Búsqueda de biclusters con significancia biológica basada en la identificación de patrones de comportamiento

Luna Taylor Jorge Enrique*, León Hirales Claudia María*

**Instituto Tecnológico de La Paz*
Boulevard Forjadores no. 4720, CP 23080
La Paz, Baja California Sur, México
email:eluna@itlp.edu.mx, email:cmleon07@gmail.com

Resumen: El incremento en el uso de la tecnología de microarreglos de ADN ha generado un gran volumen de datos biológicos, lo cual demanda un desarrollo paralelo de métodos computacionales para la interpretación funcional de estos resultados.

En este artículo se presenta una propuesta para el análisis de datos de expresión de genes, basada en la adaptación de un algoritmo genético multi-objetivo biclustering (MOGA), combinado con una función presentada recientemente para evaluar el comportamiento de los genes de un bicluster. Esta función fue modificada para identificar patrones de comportamiento de espejo en la expresión de los genes.

El método se evaluó en relación al porcentaje de biclusters con significancia estadística que es capaz de descubrir, y se comparó con otros métodos reconocidos en la literatura, obteniendo resultados competitivos.

Palabras claves: biclustering, expresión de genes, significancia estadística, patrones de comportamiento en espejo.

1. INTRODUCCIÓN

El uso de microarreglos de ADN es una técnica utilizada ampliamente para la obtención de datos de expresión genética. Una matriz de datos de expresión genética es una matriz de *m x n*, cuyas filas normalmente representan los genes, y las columnas a las condiciones experimentales. Cada elemento *(i,j)* es un valor real que representa el nivel de expresión del gen *i* bajo la condición experimental *j*. Cada fila corresponde a los niveles de expresión de un gen particular, a través de todas las condiciones experimentales, y cada columna corresponde a los niveles de expresión de todos los genes bajo una condición experimental específica. Un bicluster se define como un subconjunto de genes que exhiben un comportamiento similar para un subconjunto de condiciones experimentales, y viceversa. Así, un bicluster es una submatriz dentro de la matriz de expresión de entrada (Dharan y Nair, 2009). (Cheng y Church, 2000), son de los primeros en proponer una función para evaluar la coherencia del comportamiento de los genes dentro de un bicluster. Su cálculo nombrado como *residuo cuadrado medio* (MSR), ha sido citado y utilizado ampliamente en los métodos biclustering. En su trabajo identifican biclusters con patrones de espejo, basados en la agregación de los valores de expresión invertidos de los genes.

(Ayadi, *et al*., 2012), presentan un algoritmo biclustering apoyado en un modelo estocástico de búsqueda local. Partiendo de un primer bicluster el método mejora progresivamente su calidad ajustando tanto genes como condiciones. Los ajustes se basan en la calidad de cada gen, y en su estado en relación al bicluster inicial y la matriz de datos.

(Nepomuceno, *et al*., 2011), presentan un método biclustering basado en la técnica de optimización Scatter Search (SS), que se apoya en un algoritmo de diversificación para generar las soluciones iniciales. El cálculo de la coherencia de un bicluster lo basan en la correlación lineal que existe entre cada par de genes. Este cálculo considera tanto patrones de desplazamiento como patrones de escalado en el comportamiento de los genes del bicluster. (Yang, *et al.,* 2005), presentan un algoritmo que nombraron FLOC, que utiliza el cálculo del MSR para la evaluación de los biclusters. Identifican patrones de comportamiento de espejo, pero al igual que Cheng y Church, lo logran invirtiendo y duplicando los datos de expresión de los genes de la matriz de entrada. (Das e Idicula, 2010), proponen el algoritmo Greedy Search-Binary PSO Hybrid, que identifica biclusters con comportamiento de espejo, y presentan resultados de la significancia biológica de los grupos de genes descubiertos de acuerdo al *p-valor* de un bicluster en relación a una categoría GO.

ISBN: 978-607-95534-5-6

2. MÉTODO PROPUESTO

El método propuesto es una adaptación del algoritmo genético multi-objetivo (MOGA) (Luna y Brizuela, 2012), al cual le incorporamos dos funciones, una de ellas modificada, del trabajo de (Nepomuceno, *et al*., 2011).
Las funciones incorporadas corresponden a los métodos de inicialización y al cálculo de la aptitud de los biclusters. Estas adaptaciones del algoritmo MOGA representan un cambio importante, e influyen notablemente en los resultados obtenidos.
A continuación se describe el algoritmo MOGA, y posteriormente se presentan las funciones agregadas y modificadas.

2.1 – Algoritmo MOGA.

El Algoritmo 1 (MOGA), presentado en la Figura 1, inicia con la creación de una población de *n* biclusters. Posteriormente se calcula el frente de Pareto de cada uno, basado en el concepto de dominancia de acuerdo al método presentado en (Deb, *et al*., 2002). Para que un bicluster pertenezca al frente de Pareto uno, no debe ser dominado por algún otro bicluster de la población. Los biclusters del frente uno se descartan para proseguir con la identificación de los biclusters del frente dos, y así sucesivamente. Una vez que se tienen calculados los frentes de Pareto se lleva a cabo una selección de los mejores biclusters por medio de torneo binario. Con los biclusters seleccionados se lleva a cabo un proceso de cruza, para posteriormente aplicar una mutación sobre un porcentaje de la población hija. Después de la mutación se combinan las poblaciones padre e hija, resultando una población de tamaño *2n*, y se vuelven a calcular los frentes de Pareto. Esta población combinada se ordena de menor a mayor frente, y se toman los primeros *n* biclusters para ser considerados como la siguiente población del algoritmo. Estos procesos se repiten mientras no pase un número dado *ng* de generaciones sin que mejore el tamaño de los biclusters que se encuentran bajo un umbral establecido de aptitud.

2.2- Fase de inicialización (Nepomuceno, et al., 2011)

Inicia generando una semilla a partir de la cual se crean las soluciones potenciales (biclusters), basadas en la Ecuación 1.

$$X'_{1+kh} = 1 - X_{1+kh} \text{ para } k = 0, 1, 2, 3, \ldots, \left\lfloor \frac{n}{h} \right\rfloor \qquad (1)$$

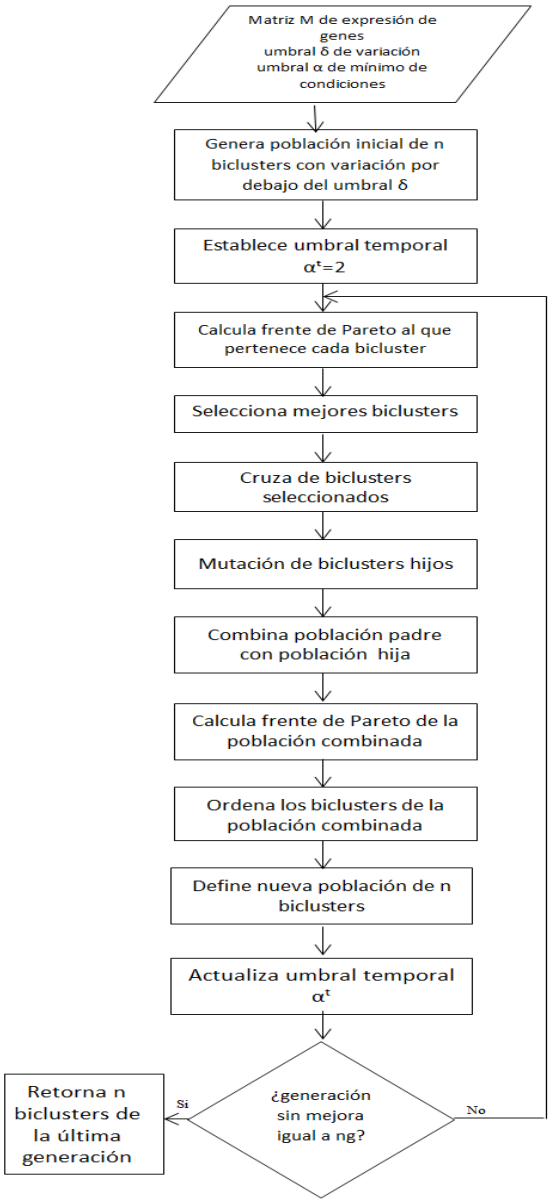


Fig. 1. Algoritmo 1 (MOGA).

Dónde:

- *n* es el número de bits de la cadena binaria.
- *h* es un número aleatorio entero menor que *n/5*.
- X_i representa el valor actual del bit *i* de la cadena binaria.
- X'_i es el nuevo valor del bit *i* de la cadena.

2.3 – Función de aptitud. Cálculo de correlación promedio (Nepomuceno, et al., 2011).

Esta función se basa en la correlación promedio de un bicluster. La aptitud del bicluster se calcula de acuerdo a la Ecuación 2.

$$p(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} p(g_i, g_j) \qquad (2)$$

Dónde:

- $p(g_i,g_j)$ es el coeficiente de correlación entre el gen i y el gen j.
- N es el número de genes del bicluster.

El Algoritmo 2 muestra el pseudocódigo de la implementación de la función de aptitud adaptada al método MOGA.

Algoritmo 2. Cálculo de la correlación promedio de un bicluster.

```
1. Se extraen los valores de expresión del bicluster a
     partir la matriz de datos  y se guardan en dC.
2.  recorre con g los genes
3.    recorre con c las condiciones
4.      suma += dC[g,c]
5.      cuadrado += dC[g,c]^2
6.    fin recorre
7.    mG[g] = suma / numCon
8.    cuadG[g] = cuadrado
9.  fin recorre
10. recorre con g los genes
11.   recorre con j los genes desde g+1
12.     recorre con c las condiciones
13.       productos[c] = dC[g,c] * dC[j,c]
14.       sumProd += Productos[c]
15.     fin recorre
16.     cov=(sumProd/numCon)-(mG[g]*mG[j])
17.     desEstG1=(cuadG[g]/numCon)-mG[g]^2
18.     desEstG2=(cuadG[j]/numCon)-mG[j]^2
19.     si desEstG1 y desEstG2 != 0
20.       correlacion=(cov/(desEstG1*desEstG2))
21.       sumCorr += correlacion
22.     fin si
23.     sumProd = 0
24.   fin recorre
25. fin recorre
26. combinatoria = (numG * (numG – 1) ) / 2
27. corrProm = sumCorr * (1/combinatoria)
28. retorna valor absoluto de correlacion * -100
```

2.4 – Función de aptitud modificada.

Con el objetivo de identificar biclusters con patrones de comportamiento de espejo se modificó la función de correlación de Nepomuceno. La modificación consiste en tomar el valor absoluto de la correlación entre cada par de genes, antes de realizar la sumatoria para obtener la correlación media del bicluster (Ecuación 3).

$$p(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \left| p(g_i, g_j) \right| \qquad (3)$$

3. EXPERIMENTOS Y RESULTADOS

Los experimentos se enfocaron en determinar el porcentaje de biclusters con significancia estadística que es capaz de descubrir el algoritmo, según las anotaciones ontológicas (GO).
Se evaluó el desempeño del algoritmo bajo diferentes combinaciones de las funciones de inicialización y de aptitud. Las opciones evaluadas son:

- Utilización de la función de inicialización original del MOGA (creación inicial de los biclusters con dos genes y dos condiciones).
- Utilización de la función de inicialización propuesta por Nepomuceno (creación de biclusters con máxima diversificación).
- Aplicación del MSR como función de aptitud.
- Aplicación de la función de correlación de Nepomuceno como función de aptitud.
- Aplicación de la función de correlación modificada como función de aptitud.

La significancia estadística de los biclusters descubiertos se determinó bajo dos consideraciones distintas. Primero se consideró un bicluster como significativo si tiene un *p-valor* por debajo de los umbrales tradicionalmente establecidos. Posteriormente se aplicó una restricción más fuerte para considerar significativo un bicluster. Además de tener un *p-valor* por debajo de los umbrales, un biclusters debe de incluir 10 de sus genes dentro una categoría GO, y por lo menos el 50% de sus genes también perteneciendo a dicha categoría. Para realizar estas evaluaciones se utilizó el software AGO (Al-Akwaa y Kadah, 2009).
Además de evaluar las diferentes propuestas del algoritmo, los resultados se compararon con tres métodos reconocidos en la literatura: CC (Cheng y Church, 2000), ISA (Ihmels, *et al*., 2004) y OPSM (Ben-Dor, *et al*., 2002). Los resultados de los algoritmos OPSM, ISA y CC los generamos utilizando el software "Biclustering Analysis Toolbox" (BicAT) (Barkow, *et al*., 2006).
El algoritmo propuesto recibe como entrada la matriz de expresión, un umbral máximo de similitud (aptitud), y el número de condiciones mínimas que se espera en cada bicluster.
Las pruebas se realizaron utilizando la matriz de expresión de (Gash, *et al.*, 2000), compuesta por 2993 genes y 173 condiciones.

3.1 Pruebas con MSR e inicialización por diversificación.

Para las evaluaciones con MSR como función de aptitud se utilizaron umbrales entre 75 y 100, obteniéndose los mejores resultados con un MSR de 100. De la misma forma los mejores resultados se obtuvieron al definir 10 genes y 10 condiciones como mínimo en cada bicluster generado.
La Figura 2 muestra una gráfica comparativa del porcentaje de biclusters significativos descubiertos por el algoritmo y los demás métodos evaluados. El eje de las Y corresponde al porcentaje de biclusters significativos bajo diferentes niveles de *p-valor.* El eje de las X presenta los diferentes métodos evaluados.
En el segundo bloque de esta gráfica se pueden observar los resultados obtenidos por el algoritmo utilizando la función MSR. Para un nivel de *p-valor* de 5% obtiene el 100% de biclusters significativos, y para *p-valores* más bajos se desploma el porcentaje hasta caer por debajo del 10%.

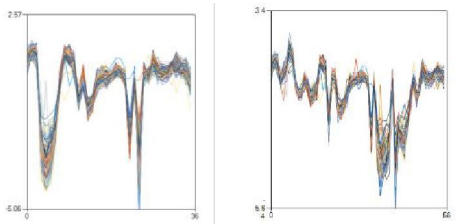
3.2 Pruebas con la función de correlación.

Las pruebas con la función de correlación se realizaron con cuatro valores distintos como umbral de correlación. Se evaluó el comportamiento de esta función en combinación con la inicialización original del MOGA, y con la inicialización con diversificación. Los mejores resultados se lograron con la opción de diversificación.
En la Figura 2 se observa que estas pruebas obtuvieron el 100% de biclusters significativos para todos los niveles de *p-valor,* superando a los demás métodos evaluados.
Dado estos resultados se optó por probar el algoritmo con restricciones más fuertes de significancia. Para esto, los biclusters deben de tener por lo menos 10 de sus genes, y por lo menos el 50% de sus genes, relacionados a una función o proceso biológico.

En la Tabla 1 y en la Figura 3 se presentan los resultados obtenidos. Se puede observar que los mejores resultados se obtuvieron con un umbral de correlación de 0.98, logrando un 90% de biclsuters significativos con la inicialización con diversificación, y un 70% con la inicialización original del MOGA.
En la Figura 4 se muestran las gráficas del comportamiento de los genes de dos biclusters descubiertos en estas pruebas. La Figura 4A muestra un bicluster con correlación de 0.97, y la Figura 4B muestra la gráfica para un bicluster con correlación de 0.98.

Tabla 1. Porcentaje de biclusters significativos utilizando la función de correlación, aplicando condiciones simple y fuerte de significancia.

Pruebas para Correlación			
Variación	% significancia estadística	% significancia estadística fuerte	Método de inicialización
.90	100	0	Diversificación
.95	100	38.4615	Diversificación
.97	100	53.8462	Diversificación
.98	100	90	Diversificación
.98	100	70	Inicio MOGA



A) B)

Fig. 4. Gráficas del comportamiento de los genes de dos biclusters descubiertos. A) Bicluster con correlación de 0.97. B) Bicluster con correlación de 0.98.

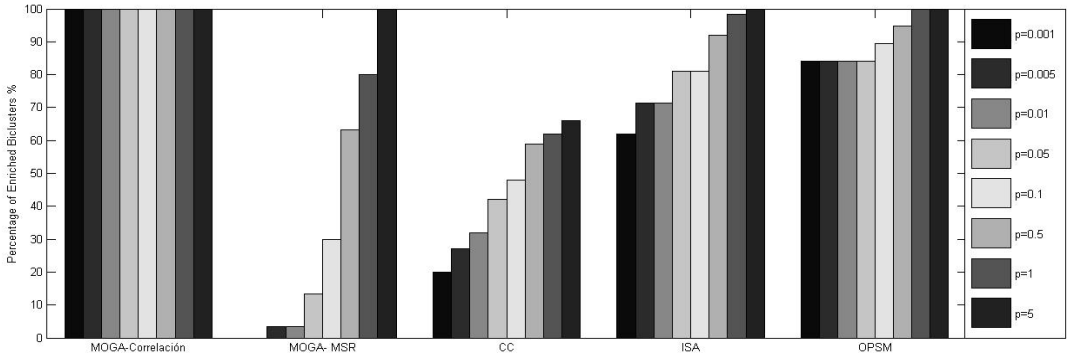


Fig. 2. Porcentaje de biclusters significativos descubiertos por los algoritmos, utilizando el *p-valor* como única condición de significancia.

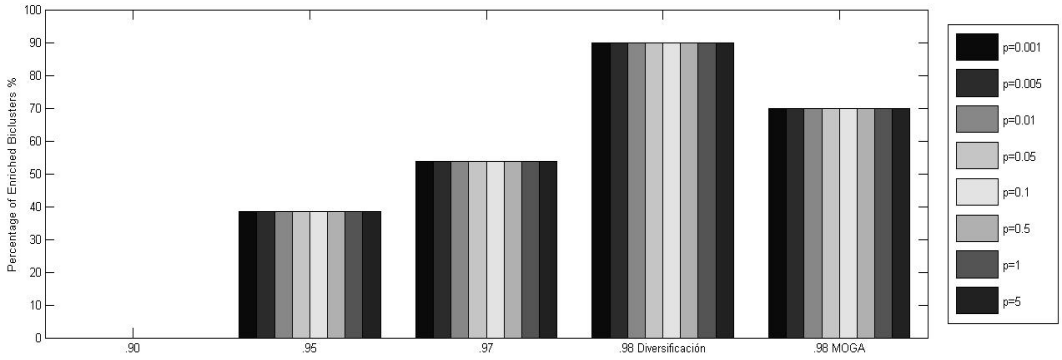


Fig. 3. Porcentaje de biclusters significativos descubiertos por el método propuesto, utilizando diferentes umbrales de correlación.

3.3 Pruebas para patrones de espejo.

La función de aptitud modificada es capaz de identificar biclusters cuyos genes presentan patrones de comportamiento de espejo. La Figura 5 muestra la gráfica de un bicluster descubierto, cuyos genes presentan un patrón de comportamiento de espejo.

Se realizaron pruebas para evaluar si los genes que muestran este tipo de comportamiento están relacionados biológicamente. Como resultado de estas pruebas se descubrieron biclusters con significancia estadística, cuyos genes se comportan con un patrón de espejo. La Tabla 2 presenta los *p-valores* y los procesos biológicos asociados a los biclusters descubiertos por el algoritmo.

En las Figuras 6, 7 y 8 se muestran las gráficas y los genes descubiertos con comportamiento de espejo, asociados a diferentes procesos biológicos, según las anotaciones ontológicas.

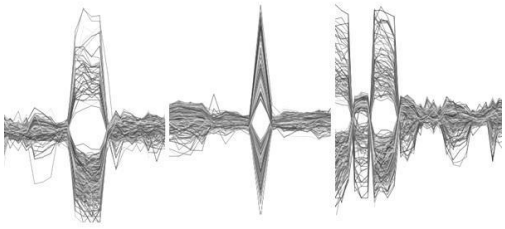


Fig. 5. Gráfica del comportamiento de los genes de un bicluster descubierto por el algoritmo, que presentan comportamiento de espejo.

Tabla 2. Biclusters significativos biológicamente cuyos genes presentan comportamiento de espejo.

Bicluster	Nombre del proceso	p-valor
Bicluster 2	Asamblea ribosoma y biogénesis	2.48e-062
Bicluster 2	Traslación	4.45e-030
Bicluster 10	Traslación	5.39e-034
Bicluster 11	Traslación	4.78e-022
Bicluster 13	Traslación	1.90e-092
Bicluster 15	Asamblea ribosoma y biogénesis	1.48e-071

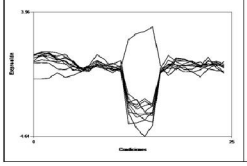
N° Bicluster	Nombre	Ontología
3	Procesamiento de rRNA	Proceso biológico
Grafica		Genes
		YBR142W YBR247C YDR365C YER006W YGL030W YGL120C YLR002C YLR293C YMR229C YMR302C YPL043W

Fig. 6. Genes con comportamiento de espejo descubiertos por el algoritmo, asociados al procesamiento de rRNA.

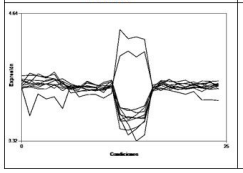
N° Bicluster	Nombre	Ontología
3	Regulación de la transcripción del RNA polimerasa promotor II	Proceso biológico
Grafica		Genes
		YDR190C YFL039C YGL115W YGR200C YJL081C YJL089W YPL086C YPL235W YPR009W YPR023C YPR072W

Fig. 7. Genes con comportamiento de espejo descubiertos por el algoritmo, asociados a la función de regulación de la transcripción del RNA polimerasa promotor II.

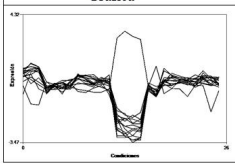
N° Bicluster	Nombre	Ontología
7	Inicio traslacional	Proceso biológico
Grafica		Genes
		YAL035W YBR079C YER025W YGR054W YGR162W YJL138C YJR007W YKR059W YMR146C YMR260C YMR309C YNL062C YOR361C YPL119C YPL237W YPR163C

Fig. 8. Genes con comportamiento de espejo descubiertos por el algoritmo, asociados a la función de inicio traslacional.

4. CONCLUSIONES

En este artículo presentamos una propuesta basada en la adaptación de las funciones de inicialización y cálculo de aptitud propuestas en (Nepomuceno, *et al.*, 2011), dentro del algoritmo MOGA presentado en (Luna y Brizuela, 2012). Los experimentos realizados muestran porcentajes muy altos de biclusters significativos descubiertos por el algoritmo, superando significativamente a tres métodos reconocidos y citados ampliamente en la literatura.

Por otro lado, se modificó la función de aptitud con el objetivo de identificar genes que se correlacionan de forma inversa, esto es, grupos de genes que muestran gráficas de comportamiento en espejo. Las pruebas realizadas muestran que estos grupos de genes correlacionados inversamente, pueden estar interrelacionados dentro de alguna función o proceso biológico. Sobre esto último podemos concluir que el algoritmo propuesto aplicando la función modificada, es capaz de identificar genes relacionados biológicamente, que las funciones tradicionales para medir la coherencia de los biclusters, como el MSR, no pueden detectar sin invertir y duplicar los datos de expresión de entrada.

Un reto importante que queda por superar es disminuir el porcentaje de traslape entre los genes que presentan los diferentes biclusters. Esto se debe lograr sin disminuir sensiblemente el porcentaje de biclusters significativos generados.

REFERENCIAS

Al-Akwaa F.M. y Kadah Y.M. (2009). An automatic gene ontology software tool for bicluster and cluster comparisons. *In Proceedings of the 6th Annual IEEE conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB'09).* 163-167.

Ayadi W., Elloumi M. y Hao J. K. (2012). Pattern-driven neighborhood search for biclustering of microarray data. *Bioinformatics.*13(Suppl 7):S11

Barkow S., Bleuler S., Prelic A., Zimmermann E. y Pand Z. (2006). BicAT: biclustering analysis toolbox. *Bioinformatics,* 22:1282-1283.

Ben-Dor A., Chor B., Karp R y Yakhini Z. (2002). Discovering local structure in gene expression data: The order-preserving submatrix problem. In *Proceedings of the 6th International Conference on Computacional Biology (RECOMB'02).* pp. 49-57.

Cheng Y. y Church G. M. (2000). Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology (ISMB'00),* pp. 93-103.

Deb K., Agarwal S, PPratap A, Meyarivan T. (2002). A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evol. Comput.* (6):182-197.

Dharan S. y Nair A. S. (2009). Biclustering of gene expression data using reactive greedy randomized adaptive search procedure. *BMC Bioinformatics*, 10(Suppl. 1):S27

Das S. y Idicula S.M. (2010). Greedy Search-Binary PSO Hybrid for Biclustering Gene Expression Data. *Inter-national Journal of Computer Applications.* 2(3):0975-8887.

Gasch A.P., Spellman P.T., Kao C.M,. Carmel-Harel O., Eisen M.B., Storz G., Botstein D. y Brown P.O. (2000). Genomic Expression Programs in the Response of Yeast Cells to Environmental Changes. *Molecular Biology of the Cell.* 11:4241-4257.

Ihmels J., Friendlander G., Bergmann S., Sarig O., Ziv Y. y Barkai N. (2002). Revealing modular organization in the yeast transcriptional network. *Nature Genetics*, 31:370-377.

Ihmels J., Bergmann S. y Barkai N. (2004). Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20:199-2003.

Luna-Taylor J.E. y Brizuela C.A. (2012). A multiobjetive genetic algorithm for the biclustering of gene expression data. *In proceeding of the 3rd international supercom-puting Conference in Mexico ISUM 2012.*

Nepomuceno J. A., Troncoso A., Aguilar-Ruiz J., (2011). Biclustering of gene expression data by correlation-based scatter search. *BioData Mining*. 4:3.

Prelic A., Bleuler S., Zimmermann P., Wille A., Buhlmann P., Gruissem W., Hennig L., Thiele L. y Zitzler E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22:1122-1129.

Yang J., Wang H., Wang W. y Philio S. Y. (2005). An Improved Biclustering Method for Analyzing Gene Expression Profiles. *International Journal on Artificial Intelligence Tools - IJAIT* , 14(5):771-790.