Desarrollo de herramientas computacionales para la búsqueda de secuencias reguladoras de la transcripción en procariotas

Carlos Andrés Pérez*

Antonio Pérez**

Juan Falgueras**

Fecha de recepción: 5-3-2007 Fecha de selección: 22-10-2007 Fecha de aceptación: 7-9-2007

ABSTRACT

This paper deals with a PERL program for tracking the DNA sequences that join transcription factors regulating the genetic expression of prokaryotic cells. The gene sets where obtained from their expressions under the same environmental conditions. The model organism was the *lactococcus lactis*, an organism for which we possess its sequenced genome in gene bank format. The program found a larger number of possible regulatory sequences in the 5' flanking region of the genes. The number of possible regulatory sequences was also determined for the amount of genes that made up the set. The program also located flanking gene sequences which could also be involved in its regulation but at a translational level. The comparison of the results with patterns obtained experimentally, was done by means of nucleotide position weight matrices, getting around 50% regulating sequences which coincide with those reported in the databases, indicating good program prediction capability, if you account for the fact that most prokaryotic cell regulating sequences have yet to be characterized by experimental methods.

* Centro de Investigación en Ciencias Básicas, Ambientales y Desarrollo Tecnológico (CICBA), Universidad Santiago de Cali.

** Universidad de Málaga, España

KEYWORDS
Bioinformatics, PERL, transcription, translation, position weight matrix, transcription factors.

RESUMEN
En la presente investigación se elaboró 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 en las mismas condiciones ambientales. El organismo modelo con el que se trabajó 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 localizó secuencias flanqueadoras de genes que podrían estar involucradas en su regulación, pero traduccionalmente.

La comparación de los resultados con patrones obtenidos de manera experimental se hizo mediante matrices de pesos de posición de nucleótidos y se lograron aproximadamente 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 aún no han sido caracterizadas por métodos experimentales.

PALABRAS CLAVE
Bioinformática, PERL, transcripción, traducción, matrices de pesos, factores de transcripción.

Clasificación Colciencias:
Tipo 1.

1. INTRODUCCIÓN

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[1] 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 aun no caracterizados. La gran velocidad a la que se están secuenciando genomas bacterianos hace que el desarrollo de estas herramientas computacionales sea prioritario, ya que una vez obtenido el genoma y definido el transcriptoma la importancia recae sobre la regulación que puede dar fenotipos tan distintos en genotipos similares.

Con el desarrollo de proyectos de secuenciación de genomas y la capacidad de análisis que nos proporcionan las herramientas informáticas entendemos la estrecha relación entre el mundo molecular de los ácidos nucleicos y las proteínas con el entorno ambiental de los organismos. Una de las metodologías más destacables que permiten conocer la expresión génica de un organismo, ante variaciones ambientales, es la tecnología de arreglos de ADN, que consiste en una colección de moléculas situadas sobre un sólido que forman una matriz bidimensional. Los fragmentos cortos de ADNc o productos de PCR tienen la posibilidad de hibridizar con los fragmentos homólogos de las muestras a analizar, que están marcados por métodos fluorescentes o enzimáticos. Esta metodología permite determinar la expresión de un buen número de genes en un momento dado y medir sus niveles de expresión, complementando los estudios sobre transcriptómica en el momento de utilizar herramientas bioinformáticas.

Aunque suele ser habitual el uso de anotaciones de función y procesos biológicos en los experimentos de arreglos de ADN,[2] no lo es tanto en la comparación de regiones reguladoras de la expresión; cada sistema de genes define un contexto particular, y buscar para todas las contingencias a las cuales la célula puede responder puede ser una actividad bastante compleja. Además, los patrones de expresión observados pueden resultar de una cascada de regulación o múltiples factores que actúan simultáneamente, aumentando la di-

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

2 http://www.pubmedcentral.nih.gov/articlerender.fcgi?tool=pubmed&pubmedid=15980548

ficultad de identificar todos los sitios relevantes.

La presente investigación se ha centrado en la elaboración de un programa computacional en lenguaje Perl, que compara secuencias flanqueadoras de genes, de 100 nucleótidos de longitud, a partir de sus extremos 5´ y 3´, con el objetivo de encontrar regiones comunes de regulación de la transcripción. Las comparaciones realizadas han sido a partir de genes que se han expresado en las mismas condiciones ambientales en experimentos de microarrays y utilizan para ello el genoma de la bacteria *Lacotococcus lactis*.

Todas las secuencias flanqueadoras del grupo de genes han sido alineadas entre sí, para hallar regiones comunes de longitud igual o mayor que 7 nucleótidos y con una identidad igual o mayor que el 75%. Una vez se aplica este primer filtro, el programa toma todas las secuencias resultantes y las compara con el fin de obtener secuencias de patrones comunes que serán contrastadas con datos experimentales.

1.1 Identificación de sitios de unión de proteínas reguladoras de la transcripción en genomas bacterianos

Un método general para identificar sitios de unión a factores de transcripción es determinar un grupo de genes corregulados (agrupación de genes con base en sus perfiles de expresión,[3] o a la anotación funcional) y buscar patrones de secuencias comunes en sus regiones reguladoras situadas aguas arriba. Una aproximación alternativa es comparar las regiones reguladoras de genes ortólogos, en diversas especies, e identificar motivos conservados funcionalmente.[4] Estas dos aproximaciones se han utilizado con éxito para analizar genomas bacterianos, pero ambos tienen limitaciones. Por ejemplo, la agrupación de genes con base en su perfil de expresión puede ser un proceso poco exacto y objetivo; cada sistema de genes define un contexto particular, y buscar para todas las contingencias a las cuales la célula puede responder puede ser una actividad bastante compleja. Además, los patrones de expresión observados pueden resultar de una cascada de regulación o múltiples factores que actúan simultáneamente y aumentan la dificultad de identificar todos los sitios relevantes. La comparación interespecie está limitada por la disponibilidad de especies separadas por distancias evolutivas apropiadas. Además, los algoritmos de alineamiento múltiple no tienen una consideración significativa de las relaciones filogenéticas. Finalmente, cuando las secuencias conservadas son identificadas, se deben agrupar los sitios potenciales para cada gen en sus regulones,[5] actividad bastante dispendiosa.

Los algoritmos computacionales utilizados para encontrar sitios comunes de un grupo de genes pueden ser catalogados como de **búsqueda**

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 Van Nimwegen, E., Zavolan, M., Rajewsky, N. & Siggia, E. D. (2002) *Proc. Natl. Acad. Sci. USA* **99,** 7323–7328.

directa, es decir, buscan en todas las secuencias por procesos iterativos y conjeturan un patrón.[6] El esfuerzo computacional crece exponencial según la longitud del sitio, mientras más grandes sean, los patrones se tornan difusos y pueden no converger con el óptimo global. Los algoritmos también difieren en la manera en que determinan la significación estadística de un motivo: **extrínsecamente**, por el contraste de la frecuencia del motivo obtenido del agrupamiento de genes respecto a su frecuencia en el resto del genoma;[7] o **intrínseco**, por la diferencia entre el número de ocurrencias del motivo y su valor esperado.

La utilización de genomas enteros dificulta el trabajo de los programas debido a la cantidad de datos implicados, la ausencia de comparaciones convenientes para los métodos extrínsecos, la multiplicidad de patrones y las limitaciones de fondo que tienen los modelos simples en el manejo de las probabilidades.

2. MATERIALES Y MÉTODO

El programa desarrollado **alineamiento.pl**, solicita al usuario los nombres de los archivos en donde se encuentra el genoma secuenciado del organismo con el que se está trabajando (*Lactococcus lactis*; en formato genbank*)* el listado de genes con expresión o anotación correlacionada, longitud de las secuencias flanqueadoras de los genes en sus extremos 5´y 3´, porcentaje mínimo de coincidencia de las secuencias flanqueadoras alineadas entre sí y longitud mínima del alineamiento.

Para la ejecución del programa se deben crear las carpetas 5´y 3´, en las que se guardarán las secuencias flanqueadoras correspondientes. Para la obtención de los mejores alineamientos locales el programa se apoya en el software **lalign.exe,**[8] el cual es ejecutado comparando cada una de las secuencias entre sí 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 que el 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 siete nucleótidos, que predominantemente hacen contacto con los factores de transcripción[9]

6 Stormo, G. & Hartzell, G. W., 3rd (1989). *Proc. Natl. Acad. Sci.* USA 86,1183–1187.

7 Van Helden, J., Andre, B. & Collado-Vides, J. (1998). *J. Mol. Biol.* 281, 827–842.

8 Programa para el alineamiento local de dos secuencias a partir de regiones comunes. http://fasta.bioch.virginia.edu/

9 Robison, K., McGuire, A. M.&Church, G. M. (1998). *J. Mol. Biol.* 284, 241–254.

y que por cuestiones de evolución neutral pueden variar en uno o dos nucleótidos.

Una vez seleccionados los resultados con estas características, el programa compara todas las secuencias entre sí y busca patrones comunes. Los reportes de resultados son guardados en los archivos **alineamientosComunes3´.txt** y **alineamientosComunes5´.txt**.

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

2.1 Pograma desarrollado

El programa puede obtenerse en la dirección electrónica:

http://www.usc.edu.co/investiga/cicba/alineamiento.txt

2.2 Comparación de patrones

Una vez se han obtenido los patrones comunes de las secuencias flanqueadoras de un mismo listado de genes, se procede a compararlos con datos experimentales. En el ámbito computacional aún no se han elaborado matrices de pesos para sitios de unión de factores de transcripción de *Lactococcus lactis,*[10] sin embargo, en el sitio *Virtual Footprint* (http://www.prodoric.de/vfp/vfp_promoter.php),se pueden comparar con matrices derivadas de diferentes genomas bacterianos.

2.3 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 está conformado por genes tomados al azar, con el fin de tener un control negativo.

2.4 Método para determinar el valor de cada nucleótido en las matrices de pesos

Este método es derivado de la teoría de la información,[11] que consiste en calcular el vector *RSequence(l)* mediante la fórmula:

$$R_{Sequence}(l)=2+\sum_{b=A}^{r} f(b,l)\cdot\log_2 f(b,l)$$

F(b, l)[12] 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 fórmula:

$$m(b,l) = f(b,l).\, R_{Sequence}(l)$$

Donde f(b,l) es igual a:

$$f(b,l)=\frac{1-n}{n+2}$$

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

---

10 *Lactococcus lactis* IL1403 sequencing project, http://spock.jouy.inra.fr/

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

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

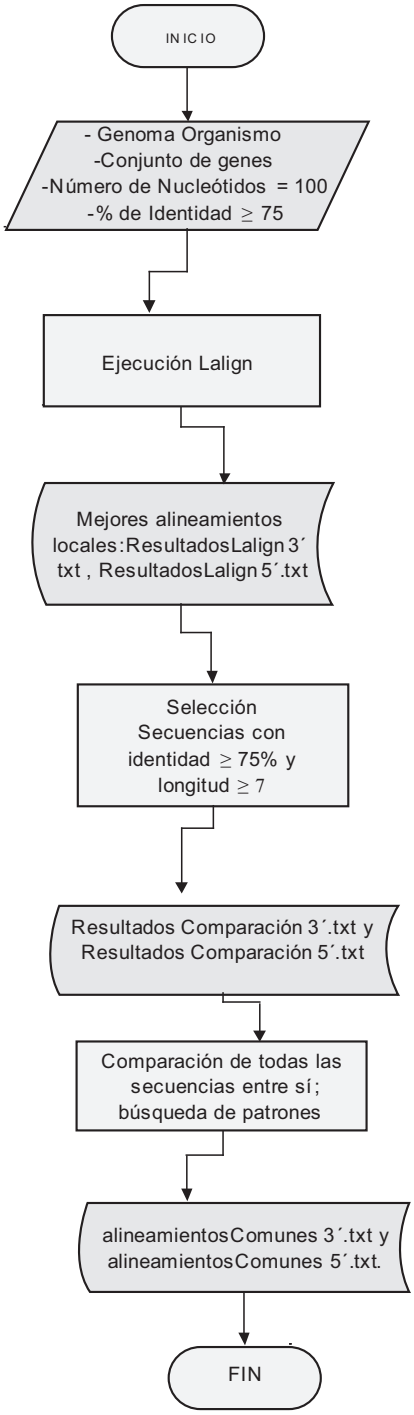

Figura 1. Diagrama de flujo de los procesos que realiza el programa alineamiento.pl.

2.5 Bases de datos y sitios electrónicos de referencia

Las anotaciones funcionales de los genes de *Lactococcus lactis* y su posición en el cromosoma se encontraron en la base de datos de genomas microbianos para análisis comparativos (MBGD) del Instituto Nacional de Biología Básica y Ciencias Naturales del Japón.[13]

Las matrices de pesos se obtuvieron a partir de las secuencias de unión a factores de transcripción, reportadas en el sitio *prokaryotic database of gene regulation* (Prodoric).[14]

Para la comparación entre los pesos de las secuencias generadas por el programa y las reportadas en la base de datos, se utilizó la herramienta virtual *footprint*, que busca y reorganiza patrones de trascripción iguales o similares a los de las secuencias flanqueadoras y que se encuentran en los genomas bacterianos antes descritos.[15]

3. 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 calculado el número de patrones obtenidos por conjunto de genes y su longitud promedio (Tabla 1).

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 y es 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 y su número es 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 flanqueadoras 5´, se obtuvieron secuencias similares (Tabla 2).

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,

13 http://mbgd.genome.ad.jp/
14 http://www.prodoric.de/
15 http://www.prodoric.de/vfp/vfp_promoter.php

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 de la región 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.

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

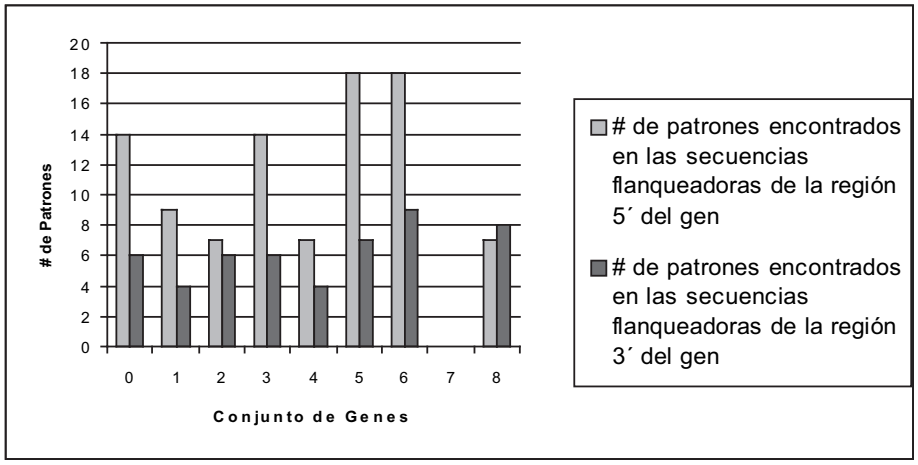

Figura 2. Histograma para la comparación del número de patrones obtenidos de las secuencias flanqueadoras 5´ y 3´.

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.

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

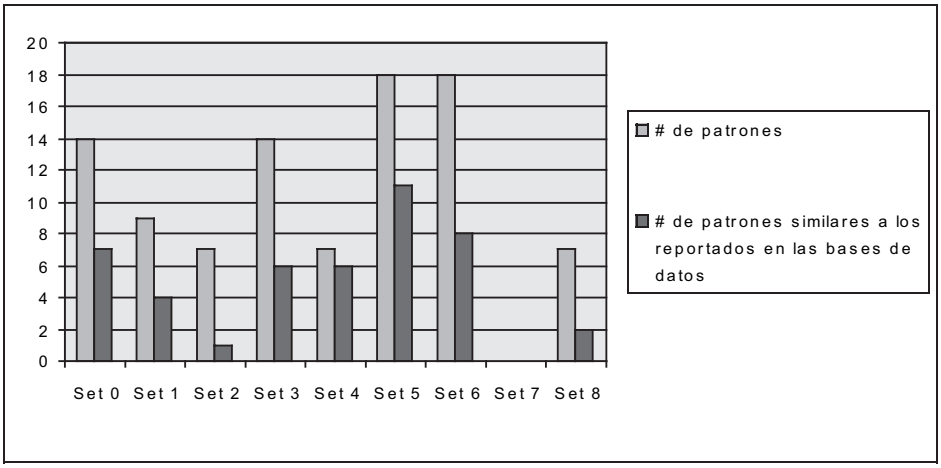

Figura 3. Histograma para la comparación del número de patrones obtenidos de las secuencias flanqueadoras 5´ y las reportadas en las bases de datos de sitios de unión a factores de transcripción.

debido a que este conjunto se elaboró con genes seleccionados al azar y no por expresarse en las mismas 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, lo que indica que el 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 de lo que ocurre en eucariotas, cuyas regiones de regulación génica pueden encontrarse en sitios aguas debajo de los genes o regiones intrónicas.[16] Por esto 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.

Los conjuntos con los que se trabajó estaban conformados por un número distinto de genes. La distribución de los datos muestra una tendencia lineal, que indica 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

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

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

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

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

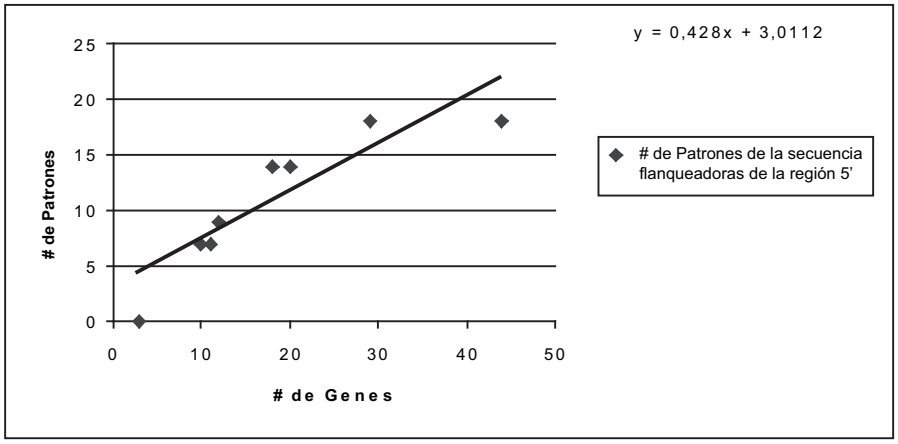

Figura 4. Línea de tendencia de la relación entre el número de patrones de secuencias flanqueadoras de la región 5´ de cada conjunto de genes Vs. Número de genes y su función lineal y(x).

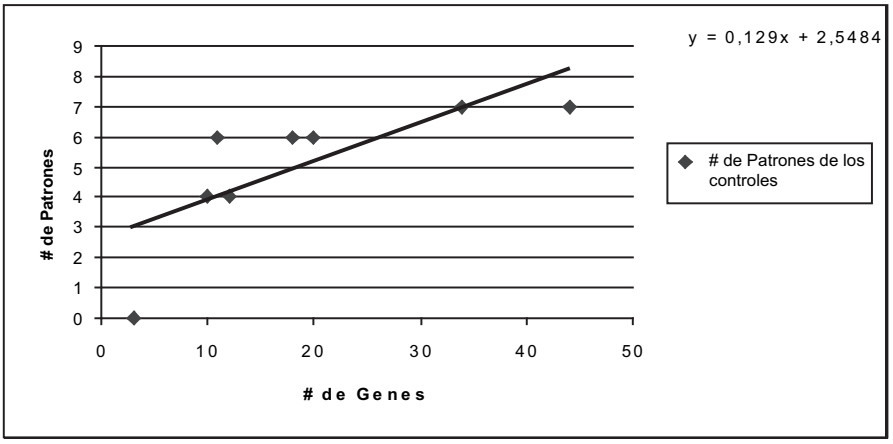

Figura 5. Línea de tendencia de la relación entre el número de patrones de los controles Vs. Número de genes y su función lineal y(x).

de los controles, lo que indica que la tendencia del programa es obtener mayor número de patrones de las secuencias que flanquean aguas arriba 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.

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 reguladora 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 glutation reductasa y la alquilhidroperóxido 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	NharR de *Escherichia coli*	Regula el 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

Continúa

Continuación Tabla 4

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 reguladora en las bases de datos
1	zitR, zitS y ymgI	CAAAAATC	AbrB de *Bacillus subtilis*	Regula la transcripción de transportadores de zinc,	7.63	7.44
4	ycbB, pth, ycbD y scrK	ATAAAAAATTTTC	comK de *Bacillus subtilis*	Regula la inducción transcripcional del gen comK y de otros reguladores transcripcionales como comC, comE, comF y comG	8.43	10.13
4	ybgD	ACGTTACGATGAAG AAACGATTTATAAA GTAGCGGCTGCTT TTGAA	OxyR (SELEX) de *Escherichia coli*	Regulador transcripcional	12.89	18.17
4	ybgD, scrK y comC	GATTTACTTATTTTC	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 fase de crecimiento	3.02	2.87
6	citR, citE, y malF	AGTACCGATG	SpoIIID de *Bacillus subtilis*		3.65	3.99
6	rgpC y ycbH	GAAAAAA	OmpR de *Escherichia coli*		8.06	8.22
6	ycbD, ycbJ, ycbF, rgpE, pi208	AGAAAATC	Regulador Fur (8mer) de *Escherichia coli*		2.01	1.97

4. CONCLUSIONES

El programa desarrollado localiza regiones reguladoras de la transcripción. Los patrones encontrados fueron los más conservados para regular expresión de genes en las mismas condiciones ambientales en un mismo individuo. Al aumentar el número de genes que se expresan en las mismas 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 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*, lo que da a conocer 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´, que indican 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 cambiaría 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, que permiten la construcción de hipótesis sobre la manera como se relacionan los conjuntos de genes que se activan en 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.

5. 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). Information content of binding sites on nucleotide sequences. *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.

CURRÍCULOS

Carlos Andrés Pérez G. (caperez@usc.edu.co) es biólogo con énfasis en genética, especialista en Simulación Molecular, magíster en Bioinformática, diplomado de estudios avanzados y doctorando en Biotecnología. Director del Centro de Investigación en Ciencias Básicas, Ambientales y Desarrollo Tecnológico (CICBA) de la Universidad Santiago de Cali y el grupo de Investigación en Biotecnología y Medio Ambiente (GIBMA), categorizado por Colciencias. Sus líneas de investigación son caracterización filogenética de *Guadua angustifolia* y desarrollo de herramientas computacionales para la localización de regiones de ADN que se unen a factores de transcripción.

Antonio J. Pérez (aperezp@uma.es) es doctor en Bioinformática. Líneas de trabajo: Análisis de secuencia y predicción de funciones proteicas en el Instituto Nacional de Bioinformática, Nodo GNV5 de Integración Universidad de Málaga - España.

Juan Falgueras Cano (juanfc@lcc.uma.es) es físico de la Universidad de Sevilla, trabaja con ordenadores desde el año 1983. Doctor en Informática del Departamento de Lenguajes y Ciencias de la Computación de la Universidad de Málaga – España, en la que se desempeña como profesor titular. Sus áreas de investigación son: Matemáticas aplicadas a la física, Ingeniería del Software y Sistemas de Información, específicamente en la evaluación formal de la usabilidad en Interfaces de usuario adaptativas.