pb del mismo sitio. Un porcentaje mayor al 95% de los genes está regulado por al menos dos promotores.

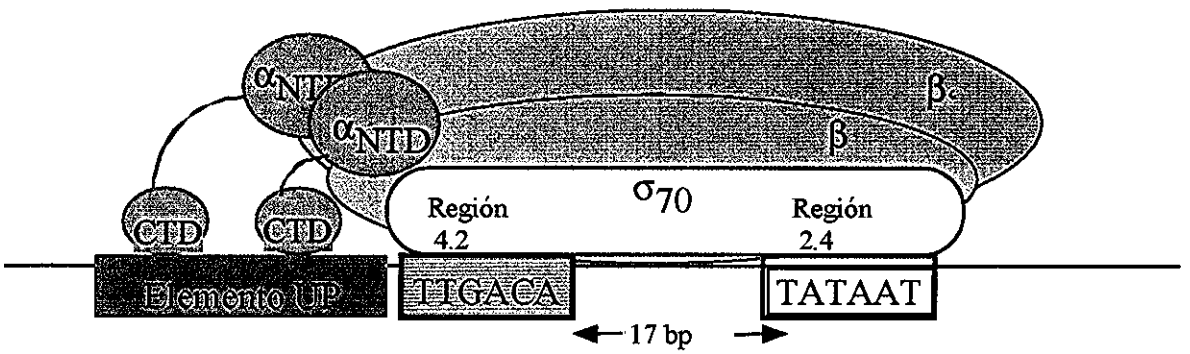


Figura 1: Ejemplo de una holoenzima de la RNA-polimerasa y su localización en un promotor σ70.

De manera adicional a la unión al promotor, se necesitan una serie de proteínas que ejerzan una función reguladora mediante su unión al DNA. La localización de los sitios de unión de los reguladores transcripcionales es variable en los diferentes genes, sin embargo, está determinada por la función que éstos ejercen [1, 4].

La organización de los genes en *E.coli* y en procariotes en general, es formando sistemas coordinados que se denominan operones y/o regulones, lo que facilita una respuesta sincronizada.

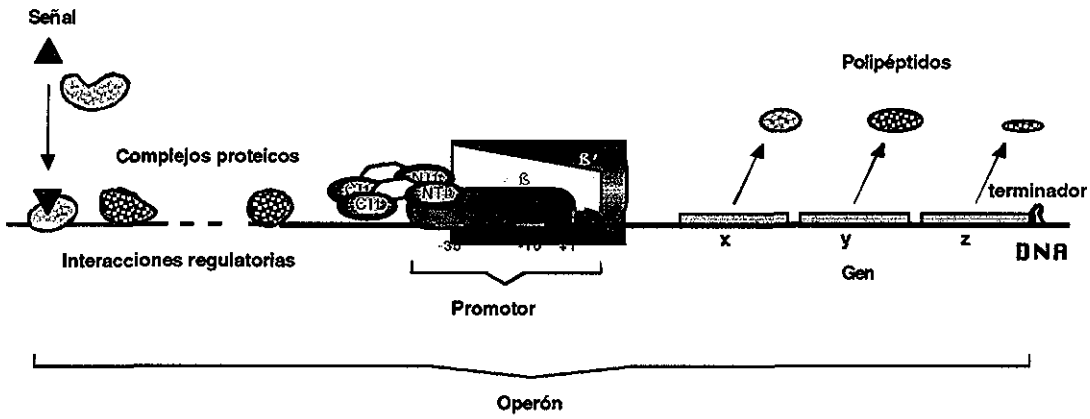


Figura 2: Diagrama que representa un ejemplo de la organización de los genes en operones y de la región reguladora río arriba de éstos.

Los datos experimentales recopilados, han facilitado la agrupación de genes co-regulados para, como siguiente paso, analizar y caracterizar el mecanismo responsable de la respuesta de todos ellos. Con el análisis de las regiones río arriba de los genes seleccionados, podemos extraer la información para detectar la identidad de los sitios de unión de los reguladores, su localización, la presencia conjunta con otros reguladores, así como conocer si la función se realiza de manera coordinada, y de todos los datos que conlleven a un mejor entendimiento de la regulación transcripcional [5].

En resumen, la modulación de la afinidad de la holoenzima por los promotores está mediada por el efecto que ejercen los sitios de unión de los reguladores transcripcionales permitiendo la expresión o no de cada gen.

Las proteínas reguladoras se clasifican dependiendo de la función que realizan. Así, encontramos reguladores transcripcionales positivos, negativos o duales. Cada uno de los sitios de unión de estas proteínas al DNA es una secuencia que permite, desde el lugar donde se encuentra, el efecto de activación o de represión específico del regulador. Cada organismo posee una extensa relación de reguladores transcripcionales que presentan secuencias de unión asociadas. Como consecuencia de la gran cantidad de información disponible de los sitios de unión, nos enfocamos en un principio a caracterizar de nuevo la existencia de una relación de posición del sitio de unión respecto al tipo de acción del regulador [4]. En la siguiente figura se muestra dicha relación en el caso de activadores y de represores transcripcionales en *E.coli* K12.

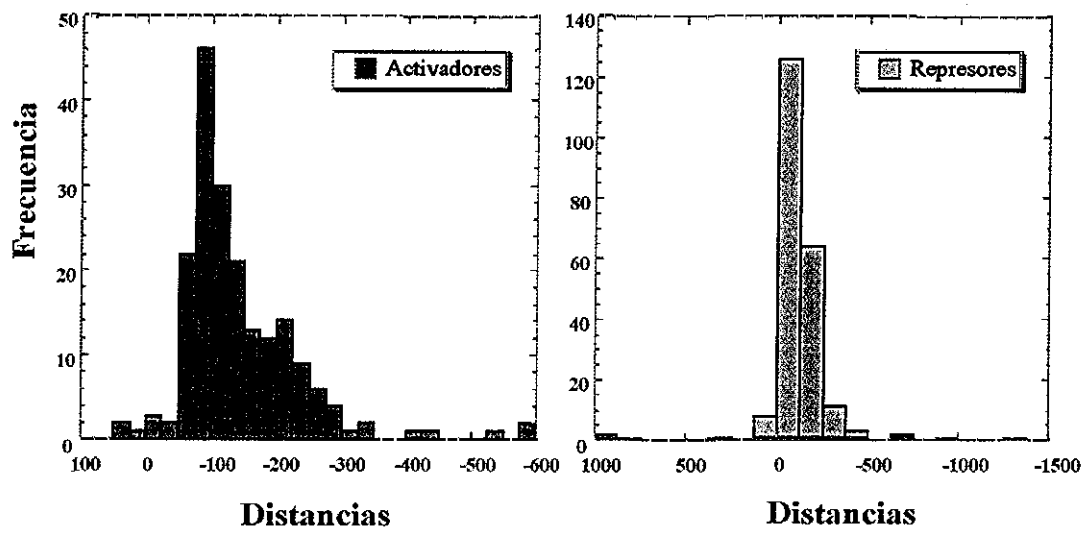


Figura 3: Histograma de las distancias de las coordenadas de los sitios de unión de los reguladores (Activadores y Represores), al inicio de la traducción del gen.

El porcentaje mayor de los sitios de unión de represores en el genoma de *E.coli* se localiza en las coordenadas cercanas al sitio de inicio de la traducción. La posición de aproximadamente 20 pb respecto al inicio de la transcripción se considera una región prohibida para la unión de los activadores. Las secuencias de unión de los activadores se encuentran preferentemente más lejos del ATG al igual que las de los duales, cuya frecuencia de aparición es mucho menor. Cada organismo posee un gran número de reguladores transcripcionales que presentan secuencias de unión asociadas. Es difícil, por ese número tan grande, el estudio de la identificación exacta del elevado número de secuencias de unión de proteínas reguladoras. Todas estas secuencias se encuentran habitualmente en la zona no codificante de los genes, río arriba de éstos, y en su mayoría a distancias no superiores a las 500 pares de bases (pb) del inicio del gen. Los dominios de unión de los reguladores al DNA se conservan habitualmente dentro de las diferentes familias y reflejan una amplia variedad de estructuras. Se presentan de varios tipos: ß plegada antiparalela, hélices-asa-hélices, dedos de Zinc [6], hélice vuelta hélice (HTH). Aunque estas estructuras pueden aparecer en proteínas o enzimas que no se unen al DNA (por ejemplo en estabilizadores), el conocimiento de ellas [7] nos puede permitir un enfoque más dirigido a la búsqueda de la secuencia [8-10]. Los contactos con el DNA de

Motivo	Ejemplos
Hélice-vuelta-hélice	CRP,Fur,BirA,PurR
Hélice-asa-hélice	DnaA
Dedos de Zinc	HypF
Hojas β	MetJ,ArcA

Tabla 1: Estructuras de unión al DNA en *E.coli*.

los reguladores transcripcionales dentro de una estructura específica pueden llevarse a cabo por puentes de hidrógeno, por contactos mediados por moléculas de agua, por interacciones de Van de Waals, etc. La localización potencial de los dominios de unión en los reguladores es una característica que fue estudiada en nuestro laboratorio [11] Tabla 1.

En Procariotes, la estructura de la mayoría (>75%) de los represores muestra el motivo HTH en la posición N-terminal, mientras que los activadores lo presentan en el extremo C-terminal, este diseño refleja una simetría que puede ser importante para la elección del algoritmo de búsqueda. Otra característica importante es que cada regulador se encuentra en su forma activa adoptando la estructura terciaria idónea, monómeros en ciertos casos, homo o heterodímeros, o bien polímeros. Los reguladores transcripcionales procariotes son en general homodímeros y reconocen sitios palindrómicos en el DNA.

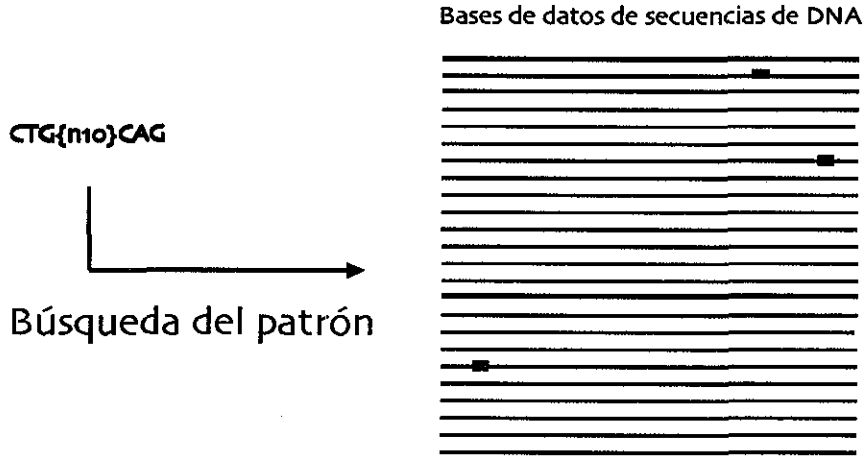
El conocimiento de estas secuencias génicas es de gran importancia, sin embargo, entre la identificación y el entendimiento de lo que dicha secuencia determina, existe un gran trecho. Realizando un estudio de los antecedentes más relevantes apreciamos que, casualmente, el comienzo del siglo 21 está acompañando el comienzo de una nueva era científica, la era genómica. En los años 70's se comenzó con la secuenciación de un gran número de genes, llegándose a la secuencia genómica completa de un ser vivo en 1995 [12]. La información disponible de todos los genomas secuenciados en la actualidad nos permite colaborar en el proceso de descifrar lo que dichas secuencias determinan. Está claro que, con el uso de la biología computacional, se abrieron posibilidades para identificar características importantes en las secuencias. El identificar los sitios de unión a DNA de proteínas reguladoras es nuestro aporte al conocimiento genómico de la bacteria *E. coli* K12 [13-15].

Es importante recordar que las secuencias de unión de los reguladores presentan una alta variabilidad, esto posibilita un control más preciso del proceso de la transcripción [16].

Una misma proteína reguladora puede ejercer su acción en genes diferentes. De manera adicional, la expresión de éstos está determinada por diferentes factores y a diversos niveles [17-20].

La identificación de la secuencia consenso para un regulador transcripcional decidimos enfocarla con dos objetivos diferentes: descubrir o determinar la secuencia de un sitio de unión a DNA de un regulador desconocido, partiendo de las secuencias río arriba de genes que sabemos se encuentran co-regulados (pattern discovery) , o bien el desarrollo de

una estrategia de búsqueda a partir de los sitios conocidos para encontrar otros genes también co-regulados (pattern search).



La Bioinformática o la Biología Computacional, ha presentado un avance espectacular en la detección de los sitios de unión. La regulación de la expresión genética es, sin embargo, uno de los misterios centrales en la era post-genómica. Por varias razones la detección computacional de los sitios de unión de los reguladores transcripcionales al DNA es un problema difícil. Se debe principalmente a las características propias que presentan estos sitios: la variabilidad, el tamaño, la localización, el número, etc. La variabilidad existente entre los sitios de cada uno de los genes de un mismo regulón es la primera dificultad. La proteína reguladora se une a secuencias que reconoce, sin embargo, las diferencias entre cada uno de los sitios de unión no nos permite una fácil identificación de éstos. La localización preferente de los sitios de unión (ver Fig3) de activadores, represores y duales es aproximadamente 400 pb río arriba del comienzo de los genes, aunque en el caso de los represores se pueden encontrar en regiones dentro del ORF al que regulan. Otra característica es el tamaño tan pequeño que presentan la mayoría de los sitios, aunque el promedio se encuentra en 20 pb, los tamaños varían entre ~ 10 y ~ 60 pb, la localización de esos fragmentos tan pequeños y tan poco repetitivos en el genoma completo no es tarea fácil. El número de sitios dentro de una misma región "upstream" es otro de los parámetros relevantes, en general se encuentran entre 1 y 2 por gen y aunque la agrupación de genes en familias nos permite cierta sobre-representación, el número de genes que pertenecen a un mismo regulón es insignificante comparado con el número total de genes del cromosoma a analizar (en nuestro caso el de *E. coli* K12). Algunos sitios presentan simetría debido al tipo de unión de la proteína reguladora, aunque no es una característica común a todos. Una aproximación al problema es bajo la hipótesis de que analizando simultáneamente varios genomas relacionados, si un gen está conservado en otra especie, la regulación del grupo de genes (regulón) de esa otra especie también debe estar conservada [43].

El conocimiento de la localización preferente de los sitios de unión, de la agrupación de genes en operones, del tamaño promedio de los sitios y de su número es una ayuda para la elección de algoritmos apropiados, aunque el conocimiento preciso de ellos por datos experimentales es quizás la mejor información de la que podemos hacer un uso más dirigido [3]. En Regulón DB se encuentra la información de los sitios de unión de 52 reguladores que a su vez se encuentran agrupados en familias (regulones) sin embargo, por estudios genómicos ya realizados [14] consideramos que el número de reguladores posibles (conocidos y predichos) es de 314 y que necesitamos información de contexto adicional para el reconocimiento de patrones o para la identificación de patrones de estos sitios.

Existen muchos métodos diferentes para cada una de las propuestas de búsqueda, entre ellos: Consensus/Patser [21-23], Gibbs Sampler y AlignACE [24-26], Pratt [27], Patrones sobre-representados [28] como Oligo y Dyad Analysis [29], usando un "diccionario" de motivos o palabras [30, 31], usando un "modelo gramatical" [32, 33], por métodos Bayesianos [34, 35], u otros [36].

A continuación detallo algunas características de algunos de los métodos más relevantes. **Consensus** (Hertz *et al.*, 1999; Stormo *et al.*, 2000) es un algoritmo de alineamiento múltiple que usa la teoría de información para encontrar el mejor alineamiento que contenga las secuencias estadísticamente significativas y **Patser** usa esas matrices para identificar los sitios en el DNA estadísticamente significativos. El programa Consensus identifica patrones con el más alto contenido informacional en un grupo dado de secuencias.

Uno de los tipos de matrices más simples es la matriz de alineamiento que nos proporciona el número de ocurrencias de cada letra en una posición determinada de un alineamiento. Otro tipo de matriz es la matriz de peso cuyos elementos reflejan un determinado peso para evaluar una secuencia problema midiendo qué tan cerca se

encuentran los "matches" de esa secuencia con los del patrón descrito por la matriz. La secuencia prueba se alinea con la matriz de peso y sus evaluaciones ("scores") son la suma de los pesos correspondientes a cada letra alineada en cada posición. Las matrices de peso pueden derivar de las matrices de alineamiento o estar determinadas experimentalmente.

Contenido informacional: Es la representación del grado de la conservación de una secuencia en una columna de una matriz de peso que representa un alineamiento de secuencias relacionadas. Si en la comparación de alineamientos de las matrices de alineamiento se asume que cada una de las secuencias son independientes e idénticamente distribuidas, entonces la probabilidad *a priori* de una secuencia de letras es el producto de las probabilidades *a priori* de cada una de las letras individuales. Esas probabilidades de las letras individuales deben ser las frecuencias totales de las letras dentro de todas las secuencias de un organismo, de un subset de secuencias o de los datos que se analizan. Se asume que los alineamientos más interesantes son aquellos en los que las frecuencias de las letras difieren mas de las probabilidades *a priori* de esas letras. El contenido informacional se obtiene por la normalización (N = número total de secuencias) de la relación de "log-likelihood" y es interesante destacar que se encuentra relacionado con la termodinámica. En particular, el contenido informacional es una medida de la discriminación entre el apareamiento dado por una secuencia de DNA funcional y por una secuencia de DNA arbitraria. La significancia estadística está dada por la "expected frequency" y puede ser usada para comparar alineamientos que presentan diferente tamaño y contienen diferente número de secuencias.

Consensus/Patser presenta una serie de problemas, como que aísla un único elemento de cada familia lo que no resulta óptimo para detectar varios elementos diferentes. Del mismo modo, el aislamiento de elementos a los que se unen los reguladores requiere de una elección *a priori*: el largo de la matriz y el número esperado de "matches".

Oligo-analysis (van Helden *et al.*, 1998) es un método basado en la sobre expresión de determinadas oligonucleótidos dentro de unas secuencias. El análisis es riguroso y exhaustivo. Este método usa una calibración para estimar la frecuencia esperada de cada oligo. En base al número de ocurrencias, las frecuencias esperadas y utilizando una binomial, se llega al cálculo de la significancia. El valor más alto de este parámetro corresponde al patrón más sobre-representado. Cuando se seleccionan únicamente los patrones para los cuales la significancia es >=0, uno espera que ocurra menos de un patrón al azar dentro de una familia. Uno espera encontrar con sig >=1 un patrón al azar cada 10 familias, con sig >=2 uno cada 100 familias, etc. La significancia es un parámetro importante ya que se puede usar para comparar. Esta característica es debida a que, es independiente del tamaño del oligonucleótido utilizado, del tamaño de las secuencias a analizar y del número de genes que se encuentren en cada una de las familias. Un problema que presenta este método es que está limitado a la detección de motivos cortos y relativamente conservados.

El descubrimiento de elementos reguladores en las secuencias no codificantes por análisis de diadas se lleva cabo mediante el algoritmo riguroso **Dyad-analysis** (van Helden *et al.*, 2000). Este método se ideó para descubrir elementos reguladores de un juego de secuencias no alineadas río arriba de genes. Se basa en la observación de que muchos sitios reguladores consisten en un par de trinucleótidos altamente conservados, separados por una región no conservada de un ancho fijo. El procedimiento consiste en contar el número de ocurrencias de cada posible par de trinucleótidos espaciados y determinar su significancia estadística.

La diada está formada por un par de palabras cortas conservadas separadas por una región de determinado tamaño y contenido variable. Las palabras pueden determinarse en un número pequeño (3 ó 4), el espacio interior puede tomar diferentes valores dependiendo del factor transcripcional (de o a 16). Se evaluó la frecuencia de cada diada en el grupo

completo de las secuencias no codificantes del organismo y se usó esa frecuencia como un estimado directo de las frecuencias de diadas esperadas en una familia de secuencias "upstream". La significancia estadística del número observado de ocurrencias se obtiene por una binomial y nos da un estimado intuitivo de la sobre representación. Fijando el umbral de la significancia en 0, el investigador podría esperar no mas de un patrón al azar dentro de cada familia y eso es independiente del largo de la secuencia, del tamaño de la palabra y del intervalo de espaciamiento considerado. La significancia se puede usar para comparar probabilidades entre patrones de diferente tamaño. Las diadas pueden presentar dos tipos de simetría, directa e inversa.

Otro método basado en la sobre-representación de determinadas secuencias es el elaborado por **Brãzma** (Brãzma A. *et al.*, 1998). El algoritmo se desarrolla para el descubrimiento de patrones que se buscan exhaustivamente para patrones regulares *a priori* desconocidos que se encuentran sobre-representados en un juego de secuencias dadas. Se aplica para el descubrimiento de patrones del conjunto completo de secuencias a analizar y para la búsqueda en regiones "upstream" de genes con patrones similares de expresión. La diferencia fundamental con otros algoritmos es que en éste no se parte de conocimientos *a priori* de cómo ni cuantos son los patrones a encontrar en la secuencia de un genoma completo.

Gibbs sampler (Lawrence *et al.*, 1993; Neuwald *et al.*, 1995) es un método muy sensible, heurístico, no riguroso, no exhaustivo y estocástico. Es un algoritmo de múltiples alineamientos locales que no asume información *a priori* de los patrones, de su localización dentro de las secuencias y que determina sus localizaciones de la información intrínseca de las secuencias mismas. Detecta palabras sobre-representadas como otros algoritmos, por ejemplo Oligo-analysis. El modelo tiene tres características fundamentales: 1) busca un número relativamente pequeño de patrones o elementos de secuencias, 2) un solo patrón está descrito por un modelo probabilístico de las frecuencias de los residuos en cada

posición, 3) la localización del patrón dentro de la secuencia está descrita por un conjunto de posiciones variables inferidas probabilísticamente. Estas características derivan de los principios bien establecidos de estructuras de proteínas y del conocimiento del origen de las variaciones de patrones en las secuencias. Presenta ciertos problemas por el tiempo que tarda en ejecutarse ya que comienza el análisis con un conjunto de posiciones arbitrarias que convergen en una matriz óptima y dependiendo de la posición de inicio dará resultados diversos por lo que conviene ejecutar varias repeticiones del mismo análisis.

Co-Bind (por Cooperative BINDING) (GuhaThakurta D. and Stormo G.D., 2001) es un algoritmo basado en que la activación transcripcional en organismos eucariotes requiere normalmente de la interacción combinatoria de múltiples factores transcripcionales. El método utiliza la estrategia de "Gibbs sampling" para modelar la cooperatividad entre dos factores transcripcionales y define la posición de matrices de peso para los sitios de unión. En aquellos casos en los que los patrones de los sitios de unión son muy débiles y no pueden ser identificados por otros métodos, Co-Bind los puede identificar eficientemente debido al sinergismo entre ellos. Aquí, dos matrices de peso son usadas para representar el sitio de unión de dos factores transcripcionales y su energía de unión combinatoria está dada por la suma de las energías de unión individuales. La probabilidad de que una molécula de factor transcripcional se una a su sitio en cada una de las secuencias del conjunto de positivos está dada por el producto de cada probabilidad individual y aquí se intenta maximizar esa probabilidad. Los principios termodinámicos en los que se basa Co-Bind pueden ser aplicados en otros casos de cooperatividad, en interacciones secuencia específicas macromoleculares, por ejemplo, en las regiones río arriba de los genes de *E. Coli*, Co-Bind puede identificar los sitios de inicio de la traducción al combinar las secuencias señales de inicio (codón de inicio) y la región de unión de ribosomas.

AlignACE (Roth *et al.*, 1998) es un programa de alineamiento local basado en el algoritmo Gibbs-sampling, para identificar motivos que se encuentran sobre-representados y está optimizado para alineamientos de secuencias de DNA. AlignACE escoge alineamientos estadísticamente significativos en la secuencia que se va a estudiar. Sin embargo, los alineamientos de las señales reguladoras son frecuentemente débiles comparados con otros elementos genómicos comunes, como por ejemplo los sitios de unión a ribosomas.

En algunos estudios, con la intención de seleccionar motivos significativos que, por la mayoría de las características, fuesen motivos reguladores, calcularon varios índices para cada matriz: MAP "score" (de maximum *a priori* "log-likelihood") (basado en la sobre-representación del motivo), "score" sitio específico (mide la especificidad de ese motivo en esa secuencia comparado con el genoma completo), posición preferencial, contenido en AT y palindromicidad. El conjunto de estos índices funcionaba como filtro a las primeras aproximaciones o estudios realizados con Gibbs-sampling.

Gibbs-sampling difiere de este algoritmo en que: 1) los modelos de patrones se cambiaron de tal manera que, las frecuencias de las bases de secuencias que no son sitios se adaptaron a las frecuencias del genoma fuente, 2) ambas cadenas de la secuencia que se analizó son simultáneamente consideradas en cada paso del algoritmo, el traslape de los sitios no está permitido aunque los sitios se encuentren en cadenas opuestas, 3) la búsqueda de motivos múltiples simultáneos fue reemplazada por un procedimiento en el que motivos simples fueran buscados y marcados, 4) Para AlignACE el valor "MAP score" es ahora el criterio en el cual se basa la última salida del algoritmo.

Hidden Markov model (HMM) (Durbin R., 1998) modelo probabilístico basado en las cadenas de Markov, de los más efectivos en modelar una familia de secuencias no alineadas o un motivo común dentro de un conjunto de secuencias no alineadas. Las

cadenas de Markov son procesos estocásticos que satisfacen las siguientes propiedades: 1) cada resultado de una sucesión de pruebas pertenece a un conjunto finito y 2) el resultado de una prueba depende a lo sumo del resultado de la prueba inmediatamente precedente y no de cualquier otro resultado previo; con cada par de estados se establecen las llamadas probabilidades de transición que pueden ordenarse en una matriz estocástica de transición. HMM entrenado, puede ser usado para discriminación o para múltiples alineamientos. Un programa ejemplo que usa este modelo es MEME.

MEME (Multiple Expectation-maximization for Motif Elicitation) (Bailey and Elkan, 1995) es un algoritmo de maximización de la esperanza (expectation maximization) para encontrar patrones en secuencias de proteínas o secuencias de DNA. Es un software que usa técnicas de inteligencia artificial para descubrir motivos de una manera automática. Es un ejemplo que deriva de Hidden Markov models (HMMs). Cuando no se tiene conocimiento previo de las secuencias, MEME es bastante acertado, aunque se obtienen beneficios cuando un conocimiento previo está disponible. Este programa explota el conocimiento previo de los motivos que están presentes en todas las secuencias de entrada, de la longitud de un motivo y de si es un palíndrome, de los patrones esperados en posiciones individuales de las secuencias. En éste como en otros programas el programa se debe correr teniendo a) identificado el largo de la secuencia patrón, b) la frecuencia esperada de los sitios de unión (un sitio por secuencia a menos que se mencione otra cosa) y c) debe incluirse en la búsqueda la cadena apropiada del DNA. Los motivos de salida de este programa pueden ser usados para análisis filogenéticos.

Un método para determinar sitios de unión de proteínas que se unen al DNA, es usando ciclos de unión y amplificación para extraer los sitios de un pool de secuencias al azar y ese método se denomina **SELEX** (de Systematic Evolution of Ligands by Exponential enrichment) (Tuerk and Gold, 1990).

La construcción de un diccionario para genomas como el que se utiliza en el algoritmo **MobyDick** (Bussmaker H.J. *et al.*, 2000) es otro de los métodos usados para identificar motivos de secuencias de DNA que controlan la expresión de los genes. La identificación de palabras está basada en un modelo heurístico probabilístico de mecánica estadística en el que se corta la cadena probabilísticamente en "palabras" y se construye un diccionario con esas palabras. Se elimina la necesidad de referencias externas para calibrar las probabilidades y encontrar los largos óptimos de los motivos automáticamente. En contraposición a los algoritmos que se basan en la frecuencia de oligonucleótidos (como Oligo-analysis o el de Brazma) éste algoritmo es aplicable no solo a fragmentos de genes co-regulados, sino al genoma completo. Hidden Markov models son algoritmos definidos en la misma vía de segmentación de datos biológicos, pero generalmente usan pocos tipos de fragmentos cada uno descrito por muchos parámetros. Con este algoritmo Bussmaker trabaja con muchos segmentos, la mayoría descritos por un solo parámetro. Un problema que presenta este método es que no es posible usarlo con fragmentos largos de 20 letras del alfabeto, para lo que generalmente se usan las matrices de peso, sin embargo hay muchos otros motivos en el genoma que presentan un número igual o menor a 10 pb para los cuales este método está especialmente diseñados para su detección.

En la mayoría de los casos se parte del conocimiento de las uniones, por datos experimentales, de los reguladores al DNA y con la ayuda de estos programas se han proporcionado los patrones de los sitios de unión. Cuando una secuencia "upstream" pertenece a un regulón, suponemos que contiene un patrón común a los hasta ese momento reportado. Los resultados del estudio de genes co-regulados con los diversos métodos, permiten, sin embargo, descubrir el patrón de un sitio de unión nuevo.

La comparación de diferentes métodos de detección de sitios de unión a DNA de reguladores transcripcionales nos permite el análisis de genomas completos de diversos organismos [37-41]. Es una herramienta útil con la que se puede comprobar si los

reguladores y el tipo de regulación tiende a mantenerse en genes que presentan un origen común (homólogos). Las predicciones pueden no ser tan fidedignas o fiables por considerar la variabilidad de los sitios de unión. Debido a que los sistemas reguladores tienden a estar conservados en la evolución, podemos usar la comparación entre especies para incrementar la confiabilidad de las predicciones de los sitios. Se denomina ortología a la relación existente entre dos genes homólogos que provienen del mismo ancestro, y divergen en la especiación. Parálogos son aquellos genes homólogos que provienen de la duplicación génica. Xenólogos son aquellos genes homólogos que provienen de otra especie por transferencia horizontal. Bajo estos parámetros, el asignar un gen a determinado regulón está reforzado no únicamente por el gen mismo, sino por el estudio evolutivo de los genes en otros genomas [42-44].

Motivación

En la literatura, hasta el momento, no disponemos de evaluaciones de algoritmos de detección de sitios de unión al DNA que nos permitan hacer comparaciones de su capacidad de análisis.

Del mismo modo, los criterios para determinar el umbral idóneo para las predicciones son heurísticos. Debido a esta motivación, decidimos enfocar nuestro esfuerzo en encontrar el umbral más adecuado para el problema de descubrir patrones y encontrar dichos sitios en la secuencia genómica de *E. Coli* K12.

Objetivos

a) Evaluar distintos algoritmos para determinar la eficiencia en la detección de sitios de unión de reguladores transcripcionales de genes co-regulados en *Escherichia coli*.

b) Obtener, con base en los datos experimentales y mediante la detección computacional, la estrategia de búsqueda de patrones que nos permita predicciones de genes considerados como co-regulados.

Relevancia

Consideramos de gran importancia la descripción de la metodología para la identificación de sitios de unión de reguladores transcripcionales, no realizada hasta el momento con la evaluación de cada uno de los algoritmos utilizados. De igual forma, es importante la determinación de un umbral adecuado para el descubrimiento de los patrones de sitios que nos permitan la predicción de nuevos genes en cada regulón.

Materiales y métodos

Las secuencias, familias, genes y proteínas de interés se obtuvieron de la base de datos RegulonDB [45, 46], y del genoma de *E. coli* en genbank [47]. Para los primeros enfoques de este trabajo utilizamos la información correspondiente a todas las familias de reguladores conocidos reportados, sin embargo, conforme avanzamos en el análisis pusimos restricciones para limpiar muestras pobres en información. No proseguimos con el estudio de proteínas que regulasen menos de tres regiones conocidas, ni de regiones en las

Reg	ups	sitios	función
AraC	5	15	activador y represor
ArcA	11	20	activador y represor
ArgR	6	12	represor
CRP	65	109	activador y represor
CysB	5	7	activador y represor
CytR	6	12	activador y represor
FIS	25	29	activador y represor
FNR	20	30	activador y represor
FruR	7	8	activador y represor
Fur	4	9	represor
GlpR	4	17	represor
IHF	14	21	activador y represor
LexA	8	9	represor
Lrp	11	22	represor
MalT	4	9	activador
NR-I	3	10	activador y represor
NagC	4	8	activador y represor
NarL	9	20	activador y represor
OmpR	4	14	activador y represor
OxyR	4	4	activador y represor
PhoB	4	7	activador
PurR	14	16	represor
SoxS	3	4	activador
TrpR	5	5	represor
TyrR	8	15	activador y represor

Tabla 2: Número de regiones y de sitios pertenecientes a cada regulón y función de la proteína reguladora.

que el intervalo de análisis no incluyese las coordenadas del sitio unión a DNA del regulador. En el momento en el que se realizó este trabajo, contamos con la información de 112 proteínas reguladoras, sin embargo, no todas ellas contienen la información necesaria para los estudios posteriores. El número total de sitios de unión de reguladores es de 505 y se muestra una tabla en el artículo con toda la información encontrada en RegulonDB (Tabla1). El conjunto de las secuencias utilizadas como de entrenamiento para el estudio de un regulón, corresponde a la colección de sitios positivos conocidos reportados en la base de datos. El resto de las regiones conocidas, reguladas por cualquier otra proteína, formó la colección de los negativos conocidos. Como consecuencia del bajo número de datos para ciertos reguladores, usamos adicionalmente la información de ciertos genes de los que no se tiene identificado el sitio de unión, aunque sabemos se encuentran regulados por la misma proteína.

El tamaño de las secuencias río arriba de los genes analizados puede determinar la eficiencia en la búsqueda de los sitios de unión (como se mencionó con anterioridad). Para el análisis de las regiones en las que se presentan los sitios de unión y teniendo en cuenta los datos de densidad de aparición de los sitios (Fig1), decidimos tomar como material de estudio el DNA de secuencias no codificantes con rangos que comprenden 200 y 400 nucleótidos arriba del ATG y 50 nucleótidos dentro del gen. Los análisis e realizaron tomando información de las regiones río arriba antes mencionadas o bien de los sitios específicos de unión de cada uno de los genes de los regulones para la búsqueda de los patrones. Con esa información, la búsqueda de los apareamientos de los patrones se llevó a cabo en las regiones "upstream" o en los sitios de unión, según lo que se quisiese analizar.

La forma básica de valoración cruzada se realiza cuando los datos analizados son escasos y consiste en dividir en dos partes el juego de datos disponibles: una parte como colección de entrenamiento y la otra como colección de ensayo. Para determinar que tan

buena es una predicción o que tanto error se permite en una predicción se utiliza el método "leave-one-out (LOO) cross-validation".

Leave-one-out (Quitar uno) es un método usado para evitar un posible sesgo introducido por la particular división entre los dos "sets" analizados, de entrenamiento y de ensayo. El número de patrones 'p' usados, se dividirán en conjuntos de ensayo de tamaño 'p-1'. Consiste entonces en dividir el conjunto original y el consiguiente análisis del elemento que en las diferentes particiones se deja fuera. El resultado del método nos evita el sesgo dado por como se están haciendo las divisiones de los datos, reflejando únicamente el dado por el algoritmo en sí y/o por la naturaleza de los datos. Combinando las regiones separadas construimos el "set" total de positivos conocidos (Verdaderos Positivos y Falsos Negativos).

Matriz de peso [21]

En la siguiente figura se muestra un ejemplo de cómo se califica una secuencia

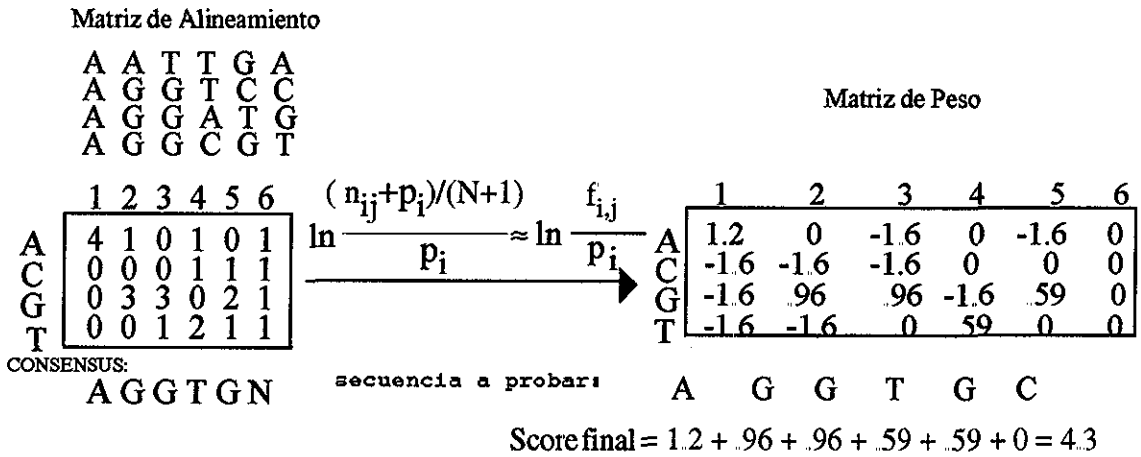


Figura 4: Representación de un alineamiento, de las matrices de frecuencia y de peso generadas y de la calificación de una secuencia blanco.

blanco usando una matriz de peso calculando la calificación o "score" final con la metodología de Contenido Informacional. El "score" final nos indica la similitud entre la

secuencia blanco y la matriz consenso o el alineamiento. En la fórmula $f_{i,j}$ es la frecuencia relativa de la base i en la posición j, y P_i es la probabilidad a priori de las letras i (0.25 para todas las bases en este ejemplo).

Las limitaciones de esta metodología son: a) que asume la contribución independiente de cada base al "score" final y b) que no toma información de contexto.

Dyad-analysis [48]

Las diadas consisten en un par de "palabras" cortas conservadas, separadas por una región de determinado tamaño y contenido variable. Las "palabras", en nuestro caso, se buscaron de un tamaño de 3 nucleótidos y la separación entre ellas fue desde 0 hasta 16 nucleótidos.

La significancia se determina con base en la frecuencia de cada una de las diadas sobre-representadas, respecto a las secuencias no codificantes de todo el genoma [29].

Criterios de evaluación

Cada uno de los umbrales obtenidos de los diferentes análisis nos indican, en el estudio de diadas, el número de "matches" de éstas en la región que estamos estudiando, y para las matrices de peso, el "score" (desde el mínimo) que identifica cada uno de los sitios reportados.

Para cada umbral, analizamos los porcentajes de sensibilidad, falsos positivos, especificidad, exactitud y valores predictivos positivos.

Los verdaderos positivos son aquellos sitios que obtenemos con los algoritmos y que se sabe por la literatura que pertenecen a los genes que se encuentran en el regulón analizado. Los falsos positivos son aquellos que obtenemos con los algoritmos y que

sabemos que pertenecen a cualquier otro regulón de los que tenemos información y que no es el analizado.

Se define sensibilidad como la proporción entre los verdaderos positivos encontrados con el método de detección y todos los positivos (VP + FN). La especificidad es el parámetro que nos indica la limpieza del método, son los verdaderos negativos, y un porcentaje del 100% nos indica una limpieza total de los resultados. Exactitud, la definimos como la proporción de todos los resultados verdaderos. El porcentaje de valores predictivos positivos nos refleja cuantos aciertos podemos determinar con referencia al número de resultados positivos (nos da la relación de verdaderos positivos frente al número total de positivos, verdaderos o falsos).

Evaluación	Fórmula
Sensibilidad	VP / (VP + FN)
Especificidad	VN / (VN + FP)
Exactitud	(VP + VN) / (VP + VN + FP + FN)
Valores predictivos positivos (PPV)	VP / (VP +FP)
"Overall performance" (OP)	(Exactitud + PPV) / 2

Tabla 3: Criterios y fórmulas de evaluación.

Para la evaluación y el manejo de los resultados obtenidos con cada uno de los algoritmos se escribieron programas en lenguaje Perl [50].

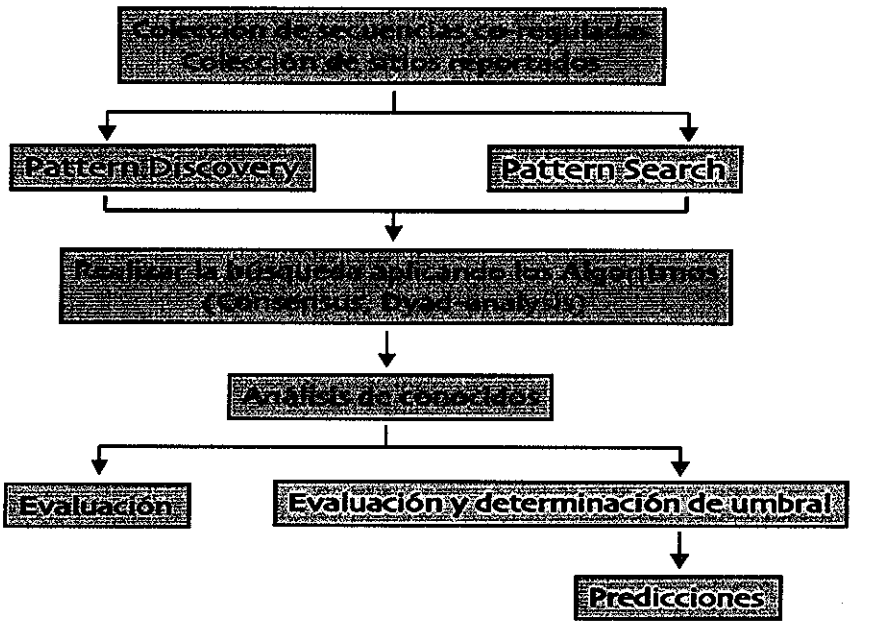
Resultados

La evaluación y la determinación de las estrategias para la detección de los sitios de unión a DNA de los reguladores transcripcionales se realizó con el análisis de dos programas computacionales publicados con ese fin: Dyad-analysis [48] y Consensus [21].

Son básicamente dos las aproximaciones computacionales para realizar los análisis: la **búsqueda de patrón (pattern search)** y el **descubrimiento de patrón (pattern discovery)**. Pattern search consiste en la búsqueda de un patrón de secuencias, cuando contamos con la información experimental de los sitios de unión al DNA de una proteína reguladora. Los algoritmos pueden utilizar, para la obtención de los patrones (diadas o matrices de peso), dicha información de las secuencias específicas correspondientes a los sitios de unión y con esos datos identificar otros genes bajo el control de la misma proteína reguladora. Con datos obtenidos de transcriptomas, por ejemplo, es factible la obtención de una colección de genes aparentemente co-regulados y un investigador puede intentar encontrar otros genes que se encuentren probablemente bajo el control de la misma proteína. La segunda aproximación, pattern discovery, se lleva a cabo en una colección de genes posiblemente co-regulados donde nos enfocamos a encontrar el sitio de unión del regulador desconocido en las regiones río arriba de los genes.

El análisis realizado contempla el uso de un juego de datos de sitios de unión a DNA determinados experimentalmente (como se explica en Materiales y Métodos y en el artículo adjunto). La ejecución de los métodos y los parámetros usados fueron estudiados para las dos aproximaciones descritas arriba, pattern discovery y pattern search. Para la primera aproximación, pattern discovery, se usó la colección completa de todas las regiones "upstream" de 200+50 y 400+50 pb de los genes co-regulados conocidos, con la intención de determinar, por la obtención de un patrón, el sitio de unión de la proteína reguladora en las regiones no codificantes de estos genes.

El esquema general del trabajo es el siguiente:



Con la aplicación del algoritmo desarrollado por Jacques van Helden [29, 48] denominado "Dyad-analysis", se obtuvieron diadas que proponemos pueden reconocer los sitios de unión de proteínas reguladoras en procariotes. Este algoritmo, hasta la realización de esta tesis, no se había utilizado para el estudio de secuencias de organismos procariotes. La identificación de diadas o patrones considerados como sobre-representados es la base del algoritmo. Los genes de los cuales se obtuvo una única diada no fueron utilizados para posteriores análisis. Con el conjunto de datos obtenidos para cada familia, realizamos la búsqueda de las coordenadas del apareamiento de cada diada en las regiones analizadas. Se seleccionaron las opciones de encontrar diadas con todas las simetrías posibles (repetida directa, repetida inversa o asimétrica) y de la búsqueda en ambas cadenas de DNA. Se utilizaron resultados en los que se obtuvo al menos una diada con una significancia igual o superior a 1.0. Con estas características analizamos 19 familias, de 86 regulones de *E. coli,* 11 de las regiones de 200+50 y 14 de las de 400+50. Dyad-analysis encuentra diadas significativas en aproximadamente el 50% de los regulones examinados.

Con la obtención de las coordenadas de los "matches" de cada diada, determinamos si éstas localizan la región y/o el sitio de unión reportado de cada gen para el regulador que caracteriza esa familia, analizando las secuencias de DNA mencionadas: 400+50 y 200+50. Por ejemplo, en RegulonDB se encuentra la información de las secuencias de unión de los genes regulados por el regulador transcripcional LexA, de 8 genes y de 9 sitios, es decir, en uno de los genes (*lexA*) aparecen dos sitios de unión al DNA. La siguiente figura (Fig5) muestra las secuencias de estos sitios de 20 pb (representados con letras mayúsculas) y en todos los casos se añadieron 10 pb a los extremos. Cada una de las secuencias contiene la indicación del gen al que pertenece y un número específico que es el del sitio de unión.

```
LexA|20|lexA|4254646  \ aaatcgccttTTGCTGTATATACTCACAGCataactgtat  \
LexA|20|lexA|4254666  \ tactcacagcATAACTGTATATACACCCAGggggcggaat  \
LexA|20|recA|2821852  \ aaacacttgaTACTGTATGAGCATACAGTAtaattgcttc  \
LexA|20|ssb|4271532  \ tgacacaaatTGACCTGAATGAATATACAGtattggaatg  \
LexA|20|sulA|1020160  \ ggatgtactgTACATCCATACAGTAACTCAcaggggctgg  \
LexA|20|uvrA|4271535  \ atgcattccaATACTGTATATTCATTCAGGtcaatttgtg  \
LexA|20|uvrB|812654  \ ttatggtgatGAACTGTTTTTTTATCCAGTataatttgtt  \
LexA|20|uvrD|3995521  \ aatcagcaaaTCTGTATATATACCCAGCTTtttggcggag  \
LexA|20|rpsU|3208371  \ ttttgaaataAGCTGGCGTTGATGCCAGCGgcaaaccgaa  \
```

Figura 5: Secuencias de unión de los genes del regulón de LexA.

En la siguiente tabla 4 se muestran las diadas obtenidas para el mismo regulador LexA. La búsqueda de los patrones (diadas) se realizó utilizando las secuencias de las regiones río arriba de 200+50 de cada gen y la búsqueda de los apareamientos de éstas se realizó en las mismas regiones. En las dos últimas columnas se presentan los porcentajes de sitios y regiones encontrados por cada una de las diadas. Una región se considera encontrada cuando cualquier diada aparea en al menos un sitio de esa región. Un sitio de pegado lo consideramos positivo cuando al menos un nucleótido de dicho sitio aparea con cualquier diada.

Reg	diadas	ups encontrados	sitios encontrados	% ups	% sitios
LexA	actn{0}gta|tacn{0}agt	5	5	62.50	55.56
LexA	atan{0}cag|ctgn{0}tat	7	7	87.50	77.78
LexA	ctgn{10}cag|ctgn{10}cag	8	9	100.00	100.00
LexA	tacn{1}gta|tacn{1}gta	4	4	50.00	44.44

Tabla 4: Diadas obtenidas con el algoritmo Dyad-analysis para el regulador LexA. "Ups encontrados" representa, con cada diada específica, cuantas regiones se encuentran. "Sitios encontrados" representa la identificación de cuantos sitios. Las dos últimas columnas muestran los porcentajes de los valores anteriores.

El conjunto de las coordenadas de los apareamientos específicos de las diadas en cada proteína nos permite determinar la localización preferente de los oligos sobre-representados.

Los mismos regulones antes mencionados se analizaron adicionalmente con el algoritmo Consensus. Consensus [49 127] determina los alineamientos y construye matrices de frecuencia, con las que se determinan las matrices de peso del tamaño seleccionado (20 pb en nuestro trabajo). Una matriz de peso por posición es generada desde una serie de secuencias alineadas. En estas matrices un valor específico es asignado a cada nucleótido en cada posición dentro de la secuencia. Mediante el uso de la frecuencia de cada base (alfabeto) en las regiones "upstream" de 200+50 y de 400+50 de todos los genes de *E. coli*, se realizó la búsqueda en una sola cadena del DNA para la construcción de las matrices de peso. Aunque también se corrió el programa con el objetivo de encontrar patrones simétricos, ningún resultado fue notablemente mejor a los obtenidos sin el uso de este requerimiento. Con el programa Patser se buscó la localización de cada matriz seleccionada en las secuencias río arriba de los genes de cada regulón o en el genoma completo según sea el análisis. En la tabla se muestran las distancias de las matrices de

peso, realizadas con los sitios del regulador LexA, y las distancias de las matrices (última fila) respecto al inicio del sitio reportado para cada gen.

Reg	gen	score	inicio	fin	sitio	distancia
LexA	uvrB	14.94	-95	-75	uvrB|812655	-19
LexA	sulA	15.64	-43	-23	sulA|1020161	-15
LexA	recA	14.71	-80	-60	recA|2821853	-20
LexA	uvrD	16.07	-77	-57	uvrD|3995521	-18
LexA	lexA	15.68	-48	-28	lexA|4254647	-19
LexA	lexA	17.45	-27	-7	lexA|4254666	-19
LexA	uvrA	16.30	-104	-84	uvrA|4271536	-19
LexA	ssb	12.99	-171	-151	ssb|4271533	-20

Tabla 5: Calificaciones obtenidas tras la aplicación del algoritmo Consensus/Patser, coordenadas de inicio y fin de los sitios definidos para cada gen y la última columna (distancia) representa la menor distancia encontrada entre la localización de la matriz y las coordenadas de los sitios.

PATTERN DISCOVERY

Para el descubrimiento de los patrones, evaluamos el número de veces en las que el sensor puede localizar un sitio de unión conocido. Los datos se obtuvieron de las regiones de 200+50 y 400+50 y no se utilizaron las secuencias de los sitios de unión de cada regulador.

Análisis de los resultados obtenidos con el algoritmo Dyad-analysis

Como se explica en unas líneas anteriores, el programa Dyad-analysis se seleccionó por encontrar parejas de palabras cortas sobre-representadas (diadas) separadas por una región variable. Determinamos que tan exacta es la definición del sitio de unión reportado analizando las coordenadas de los apareamientos de las diadas significativas resultantes. Encontramos que dichas coordenadas aparecen a lo largo de todas las secuencias analizadas, sin embargo, la mayoría de ellas se localizan en o cerca del sitio de unión

conocido (Figura 6). Observamos esa tendencia en la mayoría de las familias analizadas. En la figura 6 se muestra el ejemplo de los datos obtenidos con las 13 secuencias "upstream" de tamaño de 200+50 pb del regulador transcripcional PurR, donde " 0" representa las fronteras del sitio de unión al DNA reportado. Los números negativos indican traslape de las diadas con el sitio.

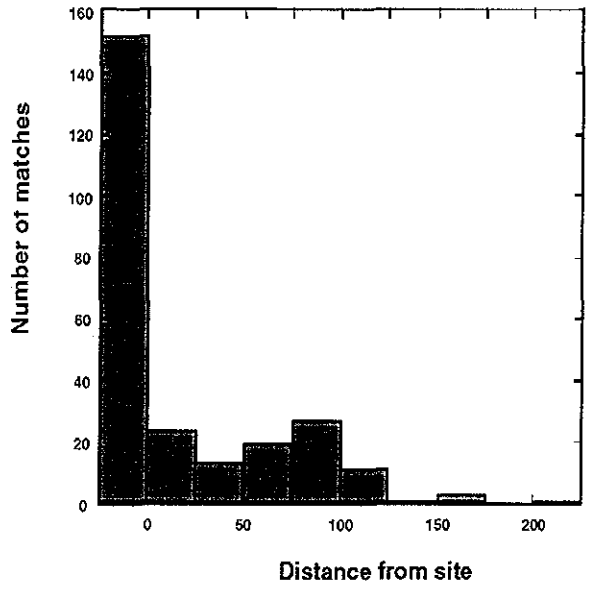


Figura 6: Histograma que refleja las mínimas distancias entre los "matches" de las diadas y las coordenadas de los sitios del regulón.

Con base en este resultado, decidimos buscar en aquellos tramos en los que las diadas se encontrasen contiguas. Mediante el conteo del número de diadas que presentan "matchs" (barrido(sweeping)), base por base dentro de las secuencias analizadas, y obtuvimos lo que denominamos "Regions of Overlapping Matches" (ROMs). Denominamos al método "Dyad-sweeping" por "Barrido de diadas" (Fig 7). La figura 7 representa los ROMs del gen *purR* realizados con el "Dyad-sweeping". Las coordenadas de los dos ROMs más altos corresponden a las coordenadas de los sitios de unión de ese regulador.

Los ROMs con el mayor número de diadas corresponden en un alto porcentaje al sitio de unión conocido en todos los regulones. A partir de la obtención de este resultado decidimos usar, como umbral estratégico de selección, el número más alto de "matches" dentro del ROM para las predicciones posteriores.

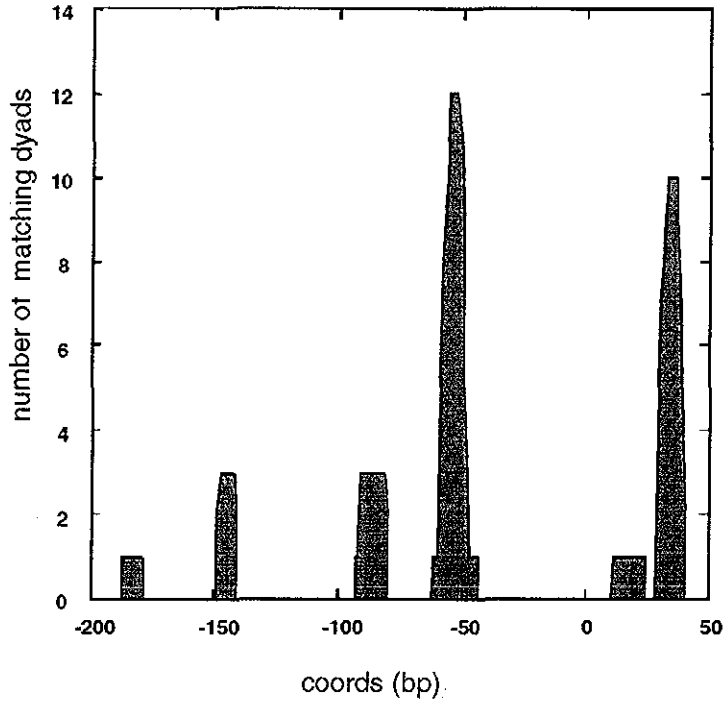


Figura 7: Dyad-Sweeping del gen purR.

Cuando al menos hubiese dos o más diadas, usamos aquellas cuyos "matches" se localizasen en estas regiones. En la tabla 4 del artículo se presenta la eficacia del método al determinar las regiones encontradas. En otras palabras, el juego de diadas seleccionadas únicamente por contribuir al ROM más alto en cada familia, es capaz de recuperar una fracción alta de todos los sitios. Es importante recordar que los ROMs hacen referencia a "matches", por lo que una única diada puede localizarse en varias posiciones -y por lo tanto sitios- en una sola región o familia. Así pues, la selección única de patrones del ROM más alto, no restringe su capacidad para encontrar más de un solo sitio por región.

En la tabla 6 se muestra esta parte del estudio realizado con los genes pertenecientes al regulador PurR para obtener las proporciones de regiones y de sitios encontrados. El

gen	coordenadas_pico		altura_pico	coordenadas_sitio		match_pico
cvpA	-74	-57	8	-73	-57	+
gcvT	-101	-89	3	-120	-104	-
glnB	-82	-73	3	-83	-67	+
glyA	-128	-116	5	-135	-119	+
guaB	-293	-277	2	-290	-274	+
purC	-171	-154	8	-170	-154	+
purE	-90	-72	5	-86	-70	+
purH	-123	-108	6	-124	-108	+
purL	-95	-77	6	-92	-76	+
purM	-78	-64	10	-79	-63	+
purR	-59	-49	5	-59	-43	+
purR	29	39	5	29	45	+
pyrC	-272	-262	5	-69	-53	-
pyrD	-101	-85	7	-101	-85	+

Tabla 6

estudio se llevó a cabo con la información obtenida a partir de las regiones de 400+50 pb de cada gen y la búsqueda se realizó en estas mismas regiones. En ella se observa que la proporción de regiones (de 13) encontradas con el ROM más alto (en la figura se representan con +) es de: 84.62. La proporción de sitios positivos es de: 85.71. Una perspectiva del trabajo sería la identificación de las secuencias correspondientes a los ROMs más altos y que no identifican a los sitios de unión.

El número de diadas que describe el conjunto de sitios conocidos en una familia reguladora es bastante variable. Por ejemplo, si usamos los sitos específicos de unión de diferentes reguladores, de TyrR se obtienen 14 diadas diferentes, mientras que de ArcA se obtienen 65. No pudimos, sin embargo, observar una correlación clara entre el número de diadas por sitio y el número total de sitios en el "training set".

La gráfica siguiente (Fig 8) representa el porcentaje de sitios y de genes encontrados, con los patrones obtenidos con la información proveniente de las regiones de

450 pb de cada uno de los genes de cada regulón y la evaluación se realizó en las mismas secuencias de 450pb. Como establecimos desde el principio, se considera como sitio encontrado aquel cuyo patrón predicho traslapa al menos una base el sitio real conocido (aunque los traslapes más bajos obtenidos cubrieron alrededor del 20% del sitio de unión).

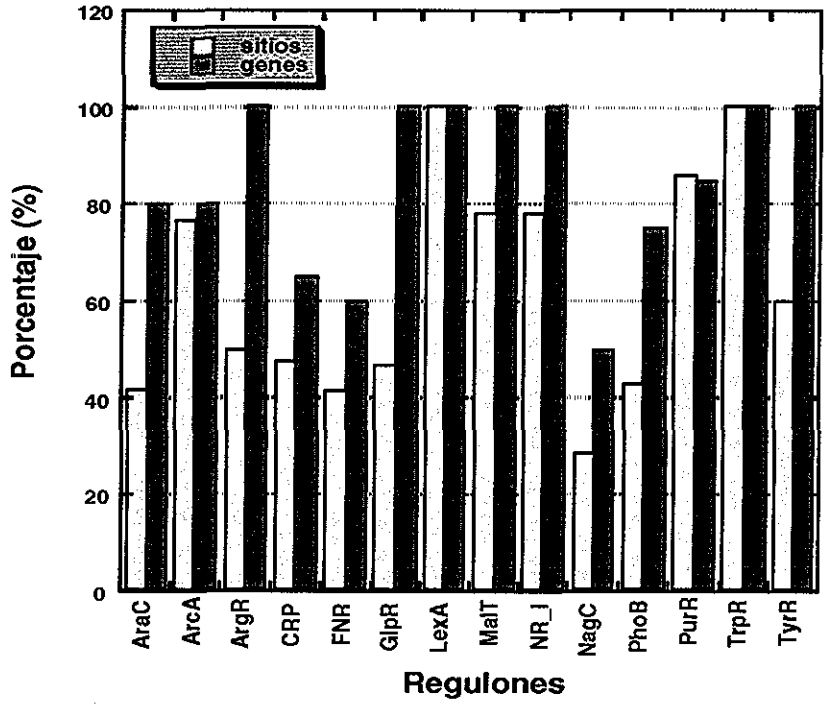


Figura 8: Porcentaje de genes y sitios encontrados con Dyad-analysis/Sweeping.

Análisis de los resultados obtenidos con el algoritmo Consensus

Como se mencionó con anterioridad, Consensus es un programa seleccionado para encontrar y alinear tramos comunes entre secuencias dentro de un conjunto de secuencias dadas. La matriz de peso generada se puede usar para buscar, con el programa compañero "Patser", las coordenadas de los sitios de apareamiento en regiones adicionales. Decidimos generar los alineamientos de tamaño 20 pb (longitud más frecuente de los sitios de unión conocidos), ignorando las cadenas complementarias. Para la búsqueda con "Patser" de los posibles apareamientos en todo el genoma, utilizamos la primera matriz del último ciclo en la que la condición es que todas las regiones analizadas contribuyan con un solo sitio. Esta

matriz es la que presenta el contenido informacional más alto. Seleccionamos aquellas secuencias con el umbral más alto. Como umbrales mínimos para búsquedas y análisis adicionales se seleccionaron los valores más bajos dentro de los genes pertenecientes al regulón.

Gen	Coordenadas sitio	Coordenadas matriz	Match_sitio	Evaluación
cvpA	-73,-57	-74,-54	16	MAX
gcvT	-120,-104	-108,-88	4	MAX
glnB	-83,-67			-
glyA	-135,-119	-136,-116	16	MAX
guaB	-290,-274	-296,-276	14	NO_MAX
purC	-170,-154	-171,-151	16	MAX
purE	-86,-70	-89,-69	16	MAX
purH	-124,-108	-125,-105	16	MAX
purL	-92,-76	-94,-74	16	MAX
purM	-79,-63	-81,-61	16	MAX
purR	-59,-43	-62,-42	16	MAX
purR	29,45	26,46	16	NO_MAX
pyrC	-69,-53	-69,-49	16	MAX
pyrD	-101,-85	-104,-84	16	MAX

Tabla 7

La Tabla 7 muestra los resultados obtenidos con la información proveniente de las regiones de 450 pb de los genes pertenecientes a la familia de PurR y la búsqueda se realizó en estas mismas regiones. Los datos de la tabla reflejan que, salvo dos regiones (NO_MAX), el resto son positivas, es decir, encontradas con el "score" máximo (MAX). En una de las regiones no hay ningún "match", otra de las regiones no se encuentra con el máximo, al igual que uno de los sitios de *purR*. Analizamos los porcentajes de genes y de sitios de todos los reguladores que se identificaron cuando la elección se llevo a cabo mediante el punto que denominamos máximo.

La tabla 8 muestra los datos obtenidos cuando se parte de la información de las secuencias de los sitios de unión reportados. El porcentaje de genes encontrado es muy alto y son muy pocos los genes que se identifican con la misma matriz (con un "score" más

pequeño) o bien que la matriz seleccionada no los identifica (genes de CRP y de FNR).

Reg	No_genes	"Touched by max"	% genes	"Touched by other"	"Not touched"	"W/O match"	% sitios
AraC	5	5	100.00	0	0	0	64.71
ArcA	10	8	80.00	2	0	0	30.00
ArgR	6	6	100.00	0	0	0	89.58
CRP	63	54	85.71	6	3	0	93.95
CysB	5	4	80.00	1	0	0	45.24
CytR	6	6	100.00	0	0	0	42.50
FIS	5	3	60.00	2	0	0	95.00
FNR	20	15	75.00	4	1	0	78.71
FruR	7	7	100.00	0	0	0	100.00
IHF	12	11	91.67	1	0	0	92.31
LexA	8	7	87.50	0	0	1	79.29
Lrp	10	7	70.00	3	0	0	92.50
NarL	9	7	77.78	2	0	0	77.78
PurR	13	12	92.31	1	0	0	96.15
TrpR	5	5	100.00	0	0	0	74.07
TyrR	8	8	100.00	0	0	0	77.84

Tabla 8

Únicamente encontramos un gen que no se puede seleccionar por el algoritmo utilizado, penúltima columna (un gen del regulador LexA). En la última columna se muestran los porcentajes del total de sitios de cada regulador incluyendo más de un sitio en cada gen si así está reportado, y es por eso que en ciertos casos los números son mayores en esta columna que en la de porcentaje de genes identificados.

En la siguiente figura (Fig 9) se muestran los sitios y genes encontrados utilizando la información de las secuencias río arriba de los genes (400 + 50 pb) y la búsqueda se realizó en las mismas secuencias "upstream" de 450 pb. En la figura anterior se utilizaron los datos de las secuencias de los sitios de unión reportados. Recordando, los genes se evalúan como positivos cuando al menos uno de los sitios de cada uno de ellos se ha encontrado. Los sitios, y por eso los porcentajes son menores, se evalúan como positivos cuando el "match" se encuentra al menos una base dentro del sitio reportado. Por ejemplo, en el regulón de ArgR, todos los genes se encuentran (100%), sin embargo, cada uno de ellos presenta dos sitios y la evaluación del descubrimiento es únicamente del 60%.

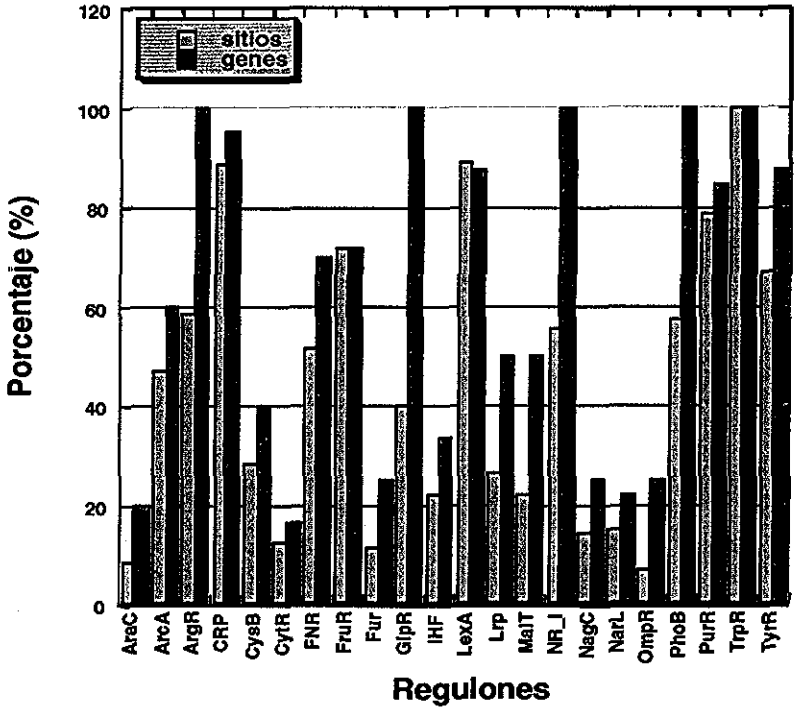


Figura 9: Evaluación del algoritmo Consensus/Patser calculando el porcentaje de genes y de sitios encontrados.

Evaluación de los métodos

La capacidad de "descubrimiento de patrones" de los dos métodos puede estimarse calculando la fracción de sitios encontrados cuando los "training sets" fueron las regiones de 200+50 ó 400+50 pb como se muestra en la Tabla 3 del artículo.

En el análisis que se realizó utilizando Dyad-analysis evaluamos si las diadas filtradas se localizan encima de los sitios conocidos. En el caso de Consensus, evaluamos si las matrices seleccionadas localizan los sitios conocidos. En este sentido obtuvimos la fracción de sitios verdaderos correctamente descubiertos por cada método. Consensus es capaz de identificar (sitio / sitio) un patrón para todas las 35 familias, mientras que Dyad-sweeping identifica el 83% de las familias.

La situación real de descubrimiento de patrones se presenta en los casos en los que se parte de las regiones de 200+50 ó 400+50 y se busca en las secuencias de los sitios reportados. Consensus encuentra el 80% de las familias y Dyad-sweeping el 60% de ellas. Los resultados se muestran en la Tabla 4 del artículo.

Consideramos relevante la comparación de los resultados de los dos métodos, aunque no era la finalidad del trabajo, para encontrar, como perspectiva de trabajo, posibles explicaciones biológicas a las diferencias. Por ejemplo, hay regulones de los que no se obtiene resultados con ninguno de los dos métodos (FIS, OxyR, SoxS). En la mayoría de los casos en los que no se obtienen con Dyad-analysis/Sweeping y se obtienen resultados con Consensus/Patser, éstos últimos son de porcentajes muy bajos. Otra diferencia es la

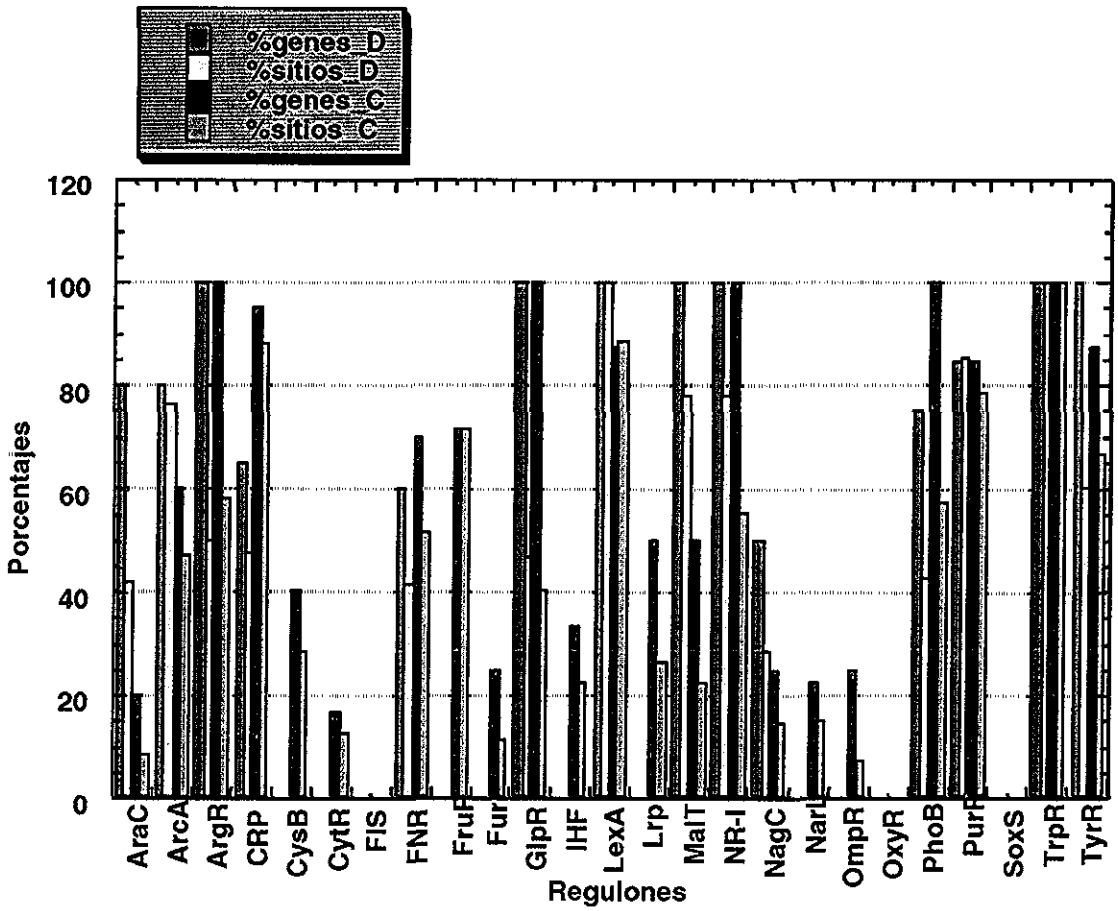


Figura 10: Porcentajes de descubrimiento de patrones con los algoritmos Dyad-analysis/Sweeping ("_D") y con Consensus/Patser ("_C").

existencia de porcentajes muy diferentes dependiendo del algoritmo utilizado.

Con el objetivo de encontrar alguna explicación alas similitudes en la falta de datos buenos, decidimos explorar si se mantenían las diferencias evaluando varias características que consideramos importantes. En la figura 10 se muestran los resultados de los porcentajes, que se indican como %genes o sitios, en el descubrimiento de patrones con los dos algoritmos agrupados. Se consideraron positivos aquellos genes en los que se identificó al menos uno de los sitios pertenecientes a ese gen. Las siguientes figuras siguen la misma distribución.

El **número de regiones** o genes analizados en cada regulón es importante debido a la cantidad de información que se puede obtener de cada familia, el **número de sitios** es de igual forma importante, así como el **promedio de genes/sitios** (figura 11) que cada regulón

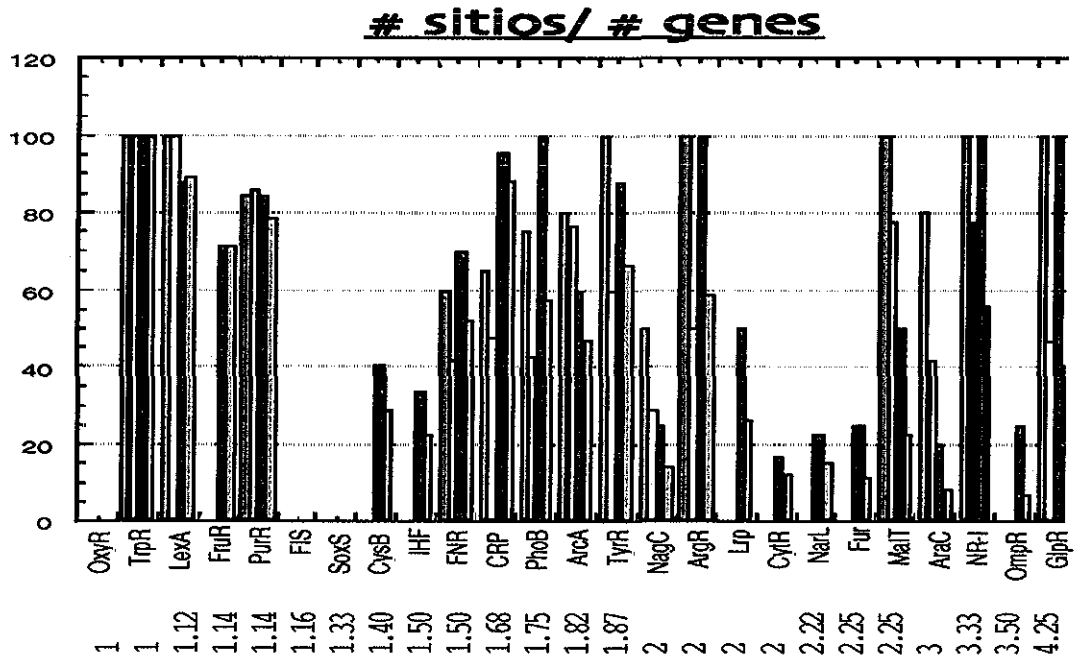


Figura 11: Porcentajes de descubrimiento de genes o sitios, ordenados por la relación entre el número de genes y el número de sitios de cada regulón.

presenta. Ninguno de estos parámetros nos explica el por qué no se obtienen resultados de ciertas familias (SoxS con muy pocos genes y FIS con muchos), o de si los resultados se

obtienen con uno solo de los algoritmos. Analizamos si existe alguna dependencia de los resultados, con la **función** (figura12) que cada proteína reguladora ejerce, es decir, si su forma de actuar es como represor, como activador o como dual y de nuevo no encontramos las explicaciones buscadas.

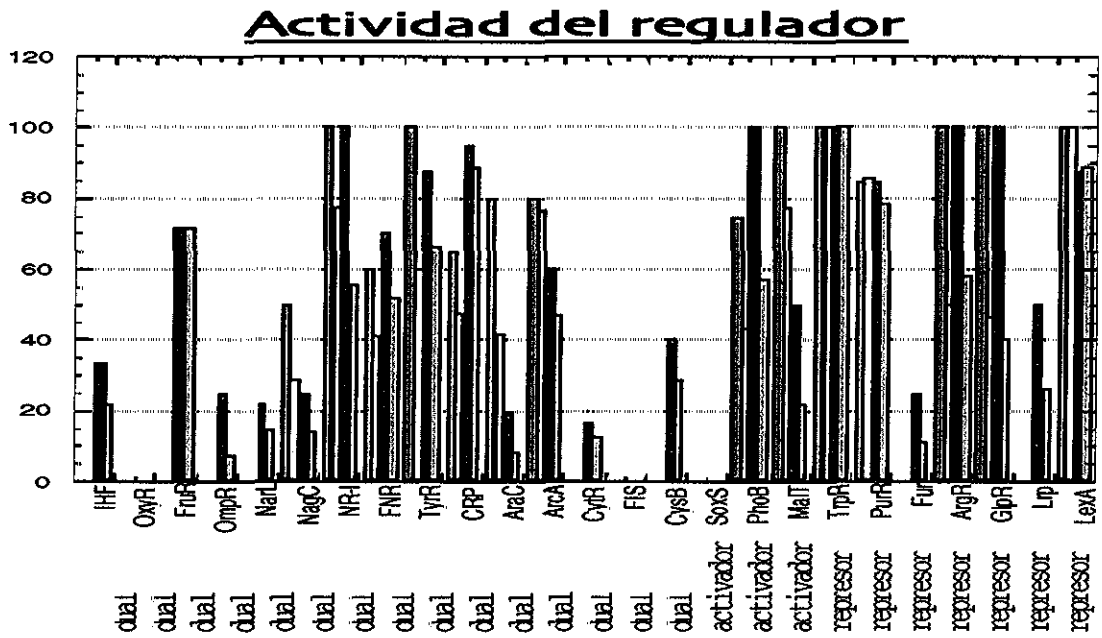


Figura 12: Porcentajes de descubrimiento de genes y sitios en regulones ordenados por el efecto (activador, represor o dual) que ejerce la proteína reguladora.

Analizamos otras informaciones incluidas en RegulónDB como: las evidencias experimentales por las que se reportan los sitios de unión o la **conformación activa** de los reguladores (en forma de dímeros, tetrámeros, etc.) y tampoco resultaron aclaradoras, así como la de **simetría** que no nos refleja diferencias significativas. Por no encontrar con los datos anteriores ninguna respuesta, decidimos explorar la metodología utilizada con cada uno de los algoritmos para encontrar explicaciones y/o posibilidades explicativas. Dyad-analysis/Sweeping se aplicó con ciertas restricciones ya mencionadas con anterioridad. El **número de diadas** tenía que ser mínimo de dos para el consecuente análisis de traslape y la **significancia** de éstas debía ser, en al menos una de las diadas incluidas, >=1. Esas características restringen aquellas familias en las que el número de diadas es "1" ó, aunque

se obtuviese un número mayor, no se llega al umbral de significancia seleccionado. Por ejemplo, de algunos regulones sin resultados no se obtuvieron diadas (CysB, FIS, Fur, IHF, NarL, OmpR, y OxyR) y de otros se obtuvo una única diada sin tener en cuenta la significancia (mayor o menor a "1") (CytR, FruR, Lrp y SoxS).

Para el análisis de los resultados con Consensus/Patser se seleccionaron las matrices del último ciclo que incluyen a todas las regiones pertenecientes a la familia estudiada. Esa matriz de frecuencia presenta el **contenido informacional** más alto y sin embargo tampoco nos permite dar una explicación a las diferencias de descubrimiento entre los regulones. La comparación de todos los valores de el contenido informacional nos reflejó datos muy semejantes (el rango se desplaza de 8.69, como contenido informacional más bajo, a 12.61

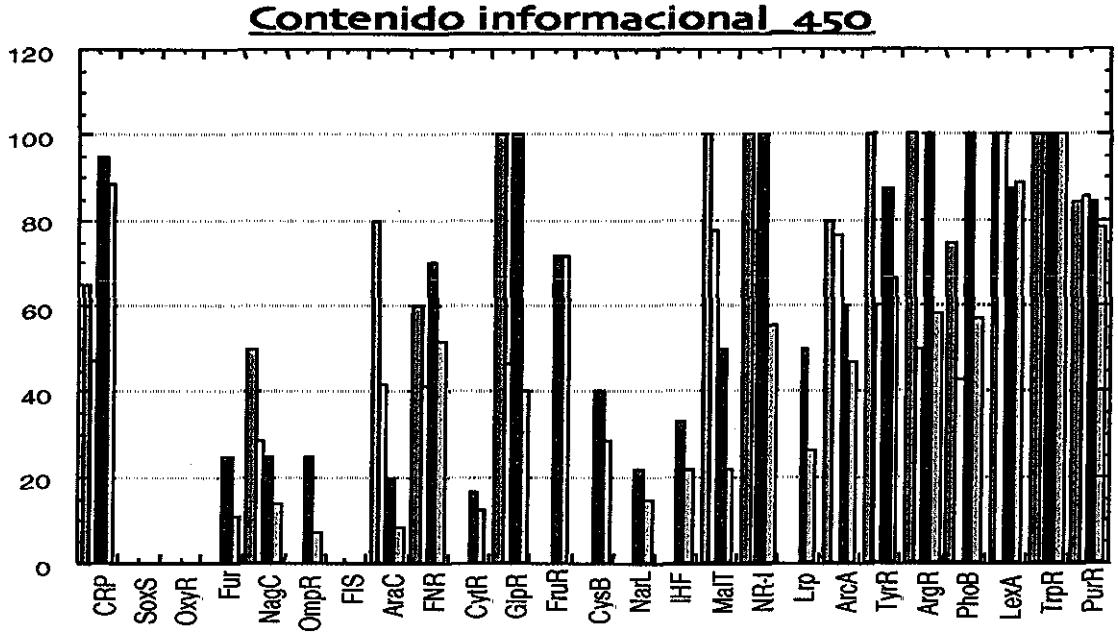


Figura 13: Porcentajes de descubrimiento de genes y sitios en regulones ordenados por el contenido informacional de las matrices obtenidas de las secuencias de 450 pb.

como el más alto). Analizamos adicionalmente el contenido informacional de las matrices obtenidas específicamente de las secuencias de los sitios para comprobar si eran muy diferentes a los valores anteriores y las diferencias de los resultados se debían a esas diferencias en las matrices. En algunas familias el contenido informacional es un valor

igual al obtenido en otro regulón, en otras el valor es menor y en otras mayor sin relación aparente aunque sea una medida de la discriminación entre la unión de una secuencia de DNA funcional y una secuencia de DNA arbitraria.

Por último, decidimos enfocar el análisis a la frecuencia esperada (**expected frequency**) que nos indica la significancia estadística y es el parámetro idóneo para la comparación entre alineamientos de matrices de diferente número de regiones. Todos los regulones que presentan una expected frequency muy alta reflejan una alta probabilidad de

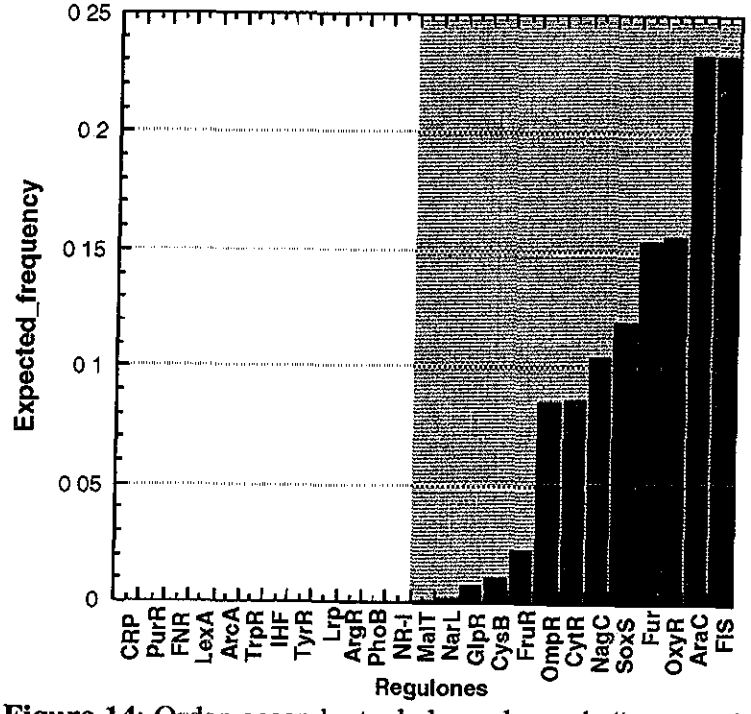


Figura 14: Orden ascendente de los valores de "expected frequency" de las matrices de los regulones indicados

encontrar un alineamiento al azar con el mismo o mayor contenido informacional, dado el peso del alineamiento y el número de secuencias que lo formaron. Como comentaremos más adelante, éste es un dato muy interesante que requiere análisis posteriores. El orden de los valores de la frecuencia esperada nos permite explicar el orden del bajo porcentaje de genes y/o sitios encontrados para algunos regulones con dicho algoritmo.

Los datos reflejan que deben ser las secuencias reguladoras de los genes de estos regulones las que no permiten encontrar, con ninguno de los dos algoritmos, los sitios específicos reportados. Las secuencias río arriba deben tener una composición o un orden

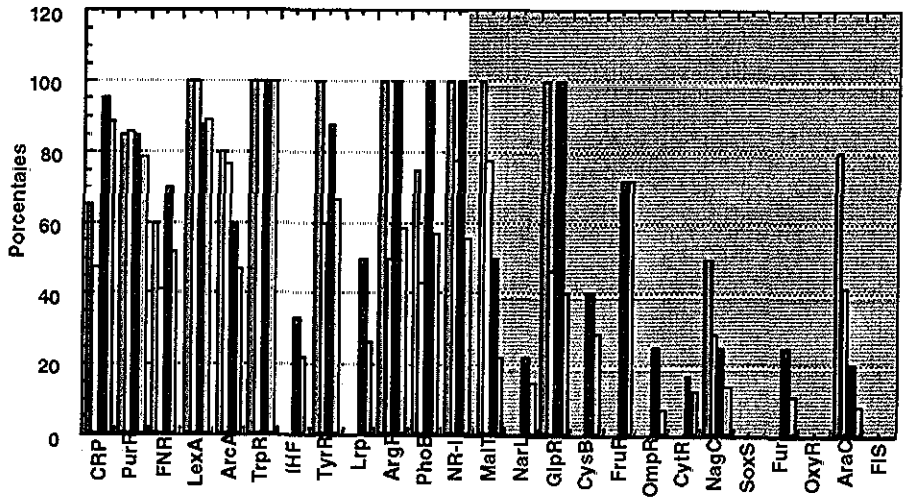


Figura 15: Porcentajes de descubrimiento de genes y sitios de los regulones en orden ascendente de su "expected frequency".

muy parecido a las secuencias que determinan los sitios ya que, cuando las matrices se obtienen de los sitios, el descubrimiento de sitios es posible sin ningún problema. Las matrices obtenidas ¿qué encuentran?

¿Matrices especiales?

Las matrices obtenidas por Consensus de las regiones 400+50 pb nos reflejó un dato muy singular que merece posteriores análisis. Con Patser se detectaron las coordenadas de esas matrices para determinar el número de genes y/o sitios que encontraban. Se observó que éstas corresponden a elementos no identificados hasta el momento y que no concuerdan con las coordenadas de los sitios de unión al DNA del regulador de ese regulón ni a las coordenadas de los sitios de ningún otro regulador. Dichas matrices tampoco identifican las secuencias de los promotores, ni se encuentran a distancias específicas de cualquiera de estos elementos.

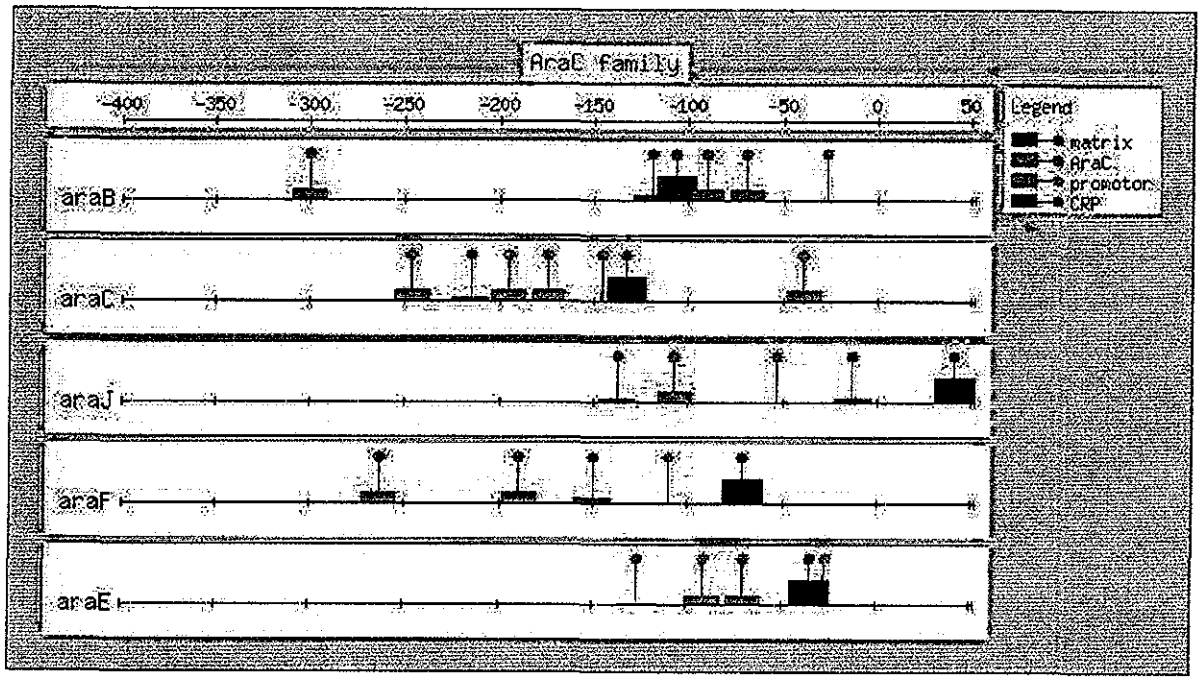


Figura 16: "Feature map" de la familia de AraC donde se representan las matrices obtenidas por Consensus en las condiciones explicadas en el texto, los sitios de AraC, los sitios de CRP y los promotores.

Analizando los valores estadísticos de dichas matrices, comprobamos que presentan un contenido informacional muy semejante a los de las matrices que encuentran altos porcentajes de genes en los que se matchea alguno de los sitios. Un parámetro que refleja una diferencia notable entre estas matrices y las otras es la "expected frequency" y en todos los casos de las mismas características ese valor es muy alto. Estas matrices, sin embargo, son muy específicas de las regiones pertenecientes a la familia, por lo tanto el estudio de estos elementos adicionales nos puede aclarar la especificidad de esas matrices a las regiones "upstream" de cada familia reguladora.

PATTERN SEARCH

La detección de nuevos miembros de los regulones requiere la elección de un umbral óptimo, o un "score" mínimo, para aceptar una secuencia como posible sitio de

unión de los genes río abajo de éstas. La información de los patrones obtenidos analizando los sitios conocidos, nos permite la evaluación de las secuencias con sitios ya reportados para determinar la capacidad de pattern search. La selección de los umbrales mínimos requiere de la evaluación de los resultados obtenidos, tomando diferentes valores estadísticos y leave-one-out (ver Materiales y Métodos) como referencia.

En la figura 3 del artículo se muestran los resultados del análisis correspondiente al regulón del regulador PurR con los datos obtenidos después de usar Dyad-analysis/Sweeping. La flecha señala el que consideramos punto óptimo. El punto óptimo para la predicción lo definimos como el promedio existente entre la exactitud y el porcentaje de valores predictivos positivos. Lo denominamos "overall performance" (OP) de "eficiencia global" y es un buen parámetro para la obtención de genes candidatos a pertenecer a la misma familia.

Analizamos la sensibilidad del método, que nos indica el porcentaje de verdaderos positivos obtenidos con el algoritmo, así como el número de falsos positivos. Determinamos del mismo modo el rango de limpieza obtenida. Los valores predictivos positivos nos señalan el porcentaje de verdaderos positivos entre todos los positivos y determinamos también la exactitud del método. El número correspondiente a la sensibilidad no lo aceptamos cuando se encontró por debajo del 60%. En las figuras 4 y 5 del artículo se muestran los resultados de sensibilidad y falsos positivos, usando Dyad-analysis/sweeping y Consensus/Patser respectivamente, para todos los regulones en el mejor valor de OP.

Con Dyad-analysis/Sweeping el porcentaje promedio de genes reportados en la búsqueda de patrones con el umbral seleccionado es del 38.59% cuando la información se obtiene de los sitios y del 34.51% cuando se obtiene de las regiones "upstream". Como se puede observar, hay ciertos regulones de los que no se obtiene ningún resultado (CytR, IHF, Lrp, etc) y es esa característica la que determina los bajos porcentajes obtenidos. Con

Consensus/Patser el porcentaje promedio es del 54.41% utilizando la información de los sitios y del 48.44% usando la de las regiones. Algunos regulones presentan un porcentaje muy bajo de genes encontrados, cercano al 20%, sin embargo de todas las familias se obtuvieron datos de la evaluación.

Predicciones

Las predicciones se realizaron usando la información proveniente de los sitios de unión al DNA de las regiones río arriba de todos los genes de *E. coli*, a partir de los umbrales óptimos seleccionados. Mediante el uso de datos adicionales, por ejemplo funciones asignadas por Monica Riley, el umbral puede ser modificado siempre y cuando no incluya un número muy alto de falsos positivos.

La tabla muestra un ejemplo de las predicciones (genes) y las funciones asignadas por Riley para cada uno de ellos, realizadas a partir de los sitios reportados para el regulador PurR con el algoritmo Consensus/Patser. Con el análisis de las funciones de cada uno de los genes, consideramos relevante el cálculo de los porcentajes de genes con funciones relacionadas a las de la misma familia y del porcentaje de los que no tenían anotación funcional y se muestran en la siguiente Tabla 7 del artículo.

Como se puede apreciar en la Tabla 9, las predicciones de algunas familias presentan funciones muy relacionadas con las ya conocidas. Lo anterior nos indica que se realizaron predicciones confiables en un alto porcentaje y en algunos casos como en ArgR en CRP o en PurR, los valores de genes con funciones relacionadas son muy altos por lo que consideramos que los genes sin anotación funcional deben encontrarse relacionados.

Gen	Función
aas	building block biosynthesis
aas	protein related
aas	type of regulation
adiY	RNA related
adiY	type of regulation
adiY	genetic unit regulated
aroF	building block biosynthesis
aroG	building block biosynthesis
aroL	building block biosynthesis
aroP	building block biosynthesis
aroP	Electrochemical potential driven transporters
aspC	carbon utilization
aspC	building block biosynthesis
b0753	function not assigned by Riley
b1841	function not assigned by Riley
cyaA	type of regulation
edd	energy metabolism, carbon
folA	building block biosynthesis
folA	central intermediary metabolism
galR	carbon utilization
galR	RNA related
galR	type of regulation
galR	genetic unit regulated
hemC	building block biosynthesis
holE	DNA related
mtr	building block biosynthesis
mtr	Electrochemical potential driven transporters
pth	protein related
tyrP	building block biosynthesis
tyrP	Electrochemical potential driven transporters
yafC	RNA related
yafC	type of regulation
yafD	function not assigned by Riley
ybeB	function not assigned by Riley
ychH	function not assigned by Riley
ydfI	central intermediary metabolism
ydiF	carbon utilization
yecH	function not assigned by Riley
yfiL	function not assigned by Riley
ygjT	function not assigned by Riley

Tabla 9 : Predicciones con funciones (Monica Riley) utilizando Consensus/Patser y el umbral escogido como idóneo del regulón de PurR. Cada una de las filas presenta una función diferente y por ese motivo un mismo gen, con varias funciones, aparece varias veces.

Mediante el uso del algoritmo Dyad-analysis/Sweeping realizamos del mismo modo las predicciones de genes pertenecientes a las todas las familias.

Las predicciones se encuentran en web y se realizaron incluyendo aquellos genes de los que RegulonDB tiene información referente al regulón al que pertenecen, y sin embargo no se tiene información de los sitios específicos a los que se une dicho regulador. A continuación muestro un ejemplo de cómo se presentan dichas predicciones indicando el algoritmo utilizado. Como se observa, para Dyad/Sweeping, en algunos casos las secuencias predichas son muy largas debido a que se seleccionaron los posibles sitios desde el primer "match" hasta el último que incluyese el pico de mayor número de diadas.

Dyad/Sweeping

```
LexA (9)
umuD        +              i: -57;f :11;m :24
CTGCTGG CAAGA ACAGA CTACT GTATA TAAAA ACAGT ATAA CTTCA GGCAG ATTAT TATGT TGTTT ATC
PurR (15)
codB        +              i: -82;f :-64; m:16
AC GAAAA CGATT GCTTT TT
prsA        +              i: -356; f:-34 4;m:2 1
GAAAACG TTTTC G
speA        -              i: -132 ;f :-119 ;m:6
GAAACCG GTTGC GC
```

Figura 17: Ejemplo de predicciones de genes que pertenecen a los regulones indicados pero no se conoce la secuencia del sitio de unión. El número entre paréntesis indica el umbral (número de "matches") para esa familia. Los signos indican si se aceptaron o no esos genes. La siguiente columna hace referencia a las coordenadas de inicio (i) y fin (f) del sitio, junto con el número de "matches" (m) que presentó. Por último se presentan las secuencias predichas.

En la siguiente figura 18 se emplearon los mismos genes que en la figura anterior, sin embargo en ella se utilizó el algoritmo Consensus/Pateser para realizar las predicciones. Las secuencias de las predicciones en este caso son de 20 pb que corresponde al tamaño de las matrices utilizadas.

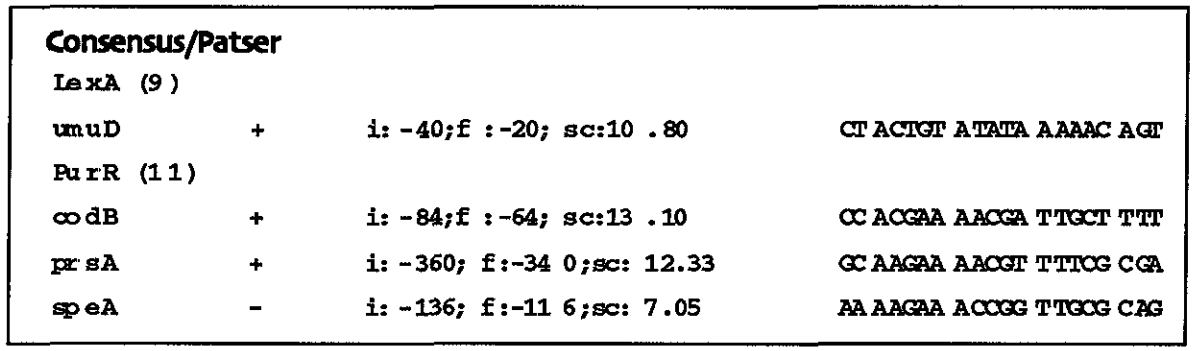


Figura 18: Ejemplo de predicciones realizadas con Consensus/Patser donde se refleja el "score" (s) obtenido y con las mismas características de la figura 17.

Durante la realización de este trabajo se publicó un artículo en el que analizó el regulón del represor LexA en el que presentaron estudios de predicciones computacionales y de análisis experimentales para corroborar sus datos. Consideramos interesante la comparación de nuestros resultados para esa familia con los allá publicados. En el artículo adjunto se muestran dos tablas con estos resultados, Tabla10 con los genes experimentalmente caracterizados pertenecientes al regulón de LexA y Tabla 11 que muestra los genes que no pertenecen al regulón. En nuestros resultados se aprecia que la mayoría de los genes pertenecientes a la familia se clasificaron como positivos con los dos algoritmos. Cuando no fue así, al menos uno de los algoritmos los califica como positivos y en solo un caso ninguno de los dos algoritmos identifica ese gen. Cuando observamos los genes que no pertenecen a la familia, podemos comprobar que son muy pocos los genes en los que la calificación de los dos métodos resulte positiva, sin embargo, no son un número despreciable y como la mayoría de las calificaciones de éstos son menores a las del grupo de negativos, por características adicionales (función, localización, etc) se podría cambiar el umbral seleccionado.

Conclusiones

La determinación del mejor uso de los algoritmos analizados para la detección de sitios o para la búsqueda de genes co-regulados, por la evaluación con mediciones estadísticas, nos refleja y establece la eficacia de los dos algoritmos en el descubrimiento de patrones y el nivel de confianza en las predicciones propuestas.

Proponemos que, debido a los resultados obtenidos, el análisis de las frecuencias de aparición de las diadas obtenidas como resultado del programa Dyad-analysis y con nuestro apoyo (Dyad-Sweeping) es un buen método de detección de los sitios de unión de los reguladores transcripcionales, siempre y cuando se obtengan datos con este algoritmo. Los porcentajes al identificar los genes se desplazan desde el 50 hasta el 100%.

La elección de un umbral adecuado para la admisión de nuevos miembros a una familia de regulación es de gran importancia debido a la necesidad de evitar un alto número de falsos positivos y asegurar el mayor número posible de verdaderos positivos. Un umbral estricto para la admisión de nuevos miembros a un regulón determina que los resultados de verdaderos positivos, por la baja proporción de genes con la que contamos, no se encuentren diluidos en un alto número de falsos positivos.

Identificamos, con matrices de peso de alto contenido informacional, unos elementos desconocidos específicos de varias familias. Estos elementos no corresponden con ninguno de los posibles elementos de los que RegulonDB tiene información, por ejemplo, no corresponden a sitios de unión de otros reguladores conocidos, no corresponden al promotor, ni se encuentran a distancias específicas de estos elementos. Es un dato importante que presenta gran relevancia debido a la especificidad de reconocimiento de los genes del regulón y debido a que no sabemos de que elemento se puede tratar, ni hay información al respecto en la literatura. Por supuesto, merece posteriores análisis.

Concluimos que la combinación de los dos algoritmos de búsqueda es la mejor estrategia. Por ejemplo, si contamos únicamente con evidencias de co-regulación, nosotros sugerimos el uso de Dyad-analysis/sweeping como primer paso para encontrar los sitios. Si Dyad-analysis encuentra diadas significativas, con la metodología del Dyad-sweeping se pueden extraer los posibles sitios de unión. El paso subsiguiente puede ser el uso de estos sitios encontrados para la elaboración de una matriz de peso con Consensus y la búsqueda posterior con Patser de los genes co-regulados.

Perspectivas

Evaluar del mismo modo la eficacia de otros métodos de detección de sitios de unión al DNA de reguladores transcripcionales (Gibbs sampler, AlignACE, MEME, Mobydick....).

Determinar las características de las secuencias de unión de ciertos reguladores de los que no se puede obtener resultados con al menos uno de los dos algoritmos.

Determinar las estrategias idóneas para el análisis de las regiones "upstream" de otros organismos. La comparación de diferentes métodos de detección de sitios de unión a DNA de reguladores transcripcionales nos permite el análisis de genomas completos de diversos organismos. La asignación de un gen a determinado regulón puede estar reforzada no únicamente por el gen mismo, sino por el estudio evolutivo de los genes en otros genomas.

Determinar por otros métodos computacionales la importancia de los elementos específicos no identificados obtenidos con matrices de peso en las regiones analizadas.

Determinar la relevancia del uso de características especiales de las diadas, del algoritmo Dyad-analysis, en la información adicional de las propiedades de simetría del sitio de unión.

Bibliografía

1. Collado-Vides J, Magasanik B,Gralla JD: **Control site location and transcriptional regulation in Escherichia coli.** *Microbiol Rev* 1991, **3**:371-394

2. Collado-Vides J: **A transformational-grammar approach to the study of the regulation of gene expression.** *J Theor Biol* 1989, **4**:403-425

3. Thieffry D, Salgado H, Huerta AM,Collado-Vides J: **Prediction of transcriptional regulatory sites in the complete genome sequence of Escherichia coli K-12.** *Bioinformatics* 1998, **5**:391-400

4. Gralla JD,Collado-Vides J:**Organization and Function of Transcription Regulatory Elements.** In *Cellular and Molecular Biology: Escherichia coli and Salmonella*. Edited by Neidhardt FC, Curtiss III R, Ingraham J, Lin ECC, Low KB, Magasanik B, Reznikoff W, Schaechter M, Umbarger HE and Riley M. Washington, D.C.: American Society for Microbiology, 1996, 1232-1245.

5. Gur-Arie R, Cohen CJ, Eitan Y, Shelef L, Hallerman EM,Kashi Y: **Simple sequence repeats in Escherichia coli: abundance, distribution, composition, and polymorphism.** *Genome Res* 2000, **1**:62-71

6. Bulyk ML, Huang X, Choo Y,Church GM: **Exploring the DNA-binding specificities of zinc fingers with DNA microarrays.** *Proc Natl Acad Sci U S A* 2001, **13**:7158-7163.

7. Gallet X, Charloteaux B, Thomas A,Brasseur R: **A fast method to predict protein interaction sites from sequences.** *J Mol Biol* 2000, **4**:917-926.

8. Kono H,Sarai A: **Structure-based prediction of DNA target sites by regulatory proteins.** *Proteins* 1999, **1:**114-131

9. Kunin V, Chan B, Sitbon E, Lithwick G,Pietrokovski S: **Consistency analysis of similarity between multiple alignments: prediction of protein function and fold structure from analysis of local sequence motifs.** *J Mol Biol* 2001, **3:**939-949.

10. Mandel-Gutfreund Y, Baron A,Margalit H: **A structure-based approach for prediction of protein binding sites in gene upstream regions.** *Pac Symp Biocomput* 2001, 139-150.

11. Perez-Rueda E, Gralla JD,Collado-Vides J: **Genomic position analyses and the transcription machinery.** *J Mol Biol* 1998, **2:**165-170

12. Koonin EV, Aravind L,Kondrashov AS: **The impact of comparative genomics on our understanding of evolution.** *Cell* 2000, **6:**573-576.

13. Rosenblueth DA, Thieffry D, Huerta AM, Salgado H,Collado-Vides J: **Syntactic recognition of regulatory regions in Escherichia coli.** *Comput Appl Biosci* 1996, **5:**415-422

14. Perez-Rueda E,Collado-Vides J: **The repertoire of DNA-binding transcriptional regulators in Escherichia coli K-12.** *Nucleic Acids Res* 2000, **8:**1838-1847

15. Robison K, McGuire AM,Church GM: **A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete Escherichia coli K-12 genome.** *J Mol Biol* 1998, **2:**241-254

16. Stormo GD: **DNA binding sites: representation and discovery.** *Bioinformatics* 2000, **1:**16-23.

17. Bucher P: **Regulatory elements and expression profiles.** *Curr Opin Struct Biol* 1999, **3:**400-407

18. Bussemaker HJ, Li H,Siggia ED: **Regulatory element detection using correlation with expression.** *Nat Genet* 2001, **2:**167-171.

19. Cohen BA, Mitra RD, Hughes JD,Church GM: **A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression [In Process Citation].** *Nat Genet* 2000, **2:**183-186

20. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP,Young RA: **Genome-wide location and function of DNA binding proteins.** *Science* 2000, **5500:**2306-2309.

21. Hertz GZ,Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **7-8:**563-577

22. Waterman MS, Arratia R,Galas DJ: **Pattern recognition in several sequences: consensus and alignment.** *Bull Math Biol* 1984, **4:**515-527.

23. Chen QK, Hertz GZ,Stormo GD: **MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices.** *Computer Applications in the Biosciences* 1995, **5:**563-566

24. Lawrence CE, Altschul SF, Boguski MS, Liu JS, Neuwald AF,Wootton JC: **Detecting subtle sequence signals: a Gibbs sampling strategy for multiple alignment.** *Science* 1993, **5131:**208-214.

25. Hughes JD, Estep PW, Tavazoie S,Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in Saccharomyces cerevisiae.** *J Mol Biol* 2000, **5:**1205-1214

26. McGuire AM, Hughes JD,Church GM: **Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes.** *Genome Res* 2000, **6:**744-757

27. Jonassen I: **Efficient discovery of conserved patterns using a pattern graph.** *Comput Appl Biosci* 1997, **5:**509-522.

28. Brazma A, Jonassen I, Vilo J,Ukkonen E: **Predicting gene regulatory elements in silico on a genomic scale.** *Genome Res* 1998, **11:**1202-1215.

29. van Helden J, Andre B,Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **5:**827-842

30. Bussemaker HJ, Li H,Siggia ED: **Building a dictionary for genomes: identification of presumptive regulatory sites by statistical analysis.** *Proc Natl Acad Sci U S A* 2000, **18:**10096-10100.

31. Bussemaker HJ, Li H,Siggia ED: **Regulatory element detection using a probabilistic segmentation model.** *Proc Int Conf Intell Syst Mol Biol* 2000, 67-74.

32. Collado-Vides J: **Grammatical model of the regulation of gene expression.** *Proc Natl Acad Sci U S A* 1992, **20:**9405-9409

33. Collado-Vides J, Gutierrez-Rios RM,Bel-Enguix G: **Networks of transcriptional regulation encoded in a grammatical model.** *Biosystems* 1998, **1-2:**103-118

34. Crowley EM, Roeder K,Bina M: **A statistical model for locating regulatory regions in genomic DNA.** *J Mol Biol* 1997, **1:**8-14.

35. Crowley EM: **A Bayesian method for finding regulatory segments in DNA.** *Biopolymers* 2001, **2:**165-174.

36. Wolfertstetter F, Frech K, Herrmann G,Werner T: **Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm.** *Comput Appl Biosci* 1996, **1:**71-80.

37. Ouzounis C, Casari G, Sander C, Tamames J,Valencia A: **Computational comparisons of model genomes.** *Trends Biotechnol* 1996, **8:**280-285

38. Gelfand MS: **Recognition of regulatory sites by genomic comparison.** *Res Microbiol* 1999, **9-10:**755-771

39. Gelfand MS, Koonin EV,Mironov AA: **Prediction of transcription regulatory sites in Archaea by a comparative genomic approach.** *Nucleic Acids Res* 2000, **3:**695-705

40. Koonin EV, Makarova KS,Aravind L: **Horizontal gene transfer in prokaryotes: quantification and classification.** *Annu Rev Microbiol* 2001, 709-742.

41. Manson McGuire A,Church GM: **Predicting regulons and their cis-regulatory motifs by comparative genomics.** *Nucleic Acids Res* 2000, **22:**4523-4530.

42. McCue L, Thompson W, Carmack C, Ryan MP, Liu JS, Derbyshire V,Lawrence CE: **Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes.** *Nucleic Acids Res* 2001, **3:**774-782.

43. Mironov AA, Koonin EV, Roytberg MA,Gelfand MS: **Computer analysis of transcription regulatory patterns in completely sequenced bacterial genomes.** *Nucleic Acids Res* 1999, **14:**2981-2989

44. Tan K, Moreno-Hagelsieb G, Collado-Vides J,Stormo GD: **A comparative genomics approach to prediction of new members of regulons.** *Genome Res* 2001, **4:**566-584.

45. Huerta AM, Salgado H, Thieffry D,Collado-Vides J: **RegulonDB: a database on transcriptional regulation in Escherichia coli.** *Nucleic Acids Res* 1998, **1:**55-59

46. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C,Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon.** *Nucleic Acids Res* 2001, **1:**72-74.

47. Blattner FR, Plunkett G, 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B,Shao Y: **The complete genome sequence of Escherichia coli K-12.** *Science* 1997, **5331:**1453-1474

48. van Helden J, Rios AF,Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **8:**1808-1818

49. Hertz GZ, Hartzell GW, 3rd,Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **2:**81-92.

50. Wall L, Schwartz RL : **Programming Perl.** *O'Reilly and Associates, Inc.*, Sebastol CA. ISBN 0-937175-64-1.

Research

Evaluation of thresholds for the detection of binding sites for regulatory proteins in *Escherichia coli* K12 DNA

Esperanza Benítez-Bellón, Gabriel Moreno-Hagelsieb and Julio Collado-Vides

Address: Program of Computational Genomics, CIFN, UNAM, A.P. 565-A, Cuernavaca, Morelos 62100, Mexico.

Correspondence: Gabriel Moreno-Hagelsieb. E-mail: moreno@cifn.unam.mx. Julio Collado-Vides E-mail: collado@cifn.unam.mx

Published: 21 February 2002

Genome Biology 2002, 3(3):research0013.1–0013.16

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/3/research/0013

Received: 16 August 2001
Revised: 20 November 2001
Accepted: 28 January 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Sites in DNA that bind regulatory proteins can be detected computationally in various ways. Pattern discovery methods analyze collections of genes suspected to be co-regulated on the evidence, for example, of clustering of transcriptome data. Pattern searching methods use sequences with known binding sites to find other genes regulated by a given protein. Such computational methods are important strategies in the discovery and elaboration of regulatory networks and can provide the experimental biologist with a precise prediction of a binding site or identify a gene as a member of a set of co-regulated genes (a regulon). As more variations on such methods are published, however, thorough evaluation is necessary, as performance may differ depending on the conditions of use. Detailed evaluation also helps to improve and understand the behavior of the different methods and computational strategies.

Results: We used a collection of 86 regulons from *Escherichia coli* as datasets to evaluate two methods for pattern discovery and pattern searching: dyad analysis/dyad sweeping using the program Dyad-analysis, and multiple alignment using the programs Consensus/Patser. Clearly defined statistical parameters are used to evaluate the two methods in different situations. We placed particular emphasis on minimizing the rate of false positives.

Conclusions: As a general rule, sensors obtained from experimentally reported binding sites in DNA frequently locate true sites as the highest-scoring sequences within a given upstream region, especially using Consensus/Patser. Pattern discovery is still an unsolved problem, although in the cases where Dyad-analysis finds significant dyads (around 50%), these frequently correspond to true binding sites. With more robust methods, regulatory predictions could help identify the function of unknown genes.

refereed research

Background

As a consequence of the availability of whole-genome expression methodologies, regulation of gene expression is at the core of current post-genomic studies [1]. Once a set of genes is clustered on the basis of similar expression profiles, a logical next step is that of searching their upstream regions for potential binding sites for transcriptional regulators. The predicted binding sites in DNA can then be mutated or used to fish out the DNA-binding regulatory protein. Different methods exist for finding binding sites [2-6], with a recent rapid increase in different methods with small variations and improvements [7-9]. However, as the computational

biology community has long been aware, a common limitation of such methods is the high rate of false-positives that they generate as a result of the low degree of conservation of the DNA sequences of binding sites.

This work is a contribution towards a more detailed evaluation of the performance of these methods, with the aim of finding the best selection of thresholds to provide reliable predictions. On the basis of our evaluations, we suggest improved methods to search for novel binding sites that give a much lower rate of false positives. We use information gathered in RegulonDB, a database on regulation of transcription in *Escherichia coli* compiled from the literature [10,11]. The database contains data on regulons - sets of genes in transcription units whose expression is regulated by the same regulatory proteins - with different types of evidence and different levels of description. For instance, at the time of writing, the database contains information on 112 regulatory proteins, but binding sites in DNA are only described for 60 of these. The data for 26 of the regulatory proteins includes information on at least three regulated genes, with at least one binding site per gene (Table 1). The total number of regulatory binding sites listed is 505.

As explained below, we distinguish between pattern discovery and pattern search and evaluate each separately. We evaluate two methodologies. One is Dyad-analysis [12], a program developed to find over-represented small words separated by a given distance. We also describe and evaluate an elaboration of this method that aims to search for probable binding sites using the dyads generated (dyad sweeping). The other method uses Consensus [13], a program that generates optimized ungapped multiple alignments for sets of known or suspected regulatory sequences and builds matrices representing the frequency of each base at each position of the aligned sequences. Its companion program 'Patser' uses the matrices generated to scan for similar new sequences. The evaluations take into account the interest in minimizing the false-positive rate, as even a very small false-positive rate can overshadow true positives because of the small number of genes expected to be part of each regulon (see below).

Description of datasets

As most regulatory sites for DNA-binding proteins are found 200 to 400 base-pairs (bp) upstream of the regulated genes [14], we built two sets of upstream regions. One contained 200 bp of the region upstream of the genes' start sites plus 50 bp downstream (200+50 set); the other contained 400 bp upstream plus 50 bp downstream of the start sites (400+50 set). Repressor sites are located near the promoter site, whereas activators tend to occupy a larger region upstream of the promoter. It is therefore potentially useful to evaluate the performance of the methods with these two different ranges of sequence. Additional information can also influence the decision of the experimentalist to select the length of upstream region to analyze. For instance, some proteins tend to have a single binding site per promoter, which has to be proximal to the promoter (for example LexA), whereas other proteins tend to have several binding sites per upstream region, with some of them farther upstream of the promoter (for example AraC, Lrp and MetJ). Another factor that influences the size of region to analyze is whether the precise site of transcription initiation (the +1 position) is known. When the promoter is known, the search can be limited to 200 bp upstream from the +1 position. If it is not known, then the reference point has to be the start codon and the 400 bp upstream of this are used - which assumes an average of 50 to 100 bp between the promoter and the beginning of the gene.

We used the total set of upstream regions containing at least one reported binding site in RegulonDB as the basic data for evaluation. In each case, upstream regions of genes regulated by the same protein (regulons) were separated from the collection and constituted the 'training sets'. For each set, the remaining upstream regions, known to be regulated by other proteins, are assumed to be the collection of 'known negatives'. Though there is still a risk that the known negatives contain genes that also pertain to the regulon we are contrasting them with, the fact that they have been the subject of experimental work allows us to think that this risk is minute.

Because of the small amount of data for each protein, we could not leave out a set of known positives to evaluate the rate of true positives, except in the case of the regulatory protein CRP. For those families having at least five upstream regions we were able to apply a 'leave one out' procedure as described below. We also have information, in some cases, on genes regulated by a given protein in the regulons analyzed, but with no reported binding site. The upstream regions of these genes were used to search for binding sites and provide further evaluation. A more detailed analysis was performed for LexA, comparing our predictions with a recent report in the literature [15].

Levels of analysis

Depending on the information available, there are basically two computational approaches to predicting binding sites for transcription initiation factors in DNA. In the best cases, there is information on experimentally determined examples of binding sites for a given regulatory protein. In such cases, the search programs can be trained using the sequences corresponding to the binding sites, and the information obtained (dyads, weight matrices) can then be used to find similar sequences, and thus other genes that might be under the control of the same regulatory protein. This is pattern searching

On the other hand, a common scenario at present is that a set of apparently co-regulated genes is identified from transcriptome experiments. In this case, a program would be

Table 1

Summary of the datasets in RegulonDB

Regulatory protein	Number of binding sites	Site size (bp)	Regions with sites	Regions without sites
Ada	2	28	2	2
AlpA	-	-	-	1
AppY	-	-	-	2
AraC	15	17	5	1
ArcA	20	61	11	9
ArgR	12	16	6	1
AsnC	-	-	-	2
AtoC	-	-	-	1
BetI	2	21	2	-
BirA	2	40	2	-
CRP	109	19	65	15
CadC	-	-	-	1
CbI	1	45	1	-
CsgD	-	-	-	2
CspA	3	5	1	1
CynR	2	60	2	-
CysB	7	42	5	1
CytR	12	40	6	2
DeoR	7	16	2	1
DnaA	8	9	2	-
DsdC	-	-	-	3
EbgR	-	-	-	1
EnvY	-	-	-	2
ExuR	-	-	-	1
FIS	29	16	25	-
FNR	30	22	20	4
FadR	6	17	4	-
FarR	2	21	1	-
FecI	1	7	1	1
FhlA	-	-	-	3
FruR	8	14	7	4
FucR	-	-	-	2
Fur	9	19	4	6
GalR	4	17	1	1
GalS	2	16	1	1
GatR	-	-	-	1
GcvA	4	29	2	-
GlpR	17	20	4	1
GntR	-	-	-	5
GutM	-	-	-	1
GutR	-	-	-	1
Hns	-	-	-	5
IHF	21	13	14	12
IclR	1	34	1	-
IlvY	4	26	2	-
KdpE	1	12	1	-
LacI	3	20	1	-
LeuO	-	-	-	1
LexA	9	20	8	1
Lrp	22	12	11	3
LysR	1	13	1	2
MalI	4	12	2	-
MalT	9	10	4	-
MarA	-	-	-	5
MarR	-	-	-	1
MelR	6	18	2	1
MetJ	5	8	2	1
MetR	3	24	2	1
Mlc	2	26	1	-
MtlR	-	-	-	1
NR_I	10	15	3	-
NadR	-	-	-	2

Table 1 *(continued)*

Regulatory protein	Number of binding sites	Site size (bp)	Regions with sites	Regions without sites
NagC	8	26	4	-
NarL	20	19	9	3
NhaR	-	-	-	1
OmpR	14	10	4	3
OxyR	4	45	4	-
PdhR	1	21	1	-
PhoB	7	17	4	1
PurR	16	16	14	3
RbsR	-	-	-	1
RcsB	2	25	2	-
RhaR	3	20	1	-
RhaS	3	17	2	-
Rob	-	-	-	4
SdiA	-	-	-	1
SoxR	2	19	2	-
SoxS	4	18	3	2
TdcA	1	15	1	-
TdcR	1	12	1	-
TorR	4	10	1	-
TrpR	5	27	5	-
TyrR	15	22	8	-
UhpA	1	39	1	-
XapR	2	13	1	-
XylR	4	16	2	-

RegulonDB contains information for the 86 regulons shown in this table. Of these, only 60 have at least three known binding sites for their corresponding regulatory protein. The second column indicates the total number of known sites, which are distributed in upstream regions (fourth column). The last column indicates the number of upstream regions for which there is experimental evidence suggesting regulation, but no direct proof of binding of the regulator to the upstream site is yet available. For instance, there are 12 known sites for ArgR located in only six regions (with two sites per region), plus one region for a different gene for which there is evidence of regulation by ArgR.

trained with a collection of upstream regions from these genes with the goal of identifying probable shared regulatory sites. This is the problem of pattern discovery. If the data come from transcriptome experiments, the collection of co-regulated genes might not be complete. Because of the noise inherent to such experiments, and/or to the limitations of clustering algorithms, a researcher might wish to try to find other genes likely to be under the control of the same protein. However, other genes regulated by the same protein might display a different pattern of expression as a result of complications such as regulation by more than one regulatory protein.

On the basis of these considerations, the analyses we present contemplate the use of experimentally determined binding sites as training sets to study pattern search, and the use of upstream regions of co-regulated genes to study pattern discovery. More precisely, we use the set of binding sites in DNA for each regulatory protein reported in RegulonDB to try to find additional genes in the genome with similar sites. We also use the data on known co-regulated genes to try to find the binding site within the genes' upstream regions.

As training sets, we ran the dyad or matrix search programs on the sequences of known regulatory binding sites and on upstream regions of 200+50 and 400+50 bp from genes regulated by a given regulatory protein. Families corresponding to a given regulatory protein were evaluated only if there were at least three sequences in the corresponding training set (40 in the collection of binding sites; 26 in the 200+50 and the 400+50 datasets). Subsequently, the dyads and matrices were evaluated against the complete collections of 200+50 and 400+50 upstream regions. This gives a total of 3 x 2 = 6 evaluations for each regulon analyzed. The evaluations included regions 200+50 or 400+50 only if there was at least one reported binding site within that range; thus, the total set of 200+50 regions contained 172 sequences, and the 400+50 set contained 189.

Dyad analysis

We used the Dyad-analysis program [12] to find dyads within each training set. The options used were to find dyads of 3 bp long separated by distances of 0 to 16 bp, with any kind of dyad (direct repeat, inverted repeat, asymmetric), searching in both DNA strands [12,16]. Further analyses were limited to the training sets where the program found at least one dyad with a significance equal to or above 1.0 (see [12] for a detailed description of significance). This left 19 families from the binding-sites training sets, 11 from the 200+50 regions, and 14 from the 400+50 regions (the program Dyad-analysis did not find any dyad in about 75% of the rejected families, and found just one in most of the rest of them).

The program Consensus was run to obtain alignments and matrices 20 bp long - the most frequent size among binding sites for regulatory proteins. To assign match scores, we used an 'alphabet' based on the frequency of each base at upstream regions of 200+50 and 400+50 of all genes in *E. coli*. The search was done in a single strand. Although we also ran the program to find symmetric patterns, no clear improvement was observed. In the Results section, we first present results of pattern discovery, then concentrate on the selection of the best thresholds, analyzing their performance on the basis of the evaluation criteria described above. Finally, we present some specific predictions.

Results

Pattern discovery

Pattern discovery starts with a collection of co-regulated genes for which no binding sites are yet known. To evaluate the methodology, we counted the number of times a sensor can locate a known binding site in a collection of 200+50 or 400+50 regions.

The Dyad-analysis program is designed to find over-represented small words. Over-represented words would be expected to occur at the binding sites, and thus the first step was to determine if the resulting dyads match the binding sites. We found that there are significant dyads all along the sequences analyzed, with most of them matching at or near the known binding sites. Figure 1 shows, using the PurR family, that most dyads were found at distances very close to or overlapping the true binding sites. We observed the same tendency for all families. We thus decided to search for stretches of contiguous matches, which we call 'regions of overlapping matches' (ROMs), in the upstream sequences being analyzed by counting (sweeping), base by base, the number of matching dyads. As seen in Figure 2, the ROMs with the highest number of matching dyads overlap the true known binding sites in the DNA. This result motivated us to use the highest number of matches within a ROM as the score. We call this method dyad sweeping.

As the highest-scoring ROMs frequently overlap reported binding sites (Figure 2, Table 2), we decided to keep, for subsequent analyses, the dyads found within the highest-scoring ROMs of each upstream region, as long as the ROM contained at least two dyads. In Table 3 it can be seen that, except in a few of the regulons, the fraction of regions with known binding sites found is quite high. In other words, the set of dyads that result after keeping only those that contribute to the highest ROM in each family is able to recover a large fraction of all the known binding sites in the family. It is important to keep in mind that a given dyad can match several positions - and therefore sites - in a single region or family. Thus, selecting only those dyads appearing in the

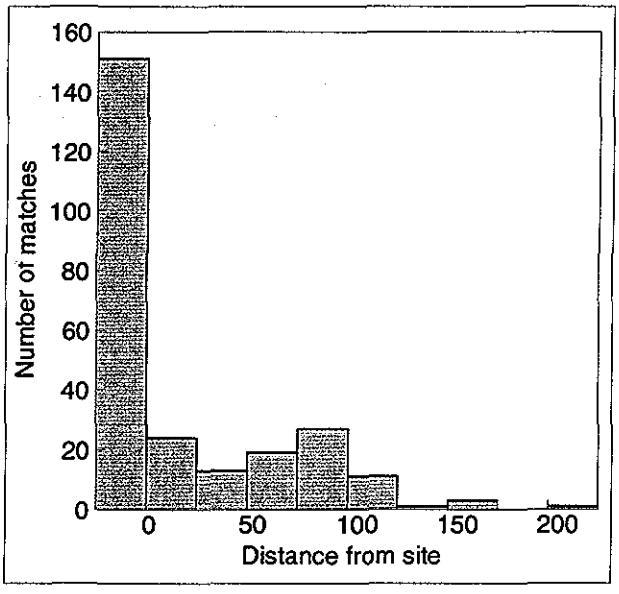


Figure 1
Position of dyads found by the Dyad-analysis program in relation to the binding sites in DNA for the whole PurR family. The graph shows the distances between all the dyads found in relation to the known binding sites of the PurR regulon. Distances below zero mean that the dyad is overlapping the binding site.

Table 2

Pattern discovery using ROMs (regions of overlapping matches) with maximal score to find binding sites in DNA

Regulatory protein	Number of genes in the regulon*	Touched by max†	Percent touched by max‡	Touched by other§	Total percent touched¶	Not touched**	Without dyads††
ArgR	6	2	33.33	4	100.00	0	0
CRP	54	42	77.78	8	92.59	4	0
FNR	17	10	58.82	0	58.82	2	5
GlpR	4	3	75.00	1	100.00	0	0
IlvY	2	2	100.00	0	100.00	0	0
LexA	8	8	100.00	0	100.00	0	0
MalI	2	2	100.00	0	100.00	0	0
MalT	4	4	100.00	0	100.00	0	0
MelR	2	2	100.00	0	100.00	0	0
NR_I	3	3	100.00	0	100.00	0	0
NarL	6	3	50.00	0	50.00	2	1
PhoB	4	4	100.00	0	100.00	0	0
PurR	12	11	91.67	0	91.67	1	0
TorR	1	1	100.00	0	100.00	0	0
TrpR	5	4	80.00	1	100.00	0	0
TyrR	8	7	87.50	0	87.50	1	0

*The total number of genes in the regulon with a known binding site (in the 400+50 upstream regions). †The number of regions where a ROM (region of overlapping matches) with the highest number of matches (max ROM) touches a known binding site ‡This value expressed as a percentage. §Number of regions where either a ROM or dyad touches a known binding site, but the max ROM does not ¶The percentage of all upstream regions in which any ROM touches a binding site. **Number of regions with dyads, but no match between known binding sites and ROMs. ††Number of regions with no dyads at all

highest peak does not restrict their ability to find more than one site per region.

The number of dyads that describe the set of known binding sites in a given regulatory family is quite variable. For instance, if we use the known binding sites as training sets, the TyrR family involves 14 different dyads whereas ArcA has 65. There is no clear correlation between the number of dyads per site and the total number of sites in the training set for any given family, or any other property of the regulatory site, such as its size.

Sequence alignment

Consensus is a program designed to find and align shared stretches of sequence among a given set of sequences. The searching method based on the results of Consensus is already available [13]; the weight matrix generated can be used to search, with the companion program Patser, for sites in other upstream regions. The search using Patser was made using the first matrix (highest informational content) obtained in the final cycle of Consensus. This cycle requires all regions to contribute at least one sequence to the matrix. Using Patser, we searched for the highest-scoring sequence in each region in the training set. The lowest value among these results was set as the minimal score and a second search was performed with this threshold in order to find new sites above this limit within each upstream region in *E. coli* for further searches and analyses.

The capacity for pattern discovery of the two methods can be estimated by calculating the fraction of binding sites found when the training sets were the 200+50 or 400+50 bp regions, as shown in Table 4. A site was considered found when the predicted pattern overlaps 20% of the binding site.

We also show the results of using the sequences of the binding sites with 10 bp extensions on each side as training sets, so we could distinguish between pattern discovery and pattern abstraction or identification. In the case of Dyad-analysis/sweeping we evaluated whether the filtered dyads overlap the set of true sites. In the case of Consensus/Patser we evaluated whether the set of sites selected by Consensus/Patser overlaps the set of known sites. Consensus/Patser is able to abstract a pattern for each of the 25 families, whereas Dyad-analysis/sweeping can only do it for 19 of the families. In 11 of these 19 families Consensus/Patser finds more sites, in two families Dyad-analysis/sweeping finds more sites, and in the remaining six both methods perform equally well.

The real pattern discovery situation is that of the 450/sites cases (see legend to Table 5 for definition), where Consensus generates matrices for 24 of the families and Dyad-analysis finds significant dyads for 11 of them. Dyad sweeping finds on average more than 70% of the binding sites (when Dyad-analysis obtains significant dyads) as compared to around 60% with Patser. Note that using shorter regions to search for DNA binding sites (200+50), improves the performance of both methods by about 5-7%.

Once Table 4 was generated, we estimated the fraction of upstream regions recovered (Table 3). A region is considered found when at least one site in that region is found. Therefore,

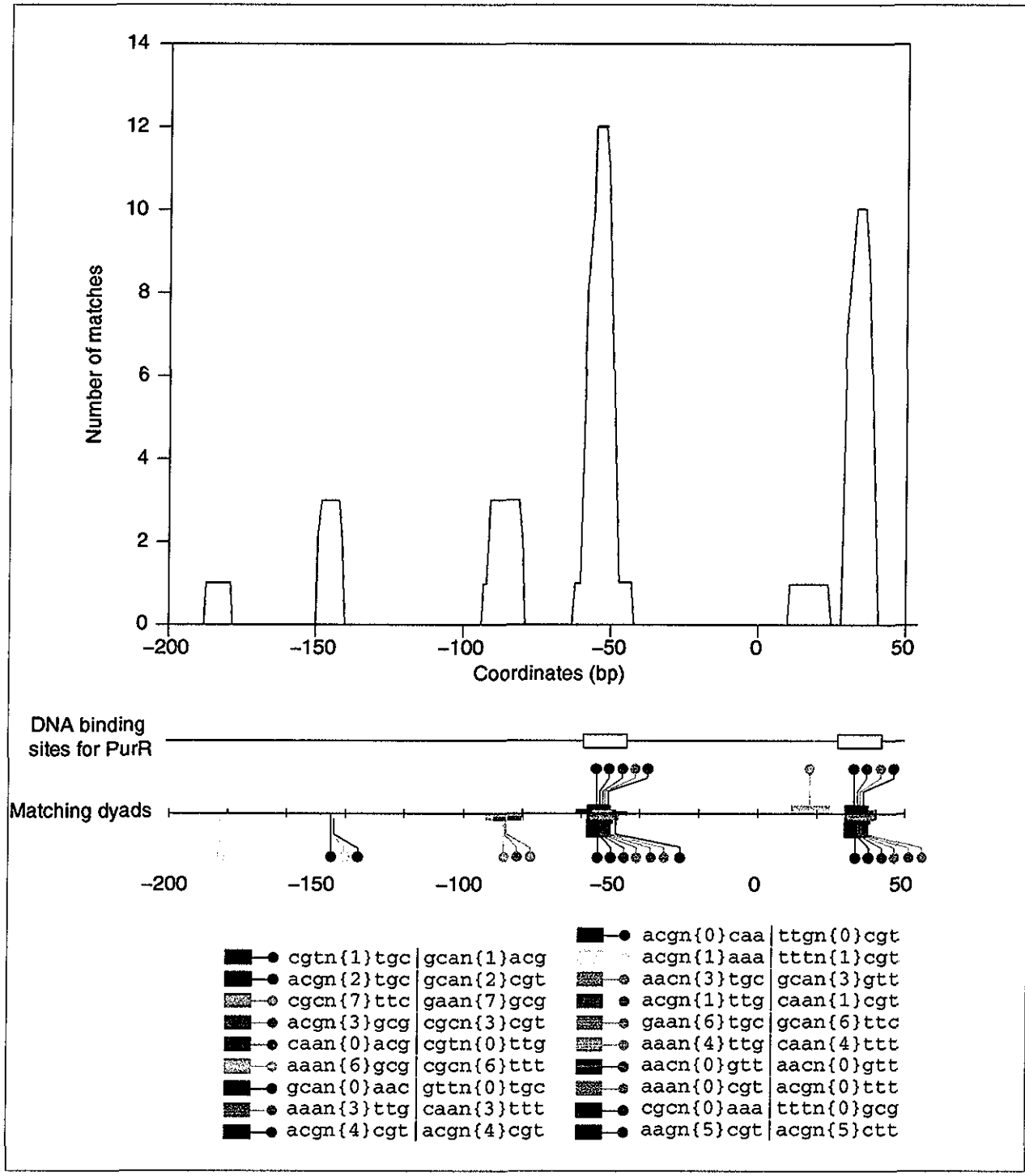


Figure 2
Dyad sweeping along the upstream region of the *purR* gene. Contiguous regions of overlapping matching dyads (ROMs) frequently overlap with the known binding sites. This example shows results after finding significant dyads in the 200+50 regions of the PurR regulon, and finding the ROMs within the same regions. The two ROMs with the highest peaks completely overlap with the two reported regulatory binding sites in this region (sites lie at positions -59 to -43 and at 29 to 45). The coordinates here are relative to the annotated first coding nucleotide of the gene. The known binding sites are illustrated as boxes below the figure. The lower line shows the different dyads coded in different colors. It can be seen, for instance, that blue dyads occur only in the two true binding sites.

Table 3

Pattern discovery at the level of upstream regions

	Consensus/Patser			Dyad-analysis/sweeping		
Regulon	Sites/sites	250/sites	450/sites	Sites/sites	250/sites	450/sites
AraC	100.00	20.00	20.00	100.00	-	80.00
ArcA	80.00	60.00	60.00	90.00	-	80.00
ArgR	100.00	100.00	100.00	100.00	33.33	100.00
CRP	95.24	93.65	95.24	90.48	66.67	65.08
CysB	100.00	60.00	40.00			
CytR	100.00	16.67	16.67			
FIS	60.00	60.00	-			
FNR	90.00	75.00	70.00	80.00	60.00	60.00
FadR	100.00	75.00	-			
FruR	100.00	14.29	71.43	71.43	-	-
Fur	100.00	100.00	25.00	75.00	-	-
GlpR	100.00	100.00	100.00	100.00	75.00	100.00
IHF	100.00	75.00	33.33	58.33	-	-
LexA	100.00	87.50	87.50	100.00	100.00	100.00
Lrp	80.00	60.00	50.00	50.00	-	-
MalT	100.00	50.00	50.00	100.00	100.00	100.00
NR_I	100.00	100.00	100.00	100.00	100.00	100.00
NagC	75.00	25.00	25.00	-	-	50.00
NarL	100.00	55.56	22.22	77.78	-	-
OmpR	100.00	-	25.00	100.00	-	-
OxyR	75.00	25.00	-			
PhoB	100.00	100.00	100.00	100.00	100.00	75.00
PurR	92.31	84.62	84.62	100.00	84.62	84.62
TrpR	100.00	100.00	100.00	100.00	100.00	100.00
TyrR	100.00	100.00	87.50	87.50	87.50	100.00
Average	94.14	68.22	61.98	88.45	82.47	85.34

For each family, we show the results with Dyad-analysis/sweeping and with Consensus/Patser. The data shown are obtained using different training sets - the 200+50 and 400+50 regions (250 and 450) and a comparison with training sets of known binding sites (sites) as a reference standard. Results are given as the number of regions where at least one binding site was found divided by the total number of regions, and expressed as percentages. Note that only the dyads extracted from the max ROMs within each region are used here. In each column heading, the first word refers to the training set and the second refers to the regions where the patterns were searched. For instance, columns headed 450/sites show the results of pattern discovery when Consensus or Dyad-analysis has as input the 450+50 bp regions, and the sensor is evaluated with the files of known sites. We counted only those regions containing known binding sites within the range covered (that is, if a known binding site is present more than 200 bp upstream of the gene start site, the corresponding 200+50 region is not counted). Averages count only the lines where the programs provided a result. Dashes mean that either there was no binding site within the region or the programs failed to provide a matrix (Consensus) or significant dyads (Dyad-analysis). A region is considered found if at least one of its binding sites is matched.

Table 4

Pattern discovery at the level of binding sites

	Consensus/Patser			Dyad-analysis/sweeping		
Regulon	Sites/sites	250/sites	450/sites	Sites/sites	250/sites	450/sites
AraC	100.00	8.33	8.33	58.33	-	41.67
ArcA	76.47	76.47	47.06	76.47	-	76.47
ArgR	75.00	58.33	58.33	100.00	16.67	50.00
CRP	90.53	87.37	88.42	66.32	46.32	47.37
CysB	100.00	42.86	28.57	-	-	-
CytR	100.00	12.50	12.50	-	-	-
FIS	55.56	66.67	-	-	-	-
FNR	93.10	65.52	51.72	58.62	41.38	41.38
FadR	83.33	50.00	-	-	-	-
FruR	100.00	14.29	71.43	71.43	-	-
Fur	100.00	66.67	11.11	77.78	-	-
GlpR	60.00	40.00	40.00	80.00	33.33	46.67
IHF	72.22	66.67	22.22	50.00	-	-
LexA	100.00	88.89	88.89	100.00	100.00	100.00
Lrp	78.95	47.37	26.32	36.84	-	-
MalT	66.67	22.22	22.22	55.56	44.44	77.78
NR_I	77.78	55.56	55.56	77.78	77.78	77.78
NagC	57.14	14.29	14.29	-	-	28.57
NarL	75.00	35.00	15.00	55.00	-	-
OmpR	71.43	-	7.14	42.86	-	-
OxyR	75.00	25.00	-	-	-	-
PhoB	71.43	57.14	57.14	71.43	71.43	42.86
PurR	92.86	85.71	78.57	92.86	78.57	85.71
TrpR	100.00	100.00	100.00	100.00	100.00	100.00
TyrR	86.67	73.33	66.67	53.33	46.67	60.00
Average	88.59	60.36	53.07	75.19	70.76	65.64

For each family, we show the results of applying Dyad-analysis/sweeping and Consensus/Patser to the problem of discovering binding sites. The results contain pattern discovery data similar to those in Table 3, but this time counting the number of binding sites found per total number of sites. Again, only dyads extracted from max ROMs are used. The names of columns are as described in Table 3.

the results differ from those in Table 4 because of the occurrence of multiple sites in some upstream regions. A clear case of this is the ArgR regulon, where each of the six regions has two binding sites. The methods detect from 17% to 58% of the sites, but find from 33 to 100% of the regions.

Detection of new members of regulons by pattern matching

Detection of new members of regulons requires the selection of an optimal threshold to accept a sequence as a predicted binding site, and the genes downstream of such sequences as new members of the regulon family. The selection of the best threshold requires the evaluation of the following parameters: sensitivity (rate of true positives), specificity (rate of true negatives), accuracy (overall rate of true results), and, very important in this case, the positive predictive value (rate of true positives among the total number of positives, true and false). Definitions of these terms are given in the legend to Table 5.

We used a leave one out (LOO) procedure to evaluate the true-positive and false-negative rates with families containing at least five reported genes with binding sites. The LOO method consists of leaving one gene at a time out of the training set; then, with the matrix or dyads built with the remaining sites, a search is made for a probable binding site within the upstream region of the gene that was left out. We combined the results of the left-out regions to build the total set of known positives for evaluation of true positives and false negatives. The evaluation of true negatives and false positives was carried out using the whole set of known positives as training sets and all the

Table 5

Definitions of parameters used in evaluating the predictions

Evaluation	Formula
Sensitivity	TP/(TP + FN)
Specificity	TN/(TN + FP)
Accuracy	(TP + TN)/(TP + TN + FP + FN)
Positive predictive value (PPV)	TP/(TP +FP)
Overall performance (OP)	(Accuracy + PPV)/2

FN, false negative; FP, false positive; TN, true negatives; TP, true positives.

remaining regions, known to be regulated by any other protein, as known negatives. Instead of calculating an average of the scores, and defining the threshold on the basis of standard deviations, we scanned the scores scale from the minimum score obtained in the collection of positives, to the maximum one, calculating the evaluation parameters noted above at each point of the scale. There is no point in searching at lower scores as there is no effect on sensitivity at such values.

In Figure 3 we show the results of the analyses of the PurR regulon using dyad sweeping. Here, the minimum number of matches evaluated was one. Note that, as the dataset of known negatives exceeds that of known positives, high accuracy coexists with a large number of true negatives. Nevertheless, at the threshold of 10 matches, despite a very low false-positive value (less than 10%), and a very high accuracy (approximately 95%) and sensitivity (90%), the positive predictive value (PPV) shows that the total true positives in the whole 'predicted' set is about 60%. This is a very important issue. As most regulatory proteins regulate just a few genes in comparison with the whole set of genes in a given organism, such a difference means that false positives might dilute reliable predictions even at very low false-positive rates. The PPV alone would leave results with very little recovery of true binding sites. Therefore, calculating an optimal point for prediction requires the use of a balanced evaluation criterion. After examining several graphs, we noticed that the average between accuracy and PPV (which we call the overall performance or OP) would be a good criterion. This makes

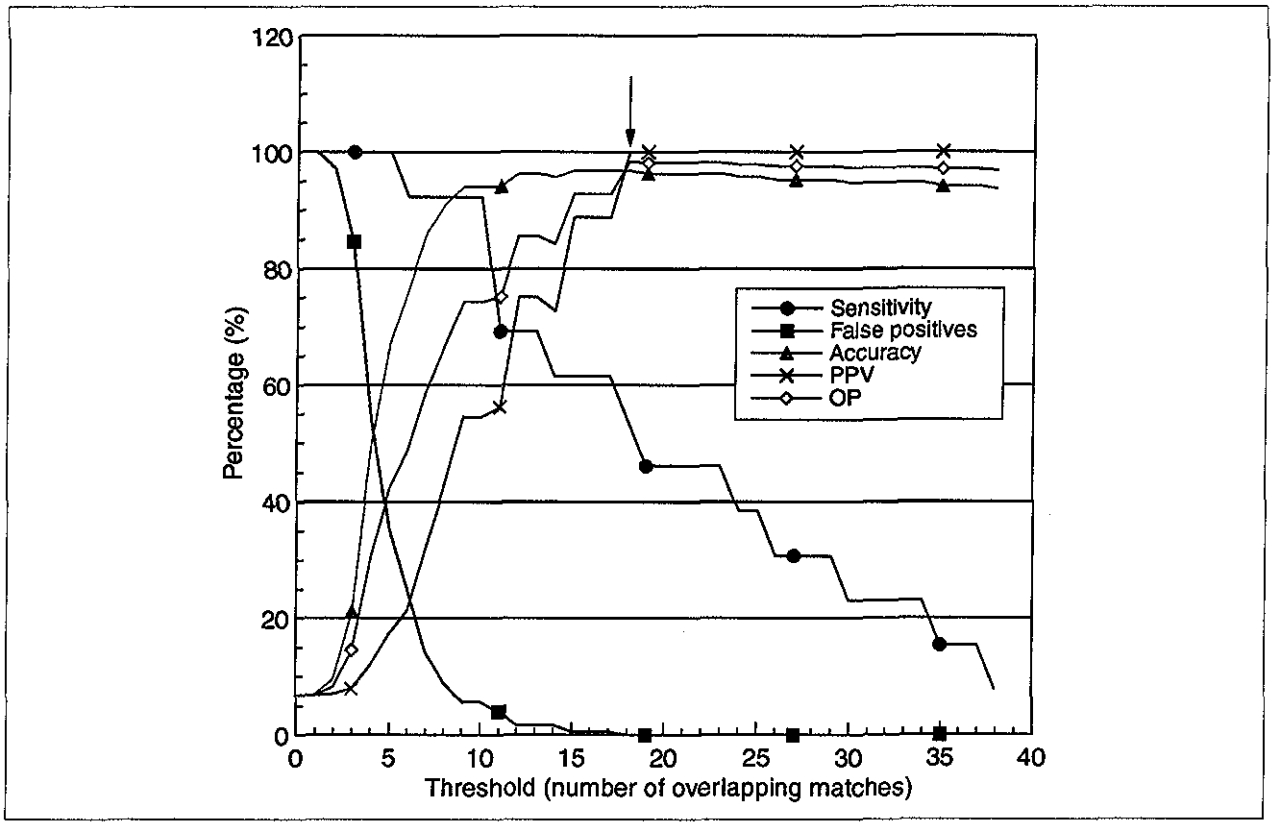


Figure 3
Evaluation of predictive capabilities as a function of the threshold using dyad sweeping. Different thresholds, defined as number of overlapping matches, were evaluated for all regulons. This graph shows the case of the PurR regulon when the dyads are obtained from the known binding sites and the evaluation is carried out on the 400+50 regions. The only dyads used in the search were those found at the ROMs with the highest value per region in the PurR regulon. The statistical parameters (see Table 5) are plotted as percentages instead of fractions. The arrow indicates the point of maximum overall performance (OP) (see text).

sense, as OP represents a trade-off between those two statistical measures. Other criteria, such as the product of accuracy and PPV, might be used instead, but OP worked well for our purposes. In a few cases, the point of highest OP leaves a very small sensitivity value (around 50% in PurR, for instance). If the sensitivity value was less than 60%, we used the last point where the sensitivity was above 60%. In Figure 4 we show the results of sensitivity and false-positive rate for all regulons at their best OP value using dyad sweeping.

The use of weight matrices derived from Consensus (with Patser) is not illustrated, as the selection of the best threshold is the same as in dyad sweeping. In Figure 5 we show the results of sensitivity and false-positive rates of each regulon at the best overall performance point of each regulon analyzed using Patser.

In Table 6 we give the fraction of sites found per family in regions of 400+50 bp when starting from different training sets using the threshold chosen as described above. Dyad-detection/sweeping still performs better at finding the sites within an upstream region, while Consensus/Patser trained with binding sites finds the sites at an average of almost 77%.

An interesting finding here was that, when trained with all the upstream 400+50 sequences, Consensus finds an alignment and matrix that clearly discriminates between the sequences used in the training set, or regulon, from any other upstream sequence in *E. coli*. However, in some families, the matrix matches at sites different from the experimentally determined DNA binding site of the regulon under analysis (Figure 6), and such sites do not correspond to any known site, motif or region annotated in RegulonDB in the upstream sequence. We also verified that they do not match conserved regions in between pairs of sites. It will be indeed interesting to find out if these sequences have any biological meaning.

Predictions

Once the optimal threshold was obtained, we proceeded to predict other members of each regulon using the complete collection of upstream regions (200+50 and 400+50) of the *E. coli* genome [17]. In order to further evaluate the predictions obtained, we used the recent annotations of cellular functions assigned by Monica Riley and her group to known *E. coli* genes [18]. About 30% of the genes in *E. coli* have no function assigned, and each gene or gene product can be

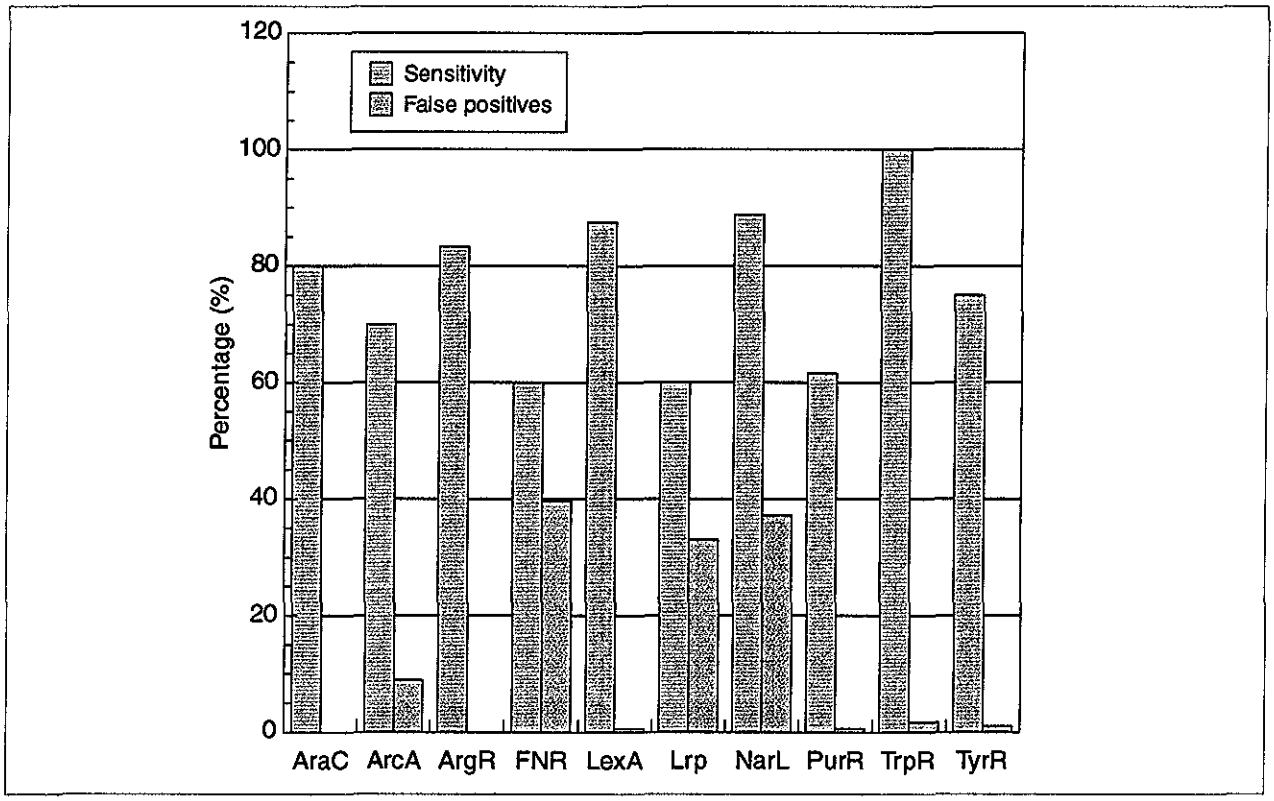


Figure 4
Performance of Dyad-analysis/dyad sweeping at the best threshold defined for each family. Sensitivity and false-positive rate (expressed as percentages) at the highest overall performance for each regulon are shown, using the binding sites as training sets, and the 400+50 regions as evaluation sets. We do not show the regulons where the methods did not provide significant results.

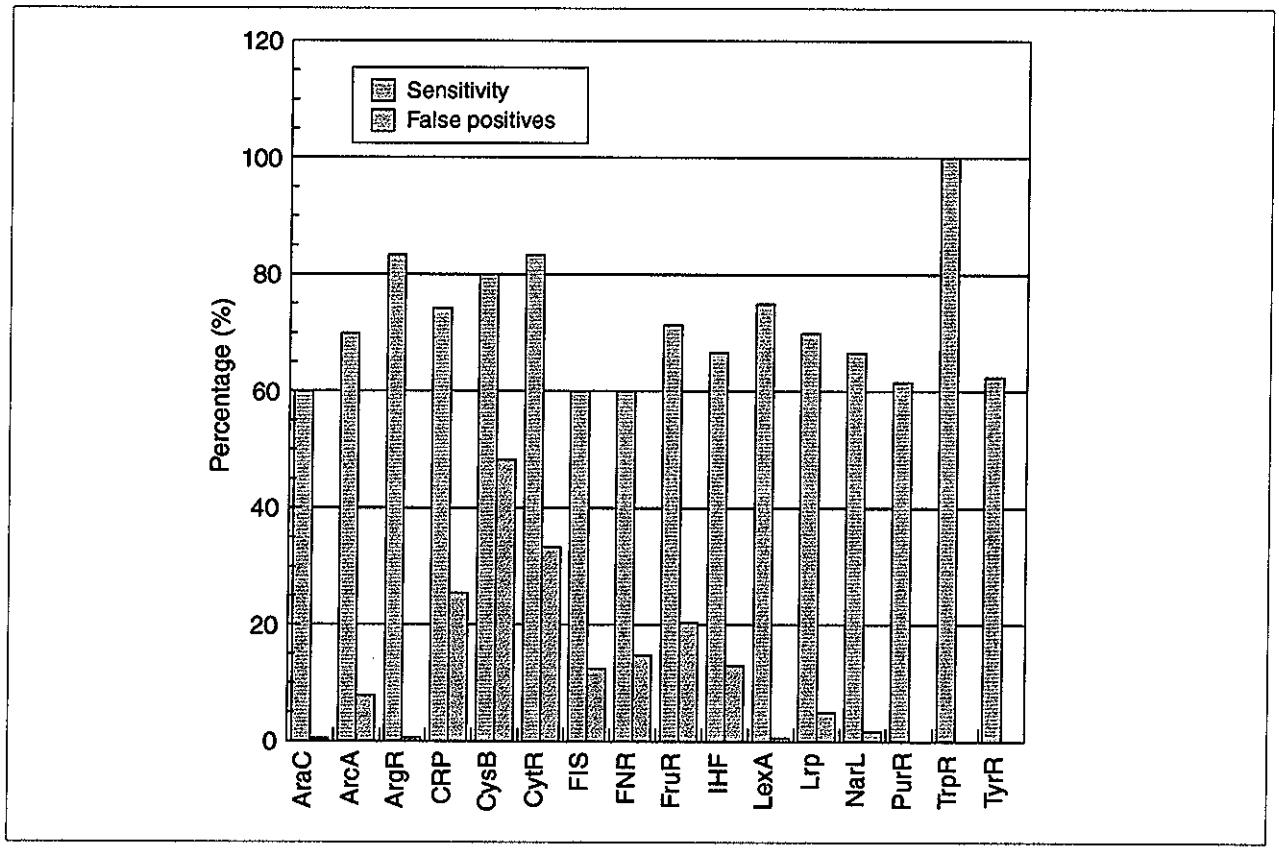


Figure 5
Performance of Consensus/Patser at the best threshold defined for each family. Sensitivity, and false-positive rate (expressed as percentages) at the highest overall performance for each regulon are shown, using the binding sites as training sets and the 400+50 regions as evaluation sets. We do not show the families where the methods did not provide significant results.

assigned to more than a single cellular role. In Table 7 we show the consistency between the functional annotations of genes experimentally demonstrated to belong to each regulon as compared with the functional annotations of the set of predicted genes. In the cases of predictions of high confidence (for example, ArgR, CRP and PurR - all with correspondences above 90%), a putative function can be reliably assigned to genes of unknown function. For instance, in the case of the PurR family, the genes without functional annotations might be assigned to macromolecule (DNA/RNA) biosynthesis. This is an example of functional gene prediction based on analysis of its regulatory elements. Annotations like 'active transporter' would require other kinds of evidence (see Additional data files). Functional annotations might be quite helpful in cleaning up wrong predictions, or adjusting the proposed thresholds, although limited by the genomic coverage of the functional assignments.

RegulonDB contains information on a few genes belonging to some of the regulons studied but with no mapped binding site for the relevant regulatory protein. As further evaluation, we show the results of dyad sweeping and Patser, trained with the known binding sites of each regulon, for all of these genes (Tables 8,9). In the tables we indicate whether the gene would be included in the corresponding predictions, the highest scoring ROM (dyad sweeping, Table 8) or pattern match (Patser, Table 9) found in the 400+50 region of the gene, and the actual sequence suggested as part of the possible binding site. Some genes would be rejected as predictions, but the small amount of data makes it impossible to appropriately evaluate this problem. A researcher might choose to use a different, perhaps lower, threshold if the intention is to find every gene for a given regulon experimentally, and such a decision would depend on how many confirmatory experiments it is possible to perform (an example is shown in the next section). Lower thresholds can also be used if the intention is to confirm new members suggested by other data, like clustering of a gene or genes with known members of a regulon. The latter case is exemplified by the results with those regulon members lacking a mapped binding site. Most contain ROMs or patterns scoring above the minimal score obtained for a known member of the

Table 6

Binding sites remaining at best threshold

	Consensus/Patser		Dyad-analysis/dyad sweeping	
Regulon	Sites/sites	450/sites	Sites/sites	450/sites
AraC	60.00	20.00	80.00	100.00
ArcA	80.00	90.00	90.00	90.00
ArgR	83.33	100.00	100.00	66.67
CRP	80.95	63.49	90.48	95.24
CysB	100.00	80.00	-	-
CytR	100.00	16.67	-	-
FIS	40.00	-	-	-
FNR	65.00	70.00	50.00	65.00
FruR	100.00	85.71	-	-
Fur	-	-	75.00	-
GlpR	-	-	100.00	100.00
IHF	83.33	33.33	-	-
LexA	87.50	87.50	87.50	100.00
Lrp	40.00	60.00	-	-
MalT	-	-	75.00	100.00
NR_1	-	-	100.00	100.00
NagC	-	-	-	50.00
NarL	77.78	22.22	-	-
PhoB	-	-	75.00	75.00
PurR	69.23	84.62	92.31	84.62
TrpR	100.00	100.00	100.00	-
TyrR	62.5	87.50	75.00	37.50
Average	76.85	66.74	62.65	76.00

For each family, we show the results with Dyad-analysis/sweeping and with Consensus/Patser accepting a match only if its score exceeds the defined best threshold. This threshold corresponds to the highest overall performance, see Figures 3 and 4. Otherwise the results are treated as explained in Table 4. The column names are as described in Tables 3 and 4

regulon (no search is performed below this lower limit), often just below our suggested threshold. Thus, if there is additional evidence that a gene belongs to a given regulon, the ROMs found can be proposed as the putative binding sites.

Comparison of results with recently examined members of the LexA regulon

A recent attempt has been made by Fernandez De Henestrosa *et al.* to locate all the members of the LexA regulon by a combined strategy that included prediction of probable binding sites and experimental confirmation [15]. Their predictions were based on similarity to known sites. Experimental confirmation showed that only 10 of the 49 predicted new members responded to LexA. The authors also give a table of previously found members of the LexA regulon, which includes a few genes not annotated in RegulonDB. We could analyze only five of their experimentally confirmed genes and 31 of their wrong predictions (predictions they later found experimentally not to be regulated by LexA) because of the lack of updating of the *E. coli* K12 genome annotations. In Table 10 we present our results for the genes noted as previously determined members of the LexA regulon in [15], plus the five new members found by this study. In Table 11 we show our results with their wrong predictions. Using dyad sweeping, we find 20 out of the 23 confirmed members of the LexA regulon, whereas we would reject 20 out of their 31 wrong predictions. With Consensus, we detect 18 of the 23 confirmed members of the regulon, while rejecting 19 of their wrong predictions.

Table 7

Correspondence between the functional annotations of predicted genes and of known genes.

Regulon	Consensus/Patser		Dyad-analysis/sweeping	
	Percent with related function	Percent without functional annotation	Percent with related function	Percent without functional annotation
AraC	73.33 (66.67)	37.50 (42.86)	53.75 (51.32)	36.51 (37.70)
ArcA	72.61 (70.75)	45.86 (47.50)	74.07 (70.83)	40.88 (43.75)
ArgR	100.00 (100.00)	18.18 (33.33)	56.25 (36.36)	51.52 (60.71)
CRP	93.90 (93.38)	40.87 (43.63)	-	-
CysB	56.78 (56.56)	43.03 (43.16)	-	-
CytR	53.90 (53.58)	38.56 (38.72)	-	-
FIS	45.52 (45.11)	41.23 (41.28)	-	-
FNR	82.85 (81.99)	45.85 (47.06)	83.77 (83.43)	41.43 (41.92)
FruR	51.96 (51.29)	36.57 (36.89)	-	-
IHF	80.95 (80.33)	46.34 (47.14)	-	-
LexA	69.70 (61.54)	42.11 (48.00)	65.79 (58.06)	38.71 (43.64)
Lrp	82.61 (81.95)	43.67 (44.58)	-	-
NarL	73.38 (72.11)	45.20 (46.35)	-	-
PurR	95.24 (92.31)	8.70 (7.14)	77.14 (70.37)	22.22 (25.00)
TrpR	85.71 (50.00)	30.00 (60.00)	46.94 (40.91)	42.35 (45.00)
TyrR	69.23 (50.00)	18.75 (27.27)	73.68 (61.54)	32.14 (40.91)

A comparison between the functional annotations of genes known to be regulated by a given protein and the functional annotations of the predicted set of genes. The percentage with related function is calculated against all the genes with functional annotations, while the percentage without functional annotations is calculated against the whole set of predicted genes. The number in parentheses excludes genes known to be part of the corresponding regulon. In cases with high correlation of functional annotations we can propose a related function for genes without functional annotations, as in the Consensus/Patser predictions of ArgR, CRP and PurR (all with correspondences above 90%). Detailed tables are provided as Additional data files.

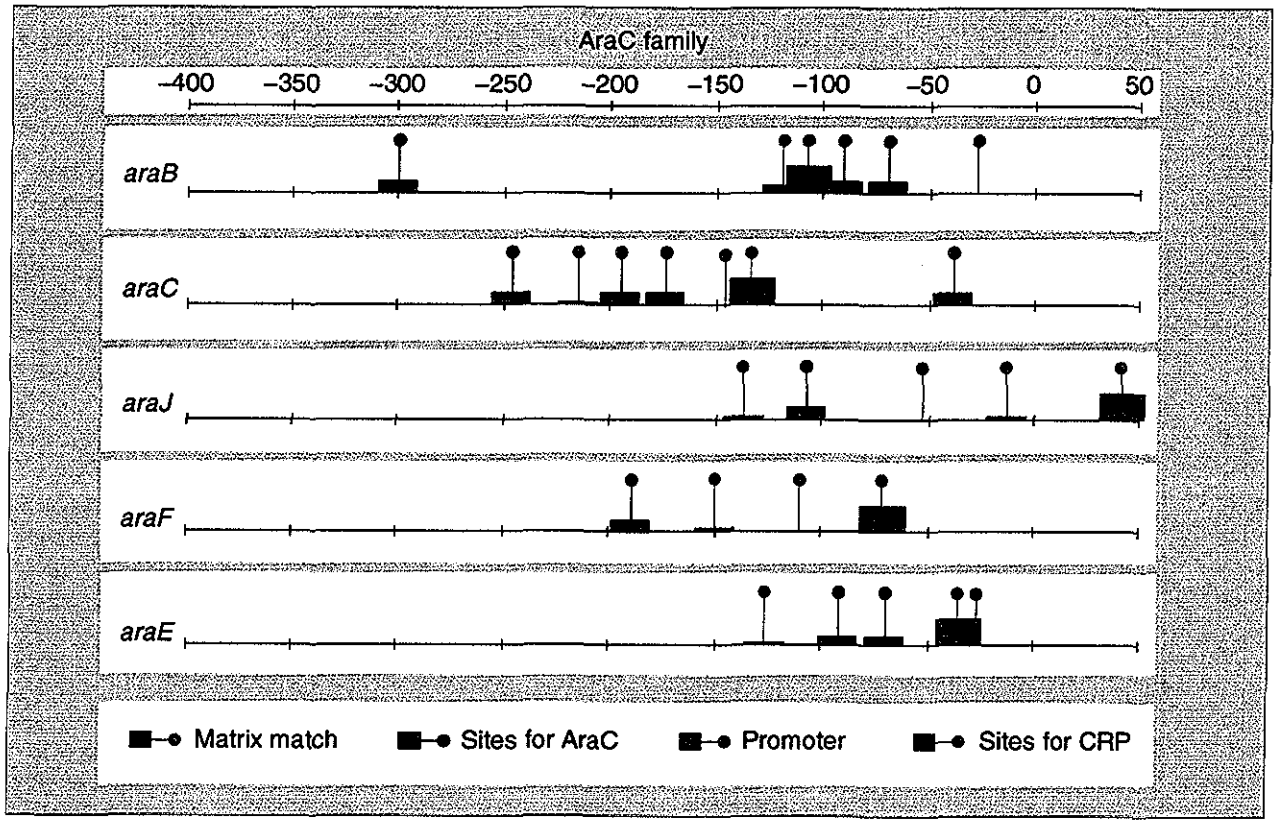


Figure 6
The positions found by Consensus/Patser. If Consensus is run to find an alignment within the 400+50 regions, the resulting matrix finds sites within each region (indicated here by the sites labeled 'matrix') that do not always match the binding sites for the relevant regulatory protein (AraC in the case illustrated here), but are very specific to the gene family. The sequence found does not correspond to known binding sites for other regulatory proteins (for example CRP) within the regions nor to the promoter.

Conclusions

Stringent evaluations of pattern discovery and pattern searching methods should be carried out to establish the confidence of a given prediction. Here we take advantage of the availability of reasonable negative samples - all other known regulons described in RegulonDB, except the one under study - in order to use standard statistical measurements of performance such as specificity and PPV. The PPV allowed us to stress how important even low rates of false positives might become in a large population. The small proportion of genes expected to be regulated by a given regulatory protein makes it important to emphasize the need for a stringent threshold to admit new members of regulons, as the true positives might be diluted in a high number of false positives. Nevertheless, if additional independent evidence is available, thresholds can be relaxed to include as many predictions as the confirmation procedure (genetic evidence of the regulatory effect, for instance) would allow. For instance, if the two computational methods were combined, only one of the genes known to be regulated by LexA (see previous section) would be rejected by both methods (*ybfE* in Table 10), while 16 of the wrong predictions are rejected by both methods (Table 11).

A very striking observation that deserves experimental analysis is the behavior of Consensus when identifying binding sites versus upstream regions. The program discovers patterns that discriminate, very specifically, the upstream regions used as training sets from the other regions. However, the patterns found do not always match the DNA binding sites. What are these specific motifs? These results imply the existence of new sequence elements specific to each family, different from those reported in the literature. We have not yet found (data not shown) any additional property that could suggest their function; their distance from the start site of transcription to known binding sites is not conserved; in some cases the predicted motif occurs upstream of the known sites in some promoters and downstream in other promoters. We have, of course, verified these observations twice, and find no additional property to associate with such families.

Table 8

Dyad-analysis/sweeping predictions in regions without binding sites reported in RegulonDB

Regulatory protein and genes shown to be regulated by it	Above threshold	Site coordinates and number of matches	Site sequence
ArcA (12)			
aceB	-	i:-144;f:-85;m:8	TTATCAAGTATTTTTAATTAAAATGGAAATTGTTTTTGATTTTGCATTTTAAATGAGTAG
fadB	-	i:-96;f:-51;m:6	ATTTCTTTAATCTTTTGTTTGCATATTTTTAACACAAAATACACAC
fumA	+	i:-77;f:-56;m:14	TATTGTTACTCGCTTTTAACAG
fumC	-	i:-92;f:-53;m:10	ATTTGTTATCAAATGGTAAATAATAAGTGAGCTAAAAGTT
glpA	-	i:-248;f:-232;m:8	TTATTTATGATTAACAG
hyaA	-	i:-168;f:-150;m:12	TACGCTTTATTAACAATAC
lpdA	+	i:-232;f:-204;m:15	TGTTTAAAAATTGTTAACAATTTTGTAAA
sucA	-	i:-323;f:-300;m:5	TGTTGTTGCAACGTAATGCGTAAA
ArgR (11)			
argD	-	i:-63;f:-51;m:5	TTTTTATGCATAT
FNR (4)			
aspA	-	i:-156;f:-144;m:3	TGATCTATTTCAC
cyoA	+	i:-25;f:2;m:5	GATCCCGTGGAATTGAGGTCGTTAAATG
icdA	+	i:-306;f:-292;m:6	ATTGAACAGGATCAC
sdhC	-	i:-340;f:-326;m:2	GATGATTAAAAATTA
LexA (9)			
umuD	+	i:-57;f:11;m:24	CTGCTGGCAAGAACAGACTACTGTATATAAAAACAGTATAACTTCAGGCAGATTATTAT GTTGTTTATC
Lrp (1)			
livK	+	i:-277;f:-269;m:2	CAGCATAAT
sdaA	-	-	-
serA	-	i:-146;f:-139;m:1	CAGCATAT
NarL (1)			
adhE	-	i:-215;f:-201;m:1	TACCCAGAAGTGAGT
caiF	-	-	-
torC	+	i:-209;f:-195;m:2	TACCCCTCCTGAGTG
PurR (15)			
codB	+	i:-82;f:-64;m:16	ACGAAAACGATTGCTTTTT
prsA	+	i:-356;f:-344;m:21	GAAAACGTTTTCG
speA	-	i:-132;f:-119;m:6	GAAACCGGTTGCGC

Sequences and positions of binding sites predicted by dyad sweeping in genes with experimental evidence for co-regulation in RegulonDB, but with no binding site experimentally identified. Genes follow the alphabetic order of the regulatory proteins, with the name of the protein separating each group. The number in parentheses after the regulator is the value of the threshold - derived from requesting best overall performance. The site coordinates are 'i' for initial base, 'f' for final position relative to the start codon. The score is given as the maximum number of matching ('m') dyads within a ROM. The number of families used was the same for any method, but we only show families where the methods provided significant results.

In the comparison of the two methods we have not found that one of them performs better in all the evaluations and scenarios considered (pattern search, pattern abstraction and pattern discovery). This implies that one could consider combining the different methods to make the best use of their respective strengths. For instance, if there is evidence of co-regulation only, we would suggest using Dyad-analysis/sweeping first to find the binding sites. If Dyad-analysis finds significant dyads, the dyad sweeping methodology can be used to extract possible binding sites. After that, the predicted sites can be used to train Consensus and search for further co-regulated genes. In cases where the DNA binding sites are known, Consensus/Patser, which are both very fast and simple to use, can give very reliable results in a short time.

The combination of computationally more confident predictions, together with additional independent evidence - for example, functional classes or operon organization - is an intelligent strategy for making more robust predictions. These more robust upstream regulatory analyses can be used to assign function to unknown genes, as illustrated here with the ArgR, CRP and PurR regulons. One can envisage highly relevant genomic applications of these predictions, such as distinguishing orthologs within families of paralogous genes, based on their differential regulation, or identifying non-orthologous gene displacement on the basis of regulatory comparisons.

The goal in computational biology is twofold: to provide, on the one hand, methods that generate useful and evaluated

Table 9

Consensus/Patser prediction in regions without binding sites reported in RegulonDB

Regulatory protein and genes thought to be regulated by it	Above threshold	Site coordinates and score	Site sequence
ArcA (8)			
aceB	+	i:-164;f:-144;sc:9.48	TTCATATTGTTATCAACAAG
fadB	-	-	-
fumA	-	-	-
fumC	+	i:-74;f:-54;sc:10.82	AATAATAAGTGAGCTAAAAG
glpA	-	i:-184;f:-164;sc:6.11	AAGAAAACATTCATAAATTA
hyaA	-	-	-
lpdA	+	i:-228;f:-208;sc:8.43	TAAAAATTGTTAACAATTTT
sucA	-	-	-
ArgR (13)			
argD	-	i:-70;f:-50;sc:11.12	TAGTGATTTTTTATGCATAT
CRP (6)			
cirA	+	i:-51;f:-31;sc:6.39	ATGTGAGCGATAACCCATTT
dsdA	-	-	-
ebgA	+	i:-91;f:-71;sc:7.53	TCGTGATCCAGTTAAAGTAA
flhD	+	i:-269;f:-249;sc:10.86	GTGTGATCTGCATCACGCAT
fucA	+	i:-399;f:-379;sc:10.18	ATATGACGGCGGTCACACTT
fucP	+	i:-205;f:-185;sc:9.74	AAGTGATGGTAGTCACATAA
glgC	-	i:-166;f:-146;sc:3.60	TCGCAATTAACGCCACGCTT
gntK	+	i:-169;f:-149;sc:11.49	ATTTGAAGTAGCTCACACTT
lpdA	-	i:-335;f:-315;sc:3.73	TGGTGATGTAAGTAAAAGAG
melA	-	i:-228;f:-208;sc:3.79	CTGCGAGTGGGAGCACGGTT
speC	-	i:-16;f:4;sc:5.55	GTTTGACCCATATCTCATGG
srlA	+	i:-91;f:-71;sc:8.89	TTGCGATCAAAATAACACTT
ubiG	-	i:-234;f:-214;sc:5.93	CAATGACCGACATCGCATAA
CysB (4)			
cysP	+	i:-237;f:-217;sc:7.04	TTTATTTGTCATTTTGGCCC
CytR (1)			
nupC	-	-	-
FNR (7)			
aspA	-	i:-367;f:-347;sc:3.88	CATGGGCAACCTGAATAAAG
cyoA	-	i:-182;f:-162;sc:4.31	TTTGTTATAACGCCCTTTTG
icdA	-	i:-105;f:-85;sc:6.30	AATCATTAACAAAAAATTGC
sdhC	-	i:-4;f:16;sc:3.89	ATTCATGATAAGAAATGTGA
FruR (5)			
aceB	+	i:-253;f:-233;sc:11.44	GATCGTTAAGCGATTCAGCA
fruB	+	i:-38;f:-18;sc:13.85	GAGGCTGAATCGTTTCAATT
ppsA	+	i:-105;f:-85;sc:6.29	TTTGCTTGAACGATTCACCG
IHF (7)			
caiT	+	i:-83;f:-63;sc:8.41	AATAATAATTATATTAAATG
ecpD	+	i:-39;f:-19;sc:8.96	ATTATTCCCTGTTTTAATTA
himA	-	-	-
himD	-	i:-136;f:-116;sc:6.43	ATTCCGAAGTTTGTTGAGTT
hycA	-	i:-73;f:-53;sc:6.57	TAATAACAATAAATTAAAAG
hypA	-	i:-155;f:-135;sc:6.75	TTAATTTATTGTTATTAAAG
narK	-	i:-106;f:-86;sc:6.66	AAATATCAATGATAGATAAA
ompR	-	i:-135;f:-115;sc:6.39	TATACTTAAGCTGCTGTTTA
sucA	-	-	-
LexA (9)			
umuD	+	i:-40;f:-20;sc:10.80	CTACTGTATATAAAAACAGT
Lrp (8)			
livK	+	i:-235;f:-215;sc:8.35	TGCCGTTATTTTATGCTGAC
sdaA	-	i:-317;f:-297;sc:4.95	ATCACCCTTTAGATATCTAC
serA	-	i:-79;f:-59;sc:6.62	TGCCGCAATATTATTTTTTG

Table 9 *(continued)*

Regulatory protein and genes thought to be regulated by it	Above threshold	Site coordinates and score	Site sequence
NarL (7)			
adhE	-	i:-160;f:-140;sc:6.11	ATAACTCTAATGTTTAAACT
caiF	-	i:-163;f:-143;sc:5.62	CAAATAATAGCGTGTCATGG
torC	-	i:-20;f:0;sc:4.98	ATAATTCTACAGGGGTTATT
PurR (11)			
codB	+	i:-84;f:-64;sc:13.10	CCACGAAAACGATTGCTTTT
prsA	+	i:-360;f:-340;sc:12.33	GCAAGAAAACGTTTTCGCGA
speA	-	i:-136;f:-116;sc:7.05	AAAAGAAACCGGTTGCGCAG

Data and analysis as described in Table 8. sc, Score as obtained from Patser. The number of families used was the same for any method, but we only show families where the methods provided significant results.

Table 10

Predictions in experimentally characterized binding sites for LexA

Gene	Consensus/Patser		Dyad sweeping	
b1728	+	sc:10.89	+	m:13
b1741	-	-	+	m:9
dinD	+	sc:14.01	+	m:20
dinG	+	sc:9.11	+	m:16
dinI	+	sc:12.22	+	m:12
dinP	-	-	+	m:14
ftsK	-	sc:7.82	+	m:11
lexA	+	sc:17.45	+	m:21
molR_1	+	sc:10.46	-	m:8
polB	-	sc:8.30	+	m:14
recA	+	sc:14.71	+	m:32
recN	+	sc:13.56	+	m:23
ruvA	+	sc:11.12	-	m:8
sbmC	+	sc:14.08	+	m:27
ssb	+	sc:12.99	+	m:29
sulA	+	sc:15.64	+	m:22
umuD	+	sc:10.80	+	m:24
uvrA	+	sc:16.30	+	m:29
uvrB	+	sc:14.94	+	m:12
uvrD	+	sc:16.07	+	m:15
ybfE	-	sc:8.18	-	m:3
yebG	+	sc:14.37	+	m:23
yjiW	+	sc:12.21	+	m:14

Fernandez De Henestrosa *et al.* [15] experimentally characterized new LexA-binding sites, which are not included in RegulonDB. The table shows our binding-site predictions with dyad sweeping and with Patser, using their corresponding best overall performance thresholds. sc, Score as obtained by Patser; m, maximum number of matching dyads. Note that most genes clearly have ROMs with 10 or more matches and with scores of Patser above 10.

Table 11

Contrasting predictions: regions known to lack LexA sites

Gene	Consensus/Patser		Dyad sweeping	
b3020	+	sc:13.66	-	m:6
brnQ	-	-	-	m:2
creA	+	sc:11.67	-	m:6
dinJ	+	sc:11.52	+	m:13
ecpD	-	-	+	m:9
hofQ	-	-	-	m:4
ilvD	-	-	-	m:6
ivbL	+	sc:13.23	+	m:21
metE	-	-	+	m:13
metR	+	sc:9.10	+	m:13
minC	+	sc:12.22	+	m:11
pshM	-	-	-	m:4
rfaJ	-	-	-	m:8
rob	-	-	-	m:6
xylE	+	sc:9.15	+	m:11
yafL	+	sc:15.21	+	m:13
ybiA	+	sc:12.07	+	m:16
ybiT	-	-	-	m:5
ycgJ	+	sc:13.42	+	m:11
ycgL	-	-	-	m:3
yciG	-	-	-	m:3
ydbH	-	-	-	m:3
ydeJ	+	sc:9.05	-	m:6
yecS	+	sc:9.89	-	m:6
yfiE	-	-	-	m:6
yfiK	-	-	-	m:6
ygjF	-	-	-	m:4
yhiX	-	-	-	m:4
yiaO	-	-	-	m:2
yigN	-	-	-	m:5
yjgN	-	sc:7.66	+	m:10

After experiment, Fernandez De Henestrosa *et al.* [15] rejected this set of genes in which they had predicted LexA sites using other computational methods. We tested the capacity of dyad sweeping and Patser to also reject these false positives. sc, Score as obtained by Patser; m, maximum number of matching dyads. Note that for both methods, most of the genes here show much smaller scores than genes belonging to the LexA regulon (see Table 10).

predictions, and, on the other hand, to use such methods as models of the biology under study. This latter virtue could generate new ways of understanding fundamental processes in gene regulation, along with, as suggested here, new properties of gene regulation at the genomic level. We cannot rely on a single methodology to solve the problems. Each algorithm should be tested on well-defined problems in order to find their strengths. Thus it should be possible to choose which method, or combination of methods, is best suited for the problem at hand.

Additional data files

Additional data files containing the functional annotations associated to the genes within each regulons, and of those genes downstream of predicted binding sites are available with the online version of this article.

Acknowledgements

E.B.B. has been supported by a fellowship from Conacyt, grant 0028. This research was supported by grants from CONACYT No. 0028 and from DGAPA to J.C.V. We acknowledge fruitful discussions with Jacques van Helden and suggestions by an anonymous referee. We also appreciate computational support from Victor del Moral.

References

1. Velculescu VE, Zhang L, Zhou W, Vogelstein J, Basrai MA, Bassett DE Jr, Hieter P, Vogelstein B, Kinzler KW: **Characterization of the yeast transcriptome.** *Cell* 1997, **88**:243-251.
2. Hertz GZ, Hartzell GW 3rd, Stormo GD: **Identification of consensus patterns in unaligned DNA sequences known to be functionally related.** *Comput Appl Biosci* 1990, **6**:81-92.
3. Lawrence CE, Reilly AA: **An expectation maximization (EM) algorithm for the identification and characterization of common sites in unaligned biopolymer sequences.** *Proteins* 1990, **7**:41-51.
4. Waterman MS, Arratia R, Galas DJ: **Pattern recognition in several sequences: consensus and alignment.** *Bull Math Biol* 1984, **46**:515-527.
5. Wolfertstetter F, Frech K, Herrmann G, Werner T: **Identification of functional elements in unaligned nucleic acid sequences by a novel tuple search algorithm.** *Comput Appl Biosci* 1996, **12**:71-80.
6. van Helden J, Andre B, Collado-Vides J: **Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies.** *J Mol Biol* 1998, **281**:827-842.
7. Hughes JD, Estep PW, Tavazoie S, Church GM: **Computational identification of cis-regulatory elements associated with groups of functionally related genes in *Saccharomyces cerevisiae*.** *J Mol Biol* 2000, **296**:1205-1214
8. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8**:1202-1215.
9. Crowley EM: **A Bayesian method for finding regulatory segments in DNA.** *Biopolymers* 2001, **58**:165-174.
10. Huerta AM, Salgado H, Thieffry D, Collado-Vides J: **RegulonDB: a database on transcriptional regulation in *Escherichia coli*.** *Nucleic Acids Res* 1998, **26**:55-59.
11. Salgado H, Santos-Zavaleta A, Gama-Castro S, Millan-Zarate D, Diaz-Peredo E, Sanchez-Solano F, Perez-Rueda E, Bonavides-Martinez C, Collado-Vides J: **RegulonDB (version 3.2): transcriptional regulation and operon organization in *Escherichia coli* K-12.** *Nucleic Acids Res* 2001, **29**:72-74.
12. van Helden J, Rios AF, Collado-Vides J: **Discovering regulatory elements in non-coding sequences by analysis of spaced dyads.** *Nucleic Acids Res* 2000, **28**:1808-1818.
13. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences.** *Bioinformatics* 1999, **15**:563-577.
14. Gralla JD, Collado-Vides J: **Organization and function of transcription regulatory elements.** In *Cellular and Molecular Biology*: Escherichia coli *and* Salmonella. Edited by Neidhardt FC, Curtiss III R, Ingraham J, Lin ECC, Low KB, Magasanik B, Reznikoff W, Schaechter M, Umbarger HE, Riley M. Washington, DC: American Society for Microbiology, 1996, 1232-1245.
15. Fernandez De Henestrosa AR, Ogi T, Aoyagi S, Chafin D, Hayes JJ, Ohmori H, Woodgate R: **Identification of additional genes belonging to the LexA regulon in *Escherichia coli*.** *Mol Microbiol* 2000, **35**:1560-1572
16. **Regulatory sequence analysis tools** [http://embnet.cifn.unam.mx/~jvanheld/rsa-tools/]
17. Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, *et al*: **The complete genome sequence of *Escherichia coli* K-12.** *Science* 1997, **277**:1453-1474.
18. Serres MH, Riley M: **MultiFun, a multifunctional classification scheme for *Escherichia coli* K-12 gene products.** *Microb Comp Genomics* 2000, **5**:205-222.