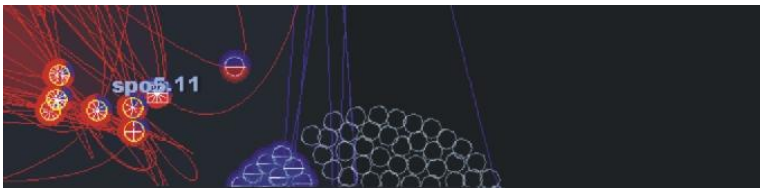
ANÁLISIS VISUAL DE DATOS DE EXPRESIÓN GÉNICA MEDIANTE BICLUSTERING (RESUMEN)

RODRIGO SANTAMARÍA



Doctor en Informática y Automática

Departmento de Informática y Automática
Facultad de Ciencias
Universidad de Salamanca

Junio 2004 - Julio 2009

Rodrigo Santamaría: *Análisis Visual de Datos de Expresión Génica mediante Biclustering (Resumen)* , Doctor en Informática y Automática, © Junio 2004 - Julio 2009

SUPERVISORES:
Luis Quintales
Roberto Therón

LUGAR:
Salamanca

PERIODO:
Junio 2004 - Julio 2009

ÍNDICE GENERAL

ÍNDICE DE FIGURAS

ÍNDICE DE CUADROS

ACRONYMS

BicAT	Biclustering Analysis Toolbox
BP	Biological Process
CC	Cellular Component
cDNA	complementary DNA
CESR	Core Environmental Stress Response
DNA	DeoxyriboNucleic Acid
FDCG	Force-Directed Clustered Graph
GEO	Gene Expression Omnibus
GO	Gene Ontology

HCE	Hierarchical Clustering Explorer
HCG	Hierarchical Clustered Graph
HCI	Human-Computer Interaction
MDS	MultiDimensional Scaling
MIAME	Minimal Information About Microarray Experiments
NCBI	National Center for Biotechnology Information
NVAC	National Visualization and Analytics Center
PCA	Principal Component Analysis
PCR	Polymerase Chain Reaction
RNA	RiboNucleic Acid
MF	Molecular Function
mRNA	messenger RNA
SESR	Specific Environmental Stress Response
SOM	Self-Organizing Maps
SVM	Support Vector Machines
TRN	Transcription Regulatory Network

Parte I

RESUMEN

1

RESUMEN

Tener una buena mente no es suficiente. Lo principal es usarla correctamente
— René Descartes, Discurso del Método, 1637

Durante la última década hemos sido testigos de una avalancha de progresos en el campo de la genómica. La iniciativa del Proyecto Genoma Humano [24] y proyectos similares han establecido las bases de la estructura genética de muchos organismos clave, identificando sus genes. Aunque no exentas de limitaciones, estos mapeos secuencia-gen han mejorado dramáticamente nuestra comprensión sobre la genómica.

las mejores estimaciones indican que sólo un 92 % del genoma humano está descifrado, y partes del genoma, tales como el DNA basura, son aún un misterio

Esto, unido a las técnicas de manipulación genética desarrolladas principalmente a partir de la reacción en cadena de la polimerasa (PCR)[1], ha dado lugar a múltiples tecnologías para determinar el comportamiento de los genes bajo distintas condiciones. De entre estas tecnologías, seguramente la más utilizada son los *microarrays*, capaces de medir el nivel de transcripción de muchas secuencias genéticas a la vez. Normalmente, un microarray cubre todo el genoma conocido de un organismo (miles de genes) y se utilizan varios microarrays para medir su transcripción bajo distintas condiciones experimentales.

Estas tecnologías generan tal cantidad de datos que, unidos a la cantidad de datos ya disponibles gracias principalmente a Internet y el uso de repositorios públicos, provocan un cuello de botella al llegar a la fase de análisis: nuestra capacidad para analizar los datos es mucho menor que nuestra capacidad para generarlos. Una solución a esta limitación es realizar análisis relativamente simples de los datos para aligerar la tarea. Sin embargo, y aunque el volumen de datos es tan grande que incluso los análisis más simples permiten hacer descubrimientos importantes, cada vez se demandan técnicas más complejas que se ajusten mejor a la realidad. Una de estas técnicas son los *algoritmos de biclustering*, una evolución de los algoritmos de clustering tradicionales.

Si no queremos simplificar en exceso el análisis, debemos entonces utilizar el máximo de nuestra competencia cognitiva para realizar dichos análisis. Una aproximación ampliamente aceptada es aumentar los niveles cognitivos involucrados en el análisis, añadiendo el pensamiento visual a la cognición abstracta [57]. Dicha aproximación dio lugar en la pasada década a la Visualización de la Información, que se ha revelado como un área de investigación clave para guiar e incrementar la utilidad de distintas técnicas de análisis, gracias a la representación y la interacción. Este ha sido el caso, por ejemplo, del análisis de grupos de genes inferidos a partir de su expresión genético y su visualización mediante heatmaps [14], que por su utilidad se ha convertido en un estándar *de facto*. Para cubrir no sólo la representación visual de los datos, si no el proceso analítico completo, nace en 2005 la Analítica Visual [52], que se está aplicando en muchas áreas de investigación y se ha convertido en una ciencia en sí misma.

1 El PCR es una técnica que utiliza la DNA polimerasa para crear múltiples copias a partir de una pequeña muestra de DNA

En general, el diseño y análisis de microarrays tiene como finalidad responder a varias preguntas, que pueden ser resumidas en [5]:

- ¿Cuál es la diferencia en el nivel de expresión génica entre distintos tipos de células y estados, cómo cambia la expresión génica con las enfermedades y los tratamientos?
- ¿Cómo se regulan los genes, cómo interacciones los genes y sus productos, cuáles son esas redes de interacción?
- ¿Cuáles son los papeles funcionales de los distintos genes y en qué procesos celulares participan?

Dependiendo del alcance de cada caso particular y del conocimiento previo disponible podemos tener, a grandes rasgos, dos tipos de instancias para estos análisis. El primero suele generar cuestiones que se responden con sí o no, tales como *¿el nivel de transcripción de los genes relacionados con cáncer son parecidos a los niveles normales para un determinado paciente?*. En estos casos, tenemos información previa (los indicadores tumorales) y el análisis es *confirmatorio*. El segundo tipo de instancia es mucho más general, por ejemplo *¿qué genes están involucrados en la respuesta celular al estrés?*. En este caso, se necesitan experimentos más complejos que cubran un espectro amplio y el análisis es *exploratorio*.

En el análisis confirmatorio (o de testeo de hipótesis), el conocimiento biológico acumulado dirige el discurso analítico, mientras que en el análisis exploratorio, sólo es una guía para enfocar el análisis y validar nuevos descubrimientos. El análisis exploratorio suele necesitar técnicas de análisis más complejas. Una de estas técnicas es el biclustering, cuyas ventajas teóricas a la hora de modelar el comportamiento de la expresión génica [32, 49, 37] han sido confirmadas por varios estudios [39, 34, 37].

Nuestro *objetivo* es comprender el proceso analítico necesario para el estudio de la expresión génica por medio de técnicas de biclustering. Para ello, este trabajo necesita en primer lugar compilar y comprender los algoritmos de biclustering disponibles. A continuación, deberemos estudiar las técnicas de visualización utilizadas para la representación de las matrices de expresión y los resultados de biclustering. Una vez detectadas las ventajas y desventajas existentes en el análisis y la visualización, identificamos dos objetivos base de investigación: el desarrollo de una nueva técnica de visualización para biclustering y la adaptación de índices para la validación de algoritmos de biclustering.

El último paso de este estudio es dar un paso atrás para poder observar el proceso completo de análisis, integrando el conocimiento adquirido en el desarrollo de una solución basada en la analítica visual para estudiar mediante biclustering los datos de expresión génica.

Los *resultados* de este estudio pueden resumirse en los siguientes puntos:

- El marco de trabajo BicOverlapper [41, 42], que aplica de manera consistente la analítica visual a la exploración de datos de expresión génica mediante algoritmos de biclustering. BicOverlapper es una herramienta abierta disponible en:
  `http://vis.usal.es/bicoverlapper`

- Un estudio general de las técnicas de visualización en el campo de la expresión génica, que desemboca en el diseño de una técnica novedosa para la visualización de biclusters.
- El desarrollo de métricas para la validación interna de algoritmos de biclustering y su aplicación a la optimización de sus parámetros.
- El desarrollo de *biclust*, un paquete para R con la implementación de varios algoritmos de biclustering. Este es un trabajo desarrollado en colaboración con Sebastian Kaiser y Friedrich Leisch de la Universidad de Munich, y está disponible en: `http://cran.r-project.org/web/packages/biclust`
- Una descripción formal del diseño de soluciones para el análisis visual de la expresión génica.
- La discusión de varios casos de estudio sobre el análisis de expresión génica mediante biclustering.

1.1 ANÁLISIS DE LA EXPRESIÓN GÉNICA

El dogma central de la biología molecular es el siguiente:

El DNA se transcribe en RNA, que se traduce en proteínas

y se puede interpretar, muy simplificadamente, como que las características observables de un organismo (fenotipos) se basan en sus genes, ya que estos dirigen la formación de proteínas y estas dan lugar, finalmente, a dichos fenotipos. Por tanto, el que una determinada característica prevalezca se debe a que la proteína asociada a la característica se encuentra en mayor cantidad, ya que había mayor cantidad de RNA para traducir. Medir la cantidad de RNA (es decir, el nivel de transcripción del DNA) nos da una idea del nivel de expresión del gen. La tecnología estándar para esta medición es el microarray. Un microarray es una cuadrícula de plástico que contiene miles de muestras genéticas que, mediante un proceso conocido como hibridación, capturan el nivel de RNA presente para cada secuencia genética en la cuadrícula.

En una matriz de expresión génica, cada columna contiene los niveles de transcripción medidos por un microarray. El procedimiento usual es realizar varios experimentos de microarray, para distintas condiciones, y después juntarlos por columnas en la matriz. En el fondo, ésta es una matriz numérica A con n filas (genes) y m columnas (condiciones) donde el elemento a_{ij} es el nivel de transcripción del gen i bajo la condición j. Cualquier método de análisis multivariante puede aplicarse a los datos en la matriz, que podemos clasificar en métodos de *filtrado* y de *clasificación*. Los métodos de filtrado se usan principalmente en testeo de hipótesis (*¿tiene cáncer este paciente?*) mientras que los métodos de clasificación son más utilizados en análisis exploratorio (*¿cuál es la respuesta de un determinado organismo al estrés?*)

aunque estrictamente hablando la expresión génica involucra muchos más procesos que la transcripción, en este contexto muchas veces ambos conceptos se utilizan indistintamente

Los métodos de filtrado funcionan normalmente mediante la definición de umbrales de expresión, y marcan los genes con expresiones mayores o menores que dichos umbrales como infra o sobre expresados. Un método de filtrado muy utilizado es el *análisis diferencial de la expresión*, que consiste en dividir en dos el conjunto de condiciones, y determinar si los genes tienen una expresión muy alta o baja en uno de los grupos (grupo de prueba) comparada con su expresión en el otro grupo (grupo de control).

Los métodos de clasificación tratan de caracterizar la estructura general de la matriz de expresión, revelando grupos (clases) de genes con comportamientos similares. Esto se puede llevar a cabo con la ayuda de conocimiento previo (clasificación supervisada) o no (clasificación no supervisada). La clasificación supervisada es menos útil para un análisis exploratorio puro, pero generalmente da mejores resultados. Algunos métodos supervisados son las máquinas de soporte vectorial (SVM), las redes neuronales, el análisis discriminante o el análisis de vecinos más cercanos. La clasificación no supervisada es independiente de cualquier dato externo, pero es menos precisa. Clustering y biclustering son ejemplos de métodos no supervisados, al igual que el análisis de componentes principales (PCA) o los mapas autoorganizados (SOM).

1.2 VISUALIZACIÓN DE LA INFORMACIÓN Y ANALÍTICA VISUAL

La *visualización de la información* es el estudio de la representación visual e interactiva de datos abstractos con el objetivo de ampliar nuestro conocimiento sobre dichos datos [57]. En la mayoría de las disciplinas, ya se relacionen con la biología, la física, la estadística, etc. utilizamos análisis numéricos que explotan nuestra inteligencia abstracta para extraer información relevante de los datos. La visualización de la información nos permite aprovechar otros tipos de inteligencia, especialmente la inteligencia visual y verbal, para mejorar nuestra comprensión sobre un determinado problema. La visualización de la información se basa en varios principios fundamentales relacionados con la psicología de la percepción, la interacción hombre-máquina (HCI), la estética y el diseño[2]. El objetivo final es dar al usuario una gran variedad de opciones para visualizar los datos representados de la forma más apropiada posible y permitiendo una alta interacción, de modo que se pueda llevar a cabo una aproximación al análisis que nos permita comenzar visualizando todo nuestro problema, para a continuación centrarnos en sus aspectos más interesantes y recuperar detalles cuando sea necesario. Este tipo de aproximación se resume en el mantra de la visualización de la información [48], un referente para el diseño de interfaces visuales:

Howard Gardner introdujo la teoría de las inteligencias múltiples en 1983

> Visión general primero, zoom y filtrado después, detalles bajo demanda

La *analítica visual* es un área de conocimiento muy reciente que atañe al proceso completo de de análisis. Engloba distintas áreas de información, pero especialmente la visualización de la información y la minería de datos, de modo que integremos el razonamiento humano y la visualización interactiva en el proceso de análisis [27].

La analítica visual se funda oficialmente alrededor de 2005 con el propósito inicial de analizar y dar respuesta rápida a situaciones de emergencia

El proceso analítico tiene cuatro fases [52]:

1. *Recuperación de información* relevante, incluyendo acceso a recursos, parseo de formatos, filtrados, etc. Involucra a las ciencias de la computación, las matemáticas, la estadística y la minería de datos.

2. *Re-representación de los datos* para que su visualización nos permita establecer un discurso analítico. Involucra al diseño gráfico, la visualización de la información y la interacción hombre-máquina.

3. *Desarrollo de comprensión* del problema mediante la interacción con las visualizaciones y la generación de hipótesis. En este punto, las áreas involucradas anteriormente se ponen al servicio del área objeto de análisis (biología, física, sociología, etc.)

4. *Producción de resultados* a partir del conocimiento adquirido y de las conclusiones a las que hemos llegado, ya sea de manera concreta (gráficos, informes, etc.) o abstracta (cambios de criterio, acumulación de experiencia, confirmación de hipótesis, etc.)

El proceso es cíclico, de manera que los resultados obtenidos pueden dar lugar a la necesidad de recuperar nuevos datos, etc. Asimismo, el flujo hacia atrás entre fases también está permitido, de modo que la interacción con los datos puede, por ejemplo, llevar a generar nuevas representaciones o incluso a adquirir nuevos datos

2 compilaciones de estos principios pueden encontrarse en [7, 57]

1.3 BICLUSTERING

El primer algoritmo de biclustering apareció en 1972, pero no es hasta 2000 cuando se empieza a aplicar a datos de microarrays

El biclustering es una técnica de análisis no supervisada que se ha aplicado mucho en los últimos años para estudiar la expresión génica. Esta técnica es una evolución natural del clustering [19], que agrupa conjuntos de genes con un perfil genético similar en todas las condiciones experimentales analizadas. La capacidad analítica del biclustering para el análisis de expresión supera a la del clustering tradicional gracias principalmente a dos de sus características: *agrupamiento simultáneo de genes y condiciones* y *solapamiento*. Gracias a la primera, los grupos encontrados por un método de biclustering (*biclusters*) pueden referirse a genes que actúan de manera similar bajo sólo una serie de condiciones, no necesariamente todas, lo cual encaja con el comportamiento biológico observado (un grupo de genes pueden trabajar juntos para atender a una determinada circunstancia, pero estar desacoplados bajo otras). Mediante la segunda, podemos tener genes en más de un bicluster a la vez, lo cual es muy interesante, ya que la realidad biológica es que un gen puede tener más de una función asociada y trabajar con distintos conjuntos de genes bajo distintas condiciones.

Estas ventajas teóricas se ven apoyadas por estudios prácticos, bien en los trabajos de cada autor (para recopilaciones exhaustivas de algoritmos de biclustering, consultar [50, 32]) o en comparativas [37, 39, 34].

Formalmente, dada una matriz de expresión $A = a_{ij}$, un bicluster $B = (G, C)$ es un subconjunto de gene G y condiciones C que tienen un comportamiento similar (bien los genes en G a lo largo de todas las condiciones en C, o viceversa, o ambas). Este subconjunto define una submatriz de A con los niveles de expresión de $B = b_{ij}$. Un algoritmo de biclustering suele encontrar varios biclusters. Para dos biclusters que se solapan, definimos la matriz de solapamiento como $O(B_1, B_2) = A(G_1 \cap G_2, C_1 \cap C_2)$. Es importante reseñar que el solapamiento puede producirse sólo en genes o sólo en condiciones. Por ejemplo, los biclusters rojo y verde en la fig. 1 se solapan en las condiciones 1 y 2 y en el gen 4, pero los biclusters azul y naranja también se solapan, solo en las condiciones 5 y 6.

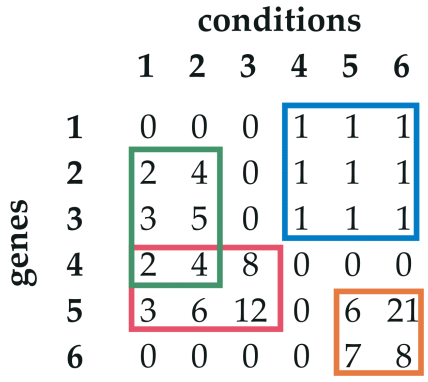


Figura 1: Ejemplo de una matriz de expresión simple, con cuatro biclusters.

Hay varios tipos de biclusters, dependiendo del tipo de comportamiento que definamos como 'similar' [32]:

- *Constante*: todos los niveles de expresión en el bicluster tienen el mismo valor, o valores fluctuando alrededor de un intervalo pequeño.
- *Coherente*: todos los niveles de expresión varían a lo largo de las filas y/o las columnas con algún tipo de coherencia, sea cual sea su valor. Esta relación suele ser aditiva o multiplicativa, de modo que el valor de expresión entre las filas/columnas del bicluster difiere en un factor aditivo o multiplicativo.
- *Evolución coherente*: en este caso, existe una variación en la tendencia pero es solo cualitativa (los niveles de expresión aumentan o disminuyen a lo largo de las filas o condiciones, pero sin seguir un factor común).

En la fig. 1, el bicluster azul es constante, el verde es coherente aditivo y el rojo es coherente multiplicativo; tanto en filas como en columnas. El bicluster naranja tiene evolución coherente por filas (ambas filas aumentan su expresión de la condición 5 a la 6), pero no por columnas (una columna aumenta del gen 5 al 6, pero la otra disminuye.

Una vez definido qué vamos a buscar, tenemos que definir cómo lo vamos a buscar. Los métodos de búsqueda de biclusters son muy variados, yendo de la adaptación del concepto de clustering a filas y columnas (*two-way clustering* [17, 51]) hasta técnicas de divide y vencerás [13, 53], búsquedas avariciosas [11, 58, 59, 3, 33, 6], enumeraciones exhaustivas [50, 49, 30, 34] o modelos estadísticos [29, 55, 45, 44, 28]. Algunos métodos comienzan a integrar conocimiento biológico en el algoritmo de búsqueda para guiar el análisis [39], convirtiéndolo, de alguna manera, en un método 'supervisado biológicamente'.

los términos GO son descripciones de los procesos biológicos, funciones moleculares y componentes celulares asociados a los genes

los motivos (motifs) *de red son estructuras básicas que se repiten frecuentemente en una red*

La *validación* de los algoritmos de biclustering se realiza en la literatura [37, 39, 34] mediante una combinación de *índices de validación externos*, como por ejemplo la medida F_1 [54], y el uso de conocimiento biológico, normalmente mediante *test de significatividad estadística* sobre términos GO y motivos de red. Un test de significatividad determina la posibilidad de que cierta característica de un grupo de elementos sea por azar. En el caso, por ejemplo, de términos GO, medimos la posibilidad de que un término compartido por varios genes en nuestro bicluster pudiera ser compartido también por ese mismo número de genes pero seleccionados aleatoriamente. Si la posibilidad es muy baja (está por debajo de un umbral de significación que suele ser menor de 0.1) entonces decimos que el bicluster *enriquece* el término GO.

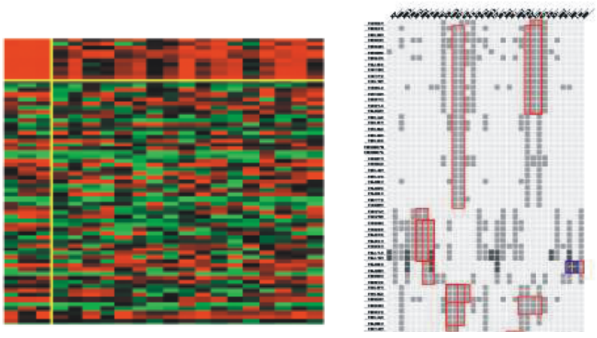
1.4 VISUALIZACIÓN DE EXPRESIÓN GÉNICA Y BICLUSTERS

Las matrices de expresión, por su alta dimensionalidad y por la complejidad de los análisis, son un reto interesante para la visualización de información. En una matriz de expresión hay tres entidades principales involucradas: genes, condiciones y niveles de expresión. Además, nos interesa visualizar patrones de comportamiento en la expresión. Hay dos técnicas de visualización fundamentales para la representación de matrices de expresión y su estructura: el mapa de calor (heatmap) y las coordenadas paralelas (ver tabla 1).

ENTITY	HEATMAP	PARALLEL COORDINATES
Gen	eje y	polilínea
Condición	eje x	eje x
Nivel de expresión	color	eje y
Estructura	reordenación	filtrado

Cuadro 1: Codificación de los datos relacionados con la expresión génica en heatmaps y coordenadas paralelas

El *heatmap* es la representación natural de las matrices de expresión, ya que emula las fluorescencias estimuladas para detectar el nivel de transcripción en las tecnologías de microarray. Un heatmap es una representación 2D de la matriz, donde las filas representan genes y las columnas condiciones. Cada nivel de expresión se representa como un cuadrado de color en la posición correspondiente a su gen y su condición.



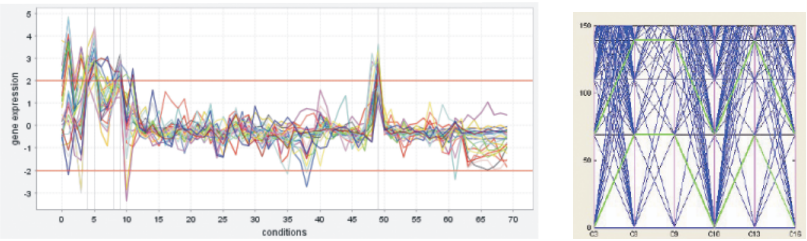
(a) (b)

Figura 2: a) Visualización de la matriz de expresión como un heatmap. Las filas y columnas se han reordenado para mostrar un bicluster con muy alta expresión en la esquina superior izquierda, delimitado por las líneas amarillas (imagen generada con *biclust* [26]). b) Detalle de la visualización de BiVoc [18] para múltiples biclusters, rodeados en rojo (imagen reproducida a partir de [18]).

El color depende del nivel de expresión, y suele seguir una escala bi o tri-color, típicamente de verde (baja expresión) a negro (expresión media) y a rojo (alta expresión). La visualización de patrones se obtiene reordenando los elementos de la matriz según, por ejemplo, los grupos encontrados por algún método de análisis (ver fig. 2a). En el caso de biclustering, la opción más utilizada es colocar en la parte superior izquierda de la matriz las genes y condiciones en el bicluster [2].

La interacción con el heatmap suele permitir cambiar las escalas de colores, reordenar en función de grupos, hacer zoom sobre la matriz, distorsionar ciertas áreas, buscar por nombres genes o condiciones, etc.

Las *coordenadas paralelas* [23] se han utilizado fundamentalmente para visualizar subconjuntos de genes. En esta técnica, cada perfil genético g_i se considera como un punto m-multidimensional $p_i = (a_{i1}, a_{i2}, ...a_{im})$ donde a_{ik} es el nivel de expresión del gen g_i bajo la condición experimental c_k. Las condiciones se representan como ejes verticales equidistantes en $x_1^c, ..., x_m^c$. Cada perfil g_i se representa como una polilínea con m puntos (x_k^c, y_k), con y_k proporcional a a_ik. El uso de patrones lineales para representar cantidades es mucho más adecuado que el uso de color, según la teoría de la percepción [57], pero la visualización no es muy adecuada para representar gran cantidad de perfiles genéticos heterogéneos, pues se recarga rápidamente. Por ello, se utiliza exitosamente sólo sobre conjuntos de genes analizados previamente, por ejemplo clusters o biclusters, y se representan sólo los perfiles de dichos genes. En el caso de biclusters, no hay una opción clara para representar las condiciones en el bicluster, algunos autores las marcan con líneas [2], otros filtran el resto de condiciones [10].



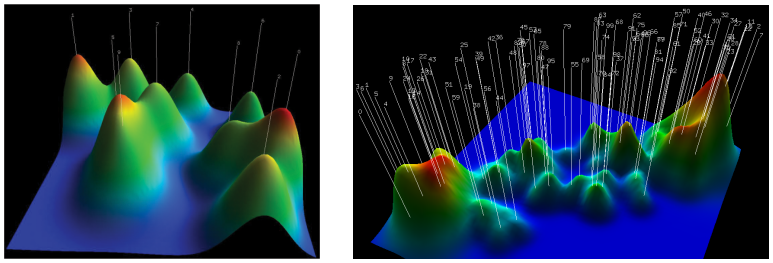
(a) (b)

Figura 3: a) Coordenadas paralelas de BicAT para un bicluster de 21 genes y 5 condiciones (generado con [2]). b) Coordenadas paralelas de BiVisu (editado a partir de [10]).

En cuanto a la *representación de múltiples biclusters*, la aproximación existente se basa en hacer una reordenación global de genes y condiciones en el heatmap para representarlos [18]. Sin embargo, esta representación tiene limitaciones geométricas cuando tenemos muchos biclusters o se solapan bastante, con lo que debe recurrir a la duplicación de filas y columnas para representarlos todos (ver fig. 2b).

Otra opción, aunque sólo aplicada a clustering, es tratar cada cluster como una entidad independiente, y realizar una proyección a 2D de cada uno según los genes que contiene [38]. En cada punto proyectado, se dibuja una 'montaña' cuyas características (color, altura, pendiente)

representan características del bicluster (desviación, número de genes, similitud; ver fig. 4).



(a) (b)

Figura 4: a) Mapa de montañas para 10 clusters. Tres de ellos tienen muy baja desviación interna (cima roja). b) El mapa de montaña de 100 clusters revela dos grupos fundamentales de biclusters, o 'superclusters', a los lados (figuras generadas con [38]).

Por otro lado, existen algunas aproximaciones para el análisis y visualización de expresión génica que se acercan mucho al concepto de analítica visual. Treeview [40] utiliza varias vistas para representar la matriz de datos, secciones de la matriz, y anotaciones genéticas. HCE [46] incorpora distintos análisis de clustering y métodos de prefiltrado, múltiples visualizaciones (heatmap, coordenadas paralelas, histogramas, mapas de dispersión) y una alta interacción con las visualizaciones. Finalmente, Hawkeye [43], aunque no para expresión génica sino para navegación por genomas, es una aproximación ya puramente basada en analítica visual. Por otro lado, Fry [16] desarrolló un esquema teórico llamado Computational Information Design que recoge la mayoría de las fases del proceso de análisis visual y lo aplica a varios problemas bioinformáticos (análisis de haplotipos, análisis de bloques isométricos, navegación por genomas, etc.)

1.5 DEFINICIÓN DEL PROBLEMA

A partir del estudio del estado del arte en el análisis mediante biclustering y su soporte visual, encontramos en primer lugar que hay una gran abundancia de algoritmos de biclustering. La heterogeneidad de los algoritmos, tanto en sus métodos de búsqueda de biclusters como en la propia definición de qué se considera bicluster, hace complicada su comparación o validación.

Los estudios existentes generalmente sólo tienen en cuenta parámetros externos para validar o comparar los algoritmos. Los métodos preferidos son 1) introducir biclusters en matrices sintéticas y determinar si los algoritmos son capaces de encontrarlos y 2) comparar los biclusters encontrados en matrices reales con datos biológicos existentes, a través de tests de significatividad estadística.

La falta de índices internos adaptados para biclustering implica que sólo podemos validar los algoritmos de biclustering mediante matrices sintéticas o con matrices reales pero sólo al nivel de lo bien que encajan con la información biológicamente existente. Aunque importante, esta validación con matrices reales no parece suficiente por una serie de razones:

- Favorecen a los métodos que encuentran relaciones biológicas ya conocidas
- El conocimiento biológico sobre muchos genes es todavía incompleto, así que, si una relación inferida por un bicluster no aparece en nuestro conocimiento previo, ¿significa que el bicluster es erróneo o que apunta a una nueva relación?
- El conocimiento biológico evoluciona rápidamente. Por ejemplo, la red transcriptómica de la bacteria *E. coli* pasó de 424 genes y 577 interacciones en 2002 [47] a 1278 genes y 2724 interacciones en 2004 [31]. Por tanto, la validación biológica que se haga con estos datos puede ser inestable en el tiempo.
- Los test de significatividad son controvertidos en círculos estadísticos [1, 22]

Por todo ello, es importante complementar estas técnicas de validación externa con otras de validación interna. El primer paso para ello es la definición de índices internos para biclustering.

Por otras parte, las comparativas existentes no tienen en cuenta la configuración inicial de parámetros de los algoritmos, usando generalmente la configuración recomendada por su autor. Sin embargo, es posible que el rendimiento de un algoritmo varíe según su configuración inicial, y por tanto sería interesante comparar los algoritmos con su mejor configuración inicial de parámetros.

Desde el punto de vista de la visualización, estas mismas características especiales del biclustering (agrupación de genes y condiciones, solapamiento) hacen difícil representar múltiples biclusters simultáneamente. La aproximación de [18] necesita la repetición de filas y columnas para visualizar muchos biclusters en un heatmap, lo que puede llevar a ambigüedades o malinterpretaciones de los datos (ver fig. 2b). La aplicación de otras técnicas, como el escalado multidimensional aplicado para clustering [38] (ver fig. 4) tampoco es satisfactorio ya que simplifica demasiado los datos, y nos va a llevar a perder fácilmente gran parte de la información sobre los solapamientos entre biclusters.

Por otro lado, las implementaciones para la visualización de biclusters mediante coordenadas paralelas son mejorables. [2] no reordena las condiciones, con lo que es muy difícil hacerse una idea del patrón representado por el bicluster en cuanto a sus condiciones (ver fig. 3a). [10] corrige esto eliminando todas las condiciones que no están en el bicluster, pero así se pierde gran parte del contexto, ya que pueden existir patrones para estos genes bajo otras condiciones que hayan quedado fuera del bicluster (ver fig. 3b).

Finalmente, existen varias herramientas de mucha calidad para el análisis visual de expresión génica mediante técnicas de clustering [40, 46]. Aunque el proceso analítico es en parte similar, no existe una aproximación satisfactoria para el análisis mediante técnicas de biclustering. Es necesario identificar cuáles son las particularidades del análisis de biclustering (principalmente, la relevancia de los solapamientos) y aplicar la experiencia de estas aproximaciones previas para implementar una solución de análisis visual de los resultados de biclustering aplicados a expresión génica.

1.6 ADAPTACIÓN DEL ESTADÍSTICO HUBERT A BICLUSTERING

El estadístico Hubert (Γ) es una medida para la validación estadística de clusters ([25], pág. 148). Mide la correlación entre dos matrices $n \times n$, X e Y. x_{ij} es la proximidad entre los objetos i y j. En clustering, y_{ij} generalmente es cero si los objetos i y j están agrupados en el mismo cluster y uno si no. El estadístico Γ normalizado es:

$$\bar{\Gamma}(X,Y) = \frac{\frac{1}{k}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}(x_{ij}-\mu_X)(y_{ij}-\mu_Y)}{\sigma_X\sigma_Y} \tag{1.1}$$

donde $k = n(n-1)/2$, μ_X y μ_Y son las medias de las matrices y σ_X y σ_Y son sus varianzas.

La adaptación de este estadístico a biclustering requiere enfrentarse a sus dos características principales: bidimensionalidad (agrupación de genes y condiciones) y solapamiento. Respecto a la bidimensionalidad, necesitamos definir dos pares de matrices, (X^g, Y^g) para genes (filas) y (X^c, Y^c) para condiciones (columnas), obteniendo dos índices, $(\bar{\Gamma}^g, \bar{\Gamma}^c)$. Para obtener el estadístico de Hubert adaptado a biclustering ($\bar{\Gamma}'$), combinamos ambos índices con una media ponderada:

$$\bar{\Gamma}' = \frac{n\bar{\Gamma}^g + m\bar{\Gamma}^c}{n+m} \tag{1.2}$$

donde n es el número de genes y m el número de condiciones. Las matrices de proximidad X^g, X^c se calculan mediante la distancia euclídea de genes y condiciones de acuerdo a los valores de la matriz de expresión A, como en el caso de clustering:

$$x^g_{ij} = \frac{1}{n}\sum_{i=1}^{n-1}\sum_{j=i+1}^{n}\sqrt{\sum_{k=1}^{m}(a_{ik}-a_{jk})^2} \tag{1.3}$$

$$x^c_{ij} = \frac{1}{m}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m}\sqrt{\sum_{k=1}^{m}(a_{ki}-a_{kj})^2} \tag{1.4}$$

Para modelar el solapamiento, la matriz Y usada para clustering se ve sustituida por las matrices de biclustering Y^g, Y^c, con:

$$y_{ij} = 1/(1+k_{ij}) \tag{1.5}$$

donde k_{ij} es el número de biclusters en el que el objeto i (un gene en Y^g, una condición en Y^c) está agrupado junto al objeto j.

1.7 PARAMETRIZACIÓN DE ALGORITMOS DE BICLUSTERING

parámetros típicos para biclustering son el máximo número de biclusters a buscar, el mínimo número de genes o condiciones a incluir, la similitud mínima permitida entre perfiles, etc.

Los índices relativos se suelen utilizar para determinar la mejor elección de parámetros para un algoritmo. Los autores generalmente proponen una configuración óptima para sus algoritmos, que suele ser la utilizada por terceros.

Los índices relativos pueden ser índices externos o internos. Independientemente del tipo de índice, el procedimiento es ejecutar el algoritmo de biclustering con distintas configuraciones paramétricas, y calcular el índice para cada resultado. La configuración con mejor valor del índice se toma como la configuración óptima.

Nuestra propuesta es aplicar el estadístico $\bar{\Gamma}'$ discutido más arriba y la medida F1 [54] para la parametrización. La medida F1 es un índice de validación externo utilizado para determinar cuánto se parecen dos biclusters. [37] extiende F1 para determinar cuánto se parecen dos conjuntos de biclusters, obteniendo la medida SS.

El algoritmo de parametrización propuesto determina la mejor opción de parámetros mediante SS y mediante $\bar{\Gamma}'$:

A: matriz de expresión génica
E: conjunto de biclusters conocido, embebido en A
P: conjunto de configuraciones de parámetros para el algoritmo de biclustering
Calcular X^g y X^c, las matrices de proximidad para los genes y condiciones en A
for cada configuración de parámetros p_i en P **do**
  Ejecutar el algoritmo de biclustering con los parámetros p_i, obteniendo el conjunto de resultados R_i
  Calcular SS_i, que determina la similitud entre E y R_i
  Calcular las matrices Y^c, Y^g para R_i siguiendo la eq. 1.5
  Calcular $\bar{\Gamma}'(X^c, Y^c)$ y $\bar{\Gamma}'(X^g, Y^g)$
  Calcular $\bar{\Gamma}'_i$ siguiendo la eq. 1.2
end for
Seleccionar el p_i correspondiente al SS_i más alto como la configuración de parámetros óptima. Marcar su índice como i1
Seleccionar el p_i correspondiente al $\bar{\Gamma}'_i$ más alto como la configuración de parámetros óptima según $\bar{\Gamma}'$. Marcar su índice como i2
Calcular $\bar{SS}$, la media de para todas las configuraciones de parámetros consideradas.
Almacenar p_{i1}, SS_{i1}, p_{i2}, SS_{i2} y $\bar{SS}$

Figura 5: Algoritmo para la parametrización de biclustering

La mejor parametrización será la de p_{i1}, ya que se realiza a partir de un índice externo, aprovechando la información a priori sobre los biclusters en la matriz. Es esperable que la parametrización p_{i2} sea peor que la de p_{i1}, ya que utiliza un índice interno independiente de la información conocida. Sin embargo, será mejor que el resultado obtenible de media con cualquier configuración.

Vamos a utilizar este procedimiento de parametrización con dos algoritmos de biclustering: Bimax [37] y el modelo de tartán de Turner [54, 55]. Usaremos para ello dos conjuntos de matrices sintéticas, la primera contendrá dos biclusters constantes con grados de solapamiento variables entre 0 % y 100 %. La segunda contiene dos biclusters no solapados, uno constante y otro coherente aditivo, con distintos niveles de ruido.

Una parte importante en la definición del procedimiento es la selección de los rangos de parámetros que se tendrán en cuenta. Deben ser suficientemente amplios para cubrir un gran espectro de combinaciones pero sin ser extremos, para evitar configuraciones inútiles. La tabla siguiente muestra los parámetros seleccionados:

Bimax		Turner	
PARÁMETRO	RANGO (SALTO)	PARÁMETRO	RANGO (SALTO)
Min. filas	3–9 (1)	t_1	0.4–0.8 (0.1)
Min. columnas	3–9 (1)	t_2	0.4–0.8 (0.1)
Umbral para binarización	1–10 % (1)		

Cuadro 2: Rangos seleccionados para los parámetros de los algoritmos de biclustering.

El procedimiento es, por tanto, ejecutar el algoritmo presentado en la fig. 5 para cada nivel de ruido o solapamiento, dependiendo del conjunto de datos. El procedimiento se realiza para los dos algoritmos propuestos. La fig. 6 muestra los resultados.

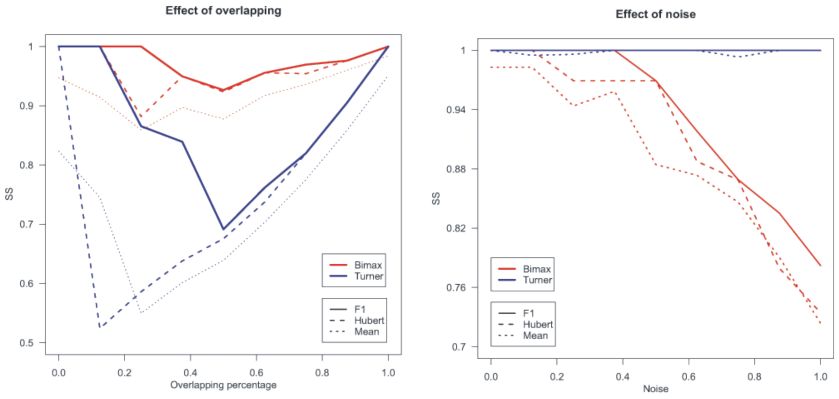


Figura 6: Gráfica con la mejor medida SS obtenida utilizando F_1 y el estadístico Hubert adaptado ($\bar{\Gamma}'$), además de la media de SS para todas las configuraciones. Izquierda: distintos niveles de solapamiento. Derecha: distintos niveles de ruido. El estudio se realiza para Bimax (rojo) y para el modelo de tartán de Turner (azul).

Observando los índices de validación, F_1 (líneas continuas) consigue los mejores resultados de SS gracias a que utiliza la información cono-

cida sobre los biclusters en la matriz. $\bar{\Gamma}'$ (líneas discontinuas) obtiene resultados subóptimos, que a veces coinciden con el mejor resultado de F_1 y mejoran la solución media (línea de puntos).

1.8 VISUALIZACIÓN DE BICLUSTERS

Nuestro objetivo es diseñar una técnica de visualización de biclusters capaz de:

- Mostrar más de diez biclusters, con un número arbitrario de elementos y un grado también arbitrario de solapamiento.
- Mantener en la misma visualización los dos niveles de información (genes y condiciones, y grupos).
- No simplificar o duplicar información. Ambas opciones pueden dar lugar a visualizaciones más directas, pero al coste de perder información o añadir ambigüedad.
- Mejorar la identificación de subgrupos de elementos que se encuentran en varios biclusters (los llamaremos superbiclusters).
- Representar varios conjuntos de resultados de distintos algoritmos de biclustering en la misma visualización.
- Permitir una alta interacción con la visualización, que ofrezca distintos puntos de vista y facilite el análisis exploratorio.

Para conseguir estos objetivos, utilizaremos una representación basada en *grafos dirigidos por fuerzas*. Estos grafos utilizan modelos físicos de fuerzas de atracción y repulsión que emulan fuerzas gravitatorias o elásticas para posicionar los nodos. En los grafos *agrupados* dirigidos por fuerzas, los nodos están de alguna forma envueltos en formas que representan grupos (ver, por ejemplo, [20, 36, 35]), aunque generalmente son grupos no solapados. La técnica de visualización resultante se llamará *Overlapper*.

Para construir el grafo, consideraremos cada gen y condición que se encuentre en al menos un bicluster como un nodo. Todos los nodos agrupados en el mismo bicluster estarán conectados entre sí por una arista, formando un subgrafo *k*-completo (ver fig. 7a).

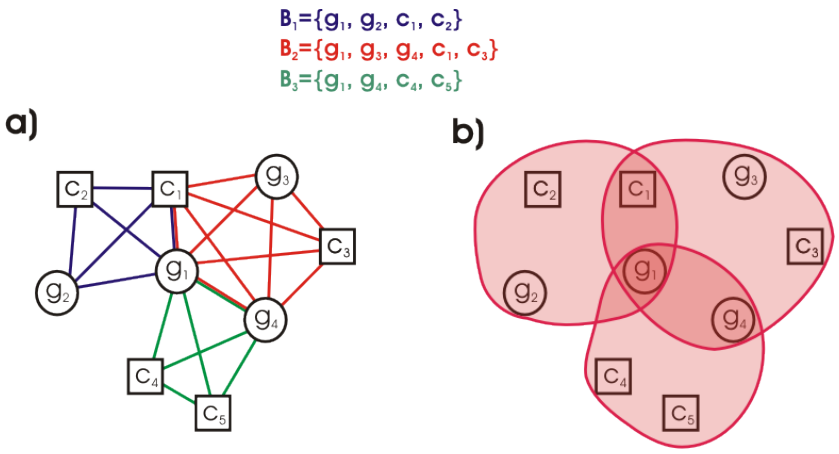


Figura 7: Ejemplo de la estructura de grafo en Overlapper.

Los nodos se representan con una forma distinta para distinguir genes (círculos) de condiciones (cuadrados), pero ambas entidades se

mantienen por lo demás al mismo nivel, facilitando la representación de agrupaciones de ambas, una característica básica del biclustering.

Para simplificar la redacción, hablaremos de nodos para referirnos tanto a genes como a condiciones

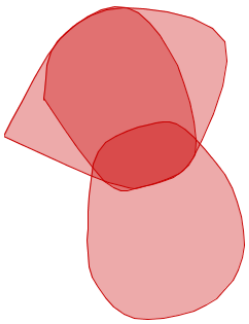
Para evitar la acumulación de aristas en la visualización [21], estas no se dibujan en la representación final. En su lugar, los nodos más externos de cada bicluster se utilizan como puntos de anclaje para dibujar una curva de spline que rodea a todos los nodos del bicluster. El área rodeada por la curva se rellena con un color transparente, de manera que las áreas solapadas tienen un color más sólido, favoreciendo así otra característica básica del biclustering, el solapamiento de grupos. Esto permite distinguir alrededor de cinco niveles de solapamiento (de un solo bicluster a cinco biclusters solapados) de manera sencilla [15].

No obstante, es esperable encontrar solapamientos mayores a cinco grupos en algunos tipos de biclusters, por lo que se establecen algunas codificaciones visuales adicionales. En primer lugar, los nodos se pueden visualizar como diagramas de sectores, con tantos sectores iguales como número de biclusters en los que se encuentra. De este modo, gracias a las fuerzas del diagrama y a los sectores, es sencillo identificar grupos de nodos relacionados (según las leyes sobre la percepción de la Gestalt [8]). En segundo lugar, los nodos que se encuentran exactamente en el mismo número de biclusters se pueden sustituir por un solo nodo (llamado nodo *dual*) con un tamaño proporcional al número de nodos elementales que comprende. Así, simplificamos la visualización al reducir el número de nodos y de aristas subyacentes.

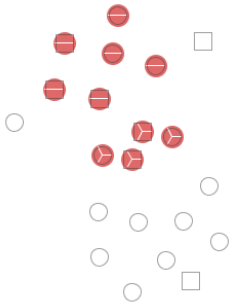
Todas estas codificaciones visuales se encuentran implementadas a modo de *capas*, de manera que el usuario puede añadirlas o eliminarlas según su criterio. En concreto, estas capas son:

- *Capa de nodos*: los nodos dibujados con formas transparentes sencillas.
- *Capa de diagramas de sectores*: los nodos dibujados como diagramas de sectores.
- *Capa de áreas*: los biclusters representados como áreas transparentes con contornos sólidos, rodeando a los nodos agrupados por el bicluster.
- *Capa de etiquetas*: nombres de los nodos y/o biclusters.
- *Capa de detalles*: información textual adicional sobre los nodos, si está disponible (por ejemplo, en el caso de los genes, su definición, organismo, localización, términos GO anotados, etc.)
- *Capa de aristas*: la estructura subyacente de aristas.

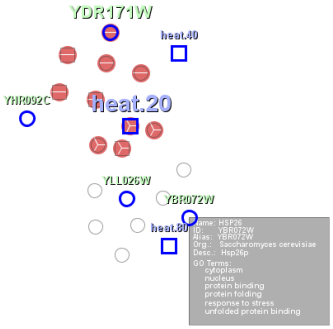
Algunos ejemplos de estas capas se pueden ver en la fig. 8. El hecho de que el usuario puede cambiar y combinar las capas como quiera favorece la implementación del mantra del diseño de interfaces [48]: comenzar con una visión general (capa de áreas), zoom y enfoque (nodos, sectores, etiquetas), detalles bajo demanda (capa de detalles).



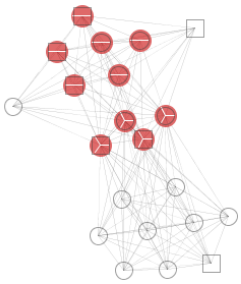
(a) Capa de áreas



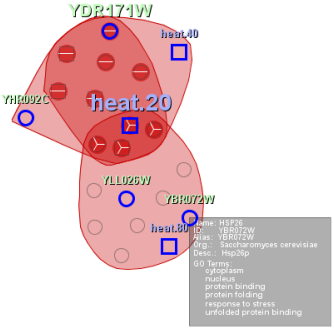
(b) Capas de nodos y sectores



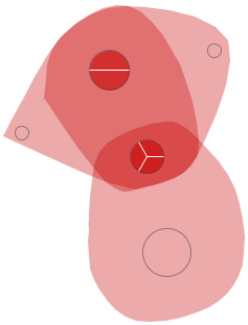
(c) Capa de etiquetas y detalles



(d) Capa de aristas



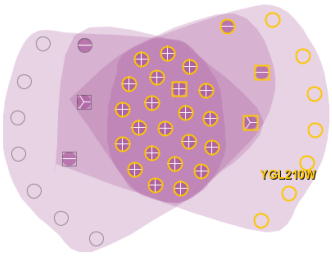
(e) Varias capas superpuestas



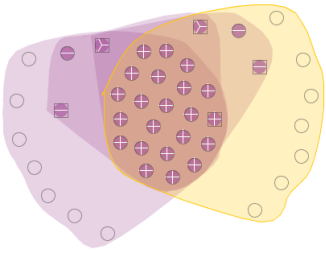
(f) Nodos duales y capa de áreas

Figura 8: Diferentes combinaciones de capas en Overlapper para tres biclusters.

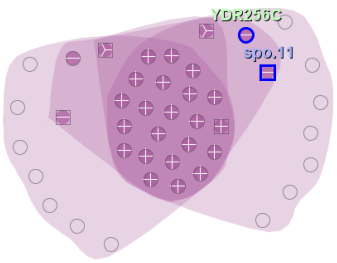
Aparte de la selección de capas, se han implementado otras muchas opciones de interacción con la visualización. En primer lugar, se ofrece una versión en miniatura de la visualización que cumple la doble función de visión general del diagrama y navegación por el mismo (ver fig. 9f). Segundo, tanto los nodos como las áreas pueden sobrepasarse con el ratón de manera que todos los nodos vecinos al nodo sobrepasado, o bien el área sobrepasada, se remarcan (figs. 9a y b). Los nodos y las áreas también se pueden seleccionar, quedando remarcados hasta que se seleccione otro elemento (figs. 9c a e). Una selección con el botón derecho del ratón muestra además los detalles disponibles sobre el elemento. Hay muchos otros detalles de interacción, que se pueden encontrar en la guía de usuario disponible con la herramienta en la que se incluye Overlapper (consultar `http://vis.usal.es/bicoverlapper`).



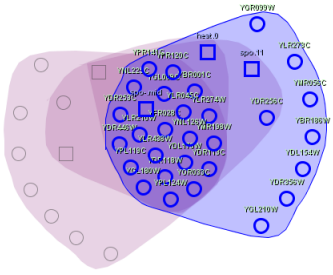
(a) Pasar por encima de un nodo



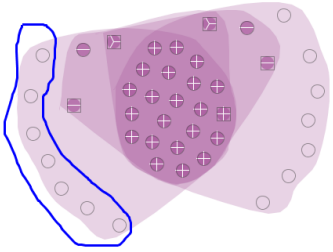
(b) Pasar por encima de un bicluster



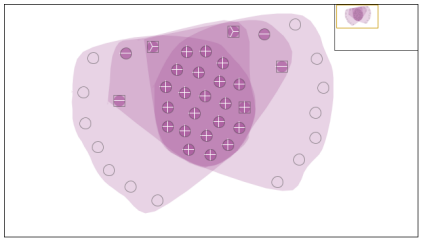
(c) Dos nodos seleccionados



(d) Un bicluster seleccionado



(e) Selección libre de nodos



(f) Visión general para navegación

Figura 9: Ejemplos de interacción con Overlapper.

1.9 ANÁLISIS VISUAL DE EXPRESIÓN MEDIANTE BICLUSTERING

Overlapper es una técnica de visualización diseñada para representar de manera simultánea múltiples biclusters. Sin embargo, para lograrlo descarta la visualización de niveles de expresión y relaja la separación visual entre genes y condiciones. Además, es una técnica de visualización de resultados, pero existen otros tipos de datos asociados al análisis de expresión mediante biclustering, especialmente los datos de entrada y el conocimiento biológico existente. Por todas estas razones, para apoyar el proceso analítico al completo, debemos soportar visualmente todos estos aspectos que no cubre Overlapper.

Las tareas del análisis de expresión génica son responder a las tres preguntas formuladas por Brazma et al. [5], y que se pueden resumir como: *buscar patrones* de expresión, *determinar relaciones* entre genes y *encontrar los roles biológicos* de los genes. En cierta medida, cada una de estas tareas se relaciona directamente con un tipo de dato y una fase del análisis. En los datos de entrada se encuentran los niveles de expresión sobre los que se buscan y confirman los patrones, que a la vez sirven para encontrar nuevas relaciones a través de los datos analizados (clusters, biclusters, etc.). Finalmente, los resultados se confirman consultando fuentes de conocimiento externo y se infieren nuevos roles (ver fig. 10). El proceso es altamente iterativo, ya que una vez obtenidas las agrupaciones volvemos a los datos de entrada para confirmar e inspeccionar los patrones de expresión, quizás volviendo a repetir el análisis; e igualmente los datos biológicos nos pueden llevar a modificar los parámetros o métodos de análisis, o a consultar los patrones de expresión y dirigir nuevos análisis.

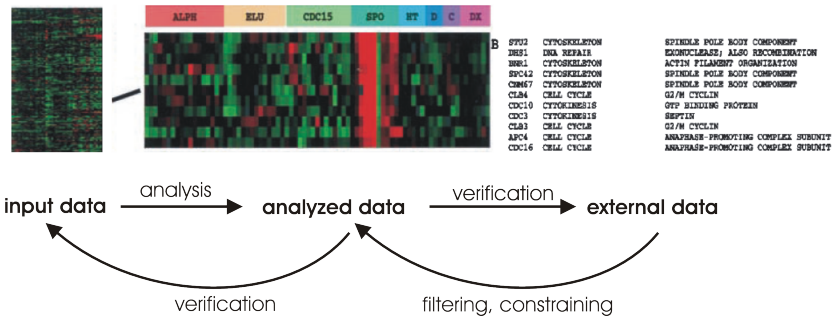


Figura 10: Ejemplo de un proceso simple de análisis: los genes y condiciones de la matriz de expresión de entrada (izquierda) se reordenan siguiendo un clustering (centro) y se etiquetan con el conocimiento disponible (derecha). La imagen superior se ha reproducido a partir de la encontrada en [14].

La conclusión de este estudio es que una herramienta que soporte el análisis visual de análisis de expresión debería implementar técnicas de visualización para cada una de los tres tipos de datos involucrados (matriz de expresión, biclusters, conocimiento biológico), de modo que se faciliten los flujos hacia delante y hacia atrás en el proceso analítico (es decir, que sea sencillo cambiar de una visualización o a otra o, mejor aún, que todas se puedan visualizar simultáneamente). Nuestra selección de técnicas de visualización se resume en la tabla 3.

VISUALIZACIÓN	DATOS	TAREA
Heatmap Coord. paralelas	Matriz de expresión	Buscar patrones
Overlapper Bubblemap	Resultados de biclustering	Buscar relaciones
Grafo de regulación Nube de palabras	Conocimiento biológico	Buscar roles

Cuadro 3: Visualizaciones, datos y tareas para el análisis de expresión génica.

Todas estas técnicas se integran en un entorno de trabajo que hemos llamado BicOverlapper[3] y que permite la interacción fluida entre las visualizaciones y el análisis mediante distintos algoritmos de biclustering (ver fig. 11). Para representar la matriz de expresión, las técnicas de visualización existentes (heatmap y coordenadas paralelas) son satisfactorias y son ampliamente conocidas por los analistas. Los cambios de diseño introducidos en estas técnicas se basan en principios de visualización de la información. En el caso del *heatmap* (fig. 11b), principalmente cambia la escala de colores para representar los niveles de expresión, a una donde es más fácil distinguir los cambios en el tono del color. Respecto a las *coordenadas paralelas* (fig. 11a), adaptamos la visualización para permitir distinguir claramente biclusters con una reordenación a la izquierda de los ejes correspondientes a las condiciones en el bicluster y el resaltado de la sección de polilínea correspondiente a dichas condiciones. Así, no se pierde el contexto del resto del patrón de expresión de los genes seleccionados, y al mismo tiempo se facilita la percepción de la agrupación en el bicluster. Además, el contexto general de la expresión de todos los genes de la matriz se mantiene en el fondo, como un degradado en grises desde la media (línea blanca) hasta cuatro desviaciones estándar de la misma.

Aparte de Overlapper (discutido en la sección 1.8), implementamos otra visualización para los biclusters, que puede servir como resumen de los resultados y ocupa poco espacio de pantalla. Esta técnica es el mapa de burbujas o *bubblemap* (fig. 11c), una adaptación en 2D para biclusters de los mapas de montaña 3D para clusters [38] discutida en 1.4. Cada burbuja representa un bicluster, su posición determinada por los genes y condiciones que agrupa. El color representa el tipo de biclustering (en caso de que se esté visualizando más de uno; en la fig. 11c tenemos 2, en morado y verde). El tamaño representa la dimensión de la submatriz que abarca el bicluster. El grado de transparencia indica la desviación interna de los niveles de expresión del bicluster.

Finalmente, las técnicas para visualizar conocimiento biológico se centran en los dos tipos de conocimiento que se utilizan en la literatura para validar y comparar resultados de biclustering: las redes de regulación transcriptómica y las anotaciones GO. Las *redes de regulación* representan cada gen como un nodo y una arista conecta dos genes si uno de ellos regula al otro. En nuestra representación utilizamos un grafo dirigido por fuerzas (fig. 11d). Las *anotaciones GO* detallan, mediante un vocabulario controlado, los procesos, funciones y componentes asociados a cada gen. Representamos dichas anotaciones en una

3 BicOverlapper es un software de libre distribución disponible en http://vis.usal.es/bicoverlapper

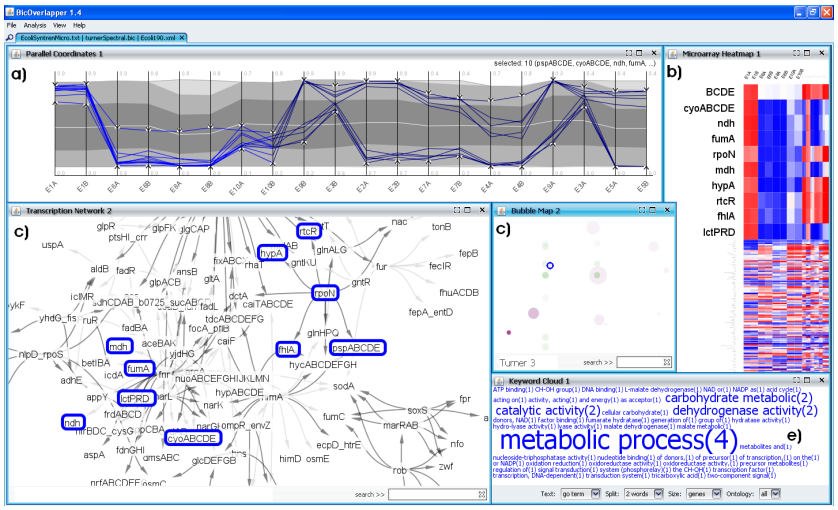


Figura 11: BicOverlapper con cinco de las seis visualizaciones que integra: a) coordenadas paralelas, b) heatmap, c) bubblemap, d) grafo de regulación transcriptómica, e) mapa de palabras. Se ha seleccionado un bicluster en c) -marcado en azul- y la selección se extiende a las entidades afectadas en cada visualización.

nube de palabras [56], con tamaños proporcionales al número de genes seleccionados con cada anotación (fig. 11e).

Todas las visualizaciones se encuentran enlazadas, de modo que la interacción con una de ellas modifica el resto. Este paradigma de múltiples vistas enlazadas mejora la capacidad de adquisición de conocimiento [57]. Hacer la conexión en lugar de dejar que el usuario la haga aligera la carga cognitiva adicional de la interfaz. Además, nuestra capacidad para detectar cambios, incluso en áreas periféricas de la visión, incrementa las posibilidades de detectar modificaciones de patrones en otras vistas enlazadas aunque estemos centrados en una en particular.

El resultado final es un entorno de trabajo que facilita el análisis de la expresión mediante biclustering. Por ejemplo, en la fig. 11 tenemos una matriz de expresión de *E. coli* sobre la que se ha realizado un análisis de biclustering con dos métodos distintos. Para el organismo, tenemos su red de regulación y las anotaciones GO de sus genes. Al seleccionar el bicluster rodeado en azul en 11c podemos comprobar sus patrones de expresión en 11a y b. Es fácil ver que es un patrón de sobreexpresión para dos condiciones (*E1A*, *E1B*) y bajo para el otras condiciones en el bicluster. También se ve rápidamente que uno de los genes no tiene una expresión muy baja realmente en estas otras condiciones, el gen *BCDE* según el heatmap. Además, sobre todo gracias a las coordenadas paralelas, observamos que el perfil del grupo se divide en dos para otras condiciones, aunque coinciden eventualmente en dos de ellas (*E3A* y *E3B*). Si comprobamos la información biológica de la que disponemos, en la red de regulación vemos que efectivamente los 10 genes del bicluster están separados en dos grupos de cinco, aunque en una zona de muchas interacciones. Esto puede estar indicando que, para ciertos procesos relacionados con las condiciones del bicluster, los dos grupos se coordinan. Finalmente, respecto a las anotaciones GO, cuatro de los diez genes están relacionados con procesos metabólicos y algunos otros con la actividad catalítica, la actividad de la dehidrogenasa y el metabolismo de carbohidratos. Podríamos a continuación seleccionar

sólo uno de los dos subgrupos del bicluster mediante las coordenadas paralelas, el heatmap o la red reguladora y ver cuáles se asocian a cada término GO.

Este ejemplo muestra el potencial de la analítica visual para facilitar y acelerar los procesos de análisis de expresión. No obstante, este ejemplo se basa en una matriz sintética de expresión generada con SyNTReN [12] para la bacteria *E. coli*. Vamos a realizar ahora una prueba con la matriz de expresión real utilizada para el estudio de Chen et al. [9] sobre la respuesta al estrés de la levadura *S. pombe*. Esta matriz contiene 10 condiciones, relacionadas con 5 tipos de estrés (dos condiciones por estrés, en diferentes instantes tras la aplicación del estrés). Los objetivos del estudio son detectar los genes que se activan o inhiben para cualquier tipo de estrés (CESR) y aquellos que sólo lo hacen para un tipo (SESR). Realizando un análisis paralelo al de Chen et al. con biclustering, obtenemos resultados similares. En este caso, contamos con la matriz de entrada[4] y las anotaciones GO, que el programa recupera de manera automática de los servicios en red de EntrezGene y QuickGO. Utilizamos el método de biclustering Bimax [37] para buscar los grupos de genes altamente expresados para cada condición. Este y otros algoritmos de biclustering, además de algunas opciones de pre y post-procesamiento, están integrados también en la herramienta[5]. El resultado son cinco biclusters, uno por cada tipo de estrés, con distintos grados de solapamiento (ver fig. 12).

El estudio original de Chen et al. se basa en un análisis de la expresión diferencial de unas condiciones de estrés respecto a otras

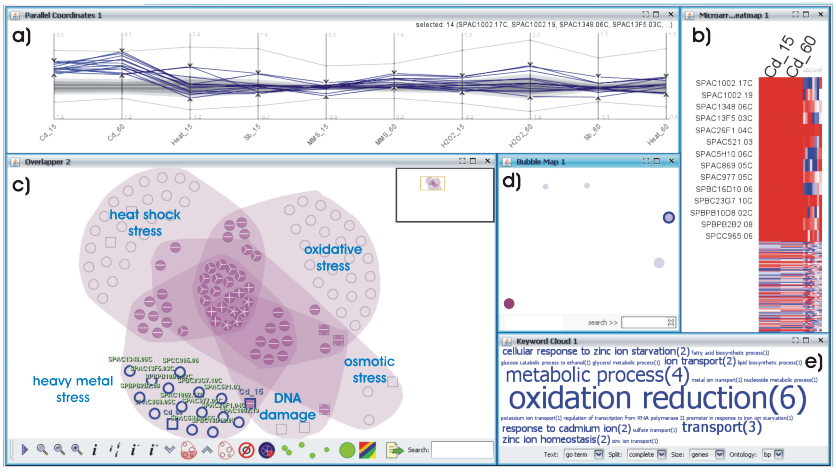


Figura 12: Análisis con BicOverlapper de cinco biclusters asociados a condiciones de estrés: a) coordenadas paralelas, b) heatmap, c) Overlapper, d) bubblemap, e) mapa de palabras. Se ha seleccionado los genes exclusivos del estrés por metales pesados en e) y la selección se extiende a las entidades afectadas en otras visualizaciones.

Es importante resaltar cómo el bubblemap (fig. 12d) resume fácilmente el número y tamaño de los biclusters, pero no representa el verdadero solapamiento, mientras que Overlapper (fig. 12c) lo representa más fiablemente. Esta visualización nos ayuda a observar que:

4 Este experimento está disponible en ArrayExpress con el nombre E-MEXP-29

5 Su implementación corresponde al paquete para R *biclust*, desarrollado en parte durante nuestra investigación.

1. Existe un núcleo central de genes activos bajo todas o casi todas las condiciones de estrés, similar al grupo CESR de Chen et al. (los genes con diagramas de cuatro o cinco sectores)

2. El estrés oxidativo, por metales pesados y por calor son los tipos de estrés que activan más genes. Además, comparten bastantes genes dos a dos (agrupaciones de genes con diagramas de dos sectores en las intersecciones).

3. No hay genes sobreexpresados exclusivamente para el daño en el DNA, y sólo dos lo están bajo estrés osmótico. Sus genes responden también a otras condiciones de estrés.

Además, Overlapper se encuentra enlazado a las otras visualizaciones, ofreciendo mediante la interacción información adicional sobre los patrones de expresión y el conocimiento biológico relacionado. Por ejemplo, si seleccionamos los genes exclusivamente relacionados con el estrés por metales pesados, obtenemos 14 genes con una expresión alta para *Cd_15* y *Cd_60*, las etiquetas de este tipo de estrés (fig. 12a y b). Además, los procesos biológicos relacionados con estos genes tienen que ver con la reducción de la oxidación y con otros procesos asociados a los iones de cadmio y zinc, lo que apoya su relación con los metales pesados.

Incluso considerando que nuestro método de análisis no es exactamente igual al de Chen et al., algunas de sus conclusiones se confirman en nuestra visualización[6]. Por ejemplo, el estrés oxidativo presenta bastante solapamiento con los estreses por metales pesados y por calor. Además, los genes relacionados con daño del DNA están casi completamente incluidos en el grupo de genes relacionados con estrés oxidativo[7].

6 El método de Chen et al. se basa en el análisis diferencial de la expresión génica bajo cada tipo de estrés, y relaciona un gen con un tipo de estrés si está sobreexpresado en una de las dos condiciones relacionadas, y no lo está para el resto de condiciones. Este método es más relajado que nuestro método de biclustering, que requiere que esté sobreexpresado en ambas condiciones del estrés.

7 En nuestro caso, están totalmente incluidos, mientras que Chen et al. identifican dos genes exclusivos del daño en el DNA

Parte II

CONCLUSIONES

2 CONCLUSIONES

El razonamiento lleva a una conclusión, pero ésta no se consolida hasta que la mente la asimila a través de la experiencia. — Roger Bacon

Los resultados principales de esta tesis son el desarrollo de una nueva técnica para la visualización de biclusters y su integración en un marco de trabajo para el análisis visual de datos de expresión génica. Esta nueva técnica de visualización permite inspeccionar múltiples biclusters en una única representación que expresa las principales características del biclustering. Dicha técnica prioriza la visualización de genes y condiciones por encima de los niveles de expresión, lo que permite flexibilizar la representación y mostrar apropiadamente las relaciones entre genes y condiciones inferidas por el biclustering. Como decimos, esta visualización se integra en un marco de trabajo que provee visualizaciones para otros datos relacionados, ya sean las matrices de expresión o conocimiento biológico externo. La combinación de visualizaciones altamente interactivas y el análisis de biclustering agilizan la generación e inspección de los biclusters desde distintos puntos de vista.

Los beneficios de las aproximaciones visuales al análisis de expresión génica son bien conocidos gracias a varias técnicas y herramientas [14, 46, 40, 38]. Algunas de ellas ya apuntaban hacia principios adoptados por la ciencia de la analítica visual. Una primera conclusión de nuestro trabajo es que es posible adaptar formalmente el análisis de expresión génica al análisis visual, con una correcta identificación de los datos, técnicas y objetivos asociados. Desarrollar esta identificación permite diseñar una solución que cubra las necesidades del analista. Tras probar nuestra solución con casos controlados, observamos que es capaz de recuperar las características ya conocidas de los datos de entrada. Cuando la probamos con casos no controlados, incluso sin ser expertos, descubrimos nuevas relaciones genéticas y otras circunstancias, gracias a la combinación de la visualización y el biclustering.

Las técnicas de biclustering son relativamente nuevas en el análisis de expresión génica, y todavía sufren de una falta de estándares y técnicas de visualización, lo cual desanima a sus potenciales usuarios. Al contrario, gran parte del éxito de otras técnicas, como el clustering, reside precisamente en tener algoritmos (clustering jerárquico y k-medias) y visualizaciones (heatmaps y dendrogramas, coordenadas paralelas) ampliamente aceptados. Para obtener el mismo éxito con el biclustering hace falta un proceso largo de validación de los algoritmos de biclustering y de los métodos para visualizar sus resultados. El biclustering y otros métodos avanzados de análisis van a ser cada vez más demandados por la comunidad científica, conforme nuestro entendimiento de la biología aumente. Por ello es importante definir las técnicas y estándares más adecuados.

Un descubrimiento relevante de nuestra investigación es que las comparaciones de métodos de biclustering normalmente no tienen en consideración la configuración inicial de parámetros de los algoritmos comparados. En nuestro trabajo hemos mostrado que, con el

diseño de los índices apropiados y de un procedimiento simple de parametrización, podemos mejorar significativamente el rendimiento de un algoritmo, permitiendo así la comparación de los algoritmos en las mejores condiciones.

Otra conclusión interesante de nuestro proceso de investigación es el hecho de que la biología es un campo muy fértil, lo cual lo hace también muy heterogéneo. Encontramos diversidad en los métodos de biclustering, en las tecnologías de microarray, en el conocimiento biológico, etc. Aunque esta diversidad es buena, ya que aporta muchos puntos de vista y soluciones a cada problema, no debería trasladarse a las estructuras subyacentes utilizadas para su distribución. Los identificadores, formatos de ficheros, convenciones de nombres, etc. deberían mantener una coherencia que sea capaz de gestionar la diversidad en las técnicas y conocimientos. Más aún, la heterogeneidad afecta también a los investigadores. Para mejorar la calidad de los análisis, es necesaria la colaboración de especialistas en el diseño e interpretación de experimentos (biólogos, médicos, etc.) y en el manejo y análisis de datos (estadísticos, informáticos, etc.)

Respecto a los descubrimientos presentados en esta tesis, podemos extraer algunas conclusiones sobre el desarrollo de soluciones basadas en el análisis visual de problemas bioinformáticos. Primero, es clave identificar y verbalizar las cuestiones que queremos resolver y las tareas requeridas para responderlas. Es también fundamental caracterizar los tipos de datos involucrados. Segundo, las decisiones sobre el diseño de técnicas de visualización deben tener en cuenta todas estas entidades detectadas (datos, tareas, cuestiones) y las características de la percepción y del razonamiento humano. Es también muy importante considerar el estado del arte de modo que no reinventemos la rueda y que aprovechemos y respetemos las técnicas ampliamente aceptadas siempre que resulten válidas. Tercero, acerca de la interacción, esta debe ser tan intuitiva como sea posible para no confundir al usuario, pero al mismo tiempo debe ser capaz de ayudar en el discurso analítico. La respuesta a cualquier interacción debe ser rápida, o el usuario sentirá que no tiene el control sobre la interfaz. Las tareas que no puedan ser automáticas deben ser optimizadas, o notificadas apropiadamente. Finalmente, hay que intentar no añadir más ruido al estado del arte. El formato de los datos, por ejemplo, sólo debería rediseñarse si los formatos existentes no satisfacen de ninguna manera nuestras necesidades. Los datos mantenidos por entidades externas y que pueden cambiar con el tiempo, deberían obtenerse a través de dichas entidades, evitando en lo posible copias locales que llevan a un mantenimiento adicional innecesario. El tiempo que pasa un analista reformateando y corrigiendo ficheros, actualizando información, etc. no es despreciable y conviene minimizarlo.

La conclusión general derivada de este trabajo de tesis es que el uso de múltiples vistas enlazadas, altamente interactivas y diseñadas para realizar tareas bien definidas, tiene un enorme potencial para facilitar el análisis y recuperar información relevante. Una parte clave del análisis es la interpretación de las relaciones complejas entre grupos, tales como las derivadas del análisis con biclustering: los genes pueden estar involucrados en varias funciones, colaborando con distintos genes en cada una, así que la visualización de estos grupos solapados es muy valiosa para el estudio de la genética funcional. Esperamos que esta tesis sirva a otros investigadores para considerar el uso del análisis y

visualización de grupos solapados y la analítica visual en sus líneas de investigación.

2.1 TRABAJO FUTURO

Hay varias oportunidades que se abren a partir de los estudios presentados en esta tesis. En primer lugar, el paquete *biclust* puede mejorar, incluyendo algoritmos de biclustering que tuvieron buenas puntuaciones en las comparativas revisadas. Además, puede incluir el índice $\bar{\Gamma}'$ propuesto, y nuestro método para encontrar la mejor configuración de parámetros para un algoritmo de biclustering.

En segundo lugar, queremos probar exhaustivamente el índice interno para validación de biclustering, cubriendo otros algoritmos de biclustering, y quizás realizando nuevas comparativas.

En tercer lugar, BicOverlapper puede crecer en distintas direcciones, de las que destacamos:

- El Desarrollo estrategias más eficientes para visualizar redes de regulación transcriptómica. Este tipo de redes suelen ser muy complejas y difíciles de visualizar. Una posible aproximación es visualizar sólo los elementos que hayamos seleccionado y sus vecinos más cercanos. Otra mejora en las redes es la detección de los motifs de red, que se han utilizado en la literatura como modo de validación y comparación. Estamos considerando también el uso de otras redes biológicas relevantes, como por ejemplo las redes metabólicas.

- Otra dirección interesante es la consideración más detallada de la visualización de la información referente a las condiciones experimentales. El estándar MIAME [4] seguramente sería un buen comienzo en el diseño de visualizaciones de factores experimentales con nubes de palabras, por ejemplo.

- Una tercera línea de crecimiento es el estudio de las métricas para corrección de malas localizaciones, y el estudio del zoom semántico, en la técnica de visualización de Overlapper. La visualización de grupos puede hacerse confusa con conjuntos muy grandes de biclusters (por ejemplo, más de 100 biclusters). Sin embargo, nuestra experiencia es que la capacidad sintetizadora de un algoritmo de biclustering, y la capacidad de percepción que tenemos para analizar grupos tan grandes pueden quedar bastante mermadas en estos casos.

- Finalmente, pero probablemente lo más importante, es la realización de estudios de usabilidad de la herramienta. Actualmente hemos revisado el software junto a algunos expertos en medicina y biología, pero la retroalimentación que se puede obtener de estos estudios seguramente nos dará nuevas pistas de los puntos débiles de la herramienta, y futuras líneas de desarrollo para solucionarlos.

BIBLIOGRAFÍA

[1] D. R. Anderson, K. P. Burnham, and W. L. Thompson. Null hypothesis testing: problems, prevalence, and an alternative. *Journal of Wildlife Management*, 64(4):912–913, 2000. URL http://www.warnercnr.colostate.edu/~anderson/PDF_files/TESTING.pdf. (Cited on page 13.)

[2] S. Barkow, S. Bleuer, A. Prelic, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, 2006. (Cited on pages 11 and 14.)

[3] A. Ben-Dor, B. Chor, R. Karp, and Z. Yakhini. Discovering local structure in gene expression data: the order-preserving submatrix problem. *Journal of Computational Biology*, 10:373–384, 2003. (Cited on page 9.)

[4] A. Brazma, P. Hingamp, J. Quackenbush, G. Sherlock, P. Spellman, et al. Minimum information about a microarray experiment (MIAME)–toward standards for microarray data. *Nat Genet*, 29(4):365–371, 2001. (Cited on page 33.)

[5] A. Brazma, J. Vilo, and E. G. Cesareni. Gene expression data analysis. *FEBS Lett*, 480:17–24, 2000. (Cited on pages 4 and 23.)

[6] A. Califano, G. Stolovitzky, and Y. Tu. Analysis of gene expression microarrays for phenotype classification. *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, 8:75–85, 2000. (Cited on page 9.)

[7] S. K. Card, J. D. Mackinlay, and B. Shneiderman. *Information Visualization: Using Vision to Think*. The Morgan Kaufmann Series in Interactive Technologies. 1999. (Cited on page 7.)

[8] D. Chang, L. Dooley, and J. E. Tuovinen. Gestalt theory in visual screen design. In *Seventh world conference on computers in education*. 2002. (Cited on page 20.)

[9] D. Chen, W. Toone, J. Mata, R. Lyne, G. Burns, et al. Global transcriptional responses of fission yeast to environmental stress. *Mol. Biol. Cell*, 14:214–229, 2003. URL http://www.sanger.ac.uk/PostGenomics/S_pombe/docs/214.pdf. (Cited on page 26.)

[10] K. O. Cheng, N. F. Law, W. C. Siu, and T. H. Lau. Bivisu: Software tool for bicluster detection and visualization. *Bioinformatics*, 2007. (Cited on pages 11 and 14.)

[11] Y. Cheng and G. M. Church. Biclustering of expression data. *Proc. Int'l Conf Intell Syst Mol Biol.*, 8:93–103, 2000. (Cited on page 9.)

[12] T. V. den Bulcke, K. V. Leemput, B.Ñaudts, P. van Remortel, H. Ma, et al. Syntren: a generator of synthetic gene expression data for design and analysis of structure learning algorithms. *BMC Bioinformatics*, 7(43), 2006. URL http://www.biomedcentral.com/1471-2105/7/43. (Cited on page 26.)

[13] D. Duffy and A. Quiroz. A permutation based algorithm for block clustering. *J. Classification*, 8:65–91, 1991. (Cited on page 9.)

[14] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 95:14863–14868, 1998. (Cited on pages 3, 23, and 31.)

[15] J.-D. Fekete and C. Plaisant. Interactive information visualization of a million items. In *IEEE Symposium on Information Visualization*. 2002. (Cited on page 20.)

[16] B. Fry. *Computational Information Design*. Ph.D. thesis, MIT, 2004. URL http://acg.media.mit.edu/people/fry/phd/. (Cited on page 12.)

[17] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. *Proc. Natural Academy of Sciences US*, 97(22):12079–12084, 2000. (Cited on page 9.)

[18] G. A. Grothaus, A. Mufti, and T. Murali. Automatic layout and visualization of biclusters. *Algorithms for Molecular Biology*, 1(15), 2006. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?&pubmedid=16952321. (Cited on pages 10, 11, and 14.)

[19] J. A. Hartigan. Direct clustering of a data matrix. *J. Am. Statistical Assoc.*, 67(337):123–129, 1972. (Cited on page 8.)

[20] J. Heer and D. Boyd. Vizster: visualizing online social networks. In *IEEE Symposium on Information Visualization*. 2005. URL http://csdl2.computer.org/persagen/DLAbsToc.jsp?resourcePath=/dl/proceedings/&toc=comp/proceedings/infovis/2005/2790/00/2790toc.xml&DOI=10.1109/INFOVIS.2005.39. (Cited on page 19.)

[21] Herman, G. Melançon, and M. S. Marshall. Graph visualization and navigation in information visualization: A survey. *IEEE Transactions on Visualization and Computer Graphics*, 6(1):24–43, 2000. (Cited on page 20.)

[22] R. Hubbard. Why we don't really know what "statistical significance"means: a mayor educational failure. *Journal of Marketing Education*, 28:114–120, 2006. URL http://jmd.sagepub.com/cgi/content/abstract/28/2/114. (Cited on page 13.)

[23] A. Inselberg. The plane with parallel coordinates. *The Visual Computer*, 1:69–91, 1985. (Cited on page 11.)

[24] International Human Genome Consortium. Initial sequencing and analysis of the human genome. *Nature*, 409:860–921, 2001. URL http://www.nature.com/nature/journal/v409/n6822/pdf/409860a0.pdf. (Cited on page 3.)

[25] A. K. Jain and R. C. Dubes. *Algorithms for Clustering Data*. Prentice Hall, 1988. (Cited on page 15.)

[26] S. Kaiser and F. Leisch. A toolbox for bicluster analysis in r. Technical Report 028, Ludwing-Maximilians-Universitat München, 2008. (Cited on page 10.)

[27] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual Data Mining: Theory, Techniques and Tools for Visual Analytics*. 2008. (Cited on page 7.)

[28] Y. Kluger, R. Basri, J. T. Chang, and M. Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13:703–716, 2003. (Cited on page 9.)

[29] L. Lazzeroni and A. Owen. Plaid models for gene expression data. Technical Report, Standford University, 2002. (Cited on page 9.)

[30] J. Liu and W. Wang. Op-cluster: Clustering by tendency in high dimensional space. In *3rd IEEE International Conference on Data Mining*, pages 187–194. 2003. (Cited on page 9.)

[31] H.-W. Ma, B. Kumar, U. Ditges, F. Gunzer, J. Buer, et al. An extended transcriptional regulatory network of Escherichia coli and analysis of its hierarchical structure and network motifs. *Nucleic Acids Research*, 32(22):6643–6649, 2004. (Cited on page 13.)

[32] S. Madeira and A. Oliveira. Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions of Computational Biology and Bioinformatics*, 1(1):24–45, 2004. (Cited on pages 4, 8, and 9.)

[33] T. M. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. *Proc. Pacific Symp. Biocomputing*, 8:77–88, 2003. (Cited on page 9.)

[34] Y. Okada, W. Fujibuchi, and P. Horton. A biclustering method for gene expression module discovery using a closed itemset enumeration algorithm. *IPSJ Digital Courier*, 3:183–192, 2007. (Cited on pages 4, 8, and 9.)

[35] H. Omote and K. Sugiyama. Force-directed drawing method for intersecting clustered graphs. In *APVis*, volume 0, pages 85–92. IEEE Computer Society, Los Alamitos, CA, USA, 2007. (Cited on page 19.)

[36] A. Perer and B. Shneiderman. Balancing systematic and flexible exploration of social networks. *IEEE Transactions onf Visualization and Computer Graphics*, 12(5):693–700, 2006. (Cited on page 19.)

[37] A. Prelic, S. Bleuer, P. Zimmermann, A. Wille, P. Bühlmann, et al. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006. URL http://www.tik.ee.ethz.ch/sop/bimax/. (Cited on pages 4, 8, 9, 16, 17, and 26.)

[38] M. Rasmussen and G. Karypis. gcluto: An interactive clustering, visualization and analysis system. Technical Report 04-021, University of Minnesota, 2004. (Cited on pages 11, 12, 14, 24, and 31.)

[39] D. J. Reiss, N. S. Baliga, and R. Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 1:662–671, 2006. URL http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1502140. (Cited on pages 4, 8, and 9.)

[40] A. J. Saldanha. Java Treeview–extensible visualization of microarray data. *Bioinformatics*, 20(17):3246–3248, 2004. (Cited on pages 12, 14, and 31.)

[41] R. Santamaría, R. Therón, and L. Quintales. Bicoverlapper: A tool for bicluster visualization. *Bioinformatics*, 24(9):1212–1213, 2008. (Cited on page 4.)

[42] R. Santamaría, R. Therón, and L. Quintales. A visual analytics approach for understanding biclustering results from microarray data. *BMC Bioinformatics*, 9(247), 2008. URL http://www.biomedcentral.com/1471-2105/9/247. (Cited on page 4.)

[43] M. C. Schatz, A. M. Phillippy, B. Shneiderman, and S. L. Salzberg. Hawkeye: an interactive visual analytics tool for genome assemblies. *Genome Biology*, 8, 2007. URL http://amos.sourceforge.net/hawkeye/. (Cited on page 12.)

[44] E. Segal, A. Battle, and D. Koller. Decomposing gene expression into cellular processes. *Proc. Pacific Symp. Biocomputing*, 8:89–100, 2003. URL citeseer.ist.psu.edu/segal03decomposing.html. (Cited on page 9.)

[45] E. Segal, B. Taskar, A. Gasch, N. Friedman, and D. Koller. Rich probabilistic models for gene expression. *Bioinformatics*, 17:S243–S252, 2001. (Cited on page 9.)

[46] J. Seo and B. Shneiderman. A rank-by-feature framework for unsupervised multidimensional data exploration using low dimensional projections. *IEEE Symposium on Information Visualization*, pages 65–72, 2004. URL http://ieeexplore.ieee.org/search/wrapper.jsp?arnumber=1382892. (Cited on pages 12, 14, and 31.)

[47] S. S. Shen-Orr, R. Milo, S. Mangan, and U. Alon. Network motifs in the transcriptional regulation network of escherichia coli. *Nature Genetics*, 31:64–68, 2002. (Cited on page 13.)

[48] B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *IEEE Visual Languages*, UMCP-CSD CS-TR-3665, pages 336–343. College Park, Maryland 20742, U.S.A., 1996. URL citeseer.ist.psu.edu/shneiderman96eyes.html. (Cited on pages 7 and 20.)

[49] A. Tanay, R. Sharan, M. Kupiec, and R. Shamir. Revealing modularity and organization in the yeast molecular network by integrated analysis of highly heterogeneus genomewide data. *Proceedings of the National Academy of Sciences of the United States of America*, 101(9):2981–2986, 2004. (Cited on pages 4 and 9.)

[50] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18:S136–S144, 2002. (Cited on pages 8 and 9.)

[51] C. Tang, L. Zhang, and M. Ramanathan. Interrelated two-way clustering: An unsupervised approach for gene expression data analysis. In *In Proceedings of the 2nd IEEE International Symposium on Bioinformatics and Bioengineering*, pages 41–48. 2001. (Cited on page 9.)

[52] J. J. Thomas and K. A. Cook. *Illuminating the Path: The Research and Development Agenda for Visual Analytics*. IEEE Press, 2005. (Cited on pages 3 and 7.)

[53] R. Tibshirani, T. Hastie, M. Eisen, D. Ross, D. Botstein, et al. Clustering methods for the analysis of DNA microarray data. Technical Report, Dept. of Health Research and Policy, Dept. of Genetics, Dept. of Biochemistry, Stanford Univ., 1999. URL citeseer.ist.psu.edu/tibshirani99clustering.html. (Cited on page 9.)

[54] H. Turner, T. Bailey, and W. Krzanowski. Improved biclustering of microarray data demonstrated through systematic performance tests. *Computational Statistics and Data Analysis*, 48:235–254, 2003. (Cited on pages 9, 16, and 17.)

[55] H. L. Turner, T. C. Bailey, W. J. Krzanowski, and C. A. Hemingway. Biclustering models for structured microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 2(4):316–329, 2005. (Cited on pages 9 and 17.)

[56] F. B. Viegas and M. Wattenberg. Tag clouds and the case for vernacular visualization. *Interactions*, 15(4):49–52, 2008. URL http://portal.acm.org/citation.cfm?id=1374501. (Cited on page 25.)

[57] C. Ware. *Information Visualization: Perception for Design*. Diane Cerra, 2nd edition, 2004. (Cited on pages 3, 7, 11, and 25.)

[58] J. Yang, W. Wang, H. Wang, and P. Yu. d-clusters: Capturing subspace correlation in a large data set. In *Proc. 18th IEEE Int'l Conf. Data Engineering*, pages 517–528. 2002. (Cited on page 9.)

[59] J. Yang, W. Wang, H. Wang, and P. Yu. Enhanced biclustering on expression data. In *Third IEEE Conf. Bioinformatics and Bioengineering*, pages 321–327. 2003. (Cited on page 9.)