UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

CENTRO DE INVESTIGACION SOBRE FIJACION DE NITROGENO

ESTRATEGIAS PARA MEJORAR LA DETECCION DE SITIOS REGULADORES *CIS* DE LA PROTEINA SP1 EN DNA DE VERTEBRADOS

T E S I S

QUE PARA OBTENER EL GRADO DE:
MAESTRO EN INVESTIGACION BIOMEDICA BASICA

P R E S E N T A :

VICTORIA F. DOMINGUEZ DEL ANGEL

CUERNAVACA, MOR. 1998

El trabajo de la tesis se realizó bajo la asesoría del Dr. Julio Collado y el Dr. Jacques Van Helden, en el Programa de Biología Molecular Computacional del Centro de Investigación sobre Fijación del Nitrógeno perteneciente a la Universidad Nacional Autónoma de México. Dicho trabajo se efectuó de Septiembre de 1996 a Octubre de 1998.

Durante el desarrollo del proyecto conté con el apoyo de la Beca del Consejo Nacional de Ciencia y Tecnología

Uno no pertenece a ninguna parte
mientras no tenga un muerto bajo tierra.
-José Arcadio Buendía-

A Victoria Del Angel,
por todo el tiempo que has estado aquí.

AGRADECIMIENTOS:

Agradezco a los Dres. Julio Collado, Jacques Van Helden, Lorenzo Segovia, Martha Vázquez y Denis Thieffry por su ayuda en la construcción y elaboración del proyecto.

A los Dres. David Romero, Guillermo Dávila, Alicia González y en especial a Enrique Merino por sus comentarios durante la revisión de la tesis.

A mis compañeros del laboratorio: Konchita, Edgar, Victor M., Victor T., Toño, Ernesto, Araceli, Alma, Arturo, Rosa María, Hely, Alberto y Ulises por todos los momentos gratos de nuestra convivencia diaria. A la gente de Nitrógeno: Mari, Edith y Lety, por su apoyo en la biblioteca; a los administrativos Felipe, Alejandra, Uriel y Lolita, por hacerme fácil lo que suele ser muy complicado; a los vigilantes Juan Lemus y Luisa, por la compañía en las horas duras de trabajo.

Agradezco a todas las personas que han estado conmigo durante mi maestría, recordándome mi casi ningún derecho a sentirme sola, deprimida o con la bandera a media asta. En especial quiero mencionar a Karine, Maru y Ana, por la sólida amistad que ni el tiempo ni la distancia han diluido. A Burama por sus charlas y su genial compañía. A Luis, por ser una de las personas más increíbles que he conocido. A la tía Espe, por su confianza y hospitalidad. Al tarahumara, por volverte importante en mi vida.

Finalizo agradeciendo a toda mi gran familia, a mis hermanos Ezequiel y Ade, por sus críticas y consejos. A mis sobrinos Olimpia, María, Tomasito, Emilio y Nicolas por toda la alegría que me han brindado. A Claudia y Tomás, por ser una gran pareja con la que siempre se puede contar.

RESUMEN:

La regulación del inicio de la transcripción es un proceso que involucra diversas interacciones entre los elementos *cis* en DNA y los elementos reguladores en *trans*. La información disponible sobre estas interacciones ha crecido exponencialmente y junto con esto, se han desarrollado herramientas computacionales que buscan sitios potenciales para regulación en secuencias de DNA.

Los diferentes programas de búsqueda de patrones en secuencias, generan (en la mayoría de los casos) un gran número de sitios potenciales, cuya comprobación experimental se vuelve prácticamente imposible.

Nuestra propuesta en este trabajo es buscar propiedades intrínsecas de un sistema en particular para integrarlas en los algoritmos de búsqueda y de esta manera mejorar su poder predictivo.

El sistema en el cual se trabajó fueron los sitios de la proteína Sp1, la cual regula un amplio espectro de genes en organismos vertebrados y cuenta con un gran número de secuencias reportadas en la literatura. Para la búsqueda de patrones se utilizó el método de las matrices de peso, que asigna valores basados en la teoría de la información.

A partir de los sitios de Sp1 se integraron propiedades de dos tipos al algoritmo de búsqueda; la primera utiliza informacion interna de los sitios con los cuales se construyó la matriz (especificidad por organismo) y la segunda utiliza información contextual (posición de los sitios).

El Inicio de la Transcripción en Organismos Eucariontes

El inició de la transcripción es un mecanismo clave en la regulación de la expresión génica. Para su activación se requiere que la RNA polimerasa reciba señales de activadores promotor - específicos e inicie la transcripción en regiones precisas del genoma (Tansey, 1997). En procariontes, el factor σ es el que lleva a cabo la acción de acoplamiento con el DNA, interactuando con el promotor, asociándose con el centro de la polimerasa y sirviendo de blanco (junto con otros componentes) para factores transcripcionales. En eucariontes cada organismo presenta tres tipos de polimerasas (I, II y III). La polimerasa II transcribe (entre otros) el RNA mensajero, interactúa con su equivalente σ, TFIID, que está constituido por una proteína que se une a la caja TATA, *TATA box-binding protein* (TBP) y por los factores asociados a TBP, *TBP-associated factors* (TAF's).

La maquinaria de transcripción necesita de manera general los siguiente elementos:

i)Estructura del promotor: Elemento TATA, Elemento Iniciador, UAS/URS y el elemento Poli(dA:dT)

ii)RNA Polimerasa: Dominio repetido en el carboxilo terminal, Holoenzima.

iii)Factores generales de la Transcripción: GTF's incluye TBP (con sus múltiples interacciones),TFIIB, TFIIE, TFIIF y TFIIH

iv)Coactivadores transcripcionales: TAF's, SRB/mediador, TFIIA, etc..

v)Represores generales de la transcripción.

Estructura del promotor:

Los elementos en *cis* se pueden dividir en elementos centrales (*core elements*) y elementos reguladores. En los elementos centrales se ensamblan el

complejo de preiniciación de la transcripción *preinitiation complex* (PIC) que incluyen a la secuencia TATA, la secuencia de Inr y el *downstream promoter element* (DPE), este último se encuentra aproximadamente a 30 pares de bases río abajo del inicio de la transcripción.

Los elementos reguladores son gene específicos, se localizan, en la mayoría de los casos río arriba de los elementos centrales. Este control puede ser activando, *upstream activation sequences* (UAS) o reprimiendo *upstream repression sequences* (URS), ya sea por *enhancer* o represores respectivamente.

El elemento **TATA**: es una secuencia a la cual se va unir el factor TBP, se encuentra en un gran número de promotores a una distancia más o menos fija (entre 25 y 30 pares de bases antes del inicio de la transcripción). El consenso de la caja TATA se define como TATAAA. Algunas variantes en este consenso puede provocar disminución de la actividad de la transcripción (como ejemplo TGTAAA). En levadura se han observado combinaciones de diferentes cajas TATA´s para un mismo promotor, (la proteína presentan diferente afinidad dependiendo de la caja) éstas se utilizan según los niveles de transcritos que se necesiten, tal es el caso del promotor *HIS3* (Hampsey, 1998).

Otro tipo de promotores son los conocidos con el nombre de *"TATA-less promoter"* cuya caja TATA es muy poco conservada y muestra una interacción muy débil con el factor TBP por lo que el papel de reclutamiento para el ensamblamiento del PIC, corre a cargo de otro(s) elemento(s) de la maquinaria central de reconocimiento.

El elemento **Iniciador** es una secuencia de DNA que está contenida en el inicio de la transcripción (Lewin, 1997). No se ha definido un consenso y se sugiere que puede ser determina por distancias fijas a partir de la caja TATA. Diferentes proteínas pueden unirse al elemento Inr, como por ejemplo CIF, YY1, E2F, TFII-I y USF. En un principio Inr se encontró en promotores *"TATA-less promoter"* posteriormente se vio que se encontraba en promotores con o sin TATA. Su función no es muy clara, pero el análisis en algunos promotores, como

es el caso de GAL 80 en levadura, demostró que la transcripción puede ser iniciada por dos rutas diferentes, una dependiente de Inr y la otra dependiente de TATA.

Elementos **UAS y URS.** Los elementos UAS *upstream activation sequences,* funcionan como sitios de unión para activadores transcripcionales específicos (*enhancers*), pueden actuar en cualquier orientación y a distancias variables con respecto al inicio de la transcripción. Una vez que es asociado el activador al elemento UAS, se facilita el ensamblamiento de PIC e interactúa con este a través de los GTF's o por medio de coactivadores. La eliminación de este elemento causa una disminución drástica en los niveles de transcrito.

En los elementos URS *upstream repression sequences,* se unen los represores gene específicos de la transcripción, los cuales, afectan la transcripción por diferentes mecanismos: interferencia con el UAS o interferencia con el dominio de activación (ya sea contactando con la proteína activadora o compitiendo por la interacción con la maquinaria de la transcripción). El complejo URS-represor puede mediar la represión de manera indirecta. El más común de estos procesos es el reclutamiento de proteínas (Ssn6-Tup1, Sin3-Rpd3) que están involucradas con las histonas e impiden la transcripción por causas estructurales.

Elementos **Poli(dA-dT)**: Las secuencias homopoliméricas dA - dT que se encuentran frecuentemente en levadura, tienen una estructura característica que impide el ensamblamiento o estabilidad en el nucleosoma. Su papel por tanto es estimular la transcripción a causa de la estructura intrínseca de la secuencia (Winter, 1989).

RNA Polimerasa II:

Las diferentes RNA polimerasas II pueden variar en su composición de subunidades, desde 8 hasta 12, dependiendo del organismo. Estas subunidades son codificadas por un conjunto de genes *RPB*. Las dos subunidades más grandes y conservadas en los grupos de eucariontes son las codificadas por RPB1

(~200 kDa) y RPB2 (~150 kDa), las cuales son homólogas a las subunidades β´y β respectivamente de la RNA polimerasa en bacterias. RPB3 está relacionado en secuencia de aminoácidos, tamaño y estequeometría con la subunidad α de bacterias. Con la subunidad σ, se encuentra similitud en estructura y función con algunos GTF´s. RPB1 y RPB2 confieren selectividad para encontrar el inicio de la transcripción y están involucradas en el proceso de elongación.

Característico de la RNA pol II es la presencia de heptapéptidos, repetidos en *tandem*, en la región carboxilo terminal. Este ***carboxil-terminal repeat*** (CTD) contiene una secuencia consenso de Tyr-Ser-Pro-Thr-Ser-Pro-Ser que es altamente conservada entre organismos eucariontes. La subunidad RPB1 incluye en levadura, 26 o 27 repeticiones, el CTD en *C. elegans* tiene 34 repeticiones, *Drosophila* tiene 43 repeticiones y el CTD en humano tiene 52.

El CTD está relacionado con las dos formas en que se encuentra la polimerasa *in vitro*, que se designa IIO cuando CTD está fosforilado y IIA cuando no está fosforilado. La forma IIA está presente en el ensamblamiento de PIC, mientras que la IIO se encuentra en el complejo de elongación. TFIIH y otras cinasas median la fosforilación. La estructura desfosforilada está mediada por TFIIB y TFIIF. TFIIF estimula la fosfatación mientras que TFIIB inhibe la estimulación de TFIIF. La CTD fosforilada, IIO, está involucrada en otros procesos como el procesamiento del pre-mRNA y en su forma hiperfosforilada facilita las interacciones electrostáticas de las cargas positivas en ciertas proteínas de *splicing*. Otros procesos donde se encuentra implicada, son en la adición de cap en el extremo 5´ y en la adición de poly(A) en el extremo 3´.

Factores Generales de la Transcripción:

Los GTF´s incluyen TBP, TFIIB, TFIIE, TFIIF y TFIIH (figura 1)

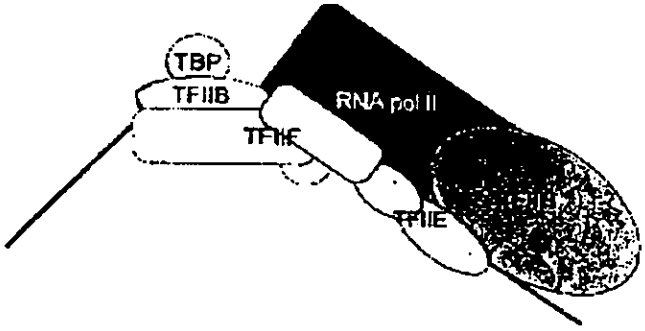


Figura 1. Ensamblaje del Complejo de Pre iniciación (PIC). El primer paso es el reconocimiento de TBP al DNA, seguido de la asociación de TFIIB, RNA polII/TFIIF, TFIIE y TFIIH

TBP es una subunidad del complejo TFIID. En levadura, TBP es un monómero de 27 kDa, es esencial para la expresión y reconocimiento del promotor en gran número de genes. El reconocimiento de TBP al DNA se da a través del surco menor. El plegamiento de la proteína es semejante a una silla de montar (Phillips, 1993). Su región C-terminal consiste en dos regiones repetidas, que muestran un eje de simetría (dominios α y β). Cada dominio muestra dos hélices alfa y cinco hojas beta antiparalelas conectadas en el siguiente orden S1-H1-S2-S3-S4-S5-H2. El dominio de unión es una curva formada de hojas-β´s y cuyo lado convexo, donde se muestran las α hélices, queda libre para interaccionar con otros factores.

El reclutamiento de TBP en el promotor ocurre en dos pasos: el primero involucra una lenta asociación de TBP-TATA (paso limitante en la activación), seguido de un rápido cambio de conformación. Frecuentemente TBP se encuentra como dímero, su disociación es lenta pero necesaria para la unión a la secuencia TATA.

Una vez efectuada la unión a la secuencia TATA, se presentan distintas interacciones con elementos del PIC, TBP-TFIIA, TBP-TFIIB, TBP-TAF's, etc.

La interacción con TFIIA se da en la superficie convexa de TBP, en las hélices localizadas en la parte terminal (H2, H2'). Este complejo estimula la activación para algunos activadores transcripcionales.

TFIIB interactúa con la región carboxilo terminal de TBP. La formación del complejo DNA-TBP-TFIIB (figura 2) es importante para el ensamblamiento de PIC, ya que intervienen en el reclutamiento de otros GTF's.

TFIIB: El monómero de TFIIB en levadura, es de 38 kDa y es codificado por el gene *SUA7* . Este factor entra al PIC después de TBP y es un pre-requisito para el reclutamiento de la RNA pol II. TFIIB interactúa directamente con TBP, RNA pol II, con las subunidades RAP30 y RAP74 de TFIIF, con TAF□40 de TFIID y es blanco de muchos activadores transcripcionales gene específicos. En su extremo amino terminal, se encuentra un motivo de unión a zinc, *zinc ribbon*, el cual es filogenéticamente la parte más conservada de la proteína y la que interactúa con el complejo TFIIF-RNA pol II. Su extremo carboxilo está constituido por dos secuencias repetidas imperfectas, con 5 hélices α, las cuales se pliegan y forman un motivo resistente a proteasas (en este motivo se da la unión con TBP). Con TBP juegan un papel crítico en la selección del inicio de la transcripción. La distancia del sitio catalítico de RNA pol II - TBP es ~110 A°, que es equivalentes a 32 pares de bases en el DNA de forma B. Esta es la distancia aproximada entre la caja TATA y el inicio de la transcripción en gran parte de los promotores conocidos en eucariontes.

TFIIF: consta de dos subunidades identificadas como RAP30 y RAP74. Tiene características que están presentes en los factores σ de bacterias. TFIIF suspende las interacciones inespecíficas entre la RNA polimerasa II y el DNA. Contribuye a la estabilización del PIC. Presenta interacciones funcionales con TFIIB, las cuales están involucradas en el cambio de conformación de la RNA pol II para los procesos de iniciación y elongación.

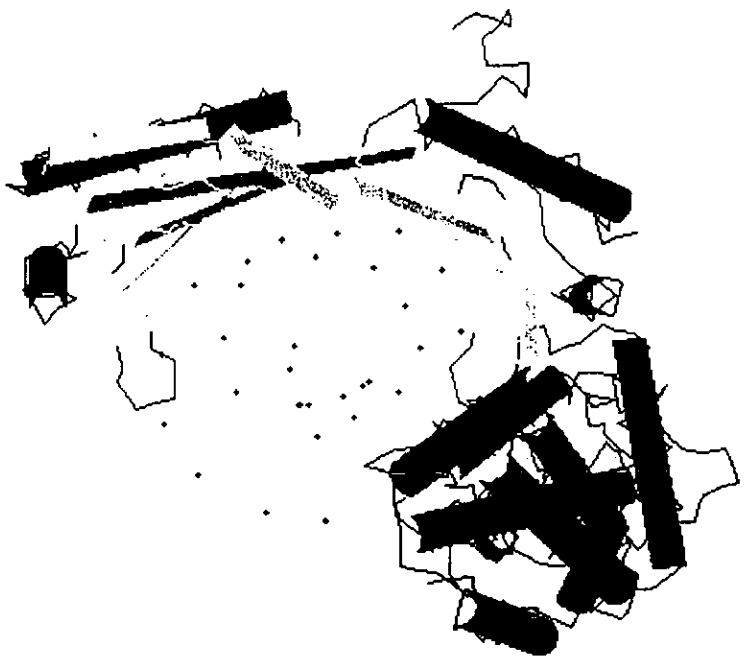


Figura 2. Complejo proteico DNA-TBP-TFIIB. Importante en el ensamblaje de PIC para el inició de la transcripción.

TFIIE: Este factor se integra al PIC después de la RNA pol II y antes que TFIIH. Interactúa con la forma desfosforilada (II A) de la pol II y con TFIIF. Ayuda al reclutamiento de TFIIH y es blanco de algunos activadores gene específicos. TFIIE está compuesto de una subunidad de 56-kDa (TFIIE-α) y otra de 34kDa (TFIIE-β) formando un heterotetrámero $\alpha_2\beta_2$. Ambas subunidades tienen motivos estructurales bien definidos, por ejemplo *zinc ribbon* (C-X_2-C-X_{21}-C-X_2-C), una secuencia en TFIIE-α con el consenso de una proteína cinasa y un consenso en TFIIE-β para unión al DNA. El complejo TFIIE-TFIIB interactúa para estabilizar la conversión de DNA en región iniciadora. A través del reclutamiento y activación de TFIIH, controla la formación de PIC.

TFIIH: Es el más complejo de los GTF's, consiste en nueve subunidades con una masa total de aproximadamente 500 kDa (comparable a la masa de la Pol II). Es el único GTF que se conoce con actividad enzimática, entre las cuales se encuentra ATPasa - helicasa DNA-dependientes y cinasa CTD. TFIIH juega un papel crítico en los estados de iniciación y postiniciación de la transcripción. Está involucrado en la hidrólisis de ATP y en la estructura del pretemplado del DNA. Aparece junto con TFIIE y ATP mediando la formación del complejo abierto. También regula la transición del inicio de la transcripción a la elongación. Funciona como componente esencial de NER *nucleotide excision repair*, el cual es un mecanismo capaz de remover daños creados a lo largo de una cadena de DNA (como asociaciones entre primigenias) (Alberos, 1994). TFIIH tiene un doble papel al intervenir en el inicio de la transcripción y en la reparación del DNA (ésto va de acuerdo con la idea que las regiones con una frecuencia alta de reparación son altamente transcritas) y se ha sugerido que en cada uno de estos mecanismos tiene una conformación diferente: cuando está asociada con la RNA pol II, la parte central de TFIIH interactúa con TFIIK, pero ante la presencia de un DNA dañado, existe un cambio de conformación que remueve a TFIIK e integra a las proteínas del complejo NER.

Coactivadores Transcripcionales:

Estas moléculas son requeridas como mediadores o adaptadores de la activación transcripcional. Se diferencía de los GTF's en que son dispensables para los niveles basales de la transcripción y ninguno aparece unido a la secuencia de DNA de manera específica.

Los coactivadores aparecen como puente para las interacciones entre proteínas activadoras gene-específicas y los GTF's; también interaccionan con proteínas que intervienen en la remodelación de la cromatina. Entre las diferentes clases de coactivadores se encuentran los TAF's (componentes de TFIID), SRB/mediador, TFIIA, SAGA (acetilación en histonas) y los complejos relacionados con la remodelación de cromatina (figura 3).

Factores asociados a TBP.- La asociación de TBP con los diferentes TAF's, crean una alta versatilidad de complejos proteicos que interactúan con: el promotor, diferentes reguladores transcripcionales y la RNA polimerasa.

Los diferentes TAFs se asignan dependiendo de sus pesos moleculares. Dentro del complejo TFIID, existen interacciones de TBP con diferentes TAF's y una multitud de interacciones TAF-TAF que forman una red proteica compleja, altamente versátil, que facilita la activación de la transcripción a través de una gran variedad de mecanismos (Chen, 1994)

Los elementos TAF's están involucrados en diferentes actividades: (i) Reconocimiento de promotores, (ii) Topología del promotor, (iii) Catálisis y (iv) Blanco para los dominios de activación (Tansey, 1997).

(i) En el reconocimiento del promotor, TFIID interacciona con varios elementos centrales del promotor, que incluyen la caja TATA y elementos río abajo. Los TAF's pueden interaccionan con TBP y con el iniciador (Inr). Las diferentes combinaciones de TAF's y TBP para construir TFIID, permiten al complejo reconocer una gran diversidad de arreglos en promotores.

(ii) Diversos estudios sugieren que TFIID induce alteraciones en la topología del promotor. Esta idea es consistente con observaciones hechas en algunos TAF's,

los cuales presentan homología a nivel de secuencia y estructura con proteínas histónicas (forman complejos similares a los nucleosomas). El compactamiento del DNA permite múltiples contactos entre TFIID y la parte central de la maquinaria basal de la transcripción. Interaccionan con otros factores basales que están involucrados con los cambios en la topología del promotor.

(iii) El complejo enzimático posee actividades catalíticas, las dos conocidas son: acetilación de histonas y fosforilación de factores basales. La acetilación de histonas es un importante paso en la conversión de cromatina inactiva a una forma transcripcional activa (aparentemente al acetilar los residuos de lisina se debilitan las interacciones histonas-DNA), TAF□250 posee actividad de *histone acetyl-transferase* (HAT) y juega un importante papel en esta conversión. TAF□250 posee también actividad de cinasa y fosforila al factor basal TFIIF. TFIIF es un importante blanco para la fosforilación, está íntimamente asociada con la polimerasa II, ayuda al reclutamiento del complejo y está involucrado en el paso iniciación/elongación.

(iv) *In vitro,* se ha visto que los TAF´s sirven como blanco para el reconocimiento de una gran variedad de dominios de activación, dando lugar a una activación dominio-específica entre proteínas reguladoras y TFIID. Con algunos factores se ha visto también efectos sinérgicos (Chen, 1994). Diferentes estructuras y ensamblamientos de TFIID, revelan diferentes requerimientos en coactivadores para diversos factores de transcripción promotor - específico como pueden ser: Sp1, *neurogenic element-binding transcription factor 1* (NFT-1) y *CCAAT-binding transcription factor* (CTF).

Otras características que se han encontrado en los TAF´s son: al analizarse en diversos organismos a nivel de estructura y función, éstas son muy conservadas a lo largo de la evolución. Interactúan de manera individual con los dominios de activación para algunos factores transcripcionales gene-específicos. Cuando activadores específicos contactan con TAF´s, se transmiten señales a la

maquinaria basal (los TAF´s TAFII110 y TAFII40 interactúan con los elementos basales TFIIA y TFIIB). Se asocian con TBP en su superficie convexa.

Se han hecho análisis experimentales para el estudio del ensamblamiento de los diferentes TAF´s, y de manera general se ha visto que existen dos tipos de TAF´s: los que contactan directamente con TBP (TAFII250, TAFII150 y TAFII30 alfa) y los que contactan con otros TAF´s (los necesitan para su reclutamiento y anclaje). Por ejemplo TAFII150, TAFII110, TAFII60, TAFII130 alfa y TAFII30 beta se unen eficientemente al TAFII250. Posteriormente, TAFII110 ayuda al reclutamiento de TAFII80, mientras que TAFII60, permite que TAFII40 interactúe con TFIID.

La integración del complejo TFIID, sigue una sucesión de pasos, que en su mayoría empiezan con la formación dimérica entre TFIID y TAFII250. Para algunos TAF´s el orden es muy importante, tal es el caso de TAFII110, el cual no se integra al complejo TBP-TAF en ausencia de TAFII250. Otros ejemplos son: TAFII40 que no se integran si no se integró antes TAFII60, y TAFII80, que sólo se integra si se han ensamblado TAFII110, TAFII150, TAFII30 alfa.

Experimentos *in vitro* mostraron que una activación a través de Sp1, necesita como mínimo un complejo cuádruple que conste de TBP y de tres TAF´s (TAFII110, TAFII150 y TAFII250). Sp1 interactúa (de manera directa) únicamente con TAFII110.

SRB/Mediador. Es un intermediario molecular, que media la interacción entre activadores y los componentes centrales de la maquinaria transcripcional. Las funciones de SRB/mediador: i) estimulan la transcripción basal, ii) responde a los activadores transcripcionales y iii) estimula la fosforilación de la RNA pol II por TFIIH. SRB/mediador juega un papel más general en la activación de la transcripción, a diferencia de los TAF´s que son coactivadores gene-específicos. Otra diferencia con los TAF´s, es la posibilidad de conferir efectos negativos en la expresión génica, es por esto que se le considera un regulador (en lugar de activador) de la transcripción (Bjorklund, 1996).

TFIIA: Se asocia con PIC a través de TBP. Estimula la transcripción de varias maneras; estabilizando la interacción TBP-TATA, interactuando con específicos reguladores transcripcionales (TAF_{II} 110 y los coactivadores PC4 y HMG2) y jugando un importante papel como anti represor. TFIIA se asocia con el complejo TBP-TATA del lado opuesto a TFIIB y su posición dentro del complejo se encuentra localizada en el lado *"upstream"* de la caja TATA.

Histona Acetil-transferasa: Entre los mecanismos de coactivación transcripcional podemos encontrar a la acetilación de histonas en el nucleosoma. Un complejo proteico que lleva a cabo este mecanismo es el SAGA (Spt-Ada-Gcn5-Acetiltranferasa), el cual se encuentra integrado por los productos de los genes *GCN5, ADA2, ADA3, STP7, SPT20/ADA5.* Presumiblemente, la acetilación debilita las interacciones entre el DNA y las histonas, relajando el efecto represivo del nucleosoma. Esto puede explicar la correlación directa que existe entre la acetilación de las histonas y la activación génica (Wolffe, 1996).

Remodelación del Complejo - Cromatina: Existen otros complejos que son dependientes de ATP, como SWI/SNF, que alteran la organización nucleosomal del DNA. Las modificaciones hechas por la proteína permiten al PIC ensamblarse a la región promotora y de esta manera ayuda a la activación transcripcional. Otros factores remodeladores de la cromatina en *Drosophila* son NURF y CHRAC; pero en general estos coactivadores se pueden encontrar desde levadura hasta humano.

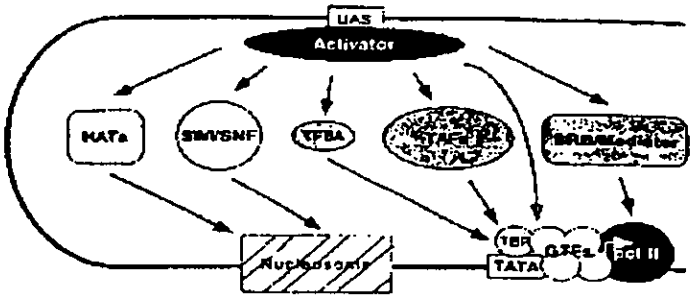


Figura 3. Esquema de los mecanismos de activación. Con sus diferentes elementos: GTF's, activadores, coáctivadores y maquinaria transcripcional (Hampsey, 1998).

Represores generales de la Transcripción

Los represores transcripcionales son comparables a los coactivadores, pero con el efecto contrario. Se pueden reconocer dos clases: una clase opera a través de los elementos centrales de la maquinaria transcripcional (Mot 1 y el complejo Ccr-NOT), la segunda se encuentra funcionalmente ligada a la cromatina (BUR y Sot4-Spt6), donde también se incluyen histonas, proteínas relacionadas a las histonas y con la desacetilación de histonas (Hampsey, 1997) (figura 4).

Mot1 es identificado como un ADI (*ATP-dependent inhibitor)* que inhibe el reconocimiento de TBP al DNA de manera ATP-dependiente. Tiene 170 -kDa y se une a TBP formando un complejo distinto al que originalmente se forma con TFIID, denotado como B-TFIID, el cual no responde a los activadores transcripcionales y posee actividad de ATPasa. Mot1 puede funcionar como represor o activador de la transcripción. Funcionalmente interactúa con Spt3, TFIIA y el complejo NOT.

Complejos **Ccr4-NOT**. La represión se da por la interacción con la maquinaria central de la transcripción. Existen diferentes proteínas NOT (*negative on TATA),* generalmente afectan la transcripción basal y funcionan como represores en promotores con una secuencia TATA débil. Se relaciona con MOT1, Spt3 y TFIIA.

Algunas proteínas NOT forman parte del complejo Ccr4, el cual es de 1.2Mda e incluye otros polipéptidos como Caf1 y Dbf2.

Proteínas **BUR** (*bypass of UAS requirement*). Se encuentran ampliamente distribuidas en levadura y reprimen la transcripción a través de dos mecanismos diferentes, el primero es a nivel de cromatina y el otro es por interacción con los GTF's. De manera general los supresores BUR1, BUR2, BUR4 y BUR5 están funcionalmente relacionados con la cromatina; se conoce que BUR5 es una subunidad del nucleosama (H3), BUR1 se identificó como una proteína cinasa que fosforila H4 y Rpb1 para la represión del gene *SUC2*. BUR3 y BUR6 reprimen la transcripción global de manera cromatina-independiente, interactuando con TBP y bloqueando la asociación de éste con TFIIA y TFIIB. BUR6 tiene su homólogo en humano, **DRAP1** que bloquea el inicio de la transcripción interactuando con los mismos factores.

Spt4-Spt6. Los genes *SPT4, SPT5* y *SPT6* pertenecen a la familia *SPT* que codifican para histonas y proteínas que afectan las funciones de la cromatina. De esta familia los implicados en la regulación son: Spt4, Spt5 y Spt6 ya que afectan la estructura de la cromatina. Interactúan con SWI/SNF, complejo remodelador de la cromatina. Spt6 tiene interacción directa con las histonas, participa en el ensamblamiento del nucleosoma como aceptor/donador de histonas durante la reorganización de la cromatina.

Desacetilación de Histonas. El mecanismo opuesto a la acetilación de histonas es el *histone deacetylases* (HDAs), el cual funciona como represor transcripcional. Entre las proteínas involucradas en este mecanismo se encuentran Sin3, Rpd3, SAP18 y SAP30.

Ssn6-Tup1. Estas proteínas forman un complejo que reprime la transcripción. No hay unión al DNA pero su efecto es a través de las histonas. Tup1 interacciona directamente con el extremo amino-terminal de las histonas H3 y H4 y afectando la estructura de la cromatina en el proceso de la transcripción.

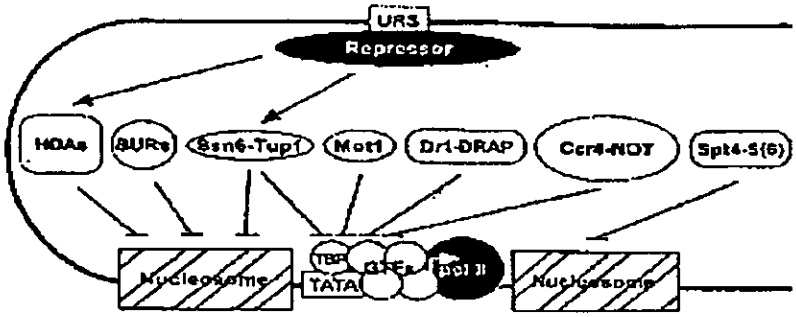


Figura 4. Esquema general de la represión transcripcional.Se muestran los dos tipos de mecanismos, a través de la interacción con el PIC o con los nucleosomas (Hampsey,1998).

Generalidades de la Proteína Sp1.

Los genes en organismos eucariontes son regulados por factores transcripcionales, los cuales interactúan directamente con los elementos reguladores en *cis*. Un factor transcripcional muy importante es Sp1 (*Specificity protein 1),* el cual activa un amplio rango de genes en virus y en organismos vertebrados (Lania, 1997).

Al purificar la proteína Sp1 humana, los resultados mostraron a un polipéptido de 778 aminoácidos. Consta de cuatro dominios de activación y cerca del carboxilo terminal se encuentra su dominio de reconocimiento al DNA, el cual contiene tres unidades de repetición en tándem que corresponden a **dedos de zinc.**

Sp1 se analizó por primera vez en los promotores tempranos del virus SV40, su activación depende de seis cajas con 10 pares de bases cada una

repetidas de manera consecutiva. Estos decanucleótidos son asimétricos y con un alto contenido de GC (Jones, 1986). Ensayos de digestión con DNasa I muestran que Sp1 protege las 6 cajas (Courey, 1989).

Sp1 puede activar uniéndose sólo a una caja sin importar su orientación con respecto al +1 del inicio de la transcripción. Frecuentemente se encuentran varios Sp1's interactuando entre sí, pueden estar adyacentes o a distancias considerables. Se une al surco mayor del DNA. Entre los genes que activa se encuentran un gran número de genes constitutivos de vertebrados, los cuales, por lo general tienen una caja TATA muy débil (*TATA less promoter*) y un sitio para el elemento iniciador (Inr) (Boam, 1995), dándole un nuevo mosaico de combinaciones y arreglos para regular la expresión temporal y espacialmente.

Los niveles de expresión de Sp1 pueden variar mucho de un tipo celular a otro (del orden de 100 veces). Se encuentra en altos niveles en células con estados tardíos de diferenciación (maduración de linfocitos, espermátidas, etc.). En general son células tejido específicas que involucran una diferenciación terminal. En el resto, Sp1 se encuentra en cantidades limitadas. Cambios fisiológicos pueden estimular los niveles de Sp1, alterando los patrones de expresión génica.

El gene de Sp1 es espacialmente regulado durante el desarrollo en vertebrados. Presenta autorregulación positiva, su región promotora contiene un arreglo extendido de sitios de pegado para Sp1.

Sp1 pasa por dos tipos de modificación post-traduccionales: glucosidación y fosforilación. En la glucosidación, a la molécula de Sp1 se le unen de 5-10 residuos de monosacáridos (posiblemente en las serinas o treoninas) que ayudan al paso por los poros nucleares, facilitando así su exportación al núcleo. La otra modificación es por fosforilación, las cinasas sólo fosforilan a Sp1 cuando se encuentra unido al DNA. Estos sitios de fosforilación (principalmente en los residuos de serinas) se dan en dos de las regiones activadoras ricas en glutaminas. Se plantea que la fosforilación induce un cambio conformacional que ayuda en el mecanismo de activación. Este tipo de regulación puede servir para

prevenir la formación de complejos no productivos entre Sp1 y el resto de los elementos de la transcripción (McKnigth, 1992).

Mecanismos de Activación para Sp1.

Existen varios modelos propuestos para la regulación positiva del inicio de la transcripción. La proteína Sp1 se ha utilizado como modelo para el estudio de estos mecanismos y sus diferentes interacciones. Sp1 presenta interacciones heterólogas (con diferentes proteínas y elementos de la maquinaria de transcripción) (Su, 1991) y homólogas (entre sitios proximales y distales). Sus diferentes dominios de activación pueden actuar con coactivadores, elementos del aparato basal de la transcripción, factores sitio específico y con otros Sp1's.

Sp1 forma complejos multiméricos importantes para la activación de la transcripción (tetrámeros). Dependiendo del contexto del promotor, Sp1 puede presentar activación simple o activación sinérgica (superactivación) (Pascal, 1991).

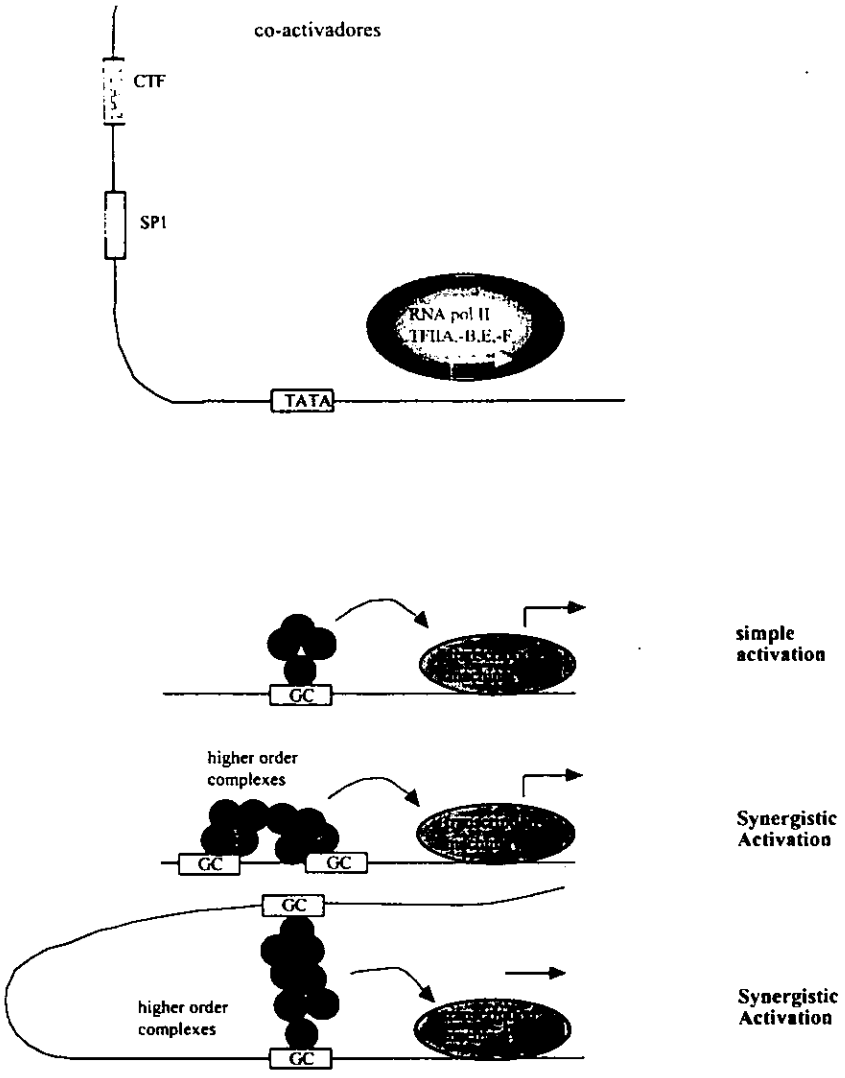


Figura 5. Diferentes interacciones y mecanismos de la proteína Sp1. Interacción indirecta (por coactivadores) o directa. Puede activar de manera simple o sinergica

En activación simple, cuando Sp1 interactúa con el aparato basal de la transcripción, lo hace a través de una interacción específica con TAFII110 (Chen, 1994) y con la intervención de TAFII250 y TAFII150 para ayudar a la estabilización de TFIID. En otros casos interactúa junto con otros factores sitio específicos (figura 5) como CTF y E2.

Para tener un efecto sinérgico en la activación, que involucre a Sp1, se requiere la interacción, ya sea proximal o distal, entre por lo menos dos sitios para esta proteína. Cuando los sitios son proximales, la interacción se hace de manera directa y cuando son distales, el DNA forma un *loop* que permite el contacto entre los Sp1's (Mastrangelo, 1991). Lo que el sistema presenta cuando se tienen múltiples sitios de Sp1 unidos al DNA, es una fuerte correlación entre la capacidad de Sp1 para formar complejos de homomultímeros y su capacidad para activar la transcripción sinergisticamente (Pascal, 1991).

Relación entre la estructura y función de los elementos en Sp1.

La regulación de la transcripción se logra a través de dos tipos de interacciones: proteína/DNA y proteína/proteína. La caracterización entre la estructura y la función de estas proteínas (activadores transcripcionales) constituyen un elemento básico para el entendimiento de la regulación. De manera general, las proteína reguladoras cuentan con dos estructuras bien definidas e independientes; la estructura de activación y la estructura de reconocimiento al DNA (Ptashne, 1992).

El Factor Sp1 provee un excelente modelo para el estudio de estas interacciones regulatorias. Su estructura de activación (figura 6) está constituida por 4 dominios (separados unos de otros): A, B, C y D. Dos de estos dominios A y B contienen una parte rica en glutamina (~30 % de glutamina) y otra rica en serina/treonina (~50 % de serina/treonina), son dos dominios de activación muy potentes y están involucrados en la formación de tetrámeros. El dominio D tiene

un papel importante en la transcripción sinérgica, forma complejos de orden superior entre diferentes sitios para Sp1 (Pascal, 1991).

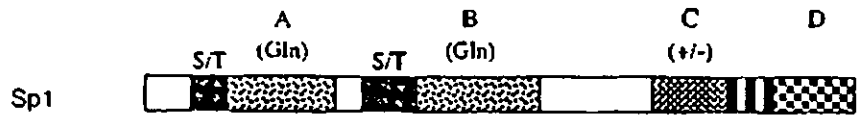


Figura 6. Diagrama básico de la distribución de los diferentes dominios de Sp1. Señala los dominios de activación A, B, C y D. La localización de los tres dedos de Zinc se marca con 3 barras negras verticales entre los dominios C y D. Se indican las regiones de la proteína que contienen altos porcentajes de serina y treonina S/T, glutamina (Gln) o cargadas (+/-)(Yieh, 1995).

Se propone que los dominios de activación interactúan con los coactivadores (TAFII-110) (Chen, 1994), con los componentes del aparato basal de la transcripción o con otros factores sitio específicos (E2F) (Karlseder, 1996).

Para su unión con el DNA no requiere formar dímeros, pero con sus estructuras de activación puede formar multímeros, los cuales son importantes para el efecto sinérgico en el inicio de la transcripción (figura 7).

Por medio de los análisis de deleción, se identificaron las cuatro regiones involucradas en la capacidad para activar la transcripción (Pascal, 1991). Como se mencionó anteriormente, los dominios A y B son ricos en glutaminas (dominio ampliamente extendido en otros factores) y al ser removidos, se observa un abatimiento en la activación; ambos dominios se utilizan para la multimerización e interacción con cofactores. La deleción de la región D, impide la formación de complejos de alto orden y tiene un severo efecto en la capacidad de la proteína para activar la transcripción sinérgica (figura 7).

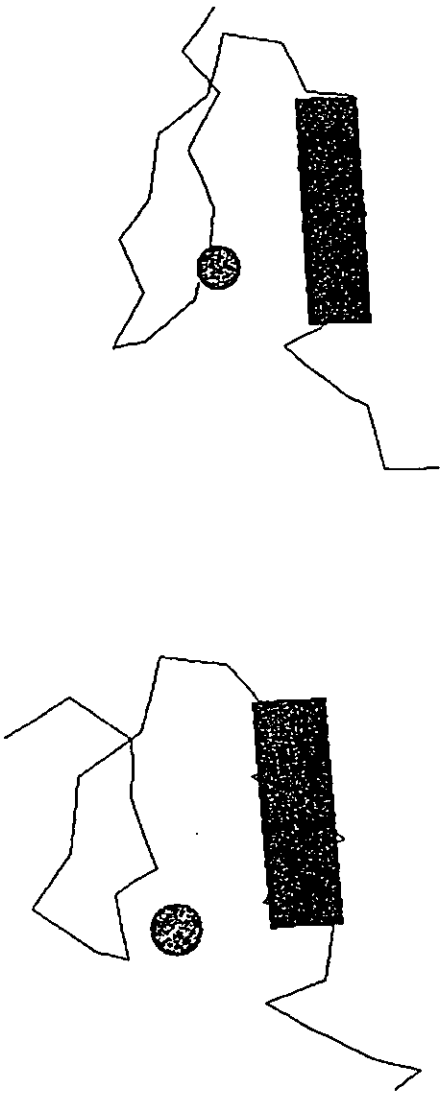


Figura 9.- Estructura tridimensional del segundo y tercer dedo de zinc pertenecientes a la proteína Sp1.

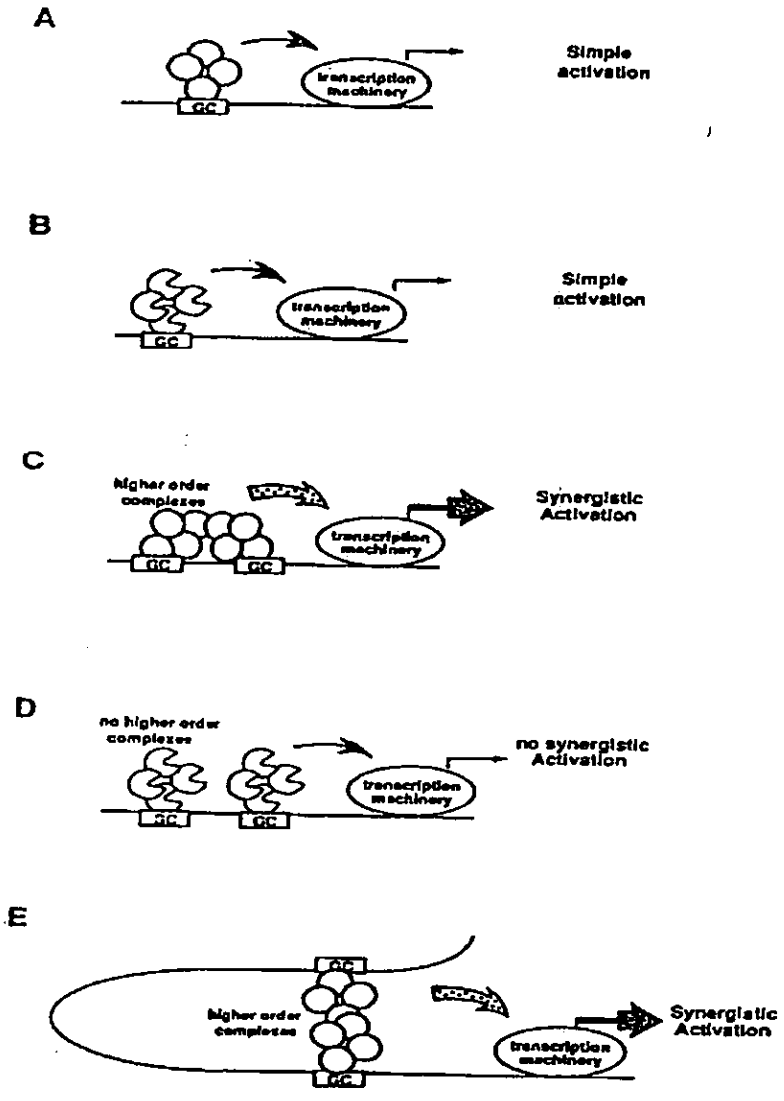


Figura 7. Modelo de activación para la proteína Sp1. Se muestran formas tetraédricas que contactan con el DNA en un solo sitio. Se tienen dos niveles de activación: simple (A,B y D) y sinérgica (C y D). Los Sp1 en forma de círculo representan toda la proteína, los incompletos (B y D) carecen del dominio D (Pascal, 1991)

Para la interacción proteína/DNA, Sp1 contiene 3 dedos de zinc del tipo Cys2-His2, cuya estructura consenso es Cys-X2_4-Cys-X12-His-X3-His (Fig. 8) (Yokono, 1998). Se une al surco mayor del DNA y reconoce un decanucleótido rico en GC, conocido como la caja GC 5´-(G/T)GGGCGG(G/A)(G/A)(C/T)-3´ (Kuwahara, 1993). La interacción con el DNA para cada uno de los dedos de zinc es diferente y produce un cambio local pero con una distorsión estructural significativa en la región 3´ de la cadena rica en guanina (Kuwahara, 1993). Los análisis muestran que los dedos dos y tres, que interactúan con la porción 5´ de la caja GC, contribuyen de una manera más fuerte en la afinidad total con el DNA que el primer dedo.

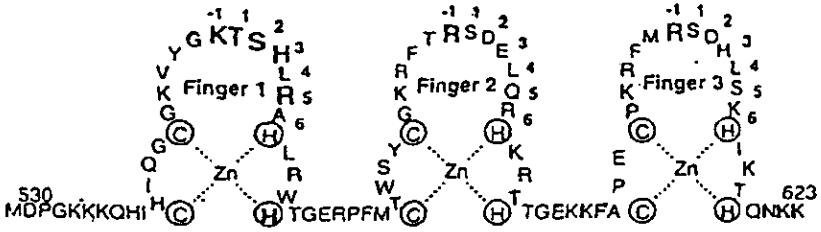


Fig. 8. Secuencias de aminoácidos en Sp1 (530-623). Las cisteínas e histidínas ligadas al átomo de zinc se encuentran encerradas en círculo. Los números muestran la posición en la hélice (Yokono, 1998)

Se definió por NMR la estructura tridimensional de los dos últimos dedos de zinc, sp1f2 y sp1f3 (figura 9).

El modelo estructural, muestra que cada dedo Cis2-His2 consiste en dos cadenas beta antiparalelas, ligadas con un *loop* que contiene los dos residuos *cis* seguido de una hélice alfa que contiene dos residuos de his orientados para coordinar el átomo de Zinc de manera tetraédrica (Narayan, 1997). Sp1f3 pertenece a una clase bien definida para los dedos de zinc, Cys-X2-Cys. Sp1f2 presenta un

arreglo Cys-X4-Cys, pero esta diferencia no altera la geometría local o la afinidad por iones metálicos.

Existe una parte central hidrofóbica formada por la hoja beta y la hélice alfa. Dentro de esta estructura se conservan los residuos Cys y His hacia el interior del dominio en posición, coordinada con el átomo de Zinc. Para el reconocimiento al DNA, se acopla el alfa hélice con el surco mayor. Cada dedo de zinc contacta con la doble cadena, aunque más específicamente con la cadena rica de G`s. (figura 10)

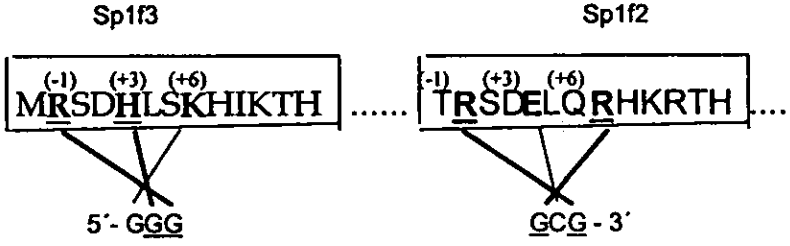


Fig 10. Interacciones propuestas entre la proteína Sp1 y una de las cadenas de DNA (Narayan,1997)

Familia de Reguladores Transcripcionales Sp

El aislamiento de cDNA´s mostraron proteínas que estaban relacionadas a nivel de secuencias y estructuras con Sp1. Estas proteínas se denominaron Sp2, Sp3 y Sp4 y junto con Sp1 forman una familia. Estas proteínas comparten dominios similares como los dedos de zinc y los de activación: glutamina y serina/treonina. Sp2, Sp3 y Sp4 reconocen la caja GC con especificidad y afinidad similar a Sp1.

Las proteínas Sp2 y Sp3 se identificaron en los promotores de las células T reconocedoras de antígeno. Ambas proteínas reconocen la caja GT, solo que Sp2

se une al motivo GT débilmente, mientras que Sp3 se une casi con la misma afinidad que Sp1 (Lania, 1997).

Sp4, es capaz de transactivar promotores virales y sintéticos que contengan cajas GC y GT, tanto en células de mamíferos como en células de *Drosophila* SL2. Experimentalmente Sp4 presenta diferencias con Sp1, ya que no es capaz de activar sinergicamente cuando se encuentra con sitios adyacentes (presumiblemente por falta del dominio D). Los análisis muestran que Sp4 es altamente expresado durante el desarrollo del cerebro y fenotipos Sp4(-) en ratones machos presentan baja actividad sexual.

Sp3 presenta de manera general una estructura muy parecida a Sp1, pero su actividad es nula en muchos genes que son activados por Sp1. De hecho Sp3 produce efectos represores en algunos promotores, ya que compite por el mismo sitio de reconocimiento al DNA (Hagen, 1994). En otros genes se ha visto un efecto de bloqueo estequeométrico o por interacción con elementos del aparato basal de la transcripción. Lo anterior muestra que la actividad de Sp3 es bifuncional, ya que puede regular de manera positiva o negativa de manera promotor - específica.

Otro factor homólogo a Sp1 es el gene *buttonhead* (btd) en *Drosophila*, el cual juega un papel importante durante su desarrollo. Mutantes en este gene presentan anormalidades en el desarrollo en el segmento de la cabeza. Experimentos transgénicos donde se expresa el gene Sp1 del humano, reconstruyen parcialmente el fenotipo de las btd- (Wimmer, 1993).

Antecedentes Directos y Objetivo

El estado del arte para la detección de sitios reguladores en el DNA presentan un gran desarrollo en las herramientas computacionales. Al realizar una búsqueda con algoritmos que están basados en estas herramientas, los resultados generan un gran número de sitios potenciales en el DNA, (gran parte de ellos falsos positivos) y su comprobación experimental se vuelve prácticamente imposible.

En esta tesis se propone la utilización de filtros que tomen en cuenta las características particulares del sistema biológico donde se llevará a cabo la predicción. Los filtros que se utilizan toman en cuenta la simetría de sitios, distancias con respecto al inicio de transcripción o traducción y algunas otras características propias del sistema. Esta idea ya ha sido puesta en práctica en otros trabajos, por ejemplo en el cromosoma III de levadura (Fondrat, 1996). En este trabajo se diseñaron filtros que toman en cuenta la propiedad palindrómica de algunos sitios así como la distancia con respecto al inicio de la traducción (ATG). En bacterias, dado la conservación de posiciones relativas al inicio de la transcripción, ésta se utiliza como herramienta para la filtración de posibles sitios (Rosenblueth, 1996).

Al pasar a otro tipo de organismos (vertebrados por ejemplo) el reconocimiento presenta adicionalmente los siguientes problemas: i) Las secuencias reguladoras son de menor longitud (menos de 20 pares de bases), ii) las distancias con respecto al origen de transcripción presentan rangos más amplios y iii) pueden presentar gran variación en los nucleótidos de reconocimiento (*mismatches*).

El objetivo central de la tesis fue darle a las predicciones de sitios una mayor certidumbre, a través de la incorporación de filtros biológicos basados en las características propias del sistema.

Lo anterior nos llevó a escoger un modelo para el cual se tuviese una gran cantidad de información, tanto en numero de sitios, como número de promotores, así como interacciones con otras proteínas y posiciones con respecto al inicio de la transcripción. Para tener una perspectiva de la información disponible en la regulación transcripcional, se utilizó la base de datos Transfac (Heinemeyer, 1998), una base de datos sobre regulación transcripcional en eucariontes, la cual contaba con 199 sitios de Sp1 en 99 promotores (en su versión actualizada cuenta con 213). En base a este criterio se seleccionó la proteína Sp1 (*Specific protein* 1). Esta proteína pertenece a la familia de los reguladores Sp, la cual, activa un amplio rango de genes en organismos vertebrados que han sido descritos por métodos experimentales (*footprinting, saturation-mutagenesis*).

Referente a los algoritmos de búsquedas, existe una gran cantidad de programas para hacer búsquedas de secuencias reguladores en DNA. En nuestro laboratorio el programa que hasta el momento ha dado mejores resultados para este propósito es Patser (Hertz, 1995). Este programa utiliza una **matriz de peso** como herramienta para el reconocimiento de sitios en DNA. Una matriz de peso es un arreglo bidimensional de renglones y columnas que se construye a partir de la información disponible sobre las secuencias de reconocimiento para una proteína reguladora. Asigna diferentes valores a cada columna de manera independiente y representa la frecuencia de cada nucleótido para una posición en particular. El algoritmo de Patser está sustentado en la teoría de la Información (ver material y métodos) lo cual ofrece un marco teórico sólido para la interpretación de los resultados.

Resultados y Discusiones.

1) Construcción de una matriz de peso para los sitios de reconocimiento de Sp1.

¿Cuál es el poder discriminatorio de una matriz?

Para poder responder a la pregunta anterior se necesitan hacer pruebas de sensibilidad (identificación de sitios verdaderos positivos) y especificidad (discriminación de falsos positivos) en las secuencias promotoras. Se tiene una matriz reportada para Sp1 (Chen, 1995), que se construyó con 126 sitios extraídos de la base de datos TRANSFAC.

T	28	0	0	0	7	2	9	0	5	29
A	8	17	0	0	19	1	0	26	18	5
C	7	0	0	2	100	4	1	11	7	75
G	83	109	126	124	0	119	116	89	96	17

En la construcción y prueba de una matriz para cualquier proteína, es necesario contar con la información precisa sobre las secuencias de reconocimiento, así como del contexto del promotor que se está regulando. Al revisar los datos de la base Transfac para la proteína Sp1, se encontraron algunas inconsistencias en la información para algunos sitios, entre los cuales podemos mencionar: i) ubicación en el promotor; la posición inicial y final para algunos sitios no coincide con la secuencia que se reporta, ii)el tamaño del sitio; la posición inicial y posición final no es congruente con el tamaño de la secuencia, iii)reiteración de la información (un mismo sitio puede estar reportado varias veces).

Lo anterior nos hizo dudar de la confiabilidad de los datos en Transfac, por lo que se realizó una búsqueda bibliográfica sobre genes regulados por la proteína Sp1 y se recolectó información de los elementos *cis* y *trans* en cada

promotor, con el objeto de construir una nueva matriz para esta proteína. Con ésto se tuvo un mayor control en el origen de los datos, además que se extendió la información sobre los sitios y se incluyeron datos sobre el contexto.

Se construyó una colección inicial que consta de:

19 promotores

82 sitios para proteínas reguladoras.

24 tipos diferentes de sitios para proteínas reguladoras

(Altschmied, 1989; Ammendola, 1990; Araki, 1991; Birnbaum, 1995; Boyer, 1990; Cheung, 1993; de Groot, 1991; Fischer, 1993; Goding, 1987; Jones, 1985; Kasai, 1992; Karlseder, 1996; Kuwahara, 1993; Li, 1996; Liu, 1992; Morgan, 1988; Morgan, 1989; Schmidt, 1989; Seal, 1991; Tamaki, 1995; Tamura, 1991; Tansey, 1997; Tebb, 1989; Therrien, 1994; von der Ahe, 1988; Wu, 1987; Yoshida, 1989).

PROMOTOR	ORGANISMO	PROMOTOR DE
E1B	Human adenovirus	
uPA_pig	Sus scrofa	plasminogen activator; urokinase
Col1a2	Homo sapiens	collagen alpha; collagen type I
H4	Homo sapiens	histone H4
IVa2	Adenovirus type 2,	complete genome
snRNAu2	Xenopus laevis	nuclear RNA; tandem repeat
NF-1	Rattus norvegicus	NFI-A.
TP-1	Homo sapiens	triosephosphate isomerase
U1snRNA	Gallus gallus	nuclear RNA
TK_HSV1	Human herpesvirus 1	thymidine kinase
JunD	Mus musculus	JUN-D protein.
hsp70	Homo sapiens HSP70 gene;	heat shock protein
VitII	Gallus gallus	vitellogenin
c-Jun	Mus musculus	c-jun oncogene
lyszyme	Gallus gallus	lysozyme
ANT2	Homo sapiens	adenine nucleotide translocator-2
g_fib	Rattus norvegicus	fibrinogen; gamma-fibrinogen
TKmouse	Mus musculus	thymidine kinase
E1A	Human adenovirus	type 5

Visualmente los elementos reguladores para cada uno de los 19 promotores se pueden ver en los mapas 1, 2 y 3. Es importante señalar que todas las proteínas en los 19 promotores presentan evidencias experimentales, como pueden ser: *footprinting, gel shift* y ensayos de deleción, entre otros.

En esta colección se tiene un total de 33 sitios de reconocimiento para Sp1. Estos se utilizaron para construir una nueva matriz de búsqueda con el programa **Wconsensus**. La matriz que se obtuvo se presenta a continuación:

```
MATRIX
number of sequences = 33
width = 15
crude information = 8.020
unadjusted information = 10.314
sample size adjusted information = 9.289
ln(probability) = -183.525   probability = 1.97778E-80
ln(expected frequency) = -136.741
expected frequency = 4.11368E-60
T |  3   3   2   5   0   0   0   0   0   2   3   4   7  12   5
A | 10  16   8   8   5   0   0   2   0   0   3   6   4   4  13
C |  4   6   8   3   0   0   0  31   0   0   4   5  14   7   7
G | 16   8  15  17  28  33  33   0  33  31  23  18   8  10   8
```

(el significado de los parámetros se explica en la sección de material y métodos donde se describe la construcción de una matriz de búsqueda).

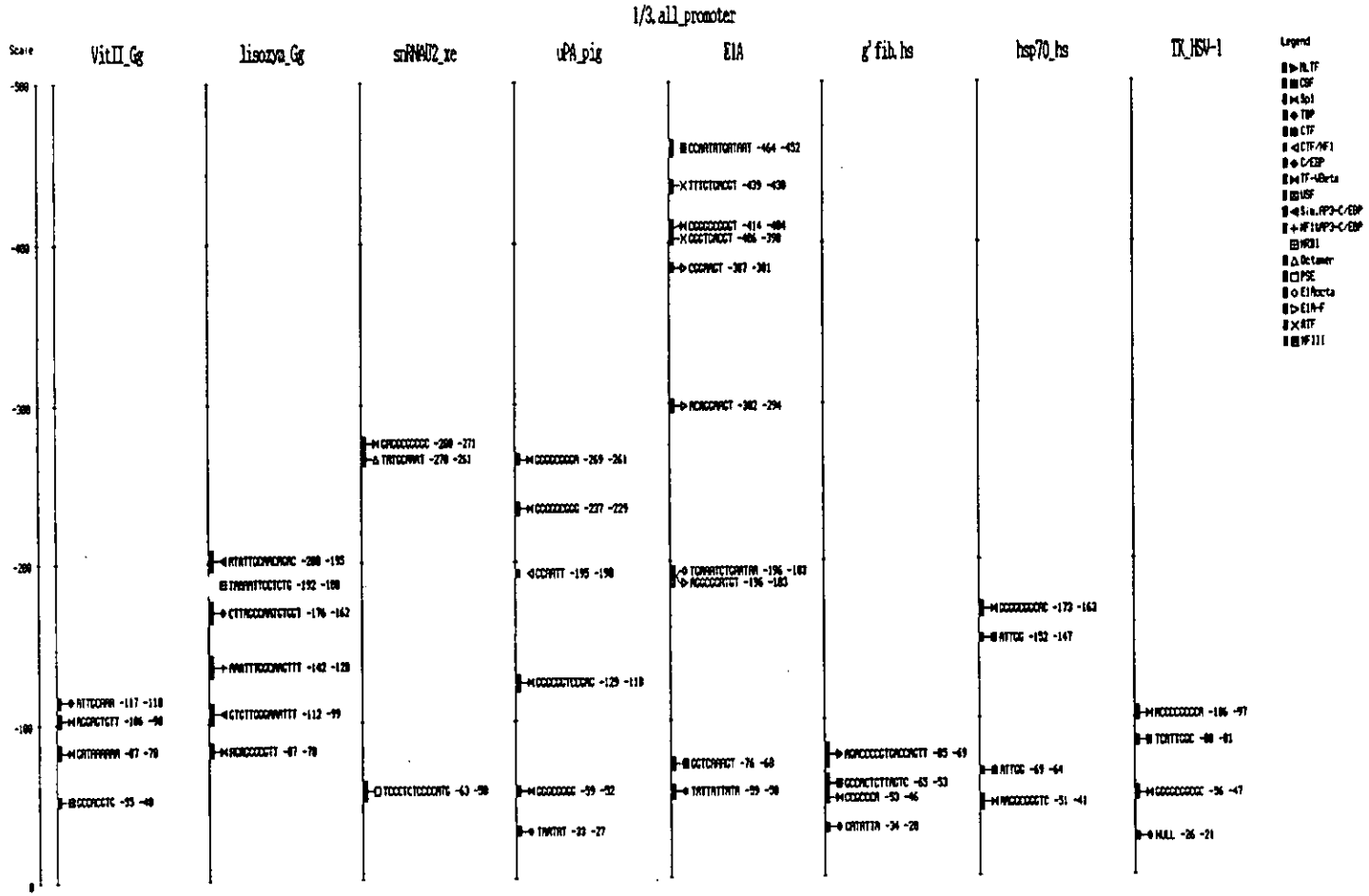
2) Comparación entre las matrices generales H (Hertz) y V(Victoria):

Hasta este momento, no se ha respondido la pregunta inicial planteada en la sección anterior, ¿qué tan bueno es el poder de discriminación en una matriz de peso?. La respuesta a está pregunta nos motivó para hacer los siguientes análisis:

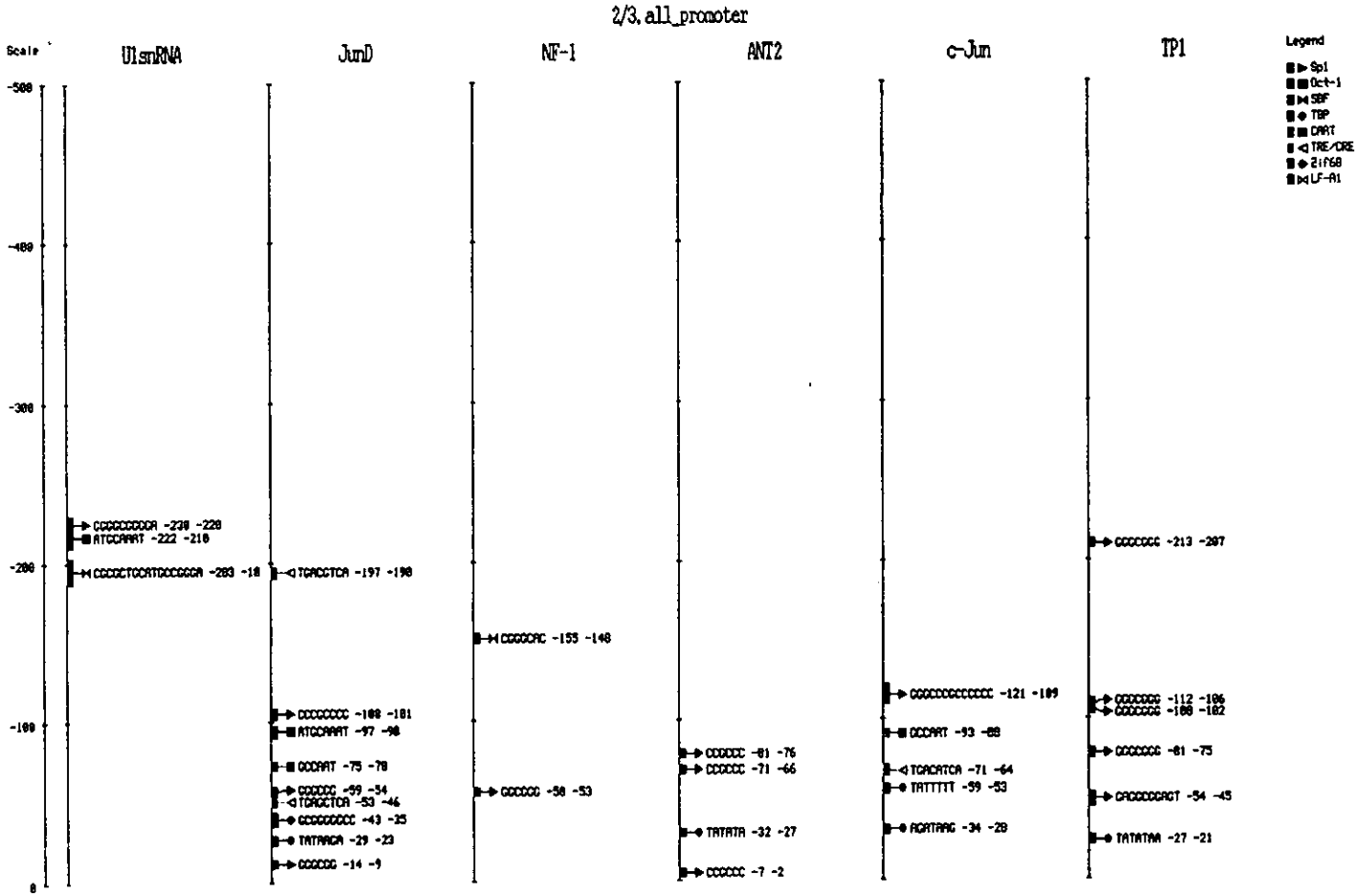
a) La selectividad que puede tener una matriz con los sitios a partir de los cuales fue generada.

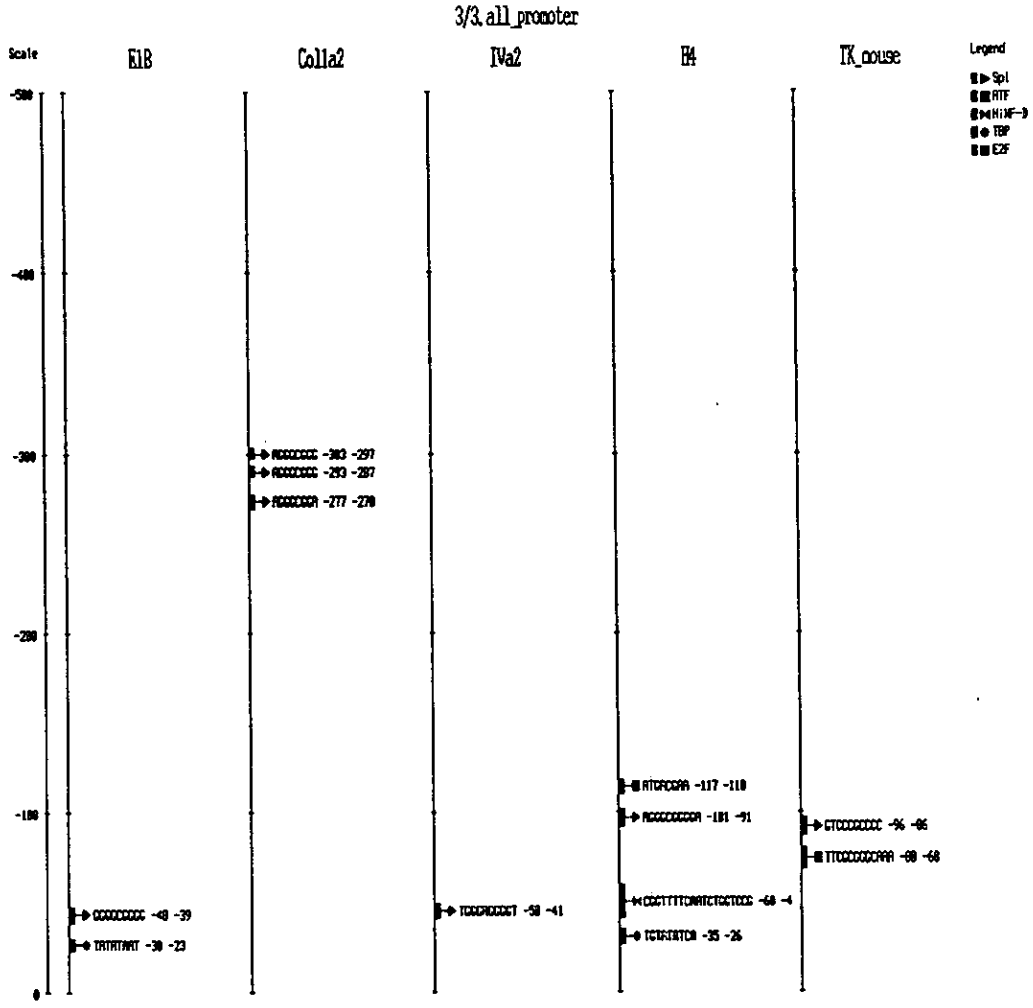
b) El poder de discriminación en una matriz por promotor

a) Para el análisis de sensibilidad y especificidad se utilizaron dos matrices: la matriz H (creada con la colección de sitios en Transfac) y la matriz V (con los 33 sitios de Sp1 provenientes de nuestra base de datos). Con ambas matrices se buscaron sitios para Sp1 en las regiones promotoras de los 19 genes


1/3_all_promoter
Score
VitII_Gg
lisozym_Gg
smRNAlf2_xe
uPA_pig
E1A
g' fib_hs
hsp70_hs
TK_HSV-1
Legend

MAPA 1


2/3. all_promoter
Scale
U1snRNA
JunD
NF-1
ANT2
c-Jun
TP1
Legend
Sp1
Oct-1
SBF
TBP
CART
TRE/CRE
2if68
LF-A1
-500
-400
-300
-200
-100
0
CCCCCCCCR -230 -220
ATCCARRT -222 -210
CCCCCTCCRTCCCCCCR -203 -18
TGACGTCR -197 -190
CCCCCRC -155 -148
CCCCCCC -108 -101
ATCCARRT -97 -90
CCCRRT -75 -70
CCCCCC -59 -54
TGRCGTCR -53 -46
CCCCCCCCC -43 -35
TATARCR -29 -23
CCCCCC -14 -9
GCCCCC -58 -53
CCCCCC -81 -76
CCCCCC -71 -66
TATATR -32 -27
CCCCCC -7 -2
CCCCCCCCCCCC -121 -109
CCCRRT -93 -88
TGACATCR -71 -64
TATTTTT -59 -53
RCRTARG -34 -28
CCCCCC -213 -207
CCCCCCC -112 -106
CCCCCCC -108 -102
CCCCCCC -81 -75
CRCCCCCRGT -54 -45
TATATAR -27 -21



Mapa 1-2-3. Muestra las distribución de las proteínas reguladoras en las 19 regiones promotores

(que constan de 600 pares de bases cada una). Mostrando los siguientes resultados.

Para un total de 11400 pb, la matriz H encontró, con un umbral de 1.75 c. I. (contenido Informacional), 178 sitios para la proteína Sp1, mientras que la matriz V encontró, con un umbral de 3.28 c. I, encontró 103 sitios para la misma proteína.

En la corrida de la matriz V se tomó como umbral el valor más bajo con el que la matriz calificó a los sitios experimentalmente de Sp1. En el caso de la matriz H se observa un conjunto de tres sitios experimentales con valores de c. I. muy por abajo del resto de los sitios, por lo que el umbral se fijó entre el promedio del tercer y cuarto valor de los sitios experimentales, esto con el propósito de ganar sensibilidad.

Los *scores* para cada uno de los sitios se muestran a continuación:

PROMOTOR	VALORES DADOS POR LA MATRIZ H	VALORES DADOS POR LA MATRIZ V
ANT2	3.60	3.59
ANT2	3.76	4.13
ANT2	8.08	6.36
C-JUN	5.51	7.17
COL1A2	4.51	7.77
COL1A2	4.95	6.64
COL1A2	5.24	8.46
E1A	4.63	7.15
E1B	8.09	9.48
G.FIB	7.22	8.31
H4	4.96	9.42
IVA2	7.23	7.21
JUND	0.83	5.12
JUND	3.6	3.28
JUND	5.89	8.22
NF-1	7.03	6.62
TK.HSV-1	6.02	4.56
TK.HSV-1	7.94	6.48
TK_MOUSE	7.36	10.5
TP1	2.66	5.76
TP1	4.19	4.57
TP1	5.37	6.85
TP1	6.89	7.88
U1SNRNA	6.34	8.53
VITII.GG	2.87	7.21
HSP70.HS	3.81	8.64
HSP70.HS	5.3	6.3
LYSOZYME.GG	-2.4	7.14
SNRNAU2.XE	7.18	10.99
UPA.PIG	0	5.59
UPA.PIG	6.34	9.74

UPA.PIG	6.7	6.82
UPA.PIG	7.36	5.59

TABLA 1.- Esta tabla muestra dos diferentes valores (dados por la matriz H y V) para los sitios de reconocimiento para Sp1. Los valores más bajos están en color rojo.

Analizando la tabla anterior, los valores más bajos con los que la matriz H calificó fueron:

lyszyme	Gallus gallus	-2.40
uPA_pig	Sus scrofa	0.00
JunD	Mus musculus	0.83
TKmouse	Mus musculus	2.66

Para la matriz V, los promotores con los valores más bajos fueron:

JunD	Mus musculus	3.28
ANT2	Homo sapiens	3.59

Con los valores obtenidos por ambas matrices se hizo un análisis bidimensional a través de gráficas; lo anterior nos hizo visualizar **el poder discriminatorio en ambas matrices para un mismo conjunto de datos y las correlaciones entre los valores de los sitios experimentales para Sp1** (a través de la dispersión entre los puntos).

Utilizando el programa **XYgraph,** se graficaron los valores de la matriz H (eje de las abscisas) contra los valores de la matriz V (eje de las ordenadas).

Se construyeron dos gráficas, en la primera se muestran los valores de los sitios experimentales (gráfica 1) y en la segunda todos los valores con un *score*

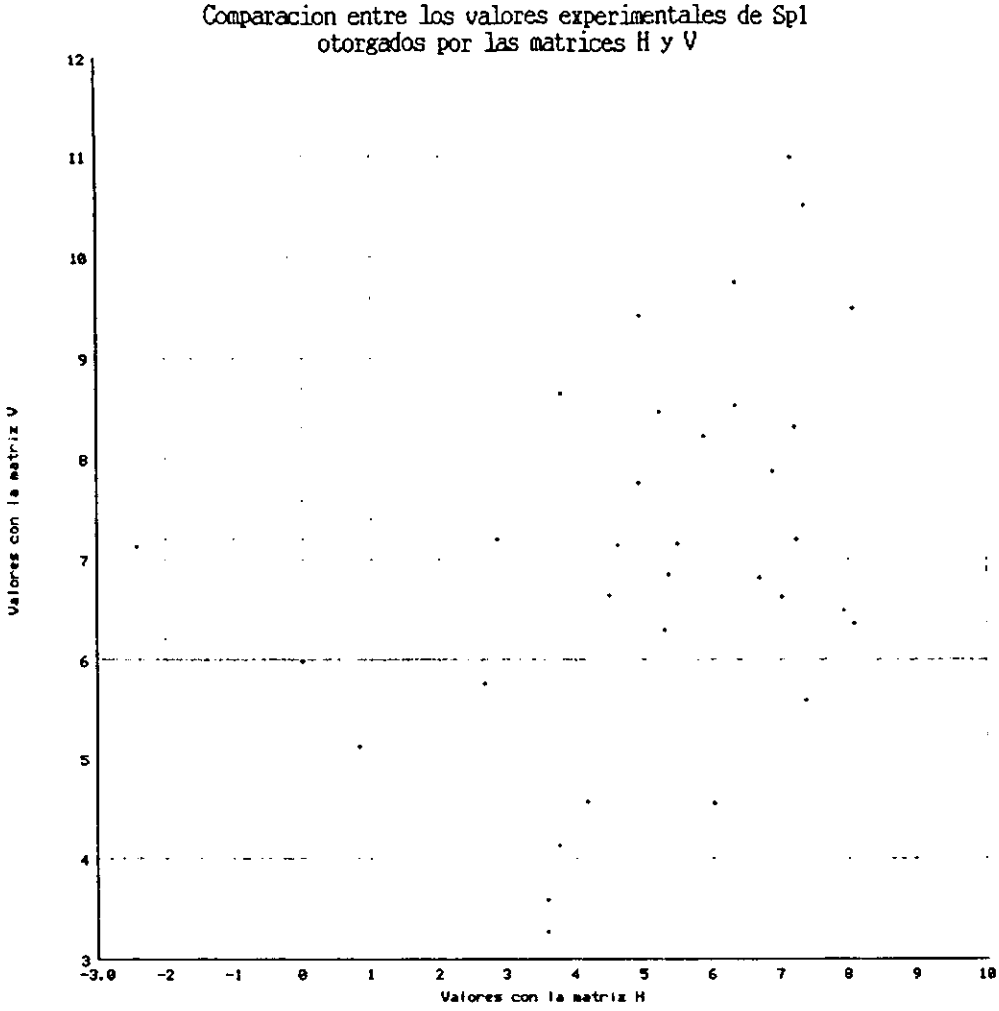
más alto de 3.27 (el umbral por arriba del cual se encuentran todos los sitios experimentales de Sp1 para la matriz V) (gráfica 2).

Se hizo un ejercicio de sobreposición de las gráficas con el propósito de observar:

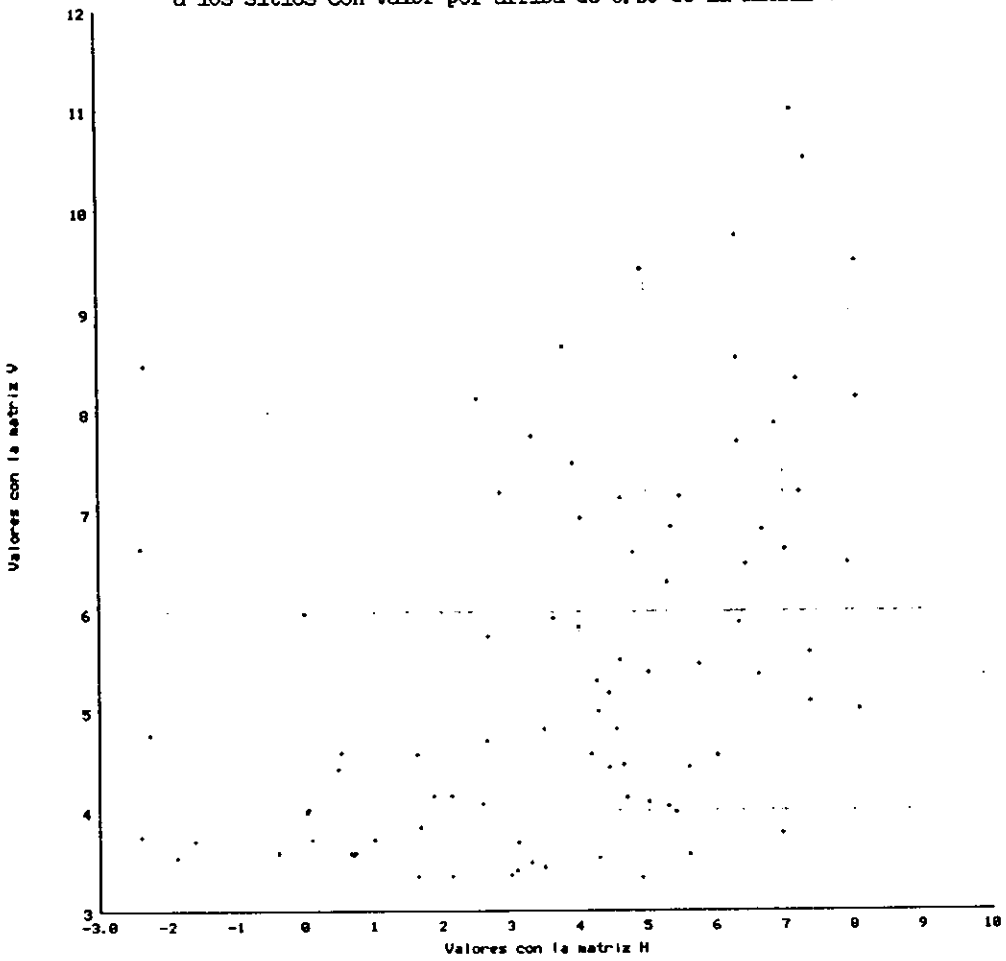
- Dispersión de los puntos; esto es solo a nivel cualitativo y se observa el ancho de las gráficas.
- División por ejes; se dividen las gráficas de tal manera que se trazan dos ejes perpendiculares a la mitad de los ejes X - Y y se forman cuatro cuadrantes; en uno de los cuales (superior derecho) se encuentran todos o la gran mayoría de los sitios experimentales de la proteína de interés.

De manera general, lo que se observa en las gráficas es lo siguiente: i) la matriz V presentan valores más altos de contenido informacional (10.99 / 10.50 con la matriz V y 8.09 / 8.08 con la matriz H). ii) El intervalo de los valores fue menor en los sitios calificados con la matriz V (3.28 - 10.99) que con la matriz H (-2.40 - 8.09). iii) Algunos sitios (presuntos falsos - positivos) son reconocidos con calificaciones más altas por la matriz H que por la matriz V. Esto se infiere después de trazar los cuatro cuadrantes imaginarios y ver la distribución de los puntos. La mayor parte de los sitios con evidencia experimental caen en el cuadrante superior derecho. Los 70 sitios restantes (en donde se asume la existencia de un gran número de falsos positivos) están ubicados en los cuadrantes inferior izquierdo e inferior derecho; los falsos positivos localizados en este último cuadrante están siendo reconocidos con valores altos con la matriz H y bajos con la matriz V.

De los tres puntos anteriores se pudo inferir lo siguiente: **la matriz V reconoce mejor a los 33 sitios de Sp1 que la matriz H.** Esto muestra la gran sensibilidad del método para reconocer los sitios a partir de los cuales fue construida la matriz de búsqueda. Lo anterior hace pensar que el poder


Comparacion entre los valores experimentales de Sp1
otorgados por las matrices H y V
Valores con la matriz V
Valores con la matriz H
12
11
10
9
8
7
6
5
4
3
-3.0
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Gráfica 1


Comparacion entre los valores asignados por las matrices H y V
a los sitios con valor por arriba de 3.28 de la matriz V
Valores con la matriz V
Valores con la matriz H
12
11
10
9
8
7
6
5
4
3
-3.0
-2
-1
0
1
2
3
4
5
6
7
8
9
10

Gráfica 2

predictivo de una matriz es limitado, ya que el reconocimiento de nuevos sitios de unión al DNA, dependen del parecido de la secuencia con el *training set* a partir del cuál se construyó la matriz.

b) El poder de discriminación en una matriz por promotor

Se inició un análisis por promotores, ambas matrices recorrieron todas las secuencias en su longitud y todos los *scores* obtenidos fueron graficados (gráficas Apéndice). De manera general los 19 promotores se pueden dividir en tres grupos: a) aquellos donde la secuencia de reconocimiento para Sp1 se distingue claramente del resto de los sitios, ejemplo thymidine kinase en *Mus musculus* y VitII en *Gallus gallus*, b) otros donde los sitios de Sp1 se diluyen con el resto de los valores formando un continuo, en este grupo se encuentran ANT2 *Homo sapiens* y JunD *Mus musculus*, c) por último se encuentran los casos donde se observa una nube de puntos fuera del resto de los sitos, en esta nube están contenidos los sitios experimentales y otros, que por su *score* podrían ser predicciones interesantes como sitios de reconocimientos para Sp1, ejemplos TP1 *Homo sapiens* y g_fib en *Rattus norvegicus*.

En este análisis de gráficas bidimensionales se pueden ver las correlaciones de ambas matrices en una región promotora. También nos deja ver lo conspicuo que pueden ser los sitios de Sp1 en algunos promotores al compararlos con el resto de la secuencia. Esta observación podría utilizarse como criterios de predicción; pero la existencia de promotores donde los valores de los sitios se diluyen con el resto de puntos en la gráfica limita su utilización.

De manera general con los dos análisis anteriores se puede dilucidar que existe una correlación directa en ambas matrices y que la asignación de *scores* por una matriz de peso **no es suficiente para discriminar sitios de Sp1 de otros sitios en secuencias promotoras**; aunque en algunos casos se pueden distinguir muy bien un sitio de Sp1 del resto en un promotor; en otros el número de predicciones es muy alto y su comprobación experimental se vuelve muy difícil.

Lo anterior hace razonable pensar en la búsqueda de nuevas propiedades biológicas que mejoren la especificidad de las matrices y la predicción de sitios. En la siguiente parte del trabajo se proponen dos propiedades:

-Utilización de información contextual **(posición)**

-Aumento de la especificidad en la matriz **(creación de matrices por organismos)**

3) Mejoramiento del poder discriminatorio de las matrices.

3.1 Análisis de la posición de los sitios

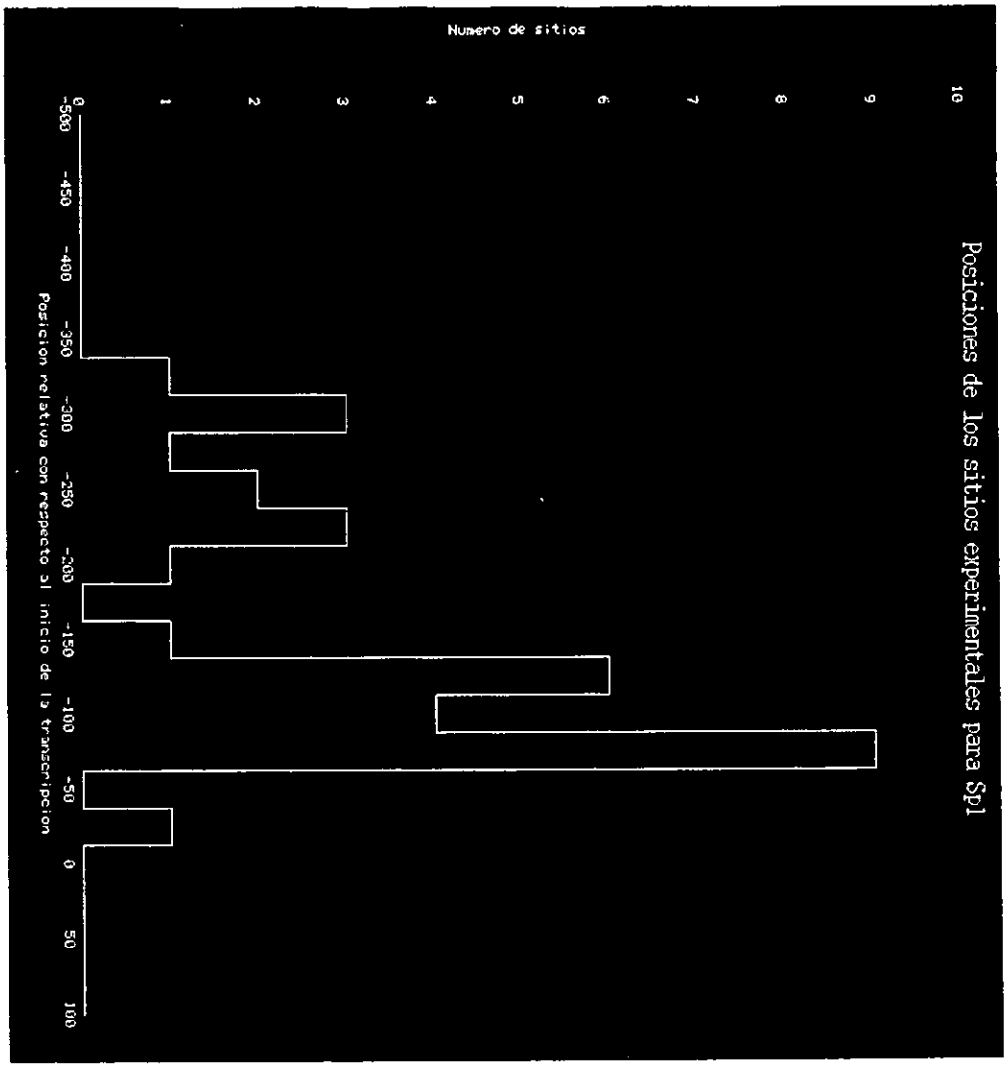
¿Existe una ubicación preferencial para los sitios de SP1 en una región promotora?

Las observaciones previas a éste trabajo muestran que sí existen posiciones preferenciales para Sp1 con respecto al inicio de la Transcripción (Pérez-Rueda, 1998). Esto se quiso corroborar con la colección de 33 sitios para Sp1. El primer análisis fue ver si en esta colección hay posición(es) preferencial(es), la ubicación dentro del promotor, y si existe independencia entre el valor del contenido informacional y la posición con respecto al inicio de la transcripción. Para esto hay que demostrar que los sitios experimentales y los sitios predichos (aquellos que hayan pasado el umbral), se distribuyen de manera diferente a lo largo de la región promotora.

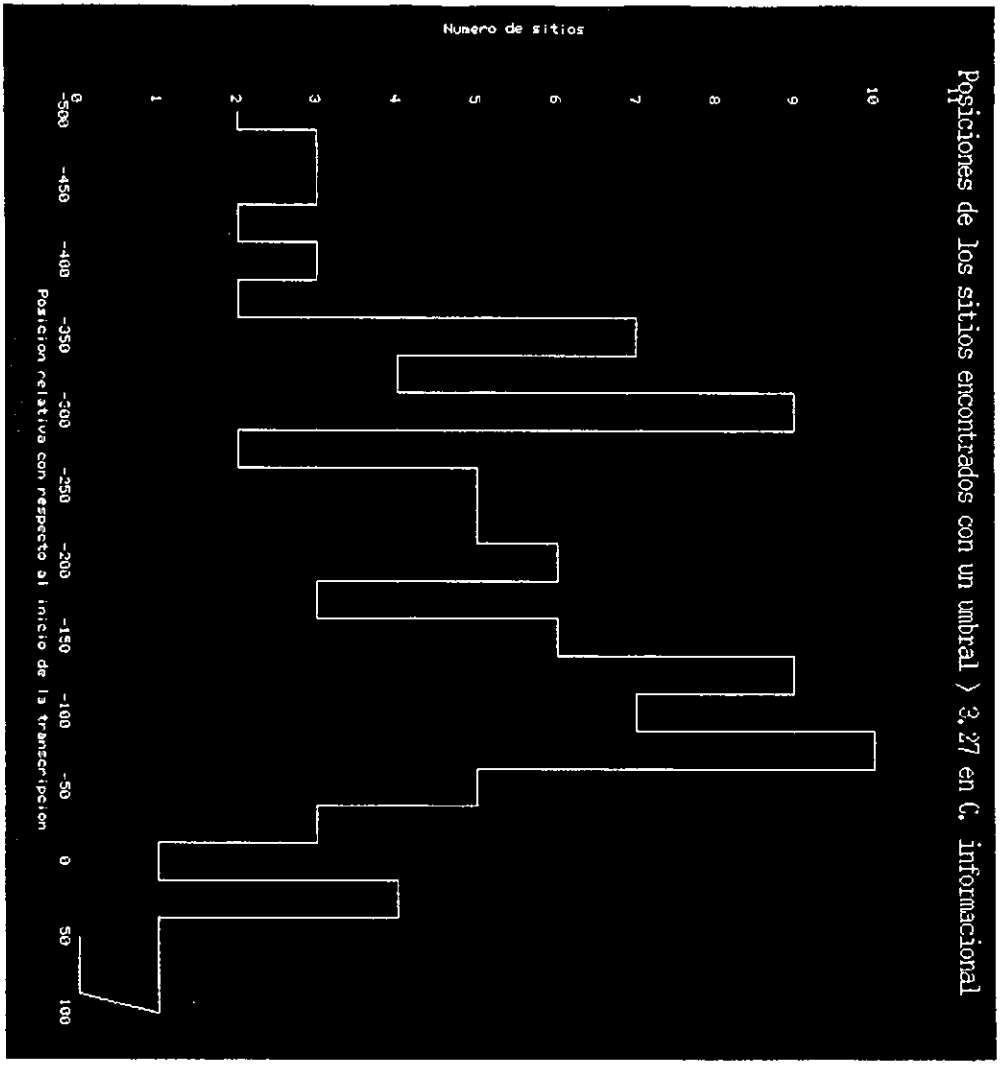
Al momento de graficar, se pudo ver lo siguiente:

a) Los sitios experimentales para Sp1 tienen posiciones preferenciales con respecto al inicio de la transcripción (como se observa en la gráfica 3).

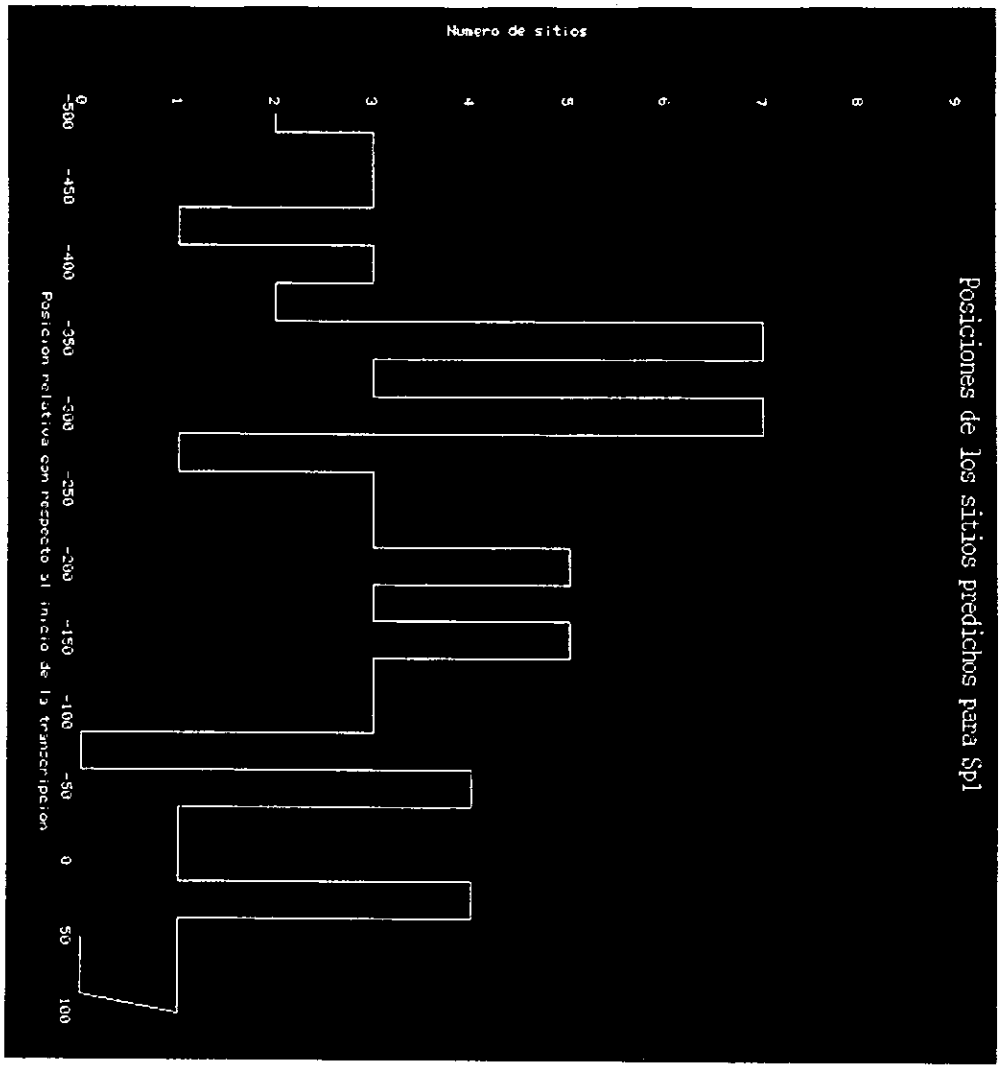
b) El valor asignado por la matriz es independiente de la ubicación posicional para los sitios; es decir en la gráfica donde se encuentran los sitios con un *score* por arriba del umbral (gráfica 4), se muestra una distribución diferente de los sitios experimentales. Esto se confirma con la gráfica 5, donde se muestra solo los sitios sin evidencia experimental pero que pasaron el umbral de los sitios experimentales.



Gráfica 3 Histograma de las posiciones con respecto del inicio de la transcripción para 33 sitios de Sp1 en 19 promotores



Gráfica 4. Histograma de las posiciones con respecto del inicio de la transcripción. de los sitios con un umbral mayor a 3.27 en los 19 promotores



Gráfica 5. Histograma de las posiciones con respecto del inicio de la transcripción, de los sitios en los 19 promotores con la matriz V

Una vez que se comprobó una distribución diferente entre los sitios experimentales y los predichos se procedió a la creación de un índice que tomase en cuenta la posición absoluta y la calificación proporcionada por una matriz (esto se discute ampliamente en la sección de material y métodos).

Los 33 sitios presentaron los siguientes valores de frecuencia absoluta (tercera columna) y frecuencia relativa (cuarta columna) a lo largo de 600 pares de bases (-500 a +100).

Intervalo de posición		Frec. absoluta	Frec. relativa
-500	-475	0	0.00000
-475	-450	0	0.00000
-450	-425	0	0.00000
-425	-400	0	0.00000
-400	-375	0	0.00000
-375	-350	0	0.00000
-350	-325	0	0.00000
-325	-300	1	0.03125
-300	-275	3	0.09375
-275	-250	1	0.03125
-250	-225	2	0.06250
-225	-200	3	0.09375
-200	-175	1	0.03125
-175	-150	0	0.00000
-150	-125	1	0.03125
-125	-100	6	0.18750
-100	-75	4	0.12500
-75	-50	9	0.28125
-50	-25	0	0.00000
-25	0	1	0.03125
0	25	0	0.00000
25	50	0	0.00000
50	75	0	0.00000
75	100	0	0.00000

Se modifican los *scores* con un cosiente (ver material y métodos) y a al nuevo valor se le suma (dependiendo de su posición) el valor que le corresponde para la frecuencia relativa.

Los 33 sitios experimentales presentaron los siguiente valores con el nuevo índice:

Promotor	Indice: score/position
snRNAU2_xe	1.66293
E1B	1.63483
TK_mouse	1.62422
H4	1.53251
hsp70_hs	1.51489
g_fib.hs	1.46778
uPA_pig	1.45321
TP1	1.40638
JunD	1.36118
U1snRNA	1.28044
uPA_pig	1.25503
Col1a2	1.23919
NF-1	1.22647
VitII_Gg	1.21696
Col1a2	1.20317
lysozyme_Gg	1.14447
c-Jun	1.1175
TK_HSV-1	1.11273
TP1	1.10306
Col1a2	1.04183
ANT2	1.0331
Iva2	1.02946
E1A	1.0209
JunD	1.0123
TP1	1.00993
ANT2	0.96518
TK_HSV-1	0.93234
hsp70_hs	0.930782
uPA_pig	0.885092
TP1	0.840018
uPA_pig	0.829406
TP1	0.663454
JunD	0.657742

A partir de estos valores se fijo un nuevo umbral en 0.75; el cual se obtuvo con el promedio del segundo y tercer valor de los índices más bajos (uPA_pig - 0.829406, TP1 - 0.663454). Con este nuevo umbral se trató de ganar especificidad (omitiendo los dos valores con el índice más bajo) y sensibilidad (Adams, 1995).

El nuevo umbral dio como resultado una disminución de posibles sitios para Sp1. De los 70 sitios predichos que tuvieron una calificación mayor a 3.27 en contenido informaciones, solo quedaron 22 sitios al integrar el valor de distancia para el nuevo índice (ver tabla 2).

Estos 22 sitios cumplen con dos características (posición y contenido informacional) y los convierte en predicciones *bona fide* para ser sitios para Sp1 (mapa 4)

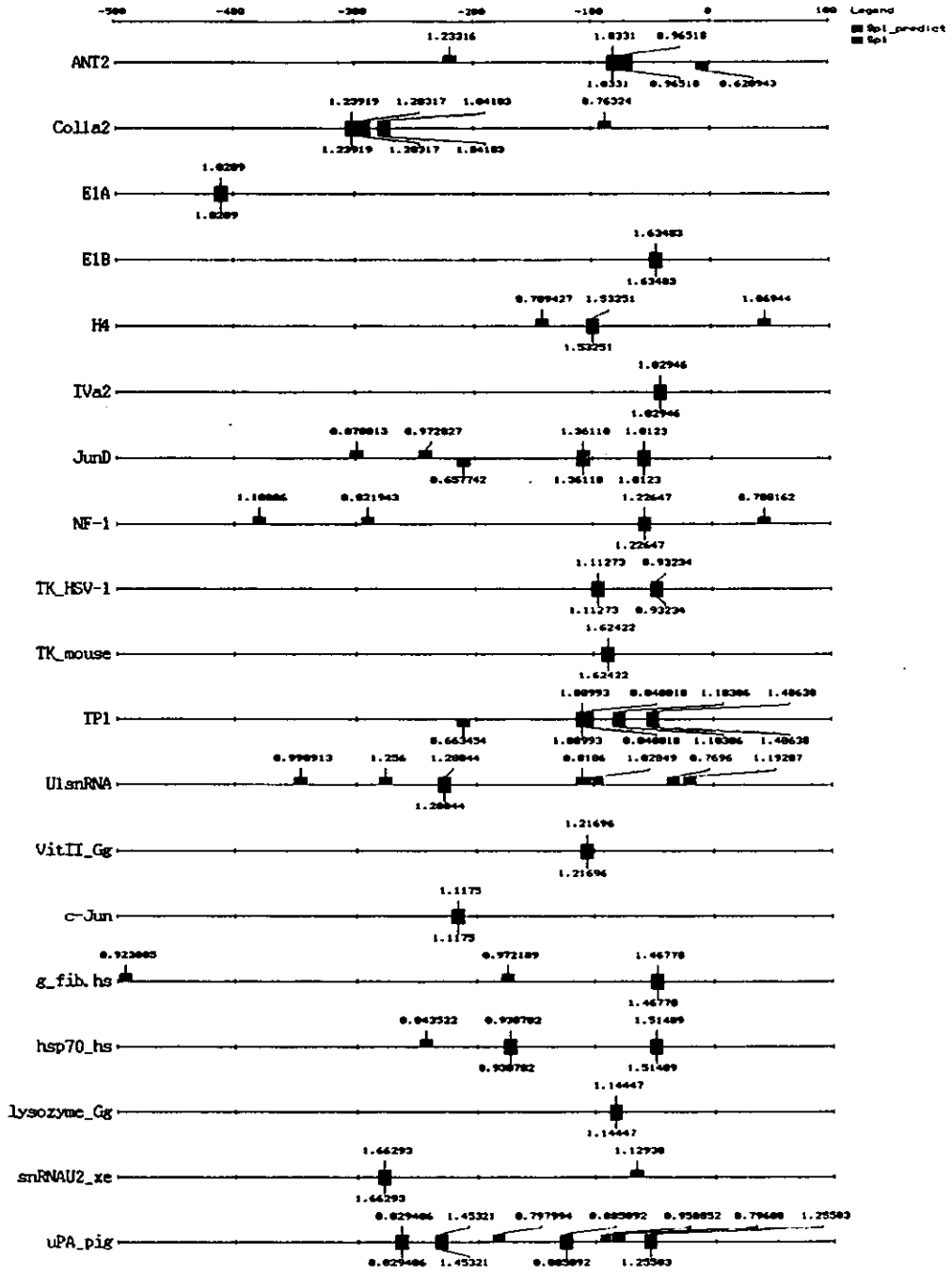
Esto motivó a realizar una revisión bibliográfica en los promotores donde se encontró al menos una predicción *bona fide* y tratar de explicar la razón por la cual no se ha documentado un sitio experimental Sp1. Podemos encontrar por lo menos tres posibles casos: i) la falta de evidencia experimental en esas región del promotor, ii) que ese sitio sea reconocido por otra proteína o iii) que se haya estudiado la región sin encontrar el sitio.

Los resultados de esta revisión se presentan a continuación y se pueden dividir en dos grupos, a) aquellos donde no se ha hecho análisis experimental en la región del promotor donde se hizo la predicción (estos se señalan con asterisco cuando se inicia su explicación) y b) donde sí se hizo una análisis experimental:

ANT2 - El nombre del gen es *Adenide nucleotide translocator* (es una proteína translocadora ATP/ADP). Se hicieron estudios en detalle, con matuteaseis y gel de retardo, en la región proximal (-87/+8). Estudios previos (Li, 1996) en -235/+46 mostraban la presencia de 7 cajas GC, 4 de ellas se encuentran en la región -235. Al remover esta región la expresión del gen reportero (CAT) decrementa su expresión aproximadamente en un 50-60%. El dicho detectado en este promotor, es un GC box, grandes probabilidades que este sitio actúe de manera sinérgica por la distancia a la que se encuentra.

PROMOTOR	Posición	*Score*	*score* modif.	Indice: *score* + posic.
ANT2	-225	7.98	1.13941	1.23316
Col1a2	-93	4.47	0.63824	0.76324
H4	-147	5.31	0.758177	0.789427
H4	40	7.49	1.06944	1.06944
JunD	-247	6.37	0.909527	0.972027
JunD	-304	5.88	0.839563	0.870813
NF-1	-295	5.10	0.728193	0.821943
NF-1	-384	7.71	1.10086	1.10086
NF-1	39	5.52	0.788162	0.788162
U1snRNA	-103	5.89	0.840991	1.02849
U1snRNA	-115	4.42	0.6311	0.8186
U1snRNA	-25	8.13	1.16082	1.19207
U1snRNA	-281	8.14	1.16225	1.256
U1snRNA	-351	6.94	0.990913	0.990913
U1snRNA	-39	5.39	0.7696	0.7696
g_fib.rat	-178	6.59	0.940939	0.972189
g_fib.rat	-497	6.47	0.923805	0.923805
hsp70_hs	-248	5.47	0.781022	0.843522
snRNAU2_xe	-71	5.94	0.84813	1.12938
uPA_pig	-187	5.37	0.766744	0.797994
uPA_pig	-87	4.70	0.67108	0.79608
uPA_pig	-96	5.84	0.833852	0.958852

Tabla 2.- Sitios predichos a partir del índice *score matrix /position*



Gráfica 4. Histograma de las posiciones con respecto del inicio de la transcripción, de los sitios con un umbral mayor a 3.27 con la matriz V en los 19 promotores

*Col1A2 - Es la región reguladora del gen Colágeno tipo I. En este promotor se han hecho experimentos de deleción-mutación en la región que va de -323 y -186. Se encontraron dos sitios distales de Sp1 en -303 y -271 (Tamaki, 1996). Se predijo un sitio *bona fide* para Sp1 en la posición -93, está predicción es muy interesante ya que sugiere una posible interacción de los sitios distales con su homónimo proximal para dar lugar a una activación sinérgica.

*H4 - Los genes de histonas codifican para proteínas involucradas en la estructura y función de la cromatina eucariotica (Birnbaum, 1995).

Dentro del promotor *H4* se analizaron dos regiones proximales. La región I que va de -124 a -86 y la II que va de -64 a -24 (ambas con respecto al inicio de la transcripción). La primera es reconocida por las proteínas Sp1 y ATF y la segunda por HiNF y TBP. Posteriormente se llevo a cabo un análisis con *footprinting* de la región II y se extendió de -156 a -72. El análisis no detecto sitio alguno de pegado para otras proteínas, aunque el método predijo un sitio para Sp1 en la región -147.

Otra predicción para Sp1 está en +40. En esta región no se han hecho análisis, pero es posiblemente que en esta zona Sp1 actúe como represor. Ya se ha encontrado en otros trabajos a Sp1 con un papel represor (Li, 1996).

*JunD - El gen JunD pertenece a la familia de los reguladores Jun/Fos, los cuales se encargan de traducir las respuestas de varios estímulos extracelulares alterando la expresión de genes. Al caracterizarse la región promotora (de Groot, 1991) entre el intervalo de -190 a -52, se hicieron análisis de *DNA footprinting* y se detectaron tres sitios de unión para la proteína Sp1. Se predijeron dos sitios *upstream* para esta proteína en -247 y -304.

*NF-1 - Se establecieron las bases funcionales del promotor de NF-1 (factor de transcripción). Se hicieron análisis con las técnicas gel de retardo y protección

con DNasa I (Ammendola, 1990) en 198 pares de bases *upstream* del inicio de la transcripción. Se determinó experimentalmente un sitio de reconocimiento para la proteína Sp1 en -62. Se encontraron otros tres posibles sitios, dos *upstream* de esta región (que pudieran tener efecto sinérgico con el sitio proximal) y un tercero *dowstream* (que puede actuar como represor).

*U1snRNA - Los snRNA *small nuclear RNA* son los RNA que están involucrados en el *splicing* de los pre-RNa mensajeros en células eucarionticas. De este grupo los U1, U2, U3, U4 y U5 snRNA son sintetizados por la polimerasa II. Se analizó la región de -230 a -180 (con mutaciones y ensayos de substitución) y se determino el papel que tenían los 3 diferentes elementos involucrados para iniciar la transcripción (Cheung, 1993). Los elementos que reconocen esta región son Sp1, Oct-1 y SBF. En los extremos de esta región se encontraron 6 secuencias que se predijeron como sitios de reconocimiento para Sp1; distribuídos tanto en regiones proximales como en distales.

*G_fibrinógeno - El fibrinógeno es producto de tres genes (alfa, beta y gamma). Se realizó un estudio en la región promotora del gene Gama Fibrinógeno (Morgan, 1998). El análisis de deleción - mutación y DNasa I *footprint* se concentró en la región de -88 a +36, debido a que estudios preliminares indicaron que la secuencia de -847 a -88 no incrementa significativamente la transcripción. En las predicciones de posibles sitios de reconocimiento para Sp1, se encontraron dos sitios distales a -178 y -497.

hsp70 - Es un gene que codifica para una proteína de choque térmico. Se hizo un estudio del papel que desempeña Sp1 en la activación de la transcripción. En esta región promotora se hizo una predicción ubicada en -248 y haciendo una revisión bibliográfica, se encontró que este sitio ya había sido detectado con un análisis de DNasa I footprinting (morgan, WD, 1989).

uPA - Los activadores de plasminógenos (PA) juegan un papel muy importante en fibrinólisis, inflamación y remodelación de tejidos durante el desarrollo. Un tipo de activador plasminógeno es el urocinasa (uPA). Se analizó con ensayos de *DNasa I footprinting* y Gel de retardamiento los elementos en *cis* y en *trans* de la región promotora que va de 0 a -500 del inicio de la transcripción (vonder Ahe, 1988). En este análisis se encontraron sitios para las proteínas TBP, CTF/NF1 y Sp1 (con cuatro sitios). Se predijeron otros 3 sitios para este último factor a -187, -96 y -87. Entre -96 y -87 existe un traslapamiento por lo que se puede ver como un solo sitio.

snRNAU2 - Esta secuencia codifica para el *small nuclear* U2 en *Xenopus laevis*. Se hizo una exhaustiva caracterización de la secuencia promotora con experimentos de competición y DNasa I *footprinting* (Tebb, 1989). Se hizo la predicción para un sitio de SP1 en la zona proximal -71. Esta predicción tiene un traslape con otra factor de transcripción en -63, identificado con el nombre de PSE *proximal sequence element*. En este caso en particular, se puede especular que Sp1 no se pega en esta región porque compite con otra proteína por el sitio.

3.2 Construcción de matrices por organismo

¿Existen diferencias entre las matrices de diferentes organismos para una misma proteína?. La respuesta a esta pregunta nos motivó a realizar los dos siguiente análisis:

1) Diferencias en la composición de nucleótidos en las matrices de Sp1 para diferentes organismos.

2) Analizar las correlaciones para los valores asignados por diferentes matrices.

Para ello se buscó información en la base de datos TRANSFAC (Heinemeyer, 1998) sobre promotores regulados por Sp1 en diferentes organismos. Los que cuentan con mayor información son:

- Humano (*Homo sapiens*) con 62 sitios en 32 secuencias (promotoras + codificantes) - Ratón (*Mus musculus*) con 29 sitios en 22 secuencias y
- Rata (Rattus norvergicus) con 27 sitios en 22 secuencias. A partir de estos sitios se construyeron tres matrices diferentes (con wconsensus) y se inicio un análisis comparativo que involucran los dos aspectos arriba mencionados y se especifican a continuación:

1) Se construyeron tres matrices y se compararon unas con otras (ratón - humano - rata) utilizando una prueba de $\chi 2$ posición por posición, analizando sus diferencias y si estas eran o no significativas (ver material y métodos)

2) A continuación se analizó el reconocimiento de las distintas matrices en las secuencias de los organismos, se inició con el barrido de cada matriz a todo lo largo de los sitios promotores por organismo. Una vez obtenido los *scores* a lo largo de la secuencia, se hicieron gráficas bidimensionales para ver la dispersión de puntos y en base a esto inferir cualitativamente la diferencia entre los *scores* de las matrices. Se construyeron 18 gráficas y las comparaciones se hicieron como señala el diagrama:

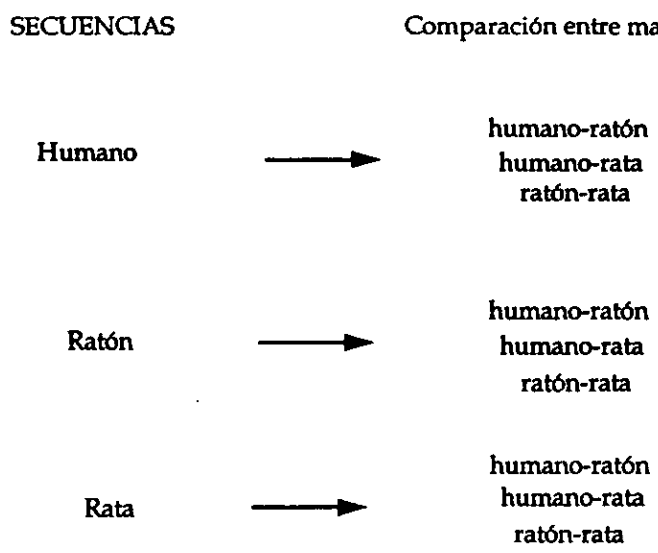


Diagrama 11. En los tres diferentes organismos se hicieron comparaciones tomando pares de matrices. Por cada par de matrices se hicieron dos gráficas dando como resultado 18 gráficas.

Se hizo una doble comparación para cada pareja de matrices. En la primera comparación se graficaron los valores para todos los sitios a lo largo de la secuencia y en el segundo solo se comparan los valores de los sitios reales para Sp1.

1) Las tres matrices para Sp1 a partir de las secuencias de tres organismos diferentes fueron las siguientes:

Humano

T	6	8	0	0	1	4	1	3	0	5	10	14	19	8	11
A	16	6	11	0	0	17	1	0	9	10	4	8	11	7	6
C	9	4	0	0	0	40	3	0	6	9	36	21	11	17	15
G	30	43	50	61	60	0	56	58	46	37	11	18	20	29	29

Ratón

T	3	4	0	0	0	2	1	2	2	2	6	10	8	8	6
A	4	3	4	0	0	0	0	0	8	2	0	1	7	5	2
C	2	2	0	0	0	26	2	0	1	1	18	10	3	4	5
G	20	20	25	29	29	1	26	27	18	24	5	8	11	12	16

Rata

T	2	9	0	0	0	2	1	2	2	0
A	1	2	6	0	0	3	0	0	5	8
C	6	2	0	0	0	21	0	1	2	3
G	18	14	21	27	27	1	26	24	18	16

Se inició el análisis $\chi 2$. En este análisis se tiene 4 categorías (los cuatro diferentes nucleótidos), los grados de libertad son tres (ver formula 9) y se escoge un valor para α de 0.01, para este caso particular, el valor de la $\chi 2$ en tablas es: **11. 34.**

Comparación entre las matrices de humano y ratón.

Pos	dif_chi	chi_histo
1	3.75	==
2	0.02	
3	0.25	
4	0.00	
5	0.48	
6	**11.74**	======
7	0.91	
8	0.15	
9	7.44	====
10	5.05	===
11	2.14	=
12	2.84	=
13	1.35	=
14	4.64	==
15	0.97	

;

Posteriormente se hizo una comparación de estas dos matrices con la matriz de *Rattus norvegicus* (rata). Esta matriz, tiene una longitud menor a nucleótidos (10 nucleótidos) y para llevar a cabo su comparación se hizo una disminución en la longitud de las matrices de humano y ratón para que la región conservada (el *core* del sitio) tuvieran la misma posición.

Comparación entre las matrices de humano y rata

```
;pos     dif_chi chi_histo
1        6.70    |===
2        5.11    |===
3        0.21    |
4        0.00    |
5        0.45    |
6        4.99    |==
7        2.16    |=
8        2.54    |=
9        5.00    |===
10       4.00    |==
```

Comparación entre las matrices de ratón y rata

```
;pos     dif_chi chi_histo
1        4.04    |==
2        3.11    |==
3        0.68    |
4        0.00    |
5        0.00    |
6        3.46    |==
7        1.93    |=
8        1.11    |=
9        0.96    |
10       8.14    |====
```

Al comparar las matrices en sus diferentes posiciones, la única diferencia significativa que se encontró fue para las matrices humano-ratón en la sexta posición, **11.74 > 11. 34**. En esta posición, la matriz de ratón conserva casi de manera perfecta el nucleótido C; mientras que la matriz de humano se encuentran los nucleótidos A y C.

2) Las gráficas se agrupan en tres conjuntos, cada uno corresponde a las secuencias de un organismo diferente (ratón, rata y humano). Una vez separadas por organismo, estas se clasifican nuevamente en grupos de dos.

En el primer grupo se encuentran los valores correspondientes al barrido completo de las secuencias en el organismo y en el segundo se graficaron los valores para los sitios con evidencia experimental de Sp1.

I) Gráficas en las secuencias de *Homo sapiens*:

En las primeras tres gráficas (gráficas 6, 7 y 8), donde se observan las correlaciones que existen para todos los valores a lo largo de las secuencias, se observan diferentes grados de dispersión. La que presenta una dispersión menor, es la gráfica donde se compara las matrices de ratón (eje X) contra la de rata (eje Y), después están los valores graficados para las matrices de rata (eje X) contra la humano (eje Y).

La gráfica 6 que contiene los valores para las matrices de ratón (eje X) y humano (eje Y) es la que presenta una mayor dispersión entre los puntos. Esta gráfica muestra en la parte superior dos nubes de puntos, indicando la existencia de secuencias con valores altos para la matriz de humano calificados con valores bajos por la matriz de ratón.

Al pasar al otro conjunto de gráficas para los sitios experimentales (gráficas 9, 10 y 11), se corrobora el deslizamiento en la gráfica de las matrices ratón-humano; este es de aproximadamente 5 unidades en el contenido informacional. De manera más precisa algunos sitios son calificados en el intervalo de 5-10 para la matriz de humano y de 5 a 0 para la de ratón. Al revisar estas secuencias, se observa que en su sexta posición presentan una **Adenina** en lugar de una **Cisteína**. Es decir, se tiene en la parte central del sitio **GGGAGG**, en lugar de **GGGCGG**.

II) Gráficas en las secuencias de *Mus musculus*:

En el segundo conjunto de gráficas, se evalúan los sitios en 22 secuencias de *Mus musculus*. La distribución de los puntos a lo largo de todas la secuencia muestra el mismo patrón de dispersión que en las secuencias de *Homo sapiens*. Mostrando la gráfica 14, se observan también dos picos en el nivel superior de la gráfica.

Del segundo conjunto de gráficas con los sitios verdaderos para los promotores de ratón, se observa la gráfica 15, donde las matrices de ratón-humano muestra una buena correlación entre ambas matrices.

III) Gráficas en las secuencias de *Rattus norvegicus,*

En el tercer conjunto, donde se evalúan los sitios de *Rattus norvegicus,* los anchos de las nubes de puntos (donde se muestra el grado de dispersión entre los puntos) son muy similares entre sí. Esto también se observa con los valores de los sitios experimentales. La gráfica más interesante de este conjunto es la gráfica 18, donde las matrices humano-ratón muestran una menor dispersión que en los otros promotores. De la misma manera, en la parte superior de las gráficas donde se encuentran dos picos de puntos bien definidos para las otras especies, en las secuencias de rata se muestran los picos diluidos.

Con lo anterior se infiere que, el conjunto de secuencias en rata, presenta una composición distinta que en las especies anteriores.

De manera general lo que se puede dilucidar de este análisis es que la matriz humana puede calificar bien los sitios para Sp1 en humano y rata, pero no los de ratón. De la misma manera esto se puede decir de la matriz de ratón, califica muy bien los sitios para ratón y rata pero no así los de humano. Esto se debe a la diferencia que existe en la sexta posición (la matriz en humano reconoce GGGAGG y la de ratón no) en ambas matrices; lo cual ocasiona que la matriz de ratón no reconozca con un buen *score* algunos sitios para humano (perdiendo sensibilidad) y la matriz humana reconozca sitios falsos de Sp1 en las secuencias de ratón (perdiendo especificidad).

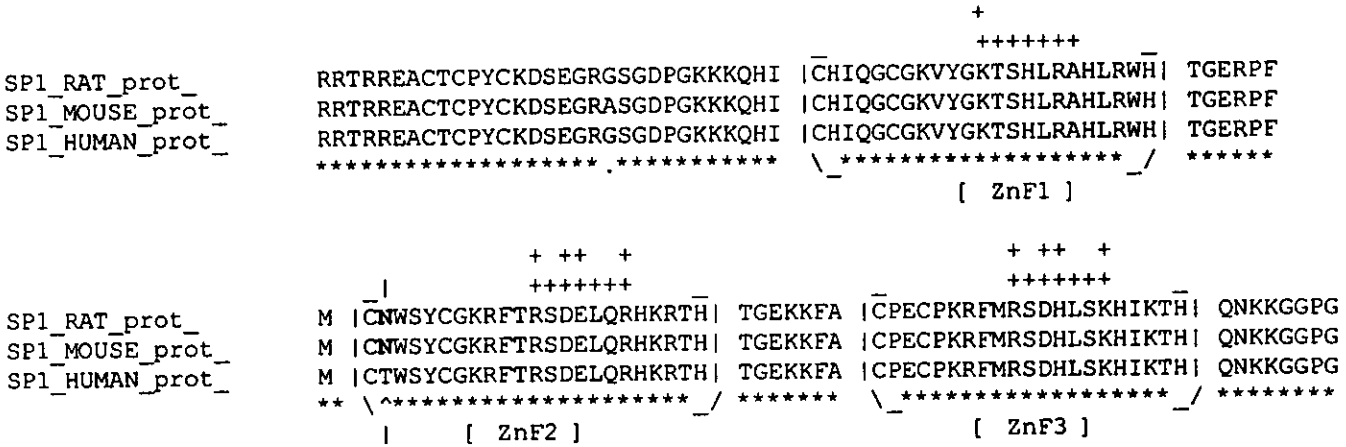
3.3 Comparación del dominio carboxilo terminal para Sp1 en diferentes organismos

Con base en los resultados obtenidos comparando las distintas matrices (por organismos) de Sp1, se examinaron los dominios de reconocimiento al DNA de las proteínas Sp1 en humano, ratón y rata; esto con la finalidad de encontrar diferencias en esta región que pudiesen explicar el reconocimiento diferencial en los sitios para una misma proteína en dos distintas especies (humano y ratón).

Se hizo un alineamiento con la estructura primaria del dominio de reconocimiento de Sp1 al DNA. Se escogieron para esta comparación las secuencias de Sp1 para humano, ratón y rata. El alineamiento en la región carboxilo terminal va del aminoácido 507 al 626 (con respecto al número de aminoácidos en *Hommo sapiens* para sp1). En esta región, se encuentran contenidos los tres dedos de zinc que se encargan del reconocimiento de la caja GC sobre el DNA (fig.12)

Como se puede ver en el alineamiento, las regiones de la hélice alfa que participan en el reconocimiento al DNA, son idénticas. En los motivos *zinc fingers*, se encuentra sólo una diferencia (marcada con ^ y negritas) en el segundo aminoácido del segundo dedo de Zinc, (una Asparagina por una Treonina, ambos aminoácidos polares). El segundo dedo de Zinc es el que participa en el reconocimiento de la quinta, sexta y séptima posición en la caja GC.*

Esto podría sugerir un pequeño cambio en la estructura del segundo dedo de zinc que provocará un reconocimiento diferencial entre la secuencia GGGCGG y GGGAGG. Mientras que Sp1 en humano parece reconocer con alta afinidad GGGCGG y GGGAGG, el Sp1 de ratón solo reconoce GGGCGG. Para comprobar esto, se tienen que hacer experimentos con ambas proteínas Sp1 y medir la afinidad (constante de disociación) que tienen por ambas secuencias.



Figuna 12. Alineamiento de la región carboxilo terminal de las proteinas Sp1 correspondientes a *Rattus norvegicus* , *Homo sapiens* y *Mus musculus*. Entre corchetes estan los aminoacidos que forman cada uno de los dedos de Zinc (fig.8) Los aminoacidos con el signo '+' en la parte superior corresponde a la región de la alpha hélice que participa en el reconocimiento en el DNA. Doble signo '++' son los aminoacidos que contactan directamento con los nucleotidos.

Lo que podría apoyar esta hipótesis son los análisis hechos en las familia de las proteínas Sp (Lania, 1997). La proteína Sp3 y Sp4 reconocen las secuencias GGGCGG y GGGTGG con una alta afinidad, semejante a la de la proteína Sp1. Sin embargo la proteína Sp2 reconoce con un alta afinidad la secuencia GC y con baja afinidad la secuencia GT. De manera general esta secuencia presenta una baja similitud fuera del dominio de reconocimiento al DNA, pero en el dominio que constituye a los tres dedos de zinc se encuentra una alta similitud. En la secuencia donde se forma el segundo dedo de zinc (que es la que contacta con los nucleótidos GC) Sp2 presenta una ligera modificación con respecto al resto de los miembros de familia; las proteínas Sp1, Sp3 y Sp4 presentan en una posición precisa un aminoácido cargado, lisina (Sp1) y arginina (Sp3-Sp4) (ver figura 13) mientras que la proteína Sp2 presenta un aminoácido hidrofóbico.

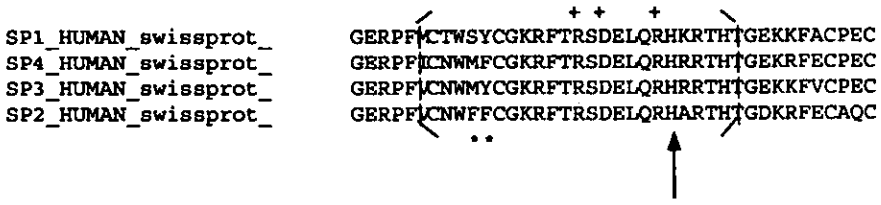


Figura 13. Alineamiento del dominio carboxilo terminal en el segundo dedo de Zinc. Los + señalan los aminoácidos que contactan directamente con el DNA. Los * señalan las diferencias. La flecha señala la posición donde se encuentran los aminoácidos cargados positivamente para Sp1, Sp3 y Sp4.

De manera general este es otro caso que podría proporcionar un argumento a favor del estudio de afinidad y estequiométrico en los Sp1 de diferentes organismos, para dilucidar las posibles diferencias en el reconocimiento de la proteína a las secuencias de DNA y las eventuales implicaciones que se pudiera tener a nivel de la estructura debido a pequeños cambios en las secuencias de aminoácidos.

CONCLUSIONES Y PERSPECTIVAS

Los métodos de búsqueda para sitios reguladores en *cis* en factores de transcripción, encuentran (en la mayoría de los casos) un gran número de sitios (gran parte de ellos presuntos falsos positivos) cuya corroboración experimental se vuelve prácticamente imposible debido a lo exhaustivo del trabajo.

La manera en la cual pretendemos mejorar la discriminación de estos presuntos sitios falsos positivos es buscando información biológica del sistema e integrarla como filtro para una mayor certidumbre en las predicciones.

En el caso particular de la proteína Sp1, se analizaron tres posibles tipos de filtros biológicos.

- El primer análisis involucró un perfil de los sitios de Sp1 sobre el resto de la secuencia, esto muestra tres tipos grupos para una colección de 19 promotores. En el primer grupo los sitios de Sp1 forman una colección aislada del resto de la secuencia. En el segundo, los sitios de Sp1 junto con otros sitios forman un conjunto independiente con respecto al resto del promotor. El último grupo muestra la dilución de los sitios de Sp1 con el resto de la secuencia.

  Esto nos indica como el poder discriminatorio de una matriz, puede dar interpretaciones diferentes dependiendo del contexto mismo del promotor. Un análisis que sería interesante para continuar en esta dirección es la búsqueda de algún tipo de propiedad convergente entre los promotores de un mismo grupo, como puede ser la función o el momento del desarrollo en que son expresados estos genes.

- El segundo análisis involucró las posiciones preferidas de Sp1 con respecto al inició de la transcripción en una colección de 19 promotores. Se encontraron las posiciones favoritas y en base a estas, se construyó un índice que integra la posición y el *score* (dado por una matriz). Con esto se eliminó un gran número de sitios y se obtuvo mayor certidumbre en la predicción.

La manera en que fue construído el índice y los valores que toman los diferentes coeficientes son heurísticos, pero la manera en que puede ir adquiriendo mayor robustez es a través de experimentos que corroboren el valor predictivo de nuestro método.

- El último análisis hecho fue una comparación de matrices de búsqueda en sitios Sp1 para diferentes organismos (ratón, rata y humano). Comparando las diferentes matrices se encontró que en las matrices de humano y de ratón existía una diferencia significativa en un nucleótido, dando un reconocimiento diferencial para la secuencia GGGCGG (reconocida por ambas proteínas) y GGGAGG (reconocida por la proteína Sp1 humana). También nos muestra como la matriz de humano reconoce mejor a los sitios de su propia especie que a los de ratón y viceversa.
- A raíz de este resultado se buscó una justificación que explicara la causa del reconocimiento diferencial. Es en esta dirección que se ahondo en la estructura primaria de ambas proteínas y se hizo una comparación. Se encontró un aminoácido diferente (una Asparagina en lugar de una Treonina). Una perspectiva en esta dirección sería hacer un análisis estequeométrico buscando diferencias a nivel de estructura que pudiera estar ocasionando un reconocimiento diferencial entre los sitios.

De manera general con este trabajo se quiso mostrar lo importante que puede ser el conocer y aplicar los diferentes métodos computacionales incorporando características propias del sistema bajo el cual se quieren hacer predicciones.

Material y Método

-Recopilación de datos sobre sitios regulados por Sp1

Se inició una búsqueda bibliográfica sobre genes regulados por la proteína Sp1. A partir de esta búsqueda se recolectó información general de los elementos *cis* y *trans* para cada promotor : i)Nombre del promotor, ii)proteína, iii) posición inicial de pegado, iv) posición final de pegado (con respecto al +1 del inicio de la transcripción) v) referencia bibliográfica (número de MEDLINE) vi) tipo de evidencia experimental y vii) secuencia.

Con la información obtenida se construyó una mini-base de datos utilizando el manejador de Sybase en el lenguaje SQL. La colección inicial que consta de :

19 promotores

82 sitios para proteínas reguladoras.

24 tipos diferentes de sitios para proteínas reguladoras

La información de esta recopilación se puede consultar en la sig. dirección electrónica: http://www.cifn.unam.mx/~victoria/sp1.html.

Se generó un mapa físico de los 19 promotores con el programa **Feature_map** (van Helden, 1998).

-Construcción de matrices y detección de sitios

Con secuencias de pegado al DNA para una misma proteína se construye una matriz de peso para el reconocimiento de sitios bajo el siguiente protocolo:

a) Se extienden (en ambos extremos) las secuencias de reconocimiento al DNA, b) se construye un alfabeto basado en las frecuencias absolutas para cada nucleótido en la región promotora y c) se corre el programa wcosensus. **Wconsensus** realiza alineamientos locales múltiples. Su archivo resultante es una matriz de peso *weight matrix*, que va construyéndose por ciclos; en el primer ciclo se construyen las primeras matrices, a las que se incorporan progresivamente

nuevas secuencias y se van salvando las matrices cuyo valor de contenido informacional se va incrementando.

Los archivos de salida de Wconsensus contiene tres medidas asociadas al contenido informacional del alineamiento:

-*Unadjusted information* es simplemente el contenido informacional del alineamiento.

-*Sample size adjusted information* se obtiene restando al *unadjusted information*, el promedio del contenido informacional esperado para un alineamiento al azar.

-*Crude information* se obtiene restando para cada columna del *Sample size adjusted information* el valor de la desviación *standar*.

-*Expected frequency*, se refiere a la probabilidad de encontrar el valor del contenido informacional en un alineamiento al azar.

Una vez hecha la matriz se realizan las búsquedas para detectar otros posibles sitios de reconocimiento para la proteína. Se utiliza el programa **Patser.** Patser barre todas las posiciones y evalúa la longitud igual al número de columnas en la matriz y compara el valor o *score* versus un valor umbral de *score*. Si dicho valor es más alto que el umbral, se considera la secuencia como un sitio potencial de pegado para la proteína reguladora que reconoce a los sitios utilizados en generar dicha matriz de peso, Sp1 en nuestro caso. Entre otras, una opción al usuario es correr la búsqueda en ambas direcciones. El programa convierte los valores de la matriz utilizando la fórmula: $\log_2 \frac{f_{b,i}}{Pb}$ ...(1) (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.

Ambos algoritmos se basan en la **Teoría de la información**. De manera general la teoría de la información se utiliza al estudiar sitios de pegado al DNA o RNA para proteínas. Para cuantificar la variabilidad y medir la información

que nos proporciona un patrón y poderlo distinguirlo del resto de la secuencia se utiliza el de ***contenido informacional***. El contenido informacional (también llamado **Rsecuencia**) 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 $Pi = 0.25$). Se utiliza el logaritmo base 2 porque el m ximo valor para la incertidumbre es de dos bits.

Cuando las frecuencias de las bases no son exactamente 0.25, 0.50 0 1.0, el cálculo de incertidumbre está en función de la frecuencia $f(b,l)$ de cada base b en la posición l:

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

donde $e(n(l))$ es una corrección para la muestra pequeña de tamaño n en la posición l. El Contenido Informacional (conservación de secuencia) 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 de una secuencia de pegado al DNA sumando todos los valores de *Rsecuencia* de las posiciones del sitio.

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

La conservación de la secuencia de pegado para una proteína va en relación directa con el tamaño del genoma G y el número de veces que este sitios se encuentra dentro del genoma γ. Esto se define con otra medida que se llama **Rfrecuencia**, 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.

-Análisis bidimensional de sitios en gráficas

Se construyeron las gráficas XY con el programa **XYgraph.** Disponible en la dirección electrónica:
http://copan.cifn.una.mx/Computational_Bio-logy/yeast-tools
(van Helden, 1998).

-Creación de un índice incorporando la distancia

El propósito de crear un nuevo índice fue mejorar la discriminación de algunos sitios mediante la inclusión de la posición respecto al inicio de transcripción.

Los criterios para definir dicho índice son los siguientes, por un lado queremos definir un índice que integre el valor definido por la matriz de peso, más una contribución por posición. Esto representa un problema de sumar o multiplicar valores con unidades diferentes. Un segundo problema es la escases de posiciones válidas, ya que no se quiere castigar sitios en posiciones que no se encuentran representadas en nuestra colección de 33 sitios. Se opto por una suma

que considera valores adimensionales, pues se logra que un sitio predicho en una posición altamente representada en la colección de los 33 sitios, se vea beneficiado, pero sin anular a los sitios que se encuentran en posiciones no representadas (que sería el caso al utilizar la multiplicación).

Una manera de resolver estos criterios es el siguiente índice SP (score posición):

$$SP = \alpha_1 \frac{Freq.absolute}{Freq.total} + \alpha_2 \frac{Score - site}{Score - Average} = Indice.score / position \ldots.(7)$$

donde:

α_1 y α_2: son los coeficientes que se utilizan para modificar los valores resultantes de los cocientes, ya sea para hacer números comparables entre dos valores adimensionales o para ponderar un parámetro con respecto a otro. En nuestro caso particular los cocientes tienen valores del mismo orden por lo que se les asignó a los coeficientes valores de uno.

Freq absolute es la frecuencia de sitios conocidos en cierto intervalo de posiciones (tomamos intervalos de 25pb).

Freq total es la frecuencia o número total de sitios conocidos, (en este caso 33).

Score-site: es el score o calificación de la secuencia del sitio definido por la matriz de peso.

Score-average: es el score promedio de todos los (33) sitios conocidos.

Obsérvese que en efecto, SP es un número adimensional obtenido de una suma de cocientes. En segundo lugar, obsérvese que los valores de (Freq.absolute/Freq.total) se encuentran en un intervalo entre 0 y 1, por lo que un sitio en una posición no representada en la colección de los sitios experimentales, no es castigado, mientras que un sitio predicho en una posición existente en la colección tendrá un SP incrementado debido a que (Freq.absolute/Freq.total) es mayor que cero. Queremos puntualizar, que la forma en la cual se ha definido el índice SP, tiene cierto grado de arbitrariedad,

pero heurísticamente fue la mejor manera en que se pudo definir este índice, ya que con la suma se mantiene la información y los valores de los coeficientes se pueden ponderizar dependiendo del sistema.

-Búsqueda de diferencias significativas en matrices

El análisis si hizo con el programa differential-profile (Jacques van Helden, *unpublished*) que utiliza el estadístico $\chi 2$, el cual proporciona información sobre el grado de discrepancia entre dos frecuencias (Mode, 1980).
La formula es:

$$\chi^2 = \sum_{j=1}^{k} \frac{(o_j - e_j)^2}{e_j} \quad \text{....(8)}$$

Cuando las frecuencias coinciden completamente $\chi 2$ vale 0. Cuando aumenta la discrepancia entre las frecuencias el valor de $\chi 2$ irá en aumento

La distribución muestral de $\chi 2$ depende en gran medida de los grados de libertad (v). Este valor es la substracción de la unidad al Número de categorías (κ). Es decir: $\nu = \kappa - 1$... (9) .Para el caso particular de las bases nucleotídicas el Número de categorías son 4.

A partir de esta información se hace un contraste de significación sobre una hipótesis nula *Ho*. Para este análisis la hipótesis nula es que las frecuencias de nucleótidos en una posición para las matrices de Sp1, son similares. Para el contraste de esta hipótesis se utiliza un valor de significatividad entre 0.99 o 0.95 de la distribución $\chi 2$, con una α (riesgo a equivocarse) de 0.05 y 0.01.

Estos valores para $\chi 2$tabla se contrastan con la $\chi 2$ muestral, en el supuesto que: $\chi 2 > \chi 2$tabla se rechaza la *Ho*; el caso contrario se acepta *Ho*.

- Alineamiento del dominio de reconocimiento de la proteína.

Se hizo un alineamiento del dominio de reconocimiento al DNA para la proteína Sp1 de tres diferentes organismos. Para ello se utilizó el programa CLUSTAL W (Thompson, 1994). El método básico de este alineamiento se lleva a cabo en tres etapas: i)se crea una matriz diagonal de distancias (normalizada a uno) donde se muestran las diferencias en el número de residuos para cada pareja de secuencias, ii)se construye un árbol sin raíz, donde el largo de las ramas representan el valor de divergencia estimado y iii)se construye un alineamiento progresivo basado en el orden de las ramas del árbol guía.

BIBLIOGRAFÍA

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

Altschmied J., Muller M., Baniahmad A., Steiner C., Renkawitz R. 1989. Cooperative interaction of chicken lysozyme enhancer sub-domain partially overlapping with a steroid receptor binding site. Nucleic Acids Res. 17:4975.

Ammendola R., Gounari F., Piaggio G., De Simone V., Cortese R. 1990. Transcription of the Promoter of the Rat NF-1 Gene Depends on the Integrity of an Sp1 Recognition Site. Mol. Cell. Biol. 10:387-390.

Araki E., Murakami T., Shirotani T., Kanai F., Shinohara Y., Shimada F., Mori M., Shichiri M., Ebina Y. 1991. A cluster of four Sp1 binding sites required for efficient expression of the human insulin receptor gene. J. Biol. Chem. 266:3944-3948.

Birnbaum MJ, Wright KL, van Wijnen AJ, Ramsey-Ewing AL, Bourke MT, Last TJ, Aziz F, Frenkel B, Rao BR, Aronin N. 1995. Functional role for Sp1 in the transcriptional amplification of a cell cycle regulated histone H4 gene. Biochemistry; 34(23):7648-58

Bjorklund S, Kim YJ. 1996. Mediator of transcriptional regulation. Trends Biochem Sci; 21(9):335-7

Boyer T. G., Maquat L. E. 1990. Minimal sequence and factor requirements for the initiation of transcription from an atypical, TATATAA box-containing housekeeping promoter. J. Biol. Chem. 265:20524-20532.

Chen JL, Attardi LD, Verrijzer CP, Yokomori K, Tjian R. 1994. Assembly of recombinant TFIID reveals differential coactivator requirements for distinct transcriptional activators.Cell; 79(1):93-105

Chen QK, Hertz GZ, Stormo GD. 1995. MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. Comput Appl Biosci; 11(5):563-6

Cheung CH, Fan QN, Stumph WE. 1993. Structural requirements for the functional activity of a U1 snRNA gene enhancer.Nucleic Acids Res; 21(2):281-7

Courey AJ, Holtzman DA, Jackson SP, Tjian R. 1989. Synergistic activation by the glutamine-rich domains of human transcription factor Sp1. Cell; 59(5):827-36

Courey AJ & Tjian R. 1992. Mechanisms of Transcriptional Control as Revealed by Studies of Human Transcription Factor Sp1. In : Transcriptional Regulation. Cold Spring Harbor Laboratory Press. pp: 743-769.

de Groot R. P., Karperien M., Pals C., Kruijer W. 1991. Characterization of the mouse junD promoter - high basal level activity due to an octamer motif. EMBO J. 10:2523-2532.

Fischer KD, Haese A, Nowock J. 1993. Cooperation of GATA-1 and Sp1 can result in synergistic transcriptional activation or interference. J Biol Chem; 268(32):23915-23

Fondrat C, Kalogeropoulos A. 1996. Approaching the function of new genes by detection of their potential upstream activation sequences in *Saccharomyces cerevisiae*: application to chromosome III.Comput Appl Biosci; 12(5):363-74

Goding C. R., Temperley S. M., Fisher F. 1987.Multiple transcription factors interact with the adenovirus-2 EII-late promoter: evidence for a new CCAAT recognition factor. Nucleic Acids Res. 15:7761-7780.

Hampsey M. 1998. Molecular genetics of the RNA polymerase II general transcriptional machinery. Microbiol Mol Biol Rev; 62(2):465-503

Hagen G, Muller S, Beato M, Suske G. 1994. Sp1-mediated transcriptional activation is repressed by Sp3. EMBO J; 13(16):3843-51

Heinemeyer T, Wingender E, Reuter I, Hermjakob H, Kel AE, Kel OV, Ignatieva EV, Ananko EA, Podkolodnaya OA, Kolpakov FA, Podkolodny NL, Kolchanov NA. 1998. Databases on Transciptional Regulation: TRANSFAC, TRRD, and COMPEL. Nucleic Acids Res. 26, 362-367.

Hertz G.Z, Hartzell GW, Stormo GD. 1990. Identification of consensus patterns in unaligned DNA sequences known to be functionally related. Comput Appl Biosci 6:81-92.

Hertz G.Z, Stormo, 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 Genome Research, World Scientific Publishing, Singapore: 201-216.

Jones KA, Kadonaga JT, Luciw PA, Tjian R. 1986. Activation of the AIDS retrovirus promoter by the cellular transcription factor, Sp1. Science; 232(4751):755-9

Jones KA, Yamamoto KR, Tjian R. 1985. Two distinct transcription factors bind to the HSV thymidine kinase promoter in vitro. Cell; 42(2):559-72

Kasai Y, Chen H, Flint SJ. 1992. Anatomy of an unusual RNA polymerase II promoter containing a downstream TATA element. Mol Cell Biol; 12(6):2884-97

Karlseder J, Rotheneder H, Wintersberger Earlseder. 1996. Interaction of Sp1 with the growth- and cell cycle-regulated transcription factor E2F. Mol Cell Biol; 16(4):1659-67

Kuwahara J, Yonezawa A, Futamura M, Sugiura Y. 1993. Binding of transcription factor Sp1 to GC box DNA revealed by footprinting analysis: different contact of three zinc fingers and sequence recognition mode. Biochemistry; 32(23):5994-6001

Lania L, Majello B, De Luca P. 1997. Transcriptional regulation by the Sp family proteins. Int J Biochem Cell Biol; 29(12):1313-23

Li R, Hodny Z, Luciakova K, Barath P, Nelson BD. 1996. Sp1 activates and inhibits transcription from separate elements in the proximal promoter of the human adenine nucleotide translocase 2 (ANT2) gene. J Biol Chem; 271(31):18925-30

Liu B., Hammer G. D., Rubinstein M., Mortrud M., Low M. J.1992. Identification of DNA elements cooperatively activating proopiomelanocortin gene expression in the pituitary glands of transgenic mice. Mol. Cell. Biol. 12:3978-3990.

Mastrangelo IA, Courey AJ, Wall JS, Jackson SP, Hough PV. DNA looping and Sp1 multimer links: a mechanism for transcriptional synergism and enhancement. Proc Natl Acad Sci U S A; 88(13):5670-4.

Mode EB. 1980. Elementos de Probabilidad y Estadística. Editorial RevertÈ Mexicana, S. A. pp: 79-87.

Morgan JG, Courtois G, Fourel G, Chodosh LA, Campbell L, Evans E, Crabtree GR. 1988. Sp1, a CAAT-binding factor, and the adenovirus major late promoter transcription factor interact with functional regions of the gamma-fibrinogen promoter.Mol Cell Biol; (6):2628-37

Morgan WD. 1989. Transcription factor Sp1 binds to and activates a human hsp70 gene promoter.Mol Cell Biol; 9(9):4099-104

Narayan VA, Kriwacki RW, Caradonna JP. 1997. Structures of zinc finger domains from transcription factor Sp1. Insights into sequence-specific protein-DNA recognition. J Biol Chem; 272(12):7801-9.

Pascal E, Tjian R. 1991. Different activation domains of Sp1 govern formation of multimers and mediate transcriptional synergism. Genes Dev;(9):1646-56.

Perez-Rueda E, Gralla JD, Collado-Vides J. 1998.Genomic position analyses and the transcription machinery. J Mol Biol; 275(2):165-70

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

Rosenblueth DA, Thieffry D, Huerta AM, Salgado H, Collado-Vides J. 1996. Syntactic recognition of regulatory regions in Escherichia coli. Comput Appl Biosci;12(5):415-22

Schmidt M. C., Zhou Q., Berk A. J. 1989. Sp1 Activates Transcription without Enhancing DNA-Binding Activity of the TATA Box Factor. Mol. Cell. Biol. 9:3299-3307.

Seal SN, Davis DL, Burch JB. 1991. Mutational studies reveal a complex set of positive and negative control elements within the chicken vitellogenin II promoter. Mol Cell Biol (5):2704-17

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

Su W, Jackson S, Tjian R, Echols H. 1991. DNA looping between sites for transcriptional activation: self-association of DNA-bound Sp1. Genes Dev; 5(5):820-6

Tamaki T, Ohnishi K, Hartl C, LeRoy EC, Trojanowska M. 1995. Characterization of a GC-rich región containing Sp1 binding site(s) as a

constitutive responsive element of the alpha 2(I) collagen gene in human fibroblastsJ Biol Chem; 270(9):4299-304 .

Tamura T.-A., Mikoshiba K. 1991. Role of a GC-rich motif in transcription regulation of the adenovirus type 2 IVa2 promoter which lacks typical TATA-box element. FEBS Lett. 282:87-90.

Tansey WP, Herr W. 1997. TAFs: guilt by association?. Cell; 88(6):729-32

Tebb G, Mattaj IW. 1989. The Xenopus laevis U2 gene distal sequence element (enhancer) is composed of four subdomains that can act independently and are partly functionally redundant. Mol Cell Biol; 9(4):1682-90

Therrien M, Drouin J. 1991. Pituitary pro-opiomelanocortin gene expression requires synergistic interactions of several regulatory elements. Mol Cell Biol (7):3492-503

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.Nucleic Acids Res; 22(22):4673-80

van Helden J, Andre B, Collado-Vides J. 1998. Extracting regulatory sites from the upstream región of yeast genes by computational analysis of oligonucleotide frequencies.J Mol Biol; 281(5):827-42

von der Ahe D., Pearson D., Nakagawa J.-i., Rajput B., Nagamine Y. 1988. Multiple nuclear factors interact with promoter sequences of the urokinase-type plasminogen activator gene Nucleic Acids Res. 16:7527.

Winter E, Varshavsky A. 1989. A DNA binding protein that recognizes oligo(dA) . oligo(dT) tracts. EMBO J; (8): 1867-1877.

Wolffe AP. 1996. Histone deacetylase: a regulator of transcription.Science; 272(5260):371-2

Wu L., Rosser D. S. E., Schmidt M. C., Berk A. J. 1987. A TATA box implicated in E1A transcriptional activation of a simple adenovirus 2 promoter. Nature 326:512-515 (1987).

Yieh L, S nchez HB, Osborne TFYieh L, S nchez HB, Osborne TF. 1995.Domains of transcription factor Sp1 required for synergistic activation with

sterol regulatory element binding protein 1 of low density lipoprotein receptor promoter. Proc Natl Acad Sci U S A; 92(13):6102-6.

Yokono M, Saegusa N, Matsushita K, Sugiura Y. 1998. Unique DNA binding mode of the N-terminal zinc finger of transcription factor Sp1. Biochemistry; 37(19):6824-32

Yoshida K, Narita M, Fujinaga K. 1989. Binding sites of HeLa cell nuclear proteins on the upstream región of adenovirus type 5 E1A gene. Nucleic Acids Res;17(23):10015-34

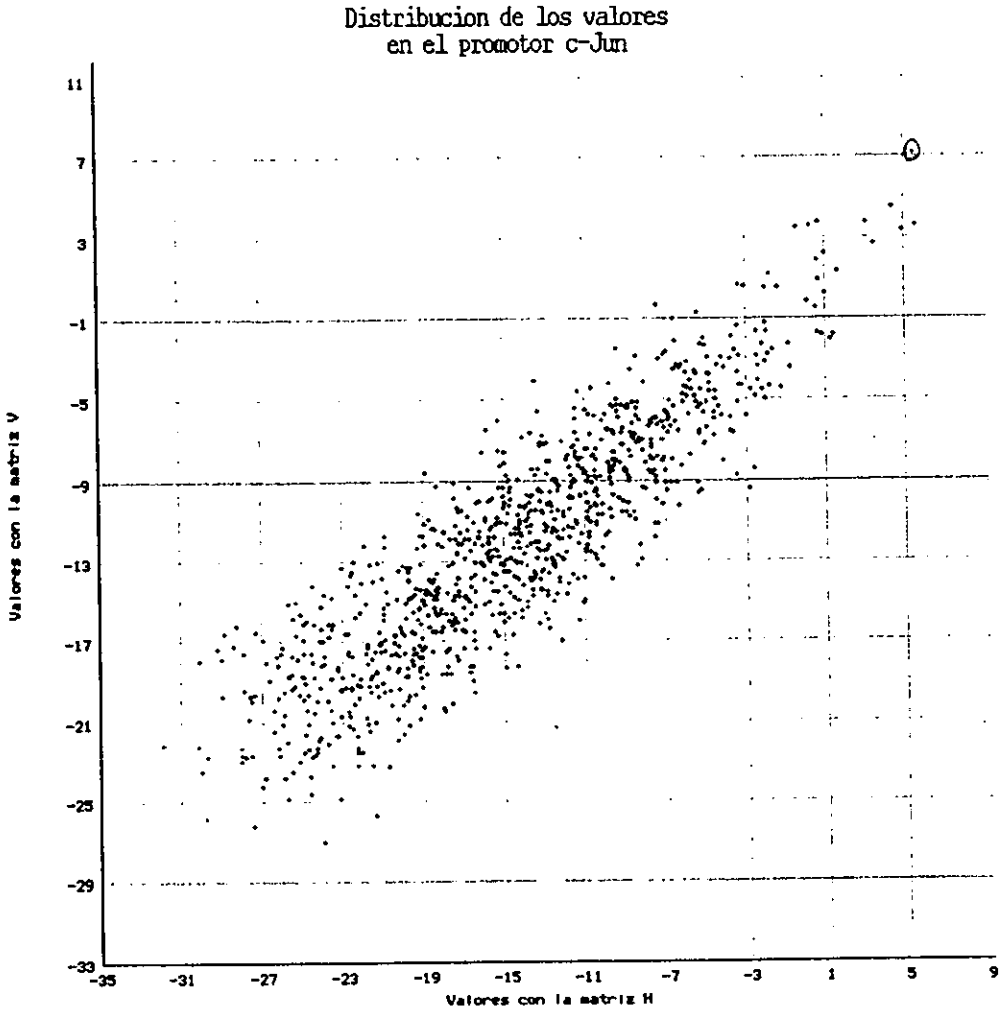
Apéndices:

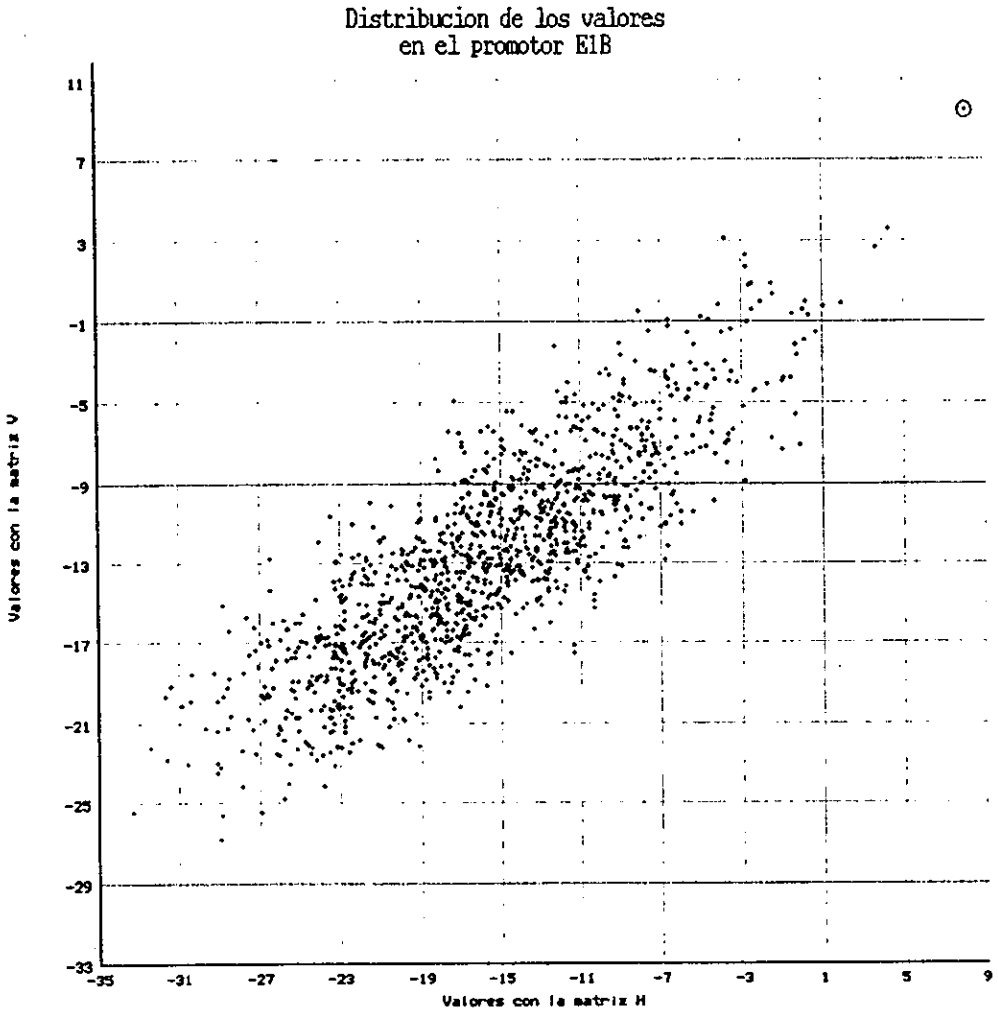
En esta sección se presentan las gráficas en el siguiente orden:

I) Aparecen las gráficas donde se comparan los valores asignados por las matrices H y V sin umbral, para cada uno de los 19 promotores. Este grupo de gráficas se subdividen en tres clases dependiendo de la ubicación de los sitios de Sp1 (pag. 71-89).

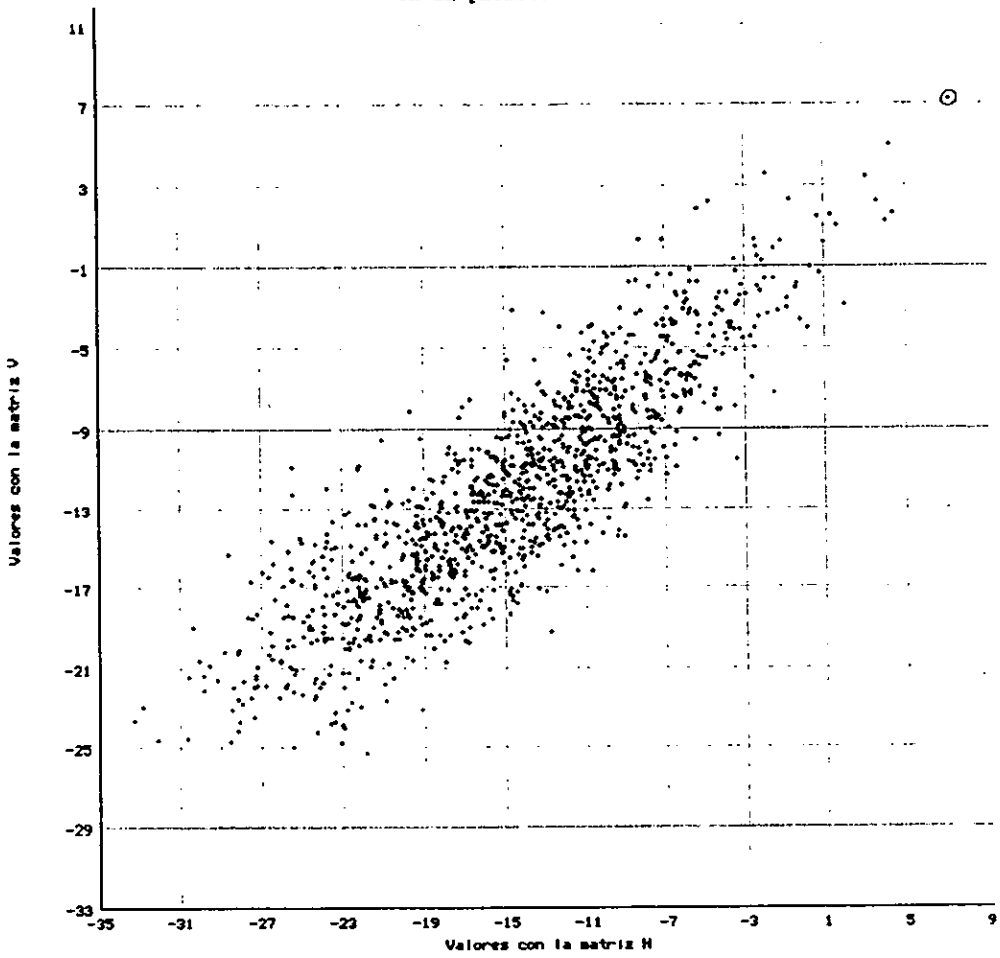
II) Las gráficas donde se muestran las comparaciones entre las diferentes matrices por organismos en los promotores de Humano, Ratón y Rata (pag. 91-98).

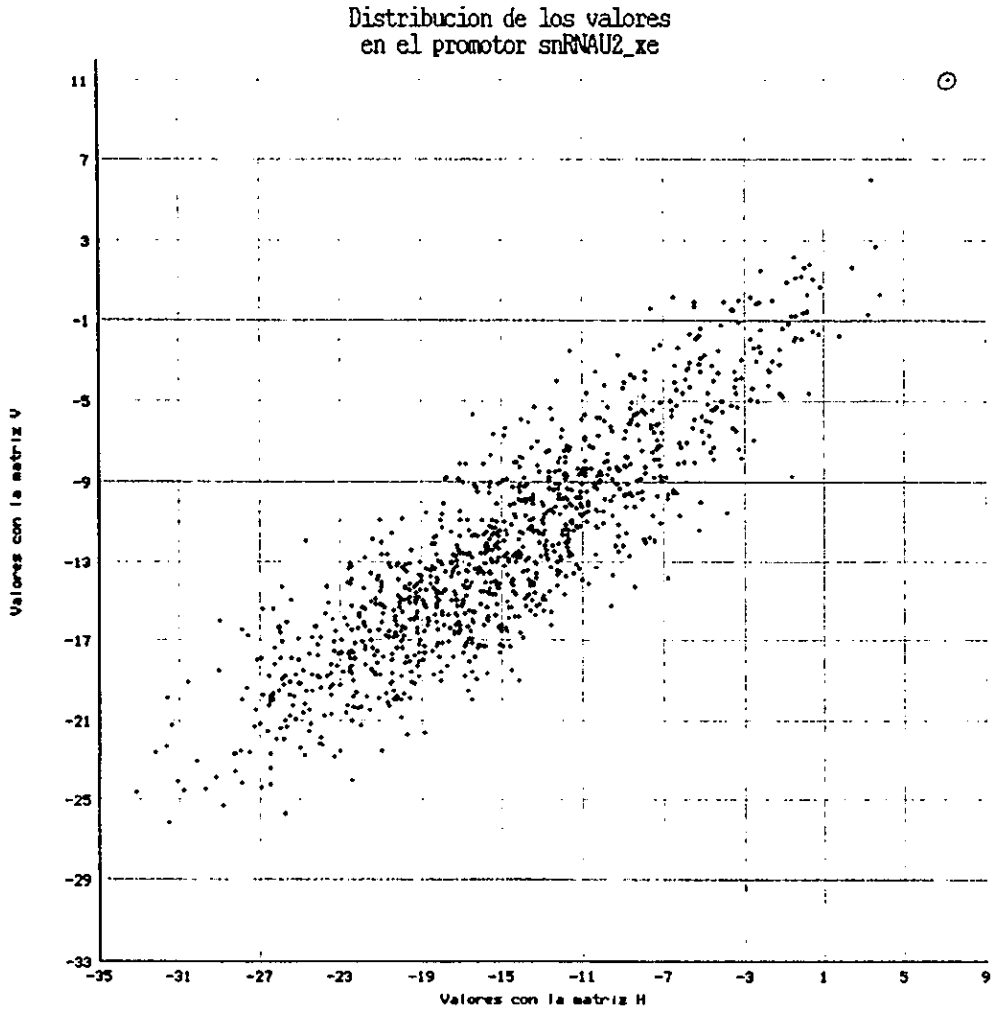
Promotores donde la secuencia de reconocimiento para Sp1 se distingue claramente del resto de los sitios

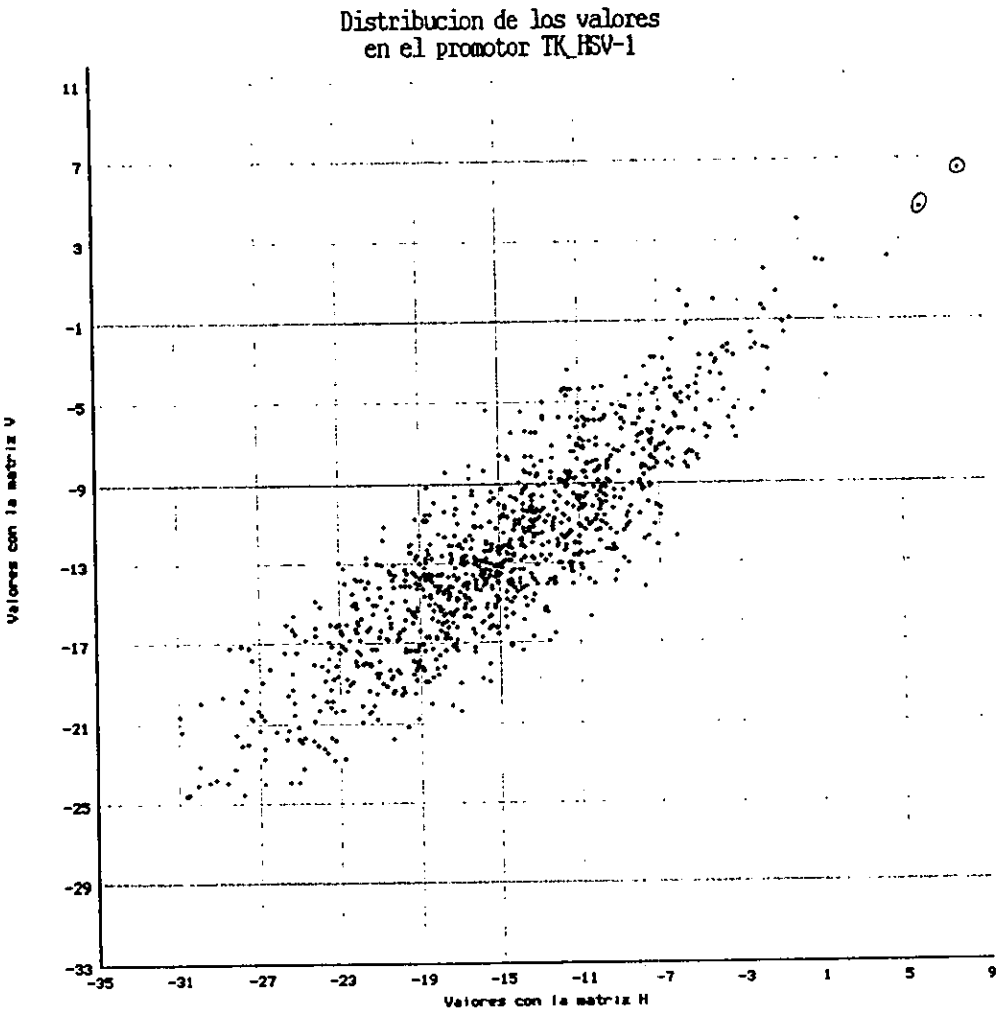

Distribucion de los valores
en el promotor c-Jun
Valores con la matriz V
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9
Valores con la matriz H

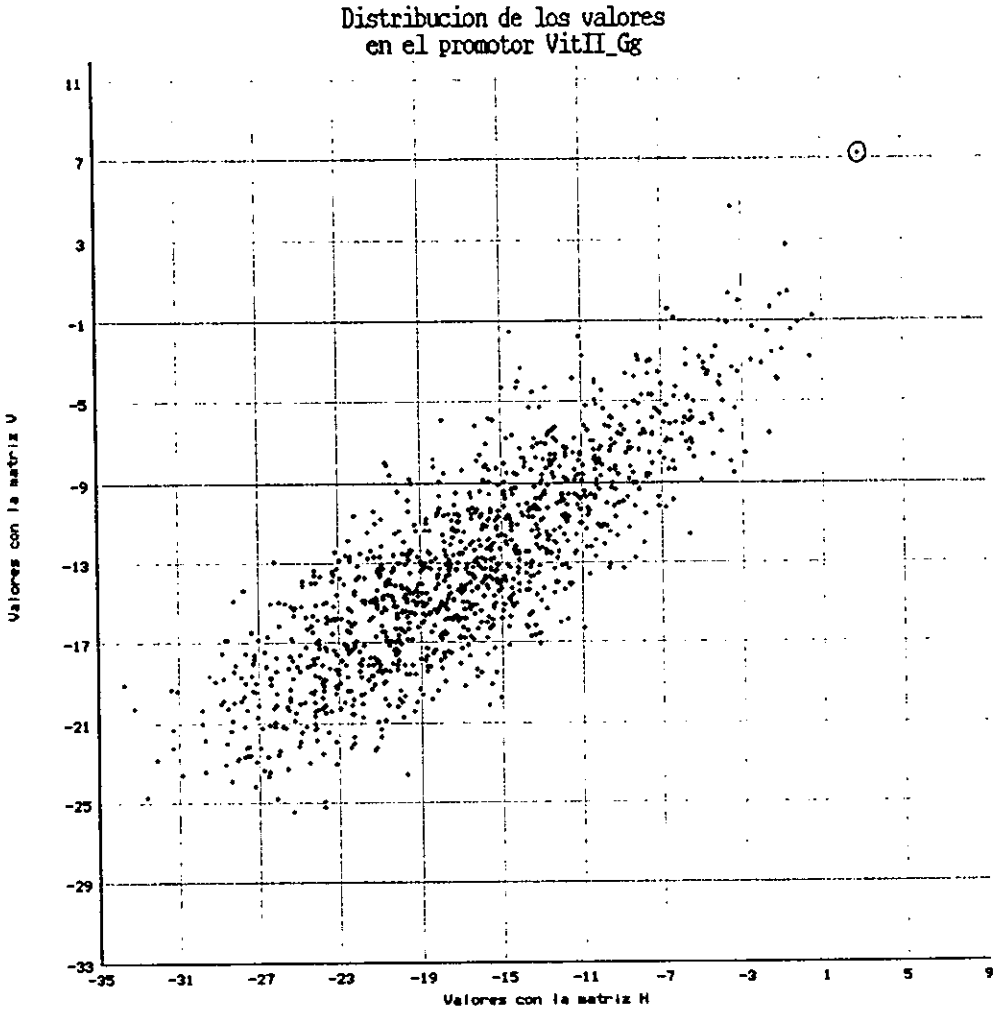

Distribucion de los valores
en el promotor E1B
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9
Valores con la matriz V
Valores con la matriz H

Distribucion de los valores
en el promotor IVa2

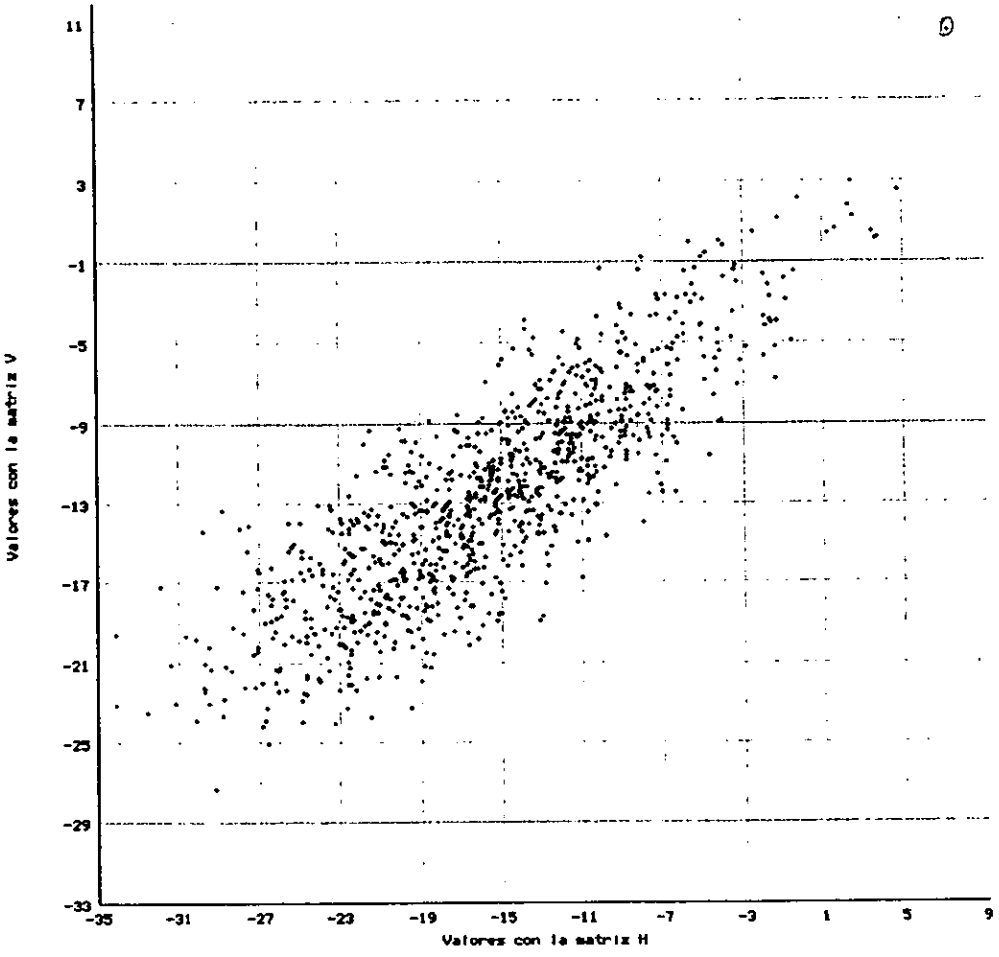



Distribucion de los valores
en el promotor snRNAU2_xe
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

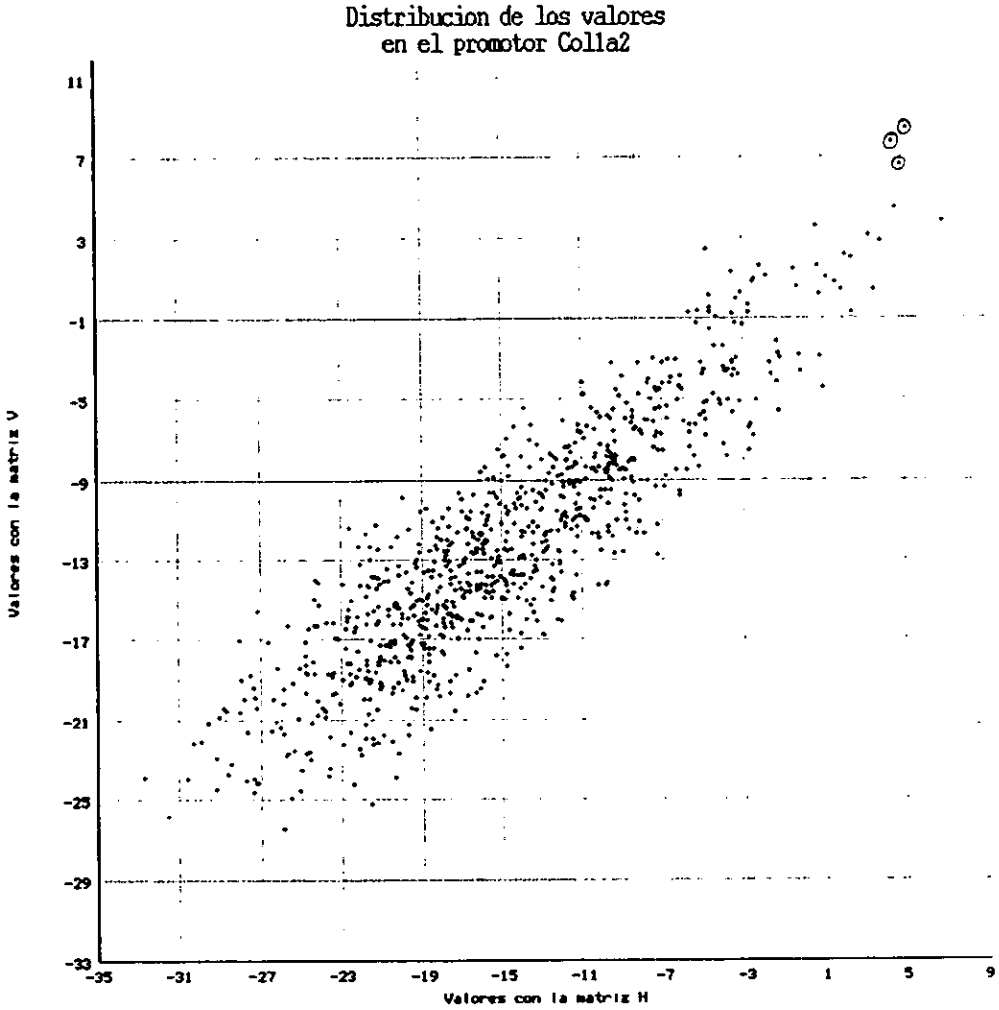

Distribucion de los valores
en el promotor TK_HSV-1
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

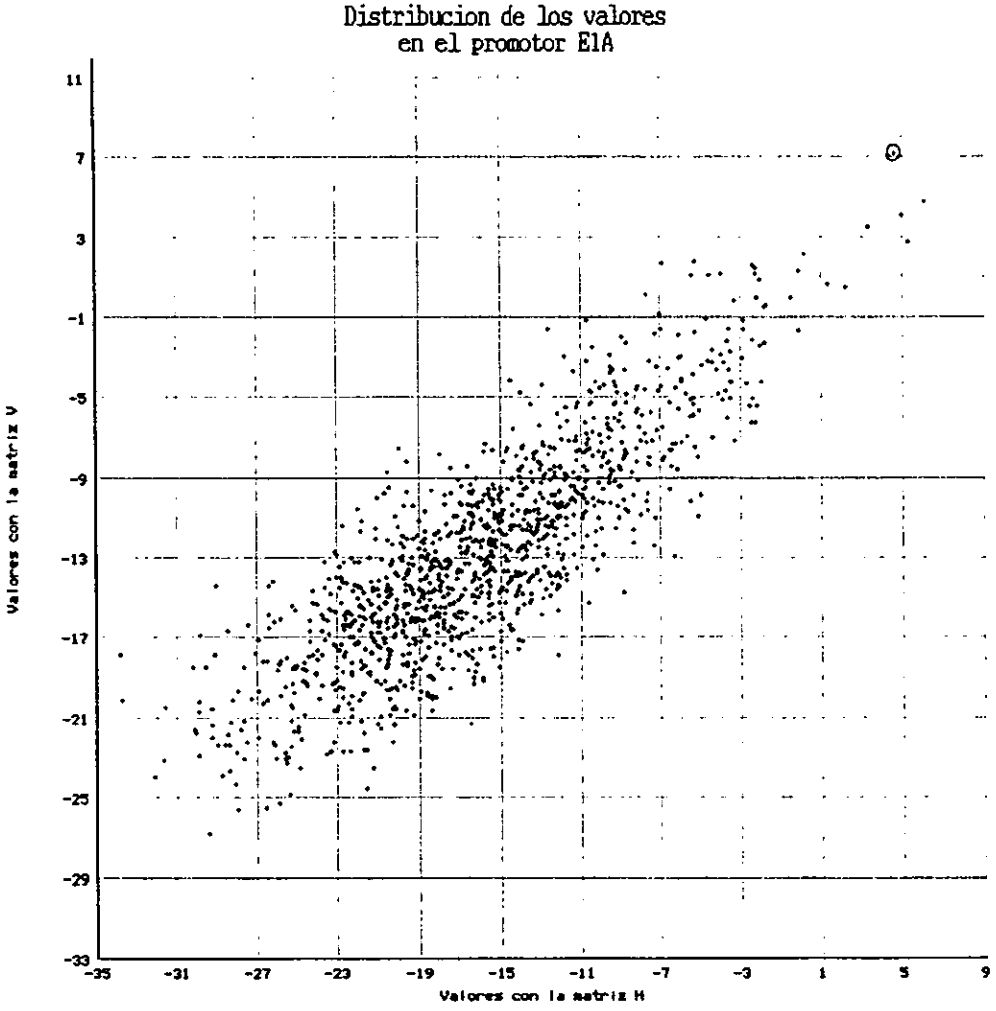

Distribucion de los valores
en el promotor VitII_Gg
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

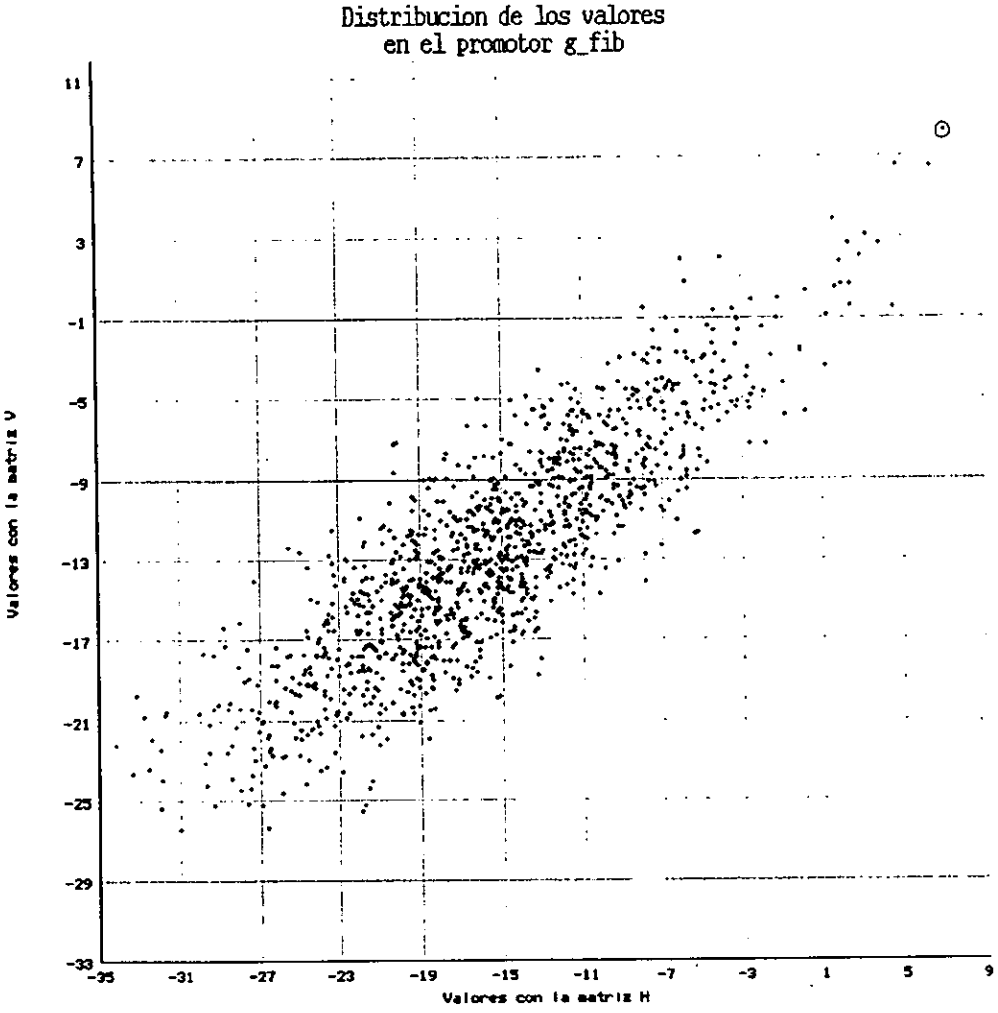
Distribucion de los valores
en el promotor TK_mouse

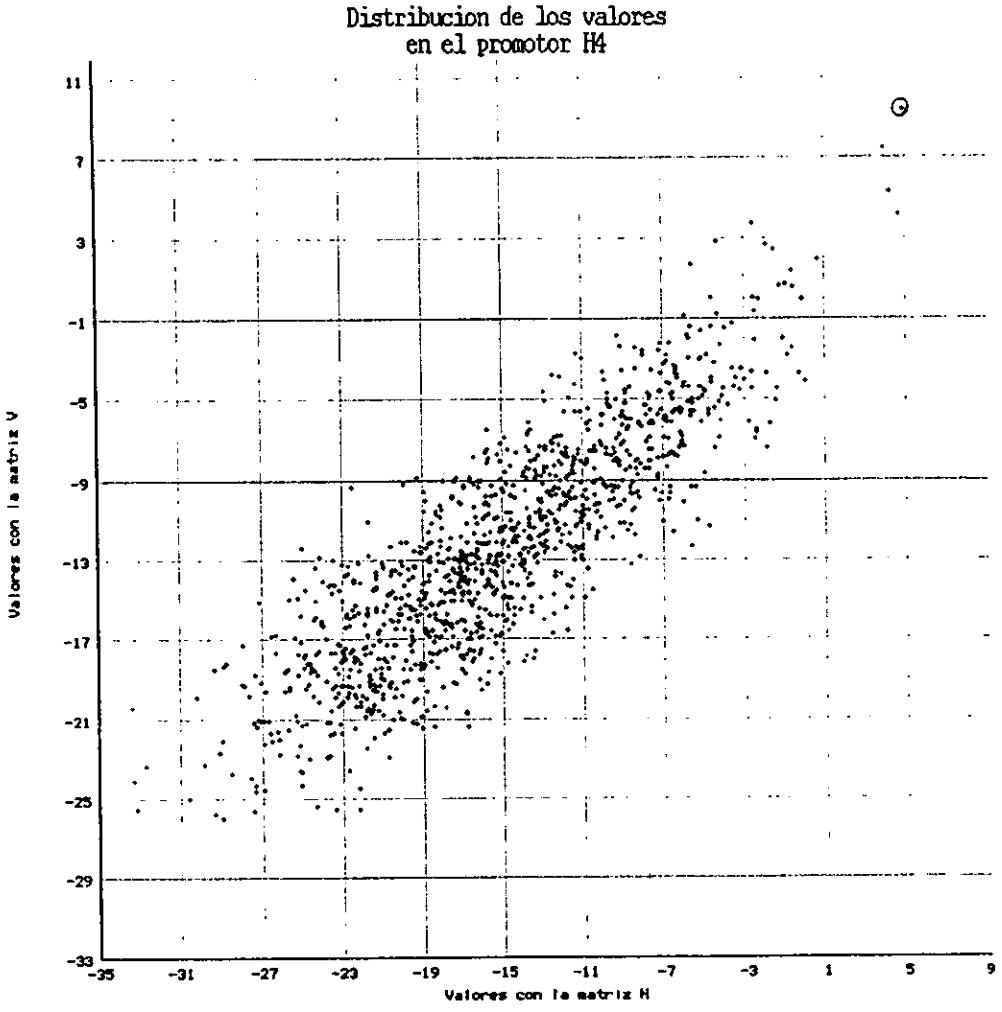


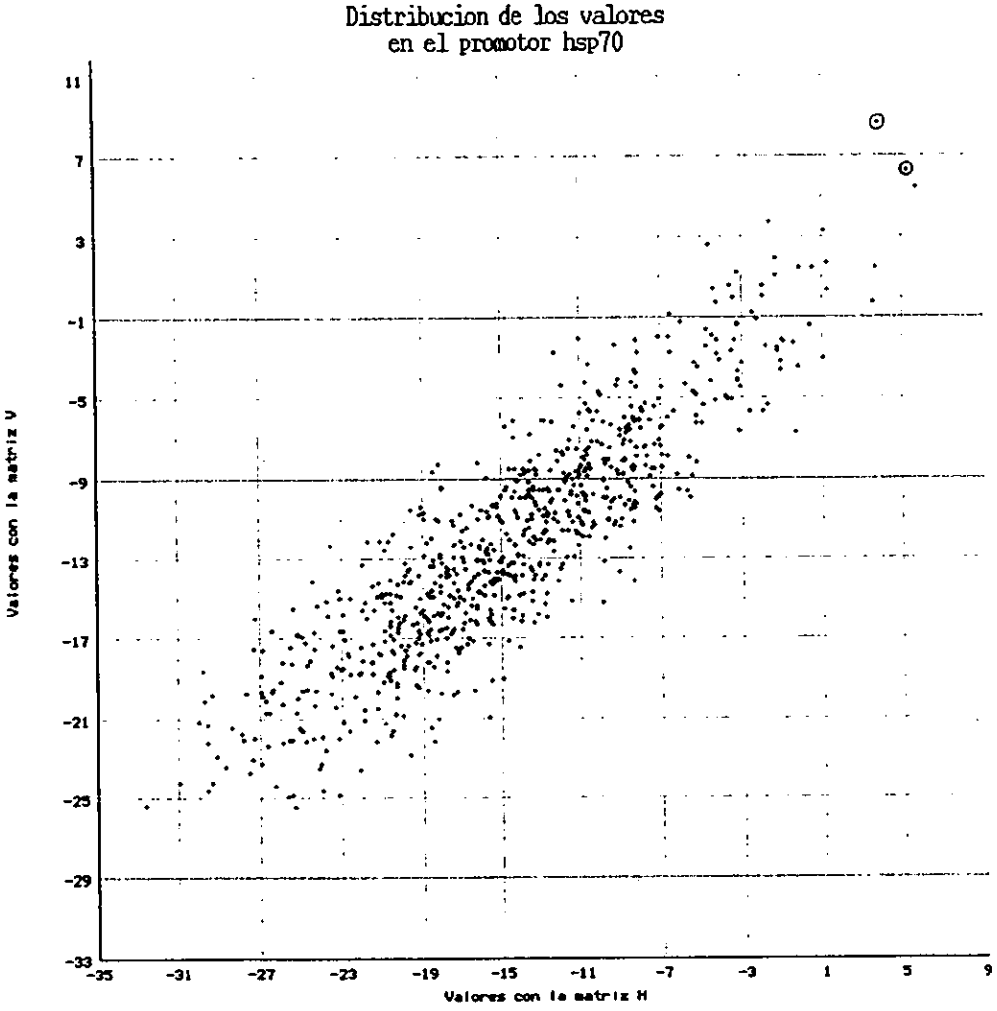
Promotores donde los sitios de Sp1 junto con otros sitios, forman un grupo independiente del resto de los sitios de la secuencia.

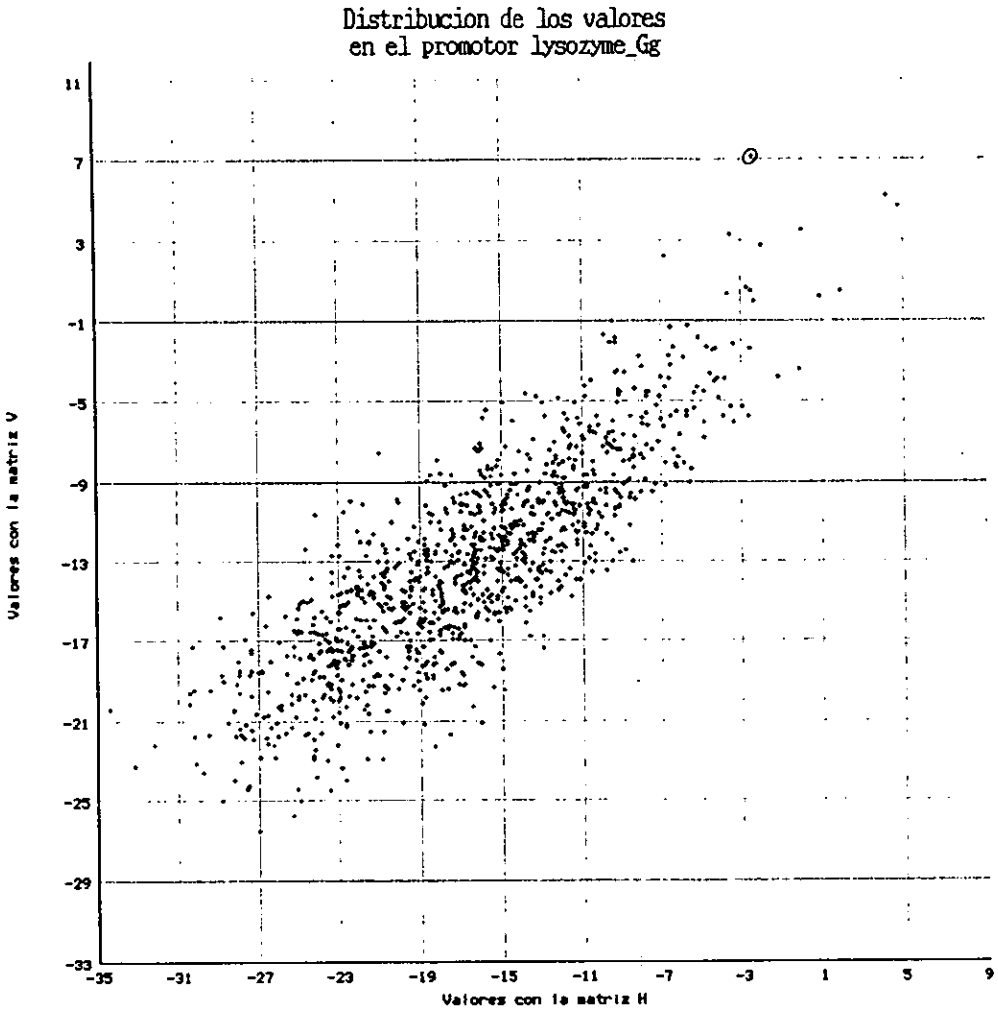

Distribucion de los valores
en el promotor Colla2
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9



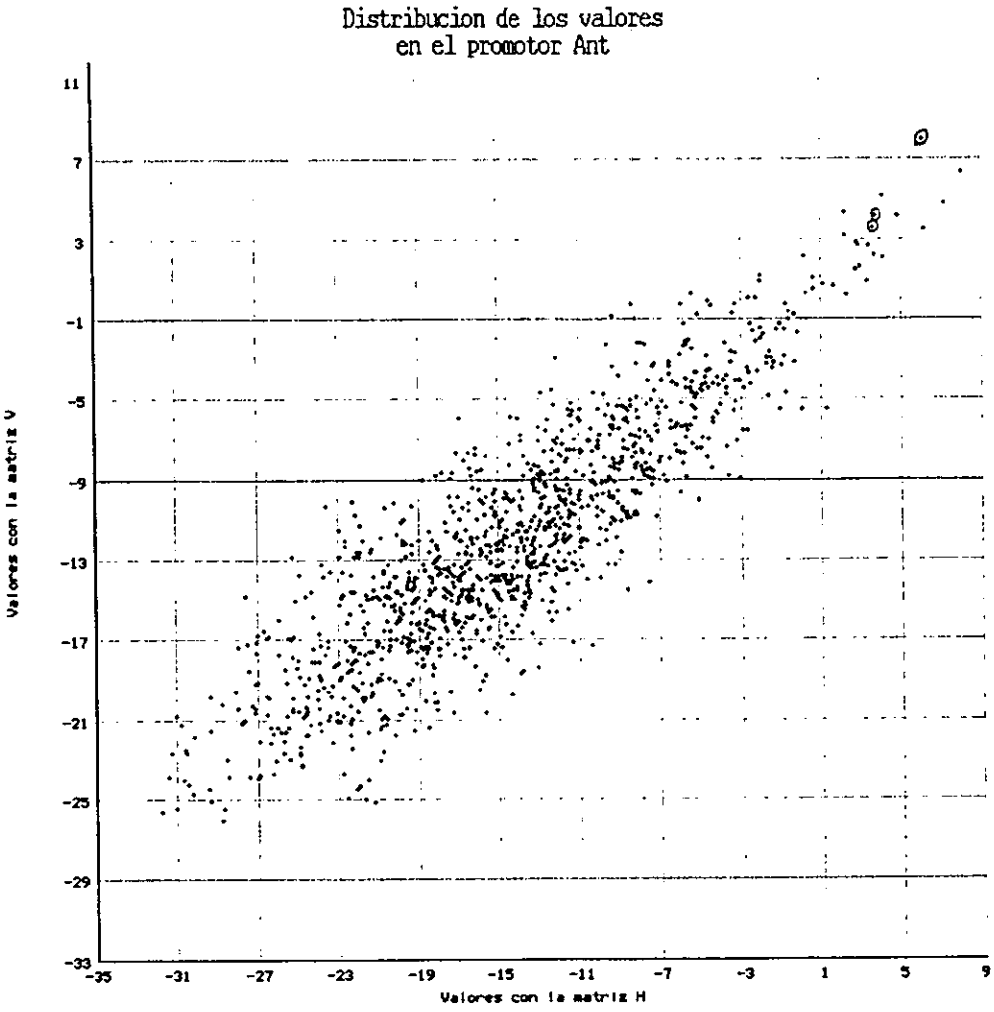

Distribucion de los valores
en el promotor g_fib
Valores con la matriz V
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9
Valores con la matriz H

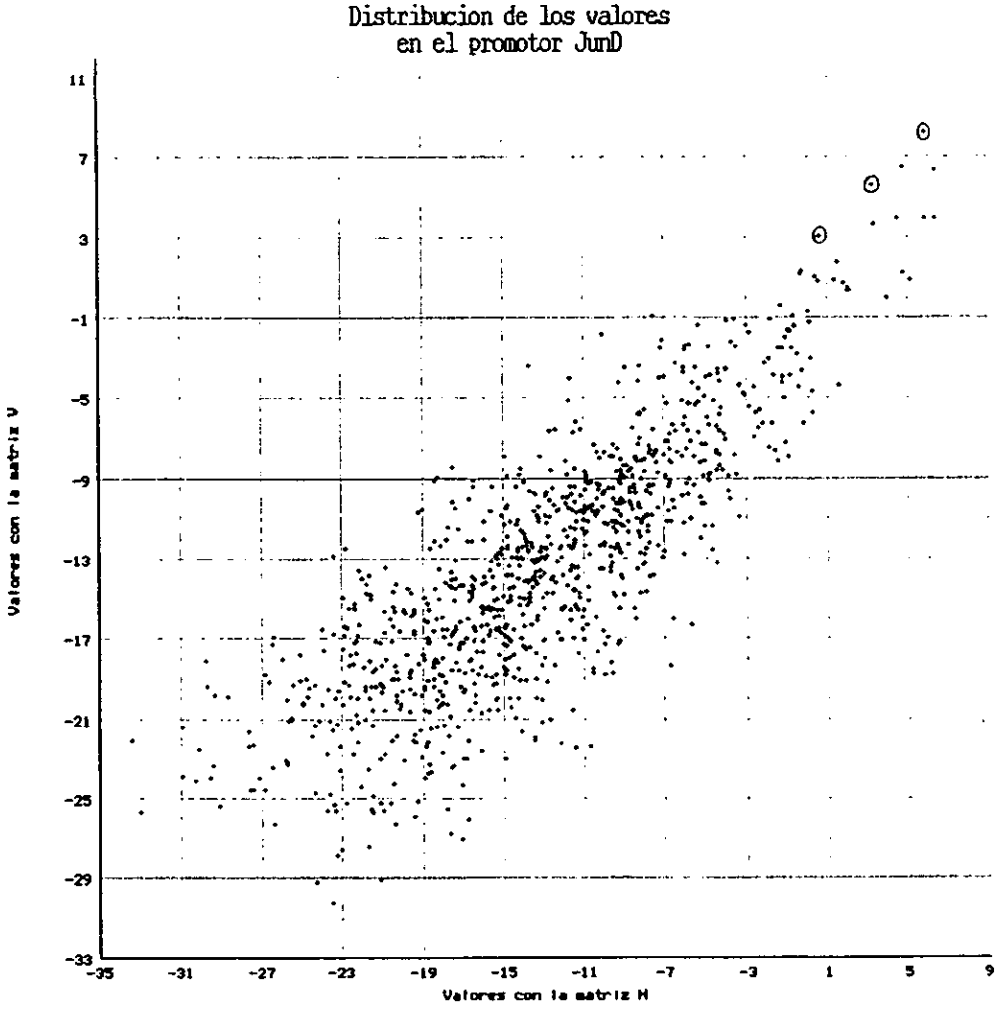

Distribucion de los valores
en el promotor H4
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

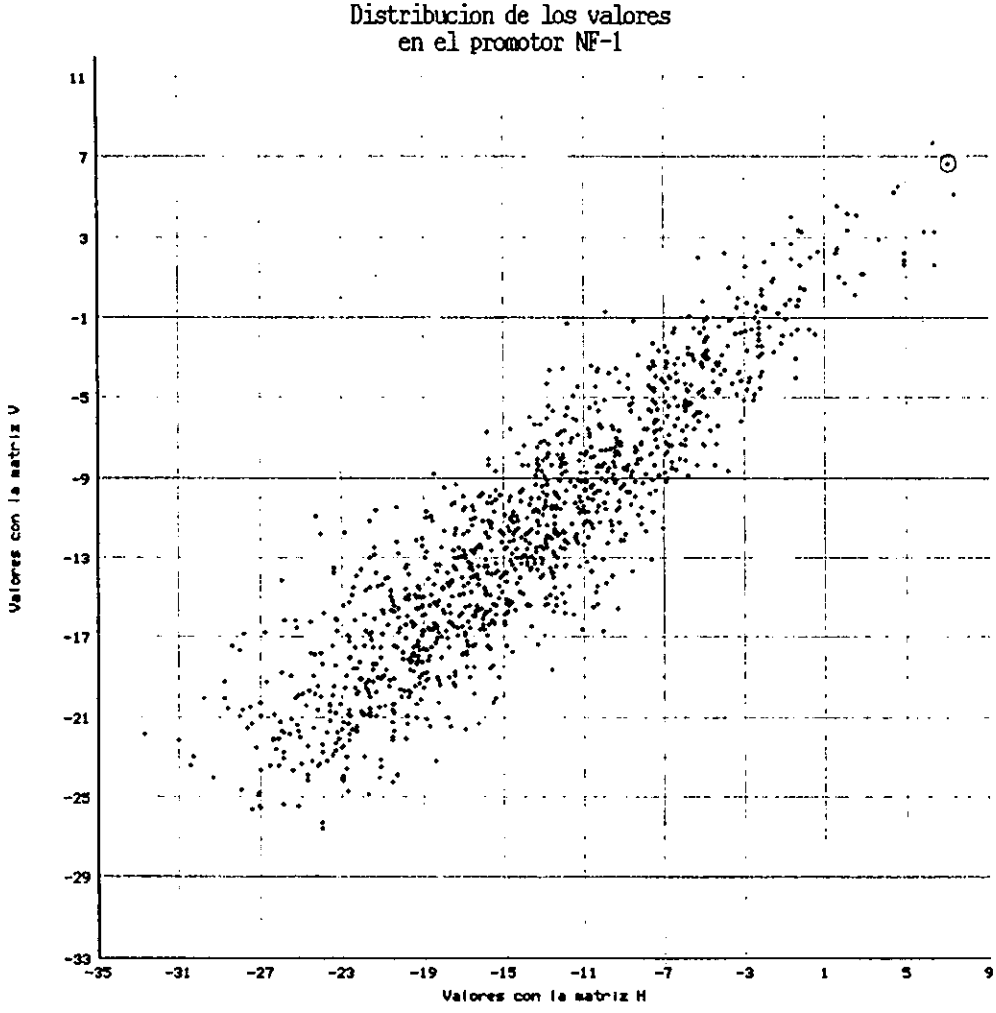

Distribucion de los valores
en el promotor hsp70
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

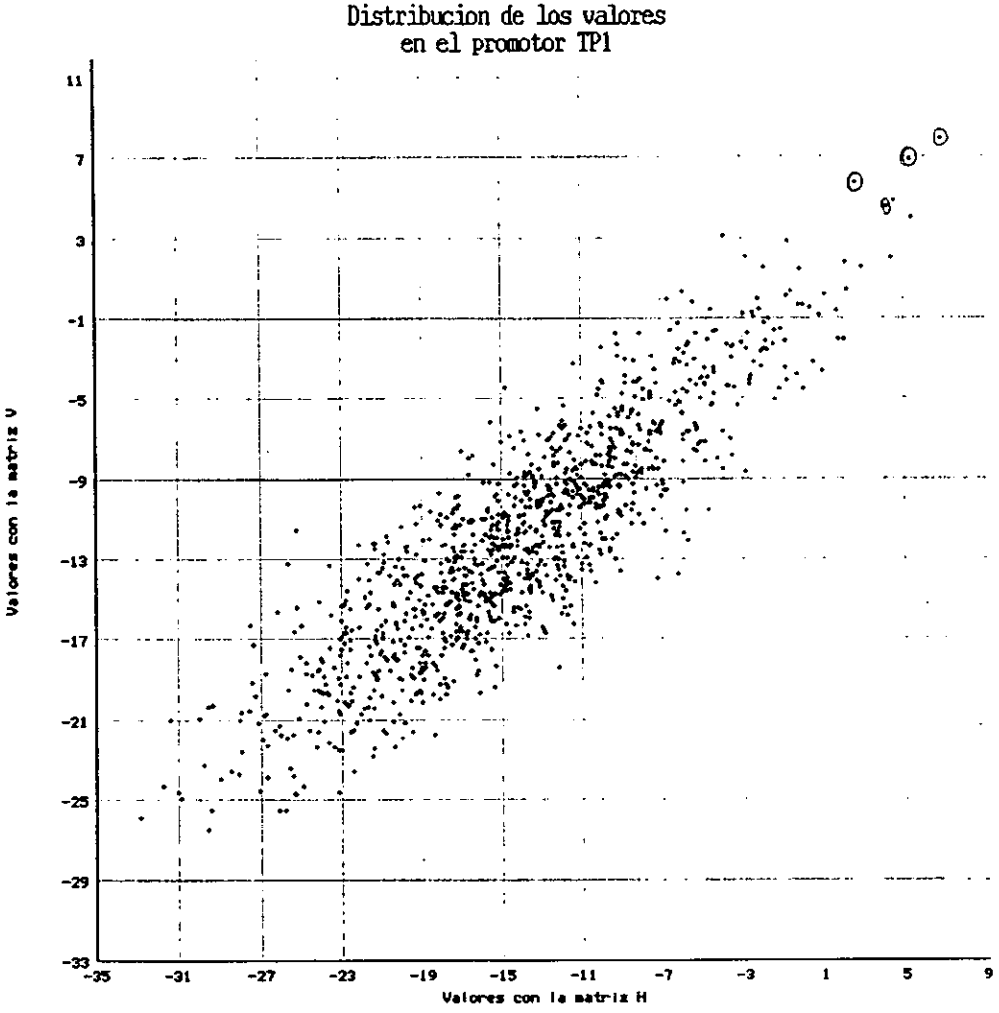

Distribucion de los valores
en el promotor lysozyme_Gg
Valores con la matriz V
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9
Valores con la matriz H

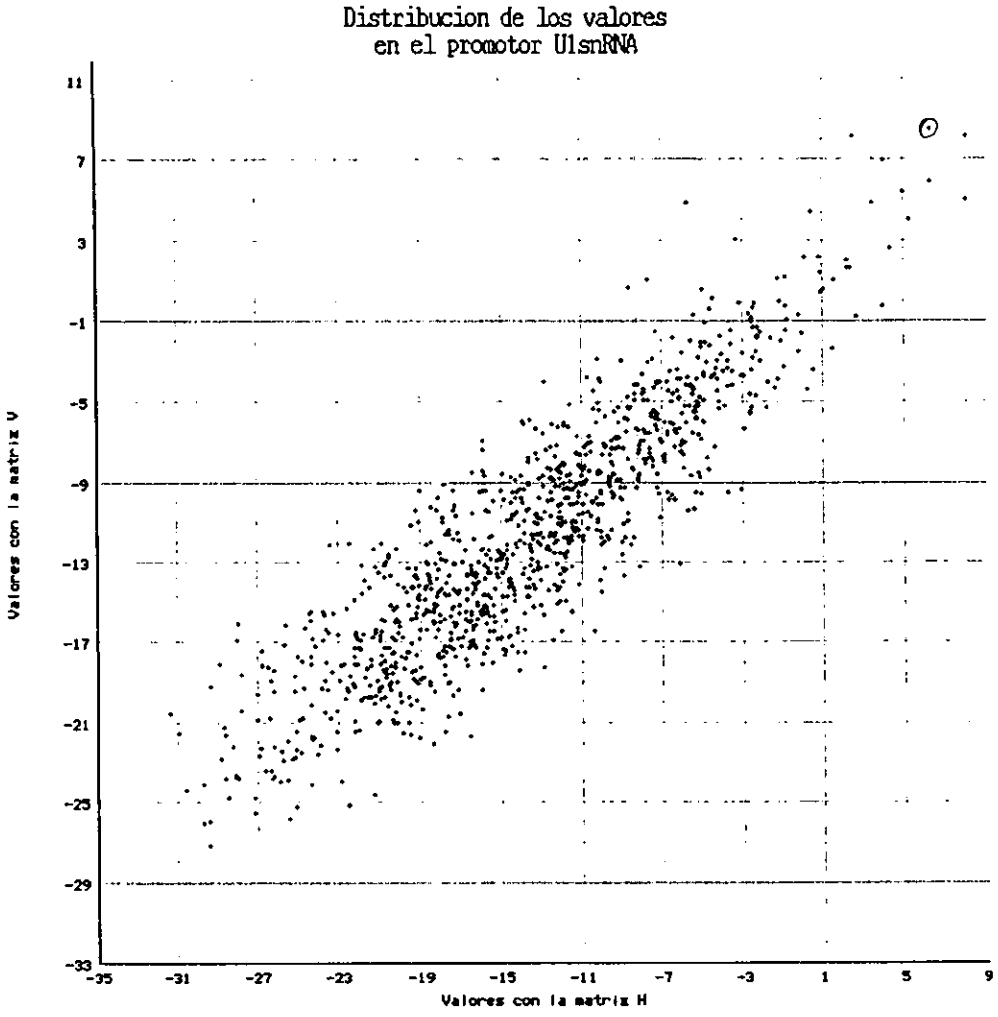
Promotores donde los valores para los sitios de Sp1 se diluyen con el resto de los sitios de la secuencia formando un continuo

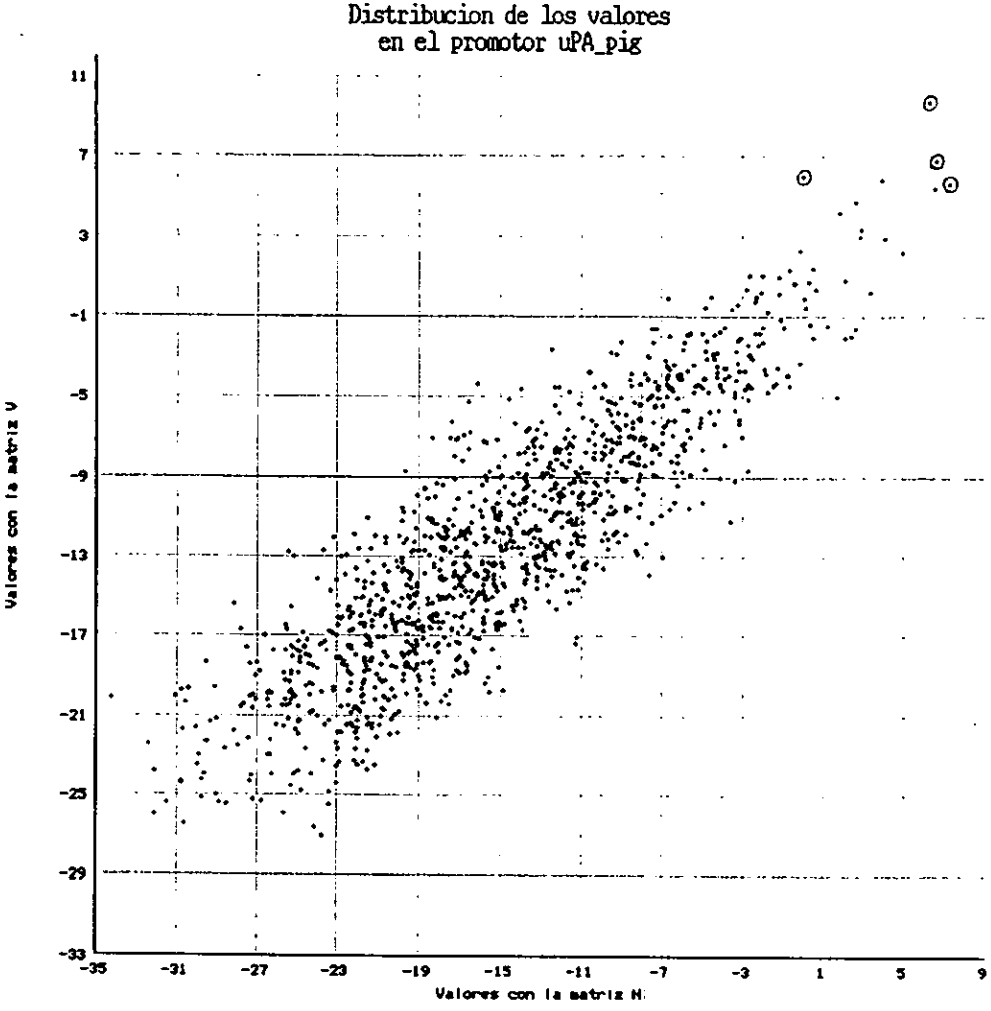

Distribucion de los valores
en el promotor Ant
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9


Distribucion de los valores
en el promotor JunD
Valores con la matriz V
Valores con la matriz M
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

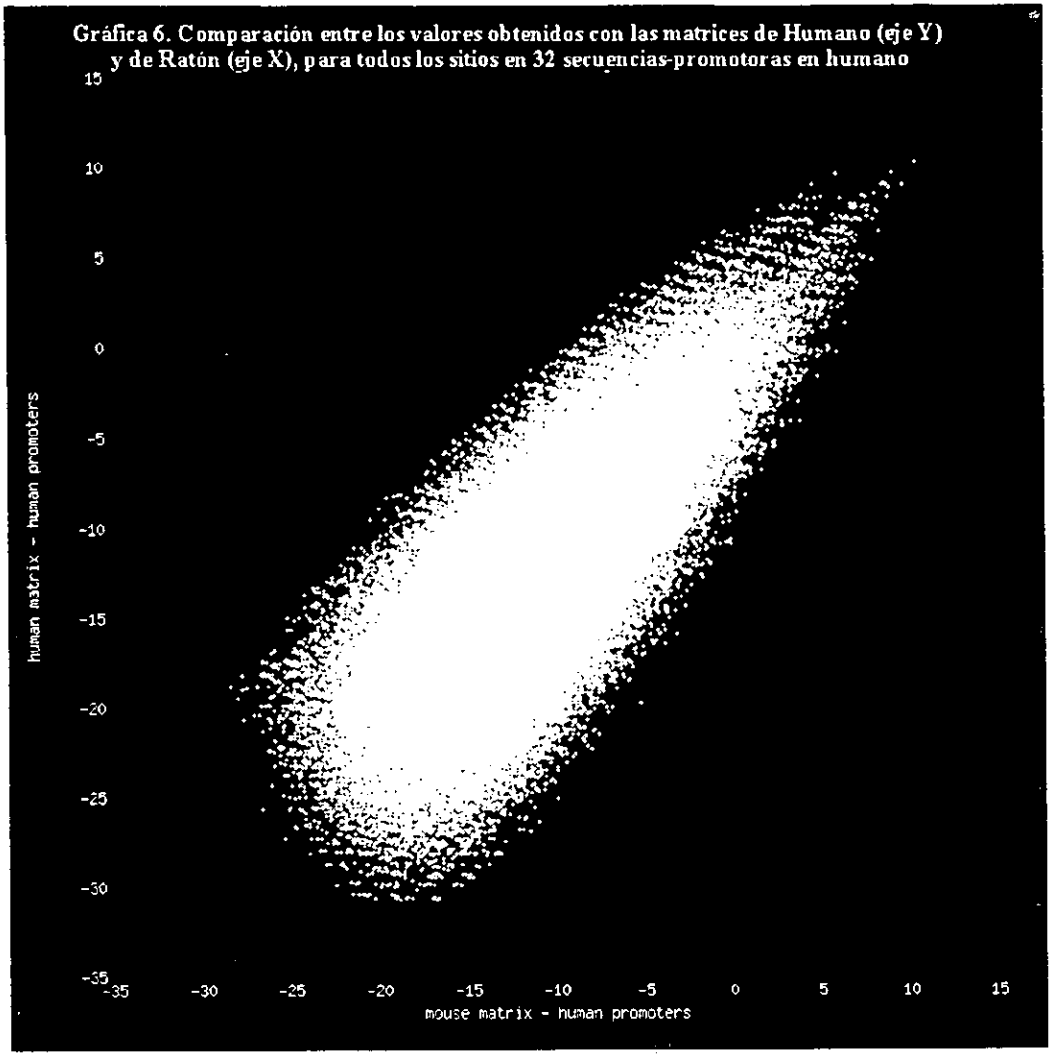

Distribucion de los valores
en el promotor NF-1
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9


Distribucion de los valores
en el promotor TP1
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

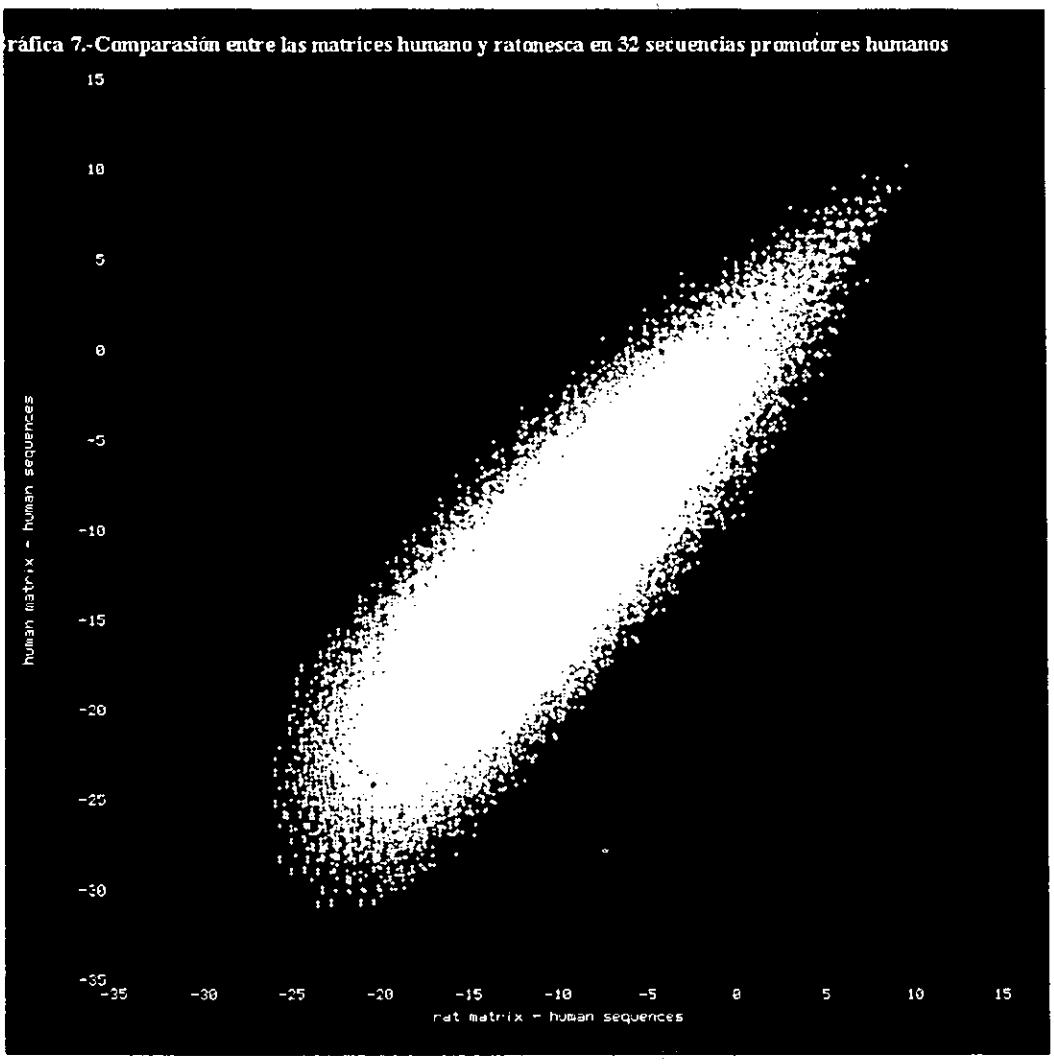

Distribucion de los valores
en el promotor U1snRNA
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9
Valores con la matriz V
Valores con la matriz H


Distribucion de los valores
en el promotor uPA_pig
Valores con la matriz V
Valores con la matriz H
11
7
3
-1
-5
-9
-13
-17
-21
-25
-29
-33
-35
-31
-27
-23
-19
-15
-11
-7
-3
1
5
9

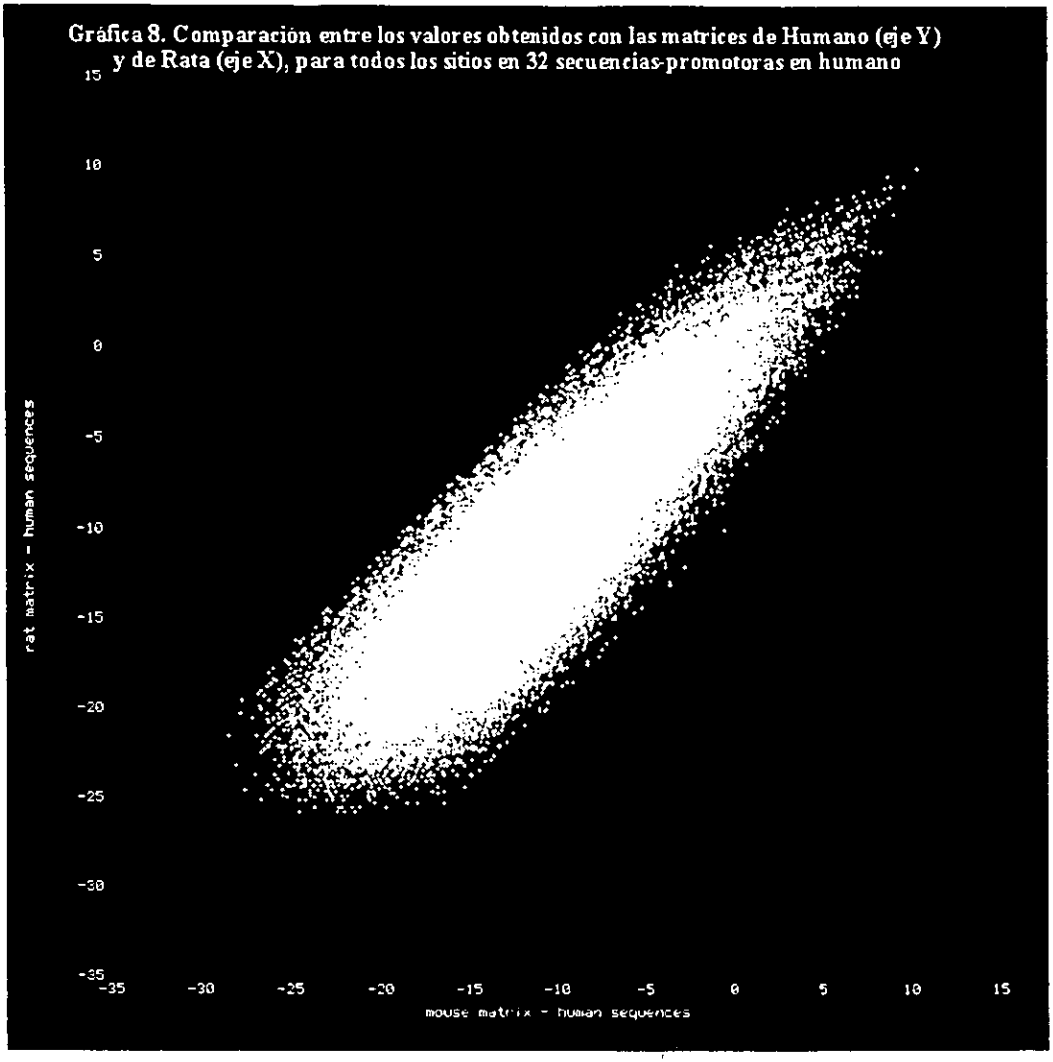
Gráfica 6. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Ratón (eje X), para todos los sitios en 32 secuencias-promotoras en humano



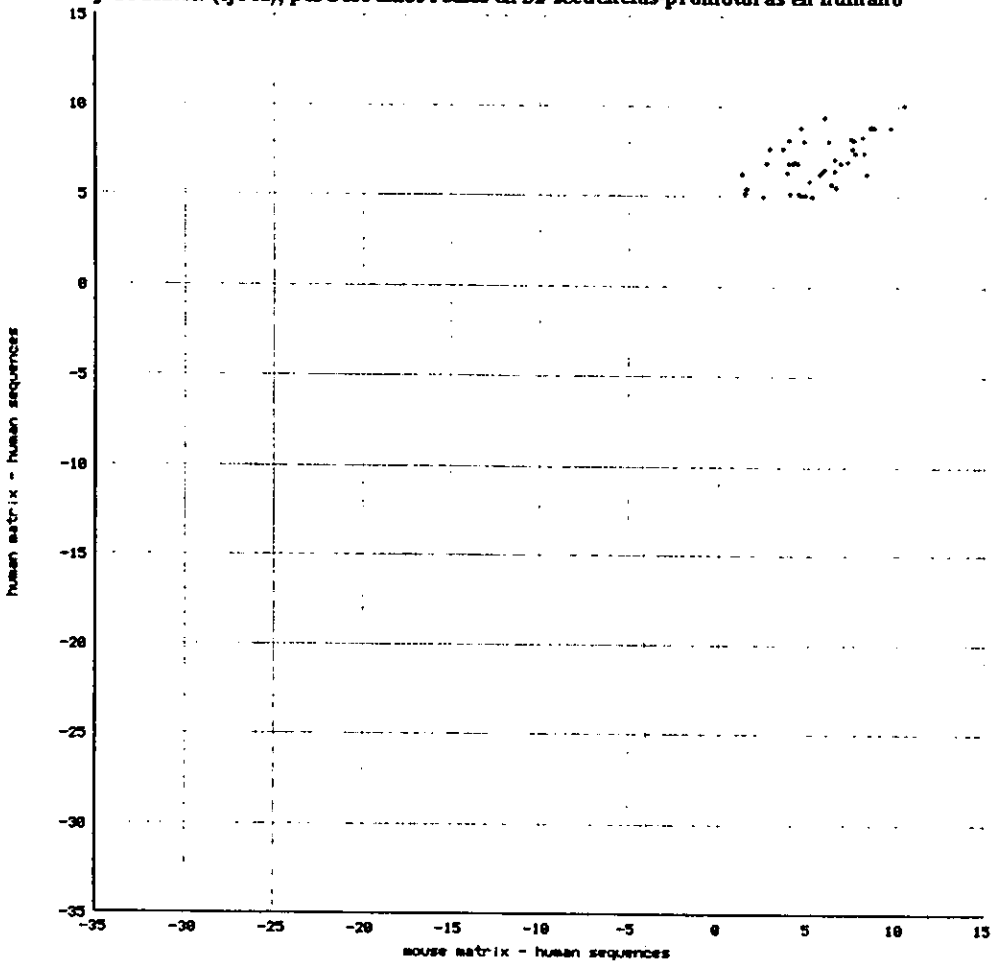
ráfica 7.-Comparasión entre las matrices humano y ratonesca en 32 secuencias promotores humanos



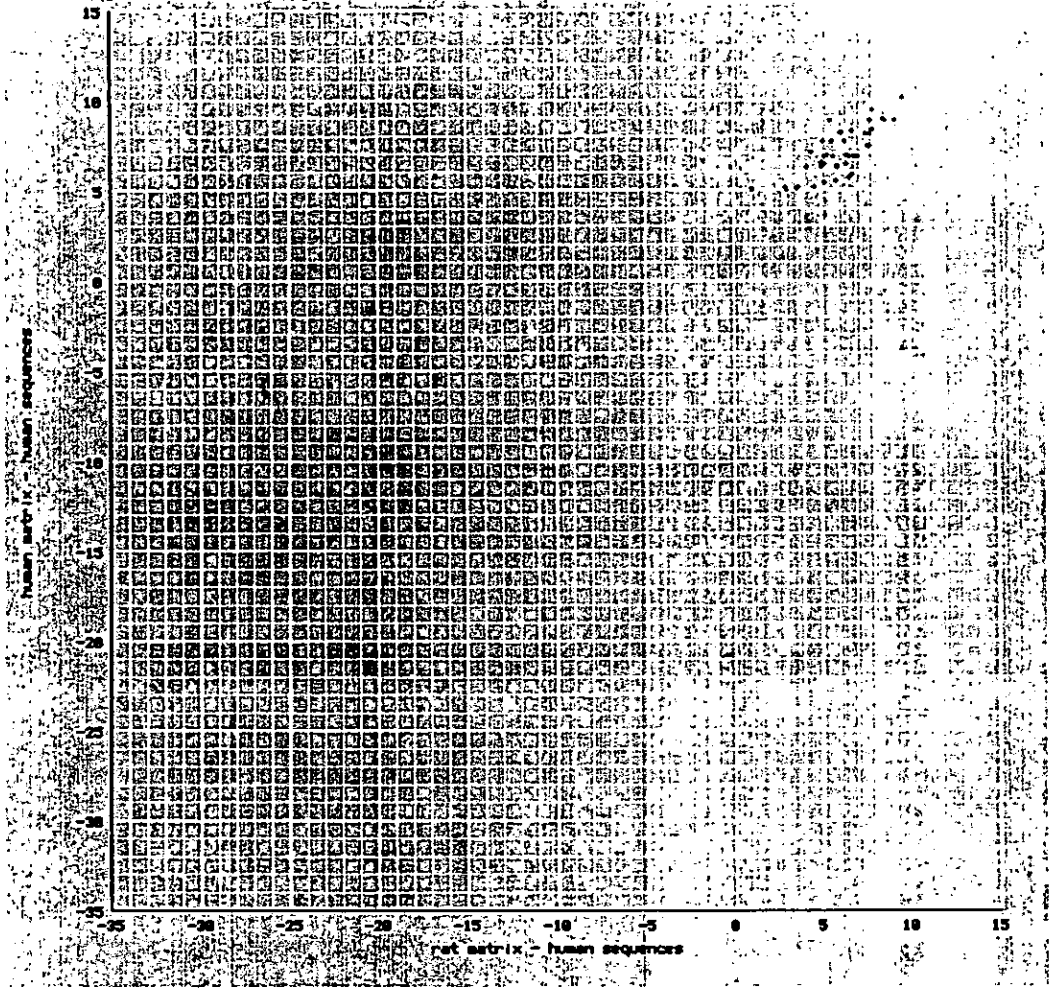
Gráfica 8. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Rata (eje X), para todos los sitios en 32 secuencias-promotoras en humano



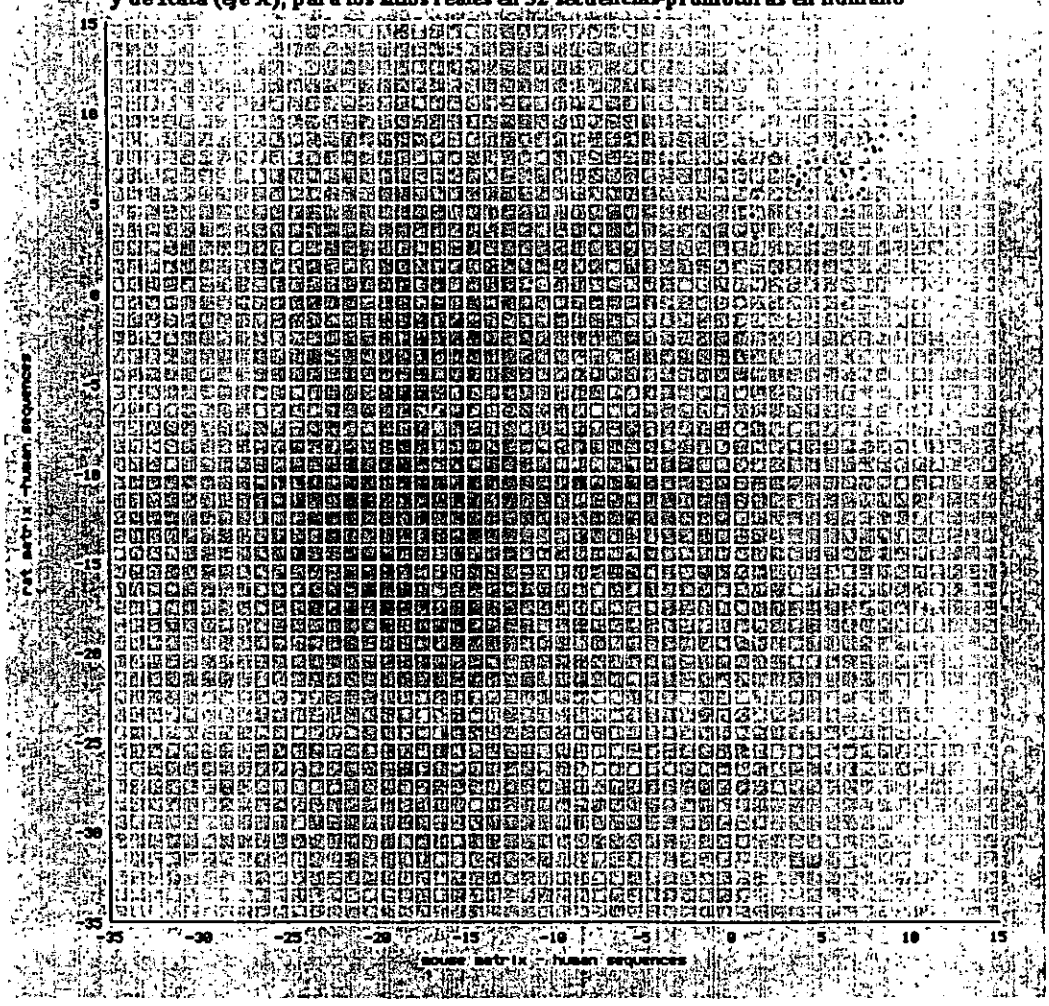
Gráfica 9. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Ratón (eje X), para los sitios reales en 32 secuencias promotoras en humano



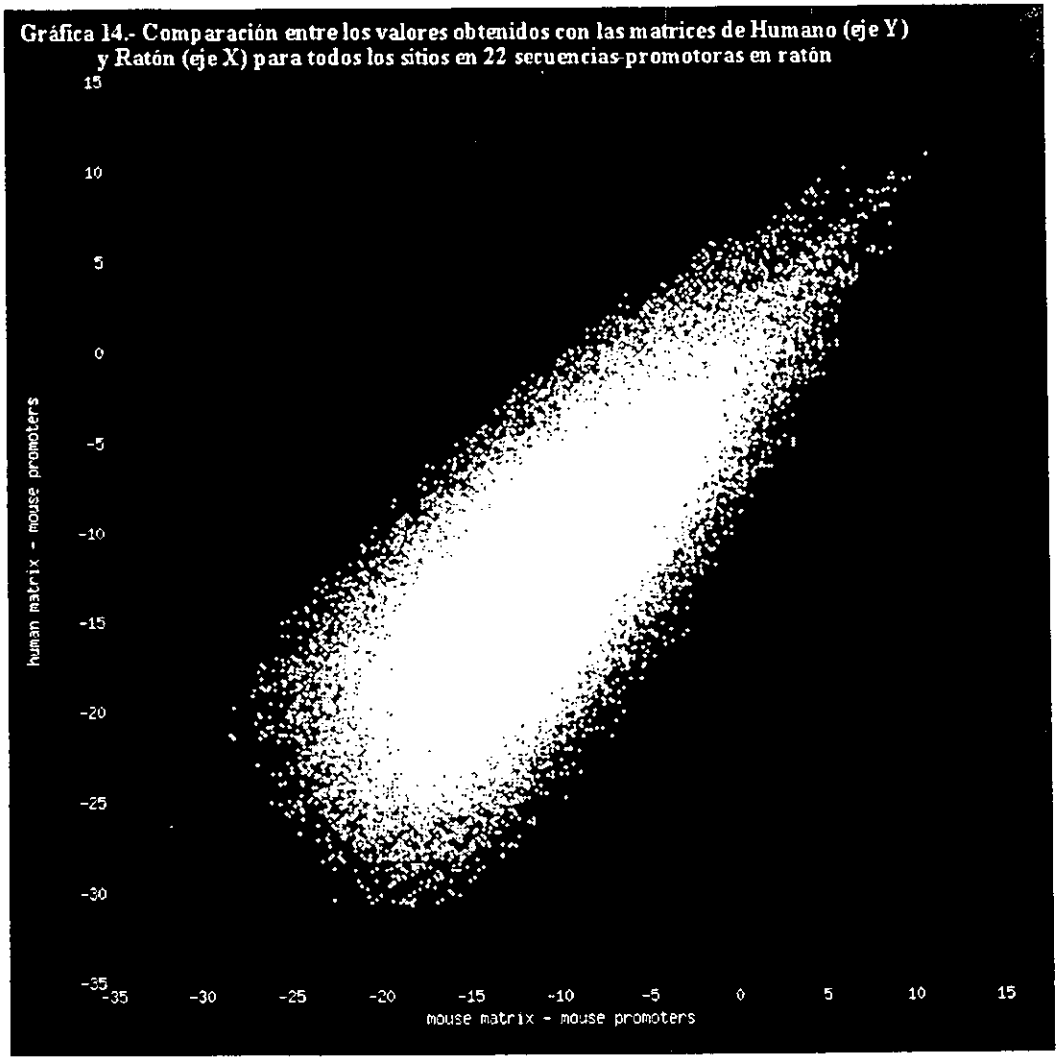
Gráfica 10. Comparación entre los valores obtenidos con las matrices de Rata (eje Y) y de Ratón (eje X), para los sitios reales en 32 secuencias-promotoras en humano



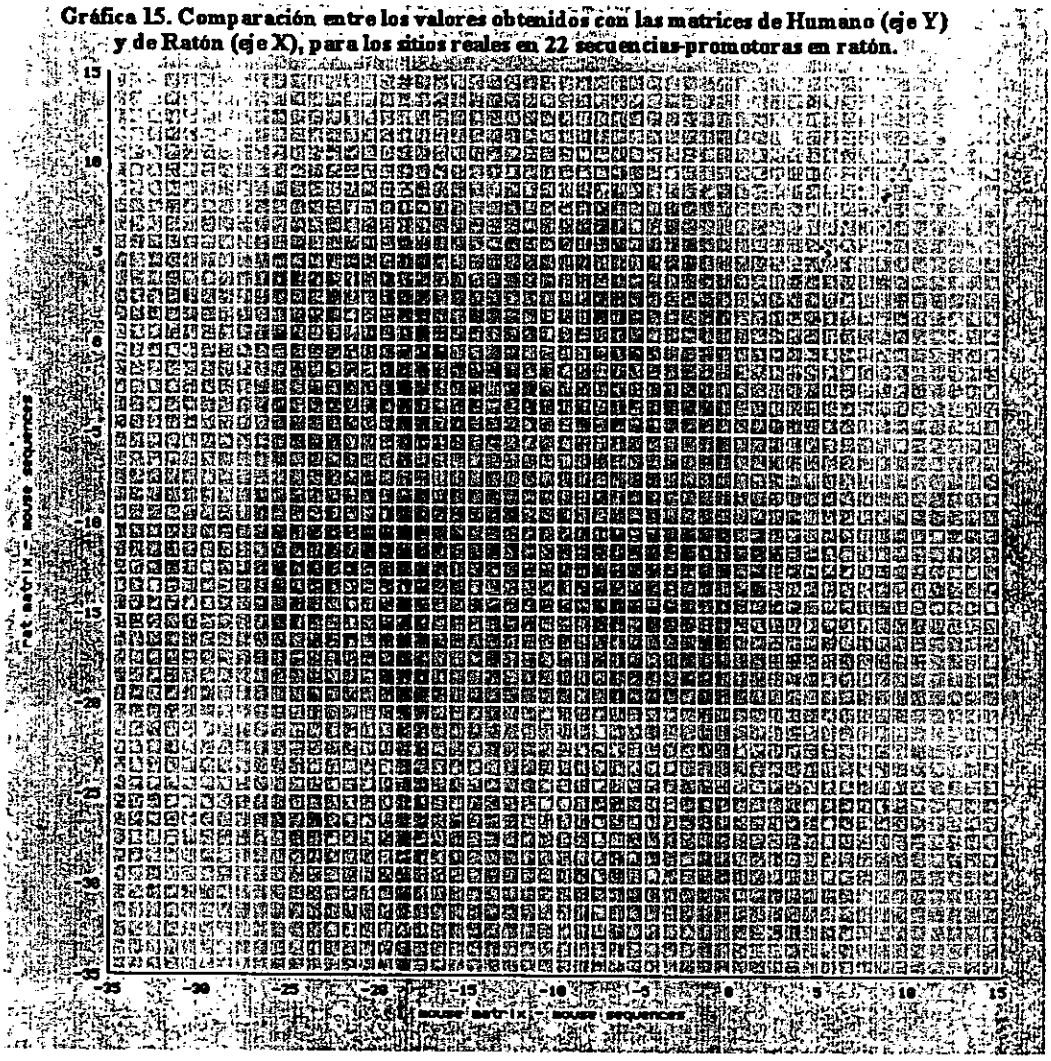
Gráfica 11. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Rata (eje X), para los sitios reales en 32 secuencias-promotoras en humano



Gráfica 14.- Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y Ratón (eje X) para todos los sitios en 22 secuencias promotoras en ratón



Gráfica 15. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Ratón (eje X), para los sitios reales en 22 secuencias-promotoras en ratón.



Gráfica 18. Comparación entre los valores obtenidos con las matrices de Humano (eje Y) y de Ratón (eje X), para todos los sitios en 22 secuencias-promotoras en rata.