UNIVERSIDAD NACIONAL AUTÓNOMA

DE MÉXICO

FACULTAD DE CIENCIAS

"RECONOCIMIENTO COMPUTACIONAL DE ELEMENTOS CIS DE LA PROTEÍNA GAL4 EN *Saccharomyces cerevisiae*: COMPARACIÓN DE ALGORITMOS Y APLICACIONES"

T E S I S

QUE PARA OBTENER EL TÍTULO DE:

B I Ó L O G A

P R E S E N T A

VICTORIA FABIA DOMÍNGUEZ DEL ANGEL

1996

TESIS CON
FALLA DE ORIGEN

VNIVERSIDAD NACIONAL
AVPNOMA DE
MEXICO

M. en C. Virginia Abrín Batule
Jefe de la División de Estudios Profesionales de la
Facultad de Ciencias
P r e s e n t e

Comunicamos a usted que hemos revisado el trabajo de Tesis: "Reconocimiento Computacional de Elementos Cis de la Proteína GAL4 en Saccharomyces cerevisiae: Comparación de Algoritmos y Aplicaciones."

realizado por Domínguez Del Angel Victoria F.

con número de cuenta 9052158-2 , pasante de la carrera de Biología

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis
Propietario Dr. Pedro Julio Collado Vides

Propietario Dr. Pedro Miramontes Vidal

Propietario Biólogo Alejandro Pelaez Goycochea

Suplente M. en C. Víctor Manuel Valdés López

Suplente Biólogo Alvaro Chaos Cador

FACULTAD DE CIENCIAS

Consejo Departamental de Biología
M. en C. Alejandro Martínez Mena

COORDINACION GENERAL
DE BIOLOGIA

La experiencia es la única fuente de la verdad:
sólo ella puede enseñarnos algo nuevo;
sólo ella puede darnos certeza .
Henri Poincaré

Al Ingeniero Ezequiel Domínguez Cervantes,

por darme su naturaleza soñadora..

Quiero agradecer a todas las personas que contribuyeron de manera directa e indirecta a la realización de esta tesis. En particular agradezco:

A la Universidad Nacional Autónoma de México, por darme la oportunidad de ser universitario.

A los cuatro sinodales, por sus críticas y valiosos comentarios: Pedro, Alejandro, Víctor y Alvaro.

A mi director de tesis Dr. Julio Collado-Vides, por su apoyo y paciencia en todo este tiempo.

A Denis Thieffry, por ser un verdadero maestro y amigo.

A Karine, por su amistad incondicional.

A todos mis compañeros del laboratorio: Araceli, Ernesto, Concepción, Alma, Víctor, Miryam y Hely por su convivencia y colaboración en el trabajo.

A mis hermanos: Claudia, (por no cerrarme las puertas en los momentos difíciles) Adelina, (por ser el mejor ejemplo de eficiencia y dedicación) y Ezequiel (por mostrarme mis defectos y ayudarme a ser mejor persona).

A mis sobrinos: Olimpia, Tomasito, María y Emilio por toda la alegría que me han brindado.

A mis tíos Alma y Cesar, que con sus consejos, apoyo y ejemplo nos han fortalecido como familia.

A Dios, porque sólo con su existencia puedo justificar las cosas que he logrado.

A mi madre, por ser lo más valioso que tengo en mi vida.

RESUMEN

Existe en el área de la biología computacional una gran variedad de programas que realizan alineamientos de secuencias de DNA y búsquedas de patrones. Esto nos motivó a llevar a cabo un estudio comparativo de los diferentes programas. Para la evaluación se hicieron pruebas con las secuencias de DNA donde se une la proteína reguladora Gal4 en *Saccharomyces cerevisiae* y se evaluó cuales eran los programas con los que se obtenían mejores resultados.

Después, se determinó el mejor programa de búsqueda para el análisis de las secuencias reguladoras del cromosoma II de *Saccharomyces cerevisiae*. Siguiendo la estrategia, llamada *Top down*, se utilizaron secuencias con funciones conocidas (en este caso las de Gal4) y se hicieron búsquedas para encontrar posibles funciones. Siguiendo otra estrategia llamada *Bottom up*, se analizó la distribución estadística de cadenas de DNA de tamaño definido en secuencias reguladoras, con el propósito de encontrar algún significado biológico.

Posteriormente se hizo un análisis *Bottom up* en las regiones reguladoras de los cromosomas III y XI con cadenas de 5-n-5meros, 6-n-6meros y 8-n-8meros, donde n es la distancia de separación que tomó valores de 4, 5, (para encontrar las cadenas fuera de fase) 9 y 10 (para encontrar las cadenas en fase). Se finalizó con un análisis a partir de la distribución de las cadenas de 5-n-5 meros, cuantificando las cadenas con simetría directa e invertida que estuvieran sobre y sub-representadas en sus diferentes distancias y observar las posibles diferencias en la distribución de los n-meros con respecto a su fase de unión.

ABSTRACT

A series of programs aiming to align DNA sequences or to find specific sequence patterns have already been developed. This motivated us to perform a systematic comparative study of the main available programs. As a testing set of sequences, we used a collection of characterized DNA binding sites for the Gal4 protein of *Saccharomyces cerevisiae*. This comparative study led us to select the best set of programs for alignment of DNA sequences and for the search of patterns involved in gene regulation.

These programs were then used to analyze the regulatory regions of the chromosome II of *S. cerevisiae*. A first strategy, called "Top-down", consisted in using a set of sequences with known biological function (binding of Gal4) and the corresponding consensus matrix to perform a preliminary search for new sites in putative regulatory regions. A second strategy, called "Bottom-up", consisted in a statistical analysis of the distribution of small DNA sequences of given size in regulatory regions. More specifically, in this respect, we used 3-11n-3mers, e.g., defined triplet followed 11 any nucleotides, followed again by a defined triplet, which is structurally similar to Gal4 binding site.

Moreover, still following a Bottom up strategy, we studied the distributions of all possible 5-n-5mers, 6-n-6mers and 8-n-8mers in the regulatory regions on chromosomes III and XI (n either equal to 4 or 5, leading to repeats out of phase, or to 9 or 10, leading to repeats in phase)

Finally, we looked for the 5-n-5mers which (1) include direct or inverted repeats, mimicking the organization of potential binding sites for regulatory proteins, (2) are over- or under-represented, thus suggesting some biological function.

ÍNDICE

I. ANTECEDENTES:

Los millones de células que constituyen a un organismo multicelular provienen todas de una célula única: el óvulo fecundado. Conforme se divide y genera nuevas células, estas van diversificándose y adquiriendo funciones especializadas. Uno de los mecanismos que ayuda al prendido y apagado de genes es la acción de un grupo de proteínas conocidas como **proteínas reguladoras**; las cuales son capaces de unirse de manera específica a secuencias de DNA y con ello bloquear o favorecer el inicio de la transcripción.

En nuestro laboratorio se empezó colectando información biológica sobre el inicio de la transcripción en organismos procariontes, principalmente *Escherichia coli*. En un estudio realizado por J. Collado-Vides y J. D. Gralla (en prensa) se analizaron las regiones de pegado de las proteínas reguladoras en una colección de promotores $\sigma 70$ (observándose zonas diferentes para la represión y activación) y $\sigma 54$. Otros trabajos en este sentido son: la formalización (utilizando herramientas gramaticales) del mecanismo de regulación en el inicio de la transcripción, el análisis de las propiedades de las proteínas reguladoras y la construcción de una base de datos que incluya toda esta información (Regulón D.B.).

Actualmente un trabajo similar se está iniciando en organismos eucariontes, tomando como modelo la regulación en ***Saccharomyces cerevisiae***. Para abordarlo, es necesario conocer los diferentes programas utilizados en los análisis de secuencias. De manera general lo que esta tesis persigue es definir y optimizar estrategias que permitan integrar y analizar de la manera más eficiente la información contenida en un gran número de regiones reguladoras.

En el segundo capítulo, la primera sección introduce una visión general del mecanismo y de los diferentes elementos involucrados en el inicio de la transcripción en eucariontes. La segunda sección presenta una breve explicación de la biología de *Saccharomyces cerevisiae* y su proteína reguladora Gal4, resaltando la importancia que han tenido como modelos en los estudios de la regulación en eucariontes.

En el tercer capítulo, se plantea de manera más detallada los objetivos de la tesis.

El cuarto capítulo presenta una primera sección con el esquema de trabajo a realizar. La segunda sección muestra una descripción general sobre los programas e incluye una introducción a la Teoría de la Información, fundamento teórico en el que están basados varios programas que se utilizan en este trabajo.

Los resultados y discusiones están contenidas en el quinto capítulo. En la primera sección se muestran los resultados comparativos entre los diferentes programas, empezando por los de alineamiento y continuando con los de búsqueda. En esta sección se escoge un programa de

búsqueda y otro de alineamiento, para ser utilizados en la realización de estrategias de análisis en regiones reguladoras. La segunda sección presenta los resultados obtenidos de las estrategias de análisis.

En las conclusiones y perspectivas, capítulo sexto, se resumen los resultados más relevantes y se reflexiona sobre las posibles direcciones que se podrían tomar para hacer más eficientes los análisis en regiones reguladoras.

II. INTRODUCCIÓN.

II.1 MECANISMOS DE REGULACIÓN GENICA.

II.1.1. Control de la Expresión:

Las diferencias en estructura y función de los diferentes tipos celulares de un organismo multicelular dependen de los distintos niveles del control de la expresión génica; este control se puede dar a seis diferentes niveles entre la lectura del DNA y su transformación en proteínas. (Figura 1).

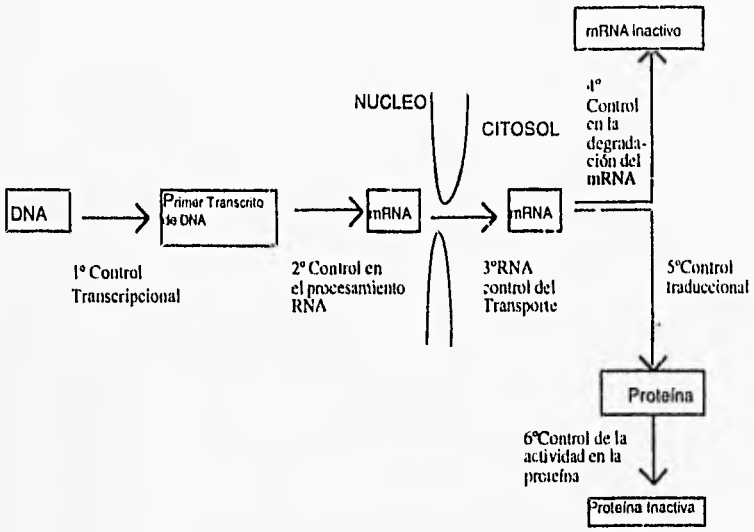


Figura 1. Representación esquemática donde se muestran los seis niveles de control de la expresión génica.

En cualquiera de los seis niveles se da algún tipo de regulación. De los seis, el más importante como punto de control para muchos genes es el primero: el inicio de la transcripción.

II.1.2. Elementos para Inicio de la Transcripción en Eucariontes:

a) Aparato basal:

En el núcleo existen 3 diferentes polimerasas RNA para cada tipo de RNA. El mRNA es el único que va a ser traducido a proteínas y es transcrito por la **Polimerasa II RNA.**

La polimerasa II RNA es una proteína de más de 500 kdaltons, de entre 8-14 subunidades, tres de las más grandes tienen homología con las subunidades de la polimerasa RNA de bacteria. Esta polimerasa tiene un dominio carboxilo terminal (CTD), que consiste de una secuencia repetitiva de aminoácidos: Tyr-Ser-Pro-Thr-Ser-Pro-Ser. Esta secuencia tiene un número variable de repeticiones en los organismos dependiendo de su escala evolutiva, levadura tiene ~26 repeticiones y mamíferos ~50.

Para la construcción del complejo basal de transcripción se necesitan los factores de transcripción (**TF**), existen prioridades entre unos y otros, pero de manera general son seis. El primer paso para el ensamblaje lo inicia el factor **TFIID**, que se compone principalmente de dos elementos: la *TATA-binding protein* (**TBP**) y de por lo menos otras ocho distintas subunidades conocidas con el nombre de **TAFs** (*TBP-associated factor*). TBP es de las proteínas mas conservadas a lo largo de la escala evolutiva (su dominio funcional muestra más de un 80% de identidad en una gran variedad de especies eucarióticas). Se va unir al promotor conocido con el nombre de caja TATA, que consta de ocho pares de bases y su consenso con sólo A•T situado en el surco menor aproximadamente en el sitio -25 con respecto al +1 o inicio de la transcripción. Tiene una forma que asemeja una silla de montar, donde la parte cóncava es la que se une al DNA y su parte convexa es la que va interactuar con otros factores, TAFs, para formar distintos complejos multiproteicos (TFIID o proteínas reguladoras negativas) . Al unirse a la caja TATA, provoca una distorsión hacia adelante del pegado en la estructura del DNA, facilitando su desenrollamiento y dándole orientación a la transcripción. Esta primera interacción va a ser estabilizada por un segundo factor TFIIA, el cual remueve la unión de TBP con TAFs reguladores negativos evitando la represión.

El siguiente factor **TFIIB** con sus dos dominios, presenta un papel doble dentro del complejo de iniciación. Su dominio COOH-terminal, se une al complejo TBP-DNA; mientras que el dominio NH2-terminal (posible dedo de zinc) se encarga (con ayuda de TFIIF) de reclutar a la polimerasa en el complejo, sirviendo de puente proteico entre TBP y la polimerasa.

TFIIF está formado por dos subunidades (la más pequeña presenta gran similitud con el factor σ70 de procariontes) y a través de este factor, se lleva a cabo la incorporación de la

polimerasa al complejo dando pie para la iniciación de la transcripción. Para finalizar el complejo de iniciación son necesarios otros dos factores: TFIIE y TFIIH.

La función de **TFIIE** no es aún muy clara, pero parece necesario para el posterior reclutamiento de TFIIH y TFIIJ. **TFIIH** posee actividad de quinasa y es capaz de fosforilar la CTD de la polimerasa.

Unidos todos estos factores a la polimerasa genera el complejo cerrado, el cual, se va a transformar en complejo abierto para iniciar la transcripción.

b) Elementos reguladores en *cis*:

El término *cis-acting* se refiere a las regiones reguladoras que actúan más o menos cerca de los genes en la misma molécula de DNA. Los loci reguladores activados en *cis* conocidos son: promotores, enhancer, boxes, dominios de doblado y heterocromatina facultativa.

Dentro de los primeros 100 pares de bases hacia arriba del inicio de la transcripción existen secuencias que tienen un importante efecto en los niveles de la transcripción, por influir directamente en la frecuencia de iniciación. Los de distribución más general son: la caja TATA (anteriormente discutida), la **caja GC**, la **caja CAAT** y el **octámero**. Todos los promotores requieren uno o más de estos elementos para funcionar adecuadamente.

La caja GC presenta un consenso con la secuencia GGGCGG. Usualmente se encuentra entre los 40-70 pares de bases hacia atrás del inicio de la transcripción (aunque este sitio puede variar). A ella se une el factor **SP1** cubriendo aproximadamente 20 pares de bases. En el promotor del gene para timidina-kinasa se presume que SP1 interactúa con el factor que se une a la CAAT box en uno de sus lados y con TFIID en el otro.

GGCCAATCT es la secuencia consenso de la caja CAAT, la cual puede ser reconocida por varios factores con afinidad variada, entre estos están CP1, la familia CTF, NF1, los cuales van a ayudar al complejo de transcripción en su ensamblamiento. A este misma secuencia se unen otros factores como CDP, el cual funciona como represor previendo que los activadores puedan reconocer estas secuencias.

En el caso de la secuencia Octámero cuyo consenso es ATTTGCAT, también sirve como reconocimiento para más de un factor. El de distribución más generalizada es **Oct-1**, pero en células linfoides, en el octámero para activar los genes de inmunoglobulina k ligth, en lugar de unirse Oct-1 se une el factor **Oct-2** funcionando como un factor tejido de específico.

Los elementos enhancer fueron primeramente descubiertos en virus. Son regiones en el DNA que influyen a distancia en los elementos transcripcionales (aparato basal) de un gen. Pueden estar hacia adelante o hacia atrás del gen que van a regular y pueden funcionar igualmente bien cuando se insertan en la posición opuesta (3' - 5').

Todos los genes eucariontes, desde hongos unicelulares hasta mamíferos, están asociados a unas proteínas con el nombre de **histonas** formando una estructura llamada **nucleosoma**. En promedio hay un nucleosoma por cada 200 pares de bases, estos a su vez están empaquetados en otras estructuras formando una fibra de cromatina de 30-n.m. Las fibras se organizan en una serie de *looped domain*, dominios doblados, que contienen entre 20 000 y 80 000 pares de

bases. Se tiene evidencia que la transcripción en eucariontes es iniciada en regiones donde la cromatina está descondensada, es decir esta zona presentan dominios de cis-acting que necesita estar en estado relajado para ser reconocidas por proteínas reguladoras.

c) Elementos Reguladores en *trans*.

En organismos eucariontes los elementos *trans-acting* en un cromosoma son capaces de influir en genes localizados en otros cromosomas. Típicamente estos elementos pueden actuar en otros cromosomas por la fácil difusión que tiene su producto. Estos pueden ser cofactores, RNA y en su mayoría son proteínas.

Estas proteínas reguladoras pueden reconocer secuencias específicas de nucleótidos en el interior de la doble hélice. Cada apareamiento de las bases expone en la superficie del DNA distintos patrones: donadores y aceptores de protones, parches hidrofóbicos (grupo methyl). Pero sólo en el surco mayor se puede distinguir el patrón para los cuatro tipos de apareamientos (Figura 2). Por esta razón las proteínas reguladoras generalmente se pegan al surco mayor.

Las regiones específicas de DNA son reconocidas por la superficie complementaria en la proteína conocida como dominio. Estos dominios tienen motivos que se han caracterizado muy bien y se han conservado a lo largo de la evolución en las diferentes proteínas reguladoras en eucariontes. Se clasifican en los cuatro tipos que se presentan a continuación:

Hélice-vuelta-hélice, este dominio se construye a partir de dos o más α hélices conectadas entre si por una corta cadena de aminoácidos haciendo una vuelta (Figura 3). La hélice más larga y cercana al carboxilo terminal es la hélice de reconocimiento, la cual se une al surco mayor del DNA. La primera vez que se descubrió fue en las secuencia nucleotídica de algunos genes homeoticos, en *Drosophila*, los cuales contienen una secuencia muy parecida de 60 aminoácidos a la que se le denomino homeodominio; cuando se caracterizo tridimensionalmente se observo una gran similitud con el motivo hélice-vuelta-hélice de los organismos procariontes. Se unen al DNA como monómeros a una secuencia 5' -A-T-T-A-3' y en todos los homeodominios hay cuatro residuos invariantes: Asn 51, Arg 53, Trp 48, Phe 49. La mayoría de los homeodominios se encuentran en la región amino-terminal de la proteína.

Algunas proteínas con este dominio se presentan a continuación:

	Función Biológica
	desarrollo embrionario de: insectos, vertebrados y plantas
	diferenciación en tejidos: hígado, riñón, hipófisis, linfocitos
	Determinación para el tipo sexual en levadura (feromonas)
	funciones múltiples (inmunoglobulinas, histonas y virus del herpes).

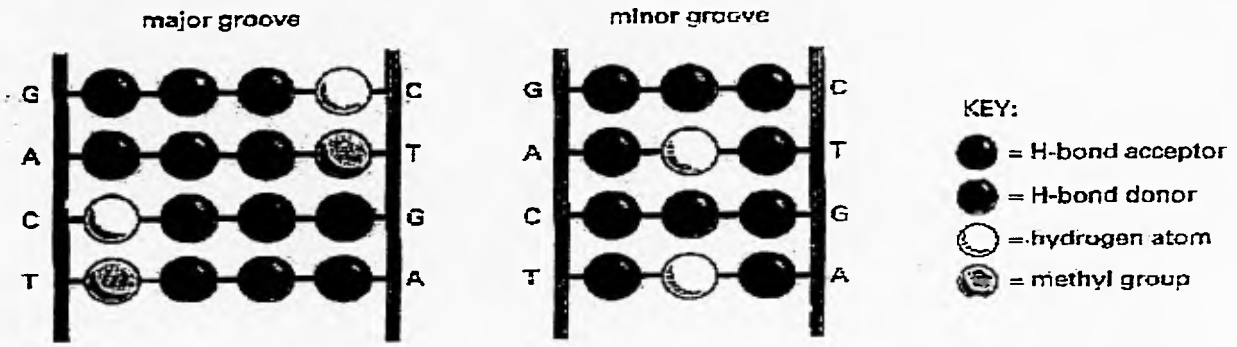


Figura 2.-Muestra esquemática de las diferencias en el reconocimiento del DNA que presentan los diferentes surcos. En el surco mayor se notan las diferencias en orientación de G-C /C-G y en A-T /T-A

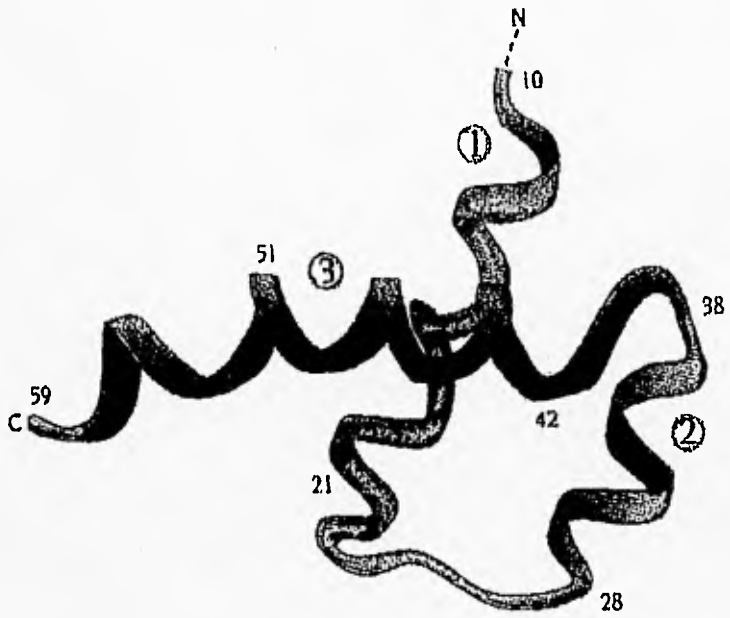


Figura 3. Estructura de pegado al DNA Hélice-Vuelta-Hélice. Las tres hélices se encuentran unidas por dos vueltas y la hélice más grande (la número 3) se le denomina de reconocimiento. Branden y Tooze, 1991.

Dedos de zinc, es un motivo ampliamente distribuido en eucariontes, el nombre se debe por un grupo más o menos conservado de aminoácidos que se une al ion zinc formando un motivo independiente dentro de la proteína. Se descubrió originalmente en TFIIIA (activador de genes ribosomales). Su estructura consiste en una antiparalela β plegada seguida por una hélice α coordinadas juntas a través de un átomo de zinc, se pega al surco mayor del DNA. Hay tres distintas familias de dedos de zinc (Branden, C. y Tooze, J. 1991):

Clásico, el átomo de zinc está ligado a dos cisteínas de la β plegada y a dos histidinas de la hélice α (figura 4a) , en esta familia el dedo de zinc tiende a repetirse en tándem (figura 4b). Se pueden encontrar desde uno hasta más de 30 dedos e interactúan en tándem también con el surco mayor del DNA.

4 Cisteínas , está ampliamente distribuido en receptores para glucocorticoides, contiene dos átomos de zinc y cada uno esta ligado a cuatro cisteínas formando un dominio globular con dos unidades de zinc. Esta familia, como el motivo hélice-vuelta-hélice en procariontes, forman dímeros; a través de las hélices α más cercanas al amino terminal, cada subunidad interactúa con el surco mayor en una secuencia palindrómica de DNA, mientras que las otras dos hélices ayudan a la formación del dímero (Figura 4c).

1 Histidina-3 Cisteínas, se ha observado en proteínas de virus, su secuencia de aminoácidos es Cys-X_2-Cys-X_4-His-X_4-Cys; el átomo de zinc está coordinado por tres cisteínas y una histidina (Figura 4d).

Como ejemplos tenemos las siguientes proteínas:

Pro[illegible]	Función Biológica
	regulación de RNA ribosomal de células eucariontes
	diferenciación de glóbulos rojos (globinas)
	desarrollo embrionário en *Drosophila*
	desarrollo de cerebro para vertebrados
	metabolismo de galactosa para levadura (galactosa-permeasa)
	mediadoras hormonales para animales (genes múltiples)
	diferenciación del intestino; insectos y vertebrados (TAT, HNF1)

Cierres de leucina, es un dominio que media la dimerización y el pegado al DNA. Cada 3.6 aminoácidos se forma una hélice α y en cada dos vueltas aparece una leucina. Esta última va interactuar paralelamente con las leucinas de otra hélice α formando construyendo de esta manera un puente para la dimerización (Figura 5a). Hecho esto, el resto de la región N terminal de ambas hélices reconocen la parte media de una región palindrómica en el surco mayor del DNA, pegándose en direcciones divergentes. La explicación anterior se puede visualizar en forma de "Y" , donde el tallo corresponde a la región dimerizada (por las uniones entre

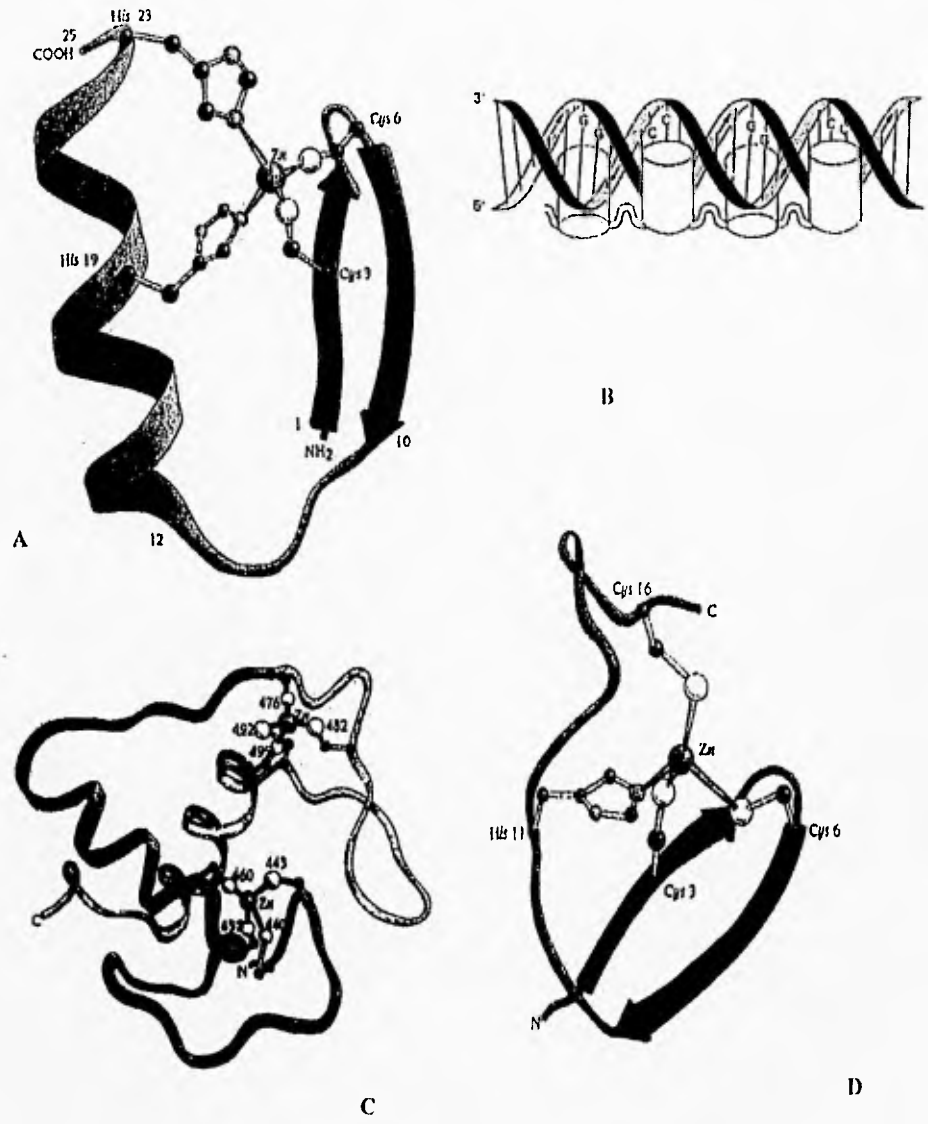


Figura 4. Representación esquemática de las distintas familias de dedos de zinc.
La figura 4a Muestra al clásico dedo de zinc, cuyo átomo de zinc se encuentra unido a dos cisteínas y a dos histidinas. 4b. Esquematiza la manera en tandem que interactúa el dedo clásico de zinc con el DNA . 4c Representa la estructura más distribuida de unión al DNA entre los receptores de glucocorticoides, donde cada átomo de zinc se encuentra unido a cuatro cisteínas. 4d. Diagrama correspondiente al dedo de zinc encontrado en el retrovirus del HIV.

A

B

Figura 5. Esquematización de los cierres de leucina.5a Muestra la interacción entre las moléculas de leucinas de las dos hélices α. 5b Visualización del dominio 'Y', la manera en que se representa la dimerización y el reconocimiento al DNA.

leucinas) y los brazos bifurcados de la "Y" a la región de unión con el DNA (Figura 5b). Las proteínas reguladoras que contienen cierres de leucina pueden formar homodímeros o heterodímeros.
Como ejemplo tenemos las siguientes proteínas:

Proteínas Reguladoras	Función Biológica
	diferenciación de tejidos hepáticos y adiposo (albúmina, TAT)
	controla la proliferación celular (genes múltiples)
	metabolismo de aminoácidos para levadura
	mediación extra e intracelular (TAT, receptores de células T)

Hélice-bucle-hélice, este motivo consiste en una corta hélice α conectada por un bucle a una segunda y más larga hélice α, estas dos hélices se unen ambas al DNA y a otro monómero (Figura 6), formando homodímeros o heterodímeros. Por ejemplo:

Proteínas Reguladoras	Función Biológica
MyoD, miosina	diferenciación de músculo (actina y miosina)
Myc, Max, Max	control para la proliferación y diferenciación celular

II.2 *Saccharomyces cerevisiae* Y LA PROTEÍNA GAL4:

Saccharomyces cerevisiae (Del latín *saccharon*, azúcar, + del griego *mykes*, hongo). Pertenece a la subclase Hemiascomycetidae y lo incluyen en el orden Endomycetales, los cuales se caracterizan por ser morfológicamente simple y no producir ascocarpo (pared celular gruesa).

Es típicamente unicelular de forma oval, contiene una vacuola grande y un núcleo excéntrico. Es un organismo heterotálico con dos tipos celulares bien definidos : a y α

Su reproducción es generalmente asexual, ocurre por gemación, produce largas cadenas de yemas. En condiciones ambientales desfavorables presenta reproducción sexual, formando un asci que contiene en su interior cuatro esporas (Figura 7).

S. cerevisiae es uno de los organismos experimentales más importante en la investigación. Se considera un sistema modelo para el entendimiento de la biología de los organismos eucariontes.

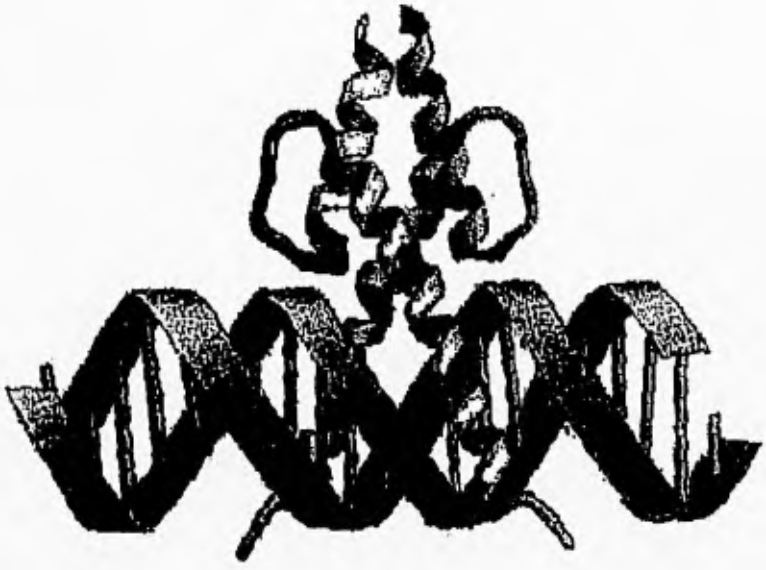


Figura.6. Esquematización del dominio hélice-bucle-hélice. Las primeras dos hélices de cada dominio se utilizan para la dimerización. Las segundas se utilizan para la unión con el DNA.

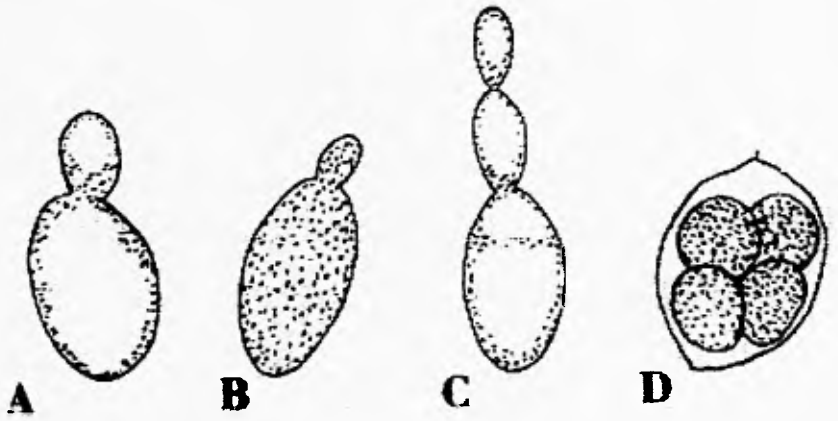


Figura 7.- Esquematización del proceso de gemación en *Saccharomyces cerevisiae.*
A - C Muestra la formación de yemas. **D** es el ascus con las cuatro ascosporas..

Las ventajas que posee son múltiples. Por un lado, las herramientas que se han ocupado para los estudios clásicos en biología molecular (tecnología del DNA recombinante) que se llevan a cabo para sistemas procariontes, son utilizables en el caso de levadura. *S. cerevisiae* ha sido el microorganismos eucariótico por excelencia para desarrollar esta tecnología, es de vida libre, tiene un crecimiento rápido y es inocuo; además se usa desde hace mucho tiempo para hacer pan y algunas bebidas alcohólica (incluyendo la cerveza y el vino).

Su genoma se incluyó en los proyectos de secuenciación y a partir del mes de abril de 1996 las secuencias de sus 16 cromosomas son completamente accesibles por la red electrónica (Internet). Esto lo convierte en el primer organismo eucarionte completamente secuenciado. Existe un gran desconocimiento sobre la función biológica en la mayoría de su genoma y sobre todo en sus regiones reguladoras.

S. cerevisiae tiene la capacidad de utilizar diferentes fuentes de carbono, una de ellas es galactosa, la cual tiene que pasar por la ruta metabólica de Leloy (Figura 8) para ser transformada en Glucosa-6-phosphato y entrar a la glucólisis.

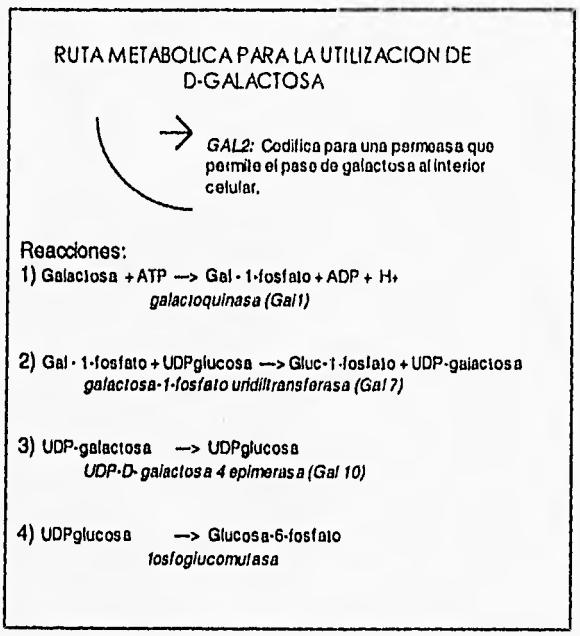


Figura 8. Utilización de las diferentes reacciones necesarias para el metabolismo de galactosa como fuente de carbono.

Los genes que codifican las enzimas necesarias en el metabolismo de Leloy son inducibles por la proteína reguladora Gal4. Gal4 (un regulador transcripcional de 881 aminoácidos) también activa la transcripción de los genes requeridos para el catabolismo de melibiosa. Su organización funcional es la siguiente:

Dominio de unión al DNA: Se une al DNA como dímero simétrico y reconoce una secuencia palindrómica en el surco mayor de 17 pares de bases. La región del aminoácido 1 al 64 se ocupa para el reconocimiento del DNA y del 65 al na 94 para la dimerización. La Primera fracción está constituida por tres distintos módulos: un compacto -metal-dominio- (residuos 8-40), una cadena extendida (41-49) y un elemento dimerizador inicial hélice α (Figura 9). El reconocimiento lo da un dominio de unión a un metal (zinc) que contiene seis residuos de cisteína coordinadas tetrahédricamenente por dos átomos de Zn 2+, lo que crea una región compacta y globular más o menos aislada; se une al surco mayor en los extremos del sitio y contacta directamente con tres pares de bases del DNA. Los residuos 41-49 forman un segmento extendido. Algunos aminoácidos con carga positiva contactan con los fosfatos de la cadena de DNA. Los aminoácidos 50-64 forman una hélice α la cual inicia el entrecruzamiento de manera paralela con la hélice α del otro monómero de Gal4.

Dominio de activación transcripcional: Los residuos 94-106, 148-196 y 768-881, forman tres zonas de regiones ácidas que se han identificado para la activación y son ricas en carga negativas. Mutaciones que reducen las cargas negativas, reducen la habilidad para realizar la activación. Por el contrario mutaciones que incrementan estas cargas, aumentan la activación (Ptashne, 1992). En principio se sugirió que este dominio ácido de activación (AAD) era una α-hélice; pero estudios más recientes presentaron a este dominio como una estructura hoja β, la cual en una de sus caras contiene muchos residuos hidrofóbicos para interactuar con el aparato basal de transcripción (Van Hoy, M. *et al.* 1993)

Dominio de unión con la proteína Gal 80: Es una región cercana al carboxilo terminal que abarca de la 851 a la 881. Gal 80 es una proteína que al interactuar con Gal 4, bloquea la acción activadora de la proteína impidiendo la unión del dominio activador con el aparato de transcripción (Leuthe, K. K. *et al.* 1992).

Figura 9. Esquema del dominio de unión al DNA de la proteína reguladora Gal4. Se puede observar la doble hélice del DNA representada por las cadenas más anchas. Los números en color negro ubican las bases que interactúan con la proteína en el surco mayor (la numeración se inicia en el punto medio de las cadenas). Las cadenas delgadas representan los dominios de las dos proteínas que se unen al DNA. Los aminoácidos del 8 al 40 representan el modulo de reconocimiento, del 41-49 se da el enlace con el DNA y del 50 al 64 se da la dimerización

II.3 INTRODUCCIÓN A LA TEORÍA DE LA INFORMACIÓN:

Cuando se estudian los sitios de pegado de moléculas al DNA o RNA y se realizan alineaciones, varios sitios son reconocidos por la misma macromolécula reconocedora y frecuentemente se escoge la base más común para cada posición y se crea una secuencia consenso.

Para cuantificar la variabilidad y medir la información que nos proporciona un patrón y poderlo distinguirlo de otras secuencias a las cuales las proteínas no se unen, un método ampliamente utilizado es el de **contenido informacional**. El contenido informacional (también llamado $R_{secuencia}$) toma en cuenta la variabilidad de las posiciones individualmente dentro de un alineamiento. Se basa de manera general en la fórmula para medir incertidumbre (creada por Shanon) que es la siguiente: $H = -\sum_{i=1}^{M} Pi \log Pi$ ... (2); donde Pi es la probabilidad de aparición para M posibles símbolos.

La base del logaritmo determina las unidades. Cuando se utiliza base dos, las unidades son los bits. Un bit de información resuelve la incertidumbre para tomar una decisión ente dos símbolos equiprobables. En el caso de secuencias para nucleótidos, M = 4 posibles bases, la decisión para escoger una sola de ellas es de 2 bits (cuando las cuatro bases aparecen con igual probabilidad $Pi = 0.25$).

Cuando las frecuencias de las bases no son exactamente 25, 50 o 100 por ciento. El cálculo de incertidumbre está en función de la frecuencia $f(b,l)$ de cada base b en la posición l:

$$H(L) = -\sum_{b=A}^{T} f(b,l) \log_2 f(b,l) \ldots (3)$$

El contenido informacional (conservación de secuencia) $R(L)$ es entonces:

$$R(L) = 2 - H(L) \quad \ldots . (4)$$

La información para los sitios va siendo aditiva e independiente entre las posiciones. Se puede obtener la conservación total $R(s)$ de una secuencia de pegado al DNA sumando todos los valores de $R(L)$ de las posiciones del sitio.

$$R(s) = \sum_{l} R(L) \quad \ldots . (5)$$

La conservación de un sitio de pegado para una proteína va en relación directa con el tamaño del genoma G y el número de veces que este sitio γ se encuentra dentro del genoma. Esto se define con otra medida que se llama $R_{frecuencia}$ $R(f)$, que se traduce como la información mínima necesaria para encontrar un cierto número de sitios dentro del genoma. La fórmula es:

$$R(f) = \log_2 G - \log_2 \gamma = -\log_2 \frac{\gamma}{G} = -\log_2 f \quad \text{(bits por sitio)} \quad \ldots(6)$$

donde f es la frecuencia de los sitios en el genoma.

III. OBJETIVOS Y PLANTEAMIENTO DE LA INVESTIGACIÓN:

Se plantean dos objetivos centrales. El primero es conocer y comparar los diferentes programas que se utilizan para realizar alineaciones y búsqueda de patrones en secuencias, con el propósito de elegir un protocolo que nos facilite el análisis en las regiones reguladoras. El segundo es aplicar los programas y elaborar estrategias que permitan analizar e integrar propiedades de las secuencias *cis* reguladoras en eucariontes.

III.1 SELECCIÓN DE MÉTODOS:

Existen diferentes programas con los que se hacen alineaciones y búsquedas de secuencias. En nuestro laboratorio se pretende conocer mejor los fundamentos de estos programas, realizar pruebas y elegir los que generen mejores resultados para que formen parte de nuestro protocolo de análisis.

Las pruebas se realizarán con la proteína reguladora Gal4 de *S. Cerevisiae*. La información preliminar se obtendrá de la base de datos TRANSFAC, la cual cuenta con secuencias de pegado para proteínas reguladoras obtenidas a partir de diferentes métodos experimentales: *footprinting reactions, competition, gel retardation* y *direct gel shift* entre otros. Esto hace que los tamaños de los sitios para una misma proteína no sean uniformes. Este es el caso de Gal4.

Partiendo de lo anterior se elabora la siguiente estrategia que permite evaluar los diferentes programas para análisis múltiple de secuencias y búsqueda de patrones.

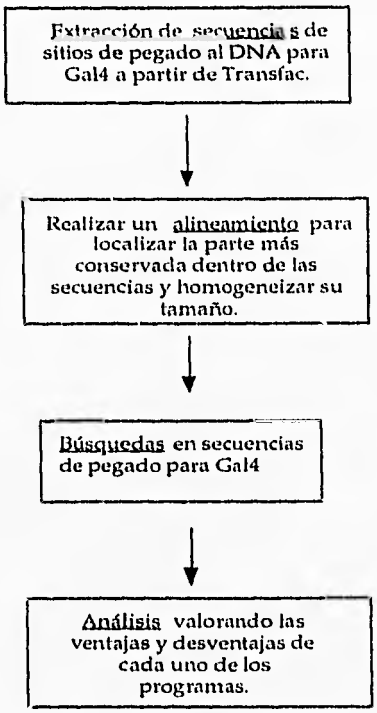


Figura 10. Estrategia a seguir para la evaluación de los diferentes programas utilizados en búsquedas de secuencias.

Extracción de secuencias: se realiza a partir de la base de datos TRANSFAC. Posteriormente con los programas de **alineación** se busca el sitio y longitud precisa donde la proteína se pega al DNA. A partir de la alineación, se utilizan los diferentes programas para realizar **búsquedas** en archivos de secuencias de regiones reguladoras que contienen sitios precisos de pegado para GAL4 y en este punto los programas muestran el grado de reconocimiento de los sitios originales de pegado para la proteína. Finalmente se comparan los resultados obtenidos de los diferentes programas y se **analizan** las diferentes ventajas y desventajas para cada uno de los programas teniendo como parámetros entre otros los siguientes criterios: número total de secuencias encontradas, número de falsos positivos reconocidos, tiempo de ejecución y plataforma en la que corre. De esta manera se escoge un método para utilizarse en los subsiguientes análisis en otras secuencias reguladoras de *S. cerevisiae*.

III.2 APLICACIONES:

Teniendo una metodología seleccionada se muestran ejemplos de posibles aplicación para el analisis de dominios reguladores en los cromosomas de *Saccharomyces cerevisiae*. Se utilizaron dos estrategias diferentes de análisis: la primera, llamada *Top-down* toma la información biológica obtenida de los sitios de pegado de las proteínas reguladoras para relizar búsquedas en las regiones reguladoras con el propósito de hacer predicciones. La segunda estrategia llamada *Bottom-up* utiliza la distribución estadística de cadenas de DNA de tamaño definido en secuencias reguladoras, con el propósito de encontrar algún significado biológico.

Este análisis se llevo a cabo inicialmente en el cromosoma II de *Saccharomyces cerevisiae*. Posteriormente se aplicó la estrategia *Bottom-up* en otros dos cromosomas, el III y XI.

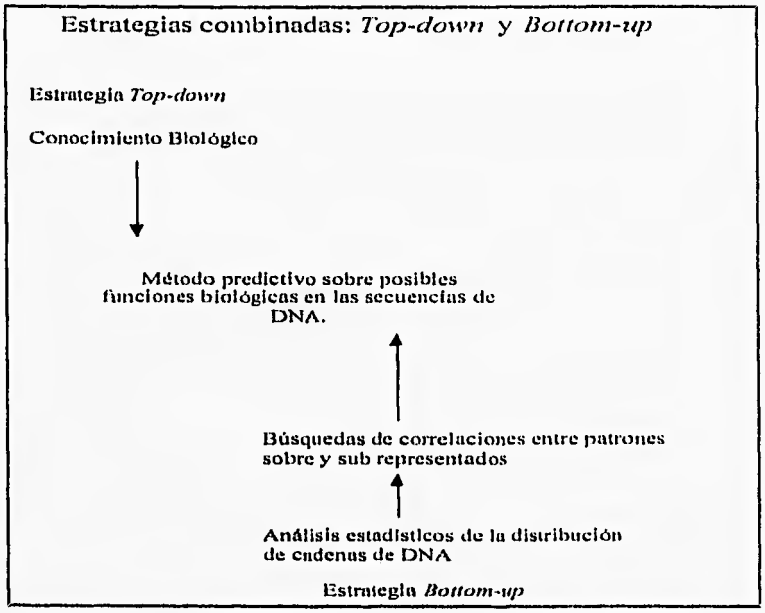


Figura 11. Estrategias *Top down* y *Bottom up* para análisis en regiones reguladoras.

IV. MATERIAL Y MÉTODOS

IV.I ESQUEMA GENERAL:

Para realizar el presente trabajo se extrajeron 10 secuencias de la base de datos Transfac, estas se sometieron a programas de alineación para reconocer las regiones de mayor similitud y caracterizar el tamaño del dominio de pegado en el DNA para Gal4, el cual ya ha sido reportado previamente en la literatura (Giniger *et al.* 1985).

Secuencias obtenidas de Transfac para GAL4

Secuencias (ID's de GENE BANK)	Posición	Sitio
k02115	368-384	CGGATTAGAAGCCGCCG
k02115	387-404	CGGGTGACAGCCCTCCGA
k02115	405-421	AGGAAGACTCTCCTCCG
k02115	469-491	CGCGCCGCACTGCTCCGAACAAT
x03102	206-226	TTCGGCCATATGTCTTCCG
x01667	561-577	CGGCGCACTCTCGCCCG
x00215	653-673	ATACTTCGGAGCACTGTTGAGCG
x00215	740-761	AGCGCTCGGACAACTGTTGACC
m81879	328-350	CACCGGCGGTCTTTCGTCCGTGC
m81879	411-433	TATCGGGGCGGATCACTCCGAAC

Lo primero es hacer todas las secuencias de un mismo tamaño (requerimiento para algunos programas), se extendieron todas las secuencias a 24 nucleótidos de longitud. Algunas se encuentran muy proximas entre sí y con ese tamaño se evita el traslapamiento

Los sitios extendidos se someten a programas dealineación que posteriormente permiten definir un patrón o una matriz. Con esta matriz (o patrón) se realiza la búsqueda en regiones reguladoras que contienen (experimentalmente comprobado) sitios de pegado para Gal4.

Una estrategia habitual en biología computacional es separar dos conjuntos de datos, el conjunto con el que se define el sistema de reconocimiento *"training set"* y el conjunto que sirve para probar el sistema *"testing set"*.Las secuencias de prueba son sitios de pegado para GAL4 que no están incluidas en la base de datos TRANSFAC, y fueron:

Secuencia	Posición	Sitio
m81879	216	CGGAAAGCTTCCTTCCGG
m81879	347	CGGAGATATCTGCGCCGT
m81879	418	CGGATCACTCCGAACCGA

Se utilizaron dos medidas estadísticas que evalúan la eficiencia para reconocer los sitios a lo largo de las secuencias (Adams *et al.* 1995), sensibilidad y especificidad:

$$Sensibilidad = \frac{Verdaderos_{positivas}}{Verdaderos_{positivos} + Falsos_{negativos}}$$

$$Especificidad = \frac{Verdaderos_{negativos}}{Verdaderos_{negativos} + Falsos_{positivos}}$$

En las aplicaciones primero se analizan las regiones reguladoras del cromosoma II.

Se inicia aplicando la estrategia *Top-down*. Se utilizaron 12 sitios de Gal4 definidos experimentalmente donde se une la proteína al DNA.

Con estas secuencias se construye la alineación (con el programa que haya dado el mejor resultado) para obtener una matriz o un patrón consenso el cual permite hacer una búsqueda computacional dentro de las regiones reguladoras del Cromosoma II, localizando los genes que posiblemente sean regulados por esta proteína.

Los diferentes métodos de búsqueda escogen una secuencia consenso o matriz tomando las diferentes posiciones de manera independiente. Los sitios de pegado de algunas proteínas conservan propiedades biológicas, como pueden ser simetría o correspondencia entre nucleótidos adyacentes.

A los resultados obtenidos se les aplican dos filtros, basados en algunas de las propiedades que se han observado en los sitios de las secuencias originales de pegado para GAL4:

Propiedad A: Los seis nucleótidos ubicados en los sitios extremos de la alineación (tres del lado izquierdo y tres del derecho), son los más conservados y sólo presentan un cambio en alguno de los seis nucleótidos, como se observa en las siguientes secuencias:

Acc. No.	Secuencia	Nu. de nucleótidos conservados Izquierda / Derecha
x03102-1	\ CGGCCATATGTCTTCCG \	3/3
x01667-1	\ CGGCGCACTCTCGCCCG \	3/3
x00215-2	\ CGGACAACTGTTGACCG \	3/3
x00215-1	\ CGGAGCACTGTTGAGCG \	3/2
m81879-4	\ CGGGGCGGATCACTCCG \	3/3
m81879-2	\ CGGCGGTCTTTCGTCCG \	3/3
m81879-3	\ CGGAGATATCTGCGCCG \	3/3
m81879-1	\ CGGAAAGCTTCCTTCCG \	3/3
k02115-4	\ CGCGCCGCACTGCTCCG \	2/3
k02115-3	\ AGGAAGACTCTCCTCCG \	2/3
k02115-2	\ CGGGTGACAGCCCTCCG \	3/3
k02115-1	\ CGGATTAGAAGCCGCCG \	3/3

Propiedad B: La secuencia que tiene el cambio del sitio conservado en el extremo derecho es transcrito en la secuencia complementaria del cromosoma, mientras que las dos secuencias que tienen el cambio en el extremo conservado izquierdo son transcritos en la cadena 5' - 3'.

El análisis Top-down termina comparando el número y calificación de los sitios de Gal4 encontrados en las regiones codificadora vs los candidatos encontrados en las regiones reguladoras. El evaluar las posibles diferencias entre ambas regiones sirve como control .

El tamaño para cada una de las regiones reguladoras es de 500 pares de bases. En regiones codificadoras se tomaron 500 nucleótidos a partir del +50 de los **CDS** (coding domain sequence) y en regiones reguladoras se toman 500 a partir del -500 de los CDS.

Para la estrategia *Bottom-up* se analiza estadísticamente la composición de la secuencia contenida en el cromosoma II de *S. cerevisiae*. Para ello se obtiene un archivo de todos los posibles 3-11n-3 meros (donde 11 es un espacio donde puede ir cualquier base) que se pueden crear a partir del alfabeto de 4 nucleótidos (A, T, G y C). Se escogió este patrón porque está muy relacionado con el sitio de pegado para la proteína Gal4. Si se escoge otro patrón, como puede ser el de todos los posibles 17-meros, sería un archivo inmensamente grande (272 Gb), mientras que el de 3-11-3 meros, ocupa 109 kb, ejemplifica la parte más conservada del sitio de pegado (las partes extremas en ambos lados) y los 11 nucleótidos la parte media con mayor variabilidad.

Después se utiliza un programa que calcula el número de posibles apariciones para cada uno de los patrones de 3-11-3 n-meros tomando en cuenta la composición y la probabilidad de aparición para cada uno de los nucleotidos en las regiones reguladoras y codificadoras del cromosoma II. La composición y probabilidad de aparición es la siguiente:

Cromosoma II:

Regiones-reguladoras: 416
#Total de Nucleótidos(cadena-sencilla):207 977.00

# de contenido de:		Probabilidad en las regiones reguladoras:
A	66072	0.304189
T	61784	0.304403
G	40262	0.178818
C	39859	0.176476

Regiones-codificadoras : 416
#Total de Nucleótidos (cadena-sencilla):208 000.00

# de contenido de:		Probabilidad en las regiones codificadoras:
A	66545	0. 319927
T	57912	0. 278423
G	42763	0. 205591
C	40780	0. 196057

La fórmula que se utiliza para calcular la frecuencia esperada de cada uno de los patrones es la siguiente:

$$\boldsymbol{P}(x) = [1 - (1 - p)^{a}]n \quad \ldots(1)$$

Donde **p** representa la probabilidad de aparición de un sitio. **(1 - p)** es la probabilidad de la negación de p, $(1 - p)^{a}$ es la probabilidad de no encontrar un sitio en una secuencia de tamaño **a**; entonces la negación es la probabilidad de encontrarlo al menos una vez en una región reguladora, $[1 - (1 - p)^{a}]$. Al multiplicar por el número total de regiones reguladoras obtenemos un solo valor para cada n-mero sobre la colección de regiones reguladoras en el cromosoma. ***P(x)*** es una función que se define con las probabilidades dadas para cada uno de los nucleótidos.

El valor de "***P observada***", se obtiene simplemente contando el número de regiones reguladoras en los que el n-mero ocurre al menos una vez. Este conteo se lleva a cabo en un programa utilizando el comando'grep'.

Posteriormente se calcula y gráfica el valor del cociente de la frecuencia calculada entre el de la frecuencia observada. Esto da una idea global sobre la distribución de los patrones, poniendo una especial atención en aquellos que pudieran tener alguna desviación en su distribución, ya sea como subrepresentados (valor mayor a uno) o sobre representados (valor menor a uno). Esto se realizarán en las regiones reguladoras y en las codificadoras.

Concatenación de las estrategias: todos los posibles n-meros serán calificados por la matriz de búsqueda de GAL4 que ha sido modificada, extrayendo aquellos que tengan una calificación suficiente para ser posibles sitios de pegado.

Después se observa la distribución de estos n-meros en regiones reguladoras y codificadoras, para encontrar alguna relación entre la función y la ocurrencia en el DNA.

Esta concatenación finaliza con la comparación de distribuciones de frecuencia entre regiones codificadoras y reguladoras de los 3-11-3 meros reconocidos por la matriz de Gal4.

Se elaboró tambien una estrategia *Bottom-up* para los cromosomas III y XI, cuya composición de nucleótidos es la siguiente:

Regiones-reguladoras en cromosoma III: 215
#Total de Nucleótidos(cadena-sencilla): 315 338.00

# de contenido de:		Probabilidad en las regiones reguladoras:
A	98164	0.31
T	95603	0.30
G	59440	0.19
C	62131	0.20

Regiones-reguladoras en cromosoma XI: 324
#Total de Nucleótidos (cadena-sencilla): 646 449.00

# de contenido de:		Probabilidad en las regiones codificadoras:
A	200353	0. 31
T	199930	0. 31
G	122386	0. 19
C	123779	0. 19

El tamaño de cada región reguladora es de 500 pares de bases, se toman a partir del -500 de cada CDS (igual que en el cromosoma II).

Se calculó la frecuencia de 5-n-5meros, 6-n-6meros y 8-n-8meros. Donde n es la distancia de separación. Esta tomó valores 4, 5, (para que el complemento estuviera fuera de fase) 9 y 10 (para encontra al complemento en fase).

Se calculó su distribución esperada y se dividió entre la observada (al igual que con todos los 3n-11-3n meros).

Con este análisis se buscan patrones sobre y subrepresentados que tengan relación con sitios de pegado de proteínas reguladoras. La búsqueda se realiza con los patrones simples permitiendo una variación de alguna base dentro de la cadena (*mismatches*).

Posteriormente se continuó con un análisis a partir de la distribución de las cadenas de 5-n-5 meros. Básicamente se cuantificó las cadenas con simetría directa e invertida que estuvieron sobre y subrepresentados en sus diferentes distancias. Para observar las posibles diferencias en la distribución de los n-meros con respecto a su fase de pegado.

IV.2 DESCRIPCIÓN GENERAL DE LOS PROGRAMAS Y SU UTILIZACIÓN.

Para realizar el análisis de las secuencias se tiene que seguir una sucesión ordenada en la utilización de los programas (Figura 12). Entre algunos de ellos se observan conexiones directas. Los podemos clasificar en tres grupos:

-Programas de alineación;
-Programas de Desplegamiento y visualización de alineación;
-Programas de Búsqueda.

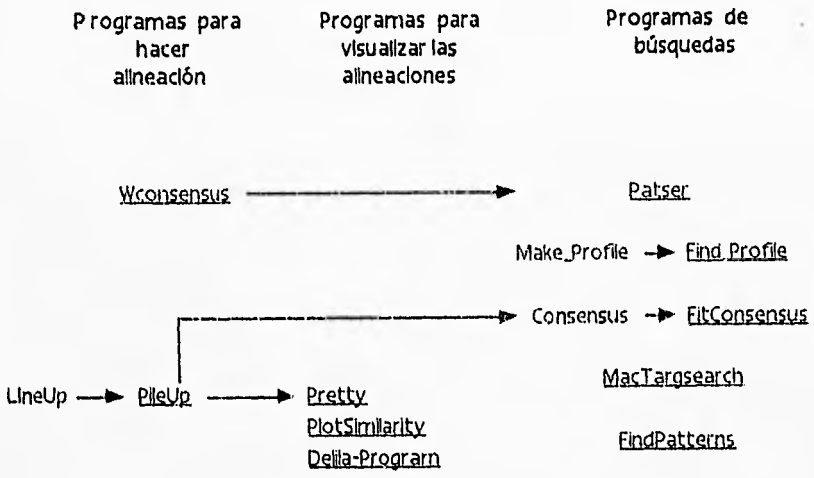


Figura 12 Esquema general de los programas utilizados para realizar el análisis de secuencias

En la figura 12, se muestra la secuencia para el uso de programas.

Los programas de alineación son dos Wconsensus y PileUp. Wconsensus despliega automáticamente el alineación y crea una matriz de búsqueda que será utilizada de manera directa por el programa Patser. PileUp utiliza las secuencias editadas en el programa LineUp, realiza el alineación y para desplegarlo puede utilizar tres programas: Pretty, PlotSimilarity (que utilizan directamente el archivo de salida de PileUp) y Delila..

Los programas de búsqueda de patrones en secuencias son: Patser, Find_Profile, FitConsensus, MacTargsearch, FindPatterns. Los cuatro primeros programas utilizan matrices para realizar las búsqueda, mientras que el último utiliza secuencias consensos. En algunos programas se pueden editar las matrices y ser utilizadas directamente, como MacTargsearch y Patser mientras que en otros se necesita de un programa intermedio que construya la matriz y con el archivo de salida correr los programas de búsqueda. Tal es el caso de Find_Profile (que utiliza a Make_Profile) y FitConsensus (que utiliza a Consensus).

En la tabla 1 se muestra una descripción general de los programas, sus archivos de entrada y salida, sus fuentes y referencias.

Los detalles de cómo fueron utilizados se incluyen en el apéndice A.

Alineamiento	LineUp	Editado por el programa LineUp.	Las secuencias alineada	Crea una alineación múltiple de secuencias a través de comparar pares de secuencias progresivamente.	Paquetería de GCG. Feng y Doolitle, 1987.
	Wconsensus	Secuencias desalineadas del mismo tamaño.	Presenta las secuencias alineadas y una matriz de peso.	Realiza alineaciones locales múltiples, construye matrices de frecuencia para representar un patrón consenso y calcula su Contenido Informacional. No se necesita definir el tamaño de la alineación.	University Colorado. Stormo, 1990 y Hertz, et al. 1990.
Búsquedas	MacTargetsearch	Una matriz de búsqueda.	Muestra los sitios y su posición con los valores de similitud más altos.	Utiliza una matriz que evalua el valor de similitud para calificar todos los posibles patrones a lo largo de la secuencia.	University of Pennsylvania School of Medicine. Goodrich, J.A. et. al. 1990.
	FitConsensus	Una matriz de porcentajes creada por el programa	Presenta la posición de los sitios con mayor porcentaje de	Utiliza una matriz de porcentajes que califica los sitios a los largo de la secuencia.	Paquetería de GCG Mulligan, et al. 1984

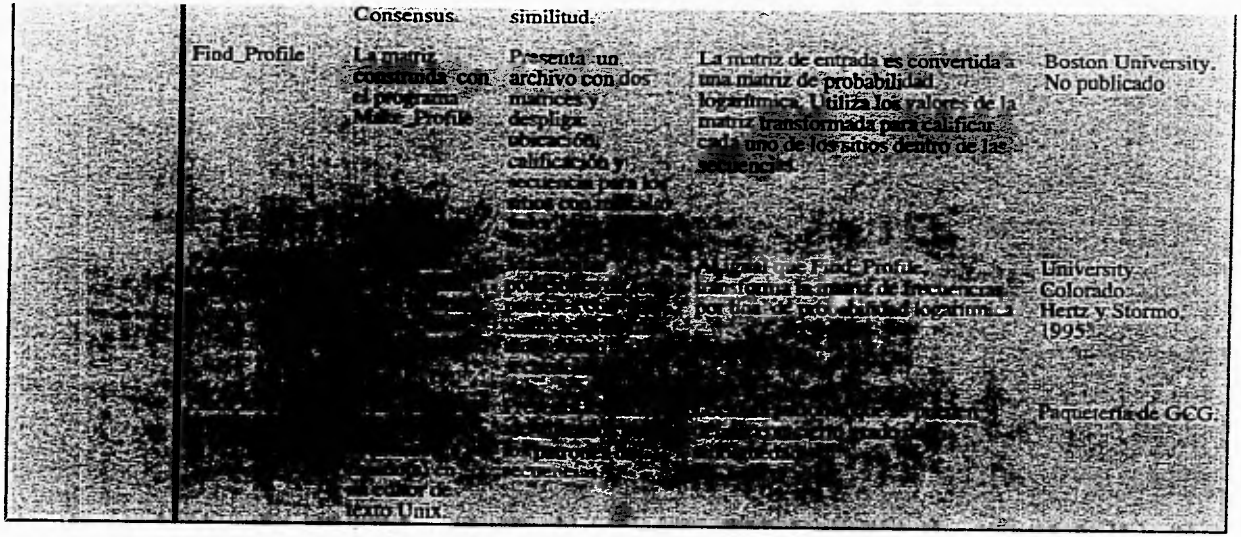


Tabla 1. Descripción general para los tres tipos de programas: Alineamiento, visualización y búsqueda.

V. RESULTADOS Y DISCUSIONES:

V.1 COMPARACIÓN ENTRE LOS ALGORITMOS.

Para el análisis y comparación de los diferentes programas se divide el trabajo en dos partes, en la primera se incluyen los programas que realizan alineaciones y en la segunda los que realizan búsquedas de patrones en archivos de secuencias.

a) Programas para hacer alineaciones múltiples:
Todos utilizaron como prueba las 10 secuencias obtenidas de la base de datos TRANSFAC

.

Se inicia con la paquetería de GCG, utilizando el editor de texto **lineup** y posteriormente **PileUp** (en este ultimo crea un archivo múltiple "msf").

Este archivo fue a su vez de archivo de entrada para el programa **PRETTY** :

```
                      1                                   31
pileup.msf{m81879-2}  ....tATCGG gGCggatcaC TCCGAACc.. .
pileup.msf{k02115-4}  ......TCGc gcCgCacTgC TCCGAACaat .
pileup.msf{k02115-2}  ....GAgCGG gtgACaGcCC TCCGAAgG.. .
pileup.msf{k02115-1}  ..aaGtaCGG AttAgaagCC gCCGAg.... .
pileup.msf{x01667-1}  ...ttAcCGG cGCACTcTCg cCCGAAC... .
pileup.msf{m81879-1}  ....cAcCGG cGgtCTtTCg TCCGtgCG.. .
pileup.msf{k02115-3}  .......aGG AagACTcTCC TCCGtgCGtc c
pileup.msf{x00215-2}  .AgcGcTCGG AcaACTGTtg aCCGt..... .
pileup.msf{x00215-1}  tAtactTCGG AGCACTGTtg agCG...... .
pileup.msf{x03102-1}  ...catTCGG ccataTGTCt TCCGAAa... .
           Consensus  -A--GATCGG AGCACTGTCC TCCGAACG-- -
```

Con Pretty se desplegó el consenso del alineación múltiple, y a partir de éste se seleccionan los sitios dentro de la secuencia que serán depurados "rasurados". El rasurameinto es importante porque nos señala la región más conservada dentro del alineación; el proceso se puede realizar empíricamente, restringiéndose sólo a las posiciones en donde las secuencias contribuyen con al menos una letra. Para este caso sería entre las posiciones 8 y 24 (señalados con las flechas) en donde se podría hacer el rasuramiento.

Otra manera de hacer el rasurado de secuencias es calculando el grado de conservación de las bases a lo largo del de la alineación, y se restringe entre los valores con mayor grado de similitud. Para esto en GCG se cuenta con el programa **Plotsimilarity** que calcula la similitud en cada uno de los sitios, utiliza como archivo de entrada el archivo MSF creado en PileUp (Figura 13)

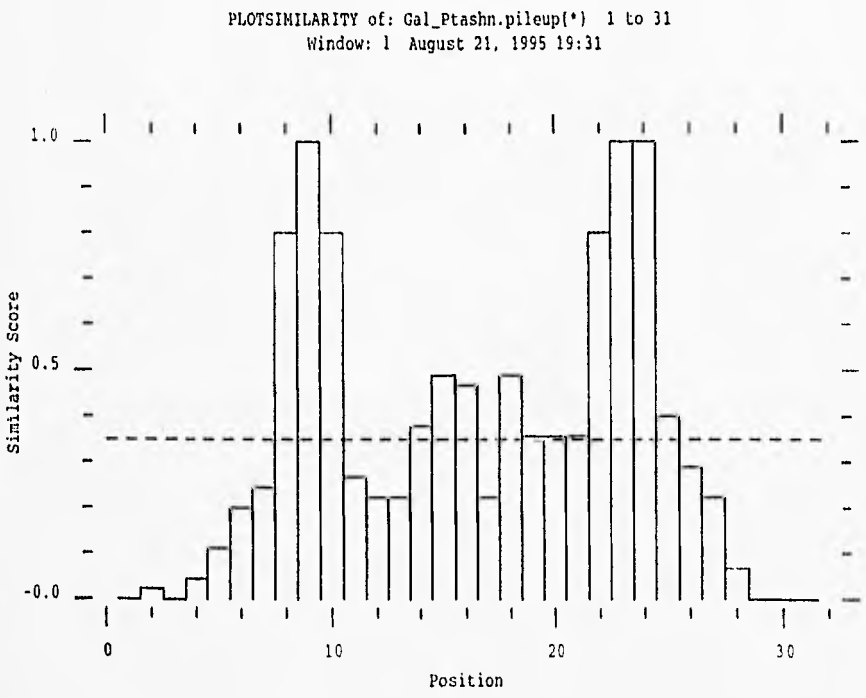


Figura 13: Desplegamiento de un histograma de frecuencia creado con el programa PLOTSIMILARITY, generado a partir de la alineación múltiple de los sitios de pegado para Gal4.

Sequence Logo

Generated by http://www.bio.cam.ac.uk/seqlogo/logo.cgi

Computed using alpro and makelogo (Schneider & Stephens, 1990)

Figura 14: Visualización del contenido informacional en cada una de las posiciones a lo largo de la alineación de los elementos *cis* de Gal4.

La figura 13 presenta los diferentes valores de similitud (Similarity Score), y se observa que los valores más altos se tienen entre el nucleótido 8 y 24 dando un buen punto de corte entre estos puntos.

También se utilizó otro programa que se encuentra en Internet *World Wide Web* con el nombre de "Delila". El objetivo es construir una secuencia Logo, que visualiza -a través de gráficas- la información contenida en un fragmento de DNA, RNA o proteína (Shaner, M.C. *et al.* 1993). A partir de una alineación, el programa Delila grafica la conservación de los nucleótidos para cada posición. La base (o nucleótidos) más conservada se representará en la parte superior del archivo de salida, comúnmente llamado "Logo" con una altura mayor. La conservación para cada nucleótido se mide, (utilizando la Teoría de la Información desarrollada por Claude Shanon) en bits de información. Entonces la escala para el archivo logo corre de cero a dos bits.

El 'logo' de Gal4 confirma la conservación de la alineación en los sitios que se escogieron para el corte. (Figura 14).

Wconsensus: Para este programa se utilizó como archivo de entrada las mismas 10 secuencias extendidas a 24 nucleótidos de longitud:

```
x03102-1 \ CATTCGGCCATATGTCTTCCGAAA \
x01667-1 \ TTACCGGCGCACTCTCGCCCGAAC \
x00215-2 \ AGCGCTCGGACAACTGTTGACCGT \
x00215-1 \ TATACTTCGGAGCACTGTTGAGCG \
m81879-2 \ TATCGGGGCGGATCACTCCGAACC \
m81879-1 \ CACCGGCGGTCTTTCGTCCGTGCG \
k02115-4 \ TCGCGCCGCACTGCTCCGAACAAT \
k02115-3 \ AGGAAGACTCTCCTCCGTGCGTCC \
k02115-2 \ GAGCGGGTGACAGCCCTCCGAAGG \
k02115-1 \ AAGTACGGATTAGAAGCCGCCGAG \
```

Como parámetros se le dio que en el primer ciclo guardará las 100 mejores pares de secuencias alineadas (sin inversas). El programa se ejecutó con cuatros valores distintos de desviación estándar: 0.5, 1.0, 1.5 y 2.0 (recomendados por G. Hertz) y los cuatro dieron como mejor resultado (el de menor frecuencia esperada) la misma alineación, que es el siguiente (Figura 15):

```
MATRIZ 1
número de secuencias = 10
longitud = 17
contenido de información = 16.324
ln(probabilidad) = -71.2028   probabilidad = 1.19399E-31
frecuencia esperada = 1.28204E-22
A |   1   0   0   4   1   2   6   1   4   1   0   1   0   2   0
0   0
C |   9   0   1   3   3   4   0   7   0   3   2   6   5   1   9
10   0
G |   0  10   9   3   4   3   2   2   0   4   1   1   4   1   1
0  10
T |   0   0   0   0   2   1   2   0   6   2   7   2   1   6   0
0   0
  1|6   :   1/5     CGGCCATATGTCTTCCG
  2|3   :   2/5     CGGCGCACTCTCGCCCG
  3|1   :   3/7     CGGACAACTGTTGACCG
  4|7   :   4/8     CGGAGCACTGTTGAGCG
  5|10  :   5/4     CGGGGCGGATCACTCCG
  6|2   :   6/4     CGGCGGTCTTTCGTCCG
  7|8   :   7/2     CGCGCCGCACTGCTCCG
  8|4   :   8/1     AGGAAGACTCTCCTCCG
  9|5   :   9/4     CGGGTGACAGCCCTCCG
 10|9   :  10/6     CGGATTAGAAGCCGCCG
```

Figura 15 Archivo de salida del programa Wconsensus. Donde se muestra: los parámetros bajo los cuales se corrió el programa, la matriz con su contenido informacional y el despliegue de la alineación.

Los programas Wconsensus y PileUp dieron como resultado la misma alineación, pero Wconsensus presenta ventajas para encontrar el patrón de secuencia común en el pegado de proteínas. Wconsensus genera alineaciones locales múltiples, su objetivo es construir una matriz que describa el patrón más significativo que comparten un grupo de secuencias relacionadas funcionalmente. Además con Wconsensus no se tiene que hacer ningún rasuramiento manual en las secuencias y calcula automáticamente el contenido informacional para cada matriz. En el caso de que el sitio sea simétrico el programa tiene la opción de incluir automáticamente las cadenas complementarias enriqueciendo a la alineación.

En GCG se utilizan al menos 2 ó 3 programas sucesivamente (consume más tiempo) para realizar una sóla alineación de tipo global múltiple. El objetivo en esta alineación será optimizar al máximo la similitud a lo largo de toda la longitud de las secuencias (insertando huecos), esto se utiliza mucho para encontrar relaciones evolutivas entre proteínas o secuencias codificadoras de DNA. En PileUp la inserción de huecos intersecuenciales puede resultar contraproducente para el alineamiento local (se busca sólo un dominio común no maximizar el parecido entre las secuencias); para evitarlos, es necesario castigar los huecos con valores muy altos; con ello se ayuda a identificar mejor los dominios conservados, se necesita

también rasurar manualmente a las secuencias y posteriormente editarlas (procurando que sean del mismo tamaño). En las alineaciones resultantes es muy útil calificar la similitud en los sitios para tener una idea de la calidad de la alineación. Para ello se requiere usar el programa Delila o PlotSimilarity.

b) Programas de Búsqueda:

A partir de los alineaciones se construyeron patrones y matrices para búsquedas en diferentes archivos de secuencias. Se utilizaron varios programas que se probaron en cinco regiones reguladoras de genebank **(ID de genebank)** en donde se realizaron búsquedas de sitios de pegado para Gal4 (los cinco ID's contienen sitios para su pegado).

Primero se comparan los cuatro programas que utilizan matrices de búsqueda para encontrar patrones. Los corrimientos completos de los programas se presentan en el apéndice B y los resultados a comparar se presentan en las siguientes cinco tablas (una para cada ID). En aquellos programas que no hacen búsqueda automática en la secuencia complementaria, se tuvo que utilizar el programa **'Reverse'** de GCG, que obtiene, a partir de un archivo de secuencias, la cadena complementaria invertida. Teniendo esta cadena se realiza nuevamente la búsqueda en el archivo complementario-invertido y se recalculan las posiciones.

A continuación se presentan cinco tablas con el resultado de las posiciones y calificaciones de los sitios que los programas reconocieron como posibles sitios de pegado para Gal4. Los programas que se compararon fueron: MacTargsearch, FitConsensus, FindProfile y Patser.

Tabla 2: X03102. Este archivo contiene 2812 pares de bases, y contiene la región reguladora y codificadora para una proteína alpha-galactosidasa necesaria para el metabolismo de melidosa.

	Fit Consensus			**Patser**	
	Posición	**- Score**		**Posición**	**- Score**
	208	- 57.65		208	- 18.45
	208c	- 53.53		208c	- 15.48
	691c	- 52.35		691c	- 14.96
	691	- 51.76		691	- 14.90
				1095	- 14.10

Tabla 3:X01667. Tiene 2475 pares de bases, contiene la región reguladora y codificadora para la proteína reguladora Gal80. funciona para reprimir la transcripción de los genes activados por Gal4.

Mac [illegible]	Fit Consensus	Find Profile	Patser
Posición - Score	Posición - Score	Posición - Score	Posición - Score
[illegible]	561 - 63.53	[illegible]	561 - 21.92
[illegible]	561c - 56.47	[illegible]	561c - 18.15

Tabla 4: X00215. Contiene 1008 pares de bases, y en esta región se encuentra el inicio de la transcripción para la proteína Gal7. Este archivo pertenece a la secuencia del Cromosoma II.

Mac [illegible]	Fit Consensus	Find Profile	Patser
Posición - Score	Posición - Score	Posición - Score	Posición - Score
[illegible]	746 - 61.18	[illegible]	746 - 20.87
[illegible]	746c - 60.08	[illegible]	659 - 19.61
[illegible]	659 - 58.24	[illegible]	746c - 19.28
[illegible]	659c - 55.29	[illegible]	659c - 16.64
[illegible]	753c - 52.35	[illegible]	753c - 13.82

Tabla 5:M81879. Son 2845 pares de bases en el archivo; se encuentra ubicado dentro del brazo derecho del cromosoma XII, contiene la región reguladora y codificadora para el gen Gal2 (permeasa para galactosa).

[illegible]	Fit Consensus	Find Profile	Patser
Posición - Score	Posición - Score	Posición - Score	Posición - Score
[illegible]	331 - 61.18	[illegible]	331 - 21.77
[illegible]	217 - 56.47	[illegible]	217 - 17.62
[illegible]	350c - 55.88	[illegible]	350c - 17.45
[illegible]	331c - 55.29	[illegible]	414 - 17.23
[illegible]	419 - 55.29	[illegible]	350 - 16.74
[illegible]	350 - 54.71	[illegible]	331c - 16.70
[illegible]	414 - 54.12	[illegible]	419 - 16.55
[illegible]	419c - 52.35	[illegible]	217c - 15.06
[illegible]	414c - 51.18	[illegible]	414c - 14.74
[illegible]	217c - 51.18	[illegible]	419c - 14.06

Tabla 6: K02115. Contiene 907 pares de bases, incluye la región promotora para los genes Gal1-Gal10. Se encuentra dentro del cromosoma II.

Fit Consensus		Patser	
Posición	Score	Posición	Score
387	61.76	387	21.35
405	60.59	405	20.35
387c	59.41	387c	18.70
469	55.29	469	17.83
368	52.94	405c	17.37
368c	52.94	469c	15.61
469c	51.18	368	15.55
		368c	13.47

En total los cinco archivos de secuencias tienen 10047 pares de bases. La búsqueda se realizó en ambas direcciones: 5'- 3' y 3'- 5' en 20094 nucleótidos. Dentro de las secuencias hay 26 sitios reales para el pegado de Gal4. Todos los sitios que se encuentran en **negritas** dentro de la tabla son falsos positivos.

De manera general los algoritmos que tienen el orden de calificación más parecido entre sí fueron Find_Profile y Patser. Ambos utilizan el Contenido Informacional para la evaluación de sitios.

En seguida se presenta una tabla de toma en cuenta los resultados evaluatorios globales para cada uno de los cuatro programas que utilizan matriz de peso como algoritmo de búsqueda (tabla 7).

TABLA DE CRITERIOS COMPARATIVOS ENTRE PROGRAMAS DE BÚSQUEDA

CRITERIOS	MacTargsearch 2.0	GCG 8.1 FitConsensus	PROFILE	Matrix Programs
Sensibilidad	1.0	[illegible]	1.0	1.0
E	0.99	[illegible]	0.99	[illegible]
D	Forward & Reverse	[illegible]	Forward	[illegible]
T	4.3 minutos Macintosh. Power Macintosh 6100/60.	[illegible]	0.06 UNIX. SPARC 1000.	[illegible] UNIX. SPARC 1000.
A	s/programa de alineación	[illegible]	s/programa de alineación	[illegible]
[illegible]	Porcentaje de ocurrencias.	[illegible]	# de Frecuencias esperadas	[illegible] Frecuencias esperadas
R	Score Similarity	[illegible]	Contenido Informacional	Contenido Informacional

Tabla 7. Evaluación con algunos criterios comparativos de ejecución, para programas que utilizan matrices de búsqueda para encontrar sitios de pegado al DNA de proteínas reguladoras. La sensibilidad y especificidad se calculó con la fórmula presentada en material y método (ver páginas 25 y 26). Los programas que no incluyen el tiempo de ejecución, se calculó con el comando en UNIX: time.

Para el programa "FindPatterns" (el programa que realiza las búsquedas a partir de una secuencia consenso) se construyó un archivo con dos patrones para realizar la búsqueda; estos patrones se hicieron con el programa Consensus a un 75 y 100 por ciento de certidumbre:

Archivo de búsqueda:

Patrones:
CGGVBVRSWBYYSWCCG
MGSVNNDVWNBNBNSCG

El archivo de salida reconoció, con el patrón de 100% de certidumbre, dos de los trece sitios que debería reconocer y con el patrón de 75 % de certidumbre, todos los sitios para Gal4, faltándole sólo por reconocer el sitio 419 y su complementaria de la secuencia m81879. Reconoció dos sitios (con sus respectivos complementarios) como falsos positivos: 456 del ID K02115 y 755 del ID X01667, estos falsos positivos son totalmente diferemes a los reconocidos por los otros programas.

Cuando se estudia los sitios de pegado en el DNA o RNA, la convención práctica es : realizar una alineación de secuencias con los sitios de pegado para una proteína (o macromolécula reconocedora), escoger la base más común para cada posición y crear una secuencia consenso. La mayoría de los autores, mencionan que las secuencias consenso (Schneider, T. D. *et al.* 1986; Hawley D. K. y W.R. McClure, 1983) no son muy confiables al tratar de buscar nuevos sitios debido esencialmente a que se pierde información cuando la frecuencia relativa de una base específica en una posición es ignorada. En una secuencia de pegado al DNA existen diferentes grados de conservación por lo que es muy importante conocer la variabilidad permitida dentro de las secuencias con la que se mantiene la funcionalidad del sitio.

El programa **FindPatterns** es el único programa de búsqueda de los presentados en este trabajo que utiliza una secuencia consenso en lugar de una matriz. No reconoció a una de las trece secuencias de prueba en dirección 5' - 3' ni a su secuencia complementaria (FitConsensus tampoco reconoció una secuencia al buscar en la cadena complementaria, pero fue una secuencia diferente). La búsqueda de los patrones fue buena como en otros programas (0.92 de sensibilidad y 0.99 de especificidad) pero es importante hacer la secuencia consenso con flexibilidad y lo suficientemente representativa del sitio. Se puede variar su certidumbre de manera que se reconozcan un mayor número de secuencias.

Otra diferencias con respecto a los otros programas fue las secuencias falsas positivas que reconoció. Los patrones de esas secuencias no fueron reconocidos por ningún otro programa. Pero definitivamente la desventaja más clara de la utilización de las secuencias consenso es la pérdida de información para la calidad del reconocimiento en los diferentes sitios de pegado,

las macromoléculas que se unen al DNA van a tener mayor afinidad por algunos sitios y esta información normalmente se pierde al tratar con secuencias consensos.

En los programas que utilizaron matrices, el que tuvo un menor número de falsos positivos fue **MacTargsearch**, (identificando a todos los sitios de pegado y sólo tres falsos positivos) . Pero en otros criterios (como tiempo, carecer de programa de alineamiento) lo hace estar en desventaja con respecto a los otros programas. Al comparar el orden de calificación en las diferentes posiciones, con el programa que tuvo una mayor semejanza fue con **FitConsensus.**

FitConsensus fue el único programa con una matriz de búsqueda que no reconoció la cadena complementaria de una secuencia de pegado. Consume más tiempo que el resto de los programas y además como en casi todos los programas en GCG el formato no es estándar, por lo que para cada algoritmo se construye un archivo de entrada diferente o se reformatean las alineaciones creados en el archivo múltiple.

El paquete **Profile** es muy rápido, el tiempo que tarda en correr en CPU es de 0.35 minutos (el tiempo lo calcula el comando 'time' de unix). Los archivos múltiples para búsquedas deben estar en formato ig y no tiene problemas de edición. Su búsqueda es sólo en una dirección y el archivo de salida despliega los patrones de secuencias encontradas. Las calificaciones tienden a redondearse a sólo dos números, por lo que en secuencias mayores de dos dígitos esto puede crear diferencias grandes, por ejemplo, una calificación de 12.51 se redondea a 13 y uno de 12.49 se redondea a 12.

Patser y **Profile** calculan el contenido informacional. Las secuencias, presentan el mismos orden (tomando en cuenta las calificaciones) y dan valores muy similares, y al revisar la dirección complementaria amplía su intervalo de reconocimiento e incluye como sitios de pegados a cuatro secuencias desconocidas. Patser hace un ajuste en el cálculo (que depende de el tamaño y número de las secuencias que participan en la alineación) y presenta calificaciones más altas para los sitios.

Los programas **Wconsensus** junto con **Patser**, de manera general presentan ventajas con respecto a los otros programas: lo primero es que utiliza las matrices como algoritmo de búsqueda y son mejores que las alineaciones comunes (asigna diferentes valores a las variaciones de la secuencia consenso dependiendo de la conservación para cada nucleótido). Calculan el contenido informacional del sitio y está relacionado con la información necesaria para que una proteína reconozca un sitio de pegado del resto del genoma del organismo, por lo tanto, los valores predictivos pueden compararse con otros datos de la biología de la regulación obtenidos de manera independiente (Hertz *et al.* 1995).

Su velocidad es superior con respecto a los otros programas y un aspecto muy importante a su favor es que en sus parámetros se le designa la frecuencia para cada nucleótido. En el caso de *E.coli* y otros organismos este punto no es muy importante, ya que la frecuencia para cada base es la misma; pero en el caso de *S. cerevisiae* esto representa una modificación importante en el valor calculado para cada sitio (dos terceras partes su genoma son A-T). Es por lo

anterior que los programas a utilizar en las segunda parte del proyecto será la matriz creada con Weonsensus para Gal4 y el programa de búsqueda Patser.

V.2 PREDICCIONES EN EL CROMOSOMA II

La estrategia ***Top-down*** busca en regiones reguladoras del cromosoma II posibles sitios de pegado para la proteína Gal4. Utiliza la matriz construida por el programa Weonsensus para que el programa Patser realice las búsquedas en las 416 regiones reguladoras (de 500 pares de bases cada una) del cromosoma II.

Terminada la búsqueda, se muestra una lista de las regiones reguladoras que tienen sitios de pegado para Gal4. Cada región reguladora está asociada a un CDS (coding domain sequence) con el nombre gen al que corresponde. Al final se editó un archivo de salida con este resultado. (Figura 16):

File containing the matrix: matrix
File containing the sequence information: greg.regions500.2hertz
Print scores greater than or equal to 16.00

***** information for the alphabet from file "alphabet". *****
letter 1: A prior frequency = 0.315586
letter 2: T prior frequency = 0.315808
letter 3: G prior frequency = 0.185518
letter 4: C prior frequency = 0.183088

width of the summary matrix: 17

A |	1	0	0	6	2	4	6	2	4	1	0	1	0	2	0	0	0
T |	0	0	0	0	2	1	3	0	8	3	8	2	2	7	0	0	0
G |	0	12	11	3	5	3	3	2	0	4	1	2	4	2	1	0	12
C |	11	0	1	3	3	4	0	8	0	4	3	7	6	1	11	12	0

Región Reguladora	Posición	Calificación	Gen
80761_81825	310	16.59	*AAR2 MATa1-mRNA splicing factor* c
94377_95159	56	16.03	Homolog to thiol-specific antioxidant c
164627_165039	288	19.83	Probable snRNP-related protein w
188126_189703	153	19.88	ACH1 Acetyl-CoA hydrolase (EC 3.1.2.1) c
211474_214401	280	18.46	PDR3 Pleiotropic drug resistance protein 3 w
242811_245018	106	17.63	c
246568_248211	64	17.05	Probable benomyl/methotrexate resistance protein c
268431_269528	221	21.00	GAL7 Galactose-1-phosphate uridylyltransferase (EC2.7.7.12) c

268431_269528	308	22.25	
270257_272353 c	167	19.98	GAL10 UDP-glucose-4-epimerase (EC 5.1.3.2)
270257_272353	231	22.92	
270257_272353	249	25.62	
273022_274605	47	17.26	GAL1 Galactokinase (EC 2.7.1.6) w
273022_274605	66	25.55	
273022_274605	84	25.64	
273022_274605	148	22.25	
286881_288020	18	16.34	[MRF1] Probable (mitochondrial) ssDNA-binding protein c
288020_288357	355	16.34	
366107_366736	79	16.40	TIP1 Temp. shock-inducible protein precursor SRP1/TIP1 c
446663_447226	69	18.40	
447801_448535	153	16.45	/ Homology to chitin synthase / c
639555_639980	166	16.48	
656507_657308	305	16.74	
659161_660258	175	18.64	PDB1 Pyruvate dehydrogenase (lipoamide), ß chain precursor (EC 1.2.4.1) c
696597_698060	234	17.21	*Probable sugar transport protein c*
698675_699388	135	16.91	*Probable ATP/GTP-binding protein w*
721395_722108	321	17.51	RIB5 Riboflavin synthase a-chain (EC 2.5.1.9) c
724393_726456	120	17.67	/ Homology to serine-type carboxypeptidase PCR1 (S. cerevisiae) / w
754225_754623	262	17.27	/ Homology to human bcr (break-poin cluster) protein / c

Figura 16. Archivo con las búsquedas en regiones reguladoras del Cromosoma II de la proteína Gal4. Se aceptan todos los sitios con una calificación arriba de 16.00 . En la primera columna delimita la ubicación de los CDS dentro del cromosoma (el CDS al que pertenece la región reguladora). La segunda columna presenta la posición de pegado dentro de los 500 nucleótidos de la región reguladora. La tercera columna, la calificación que le dio la matriz de peso al sitio. La cuarta el nombre del gen que regula. El nombre del gen se encuentra asociado a una letra 'c' o 'w', que significa en la cadena que se encuentre el CDS, ya sea en la complementaria 'c', o en la directa.

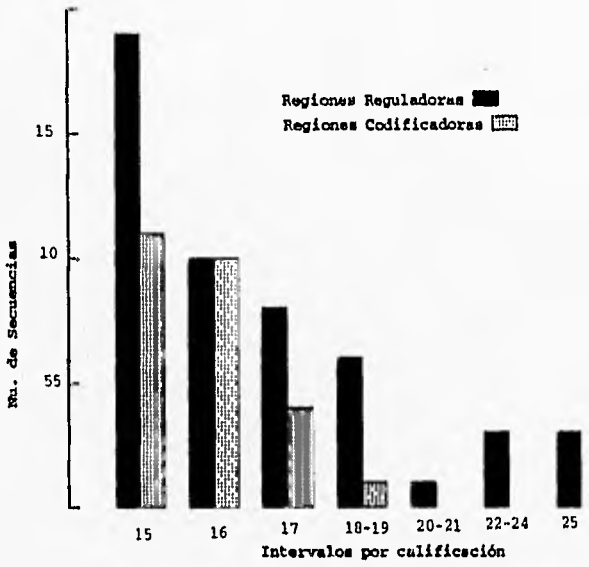
Al analizar la figura 16 con los genes que posiblemente sean regulados por Gal4, aparecen algunas proteínas que podrían estar involucradas en la asimilación de azúcares. Este resultado puede ser interesante ya que Gal4 regula a los genes involucrados en la asimilación de dos fuentes de carbono: galactosa y melidosa.

Posteriormente al archivo de salida, se le aplicaron dos filtros basados en las características de los sitios: conservación de número y lugar de las seis bases en los extremos. Las tres regiones reguladoras que aparecen con letras itálicas no pasaron los filtros biológicos impuesto.

En el cromosoma II se encuentran los genes para tres proteínas (Gal1, Gal10 y Gal7) que son reguladas por Gal4. Las regiones reguladoras de estos tres genes tuvieron sitios de pegado para Gal4 reconocidos con calificaciones muy altas por la matriz.

Después se corrió la matriz en el archivo de regiones codificadoras, con el propósito de hacer comparaciones entre regiones codificadoras y reguladoras y de observar posibles diferencias en número y calificación de los sitios para el pegado de la proteína Gal4.

Se graficó la frecuencia para los intervalos con valores arriba de 15.00 de calificación. En la gráfica comparativa de secuencias codificadoras y reguladoras se observa una clara diferencia de distribución y calificación de sitios, teniendo un intervalo más alto en número y calificación las reguladoras.(Gráfica 1).



GRAFICA 1: Comparación de la frecuencia y calificación de sitios de Gal4. La comparación se hizo entre regiones reguladoras y codificadoras para sitios con una calificación mayor a 15.

Para la estrategia ***Bottom-up*** se lleva a cabo un análisis estadístico y es necesario determinar un patrón específico con el que se analiza las regiones. El patrón fue: 3-11-3 meros (pagina 27 en Material y Método).

Se construyeron dos programas cuyos archivos de salida eran: frecuencia esperada *compute_stats*[1] y frecuencia observada *run_agrep.perl*[1] para todos los posibles 3-11-3 meros de regiones codificadoras y de reguladoras. Para el cálculo de la frecuencia esperada los patrones se consideran como hexámeros.

Ejemplo de archivos en regiones codificadoras del Cromosoma II (Figura 17):

Frecuencia esperada		Frecuencia observada	
AAAAAA	415	AAA.....AAA	407
AAAAAC	415	AAA.....AAC	414
AAAAAG	415	AAA.....AAG	412
AAAAAT	415	AAA.....AAT	412
AAAACA	415	AAA.....ACA	413
AAAACC	413	AAA.....ACC	407
AAAACG	412	AAA.....ACG	408
AAAACT	415	AAA.....ACT	415
AAAAGA	415	AAA.....AGA	415
AAAAGC	412	AAA.....AGC	411
AAAAGG	411	AAA.....AGG	407
AAAAGT	415	AAA.....AGT	409
AAAATA	415	AAA.....ATA	413
AAAATC	415	AAA.....ATC	414
AAAATG	415	AAA.....ATG	414
AAAATT	415	AAA.....ATT	414
AAACAA	415	AAA.....CAA	415
AAACAC	413	AAA.....CAC	409
AAACAG	412	AAA.....CAG	411
AAACAT	415	AAA.....CAT	415
AAACCA	413	AAA.....CCA	408
AAACCC	402	AAA.....CCC	365

Figura 17. Despliegue de una pequeña parte de los archivos de salida de compute_stats y run_agrep.perl. Compute_stats calcula el número de veces que se espera encontrar el archivo de 3-11-3. N es cualquier base por lo que se calcula como hexámeros.

[1] Programas elaborados en colaboración con el laboratorio de Temple F. Smith. BioMolecular Engineering Research Centre, Boston University.

Después se obtuvo el cociente entre el valor esperado y el observado para todas las diferentes secuencias de 6-meros (3-11-3 meros):

Ejemplo de una pequeña parte de el archivo:

```
CCCCCC 1.63
CCCGGG 1.56
GGGCCC 1.46
CCCCCG 1.35
CCCACC 1.34
CCCCGG 1.31
CGGCCC 1.30
CCGCCC 1.30
CCCGCG 1.30
CCCGCC 1.30
```

La distribución de estos cocientes se muestran en las gráficas 2 y 3.

De manera genteral la mayoría de los hexámeros, tanto en regiones codificadoras como en reguladoras, ocurren con la frecuencia esperada (el cociente igual a 1).

Hubo algunos cocientes que muestran valores menores a uno (es decir que son hexámeros **sobre representados**) o valores mayores a uno (hexámeros **sub representados**).

Es **importante analizar si los hexámeros localizados en los extremos de las gráficas tienen algún significado biológico.**

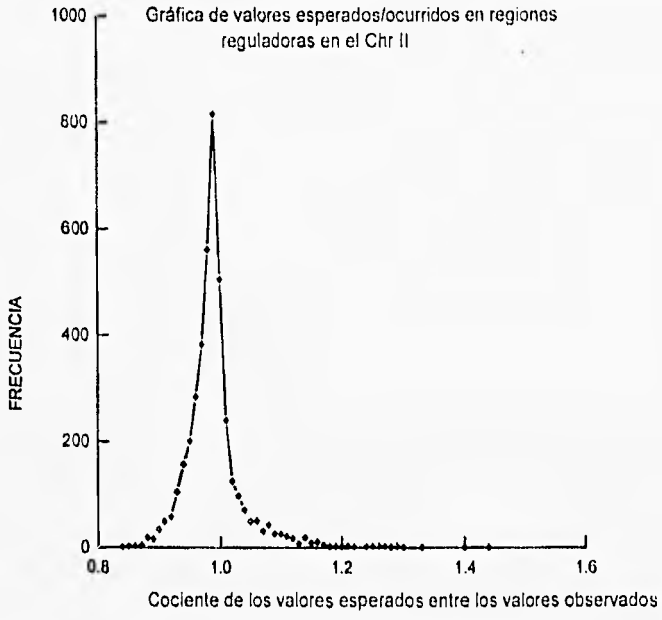
Para ello se realizaron pruebas con la proteína Gal4.

Primero se modificó la matriz de Gal4 para que pudiera calificar los patrones 3-11n-3 meros:

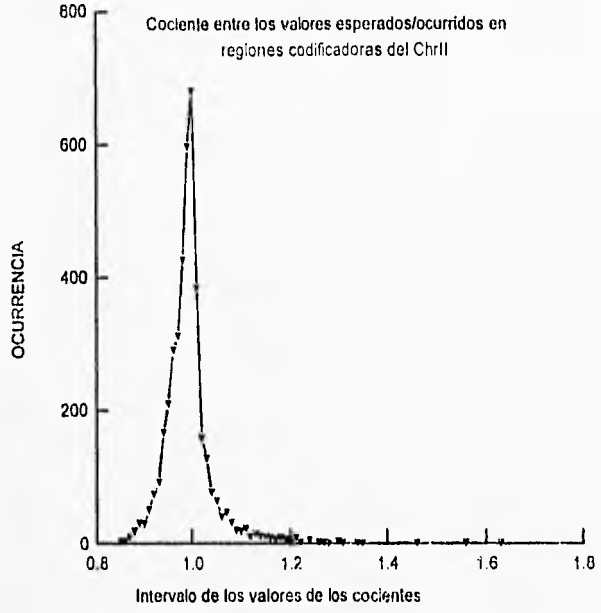
```
A |  1   0   0   3   3   3   3   3   3   3   3   3   3   3   0   0   0
C | 11   0   1   3   3   3   3   3   3   3   3   3   3   3  11  12   0
G |  0  12  11   3   3   3   3   3   3   3   3   3   3   3   1   0  12
T |  0   0   0   3   3   3   3   3   3   3   3   3   3   3   0   0   0
```

después se determina el umbral con el que la nueva matriz, identifica a todas las secuencias de pegado para Gal4 sobre el resto de los archivos a los cuales pertenece, dando el siguiente resultado:

```
x03102 position= 208  score= 14.95
x01667 position= 561  score= 14.95
x00215 position= 659  score= 12.36
```



GRAFICA 2.- Distribución de los valores de los cocientes de las frecuencias esperadas y ocurridas de todos los posibles hexámeros en regiones reguladoras. En el eje de las abscisas se observa que la mayoría de los hexámeros tienen una frecuencia esperada al azar



GRAFICA 3.- Distribución de los valores de los cocientes de las frecuencias esperadas y ocurridas de todos los posibles hexámeros en regiones codificadoras. En el eje de las abscisas se observa que la mayoría de los hexámeros tienen una frecuencia esperada al azar es decir valor igual a uno.

```
x00215 position= 746  score= 14.95
m81879 position= 217  score= 14.95
m81879 position= 331  score= 14.95
m81879 position= 350  score= 14.95
m81879 position= 414  score= 14.95
m81879 position= 419  score= 14.95
k02115 position= 368  score= 14.95
k02115 position= 387  score= 14.95
k02115 position= 405  score= 12.36
k02115 position= 469  score= 12.36
```

Con un umbral de 12.36 reconoce a todos los 3-11n-3 meros funcionales para Gal4.

Con este umbral y la matriz modificada, se ejecuta el programa Patser para reconocer en el archivo todos los 3-11-3 meros que pueden ser sitios para el pegado de Gal4 en las regiones reguladoras y codificadoras del cromosoma II.

Obtenidos los sitios que pueden pasar como posibles sitios de pegado, se busca el valor que obtuvo su cociente (esperados/ocurridos), tanto en regiones reguladoras como en codificadoras (Figura 18):

Regiones Codificadoras		Regiones Reguladoras	
Patrón	Cociente	Patrón	Cociente
CGG.....CCC	1.30	GGG.....CCG	1.29
CCG.....CCG	1.26	CGG.....CGG	1.27
GGG.....CCG	1.16	CGG.....CCC	1.26
CGT.....CCG	1.13	CCG.....CCG	1.21
CGG.....CGG	1.09	CGG.....CCG	1.19
CGC.....CCG	1.09	AGG.....CCG	1.10
CGG.....CCG	1.07	CGT.....CCG	1.08
CGG.....ACG	1.04	CGG.....CCA	1.08
CGG.....TCG	1.03	CGA.....CCG	1.08
CTG.....CCG	1.02	CAG.....CCG	1.08
CGG.....CCT	1.02	CGG.....ACG	1.06
CGA.....CCG	1.02	CGG.....TCG	1.04
CAG.....CCG	1.00	TGG.....CCG	1.03
CGG.....CAG	0.98	CGG.....GCG	1.00
AGG.....CCG	0.98	CGC.....CCG	1.00
CGG.....GCG	0.97	CGG.....CCT	0.98
CGG.....CCA	0.97	CTG.....CCG	0.97
TGG.....CCG	0.91	CGG.....CTG	0.97
CGG.....CTG	0.88	CGG.....CAG	0.94

Figura 18. Muestra los cocientes, tanto en regiones reguladoras como en codificadoras, de todos los posibles 3-11-3 meros que pueden ser considerados sitios para Gal4. Los 3-11-3 meros subrayados se encuentran en los extremos de las secuencias que son reconocidas experimentalmente por Gal4.

El 3-11-3 meros (hexámero) más común y por el que tiene mayor afinidad la proteína Gal4 es: CGGCCG que se encuentra subrepresentado tanto en regiones codificadoras como en reguladoras; aunque en regiones reguladoras se encuentra todavía menos representado (1.19) que en codificadoras (1.07). La comparación de estos cocientes para regiones codificadoras y reguladoras se presenta en la gráfica 4.

Al hacerse el análisis utilizando la estrategia Top-down trabajando directamente las secuencias de pegado de la proteína y extrayendo sus posibles propiedades biológicas nos da una visión general de lo que se puede hacer dentro de los proyectos de genoma. Se cuenta con una grandes cantidades de secuencias y hace falta más información sobre su función. Entonces para corroborar la eficiencia (o ineficiencia) de esta estrategia faltarían datos, tanto de regiones del pegado de la proteína (que nos servirían para confirmar las posibles propiedades biológicas de las secuencias asociadas a está proteína), como de regiones codificadoras.

La estrategia Bottom-up adolece de la misma situación, la distribución de 3-11-3 meros en regiones codificadoras y reguladoras, presentan en las gráficas extremos (ya sea sobre o sub representadas) muy interesantes, que podrían estar hablando de secuencias con una función biológica importante.

Un ejemplo concreto de análisis de los patrones localizados en los extremos fue la proteína Gal4, que al concatenar la estrategia Botom-up y Top-down, determinó que el hexámero más comúnmente asociado a este sitio de pegado estuvo subrepresentado, tanto en secuencias reguladoras como en codificadoras. El resultado muestra la convergencia de ambas estrategias para integrar la información disponibles en artículos y en bases de datos.

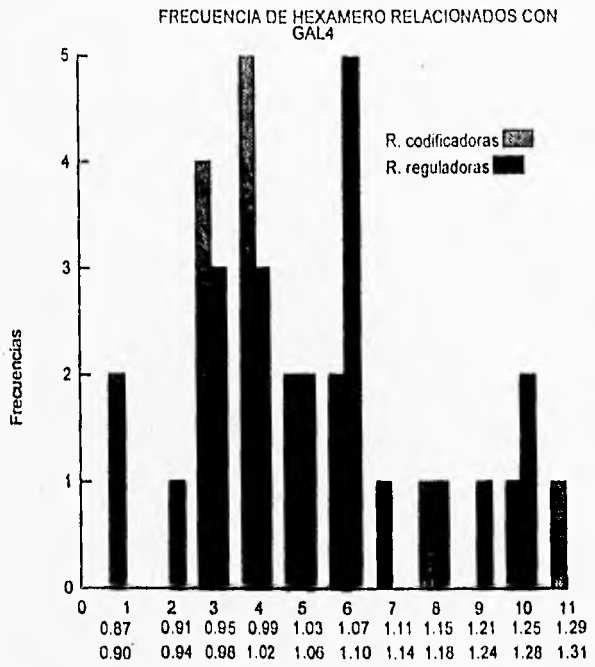
Sería muy arriesgado pensar que todas las secuencias que no estén distribuidas como se espera, tengan una función biológica asociada, pero puede darnos una idea general de como es la distribución de patrones al interior del cromosoma.

V.3 PREDICCIONES EN LOS CROMOSOMAS III Y XI

La distribución de los 5-n-5meros, 6-n-6meros y 8-n-8meros fueron muy semejantes entre sí en ambos cromosomas como se muestra en la gráfica 5.

La distribución de la mayoría de las cadenas de DNA dentro de las regiones reguladoras son, como se esperaría, al azar. Las cadenas más interesantes son aquellas que se encuentran sobre y subrepresentadas en estas regiones.

Cuando se buscó (permitiendo una variación en el patrón) si algunas de estas cadenas podían tener algún significado biológico se encontró lo siguiente:



GRAFICA 4.- Comparación de la distribución de los hexámeros relacionados con Gal4 en regiones reguladora y codificadoras.

Dento de los patrones sobre representados se encontró que uno estaba relacionado con la proteína de *S. cerevisiae* AP1, la cual está involucrada con la respuesta a tensión por oxidación.

Dentro de los patrones sub representados se encontró uno que tenía relación con el pegado de la proteína Mat-alpha2, la cual está muy relacionada con la regulación del mecanismos de diferenciación sexual en levadura

El siguiente análisis se hizo a partir de la distribución de las cadenas de 5-n-5 meros con las diferentes distancias de separación. Se cuantificó a todas las cadenas con simetría directa e invertida que estuvieron dentro de los intervalos sobre y subrepresentados. Obteniendo los siguientes resultados (Tabla 8 y 9):

Cromosoma III:

	SIMETRÍA DIRECTA		SIMETRÍA INVERTIDA	
	Región sub representada	[redacted]	Región sub representada	[redacted]
	9	[redacted]	7	[redacted]
	25	[redacted]	27	[redacted]
	22	[redacted]	46	[redacted]
	37	[redacted]	61	[redacted]

Tabla 8. Número de 5-n[4,5,9 y 10]-5 meros con simetría directa e invertida encontrados en las zonas de sobre y subrepresentación del cromosoma III

Cromosoma XI:

	SIMETRÍA DIRECTA		SIMETRÍA INVERTIDA	
	Región sub representada	[redacted]	Región sub representada	[redacted]
	6	[redacted]	9	[redacted]
	30	[redacted]	17	[redacted]
	30	[redacted]	19	[redacted]
	45	[redacted]	35	[redacted]

Tabla 9. Número de 5-n[4,5,9 y 10]-5 meros con simetría directa e invertida encontrados en las zonas de sobre y subrepresentación del cromosoma XI.

Posteriormente se obtuvo el cociente entre el número de patrones encontrados en regiones subrepresentadas y sobre representadas para analizar algún posible sesgo en los datos que estuviera indicando alguna función biológica (Tabla 10):

	CROMOSOMA III Simetría invertida		CROMOSOMA XI Simetría invertida	Simetría directa
	7		9	
	1.93		1.31	
	1.64		0.54	1.83
	1.95		0.95	1.66

Tabla 10. Valores de los cocientes de los números de los 5-n-5 meros (con simetría directa e invertida) encontrados en zonas de sub-representadas y sobre-representada en los cromosomas III y XI.

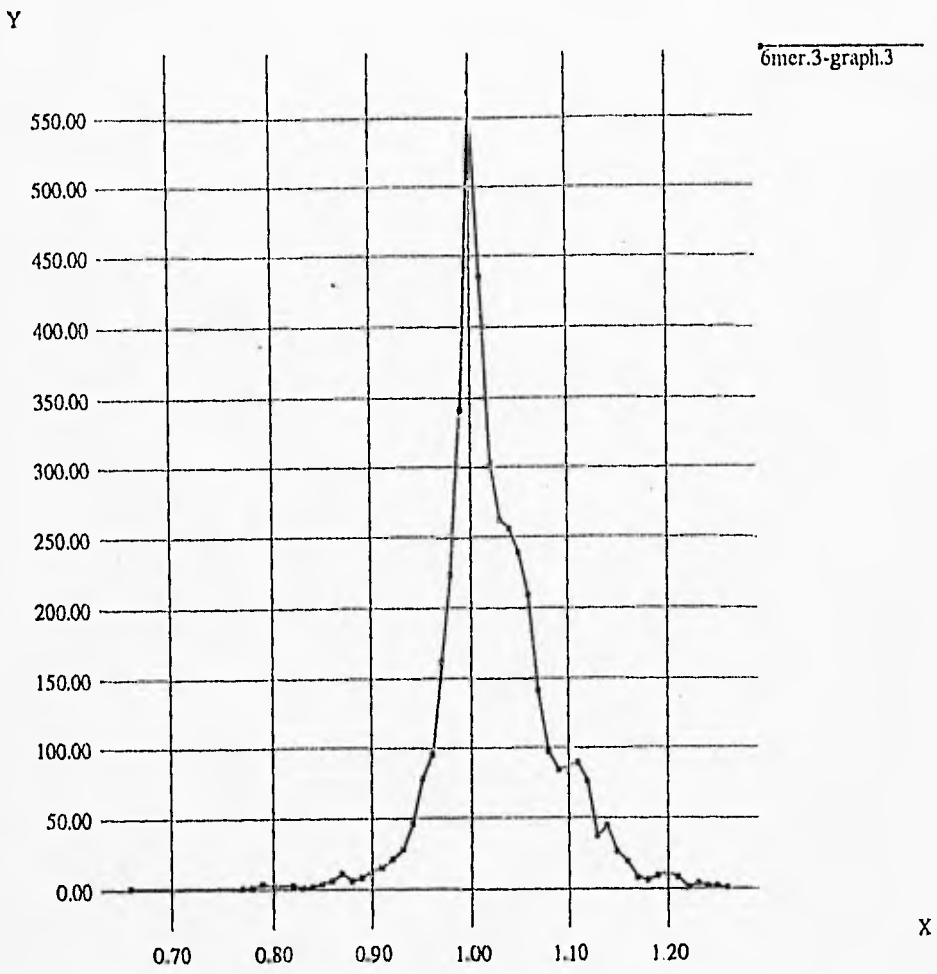
Al observar los valores de los cocientes en la Tabla 10, se puede advertir las diferencias entre las cadenas de 5-n-5 meros que se encuentran en fase (con los intervalos de 9 y 10) y los que se encuentran fuera de fase (4 y 5) tanto en el cromosoma III como en el XI.

Los valores más cercanos a uno indican una distribución equivalente en zonas de sobre representación como sub-representaciones. Mientras que los valores más alejados a uno indican una desviación en la distribución asociado posiblemente a una función biológica.

Curiosamente los resultados más alejados de una distribución equivalente son los patrones que pueden estar asociados a una función biológica. Un gran número de proteínas involucradas en la regulación de la transcripción, se pegan al DNA como dímeros por lo que es necesario que su sitio de unión al DNA se encuentre en fase (en la misma cara del DNA) y la dimerización se da como una imagen especular, por lo que los sitios presentan simetría invertida, tal es el caso de Gal4, GCN4 y Mat-alpha2.

La distribución tiene un sesgo para los patrones en fase (más marcado para los que presentan simetría invertida que directa), se encuentran más sub-representados que sobre-representados. Esto podría deberse a la importancia que tienen estos sitios con respecto al resto de la secuencia, deben existir pocos y en lugares precisos para asegurar especificidad y sean distinguibles para las proteínas reguladoras.

Los 5-n-5 meros que no están en fase y que difícilmente se les puede atribuir una función biológica, presentan un valor de cociente cercano a uno, es decir, se pueden encontrar distribuidos indistintamente tanto sobre como sub-representados.



GRAFICA 5.- Distribución general presentada por todos los 5-n-5 meros, 6-n-6 meros y 7-n-7 meros en las regiones reguladoras de los cromosomas III y XI.

V. CONCLUSIONES Y PERSPECTIVAS.

La ingeniería computacional avanza impresionantemente y tiene un impacto relevante en biología. Los proyectos de secuencias han acumulado explosivamente secuencias de DNA y sin un buen soporte en cómputo sería prácticamente imposible realizar esta faena.

En nuestro laboratorio se analizan las regiones reguladoras con el propósito de hacer búsquedas y predicciones de sitios *cis*-acting para el pegado de proteínas. Para realizar este trabajo, existe una gran variedad de programas de cómputo y no es fácil tomar una decisión sobre la utilización de algún programa en específico.

Es por esto que se hizo una revisión general de algunos de los programas más utilizados en esta área y se hicieron pruebas de reconocimiento de los sitios de pegado de la proteína reguladora GAL4 de uno de los organismos más importante para la investigación: *Sacharomyces cerevisiae*.

En esta revisión no se busca calificar tajantemente a los programas (cada uno ofrece diferentes ventajas y desventajas), pero con base en las necesidades particulares de nuestro laboratorio se elige el que mejor se acopla y obtenga mejores resultados.

Los métodos en está área de trabajo han cambiado con el tiempo, primero, generando secuencia consenso, después, simples matrices construidas con alineaciones hechas a ojo y ahora cuenta con métodos que construyen una alineación donde existe una correlación directa entre la evaluación estadística del sitio y su reconocimiemo termodinámico por una macromolécula.

El programas **Wconsensus** junto con **Patser,** presentan ventajas con respecto a los otros programas. Wconsensus utiliza las matrices como algoritmo de búsqueda. Patser se sustenta en la teoría de la información lo que ofrece un marco teórico sólido para la interpretación de resultados.

Ambos programas ayudaron en la construcción de las estrategias **Top down** y **Bottom Up** para realizar un análisis más global de secuencias y hacer predicciones sobre posibles sitios reguladores.

En el cromosoma II con la proteína Gal4 la estrategia *Top down* presenta limitantes en sus alcances predictivos, debido básicamente a la falta de información biológica en las secuencias. Mientras que la estrategia *Bottom up* muestra resultados de tipo estadístico que podrían ser interesantes ya que reflejan cierto sesgo en la distribución de cadenas asociado muy probablemente a que intervienen en mecanismos biológicos.

Al hacerse el análisis *Bottom up* en las regiones reguladoras del cromosoma III y XI, se descubrió que algunos patrones sobre y sub-representados eran reconocidos como sitios de pegado para algunas proteínas reguladoras, como por ejemplo Mat-alpha2. El último análisis *Bottom up* se hizo cuantificando las cadenas de 5-n-5 meros en las zonas de sub y sobre

representación y se vió que las cadenas que podían funcionar para el pegado de proteínas reguladoras (por encontrarse en fase), se encontraban preferentemente sub representadas.

Las estrategias presentadas nos brindan una perspectiva global del trabajo futuro que se puede llevar a acabo para integrar los trabajos de una nuevo análisis en biología molecular en *S. cerevisiae* y en otros organismos, pero su estudio necesita integrar más elementos de regulación que sólo el sitio de pegado para la proteína reguladora. Algunas características biológicas que pueden enriquecer los análisis son: posiciones relativas de sitios con respecto al CDS, combinación con diferentes elementos involucrados en la transcripción (Rosenblueth *et al.* 1995) y estructura tridimensional de pegado.

BIBLIOGRAFÍA:

Adams, M.R., Das, S. ySmith, T.F. 1995. Multiple Domain Protein Diagnostic Patterns. En prensa.

Alberts, B., Bray, D., Lewis, J., Raff, M., Roberts, K. y Watson, J.D. 1994. Molecular Biology of the Cell. Third edition. Garland Publishing, Inc. New York & London. pp. 401-432.

Bold, H.C., Alexopoulos, C.J., Delevoryas, T. 1987. Morphology of Plants and fungi. Harper & Row, Publishers, New York. pp. 912

Branden, C. y J. Tooze. 1991. Introduction to Protein Structure. Garland Publishing, Inc. New York and London. pp. 113-126.

Botstein, D.yFink, G.R. 1988.Yeast: an experimental organism for modern biology. *Science* 240:1439-1443.

Bram, R.J., Lue, N.F.yKornberg, R.D. 1986.A GAL family of upstream activating sequences in yeast: roles in both induction and repression of transcription. *EMBO J* 5:603-608.

Bucher, P.yTrifonov, E.N. 1986.Compilation and analysis of eukaryotic POL II promoter sequences. *Nucleic Acids Res* 14:10009-10026.

Buratowski, S. 1994.The basics of basal transcription by RNA polymerase II. *Cell* 77:1-3.

Chouard, T. y Yaniv, Moshe.1994. Le Contrôle de L'Expression des Génes. *La Recherche* 266:626-635.

Collado-Vides, J., Magasanik, B.yGralla, J.D. 1991.Control site location and transcriptional regulation in Escherichia coli. *Microbiol Rev* 55:371-394.

Comai, L., Zomerdijk, J.C., Beckmann, H., Zhou, S., Admon, A.yTjian, R. 1994.Reconstitution of transcription factor SL1: exclusive binding of TBP by SL1 or TFIID subunits. *Science* 266:1966-1972.

Danchin, A. 1993. La Secuenciación de Pequeños Genomas.*Mundo Científico* 134:378-386.

Dujon, B., Alexandraki, D., Andre, B., et al. 1994.Complete DNA sequence of yeast chromosome XI. *Nature* 369:371-378.

Feng, D.F. y Doolitle, R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. *J.Mol. Evol.* 25(4):361-360.

Giniger, E., Varnum, S.M.yPtashne, M. 1985.Specific DNA binding of GAL4, a positive regulatory protein of yeast. *Cell* 40:767-774.

Goodrich, J.A., Schwartz, M.L.yMcClure, W.R. 1990.Searching for and predicting the activity of sites for DNA binding proteins: compilation and analysis of the binding sites for *Escherichia coli* integration host factor (IHF). *Nucleic Acids Res* 18:4993-5000.

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

Hertz, G.Z., Hartzell, G.W.yStormo, G.D. 1990.Identification of consensus patterns in unaligned DNA sequences known to be functionally related. *Compnt Appl Biosci* 6:81-92.

Hertz, G.Z. yStormo, G.D. 1995. "Identification of Consensus Paterns in Unaligned DNA and Protein Sequences: a Large-Deviation Statistical Basis for Penalizing Gaps ". In : Lim, H. A. & Cantor, C. R. Bioinformatics and GenomeResearch, World Scientific Publishing, Singapore.,.

Hertz, G.Z. yStormo, G.D. 1995. *E. coli* Promotor Sequences: Analysis and Predection (en prensa).

Huibregtse, J.M., Good, P.D., Marczynski, G.T., Jaehning, J.A.yEngelke, D.R. 1993.Gal4 protein binding is required but not sufficient for derepression and induction of GAL2 expression. *J Biol Chem* 268:22219-22222.

Jones, E.W., Pringle, J.R. y Broach, J.R. 1992. The Molecular and Cellular Biology of the Yeast *Sacharomyces*.Gene Expression. Vol. II. Cold Spring Harbor Laboratory Press. pp.193-283

Knüppel, R., Dietze, P., Lehnberg, W., Frech, K. y Wingender, E. 1994. TRANSFAC Retrieval Program: A Network Model Database of Eukaryotic Transcription Regulating Sequences and Proteins. *J. Comput. Biol* 3:191-198.

Kraulis, P.J., Raine, A.R., Gadhavi, P.L.yLaue, E.D. 1992.Structure of the DNA-binding domain of zinc GAL4 [see comments]. *Nature* 356:448-450.

Kauffman, S.A. 1993. The Origins of Orden. Self-Organization and Selection in Evolution. Oxford University Press. pp. 709.

Latchman, D.S. 1990.Eukaryotic transcription factors. *Biochem J* 270:281-289.

Lehninger, A. L., Nelson, D.L. y Cox, M.M. 1993. Principles of Biochemistry. Worth Publishers. 2nd. ed. pp 1114.

Leuther, K.K.yJohnston, S.A. 1992.Nondissociation of GAL4 and GAL80 in vivo after galactose induction. *Science* 256:1333-1335.

Lewin, B. 1994. Genes V. Oxford University Press. pp 1272.

Marmorstein, R., Carey, M., Ptashne, M.yHarrison, S.C. 1992.DNA recognition by GAL4: structure of a protein-DNA complex [see comments]. *Nature* 356:408-414.

Mulligan, M.E., Hawley, D.K., Entriken, R.yMcClure, W.R. 1984.Escherichia coli promoter sequences predict in vitro RNA polymerase selectivity. *Nucleic Acids Res* 12:789-800.

Oliver, S.G., van der Aart, Q.J., Agostoni-Carbone, M.L., et al. 1992.The complete DNA sequence of yeast chromosome III [see comments]. *Nature* 357:38-46.

Ptashne, M. 1992. A Genetic Switch, Phago λ and Higher Organism. 2nd. edition. Cell Press & Blackwell Scientific Publication. pp 113-183.

Rhodes, D.yKlug, A. 1993.Zinc fingers. *Sci Am* 268:56-9, 62-5.

Rosenblueth, D.A.,Thieffry, D. Huerta, A.M., Salgado, H.yCollado-Vides. 1995. Sintactic Recognition of Regulatory Regions in *Escherichia coli*.(en prensa *CABIOS*)

Schneider, T.D. 1988. Information and Entropy of Patterns in Genetic Switchs. Maximum-Entropy and Bayesian Methods in Science and Engineering. Vol 2. Erickson, G. J. and Smith, C.R. (*eds*). pp 147-154.

Schneider, T.D., Stormo, G.D., Gold, L.yEhrenfeucht, A. 1986.Information content of binding sites on nucleotide sequences. *J Mol Biol* 188:415-431.

Schneider, T.D. New Approaches in Mathematical Biology: Information theory and Moleculas Machines. Trieste Conference on Chemical Evolution, IV; Physics of the Origin and Evolution of Life. Kluwer Academic Publishers. (En prensa).

Schneider, T.D.1996. Information Theory Primer. *ftp://ftp.ncifcrt.gov/pub/delila/primer.ps.*

Shaner, M.C., Blair, I.M. y Schneider, T.D. 1993. Sequence Logos: A Powerful, Yet Simple, Tool. "Alternative Approaches to Sequence Representation". In Hawaii International Conferencia on System Science. pp 813-821.

Spencer, J.F.T., Spencer, D.M. y Smith A.R.W. 1983. Yeast Genetics. Fundamental and Applied Aspect. Springer Verlag, New York Inc.

Staden, R. 1984.Computer methods to locate signals in nucleic acid sequences. *Nucleic Acids Res* 12:505-519.

Staden, R. 1989.Methods for calculating the probabilities of finding patterns in sequences. *Comput Appl Biosci* 5:89-96.

Stormo, G.D. 1988.Computer methods for analyzing sequence recognition of nucleic acids. *Annu Rev Biophys Biophys Chem* 17:241-263.

Stormo, G.D. 1990.Consensus patterns in DNA. *Methods Enzymol* 183:211-221.

Struhl, K. 1994.Duality of TBP, the universal transcription factor. *Science* 263:1103-1104.

Tjian, R. 1995.Molecular machines that control genes. *Sci Am* 272:54-61.

Van Hoy, M., Leuther, K.K., Kodadek, T.yJohnston, S.A. 1993.The acidic activation domains of the GCN4 and GAL4 proteins are not alpha helical but form beta sheets [see comments]. *Cell* 72:587-594.

Willoughby, S. S. 1985. Probabilidad y Estadistica. 13a. ed. Publicaciones Cultural S.A. de C.V. pp. 69-80.

Wingender, E. 1993. Gene regulation in Eukaryotes. VCH. pp. 430

Yockey, H.P. 1992. Information theory and molecular biology. Cambridge Universoty Press. pp. 13-55.

APÉNDICE:

A)

EXPLICACIÓN GENERAL DE LOS PROGRAMAS:

LineUp:
Este programa pertenece al paquete Wisconsin Sequence Analysis Package™, conocido comúnmente como GCG (Genetics Computer Group). Este es un editor en pantalla útil para editar varias secuencias (máximo 30) simultáneamente. El archivo de salida de este programa es un archivo con múltiples secuencias (msf : *multiple sequence file*) que contiene todas las secuencias incorporadas. Este programa no produce alineaciones, solamente sirve para desplegar las secuencias en pantalla y producir el archivo de entrada para el programa de alineación Pile Up.

PileUp:
PileUp crea una alineación múltiple de secuencias a través de comparar pares de secuencias progresivamente. El algoritmo usado es una simplificación del reportado por Freng y Doolitle en 1987. El procedimiento de la alineación múltiple global se inicia con la comparación de las dos secuencias más similares, generando un grupo o *cluster*. Este grupo puede entonces alinearse a la siguiente secuencia o grupo de secuencias alineadas más relacionadas. La alineación final es obtenido por una serie de comparaciones progresivas entre los diferentes clusters, que incluyen secuencias y grupos cada vez más diferentes del par de secuencias inicial. Después de la alineación, las secuencias son agrupadas por su similaridad para producir un dendograma o árbol representando las relaciones de los grupos.

PileUp puede alinear hasta 500 secuencias con una longitud máxima de 7000 caracteres cada una.

Pretty:
Pretty despliega la alineación múltiple creada por PileUp y muestra el consenso (el nucleótido más conservado para cada sitio). Este programa utiliza como archivo de entrada cualquier alineación en formato MSF (*multiple sequence format*). Para desplegar la secuencia consenso se le tiene que dar el comando en línea -CON dando como símbolo consenso el que tenga mayor ocurrencia dentro de una columna. Con la opción -CASE las letras más representativas dentro de la alineación se pondrán en mayúscula para su mejor visualización.

PlotSimilarity:

PlotSimilarity calcula el promedio de identidad en un grupo alineado de secuencias para cada una de sus posiciones. La ventana de comparación se va deslizando a lo largo de la secuencia, sacando los valores de similitud sitio por sitio para su posterior graficación.

Su archivo de entrada puede ser una alineación múltiple creada desde PileUp y una importante restricción es que todas las secuencias deben tener el mismo tamaño.

Wconsensus:

Este programa se puede traer por el ftp anónimos de la Universidad de Colorado y se encuentra en un directorio junto con otros paquetes del programas para análisis múltiple de secuencias.

Wconsensus determina un patrón consenso para un grupo desalineado de secuencias. Se diferencia de sus otros programas homónimos (Consensus y Lconsensus escritos por Stormo and Hartzell que se encuentran en el mismo directorio) que puede determinar el ancho del patrón de búsqueda y no adiciona huecos.

El algoritmo realiza alineaciones locales múltiples y se basa en la construcción de matrices de frecuencias que represente un patrón consenso de secuencias . Un ejemplo de archivos de salida de este algoritmo es el siguiente:

```
MATRIX I
number of sequences = 7
width = 21
unadjusted information = 26.046
ln(probability) = -79.9736   probability = 1.85313E-35
ln(expected frequency) = -56.7802   expected frequency = 2.19123E-25

A| 5  0  0  0  0  5  0  5  1  4  2  6  2  3  0  7  0  0  4  0  4
C| 0  6  0  0  0  0  0  0  0  0  2  0  1  3  7  0  0  1  0  1  0
G| 1  1  0  7  1  0  0  1  0  1  0  1  0  0  0  0  7  1  1  1  0
T| 1  0  7  0  6  2  7  1  6  2  3  0  4  1  0  0  0  5  2  5  3

 1|6   :   1/3     ACTGTATGAGCATACAGTATA
 2|7   :   4/4     ACTGTATATTCATTCAGGTCA
 3|5   :   5/3     ACTGTTTTTTTATCCAGTATA
 4|1   :   6/1     TCTGTATATATACCCAGCTTT
 5|3   :   8/12    ACTGTATATAAAAACAGTATA
 6|2   :  10/4     GCTGTATATAAAACCAGTGGT
 7|4   :  11/2     AGTGGTTATATGTACAGTATT
```

Una matriz de peso (*weigth matrix*) es un arreglo bidimensional de los valores de ocurrencia para un grupo de caracteres en cada una de las posiciones. Para secuencias de DNA el arreglo

es de 4 x L donde el 4 representa cada uno de los renglones para los 4 nucleótidos (A, C, G, T) y L columnas que corresponden a las posiciones del nucleótido a lo largo de un patrón.

Dentro de Weonsensus se van construyendo por ciclos; en el primer ciclo se construyen las primeras matrices, después a éstas se les van adicionando secuencias, pero al incorporarse nuevas secuencias sólo se guardan aquéllas cuyo contenido informacional vaya siendo más alto (Figura A1). El número máximo de secuencias guardadas se puede determinar con la opción -q. A las matrices salvadas, en los siguientes ciclos, se les va incorporando progresivamente secuencias y se van eliminando aquellas que no van aumentando su contenido informacional conforme se adicionan secuencias.

El programa usa tres criterios diferentes para decidir cuando parar y dejar de adicionar secuencias:

1) Cuando cada secuencia ha contribuido para la matriz, es decir, todas las secuencias están presentes al menos una vez en la alineación.

2) Al designar el número máximo de secuencias guardadas en la matriz (con la opción -n).

3) El algoritmo decide el número de ciclos para encontrar la alineación más significativa (esto se puede modificar con la opción -t).

La significatividad en matrices depende de sus valores para el contenido informacional (indica que tan conservadas están las posiciones dentro de la alineación) y la mejor alineación es el que tiene el valor más bajo de la frecuencia esperada (*expected frecuency*). Lo que indica el valor de frecuencia esperada es la probabilidad de encontrar este mismo valor en una alineación al azar de igual tamaño y con el mismo número de secuencias.

Al realizar las alineaciones se tienen que variar los valores para la corrección de la desviación estándar (con la opción -s) para el cálculo de frecuencia esperada; los valores que recomienda el autor son de 0.5, 1, 1.5 y 2.0 (Wconsensus's program).

Para cuando la secuencia es simétrica, da la opción de añadir en la alineación los inversos de las secuencias: -c1 que incluye ambas cadenas como secuencias separada o -c2 que incluye ambas cadenas como una sola secuencia.

El programa imprime dos listas de matrices; en la primera se incluyen las que tuvieron los valores más altos en el ajuste de la información (*adjusted information*) dentro de los ciclos; se ordenan por decremento en la significancia estadística (es decir incrementando la frecuencia esperada). En general, en la primera lista se encuentran las alineaciones más interesantes. La segunda lista se salva después de finalizar el ciclo del programa y es ordenada también por el decremento en la significancia estadística; la última lista se usa cuando se desea que cada secuencia contribuya con un patrón en la alineación final (es decir, cuando no se utilizó la opción -n).

En el archivo de salida los patrones contenidos en la matriz son listados a partir del orden de las secuencias en el archivo de entrada. Se tienen dos columnas. En la primera, los valores son enteros y se encuentran separados "a | b". El primer entero indica la secuencia a la que pertenece el patrón, y el segundo entero indica durante qué ciclo se adicionó el patrón a la

matriz. En la segunda columna se encuentra "a/b" donde el primer entero indica la secuencia a la que pertenece la palabra y el segundo indica donde empieza el patrón dentro de la secuencia.

MacTargsearch:

MacTargsearch es una adaptación del programa TARGSEARCH (Mulligan, 1984), que se utilizó para predecir secuencias promotoras σ 70. Se inició buscando promotores para *E. coli*, secuencias -10 y -35 y calificando a los espacios que había entre ellas.

Fue compilado en Microsoft BASIC para Apple Macintosh y para correr requiere de una Macintosh con al menos un megabyte de memoria y de preferencia conectada a una ImageWrite (Goodrich, 1990).

En este programa se construye un *target file* que consta de dos regiones y un espacio entre ambas, (se conserva el modelo la región -35 /espacio/ -10 necesario para el reconocimiento de la RNA polimerasa).

El patrón o target, se crea a partir de un grupo de secuencias, previamente alineadas, el cual se introducen en la matriz de peso utilizada para evaluar los sitios.

Ésta contendrá los valor asociados a cada una de las cuatro bases en cada posición de la secuencia. Como se observa en el siguiente ejemplo de IHF:

Shark HD:Desktop Folder:* Shark Users *:Shark- Vicky:MacTargsearch 2.0:IHF Si

Length of Region 1 : 24 Length of Region 2 : 24 Number of Spacers : 1 Clear

Maximum Summation Score : 682 Baseline Score : 323 Print Save Close

Region 1

bp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	4	7	7	9	10	10	8	6	15	7	12	16	13	6	6	3	11	14	9	8	13	14	20	21
C	6	3	10	5	5	4	4	4	2	1	6	1	3	8	5	9	7	5	1	4	2	4	0	0
G	8	4	4	5	2	7	3	7	4	5	2	3	2	1	3	2	4	1	3	2	2	1	0	1
T	9	13	6	8	10	6	10	10	6	14	7	7	9	12	13	13	5	7	14	13	10	8	7	5

Spacer

Length 0

Score 0

Region 2

bp	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24
A	5	0	27	21	5	15	19	6	2	0	20	7	12	9	7	6	9	17	9	13	8	9	8	11
C	0	27	0	0	9	4	3	6	0	1	0	6	6	7	6	9	2	3	4	3	10	6	2	4
G	0	0	0	3	2	3	3	13	1	0	7	4	5	3	3	10	3	2	5	2	4	10	4	4
T	22	0	0	3	11	5	2	2	24	26	0	10	4	8	11	2	13	5	9	9	5	2	11	6

En este archivo-patrón está contenido la **presencia** de las cuatro bases para cada posición en las 27 secuencias de IHF compiladas.
Durante la búsqueda - para promotor, DNA binding site, etc. - dentro de una secuencia de DNA se determina el **"similarity score"** o valor de similitud para todos los posibles sitios en una dirección y en su reversa:

$$S.S=\left(\frac{\text{Suma del valor de las bases por sitio + valor del espacio - valor basal}}{\text{maximo valor - valor basal}}\right)100$$

El valor basal se refiere a la presencia de bases al azar en una secuencia, y es igual al 25% de la suma total de los valores para todas las bases y su valor del espacio que le corresponda.
El máximo valor es la suma de los valores más altos para una base en cada sitio y del valor más alto de los espacios.

Consensus:
Este programa pertenece a GCG y está incluido, junto con otros programas, en la sección de reconocimientos de patrones (Pattern Recognition).
Calcula un consenso para un grupo de secuencias nucleotídicas. Construye una tabla con los porcentajes de G, A, T y C en cada posición, y teniendo la contribución de cada nucleotido escribe la expresión consenso.
Su archivo de entrada debe ser un texto que contenga un grupo alineado de secuencias (se puede utilizar pile-up) que sean del mismo tamaño y no mayor de 130 caracteres (es decir sólo una secuencia por línea). Para su sintaxis la primera línea debe incluir una diagonal seguida de dos puntos en línea. No acepta espacios en blanco ni algún otro carácter fuera de la secuencia. Ejemplo:

```
/          ..
.........AAATAGGAT
.........TTGTAGGTG
..........TGTAGGTG
TTTATTGTATGTGAAGATT
```

La secuencia consenso depende de la tabla de porcentajes y del nivel de certidumbre que se desee representar. El programa por default presenta un porcentaje de certidumbre de 75.0% . Este valor se obtiene por columna (posición) con una suma progresiva iniciándose con el valor más alto, seguido del segundo, y así sucesivamente hasta que la suma sea igual o mayor que el nivel de certidumbre elegido. Si dos nucleotidos tienen el mismo valor se designa arbitrariamente cualquiera de ellos.
El archivo de salida se presenta como en el siguiente ejemplo:

CONSENSUS of: acceptor.dat

*****

%G	15	22	10	10	10	6	7	9	7	5	5	24	1	0
%A	15	10	10	15	6	15	11	19	12	3	10	25	4	100
%T	52	44	50	54	60	49	48	45	45	57	58	30	31	0
%C	18	25	30	21	24	30	34	28	36	35	27	21	64	0
Total	114	114	115	127	127	127	128	128	128	130	131	131	131	131

%G	100	52	24	19
%A	0	22	17	20
%T	0	8	37	29
%C	0	18	22	32
Total	131	131	131	131

*****

CONSENSUS sequence to a certainty level of 75.0 percent at each position:

BBYHYYYHYY YDYAGVBH

Para la ambigüedad en nucleótidos se basa en la tabla IUB, utilizada para todos los programas den GCG y en GeneWorks.

FitConsensus:

Este programa es el complemento de *Consensus program*. Utiliza la tabla de porcentajes para cotejar y calificar todos las posibles alineaciones en una secuencia nucleotídica. En los parámetros se especifica el numero de *fits* (patrones encontrados) que se desean, y muestra aquellos fragmentos, que a los largo de la secuencia, tuvieron mayor calificación.
La manera como se presentan los resultados es en orden ascendente a partir de su posición en la secuencia.

Para calificar, una vez que se tiene el fragmento, se coteja según su secuencia, los valores que le corresponden en la tabla de porcentajes, se suman y ese valor se divide entre la longitud de la matriz (18 en el caso de la tabla consensus del ejemplo anterior) y esa es la calificación para ese fragmento de la secuencia.

Profile:

Este paquete se ejecutan bajo plataforma UNIX y consta principalmente de dos programas: make_profile y find_profile.

Con make_profile se construye una matriz que permite encontrar patrones en secuencias. Primero se necesita tener un archivo de entrada con los sitios, previamente alineados que contienen a la matriz .

Se escribe la orden que ejecuta el programa y las características del archivo. En segundos crea un archivo de salida como se ve en el siguiente ejemplo con los sitios de pegado para la proteína Ara C:

* Matrix file started at index 0. Info Content = 6.375004

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
A	1	3	4	2	1	2	2	5	2	1	1	4	4	2	2	4	2	2	5	2	3
C	1	0	2	2	5	4	1	0	0	2	3	1	2	4	5	2	1	2	3	2	3
G	3	3	2	5	2	2	3	0	1	0	2	0	2	0	1	1	1	3	0	3	0
T	4	3	1	0	0	1	2	4	6	6	2	4	1	3	0	2	5	2	1	2	3

* End of file

El archivo de salida presenta el calculo del Contenido informacional total de la matriz (con la misma fórmula que el programa consensus).

Find_profile necesita como archivo de entrada el archivo de salida de make_profile. A partir de los valores de la matriz se construye una tabla logarítmica de probabilidad, sumando 1 a todos los valores (para evitar los valores de cero) y aplicando posteriormente la siguiente formula (obtenida directamente del código fuente) : $\log_2 \frac{fb}{Pb}$.

Convertida la matriz, se suman todos los valores que le corresponden a la secuencia dependiendo de su sitio dentro de la posición

Los resultados se presentan de la siguiente manera:

Read in matrix:

3	4	2	1	2	2	5	2	1	1	4	4	2	2	4	2	2	5	2	3	0
0	2	2	5	4	1	0	0	2	3	1	2	4	5	2	1	2	3	2	3	0
3	2	5	2	2	3	0	1	0	2	0	2	0	1	1	1	3	0	3	0	0
3	1	0	0	1	2	4	6	6	2	4	1	3	0	2	5	2	1	2	3	0

Converted matrix to :

0.62 -0.12 -0.70 -0.12 -0.12 0.88 -0.12 -0.70 -0.70 0.62 0.62 -0.12 -0.12 0.62 -0.12 -0.12 0.88 -0.12 0.30 -1.70

-0.12 -0.12 0.88 0.62 -0.70 -1.70 -1.70 -0.12 0.30 -0.70 -0.12 0.62 0.88 -0.12 -0.70 -0.12 0.30 -0.12 0.30 -1.70

-0.12 0.88 -0.12 -0.12 0.30 -1.70 -0.70 -1.70 -0.12 -1.70 -0.12 -1.70 -0.70 -0.70 -0.70 0.30 -1.70 0.30 -1.70 -1.70

-0.70 -1.70 -1.70 -0.70 -0.12 0.62 1.11 1.11 -0.12 0.62 -0.70 0.30 -1.70 -0.12 0.88 -0.12 -0.70 -0.12 0.30 -1.70

```
Best score for matrix = 13.67

Match in sequence      2 at position    46, score =   4.5,
'...GACGCCGTGCAAATAATCAATG..."
Match in sequence      2 at position    67, score =   6.2,
"...GTGGACTTTTCTGCCGTGATTA..."
Match in sequence      2 at position   101, score =   5.2,
'...TTACGCGTTTTTGTCATGGCTT..."
Match in sequence      2 at position   481, score =   4.2,
"...TTTGCTGTTCCCGCCAGGAGAG..."
Match in sequence      2 at position   531, score =   3.2,
"...CTCGCGAATGGTATCACCAGTG..."
Match in sequence      2 at position   923, score =   3.8,
"...TTAGGGATTAGCGTCTTAAGCT..."
Match in sequence      2 at position  1223, score =     3,
"...CAGGCACATTATGCAAGCATTG..."
Match in sequence      3 at position   113, score =   4.5,
'...GACGCCGTGCAAATAATCAATG..."
Match in sequence      3 at position   134, score =   6.2,
"...GTGGACTTTTCTGCCGTGATTA..."
Match in sequence      4 at position   376, score =   4.2,
"...TTTGCTGTTCCCGCCAGGAGAG...'
Match in sequence      4 at position   426, score =   3.2,
"...CTCGCGAATGGTATCACCAGTG..."
Match in sequence      4 at position   818, score =   3.8,
"...TTAGGGATTAGCGTCTTAAGCT..."
Match in sequence      4 at position  1143, score =     3,
"...GAACACTTTATTACCCAACCAC...'

Total of 13 matches
```

Para el archivo de salida se pueden designar un umbral a partir del cual muestre las secuencias con los valores más altos.

Patser:

Patser está incluido (junto con Wconsensus) dentro de un paquete que se encuentra en la Universidad de Colorado con programas para análisis múltiple de secuencias.

Con este programa se pueden encontrar patrones de interés en grupos de secuencias. Patser va calificando, avanzando en ambas direcciones en el DNA, todos los posibles patrones que resulten. El archivo de entrada tiene que ser una matriz que se haya creado en algún programa del paquete (como puede se Consensus o Wconsensus) a partir de una alineación múltiple.

El programa convierte los valores de la matriz utilizando la fórmula: $\log 2 \frac{f_{b,i}}{Pb}$ (Stormo, G.D. 1990); donde $f_{b,i}$ es la frecuencia de la base b en la posición i y Pb representa

la frecuencia genómica en cada base. La formula construye una nueva matriz a partir de la cual se realizan las búsquedas.

El archivo de búsqueda puede contener algunos comentarios seguidos de ' ; ', ' % ' o ' # ' para que no sean tomados en cuenta por el programa. A cada secuencia le debe corresponder un nombre y estar encerrada entre diagonales invertidas (\).

Tiene diferentes opciones para determinar a partir de que umbrales se quieren archivar las calificaciones de los patrones (la más usual es '-l ').

Puede buscar en sólo una dirección o en reversa; cuando el sitio es simétrico conviene hacerlo en ambas direcciones (la opción -c califica la secuencia complementaria).

FindPatterns:

FindPatterns se encuentra también en la paquetería de GCG y pertenece a la sección Database Searching, localiza cortos patrones (como lugares de corte para enzimas de restricción o secuencias consenso) en grupos de secuencias.

El archivo de entrada es un pattern.dat (creado previamente en algún editor de texto en UNIX) con el siguiente formato, ejemplo:

Name	Offset	Pattern	Overhang	Documentation ..
BamHI	1	GGATCC	0	!
EcoRI	1	GAATTC	0	!
Promotor	1	TAATA(N){20,30}ATG	0	!

Este archivo contiene los patrones que se desean encontrar en las secuencias. Cada patrón no debe ser más largo de 132 caracteres y puede buscar hasta 2000 patrones en una secuencia de nucleotidos.

B)

ARCHIVO DE SALIDA DE LOS PROGRAMAS DE BÚSQUEDA:

MacTargsearch:

Target File: Shark HD:Desktop Folder:* Shark Users *:Shark- Vicky:MacTargsearch 2.0:Gal4:Gal4_p1
Sequence File: Shark HD:Desktop Folder:* Shark Users *:Shark- Vicky:MacTargsearch 2.0:Gal4:k02115.ig
Sequence Length: 907 bp linear (908 characters in sequence file)
Searched Region: All

Lower S.S. Limit: 55 Imput Checking: Yes Ignore Lowercase: Yes

BP/Dir	Region 1	Spacer	Region 2	Sim. Score
(cons.)	CGGAGCACTGTCCTCCG (116)	0bp (0)	(0)	100.0
387 F	CGGGTGACAGCCCTCCG (105)			85.0
405 F	AGGAAGACTCTCCTCCG (103)			82.3
403 R	CGGAGGGCTGTCACCCG (101)			79.6
421 R	CGGAGGAGAGTCTTCCT (94)			70.1
469 F	CGCGCCGCACTGCTCCG (94)			70.1
368 F	CGGATTAGAAGCCGCCG (90)			64.6
384 R	CGGCGGCTTCTAATCCG (90)			64.6
485 R	CGGAGCAGTGCGGCGCG (87)			60.5

Target File: Gal4:m81879.ig
Sequence Length: 2845 bp linear (2846 characters in sequence file)
Searched Region: All
Lower S.S. Limit: 55 Imput Checking: Yes Ignore Lowercase: Yes

BP/Dir	Región 1	Spacer	Región 2	Sim. Score
(cons.)	CGGAGCACTGTCCTCCG (116)	0bp (0)	(0)	100.0
331 F	CGGCGGTCTTTCGTCCG (107)			87.8
217 F	CGGAAAGCTTCCTTCCG (96)			72.8
366 R	CGGCGCAGATATCTCCG (95)			71.4
347 R	CGGACGAAAGACCGCCG (94)			70.1
419 F	CGGATCACTCCGAACCG (94)			70.1
350 F	CGGAGATATCTGCGCCG (93)			68.7
414 F	CGGGGCGGATCACTCCG (92)			67.3
435 R	CGGTTCGGAGTGATCCG (89)			63.3

233 R CGGAAGGAAGCTTTCCG (87) 60.5

430 R CGGAGTGATCCGCCCCG (87) 60.5

Target File: Gal4:x00215.ig
Sequence Length: 1008 bp linear (1009 characters in sequence file)
Searched Región: All
Lower S.S. Limit: 55 Imput Checking: Yes Ignore Lowercase: Yes

---

BP/Dir	Región 1	Spacer	Región 2	Sim. Score	
(cons.)	CGGAGCACTGTCCTCCG (116)	0bp (0)	(0)		100.0
746 F	CGGACAACTGTTGACCG (104)				83.7
762 R	CGGTCAACAGTTGTCCG (102)				81.0
659 F	CGGAGCACTGTTGAGCG (99)				76.9
675 R	CGCTCAACAGTGCTCCG (94)				70.1
769 R	*CGGATCACGGTCAACAG (89)*				63.3

Target File:Gal4:x01667.ig
Sequence Length: 2457 bp linear (2458 characters in sequence file)
Searched Región: All
Lower S.S. Limit: 55 Imput Checking: Yes Ignore Lowercase: Yes

---

BP/Dir	Región 1	Spacer	Región 2	Sim. Score	
(cons.)	CGGAGCACTGTCCTCCG (116)	0bp (0)	(0)		100.0
561 F	CGGCGCACTCTCGCCCG (108)				89.1
577 R	CGGGCGAGAGTGCGCCG (96)				72.8

Target File: Gal4:x03102.ig
Sequence Length: 2812 bp linear (2813 characters in sequence file)
Searched Región: All
Lower S.S. Limit: 55 Imput Checking: Yes Ignore Lowercase: Yes

---

BP/Dir	Región 1	Spacer	Región 2	Sim. Score

(cons.)	CGGAGCACTGTCCTCCG (116)	0bp (0)	(0)	100.0
208 F	CGGCCATATGTCTTCCG (98)			75.5
224 R	CGGAAGACATATGGCCG (91)			66.0
707 R	*CGTCGGAATCTCTGCCG (89)*			63.3
691 F	*CGGCAGAGATTCCGACG (88)*			61.0

FitConsensus:

k02115.fit

```
::::::::::::::
 FITCONSENSUS of: k02115.em  Check: 2280  from: 1  to: 907

 FROMIG of: /export/home/gcg1/Vick/Gal/allgal2
  toig of: k02115  check: 2280  from: 1  to: 907
 locus       yscga1        907 bp ds-dna              pln       14-apr-1992
 definition  yeast (s. cerevisiae) gal1-gal10 inducible promoter and genes.
 accession   k02115
 keywords    gal1 gene; gal10 gene; epimerase; galactokinase; . .


  position:   368   387   405   469
     frame:     2     3     3     1
   quality: 52.94 61.76 60.59 55.29


::::::::::::::
m81879.fit
::::::::::::::
 FITCONSENSUS of: m81879.em  Check: 9482  from: 1  to: 2845
 locus       yscga12a     2845 bp ds-dna             pln       15-apr-1992
 definition  saccharomyces cerevisiae galactose permease (gal2) gene,
complete cds.
 accession   m81879 . . .


  position: 217   331   350   414   419
     frame:   1     1     2     3     2
   quality:56.47 62.94 54.71 54.12 55.29


::::::::::::::
x00215.fit
```

```
::::::::::::::
FITCONSENSUS of: x00215.em  Check: 6177  from: 1  to: 1008

locus        scgal7       1008 bp     dna              pln       06-dec-1983
definition   yeast gal7 gene transcriptional initiation Región encoding
             galactose-1-phosphate uridylyl transferase (strain d585-11c).


position: 659   746
   frame:   2     2
 quality: 58.24 61.18


::::::::::::::
x01667.fit
::::::::::::::
FITCONSENSUS of: x01667.em  Check: 7249  from: 1  to: 2457

locus        scgal80      2457 bp     dna              pln               06-
jul-1989
definition   yeast regulatory protein gal80.
accession    x01667
keywords     direct repeat; inverted repeat; regulatory protein.


  position: 561
     frame:   3
   quality: 63.53


::::::::::::::
x03102.fit
::::::::::::::
FITCONSENSUS of: x03102.em  Check: 5510  from: 1  to: 2812
locus        scmel1       2812 bp     dna              pln       06-jul-1989
definition   yeast mel1 gene for alpha-galactosidase.
accession    x03102
keywords     alpha-galactosidase; inverted repeat; protein binding site;


  position:   208   691
     frame:     1     1
   quality: 57.65 51.76
```

En dirección reversa complementaria los resultados fueron:

```
k02115.fit
::::::::::::::
```

```
FITCONSENSUS of: k02115.rev  Check: 933  from: 1  to: 907

  REVERSE-COMPLEMENT of: k02115.gb_pl  check: 2280  from: 1  to: 907
LOCUS       YSCGAL        907 bp ds-DNA              PLN       14-APR-1992
DEFINITION  Yeast (S. cerevisiae) GAL1-GAL10 inducible promoter and genes.
ACCESSION   K02115
KEYWORDS    GAL1 gene; GAL10 gene; epimerase; galactokinase;
            mutational analysis; promoter Región; regulatory Región; . . .


  position: 423   436   449   468   490   505   524
     frame: 3     1     2     3     1     1     2
   quality: 51.18 45.29 34.12 44.71 45.88 59.41 52.94
```

::::::::::::::

m81879.fit

::::::::::::::

```
 FITCONSENSUS of: m81879.rev  Check: 2379  from: 1  to: 2845

  REVERSE-COMPLEMENT of: m81879.gb_pl  check: 9482  from: 1  to: 2845
LOCUS       YSCGAL2A     2845 bp ds-DNA              PLN       15-APR-1992
DEFINITION    Saccharomyces cerevisiae galactose permease (GAL2) gene,
complete cds.
ACCESSION   M81879

  position: 2411  2416  2480  2499  2613
     frame: 2     1     2     3     3
   quality: 52.35 51.18 55.88 55.29 51.18
```

::::::::::::::

x00215.fit

::::::::::::::

```
 FITCONSENSUS of: x00215.rev  Check: 3848  from: 1  to: 1008

  REVERSE-COMPLEMENT of: x00215.gb_pl  check: 6177  from: 1  to: 1008
LOCUS       SCGAL7       1008 bp    DNA             PLN       06-DEC-1983
DEFINITION  Yeast GAL7 gene transcriptional initiation Región encoding
            galactose-1-phosphate uridylyl transferase (strain D585-11C).
ACCESSION   X00215
KEYWORDS    transferase. . . .

 position:   240   247   334
    frame:     3     1     1
  quality: 52.35 60.00 55.29
```

::::::::::::::
x01667.fit
::::::::::::::

```
 FITCONSENSUS of: x01667.rev  Check: 4644  from: 1  to: 2457

  REVERSE-COMPLEMENT of: x01667.gb_pl  check: 7249  from: 1  to: 2457
LOCUS       SCGAL80      2457 bp    DNA             PLN       06-JUL-1989
DEFINITION  Yeast regulatory protein GAL80.
ACCESSION   X01667
KEYWORDS    direct repeat; inverted repeat; regulatory protein.
SOURCE      yeast. . . .


 position: 1881
     frame: 3
   quality: 56.47
```

::::::::::::::
x03102.fit
::::::::::::::

```
 FITCONSENSUS of: x03102.rev  Check: 8248  from: 1  to: 2812

  REVERSE-COMPLEMENT of: x03102.gb_pl  check: 5510  from: 1  to: 2812
LOCUS       SCMEL1       2812 bp    DNA             PLN       30-MAR-1995
DEFINITION  Yeast MEL1 gene for alpha-galactosidase.
ACCESSION   X03102
KEYWORDS    alpha-galactosidase; inverted repeat; protein binding site;
            signal peptide. . . .

 position: 2106  2589
     frame:  3     3
   quality: 52.35 53.53
```

Profile:

Read in matrix:

A:	1	0	0	4	1	2	6	1	4	1	0	1	0	2	0	0	0	0
C:	9	0	1	3	3	4	0	7	0	3	2	6	5	1	9	10	0	0
G:	0	10	9	3	4	3	2	2	0	4	1	1	4	1	1	0	10	0
T:	0	0	0	0	2	1	2	0	6	2	7	2	1	6	0	0	0	0

Converted matrix to :

-0.81 -1.81 -1.81 0.51 -0.81 -0.22 1.00 -0.81 0.51 -0.81 -1.81 -0.81 -1.81 -
0.22 -1.81 -1.81 -1.81 -1.81
1.51 -1.81 -0.81 0.19 0.19 0.51 -1.81 1.19 -1.81 0.19 -0.22 1.00 0.78
-0.81 1.51 1.65 -1.81 -1.81

```
 -1.81   1.65   1.51   0.19   0.51    0.19  -0.22  -0.22  -1.81   0.51  -0.81  -0.81   0.51
-0.81  -0.81  -1.81   1.65  -1.81
 -1.81  -1.81  -1.81  -1.81  -0.22  -0.81  -0.22  -1.81   1.00  -0.22   1.19  -0.22  -0.81
1.00  -1.81  -1.81  -1.81  -1.81

Match in sequence      1 at position   207, score =     11,
"...CGGCCATATGTCTTCCGA..."
Match in sequence      1 at position   690, score =    7.2,
"...CGGCAGAGATTCCGACGG..."
Match in sequence      1 at position  1094, score =    6.4,
"...CTGAGTTCACTCGTCCAG..."
Match in sequence      1 at position  2812, score =     17,
"...CGGAGCACTGTCCTCCGA..."
Match in sequence      2 at position   560, score =     14,
"...CGGCGCACTCTCGCCCGA..."
Match in sequence      3 at position   658, score =     12,
"...CGGAGCACTGTTGAGCGA..."
Match in sequence      3 at position   745, score =     13,
"...CGGACAACTGTTGACCGT..."
Match in sequence      4 at position   216, score =    9.9,
"...CGGAAAGCTTCCTTCCGG..."
Match in sequence      4 at position   330, score =     14,
"...CGGCGGTCTTTCGTCCGT..."
Match in sequence      4 at position   349, score =      9,
"...CGGAGATATCTGCGCCGT..."
Match in sequence      4 at position   413, score =    9.5,
"...CGGGGCGGATCACTCCGA..."
Match in sequence      4 at position   418, score =    8.8,
"...CGGATCACTCCGAACCGA..."
Match in sequence      5 at position   367, score =    7.8,
"...CGGATTAGAAGCCGCCGA..."
Match in sequence      5 at position   386, score =     14,
"...CGGGTGACAGCCCTCCGA..."
Match in sequence      5 at position   404, score =     13,
"...AGGAAGACTCTCCTCCGT..."
Match in sequence      5 at position   468, score =     10,
"...CGCGCCGCACTGCTCCGA..."
```

INVERSA:

```
Best score for matrix = 18.72
Match in sequence      1 at position  2105, score =    7.2,
"...CGTCGGAATCTCTGCCGC..."
Match in sequence      1 at position  2588, score =    7.8,
"...CGGAAGACATATGGCCGA..."
Match in sequence      2 at position  1880, score =     10,
"...CGGGCGAGAGTGCGCCGG..."
Match in sequence      3 at position   239, score =    6.1,
"...CGGATCACGGTCAACAGT..."
Match in sequence      3 at position   246, score =     12,
"...CGGTCAACAGTTGTCCGA..."
Match in sequence      3 at position   333, score =    8.9,
"...CGCTCAACAGTGCTCCGA..."
Match in sequence      4 at position  2410, score =    6.3,
"...CGGTTCGGAGTGATCCGC..."
```

```
Match in sequence      4 at position  2415, score =      7,
"...CGGAGTGATCCGCCCCGA..."
Match in sequence      4 at position  2479, score =    9.7,
"...CGGCGCAGATATCTCCGC..."
Match in sequence      4 at position  2498, score =      9,
"...CGGACGAAAGACCGCCGG..."
Match in sequence      4 at position  2612, score =    7.3,
"...CGGAAGGAAGCTTTCCGA..."
Match in sequence      5 at position   422, score =    7.9,
"...CGGAGCAGTGCGGCGCGA..."
Match in sequence      5 at position   486, score =    9.6,
"...CGGAGGAGAGTCTTCCTT..."
Match in sequence      5 at position   504, score =     11,
"...CGGAGGGCTGTCACCCGC..."
Match in sequence      5 at position   523, score =    5.7,
"...CGGCGGCTTCTAATCCGT..."
```

PATSER

File containing the sequence information: allgal
Also score the complementary strands
Print scores greater than or equal to 13.47

```
***** Information for the alphabet from file "alphabet". *****
letter   1: A (complement: T)  prior frequency = 0.250000
letter   2: G (complement: C)  prior frequency = 0.250000
letter   3: C (complement: G)  prior frequency = 0.250000
letter   4: T (complement: A)  prior frequency = 0.250000

width of the summary matrix: 17
A |   1   0   0   4   1   2   6   1   4   1   0   1   0   2   0   0   0
G |   0  10   9   3   4   3   2   2   0   4   1   1   4   1   1   0  10
C |   9   0   1   3   3   4   0   7   0   3   2   6   5   1   9  10   0
T |   0   0   0   0   2   1   2   0   6   2   7   2   1   6   0   0   0

x03102      position= 208          score= 18.45
x03102      position= 208C         score= 15.48
x03102      position= 691          score= 14.90
x03102      position= 691C         score= 14.96
x03102      position= 1095         score= 14.10
x01667      position= 561          score= 21.92
x01667      position= 561C         score= 18.15
x00215      position= 659          score= 19.61
x00215      position= 659C         score= 16.64
x00215      position= 746          score= 20.87
x00215      position= 746C         score= 19.28
x00215      position= 753C         score= 13.82
m81879      position= 217          score= 17.62
m81879      position= 217C         score= 15.06
m81879      position= 331          score= 21.77
m81879      position= 331C         score= 16.70
```

```
m81879      position= 350          score= 16.74
m81879      position= 350C         score= 17.45
m81879      position= 414          score= 17.23
m81879      position= 414C         score= 14.74
m81879      position= 419          score= 16.55
m81879      position= 419C         score= 14.06
k02115      position= 368          score= 15.55
k02115      position= 368C         score= 13.47
k02115      position= 387          score= 21.35
k02115      position= 387C         score= 18.70
k02115      position= 405          score= 20.35
k02115      position= 405C         score= 17.37
k02115      position= 469          score= 17.83
k02115      position= 469C         score= 15.61
```

FINDPATTERNS

```
          k02115.em  ck: 2280  len: 907   !

Gal4_1                CGGVBVRSWBYYSWCCG
          387: CCGAG CGGGTGACAGCCCTCCG AAGGA

Gal4_2                MGSVNNDVWNBNBNSCG
          368: AAGTA CGGATTAGAAGCCGCCG AGCGG
          387: CCGAG CGGGTGACAGCCCTCCG AAGGA
          405: TCCGA AGGAAGACTCTCCTCCG TGCGT
          456: TGAAA CGCAGATGTGCCTCGCG CCGCA
          469: TGCCT CGCGCCGCACTGCTCCG AACAA

Gal4_2 /Rev           CGSNVNVNWBHNNBSCK
          469: TGCCT CGCGCCGCACTGCTCCG AACAA

          m81879.em  ck: 9482  len: 2,845 !

Gal4_2                MGSVNNDVWNBNBNSCG
          217: CAATT CGGAAAGCTTCCTTCCG GGATG
          331: GGCAC CGGCGGTCTTTCGTCCG TGCGG
          350: CCGTG CGGAGATATCTGCGCCG TTCAG
          414: AGTAT CGGGGCGGATCACTCCG AACCG

Gal4_2 /Rev           CGSNVNVNWBHNNBSCK
          217: CAATT CGGAAAGCTTCCTTCCG GGATG
          414: AGTAT CGGGGCGGATCACTCCG AACCG

          x00215.em  ck: 6177  len: 1,008 !

Gal4_1                CGGVBVRSWBYYSWCCG
          746: GCGCT CGGACAACTGTTGACCG TGATC
```

```
Gal4_2                  MGSVNNDVWNBNBNSCG
          659: TACTT CGGAGCACTGTTGAGCG AAGGC
          746: GCGCT CGGACAACTGTTGACCG TGATC

          x01667.em  ck: 7249   len: 2,457 !

Gal4_2                  MGSVNNDVWNBNBNSCG
          561: TTTAC CGGCGCACTCTCGCCCG AACGA

Gal4_2 /Rev             CGSNVNVNWBHNNBSCK
          561: TTTAC CGGCGCACTCTCGCCCG AACGA
          755: ATCTT CGGTCTCAACCGTGCCT AATGC

          x03102.em  ck: 5510   len: 2,812 !

Gal4_2                  MGSVNNDVWNBNBNSCG
          208: TCATT CGGCCATATGTCTTCCG AAAGA

     Total finds:          20
    Total length:      10,029
 Total sequences:           5
        CPU time:       00.35
```

En búsqueda en posiciones inversas-complementarias:

```
FINDPATTERNS on *.rev allowing 0 mismatches


          k02115.rev   ck: 933    len: 907    !    REVERSE-COMPLEMENT  of:
k02115.gb_
pl  check: 2280  from: 1  to: 907

Gal1 /Rev               CGGWSRRVWSYBVBCCG
          505: TCCTT CGGAGGGCTGTCACCCG CTCGG

Gal2                    MGSVNNDVWNBNBNSCG
          423: TTGTT CGGAGCAGTGCGGCGCG AGGCA

Gal2 /Rev               CGSNVNVNWBHNNBSCK
          423: TTGTT CGGAGCAGTGCGGCGCG AGGCA
          436: TGCGG CGCGAGGCACATCTGCG TTTCA
          487: ACGCA CGGAGGAGAGTCTTCCT TCGGA
          505: TCCTT CGGAGGGCTGTCACCCG CTCGG
          524: CCGCT CGGCGGCTTCTAATCCG TACTT



          m81879.rev   ck: 2379   len: 2,845 !    REVERSE-COMPLEMENT  of:
m81879.gb_
```

```
pl  check: 9482  from: 1  to: 2845

Gal2                    MGSVNNDVWNBNBNSCG
         2,416: CGGTT CGGAGTGATCCGCCCCG ATACT
         2,613: CATCC CGGAAGGAAGCTTTCCG AATTG

Gal2 /Rev               CGSNVNVNWBHNNBSCK
         2,416: CGGTT CGGAGTGATCCGCCCCG ATACT
         2,480: CTGAA CGGCGCAGATATCTCCG CACGG
         2,499: CCGCA CGGACGAAAGACCGCCG GTGCC
         2,613: CATCC CGGAAGGAAGCTTTCCG AATTG



       x00215.rev    ck: 3848    len: 1,008  !      REVERSE-COMPLEMENT  of:
x00215.gb_
pl  check: 6177  from: 1  to: 1008

Gal1 /Rev               CGGWSRRVWSYBVBCCG
           247: GATCA CGGTCAACAGTTGTCCG AGCGC

Gal2 /Rev               CGSNVNVNWBHNNBSCK
           247: GATCA CGGTCAACAGTTGTCCG AGCGC
           334: GCCTT CGCTCAACAGTGCTCCG AAGTA



      x01667.rev    ck: 4644    len: 2,457  !      REVERSE-COMPLEMENT  of:
x01667.gb_
pl  check: 7249  from: 1  to: 2457

Gal2                    MGSVNNDVWNBNBNSCG
         1,687: GCATT AGGCACGGTTGAGACCG AAGAT
         1,881: TCGTT CGGGCGAGAGTGCGCCG GTAAA

Gal2 /Rev               CGSNVNVNWBHNNBSCK
         1,881: TCGTT CGGGCGAGAGTGCGCCG GTAAA



      x03102.rev    ck: 8248    len: 2,812  !      REVERSE-COMPLEMENT  of:
x03102.gb_
pl  check: 5510  from: 1  to: 2812

Gal2 /Rev               CGSNVNVNWBHNNBSCK
         2,589: TCTTT CGGAAGACATATGGCCG AATGA

     Total finds:          20
    Total length:      10,029
 Total sequences:           5
        CPU time:       00.33
```