Evaluación de biclusters en un entorno evolutivo

Beatriz Pontes
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Sevilla
bepontes@lsi.us.es

Raúl Giráldez
Escuela Politécnica Superior
Univ. Pablo de Olavide
giraldez@upo.es

Federico Divina
Escuela Politécnica Superior
Univ. Pablo de Olavide
fdivina@upo.es

Francisco Martínez-Álvarez
Dpto. Lenguajes y Sistemas Informáticos
Universidad de Sevilla
fmartinez@lsi.us.es

Resumen

La mayoría de las heurísticas utilizadas para la búsqueda de biclusters en microarrays hacen uso del residuo cuadrático medio (MSR) como medida de evaluación de las distintas soluciones generadas. El uso de MSR permite obtener biclusters importantes, sin embargo, recientemente se ha demostrado que dicha medida no es válida para reconocer determinados tipos de biclusters. En este trabajo se propone una nueva medida para la evaluación de biclusters, denominada *Error Virtual* (VE), y que ha sido incorporada en el contexto de un algoritmo de búsqueda evolutivo, con el fin de comprobar su validez. Los resultados experimentales que se obtienen aplicando dicha heurística muestran que el uso de esta nueva medida produce biclusters interesantes, que no podrían haber sido generados con el uso del residuo cuadrático medio.

1. Introducción

Gracias a los avances tecnológicos actuales, recientemente se ha hecho posible la secuenciación completa de los genomas de algunas especies. Dichos genomas constituyen una enorme fuente de información que necesita ser analizada. La tecnología Microarray permite el estudio de genomas completos de forma aislada, así como de combinaciones, de forma que sea posible extraer información de relaciones entre diferentes especies [9].

A partir de los datos obtenidos mediante experimentos microarray, se construyen matrices numéricas que permitan el análisis computacional de dichos datos. Existen varias técnicas para obtener conocimiento a partir de los datos de un microarray, dependiendo de la aplicación concreta en estudio. Entre ellas cabe citar las técnicas de clustering [11], donde se persigue agrupar genes que presenten un comportamiento similar bajo todas las condiciones experimentales presentes en el microarray. Estas agrupaciones se determinan según la similitud entre un conjunto de genes, para todas las condiciones de forma simultánea [10]. Sin embargo, dependiendo del tipo de microarray que se analice, puede resultar interesante agrupar genes que muestren un comportamiento similar frente a subconjuntos del total de las condiciones. Esta idea ha motivado la reciente aparición de una línea de investigación denominada biclustering. Las técnicas de biclustering son una variación de las técnicas de clustering, que permiten la generación de biclusters en los que los genes sigan una misma tendencia frente a subconjuntos de condiciones del microarray original, y presentando, por tanto, una mayor complejidad que el clustering [7].

Cheng y Church fueron los primeros en apli-

car biclustering sobre datos genómicos [5], proponiendo para ello un algoritmo de búsqueda voraz, combinado con una medida de evaluación de biclusters, denominada residuo cuadrático medio *Mean Squared Residue* (MSR). Dicha medida ha sido utilizada e incorporada en otros trabajos de investigación, en los que se hace uso de diferentes heurísticas de búsqueda [2, 12]. Sin embargo, otros autores han basado la búsqueda de biclusters en modelos discriminativos, sin hacer uso de una medida concreta para la evaluación de los resultados [8]. Entre los distintos métodos de biclustering propuestos, son de especial interés aquellos basados en el uso de heurísticas evolutivas [4, 7], en las que frecuentemente se incorpora el valor del residuo MSR como parte principal de la función objetivo que se vaya a optimizar.

El uso de MSR como objetivo principal en la búsqueda de biclusters ha permitido la obtención de biclusters interesantes. Sin embargo, existen ciertos tipos de biclusters que el residuo no es capaz de reconocer como buenas soluciones [1]. Es por ello por lo que se propone en este trabajo una nueva medida de evaluación, denominada *Error Virtual* (*Virtual Error*, VE), que cubra las deficiencias presentadas por el residuo. Para poder comprobar la eficacia de dicha medida, VE ha sido incorporado en un algoritmo de búsqueda evolutivo, que ya había sido utilizado con anterioridad junto con el residuo MSR. Los resultados obtenidos muestran que gracias al uso de VE se obtienen biclusters interesantes que no podrían haber sido generados con el uso de MSR.

El contenido de este artículo está organizado de la siguiente manera: la sección 2 presenta la motivación principal de este trabajo, describiendo posteriormente, en la sección 3, la medida de evaluación propuesta. El algoritmo evolutivo en el que ha sido incluida dicha medida es descrito en la sección 4, mostrando un análisis de los resultados obtenidos en la sección 4. Por último, la sección 6 resume las principales conclusiones de este trabajo.

2. Motivación

La mayoría de las técnicas de biclustering basadas en la evaluación se basan en la utilización del residuo, *Mean Squared Residue* (MSR)[5]. MSR trata de cuantificar la coherencia numérica presentada por los genes y condiciones de un bicluster $\mathcal{B}$, compuesto por $\mathtt{I}$ filas y $\mathtt{J}$ columnas, y se define como sigue:

$$\mathtt{MSR}(\mathcal{B}) = \frac{1}{\mathtt{I} \cdot \mathtt{J}} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (\mathtt{b}_{ij} - \mathtt{b}_{iJ} - \mathtt{b}_{Ij} + \mathtt{b}_{IJ})^2 \quad (1)$$

donde $\mathtt{b}_{ij}$, $\mathtt{b}_{iJ}$, $\mathtt{b}_{Ij}$ y $\mathtt{b}_{IJ}$ representan el elemento de la fila $\mathtt{ith}$ y la columna $\mathtt{jth}$, la media de la fila $\mathtt{i}$-ésima y la columna $\mathtt{j}$-ésima, y la media de la submatriz, respectivamente. Cuando los niveles de expresión de los distintos genes siguen una evolución coherente a través de las condiciones contenidas en el bicluster $\mathcal{B}$, entonces, el valor el residuo será nulo (MSR($\mathcal{B}$)= 0). De forma general, cuanto menor sea el valor de MSR, mejor será la calidad del bicluster. Por lo tanto, cuando se trate de biclusters donde los genes no presenten variación alguna, o de biclusters triviales (un solo gen o condición), el valor del residuo será también muy bajo. Para evitar como buenas estas submatrices, se hace uso de otras medidas, en combinación con el residuo, como pueden ser la varianza de gen [7, 5].

Como se demuestra en [1], el residuo MSR no es una buena medida de evaluación, sobretodo cuando se aplica a distintos tipos de biclusters. En este trabajo, el autor realiza un estudio en profundidad sobre las principales características inherentes a los biclusters, definiendo formalmente los dos tipos de patrones ya expuestos con anterioridad: desplazamiento y escalado. Los distintos genes contenidos en un bicluster pueden presentar cualquiera de ellos, o más frecuentemente, ambos a la vez. El valor del residuo es útil para reconocer patrones de desplazamiento en un conjunto de genes, no siendo adecuado su uso cuando éstos presentan patrones de escalado.

La figura 1 muestra dos biclusters cuyos genes presentan la misma tendencia de expresión bajo las distintas condiciones. Cada una

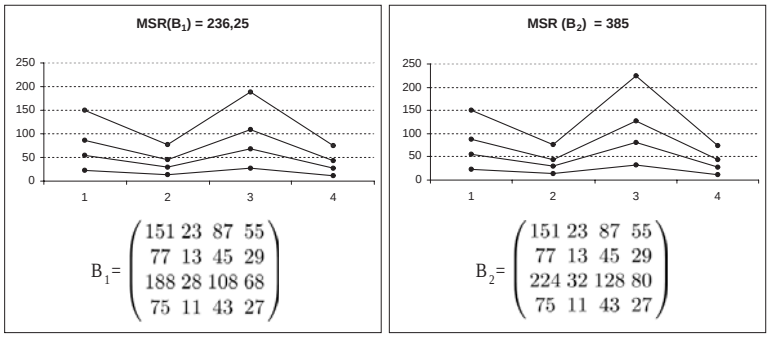

Figura 1: Ejemplos de biclusters similares con distintos valores de MSR

de las líneas del gráfico representan los niveles de expresión de cada uno de los genes, correspondiéndose con los valores numéricos que aparecen en las matrices situadas en la parte inferior de la figura, donde las columnas hacen referencia a los distintos genes y las filas a las condiciones experimentales.

A pesar de que en ambos biclusters los genes presentan el mismo comportamiento bajo todas las condiciones, el valor del residuo varía de forma significativa de uno a otro, indicando que existe una diferencia entre la bondad de ambos. Los valores de los residuos de los biclusters presentados en la figura son 236.25 y 385, respectivamente, siendo la única diferencia entre ambos la variación del valor concreto de expresión de los genes bajo la tercera condición, pero sin variar la tendencia de ninguno de ellos. Comparando ambos biclusters gráficamente, no es posible concluir cual de los dos es mejor, ya que el hecho de que todos los genes tengan un nivel de expresión superior bajo una de las condiciones no implica que su calidad sea inferior.

Todo esto sirve de motivación para el diseño de una nueva medida que permita la evaluación de biclusters obtenidos a partir de microarrays. Esta nueva medida debe poder evaluar la bondad de cada bicluster con independencia de los valores numéricos concretos contenidos en él. El objetivo es, por tanto, poder cuantificar el comportamiento de los distintos genes contenidos en cada submatriz, y es por ello por lo que se presenta una nueva medida, basada en el uso de patrones de comportamiento, y denominada *Error Virtual*, VE.

3. Error Virtual

La idea principal en la que se basa el VE es crear un patrón para cada bicluster que represente la tendencia general de todos los genes contenidos en él. Dicho patrón debe ser creado de forma que sea un buen representante del comportamiento de los genes frente a las condiciones experimentales, cuando todos ellos varíen de forma similar a través de las condiciones, con independencia de los valores numéricos concretos. VE se basa en la creación de un patrón de comportamiento para cada bicluster, por lo tanto, la calidad de dicho patrón dependerá de la forma en que éste sea creado.

En la siguiente definición se especifica cómo el patrón utilizado para el cálculo de VE es creado, a partir de un bicluster $\mathcal{B}$, formado por $\mathtt{I}$ condiciones (o filas) y $\mathtt{J}$ genes (o columnas), y donde cada uno de los elementos está representado por $\mathtt{b}_{ij}$, donde $1 \leq \mathtt{i} \leq \mathtt{I}$ y $1 \leq \mathtt{j} \leq \mathtt{J}$.

Definición 1: Patrón de comportamiento. *Dado un bicluster $\mathcal{B}$ que contenga $\mathtt{I}$ condiciones y $\mathtt{J}$ genes, se define el patrón de comportamiento como una colección de $\mathtt{I}$ elementos $\mathtt{P}_i$, donde cada uno de ellos viene dado por:* $\mathtt{P}_i = \frac{\sum_{j \in J} b_{ij}}{J}$, *donde* $\mathtt{b}_{ij} \in \mathcal{B}$, $1 \leq \mathtt{i} \leq \mathtt{I}$ *y* $1 \leq \mathtt{j} \leq \mathtt{J}$.

De esta forma, cada uno de los puntos del patrón representa un valor significativo de todos los genes frente a una condición determinada.

Una vez que el patrón ha sido creado, el objetivo es cuantificar en qué medida los distintos genes del bicluster se ajustan a él. En este sentido, se hace necesario el uso de una técnica que permita comparar de forma apropiada cada uno de los genes y el patrón. Esta técnica debe realizar un suavizado previo de los valores de expresión de cada gen, ya que el objetivo es comparar los comportamientos y no los valores numéricos concretos. Esta idea puede verse claramente analizando la figura 1, donde los genes en ambos biclusters presentan la misma conducta pero con diferentes valores concretos de expresión.

Definición 2: Estandarización. *Sea $\mathcal{B}$ un bicluster con $\mathtt{J}$ genes e $\mathtt{I}$ condiciones. Sean $\mathtt{b}_{ij}$ cada uno de los elementos de $\mathcal{B}$, $1 \leq \mathtt{i} \leq \mathtt{I}$ y $1 \leq \mathtt{j} \leq \mathtt{J}$. Se define el bicluster estandarizado de $\mathcal{B}$ como un nuevo bicluster $\mathcal{B}'$, cuyos elementos $\mathtt{b}'_{ij}$ cumplen que $\mathtt{b}'_{ij} = \frac{b_{ij}-\mu_{g_j}}{\sigma_{g_j}}, 1 \leq \mathtt{i} \leq \mathtt{I}, 1 \leq \mathtt{j} \leq \mathtt{J}$, donde $_{g_j}$ y μ_{g_j} representan la desviación estándar y la media aritmética de todos los valores de expresión del gen $\mathtt{j}$, respectivamente.*

Gracias a la estandarización se llevan a cabo dos tareas diferentes. La primera de ellas es llevar el valor de expresión de todos los genes a un mismo rango (alrededor de 0 en este caso), para poder realizar una comparación más justa. La segunda de ellas es homogeneizar los valores de expresión de cada gen, modificando de esta forma sus valores bajo todas las condiciones, y suavizando su representación gráfica.

Es importante tener en cuenta que para poder comparar los valores de expresión génica a los valores contenidos en el patrón de comportamiento creado, todos ellos deben pertenecer al mismo rango de valores. Por lo tanto, el patrón de comportamiento debe ser también estandarizado, creando de esta forma un nuevo patrón llamado patrón virtual. Este proceso se muestra en la ecuación 2, donde $\mathtt{P}_i$ denota el valor del patrón para la condición i, y donde $\overline{\mathtt{P}}$, $_P$ denotan la media y la desviación de los valores del patrón, respectivamente.

$$\mathtt{P}'_i = \frac{\mathtt{P}_i - \overline{\mathtt{P}}}{_P} \tag{2}$$

Definición 3: Error Virtual. *Dado un bicluster $\mathcal{B}$ con $\mathtt{I}$ condiciones y $\mathtt{J}$ genes, y un patrón $\mathtt{P}$ que contiene $\mathtt{I}$ valores, se define VE como la media de las diferencias numéricas entre cada gen estandarizado y los valores del patrón estandarizado para cada condición:*

$$\mathtt{VE}(\mathcal{B}) = \frac{1}{\mathtt{I} \cdot \mathtt{J}} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (\mathtt{b}'_{ij} - \mathtt{P}'_i) \tag{3}$$

VE($\mathcal{B}$) se corresponde con la medida propuesta, siendo los biclusters con valores más bajos de VE considerados de mejor calidad que aquellos que tengan un valor más alto. Esto se debe al hecho de que VE calcula las diferencias entre los genes estandarizados y el patrón estandarizado, por lo tanto, cuando más parecidos sean los genes, menor será el valor de la medida VE.

Es importante destacar que el comportamiento definido por patrones de desplazamiento en biclusters no incrementa el valor de VE, ya que la estandarización de los genes permite que VE compare el comportamiento en el mismo rango de valores. En el caso de comportamiento definido por patrones de escalado, éste tiene un mínimo efecto en VE, ya que la estandarización realiza un suavizado de los valores que disminuye las diferencias numéricas entre los distintos genes que presenten un comportamiento escalado. A modo de ejemplo, los biclusters representados en la figura 1 tienen valores de VE cercanos a cero ($\mathtt{VE}(\mathtt{B}_1) = 2{,}77 \times 10^{-17}$ y $\mathtt{VE}(\mathtt{B}_2) = -1{,}39 \times 10^{-17}$). Estos valores indican que VE considera ambos biclusters de una calidad similar. El nombre de esta medida (VE, *Error Virtual*) se debe a que el error no se calcula usando los genes originales, sino en base a unos genes "virtuales", una vez que los datos originales han sido estandarizados.

En general, esta nueva medida proporciona un valor para cada bicluster, que permite

cuantificar las similitudes entre los genes, comparando sus comportamientos a los de un patrón creado a partir de ellos. Esta comparación se lleva a cabo de forma que las tendencias de desplazamiento y escalado sean mínimamente penalizadas, mientras que las diferencias de comportamiento entre los genes incrementan de forma notable el valor de VE.

4. VE en una heurística evolutiva

Para poder comprobar la eficacia de VE como medida para evaluar y establecer un valor indicativo de la calidad de biclusters, dicha medida ha sido incorporada en un algoritmo evolutivo de biclustering (SEBI) [7]. Para ello se ha modificado dicho algoritmo evolutivo, de forma que VE aparezca dentro de la función objetivo, como se explica a continuación.

SEBI se basa en una estrategia de cubrimiento secuencial, utilizando para ello un algoritmo evolutivo, denominado EBI (*Evolutionary BIclustering*), el cual es ejecutado n veces, siendo n un parámetro definido por el usuario. EBI parte de una matriz de expresión génica como entrada, y devuelve un bicluster encontrado, que es guardado en una lista denominada *Results*, siendo EBI llamado sucesivamente para la búsqueda de más biclusters en la matriz original.

Para controlar el solapamiento cometido entre los distintos biclusters generados, se asocia un peso a cada uno de los elementos de la matriz original, de forma que cuando se genere un bicluster, dichos pesos son ajustados. El peso de un elemento depende del número de biclusters de la lista *Results* que contengan a ese elemento. Cuanto mayor sea el número de biclusters que cubran a un elemento, mayor será el peso para dicho elemento [7].

En [7], la evaluación de cada individuo (o bicluster) X venía dada por la ecuación 4,

$$\mathtt{f(X)} = \frac{\mathtt{MSR(X)}}{} + \frac{1}{\mathtt{rv(X)}} + \mathtt{wd} + \mathtt{penalty} \tag{4}$$

donde MSR(X) representa el residuo cuadrático medio de X , es el límite para MSR definido por el usuario, rv(X) es la varianza de fila de X , wd es usado para penalizar los biclusters de menor volumen, y penalty es la suma de los pesos asignados a cada elemento de la matriz de expresión pertenecientes al bicluster X . El objetivo es minimizar el valor de la función objetivo, por lo que se persigue encontrar biclusters con un valor para el residuo cuadrático medio menor que el límite establecido , con un volumen que sea el más grande posible, una varianza de fila relativamente alta, y minimizando el efecto de solapamiento entre biclusters.

En la versión de EBI adaptada para ser usada junto con VE, se ha modificado la función objetivo definida en 4 de la siguiente manera:

$$\mathtt{f(X)} = \mathtt{VE(X)} + \mathtt{wd} + \mathtt{penalty} \tag{5}$$

donde VE(X) es el valor de VE para X , wd y penalty son definidos de la misma forma que en [7], pero adaptados para ser usados con VE. Esta nueva función objetivo tiene que ser también minimizada, ya que se buscan biclusters con valores bajos de VE, gran volumen y mínimo solapamiento entre los biclusters generados. En este caso, no se hace uso de la varianza de fila, a diferencia del caso anterior 4. Esto es debido a que con el uso de VE no se necesita este sumando para evitar biclusters triviales, como ocurría haciendo uso del residuo.

Al igual que en [7], la población inicial consiste en un conjunto de biclusters que contienen un solo gen de la matriz original. La selección de padres se realiza por torneo, y se recombinan mediante un operador de cruce, con una probabilidad p_c (con un valor por defecto 0.9), mientras que los hijos producidos son mutados con una probabilidad p_m (con un valor por defecto de 0.1). Se aplica elitismo (copia del mejor individuo a la siguiente generación) con una probabilidad p_e (con un valor por defecto 0.75), devolviendo al final de todo el proceso el mejor individuo producido, según la función objetivo.

Cada individuo de la población representa un bicluster, codificado mediante una cadena binaria de longitud N + M , donde N y M son el número de filas (condiciones) y columnas (genes) de la matriz original, respectivamente. Cada uno de los primeros N bits están relacio-

nados con las filas o condiciones, en el mismo orden en que aparecen en la matriz, de forma que si uno de esos bits tiene un valor igual a 1, dicha condición pertenece al individuo (bicluster) correspondiente. De la misma manera, los restantes `M` bits se corresponden con las columnas o genes de la matriz original, definiéndose su inclusión en el bicluster con los valores establecidos a 1.

5. Pruebas experimentales

Para poder comprobar experimentalmente el uso de VE en una heurística de búsqueda evolutiva, se han realizado experimentos con dos bases de datos muy conocidas. La primera de ellas contiene datos relativos al ciclo celular de la levadura *Saccharomyces cerevisiae* [6], almacenados en un microarray que contiene 2884 genes y 17 condiciones. El segundo conjunto de datos utilizado contiene datos de expresión de células humanas *human B-cells* [3], formado por 4026 genes y 96 condiciones.

Para poder comparar los resultados generados con los ya obtenidos en trabajos anteriores (haciendo uso de MSR [7]), en este caso se ha utilizado la misma configuración de parámetros. Concretamente, el tamaño de la población es de 200 individuos, iterando sobre un total de 100 generaciones. La probabilidad de cruce utilizada es de 0.85, mientras que la probabilidad de mutación es de 0.2. Asimismo, el número de biclusters a obtener para cada conjunto de datos probados es de 100.

La figura 2 muestra seis de los cien biclusters encontrados en el conjunto de datos de la levadura (ver la tabla 1 con los resultados numéricos concretos). El bicluster etiquetado como $\texttt{yeast}_1$ es el obtenido tras la primera iteración del algoritmo. Como se puede apreciar en su gráfica, es un bicluster interesante, aunque su valor del residuo es alto: 535.8. Sin embargo, el valor de la medida VE es bajo: 0.38. Además, es importante tener en cuenta que el primer bicluster generado haciendo uso de VE es un buen bicluster, cosa que no ocurría en los trabajos previos basados en MSR. [5, 7].

En general, en la figura se puede apreciar cómo en todos los biclusters los genes presentan un comportamiento similar bajo todas las condiciones seleccionadas, siendo el valor de VE inferior a 0.38 para todos ellos. En particular, es posible encontrar casos en los que uno o varios genes aparezcan distantes del resto, aún cuando muestren el mismo comportamiento. Por ejemplo, el bicluster $\texttt{yeast}_{44}$ resulta interesante ya que se diferencian tres genes en la parte superior de la gráfica que se encuentran distanciados del resto. Esto pone de manifiesto que VE no es vulnerable a comportamientos de desplazamiento o escalado entre los genes. Este tipo de biclusters son difíciles de encontrar utilizando MSR como medida de evaluación, ya que su valor del residuo es mayor [1].

Con respecto al volumen de los biclusters obtenidos, la mayoría de ellos contienen un gran número de condiciones, al igual que en la versión de SEBI presentada utilizando el residuo MSR como término principal de la función objetivo [7]. Sin embargo, el número de genes es mayor en este caso, esto es, trabajando con VE como término principal de la función a minimizar. Por lo tanto, el uso de VE permite incluir un mayor número de genes en los biclusters, sin afectar a la calidad de los mismos.

El conjunto de datos de células humanas tiene un mayor volumen y posee datos más complejos de analizar que el de la levadura, ya comentado. Por lo tanto, también resulta más difícil extraer buenos biclusters con un valor bajo de VE a partir de él. En la figura 3 se muestran seis de los cien biclusters obtenidos a partir de este conjunto de datos (ver la tabla 1 para comprobar los resultados numéricos).

El bicluster $\texttt{hum an}_1$, que es obtenido por SEBI y VE en la primera ejecución del algoritmo evolutivo, es un bicluster interesante, tal y como se muestra en la figura, a diferencia del primer bicluster obtenido con el uso del residuo MSR. Aunque el bicluster $\texttt{hum an}_1$ presenta un bajo valor para VE (0.57), su valor del residuo es muy alto (7173.5), lo que refuerza la conclusión de que con el uso de VE es posible encontrar buenos biclusters en las primeras iteraciones.

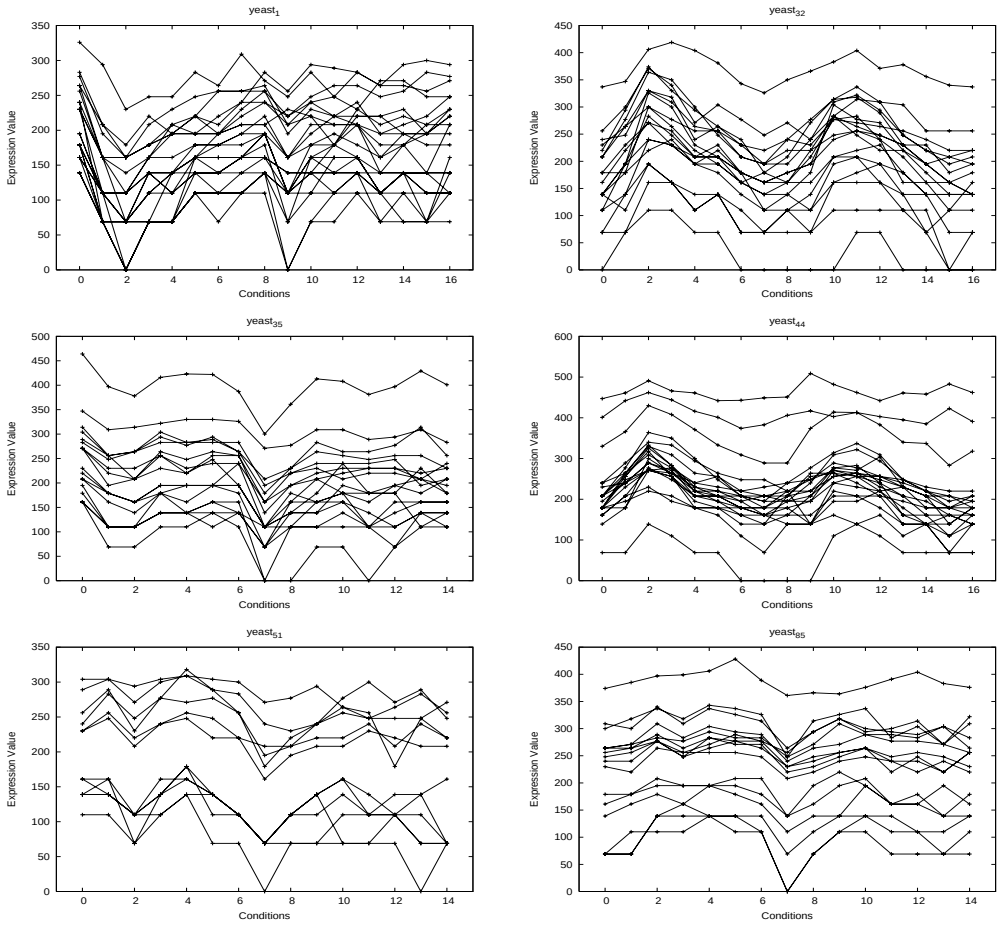

Figura 2: Biclusters obtenidos de la levadura con VE.

En cuanto a las características propias de los biclusters generados, todos ellos presentan conjuntos de genes con tendencias similares, estando contenidos en rangos de valores similares, debido al conjunto de datos analizado. Además, los biclusters obtenidos con VE no son sensibles a comportamientos de escalado y desplazamiento, como puede verse por ejemplo en el bicluster $\texttt{hum an}_{52}$, donde los genes presentan diferentes niveles de expresión de una condición a otra, variando todos ellos en una magnitud diferente. Por último, el número de genes y condiciones (volumen) son similares a aquellos producidos haciendo uso del residuo.

La tabla 1 resume algunos de los resultados obtenidos para los dos conjuntos de datos analizados, y mostrados en las figuras 2 y 3. La tabla superior se corresponde a los resultados obtenidos para la levadura, mientras que la tabla inferior muestra los de células humanas. Para cada tabla, la primera columna indica el nombre del bicluster, la segunda indica el valor VE y la tercera el valor MSR para ese bicluster. Las últimas dos columnas muestran el número de genes y condiciones presentes en cada bicluster, respectivamente.

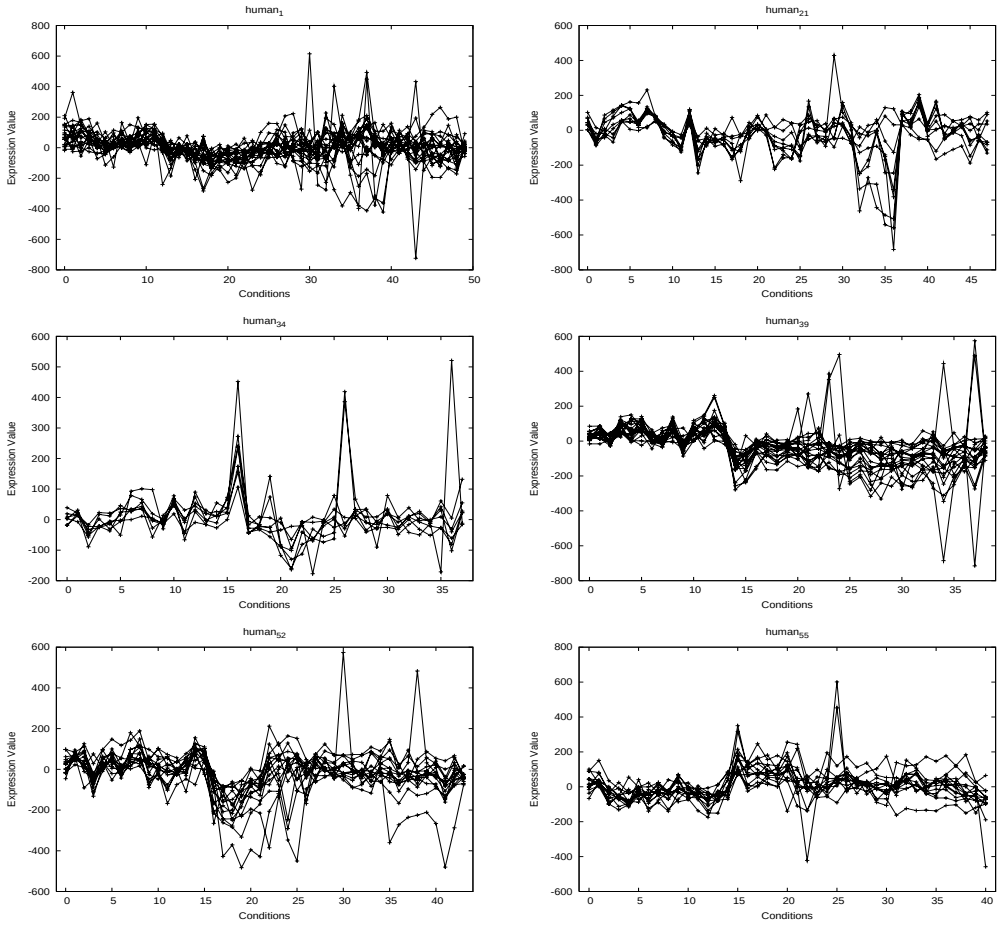

Figura 3: Biclusters obtenidos de células humanas con VE.

Como puede verse en la tabla 1, los biclusters obtenidos presentan un valor bajo de VE, a la vez que tienen un valor alto de MSR para los dos conjuntos de datos. Teniendo en cuenta que las representaciones de dichos biclusters son interesantes, éstos pueden ser rechazados cuando se utilicen técnicas basadas en el uso del residuo como principal objetivo a tener en cuenta. Por ejemplo, muchas de las técnicas basadas en el residuo establecen un límite para el MSR a partir del cual se rechaza cualquier bicluster que lo supere, no admitiéndose como una buena solución. Un valor típico de dicho límite es de 300, por lo que la mayoría de los biclusters presentados aquí habrían sido rechazados, aún siendo unas buenas soluciones.

6. Conclusiones

En este trabajo se presenta una nueva medida para la evaluación de la calidad de biclusters obtenidos a partir de datos de microarrays. Dicha medida recibe el nombre de *Error Virtual* (VE), y está basada en el concepto de patrones de comportamiento. La mayoría de las técnicas de biclustering existentes utilizan como medi-

Levadura				
Bicluster	VE	MSR	#Gen	#Cond.
$yeast_1$	0.38	535.8	23	17
$yeast_{32}$	0.29	408.9	19	17
$yeast_{35}$	0.28	380.6	18	15
$yeast_{44}$	0.30	583.5	21	17
$yeast_{51}$	0.34	346.7	12	15
$yeast_{85}$	0.36	232.1	16	15

Células Humanas				
Bicluster	VE	MSR	#Gen	#Cond.
$human_1$	0.57	7173.5	21	50
$human_{21}$	0.39	6405.4	9	48
$human_{34}$	0.43	3278.8	7	38
$human_{39}$	0.44	5786.1	21	39
$human_{52}$	0.42	5660.7	15	44
$human_{55}$	0.46	4069.5	14	41

Cuadro 1: Información de los biclusters en las figuras 2 y 3.

da de evaluación el residuo cuadrático medio (MSR), medida muy conocida pero que ha sido demostrado que no es válida para reconocer cierto tipo de biclusters.

Para poder comprobar la validez de la medida propuesta, ésta ha sido utilizada como término principal de la función objetivo en un algoritmo evolutivo. Dicho algoritmo evolutivo ha sido aplicado a dos bases de datos reales muy conocidas, extrayendo de esta aplicación las siguientes conclusiones:

El uso de VE permite la obtención de biclusters con valores de VE muy bajos, que se corresponden con buenos biclusters. Sin embargo, esos mismos biclusters no podrían haberse obtenidos con el uso del residuo (MSR), ya que su evaluación con esta otra medida da como resultados valores que superan el límite establecido para el MSR. Además, con el uso de VE no es necesario tener en cuenta la varianza de los genes, como ocurría con el residuo. Por último, VE permite la obtención de buenos biclusters sin tener en cuenta factores de desplazamiento o escalado entre los genes, siempre que sigan la misma tendencia o comportamiento.

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] J. S. Aguilar-Ruiz, D. S. Rodriguez, and D. A. Simovici. Biclustering of gene expression data based on local nearness. In *1Proceedings of EGC 2006,*, pages 681–692, Lille, France, 2006.

[3] A. A. Alizadeh, M. B. Eisen, R. E. Davis, and et al. Distinct types of diffuse large b-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[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. IEEE, 2004.

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

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

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

[8] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:136–144, 2002.

[9] C. Tilstone. Dna microarrays: Vital statistics. *Nature*, 424:610–612, 2003.

[10] J. Wang, J. Delabie, H. C. Aasheim, E. Smeland, and O. Myklebost. Clustering of the som easily reveals distinct gene expression patterns: results of a reanalysis of lymphoma study. *BMC Bioinformatics*, 3(36):doi: 10.1186/1471–2105–3–36, 2002.

[11] R. Xu and D. C. W. II. Survey of clustering algorithms. *IEEE Transactions on Neural Networks*, 16(3):645–678, 2005.

[12] J. Yang, H. Wang, W. Wang, and P. S. Yu. An improved biclustering method for analyzing gene expression profiles. *International Journal on Artificial Intelligence Tools*, 14:771–790, 2005.