LOCALIZACIÓN DE SECUENCIAS REGULADORAS DE LA TRANSCRIPCIÓN POR MÉTODOS COMPUTACIONALES

TRANSCRIPTION REGULATION SEQUENCE DETECTION BY COMPUTATIONAL METHODS

Carlos Andres Perez Galindo

Grupo de Investigación en Biotecnología y Medio Ambiente (GIBMA) – Centro de Investigaciones en Ciencias Básicas, Ambientales y Desarrollo Tecnológico (CICBA), Universidad Santiago de Cali.
pegaso107@gmail.com

RESUMEN

El aumento en la tasa de secuencias biológicas reportadas en las bases de datos, a partir de los procesos de secuenciación y por tanto del crecimiento de las listas de genes de organismos cuyo genoma ha sido secuenciado, contrasta con el poco conocimiento sobre la manera en que esos genes son regulados. En la presente investigación, se elaboro un programa en lenguaje PERL, para la localización de secuencias de ADN que se unen a factores de transcripción que regulan la expresión génica en procariotas. Los conjuntos de genes fueron obtenidos a partir de su expresión (micro arreglos) bajo las mismas condiciones ambientales. El organismo modelo con el que se trabajo fue *lactococcus lactis,* del cual se dispone su genoma secuenciado en formato del banco de genes. El programa encontró mayor número de posibles secuencias reguladoras en la región flanqueadora 5´ de los genes. El número de posibles secuencias reguladoras también estuvo determinado por la cantidad de genes que conformaron cada conjunto. El programa también localizo secuencias flanqueadoras de genes que podrían estar involucradas en su regulación, pero a nivel traduccional.
La comparación de los resultados con patrones obtenidos experimentalmente, se hizo mediante matrices de pesos de posición de nucleótidos, obteniéndose aproximadamente un 50 % de secuencias reguladoras que coincidían con las reportadas en las bases de datos, lo que indica un buen nivel de predicción del programa si se tiene en cuenta que la mayoría de secuencias reguladoras para procariotas, aun no han sido caracterizadas por métodos experimentales.

Palabras clave: Bioinformática, PERL, transcripción, traducción, matrices de pesos, factores de transcripción.

ABSTRACT

The increase in the number of biological sequences reported to the data bases, and the growth in the accompaning gene lists from the organisms whose genome has been sequenced, contrasts with the little existing knowledge of how these genes are regulated. In this study, a PERL computer program was created to detect the DNA sequences that join the transcription factors which regulate the genetic expression in prokaryote organisms. The gene sets were obtained from their expressions (microarrays) under the same environmental conditions. The model organism used was *lactococcus lactis,* whose sequenced genome is available in gene bank format. The program found a greater number of possible regulating sequences in the 5´ gene flanking region. The number of possible regulatory sequences was also determined by the number of genes which make up each set. The program also detected gene flanking sequences which might be involved in its regulation, but at the translational level.

The comparisson of the results with experimentally obtained standards was done with position weight nucleotid arrays getting approximately 50% regulating sequence coincidence with reported data which indicates a good prediction level from the program if one takes into account that the mayority of prokaryote regulating sequences still have not been characterized by experimental methods.

Keywords: Bioinformatics, PERL, transcription, translation, weight array, transcription factors

I. INTRODUCCION.

Hoy en día, se observa un aumento en la tasa de secuencias biológicas reportadas en las bases de datos, a partir de los procesos de secuenciación y por tanto del crecimiento de las listas de genes de organismos cuyo genoma ha sido secuenciado. Sin embargo, este hecho contrasta con el poco conocimiento sobre la manera en que esos genes son regulados. Por ejemplo, en *Escherichia coli*, la bacteria más estudiada, aproximadamente 1 / 5 de las 300 a 350 proteínas reguladoras estimadas, tienen caracterizados sus sitios de unión al ADN. Para las bacterias cuyo genoma ha sido secuenciado recientemente, así exclusivamente, los sitios de unión a factores de transcripción que se alineen por homología con las secuencias identificadas en *E.coli* y *Bacillus subtilis,* pueden ser usadas para inferir propiedades regulatorias del organismo. Por tanto, es importante el desarrollo de herramientas computacionales para identificar secuencias de unión de factores de

transcripción aún no caracterizados. La gran velocidad a la que se están

la obtención de los mejores alineamientos locales el programa se apoya en el software **lalign.exe**, el cual es ejecutado comparando cada una de las secuencias entre si de cada carpeta. Los resultados son guardados en el archivo **ResultadosLalign3′.txt** y **ResultadosLalign5′.txt**. Una vez se tienen estos archivos, el programa selecciona aquellos alineamientos con una longitud y porcentaje de similitud igual o mayor al proporcionado por el usuario. Los resultados de este primer filtro son guardados en los archivos **ResultadosComparacion3′.txt** y **ResultadosComparacion5′.txt**, para cada orientación de las secuencias flanqueadoras. En la presente investigación se trabajó con un valor de identidad igual o mayor al 75 % y una longitud mínima del alineamiento de 7, debido, a que en los genomas de procariotas, los sitios de unión a factores de transcripción tienen una longitud variable de aproximadamente 30 nucleótidos, sin embargo, hay dos regiones altamente conservadas de estos sitios, de aproximadamente 7 nucleótidos, que predominantemente hacen contacto con los factores de transcripción y que por cuestiones de evolución neutral pueden variar en uno o dos nucleótidos.

Los primeros ocho conjuntos de genes, corresponden a aquellos que tuvieron un nivel similar de expresión en experimentos de micro arreglos. El conjunto 8 esta conformado por genes seleccionados al azar, con el fin de utilizarlos como control negativo.

II.I. PROGRAMA DESARROLLADO.

El programa puede obtenerse en la dirección electrónica: http://www.usc.edu.co/investiga/cicba/alineamiento.txt

II.II. CONJUNTO DE GENES DE *Lactococcus lactis* UTILIZADOS EN LA COMPROBACIÓN DEL PROGRAMA.

Los conjuntos de genes proceden de un experimento de arreglos de ADN, en que el control es la cepa utilizada en la secuenciación de su genoma y la diana es una cepa natural, utilizada en alimentación, específicamente en la producción de yogur.
El conjunto número 8, esta conformado por genes tomados al azar, con el fin de tener un control negativo.

II.III. MÉTODO PARA DETERMINAR EL VALOR DE CADA NUCLEÓTIDO EN LAS MATRICES DE PESOS.

Este método es derivado de la teoría de la información[5], el cual consiste en calcular el vector *RSequence(1)*, mediante la formula:

F(b, l)[6] es la frecuencia de cada base b

en la posición l de los sitios alineados. La matriz de pesos m(b, l) se calcula mediante la formula:

Donde f(b,l) es igual a:

Para calcular la puntuación de cada secuencia, se suma cada uno de los pesos de los nucleótidos por posición.

III. RESULTADOS.

Para la obtención de las posibles secuencias reguladoras, se partió de alineamientos locales entre regiones flanqueadoras 5´ de los genes que conforman un mismo conjunto de datos.

A partir de las alineaciones, se realizaron comparaciones entre todas las secuencias con el fin de obtener patrones comunes. Para intentar diferenciar los resultados de las secuencias flanqueadoras 5' y 3' se ha cal[5] ones obtenidos por conjunto de genes y su longitud promedio (tabla 1).

Tabla 1. Número y tamaño de patrones encontrados por conjunto de genes. El conjunto de genes de texto azul, corresponde a los patrones encontrados en las secuencias flanqueadoras 5´ del gen (100 nucleótidos aguas arriba); El conjunto de genes de texto rojo, corresponde a los patrones encontrados en las secuencias flanqueadoras de la región 3´ del gen (100 nucleótidos aguas abajo); * conjunto de genes control.

Excepto para el conjunto de genes 7 y 8, los resultados indican que hay diferencias entre los patrones de las secuencias flanqueadoras 5´ y 3´, no sólo a nivel de similitud con los reportados en las bases de datos, sino también en el número obtenido, siendo mayor el de las secuencias flanqueadoras 5´ (figura 2). El número de genes del conjunto 7 es muy reducido (3 genes) y el conjunto 8 estuvo conformado por 34 genes, todos seleccionados al azar, por tanto, los patrones obtenidos de las secuencias flanqueadoras 5´ y 3´ de este conjunto son controles, siendo su número muy similar.

Hasta el momento se carece de una base de datos de factores de transcripción para Lactococcus lactis y las reportadas no tienen la totalidad de secuencias involucradas en procesos regulatorios de la transcripción, por tanto es muy difícil que el número de patrones obtenidos coincida en su totalidad con los de las bases de datos. Sin embargo, para los diferentes conjuntos de genes, excepto el 7, obtenidos de las secuencias

[5] Schneider, T. D., Stormo, G. D. & Gold, L. (1986). Information content of binding sites on nucleotide sequences. J. Mol. Biol. 188, 415-431.

[6] Matrices de pesos: http://prodoric.tu-bs.de/vfp/vfp_help.php#pwm

Conjunto de Genes	Número de Genes	Número de Patrones	Tamaño Promedio del Patrón en Número de Nucleótidos
0	20	14	7.36
1	12	9	8.11
2	11	7	7.33
3	18	14	7.42
4	10	7	7.83
5	44	18	10.78
6	29	18	7.50
7	3	0	0
8 *	34	7	7.42
0 *	20	6	7.32
1 *	12	4	7.95
2 *	11	6	7.33
3 *	18	6	7.33
4 *	10	4	7.25
5 *	44	7	14.7
6 *	29	9	7.22
7 *	3	0	0
8 *	34	8	7.34

flanqueadoras 5´, se obtuvieron secuencias similares (tabla 2).

Tabla 2. Número de patrones encontrados en las secuencias flanqueadoras de la región 5´ que son similares a los reportados en las bases de datos (verdaderos positivos).* conjunto de genes control

Para la mayoría de conjuntos, aproximadamente el 50% del número de patrones fue similar al reportado en las bases de datos (figura 3). Los patrones del conjunto 8 podrían ser considerados como falsos positivos, debido a que este conjunto se elaboro con genes seleccionados al azar y no, por expresarse bajo las mismas

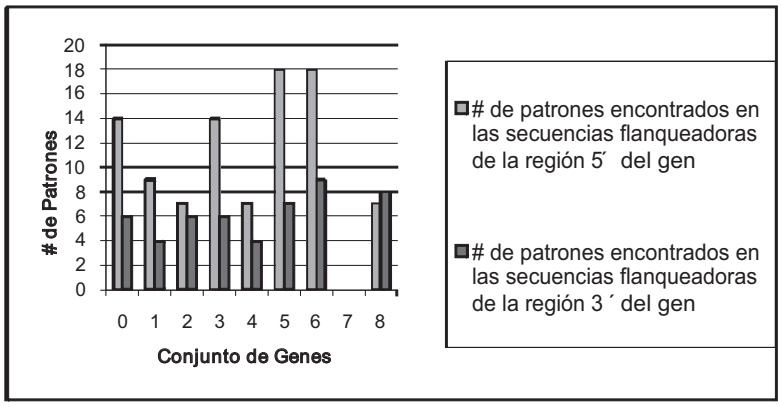


Figura 2. Histograma para la comparación del número de patrones obtenidos de las secuencias flanqueadoras 5´ y 3´.

condiciones ambientales. Sin embargo, hay que tener en cuenta el número de secuencias flanqueadoras en las que se encuentran y las puntuaciones que obtuvieron respecto a las secuencias de las bases de datos, lo que podría indicar que algunas de estas secuencias pondrían ser verdaderos positivos obtenidas por comparación aleatoria de secuencias flanqueadoras de genes.

Al realizarse una comparación entre las secuencias de los patrones obtenidos a partir de las regiones flanqueadoras 5´ con las 3´, de todos los conjuntos de genes, se encontró que muy pocas coincidían (tabla 3), al igual que comparar estos resultados con los patrones reportados en las bases de datos, indicando que el

Conjunto de Genes	*Número de Genes*	*Número de Patrones*	Número de patrones similares con los reportados en las bases de datos
0	20	14	7
1	12	9	4
2	11	7	1
3	18	14	6
4	10	7	6
5	44	18	11
6	29	18	8
7	3	0	0
8 *	34	7	2

posible número de falsos positivos es reducido, debido a que las regiones reguladoras de la transcripción se localizan aguas arriba de los genes en procariotas, muy diferentes a lo que ocurre en eucariotas, cuyas regiones de regulación génica pueden encontrarse en sitios aguas debajo de los genes o regiones intrónicas[7]. Es por esto, que los programas de predicción de regiones reguladoras de la transcripción en procariotas, utilizan las regiones flanqueadoras 5´ para su evaluación. En la presente investigación, se han utilizado las regiones flanqueadoras 3´, como controles.

Tabla 3. Patrones que coinciden tanto en las regiones flanqueadoras 5´ y 3´ de un mismo conjunto de genes (posibles

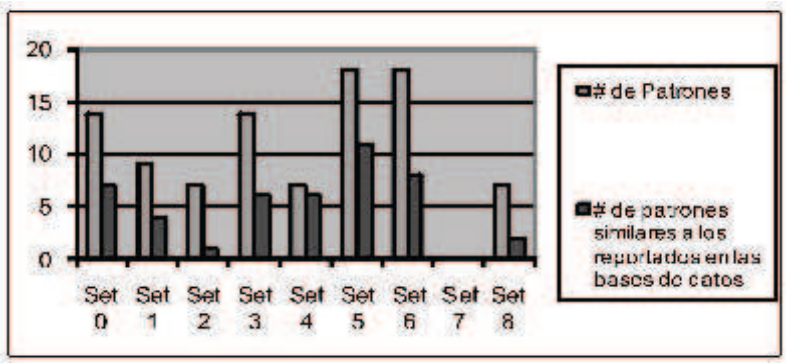


Figura 3. Histograma para la comparación del número de patrones obtenidos de las secuencias flanqueadoras 5´ y las reportadas en las bases de datos de sitios de unión a factores de transcripción.

falsos positivos). * Secuencias reportadas en la base de datos como sitio de unión a factores de transcripción.

Los conjuntos con los que se trabajo, estaban conformados por un número distinto de genes. La distribución de los datos muestra una tendencia lineal, indicando que a mayor número de genes mayor número de patrones obtenidos por el programa.

La correlación de los datos permite obtener la relación entre el número de patrones y el número de genes. Para el número de patrones obtenidos de las secuencias flanqueadoras 5´, la correlación es muy buena. El coeficiente de correlación es igual a 0.8 (figura 4).

Para el número de patrones obtenidos de las secuencias flanqueadoras 3´, la pendiente es 0.129 y el coeficiente de correlación es de 0.60 (figura 5).

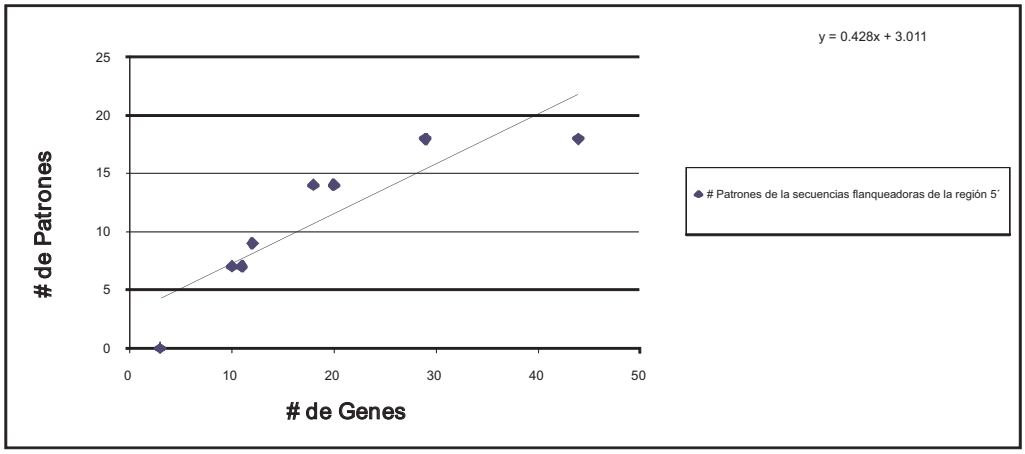
Las figuras 4 y 5, muestran que la pendiente de la gráfica es mayor para el número de patrones de secuencias flanqueadoras de la región 5´ de cada conjunto de genes Vs. Número de genes, respecto a la curva deducida de los controles, indicando que la tendencia del programa es obtener mayor número de patrones de las secuencias que flanquean aguas arriba

Conjunto	Patrones
0	TAAAAAT * GTAAAA
1	Ninguno
2	Ninguno
3	AGAAAAA
4	Ninguno
5	Ninguno
6	Ninguno
7	Ninguno

[7] Cliften P, Hillier L, Fulton L, Graves T, Miner T, Gish W, WaterstonR, Johnston M: Surveying Saccharomyces genomes to identify functional elements by comparative DNA sequence analysis. Genome Res 2001, 11:1175-1186.

a los genes con un perfil de expresión similar. Para comprobar la precisión del programa desarrollado, se buscaron las anotaciones funcionales de los genes con patrones similares, su posición en el cromosoma y la comparación, mediante matrices de pesos, de los patrones con los hallados experimentalmente en otros organismos.

Tabla 4. Algunas de las posibles regiones de regulación generadas por el programa, con su respectiva puntuación, obtenida de la matriz de pesos por posición de nucleótidos.



IV. CONCLUSIONES.

El programa desarrollado localiza regiones reguladoras de la

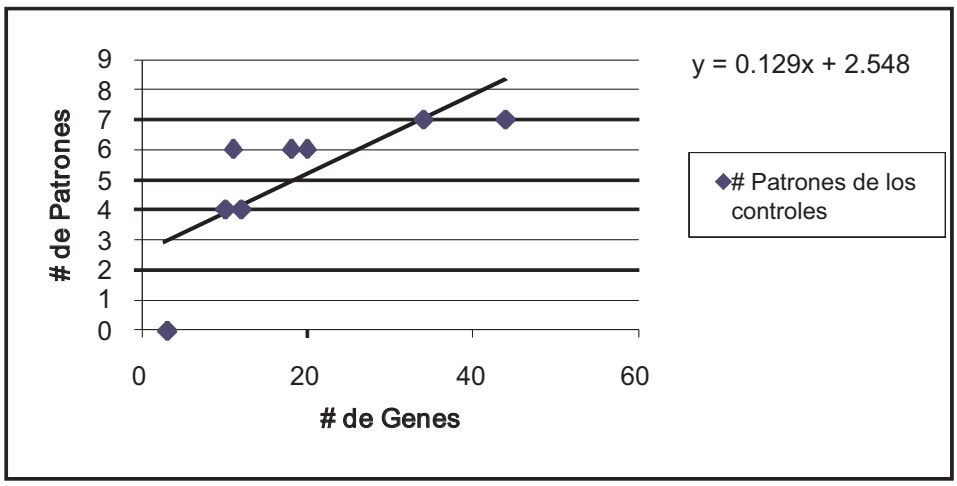


Figura 5. Línea de tendencia de la relación entre el número de patrones de los controles Vs. Número de genes y su función lineal y(x).

transcripción. Los patrones encontrados, fueron los más conservados para regular expresión de genes bajo las mismas condiciones ambientales en un mismo individuo. Al aumentar el número de genes que se expresan bajo las mimas condiciones ambientales, el programa aumenta el número de predicciones lo que indica un mayor número de proteínas involucradas en la regulación génica.

Para las secuencias flanqueadoras de genes 5´, se encontraron varios patrones para una misma secuencia y con longitudes promedio de 7 nucleótidos, lo que indica varias regiones altamente conservadas en los sitios de unión a los factores de transcripción y la participación de más de una proteína en el proceso regulatorio.

Al restringir la búsqueda de secuencias comunes de las regiones flanqueadoras 5´ de cada gen, a

Conjunto de Genes	Pertenece a la región flanqueadora 5´ de los genes...	Patrón O Sitio de Unión obtenido por el programa	Posible regulador encontrado en las bases de datos que se une al patrón	Función del Regulador reportado en las bases de datos.	Puntuación del patrón	Puntuación de la secuencia regulatoria en las bases de datos
0	oppA y oppC	TAAAAAT	OmpR) de *Escherichia coli* (strain K12)	Regula los niveles de expresión de las proteínas porinas externas de membrana OmpF y OmpC	8.62	8.62
0	citR y ywfH	TAAGCCTTTT	Región promotora del gen RhlR de *Pseudomonas aeruginosa*	Regulador transcripcional que regula la expresión génica en respuesta a la densidad celular	9.44	10.14
0	ywfH y rgpC	TAAAACAATAA	Región promotora del gen OxyR de *Escherichia coli* (strain K12)	La cual es una proteína que se produce en células expuestas a H_2O_2 o nitrosotioles, además regula la transcripción de 9 diferentes enzimas, entre las que se encuentran la glutatión reductasa y la alquilhidroperoxido reductasa.	2.93	3.17
1	hisD, ydcG y fruA	TAAAAAAG	AbrB de *Bacillus subtilis*	Regulador de la expresión de genes durante la transición de estados entre el crecimiento vegetativo, la fase estacionaria y la esporulación	7.22	7.44
1	ydcG	TCTTAAAAAAG	NhaR de *Escherichia coli*	regula al gen osmC responsable de la respuesta a diferentes condiciones de estrés	11.76	12.09
1	fruA, yghG, ywfE	TATAAAAAA	PvdS *Pseudomonas aeruginosa*	Regula un factor sigma, responsable de la transcripción del regulón Fur el cual contiene una serie de proteínas reguladoras positivas y negativas dependientes de hierro	6.02	6.29
1	yghG, ywfE	CCCAAATTAGAG	CytR de *Escherichia coli*	Reprime la transcripción de los genes que codifican a las proteínas que transportan y catabolizan nucleótidos	7.71	7.61
1	zitR, zitS y ymgI	CAAAAATC	AbrB de *Bacillus subtilis*	Regula la transcripción de transportadores de zinc.	7.63	7.44

Conjunto de Genes	Pertenece a la región flanqueadora 5 de los genes...	Patrón O Sitio de Unión obtenido por el programa	Posible regulador encontrado en las bases de datos que se unce al patrón	Función del Regulador reportado en las bases de datos.	Puntuación del patrón	Puntuación de la secuencia regulador a en las bases de datos
4	ycdG, pcnB, ycbD y sarK	ATAAAAAA TTTTC	comK de *Bacillus subtilis*	Regula la inducción transcripcional de gen comK y de otros reguladores transcripcionales como comG, comE, comF y comC	5.43	10.13
4	ybgD	ACGTTAC GATGAAC AAACGATT TATAAAGT AGCGGCT GCTTTTGA A	OxyR (SELEX) de *Escherichia coli*	Regulador transcripcional	12.89	18.17
4	ybgD, sarK y comC	GATTTACT TATTTTC	Fis de *Escherichia coli*	Regulador global de la transcripción y un facilitador de eventos de recombinación en sitios específicos, variando su regulación en respuesta a cambios en la disponibilidad de alimento y tasa de crecimiento	3.02	2.57
6	citR, citF, y malF	AGTACCG ATG	SpoIIID de *Bacillus subtilis*		3.65	3.99
6	rgnC y ycbH	GAAAAAA	OmpR de *Escherichia coli*		5.08	0.52
6	ycbD ycbJ ycbF rgpE b4238	AGAAAAT C	regulador Fur (dímer) de *Escherichia coli*		2.01	1.97

secuencias iguales o mayores de 7 nucleótidos, permitió, no sólo localizar secuencias cortas muy conservadas que predominantemente se unen a las proteínas, sino también, secuencias largas de hasta 41 nucleótidos, que las contienen y altamente conservadas de Bacillus subtilis y Escherichia coli, indicando su gran importancia biológica para los microorganismos en los procesos de regulación génica. Una comparación filogenética de estas secuencias podría indicar si la evolución de estos genes ha sido vertical u horizontal.

Las secuencias largas obtenidas por el programa, pueden considerarse no sólo como reguladoras transcripcionales, sino también, como reguladoras a otro nivel del flujo de la información genética, como por ejemplo la traducción, debido a su alta conservación y relación con los genes argF y yajE, implicados en la producción del ARN ribosomal 16S, 5S, 23S y el ARN de transferencia para alanina y asparagina.

El programa predice un número de patrones 5′, 3.3 veces mayor al número de patrones de secuencias flanqueadoras de la región 3′, lo cual apoya los datos experimentales que muestran que los sitios de unión a los factores de transcripción se localizan principalmente en la región 5′, además, solamente el 3.2 % de las secuencias control 3′ coincidieron con las secuencias 5′, indicando un bajo

número de secuencias obtenidas debido a factores aleatorios. El trabajo desarrollado tiene una gran validez, si se considera que aproximadamente el 50 % de los patrones obtenidos en las regiones flanqueadoras 5´, están reportados en las bases de datos de sitios de unión a factores de transcripción, derivados de métodos experimentales. Las secuencias comparadas han tenido pesos idénticos o similares. El segundo caso indica mutaciones de sitio específico debido a la evolución del organismo, que podrían ser utilizadas para deducir aquellos nucleótidos en las secuencias conservadas, que no son esenciales para la unión del ADN con la proteína.

Los resultados obtenidos, son un importante punto de partida, para desarrollar estudios biotecnológicos experimentales que permitan controlar la regulación génica mediante mutaciones dirigidas, debido a que el programa aporta la secuencia patrón y por tanto su localización en el genoma. La alteración de una de estas secuencias, cambiaria la respuesta del organismo a variaciones ambientales, sin necesidad de caracterizar genética y bioquímicamente un conjunto de genes, lo cual, ahorra considerablemente los recursos y el tiempo de obtención de fenotipos que se deseen para aplicaciones que puedan tener una representatividad tecnológica.

Por otra parte, las secuencias patrones y sus correspondientes factores de transcripción obtenidos por la metodología descrita, proporcionan secuencias funcionales de ADN que pueden ser comparadas por homología con organismos próximos y distantes evolutivamente, permitiendo la construcción de hipótesis sobre la manera en que se relacionan los conjuntos de genes que se activan bajo las mismas condiciones ambientales, lo cual contribuiría a los diseños experimentales para localización de secuencias reguladoras de la transcripción y caracterización genética de rutas bioquímicas.

V. BIBLIOGRAFIA.

1 Bussemaker, H. J., Li, H. & Siggia, E. D. (2000) Proc. Natl. Acad. Sci. USA 97,10096–10100.

2 Cliften P, Hillier L, Fulton L, Graves T, Miner T, Gish W, Waterston R & Johnston M. (2001) Genome Res. 11, 1175-1186.

3 Eisen, M. B., Spellman, P. T., Brown, P. O. & Botstein, D. (1998) Proc. Natl. Acad. Sci. USA 95, 14863–14868.

4 McCue, L., Thompson, W., Carmack, C., Ryan, M. P., Liu, J. S., Derbyshire,V. & Lawrence, C. E. (2001) Nucleic Acids Res. 29, 774–782.

5 Pérez – Rueda, E. & Collado – Videz, J. (2000) Nucleic Acids Res. 28, 56 – 59.

6 Robison, K., McGuire, A. M.&Church, G. M. (1998) J. Mol.

Biol. 284, 241–254.
7 Schneider, T. D., Stormo, G. D. & Gold, L. (1986). J. Mol. Biol. 188, 415-431.
[8] Stormo, G. & Hartzell, G. W., 3rd (1989) Proc. Natl. Acad. Sci. USA 86,1183–1187.
9 Van Helden, J., Andre, B. & Collado-Vides, J. (1998) J. Mol. Biol. 281, 827–842.
10 Van Nimwegen, E., Zavolan, M., Rajewsky, N. & Siggia, E. D. (2002) Proc. Natl.Acad. Sci. USA 99, 7323–7328.