VE^t: Una medida para la detección de patrones de desplazamiento y escalado en biclusters

Beatriz Pontes
Dept. de Lenguajes y Sistemas Informáticos
ETS Ingeniería Informática
Universidad de Sevilla
41012 Sevilla
bepontes@us.es

Raúl Giráldez
Escuela Politécnica Superior
Univ. Pablo de Olavide
Ctra. Utrera Km. 1
41013 Sevilla
giraldez@upo.es

Jesús Aguilar
Escuela Politécnica Superior
Univ. Pablo de Olavide
Ctra. Utrera Km. 1
41013 Sevilla
aguilar@upo.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 obtenidas. El uso de MSR permite obtener biclusters interesantes, sin embargo, algunos trabajos han demostrado que dicha medida no es válida para reconocer determinados tipos de biclusters. VE(*Error Virtual*) es una medida de evaluación de biclusters que fue desarrollada con el fin de evitar los inconvenientes de MSR. Frente a MSR, que solamente permite detectar patrones de desplazamiento, VE es capaz de distinguir también patrones de escalado en biclusters, aunque no simultáneamente al desplazamiento. En este trabajo realizamos un estudio sobre una variación de VE que permite detectar patrones de desplazamiento y escalado simultáneamente en biclusters.

1. Introducción

Los avances tecnológicos actuales hacen 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 es posible extraer información de relaciones entre diferentes especies [16].

A partir de los datos obtenidos mediante experimentos microarray, se construyen matrices numéricas que permiten 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.

Las técnicas de biclustering son una variación de las técnicas de clustering [17] que permiten agrupar genes que muestren un comportamiento similar frente a subconjuntos del total de las condiciones. Son especialmente interesantes aquellos biclusters en los que los genes siguen una misma tendencia frente a subconjuntos de condiciones del microarray original, y presentando, por tanto, una mayor complejidad que el clustering tradicional [11].

Cheng y Church fueron los primeros en aplicar 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, 19]. 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 [15]. Entre los distintos métodos de biclustering propuestos, son de especial interés

aquellos basados en el uso de heurísticas evolutivas [4, 9, 12], en las que frecuentemente se incorpora el valor del MSR como parte principal de la función objetivo que se va a optimizar. MSR ha sido utilizado recientemente como medida de evaluación en conjunción con heurísticas como la optimización basada en enjambres de partículas [10] o sistemas inmunes artificiales [8].

Aunque el uso de MSR permite la obtención de biclusters interesantes, existen ciertos tipos de biclusters que no se reconocen como buenas soluciones usando dicha medida [1]. Es por ello por lo que en trabajos anteriores hemos propuesto una medida de evaluación alternativa, llamada *Error Virtual* (*Virtual Error*, VE) [13]. VE cubre alguna de las deficiencias del residuo, aunque sigue sin ser capaz de encontrar algunos tipos de biclusters. Por ello en este trabajo se propone una variación de VE, denominada *Error Virtual Traspuesto* (VE^t), capaz de identificar tipos de biclusters que ni MSR ni VE reconocen como interesantes.

2. Patrones de comportamiento en expresión génica

Los biclusters agrupan genes cuyos valores de expresión siguen una tendencia similar con respecto a todas las condiciones que forman parte de él. Desde este punto de vista, es posible identificar biclusters en matrices de expresión haciendo uso de medidas basadas en patrones de comportamiento entre genes. En [1] se presenta un estudio en profundidad sobre las principales características inherentes a los biclusters, defiendo formalmente dos tipos de patrones: desplazamiento y escalado.Sus definiciones se basan en relaciones numéricas entre los valores de expresión de los genes en un bicluster.

Sea $\mathcal{B}$ un bicluster compuesto por I condiciones y J genes, de manera que cada elemento perteneciente al bicluster vendrá representado por $b_{ij} \in \mathcal{B}$. Así, un bicluster $\mathcal{B}$ sigue un patrón de desplazamiento perfecto cuando sus valores b_{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. Diremos que β_i es el *coeficiente de desplazamiento* para la condición i. De este modo, podemos representar los valores de expresión como $b_{ij} = \pi_j + \beta_i$.

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, que denominamos *coeficiente de escalado*. Así, decimos que un bicluster sigue un patrón de escalado perfecto cuando sus valores se pueden representar mediante la expresión $b_{ij} = \pi_j \times \alpha_i$.

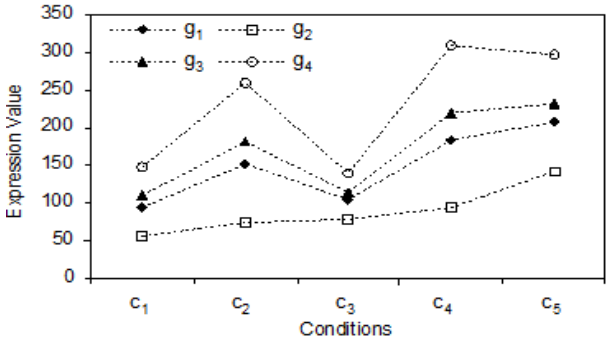
Los patrones de desplazamiento o escalado no suelen estar presentes de manera aislada en datos reales de expresión génica. Los valores de expresión en biclusters suelen contener los dos tipos de patrones de forma simultánea. En ese caso, los valores de expresión se obtienen haciendo uso de los dos coeficientes explicados anteriormente. De esta forma, podemos decir que los valores de un bicluster con patrones de desplazamiento y escalado simultáneos siguen siguiente expresión, que combina las anteriormente expuestas: $b_{ij} = \pi_j \times \alpha_i + \beta_i$.

En la figura 1 se muestra un ejemplo de un bicluster con ambos patrones. Como puede verse en la figura, identificar ambos tipos de patrones directamente sobre la gráfica no es tan inmediato como en los casos anteriores.

A simple vista, podemos decir que los genes g_1, g_3 y g_4 presentan un comportamiento similar, aunque el gen g_4 varia en la última condición. Sin embargo, el gen g_2 sigue una tendencia ascendente a lo largo de las todas las condiciones, al contrario que el resto de genes, cuyo comportamiento va variando. Esto ocurre cuando los coeficientes de desplazamiento β_i son del mismo orden que $\pi_j \times \alpha_i$. También es interesante observar que el desplazamiento hace que los genes g_1, g_2 y g_3 cambien significativamente su valor para la última condición. El coeficiente de desplazamiento para dicha condición es 83, aproximadamente igual a la expresión $\pi_j \times \alpha_i$ para los genes.

3. VE^t: Error Virtual Traspuesto

Virtual Error (VE) es una medida de evaluación diseñada para identificar patrones de



$$\mathcal{B} = \begin{pmatrix} 95 & 56 & 110 & 149 \\ 152 & 74 & 182 & 260 \\ 104 & 78 & 114 & 140 \\ 185 & 94 & 220 & 311 \\ 208 & 143 & 233 & 298 \end{pmatrix} = \begin{pmatrix} 25\times 3+20 & 12\times 3+20 & 30\times 3+20 & 43\times 3+20 \\ 25\times 6+2 & 12\times 6+2 & 30\times 6+2 & 43\times 6+2 \\ 25\times 2+54 & 12\times 2+54 & 30\times 2+54 & 43\times 2+54 \\ 25\times 7+10 & 12\times 7+10 & 30\times 7+10 & 43\times 7+10 \\ 25\times 5+83 & 12\times 5+83 & 30\times 5+83 & 43\times 5+83 \end{pmatrix}$$

$$\begin{array}{rccccc} & \pi_1 & \pi_2 & \pi_3 & \pi_4 & \\ \{\pi_j\} = & \{25 & 12 & 30 & 43\} & \\ & \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \alpha_5 \\ \{\alpha_i\} = & \{3 & 6 & 2 & 7 & 5\} \\ & \beta_1 & \beta_2 & \beta_3 & \beta_4 & \beta_5 \\ \{\beta_i\} = & \{20 & 2 & 54 & 10 & 83\} \end{array}$$

Figura 1: Bicluster con patrones de desplazamiento y escalado.

comportamiento en biclusters [13]. Frente a MSR, que solamente permite identificar patrones de desplazamiento, VE es capaz de detectar también patrones de escalado, aunque no de forma simultánea.

La idea principal en la que se basa 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, denominado *Gen virtual*, y en la comparación de dicho *Gen virtual* con el resto de los genes presentes en el bicluster.

Los biclusters con valores más bajos de VE son 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.

Usando VE como objetivo en un entorno evolutivo [13] es posible encontrar biclusters interesantes en microarrays que con las medidas disponibles hasta la fecha no seria posible encontrar. Sin embargo, aunque el comportamiento de VE hace que su valor se encuentre cercano a 0 para biclusters con patrones de desplazamiento y escalado perfecto, del orden de 10^{-15} [14], no se ha podido demostrar analíticamente que VE permita reconocer ambos patrones simultáneamente.

En este trabajo presentamos una versión mejorada de VE, denominada VE^t (*Error Virtual Traspuesto*), y que permite encontrar patrones de desplazamiento y escalado de forma simultánea en biclusters. La motivación de esta variación de VE parte de [6], donde los autores aplican varios tipos de transformaciones numéricas para ver sus repercusiones a la hora de detectar patrones de desplazamiento y escalado usando distintas métricas.

El cálculo de VE^t se realiza de una forma similar al de VE pero considerando la matriz

del bicluster traspuesta. De forma conceptual, VE^t consiste en la creación de una *Condición virtual*, en vez de un *Gen virtual* como en el caso de VE.

En la siguiente definición se especifica cómo el patrón utilizado para el cálculo de VE^t es creado, a partir de un bicluster $\mathcal{B}$.

Definición 1: Condición virtual o patrón de comportamiento. *Dado un bicluster $\mathcal{B}$ que contenga I condiciones y J genes, se define la condición virtual como una colección de J elementos P_j, donde cada uno de ellos viene dado por: $P_j = \frac{\sum_{i \in I} b_{ij}}{I}$, donde $b_{ij} \in \mathcal{B}, 1 \leq i \leq I$ y $1 \leq j \leq J$.*

De esta forma, cada punto del patrón representa un valor significativo de todas las condiciones frente a un gen determinado.

Una vez que el patrón ha sido creado, el objetivo es cuantificar en qué medida las distintas condiciones del bicluster se ajustan a él. En este sentido, se hace necesario el uso de una técnica que permita comparar de forma apropiada cada una de las condiciones y el patrón o condición virtual. Esta técnica debe realizar un suavizado previo de los valores de expresión de cada condición, ya que el objetivo es comparar los comportamientos y no los valores numéricos concretos.

Definición 2: Estandarización. *Sea $\mathcal{B}$ un bicluster según las premisas anteriores. Se define el bicluster estandarizado de $\mathcal{B}$ como un nuevo bicluster $\mathcal{B}'$, cuyos elementos b'_{ij} cumplen que $b'_{ij} = \frac{b_{ij} - \mu_{c_i}}{\sigma_{c_i}}$, donde σ_{c_i} y μ_{c_i} representan la desviación estándar y la media aritmética de todos los valores de expresión de la condición i, respectivamente.*

Para poder comparar los valores de todas las condiciones con los valores contenidos en el patrón de comportamiento creado, todos 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 estandarizado*. Este proceso se muestra en la ecuación 1, donde P_j denota el valor del patrón para el gen j, y donde $\overline{P}$, σ_P denotan la media y la desviación de los valores del patrón, respectivamente.

$$P'_j = \frac{P_j - \overline{P}}{\sigma_P} \tag{1}$$

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

$$VE^t(\mathcal{B}) = \frac{1}{I \cdot J} \sum_{i=1}^{i=I} \sum_{j=1}^{j=J} (b'_{ij} - P'_j) \tag{2}$$

3.1. Análisis

En esta sección se incluyen pruebas formales que demuestran que el valor de VE^t para biclusters con patrones perfectos (de desplazamiento, escalado o ambos simultáneamente) es igual a cero.

Teorema 1. *El valor de VE^t para un bicluster con patrón de desplazamiento perfecto es igual a cero.*

Demostración: Sea $\mathcal{B}$ un bicluster con patrón de desplazamiento perfecto, podemos expresar sus elementos como $b_{ij} = \pi_j + \beta_i$. Aplicando dos propiedades algebraicas simples[1], es posible reescribir la media y desviación de cada condición c_i como:

$$\mu_{c_i} = \mu_\pi + \beta_i \quad ; \quad \sigma_{c_i} = \sigma_\pi$$

Con estos resultados se obtienen los valores de b_{ij} estandarizados:

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{c_i}}{\sigma_{c_i}} = \frac{\pi_j + \beta_i - \mu_\pi - \beta_i}{\sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi}$$

Haciendo uso de las mismas propiedades[1] se obtienen la media y la desviación estándar para el patrón:

$$\mu_\rho = \mu_\pi + \mu_\beta \quad ; \quad \sigma_\rho = \sigma_\pi$$

[1] Siendo $f(x) = g(x) \times c_1 + c_2$, las propiedades relacionadas con la media aritmética ($\mu_{f(x)}$) y la desviación estándar ($\sigma_{f(x)}$) de $f(x)$ son: $\mu_{f(x)} = \mu_{g(x)} \times c_1 + c_2$ y $\sigma_{f(x)} = \sigma_{g(x)} \times c_1$.

Por último, los valores estandarizados del patrón son:

$$\hat{\rho}_j = \frac{\rho_j - \mu_\rho}{\sigma_\rho} = \frac{\pi_j + \mu_\beta - \mu_\pi - \mu_\beta}{\sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi} = \hat{b}_{ij}$$

Este resultado demuestra que para aquellos biclusters que estén representados por un patrón de desplazamiento perfecto, los valores de la condición virtual o patrón coinciden con los valores del bicluster, una vez estandarizados. Por lo tanto, el valor de VE^t para estos biclusters será cero. ■

Teorema 2. *El valor de VE^t para un bicluster con patrón de escalado perfecto es igual a cero.*

Demostración: Sea $\mathcal{B}$ un bicluster con patrón de desplazamiento perfecto, podemos expresar sus elementos como $b_{ij} = \pi_j \times \alpha_i$. Siguiendo el mismo razonamiento que en el teorema anterior, la media y desviación de cada condición c_i son:

$$\mu_{c_i} = \alpha_i \times \mu_\pi \quad ; \quad \sigma_{c_i} = \alpha_i \times \sigma_\pi$$

Haciendo uso de estos resultados estandarizamos b_{ij}:

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{c_i}}{\sigma_{c_i}} = \frac{\pi_j \times \alpha_i - \alpha_i \times \mu_\pi}{\alpha_i \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi}$$

Calculamos ahora la media y desviación de los valores del patrón:

$$\mu_\rho = \mu_\pi \times \mu_\alpha \quad ; \quad \sigma_\rho = \mu_\alpha \times \sigma_\pi$$

Y por último los valores estandarizados del patrón:

$$\hat{\rho}_j = \frac{\rho_j - \mu_\rho}{\sigma_\rho} = \frac{\pi_j \times \mu_\alpha - \mu_\pi \times \mu_\alpha}{\mu_\alpha \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi} = \hat{b}_{ij}$$

Al igual que en la demostración anterior, los valores del patrón estandarizado coinciden con los valores de las condiciones estandarizadas, por lo que el valor de VE^t para biclusters con patrones de escalado perfecto será cero. ■

Teorema 3. *El valor de VE^t para un bicluster con patrones de desplazamiento y de escalado perfecto es igual a cero.*

Demostración: Sea $\mathcal{B}$ un bicluster con patrón de desplazamiento y escalado perfecto, podemos expresar sus elementos como $b_{ij} = \pi_j \times \alpha_i + \beta_i$. Siguiendo el mismo razonamiento que en los dos teoremas anteriores, podemos expresar la media y desviación de cada condición c_i como:

$$\mu_{c_i} = \alpha_i \times \mu_\pi + \beta_i \quad ; \quad \sigma_{c_i} = \alpha_i \times \sigma_\pi$$

Haciendo uso de estos resultados estandarizamos b_{ij}:

$$\hat{b}_{ij} = \frac{b_{ij} - \mu_{c_i}}{\sigma_{c_i}} = \frac{\pi_j \times \alpha_i + \beta_i - \alpha_i \times \mu_\pi + \beta_i}{\alpha_i \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi}$$

Calculamos ahora la media y desviación de los valores del patrón:

$$\mu_\rho = \mu_\pi \times \mu_\alpha + \mu_\beta \quad ; \quad \sigma_\rho = \mu_\alpha \times \sigma_\pi$$

Y por último los valores estandarizados del patrón:

$$\hat{\rho}_j = \frac{\rho_j - \mu_\rho}{\sigma_\rho} = \frac{\pi_j \times \mu_\alpha + \mu_\beta - \mu_\pi \times \mu_\alpha - \mu_\beta}{\mu_\alpha \times \sigma_\pi} = \frac{\pi_j - \mu_\pi}{\sigma_\pi} = \hat{b}_{ij}$$

Al igual que en las demostraciones anteriores, los valores del patrón estandarizado coinciden con los valores de las condiciones estandarizadas, por lo que el valor de VE^t para biclusters con patrones de desplazamiento y escalado perfecto será cero.■

Gracias a estos teoremas podemos decir que VE^t es una medida que supera en efectividad a todas las medidas propuestas hasta la fecha para la evaluación de biclusters. Mientras que MSR solamente es capaz de detectar patrones de desplazamiento y VE permite también encontrar patrones de escalado, aunque pero no simultáneamente, VE^t es capaz de reconocer biclusters que contengan ambos patrones de forma separada o simultánea.

4. Análisis Experimental

En este apartado presentamos algunos estudios realizados sobre el comportamiento de

la nueva medida VE^t. En primer lugar, se exponen algunos ejemplos a partir de los cuales es posible deducir que el valor de VE^t será menor para aquellos biclusters en los que sus datos presenten un comportamiento más parecido a un patrón de desplazamiento y escalado perfecto.

Para poder estudiar el comportamiento de VE^t cuando un bicluster no presenta un patrón perfecto, añadimos un término aditivo χ_{ij} a la expresión de patrones simultáneos. Este nuevo término representa el error cometido al aproximar los datos del bicluster por un patrón de desplazamiento y escalado.

$$b_{ij} = \pi_j \times \alpha_i + \beta_i + \chi_{ij} \qquad (3)$$

Haciendo uso de esta ecuación, es posible estudiar las variaciones que se producen en el valor de VE^t en relación con los valores de χ_{ij}. Sin embargo, es importante destacar la complejidad de dicho estudio, debido a la multitud de casos posibles en relación a la distribución y rango de los valores χ_{ij} a lo largo de la matriz de datos.

Existen dos casos para los cuales el valor de VE^t no se verá afectado por poderse interpretar que los valores de χ_{ij} se incluyen en la definición de la ecuación 3. Dichos casos se dan cuando los valores de son constantes para toda la matriz o bien son constantes por filas (condiciones). En ambos casos se podría reformular la ecuación anterior y hacerla coincidir con la expresión de desplazamiento y escalado perfecto, donde el error pasa a formar parte del término β_i).

Para poder ver el comportamiento de VE^t en el caso mas genérico de los valores χ_{ij}, hemos añadido errores aleatorios dentro de un rango al bicluster de la figura 1 que representa un bicluster con patrón de desplazamiento y escalado perfecto. En concreto se han generado 100 biclusters basados en el ejemplo de la figura 1 y con errores aleatorios de tipo real comprendidos en 100 rangos de amplitud distinta. La diferencia de amplitud entre dos rangos consecutivos es de 0,1. Es decir, se han generado 100 biclusters con errores aleatorios entre 0 y 0.1, 100 biclusters con errores comprendidos entre 0 y 0.2, 100 entre 0 y 0.3, etcétera, hasta llegar a una amplitud del error entre 0 y 10.

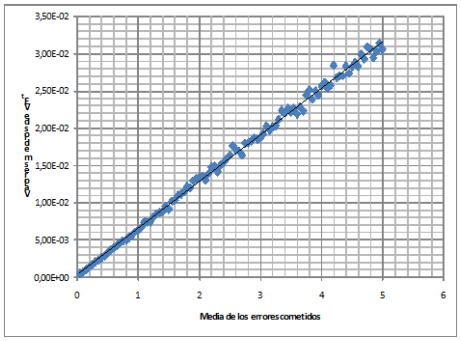


Figura 2: Comportamiento de VE^t frente al error.

Para cada grupo de 100 biclusters con el mismo rango de errores se ha calculado el valor medio de VE^t, y a partir de dichos valores hemos obtenido la gráfica de la figura 2. El eje de abscisa representa la media de los errores cometidos para cada amplitud, esto es, el valor intermedio de cada rango de errores. El eje de ordenada se corresponde con los valores medios de VE^t de cada grupo. Esta gráfica permite apreciar claramente que los valores de VE^t presentan un decrecimiento lineal cuanto menor es el error existente en los biclusters.

4.1. VE^t en datos de expresión génica

En este apartado presentamos algunos biclusters ya obtenidos a partir de datos génicos reales, comentado los valores medios de MSR, VE y VE^t.

Los biclusters que se analizan en este apartado fueron obtenidos usando como heurística de búsqueda un algoritmo evolutivo cuya función objetivo estaba basada en VE [13]. Se analizaron dos bases de datos: la primera de ellas, *Saccharomyces cerevisiae* (en adelante *Yeast*), contiene datos relativos al ciclo celular de la levadura [7]; la segunda, *human B-cells* (en adelante *Human*), contiene datos de expresión de células humanas [3].

Para los 100 biclusters obtenidos anteriormente de cada base de datos, hemos calculado sus correspondientes valores de VE^t. Posteriormente hemos realizado una ordenación creciente de los biclusters según VE^t. Mediante este simple experimento se ha podido comprobar que ninguno de los 10 primeros biclusters clasificados según VE^t están entre los 10 mejores según VE, en ninguna de las bases de

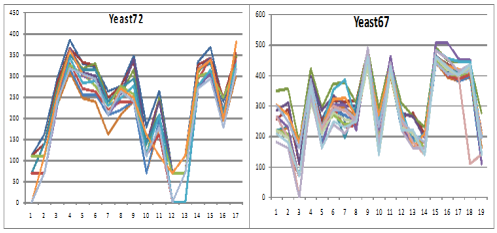


Figura 3: Biclusters de la levadura.

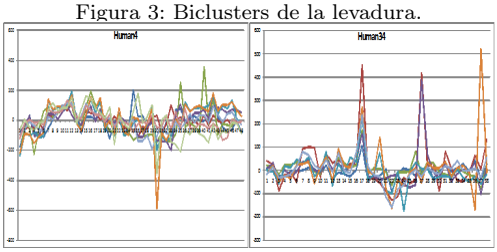


Figura 4: Biclusters de células humanas.

datos. Además, la mayoría de estos biclusters tienen un valor de MSR superior al límite establecido por Cheng y Church [5] para dichas bases de datos (300 para *Yeast* y 1200 para *Human*).

A continuación se muestran algunos ejemplos que muestras la validez de esta nueva propuesta. Se han seleccionado dos biclusters para cada base de datos de entre los 10 mejores según VE^t que presentan un comportamiento fácil de identificar a simple vista.

La figura 3 muestra dos de los biclusters de *Yeast* con un valor de VE^t bajo. Ambos biclusters son visualmente interesantes. Sin embargo, ninguno de estos dos biclusters se encuentra entre las 10 mejores evaluaciones de VE, y ambos tienen valores de MSR superiores al umbral 300. En concreto, el bicluster 72 tiene un valor MSR de 347,74 (décimo mejor según MSR) y un valor de VE igual a 0,398 (29 mejor de 100 según VE), mientras que VE^t considera que es el mejor bicluster de los 100 bajo estudio. El bicluster 67 está clasificado como el séptimo mejor según VE^t, mientras que para MSR tiene el orden 84 (valor MSR de 851,54), y tiene el puesto 82 según VE (valor 0.4815).

Para el caso de *Human*, se han representado los biclusters 4 y 34 en la figura 4. Los dos ocupan buenas posiciones siguiendo el criterio de MSR, ya que se sitúan en los puestos 2 (bicluster 34 con valor MSR 3278,81) y 3 (bicluster 4 con MSR 4007,21). Sin embargo, el límite de MSR establecido para esta base de datos es 1200, por lo que ninguno de los dos se podría haber obtenido haciendo uso de MSR. En el caso de VE, estos biclusters ocupan los puestos 14 (bicluster 34 con valor 0,4119) y 25(bicluster 4 con valor 0,4712). Tanto VE como MSR consideran el bicluster 34 mejor que el bicluster 4. Sin embargo, para VE^t el bicluster 4 es el quinto mejor de los 100 frente al bicluster 34 que ocupa la posición 7.

5. Conclusiones

En este trabajo se presenta una variante de la medida VE propuesta en trabajos previos para la evaluación de biclusters obtenidos a partir de datos génicos en microarrays. Dicha variante, llamada VE^t, añade a las ventajas de la medida original la capacidad de identificar patrones de desplazamiento y escalado simultáneamente. Este tipo de patrones no son reconocidos por ninguna medida actual para biclustering, por lo que VE^t constituye una importante aportación a este tipo de métodos.

Se ha demostrado analíticamente que VE^t es cero cuando el bicluster presenta patrones de desplazamiento y escalado, tanto independientes como simultáneos. Además, queda demostrado que VE^t presenta un crecimiento lineal según se incrementa el error cometido al aproximar los datos de un bicluster a un patrón perfecto.

Como trabajo futuro, VE^t será incluida en una heurística de biclustering, realizando comparaciones de los resultados según sean los objetivos utilizados en la optimización -MSR, VE y VE^t- de forma conjunta o independiente. Además, se incluirán validaciones biológicas usando herramientas disponibles para ello.

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] H. Cho and I. S. Dhillon. Effect of data transformation on residue. Technical report, 2007.

[7] R. Cho, M. Campbell, E. Winzeler, L. Steinmetz, A. Conway, L. Wodicka, T. Wolfsberg, A. Gabrielian, D. Landsman, D. Lockhart, and R. Davis. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2:65–73, 1998.

[8] G. P. Coelho, F. O. de Franca, and F. J. V. Zuben. Multi-objective biclustering: When non-dominated solutions are not enough. *Journal of Mathematical Modelling and Algorithms*, 8(2):175–202, 2009.

[9] 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.

[10] J. Liu, Z. Li, X. Hu, and Y. Chen. Biclustering of microarray data with mospo based on crowding distance. *BMC bioinformatics*, 10 Suppl 4(Suppl 4):S9+, 2009.

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

[12] S. Mitra and H. Banka. Multi-objective evolutionary biclustering of gene expression data. *Pattern Recognition*, 39(12):2464 – 2477, 2006. Bioinformatics.

[13] B. Pontes, F. Divina, R. Giráldez, and J. S. Aguilar-Ruiz. Virtual error: A new measure for evolutionary biclustering. In *Fifth European Conference on Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics (EvoBio 2007)*, pages 217–222, 2007.

[14] B. Pontes, R. Giráldez, F. Divina, and F. Martínez-Álvarez. Evaluación de biclusters en un entorno evolutivo. In *IV Taller nacional de minería de datos y aprendizaje (TAMIDA)*, pages 1–10, 2007.

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

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

[17] 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 394Ũ405, Madison, WI, 2002.

[18] X. Xu, Y. Lu, A. K. H. Tung, and W. Wang. Mining shifting-and-scaling co-regulation patterns on gene expression profiles. pages 89–99, 2006.

[19] 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.