SPADE: Algoritmo Evolutivo para Descubrir Patrones de Desplazamiento en Microarrays

Beatriz Pontes[1], Raúl Giráldez[2], Jesús S. Aguilar-Ruiz[2]

Resumen— **El interés por extraer conocimiento útil de datos de expresión genómica ha experimentado un enorme auge en los últimos años con el desarrollo de los microarrays. Las técnicas de biclustering son aplicadas para obtener subconjuntos de genes que se expresen de manera similar frente a determinadas condiciones en un microarray. Una manera de medir la calidad de un bicluster es detectando si los genes que contiene siguen una tendencia similar, representada por un patrón. En este artículo se aborda este problema mediante computación evolutiva, presentando un algoritmo, llamado SPADE, para la extracción de patrones de desplazamiento en biclusters. Los resultados empíricos obtenidos por SPADE muestra la calidad de nuestra propuesta.**

Palabras clave— **Datos de Expresión Genómica, Biclustering, Algoritmos Evolutivos.**

I. INTRODUCCIÓN

Un microarray es una matriz bidimensional que almacena datos de expresión genómica, donde las filas representan condiciones experimentales y las columnas se corresponden con los diferentes genes a estudiar. De esta manera, cada elemento de la matriz equivale al nivel de expresión de un gen bajo una determinada condición. Con fin de extraer información a partir de microarrays, se han utilizado diversas técnicas de clustering [3], generalmente agrupando genes teniendo en cuenta sus relaciones funcionales sobre todas las condiciones experimentales. Sin embargo, los genes que guardan una potencial relación entre sí no tienen por qué hacerlo con respecto a todas las condiciones [13].

Las técnicas de biclustering son una variante de las técnicas de clustering, donde la búsqueda se realiza simultáneamente sobre las filas y columnas en la matriz. En el caso del análisis de microarrays, dichas técnicas son utilizadas para poder identificar grupos de genes relacionados entre sí frente a subconjuntos de condiciones experimentales [7]. Éstas técnicas se basan en la idea de que no todos los genes de un microarray son relevantes para todas las condiciones, aplicando así los conceptos de clustering sobre las dos dimensiones a la vez. Dado su gran interés, los métodos de biclustering para el análisis de datos biológicos han sido ampliamente abordados en los últimos años [4], [5], [12].

1.Dpto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, Avda. Reina Mecedes s/n - 41012 - Sevilla. E-mail: bepontes@lsi.us.es.

2.Escuela Politécnica Superior, Universidad Pablo de Olavide, Ctra. Utrera Km 1 - 41013 - Sevilla. E-mail: {giraldez, aguilar}@upo.es.

Un bicluster puede contener un conjunto de genes con un comportamiento similar, aunque el valor de expresión de éstos no lo sea, es decir, dichos genes siguen un determinado patrón de comportamiento [1]. De esta manera, podemos hablar de dos tipos de patrones diferentes en un bicluster, patrones de desplazamiento y patrones de escalado. Estos patrones son utilizados para describir el comportamiento común de los genes en un bicluster, así como también pueden resultar útiles para poder añadir más genes o condiciones a un bicluster dado.

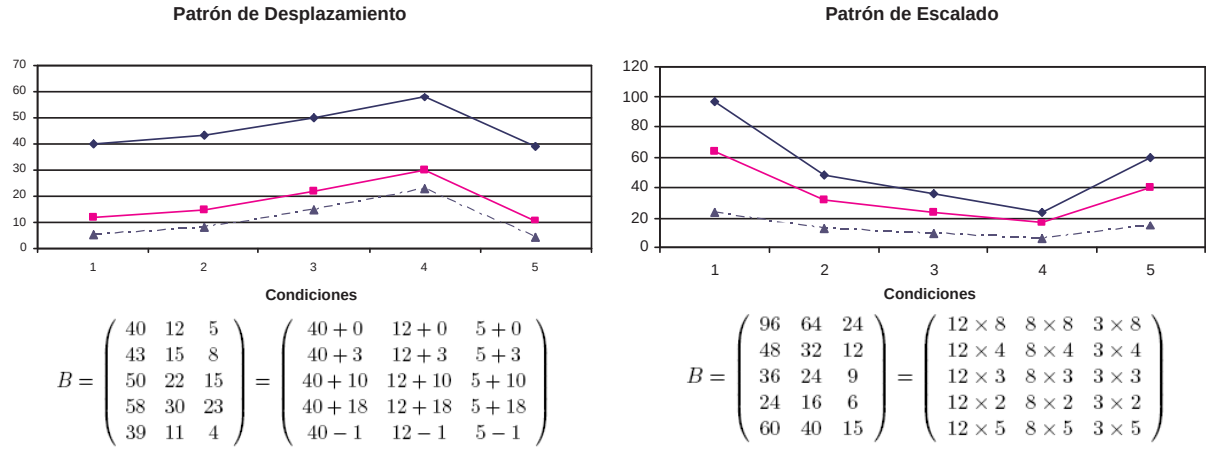
En este trabajo se presenta un nuevo algoritmo, denominado SPADE (*Shifting PAttern Discovery based on Evolutionary algorithms*), que aplica un algoritmo evolutivo para encontrar patrones de desplazamiento que representen, de la manera más precisa posible, el comportamiento de todos los genes contenidos en un determinado bicluster. Los resultados experimentales muestran que SPADE obtiene dichos patrones con gran exactitud.

El resto de este trabajo está organizado de la siguiente forma: los distintos tipos de patrones que un bicluster puede presentar son presentados en la sección II; en la sección III se detalla el algoritmo SPADE; los resultados obtenidos en los experimentos llevados a cabo son presentados en la sección IV; finalmente, la última sección resume las principales conclusiones de este trabajo.

II. PATRONES EN DATOS DE EXPRESIÓN GENÓMICA

Como ya se ha mencionado, todos los genes que forman un bicluster pueden seguir un comportamiento similar, al que llamamos patrón. Dichos patrones fueron introducidos en [7], aunque se encuentran formalmente descritos en [1], dónde se distingue entre dos tipos de patrones que pasamos a definir.

Sea $\mathcal{M}$ un microarray con N filas (condiciones c_i, con $1 \leq i \leq N$), y M columnas (genes g_j, con $1 \leq j \leq M$), de manera que cada elemento de la matriz vendrá representado por $v_{ij} \in \mathcal{M}$. Sea también $\mathcal{B} \subseteq \mathcal{M}$ un bicluster compuesto por $n \leq N$ condiciones y $m \leq M$ genes, de manera que cada elemento perteneciente al bicluster vendrá representado por $w_{ij} \in \mathcal{B}$. A partir de estas definiciones, podemos definir los patrones de desplazamiento y escalado como sigue [1]:



$$B = \begin{pmatrix} 40 & 12 & 5 \\ 43 & 15 & 8 \\ 50 & 22 & 15 \\ 58 & 30 & 23 \\ 39 & 11 & 4 \end{pmatrix} = \begin{pmatrix} 40+0 & 12+0 & 5+0 \\ 40+3 & 12+3 & 5+3 \\ 40+10 & 12+10 & 5+10 \\ 40+18 & 12+18 & 5+18 \\ 40-1 & 12-1 & 5-1 \end{pmatrix} \qquad B = \begin{pmatrix} 96 & 64 & 24 \\ 48 & 32 & 12 \\ 36 & 24 & 9 \\ 24 & 16 & 6 \\ 60 & 40 & 15 \end{pmatrix} = \begin{pmatrix} 12\times 8 & 8\times 8 & 3\times 8 \\ 12\times 4 & 8\times 4 & 3\times 4 \\ 12\times 3 & 8\times 3 & 3\times 3 \\ 12\times 2 & 8\times 2 & 3\times 2 \\ 12\times 5 & 8\times 5 & 3\times 5 \end{pmatrix}$$

Fig. 1. Patrones de desplazamiento y escalado.

▪ *Patrón de desplazamiento.* Un bicluster $\mathcal{B}$ sigue un patrón de desplazamiento cuando sus valores w_{ij} se pueden obtener sumando un cierto valor β_i, que será constante para la condición i-ésima, a un valor típico (π_j) para el gen j-ésimo. Formalmente, un bicluster presenta un patrón de desplazamiento cuando sus valores se rigen por la siguiente expresión:

$$w_{ij} = \pi_j + \beta_i + \xi_{ij} \tag{1}$$

donde w_{ij} denota el valor que presenta el gen j bajo la condición i en dicho bicluster; π_j es el valor fijo para el gen j-ésimo; $beta_i$ es el valor de desplazamiento para la condición i; y ξ_{ij} representa el error cometido por el patrón para el valor w_{ij}.

▪ *Patrones de escalado.* La definición de patrón de escalado es análoga a la del de desplazamiento, sustituyendo el valor aditivo β_i por un factor multiplicativo α_i, tal y como se muestra en la expresión:

$$w_{ij} = \pi_j \times \alpha_i + \xi_{ij} \tag{2}$$

donde w_{ij} denota el valor que presenta el gen j bajo la condición i en dicho bicluster; π_j es el valor fijo para el gen j-ésimo; $alpha_i$ es el valor de escalado para la condición i; y ξ_{ij} representa el error cometido por el patrón para el valor w_{ij}.

En ambos casos, cuando el error cometido ξ_{ij} es 0 para todos los valores del bicluster, decimos que se trata de un *bicluster perfecto*. La figura 1 muestra un ejemplo de dos bicluster que presentan un patrón de desplazamiento perfecto (a la izquierda), y un patrón de escalado perfecto (a la derecha). Como se puede observar, en el caso del patrón de desplazamiento, las líneas que representan los genes tienen la misma forma, presentando la mismas pendientes en todos sus tramos y variando solamente en el rango en el que están situadas. Sin embargo, en el caso de patrones de escalado, dichas representaciones tendrán en común la forma pero con diferentes pendientes.

III. Búsqueda de Patrones de Desplazamiento

El objetivo de este trabajo es la búsqueda del patrón de desplazamiento que represente de la manera más precisa posible el comportamiento general de un bicluster. Para ello, se afronta el problema en el contexto de los algoritmos evolutivos [6], [11], proponiendo un algoritmo denominado Spade (*Shifting PAttern Discovery based on Evolutionary algorithms*). Partiendo de un bicluster, Spade devuelve como resultado el conjunto de valores β_i para dicho bicluster, que definirá el mejor patrón encontrado.

Cada individuo ($\mathcal{I}$) está formado por una secuencia de números reales que representan los valores beta en codificación real ($\mathcal{I} = \{\beta_1, \beta_2, \ldots, \beta_i, \ldots, \beta_n\}$), que se corresponde con un posible patrón de desplazamiento. Todos los individuos tendrán la misma longitud e igual al número de condiciones del bicluster.

Spade está basado en un esquema típico de algoritmo evolutivo. La población inicial puede ser generada de dos formas diferentes: aleatoriamente o mediante un proceso basado en generar mutaciones de los valores originales del bicluster. Esta última manera es la que se ha aplicado para realizar los experimentos que se presentan en este artículo. En cada iteración, se evalúa cada individuo de acuerdo con la función objetivo que se define en la ecuación 3, y que está basada en el error medio absoluto (MAE) de cada individuo, esto es, la media de los valores $|\xi_{ij}|$ (ecuación 4).

$$\phi(\mathcal{I}) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m}|\xi_{ij}|}{n \times m} \tag{3}$$

$$\xi_{ij} = w_{ij} - \pi_j - \beta_i \tag{4}$$

En el contexto de este trabajo se han implementado también otras alternativas para la función objetivo, tales como el error medio cuadrático, produciendo resultados similares.

Al final de cada generación, una copia del mejor individuo se introduce sin modificación alguna en la generación siguiente (elitismo). Tras esto, se selecciona un conjunto de individuos mediante el método de la ruleta [9], que serán también copiados a la siguiente generación. El tamaño de dicho conjunto dependerá del porcentaje de réplicas, que será configurable. A estas réplicas se les aplica el operador de mutación según la probabilidad establecida. Por último, el resto de la población se forma a partir de la aplicación del operador de cruce. Dicho operador crea nuevos individuos a partir de dos padres y un número determinado de puntos de cruce, intercambiando las secuencias en cada uno de esos puntos. De esta manera se generan individuos en los que los valores beta proceden de los dos padres. El operador de mutación se aplica a esta descendencia dependiendo de la probabilidad de mutación. De esta manera, cada vez que un individuo es elegido para ser mutado, se elige también un valor beta que será alterado. El nuevo valor será calculado en función del valor anterior y la media de los errores cometidos para ese beta, pudiendo ser éste tanto positivo como negativo, dependiendo del rango de los valores del bicluster. Tras un número predeterminado de generaciones, el algoritmo devuelve el mejor conjunto de valores beta encontrado.

IV. Experimentos

Para comprobar el rendimiento del método propuesto para la búsqueda de patrones de desplazamiento, se han llevado a cabo diferentes pruebas sobre biclusters obtenidos en trabajos previos [10]. Estos biclusters fueron obtenidos tras la aplicación de un algoritmo evolutivo a partir de dos bases de datos muy conocidas y utilizadas en este contexto: el ciclo celular de la levadura, *Saccharomyces cerevisiae* [8]; y datos de expresión de células humanas, B-cells [2]. En esta sección se muestran los resultados obtenidos empíricamente.

En la tabla I se muestran los parámetros de configuración utilizados para los resultados que se presentan en este trabajo. El tamaño de población y número de generaciones se establecieron a 100, mientras que la probabilidad de mutación y la probabilidad de cruce fueron fijadas a 0,50 y 0,80, respectivamente. El número de puntos de cruce es igual a 2.

Spade ha sido probado utilizando diversos tipos de biclusters, con diferente número de genes y/o condiciones, así como mostrando diferentes grados de comportamiento de desplazamiento. Ya que todos ellos fueron obtenidos a partir de datos reales, ninguno muestra un comportamiento de desplazamiento perfecto, presentando también tendencias de escalado. En el caso de ejecutar Spade sobre biclusters que presenten un desplazamiento perfecto, el algoritmo hubiera obtenido una solución con un error igual a cero desde la primera iteración, debido al método de inicialización de la población.

TABLA I
Parámetros de SPADE

Parámetro	Valor
Número de generaciones	100
Tamaño de la población	100
Probabilidad de cruce	0.8
Probabilidad de réplicas	0.2
Probabilidad de mutación	0.5
Número de Puntos de Cruce	2

El algoritmo fue ejecutado varias veces para todos los biclusters obtenidos en [10], mostrando a continuación algunos de los resultados obtenidos.

A. Levadura

En la figura 2 se puede observar el resultado de aplicar Spade sobre tres de los biclusters analizados. En dicha figura se presentan dos gráficas por cada bicluster: la primera de ellas se corresponde con el bicluster original (a la izquierda), mientras que la segunda muestra el patrón obtenido para dicho bicluster (a la derecha). En estas gráficas es posible apreciar cómo la calidad del patrón encontrado depende del grado en que los genes de cada bicluster siguen una misma tendencia de desplazamiento.

De forma general, es posible decir que el patrón busca una uniformidad en el comportamiento, ignorando formas o tendencias locales. Por ejemplo, en el bicluster 99_1, el comportamiento global de los genes queda representado con gran exactitud por su patrón. Por tanto, un patrón se ajustará en mayor grado a todos los genes cuanto más uniforme sea el comportamiento que éstos muestran. El bicluster 64_1 tiene asociado el error final más bajo ($MAE = 10{,}8$), por lo que el patrón encontrado por Spade representa de una manera más precisa el comportamiento de todos los genes que en otros biclusters analizados. El valor más alto de la función objetivo se obtiene para el bicluster 79_1, ($MAE = 11{,}7$), debido a que los genes que lo componen presentan diversas variaciones entre ellos, a pesar de estar compuesto por un número bajo de genes. En cualquier caso, las diferencias existentes entre los distintos errores cometidos no son significativas. Finalmente, cabe destacar también la rápida convergencia Spade presentó para la mayoría de los biclusters.

B. Células Humanas

La figura 3 presenta los resultados obtenidos para tres de los cien biclusters analizados para la base de datos de células humanas. La estructura

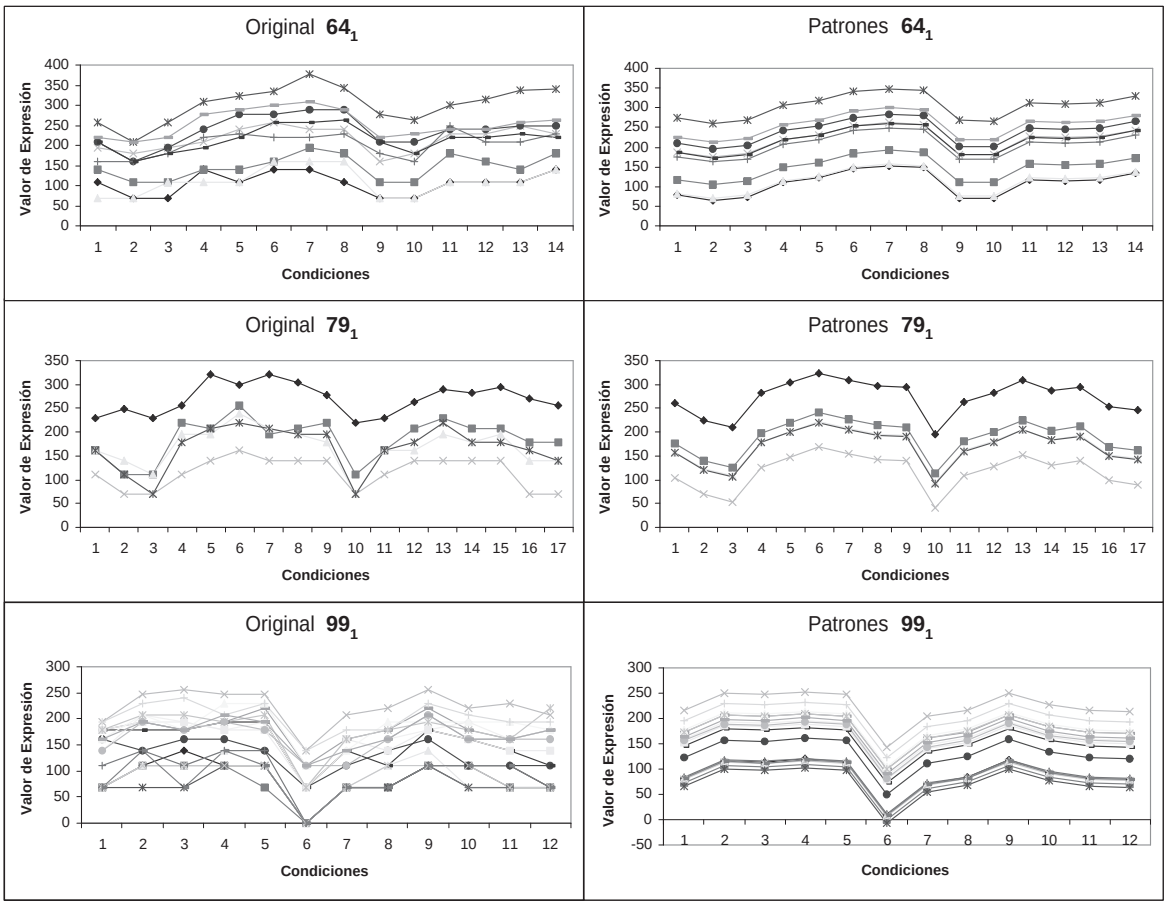


Fig. 2. Tres biclusters analizados por SPADE para la bases de datos de la levadura.

es similar a la de la figura anterior. Teniendo en cuenta los resultados ya presentados de la levadura, a continuación se destacan algunas diferencias.

En primer lugar, es posible observar la presencia de valores negativos, aunque esto no es relevante para nuestro algoritmo. Además, los biclusters obtenidos de células humanas están compuestos por un número mayor de condiciones que en el caso anterior, por lo que es previsible que los valores de la función objetivo sean mayores. Por último, los genes que se analizan en este caso se encuentran más próximos unos de otros, obteniéndose por lo tanto patrones más cercanos, como puede apreciarse comparando las figuras 2 y 3, donde el rango de los patrones obtenidos es mucho más amplio en el primer caso.

También es destacable que los biclusters analizados para el caso de las células humanas son más heterogéneos que en el caso de la levadura. Así, en la figura 3 mostramos diferentes tipos de biclusters. El bicluster 101_1 está formado solamente por 3 genes y por un gran número de condiciones (72), mientras que el bicluster 50_1 tiene un tamaño medio, ya que consta de 11 genes y 58 condiciones. Sin embargo, el valor del error dependerá de la similitud entre las curvas de genes y no del tamaño del bicluster. No obstante, a mayor tamaño, más difícil resultará que los genes sigan un comportamiento similar de manera general, por ello, como era de prever, los valores de la función objetivo son mayores que en el caso de la levadura, debido principalmente al mayor número de condiciones que presentan los bicluster de este segundo caso de estudio.

Por otro lado, al igual que en el caso de la levadura, no existe mucha diferencia entre los errores cometidos para los distintos biclusters (el mejor valor del MAE es 24,7 para el 101_1, y el peor es 27,4 para el 31_1). En este caso, la convergencia del algoritmo no resultó tan rápida como en el caso anterior, aunque se consigue un valor estable en las últimas generaciones.

V. CONCLUSIONES

Las técnicas de biclustering se aplican sobre datos de expresión genómica para agrupar genes y condiciones de un microarray simultáneamente. Los genes que componen un mismo bicluster pueden caracterizarse por tener un comportamiento similar para las condiciones presentes. Es posible modelar dicho comportamiento mediante un patrón. Un caso particular de patrón de comportamiento es el de desplazamiento, que se caracteriza por considerar que la representación gráfica del bicluster presenta curvas paralelas para los genes, donde

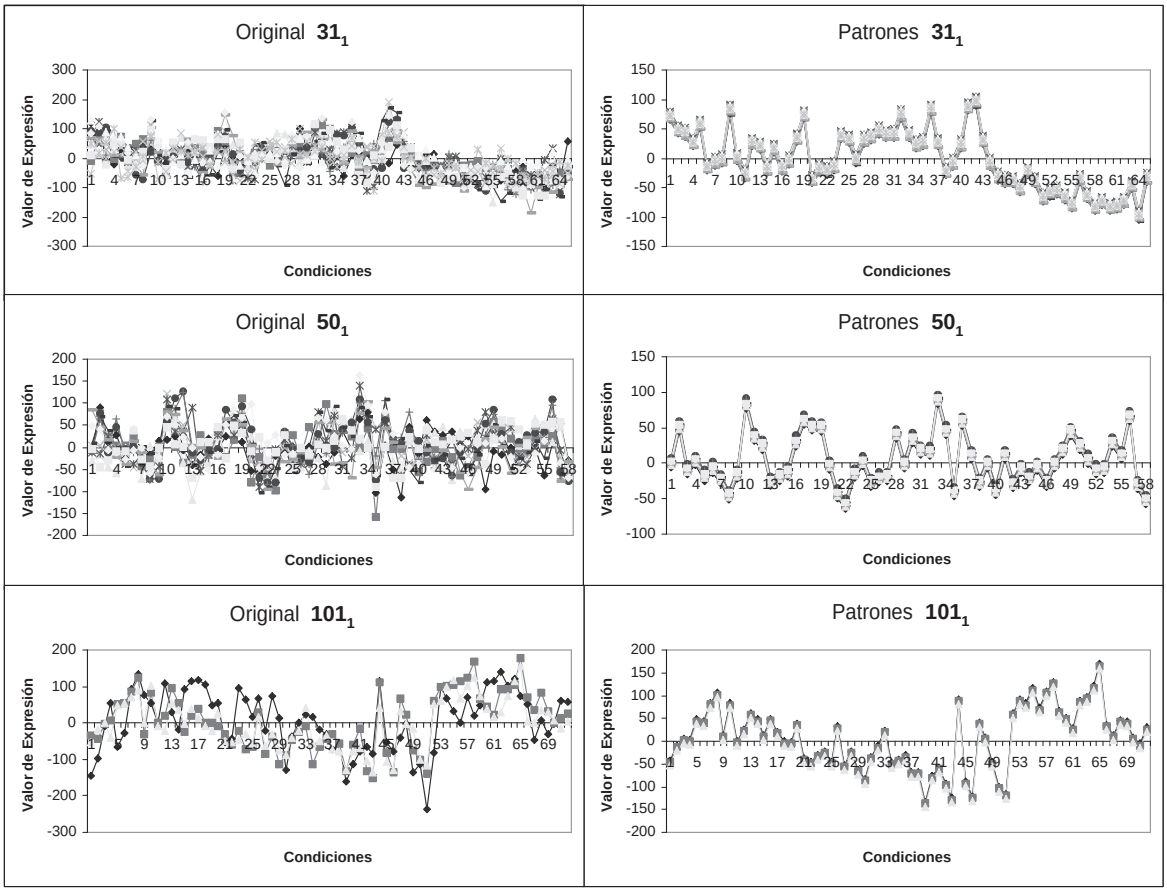


Fig. 3. Tres biclusters analizados por SPADE para la bases de datos de las células humanas.

sólo varía el rango en el que éstos tienen su nivel de expresión. En este artículo se presenta una nueva herramienta, denominada SPADE, que busca patrones de desplazamiento que representen la tendencia general de los genes de un bicluster. Partiendo de un bicluster, SPADE aplica un algoritmo evolutivo para obtener los coeficientes que definan el patrón. Los resultados obtenidos tras llevar a cabo gran número de pruebas empíricas confirman la calidad de nuestra propuesta para encontrar este tipo de patrones, produciendo soluciones muy precisas. Los trabajos futuros en este contexto se basarán en la búsqueda de ambos tipos de patrones, desplazamiento y escalado, simultáneamente.

AGRADECIMIENTOS

Este trabajo ha sido subvencionado por la Comisión Interministerial de Ciencia y Tecnología, CICYT, con los proyectos TIN2004–00159 y TIN2004-06689C0303.

REFERENCIAS

[1] J. S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21:3840 – 3845, 2005.

[2] A. A. Alizadeh. et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503 – 511, 2000.

[3] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281 – 297, 1999.

[4] S. Bleuler, A. Prelić, and E. Zitzler. An ea framework for biclustering of gene expression data. In *Congress on Evolutionary Computation (CEC-2004)*, pages 166 – 173, Piscataway, NJ, 2004.

[5] K. Bryan, P. Cunningham, and N. Bolshakova. Biclustering of expression data using simulated annealing. In *18th IEEE Symposium on Computer-Based Medical Systems*, pages 383 – 388, Dublin, Ireland, 2005.

[6] L. D. Chambers et al. *Practical Handbook of Genetic Algorithms, volume III*. CRC Press, 1999.

[7] Y. Cheng and G. M. Church. Biclustering of expression data. In *In Proceedings of the 8th International Conference on Intellingent Systemns for Molecular Biology*, pages 93 – 103, La Jolla, CA, 2000.

[8] R. Cho. et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65 – 73, 1998.

[9] K. A. DeJong. *An analysis of the behavior of a class of genetic adaptive systems*. PhD thesis, University of Michigan, 1975.

[10] F. Divina and J. S. Aguilar-Ruiz. Biclustering of expression data with evolutionary computation. *IEEE Transactions on Knowledge & Data Engineering*, 18(5):590 – 602, 2006.

[11] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning*. Addison Wesley, 1989.

[12] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Trans. on Comput. Biology and Bioinformatics*, 1:24 – 25, 2004.

[13] H. Wang, W. Wang, J. Yang., and P. S. Yu. Clustering by pattern similarity in large data sets. In *ACM SIGMOD International Conference on Management of Data*, page 394 – 405, Madison, WI, 2002.