Biclustering sobre datos de expresión génica basado en búsqueda dispersa

Juan Antonio Nepomuceno Chamorro,
janepo@us.es

Supervisado por los profesores doctores:
Alicia Troncoso Lora, Jesús S. Aguilar Ruiz

Tesis enviada a la comisión académica del programa de doctorado de Ingeniería Informática, bajo la línea de investigación Ingeniería y Tecnología del Software, siguiendo los requerimientos para su lectura en el Departamento de Lenguajes y Sistemas Informáticos.

(Tesis Doctoral)

Índice general

Índice de figuras

Índice de tablas

Parte I

Introducción

Capítulo 1

Introducción

Lee mucho, mantén tu mente activa, mantén tus antenas desplegadas: cuando te informen de algo interesante, indaga. Como Luis Pasteur, la suerte favorece a la mente preparada.
Cartas a una joven matemática. Ian Stewart (pág. 168)

1.1. Motivación

La Física durante el siglo XX avanzó de tal manera que motivó tanto el desarrollo de otras materias, como la Matemática o la Ingeniería, como un avance de tipo social. Basta tener en cuenta la energía atómica y todas las consecuencias que de ella se produjeron. La Biología está llamada a cumplir ese mismo papel durante el siglo XXI. La irrupción de nuevas tecnologías motiva nuevos retos y éstos, a su vez, implican el desarrollo e incluso el nacimiento de otras disciplinas.

La Informática juega un papel central en este proceso de avance científico y tecnológico. Se puede establecer una doble vía de interacción entre Biología e Informática. Así por ejemplo, la Computación Evolutiva usa una metáfora biológica como fuente de inspiración y motiva el desarrollo de las técnicas computacionales que conocemos como bioinspiradas. Por otro lado, el análisis de los datos de tipo ómico, de los que hablaremos en el siguiente capítulo, motiva el desarrollo la Bioinformática como disciplina. Las necesidades de definición de modelos predictivos basados en los nuevos tipos de datos motiva el nacimiento de nuevas técnicas de Minería de Datos. Y muchos más ejemplos que pueden ilustrar esta doble vía de comunicación entre la Biología y la Informática como fuente de conocimiento emergente.

La motivación del trabajo presentado en esta tesis se puede dibujar en un triángulo con tres vértices: *Minería de Datos*, *Computación Evolutiva*

y *Bioinformática*. La Minería de Datos (o *Data Mining*) surge en los años noventa en el contexto del descubrimiento de conocimiento en las bases de datos (*Knowledge Discovery in Databases* o KDD). Tiene como principal labor el descubrimiento de conocimiento previamente desconocido a partir de grandes volúmenes de datos. Fusiona aspectos de la Estadística y el Aprendizaje Automático y su principal labor es llevar a cabo un proceso de ingeniería que dé el paso de la Información al Conocimiento. La Computación Evolutiva utiliza la metáfora de la teoría de la evolución darwinista como marco de referencia de una serie de algoritmos de optimización. Estos algoritmos son de tipo aproximado es decir, no encuentran la solución exacta del problema de optimización sino una aproximación a la misma. Permiten tratar aquellos problemas de optimización computacionalmente intratables o *NP-completos* para los que no existe un algoritmo exacto que los resuelva. La Bioinformática consiste en el estudio y desarrollo de técnicas que permitan, entre otros aspectos, el descubrimiento de nuevos biomarcadores o indicios que permitan tanto el pronóstico como el diagnóstico de enfermedades. El desarrollo de metodologías, tanto predictivas como de clasificación, usando *datos ómicos* constituye una de las principales tareas de la Bioinformática en la actualidad.

Los datos de expresión génica obtenidos mediante la tecnología de microarray son un tipo particular de datos ómicos. Esta tecnología permite monitorizar el nivel de expresión de un grupo de genes de una muestra obtenida en el laboratorio. De esta forma se pueden generar matrices de expresión donde cada elemento refleja el nivel de expresión de un gen bajo una determinada condición experimental. Este tipo de datos constituyen una fotografía del proceso de *transcripción*, mediante el que el ADN se transforma en ARN, por eso son datos transcriptómicos, por lo que constituyen una herramienta muy interesante para el estudio de procesos biológicos y el descubrimiento de biomarcadores en general. Los conjuntos de datos, o *datasets*, de este tipo tienen como característica un número muy elevado de genes, del orden de miles, frente a un número reducido de condiciones, decenas, lo que determina el tipo de análisis que se efectúen sobre ellos. Además, por otro lado, la propia naturaleza del problema biológico condiciona la tarea misma que se debe realizar.

El *Biclustering* es una técnica de aprendizaje no supervisado que agrupa tanto genes como condiciones. Este doble agrupamiento lo diferencia del *clustering* tradicional sobre este tipo de datos ya que sólo agrupa o bien genes o condiciones. El objetivo es el descubrimiento de patrones locales más que la división en grupos de los datos para una posterior clasificación. Mediante el biclustering se pueden descubrir grupos de genes que se expresen de una

manera similar bajo unas determinadas condiciones. Es decir, se trata de un problema de búsqueda de patrones para determinar qué genes se activan o desactivan cuando ocurren unas condiciones concretas objeto de estudio. Computacionalmente es un problema intratable o NP-completo.

La motivación de la tesis presentada es la elaboración de una técnica de biclustering mediante una metaheurística evolutiva, concretamente la búsqueda dispersa (o *scatter search*). Se trata de un problema de Minería de Datos, el biclustering, abordado como un problema de optimización que se resuelve mediante una técnica evolutiva, la búsqueda dispersa, en el contexto del análisis de datos de expresión génica. Por lo tanto, el trabajo desarrollado constituye un claro ejemplo del triángulo antes descrito que tiene la Minería de Datos, la Computación Evolutiva y la Bioinformática como vértices.

1.2. Planteamiento

El punto de partida de la tesis presentada es el trabajo presentado en [1]. En este artículo se demuestra que la medida que usualmente se utiliza como criterio de evaluación en los algoritmos de biclustering, el residuo cuadrático medio (o *Mean Squared Residue*), no detecta cierto tipo de patrones importantes desde el punto de vista biológico. Estos patrones son aquellos patrones con un escalado pronunciado entre los genes.

El inicio del trabajo lo constituye el estudio de los criterios de búsqueda de biclusters así como el desarrollo de un algoritmo que permita su posterior uso. Tras varias aproximaciones, se plantea un algoritmo de biclustering basado en una metaheurística de búsqueda dispersa. La propuesta *SSCorr* presenta la adaptación del esquema de búsqueda dispersa al contexto del biclustering. Se plantea el problema como un problema de optimización cuya función objetivo basa su criterio de calidad en la correlación lineal. Los resultados se analizan de una manera estándar utilizando tres conjuntos de datos y, por otro lado, se presentan la comparación con los algoritmos usados como marco de referencia en la mayoría de los algoritmos de biclustering.

Los buenos resultados obtenidos, junto con ciertas mejoras en la función de calidad del algoritmo *SSCorr*, motivan una segunda propuesta *BISS* que captura patrones de activación e inhibición de genes. En este paso se hace un especial énfasis en el análisis de los resultados y sobre todo en la comparación con otras propuestas de la bibliografía, tanto los algoritmos generalmente utilizados como marco de referencia como otros algoritmos de biclustering basados en correlaciones lineales. Una de las tareas más complicadas en los trabajos de biclustering lo constituye las comparaciones entre algoritmos. Como se tratan de técnicas no supervisadas, donde no hay una medida de

la precisión, se debe utilizar el conocimiento experto para la comparación entre los resultados de los algoritmos. Para manejar dicho conocimiento se emplean repositorios de genes como *Gene Ontology* (GO), mediante los que se puede asociar una funcionalidad biológica a un grupo de genes. Dicho estudio recibe el nombre de *estudio de enriquecimiento de genes*. En función de si los biclusters encontrados representan o no una función se establece un ranking entre los resultados obtenidos por los distintos algoritmos. Dicho ranking se establece según el porcentaje de biclusters enriquecidos que cada algoritmo encuentra. El análisis biológico de los resultados constituye una de las principales motivaciones para el trabajo que se presenta en BISS. En este trabajo no sólo se amplía la propuesta anterior y se realiza un estudio comparativo más especializado, sino que se profundiza en la comparación de los resultados desde un punto de vista biológico.

La experiencia adquirida en las propuestas de SSCorr y BISS, junto con un análisis de bibliografía existente, plantea una nueva propuesta, denominada como *GoldBinch*, que integra la información biológica almacenada en GO, no ya como información que se use a posteriori en la fase de análisis de los resultados, sino como criterio de la búsqueda en sí. Desde nuestro conocimiento, no existen algoritmos de biclustering que integren información biológica como parte del criterio de búsqueda de resultados. La integración de información de distintas fuentes de datos, usando los repositorios públicos existentes, es una de las últimas tendencias en bioinformática. Se plantea un algoritmo de búsqueda dispersa para biclustering cuya función objetivo añade un término extra que refleja, por cada bicluster que se evalúe, la calidad de ese grupo de genes según su información almacenada en GO. Se estudian dos posibilidades para dicho término de integración de información biológica, se comparan entre sí y se comprueba que los resultados son mejores cuando se usa información biológica. Junto con la evaluación estándar usada en biclustering, se presenta también un análisis de tipo cualitativo que permite apreciar las diferencias entre las dos formas propuestas de integrar de la información biológica.

1.3. Objetivos

Los objetivos concretos de la presente tesis son:

- Estudio de los patrones de desplazamiento y escalado como punto de partida. Elección de un criterio de calidad para el proceso de optimización que permita la búsqueda de biclusters con dichos patrones.

- Desarrollo de un algoritmo de biclustering basado en una metaheurística de búsqueda dispersa.
- Estudio de otro tipo de patrones biológicos no contemplados anteriormente.
- Comparación del algoritmo propuesto con otros algoritmos de la bibliografía, tanto de tipo general como aquellos basados en la correlación lineal.
- Desarrollo de un algoritmo de biclustering que integre información biológica junto un análisis de los resultados tanto cuantitativo como cualitativo.

1.4. Principales contribuciones

Las propuestas realizadas en esta tesis han dado lugar a resultados publicados tanto en revistas como en conferencias nacionales e internacionales.

Los resultados publicados en revistas son:

- Un algoritmo de biclustering basado en un esquema de búsqueda dispersa. La función objetivo se basa en correlaciones lineales entre los genes y captura patrones de desplazamiento y escalado. Los algoritmos de biclustering existentes en la bibliografía hasta ese momento se basaban en el residuo (MSR) como medida de calidad. Los resultados experimentales y las comparaciones con otros algoritmos se presentan siguiendo las pautas generales de otros artículos de la bibliografía. La propuesta SSCorr descrita en los capítulos VII y IX se presenta en este trabajo:

    - *"Biclustering of Gene Expression Data by Correlation-Based Scatter Search"*. Juan A. Nepomuceno et al. BioData Mining, 2011, 4, 3. DOI: 10.1186/1756-0381-4-3, Impact Factor[1]: 1.54. Cuartil: Q2 (Categoría: *Mathematical and Computational Biology*). Citas: (23 citas según Google Scholar (01-05-2015)).

- Un algoritmo que mejora la función objetivo del algoritmo anterior, así como otros aspectos del esquema de búsqueda dispersa. Se capturan de esta forma comportamientos de activación-inhibición entre genes no contemplados anteriormente. El estudio experimental presentado hace

[1]La revista no tenía índice de impacto el año de publicación del artículo.

especial hincapié en el contexto biológico. La comparativa realizada se ha efectuado respecto a los algoritmos clásicos de biclustering así como también respecto a otros algoritmos basados en la correlación. La propuesta BISS descrita en los capítulos VII y X se presenta en este trabajo:

  - *"Scatter Search-based identification of local patterns with positive and negative correlations in gene expression data"*. Juan A. Nepomuceno et al. Applied Soft Computing. Impact Factor: 2.6. Cuartil: Q1 (Categoría: *Computer Science and Artificial Intelligence*).

- Un algoritmo de biclustering que integra información biológica en el criterio de búsqueda. El algoritmo se basa en un esquema de búsqueda dispersa que permite intercambiar varias medidas de integración de información biológica. El estudio experimental se lleva a cabo con un doble objetivo: mostrar que la integración de información biológica mejora los resultados, así como estudiar las diferencias entre dos posibles formas de integrar la información. Se realiza un estudio tanto cuantitativo como cualitativo de los resultados. Así mismo, los resultados se comparan con los algoritmos clásicos de biclustering. La propuesta GoldBinch descrita en los capítulos VIII y XI se presenta en este trabajo:

  - *"Integrating biological knowledge based on functional annotations for biclustering of gene expression data"*. Juan A. Nepomuceno et al. Computer Methods and Programs in Biomedicine, 2015 May; 119(3):163-80. DOI: 10.1016/j.cmpb.2015.02.010, Impact Factor: 1.093. Cuartil: Q1 (Categoría: *Computer Science, Theory and Methods*).

Así mismo, los resultados publicados en congresos nacionales e internacionales son:

- Un estudio de métodos clásicos de optimización local para la detección de patrones de desplazamiento y escalado. Las publicaciones asociadas son:

  - *"Biclusters Evaluation Based on Shifting and Scaling Patterns''*. IDEAL 2007: 8th International Conference on Intelligent Data

Engineering and Automated Learning. Birmingham, UK, December, 16-19, 2007, pp. 840-849.

- *"Patrones en Biclusters usando Técnicas de Optimización sin Restricciones"*. MAEB 2007: Metaheurísticas, Algoritmos Evolutivos y Bioinspirados. Tenerife, España, pp. 413-417.
- *"Evaluación de Biclusters basada en patrones de desplazamiento y escalado"* Juan A. Nepomuceno et al. I Workshop Español sobre Extracción y Validación de Conocimiento en Bases de Datos Biomédicas (EvaBio 2007). Salamanca, Spain.

- Un algoritmo de biclustering basado en un esquema híbrido, entre una búsqueda dispersa y un algoritmo genético, que usa el residuo cuadrático (MSR) como criterio de búsqueda de biclusters. Las publicaciones asociadas son:

  - *"A Hybrid Metaheuristic for Biclustering Based on Scatter Search and Genetic Algorithms"*. PRIB 2009: Pattern Recognition in Bioinformatics, 4th IAPR International Conference. Sheffield, UK, September 7-9, 2009, pp 199-210.
  - *"Un algoritmo de Biclustering Basado en Búsqueda Dispersa y Algoritmos Genéticos"* Juan A. Nepomuceno et al. II Workshop Español sobre Extracción y Validación de Conocimiento en Bases de Datos Biomédicas (EvaBio 2009). Sevilla, Spain.

- Estudio del solapamiento entre los biclusters obtenidos por el algoritmo anteriormente presentado. Se analiza el papel de un factor de corrección del solape en la función objetivo.

  - *"An Overlapping Control-Biclustering Algorithm from Gene Expression Data"*. ISDA 2009: International Conference on Intelligent Systems Design and Applications. Pisa, Italy, November 30-December 2, 2009, pp 1239-1244.

- Un algoritmo de biclustering inspirado en un esquema de búsqueda dispersa que utiliza la correlación como mecanismo de búsqueda de biclusters.

  - *"Evolutionary metaheuristic for biclustering based on linear correlations among genes"*. SAC 2010: Proceedings of the 2010 ACM Symposium on Applied Computing (SAC). Sierre, Switzerland, March 22-26, 2010, 1143-1147.

- Un primer algoritmo de biclustering basado en una búsqueda dispersa. El algoritmo es una modificación de algoritmo anterior de manera que se incorpora un método de la mejora en el esquema. Las publicaciones asociadas son:

  - *"Correlation-Based Scatter Search for Discovering Biclusters from Gene Expression Data"*. EvoBIO 2010: Proceedings of the 8th European Conference on Evolutionary Computation, Machine Learning and Data Mining. Istanbul, Turkey, April 7-9, 2010, pp 122-133.
  - *"Búsqueda Dispersa aplicada al descubrimiento de patrones en datos de expresión genética"*. MAEB 2010: Metaheurísticas, Algoritmos Evolutivos y Bioinspirados, Valencia, España.

- Estudio de una búsqueda local que mejore los resultados del algoritmo propuesto anteriormente.

  - *"A local search in Scatter Search for improving Biclusters"*. Juan A. Nepomuceno et al. 3th Congress on Natural and Biologically Inspired Computing (NaBIC 2011). Salamanca, Spain, 2011, pp 521-526.

- Estudio de la estructura interna de los biclusters obtenidos así como su utilización para la generación de redes de co-expresión de genes. Publicaciones asociadas:

  - *"Inferring gene coexpression networks with Biclustering based on Scatter Search"*. Juan A. Nepomuceno et al. 11th International Conference on Intelligent Systems Design and Applications (ISDA 2011). Córdoba, Spain, 2011, 1091 -1096.
  - *"Inferring gene co-expression networks with Biclustering based on linear correlations among genes"*. Juan A. Nepomuceno et al. Póster en Benelux Bioinformatics Conference (BBC 2011). CRP Santé Luxemburgo.

- Una primera aproximación a un algoritmo de biclustering que integre información biológica como criterio de búsqueda.

  - *"GoldBinch: a scatter search-based biclustering of gene expression data algorithm that integrates biological knowledge with functional annotations"*. Juan A. Nepomuceno et al. Póster en XII Symposium on Bioinformatics 2014. Sevilla, Spain.

1.5. Estructura de la memoria

La memoria de la tesis se organiza por capítulos de la siguiente manera.

En el capítulo II se presenta el contexto de la tesis. Se introduce el dogma central de la Biología Molecular, la revolución de los datos ómicos y qué se entiende por Bioinformática en la actualidad. Así mismo, se proporciona un glosario de términos con los conceptos de Biología necesarios para entender el marco de trabajo.

El estado del arte se compone de tres apartados, los capítulos III y IV dedicados a los datos de expresión génica y al estado del arte de biclustering respectivamente, así como unos apuntes sobre trabajos que integran información biológica en el capítulo V. En el capítulo III se presenta la tecnología de microarray y se muestra cómo es el flujo de trabajo normal con estos datos. Este capítulo tiene como objetivo la exposición de la dificultad de trabajo con este tipo de datos así como la experiencia adquirida en su manejo. La idea central del capítulo es contextualizar el biclustering como una técnica de análisis de datos de expresión génica a alto nivel. En el capítulo IV se presentan el problema del biclustering así como un resumen exhaustivo de los principales algoritmos desarrollados estos últimos años. El capítulo V contiene unos pequeños apuntes sobre técnicas en las que se integra información biológica en su funcionamiento. Así mismo, se apuntan varios trabajos sobre medidas de similitud entre genes basadas en la información almacenada en la ontología de genes GO.

En el capítulo VI se explica el motor de búsqueda de los algoritmos que se presentan en la tesis. Se basa en una metaheurística de búsqueda dispersa adaptada al problema del biclustering. Se explica en detalle su funcionamiento, así como los distintos procedimientos internos y su adaptación al problema.

En el capítulo VII se presentan las dos propuestas SSCorr y BISS que utilizan la correlación lineal como criterio para la búsqueda de biclusters. SScorr es una primera aproximación que encuentra patrones de desplazamiento y escalado. BISS modifica a SSCorr de manera que incluye también patrones de activación-inhibición. Estos patrones reflejan genes que tengan un patrón de expresión complementario, es decir, la inhibición implica la activación de otro gen.

En el capítulo VIII se presenta la propuesta GoldBinch, que incorpora información biológica como criterio de búsqueda de biclusters. Además de la matriz de expresión, se incorpora como datos de entrada un fichero de anotaciones directa que asocia a cada gen un término con un significado biológico. Por ejemplo, cada gen se asocia con un conjunto de términos GO.

De esta forma se incorpora, a través de una medida de similitud funcional entre genes, información biológica almacenada en repositorios públicos como GO o *Kyoto Encyclopedia of Genes and Genomes* (KEEG). En este capítulo se exponen también las modificaciones necesarias en la búsqueda dispersa relativas a la incorporación de este criterio de búsqueda.

Los capítulos IX, X y XI muestran los experimentos realizados y los resultados obtenidos con las propuestas anteriores. En el capítulo IX se expone la experimentación llevada a cabo con la propuesta SSCorr. Dicha experimentación sigue las pautas estándares de los artículos de biclustering. En el capítulo X, por contra, se hace especial énfasis en el contexto biológico del problema. La validación de los resultados se lleva a cabo teniendo en cuenta todas las ramas de la ontología GO así como refinando el concepto de bicluster enriquecido. Es decir, se validan los resultados teniendo especialmente en cuenta el contexto del problema. Así mismo, se incluye comparativas con algoritmos recientes de biclustering basados también en la correlación y que no se habían estudiado en la experimentación del capítulo anterior.

En el capítulo XI se presenta el estudio experimental realizado con la propuesta GoldBinch. Los experimentos se han realizado con un doble objetivo: verificar que la integración de información biológica mejora los resultados del proceso de biclustering y, por otro lado, comparar las dos posibilidades de integración contempladas en GoldBinch. Los resultados han sido estudiados en función del *enriquecimiento* de los biclusters obtenidos. Con vistas a mostrar las diferencias entre las dos medidas de integración biológica, también se ha considerado la naturaleza de dicho enriquecimiento, en concreto el número de términos GO asociados a cada bicluster enriquecido. Así mismo, y partiendo de esta misma motivación, se añade un estudio cualitativo de algunos biclusters de manera que se observe claramente las diferencias entre las dos medidas de integración. La experimentación llevada a cabo también incluye una comparativa con los algoritmos clásicos de biclustering.

Finalmente, en el capítulo XII se exponen las conclusiones la tesis, así como las futuras líneas de investigación que el trabajo presentado invita a explorar.

Capítulo 2

Sobre Bioinformática en el siglo XXI

La consecuencia de todos estos progresos es que en el centro mismo de la Biología y la Medicina ha surgido una nueva ciencia a la que podríamos llamar criptografía del ADN. Hemos interceptado un mensaje muy sofisticado que reviste una importancia crucial para el futuro de la especie humana. Está escrito en un código extraño y aparentemente indescifrable, engañosamente en su uso de tan sólo cuatro letras, pero lo bastante complejo como para que sean necesarias varias décadas antes de que una combinación de ingenio, investigación de laboratorio y un complicado análisis por medio de las más potentes supercomputadoras acaben por desvelar todos los secretos del código. Pero ¡qué fantástica aventura!
El lenguaje de la Vida: el ADN y la Revolución de la Medicina Personalizada. Francis Collins (pág. 36)

2.1. Introducción

Recientemente aparecía un artículo en prensa[1] sobre la creatividad. En este artículo se exponían una serie de citas sobre cómo surgían las ideas y se abría la reflexión con un ejemplo concreto: la experiencia de un concierto en Estocolmo en el que, según el autor, se había conseguido una experiencia musical trascendente. El intérprete era un virtuoso del violín perteneciente a la escuela de violinistas de San Petersburgo, fundada en 1868 por Leopoldo Auer, que tocaba una obra de Bach de 1720 compuesta para violín y lo

[1] `http://cultura.elpais.com/cultura/2014/11/19/babelia/1416418422_239455.html`

hacía con un Stradivarius de 1717. La triple conjunción de tres fenómenos independientes entre sí, de distintos ámbitos, un lutier italiano, un compositor alemán y el fundador de una escuela de intérpretes ruso, constituía la explicación, según el autor, de aquel proceso de creación único. El artículo continuaba con una colección de citas sobre los procesos creativos, de la que destacamos:

"En la frontera se cree peor y se crea mejor"

El siglo XXI comenzó con dos revoluciones tecno-científicas que están llamadas a condicionar el discurrir de las ciencias durante todo el siglo. Por un lado la revolución de las nuevas tecnologías, surgida en el seno de la Informática como disciplina, por otro lado la revolución que supuso el Proyecto Genoma Humano[2] en la Biología. La confluencia de ambas revoluciones transciende la idea de no sólo aplicar una determinada técnica o teoría a un problema, sino que constituye el nacimiento de nuevas preguntas totalmente diferentes. El verdadero avance científico no vendrá de la aplicación de conocimientos previos a ámbitos nuevos sino de la confluencia de todos estos conocimientos en nuevas disciplinas. No se trata en el caso que nos ocupa de informáticos que trabajen con problemas de tipo biológico, ni de biólogos que usen herramientas informáticas, sino del nacimiento de una nueva disciplina, que al igual del ejemplo del concierto de Estocolomo, nace gracias a la conjunción de disciplinas independientes.

2.2. Dogma central de la Biología Molecular

En los años cincuenta Watson y Crick describieron la estructura de doble hélice que presentan las moléculas de ADN. Se sentaban de esta forma las bases de la genética molecular y se identificaba esta molécula, con su característica de complementariedad entre sus dos hebras, como clave en la herencia. El ADN, o ácido desoxirribonucleico, se puede caracterizar como una secuencia de cuatro letras A, T, G, C, que se corresponden con las cuatro bases nitrogenadas Adenina, Timina, Guanina y Citosina. Esta secuencia se compone de genes, fragmentos que almacenan información para generar un producto simple como por ejemplo una proteína, y de partes sin información o ADN basura. Se dice genoma al conjunto de todos los genes de un organismo. A finales del siglo XX, el Proyecto Genoma Humano logra determinar el genoma completo de nuestra especie identificando, y

[2] `http://es.wikipedia.org/wiki/Proyecto_Genoma_Humano`

localizando en sus respectivos cromosomas, a más de 20.000 genes codificantes. El determinismo genético postulaba que un gen generaba una proteína y por lo tanto una funcionalidad. De esta manera se pensaba que, una vez que se había conseguido descifrar el código genético para nuestra especie, se podrían comprender los fundamentos de la mayoría de enfermedades y procesos biológicos en general. Sin embargo este optimismo con el que se comenzaba el siglo XXI pronto tuvo que ser matizado y el dogma central de la biología molecular, que postula que un gen codifica una función, tuvo que ser redefinido.

El ADN entre dos individuos de una misma especie es prácticamente idéntico, en nuestros caso tan sólo nos diferenciamos unos de otros en un 0,1 %. Sin embargo existe gran variabilidad y diversidad entre nosotros. El ADN se lee y se transforma en ARNm, o ácido ribonucléico mensajero, que da lugar en los ribosmas a los aminoácidos que constituyen las cadenas de las proteínas. Éstas son las encargadas de efectuar los distintos procesos y actuar en las funciones moleculares. Se dice que un gen se expresa si se ha leído y caso contrario se dice que está silenciado. Aunque tengamos la misma información genética, ésta se puede compactar de multitud de maneras para formar los cromosomas que son las moléculas que almacenan en el núcleo de nuestras células el material genético. Las histonas son proteínas que sirven de nudos para que el ADN se pueda comprimir y formar los cromosomas. Existen muchos factores de tipo químico que afectan a este proceso de compactación, lo que influye también en que un determinado gen se pueda o no expresar. Por ejemplo, la acetilación de las histonas afecta a estas proteínas y provocará que el ADN compactado no se exprese, o por ejemplo la metilación del ADN hará que éste se silencie. Es decir, que el ADN se exprese o no dependerá de una serie de factores o condiciones tanto genéticos como no genéticos. El conjunto de genes que pueden ser leídos, es decir que se expresan, depende de una serie de condiciones ambientales.

El dogma central de la biología molecular que postulaba que un gen era sinónimo de una proteína, y por tanto de una función, se redefine a una visión de tipo holística o basada en sistemas complejos. Una funcionalidad viene determinada no por un gen sino por un conjunto de genes que se expresan según unas determinadas condiciones. Estas condiciones determinarán que un proceso se active y su variabilidad estará relacionada con la expresión o no de los genes en cuestión. Así por ejemplo, al leer el genoma de una persona se podrá tan sólo afirmar la predisposición que tiene a sufrir o no una enfermedad pero no estará determinado genéticamente a padecerla.

2.3. Ciencias ómicas

Las ciencias ómicas son el resultado del desarrollo de distintas *tecnologías de alto rendimiento* que producen gran cantidad de datos en diferentes campos de la biología molecular. Esta generación masiva de datos, tan sólo posible gracias a los avances en biotecnología de estos últimos años, permiten una visión global de sistemas de los procesos biológicos.

En la actualidad se pueden secuenciar genomas y transcriptomas completos y gran parte del epigenoma, proteoma o metaboloma. Así se puede hablar de *genómica*, o ciencia del estudio del genoma, sin duda la más desarrollada y punto de partida de todas las demás. Los transcritos son los distintos tipos de ARN que se generan a partir de la expresión génica y que intervienen en la sínstesis de proteínas, ARNm o mensajero, ARNt o transferente, microRNA, etc. La *trasncriptómica* es la ciencia que estudia el transcriptoma. A veces la *proteómica*, estudio del proteoma o conjunto de proteínas, y la *metabolómica*, o estudio de los metabolitos, se incluye como parte de la trasncriptómica. La *epigenómica* se relaciona con el estudio de las condiciones epigenéticas, que son las condiciones de tipo no genético que intervienen en la expresión génica.

El reto en la actualidad se encuentra en convertir en conocimiento biológico toda esta gran cantidad de información generada por las tecnologías de alto rendimiento. Es decir, dar el salto desde el análisis de datos ómicos a nuevas preguntas y respuestas biológicas. Desde esta perspectiva se habla de *medicina personalizada* como una medicina de precisión en la que, en función de los datos de cada individuo, se puedan desarrollar métodos de prognosis, que determinen la propensión a padecer un transtorno o enfermedad, y en la que la elaboración de fármacos sea de manera personalizada.

2.4. Bioinformática

En la actualidad queda claro que el papel que juega la Informática en la revolución de la era postgenómica es algo más que la tradicional definición de bioinformática como servicio de apoyo o soporte de los experimentos realizados en los laboratorios. Como hemos visto en el apartado anterior, el objetivo no es el mero análisis de datos sino la generación de nuevo conocimiento a partir de los mismos. Bajo esta perspectiva podríamos dar una definición más actual de bioiformática de la siguiente manera.

> Bioinformática es el diseño, implementación y aplicación de tecnologías computacionales, métodos y herramientas para la explo-

> tación de los datos de tipo ómico. Su principal misión consiste en el modelado de sistemas complejos y de patrones de información con propósitos predictivos.

Esta nueva definición queda aún más clara, si cabe, si tenemos en cuenta la confluencia que en la actualidad está teniendo lugar entre datos de tipo ómico, o biosanitarios en general, y las nuevas técnicas de análisis de datos generalmente conocidas como *Big Data* [89]. Este tipo de trabajos son, además de un ejemplo del nuevo discurrir de la bioinformática para los próximos años [89], un claro ejemplo de la necesidad de entender la disciplina como algo más que el mero servicio de apoyo o soporte a los experimentos realizados en los laboratorios de biología molecular.

2.4.1. Descubrimiento de biomarcadores

En el ámbito de la Biomedicina, la Bioinformática se entiende como la disciplina cuya principal misión es el descubrimiento de nuevos biomarcadores usando para ello datos de tipo ómico.

Se define como *biomarcador* aquella característica que permita medir y evaluar objetivamente una determinada respuesta terapéutica a una enfermedad, fármaco o proceso biológico en general. Los biomarcadores se pueden clasificar en tres grandes grupos: aquellos que sirven para detectar el desarrollo de una enfermedad, aquellos que predicen la respuesta a un determinado tratamiento y, por último, aquellos que se pueden utilizar directamente como sustitutos de dicho tratamiento o también conocidos como dianas terapéuticas [7]. Alternativamente, se habla de biomarcadores *predictivos* o de *prognosis*. Es decir, los que permiten conocer la respuesta a un determinado tratamiento y los que permiten establecer el riesgo de padecer una enfermedad. Se puede entender un biomarcador como el indicador de una enfermedad. Por ejemplo, un biomarcador puede ser la concentración anormal de una proteína que indique el desarrollo de una enfermedad.

Los datos ómicos permiten el estudio y descubrimiento de nuevos biomarcadores. A través del estudio de este tipo de datos se puede establecer las consecuencias a nivel del individuo partiendo de información a nivel molecular. Es decir, establecer relaciones de *genotipo-fenotipo*. La figura 2.1 muestra las distintas maneras que los datos ómicos se pueden utilizar para estudiar estas relaciones[3].

Los datos de expresión génica obtenidos mediante la tecnología de microarray son un tipo particular de datos ómicos. Cuando se trabaja con ellos se

[3]La imagen ha sido tomada de la revista Nature del artículo *"'Methods of integrating data to uncover genotype-phenotype interactions"* con DOI: 10.1038/nrg3868

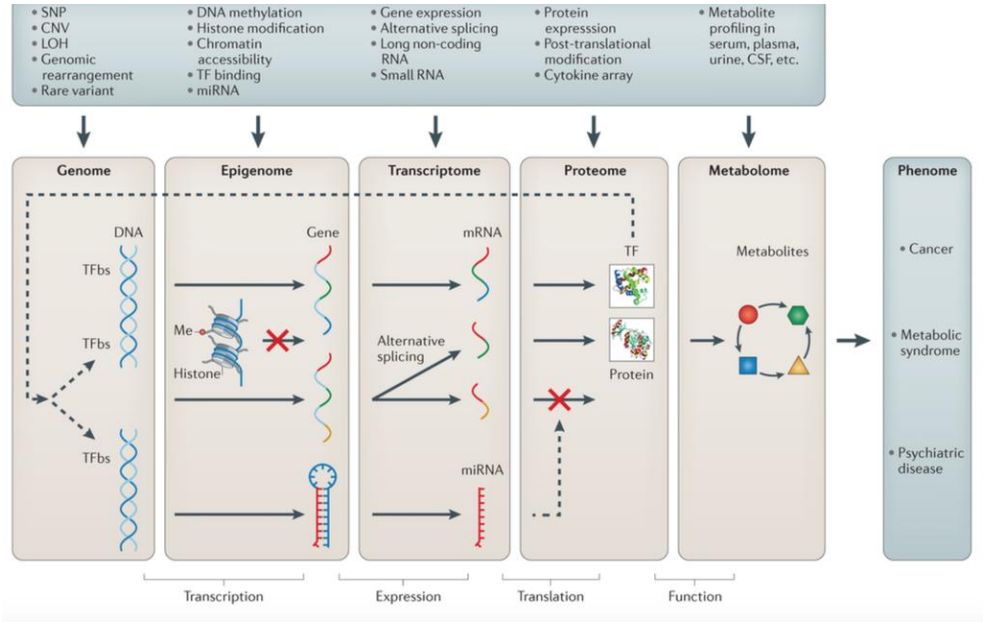

Figura 2.1: Métodos de integración de datos para el descubrimiento de interacciones genotípicas-phenotípicas.

debe tener en cuenta el contexto de trabajo y tener en consideración algunos aspectos fundamentales como:

- ¿Qué enfermedad o estudio biológico motivan la generación de estos datos?
- ¿Qué tipo de chips o arrays se han utilizado?
- ¿Qué información se pretende encontrar al diseñar los chips o placas del experimento?

2.5. Rudimentos de Biología

Como hemos visto con anterioridad, la visión reduccionista dada por el dogma central de la biología debe ser modificada por una visión de tipo holístico o basada en Sistemas Complejos. El sistema en sí no puede ser entendido por la simple suma de las partes que lo componen sino mediante la emergencia de un nuevo conocimiento generado por sus interacciones. Sin embargo hay que conocer bien las partes del mismo para su comprensión.

Veamos a continuación una serie de términos básicos o rudimentarios que se deben conocer relacionados con el dogma central de la biología.

2.5.1. Términos básicos

Las siguientes definiciones están extraídas del libro *El lenguaje de la Vida* de Francis Collin.

- **Gen**: unidad física y funcional de la herencia, que se pasa de padres a hijos y contiene la información necesaria para especificar caracteres específicos. Los genes están dispuestos uno a continuación del otro en unas estructuras denominadas cromosomas. Un cromosoma contiene una única molécula de ADN de gran longitud, en la cual cada gen ocupa una porción muy pequeña. Los seres humanos tenemos aproximadamente 20.000 genes codificadores de proteínas en nuestros cromosomas.

- **ADN** (ácido desoxirribonucleico): nombre químico de la molécula portadora de las instrucciones genéticas en los seres vivos. La molécula de ADN consiste en dos hebras que giran una alrededor de la otra formando una doble hélice. Cada hebra es una cadena larga formada por la alternación de un azúcar (desoxirribosa) y un grupo fosfato. A cada se une una de cuatro bases: A,T, C o G. Las dos hebras se mantienen unidas por enlaces entre las bases (A se une a T y C a G). La secuencia de bases contiene las instrucciones para hacer moléculas de ARN y proteínas.

- **Cromosomas**: un cromosoma es el resultado del empaquetamiento del ADN, junto con las proteínas auxiliares determinadas, previo a la división celular para su segregación posterior en células hijas. Los cromosomas se encuentran en el núcleo de las células y diferentes especies tienen diferente número y morfología de cromosomas. Los humanos tenemos 23 pares de cromosomas: 22 pares de autosomas y un cromosoma sexual (X ó Y). Cada uno de los progenitores aporta un cromosoma a cada par, de manera que los hijos reciben la mitad de los cromosomas de la madre y la mitad del padre.

- **Código genético** (ACGT): instrucciones contenidas en un gen que le dicen a la célula cómo hacer una proteína específica. A, T, G, y C son las "letras" del código genético y representan las bases nitrogenadas adenina, timina, guanina y citosina, respectivamente. Junto con un azúcar y un enlace fosfato, estas bases constituyen los nucleótidos, que son la unidad fundamental del ADN. En cada gen se combinan las cuatro bases en diversas formas, para crear palabras de tres letras que

especifican qué aminoácido es necesario en cada paso de la elaboración de la proteína.

- **ARN** (ácido ribonucleico): molécula parecida al ADN. A diferencia de éste, el ARN tiene una sola hebra, que está compuesta por un azúcar distinto (la ribosa) y grupos fosfatos alternados. A cada molécula de azúcar se une una de las siguientes cuatro bases: A (adenina), U (uracilo), C (citosina) o G (guanina). En la célula existen varios tipos de ARN: ARN mensajero, ARN ribosómico y ARN de transferencia. Recientemente se ha descubierto que algunos ARN intervienen en la regulación de la expresión génica.

- **ARNm o mensajero**: molde para la síntesis de proteínas en los ribosomas del citoplasma. Cada juego de tres bases, llamado codón, especifica un aminoácido en la secuencia que comprende la proteína. La secuencia de la cadena única del ARNm está basada en la secuencia de una cadena complementaria de ADN.

- **Aminoácido**: grupo de veinte moléculas pequeñas distintas que se unen formando cadenas largas o polipéptidos. Una proteína está constituida por uno o más polipéptido. La secuencia de aminoácidos hace que el polipéptido se pliegue de una forma determinada que es biológicamente activa. Las secuencias de aminoácidos de las proteínas están codificadas en los genes.

- **Proteína**: una molécula compuesta por una o más cadenas de aminoácidos. La secuencia de éstos es una traducción de la secuencia de ADN del gen que codifica la proteína. Las proteínas desempeñan una amplia gama de actividades vitales en la célula: estructurales (citoesqueleto), mecánicas (músculo), bioquímicas (enzimas) y de señalización celular (hormonas). Las proteínas también son una parte esencial de la dieta.

- **ADN no codificador**: ADN que no codifica aminoácidos. La mayoría del ADN no codificador se encuentra entre genes en los cromosomas y su función es desconocida. Otras secuencias de ADN no codificadoras son los intrones, que se encuentran dentro de los genes. Algunos segmentos de ADN no codificador intervienen en la regulación de la expresión génica.

- **Citoplasma**: líquido gelatinoso del interior de la célula. Está compuesto por agua, sales y varias moléculas orgánicas. Algunos orgánulos intracelulares, como el núcleo y las mitocóndrias, están envueltos en membranas que los separan del citoplasma.

- **Genoma**: conjunto completo de instrucciones genéticas de una célula. En los humanos, el genoma está formado por 23 pares de cromosomas que se encuentran en el núcleo más un pequeño cromosoma en las mitocondrias de la célula. Estos cromosomas contienen en conjunto aproximadamente 3.100 millones de bases.

- **Carácter poligénico**: carácter, o característica, cuyo fenotipo está influido por más de un gen. Lo caracteres que exhiben una distribución continua, como la estatura o el color de la piel, son poligénicos. La herencia de los genes poligénicos no obedece los típicos cocientes fenotípicos de la herencia mendeliana, aunque cada uno de los genes que contribuyen a definir un rasgo se hereda del modo que describió Mendel. Muchos caracteres poligénicos también están influidos por el ambiente y reciben el nombre de multifactoriales.

- **Gen supresor de tumores**: Gen cuya función normal consiste en dirigir la producción de una proteína que forma parte del sistema que frena la división celular. La proteína supresora de tumores mantiene la división celular bajo control; sin embargo, cuando muta y no puede realizar bien su trabajo, puede desencadenarse un crecimiento celular incontrolado que contribuye al desarrollo del cáncer.

- **Oncogén**: Gen mutado que puede hacer que las células normales se conviertan en células cancerosas. En su forma normal, sin mutación, reciben el nombre de protooncogenes, y participan en la regulación de la división celular. Los oncogenes actúan como el acelerador de un coche, empujando las células a dividirse.

- **Mutación**: alteración estructural permanente en el ADN. En la mayoría de los casos pueden no tener ningún efecto o causar daño, pero en ocasiones una mutación puede mejorar la probabilidad de supervivencia de un organismo. Las mutaciones pueden originarse en errores de copia del ADN, exposición a radiaciones ionizantes, exposición a sustancias químicas mutágenas o infección por virus. Las mutaciones en la línea germinal son las que se producen en el óvulo o espermatozoide, y son transmitidas a la descendencia. Las mutaciones somáticas, en cambio, se producen en las células del cuerpo y no se transmiten.

- **Prognosis**: La evolución probable o prevista de una enfermedad.

- **Diagnosis**: La identificación de la naturaleza y causa de un determinado trasntorno.

- **Fenotipo**: Manifestación visible del genotipo. Características observables de un determinado organismo.
- **Genotipo**: Información genética de un determinado organismo.

Se deben diferenciar y no confundir los términos genético, génico y genómico.

- **Genético**: relativo a la genética, es decir, relacionado con la herencia
- **Génico**: relativo a los genes.
- **Genómico**: relativo al genoma, es decir, genes, proteínas y todo lo relacionado con la genómica.

Parte II

Estado del arte

Capítulo 3

Datos de expresión génica

...una cosa es saber de física de una forma abstracta y otra muy distinta lidiar con problemas directamente relacionados con datos experimentales, como los que provenían de la nueva tecnología que se estaba desarrollando en Los Álamos.
Aventuras de un matemático. Memorias de Stanislam M. Ulam (pág. 160).

3.1. Introducción

Todas las células de nuestro cuerpo tienen en su núcleo el mismo ADN sin embargo existe una gran variedad tanto morfológica como funcionalmente entre ellas. Se considera el *dogma central de la Biología Molecular* a la explicación tradicional del proceso mediante el cual la información almacenada en las hebras de ADN da lugar a una proteína o producto funcional. Básicamente se puede resumir diciendo que el ADN da lugar al ARN mediante el proceso de *transcripción* y éste, mediante el proceso de *translación*, origina los aminoácidos que, mediante plegados y composición de las cadenas que forman, construyen las proteínas. Las proteínas son las unidades funcionales que están implicadas en las distintas funciones biológicas que tienen lugar en nuestro organismo. Se llama *expresión de un gen* al proceso completo mediante el cual un gen da lugar a un producto funcional. Por lo tanto, se puede decir que mediante la transcripción de un trozo de ADN se origina una cadena de ARN que da lugar a una cadena de aminoácidos mediante el proceso de translación. La translación se lleva a cabo mediante los ribosomas, en el citoplasma de la célula, y las proteínas se forman al plegarse las cadenas de aminoácidos producidas según las propiedades físico-químicas del entorno. La figura muestra un esquema del dogma central de la Biología

Molecular.

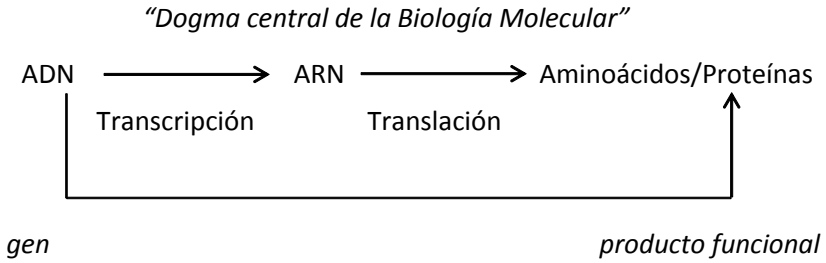

Figura 3.1: Dogma dentral de la Biología Molecular.

Los genes son fragmentos de ADN que actúan como unidades físicas y funcionales de la herencia y cuando se activan codifican una proteína. Sin embargo su activación depende de una serie de circunstancias. Es decir, la regulación de la expresión de un gen viene determinada mediante una serie de condiciones que aportan la información necesaria para que dicho gen se active. Se llama *regulación génica* al proceso que origina que un gen se active o no. La mayor parte de los procesos de regulación génica tienen lugar durante el proceso de transcripción y las condiciones que los originan reciben el nombre de factores de transcripción. El estudio de dichas condiciones es fundamental para entender la mayoría de los procesos biológicos, enfermedades, etc, de origen genético. El esquema descrito en la figura es un simplificación. El dogma central de la Biología Molecular representa la visión clásica según la cual un gen da lugar a una proteína. En la actualidad se sabe que en la mayoría de los casos no es tan simple y que en general un grupo de genes actuando conjuntamente regulan que una determinada función se determine o no, es decir que varios genes codifican una o varias proteínas. Las situaciones en las que un gen codifica un producto funcional constituye la excepción en lugar de la regla. Esta visión, que podemos llamar de tipo *holístico* o de sistemas, es la más aceptada en la actualidad. Bajo esta perspectiva se habla de caracteres *poligénicos* o *multifuncionales* en contraposición a la situación en la que un sólo gen determina una determinada funcionalidad.

Como ya hemos comentado los genes son fragmentos de ADN que se disponen de forma que constituyen las unidades funcionales de herencia. En el caso de los humanos tenemos por ejemplo aproximadamente 20.000 genes que codifican proteínas. El resto del ADN tiene información que no codifica proteínas y que hasta ahora recibía el nombre de *ADN basura*. Es decir, tan sólo servía como separador de los trozos de ADN, los genes, que sí codifican proteínas. Recientemente se ha descubierto que estas amplías regiones del

ADN, la mayor parte, no cumple una función testimonial sino que interviene directamente en los procesos reguladores de genes. El papel de las regiones de ADN, no codificantes o basura, juega un papel fundamental en el proceso de expresión génica. En cualquier caso los factores ambientales son primordiales para estudiar cuándo un gen, o un grupo de genes, se activan y codifican un grupo de proteínas que determinan una función biológica.

3.2. Datos de microarrays

Las tecnologías de alto rendimiento, o *high throughput technologies*, permiten medir el nivel de expresión de cientos de genes simultáneamente. El uso del término microarray suele llevar a confusión porque suele ser utilizado indistintamente tanto para la tecnología, como para los experimentos que se realizan mediante ella así como los datos generados. Básicamente se puede decir que es una tecnología que permite generar conjuntos de datos que miden el nivel de expresión de un grupo genes bajo estudio.

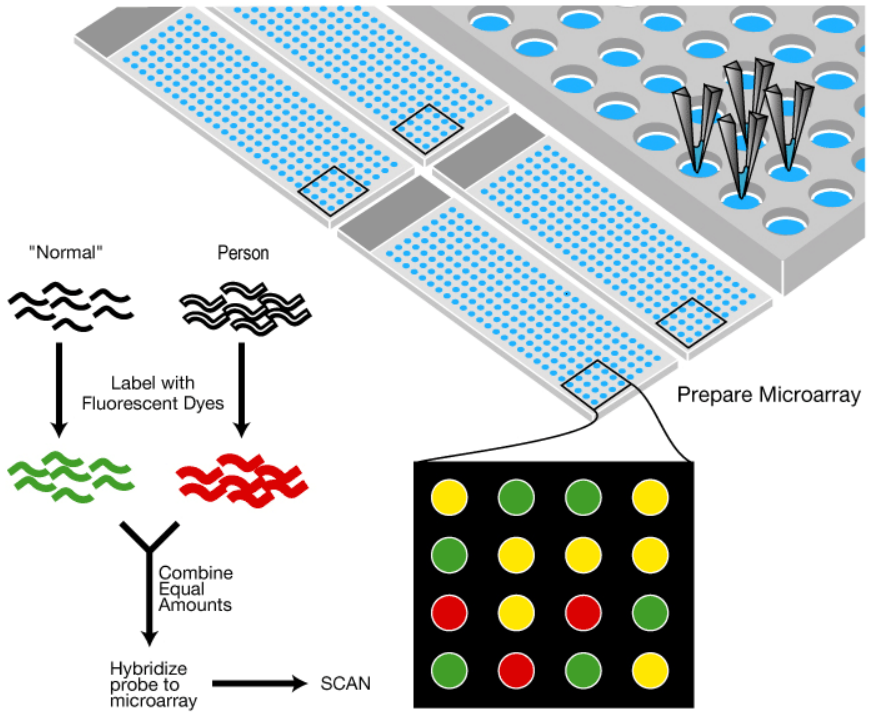

Figura 3.2: Plataforma de microarray.

La *tecnología de microarray* permite generar *chips* o rejillas con miles de *spots* o celdillas en los que se encuentran cadenas de nucleótidos de los

que se puede medir su nivel de abundancia o transcripción. Existen varios tipos de microarrays, de tecnologías de microarrays, de un canal o dos, etc. Básicamente, y de manera muy general, la idea consiste en preparar cada hueco con una hebra de ADN o *probe set* que identificamos con un gen. En la realidad una muestra contiene miles de hebras iguales. Una vez preparada la rejilla con una serie de huecos o *spots* del microarray, se toma la muestra del experimento que se quiere estudiar, se tiñe con una sustancia y se vierte sobre dicha rejilla. Se dice que se hibrida hueco a hueco. Por complementariedad de ADN se unen las hebras de cada hueco con las de la muestra en estudio. De esta forma cada hueco tiene una concentración de hebras de ADN que se puede medir según la intensidad del color que se ha generado. Se procesa dicha intensidad y se le asigna un valor numérico. Además de la dificultad de tipo técnico, se añade un procesado en profundidad de los datos generados: correcciones de errores, procesamiento de imágenes, etc. Todas esas técnicas es lo que se conoce como *Análisis de Microarrays a bajo nivel*.

Se conoce como *plataforma de microarray* a la arquitectura concreta del experimento que se ha realizado. Hay varios tipos de plataformas: *DNA microarrays*, *Affymetrix*, etc. Cuando se usa el término *experimento de microarray* se refiere a los distintos estudios experimentales que se han desarrollado para una misma plataforma. Generalmente en un mismo experimento hay involucradas varios chips o rejillas. La figura 3.2 muestra un esquema de cómo se genera una rejilla o chip de ADN. La reunión final de todos los chips es el resultado final del experimento y se suele denominar de manera general como *datos de microarray*. El conjunto de datos generado, compuesto por el nivel de expresión de los genes estudiados a los largo de cada chip, constituye la *matriz de expresión*. Las filas de la matriz de expresión son los *patrones de expresión* de los genes y las columnas los *perfiles de expresión* de cada condición. De esta forma, una columna completa se corresponde con todos los valores de expresión de los genes para un mismo chip o rejilla. Al unir todos los chips se forma una matriz en la que cada fila mide cómo varía la expresión de los genes. Cada elemento de la matriz de expresión representa el nivel de expresión de un gen bajo una determinada condición experimental[1].

La hipótesis fundamental asociada a los estudios con datos de expresión génica nos dice que si dos genes tienen perfiles de expresión similares entonces comparten una misma funcionalidad o intervienen en un mismo proceso biológico. Es decir, genes co-expresados implica genes co-regulados. De esta

[1] DNA Microarray Methodology - Flash Animation:
http://www.bio.davidson.edu/genomics/chip/chip.html

forma la tecnología de microarray es un herramienta muy útil no sólo para la investigación en biología molecular sino también en estudios clínicos.

3.3. Proceso de trabajo

La figura 3.3 muestra cómo es el flujo de trabajo estándar cuando se trabaja con datos de expresión obtenidos mediante la tecnología de microarray. Una hipótesis de trabajo genera una pregunta biológica sobre, por ejemplo, un determinado proceso de respuesta de unos pacientes ante una enfermedad, etc. Se elabora un diseño experimental para tratar de responder a dicha pregunta, por ejemplo datos de estudio y de control, tras lo que se prepara el experimento de microarray. Se eligen una serie de placas o rejillas, se lleva a cabo la extracción del ARN de las muestras que se van a estudiar y se hibridan o combinan las muestras en las placas. El análisis de la intensidad de color que se ha producido en las imágenes generadas permite obtener un valor numérico para cada celdilla. Estos datos numéricos generados deben ser sometidos a un control de calidad o de normalización, mediante los que se eliminan duplicados, se estandarizan todos los valores y se reúne toda la información.

La matriz obtenida mediante el proceso anterior debe ser preprocesada para eliminar genes duplicados, realizar el tratamiento de los valores perdidos, el proceso de etiquetado o anotación de los genes, etc, tras lo cual se obtiene la matriz de expresión sobre la que se puede llevar a cabo el análisis de alto nivel o de minería sobre los datos. Básicamente se pueden realizar estudios de Minería de Datos, elaboración de modelos predictivos o clasificadores, exploración de los datos o aplicar técnicas de clustering así como estudios de anotación funcional de genes cuyo objetivo es su clasificación. En general tras el estudio se realiza un contraste con la información ya existente almacenada en bases de datos con información biológica, como por ejemplo la ontología de genes GO, en lo que se conoce como proceso de enriquecimiento de genes de tal manera que se puedan establecer unas conclusiones robustas a la pregunta planteada.

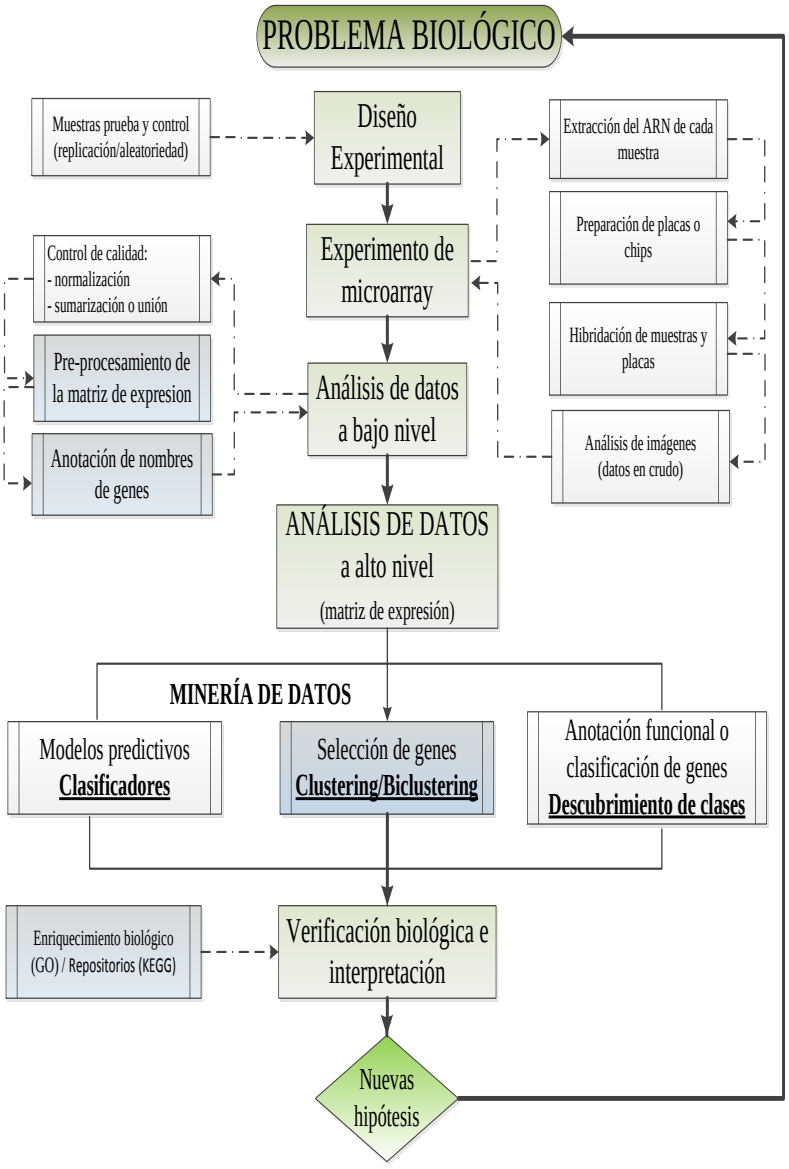

Figura 3.3: Flujo de trabajo con datos producidos mediante microarrays.

3.3.1. Repositorios públicos

En la actualidad existen varios repositorios públicos que almacenan datos de expresión génica. Destacan el repositorio GEO *Gene Expression Omnibus*, asociado al *National Center of Biotechnology Information* (NCBI) pertene-

ciente al gobierno de EEUU, así como *ArrayExpress* asociado al *European Bioinformatics Institute* (EBI), que es la versión europea del anterior.

En estos repositorios se almacena tanto los datos generados por cada experimento de microarray como la información relacionada con estos. El protocolo MIAME, *Minimum Information About a Microarray Experiment*, es una guía que indica qué información debe ir asociada a cada experimento de microarray. Básicamente la información que se guarda son los datos en crudo sobre cada chip del experimento, que suelen ser ficheros con formato CEL o GPR, la matriz de expresión producto de la normalización de los anteriores, información sobre las muestras y el diseño del experimento realizado, información sobre cada gen. Generalmente también se suele notificar los protocolos que se han seguido para el manejo de los datos. Cada gen está asociado a un *probe set* o prueba, que es el fragmento del ADN que se pone en cada celdilla. Por cada gen se almacena de manera adicional sus distintas nomenclaturas, sus anotaciones en la ontología de genes GO así como relaciones relevantes con otros conceptos biológicos.

El repositorio GEO[2] además de almacenar datos, integra herramientas web que ayudan a su análisis, procesamiento, visualización, envío, etc. La figura 3.4 muestra cómo es el acceso a la información de un registro de un experimento de microarray en este repositorio. La organización de la información se organiza de manera que se pueden encontrar tanto los ficheros subidos por los usuarios, como los ficheros generados automáticamente por la plataforma a partir de éstos. Los usuarios cuando suben la información de un experimento deben diferenciar entre *series*, *plataformas* y *muestras*. Las series recogen la información de todas las muestras utilizadas, tipo de placas, descripción, etc. Por ejemplo, un fichero con el nombre GSE3541 es una serie que recoge esta información. Un fichero con el nombre GPL341 es un ejemplo de fichero para una plataforma. Permite agrupar aquellas muestras de una serie que se han generado siguiendo unas mismas pautas. Contiene un número de identificación y una serie de características generales. No es obligatorio pero puede incluir la matriz de expresión (la matriz de datos) que ha generado el usuario para su investigación. Los ficheros para las muestras, por ejemplo GSM81022, contienen los datos numéricos de los perfiles de expresión de los genes para dicha muestra en concreto. Se asocian a una serie y constituyen los datos en crudo con los que se forman las columnas, o condiciones experimentales, de la matriz de expresión.

Los conjuntos de datos o *datasets* de GEO se generan automáticamente por el repositorio. Se construye la matriz de expresión de genes, donde

[2]http://www.ncbi.nlm.nih.gov/geo/

los genes son filas y las columnas condiciones experimentales, y se incluye una serie de información útil, como por ejemplo, distintas notaciones para el nombre de los genes, sus términos GO asociados, etc. Los *datasets* son ficheros con la nomenclatura GDSxxxx. De manera adicional se puede encontrar en ellos la descripción del experimento, los autores, las referencias en las que se usan los datos del experimento de microarray, así como la matriz de expresión con formato SOFT. La figura 3.4 es un ejemplo del *dataset* GDS4154. Se puede observar en la figura la referencia el registro de la serie utilizada, GSE19587, subido por los usuarios, el registro generado por el repositorio, GDS4154, así como una visualización parcial de las muestras.

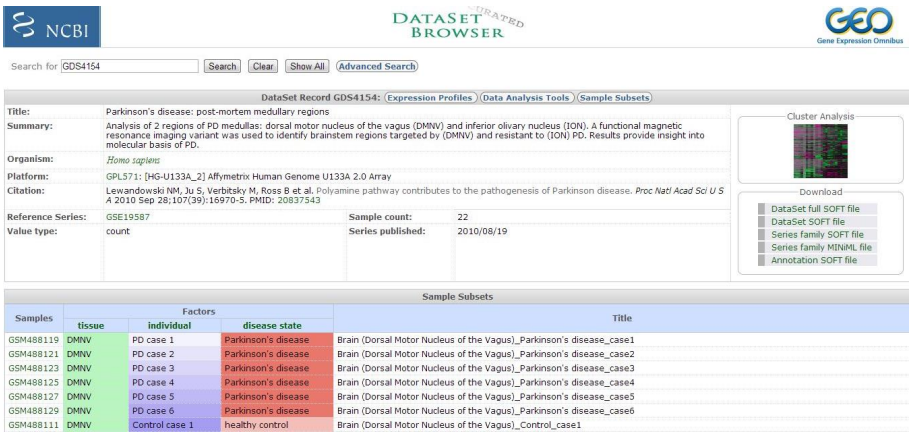

Figura 3.4: Navegador del repositorio GEO: acceso a datos.

Los ficheros con formato SOFT son ficheros de textos planos en los que se puede encontrar la matriz de expresión asociada a cada conjunto de datos o fichero GDSxxxx. Por cada fila de la matriz se puede observar la prueba o *probe set*, el gen asociado en varios formatos, así como información relevante sobre dicho gen: localización en el cromosoma en el que se encuentra, términos GO asociados, etc. Para cada columna de la matriz la información sobre cada muestra con su notación GSMxxx así como información sobre cada muestra en el texto adicional del fichero.

3.3.2. Procesamiento de los datos

Las matrices de expresión que se pueden obtener a través de un registro en formato SOFT deben ser preprocesadas para su uso. Una vez que los datos en crudo se han unido y procesado siguen quedando una serie de pasos que realizar. Así por ejemplo, un mismo gen puede aportar varios perfiles de expresión que deben ser homogeneizados de forma que se constituya una

sola fila en la matriz de expresión procesada. Se pueden fusionar según la media de los valores y conseguir así una sola fila de la matriz, mediante la mediana, etc. Por otro lado, los valores nulos, o *missing values*, deben ser reemplazados y se puede llevar a cabo poniendo cero, el valor medio de los restantes elementos de esa fila, el valor de la mediana o por el valor medio de aquellas filas más parecidas. En general se establece un umbral, y si una fila de la matriz tiene una cantidad de valores nulos superiores a ese umbral, se elimina esa fila de la matriz. Finalmente, y tras una estandarización de los valores de los datos para que todos se encuentren en rangos de valores similares, aquellas filas que presenten un comportamiento plano, sin picos, se dirá que son perfiles de genes que no están diferencialmente expresados y se filtrarán para el estudio.

El procesamiento de los datos de la matriz de expresión se puede llevar a cabo con distintas herramientas, como por ejemplo librerías R del paquete *Bioconductor* o herramientas web como *Babelomics* [62].

3.3.3. Genes: nomenclatura y anotaciones

Uno de los grandes problemas que se encuentran al trabajar con datos de microarray es la nomenclatura usada para los genes. Existen distintas nomenclaturas que dependen de las bases de datos utilizadas, los laboratorios, plataformas, etc. Así un mismo gen se puede encontrar con varios nombres siendo esto uno de los problemas más comunes cuando se define los datos objetos de estudio. Hay que diferenciar entre las etiquetas puestas para cada fila de la matriz de expresión, los *probe sets*, los genes a los que representan y la nomenclatura de estos. Se pueden encontrar genes notados para la plataforma *ORFF*, *Ensembl*, *Entrez*, *Uniprot*, etc. En el caso de trabajar con datos de *Homo sapiens*, la web *Hugo Gene Nomenclature Committe*[3] puede utilizarse a modo de diccionario para establcer la equivalencia entre los distintos nombres de genes. En general, la mejor opción de trabajo es seleccionar *gene symbol* como nomenclatura estándar ya que constituye la opción más general de trabajo y suele ser reconocida por la mayoría de herramientas.

3.3.4. Ontología de Genes: GO

Tras el análisis a alto nivel de los datos es necesario la interpretación biológica de los resultados obtenidos. Si el estudio realizado ha sido una exploración de los datos siguiendo una técnica de Minería de Datos de tipo

[3] http://www.genenames.org/

no supervisado, como por ejemplo ocurre con las técnicas de biclustering, será además de vital importancia la verificación de los resultados mediante conocimiento experto. Para esta labor se utilizan repositorios biológicos como GO[4] (*Gene Ontology*), KEGG[5] (*Kyoto Encyclopedia of Genes and Genomes*) o por ejemplo Reactome[6]. Mediante estos repositorios se establecen relaciones entre los genes y los procesos, rutas metabólicas o funcionalidades en los que los genes intervienen.

La ontología de genes GO es sin lugar a dudas el repositorio más utilizado. Constituye un vocabulario mediante el que se relaciona cada gen con el término funcional para el cual ha sido previamente anotado, ya sea una función molecular, un proceso biológico o el lugar de la célula en el que interviene. Por este motivo, GO tiene una estructura de árbol con tres diferentes ramas o subontologías: CC (*Cellular Component*) para relacionar genes y las partes de la célula o del entorno fuera de estas, MF (*Molecular Function*) para productos o actividades en las que intervienen los genes a nivel molecular y BP (*Biological Process*) para los procesos biológicos.

Se conoce con el nombre de enriquecimiento biológico al estudio mediante el que se establece la relevancia de un grupo de genes en un determinado término. Existen multitud de herramientas vía web para dado un grupo de genes determinar el conjunto de términos GO estadísticamente significativos asociados. Sin embargo, si se necesita analizar el resultado de un algoritmo completo estas herramientas no sirven si no permiten un acceso mediante línea de comandos para poder automatizar las consultas.

[4] http://geneontology.org/
[5] http://www.genome.jp/kegg/
[6] http://www.reactome.org/

Capítulo 4

Biclustering de datos de expresión génica

Mientras pensamos en más formas de usar una herramienta, imaginamos más objetivos que puede ayudarnos a conseguir. National Geographic enero 2013 pág. 23. Genes inquietos, David Dobbs.

4.1. Introducción

Biclustering es una técnica de Aprendizaje No Supervisado que se encarga de la búsqueda de patrones locales en datos de expresión génica. El objetivo es el agrupamiento de ejemplos o instancias bajo un determinado subconjunto de atributos o clases que, aplicado en el contexto de los datos de expresión génica, significa agrupar genes bajo un subconjunto de condiciones experimentales. Este problema se puede encontrar en la bibliografía y en otro contexto con otros nombres como clustering de subespacios [44] o co-clustering [28]. Sin embargo es en el contexto biológico donde adquirió su mayor importancia y se desarrolló ampliamente considerándose el artículo fundacional en este campo el de Cheng y Church del año 2000 [24]. Intuitivamente se puede entender como la búsqueda de submatrices que reflejen cierto tipo de patrones en la matriz de datos. La principal diferencia entre biclustering y clustering reside en que encuentra patrones locales, que no dependen de todas las condiciones, y en el solapamiento entre los mismos. La naturaleza de los datos de expresión génica hace que este hecho sea relevante debido a su significado biológico.

Un experimento de microarray se realiza con multitud de muestras y cada una de ellas serán los atributos o clases que reflejan unas condiciones

experimentales que son objeto de estudio. La elección de dichas muestras dependen de la hipótesis biológica que se estudia, generalmente son condiciones temporales en un ciclo celular, muestras de pacientes, etc. El interés radica en determinar qué muestras implican una función biológica que vendrá determinada por un grupo de genes coexpresados. Por ello el objetivo no es agrupar genes con un comportamiento similar bajo todas las condiciones sino tan sólo bajo un subconjunto de las mismas. De esta forma se discriminará y se identificará cómo influyen y bajo qué condiciones un grupo de genes se co-expresa y cuales no. Se asume como hipótesis que un grupo de genes co-expresados se co-regulan y por tanto comparten una misma funcionalidad biológica. Por tanto, debido a la naturaleza de los datos, la técnica de biclustering es más adecuada que el clustering tradicional que agrupa bajo todas las condiciones experimentales.

Por otro lado, a diferencia con lo que ocurre en la mayoría de algoritmos de clustering, las soluciones encontradas por un algoritmo de biclustering admiten solapamiento entre ellas, es decir, que un mismo gen puede pertenecer a la vez a varios grupos. Intuitivamente se puede visualizar como que las submatrices obtenidas como resultado no tienen por qué ser disjuntas. Desde un punto de vista biológico este hecho es importante debido a que un mismo gen o grupo de genes puede actuar en varios procesos y funcionar así como catalizador o inhibidor de los mismos. El solapamiento entre genes, en el sentido de un mismo gen interviniendo en varios procesos, es un hecho fundamental para el descubrimiento de biomarcadores en datos de expresión génica.

En la actualidad existe una gran variedad de algoritmos de biclustering que pueden ser clasificados de multitud de formas: según la técnica algorítmica en la que se basan, según el tipo de resultados que encuentran, etc. El saber qué algoritmo de biclustering es el más adecuado en cada momento es una tarea difícil de resolver, así como la elección de qué parte de los resultados son relevantes para el problema que se estudia, etc. En función de qué algoritmo se utilice se encontrará un tipo de patrones, un número de soluciones muy variables en número o en tamaño, un tiempo de ejecución variable, etc. En general los resultados de un algoritmo de biclustering deben de ser analizados a su vez para ver cómo extraer información de ellos. Se requiere un conocimiento en profundidad del problema en cuestión que se esté estudiando para poder sacar partido de la información que se obtiene con este tipo de técnicas. Uno de las grandes preguntas abiertas es cómo comparar distintos algoritmos y según qué criterio. En cada artículo se puede encontrar distintos tipos de análisis siendo el más aceptado el que se basa en información suministrada por expertos. En concreto por la información

almacenada en ontologías como (GO).

Recientemente se han publicado trabajos en los que los algoritmos de biclustering se aplican a otro tipo de datos biológicos, que fusionan información de varias fuentes de datos o que se centran en su aplicación más que en los algoritmos en sí [79, 33].

4.2. Definiciones

Se define un microarray como una matriz de números reales donde las filas representan a los genes y las columnas a las condiciones experimentales. En el conjunto de datos los genes son los ejemplos o instancias y las condiciones experimentales o muestras son los atributos o clases. Se trabajará generalmente con conjuntos de datos no etiquetados en el sentido que no hay una etiqueta previa para cada instancia o gen. En cualquier caso al ser el biclustering una técnica no supervisada el etiquetado en el conjunto de datos no es relevante.

Definición **1 (Microarray)** *Formalmente un microarray se puede definir como una 3-tupla* $M = (G, C, f)$ *donde* $G = \{g_1, \ldots, g_n\}$ *es el conjunto de genes,* $C = \{c_1, \ldots, c_m\}$ *es el conjunto de condiciones y* $f : G \times C \longrightarrow \mathbb{R}$ *es la función que asocia a cada par* (g_i, c_j) *un valor real.*

$$M = \begin{pmatrix} m_{1,1} & \ldots & m_{1,m} \\ \vdots & \vdots & \vdots \\ m_{n,1} & \ldots & m_{n,m} \end{pmatrix}$$

Un problema de clustering consiste en realizar una partición del espacio de datos siguiendo un determinado criterio, es decir, las soluciones son disjuntas entre si y su unión es el total. Se satisface una condición de homogenediad de las soluciones, o similitud entre los objetos de un mismo cluster, y de separación entre ellas o distancia entre clusters. Esta doble condición motiva que las técnicas de clustering se pueden agrupar en dos tipos: de tipo aglomerativo o de división. Partiendo de una solución a la que se le agregan otras o partiendo del conjunto de datos al completo al que se establecen fronteras entre las distintas partes. En el contexto de los datos de microarray un algoritmo de clustering agrupa un conjunto de genes que tengan un comportamiento similar a lo largo de todas las condiciones de la matriz de expresión.

DEFINICIÓN **2 (Solución de un problema de Clustering)** *Dado un microarray M, las soluciones de una técnica de clustering es un conjunto* $\mathbf{S} = \bigcup_{i=1}^{n} S_i$ *tal que se verifica:*

- $\bigcup_{i=1}^{n} S_i = M$
- $S_i \bigcap S_j = \emptyset, \forall i, j \in N$

Aunque recientememte algunos algoritmos de clústering sobre redes admiten resultados solapados, la mayoría de los algoritmos de clustering no permiten solapamientos entre los resultados, es decir, en el caso de datos de expresión un mismo gen no puede pertenecer a dos clusters. Este hecho, junto con la imposibilidad de diferenciar qué subconjunto de condiciones son las relevantes para la activación o no de un grupo de genes, motiva la definición del biclustering.

Un algoritmo de biclustering agrupa un conjunto de genes que presentan un comportamiento similar a lo largo de un subconjunto determinado de condiciones de la matriz de expresión. Intuitivamente se puede decir que se buscan submatrices de genes coexpresados bajo unas determinadas condiciones. Se trata de realizar una búsqueda de patrones locales según un criterio preestablecido. Las soluciones o *biclusters* no constituyen una partición del espacio y admiten solapamiento, es decir, se satisface el criterio de homogeneidad de soluciones pero no el de separación de las mismas. Además, se admiten soluciones solapadas y no se establece una partición del conjunto de datos.

DEFINICIÓN **3 (Solución de un problema de Biclustering)** *Dado un microarray M, las soluciones de una técnica de biclustering es un conjunto* $\mathbf{B} = \bigcup_{i=1}^{n} B_i$ *tal que se verifica:*

- $\bigcup_{i=1}^{n} B_i \neq M$
- *Existe* $i, j \in N$ *tal que* $B_i \bigcap B_j \neq \emptyset$

DEFINICIÓN **4 (biclusters)** *Dado el conjunto de soluciones de una técnica de biclustering* $\mathbf{B} = \bigcup_{i=1}^{n} B_i$*, se llama* bicluster *a cada elemento de dicho conjunto B_i. Dado un microarray $M = (G, C, f)$, cada biclúster se define como una 3-tupla $B = (I, J, f)$ tal que $I \subseteq G$, $J \subseteq C$ y f la función que asocia a cada término del microarray un valor real.*

$$\begin{pmatrix} b_{1,1} & \dots & b_{1,j} \\ \vdots & \vdots & \vdots \\ b_{i,1} & \dots & b_{i,j} \end{pmatrix}$$

4.2.1. NP-completitud

Intuitivamente diremos que un problema es NP-completo si no se puede resolver de ninguna manera en un tiempo "razonable", entendiendo tiempo como número de pasos de computación. Los problemas NP-completos son problemas que se plantean como problemas de decisión, que se pueden responder de manera afirmativa o negativamente, y se usan como un mecanismo para establecer la complejidad computacional del problema que se esté estudiando. Si se puede establecer una relación entre el problema de estudio y alguno de los problemas conocidos como NP-completos, se puede afirmar que no puede existir un algoritmo exacto que resuelva nuestro problema. Es decir, tan sólo podremos construir algoritmos que den una solución aproximada pero nunca la solución exacta o única del problema.

En su formulación estándar el problema del biclustering es un problema NP-completo. Antes de su irrupción en el campo del análisis de datos de expresión génica, fue estudiado en primero lugar por Morgan y Sonquist [67], y con posterioridad por Hartigan [44] y por Mirkin [63]. Su NP-completitud se estudia a través de su reducción a un problema clásico de grafos y combinatoria como es el problema del clique. La matriz de expresión se puede trasformar en un grafo bipartito de tal forma que por un lado los genes se transforman en una familia de nodos y por otro, las condiciones en otra familia de nodos. Las aristas que unen genes con condiciones llevan asociado un peso que será el valor de expresión de cada elemento de la matriz. De esta forma encontrar biclusters en la matriz de expresión se transforma en buscar bicliques de tamaño máximo en el grafo construido. Dadas dos familias de nodos, un biclique es un tipo de grafo bipartito donde cada vértice conecta un vértice de la primera familia con un vértice de la segunda de tal forma que se constituye un clique. Es decir, un grafo completo o aquel en el que todos los vértices son dos a dos adyacentes. El problema del clique es un problema clásico de combinatoria que se sabe que es NP-completo. Dicho problema trata de responder si es posible o no, dado un grafo y un tamaño en concreto, encontrar un grafo completo de dicho tamaño en el grafo. De esta forma la NP-completitud se establece reduciendo el problema del biclustering al problema del biclique.

En el trabajo presentado en [4] se estudian las distintas formulaciones del problema del biclustering y se demuestra la NP-completitud a través del problema del etiquetado de un grafo. Se define el problema del biclustering asociado al etiquetado de un grafo como: "dado un grafo bipartito G y un número entero k, determinar si G tiene un conjunto de grafos bipartitos completos, o biclusters, de al menos tamaño k" y se establce su

NP-completitud.

4.2.2. Patrones

Los distintos algoritmos de biclusterting definen heurísticas que buscan determinados patrones cuya definición trata de capturar información biologica relevante. En el trabajo presentado en [60] se establece una clasificación de estos patrones, que posteriormente es ampliada en [34].

Biclúster con valores constantes

Son aquellos biclusters que tienen todos los valores iguales, por lo que sus elementos se pueden describir de la forma $b_{i,j} = \mu$, siendo μ el valor constante para todos sus elementos. En general este tipo de biclusters no se encuentran en los datos debido a la presencia de ruido, por lo que su definición se modifica de tal manera que

$$b_{i,j} = \mu + \epsilon_{i,j}$$

siendo en este caso μ la media de los valores del bicluster y $\epsilon_{i,j}$ el error o diferencia de cada valor respecto a la media en cada valor.

Biclúster con valores constantes por filas o columnas

Son aquellos biclusters que presentan valores constantes o bien por filas o por columnas. Cada fila, o columna, tiene un mismo valor que se repite. Si la matriz de datos se preprocesa normalizando según la media de cada fila, o columna, serían biclusters con valores constantes. Cada elemento se puede representar como:

- $b_{i,j} = \mu + \alpha_i$, constante por filas, de manera aditiva.
- $b_{i,j} = \mu \times \alpha_i$, constante por filas, de manera multipliativa.
- $b_{i,j} = \mu + \beta_j$, constante por columnas, de manera aditiva.
- $b_{i,j} = \mu \times \beta_j$, constante por columnas, de manera multiplicativa.

Intuitivamente son aquellos biclusters que presentan patrones planos que son paralelos unos con otros.

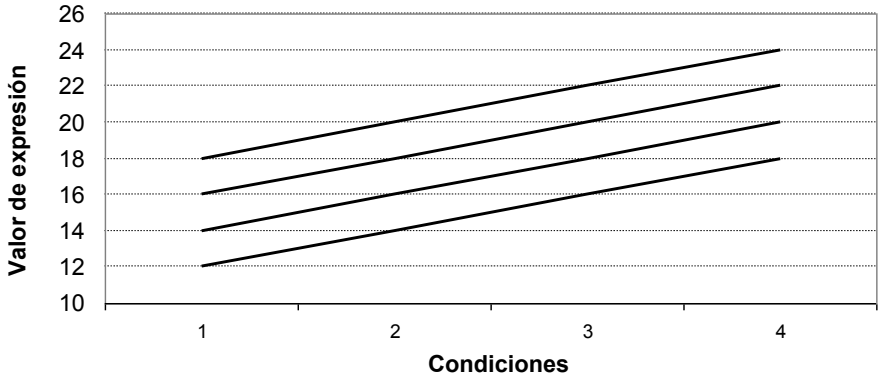

Figura 4.1: Bicluster con valores coherentes.

Biclúster con valores coherentes

Son aquellos biclusters que no tienen un valor constante por fila o columna porque cada valor recibe una contribución por fila y por columna. Se pueden representar como:

$$b_{i,j} = \mu + \alpha_i + \beta_j$$

siendo α_i el ajuste por filas y $\beta_j$el ajuste por columnas. El siguiente ejemplo muestra un biclúster que toma como valor base 5 y recibe una modificación según en qué fila y columna se encuentre.

$$\begin{bmatrix} 12 & 14 & 16 & 18 \\ 14 & 16 & 18 & 20 \\ 16 & 18 & 20 & 22 \\ 18 & 20 & 22 & 24 \end{bmatrix} = \begin{bmatrix} 5+5+2 & 5+5+4 & 5+5+6 & 5+5+8 \\ 5+7+2 & 5+7+4 & 5+7+6 & 5+7+8 \\ 5+9+2 & 5+9+4 & 5+9+6 & 5+9+8 \\ 5+11+2 & 5+11+4 & 5+11+6 & 5+11+8 \end{bmatrix}$$

La figura 4.1 muestra la representación gráfica del biclúster donde cada columna representa una línea del dibujo. Puede observarse claramente que son patrones paralelos aunque no planos, como ocurría en el caso anterior.

Biclúster con evoluciones coherentes

Los patrones con evoluciones coherentes son aquellos que capturan la idea de un mismo patrón de activación o inhibición entre todos los genes. Es decir, que todos los genes siguen una misma tendencia. Los genes crecen o decrecen a lo largo de las condiciones experimentales de la misma manera.

Se dice que siguen un patrón de desplazamiento si todos los genes siguen el mismo comportamiento con la misma intensidad y de escalado si lo hacen con intensidades equivalentes.

Un grupo de genes de un biclúster sigue un patrón de desplazamiento cuando su valor $b_{i,j}$ varía en la adición de una constante β_i. Análogamente, un *bicluster* sigue un patrón de escalado cuando sus valores $b_{i,j}$ varían en la multiplicación de una constante α_i. Formalmente, siguen la fórmula:

- $b_{i,j} = \pi_j + \beta_i$, patrón de desplazamiento.
- $b_{i,j} = \pi_j \times \alpha_i$, patrón de escalado.

Un ejemplo de un biclúster siguiendo patrones de desplazamiento es:

$$\begin{bmatrix} 30 & 10 & 5 & 15 \\ 33 & 13 & 5 & 15 \\ 40 & 20 & 15 & 25 \\ 50 & 30 & 25 & 35 \end{bmatrix} = \begin{bmatrix} 30+0 & 10+0 & 5+0 & 15+0 \\ 30+3 & 10+3 & 5+3 & 15+3 \\ 30+10 & 10+10 & 5+10 & 15+10 \\ 30+20 & 10+20 & 5+20 & 15+20 \end{bmatrix}$$

Un ejemplo de un biclúster siguiendo patrones de escalado es:

$$\begin{bmatrix} 48 & 80 & 16 & 96 \\ 24 & 40 & 8 & 48 \\ 18 & 30 & 6 & 36 \\ 12 & 20 & 4 & 24 \end{bmatrix} = \begin{bmatrix} 6\times 8 & 10\times 8 & 2\times 8 & 12\times 8 \\ 6\times 4 & 10\times 4 & 2\times 4 & 12\times 4 \\ 6\times 3 & 10\times 3 & 2\times 3 & 12\times 3 \\ 6\times 2 & 10\times 2 & 2\times 2 & 12\times 2 \end{bmatrix}$$

Los biclusters anteriores se representan en las figuras 4.2 y 4.3. La primera figura (Fig. 4.2) representa el biclúster anterior con patrones de desplazamiento. Estos patrones muestran genes con la misma forma y la misma pendiente. Los valores iniciales son diferentes para cada gen pero la forma de todos es la misma. En la figura 4.3 se representa el biclúster anterior con patrones de escalado. En este caso, los genes tienen la misma forma pero las pendientes no son las mismas. Los cambios entre cada gen son más marcados en este segundo caso.

Bicluster con evoluciones inversas coherentes

Aquellas situaciones en las que un grupo de genes se activan o expresan cuando un determinado gen, o un grupo, se desactivan son interesantes desde un punto de vista biológico. Este tipo de comportamientos son comunes

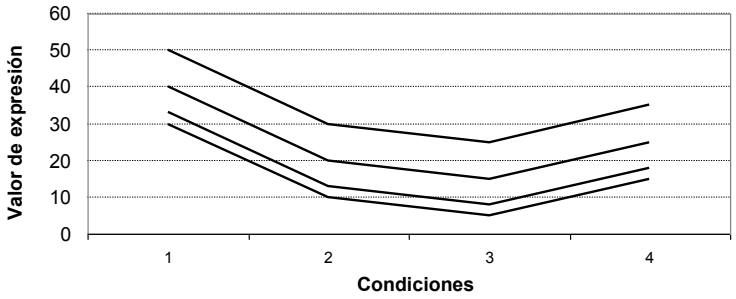

Figura 4.2: Bicluster con patrones de desplazamiento.

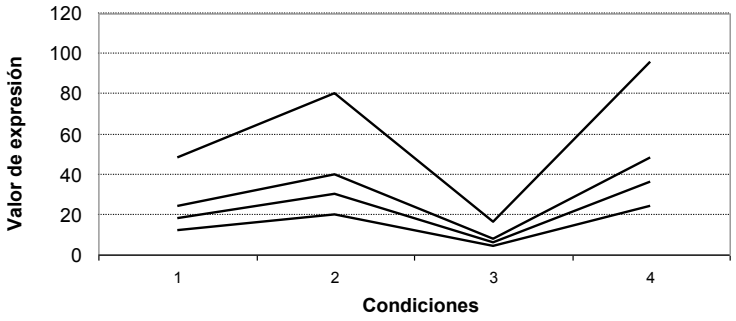

Figura 4.3: Bicluster con patrones de escalado.

en muchas rutas metabólicas como se muestra en [83] y en [36] y se ha estudiado con detalle en el trabajo presentado en [94] donde se les llama patrones de *inhibición-activación*. Se pueden considerar un caso particular de los patrones con evoluciones coherentes, en los que el factor de escalado divide en lugar de multiplicar, y el factor desplazamiento resta en lugar de sumar. La figura 4.4 muestra un biclúster formado por tres genes en el que uno de ellos actúa como inverso de los otros dos. Es decir, presenta un comportamiento inverso a los otros dos.

4.3. Principales algoritmos de Biclustering

Se pueden encontrar en la literatura una considerable cantidad de métodos de biclustering que se diferencian unos de otros según la técnica algorítmica en la que se basan, el tipo de resultado que obtienen, los mecanismos de validación que llevan a cabo, etc. La clasificación de estos no es un problema sencillo pues se pueden establecer diferentes criterios de ta-

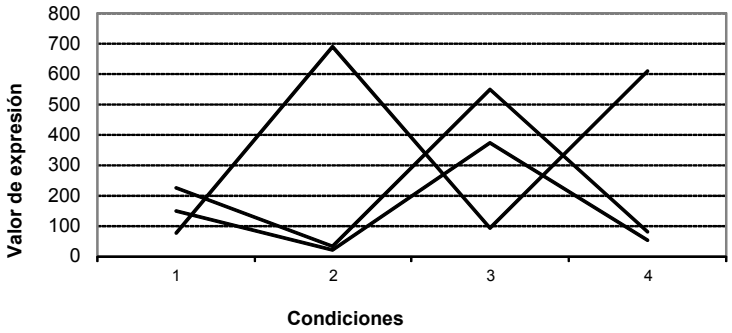

Figura 4.4: Bicluster con patrones de inhibición-activación.

xonomía y en función de ellos hacer varias familias. De hecho, uno de los grandes problemas en biclustering consiste en establecer cuál es el mejor algoritmo a utilizar para un análisis de datos en concreto. Debido a la estructura del problema, comentada previamente, no existe una metodología intrínseca de comparación entre las soluciones, a diferencia de cómo si ocurre en Clustering: distancias entre grupos, estructura de la partición, etc. Las comparativas que se establecen en la mayoría de artículos se basan en conocimiento experto del análisis de datos realizado, siendo el análisis de enriquecimiento biológico la metodología de comparación más aceptada. Por todo ello, a la hora de elegir qué algoritmo utilizar tiene más importancia la disponibilidad del software, o el tiempo de ejecución, que otro tipo de consideraciones más formales.

4.3.1. Algoritmos clásicos

Bajo la denominación *Algoritmos Clásicos* enmarcamos una serie de algoritmos de biclustering que suelen ser utilizados en las comparativas de los artículos. Estos algoritmos son frecuentemente usados debido a que fueron establecidos por su uso como marco de trabajo y sobre todo a la fácil disponibilidad de su código.

El algoritmo de Cheng y Church (***ChCh***) [24] es considerado como el trabajo fundacional dado que introduce la necesidad de la búsqueda de submatrices en datos de expresión génica para la extracción de información biológica relevante. Así mismo, muestra la no adecuación a estos datos del Clustering tradicional. Es un algoritmo voraz determinístico que busca aquellos biclusters que minimicen el valor de una medida llamada Mean Squared Residue (MSR), o Residuo Cuadrático Medio (eq. 4.1). El proceso comienza con la matriz completa del microarray quitando filas y columnas de forma

que se genera un bicluster y se evalúa el valor de su residuo. Si este valor es menor que un cierto umbral establecido como parámetro entonces se añaden aquellas filas y columnas que no aumenten dicho valor. Una vez encontrado el bicluster, los valores que encuentra son cambiados en el microarray por valores aleatorios y se repite el proceso de búsqueda tantas veces como biclusters se desee encontrar. Se debe introducir como parámetro de entrada tanto el umbral de búsqueda del residuo como el número de biclusters que se desea como resultado. Si I y J son el conjunto de filas (genes) y columnas (condiciones) de un bicluster, la medida MSR se define de la siguiente manera:

$$MSR = \frac{1}{|I||J|} \sum_{i \in I, j \in J} (a_{ij} - a_{iJ} - a_{Ij} + a_{IJ})^2 \tag{4.1}$$

donde a_{ij} es el valor del elemento que ocupa la fila i y la columna j (valor de expresión del gen i bajo la condición j), a_{iJ} es la media de todos los elementos que ocupan la fila i, a_{Ij} es la media de los que ocupan la columna j y a_{IJ} es la media del bicluster al completo.

El algoritmo ***FLOC*** [91] es una mejora que permite obtener los biclusters de manera simultánea durante el proceso y que permite manejar los valores perdidos o ausentes en el microarray.

El algoritmo ***ISA*** [12], *Iterative Signature Algorithm*, es un algoritmo voraz no determinista que busca biclusters que verifiquen que el valor medio de cada columna esté por debajo de un cierto umbral establecido para las columnas y, análogamente, el valor medio de cada fila esté por debajo de un umbral establecido para las filas. El proceso parte de un bicluster construido aleatoriamente que actúa como semilla y se añaden y quitan filas y columnas hasta que se alcanza convergencia. El proceso se repite con distintas semillas y de esta forma se obtienen distintos biclusters.

El algoritmo ***OPSM*** [11], *Order-preservering submatrix problem*, es un algoritmo voraz determinista que busca biclusters que preservan una determinada ordenación entre filas y columnas. La idea subyacente en el modelo de ordenación es encontrar submatrices cuyas filas sigan una misma tendencia de crecimiento o decrecimiento y por lo tanto conserven la ordenación. Se puede probar que dicho modelo captura los patrones más usuales como valores constantes, de desplazamiento o de escalado. Se introduce así mismo una corrección basada en una probabilidad para relajar las condiciones de dicho modelo y contemplar pequeñas variaciones, posibles errores en los datos, etc. El algoritmo trabaja iterativamente construyendo biclusters parciales los cuales se van incrementando en tamaño. Sólo se conservan los mejores biclusters construidos en cada paso de manera que la repetición del proceso

permite solapamiento entre los distintos biclusters que se encuentran.

El algoritmo ***xMotifs*** [68], *Conserved gene expression motifs*, es un algoritmo voraz no determinista que encuentra biclusters con valores constantes. Básicamente la idea consiste en hacer una partición para cada fila, dicha partición se realiza según la significancia estadística de cada intervalo generado comparado con una distribución uniforme. Una vez realizada se elige una columna, que actúa como semilla, y otro conjunto de columnas, que actúan como discriminantes, estableciendo así las cotas de un posible bicluster. A continuación se añaden filas y si se verifica que las nuevas filas tienen el mismo estado que la fila original se añaden al bicluster. El proceso se repite para cada fila pudiendo generar, por lo tanto, biclusters solapados. Una restricción importante al aplicar este algoritmo es el número de columnas del microarray, así por ejemplo en la implementación disponible en BiCAT [10] sólo admite microarrays que tengan menos de 64 condiciones. En general la mayoría de matrices de expresión tienen menos de 64 columnas pero en algunos casos, sobre todo experimentos de tipo temporal, puede sobrepasar este número de columnas.

El algoritmo ***SAMBA*** [85], *Statistical-Algorithmic Method for Bicluster Analysis*, es un algoritmo voraz que enumera de manera exhaustiva todos los biclusters que se pueden construir teniendo en cuenta un modelo basado en grafos bipartitos. El algoritmo se basa en técnicas de teoría de grafos y en técnicas de muestreo estadístico. Se define el microarray como un grafo bipartito que relaciona genes y condiciones en el que cada arista indica la relación entre ambos. Siguiendo un modelo estadístico se asignan pesos a cada arista. El algoritmo funciona añadiendo y quitando nodos de forma que se encuentran aquellos subrgrafos con un peso máximo. Bajo las restricciones del modelo el problema deja de ser NP-completo y el algoritmo funciona en tiempo polinomial.

4.3.2. Algoritmos basados en metaheurísticas

Se conocen como metaheurísticas a una familia de esquemas algorítmicos que encuentran soluciones aproximadas a problemas de búsqueda. Si estamos ante un problema cuya complejidad computacional impide la existencia de un algoritmo que encuentre una solución de manera eficiente, la aplicación de un esquema metaheurístico es una alternativa para su resolución. Se debe tener en cuenta que ante problemas para los que sí existe un algoritmo exacto, o clásico, éste siempre mejora la solución aportada por el algoritmo aproximado. Por lo tanto una metaheurística se debe aplicar sólo ante problemas computacionalmente costosos. Bajo el término metaheurística se

encuentran una variedad de esquemas algunos de ellos bioinspirados, como los algoritmos genéticos, los algoritmos de enjambres de partículas o lo algoritmos meméticos, aunque otras técnicas que no pueden ser calificadas como bioinspiradas como por ejemplo búsquedas tabú o de enfriamento simulado.

Como ya hemos comentado el problema del biclustering en su formulación estándar, no sujeta a ningún tipo de reformulación adicional o tratamiento previo, es un problema NP-completo. Existe por lo tanto una familia de algoritmos de biclustering basados en la aplicación de una metaheurística que optimiza una función de búsqueda que mide la calidad de los biclusters. La mayoría de estos algoritmos usan funciones objetivo basadas en el residuo cuadrático MSR (eq. 4.1), previamente comentado. El motivo de caracterizar de esta forma la calidad de los biclusters se debe básicamente a dos motivos. Por un lado el residuo interviene en el algoritmo de Cheng y Church que fue el artículo fundacional del biclustering y por otro lado, y con más fuerza, todos estos algoritmos establecen comparativas entre ellos de forma que es fundamental el uso de la misma función objetivo para poder establecer qué aproximación es mejor.

Las primeras aproximaciones se basan en algoritmos genéticos que son considerados algoritmos bioinspirados debido a que su idea intuitiva se basa en la teoría de la selección natural enunciada por Darwin, base de la Computación Evolutiva. La idea básica es la supervivencia de los mejor adaptados. De esta forma se codifican individuos que representan soluciones al problema considerado y cada cual se pondera con un valor que viene determinado por la evaluación de la función objetivo. Estás soluciones dan lugar a otras y sólo permanecen en el proceso aquellas con mejor ponderación. Al final la mejor o las mejores soluciones resultantes son la solución buscada del problema. Para llevar a cabo un esquema evolutivo como el descrito es necesario un mecanismo mediante el cual varios individuos se relacionen entre ellos y den lugar a nuevos individuos que también codifiquen soluciones al problema y, por otro lado, un mecanismo que evite que el proceso quede atrapado en una solución parcial. Los Algoritmos Genéticos son un tipo particular de Algoritmos Evolutivos en los que cada solución viene codificada como un gen que suele ser representado como una cadena con información binaria. Hay un mecanismo de generación de nuevas soluciones basado en operadores de cruces entre genes, de un punto de corte, dos puntos de corte, uniformes, etc., y otro que evita caer en soluciones parciales o mínimos locales que recibe el nombre de operador de mutación.

El algoritmo SEBI [29], *Sequential Evolutionary BIclustering*, presenta un esquema evolutivo basado en algoritmos genéticos en el que la función de optimización utilizada se basa en el residuo, MSR. El algoritmo consiste

en un algoritmo evolutivo de recubrimiento secuencial en el sentido que se ejecuta un algoritmo genético por cada bicluster que se obtiene. La función objetivo lleva asociada un término que controla que no se encuentre de forma reiterada la misma solución de manera que penaliza en la búsqueda aquellos biclusters ya obtenidos. En este trabajo se denomina a dicho factor como de control del solape. Se debe tener en cuenta que en este sentido es un mecanismo algorítmico para el control de no encontrar de forma reiterada siempre la misma solución. En general, el solapamiento entre resultados de un algoritmo de biclustering es en general deseable dado el significado biológico que representa. La función objetivo del algoritmo SEBI optimiza de tal forma que, además de minimizar el residuo y controlar la reiteración o solape de soluciones que se encuentran, maximiza el tamaño de los biclusters encontrados. Se trata de una combinación de varios términos en la suma que son controlados mediante unos parámetros.

Siguiendo este mismo enfoque se puede plantear el problema como un problema de optimización multiobjetivo. En [66, 9] se utiliza el algoritmo NSGA-II, un algoritmo evolutivo de optimización multiobjetivo, maximizando por un lado el tamaño de los biclusters y minimizando el valor del residuo. En la optimización multiobjetivo se encuentra todo un conjunto de soluciones candidatas y en función de a qué término se le conceda prioridad se elige una u otra. Ambos trabajos son de los mismos autores y presentan la idea expuesta aunque con distintos estudios experimentales en los que se enfatizan distintos aspectos. En [30] se adopta un mismo enfoque y se presenta un algoritmo de biclustering, llamado SMOB, que es una versión multiobjetivo del algoritmo SEBI también de los mismos autores.

En [16] se presenta un algoritmo evolutivo basado en Algoritmos Genéticos. Se aborda el problema como una optimización del tamaño de los biclusters sujeta a una restricción dada por la calidad de los mismos. Dicha restricción se establece según un umbral del residuo de cada bicluster. De esta forma se establece una búsqueda local por cada bicluster encontrado de modo que lo transforma hasta que se cumple la restricción impuesta. Esta búsqueda local acelera el proceso de búsqueda y se concluye que de esta forma se mejora el tiempo de ejecución de otras técnicas. Por lo tanto, el objetivo no es sólo encontrar buenos biclusters sino que el proceso tenga un rendimiento más efectivo.

En [37], teniendo en cuenta la misma motivación, se presenta un algoritmo genético que incorpora una búsqueda local en el proceso. En este caso no se establece como una restricción a satisfacer sino que mejora el valor de la función objetivo. Los mismos autores en [38] presentan un algoritmo de biclustering basado en un algoritmo multiobjetivo memético. Los Algoritmos

Meméticos son un tipo de algoritmos genéticos que incorporan una búsqueda local en el proceso de búsqueda. La intuición en la que se basan es que en el esquema de selección entre individuos no sólo importa la información heredada de los padres sino las condiciones ambientales que les proporciona una mejor adaptación. Es por ello que un algoritmo basado en el esquema de un algoritmo memético es básicamente un algoritmo genético al que se incorpora un mecanismo de mejora de soluciones.

Cuando se comenta búsqueda local hay que diferenciar dos tipos de búsquedas, aquellas que son independientes del significado de la función objetivo y aquellas que dependen de su semántica o significado. Las primeras se basan en un rastreo de tipo aleatorio de las soluciones próximas en el espacio de búsqueda a la que se quiere mejorar mediante la búsqueda local. No dependen del problema, las hay de varios tipos y son aplicables a cualquier problema de optimización. Las segundas mejoran la solución generando una nueva teniendo en cuenta modificaciones basadas en la función concreta que se optimiza. Las búsquedas locales antes descritas son la mayoría del segundo tipo y se basan en la mejora del residuo MSR.

En [27], por ejemplo, se aborda el problema basándose en una metaheurística conocida como GRASP Reactivo, *Reactive Greedy Randomized Adaptive Search Procedure*, que contiene una búsqueda local del primer tipo. El algoritmo de biclustering se basa en la generación de una solución que ejerce de semilla, que se encuentra mediante un algoritmo de clustering, y es el punto de salida de la metaheurística. Este tipo de algoritmos, GRASP, se basa en ordenaciones de las soluciones candidatas según un parámetro y la búsqueda voraz junto a mejoras locales de las soluciones.

Se pueden encontrar diversos trabajos que aplican las metaheurísticas más conocidas para abordar el problema del biclusterig. Todos ellos tienen en común el basarse en la optimización de una función en la que interviene el residuo cuadrático. Así por ejemplo, en [19] se aplica una metaheurística de Recocido Simulado, *Simulated Annealing*, que debe su nombre a los procesos de enfriamiento de materiales, los cuales siguen un proceso de equilibrio entre sus partículas y ahí se encuentra la idea de optimización. En [54] se aplica una metaheurística de Nubes de Partículas, *Particle Swarm Optimization*, que recibe el nombre del comportamiento de los enjambres de animales como, por ejemplo, enjambres de abejas. O, como también por ejemplo, en [53] que se aplica un algoritmo de Estimación de Distribuciones, *Estimation of Distribution Algorithm*, cuyo funcionamiento se basa en la definición de un grafo del espacio de búsqueda y la definición de una distribución de probabilidad. Dicha distribución se ajusta al grafo subyacente en el espacio de búsqueda y de esta forma se guía de manera eficiente el proceso de

optimización.

El problema del Residuo Cuadrático Medio como medida de calidad

Como hemos comentado, las aproximaciones anteriores se basan en la optimización de funciones basadas en el residuo cuadrático medio o MSR definido en el artículo fundacional de Cheng y Church. Sin embargo, en el caso de biclusters con evoluciones coherentes, la medida MSR presenta problemas. En el trabajo presentado en [1] se demuestra que la medida MSR depende de la varianza de los patrones de escalado, lo que implica que aquellos biclúster donde se observen un valor alto para la varianza del factor de escalado no son capturados por el residuo. Geométricamente esta situación se corresponde con biclusters que presenten picos muy pronunciados en su crecimiento o decrecimiento. Este tipo de comportamientos son los más relevantes biológicamente pues se corresponden con funciones de activación entre genes.

Este hecho motiva una serie de trabajos cuyo principal objetivo es el estudio de nuevas medidas de calidad para la evaluación de biclusters que corrijan los defectos del residuo como medida de ponderación [75, 74, 76].

4.3.3. Algoritmos basados en correlación

La correlación lineal entre genes ha sido utilizada como medida de calidad de biclusters en varios trabajos. Se basa en que los principales patrones se recogen en un biclúster que tenga sus genes correlacionados entre sí. Dos genes están correlacionados si uno se puede expresar como combinación lineal del otro y por tanto muestran ambos un mismo comportamiento. La correlación es ampliamente utilizada en el análisis de datos de expresión y otros tipos de datos ómicos. Por lo tanto se establece que genes co-expresados son equivalentes a genes correlacionados y se trabaja en consecuencia de esta forma.

Siguiendo este nuevo enfoque, que resuelve el problema del residuo cuadrático medio como medida de calidad, han surgido varios trabajos que se basan en la correlación como objetivo a optimizar.

En [88] se presenta un algoritmo que define una medida basada en la correlación de Pearson, no sólo entre genes sino también entre condiciones, y un proceso de búsqueda basado en un SVD, *Singular Value Decomposition*. Este tipo de proceso de búsqueda es parecido al utilizado en el *Spectral Biclustering* [49], que se basa en un proceso de cálculo de autovectores asociados a los autovalores de la matriz de los biclusters. El algoritmo propuesto es una

modificación que encuentra biclusters correlacionados gracias a la definición de la nueva medida.

BCCA [14], *Bi-Correlation Clustering Algorithm*, es un algoritmo voraz no determinista que realiza la búsqueda de biclusters teniendo en cuenta la correlación de Pearson como medida de similitud entre genes. El algoritmo depende de un parámetro de correlación según el cual se definen los biclusters a encontrar. Es exhaustivo y requiere un tiempo de ejecución considerable aunque se puede ejecutar en una versión no exhaustiva. Si se ejecuta de esta segunda forma el algoritmo actúa de manera que los resultados obtenidos no están solapados.

BICLIC [92], *BIclustering by Correlated and Large number of Individual Clustered seeds*, utiliza la correlación como medida y se basa en la expansión de una semilla. Primero se aplica un clustering a cada dimensión para construir un biclúster semilla y luego, mediante un proceso de expansión, se elabora una búsqueda en toda la matriz. Este algoritmo a diferencia de por ejemplo BCCA, obtiene tanto patrones correlados positiva como negativamente.

El algoritmo presentado en [34] se basa en la correlación de Spearman por filas y columnas en lugar de la correlación de Pearson. Se define una función de calidad basada en la correlación y se elabora un proceso de búsqueda basado en una metaheurística de (EDA) *Estimation of distribution algorithms*, algoritmos de estimación de la distribución en español. Este algoritmo como el anterior encuentra patrones correlacionados postiva y negativamente.

El uso de la correlación como medida de calidad para la búsqueda de biclusters se ha generalizado en muchos trabajos, aceptando que genes correlacionados implica co-expresados y por lo tanto co-regulados [6, 88, 51, 64, 15, 15, 92].

4.3.4. Otros algoritmos

Se pueden también destacar otros algoritmos que debido a las técnicas de búsqueda que utilizan han marcado tendencia. El algoritmo presentado en [50], ***Plaid Model biclustering***, es un algoritmo que modela estadísticamente la matriz de entrada como una superposición de capas, donde cada una se corresponde con un bicluster. De manera iterativa se ajustan los parámetros asociados a cada capa para estudiar el valor de su residuo (MSR) y de esta manera se obtienen los biclusters más relevantes. El algoritmo presentado en [49], ***Spectral biclustering***, utiliza técnicas basadas en el cálculo de autovectores, técnicas comunes en el álgebra lineal, para capturar biclusters que presenten un valor de varianza por debajo de un determinado

umbral.

Los trabajos presentados en [43] y [39] siguen de manera independiente un enfoque parecido. Se caracterizan los principales patrones como hiperplanos en espacios de alta dimensionalidad. Es decir, se considera el espacio generado por todas las condiciones y se trata de buscar variedades lineales en ese espacio. Este enfoque de tipo geométrico conlleva el uso de técnicas clásicas generalmente utilizadas en otro tipo de campos, como por ejemplo las técnicas de procesamiento de imágenes, aplicando de esta forma el algoritmo de detección de líneas de la *transformada de Hough*. El artículo presentado en [43] es el primero que usa por primera vez este enfoque, sin embargo [39] usando ideas parecidas presenta un estudio experimental más riguroso y que sigue un enfoque más cercano al problema biológico.

Algunos algoritmos de biclustering requieren una discretización de los datos, siendo este paso de vital importancia según qué tipo de resultados se deseen obtener. El algoritmo de ***BiMAX*** presentado en [77] es el principal representante de este grupo de algoritmos. Este algoritmo se encuentra disponible entre los algoritmos de la herramienta software presentada en [10], lo que suele considerarse referente entre aquellos algoritmos que requieren discretización previa. Requiere un alto coste computacional de cálculo y otros algoritmos como [81] o [51] trabajan también con datos discretos y mejoran su coste computacional.

Los algoritmos ***eCCC-biclustering*** y ***CCC-biclustering***, presentados en los artículos [58] y [59] respectivamente, trabajan con datos discretos pero se especializan en datos de expresión temporales. Es decir, las columnas o condiciones de la matriz de expresión deben respetar el orden en el que se encuentran puesto que representan datos de tipo temporal. En estos artículos por ejemplo los datos son momentos del ciclo celular de la levadura. De esta manera la complejidad computacional del problema del biclustering se reduce pasando a ser un problema computacionalmente tratable, no es NP-completo. Gracias a la discretización de los datos, y la restricción de columnas continuas en los resultados, se pueden aplicar técnicas de tratamientos de cadenas. Ambos algoritmos se basan en el conocido algoritmo de Ukonnen de búsquedas de cadenas similares que se aplica de manera local iterativamente. La diferencia entre ambos algoritmos, presentados por los mismos autores, consiste en la tolerancia a errores en los patrones encontrados. Se considera positivo admitir ciertos errores en la identificación de las cadenas y de esta manera se pueden corregir errores cometidos en el paso de discretización.

4.4. Metodología de comparación y validación

Debido a la gran cantidad de algoritmos de biclustering que se pueden encontrar estos últimos años en la literatura, es de gran importancia abordar el problema de la comparación entre los mismos y los mecanismos para determinar qué algoritmo es el más adecuado según la naturaleza de los datos que se manejen. Los principales *surveys* presentes en la literatura [60, 86, 20] no abordan el estudio comparativo entre los distintos algoritmos sino tan sólo su clasificación. Se pueden clasificar los distintos algoritmos según la técnica de búsqueda y según los patrones que encuentra.

El problema del biclustering es un problema de aprendizaje no supervisado, por lo tanto la manera de determinar si un algoritmo obtiene buenos resultados consiste en recurrir a conocimiento externo al problema. Por lo tanto, se debe tener un conocimiento profundo de los datos de estudio y se trata de determinar la calidad de los resultados obtenidos. El artículo [77] plantea una comparación entre algoritmos en función del porcentaje de biclusters enriquecidos que cada algoritmo devuelva. Se dice que un *biclúster está enriquecido* si es estadísticamente significativo respecto a un conjunto de genes de referencia. Se toma como referencia la ontología de genes GO [5] y se trata de determinar si el grupo de genes de un biclúster está presente en un término de dicha ontología. Para ello se realiza un contraste de hipótesis, normalmente el test de Fisher, que como se repite respecto a todos los términos GO presentes en las anotaciones se lleva a cabo un contraste múltiple de hipótesis. Este test múltiple se debe corregir mediante otro proceso estadístico, normalmente el método de Bonferroni. Como cada algoritmo obtiene un número distinto de resultados la comparativa se establece en función del porcentaje de biclusters enriquecidos. Hay que tener en cuenta que estos estudios experimentales suelen presentar problemas de varios tipos: cómo son los ficheros de anotaciones de términos GO utilizados, como se determina que un grupo de genes esté presente o no en un término, de nomenclatura de genes, etc.

En biclustering es más complicado que en clustering la definición de unas medidas de validación en función de la estructura de los resultados. Téngase en cuenta que se puede medir la homogeneidad de los resultados pero no la separación de los mismos dado que no la cumplen. Los criterios de homogeneidad se establecen en función de una medida de tipo estadístico, generalmente el residuo o la correlación, y se analizan los resultados de esta manera. De manera alternativa podemos encontrar en muchos artículos un método de comparación entre algoritmos basados en datos sintéticos [32]. Se genera una matriz de manera artificial y se almacena dentro de ella un

conjunto de biclusters que se estimen buenos. Se estudian y se comparan los distintos algoritmos entre sí según su capacidad de detectar estos biclusters. Se suele utilizar el *índice de Jaccard* [32].

El problema que presenta un enfoque basado en datos sintéticos es que la distribución de los datos debe ajustarse lo máximo a la realidad. Es decir, no pueden estar resaltados los biclusters que se buscan debido a que los valores que los rodean son excesivamente diferentes en rango, tendencia, etc. De esta forma se daría un sobreajuste entre los resultados a buscar y el mecanismo de búsqueda del algoritmo y la validación no sería realista. Por otro lado, determinar qué significa que un biclúster sea o no bueno introduce también de manera inevitable un sobreajuste en la validación.

Algunos trabajos establecen la comparativa entre algoritmos abriendo el espectro de la validación de los genes [52], trabajando por ejemplo con redes de proteínas y no sólo con GO, o utilizando índices internos de carácter descriptivo y técnicas de visualización de datos [82].

4.5. Nuevas tendencias

Existen gran cantidad de algoritmos de biclustering como hemos podido ver en este capítulo. La comparación entre ellos, y determinar qué algoritmo ofrece un mejor rendimiento para unos datos y un problema en concreto, no es una tarea fácil. De hecho se puede considerar todavía un problema abierto en este campo. Actualmente se considera que son más interesantes las aplicaciones, e integración del biclustering dentro de estudios más generales, que el desarrollo de un algoritmo en concreto que compita con los demás en tiempo, precisión o sensibilidad. No obstante el desarrollo de nuevas medidas de calidad, para una mejor detección de los patrones que se buscan, y el desarrollo de algoritmos, sigue siendo campos de investigación relevantes. Así por ejemplo hay estudios sobre el funcionamiento de nuevas medidas, como la información mutua propia de la teoría de la información, en el problema de la detección de biclusters [42]. Otros trabajos, como [46], mejoran técnicas clásicas de biclustering mejorando su tolerancia a ruido, escabilidad o flexibilidad. En este caso se mejora el algoritmo de OPSM mediante técnicas de tratamiento de local de cadenas.

Se conoce como tuberías de técnicas o *pipelines* al encadenamiento de distintas técnicas para la resolución de un problema. De esta manera un algoritmo de biclustering se convierte en una pieza que se integra en un esquema algorítmico más general. Por ejemplo, en [22] se aborda un problema de clasificación en el que juega un papel relevante el biclustering. Se define el concepto de metabiclúster que se utiliza como pieza fundamental en la

clasificación de datos. Este trabajo es un ejemplo interesante de aplicación de biclustering en el marco de un problema general de desarrollo de modelos de prognosis de enfermedades. Así mismo, en [56] el biclustering se enmarca en un problema de análisis de datos de espectrometría de masas. En [47] se integra el biclustering en un marco de descubrimiento de módulos reguladores en el que se utilizan datos de expresión génica y de secuenciación, junto con datos auxiliares de tipo filogenéticos. En [26] se utiliza el biclustering para mejorar un método de detección de subredes de proteínas que sirven como marcadores para desarrollar modelos de prognosis en cáncer.

En general, se debe tener en cuenta que el problema de detección de módulos de una red es un problema de búsqueda que utiliza ideas similares a los algoritmos de biclustering. En el marco de las redes de genes interesa módulos que no sean disjuntos entre sí y, por otro lado, el interés radica en la búsqueda de patrones locales más que en la descripción de la red en su conjunto [90]. Estas ideas constituyen una oportunidad de trasladar las ideas propias del biclustering al problema de búsqueda de módulos de una red con el objetivo de definir nuevos marcadores. Análogamente, las ideas propias del biclustering se pueden trasladar al estudio de otros tipos de datos como por ejemplo el estudio de datos de microRNA [73]. Se trata de asociar cadenas de microRNAs con sus genes objetivos, o *target genes*, por lo que tras un procesamiento inicial de los datos de entrada se puede generar una matriz que enfrente microRNA y genes candidatos, teniendo como objetivo agrupar cadenas de microRNAs con comportamiento similar para un grupo de genes.

Finalmente, la integración entre distintas fuentes de datos es uno de las tendencias actuales en Bioinformática [7]. La proliferación de repositorios públicos de fácil acceso, así como proyectos como GO, KEGG, etc, facilitan la fusión de distintas fuentes de datos. Siguiendo estas ideas se puede realizar el estudio de cómo afectan a los algoritmos de biclustering conocidos la integración de información adicional proveniente de otro tipo de datos. De esta manera se consigue introducir un sesgo en la búsqueda y se utiliza información extra que ayude a encontrar los biclusters.

Capítulo 5

Apuntes sobre integración de información biológica

Es misión de la ciencia traspasar las fronteras del conocimiento, indagar más allá y ampliar los límites del saber humano, añadiendo más datos que nos permitan comprender el complejo mundo en que estamos inmersos.
Explorando los genes. Del Big-Bang a la nueva Biología. Nicolás Jouve (pág. 265).

5.1. Introducción

En el contexto del biclustering se suelen usar repostorios biológicos públicos como *the Gene Ontology project* (GO) o *Kyoto Encyclopedia of Genes and Genomes* (KEGG) para tareas de validación y comparación entre los resultados de distintos algoritmos. En el caso de GO, por ejemplo, se suele usar el enriquecimiento de genes como criterio para elaborar rankings de comparación entre algoritmos. Dados los biclusters obtenidos por cada algoritmo, se estudia el porcentaje de ellos que son estadísticamente significativos en alguno de los términos de la ontología. Es decir, si existe un subgrupo de genes de ese biclúster que están relacionados con algún término GO. Según ese porcentaje se establce el ranking entre los resultados obtenidos por los distintos algoritmos [77]. GO es una ontología con una estructura jerárquica con tres ramas o dominios: BP o procesos biológicos, CC o componentes celulares y MF o funciones moleculares. Cada término de la ontología tiene asociado un grupo de genes, que son anotados según la información obtenida experimentalmente, mediante análisis informáticos, etc. Los términos más bajos en la jerarquía son más específicos que los más altos, que engloban

a mayor número de genes bajo una descripción más general. Los ficheros de anotaciones de genes relacionan términos de GO con el conjunto de genes que están relacionados o anotados para dicho término. Este tipo de ficheros son los que generalmente se usan en las labores de validación y comparación de los resultados de los algoritmos de biclustering.

Toda la información almacenada en estos repositorios se utiliza por lo tanto a posteriori del proceso de búsqueda propio del biclustering. Por otro lado, en biclustering, al igual que en el resto de técnicas de análisis de datos de expresión génica, se asume que si un grupo de genes están co-expresados entonces comparten una determinada funcionalidad común. Esta hipótesis de partida ha sido criticada por algunos autores que ponen en cuestión la completitud de los datos. Se argumenta que pueden existir grupos de genes que se co-expresen en simultáneo, para una misma condición, como resultado de procesos paralelos pero independientes y no por ello tienen por qué compartir una misma funcionalidad o proceso biológico [18]. Es decir, la información que se pueda extraer de los datos de expresión nos da una guía de investigación, o una pista, pero no es concluyente. En otras palabras, genes co-expresados pueden ser genes co-regulados pero se necesita información extra para llegar a la conclusión.

Este tipo de ideas motivan el uso de información complementaria para un análisis más profundo de los datos de expresión génica, no ya como herramienta de validación a posteriori, sino insertando toda la información en las mismas técnicas o algoritmos. En general la integración de la información biológica proviniente de distintas fuentes de datos es uno de los retos y líneas de trabajo más prometedoras actualmente en Bioinformática [7].

5.2. Integración de información biológica

En el campo del biclustering no se ha iniciado el desarrollo de algoritmos híbridos que integren información a priori. Sin embargo, en clustering se pueden encontrar ya algunas propuestas que integran información biológica en el proceso de búsqueda. Por ejemplo, el algoritmo presentado en [87] se basa en el algoritmo de clustering del K-means e integra información de GO mediante fichero de anotaciones directa. Se define una distancia que se basa en la idea de co-expresión y similitud funcional entre los genes. En clasificación, por ejemplo, el algoritmo presentado en [8], que clasifica grupos de genes, integra una medida basada en GO como parte del flujo de trabajo para llevar a cabo dicha clasificación. En concreto se usa la correlación de Pearson para medir la co-expresión entre los genes y, por otro lado la medida basada en GO para definir un concepto de distancia. Dicha distancia se usa

para construir grupos de genes para los que se elabora un ranking de genes candidatos. En el campo de la selección de atributos el método presentado en [57] integra información proveniente de KEGG en lugar de GO. El algoritmo es un algoritmo evolutivo que gracias a la información de KEGG mejora los resultados de los algoritmos clásicos de selección de atributos.

En general se pueden encontrar otros trabajos que integran información proveniente de distintas fuentes de datos, no sólo GO o KEGG, para mejorar el rendimiento. Por ejemplo, en el trabajo presentado en [55] se fusiona información de la interacción de redes de proteínas, datos de tipo genómico e información extraída de la literatura (*literature mining*) de manera simultánea. El algoritmo COALESCE [47] busca módulos de regulación de genes usando datos de expresión génica junto con datos de secuenciación de ADN. El algoritmo propuesto en [26] usa redes de interacción de proteínas y datos de expresión génica para descubrir biomarcadores relacionados con diversos tipos de cáncer. Dichos biomarcadores son grupos de genes que además de estar co-expresados tienen una alta conectividad en las redes de proteínas. Más recientemente, y en el campo del biclustering, aunque aplicado a datos de microRNA en lugar de a datos de co-expresión, el algoritmo presentado en [73] usa en uno de sus pasos una medida basada en GO para elaborar un ranking entre los biclusters que encuentra.

5.2.1. Medidas de similutud funcional entre genes basadas en GO

Se pueden definir medidas de distancias entre genes según la información sobre dichos genes almacenada en GO. Existen multitud de medidas de similitud semántica para comparar términos de GO donde los genes están anotados. Estas medidas no miden la distancia entre dos genes sino entre los propios términos de la ontología y a través de estos se puede establecer la medida entre los genes. Estas medidas se pueden clasificar básicamente en dos grupos: medidas basadas en la topología de la ontología, o *edge-based measures*, y medidas basadas en la información almacenada dicha ontología, o las *information content (IC)-based measures*. El primer grupo asume que la especifidad de un término se puede inferir directamente de su profundidad en el grafo de GO, se basa en la topología del grafo de GO. El segundo grupo por otro lado se basa en la frecuencia de aparición de un término en dicho grafo más que en el lugar en el que se encuentra.

En general un mismo gen suele estar anotado en más de un término GO por lo que, si el objetivo es establecer distancias entre genes, es más adecuado usar medidas de similitud que comparen grupos de términos GO

en lugar de términos individuales entre sí. En la referencia [72] se presenta una comparación entre estas medidas y se concluye que $SimGIC$ es la que obtiene mejor resultado. Esta medida se usa para computar la similitud entre dos genes usando el contenido de información (*information content* o IC) asociado con cada término GO asociado a cada gen. Para poder computar el IC se necesita no sólo los ficheros de anotaciones extraídos de GO sino también su estructura de grafo (que se puede obtener como un fichero con la extensión .obo). Otras medidas de similitud se basan en la construcción de matrices binarias que recogen la estructura de GO. Estas matrices tiene los genes como filas y los términos GO como columnas y cada elemento indica si un gen está o no anotado en un término GO. Estas medidas se inspiran en ideas propias de teoría de la información. La medida presentada en [65] se puede clasificar como una medida de similitud de semántica pero no necesita ni un preprocesamiento previo, para construir una matriz, ni un fichero diferente del fichero de anotaciones con la estructura del grafo de GO. Esta medida se basa en el solapamiento entre los términos GO asociados a dos genes y de esa forma se mide la similitud entre ellos. Para esta medida es esencial la manera en la que el fichero de anotaciones, que relaciona genes y sus términos GO asociados, esté construido. Cuando el fichero de anotaciones tiene en cuenta un término y todos sus términos padre en la ontología, se consigue reflejar la estructura jerárquica de la ontología.

Parte III

Propuestas

Capítulo 6

Biclustering usando la Búsqueda Dispersa como motor de búsqueda

Johnny[1] *me dijo un día:"Este cálculo necesitará más multiplicaciones que las que haya hecho antes toda la humanidad". Pero cuando estimé el número de multiplicaciones hechas por los niños de los colegios de todo el mundo en los últimos cincuenta años, descubrí que el número que yo había estimado ¡era unas diez veces mayor!*
Aventuras de un matemático. Memorias de Stanislam M. Ulam (pág. 210).

6.1. Introducción

Dada la matriz de expresión de un conjunto de genes, se trata de encontrar aquellos biclusters que proporcionen información relevante. La exploración de los datos se puede plantear como un mecanismo de búsqueda al que se le proporciona un criterio que mide la calidad de cada bicluster. Bajo esta perspectiva, el problema de *biclustering* puede ser abordado definiendo un esquema algorítmico que optimice una determinada función objetivo. Dicha función proporcionará una manera de decidir sobre la calidad de cada bicluster y de esta manera, siguiendo el valor que proporciona dicha función, se puede guiar la exploración. Debido a la complejidad computacional del problema se utilizará un esquema metaheurístico.

[1]Se refiere a John von Neumann, compañero y amigo de Stanislam M. Ulam.

La *Búsqueda Dispersa* o *Scatter Search* es una metaheurística de optimización basada en la evolución de un pequeño conjunto de soluciones. Como toda metaheurística, es un esquema de búsqueda que establece una serie de reglas para alcanzar la mejor solución del problema estudiado. Se trata de un algoritmo aproximado que encuentra una solución cercana a la solución óptima sin llegar a alcanzarla. Sólo debe aplicarse a aquellos problemas que no sean computacionalmente tratables y para los que no existe un algoritmo exacto que los resuelva. Las metaheurísticas basadas en poblaciones usan la evaluación de cada posible solución generada como mecanismo de optimización. En nuestro caso cada solución codificará un bicluster y por lo tanto el algoritmo generará un gran número de biclusters para poder decidir cuál es el mejor de todos ellos. La combinación entre distintas soluciones generan otras nuevas y aquella que tenga el mejor valor respecto a la función de evaluación al final del proceso se considera el óptimo del problema. La principal diferencia de la búsqueda dispersa con otras metaheurísticas basadas en poblaciones, como los algoritmos genéticos, es que tan sólo evoluciona un pequeño grupo de soluciones en lugar de toda la población. Este pequeño conjunto de soluciones tiene que recoger la mayor información posible del espacio de búsqueda y de esta forma se consigue alcanzar la solución óptima con más eficiencia. Para su construcción por un lado se elige un primer grupo de soluciones que tienen los mejores valores para la función objetivo de entre todas aquellas generadas en la población inicial. Por otro lado, se eligen aquellas soluciones lo más distintas posibles respecto del primer grupo teniendo en cuenta la codificación del problema. Se dice de ellas que son lo más dispersas posibles. De esta manera se garantiza la exploración de soluciones que cubren distintas zonas del espacio de búsqueda. Mediante este mecanismo de elección de calidad y diversidad se puede aplicar el esquema evolutivo a un pequeño grupo en lugar de a toda la población inicial de soluciones generadas.

6.2. Búsqueda Dispersa: esquema básico

En la figura 6.1 se puede observar el esquema básico de un algoritmo de biclustering basado en una metaheurística de búsqueda dispersa. La entrada es la matriz de expresión del conjunto de genes y el número de biclusters que se van a buscar. Como se verá más adelante, según qué versión del algoritmo vayamos aplicar se tienen que añadir algunos parámetros de entrada adicionales. Cada bicluster se obtiene mediante un esquema de búsqueda dispersa que se aplica tantas veces como número de biclusters se haya introducido como parámetro de entrada.

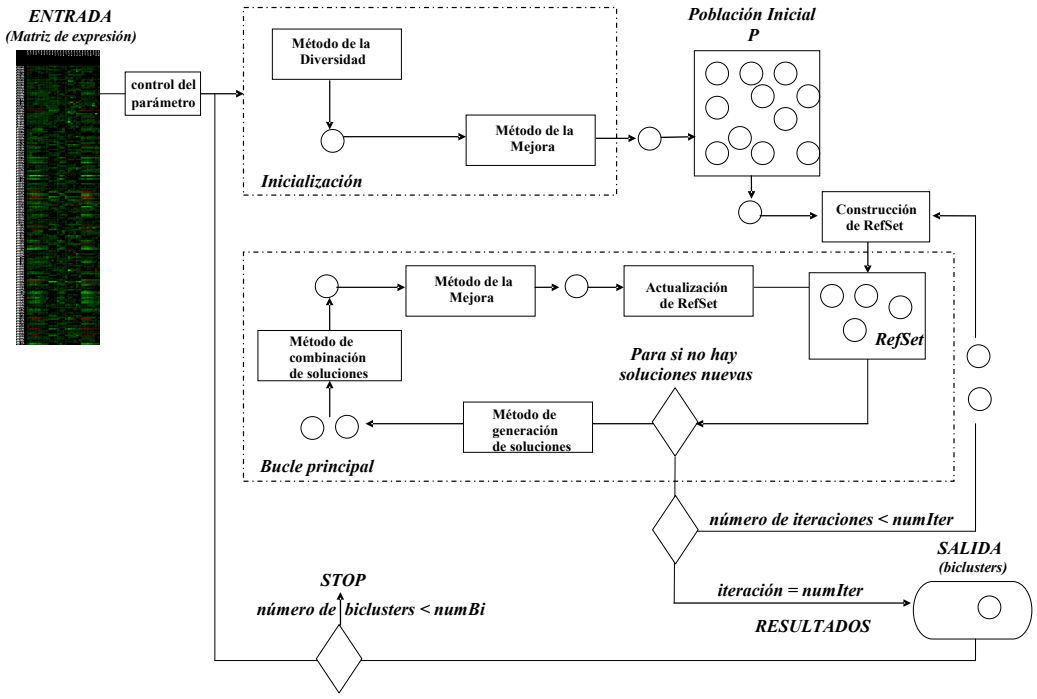

Figura 6.1: Esquema Búsqueda Dispersa o *Scatter Search* para Biclustering.

Para obtener un bicluster el esquema de búsqueda dispersa comienza construyendo una población inicial de posibles soluciones que actúa como inicialización del proceso. Cada solución codifica un bicluster al que se le asocia un peso que viene determinado por el valor que le otorga la función objetivo o de optimización. El número de soluciones de esta población inicial es una decisión de diseño, siendo el valor estándar 200 en la mayoría de los esquemas clásicos de búsqueda dispersa [61]. Las soluciones no se generan de forma aleatoria sino de forma que sean tan diferentes entre sí como sea posible. Esto se consigue aplicando el *método de diversificación*. Así mismo, se aplica una búsqueda local mediante el *método de la mejora*, de forma que el punto de partida sean soluciones con la mejor calidad posible.

Una vez construida la población inicial se genera un pequeño conjunto de soluciones que sea lo más descriptivo posible del espacio de búsqueda. Dicho conjunto consta de un reducido número de soluciones, en los esquemas clásicos tan sólo 10 soluciones [61], y recibe el nombre de *conjunto de referencia*. El espacio de búsqueda es el conjunto de todas las posibles soluciones que se puedan generar; es decir, todos los posibles biclusters. Debido a la complejidad computacional del problema del biclustering, que como vimos

Algorithm 1 Esquema de búsqueda dispersa para biclustering

INPUT microarray D, número de biclusters que se encuentran $numBi$, parámetro para controlar el tamaño M

OUTPUT Conjunto *Results* con $numBi$ biclusters.

begin
1: $num \leftarrow 0$, $Results \leftarrow \emptyset$
2: **while** ($num < numBi$) **do**
3: $Population \leftarrow$ DiversificationGeneration()
4: $Population \leftarrow$ Improvement($Population$,ρ)
5: $Reference\ Set \leftarrow$ Build($Population$)
6: $Population \leftarrow Population \setminus Reference\ Set$
7: $i \leftarrow 0$
8: **while** ($i < numIter$) **do**
9: **while** (NOT stable) **do**
10: $A \leftarrow$ SubsetGeneration($Reference\ Set$)
11: $B \leftarrow$ SolutionCombination(A)
12: $B \leftarrow$ Improvement(B,ρ)
13: $Reference\ Set \leftarrow$ Update(B,$Reference\ Set$)
14: **end while**
15: $Reference\ Set \leftarrow$ Rebuild($Population$,$Reference\ Set$)
16: $Population \leftarrow Population \setminus Reference\ Set$
17: $i \leftarrow i + 1$
18: **end while**
19: $Bicluster \leftarrow$ the best one from $Reference\ Set$
20: $Results \leftarrow Results \cup \{Bicluster\}$
21: $num \leftarrow num + 1$
22: **end while**
end

en el capítulo anterior se trata de un problema NP-completo, no se pueden comprobar una a una cada posible solución del espacio de búsqueda sino que la búsqueda se debe realizar de manera aproximada sin recorrer todo el espacio de búsqueda. Para ello la búsqueda dispersa construye un conjunto de soluciones representativas que recibe el nombre de conjunto de referencia. Las soluciones que constituyen el conjunto de referencia se eligen según dos criterios. Un primer grupo que recoge las mejores soluciones de entre todas las de la población inicial y otro segundo grupo con aquellas más diferentes de la población respecto de las elegidas en el primero. De esta forma se consigue no sólo elegir las mejores soluciones de la población, que harán que el proceso de búsqueda se aproxime a la mejor solución posible, sino que se introduce diversidad en la búsqueda, de forma que se explore la mayor parte posible de todo el espacio de búsqueda y se trata, de este modo evitar que el proceso caiga en lo que se conoce como óptimos locales. Esta forma de construir el conjunto de referencia consigue un equilibrio entre soluciones

buenas que *intensifiquen* el proceso de búsqueda y, por otro lado, soluciones lo más dispersas posible, y de ahí viene el nombre de búsqueda dispersa, que logren la *dispersión* en el proceso de búsqueda. Es importante advertir que la población inicial es actualizada quitando de ella aquellas soluciones que son incluidas en el conjunto de referencia.

Una vez construido el conjunto de referencia por primera vez se inicia el proceso de estabilidad del mismo. Este proceso es el bucle interno del algoritmo (ver Algoritmo 1) donde se realizan una serie de modificaciones en dicho conjunto hasta que no se pueden encontrar mejores soluciones desde el punto de vista de la función objetivo. En concreto, se combinan todas las soluciones de dicho conjunto entre sí mediante el método de *generación de soluciones* y se generan soluciones nuevas que introduzcan nuevos biclusters en el proceso mediante el *método de la combinación*. De esta forma las diez soluciones del conjunto de referencia han generado un número nuevo de soluciones novedosas que antes de ser consideradas son mejoradas aplicando la búsqueda local que codifica el método de la mejora. Teniendo en cuenta las soluciones originales del conjunto de referencia y todas las nuevas que han sido generadas se genera un nuevo conjunto de referencia mediante el *método de actualización del conjunto de referencia*, que elige las diez mejores según la función objetivo. Si este nuevo conjunto es igual que el original se dirá que ha alcanzado la estabilidad en el proceso y el algoritmo continúa. En caso contrario, se repite otra vez el proceso hasta que la estabilidad sea alcanzada y el conjunto generado sea igual que el del paso anterior. Hay que tener en cuenta que el proceso converge siempre a una situación estable siendo el número de pasos del mismo un indicativo del grado de intensificación de la búsqueda.

Una vez alcanzado un conjunto de referencia estable se reconstruye y se repite el proceso anterior. Esta operación se realiza un número establecido de veces, que en los esquemas clásicos de búsqueda dispersa suele ser veinte [61]. La reconstrucción del conjunto de referencia se realiza considerando las cinco mejores soluciones del conjunto de referencia junto con otras cinco que se eligen de la población inicial y son lo más dispersas posibles respecto de las cinco anteriores. Tras ello se actualiza la población inicial. Una vez repetido el proceso las veinte veces, se toma la mejor solución del conjunto de referencia, el bicluster con el mejor peso, y se almacena en el conjunto de resultados. Este bicluster será uno de los biclusters obtenidos por el algoritmo. El esquema anterior se repite tantas veces como señale el parámetro de entrada introducido, que indica el número de biclusters que se desean obtener.

El pseudocódigo presentado en el algoritmo 1 recoge el esquema anterior

y permite leer rigurosamente cómo funciona el proceso. Se puede observar con claridad que el algoritmo es un esquema de búsqueda secuencial de biclusters en el sentido que un mismo proceso se repite un número de veces, una por cada bicluster que se quiera encontrar. El bucle general, línea 2 a la 22, se repite tantas veces como indique el parámetro de entrada número de biclusters buscados. Entre las líneas 8 y 18 se describe el bucle principal del esquema de búsqueda dispersa, que se encarga de reconstruir el conjunto de referencia y que anteriormente hemos comentado que en los esquemas clásicos se suele repetir veinte veces. Dentro de este bucle se encuentra el bucle de estabilidad, líneas 9 a la 14, en el que evoluciona el conjunto de referencia hasta que alcanza la estabilidad, como hemos descrito anteriormente. Se puede observar en las líneas 19 y 20 cómo se almacena en el conjunto de resultados el mejor bicluster del conjunto de referencia resultante del esquema de búsqueda dispersa (línea 3 a la 18). Es importante observar que la población inicial tan sólo actúa como arranque del proceso pero tan sólo sirve como apoyo del mismo, líneas 3 a la 6. La intensificación de la búsqueda se consigue en el bucle de estabilidad, líneas 9 a la 14, mientras que la diversificación de la misma se logra principalmente cuando se construye o reconstruye el conjunto de referencia, líneas 5 y 15. Las líneas 6 y 16 son fundamentales para que la población inicial se vaya actualizando y aporte información nueva en los sucesivos pasos del proceso.

6.3. Búsqueda Dispersa: métodos y detalles

Describimos a continuación los diferentes métodos que hemos descrito tanto en la figura 6.1 como en el pseudocódigo del algoritmo con más detalle (ver Algoritmo 1).

6.4. Codificación de las soluciones

Se puede describir un bicluster enumerando los genes o filas de la matriz del microarray que lo constituyen junto con sus columnas o condiciones correspondientes. Con tan sólo esa descripción se puede construir dicho microarray. Siguiendo esta idea, y basándonos en trabajos previos, se ha codificado cada bicluster como una cadena de ceros y unos. Con más detalle, si un bicluster B es una matriz M compuesta de $n \leq N$ filas o genes y $l \leq L$ columnas o condiciones, se codifica como una cadena binaria de tamaño $N + L$ donde los N primeros bits codifican los genes y los L segundos las condiciones. La figura 6.2 muestra un pequeño ejemplo.

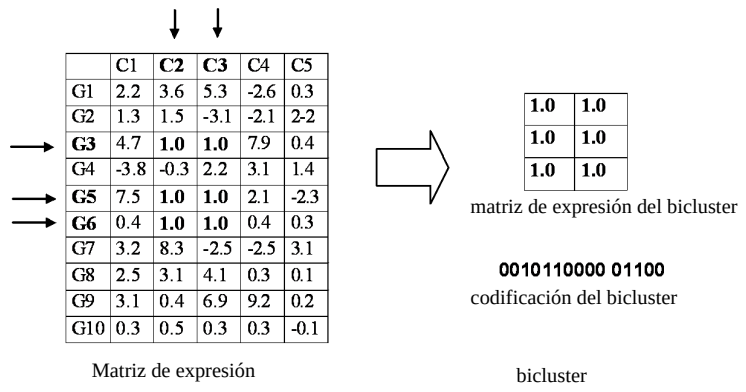

Figura 6.2: Microarray y bicluster {G3,G5,G6—C2,C3} con su codificación.

6.4.1. Método de diversificación

El método de la diversificación se encarga de la construcción de la población inicial de las soluciones de forma que estas sean lo más dispersas o diferentes entre sí. De esta manera la población inicial de soluciones consigue describir mejor el espacio de soluciones que un mero conjunto de soluciones aleatorias. Se trata de trabajar soluciones que describan de la mejor manera posible el espacio de soluciones. Se ha seguido un método para generar biclusters lo más dispersos posibles entre sí, es decir, lo más lejanos según el concepto de distancia utilizado. Debido a la codificación binaria de soluciones, se emplea la distancia Hamming que mide el grado de similitud entre cadenas de 0s y 1s. Dicho método se basa en la generación de cadenas de ceros y unos lo más diversas, o diferentes entre si, descrito en [61]. Este método se considerada clásico en los esquemas basados en búsquedas dispersas que utilizan codificaciones binarias. Se generan por separado la subcadena para los genes y la subcadena para las condiciones.

Se toma como semilla una cadena binaria x de ceros y unos en la que x_i, con $i = 1, \ldots, n$ y n el número de bits, representa el valor o bit que hay en cada posición. Se genera una nueva cadena binaria x' de tal manera que cada nuevo bit es generado mediante la regla:

$$x'_{1+kh} = 1 - x_{1+kh} \text{ for } k = 0, 1, 2, 3, \ldots, \lfloor n/h \rfloor \tag{6.1}$$

siendo $\lfloor n/h \rfloor$ el entero más grande mayor o igual que n/h y h es el entero menor que $n/5$, los valores que se salen del rango simplemente no provocan ningún nuevo bit a tener en cuenta. El resto de bits de la cadena de x'

permanecen iguales al valor que tuvieran en x. Después de haberse generado todas las posibles soluciones, si hacen falta más soluciones basta repetir el proceso con una nueva semilla. Se suele tomar como nueva semilla la última cadena generada con anterioridad.

Es importante tener en cuenta que la codificación elegida para el problema lleva asociada implícitamente la definición de una métrica. Esta métrica o distancia definirá la topología del espacio de soluciones o de búsqueda. Así el significado de soluciones dispersas o alejadas entre sí está ligado a esta distancia.

Como hemos comentado anteriormente, se ha considerado como distancia la *distancia Hamming* que se define como el número de cambios necesarios para transformar una cadena de bits en otra. Es decir, dadas dos cadenas binarias, la distancia Hammning entre ambas indica lo parecidas o distintas que son entre sí. Por ejemplo, la distancia Hamming entre 10$\underline{1}\underline{1}$01 y 1001001 es 2 pues con tan sólo dos cambios de bits se puede generar una teniendo en cuenta la otra.

6.5. Construcción y reconstrucción del conjunto de referencia

El conjunto de referencia lo constituye un pequeño grupo de soluciones cuya evolución será la guía del proceso en la búsqueda del óptimo. Estas soluciones tienen que ser lo mejor posible tanto desde punto de vista de la intensificación de la búsqueda como de la diversidad de la misma. Se entiende como intensificación en el proceso de búsqueda como la convergencia hacia el óptimo, de manera que el proceso se acelera cuando se acerca a la solución. Se entiende diversidad como la capacidad del proceso de explorar la mayor parte de regiones del espacio de búsqueda. Ambas deben mantener un equilibrio porque por un lado el proceso debe acercarse al óptimo con mayor rapidez pero sin dejar de explorar otras regiones del espacio en las que pudieran encontrase mejores soluciones. Un proceso evolutivo desequilibrado puede o bien no encontrar soluciones óptimas, porque es demasiado diverso, o bien caer en óptimos locales porque encuentra la mejor solución posible pero restringida a una región muy limitada del espacio de búsqueda. Cuando se construye o se reconstruye el conjunto de referencia se tienen en cuenta dos grupos de soluciones: las que aportan intensificación a la búsqueda y las que aportan diversidad.

La primera vez que se construye el conjunto de referencia, línea 5 del algoritmo **??**, se introduce un primer grupo de soluciones que son las mejores

de la población inicial, desde el punto de vista de su peso, y un segundo grupo, elegidas de la población inicial, que son lo más dispersas posible respecto a las primeras. Es decir, si el tamaño del conjunto de referencia es 10 (tamaño estándar en los esquemas clásicos de búsqueda dispersa) se eligen las 5 mejores de la población y aquellas 5 que sean lo más diferentes, dispersas o lejanas de la población, en el sentido de la distancia Hamming.

Las siguientes veces que se reconstruye el conjunto de referencia se realiza un proceso análogo, línea 15 del algoritmo 1. Se consideran para el nuevo conjunto de referencia las mejores soluciones del conjunto de referencia previo, si el tamaño es el estándar se eligen las 5 mejores soluciones de las 10 que hay, y por otro lado, se consideran las 5 soluciones más dispersas de la población inicial respecto de las previamente consideradas.

Es importante tener en cuenta que cuando se eligen las soluciones de la población inicial deben eliminarse de la misma. Caso contrario el proceso no funcionaría correctamente desde el punto de vista de la diversidad.

6.6. Método de generación de nuevas soluciones

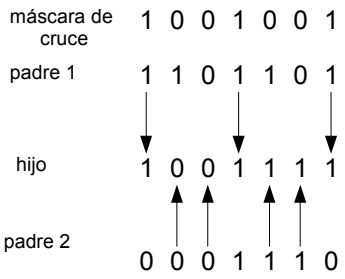

Figura 6.3: Operador de cruce uniforme.

A partir de las soluciones existentes en el conjunto de referencia, se pueden definir distintos mecanismos de generación de nuevas soluciones, o *métodos de generación de nuevas soluciones*, según cómo sea la codificación de las soluciones [61]. Debido a la codificación binaria de los biclusters, y a los estudios previos realizados, utilizamos el *operador de cruce uniforme* generalmente utilizado en los algoritmos genéticos. Se pueden utilizar otros operadores más sofisticados que aceleren el proceso, de hecho las principales mejoras en los esquemas de búsqueda dispersa tienen lugar en el refinamiento en la generación de nuevas soluciones.

Dadas dos soluciones, se genera una nueva solución mediante el operador de cruce uniforme. Se generan todas las posibles parejas teniendo en cuenta las soluciones existentes en el conjunto de referencia. Por lo tanto, se generan $S * (S-1)/2$ nuevas soluciones siendo S el tamaño del conjunto de referencia. El operador de cruce uniforme genera aleatoriamente una máscara y la solución generada se compone de los valores de una de las soluciones originales cuando hay un 1 en la máscara y de la otra solución cuando hay un 0, véase la figura 6.3.

6.7. Método de combinación y actualización del conjunto de referencia

Se combinan todas las soluciones generando de esta maneras nuevas soluciones que son mejoradas mediante el método de la mejora. Entre las soluciones existentes en el conjunto de referencia y todas las generadas se eligen las mejores soluciones según el valor de la función objetivo. De esta manera se actualiza el conjunto de referencia y se intensifica la búsqueda. Si el conjunto de referencia generado difiere del anterior, es decir, el método de la combinación ha sido capaz de generar mejores soluciones, entonces el proceso continúa, caso contrario se alcanza la estabilidad del conjunto de referencia y continúa una nueva iteración de la búsqueda dispersa.

6.8. Método de la mejora

El método de la mejora consiste en una búsqueda local que dada una solución genera una nueva solución con un mejor valor de la función objetivo. De esta manera se acelera la convergencia del proceso, ya que se trabajan con las mejores soluciones posibles del espacio durante el proceso de búsqueda.

La búsqueda local puede depender o no de la semántica de la función objetivo que se utilice en concreto. Si se liga a la función objetivo, se puede asegurar un funcionamiento rápido y efectivo en todos los casos. Se utiliza el conocimiento concreto del problema para alcanzar un óptimo local cercano a la función con la que se trabaje. Otra opción consiste en la definición de una búsqueda local "ciega" mediante una serie de permutaciones y, de esta forma, se desliga la función objetivo del problema en cuestión del proceso completo de búsqueda. En cualquier caso es necesario el método de la mejora como acelerador de la convergencia y garantía del correcto funcionamiento del proceso completo.

Capítulo 7

Criterio de búsqueda basado en correlación

Recuerdo una noche en que regresé tarde de Londres. En aquellos tiempos, como medida de economía, apagaban a medianoche el alumbrado urbano. Contemple el firmamento, atravesado por la Vía Láctea, como nunca lo había visto. No habrá faroles en mi isla desierta, así que veré bien las estrellas.
Agujeros negros y pequeños universos y otros ensayos. Stephen Hawking (pág. 113)

7.1. Introducción

En este capítulo se presentan dos propuestas de un algoritmo de biclustering basado en un esquema de búsqueda dispersa y cuya función objetivo está basada en el uso de correlaciones lineales. Se utiliza una medida de correlación lineal como medida de calidad para evaluar genes co-expresados y se integra como parte de la función de peso de la metaheurística. Ambas propuestas usan el esquema de búsqueda dispersa expuesto en el capítulo anterior y difieren, básicamente, en la capacidad de encontrar o no patrones de activación-inhibición. Dichos patrones se corresponden con grupos de genes correlacionados negativamente, es decir, genes que presentan una misma tendencia de crecimiento-decrecimiento.

7.2. Propuesta SSCorr: correlaciones lineales I

Veamos cómo se define la función objetivo y el método de la mejora de la primera propuesta de biclustering basado en correlaciones lineales usando

un esquema de búsqueda dispersa.

7.2.1. Evaluación de biclusters: función objetivo

Dos genes muestran entre sí un patrón de desplazamiento y escalado si verifican la siguiente expresión:

$$g_Y = \alpha g_X + \beta \quad \alpha, \beta \in \mathbb{R} \tag{7.1}$$

Es decir, ambos genes dependen linealmente el uno del otro. Se puede elegir su coeficiente de correlación lineal para medir su grado de interdependencia. El coeficiente de correlación entre dos variables X y Y se define como:

$$\rho(X,Y) = \frac{cov(X,Y)}{\sigma_X \sigma_Y} = \frac{\sum_i^n (x_i - \overline{x})(y_i - \overline{y})}{n\sigma_X \sigma_Y} \tag{7.2}$$

donde $cov(X,Y)$ es la covarianza entre las variables X e Y, $\overline{x}$ y $\overline{y}$ son la media de los valores de dichas variables y σ_X y σ_Y sus desviaciones estándar.

Dado un bicluster B compuesto por N genes, $B = [g_1, \ldots, g_N]$, la correlación media de B, $\rho(B)$, se define de la siguiente manera:

$$\rho(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \rho(g_i, g_j) \tag{7.3}$$

donde $\rho(g_i, g_j)$ es el coeficiente de correlación entre el gen i y el gen j. Sólo se deben tener en cuenta $\binom{N}{2}$ elementos dada la simetría del coeficiente de correlación, es decir $\rho(g_i, g_j) = \rho(g_j, g_i)$.

La figura 7.1 muestra dos biclusters, uno con genes con un régimen distinto de expresión entre ellos (figura a la izquierda), es decir no siguen un mismo patrón de comportamiento, y otro en el que se puede observar que siguen un patrón perfecto de desplazamiento y escalado (figura a la derecha). El biclúster de la derecha, donde se puede observar que todos los genes crecen y decrecen de una misma forma, patrón de desplazamiento, aunque con distinta intensidad, patrón de escalado, tiene una correlación media igual a 1. En el caso de la figura a la izquierda, donde no se observa ningún patrón de comportamiento común, la correlación media es casi cero, en concreto 0,003. Se consideran por tanto biclusters de calidad a aquellos que tengan una correlación media cercana a uno y caso contrario se considerarán biclusters de poca calidad. Como consecuencia de esta idea, la función objetivo que se emplea en el proceso de optimización es la siguiente:

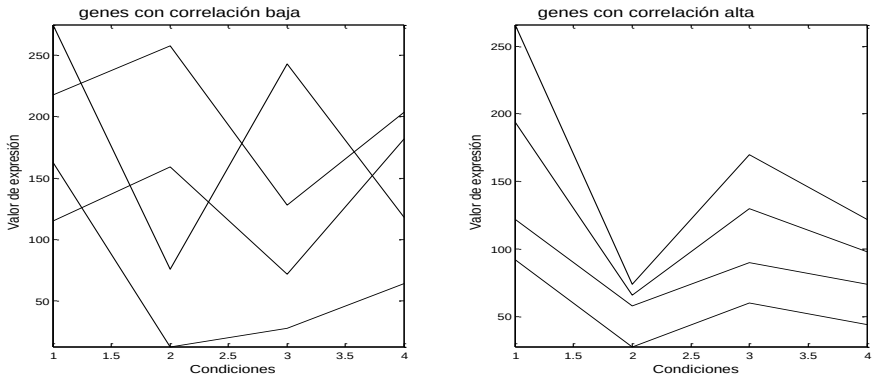

Figura 7.1: A la izquierda un bicluster formado por genes que no siguen un mismo patrón de expresión. A derecha otro bicluster con patrones de desplazamiento y escalado.

$$f(B) = (1 - \rho(B)) + \sigma_\rho + M_1 \left(\frac{1}{nG}\right) + M_2 \left(\frac{1}{nC}\right) \tag{7.4}$$

donde nG y nC son los números de genes y condiciones del biclúster B respectivamente, M_1 y M_2 son parámetros de control del volumen del biclúster y σ_ρ es la desviación estándar de los valores $\rho(g_i, g_j)$ de (7.3). La desviación estándar se incluye en la función objetivo para evitar casos en los que el biclúster tenga un valor alto para la correlación media pero a la vez esté formado por pares de genes poco correlacionados entre sí. Es decir, para controlar que los biclusters estén formados por patrones de genes lo más homogéneos posibles. Por otro lado, se debe tener en cuenta que se ha considerado $(1 - \rho(B))$ puesto que estamos resolviendo un problema de minimización. Los mejores biclusters serán, por lo tanto, aquellos que tengan un valor para la función objetivo lo más cercano a cero posible.

La correlación que se ha utilizado es la *correlación de Pearson*. Se debe tener en cuenta que la optimización es resistente a posibles ruidos (*outliers*) en los datos debido a que se inserta en un proceso de optimización donde se evalúan multitud de posibles soluciones. Se pueden usar otras definiciones de la correlación pero implica un mayor coste computacional para cada evaluación de la función objetivo.

7.2.2. Método de la mejora

Como se ha comentado en el capítulo anterior la búsqueda dispersa incorpora un mecanismo de mejora de las soluciones generadas de manera que se acelere el proceso de búsqueda. Se mejora el valor de su función objetivo,

mediante una búsqueda local en el entorno de cada solución, y se intercambia la solución existente por la nueva y de esta manera se introduce más calidad en la optimización.

En general se suelen usar como búsquedas locales "mecanismos ciegos", basados en permutaciones de las cadenas que modifican las soluciones, de manera que sean procesos independientes del significado de la función objetivo. Sin embargo, si se puede definir un algoritmo que dependa de la semántica de la función objetivo y optimice localmente cada solución, se suele llevar a cabo porque el proceso acelera su convergencia de manera considerable.

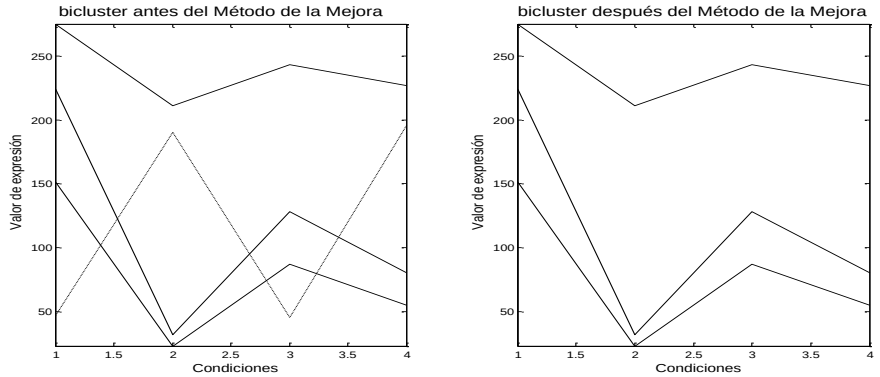

Figura 7.2: Un biclusters antes (izquierda) y después (derecha) de ser sometido al método de la mejora.

La función objetivo del caso que nos ocupa (ecuación 7.4) busca genes correlacionados positivamente que, como hemos comentado, reflejan comportamientos de desplazamiento y escalado. Se puede definir como mecanismo de mejora para cada biclúster un mecanismo que elimine aquellos genes que no estén correlacionados positivamente. La figura 7.2 muestra un biclúster compuesto por cuatro genes: tres altamente correlacionados positivamente y otro que tiene correlación negativa con los tres anteriores. La correlación media de dicho biclúster es de 0,0083 pero tras la mejora efectuada, que consiste en eliminar el gen correlacionado negativamente, obtenemos un biclúster con correlación media igual a 1. Aunque el volumen haya disminuido la correlación media ha mejorado en este caso y la función objetivo, en su conjunto, mejora su valor. Se debe tener en cuenta que el mecanismo optimiza tan sólo uno de los términos de la función objetivo, la correlación media, empeorando el término correspondiente al número de genes del bicluster. Si un biclúster se somete al método de la mejora pero no mejora el valor de la función objetivo, aunque haya mejorado su correlación media,

simplemente no se intercambiará por el nuevo biclúster construido.

A continuación se presenta el pseudocódigo del algoritmo del método de la mejora expuesto anteriormente.

Algorithm 2 Método de la mejora para SSCorr

INPUT Bicluster $B = [g_1, \ldots, g_N]$
OUTPUT Bicluster $B' \subseteq B$ tal que $\rho(g_i, g_j) \geq 0 \quad \forall g_i, g_j \in B'$

```
begin
  i ← 1, B' ← {g_i}, R ← {}
  while (i < N) do
    j ← i + 1
    while (j ≤ N) do
      if (ρ(g_i, g_j) > 0) then
        if (g_j ∉ R) then
          B' ← B' ∪ {g_j}
        end if
      else
        R ← R ∪ {g_j}
      end if
      j ← j + 1
    end while
    i ← i + 1
  end while
end
```

7.3. Propuesta BISS: correlaciones lineales II

Veamos cómo se define la función objetivo y el método de la mejora de la segunda propuesta de biclustering basado en correlaciones lineales usando un esquema de búsqueda dispersa.

7.3.1. Evaluación de biclusters: función objetivo

Los patrones de desplazamiento y escalado reflejan la mayoría de los procesos biológicos en los que la activación de un gen implica la activación adicional de otros genes. Hemos visto que se puede utilizar la correlación lineal como forma de medir esta co-expresión múltiple de genes en los datos de expresión. Sin embargo en la primera propuesta realizada en el apartado anterior no se contemplan los patrones de activación-inhibición. Estos patrones consisten en que la desactivación de un gen marca la señal para la activación de otros e implica que estos genes tengan un régimen de expresión complementario. Cuando uno se expresa el otro se inhibe y viceversa. Este

tipo de patrones son muy relevantes en diversos procesos biológicos y se le está prestando una especial atención recientemente [94].

Se modifica la ecuación de la correlación media definida en la ecuación 7.3 de manera que se capturen genes que estén correlacionados tanto positiva como negativamente. Para ello se define:

$$\rho_{|\cdot|}(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\rho(g_i, g_j)| \tag{7.5}$$

donde $\rho(g_i, g_j)$ es el coeficiente de correlación lineal entre el gen i y el j. La principal diferencia con la definición de la ecuación 7.3 consiste en el valor absoluto de manera que no se pierdan aquellos genes con correlación negativa. En la propuesta anterior un par de genes con correlación negativa alta podían eliminarse con un par de genes con correlación positiva alta. Por ejemplo, un par de genes con correlación, por ejemplo, igual a 0,9 podrían compensar su valor con otro par de genes con correlación igual a $-0{,}9$. De esta forma el efecto de pares de genes con correlación negativa no sólo se estaría perdiendo sino que estaría disminuyendo la calidad de los biclusters cuando lo que tenía que hacer sería aumentarla. La ecuación 7.5 captura correlaciones negativas entre genes sin embargo la ecuación 7.3 no lo hace.

En consecuencia, la función objetivo se define de la siguiente manera:

$$f(B) = (1 - \rho_{|\cdot|}(B)) + \sigma_\rho(B) + M\left(\frac{1}{vol(B)}\right) \tag{7.6}$$

siendo la principal diferencia con la función objetivo del apartado anterior la simplificación del parámetro asociado al volumen, además de por supuesto el cambio esencial del cálculo de la correlación media considerando el valor absoluto.

7.3.2. Método de la mejora.

El método de la mejora definido en esta segunda propuesta difiere con el de la sección anterior ya que no consiste en eliminar genes con correlación negativa. La idea consiste en definir un umbral de correlación y todo aquel gen que tenga una correlación por debajo de dicho umbral se elimina. Dado dicho umbral, si un gen tiene un valor de la correlación por debajo de dicho umbral con al menos un gen del resto del biclúster, entonces se elimina del mismo.

La figura 7.3 muestra un ejemplo de un biclúster compuesto por cuatro genes que tienen una correlación media de 0,70 usando la ecuación 7.5. Se puede observar que los genes 2 y 3 presentan un patrón de escalado y que

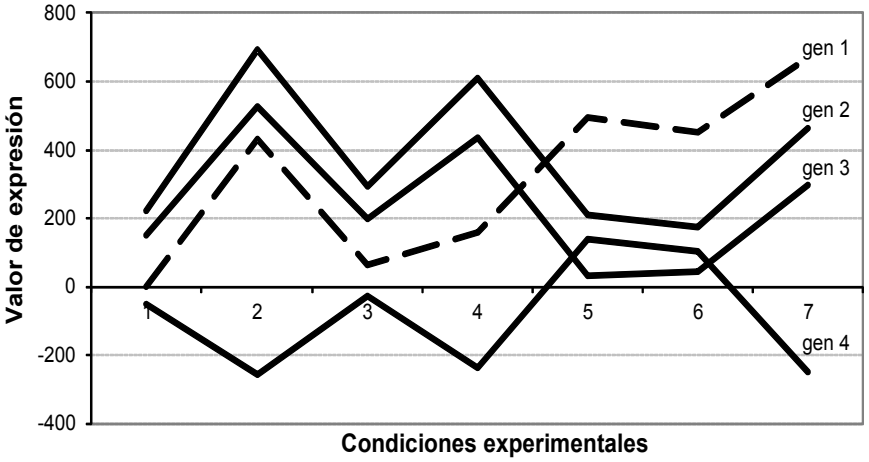

Figura 7.3: Un biclúster antes (línea discontinua) y después (línea resaltada) de haber aplicado el método de la mejora.

el gen 4 presenta un patrón de activación-inhibición respecto a estos dos genes. El gen 1 no presenta ningún patrón de comportamiento coherente con el resto de genes. La tabla 7.1 muestra el valor de la correlación entre estos cuatro genes. Si se establece como umbral de correlación 0,5, el gen 1 se elimina ya que su correlación con los genes 2, 3 o 4 es menor que 0,5. Una vez que se elimina el gen 1 del biclúster, el nuevo biclúster compuesto por los genes 2, 3 y 4 tiene un valor para la correlación media dada por la ecuación 7.5 de 0,98.

	g1	g2	g3	g4
g1	1	0.14	-0.01	-0.05
g2		1	0.98	-0.89
g3			1	-0.93
g4				1

Tabla 7.1: Correlación entre los genes 1, 2, 3 y 4.

En el Algoritmo 3 se presenta el pseudocódigo del método de la mejora expuesto.

Algorithm 3 MÉTODO DE LA MEJORA PARA BISS

INPUT Bicluster $B = [g_1, \ldots, g_N]$, umbral de correlación mínima ρ
OUTPUT Bicluster $B' \subseteq B$ tal que $|\rho(g_i, g_j)| \geq \rho \quad \forall g_i, g_j \in B'$

```
begin
 1: i ← 1, B' ← {g_i}, R ← {}
 2: while (i < N) do
 3:    j ← i + 1
 4:    while (j ≤ N) do
 5:       if (|ρ(g_i, g_j)| ≥ ρ) then
 6:          if (g_j ∉ R) then
 7:             B' ← B' ∪ {g_j}
 8:          end if
 9:       else
10:          R ← R ∪ {g_j}
11:       end if
12:       j ← j + 1
13:    end while
14:    i ← i + 1
15: end while
end
```

Procedimiento de elección del umbral de correlación.

El método de la mejora anteriormente propuesto introduce un nuevo parámetro para usarse como umbral de correlación. En el algoritmo de biclustering propuesto dicho parámetro se ajusta automáticamente.

Se define un procedimiento que calcula para cada conjunto de datos el mejor valor del umbral de correlación. Dicho procedimiento consiste en un muestreo experimental automatizado para el conjunto de datos con el que se trabaje. Concretamente, se realiza un estudio para distintos valores del parámetro, desde 0,1 hasta 0,9, y se elige aquel parámetro que obtenga los mejores resultados. Para cada uno de ellos se generan 100 biclusters de manera aleatoria, se le aplica el método de la mejora y se ve qué número de biclusters efectivamente se mejora y genera un nuevo biclúster con un valor más pequeño para su función objetivo. Aquel parámetro que obtenga un mayor número de biclusters mejorados es el que se elige como umbral de correlación para el método de la mejora. Este procedimiento tiene lugar al principio del proceso de búsqueda dispersa, antes de generar la población inicial.

Capítulo 8

Criterio de búsqueda basado en la integración de información biológica

La ortodoxia se ha resistido con uñas y dientes -en gran medida sigue resistiéndose- a aceptar la teoría de Margulis por el sencillo hecho de que no encaja con sus prejuicios darwinistas.
Deconstruyendo a Darwin. Javier Sampedro (pág. 42)

8.1. Introducción

En este capítulo se presenta un algoritmo de biclustering que integra información de tipo biológico durante el proceso de búsqueda. De esta manera se utiliza el conocimiento almacenado en repositorios públicos para introducir un sesgo que ayude al algoritmo a encontrar los mejores biclusters desde un punto de vista biológico. La integración de información es una de las tendencias actuales en Bioinformática [7] y ha sido investigada en distintos campos como el clústering tradicional [87], la selección de características [57], etc, sin embargo, desde nuestro conocimiento, aún no ha sido estudiada en biclustering.

8.2. Propuesta GoldBinch: integración de información biológica

El algoritmo propuesto se basa en un esquema de búsqueda dispersa que permite diferenciar claramente, como en los capítulos previos, el proceso de

búsqueda y el criterio que se usa. La función objetivo integra un término que permite ponderar, en base a la información almacenada en repositorios públicos como GO o KEGG, la calidad del grupo de genes analizados. Esta información condiciona el valor de la función objetivo y por lo tanto, de esta manera, se introduce el sesgo deseado en el proceso de búsqueda. La función objetivo admite distintas configuraciones según se defina el término de integración, por lo que se pueden estudiar y comparar distintas propuestas. Es importante destacar que con el objetivo de poder analizar las distintas posibilidades de integración el método de la mejora del esquema de búsqueda dispersa se diseña totalmente independiente de la semántica de la función objetivo.

8.2.1. Datos de entrada

Los datos de entrada del algoritmo son básicamente la matriz de expresión de los genes y un fichero de anotaciones directas para los genes. La matriz de expresión se compone de los patrones y perfiles de expresión de genes, que se corresponden respectivamente con las filas y columnas de la matriz. Cada elemento representa el valor de expresión de un gen para una determinada condición experimental. Los ficheros de anotaciones directas son ficheros planos donde cada línea asocia a un gen un conjunto de términos del repositorio biológico. Estos términos se asocian con una determinada funcionalidad asociada a dicho gen. Así por ejemplo, la expresión *TVP15 GO:0006810,GO:0016192* es una línea de un fichero de anotaciones extraído de GO, donde el gen *TVP15* se asocia con los términos *GO:0006810* y *GO:0016192*, términos GO para los que este gen está anotado. Se debe tener en cuenta que los ficheros de anotaciones directa se pueden construir o generar de varias formas. En cualquier caso el algoritmo admite como entrada cualquier tipo de fichero de anotaciones directa. De manera adicional el número de biclúster que se establece como un parámetro de entrada.

8.2.2. Función Objetivo

La función objetivo se compone de tres términos, un primer término que controla cómo tiene lugar la integración de información biológica y otros dos términos para la búsqueda de genes correlacionados y para maximizar el volumen.

La función objetivo propuesta para evaluar el biclúster B se define como:

$$f(B) = M_1 \cdot f_1(B) + M_2 \cdot f_2(B) + M_3 \cdot f_3(B) \tag{8.1}$$

donde f_1 mide el volumen del biclúster, f_2 los patrones que se encuentran y f_3 la calidad del mismo desde un punto de vista biológico, M_1, M_2 y M_3 son parámetros para controlar los pesos de los términos f_1, f_2 y f_3, respectivamente.

La medida f_1 se usa para controlar el volumen del biclúster y se define como:

$$f_1(B) = \frac{1}{nG \cdot nC} \tag{8.2}$$

donde nG es el número de genes y nC el número de condiciones del biclúster B. Este término es importante para controlar el tamaño de los biclusters durante el proceso de búsqueda y evitar información no relevante [70]. Se usa para evitar biclusters triviales con un número muy bajo de genes o condiciones, como por ejemplo sólo 2 condiciones.

La medida f_2 se basa en la media de la correlación entre los genes del biclúster. Téngase en cuenta que se calcula la correlación teniendo en cuenta sólo las condiciones del biclúster y no todas las del perfil de expresión. Este término se considera para capturar la mayoría de los patrones relevantes en biclusters tales como los de desplazamiento, de escalado [1] o aquellos que denotan patrones de activación-inhibición [14]. Se define como:

$$\rho_{|\cdot|}(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} |\rho(g_i, g_j)| \tag{8.3}$$

donde $\rho(g_i, g_j)$ es la correlación de pearson entre los genes g_i y g_j.

Nótese que sólo $\binom{N}{2}$ elementos se consideran debido a la simetría del término de la correlación. El valor absoluto se tiene en cuenta para evitar grupos de genes con una alta correlación positiva pudieran eliminar el efecto de grupos con una alta correlación negativa. Es importante tener en cuenta que en este término es donde intervienen las condiciones del biclúster, es decir, dos biclusters con los mismos genes pero con valores distintos para las condiciones tendrán valores distintos para la correlación.

Dado que la función objetivo se optimiza buscando su valor mínimo, el término f_2 se define como:

$$f_2(B) = 1 - \rho_{|\cdot|}(B) \tag{8.4}$$

La medida f_3 maneja el fichero de anotaciones directas generado a través de un repositorio con información biológica. De esta manera, se determina cómo se integra la información biológica. Téngase en cuenta que el término $f_3(B)$ evalúa el conjunto de genes del biclúster B pero no sus condiciones.

Es decir, el valor de f_3 es el mismo para dos biclusters que tengan el mismo conjunto de genes pero distintas condiciones. Este término mide la relevancia biológica de un biclúster, respecto de la información codificada en el fichero de anotaciones, pero no cómo son los patrones que lo describen, para lo que es necesario el término basado en la correlación.

A continuación se presentan dos funciones (secciones 8.2.2 y 8.2.2), que usan el fichero de anotaciones como base para calcular la relevancia biológica de los biclusters, y que se utilizan para definir el término f_3.

Medida FracGO

A continuación se define una medida basada en el enriquecimiento de genes [13] como mecanismo para medir la relevancia biológica de los biclusters. La idea se basa en medir la proporción de genes de un biclúster que están asociados a términos GO enriquecidos, denotaremos a esta medida como *FracGO*. El proceso de cálculo tiene lugar de la siguiente manera. Se consideran términos GO enriquecidos a aquellos que estén sobre-representados con p-value ajustado por debajo de un determinado umbral, conocido como umbral de significancia y que usualmente se suele elegir como 0,05. Para ellos se calcula el p-value ajustado para cada término GO en el fichero de anotaciones respecto al grupo de genes del biclúster. El universo de genes es el conjunto completo de genes presentes en la matriz de expresión. Se utiliza el test de Fisher para determinar si los términos estás estadísticamente sobre-representados y el test de Bonferroni para el cálculo del ajuste del p-value con tantas comparaciones como hipótesis se testean, que se corresponden con el número de términos presentes en el fichero de anotaciones.

De esta forma se define *FracGO* como:

$$FracGO(B) = \begin{cases} 0 & if J = 0 \\ \frac{1}{J \cdot N} \sum_{i=1}^{J} x_i & if J \geq 1 \end{cases} \tag{8.5}$$

donde J es el número de términos GO enriquecidos, en concreto aquellos que tengan un p-value ajustado por debajo de 0,05, N el número de genes en el biclúster y x_i es el número de genes del biclúster que están presentes en el término GO i del fichero de anotaciones.

Nótese que *FracGO* es igual a 1 si todos los genes del biclúster están asociados con todos los términos GO enriquecidos ($x_i = N, \forall i = \{1, ..., J\}$) y 0 cuando no hay ningún término GO enriquecido asociado. Se puede concluir que *FracGO* vale 1 cuando los biclusters son biológicamente relevante y 0 en caso contrario.

En este caso, el término f_3 se define directamente como:

$$f_3(B) = 1 - FracGO(B) \tag{8.6}$$

El biclúster B se considera como un biclúster bueno si f_3 es igual a 0 y malo si su valor es igual a 1.

Medida SimNTO

En la literatura existen varias medidas basadas en la similitud según GO entre pares de genes. Estas medidas miden el grado de semejanza entre dos genes según la información presente de estos en GO. La medida simNTO se basa en el solapamiento de términos definido en [65] y tan sólo utiliza como entrada el fichero de anotaciones directas. Esta medida es más rápida y simple que otras medidas de similitud GO, basadas en el contenido de información, o *information content* (IC), que necesitan de manera adicional la estructura del grafo de GO. SimNTO captura la estructura del grafo de GO a través de la manera que se construye el fichero de anotaciones, en concreto si éste propaga las anotaciones de los términos hasta la raiz de la ontología de GO, es decir, teniendo en cuenta todos los términos padre de cada término.

Se propone en este apartado una medida para evaluar biclústers basada en simNTO. Esta medida se basa en la media de los valores de simNTO de los pares genes que componen el bicluster. Es decir,

$$SimNTO(B) = \frac{1}{\binom{N}{2}} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} simNTO(g_i, g_j) \tag{8.7}$$

donde N es el número de genes del biclúster B y $simNTO$ la medida del solapamiento normalizada propuesta en [65]. En particular, $simNTO$ mide el solapamiento entre dos genes g_1 y g_2 respecto de los términos GO de la siguiente manera:

$$simNTO(g_1, g_2) = \frac{|annot_{g_1} \cap annot_{g_2}|}{min(|annot_{g_1}|, |annot_{g_2}|)} \tag{8.8}$$

donde $annot_{g_i}$ es el conjunto de términos GO asociados al gen g_i y $|\cdot|$ el número de elementos del conjunto. Es importante comentar que el fichero de anotaciones debe estar construido propagando las anotaciones hasta los términos superiores de la jerarquía de GO. Por lo tanto, $annot_{g_i}$ se define considerando el conjunto de todas las anotaciones directas del gen g_i y las asociadas a sus términos padre en la ontología, excluyendo la raiz de la misma. Es decir, el fichero de anotaciones debe capturar la estructura jerárquica de GO. Nótese que la medida $simNTO$ toma valores entre 0 y 1. Dos genes

que compartan las mismas anotaciones, y por lo tanto son similares en la ontología, tienen un valor igual a 1, en caso contrario su valor será 0. Por otro lado, si alguno de los genes del biclúster no tiene información en el fichero de anotaciones, es decir $annot_g$ es el conjunto vacío, se define su valor de $simNTO$ directamente como cero.

La función f_3 se define como:

$$f_3(B) = 1 - SimNTO(B) \tag{8.9}$$

Se puede apreciar que $f_3(B) = 0$ cuando B esté formado por genes similares según la información de GO. En este caso se dice que B es un biclúster de calidad según la información de GO.

8.2.3. Descripción del algoritmo

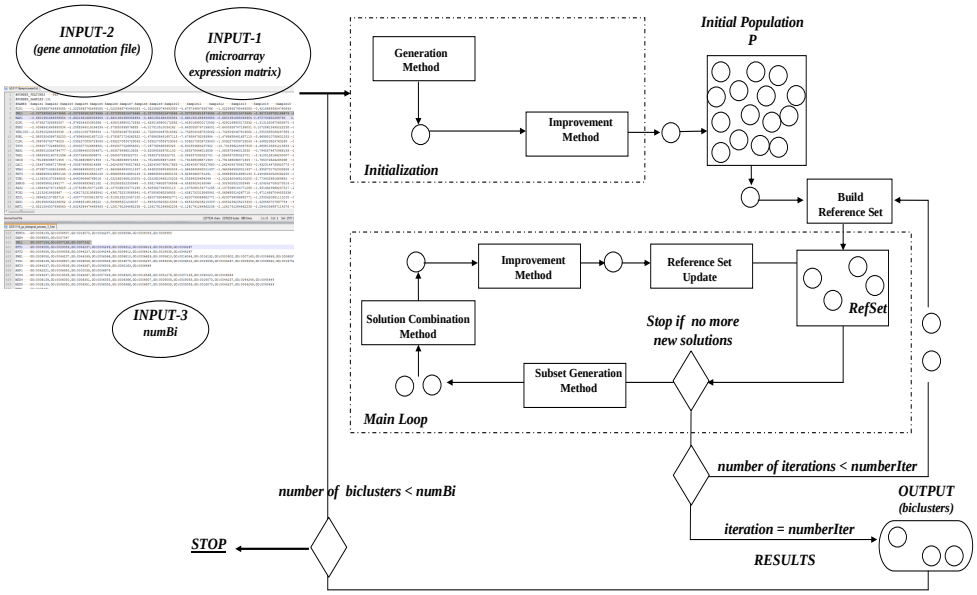

Figura 8.1: Esquema del algoritmo de biclustering propuesto basado en un esquema de búsqueda dispersa.

El algoritmo propuesto se basa en la adaptación de un esquema de búsqueda dispersa para la optimización de la función objetivo propuesta (ecuación 8.1). Debido a que el objetivo es el estudio de varias medidas de integración biológica dentro de la función objetivo, es importante definir un esquema de búsqueda que sea totalmente independiente de la función objetivo. Este hecho motiva un método de la mejora "ciego", en el sentido de ser

independiente de la función objetivo. Dicho método se basa en la generación de nuevas soluciones en un entorno de vecindad de la solución dada y su evaluación buscando una solución mejor.

La figura 8.1 muestra el esquema de funcionamiento del algoritmo. Se puede observar que la principal novedad en el esquema propuesto son los argumentos de entrada, constituidos por la matriz de expresión y por el fichero de anotaciones para los genes. Así mismo una de las principales diferencias con los esquemas de búsqueda dispersas en biclustering vistos con anterioridad lo constituye el método de la mejora.

8.2.4. Método de la mejora

El método de la mejora elegido es una búsqueda local "ciega" en el sentido que es independiente de la semántica de la función objetivo. En general este tipo de búsquedas locales son menos eficientes a la hora de mejorar localmente soluciones, en comparación con aquellas definidas "ad-hoc" o ligadas a la función objetivo que se emplee, como por ejemplo en el capítulo anterior, pero permiten intercambiar distintas funciones y poder comparar el rendimiento de cada una. El objetivo es proporcionar una búsqueda local que permita acelerar la convergencia del proceso de búsqueda y, por otro lado, que sea válida para cualquier función objetivo que se utilice. De esta forma la función objetivo puede aceptar varias configuraciones, o medidas para integrar la información biológica, sin tener que realizar ningún cambio en el esquema de búsqueda.

El método de la mejora diseñado consiste en seleccionar aquel biclúster con mejor valor para la función objetivo entre los biclusters cercanos al dado como entrada. Dado un biclúster/solución se toma su codificación binaria y se generan un conjunto de soluciones cercanas, en el sentido de la distancia *Hamming*, de entre las cuales se selecciona la mejor. Si no hay ninguna mejor permanece como salida el mismo biclúster utilizado como entrada. Téngase en cuenta que la distancia *Hamming* mide el grado de similitud entre dos cadenas, por lo que pequeñas permutaciones en una cadena binaria dan otra cadena cercana según esta distancia. La figura 8.2 muestra un ejemplo de cómo se generan las soluciones cercanas a la dada. Cada bit de la cadena original es analizado, si su valor es 0 permanece igual pero si su valor es 1 se aplica una de los siguientes casos, de manera que se generen distintas soluciones según cada caso:

- Caso 1: El bit a la derecha cambia su valor a 1 y el bit actual permanece igual.

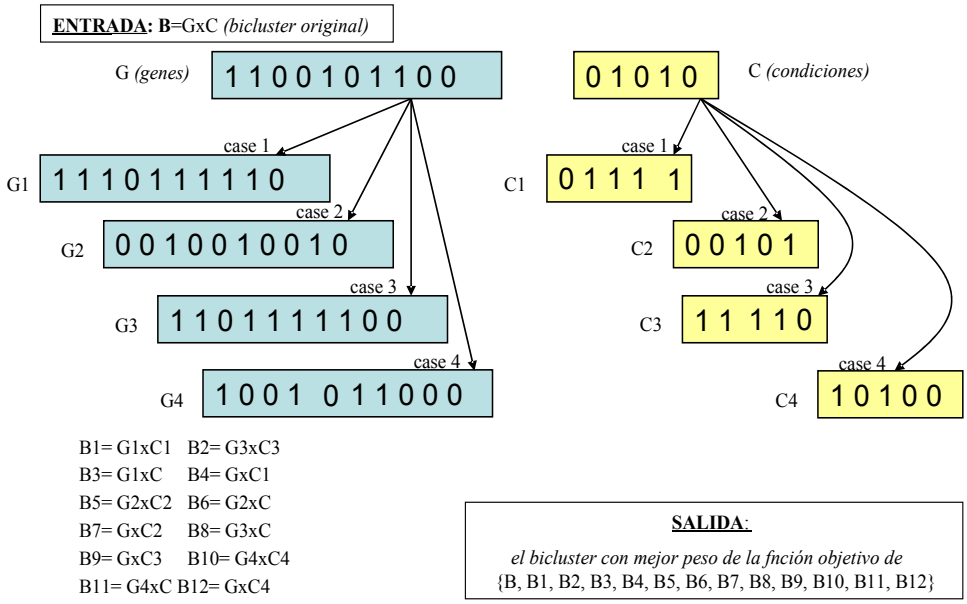

Figura 8.2: Un pequeño ejemplo que ilustra el método de la mejora.

- Caso 2: El bit a la derecha cambia su valor a 1 y el bit actual cambia a 0.
- Caso 3: El bit a izquierda cambia a 1 y el bit actual permanece igual.
- Caso 4: El bit a izquierda cambia a 1 y el bit actual cambia a 0.

Aplicando esta regla se generan doce nuevas soluciones aplicando el método, véase la figura 8.2. Si no se genera ningún nuevo biclúster con un valor mejor para su función objetivo, la salida del método de la mejora es el biclúster original.

Parte IV

Resultados

Capítulo 9

SScorr: resultados

Lo más importante es tener pasión. No se trata de lo que digan los demás que es mejor para ti, hay que escuchar lo que te dicta el corazón, los instintos. Si de verdad te apasiona algo, encontrarás la manera de hacerlo. Pero sin pasión, es inútil.
National Geographic junio 2013. Zona de riesgo, entrevista a Gerlinde Kaltenbrunner, alpinista austriaca

9.1. Introducción

En este capítulo presentamos los experimentos realizados con el algoritmo propuesto en la sección 7.2 del capítulo VII. El objetivo del estudio experimental que se presenta a continuación es mostrar la calidad de los biclusters obtenidos.

Se han elegido para los experimentos tres conjuntos de datos o *data sets*. El primero (Yeast) fue utilizado en el trabajo presentado en [25], está relacionado con el ciclo celular de la levadura, *Saccharomyces cerevisiae*, y tiene 2,884 genes y 17 condiciones, . El segundo conjunto de datos (Lymphoma) presenta un estudio en Homo sapiens sobre un tipo concreto de cáncer, linfomas o tumores sólidos hematológicos, tiene los datos de expresión de 4026 genes bajo 96 condiciones experimentales y fue generado en [3]. Estos dos conjuntos de datos fueron los utilizados en el artículo [24] y han sido frecuentemente utilizados en gran variedad de artículos de biclustering [29, 66, 75]. En este trabajo fue donde se realizó su preprocesamiento. El tercer conjunto de datos (GaschYeast) procede de un estudio de la levadura, *Saccharomyces cerevisiae*, realizado en [40]. Está compuesto por 2993 genes y 173 condiciones experimentales. Se ha usado la versión del mismo proporcionada por los datos suplementarios de [77].

Id bi.	Genes	Cond.	Volumen	$\rho(B)$	$\sigma(B)$	MSR	Varianza de los genes
biYeastN15	7	10	70	0.95	0.56	59.2	882.8
biYeastN21	11	9	99	0.92	0.47	205.2	1190.5
biYeastN24	9	9	81	0.92	0.45	142.9	1344.8
biYeastN40	13	8	104	0.89	0.45	368.2	2185.4
biYeast	22.27	6.46	133.1	0.90	0.48	321.0	1508.7
biLymphomaN1	14	14	196	0.92	0.43	3719.2	29180.0
biLymphomaN11	17	7	119	0.92	0.50	1607.9	10317.6
biLymphomaN15	21	10	210	0.86	0.43	1818.4	8351.2
biLymphomaN54	9	14	126	0.82	0.45	1292.6	6108.0
biLymphoma	10.81	11.53	123.7	0.85	0.45	2593.3	11643.07
bi1-GaschYeastN1	13	25	325	0.96	0.42	0.08	1.51
bi1-GaschYeastN10	12	22	264	0.95	0.48	0.06	1.19
bi1-GaschYeastN11	41	17	697	0.93	0.34	0.15	1.67
bi1-GaschYeastN25	19	10	190	0.93	0.43	0.19	0.89
bi1-GaschYeast	16.36	14.08	237.6	0.89	0.43	0.32	1.50
bi2-GaschYeastN1	54	39	2106	0.82	0.32	0.22	1.00
bi2-GaschYeastN4	43	32	1376	0.84	0.45	0.18	1.02
bi2-GaschYeastN9	48	24	1152	0.87	0.41	0.17	1.18
bi2-GaschYeastN27	33	28	924	0.84	0.39	0.13	0.72
bi2-GaschYeast	46.69	27.69	1269.4	0.72	0.34	0.38	1.02

Tabla 9.1: Información sobre los biclusters encontrados por el algoritmo SSCorr

Los parámetros internos del algoritmo propuesto se han elegido de la siguiente forma: el esquema de búsqueda dispersa tiene 20 como el número máximo de iteraciones, 10 el tamaño del conjunto de referencia, 200 para el número de soluciones de la población inicial y 100 para el número de biclusters que se obtienen en cada ejecución. M_1 y M_2 son parámetros de la función objetivo que guían la búsqueda para obtener biclusters de un tamaño determinado. Se utilizan valores altos para M_1 y M_2 si se quieren obtener biclusters de tamaño grande. Los resultados para el dataset de Yeast y de Lymphoma se han obtenido con valores de $M_1 = 1$ y $M_2 = 1$. Los resultados para GaschYeast se han obtenido para valores de $M_1 = 1$ y $M_2 = 1$ y para $M_1 = 10$ y $M_2 = 10$ para así mostrar la influencia de estos parámetros en el volumen de los biclusters.

9.2. Resultados

La tabla 9.1 muestra la información de cuatro biclusters seleccionados entre los obtenidos de la ejecución del algoritmo SSCorr, así como el valor medio para los 100 biclusters obtenidos (en negrita). Para cada biclúster se muestra su identificador, el número de genes y de condiciones, su volumen, el

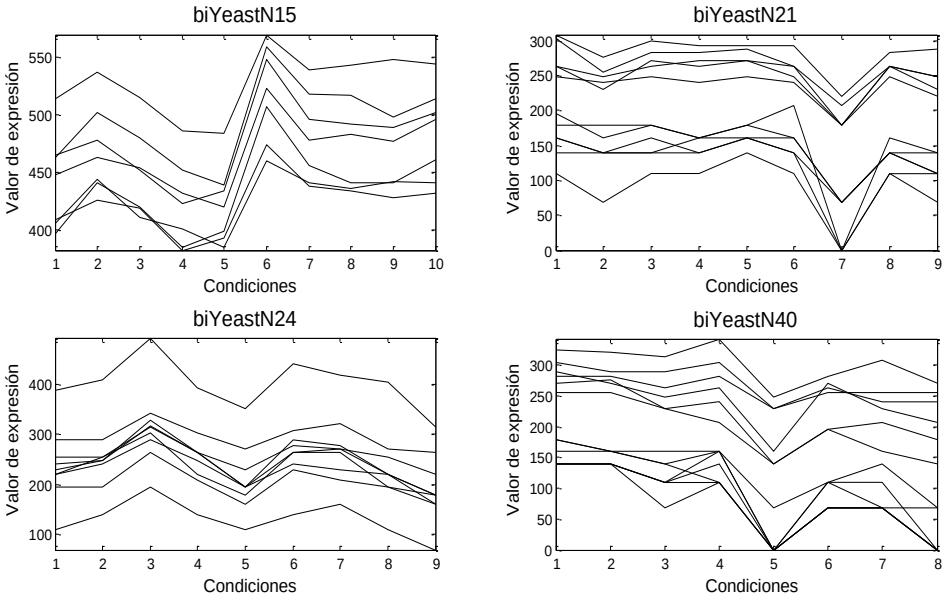

Figura 9.1: Resultados para el dataset Yeast

valor medio de la correlación $\rho(B)$ y la desviación estándar $\sigma(B)$, así como el valor del residuo cuadrático MSR como de la varianza del valor de los genes con el objetivo de poder establecer comparaciones con otros algoritmos. La varianza de los genes mide la variabilidad en el valor de expresión de cada gen.

Las figuras 9.1 y 9.2 muestran cuatro biclusters de los datos Yeast y Lymphoma cuya información se refleja en la tabla 9.1. Las figuras 9.3 y 9.4 muestran por otro lado biclusters de los datos GaschYeast. Los biclusters *bi1-GaschYeastN1*, *bi1-GaschYeastN10*, *bi1-GaschYeastN11* y *bi1-GaschYeastN25* de la figura 9.3 se han obtenido con valores $M_1 = 1$ y $M_2 = 1$, mientras que los biclusters *bi2-GaschYeastN1*, *bi2-GaschYeastN4*, *bi2-GaschYeastN9* y *bi2-GaschYeastN27* de la figura 9.4 lo han sido con los valores $M_1 = 10$ y $M_2 = 10$. Nótese que cuanto mayores son los valores de los parámetros el volumen de los biclusters. La elección de los parámetros $M_1 = M_2 = 1$ se debe a poder obtener biclusters con un número pequeño de genes y que de esta forma tengan una representación gráfica más clara para apreciar con más claridad los patrones que subyacen. Sin embargo, si el objetivo es encontrar biclusters compuestos por genes que compartan el mismo atributo GO, es más adecuado obtener un número más alto de genes. Por lo tanto, los parámetros $M_1 = M_2 = 10$ se han elegido con este objetivo, obtener biclusters compuestos por un número alto de genes.

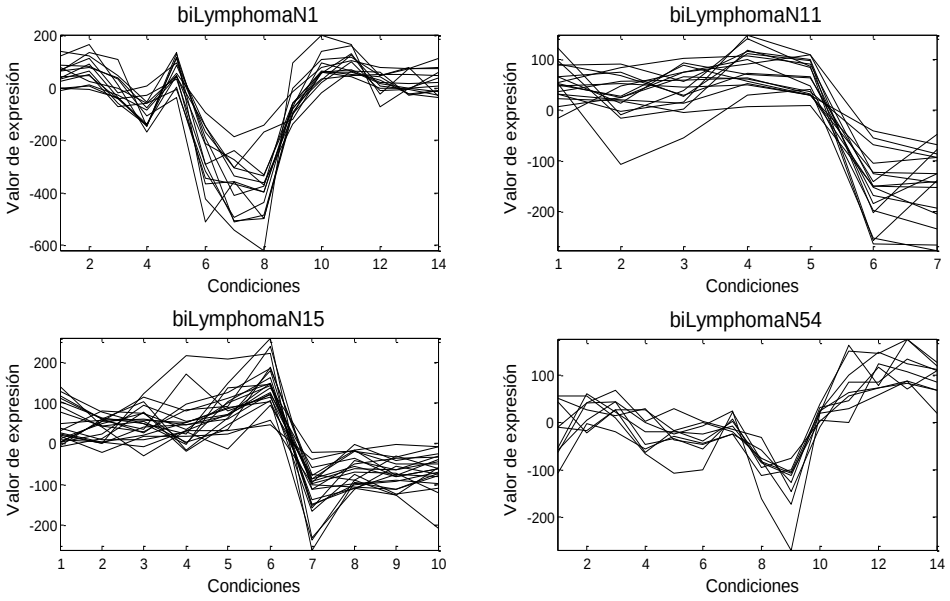

Figura 9.2: Resultados para el dataset Lymphoma

9.3. Discusión

Los cuatro biclusters presentados en la figura 9.1, obtenidos de los datos de Yeast, tienen un valor alto para la correlación media em concreto un valor de 0,90. Nótese que se pueden observar claramente patrones de desplazamiento y escalado en estos biclusters. La mayoría de los artículos que utilizan el residuo, MSR, como función de evaluación y trabajan con el dataset de Yeast, consideran que aquellos que tienen un valor de MSR por debajo de 300 son biclusters de calidad [24, 29]. Este valor como en natural depende del dataset con el que se trabaje. Los biclusters *biYeastN15*, *biYeastN21* y *biYeastN24* tienen un residuo por debajo de 300, sin embargo *biYeastN40* tiene un valor por encima de esta cota. Es interesante observar que en este caso el valor de la varianza de los genes es más alto en comparación con los otros tres biclusters.

Los cuatro biclusters de la figura 9.2, obtenidos de los datos de Lymphoma, muestran genes con un comportamiento similar y tienen un valor alto para la correlación en concreto un valor de 0,85. Sin embargo, estos biclusters no se pueden considerar buenos teniendo en cuenta otros trabajos que utilizan el mismo conjunto de datos, y que se basan en el uso del MSR como criterio de evaluación [24, 29], ya que tienen un MSR por encima de 1200. El bicluster con el valor más bajo de MSR es *biLymphomaN54* (1292.6), que sin embargo tiene el valor más bajo para la correlación (0.82). En el

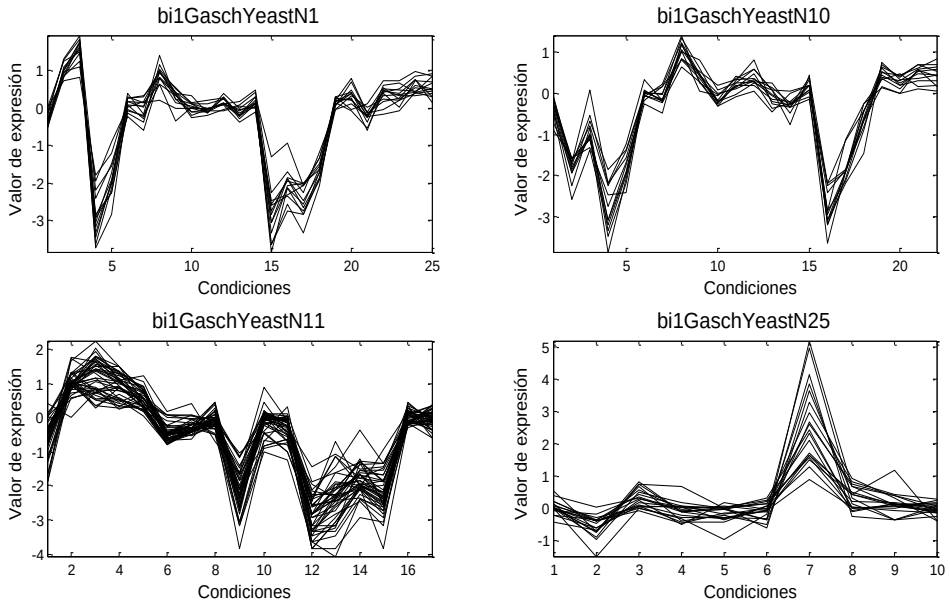

Figura 9.3: Resultados para el dataset GaschYeast ($M_1 = 1, M_2 = 1$)

trabajo presentado en [1] se probó que el residuo no es lo suficiente preciso como medida para descubrir biclusters con patrones de desplazamiento y escalado. Aquellos biclusters con un valor alto para la varianza de sus genes, es decir que presentan picos muy acentuados, no pueden ser detectados por funciones de evaluación basadas en el MSR. Los resultados de los biclusters de la figura 9.2 son un claro ejemplo de esta situación.

Las figuras 9.3 y 9.4 muestran biclusters de los datos de GashYeast. Se puede observar claramente en estas figuras los efectos de los parámetros M_1 y M_2. Cuanto más altos sean sus valores, mayor será el volumen de los biclusters. Todos los genes presentan un comportamiento similar, que es fácilmente observable en la figura. Así por ejemplo, los patrones de escalado se pueden observar claramente en el biclúster *bi1-GaschYeastN25*, donde la forma de los genes entre las condiciones 6 y 8 es la misma aunque cada gen incremente su valor de expresión con distinta intensidad. Por otro lado, el valor de la correlación media muestra que todos los biclusters de GaschYeast están compuestos por genes altamente correlacionados. Téngase en cuenta que los valores del MSR y de la varianza entre los genes están en un rango de valores de magnitud diferente de los otros dos conjuntos de datos estudiados. Esto se debe al procesamiento de este conjunto de datos.

Se debe tener en cuenta que todos los biclusters presentan patrones de desplazamiento y escalado y, por lo tanto, un valor alto para la correlación

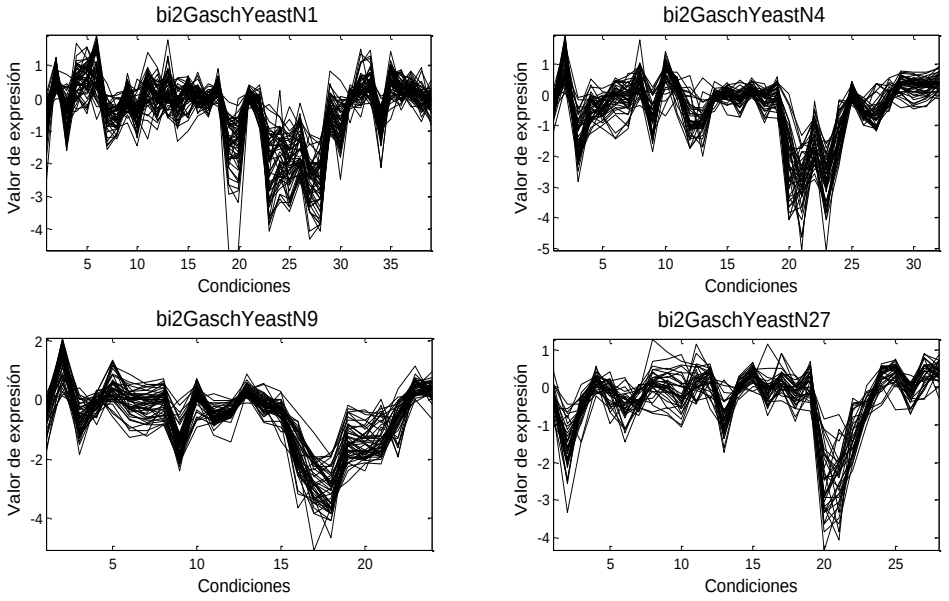

Figura 9.4: Resultados para el dataset GaschYeast ($M_1 = 10, M_2 = 10$)

media. Por otro lado, la desviación estándar es baja, es decir, el coeficiente de correlación de cada par de genes tiene un valor parecido y cercano al valor de la correlación media de los biclusters. Se puede asegurar que todos los biclusters están compuestos por parejas de genes con un valor alto de correlación entre si y que no contienen parejas de genes que no estén correlacionados.

9.4. Análisis comparativo

El rendimiento del algoritmo SSCorr se ha comparado con el de otros algoritmos de biclustering, como ChCh [24], OPSM [11], ISA [48], BiMax [77], xMotifs [68] y Samba[85] para los datos de GaschYeast y con CCC-Biclustering [59] para los de Yeast. Se han generado además un grupo de biclusters aleatorios con vistas a tener un caso trivial. La metodología que se ha seguido sigue de cerca a la propuesta en [77] donde el rendimiento de todos los algoritmos es evaluado respecto al porcentaje de biclusters enriquecidos usando GO como referencia. La ontología GO [40] se suele utilizar para estudiar si un grupo de genes pertenecientes a un biclúster está enriquecido respecto a un término específico de GO. Existen varias herramientas para llevar a cabo análisis de enriquecimiento con GO, en particular en este estudio se ha utilizado la herramienta AGO [2]. Esta herramienta se basa

en *GeneMerge* [23]. *GeneMerge* utiliza un contraste múltiple de hipótesis basado en la distribución hipergeométrica con las correcciones de Bonferroni. El enriquecimiento de un grupo de genes en un determinado término GO se establece respecto a un determinado p-value. Se dice que un biclúster está enriquecido o sobre-representado respecto a un término GO si el p-value obtenido para dicho término es menor que un valor pre-establecido. Así, el porcentaje de biclusters enriquecidos se utiliza como criterio de comparación entre algoritmos de biclustering.

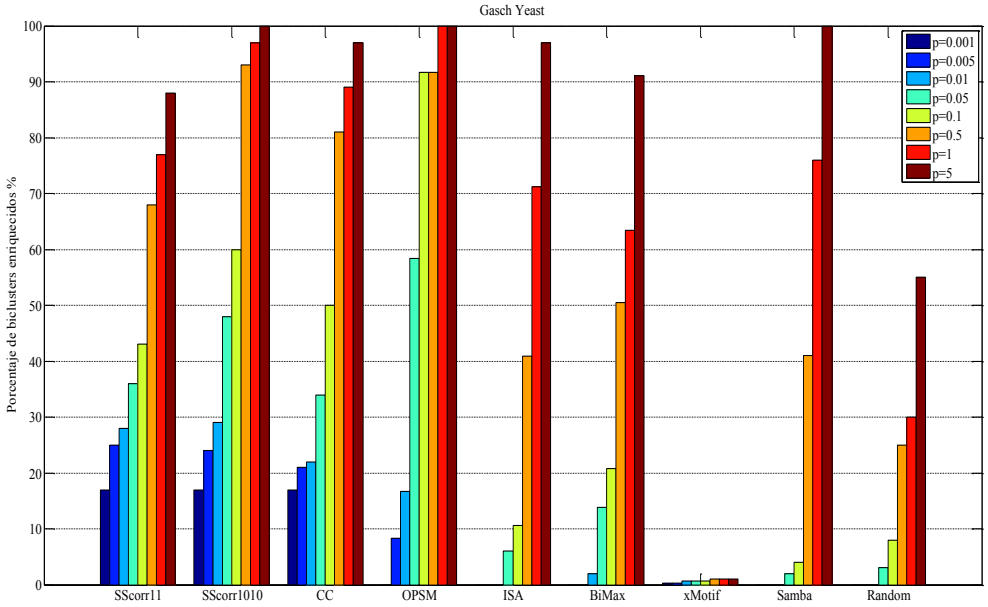

Figura 9.5: Comparación entre los distintos algoritmos de biclustering para el dataset GaschYeast

	num. biclusters	num. genes	num. condiciones	volumen
SScorr11	100	16.4(6.9)	14.8(3.1)	237.6
SScorr1010	100	46.7(8.2)	27.1(5.5)	1269.4
CC	100	82.0(130.1)	19.8(16.3)	2557.31
OPSM	12	95.6(119.6)	12.5(3.6)	849.8
ISA	66	76.3(43.9)	8.7(1.4)	645.7
BiMax	101	24.0(2.8)	3(0)	72.1
xMotifs	306	1.2(0.4)	42.3(11.4)	46.7
Samba	100	911.5(132.1)	25.1(8.2)	22344.7
Random	100	12.8(2.4)	25.0(2.1)	318.3

Tabla 9.2: Comparación entre los biclusters encontrados por distintos métodos de biclustering para los datos de GashYeast.

La figura 9.5 representa el porcentaje de biclusters enriquecidos para ca-

da algoritmo a distintos niveles de significancia (0,001, 0,005, 0,01, 0,05, 0,1, 0,5, 1 and 5para el conjunto de datos GaschYeast). En esta figura, SScorr11 significa una ejecución del algoritmo propuesto con los parámetros de penalización $M_1 = 1$ y $M_2 = 1$. Análogamente, SScorr1010 para los valores $M_1 = 10$ and $M_2 = 10$. Si nos centramos en un nivel de significancia del 0,01 la proporción de biclusters enriquecidos, usando la rama de BP para SScorr11 y SScorr1010, es aproximadamente del 21 % para ChCh, para OPSM del 17 %, para BiMax 2 % y 0 % para el resto. Se puede observar que SScorr1010 mejora los resultados del resto de los métodos para valores de significancia pequeños excepto para ChCh con 0,001 donde los resultados son prácticamente iguales. Por otro lado, ChCh tiene un porcentaje más alto que SScorr11 para valores de significancia de 0,1 a 5. Este hecho se debe a que más fácil el enriquecimiento funcional para biclusters compuestos por un número grandes de genes que aquellos compuestos por un número pequeño. Sin embargo, los niveles de significancia mayores a 0,05 son demasiado poco exigentes. La tabla 9.2 muestra la información sobre el tamaño de los biclusters obtenidos por los distintos métodos. Se puede observar que los biclusters obtenidos por ChCh tienen más genes que los obtenidos por SSCorr. En particular, SSCorr1010 encuentra biclusters con más genes que SScorr11 y además mejora los resultados obtenidos por ChCh para todos los niveles de significancia. El resto de los métodos encuentra un porcentaje menor de biclusters enriquecidos aunque OPSM también presenta resultados buenos cuando el nivel de significancia está por encima de 0,05.

En general establecer una comparación entre algoritmos de biclustering no es una tarea sencilla pues el tamaño de los biclusters o los patrones que se encuentran suelen diferir bastante entre cada algoritmo. La tabla 9.2 muestra la información de los biclusters encontrados por cada método, la media y la desviación estándar del número de genes y condiciones, así como el volumen. La desviación estándar muestra la variabilidad en el tamaño de los resultados. Por ejemplo, ChCh, OPSM y SAMBA encuentran biclusters muy desiguales en su número de genes como se puede observar en el valor alto para la desviación estándar. La figura 9.5 formula la definición de biclúster enriquecido con un criterio más restrictivo. El porcentaje de biclusters enriquecidos se calcula tras filtrar, y poner como restricción, que al menos la mitad de los genes que componen el biclúster deben participar en el mismo término de GO a la hora de calcular el enriquecimiento. Es decir, que la mitad de los genes de los biclusters considerados como buenos compartan la misma funcionalidad. La figura 9.6 muestra la comparación entre todos estos métodos con la nueva definición de enriquecimiento. Se puede apreciar que SScorr11 y SScorr1010 obtienen los mejores resultados

para valores bajos de significancia (de 0,001 a 0,01) aunque para los valores de 0,05 a 5 OPSM y Samba muestra los mejores resultados. Tanto ISA como Bimax no encuentran resultados con esta nueva definición más restrictiva de enriquecimiento.

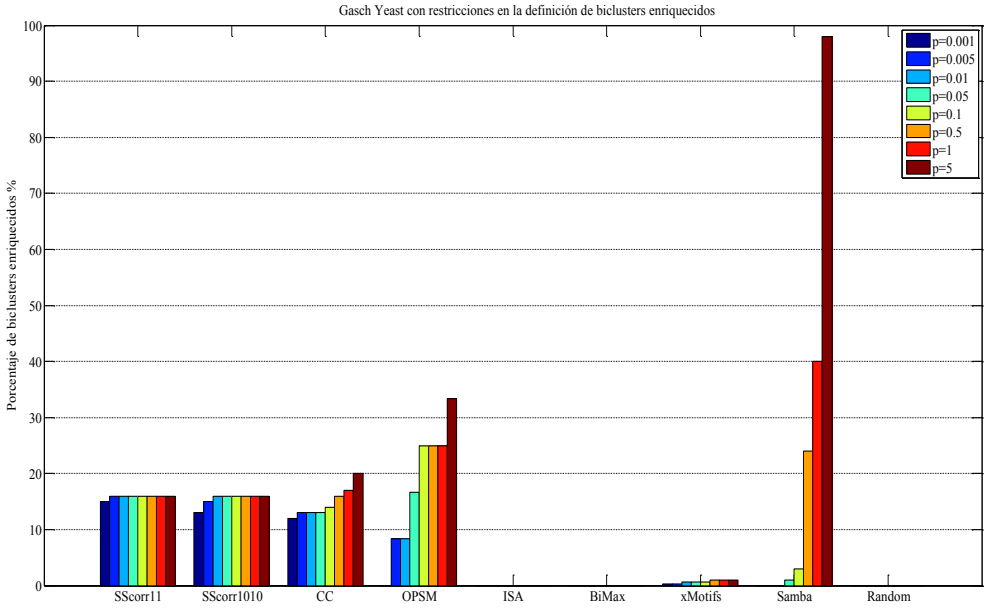

Figura 9.6: Comparación entre los distintos algoritmos de biclustering para el dataset GaschYeast con una definición más restrictiva de enriquecimiento biológico

La figura 9.7 muestra la comparación entre los resultados de SScorr11 y SScorr1010 respecto al algoritmo CCC-Biclustering propuesto en [59] usando los datos de Yeast. Los biclusters de CCC-Biclustering se han obtenido usando la herramienta BiGGEsTS [41]. Se muestran los resultados para todos los biclusters obtenidos por este algoritmo así como para un grupo de cien elegidos al azar. Se puede observar que SScorr11 y SScorr1010 obtienen un porcentaje mayor de biclusters enriquecidos. Se podía esperar este hecho debido a que SScorr11 y SScorr1010 obtienen 100 biclústers mientras que CCC-Biclustering realiza un búsqueda exhaustiva de todos aquellos biclusters maximales, restringidos a tener condiciones contiguas, y devuelve 14412 biclusters. Si se consideran tan sólo 100 biclusters de los obtenidos, los resultados de CCC-Biclustering mejoran a SSCorr para valores de significancia entre 0.001 y 0.01 y, por otro lado, obtienen resultados iguales para valores mayores o iguales que 0.05.

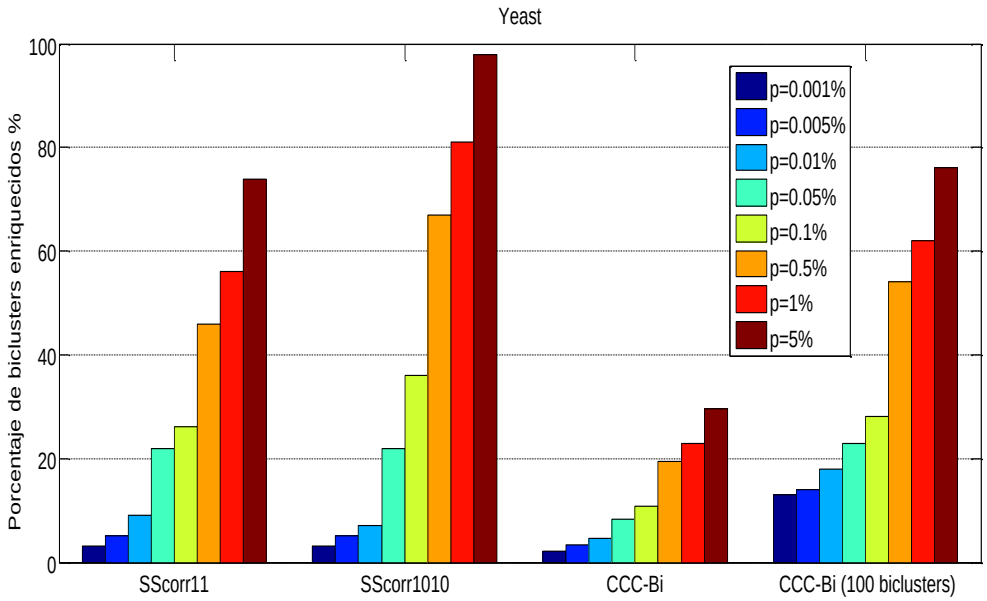

Figura 9.7: Comparación entre el algoritmo propuesto y el algoritmo CCC-Bi para el dataset GaschYeast

9.5. Estudio biológico

El estudio anterior de enriquecimiento se ha realizado con la herramienta presentada en [2] y, como ya hemos comentado, la información obtenida se muestra en las figuras 9.1, 9.2 y 9.3 y en la tabla 9.1. En este apartado vamos a hacer el estudio de concreto de varios biclusters para ver los procesos biológicos con los que están relacionados.

El análisis de *biYeastN15* identifica el proceso GO (GO: 0006412) con un p-value $5{,}81e-006$. Dicho término está relacionado con la *traducción*, mediante la que el ARNm da lugar a los aminoácidos en el ribosoma. El biclúster está compuesto por cinco genes que se co-expresan bajo siete condiciones experimentales. Otro ejemplo es el biclúster *biLymphomaN1* que identifica al término GO (GO:0016887) conocido como *ATPase activity* con un p-value de 0,0069. Este proceso se relaciona con las reacciones catalíticas que producen el ATP. Para el caso de los datos de GaschYeast se puede localizar el término (GO:0006412) relacionado con la *translación* con hasta ocho de los biclusters encontrados. Sin embargo, son más interesantes los resultados obtenidos para el biclúster *bi1GaschYeastN11*, que localiza el término GO (GO:0042254) llamado *ribosome biogenesis and assembly*, con un p-value de $1{,}38e-007$ y que está relacionado con los ribosomas y la síntesis de proteínas.

Capítulo 10

BISS: resultados

Hasta un descubrimiento tan aparentemente directo como el de un fósil a menudo es resultado de la formulación de una hipótesis encubierta pues, de otra manera, ¿por qué había de mirar alguien unos restos fósiles dos veces y acaso llevárselos para investigarlos después más detalladamente?.
Consejos a un joven científico. P.B. Medawar (pág. 104)

10.1. Introducción

En este capítulo presentamos el estudio experimental relativo a la propuesta presentada en la sección 7.3 del capítulo VII. En la sección 10.2 se describen los conjuntos de datos que se utilizan en estos experimentos. En la sección 10.3 se lleva a cabo el estudio experimental para la configuración de parámetros del algoritmo. Los resultados obtenidos se presentan en la sección 10.4. En concreto, un estudio comparativo con los principales algoritmos de biclustering usados comúnmente en la literatura como marco de referencia se muestran en la sección 10.4.1. Por otro lado, en la sección 10.4.2 se muestra la comparación con otros dos algoritmos que también utilizan la correlación como función de mecanismo de búsqueda. Así mismo, se presenta un estudio biológico cualitativo de los resultados obtenidos en la sección 10.4.3.

10.2. Descripción de los datos

Se han utilizado tres conjuntos de datos para los experimentos realizados, dos de la levadura *Saccharomyces cerevisiae* y uno de *Homo sapiens* relacionado con la enfermedad del alzheimer. La levadura se suele utilizar

como organismo modelo en muchos estudios y en consecuencia se puede encontrar mucha información relacionada en los repositorios (lo que facilita los problemas relacionados con la validación de la información, nomenclaturas de genes, etc).

Uno de los datos de la levadura, lo notaremos como *GaschYeast*, y el de Homo sapiens, lo notaremos como *Alzheimer*, se han descargado como parte del material suplementario de las referencias [77] y [78] respectivamente. Los datos de GaschYeast están formados por 2993 genes y 173 muestras y los del Alzheimer por 1663 genes y 33 condiciones. Los otros datos de la levadura se han descargado del repositorio *Gene Expression Omnibus* (GEO)[1], en concreto el registro *GDS1116* generado en el estudio llevado a cabo en la referencia [17]. Son datos temporales compuestos por 7084 genes y 131 condiciones. El preprocesamiento de estos datos de la levadura se ha llevado a cabo usando GEPAS[2] usando las opciones por defecto. En el caso de los valores perdidos en los perfiles de expresión de los genes, han sido sustituidos con la media de dicho perfil, es decir, la fila de la matriz en la que se encuentran, y en aquellos casos en los que existía más de un 80 % de valores perdidos, esa fila se ha eliminado de los datos. Tras este procesamiento la matriz de expresión está constituida por 6229 genes y 131 condiciones experimentales. Es decir, aproximadamente el 12 % de los genes son filtrados en el tratamiento.

10.3. Configuración de parámetros

	BISS			
	$M = 1$	$M = 10$	$M = 20$	$M = 40$
genes	11.0	281.4	386.5	415.1
condiciones	12.7	34.4	34.1	34.0
tamaño	138.0	10057.7	13244.8	13929.8
$\rho_{\lvert\cdot\rvert}(B)$	0.89	0.30	0.23	0.21

Tabla 10.1: Resultados obtenidos por BISS con distintos valores del parámetro M.

En esta sección se va a estudiar cómo afecta el parámetro que controla el volumen en la función objetivo (ecuación 7.6) al rendimiento del algoritmo. Se ejecuta BISS con los datos de GDS1116 con distintos valores del paráme-

[1] http://www.ncbi.nlm.nih.gov/geo/
[2] http://www.gepas.org/

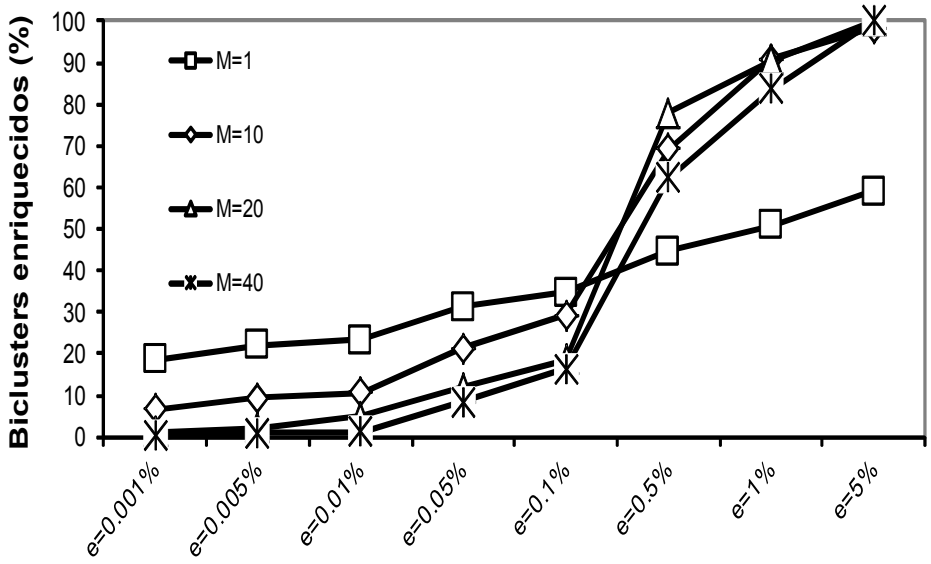

Figura 10.1: Porcentaje de biclusters enriquecidos obtenidos con BISS con diferentes valores del parámetro.

tro y se comparan los resultados obtenidos. El estudio es análogo para los otros conjuntos de datos y los resultados son similares.

Los valores elegidos para el estudio del parámetro son 1, 10, 20 and 40. La tabla 10.1 muestra el valor medio de los 100 biclusters obtenidos para el número de genes, condiciones y volumen de los biclusters, junto con la correlación media (ecuación 7.5). Se puede observar como un valor bajo en el parámetro implica un tamaño pequeño en los biclusters, en concreto un número bajo de genes. El número de genes aumenta cuando el parámetro también lo hace de 10 a 40, sin embargo el número de condiciones permanece constante. Se puede también observar que un valor bajo para el parámetro implica biclusters con valores de la correlación cercanos a 1 y por lo tanto compuesto por genes altamente correlacionados entre sí tanto negativa como positivamente. Esta situación en cambio implica biclusters con un tamaño muy pequeño, como cabía esperar, por lo que según sea el tamaño deseado se deberá elegir el valor del parámetro.

La figura 10.1 muestra la relevancia biológica de los resultados obtenidos mediante el análisis del enriquecimiento de los biclusters respecto GO para los distintos valores del parámetro M. El porcentaje de biclusters enriquecidos se ha estudiado usando las tres ramas de GO: BP, MF y CC, la figura muestra la media de los resultados obtenido para las tres. Generalmente se

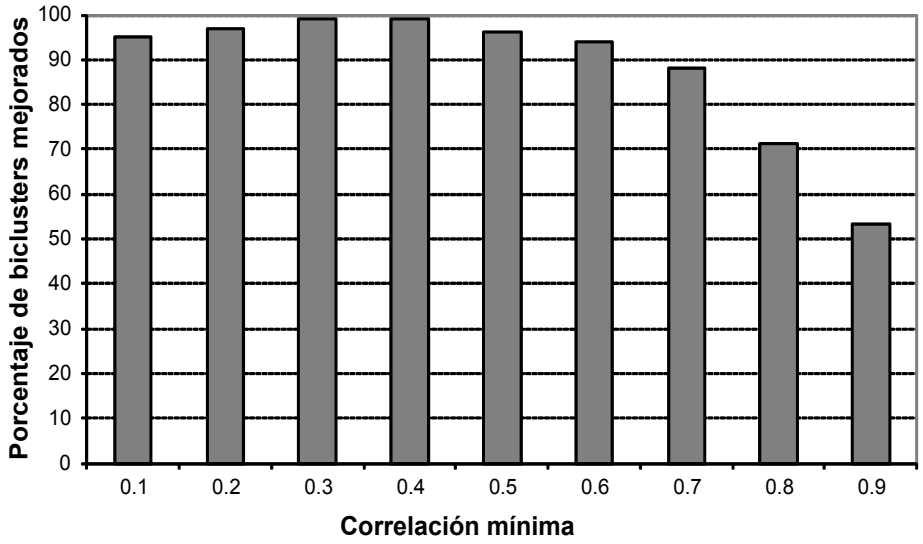

Figura 10.2: Porcentaje de biclusters enriquecidos según el valor de la función objetivo.

suele usar un nivel de significancia del 90 % ($e = 0{,}1\,\%$) o 95 % ($e = 0{,}05\,\%$), es decir, un biclúster se dice que está enriquecido si su p-value ajustado asociado a un término es menor que 0,1 o 0,05 respectivamente. Se puede observar en la figura que, para un nivel de significancia del 95 % el porcentaje de biclusters enriquecidos para $M = 1$ es mayor que para $M = 10$ (31,34 % frente 21,34 %). Sin embargo, para un nivel de significancia del 90 %, los valores para ambos parámetros son mucho más similares (34,67 % y 29,34 %).

Se puede concluir observando la tabla 10.1 y la figura 10.1 que un valor alto para la correlación implica mayor relevancia biológica, como era de esperar. Por lo tanto, el valor adecuado para el parámetro es un valor entre $M = 1$ y $M = 10$ y, de esta forma, se consigue el equilibrio entre tamaño y calidad. En consecuencia se ha elegido para la experimentación un valor del parámetro $M = 5$.

El umbral de la correlación que se utiliza en el método de la mejora se ajusta automáticamente (método presentado en 7.3.2). No obstante, la figura 10.2 muestra cómo interviene su valor en el método de la mejora. En la figura se puede ver el porcentaje de biclusters que mejoran cada valor del umbral. Con un umbral de 0,3 o 0,4 prácticamente todos los biclusters, el 99 % mejoran su valor, por lo que sería la mejor elección para dicho umbral.

10.4. Resultados

En esta sección se comparan los resultados obtenidos por BISS con los obtenidos por otros algoritmos de la literatura. Primero se presenta una comparación con un grupo de algoritmos considerados como clásicos, presentes en la herramienta BiCAT [10]: ChCh [24], ISA [12] y OPSM [11], y, por otro lado, con dos algoritmos también basados en la correlación BCCA [14] y BICLIC [93].

La configuración de los parámetros para BISS es de 5 como factor de penalización y 100 el número de biclusters que se obtienen. Se debe tener en cuenta que cada biclúster se encuentra mediante un proceso de búsqueda dispersa independiente y que, por lo tanto, cuanto mayor sea el número de resultados mejor se controlará la naturaleza aleatoria propia de toda metaheurística de tipo evolutivo.

10.4.1. Comparación con algoritmos clásicos

Los algoritmos ChCh [24], ISA [12] y OPSM [11], presentes en la herramienta BiCAT [10], se han ejecutado sobre los datos utilizando los parámetros por defecto o recomendados por los autores. El algoritmo xMotifs [68] también está presente en BiCAT pero no se ha podido aplicar a los datos del experimento debido al gran número de condiciones de las matrices. xMotifs sólo se puede aplicar a matrices de expresión con menos de 64 condiciones.

La tabla 10.2 muestra los diferentes rasgos de los biclusters obtenidos por BISS, ChCh, ISA y OPSM para los conjuntos de datos GDS1116, GaschYeast y Alzheimer. En particular, la tabla muestra la media del número de genes, condiciones y tamaño de los biclusters, junto con el tamaño del más pequeño y el mayor, así como la media del valor de la correlación y el número de pares de genes correlacionados negativamente. Los valores de $\rho_{|\cdot|}(B)$ y $\rho(B)$ son los valores de la correlación media considerando el valor absoluto (ecuación 7.5) y sin considerarlo. Si la diferencia entre ambos valores es pequeña significa que hay pocos genes con correlación negativa entre ellos, en caso contrario se tratará de biclusters en los que dos genes presenten patrones de activación-inhibición. Se puede observar que OPSM no captura correlaciones negativas para GaschYeast y Alzheimer, sin embargo sí para GDS1116. En este caso, obsérvese que los valores entre la correlación con valor absoluto y sin valor absoluto son cercanos entre sí (0,95 y 0,89), lo cual quiere decir que la correlación negativa entre pares de genes tiene un valor muy pequeño. En el caso de ISA se puede observar que tanto para GDS1116 como para GaschYeast, el número de pares de genes correlacionados negati-

vamente es pequeño. Por tanto, los valores de la correlación con y sin valor absoluto son bastante cercanos entre sí (0,74 y 0,63, 0,60 y 0,52, respectivamente). En el caso del conjunto de datos de Alzheimer ISA obtiene tres biclusters con un valor alto para la correlación negativa pero sólo tienen dos condiciones, por lo que se pueden considerar como triviales y sin información relevante. Estas observaciones muestran que tanto OPSM como ISA no son adecuados para la búsqueda de patrones de activación-inhibición. Por otro lado, se puede observar en la tabla la gran variabilidad entre los resultados que cada algoritmo aporta.

La figura 10.3 muestra más claramente la diversidad entre los tamaños de los biclusters obtenidos con BISS, ChCh, ISA y OPSM para el conjunto de datos de GDS1116. Cada punto representa un biclúster, donde los ejes x e y representan genes y condiciones respectivamente. OPSM tiene 5 biclusters con más de 400 genes, y muy pocas condiciones, de 2 a 6, que no aparecen en la figura. Se puede observar que los biclusters de BISS (cuadrados) se agrupan en dos grupos: uno con un número alto de genes y otros con menos de 50 genes. Además todos los biclusters obtenidos por BISS tienen más condiciones que los obtenidos por ChCh, ISA u OPSM. Se puede también observar que los biclusters de ISA (círculos) y los de ChCh (diamantes) tienen menos de 10 condiciones, aunque los de ISA tienen todos más de 50 genes.

La figura 10.4 presenta el porcentaje de biclusters enriquecidos en uno o más términos GO para BISS, ChCh, ISA y OPSM para los conjuntos de datos de GDS1116 y GashYeast. Este porcentaje se ha calculado para las tres ramas de GO (BP, CC y MF) y para los umbrales de significancia más usados, 0,05 % y 0,1 %. Se puede observar que BISS obtiene mejores resultados que ISA para los datos de GaschYeast y mejor que ChCh para ambos conjuntos de datos en las tres ramas de GO y para ambos umbrales. Sin embargo, los resultados de ISA y OPSM son mejores que los de BISS para GDS1116. No obstante, ambos algoritmos obtienen un número muy bajo de genes correlacionados negativamente.

Aquellos biclusters compuestos por un número muy alto de genes tienen mayor probabilidad de tener subgrupos de genes que sean significativos en términos GO. La influencia del tamaño de los biclusters en las comparaciones entre algoritmos ha sido estudiada en la literatura [14, 77, 32]. Generalmente, para evitar el efecto del tamaño de los biclusters, se suelen filtrar aquellos compuestos por un número muy alto de genes. Siguiendo la experimentación llevada a cabo en [14], se ha considerado 50 como el número máximo de genes por biclúster y en función de ello se han filtrado los resultados.

La figura 10.5 muestra el porcentaje de biclusters enriquecidos después

del proceso de filtrado. Se debe tener en cuenta que todos los biclusters de ISA para GDS1116 desaparecen al filtrase los resultados ya que todos los biclusters tenían más de 50 genes. Se puede apreciar que BISS obtiene más biclusters enriquecidos que ISA para GaschYeast y más que ChCh para ambos conjuntos de datos en las tres ramas de GO. Por otro lado, BISS mejora los resultados de OPSM para GDS1116 y obtiene resultados similares en el caso de GaschYeast.

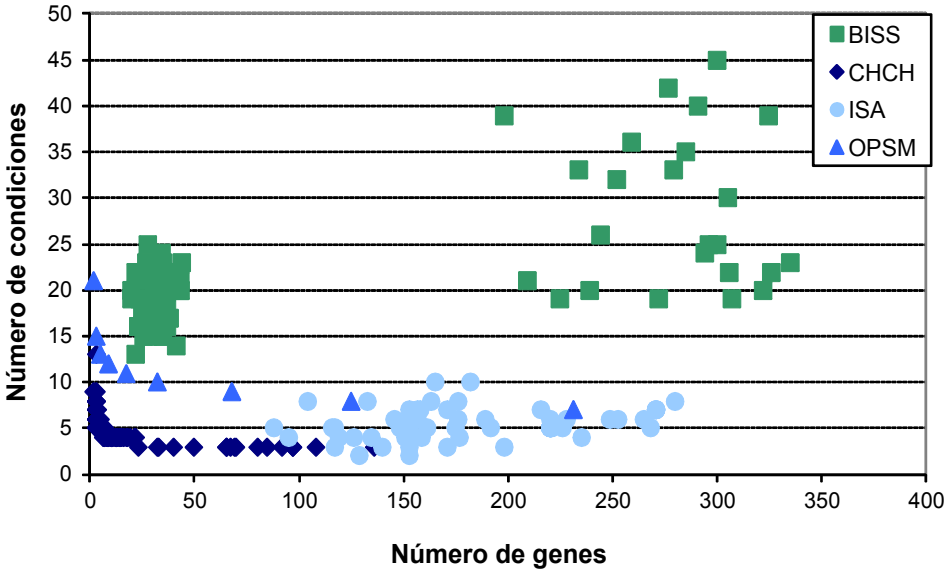

Figura 10.3: Tamaño de los biclusters obtenidos con los algoritmos BISS, ChCh, ISA y OPSM con los datos de GDS1116.

	num. bicluster	num. genes	num. condiciones	min. tamaño	media tamaño	max. tamaño	$\rho_{\|\cdot\|}(B)$	$\rho(B)$	pares de genes corr. neg.
GDS1116									
BISS	100	90.4	21.5	(22x13)	2370.29	(300x45)	0.65	0.21	4806.4
ChCh	100	19.0	4.68	(3x5)	66.9	(136x3)	0.49	0.00	266.9
ISA	60	182.1	5.5	(129x2)	1023.5	(280x8)	0.74	0.63	1489.42
OPSM	14	678.1	9.0	(2x21)	2207.1	(4186x2)	0.95	0.89	290033.4
GaschYeast									
BISS	100	96.2	26.2	(20x18)	2680.4	(292x45)	0.75	0.35	3839.7
ChCh	100	70.6	19.1	(45x9)	1407.1	(222x90)	0.3	0.0	1986.9
ISA	66	76.3	8.7	(11x11)	645.7	(136x10)	0.60	0.52	366.9
OPSM	12	95.6	12.5	(4x18)	849.8	(387x7)	0.98	0.98	0.0
Alzheimer									
BISS	100	48.7	15.1	(36x10)	741.8	(62x18)	0.83	0.16	488.7
ChCh	100	13.4	6.7	(3x7)	170.55	(152x33)	0.54	0.09	34.2
ISA	3	51.67	2.0	(43x2)	103.3	(56x2)	1.0	0.1	609.67
OPSM	13	246.5	9.8	(12x13)	925.0	(809x3)	0.96	0.96	0.0

Tabla 10.2: Resultados obtenidos por BISS y por los algoritmos clásicos, ChCh, ISA y OPSM para los datos de GDS1116, GaschYeast y Alzheimer.

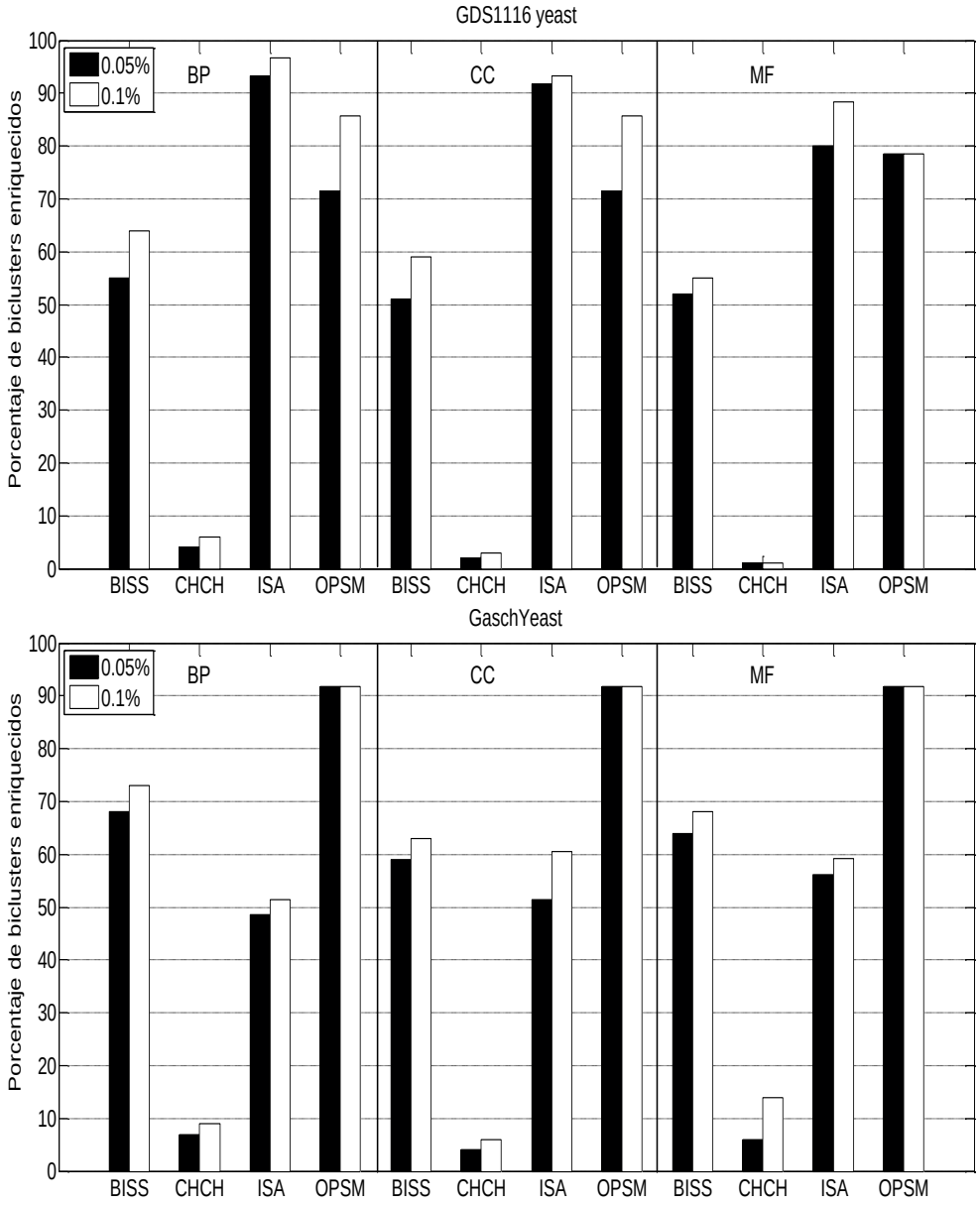

Figura 10.4: Porcentaje de biclusters enriquecidos obtenidos por BISS, ChCh, ISA and OPSM para las ramas de GO BP, CC y MF con los datos de GDS1116 y GaschYeast.

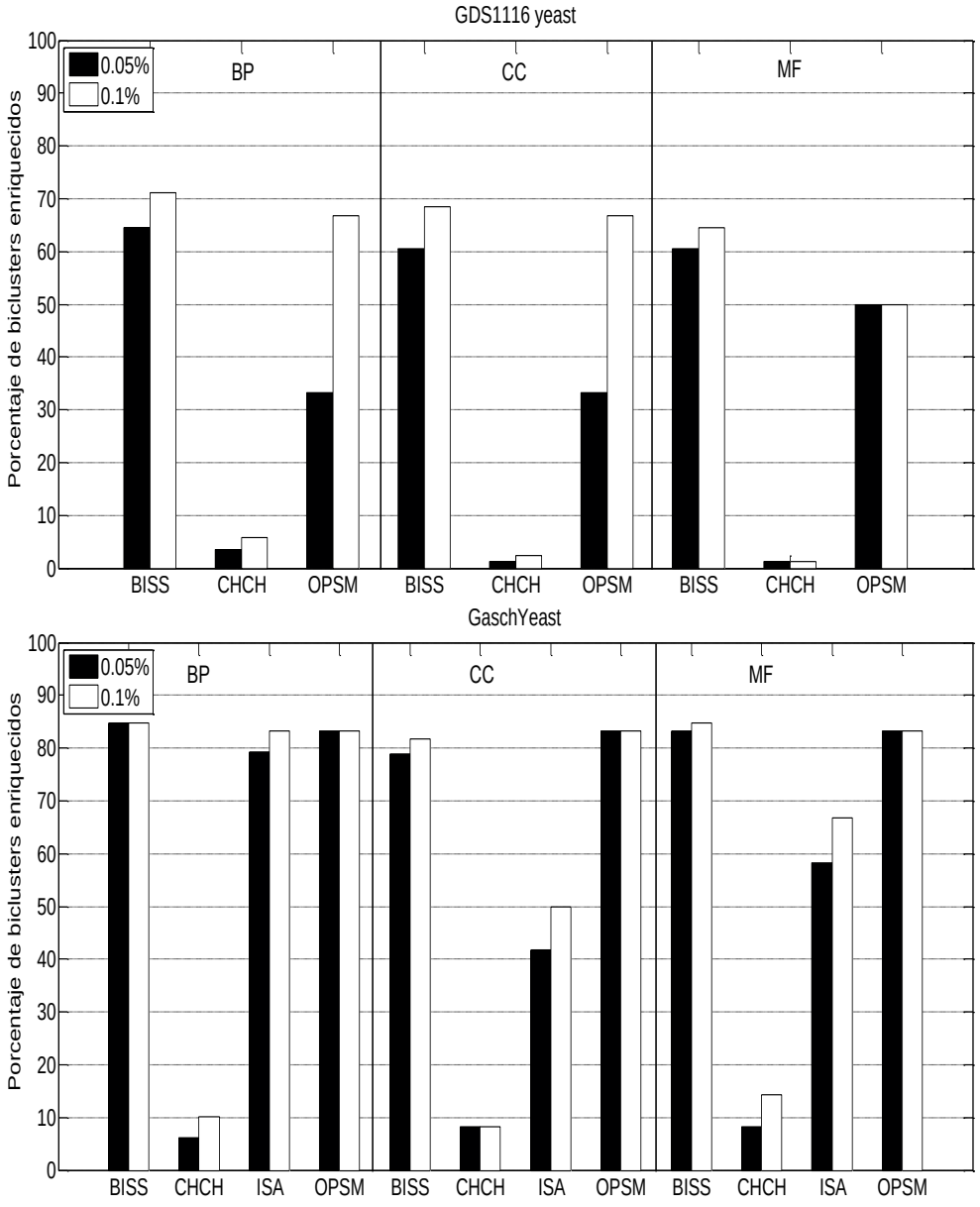

Figura 10.5: Porcentaje de biclusters enriquecidos obtenidos por BISS, ChCh, ISA y OPSM para las ramas de GO BP, CC y MF con los datos de GDS1116 y GaschYeast tras el proceso de filtrado.

10.4.2. Comparación con algoritmos basados en correlación

Se ha comparado el rendimiento de BISS con el de otros dos algoritmos que también están basados en la correlación, en concreto BCCA [14] y BICLIC [93]. BCCA no obtiene genes correlacionados negativamente, sin embargo BICLIC sí los obtiene, por lo que encuentra patrones de activación-inhibición.

El algoritmo BCCA tiene dos configuraciones posibles de parámetros, según si se buscan biclusters con solape, notaremos BCCA-yes, o sin solape, notaremos BCCA-not. En el caso de BCCA-not la configuración de parámetros necesita el número de biclusters que se buscan y el umbral de la correlación. En el caso de BCCA-yes tan sólo umbral de la correlación. Este umbral depende de la naturaleza de los datos y del tamaño deseado de los biclusters encontrados, que disminuye cuando el umbral aumenta. La principal diferencia entre los dos algoritmos BCCA es que BCCA-yes busca todos los resultados posibles mediante una búsqueda exhaustiva y por lo tanto el tiempo de ejecución es mucho mayor que el de BCCA-not. En el caso de BCCA-not se ha elegido como parámetro 100 biclusters y, tras realizar un estudio experimental se ha elegido 0,2 para GDS1116 y 0,8 para GashcYeast y Alzheimer para el umbral de correlación. En el caso de BCCA-yes se ha elegido como umbral de correlación 0,85. En este segundo caso el estudio experimental no es posible dado el alto coste computacional de la ejecución, por lo que se ha elegido este valor siguiendo la recomendación de la referencia [14].

El algoritmo BICLIC tiene tres parámetros de entrada: el umbral de correlación y el número mínimo de genes y condiciones para cada biclúster que se encuentre. Se han elegido como parámetros 0,9, 5 y 5 para GaschYeast, 0,85, 25 y 10 para GDS1116 y, por último, 0,6, 5 y 2 para los datos de Alzheimer. En el caso de los datos de GaschYeast se han elegido estos parámetros siguiendo los parámetros por defecto proporcionado por los autores del algoritmo [93]. En cambio, en los casos de GDS1116 y Alzheimer, dado que no se obtenían con esta elección resultados satisfactorios, se ha llevado a cabo un proceso experimental para la elección de los valores de los parámetros.

La tabla 10.3 muestra la información de los biclusters obtenidos con BISS, BCCA-not, BCCA-yes y BICLIC. BISS y BCCA-not obtienen 100 biclusters como era esperado. BCCA-yes obtiene 1662, 17322 y 368, mientras que BICLIC obtiene 5988, 14791 y 4405 para los datos de GDS1116, GaschYeast y Alzheimer respectivamente. Se puede observar en la tabla que BCCA tiene 0 pares de genes correlacionados negativamente, por lo que se

puede deducir que no obtiene patrones del tipo activación-inhibición a pesar de patrones interesantes desde el punto de vista biológico [94]. Por otro lado, BICLIC si captura este tipo de patrones para GaschYeast y Alzheimer pero para GDS1116. Observando la tabla se puede deducir que BICLIC obtiene pocos genes correlacionados negativamente ya que hay poca diferencia entre la correlación con valor absoluto y sin valor absoluto (0,76 y 0,61 para GaschYeast y 0,69 y 0,68 para los datos de Alzheimer).

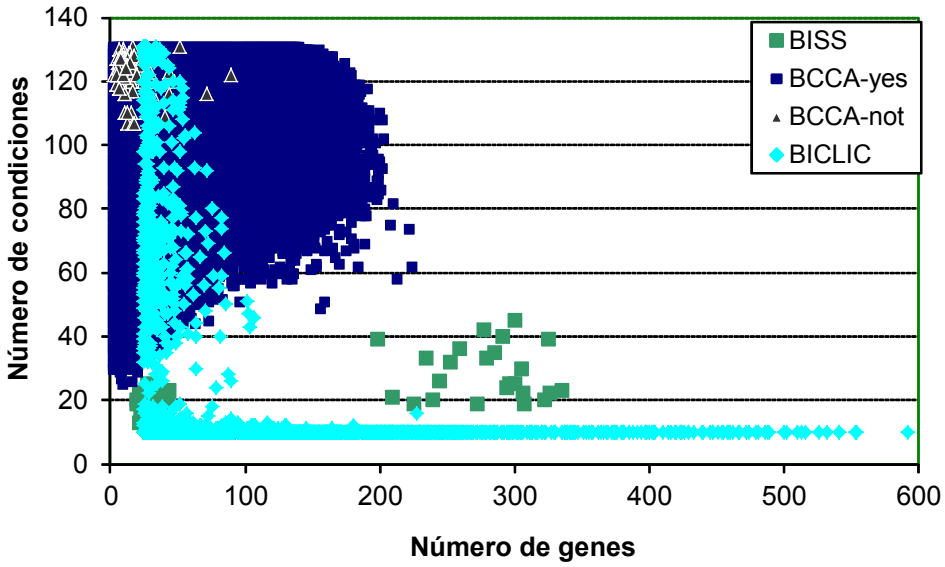

Figura 10.6: Tamaño de los biclusters obtenidos por BISS, BCCA-not, BCCA-yes y BICLIC para los datos de GDS1116.

	num. bicluster	num. genes	num. condiciones	min. tamaño	media tamaño	max. tamaño	$\rho_{\lvert\cdot\rvert}(B)$	$\rho(B)$	pares de genes corr. neg
GDS1116									
BISS	100	90.4	21.5	(22x13)	2370.29	(300x45)	0.65	0.21	4806.4
BCCA-not	100	16.5	122.7	(3x122)	2022.5	(90x122)	0.36	0.36	0.0
BCCA-yes	1662	10.4	86.3	(2x18)	943.0	(32x114)	0.85	0.85	0.0
BICLIC	5988	76.7	14.9	(25x10)	940.5	(685x10)	0.85	0.85	0.0
GaschYeast									
BISS	100	96.2	26.2	(20x18)	2680.4	(292x45)	0.75	0.35	3839.7
BCCA-not	100	2.6	22.7	(2x8)	54.0	(15x19)	0.77	0.77	0.0
BCCA-yes	17322	59.4	97.7	(2x3)	5906.5	(121x163)	0.9	0.9	0.0
BICLIC	14791	178.9	29.2	(5x6)	2249.3	(139x159)	0.76	0.61	5797.1
Alzheimer									
BISS	100	48.7	15.1	(36x10)	741.8	(62x18)	0.83	0.16	488.7
BCCA-not	100	9.1	14.8	(2x8)	97.8	(24x29)	0.86	0.86	0.0
BCCA-yes	368	8.9	16.1	(2x5)	118.8	(16x23)	0.88	0.88	0.0
BICLIC	4405	516.1	14.2	(5x13)	5888.9	(899x25)	0.69	0.68	2799.9

Tabla 10.3: Resultados obtenidos por BISS y por los algoritmos BCCA-not, BCCA-yes y BICLIC para los datos de GDS1116, GaschYeast y Alzheimer.

La figura 10.6 muestra los tamaños de los biclusters encontrados. Se puede observar que BCCA-not, BCCA-yes y BICLIC obtienen biclusters con un número de condiciones mayor y un número de genes menor que los obtenidos por BISS.

La figura 10.7 presenta el porcentaje de biclusters enriquecidos en uno o más términos GO para BISS, BCCA-not, BCCA-yes y BICLIC para GDS1116 y GaschYeast filtrando los resultados considerando sólo aquellos biclusters que tuvieran menos de 50 genes. Se puede observar que BISS mejora a BCCA-not, BCCA-yes y BICLIC en las tres ramas de GO y para ambos umbrales de significancia para GaschYeast. Por otro lado, BISS mejora a BCCA-not, BCCA-yes y BICLIC en las ramas de CC y MF con los datos de GDS1116. Sin embargo en la rama BP, los resultados de BISS y de BCCA-yes son similares para 0,1 % y sin embargo los de BCCA-yes mejoran los de BISS para 0,05 %. No obstante, hay que recordar que BCCA-yes no podía obtener correlaciones negativas y por lo tanto patrones de activación-inhibición.

Con el objetivo de enfatizar las diferencias entre los resultados obtenidos, se puede realizar el estudio teniendo una definición aún más restrictiva del significado del enriquecimiento de los biclusters. La mayoría de las veces el enriquecimiento de un biclúster respecto a un determinado término GO se debe a un pequeño subgrupo de genes. Se puede proporcionar una definición más restrictiva según la cual un biclúster está *altamente enriquecido* si tiene un número de genes por encima de un umbral, previamente establecido, significativos en un determinado término GO [2, 23]. Por ejemplo, dicho umbral se puede considerar un porcentaje del número de genes o un valor de 5 genes como se propone en la bibliografía.

La figura 10.8 muestra el porcentaje de biclusters altamente enriquecidos para BISS, BCCA-yes y BICLIC. Se puede observar que los resultados obtenidos por BISS son mejores que los obtenidos por BCCA-yes y BICLIC para las ramas de BP y CC con el conjunto de datos de GDS1116. Para la rama MF, BISS tiene 27,6 % y 28,9 % para los niveles de significancia 0,1 % y 0,05 % respectivamente, BCCA-yes tiene 27,5 % para ambos niveles y BICLIC tiene 30,4 % y 30,8 %. Además, se puede observar que los resultados obtenidos por BISS y BCCA-yes son similares y siempre mejores que los de BICLIC para los datos de GaschYeast.

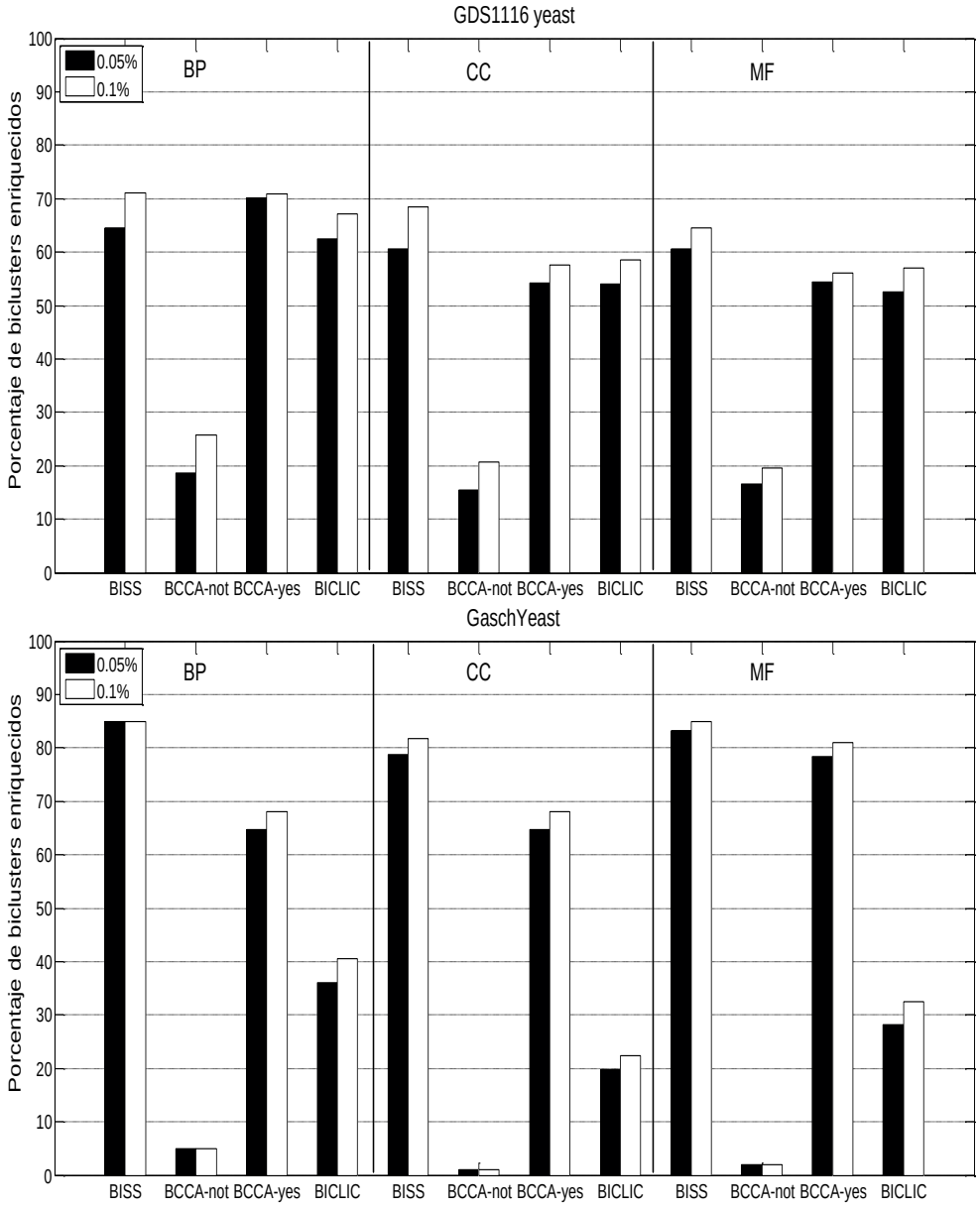

Figura 10.7: Porcentaje de biclusters enriquecidos obtenidos por BISS, BCCA-not, BCCA-yes y BICLIC para las ramas de GO BP, CC y MF con los datos de GDS1116 y GaschYeast.

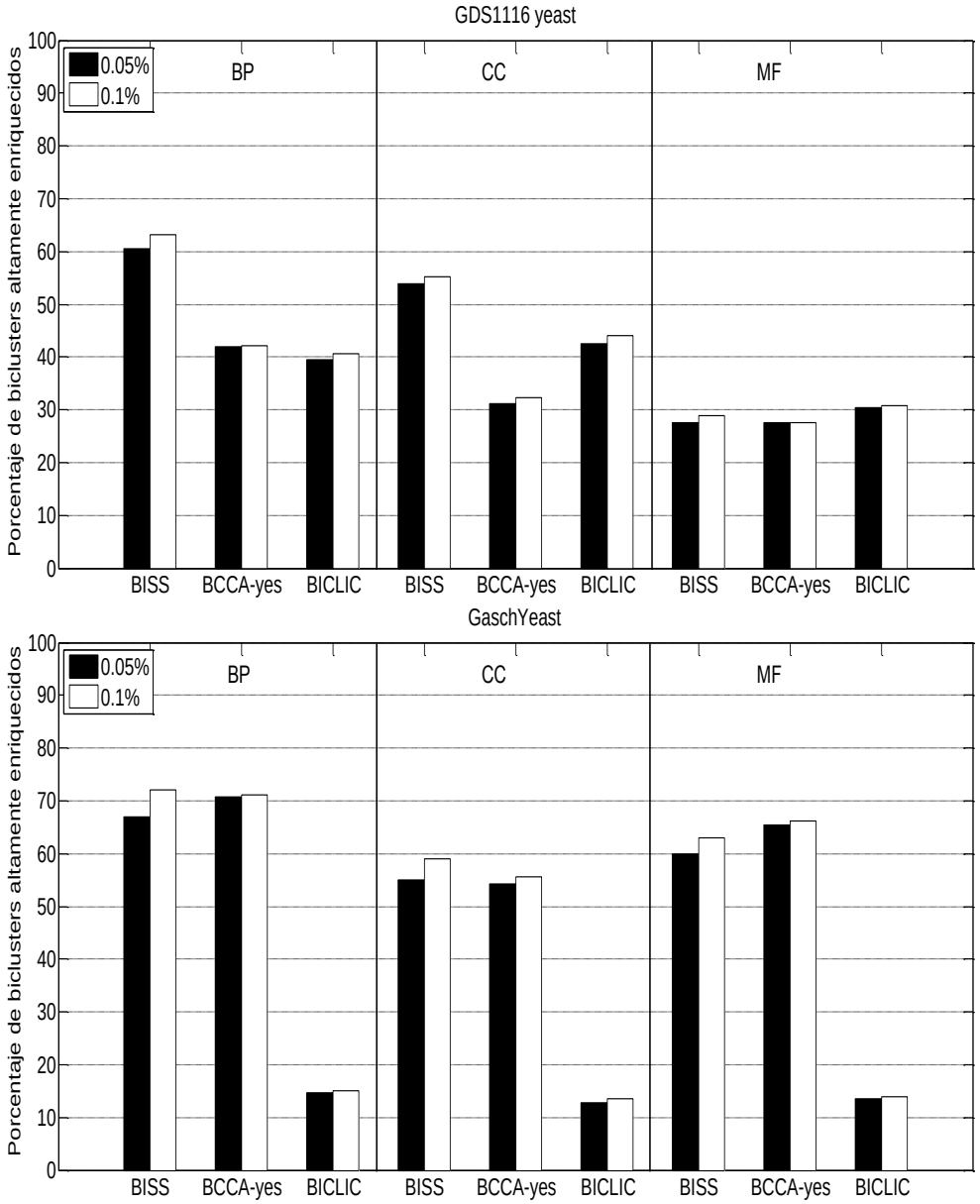

Figura 10.8: Porcentaje de biclusters altamente enriquecidos obtenidos por BISS, BCCA-not, BCCA-yes y BICLIC para las ramas BP, CC y MF con los datos GDS1116 y GaschYeast.

10.4.3. Estudio de la significancia biológica de los biclusters

En esta sección se presenta un estudio biológico cualitativo de algunos de los biclusters obtenidos por BISS.

La figura 10.9 muestra la representación gráfica de un biclúster com-

puesto por 35 genes y 13 condiciones obtenidos por BISS para los datos del Alzheimer. Observando la figura se pueden observar patrones de desplazamiento y escalado además de patrones de activación-inhibición. En concreto, se han representado dos genes correlacionados positivamente usando líneas discontinuas y un tercer gen con correlación negativa respecto a los dos genes anteriores representado usando línea negra.

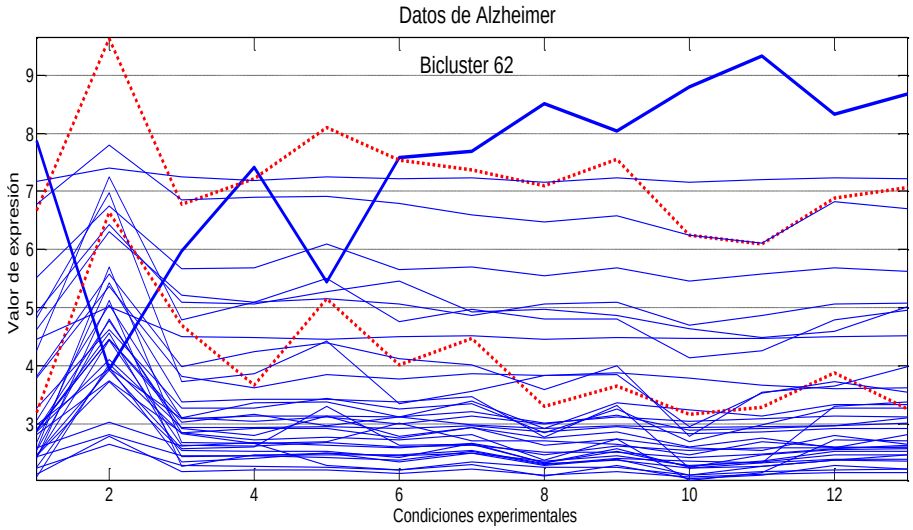

Figura 10.9: Representación de un bicluster obtenido por BISS con los datos de Alzheimer.

La tabla 10.4 muestra el estudio biológico de varios biclusters de BISS obtenidos para los datos de GDS1116 y GaschYeast. Se ha utilizado la herramienta web FuncAssociate[3] para generar la información que se refleja en la tabla. En concreto, el tamaño de los biclusters, el número de genes del biclúster asociados con el término GO correspondiente, el número de genes en dicho término, el valor p-value ajustado y la descripción del término GO para los dos primeros términos GO asociados. Se puede observar que el biclúster #1 tiene 22 y 12 genes en los términos GO:0022626 y GO:0022627 respectivamente, estando formados dichos términos por 163 y 64 genes. Estos términos GO pertenecen a la rama CC y están relacionados con la funcionalidad de los ribosomas. El biclúster #70 tiene 4 genes asociados con los términos GO:0000722 y GO:0003678. Para los datos de GaschYeast, los biclusters #1 y #62 tienen aproximadamente la mitad de sus genes relacionados en los términos GO:0034660 y GO:0034470.

[3]http://llama.mshri.on.ca/funcassociate/

	bicluster tamaño	genes del bi. en término GO	genes en término GO	p-value	término GO & descripción
GDS1116					
bicluster #1	(38,22)	22	163	$< 0{,}001$	GO:0022626 cytosolic ribosome
		12	64	$< 0{,}001$	GO:0022627 cytosolic small ribosomal subunit
bicluster #70	(24,19)	4	19	$< 0{,}001$	GO:0000722 telomere maintenance via recombination
		4	36	0,003	GO:0003678 DNA helicase activity
GaschYeast					
bicluster #1	(51,30)	12	64	$< 0{,}001$	GO:0022627 cytosolic small ribosomal subunit
		23	163	$< 0{,}001$	GO:0022626 cytosolic ribosome
bicluster #62	(44,20)	21	483	$< 0{,}001$	GO:0034660 ncRNA metabolic process
		19	426	$< 0{,}001$	GO:0034470 ncRNA processing

Tabla 10.4: Significancia biológica de varios biclusters obtenidos por BISS para los datos de GDS1116 y GaschYeast.

Capítulo 11

GoldBinch: resultados

La noble ciencia de la genealogía pierde esplendor por la extrema imperfección de sus archivos.
El origen de las especies. Charles Darwin (pág. 774)

11.1. Introducción

En este capítulo presentamos el estudio experimental relativo a la integración de información biológica presentada en el capítulo VIII. El objetivo que nos marcamos en la experimentación presentada es el análisis de cómo afecta en el rendimiento de un algoritmo de biclustering la integración de información biológica. Se comparan principalmente los resultados obtenidos con las medidas FracGO y SimNTO para la integración de información. Los resultados también se comparan con un grupo de algoritmos de biclustering considerados como clásicos, ChCh [24], ISA [12], OPSM [11] y xMotifs [68], y que suelen ser usados como marco de trabajo en la literatura. La metodología de comparación se basa en la propuesta en la referencia [77], basada en el estudio del porcentaje de biclusters enriquecidos, y que se usa comúnmente en la literatura.

11.2. Datos

Para los experimentos realizados se han utilizado dos conjuntos de datos de la levadura descargados del repositorio GEO [31] con identificadores $GDS1116$ y $GDS2914$. El primero está compuesto por 7085 perfiles de expresión y 131 muestras. Estos datos fueron generados en un estudio sobre la variabilidad genética en la evolución de poblaciones de levadura obtenidas mediante el cruce de cepas de dos tipos distintos. El segundo conjunto

de datos está formado por 15488 perfiles de expresión y 36 muestras y se trata de datos de tipo temporal relacionados con el ciclo de la levadura de muestras tratadas con dosis altas o bajas de cafeína. Estos datos en crudo se han tratado con la herramienta web Babelomics [62] siguiendo el mismo preprocesamiento para ambos conjuntos de datos. Se han filtrado aquellos perfiles de expresión con más de un 30 % de valores nulos así como, por otro lado, se han cambiado por la media de los valores del perfil de expresión correspondiente el resto de valores nulos que permanecían en los datos. Aquellos perfiles que aparecen repetidos varias veces para un mismo gen han sido fusionado mediante la media de sus valores. Tras este procesamiento de los datos en crudo, la matriz de expresión de $GDS1116$ está compuesta de 882 genes y 131 condiciones y, por otro lado, $GDS2914$ por 975 genes y 36 condiciones.

La información biológica se proporciona a través de un fichero de anotaciones directas de los términos GO obtenidos para la rama de BP de GO. Este tipo de fichero relaciona cada gen con el conjunto de términos GO en los que interviene. Se puede generar de varias formas según se pongan más o menos restricciones a GO. En este trabajo se ha tenido en cuenta la estructura de árbol de GO y los ficheros de anotaciones se han construido propagando las anotaciones hasta los términos altos de la ontología. Estos ficheros de anotaciones se han generado usando Babelomics con las opciones por defecto para este tipo de tareas. El fichero de anotaciones para los datos $GDS1116$ tiene información sobre anotaciones para 632 de los 882 genes de la matriz de expresión. Este fichero contiene 245 términos GO diferentes y el número medio de términos GO por gen es de 10,6. Por otro lado, para $GDS2914$ de los 975 genes de su matriz de expresión, el fichero de anotaciones generado contiene información para 658 genes, contiene 256 términos GO y el número medio de términos por gen es igual a 10,1.

11.3. Resultados

Se presenta a continuación los resultados obtenidos con el objetivo de mostrar que la integración de información mejora el rendimiento y se consiguen encontrar biclusters de mayor calidad. Los resultados se han obtenido usando las distintas configuraciones de la función objetivo propuestas en el capítulo VIII. Se estudian estos resultados según el tamaño de los biclusters obtenidos, el porcentaje de enriquecimiento así como el solapamiento. Así mismo, se presentan también los resultados obtenidos usando otras fuentes de información diferentes a GO, como KEGG e Interpro, y usando otras medidas de similitud funcional también basadas en la similitud entre pares

Configuración función objetivo	Numero de biclusters	(M_1, M_2, M_3)	Biclusters enriquecidos (%)		
			BP	MF	CC
1-SimNTO	10	(2,1,1)	100	100	100
1-SimNTO	50	(2,1,1)	100	98	100
1-SimNTO	100	(2,1,1)	99	97	97
1-FracGO	10	(2,1,1)	100	70	50
1-FracGO	50	(2,1,1)	100	72	64
1-FracGO	100	(2,1,1)	100	73	67
0	10	(2,1,0)	85	82	88
0	50	(2,1,0)	82	72	84
0	100	(2,1,0)	85	82	88

Tabla 11.1: Porcentaje de biclusters enriquecidos obtenidos por el algoritmo GoldBinch para diferentes valores del número de biclusters.

de genes. Cada ejecución del algoritmo encuentra 100 biclusters. La tabla 11.1 presenta el porcentaje de biclusters enriquecidos obtenidos por el algoritmo con distintos valores para el parámetro de entrada que determina el número de biclusters que se desean obtener (10, 50, y 100). Se puede observar que el porcentaje de biclusters enriquecidos no depende de esta elección y por lo tanto no es un parámetro que afecte al estudio. Se debe tener en cuenta que el algoritmo obtiene cada biclúster de manera independiente a través de un procedimiento no determinista, es por ello que se elige un valor suficientemente grande para el número de biclusters que se encuentran, concretamente el valor de 100.

Datos	Función objetivo		Tamaño	Biclusters enriquecidos (%)			Términos GO por biclus. (BP)	Tiempo (segundos)
	Medida f_3	Parámetros (M_1, M_2, M_3)		BP	MF	CC		
GDS1116	1-SimNTO	(2, 1, 1)	(11.6 × 15.6)	99	97	97	5.74	28.65
		(2, 1, 2)	(8.7 × 15.2)	87	75	84	3.67	15.78
		(2, 2, 1)	(10.1 × 5.1)	95	83	89	4.5	24.03
	1-FracGO	(2, 1, 1)	(181.6 × 19.4)	100	73	67	1	5410.98
		(2, 1, 2)	(193.9 × 19.3)	100	79	79	1	5546.12
		(2, 2, 1)	(109.2 × 3)	100	47	52	1	5682.65
	0	(2, 1, 0)	(23.5 × 14.7)	85	82	88	2.16	38.29
		(2, 2, 0)	(46.8 × 3.2)	25	17	21	0.4	11.51
GDS2914	1-SimNTO	(2, 1, 1)	(10.9 × 10.9)	87	33	33	5.45	27.76
		(2, 1, 2)	(8.0 × 10.2)	65	24	33	2.94	15.67
		(2, 2, 1)	(9.5 × 3.3)	87	25	30	5.39	22.18
	1-FracGO	(2, 1, 1)	(180.6 × 15.1)	100	97	94	1.01	5420.84
		(2, 1, 2)	(193.2 × 15.8)	100	98	94	1	5920.09
		(2, 2, 1)	(102.8 × 3)	100	72	67	1	6311.43
	0	(2, 1, 0)	(24.1 × 9.0)	2	5	6	0.02	13.14
		(2, 2, 0)	(37.1 × 3.1)	13	15	17	0.19	7.16

Tabla 11.2: Resultados obtenidos con distintas configuraciones de la función objetivo para los dos conjuntos de datos GDS1116 y GDS2914. Cada columna muestra una ejecución que obtiene 100 biclusters.

Datos	Algoritmo	Número of biclusters	Tamaño	Biclusters enriquecidos (%)			Términos GO por bicluster (BP)
				BP	MF	CC	
GDS1116	ChCh	100	(21.9 × 18.4)	10	8	13	0.30
	ISA	11	(50.7 × 6.5)	72.7	72.7	72.7	6.45
	OPSM	16	(128.1 × 10.4)	75	81.2	75	20.06
	xMotifs	-	-	-	-	-	-
GDS2914	ChCh	100	(17.2 × 8.8)	4	4	2	0.04
	ISA	5	(28.6 × 2)	20	20	0	0.60
	OPSM	11	(164.4 × 7.2)	45.45	36.4	36.4	28.64
	xMotifs	999	(61.2 × 5)	64.76	31.5	52.15	1.13

Tabla 11.3: Resultados obtenidos por los algoritmos clásicos ChCh, ISA, OPSM y xMotifs para los dos conjuntos de datos GDS1116 y GDS2914.

La tabla 11.2 resume la información de los biclusters obtenidos usando diferentes configuraciones de la función objetivo para los conjuntos de datos $GDS1116$ y $GDS2914$. La columna *Medida* indica si la función objetivo tiene en cuenta la integración de la información a través de la medida SimNTO (Eq. 8.9) o FracGO (Eq. 8.6) o si no se lleva a cabo integración de información ($f_3 = 0$ en Eq. 8.1). La columna *Parámetros* representa el peso correspondiente a cada término de la función objetivo. Es importante destacar que todos los términos varían entre 0 y 1.

El parámetro M_1 relacionado con el tamaño de los biclusters se fija a 2 para evitar encontrar biclusters con un número trivial de genes o condiciones[69], de esta manera se evita encontrar biclusters compuestos por ejemplo por tan sólo dos genes o condiciones. Las configuraciones de la función objetivo estudiadas son las siguientes: ($M_2 = 1$, $M_3 = 1$) que le da el mismo peso al término relativo a la correlación que al de la integración biológica, ($M_2 = 2$, $M_3 = 1$) donde tiene más importancia el término de la correlación entre los genes que el relativo a la integración y, por último, ($M_2 = 1$, $M_3 = 2$) donde la situación es justo la contraria. En el caso que no se lleve a cabo la integración de información biológica se tiene que $M_3 = 0$ y por lo tanto tan sólo hay dos casos de estudio según la importancia del término relativo a la correlación $M_2 = 1$ o $M_2 = 2$. Se debe tener en cuenta que M_1 no puede ser igual a cero para evitar biclusters triviales [69]. Por otro lado, si $M_2 = 0$ el número de condiciones no se podría controlar y se estaría llevando a cabo un algoritmo de clustering pues se consideran todas las condiciones en cada biclúster.

El resto de columnas de la tabla 11.2, *Tamaño*, *Biclusters enriquecidos (%)*, *Términos GO por bicluster* y *Tiempo*, muestran respectivamente el valor medio del número de genes y condiciones de los 100 biclusters obtenidos, el porcentaje de biclusters enriquecidos, la media de términos GO por biclúster y el tiempo medio para obtener cada biclúster. Generalmente el estudio del enriquecimiento de los biclusters se realiza respecto al término de BP de GO [77] [32]. Sin embargo en este estudio la significancia biológica también se estudia usando las otras dos ramas de GO: MF y CC. Haya que tener en cuenta que el número medio de términos GO por biclúster se establece teniendo en cuenta todos los biclusters, no sólo aquellos que están enriquecidos.

El algoritmo GoldBinch se ha comparado con una familia de algoritmo de biclustering considerados como clásicos. Estos algoritmos son en concreto ChCh [24], ISA [12], OPSM [11] y xMotifs [68], disponibles a través de la herramienta BiCAT [10], y que se suelen usar como marco de referencia en la literatura de biclustering [77]. La tabla 11.3 presenta el número de

biclusters, el tamaño medio de los mismos, la media del porcentaje de biclusters enriquecidos según las ramas de BP, MF y CC de GO, así como el número de términos GO por biclúster según BP, para ambos conjuntos de datos $GDS1116$ y $GDS2914$. Téngase en cuenta que tan sólo se presentan resultados de xMotfs para los datos de $GDS2914$ porque el algoritmo sólo se puede ejecutar para matrices que tengan menos de 64 columnas.

Las figuras 11.1 y 11.2 muestran el porcentaje de biclusters enriquecidos presentados en la tablas 11.2 y 11.3 para los dos conjuntos de datos, $GDS1116$ y $GDS2914$, respectivamente. Las barras de color negro representan BP, las grises MF y las blancas CC. Se debe tener en cuenta que 211, 212 y 221 representan $(M_1 = 2, M_2 = 1, M_3 = 1)$, $(M_1 = 2, M_2 = 1, M_3 = 2)$ y $(M_1 = 2, M_2 = 2, M_3 = 1)$ respectivamente.

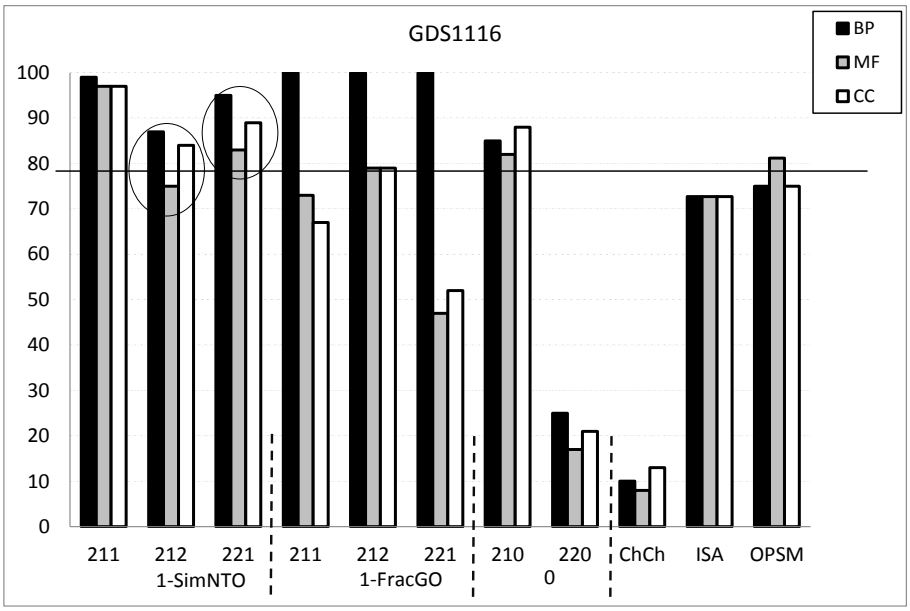

Figura 11.1: Porcentaje de biclusters enriquecidos para GDS1116.

Las figuras 11.3, 11.4 y 11.5 muestran el tamaño de los biclusters obtenidos para $GDS1116$ cuando se usan diferntes configuraciones de pesos propuestas. Cada punto representa un biclúster donde el número de genes está representado en el eje x y el número de las condiciones en el eje y. Esta misma información se puede observar en la tabla 11.4 que muestra la media, la varianza, los valores máximos y mínimos del número de genes y condiciones.

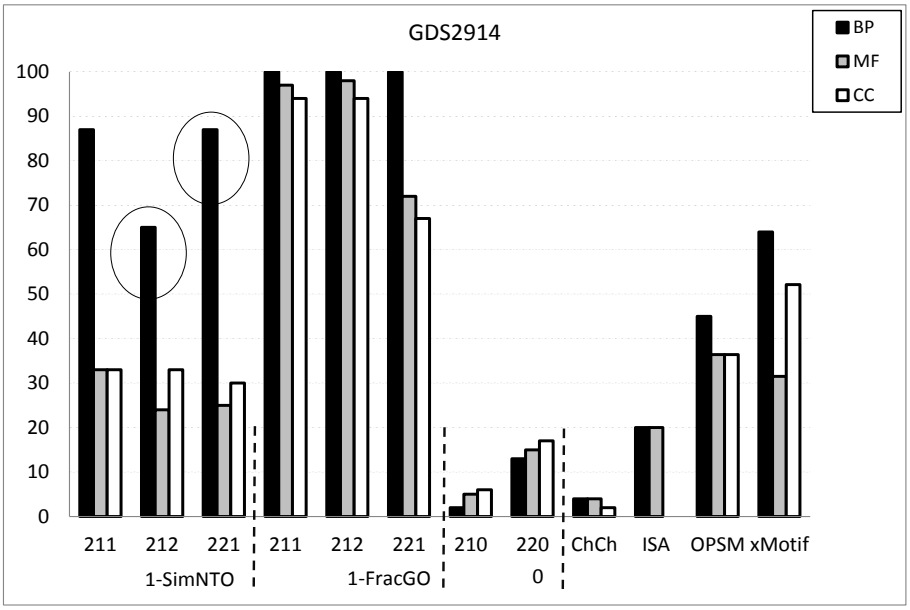

Figura 11.2: Porcentaje de biclusters enriquecidos para GDS2914.

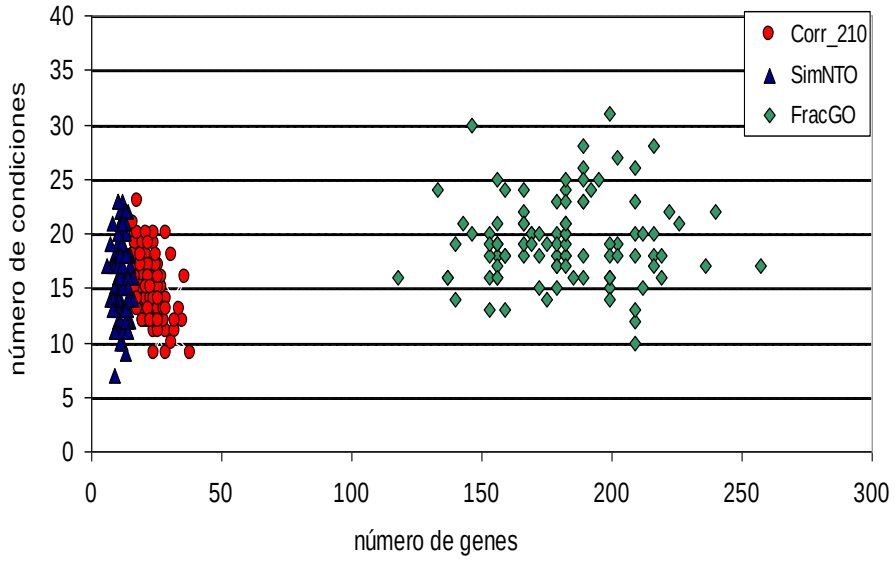

Figura 11.3: Tamaño de los biclusters obtenidos con GDS1116 cuando la correlación y la integración biológica tienen la misma importancia.

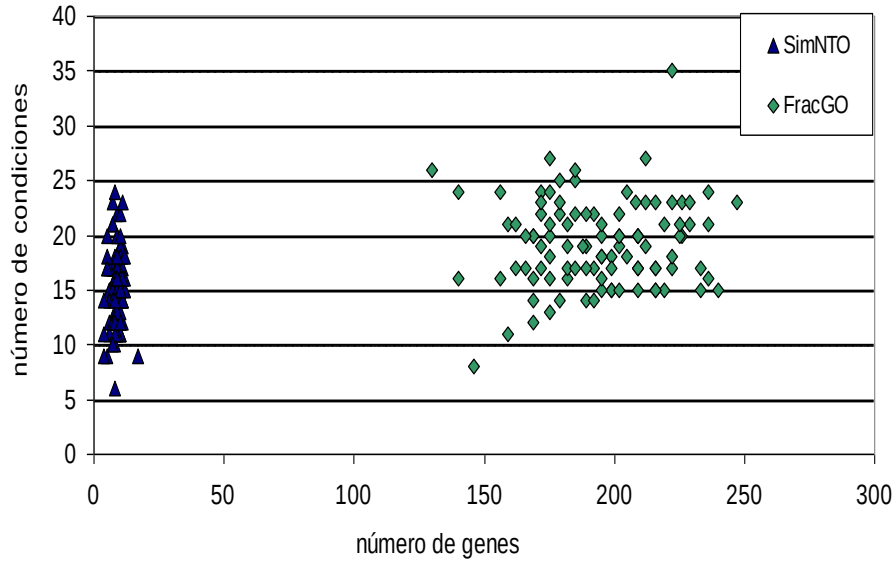

Figura 11.4: Tamaño de los biclusters obtenidos con GDS1116 cuando la integración biológica es más importante que la correlación.

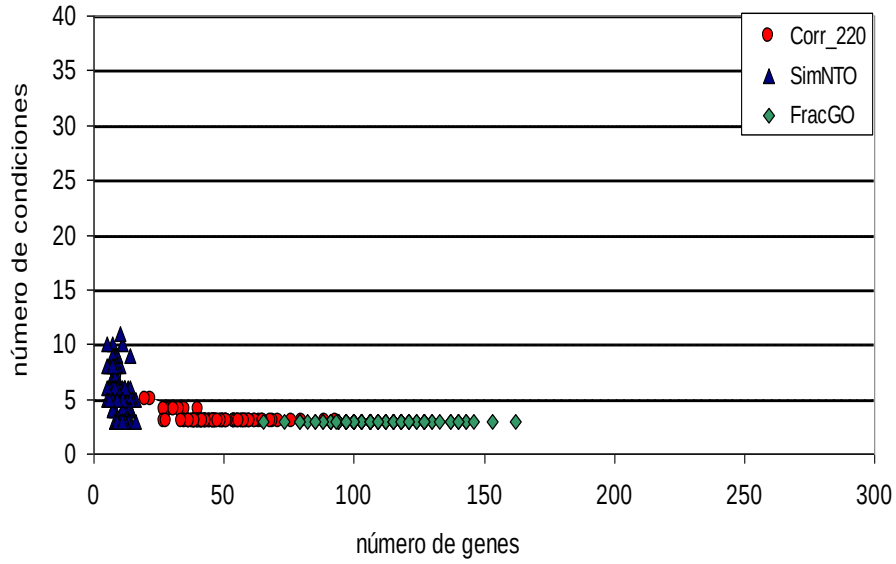

Figura 11.5: Tamaño de los biclusters obtenidos con GDS1116 cuando la correlación es más importante que la integración biológica.

	Configuración de la ejecución	Genes				Condiciones			
		media	varianza	max.	min.	media	varianza	max.	min.
Figura 11.3	SimNTO	11.6	2.1	16	6	15.6	3.4	23	7
	FracGO	181.2	25.9	257	118	19.4	3.9	31	10
	Corr-210	23.5	4.6	38	16	14.7	2.8	23	9
Figura 11.4	SimNTO	8.7	2.1	17	4	15.2	3.5	24	6
	FracGO	193.9	24.8	247	130	19.3	4.0	35	8
Figura 11.5	SimNTO	10.1	2.6	16	5	5.1	2.1	11	3
	FracGO	109.1	17.6	162	65	3	0	3	3
	Corr-220	46.8	13.8	93	9	3.1	0.6	7	3
Figura 11.8	OPSM	128.2	226.2	774	2	10.4	6.3	23	2
	CC	21.7	14.9	90	9	18.4	9.0	68	7
	ISA	50.7	38.2	97	4	6.4	2.3	9	2

Tabla 11.4: Media, varianza, máximo y mínimo del número de genes y de condiciones para biclusters representados en las figuras 11.3, 11.4, 11.5 y 11.8.

Las figuras 11.6 y 11.7 muestran el solapamiento entre los 100 biclusters obtenidos por la función objetivo basada en SimNTO, con la configuración que le da más importancia a la integración biológica que a la parte de la correlación, para los datos GDS1116. Cada elemento de la matriz de la figura es el porcentaje de solpamiento entre dos biclusters, que se define como la proporción de genes y condiciones que comparten ambos biclusters. El algoritmo GoldBinch no incorpora ningún mecanismo de control de solape entre los biclusters, sin embargo se puede observar en las figuras que todos los biclusters tienen un solapamiento por debajo del 30 %. Para otras configuraciones de la función objetivo los resultados son análogos.

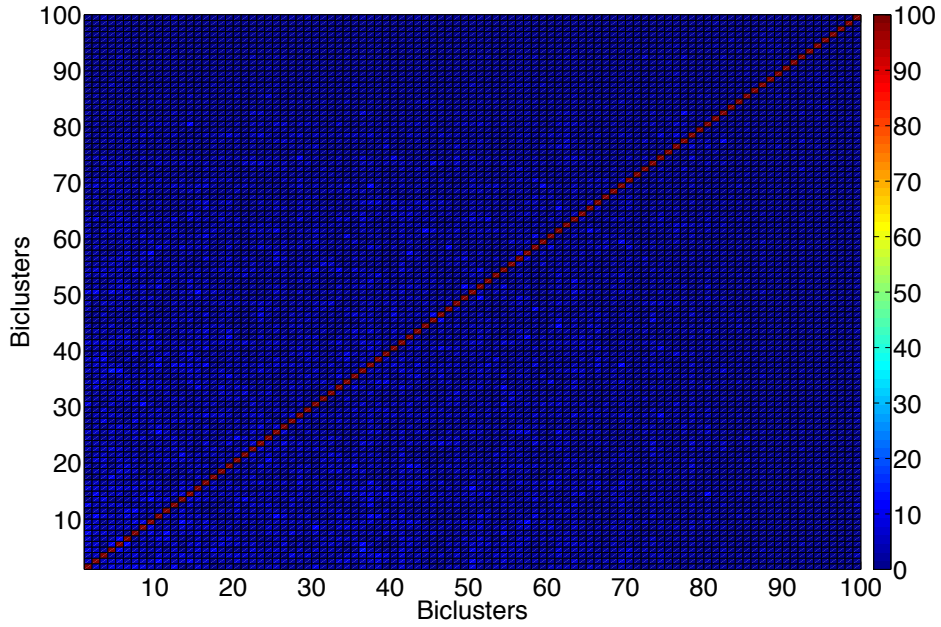

Figura 11.6: Porcentaje de solapamiento entre los biclusters obtenidos para GDS1116.

El algoritmo GoldBinch se ha comparado con una familia de algoritmo de biclustering considerados como clásicos. Estos algoritmos son en concreto ChCh [24], ISA [12], OPSM [11] y xMotifs [68], disponibles a través de la herramienta BiCAT [10], y se suelen usar como marco de referencia en la literatura de biclustering [77]. La tabla 11.3 presenta el número de biclusters, el tamaño medio de los mismos, la media del porcentaje de biclusters enriquecidos según las ramas de GO BP (*Biological Process*), MF (*Molecular Function*) y CC (*Celular Components*), así como el número de términos

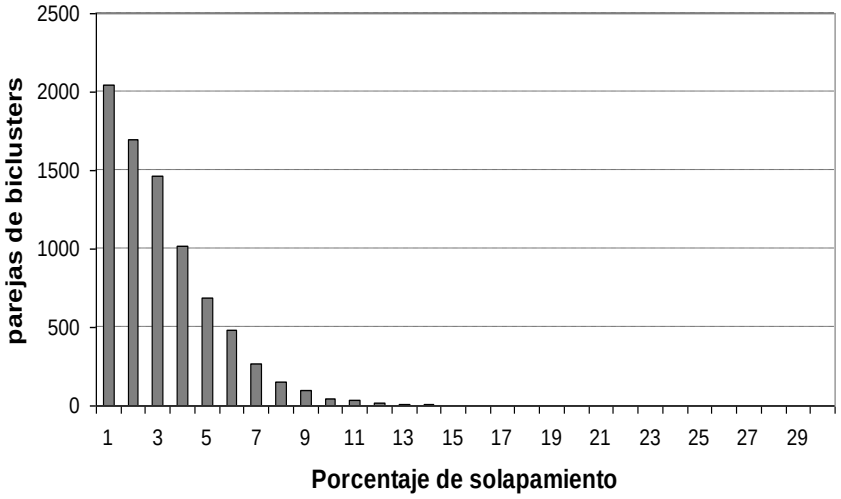

Figura 11.7: Histograma para el porcentaje de solapamiento entre los biclusters obtenidos para GDS1116.

GO por biclúster según BP, para ambos conjuntos de datos $GDS1116$ y $GDS2914$. Téngase en cuenta que tan sólo se presentan resultados de xMotfs para los datos de $GDS2914$ porque el algoritmo sólo se puede ejecutar para matrices que tengan menos de 64 columnas.

La figura 11.8 presenta el tamaño de los biclusters obtenidos por los algoritmos ChCh, ISA y OPSM para los datos $GDS1116$. La tabla 11.4 muestra de manera cuantitativa la misma información que esta figura.

La tabla 11.5 muestra los resultados obtenidos para $GDS1116$ y $GDS2914$ usando FracGO como medida de integración y como fuente los ficheros de anotaciones construidos con rutas KEGG y con Interpro. La información de estos ficheros de anotaciones, construidos sin usar la ontología GO, se muestra en la tabla 11.6. En concreto, se muestra el tamaño del fichero de anotaciones, el número de genes y el número medio de términos anotados por cada gen.

La tabla 11.7 muestra los resultados obtenidos utilizando otras medidas GO entre pares de genes distintas de SimNTO como medidas de integración. En concreto, las medidas SimGIC y SimUI. Estas medidas son medidas basadas en grafos que tienen en cuenta la estructura jerárquica de GO. SimGIC es una medida híbrida basada en el contenido de información, *information content* (IC), del grafo de GO mientras que, por otro lado, SimUI sólo se

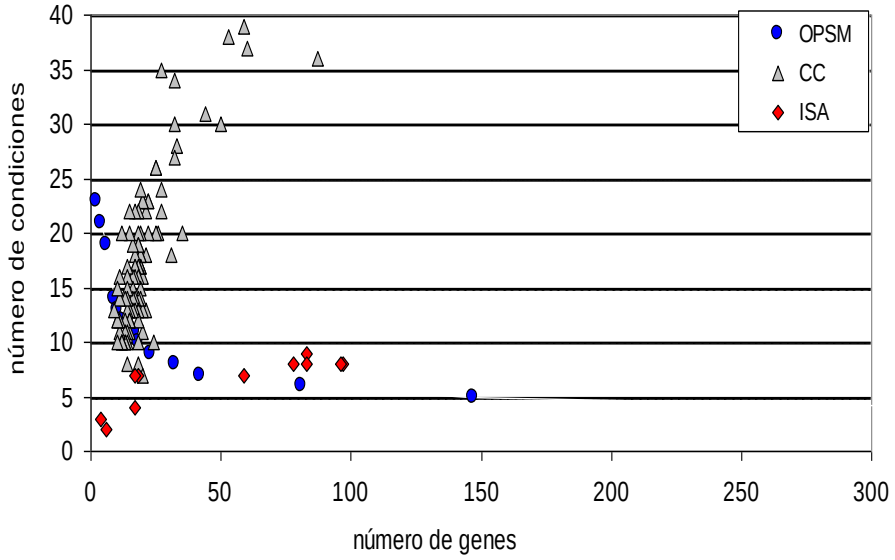

Figura 11.8: Tamaño de los biclusters obtenidos para CC, ISA y OPSM para GDS1116.

basa en contar términos en dicho grafo. El código fuente usado en los experimentos ha sido el código proporcionado en la referencia [21].

11.4. Discusión de los resultados

La integración de información biológica mejora los resultados de los algoritmos de biclustering respecto al porcentaje de biclusters enriquecidos. Se puede observar en la tabla 11.2 que los mejores biclusters son aquellos que se obtienen con las configuraciones FracGO y SimNTO de la función objetivo con BP (ver barras negras en las figuras 11.1 and 11.2). En concreto, para BP se obtiene un 100 % de biclusters enriquecidos con FracGO para ambos conjuntos de datos y para SimNTO 99 %, 87 % o 95 % con $GDS1116$ y, por otro lado, 87 %, 65 % o 87 % para $GDS2914$. La medida FracGO se basa en la proporción de genes que se encuentran en un mismo término GO enriquecido, por lo tanto si la validación se lleva a cabo usando la misma rama de GO respecto a la que se calcula FracGO, los resultados podrían estar sesgados. Por este motivo, y aunque generalmente las comparativas en los trabajos de biclustering se suelen establecer tan sólo estudiando el enriquecimiento respecto de BP, se ha incluido en la tabla 11.2 el estudio del enriquecimiento de biclusters respecto de las otras dos ramas de GO: MF y

CC. Téngase en cuenta que SimNTO se basa en un criterio distinto al del enriquecimiento de genes. Para los datos $GDS2914$, se puede observar que tanto SimNTO como FracGO mejoran los resultados en todos los casos a la configuración de la función objetivo cuando no hay integración biológica ($f_3 = 0$). Para los datos de $GDS1116$, se puede observar en la figura 11.1 que la configuración de parámetros 211 para SimNTO mejora a las configuraciones 210 y 220. Por otro lado, para FracGO los resultados para 212 son un poco mejor que los de 210 para MF y CC pero están en el mismo rango de valores (véase la línea horizontal al 79 % en la figura).

Por otro lado, el porcentaje de biclusters enriquecidos usando tanto FracGO como SimNTO mejora los resultados de los algoritmos clásicos de biclustering. Para $GDS1116$, se puede observar en la figura 11.1 que SimNTO obtiene mejores resultados que ChCh, ISA y OPSM. Por otro lado, todas las configuraciones de FracGO mejoran a ChCh y para 212 los resultados están el mismo rango de valores que ISA y OPSM para MF y CC (véase la línea horizontal en la figura al 79 %). Para $GDS2914$, la figura 11.2 muestra claramente que SimNTO mejora a ChCh, ISA, OPSM y xMotifs para BP y FracGO mejora a todos los algoritmos para BP, MF y CC. Para OPSM se obtiene un valor alto para el número medio de términos GO por biclúster (20.06 y 28.64 para $GDS1116$ y $GDS2914$ respectivamente). Este hecho se debe a que los biclusters de OPSM están compuestos por un número muy alto de genes.

El porcentaje de biclusters enriquecidos para SimNTO es mayor cuando se le da énfasis al término de la correlación (configuración 221) que cuando se enfatiza el término de la integración (configuración 212) (véase los círculos en las figuras 11.1 y 11.2). Este hecho se debe a que la medida SimNTO no se basa en la idea del enriquecimiento de genes sino en el solapamiento de términos GO entre los términos asociados a cada gen. La mejor configuración para SimNTO es 211, es decir, la misma importancia para M_2 que para M_3. Por lo tanto, se puede concluir que la mejor opción para la configuración de la función objetivo es buscar un equilibrio entre el término relativo a la correlación y el relativo a la integración biológica.

Las comparativas basadas en el enriquecimiento de los genes pasan por alto las condiciones que componen los biclusters. Estas comparativas deben ser ampliadas observando los tamaños de los biclusters para descartar situaciones en las que los resultados obtenidos sean triviales y tienen información relevante. Por ejemplo, un biclúster compuesto por un número alto de genes, que muestren enriquecimiento GO, pero que sólo esté compuesto por dos condiciones, no aporta información relevante de un comportamiento subyacente, tan sólo muestra que ese conjunto de genes se expresan en dos

condiciones. Se puede observar teniendo en cuenta las figuras 11.3 and 11.4 y la tabla 11.2, que el tamaño de los biclusters es igual cuando el término de integración es igual o más relevante que el término de la correlación (parámetros $M_2 = M_3 = 1$ o $M_2 = 1$ y $M_3 = 2$, respectivamente). Sin embargo, en la figura 11.5 se puede observar que el tamaño decrece cuando se prima el termino de la correlación (parámetros $M_2 = 2$ y $M_3 = 1$). En particular, tan sólo 5 condiciones para SimNTO como valor medio y colapsa a 3 para FracGO y el caso en el que no hay integración. Por otro lado, se debe observar que los biclusters de FracGO tienen en general un número más alto de genes que los de SimNTO.

A pesar de que el porcentaje de biclusters enriquecidos es del 100 % con FracGO para BP, el valor medio de términos GO por biclúster es de tan sólo 1 (véase tabla 11.2). La medida FracGO busca genes que compartan un mismo término GO, por esta razón FracGO encuentra biclusters con tan sólo un término GO que probablemente será un término genérico en los niveles superiores de GO. Por el contrario, el número medio de atributos GO para SimNTO es de 5,74, 3,67 y 4,5 para $GDS1116$ y, por otro lado, de 5,45, 2,94 y 5,39 para $GDS2914$. Por lo tanto, los biclusters obtenidos con SimNTO se componen por grupos de genes relacionados con un número mayor de términos GO que los obtenidos por FracGO, lo que implica que aportan información biológica más relevante.

En la tabla 11.2 se puede observar que el coste computacional de FracGO es más alto que el de SimNTO. Este hecho es lógico teniendo en cuenta su definición. Para cada conjunto de genes se tiene que calcular por cada término GO del fichero de anotaciones el p-value asociado. Además, se debe evaluar multitud de soluciones/biclusters durante el proceso de búsqueda (generación de la población inicial, mejora, estabilidad del conjunto de referencia, etc), es decir, cientos o incluso miles de posibles soluciones. Hay que tener en cuenta que cuanto menos términos haya en los ficheros de anotaciones más rápido será el tiempo de ejecución para la medida FracGO.

Los ficheros de anotaciones se pueden generar de multitud de forma, sin embargo la medida SimNTO requiere una serie de restricciones. En concreto, esta medida requiere que el fichero de anotaciones sea una ontología, como en el caso de GO, y que el fichero de anotaciones tenga en cuenta dicha ontología al completo y no sólo una parte. FracGO en cambio acepta cualquier tipo de fichero de anotaciones sea fruto o no de una ontología o, en el caso de que lo sea, acepta informaciones parciales de la misma, es decir, la información parcial entre dos niveles dados de esa ontología. La tabla 11.5 muestra un ejemplo en el que se puede utilizar la medida FracGO pero no SimNTO. En la tabla 11.6 se puede observar que el número de genes anotados en

GO para ambos conjuntos de datos es sensiblemente superior a los anotados en KEGG o InterPro. Se puede observar en las tablas 11.2 y 11.5 cómo influyen el número de genes anotados y el número de anotaciones por gen en los resultados obtenidos por FracGO. Cuando menor es el número de genes anotados, es decir menos información se tiene, peores son los resultados.

Datos	Anotaciones funcionales de	Parámetros (M_1, M_2, M_3)	Tamaño	Biclusters enriquecidos (%)			Términos GO por biclus. (BP)	Tiempo (segundos)
				BP	MF	CC		
GDS1116	KEGG	(2, 1, 1)	(73.4 × 18.1)	13	26	34	0.20	165.4
		(2, 1, 2)	(79.6 × 17.8)	13	26	27	0.21	175.6
		(2, 2, 1)	(49.8 × 3.1)	7	11	8	0.07	145.3
	InterPro	(2, 1, 1)	(14.4 × 16.2)	71	80	71	1.81	3613.3
		(2, 1, 2)	(30.4 × 17.0)	60	78	52	3.32	2380.3
		(2, 2, 1)	(43.3 × 3.3)	24	21	22	0.52	2243.8
GDS2914	KEGG	(2, 1, 1)	(69.4 × 13.3)	18	49	30	0.33	116.0
		(2, 1, 2)	(77.1 × 14.2)	24	46	23	0.48	126.6
		(2, 2, 1)	(42.6 × 3.0)	12	15	11	0.12	100.6
	InterPro	(2, 1, 1)	(17.5 × 9.1)	20	19	6	0.41	1795.3
		(2, 1, 2)	(15.5 × 9.6)	26	23	15	0.99	1825.8
		(2, 2, 1)	(39.2 × 3.1)	25	18	16	0.33	1850.6

Tabla 11.5: Resultados obtenidos con la la medida FracGO usando rutas KEGG e Interpro con los conjuntos de datos GDS1116 y GDS2914. Cada columna muestra una ejecución que obtiene 100 biclusters.

Datos	Tamaño	Fuente anotación funcional	Número de genes	Número de términos	Media term. por gen
GDS1116	(882 × 131)	GO (BP domains)	632	245	10.6
		KEGG pathways	239	53	1.8
		InterPro	575	699	2.5
		GO (MF domains)	634	135	5.5
		GO (CC domains)	703	44	3.1
GDS2914	(975 × 36)	GO (BP domains)	658	256	10.1
		KEGG pathways	190	65	1.9
		InterPro	556	653	2.2
		GO (MF domains)	615	127	4.9
		GO (CC domains)	740	46	3.2

Tabla 11.6: Anotaciones funcionales para GO, rutas KEEG e InterPro para los conjuntos de datos GDS1116 y GDS2914.

La tabla 11.7 muestra los resultados de otras medidas GO basadas en pares de genes distintas a SimNTO. SimGIC y SimUI integran la información biológica en el algoritmo usando la misma idea que SimNTO, la similitud GO entre pares de genes, pero su funcionamiento difiere del de SimNTO. Se puede observar que los biclusters obtenidos por SimNTO tienen un número de términos GO por biclusters más alto que SimGIC y SimUI. Por otro lado, SimNTO obtiene un porcentaje más alto de biclusters enriquecidos que SimGIC y SimUI. Se puede por tanto concluir que SimNTO obtiene mejores resultados que SimGIC y SimUI. Se debe tener en cuenta además que el coste computacional de SimNTO es menor debido a su naturaleza, ya que la estructura de GO la tiene en cuenta a través del fichero de anotaciones sin tener que consultar un fichero adicional con la misma.

Datos	Función objetivo		Tamaño	Biclusters enriquecidos (%)			Términos GO por biclus.	Tiempo
	Medida f_3	Parámetros (M_1, M_2, M_3)		BP	MF	CC	(BP)	(segundos)
GDS1116	1-SimGIC	(2, 1, 1)	(11.2 × 15.6)	100	100	99	4.5	1409.4
		(2, 1, 2)	(9.0 × 20.0)	100	100	100	2.0	419.0
		(2, 2, 1)	(19.6 × 4.2)	51	48	53	1.1	887.1
	1-SimUI	(2, 1, 1)	(10.6 × 15.6)	98	97	98	4.1	584.3
		(2, 1, 2)	(8.71 × 16.9)	94	90	93	3.1	275.5
		(2, 2, 1)	(10.2 × 4.6)	62	66	65	1.5	509.5
GDS2914	1-SimGIC	(2, 1, 1)	(23.0 × 9.1)	2	4	7	0.03	1572.2
		(2, 1, 2)	(14.8 × 8.8.x)	36	17	18	1.6	993.1
		(2, 2, 1)	(34.4 × 3.1)	10	8	9	0.1	993.1
	1-SimUI	(2, 1, 1)	(17.0 × 9.2)	11	5	11	0.2	348.2
		(2, 1, 2)	(8.4 × 9.8)	64	48	47	1.7	348.2
		(2, 2, 1)	(18.7 × 3.1)	21	8	19	0.6	536.4

Tabla 11.7: Resultados obtenidos usando otras medidas de GO basadas en pares de genes para la integración de información biológica. Cada columna representa una ejecución que obtiene 100 biclusters.

En resumen, y teniendo en cuenta el coste computacional para la medida FracGO, la opción de SimNTO es la más adecuada para la integración de información cuando se usa la ontología GO teniendo en cuenta toda su estructura. De hecho los biclusters que se obtienen con FracGO en estos casos estarán compuestos por genes que comparten tan sólo un término GO muy general en la ontología. Sin embargo, si la información de la ontología es tan sólo parcial, de manera que se tenga en cuenta tan sólo la información entre varios subniveles intermedios de la misma, o se utiliza como información ficheros de anotaciones generados por otras fuentes de información como KEGG o InterPro, la única opción en estos casos es FracGO.

11.5. Evaluación biológica cualitativa

El enriquecimiento de genes es el criterio comúnmente utilizado en la literatura de biclustering como marco de comparación entre algortimos ([77], [32]). En la sección anterior hemos analizado los resultados obtenidos por el algoritmo utilizando las configuraciones basadas en SimNTO y FracGO. Se ha visto que los biclusters obtenidos con FracGO están formados por grupos de genes que tan sólo son significativos en un término GO (véase la tabla 11.2) que, si es uno de los términos más altos en la jerarquía de GO, puede aportar información no relevante. En esta sección se trata de analizar no ya que el algoritmo obtenga información biológica sino de qué tipo de información se trata y si es relevante o no. A continuación vamos a tratar de confirmar la hipótesis de que los biclusters de FracGO capturan información no relevante, por ser información de términos GO muy generales, mientras que SimNTO si captura información relevante.

A continuación se realiza un evaluación biológica de tipo cualitativo usando las ideas usadas en la referencia [73]. Este estudio se basa en la significancia estadística de un grupo de genes en una determinada ruta usando Reactome [45] como referencia. Se han estudiado los cinco primeros biclusters según el valor de su función objetivo para los datos $GDS1116$. En concreto, los cinco primeros biclusters para las configuraciones 211 y 212 para SimNTO y FracGO, así como los cinco primeros biclusters para la configuración 210 para Corr. Téngase en cuenta que los biclusters de las configuraciones 221, para FracGO y SimNTO, y los biclusters para la configuración 220 para Corr tienen un número muy bajo de condiciones y no son interesantes de estudiar. Los cinco biclusters de SimNTO para 211 y 212 presentan mapeo de rutas, es decir, están compuestos por genes presentes en rutas metabólicas representadas en Reactome. En cambio, tan sólo dos biclusters de los cinco de Corr, para 220, presentan mapeo de rutas y no se encuentra ninguna in-

formación para FracGO ni para la configuración 211 ni para 212. Las tablas 11.8 y 11.9 muestran el análisis de este mapeo de rutas usando Reactome para los biclusters tercero y quinto de SimNTO con las configuraciones 211 y 212 respectivamente. La tabla 11.10 muestra la información del cuarto bicluster de Corr para 220. Cada columna de la tabla muestra el identificador de la ruta o *pathway*. En el capítulo de datos suplementarios se suministra la información de todas las tablas generadas para todos los biclusters.

Identificador pathway	Nombre pathway	FDR
247749	Eukaryotic Translation Elongation	0,001
260795	Translation	0,001
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0,001
217188	Formation of a pool of free 40S subunits	0,001
257612	Eukaryotic Translation Termination	0,001
257951	Peptide chain elongation	0,001
188965	SRP-dependent cotranslational protein targeting to membrane	0,001
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0,001
189048	Nonsense-Mediated Decay (NMD)	0,001
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0,001
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0,002
251703	Cap-dependent Translation Initiation	0,002
230274	Eukaryotic Translation Initiation	0,002
257608	Formation of the ternary complex, and subsequently, the 43S complex	0,040
233365	Metabolism of proteins	0,040
248935	Ribosomal scanning and start codon recognition	0,050

Tabla 11.8: Resultado del análisis usando Reactome para el biclúster 3 obtenido con SimNTO y con la configuración 211.

Identificador pathway	Nombre pathway	FDR
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0,014
217188	Formation of a pool of free 40S subunits	0,014
257951	Peptide chain elongation	0,014
188965	SRP-dependent cotranslational protein targeting to membrane	0,014
247749	Eukaryotic Translation Elongation	0,014
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0,014
189048	Nonsense-Mediated Decay (NMD)	0,014
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0,014
188483	S6K1 signalling	0,014
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0,014
251703	Cap-dependent Translation Initiation	0,019
230274	Eukaryotic Translation Initiation	0,020
260795	Translation	0,032
188482	mTORC1-mediated signalling	0,036
188481	S6K1-mediated signalling	0,036

Tabla 11.9: Resultado del análisis usando Reactome para el biclúster 5 obtenido con SimNTO y con la configuración 212.

Identificador pathway	Nombre pathway	FDR
231079	Gene Expression	0,002
233365	Metabolism of proteins	0,010
188483	S6K1 signalling	0,012
188482	mTORC1-mediated signalling	0,039
188481	S6K1-mediated signalling	0,039

Tabla 11.10: Resultado del análisis usando Reactome para el biclúster 4 obtenido con la medida Corr y con la configuración 210.

Las rutas encontradas para cuatro de los cinco biclusters obtenidos con SimNTO para la configuración 112 son: *Eukaryotic Translation Initiation*, *Cap-dependent Translation Initiation* y *L13a-mediated translational silencing of Ceruloplasmin expression* que están asociadas a proteínas del ribosoma (RPL9A ,RPL9B y RPL8A), además de asociadas a procesos de ensamblado del propio ribosma (27SA3 pre-rRNA y 27SB pre-rRNA). Análogamente, las rutas de cuatro de los cinco biclusters obtenidos con SimNTO para la configuración 212 son: *S6K1 signalling*, *S6K1-mediated signalling* y *mTORC1-mediated signalling*. En el caso de los biclusters obtenido con la medida Corr para la configuración 220 tan sólo se encuentran rutas para dos de ellos. Al contrario de lo que sucede con los resultados para SimNTO, los resultados para la medida Corr no pueden considerarse relevantes. En el análisis funcional, los biclusters de SimNTO muestran más genes mapeados en rutas de Reactome que los de Corr. Este hecho apoya la hipótesis de que la integración de la información biológica basado en la medida SimNTO mejora los resultados del caso en el que no hay integración de información.

Por otro lado, la información más interesante es que Reactome no muestra información para ningún biclúster de FracGO. La razón es que estos biclusters agrupan genes que se asocian con términos GO muy generales y que no aportan información relevante. La causa es que el fichero de anotaciones de GO usado en los experimentos se ha construido teniendo en cuenta toda la jerarquía de GO, es decir, propagando la información de los términos hasta la raiz. Se han analizado estos biclusters de FracGO usando *Gene Term Linker* [35] y *Revigo* [84] para confirmar la hipótesis de trabajo: que FracGO encuentra biclusters enriquecidos pero estos no aportan información interesante. En primer lugar *Gene Term Linker* filtra aquellos términos GO irrelevantes identificando metagrupos de genes con un significado biológico coherente. En segundo lugar, *Revigo* resume en una lista todos los términos

GO agrupándolos según su similitud y eliminando la información redundante. La figura 11.10 muestra los términos GO resultantes del primer biclúster de FracGO para 211. Todos los términos GO se agrupan en dos grupos: *Metabolism* y *Protein Ubiquitation*. Después de aplicar el análisis combinado usando *Gene Term Linker* y *Revigo*, se puede comprobar que todos los términos GO están agrupados en términos muy generales. Se confirma la hipótesis de trabajo que nos decía que tal y como se han proporcionado los ficheros de anotaciones en los experimentos, capturando la información de GO hasta la raiz de la ontología, los biclusters obtenidos por FracGO capturan información biológica pero dicha información no es relevante. Por este motivo no se encontraban rutas asociadas con Reactome. Se puede consultar la información suplementaria para ver todas las figuras y tablas generadas en este estudio.

Figura 11.9: Grupo de términos GO obtenidos con Revigo para el bicluster 1 para la medida FracGO y la configuración 211. Sólamente se agrupan dos términos: protein-ubiquitination y metabolism.

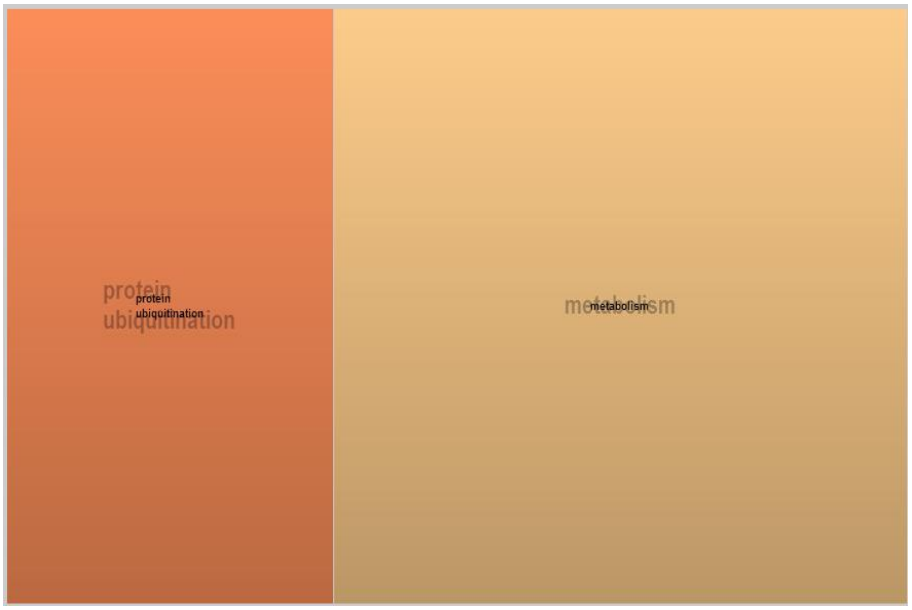

Figura 11.10: Grupo de términos GO obtenidos con Revigo para el bicluster 1 para la medida FracGO y la configuración 211. Sólamente se agrupan dos términos: protein-ubiquitination y metabolism.

Parte V

Conclusiones

Capítulo 12

Conclusiones y trabajos futuros

En mi opinión, la característica más importante de los buenos profesores es que se colocan en el lugar del alumno. No se trata simplemente de impartir lecciones claras y precisas y de corregir exámenes; el objetivo principal es ayudar al estudiante a entender la materia.
Cartas a una joven matemática. Ian Stewart (pág. 172)

12.1. Conclusiones

Los datos de expresión génica, y su particular naturaleza e importancia, motivan no sólo el desarrollo de nuevas técnicas sino la formulación de nuevos problemas como el problema del biclustering. El principal objetivo de la tesis expuesta ha sido la elaboración de un algoritmo de biclustering que permita el estudio de distintos criterios de búsqueda. La motivación inicial del trabajo desarrollado fue el artículo [1] donde se demostraba que el residuo cuadrático medio, o *Mean Squared Residue* (MSR), no detectaba cierto tipo de patrones. Estos patrones, patrones de escalado, son de gran importancia en el contexto biológico ya que reflejan genes que se expresan simultáneamente aunque con distinta intensidad. Gran parte de los algoritmos de biclustering se basan en el residuo (MSR) como criterio de búsqueda por lo que no detectan estos patrones.

Esta motivación inicial, de no perder de vista la naturaleza biológica del problema abordado, ha propiciado no sólo el estudio en profundidad de los datos estudiados sino también del contexto de los mismos. La cantidad de información biológica almacenada en repositorios de datos públicos como

Gene Ontology (GO) o *Kyoto Encyclopedia of Genes and Genomes* (KEGG) generalmente se usa en los artículos de biclustering tan sólo para validar los resultados, sin embargo en Bioinformática una de las tendencias actuales es la integración de información de datos que provienen de distintas fuentes.

Teniendo en cuenta esta motivaciones se pueden resumir básicamente las contribuciones presentadas de la siguiente forma:

- Se ha presentado un algoritmo de biclustering basado en un esquema de búsqueda dispersa que hemos denominado SSCorr. Dicho algoritmo optimiza una función cuyo criterio de calidad entre genes se basa en las correlaciones lineales entre ellos. De esta forma se aporta un criterio de búsqueda distinto del residuo. SScorr tiene interés pues incluye la búsqueda dispersa a la familia de algoritmos de biclustering basados en computación evolutiva, o metaheurísticas en general. Aunque la mayoría de los algoritmos de biclustering basados en computación evolutiva se basan en el residuo, existen algoritmos, basados en otras técnicas, que también usan la correlación como criterio de búsqueda y que surgieron de manera simultánea.

- Los resultados de SScorr se han validado usando la experimentación habitual de otros artículos de biclustering, además de usar una evaluación biológica de los resultados.

- Se ha mejorado el algoritmo anterior con una versión que hemos denominado BISS. Esta nueva propuesta captura los patrones anteriores junto con patrones de activación-inhibición que no capturaba la propuesta previa. La motivación biológica hizo marcar el énfasis en ese tipo de patrones que reflejan un comportamiento relevante desde un punto de vista biológico.

- La experimentación realizada con BISS ha sido especialmente elaborada y constituye, desde nuestro punto de vista, un avance respecto a la realizada con SSCorr. Además de realizar una comparativa con los algoritmos clásicos de biclustering, se realiza una comparación con otros dos algoritmos de biclustering que también se basan en la correlación como criterio de búsqueda. Así mismo, se le saca partido a la información de GO utilizando sus tres ramas para validar, así como se estudia con detalle las posibles restricciones en el concepto de enriquecimiento de un grupo de genes.

- Se ha presentado un algoritmo de biclustering, basado también en una búsqueda dispersa, que integra información biológica como crite-

rio de búsqueda. Se ha denominado a dicho algoritmo GoldBinch. La información almacenada en repositorios como GO o KEEG se integra dentro del criterio de búsqueda de los biclusters a través de los ficheros de anotaciones. De esta forma se introduce un sesgo en la búsqueda de tal manera que se mejoran los resultados obtenidos por el algoritmo de biclustering. Se ha estudiado varias posibilidades de integración de información a través de varias medidas de similitud entre genes como parte de la función objetivo.

- La experimentación realizada con GoldBinch ha tenido un doble objetivo. Por un lado verificar si la integración de información biológica mejora o no los resultados obtenidos del proceso de biclustering. Por otro lado, estudiar las diferencias entre las distintas posibilidades de integración de información. Para ello, primero se ha realizado una comparativa experimental basada en el porcentaje de biclusters enriquecidos, considerada estándar en biclustering. Dicha comparativa se ha realizado usando las tres ramas de GO. En un segundo paso, se ha profundizado en el análisis biológico de los resultados, primero estudiando los atributos GO asociados y, con posterioridad, realizando un estudio cualitativo de algunos biclusters. De esta forma se muestra claramente las diferencias entre las distintas posibilidades de integración de información.

Los algoritmos de biclustering son algoritmos altamente especializados para un tipo concreto de datos, los datos de expresión génica. Surgen debido a las características de los mismos: necesidad de solapamiento de los resultados, búsqueda de comportamientos locales, miles de genes frente a decenas de condiciones, etc. Resulta de vital importancia entender el contexto biológico de estos algoritmos. Dicho contexto es importante, no sólo para poder validar los resultados obtenidos, sino para poder diseñar algoritmos más eficientes. Así por ejemplo, es necesario comprender cuáles son los patrones que se deben buscar y por qué o, por ejemplo, la posibilidad de aprovechar la gran cantidad de información existente y que nos permite condicionar, y mejorar, los procesos de búsqueda. Por último, conocer el contexto del problema nos permite no sólo resolver de manera más eficiente la preguntas planteadas sino formular nuevas hipótesis de trabajo que motiven nuevas preguntas.

12.2. Futuras líneas de investigación

Las líneas de trabajo futuro que surgen a partir del trabajo presentado en esta tesis son de varios tipos.

Como continuación natural del trabajo presentado surgen las siguientes líneas:

- Realizar mejoras en el esquema de búsqueda dispersa. Se puede estudiar un nueva codificación para las soluciones y versiones más elaboradas para los métodos de la combinación, de generación de soluciones y, sobre todo, de la mejora.
- Estudiar el efecto de la integración de información utilizando otros algoritmos clásicos en biclustering.
- Realizar una comparativa entre algoritmos de biclustering más elaborada desde un punto de vista biológico. Las comparativas actuales no tienen en cuenta las particularidades de GO donde, por ejemplo, se tiene en cuenta que un grupo de genes esté o no en una función biológica pero no qué tipo de función o si lo está en varias a la vez.

El contexto del trabajo realizado es la Bioinformática y el descubrimiento de biomarcadores. Una visión de este campo permite ver claramente posibilidades de extensión y adaptación del trabajo realizado en otras temáticas:

- Los datos de microRNA son de vital importancia para entender el proceso de la regulación génica. Los conjuntos de datos generados en este contexto plantean una serie de necesidades que los algoritmos de biclustering pueden ayudar a resolver. Nos planteamos una adaptación del trabajo realizado en este nuevo contexto, para ello se seguirá muy de cerca las referencias [73, 33].
- En el contexto de las redes de genes/proteínas el solapamiento de los resultados es de vital importancia parta descubrir factores de transcripción o genes que sirvan de activadores o inhibidores de los procesos. En este contexto los algoritmos de clustering de redes se están adaptando para admitir el solape de los resultados. Nos planteamos el desarrollo de un nuevo algoritmo utilizando las ideas de los algoritmos de biclustering y la experiencia adquirida en los métodos de validación de resultados. Las referencias fundamentales que motivan estas ideas son [71, 80].

Parte VI

Apéndices

Datos suplementarios

Sobre todo en la ciencia biomédica existe una preocupante tendencia a publicar resultados irreproducibles fuera del laboratorio donde se obtuvieron, razón por la cual ha empezado a exigirse una mayor transparencia sobre las condiciones de los experimentos. A Francis Collins, director de los Institutos Nacionales de Salud, le preocupa la "salsa secreta"-procedimientos especializados, software customizado, ingredientes originales-que los investigadores no comparten con sus colegas.
National Geographic marzo 2015 pág. 63. La era de la incredulidad, Joel Achenbach.

Resumen

A continuación se proporciona el documento con los datos suplementarios asociados a la experimentación presentada en el capítulo XI. Dicha experimentación se asocia con la propuesta GoldBinch presentada en el capítulo VIII. Las tablas y figuras aportadas en este apéndice permiten observar toda la documentación generada en el estudio experimental. Estos datos suplementarios se encuentran disponibles en la dirección `http://www.lsi.us.es/~janepo/GoldBinch.html` como material adicional del artículo *"Integrating biological knowledge based on functional annotations for biclustering of gene expression data"* publicado en la revista *Computer Methods and Programs in Biomedicine* (ISSN 0169-2607).

En dicha dirección se puede encontrar también, de forma adicional, los datos, tanto en crudo como procesados, de los conjuntos de datos utilizados como de los resultados generados por estos. Así mismo, se proporciona el ejecutable del algoritmo junto con la información necesaria para la reproducibilidad de la experimentación.

Integrating biological knowledge based on functional annotations for biclustering of gene expression data (Supplementary)

1. Supplementary Information to Biological Evaluation

This studio is based on the functional analysis by considering pathway mapping and statistical significance of gene enrichment in pathways. The resource used for mapping gene in pathways is Reactome. The first five biclusters according to the fitness function value for runs with GDS1116 dataset have been studied. Concretely, the five first biclusters for SimNTO and FracGO with 211 and 212 configuration parameters and for Corr with 210. Note that biclusters for SimNTO and FracGO 221 and for Corr 220 are composed of a low number of conditions and they are less interesting to study. The five SimNTO 211 and 212 biclusters present pathways mapping, only two biclusters for Corr 220 present information but there is not any information for FracGO 211 and 212 biclusters. Tables 1, 2, 3, 4 and 5 show the mapping analysis reported by Reactome for SimNTO 211 biclusters, Tables 6, 7, 8, 9 and 10 for SimNTO 212 biclusters and Tables 11 and 12 for Corr 210 biclusters.

Gene Term Linker and Revigo tools have been used to show that FracGO biclusters present GO information very general. Firstly, the first tool filters irrelevant GO information by identifying metagroups of genes with coherent biological significance (Tables 13, 14, 15 and 16). Secondly, Revigo summarizes a list of these GO terms by finding representative subsets of terms using a clustering procedure that removes redundant terms. Figures 1, 2, 3, 4 and 5 show the significant GO term enriched for the five FracGO 211 reported biclusters. All reported enriched GO terms are clustered in general GO terms.

1.1. Gene Term Linker Information Tables

Tables 13, 14, 15, 16 and 17 show the results reported by Gene Term Linker. For each bicluster, the biggest metagroup (first row in the tables) is used as input data for Revigo tool (annotation terms in the last column). In the first column, Tables show the Silhouette Width coefficient, which measures how appropriately genes have been clustered in the metagroup. The second column represents the genes included in each metagroup. The third column represents the number of annotated genes in the input bicluster together with the total number of genes in the input list (bicluster). The fourth column represents the number of annotated genes in the reference list together with the total number of genes in the reference list (genes from the organism). Next column shows the p-value calculated using the Hypergeometric distribution and corrected for multiple testing with FDR method. Finally, the last column presents the annotation of terms obtained from the different selected biological annotation resources (Gene Ontology).

Pathway identifier	Pathway name	FDR
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0.001
217188	Formation of a pool of free 40S subunits	0.001
257612	Eukaryotic Translation Termination	0.001
257951	Peptide chain elongation	0.001
188965	SRP-dependent cotranslational protein targeting to membrane	0.001
247749	Eukaryotic Translation Elongation	0.001
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.001
189048	Nonsense-Mediated Decay (NMD)	0.001
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0.001
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.001
251703	Cap-dependent Translation Initiation	0.002
230274	Eukaryotic Translation Initiation	0.002
260795	Translation	0.002
233365	Metabolism of proteins	0.034

Table 1: Mapping analysis provided by Reactome for bicluster 1 from the SimNTO measure and 211 configuration.

Pathway identifier	Pathway name	FDR
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.001
251703	Cap-dependent Translation Initiation	0.002
230274	Eukaryotic Translation Initiation	0.002
260795	Translation	0.003

Table 2: Mapping analysis provided by Reactome for bicluster 2 from the SimNTO measure and 211 configuration.

Pathway identifier	Pathway name	FDR
247749	Eukaryotic Translation Elongation	0,001
260795	Translation	0,001
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0.001
217188	Formation of a pool of free 40S subunits	0.001
257612	Eukaryotic Translation Termination	0.001
257951	Peptide chain elongation	0.001
188965	SRP-dependent cotranslational protein targeting to membrane	0.001
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.001
189048	Nonsense-Mediated Decay (NMD)	0.001
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0.001
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.002
251703	Cap-dependent Translation Initiation	0.002
230274	Eukaryotic Translation Initiation	0.002
257608	Formation of the ternary complex, and subsequently, the 43S complex	0.040
233365	Metabolism of proteins	0.040
248935	Ribosomal scanning and start codon recognition	0.050

Table 3: Mapping analysis provided by Reactome for bicluster 3 from the SimNTO measure and 211 configuration.

Pathway identifier	Pathway name	FDR
233365	Metabolism of proteins	0,026

Table 4: Mapping analysis provided by Reactome for bicluster 4 from the SimNTO measure and 211 configuration.

Pathway identifier	Pathway name	FDR
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0.003
217188	Formation of a pool of free 40S subunits	0.003
257612	Eukaryotic Translation Termination	0.003
257951	Peptide chain elongation	0.003
188965	SRP-dependent cotranslational protein targeting to membrane	0.003
247749	Eukaryotic Translation Elongation	0.003
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.003
189048	Nonsense-Mediated Decay (NMD)	0.003
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0.003
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.004
251703	Cap-dependent Translation Initiation	0.005
230274	Eukaryotic Translation Initiation	0.005
260795	Translation	0.007

Table 5: Mapping analysis provided by Reactome for bicluster 5 from the SimNTO measure and 211 configuration.

Pathway identifier	Pathway name	FDR
188483	S6K1 signalling	0,003
188481	S6K1-mediated signalling	0,0211
188482	mTORC1-mediated signalling	0,0211

Table 6: Mapping analysis provided by Reactome for bicluster 1 from the SimNTO measure and 212 configuration.

Pathway identifier	Pathway name	FDR
188483	S6K1 signalling	0.002
231079	Gene Expression	0.003
233365	Metabolism of proteins	0.004
188482	mTORC1 mediated signalling	0.012
188481	S6K1 mediated signalling	0.012
252465	mTOR signalling	0.029
262893	PKB mediated events	0.029
237871	PI3K Cascade	0.042

Table 7: Mapping analysis provided by Reactome for bicluster 2 from the SimNTO measure and 212 configuration.

Pathway identifier	Pathway name	FDR
248935	Ribosomal scanning and start codon recognition	0.002
257608	Formation of the ternary complex, and subsequently, the 43S complex	0.010
231079	Gene Expression	0.036
256674	Activation of the mRNA upon binding of the cap-binding complex and eIFs, and subsequent binding to 43S	0.041

Table 8: Mapping analysis provided by Reactome for bicluster 3 from the SimNTO measure and 212 configuration.

Pathway identifier	Pathway name	FDR
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0.001
217188	Formation of a pool of free 40S subunits	0.001
257951	Peptide chain elongation	0.001
188965	SRP-dependent cotranslational protein targeting to membrane	0.001
247749	Eukaryotic Translation Elongation	0.001
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.001
257608	Formation of the ternary complex, and subsequently, the 43S complex	0.001
189048	Nonsense-Mediated Decay (NMD)	0.001
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0.001
248935	Ribosomal scanning and start codon recognition	0.001
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.001
251703	Cap-dependent Translation Initiation	0.002
230274	Eukaryotic Translation Initiation	0.002
260795	Translation	0.003
188483	S6K1 signalling	0.003
188481	S6K1-mediated signalling	0.022
188482	mTORC1-mediated signalling	0.022

Table 9: Mapping analysis provided by Reactome for bicluster 4 from the SimNTO measure and 212 configuration.

Pathway identifier	Pathway name	FDR
232946	GTP hydrolysis and joining of the 60S ribosomal subunit	0.014
217188	Formation of a pool of free 40S subunits	0.014
257951	Peptide chain elongation	0.014
188965	SRP-dependent cotranslational protein targeting to membrane	0.014
247749	Eukaryotic Translation Elongation	0.014
189183	Nonsense Mediated Decay (NMD) independent of the Exon Junction Complex (EJC)	0.014
189048	Nonsense-Mediated Decay (NMD)	0.014
189050	Nonsense Mediated Decay (NMD) enhanced by the Exon Junction Complex (EJC)	0.014
188483	S6K1 signalling	0.014
252688	L13a-mediated translational silencing of Ceruloplasmin expression	0.014
251703	Cap-dependent Translation Initiation	0.019
230274	Eukaryotic Translation Initiation	0.020
260795	Translation	0.032
188482	mTORC1-mediated signalling	0.036
188481	S6K1-mediated signalling	0.036

Table 10: Mapping analysis provided by Reactome for bicluster 5 from the SimNTO measure and 212 configuration.

Pathway identifier	Pathway name	FDR
231079	Gene Expression	0.002
233365	Metabolism of proteins	0.010
188483	S6K1 signalling	0.012
188482	mTORC1-mediated signalling	0.039
188481	S6K1-mediated signalling	0.039

Table 11: Mapping analysis provided by Reactome for bicluster 4 from the Corr measure and 210 configuration.

Pathway identifier	Pathway name	FDR
231079	Gene Expression	0.005
233365	Metabolism of proteins	0.009

Table 12: Mapping analysis provided by Reactome for bicluster 5 from the Corr measure and 210 configuration.

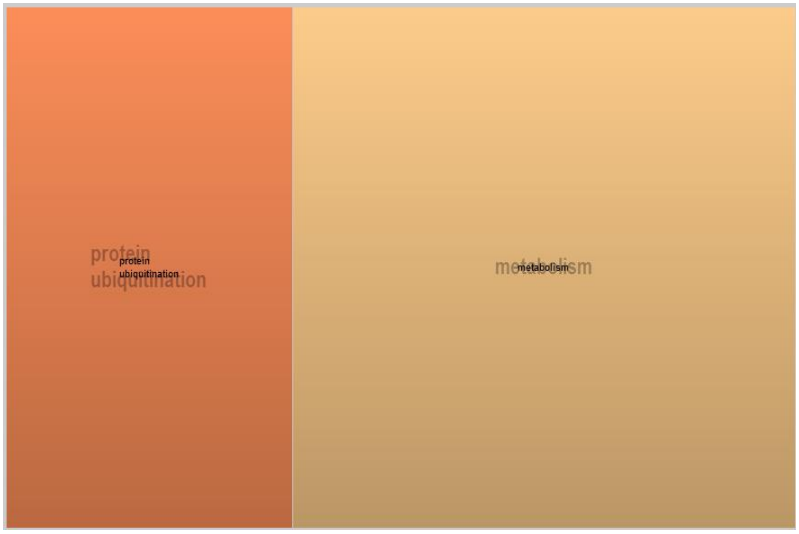

Figure 1: Cluster of GO terms reported by Revigo from bicluster 1 for the FracGO measure and 211 configuration.

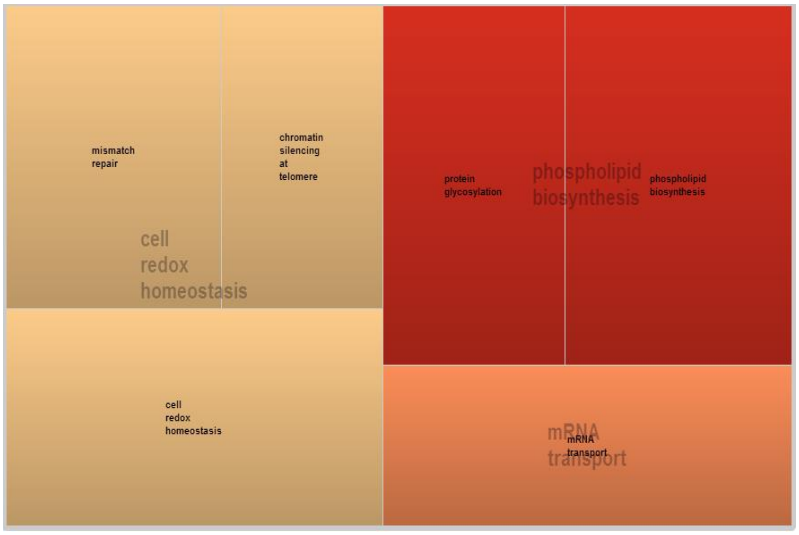

Figure 2: Cluster of GO terms reported by Revigo from bicluster 2 for the FracGO measure and 211 configuration.

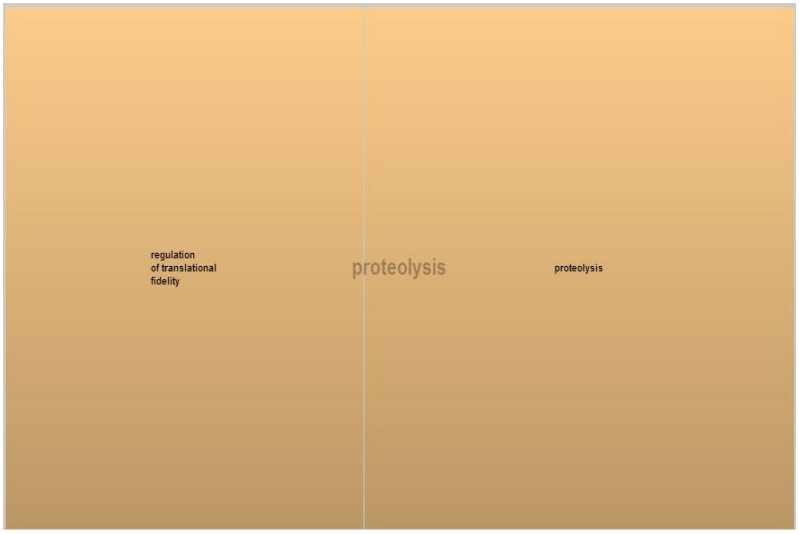

Figure 3: Cluster of GO terms reported by Revigo from bicluster 3 for the FracGO measure and 211 configuration.

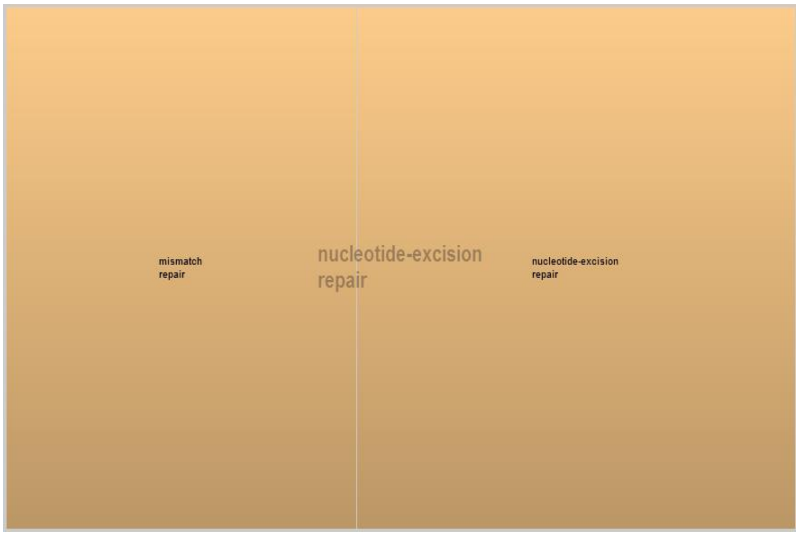

Figure 4: Cluster of GO terms reported by Revigo from bicluster 4 for the FracGO measure and 211 configuration.

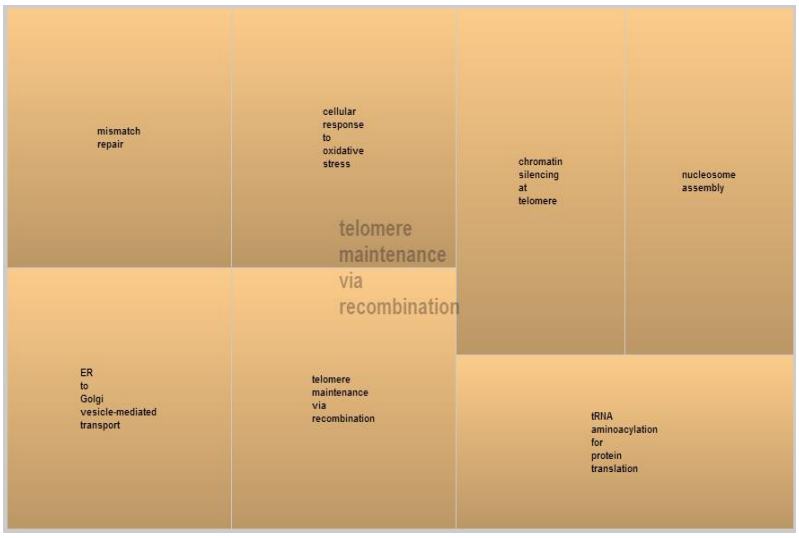

Figure 5: Cluster of GO terms reported by Revigo from bicluster 5 for the FracGO measure and 211 configuration.

Silhouette Width	Genes	list	ref. list	pValue	GO Terms
-0,008	ENO1 HEM15 FBA1 CHA1 RFC4 MSH2 RFA2 POL3 YNK1 GUK1 ADO1 SUB2 EST3 GBP2 UBP10 WBP1 SWP1 SNL1 OST6 UBC7 GNA1 ARD1 CST26 MUM3 RPC25 TRR2 URA3 CDD1 RPL40B RAD6 CDC36 SAD1 RDS3 PRP43 THO2 SSA2 MED8 ADY4 PDS5, NUP145	40(194)	369(7109)	1,96E-09	GO:0016829:lyase activity (MF) ;GO:0005739:mitochondrion (CC) ;GO:0000781:chromosome Telomeric region (CC) ;GO:0008152:metabolic process (BP) ;GO:0016746:transferase activity Transferring acyl groups (MF) ;GO:0016567:protein ubiquitination (BP) ;GO:0005198:structural molecule activity (MF)
0,348	RPS8B RPS9A RPL15A RPS4B RPS5 RPS3 RPL17B RPL2A RPL12A RPS6A RPS6B RPL32 RPL40B RPS13 RPL20A RPL42A RPS9B RLP24	18(194)	159(7109)	2,66E-02	GO:0003735:structural constituent of ribosome (MF) ;GO:0030529:ribonucleoprotein complex (CC) ;GO:0022625:cytosolic large ribosomal subunit (CC)
0,062	RPS8B RPS9A RPS4B RPS5 RPS3 RPS6A RPS13 RPS9B RPS6B ASC1,PRP43 HAS1 RLP24 RPL40B	14(194)	127(7109)	8,21E-01	GO:0003735:structural constituent of ribosome (MF) ;GO:0030686:90S preribosome (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC) ;GO:0000462:maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5,8S rRNA LSU-rRNA) (BP) ;GO:0032040:small-subunit processome (CC)
0,493	RPS9A RPS5 RPS3 RPS9B RPS13 ASC1 RPS6B RPL15A HAS1	9(194)	84(7109)	4,38E-04	GO:0003735:structural constituent of ribosome (MF) ;GO:0030529:ribonucleoprotein complex (CC) ;GO:0030686:90S preribosome (CC) ;GO:0015935:small ribosomal subunit (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC)
0,711	TEF1 HBS1 TEF4 TEF2 LSG1 RBG1 PSA1 EST3 SRP102	9(194)	91(7109)	7,91E-04	GO:0006414:translational elongation (BP) ;GO:0003746:translation elongation factor activity (MF) ;GO:0005525:GTP binding (MF) ;GO:0003924:GTPase activity (MF)
0,826	SUB2 PRP43 YRF1-2 DBP1 HAS1 YHL050C YML133C YEL077C	8(194)	77(7109)	1,12E-03	GO:0004386:helicase activity (MF) ;GO:0008026:ATP-dependent helicase activity (MF)

Table 13: Functional annotation using Gene Term LilnkerMapping analysis provided by Reactome for bicluster 1 from the SimNTO measure and 211 configuration.

Silhouette Width	Genes	list	ref. list	pValue	GO Terms
0,067	DBP5 SUB2 HEK2 SHE2 NUP145 GNA1 LAT1 ARG7 ARD1 CST26 MUM3 SAS3 SAS2 ALE1 NHP6A MSH4 RFC4 MSH2 POL2 SCS2 NUS1 SWH1 CAB5 ASP3-1 HXK2 YMR099C ADH3 GPM1 ADH2 RFA2 GBP2,UBP10 HHF1 TOP2 HTA2 RSC8 DBP6 TIF1 WSS1 YLL066C,OST1 ALG11 PSA1 CBR1 AHP1 TRR1 TRR2 PDI1 PRY2 PHO5 DSE2 SSA1 SIM1,FIG2 THI22 MED8 PDS5	57 (210)	568 (7109)	2,8E-12	GO:0051028: mRNA transport (BP);GO:0016746: transferase activity, transferring acyl groups (MF);GO:0006298: mismatch repair (BP); GO:0005635: nuclear envelope (CC); GO:0008654: phospholipid biosynthetic process (BP); GO:0000781: chromosome, telomeric region (CC); GO:0006348: chromatin silencing at telomere (BP); GO:0005694: chromosome (CC); GO:0008094: DNA-dependent ATPase activity (MF); GO:0004386: helicase activity (MF); GO:0008026: ATP-dependent helicase activity (MF); GO:0006486: protein amino acid glycosylation (BP); GO:0045454: cell redox homeostasis (BP); GO:0005576: extracellular region (CC); GO:0005198: structural molecule activity (MF)
0,519	RPL8A RPS8B RPS9A RPL13A RPL15A RPS4B RPS0A RPS5 RPS3 RPL12A RPL4A RPS6A RPL28 RPS4A RPL12B RPS2 RPS19A RPL9B RPS13 RPS7A RPS9B RPL2B	22 (210)	150 (7109)	3,6E-05	GO:0003735: structural constituent of ribosome (MF); GO:0022625: cytosolic large ribosomal subunit (CC)
0,213	RPS8B RPS9A RPS4B RPS0A RPS5 RPS3 RPS6A RPS4A RPS13 RPS7A RPS9B RPS2 RPS19A NOP56 NOP58 RPL15A RPL4A RPL28	18 (210)	143 (7109)	1,7E-02	GO:0003735: structural constituent of ribosome (MF); GO:0030686: 90S preribosome (CC); GO:0022627: cytosolic small ribosomal subunit (CC); GO:0015935: small ribosomal subunit (CC); GO:0006407: rRNA export from nucleus (BP); GO:0000462: maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5,8S rRNA, LSU-rRNA) (BP)
0,305	RPS9A RPL28 RPS2 RPS9B RPS4B RPS4A RPS6A RPS7A RPS0A RPL8A NOP56 NOP58	12 (210)	62 (7109)	1,9E-03	GO:0003735: structural constituent of ribosome (MF); GO:0030686: 90S preribosome (CC); GO:0022627: cytosolic small ribosomal subunit (CC); GO:0032040: small-subunit processome (CC); GO:0015935: small ribosomal subunit (CC)
0,739	YJR029W YBL005W-B LIF1 YIL082W-A YJL113W MSH2 PSA1 POL2 VMA1 REV7 APN1 LCL3 POP5 YPS6	14 (210)	116 (7109)	7,1E-02	GO:0006310: DNA recombination (BP); GO:0016779: nucleotidyltransferase activity (MF); GO:0004519: endonuclease activity (MF); GO:0004518: nuclease activity (MF); GO:0003887: DNA-directed DNA polymerase activity (MF); GO:0004540: ribonuclease activity (MF); GO:0006508: proteolysis (BP); GO:0008233: peptidase activity (MF); GO:0004190: aspartic-type endopeptidase activity (MF)
0,699	VMA8 VMA1 VMA5 VMA6 TUB1 ATG18 VBA3 VBA5	8 (210)	95 (7109)	6,8E-03	GO:0006811:ion transport (BP); GO:0000329:fungal-type vacuole membrane (CC); GO:0015991:ATP hydrolysis coupled proton transport (BP); GO:0015992:proton transport (BP); GO:0046961:proton-transporting ATPase activity rotational mechanism (MF); GO:0007035:vacuolar acidification (BP); GO:0005773:vacuole (CC); GO:0005774:vacuolar membrane (CC)

Table 14: Functional annotation using Gene Term LilnkerMapping analysis provided by Reactome for bicluster 2 from the SimNTO measure and 211 configuration.

Silhouette Width	Genes	list	ref list	pValue	GO Terms
-0,022	RPS23A RPL2A RPS23B RPL2B ZUO1 RPS5 IMP1 SBH2 IMP2 SRP102 LPD1 ADH1 YMR099C FBA1 PGK1 BFR1 MKT1 RBG1 OLA1 RHO2 ARF2 YPS6 YPS5 YIL082W-A YBR012W-B YJL113W	26(206)	168(7109)	1,4E-07	GO:0003735:structural constituent of ribosome (MF) ;GO:0006450:regulation of translational fidelity (BP) ;GO:0016021:integral to membrane (CC) ;GO:0016020:membrane (CC) ;GO:0005844:polysome (CC) ;GO:0005525:GTP binding (MF) ;GO:0006508:proteolysis (BP) ;GO:0004190:aspartic-type endopeptidase activity (MF)
0,617	RPS9A RPL13A RPS1A RPL17A RPS23A RPS5 RPS8A RPL17B RPL2A RPL6B RPL4A RPS7B RPS23B RPL32 RPS4A RPL8B RPL40B RPS2 RPL25 RPS13 RPS9B RPL2B	22(206)	150(7109)	2,4E-05	GO:0003735:structural constituent of ribosome (MF) ;GO:0022625:cytosolic large ribosomal subunit (CC)
0,305	RPS9A RPS5 RPS2 RPS9B RPL6B RPL4A RPL25	7(206)	32(7109)	2,8E+00	GO:0003735:structural constituent of ribosome (MF) ;GO:0015935:small ribosomal subunit (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC)
0,229	GNA1 SAS3 SAS2 TAN1 TRM11 TRM12 UBA4 TYE7 YHP1,HMO1,EEB1 LRO1 LSM5 TAD1	14(206)	134(7109)	3,0E+00	GO:0016746:transferase activity, transferring acyl groups (MF) ;GO:0008033:tRNA processing (BP) ;GO:0006351:transcription, DNA-dependent (BP) ;GO:0006355:regulation of transcription, DNA-dependent (BP) ;GO:0000790:nuclear chromatin (CC)
0,501	RPS9A RPS1A RPS5 RPS7B RPS4A RPS13 RPS9B RPS8A RPS23A RPS23B RPS2 NOP56 PRP43	13(206)	118(7109)	3,3E+00	GO:0003735:structural constituent of ribosome (MF) ;GO:0030686:90S preribosome (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC) ;GO:0000462:maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5,8S rRNA, LSU-rRNA) (BP)
0,262	YRF1-6 SUB2 PRP43 YRF1-2 DBP6 DBP2 SSA4 SAS3 SAS2 UBP10 POL2 HEK2 GBP2	13(206)	119(7109)	3,6E+00	IPR001650:Helicase, C-terminal;IPR014001:DEAD-like helicase;IPR011545:DNA/RNA helicase, DEAD/DEAH box type, N-terminal ;GO:0004386:helicase activity (MF) ;GO:0008026:ATP-dependent helicase activity (MF) ;GO:0006348:chromatin silencing at telomere (BP) ;GO:0000781:chromosome, telomeric region (CC)
0,459	RPS9A RPS7B RPL40B RPS2 RPS9B RPL8B NOP56 PRP43	8(206)	73(7109)	1,2E-03	GO:0003735:structural constituent of ribosome (MF) ;GO:0032040:small-subunit processome (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC) ;GO:0030686:90S preribosome (CC)
0,665	ALG5 WBP1 CWH41 SWP1 SBH2 CDC48	6(206)	47(7109)	2,2E-03	GO:0016021:integral to membrane (CC) ;GO:0016020:membrane (CC) ;GO:0006487:protein amino acid N-linked glycosylation (BP);

Table 15: Functional annotation using Gene Term LilnkerMapping analysis provided by Reactome for bicluster 3 from the SimNTO measure and 211 configuration.

Silhouette Width	Genes	list	ref list	pValue	GO Terms
0,232	IMD4 RPC25 CDC19 ADE13 YNK1 IMD3 ADO1 RPA34 PGM3 POL3 POL2 URK1 CPA1 SRS2 RFC2 RAD2 DOT1 NHP6A	18(205)	142(7109)	1,1E-02	GO:0003677:DNA binding (MF);GO:0006298:mismatch repair (BP) ;GO:0006289:nucleotide-excision repair (BP)
0,609	YRF1-6 YRF1-2 PRP2 DRS1 HAS1 DBP6 YLL067C YHL050C YFL066C YML133C RVB1 SRS2 YRF1-7	13(205)	100(7109)	5,1E-01	GO:0004386:helicase activity (MF) ;GO:0008026:ATP-dependent helicase activity (MF) ;GO:0003678:DNA helicase activity (MF)
0,133	ARC1 SES1,GRS1 YNL247W THS1 TGS1 CDC33 THO2 POP5 SMT3 NUS1 PMI40 OST3 KTR3	14(205)	118(7109)	6,5E-01	GO:0006418:tRNA aminoacylation for protein translation (BP);GO:0016874:ligase activity (MF) ;GO:0004812:aminoacyl-tRNA ligase activity (MF);GO:0006486:protein amino acid glycosylation (BP)
0,501	RPS9A RPL13A RPL15A RPS0A RPL16A RPL2A RPL6B RPL12A RPS1B RPS6A RPS23B RPL32 RPL9B RPS13 RPL42A	15(205)	150(7109)	2,5E+00	GO:0003735:structural constituent of ribosome (MF);GO:0005622:intracellular (CC) ;GO:0022625:cytosolic large ribosomal subunit (CC)
0,269	ENO1 CDC19 IMD3 TDH1 YGL082W HSC82 ECM33 SPS2 YMR099C PMI40 PGM3,MIG2	12(205)	133(7109)	4,3E-04	GO:0001950:plasma membrane enriched fraction (CC) ;GO:0005975:carbohydrate metabolic process (BP);
0,765	ASP3-2 ASP3-3 APE2 ASP3-4 ADE13 CPA1	6(205)	35(7109)	4,3E-04	GO:0030287:cell wall-bounded periplasmic space (CC);
0,894	SER1 ARO8 CYS4 BNA5 ARO9	5(205)	25(7109)	6,3E-04	GO:0030170:pyridoxal phosphate binding (MF)
0,677	SAS3 SAS2 DOT1 POL2 ADA2,GCN5	6(205)	38(7109)	6,8E-04	GO:0006348:chromatin silencing at telomere (BP) ;GO:0004402:histone acetyltransferase activity (MF)
0,686	RPS9A RPS0A RPS1B RPS6A RPS13 RPS23B ASC1	7(205)	64(7109)	2,3E-03	GO:0003735:structural constituent of ribosome (MF);GO:0030686:90S preribosome (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC) ;GO:0000462:maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5,8S rRNA, LSU-rRNA) (BP) ;GO:0005622:intracellular (CC)
0,809	HO APN1 LCL3 YML039W RAD2 YCL019W POL3	7(205)	67(7109)	3,0E-03	GO:0004519:endonuclease activity (MF) ;GO:0004518:nuclease activity (MF) ;GO:0003677:DNA binding (MF)
0,752	RVB1 PEX1 PCH2 HSP104 PDR5,GCN20 SRP54 RFC2	8(205)	86(7109)	3,2E-03	GO:0017111:nucleoside-triphosphatase activity (MF)

Table 16: Functional annotation using Gene Term LilnkerMapping analysis provided by Reactome for bicluster 4 from the SimNTOlib measure and 211 configuration.

Silhouette Width	Genes	list	ref list	pValue	GO Terms
-0,017	YRF1-6 YRF1-2 YRF1-7 YRF1-1 HHF2 HHF1 HHT2 HTB2 RVB1 SAS3 SAS2 DOT1 POL2 YNK1 POL3 URK1 RFC2 MSH2 CDC19 GCN5 TAN1 SCS2 NUS1 KAP122 UIP3 GBP2 RSC8 MATALPHA2 YHP1 HMLALPHA2 WSS1 RNH1 ENO1 YGL082W YDL124W HSC82 TIF2 PFA3 MUM3 DED81 GUS1 DPS1 THS1 PEX1 PCH2 HSP104 CUE1 EMP47 SRP102 BET1 RER1 SOP4 OLA1 THG1 LSG1 EFT2 YKR018C,FUN14 PTH2 NCP1 SEN54 SAM35 SVF1 OCA1 UBA4 RPN2 ATG19 HRT3,VAC14	69 (198)	597 (7109)	1,5E-21	GO:0000722:telomere maintenance via recombination (BP) ;GO:0003678:DNA helicase activity (MF) ;GO:0006334:nucleosome assembly (BP) ;GO:0000786:nucleosome (CC) ;GO:0005694:chromosome (CC) ;GO:0000788:nuclear nucleosome (CC); GO:0006333:chromatin assembly or disassembly (BP) ;GO:0006348:chromatin silencing at telomere (BP);GO:0006298:mismatch repair (BP);GO:0016746:transferase activity, transferring acyl groups (MF) ;GO:0005635:nuclear envelope (CC); GO:0001950:plasma membrane enriched fraction (CC) ;GO:0006418:tRNA aminoacylation for protein translation (BP) ;GO:0004812:aminoacyl-tRNA ligase activity (MF);GO:0017111:nucleoside-triphosphatase activity (MF); GO:0030176:integral to endoplasmic reticulum membrane (CC) ;GO:0006888:ER to Golgi vesicle-mediated transport (BP) ;GO:0005525:GTP binding (MF) ;GO:0005741:mitochondrial outer membrane (CC) ;GO:0034599:cellular response to oxidative stress (BP) ;GO:0030674:protein binding, bridging (MF)
0,196	RPS9A RPS7B RPS6A RPS6B RPS2 RPS9B RPL8A RPS0A RPL40B RPS3 RPS4B RPS23A RPL24B RPL6B RPL9A EFT2 PRP43 NOP58	18 (198)	152 (7109)	1,8E-02	GO:0003735:structural constituent of ribosome (MF);GO:0032040:small-subunit processome (CC) ;GO:0022627:cytosolic small ribosomal subunit (CC) ;GO:0015935:small ribosomal subunit (CC) ;GO:0030686:90S preribosome (CC) ;GO:0000462:maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5,8S rRNA, LSU-rRNA) (BP) ;GO:0019843:rRNA binding (MF)
0,791	THG1 YMR045C YER138C YBL005W-B YIL082W-A YBR012W-B POL3 POL2 LCL3 RNH1 LIF1 MSH2	12 (198)	69 (7109)	3,4E-02	GO:0016779:nucleotidyltransferase activity (MF) ;GO:0004518:nuclease activity (MF) ;GO:0003887:DNA-directed DNA polymerase activity (MF) ;GO:0004519:endonuclease activity (MF) ;GO:0004523:ribonuclease H activity (MF) ;GO:0006310:DNA recombination (BP) ;GO:0032197:transposition, RNA-mediated (BP) ;GO:0000943:retrotransposon nucleocapsid (CC) ;GO:0006508:proteolysis (BP) ;GO:0008233:peptidase activity (MF) ;GO:0046797:viral procapsid maturation (BP) ;GO:0015074:DNA integration (BP) ;GO:0004190:aspartic-type endopeptidase activity (MF) ;GO:0003964:RNA-directed DNA polymerase activity (MF) ;GO:0004540:ribonuclease activity (MF); GO:0032196:transposition (BP)
0,553	RPL8A RPS9A RPL13A RPL24B RPS23A RPS4B RPS0A RPS3 RPL6B RPL9A RPS7B RPS6A RPS6B RPL40B RPS2 RPL1A RPS9B	17 (198)	150 (7109)	7,6E-02	GO:0003735:structural constituent of ribosome (MF);GO:0022625:cytosolic large ribosomal subunit (CC)
0,685	YRF1-6 YBL111C PRP43 YRF1-2 DBP1 DRS1 DBP6 TIF2 YHL050C YEL077C YBL112C RVB1 RPL6B LSG1	14 (198)	129 (7109)	1,2E+00	GO:0008026:ATP-dependent helicase activity (MF) ;GO:0004386:helicase activity (MF); GO:0004004:ATP-dependent RNA helicase activity (MF) ;GO:0000027:ribosomal large subunit assembly (BP)
0,655	ROT2 STT3 CWH41 OST3 ALG11 NUS1 EUG1 HSC82 CUE1	9 (198)	113 (7109)	4,1E-03	GO:0006486:protein amino acid glycosylation (BP)

Table 17: Functional annotation using Gene Term LilnkerMapping analysis provided by Reactome for bicluster 5 from the SimNTO measure and 211 configuration.

2. Additional experiments with different parameters values

This table presents the percentage of enriched biclusters and the average of the number of GO terms per bicluster for 100 biclusters obtained by the proposed algorithm when using the SimNTO measure for GDS1116 dataset with different values of M_1, M_2 and M_3 parameters. The main parameters are M_2 and M_3 that control the average correlation and the biological integration measure, respectively. Zero values are excluded of this experiment because these values report biclusters without biological meaning. For the same reason, M_1 is set to 2 to find biclusters with a non trivial number of genes.

Several values varying from 1 to 2 for M_2 and M_3 have been considered (namely, 1.0, 1.3, 1.5, 1.8 and 2) according to the comments of the reviewer. It can be observed that the results do not show meaningful differences. Due to this fact, a representative configuration setting can be summarized as: the average correlation and the biological integration are equally important ($M_2 = 1$, $M_3 = 1$), the first is more important than the second ($M_2 = 2$, $M_3 = 1$) or vice versa ($M_2 = 1$, $M_3 = 2$).

		Enriched bi. (%)			GO terms per bi.		
(M_1, M_2, M_3)	size	BP	MF	CC	BP	MF	CC
(2,2.0,1.0)	(10.71 × 5.3)	99	86	89	4.8	1.1	2.1
(2,1.8,1.0)	(10.6 × 5.8)	95	88	92	4.7	1.1	2.0
(2,1.5,1.0)	(10.1 × 8.1)	96	93	96	4.6	1.3	2.1
(2,1.3,1.0)	(11.0 × 9.4)	99	94	95	5.1	1.1	2.1
(2,1.0,1.0)	(11.6 × 15.6)	99	97	97	5.7	1.2	2.2
(2,1.0,2.0)	(8.7 × 15.0)	97	84	88	3.8	1.0	1.9
(2,1.0,1.8)	(9.2 × 15.7)	91	83	84	4.0	1.0	1.9
(2,1.0,1.5)	(9.7 × 15.4)	98	91	93	4.6	1.0	2.0
(2,1.0,1.3)	(10.2 × 15.3)	99	95	95	4.9	1.2	2.1
(2,1.0,1.0)	(11.6 × 15.6)	99	97	97	5.7	1.2	2.2

Table 18: SimNTO-based biclusters with different values of M_1, M_2 and M_3 parameters for GDS1116 dataset.

Bibliografía

[1] J.S. Aguilar-Ruiz. Shifting and scaling patterns from gene expression data. *Bioinformatics*, 21(20):3840–3845, 2005.

[2] Fadhl M. Al-Akwaa and Yasser M. Kadah. 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*, pages 163–167, Piscataway, NJ, USA, 2009. IEEE Press.

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

[4] N. Amit. The biclustering graph editing problem. Master's thesis, Tel Aviv University, School of Mathematical Sciences, 2004. Thesis towards the M.Sc. degree. University of Tel Aviv.

[5] Michael Ashburner, Catherine A Ball, Judith A Blake, David Botstein, Heather Butler, J Michael Cherry, Allan P Davis, Kara Dolinski, Selina S Dwight, Janan T Eppig, et al. Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29, 2000.

[6] Wassim Ayadi, Mourad Elloumi, and Jin-Kao Hao. A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. *BioData Mining*, 2(1):9, 2009.

[7] Francisco Azuaje. *Bioinformatics and Biomarker Discovery: Omic Data Analysis for Personalized Medicine*. Wiley-Blackwell, 2010.

[8] Francisco Azuaje, Haiying Wang, Huiru Zheng, Frederique Leonard, Magali Rolland-Turner, Lu Zhang, Yvan Devaux, and Daniel Wagner. Predictive integration of gene functional similarity and co-expression defines treatment response of endothelial progenitor cells. *BMC Systems Biology*, 5(1):46, 2011.

[9] H. Banka and S. Mitra. Evolutionary biclustering of gene expressions. *Ubiquity*, 7(42):1–12, 2006.

[10] S. Barkow, S. Bleuler, A. Prelic, P. Zimmermann, and E. Zitzler. Bicat: a biclustering analysis toolbox. *Bioinformatics*, 22(10):1282–1283, 2006.

[11] 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(3-4):373–384, 2003.

[12] S. Bergmann, J. Ihmels, and N. Barkai. Iterative signature algorithm for the analysis of large-scale gene expression data. *Physical Review E*, 67(031902):1 –18, 2003.

[13] Gabriel F. Berriz, Oliver D. King, Barbara Bryant, Chris Sander, and Frederick P. Roth. Characterizing gene sets with funcassociate. *Bioinformatics*, 19(18):2502–2504, 2003.

[14] Anindya Bhattacharya and Rajat K. De. Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics*, 25(21):2795–2801, 2009.

[15] Anindya Bhattacharya and Rajat K. De. Bi-correlation clustering algorithm for determining a set of co-regulated genes. *Bioinformatics*, 25(21):2795–2801, 2009.

[16] S. Bleuler, A. Prelic, and E. Zitzler. An ea framework for biclustering of gene expression data. In *Proceedings of Congress on Evolutionary Computation, 2004 - CEC2004*, volume 1, 2004.

[17] Kruglyak L. Brem RB. The landscape of genetic complexity across 5,700 gene expression traits in yeast. *Proc Natl Acad Sci U S A*, 5(102):1572–7, Feb 2005.

[18] Jenny Bryan. Problems in gene clustering based on gene expression data. *Journal of Multivariate Analysis*, 90:44–66, 2004.

[19] K. Bryan. Biclustering of expression data using simulated annealing. In *Proceedings of the 18th IEEE International Symposium on Computer-Based Medical Systems*, pages 383–388. USA, 2005.

[20] S. Busygin, O. Prokopyev, and P.M. Pardalos. Biclustering in data mining. *Computers and Operations Research*, 35(9):2964–2987, 2008.

[21] Horacio Caniza, Alfonso E. Romero, Samuel Heron, Haixuan Yang, Alessandra Devoto, Marco Frasca, Marco Mesiti, Giorgio Valentini, and Alberto Paccanaro. Gossto: a stand-alone application and a web tool for calculating semantic similarities on the gene ontology. *Bioinformatics*, 30(15):2235–2236, 2014.

[22] Artur J. ; Figueiredo MÃ¡rio A. T. ; Cordeiro Madeira Sara Carreiro, AndrÃ© Valerio ; Ferreira. Towards a classification approach using meta-biclustering: Impact of discretization in the analysis of expression time series. *Journal of Integrative Bioinformatics*, 9(3):207, 2012.

[23] Cristian I. Castillo-Davis and Daniel L. Hartl. Genemerge–post-genomic analysis, data mining, and hypothesis testing. *Bioinformatics*, 19(7):891–892, 2003.

[24] Y. Cheng and G.M. Church. Biclustering of expression data. In *Proceedings of the 8th International Conference on Intelligent Systems for Molecular Biology*, volume 8, pages 93–103, 2000.

[25] R.J. Cho and et al. A genome-wide transcriptional analysis of the mitotic cell cycle. *Molecular Cell*, 2(1):65–73, 1998.

[26] Phuong et al. Dao. Inferring cancer subnetwork markers using density-constrained biclustering. *Bioinformatics*, 26(18):625–631, 2010.

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

[28] I.S. Dhillon, S. Mallela, and D.S. Modha. Information-theoretic co-clustering. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 89–98. ACM Press New York, NY, USA, 2003.

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

[30] F. Divina and J.S. Aguilar-Ruiz. A multi-objective approach to discover biclusters in microarray data. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 385–392. ACM Press New York, NY, USA, 2007.

[31] Ron Edgar, Michael Domrachev, and Alex E. Lash. Gene expression omnibus: Ncbi gene expression and hybridization array data repository. *Nucleic Acids Research*, 30(1):207–210, 2002.

[32] Kemal Eren, Mehmet Deveci, Onur Kucuktunc, and Umit V. Catalyurek. A comparative analysis of biclustering algorithms for gene expression data. *Briefings in Bioinformatics*, 14(3):279–292, 2013.

[33] Antonino Fiannaca, Massimo La Rosa, Laura La Paglia, Riccardo Rizzo, and Alfonso Urso. Analysis of mirna expression profiles in breast cancer using biclustering. *BMC Bioinformatics*, 16(Suppl 4):S7, 2015.

[34] Jose Luis Flores, Iñaki Inza, Pedro Larrañaga, and Borja Calvo. A new measure for gene expression biclustering based on non-parametric correlation. *Computer Methods and Programs in Biomedicine*, 112(3):367–397, 2013.

[35] Celia Fontanillo, Ruben Nogales-Cadenas, Alberto Pascual-Montano, and Javier De Las Rivas. Functional analysis beyond enrichment: Non-redundant reciprocal linkage of genes and biological terms. *PLoS ONE*, 6(9):e24289, 09 2011.

[36] G Fontanini, S Vignati, D Bigini, A Mussi, M Lucchi, CA Angeletti, F Basolo, and G Bevilacqua. Bcl-2 protein: a prognostic factor inversely correlated to p53 in non-small-cell lung cancer. *British Journal of Cancer*, 71(5):1003–1007, May 1995.

[37] Cristian Andrés Gallo, Jessica Andrea Carballido, and Ignacio Ponzoni. Bihea: A hybrid evolutionary approach for microarray biclustering. In *Proceedings of the 4th Brazilian Symposium on Bioinformatics: Advances in Bioinformatics and Computational Biology*, volume 5676, pages 36–47, 2009.

[38] Cristian Andrés Gallo, Jessica Andrea Carballido, and Ignacio Ponzoni. Microarray biclustering: A novel memetic approach based on the pisa platform. In *Proceedings of the 7th European Conference on Evolutionary Computation, Machine Learning and Data Mining - EvoBIO 2009*, pages 44–55, 2009.

[39] X. Gan, A.W.C. Liew, and H. Yan. Discovering biclusters in gene expression data based on high-dimensional linear geometries. *BMC Bioinformatics*, 9(209):1–15, 2008.

[40] Audrey P. Gasch, Paul T. Spellman, Camilla M. Kao, Orna Carmel-Harel, Michael B. Eisen, Gisela Storz, David Botstein, and Patrick O. Brown. Genomic expression programs in the response of yeast cells to environmental changes. *Mol. Biol. Cell*, 11(12):4241–4257, 2000.

[41] Joana P. Gonçalves, Sara C. Madeira, and Arlindo L. Oliveira. Biggests: integrated environment for biclustering analysis of time series gene expression data. *BMC Research Notes*, 2(124), 2009.

[42] Neelima Gupta and Seema Aggarwal. Mib: Using mutual information for biclustering gene expression data. *Pattern Recognition*, 43(8):2692 – 2697, 2010.

[43] R. Harpaz and R. Haralick. Exploiting the geometry of gene expression patterns for unsupervised learning. *18th International Conference on Pattern Recognition (ICPR 2006)*, 2:670–674, 2006.

[44] JA Hartigan. Direct clustering of a data matrix. *Journal of the American Statistical Association*, 67(337):123–129, 1972.

[45] Robin Haw, Henning Hermjakob, Peter D'Eustachio, and Lincoln Stein. Reactome pathway analysis to enrich biological discovery in proteomics data sets. *Proteomics*, 11(18):3598–3613, September 2011.

[46] Rui Henriques and Sara Madeira. Bicspam: flexible biclustering using sequential patterns. *BMC Bioinformatics*, 15(1):130, 2014.

[47] Curtis Huttenhower, K. Tsheko Mutungu, Natasha Indik, Woongcheol Yang, Mark Schroeder, Joshua J. Forman, Olga G. Troyanskaya, and Hilary A. Coller. Detailing regulatory networks through large scale data integration. *Bioinformatics*, 25(24):3267–3274, 2009.

[48] J. Ihmels, S. Bergmann, and N. Barkai. Defining transcription modules using large-scale gene expression data. *Bioinformatics*, 20(13):1993–2003, 2004.

[49] Y. Kluger, R. Basri, J.T. Chang, and M. Gerstein. Spectral biclustering of microarray data: Coclustering genes and conditions. *Genome Research*, 13(4):703, 2003.

[50] L. Lazzeroni and A. Owen. Plaid models for gene expression data. *Statistica Sinica*, 12(1):61–86, 2002.

[51] Guojun Li, Qin Ma, Haibao Tang, Andrew H. Paterson, and Ying Xu. Qubic: a qualitative biclustering algorithm for analyses of gene expression data. *Nucleic Acids Research*, 37(15):e101, 2009.

[52] Li Li, Yang Guo, Wenwu Wu, Youyi Shi, Jian Cheng, and Shiheng Tao. A comparison and evaluation of five biclustering algorithms by quantifying goodness of biclusters for gene expression data. *BioData Mining*, 5(1):8, 2012.

[53] F. Liu, H. Zhou, J. Liu, and G. He. Biclustering of gene expression data using eda-ga hybrid. In *IEEE Congress on Evolutionary Computation. CEC 2006*, pages 1598–1602, 2006.

[54] Junwan Liu, Zhoujun Li, Xiaohua Hu, and Yiming Chen. Biclustering of microarray data with mospo based on crowding distance. *BMC Bioinformatics*, 10(Suppl 4):S9, 2009.

[55] Kenneth Lo, Adrian Raftery, Kenneth Dombek, Jun Zhu, Eric Schadt, Roger Bumgarner, and Ka Yeung. Integrating external biological knowledge in the construction of regulatory networks from time-series expression data. *BMC Systems Biology*, 6(1):101, 2012.

[56] Hugo López-Fernández, Miguel Reboiro-Jato, Sara C Madeira, Rubén López-Cortés, JD Nunes-Miranda, HM Santos, Florentino Fdez-Riverola, and Daniel Glez-Peña. A workflow for the application of biclustering to mass spectrometry data. In *7th International Conference on Practical Applications of Computational Biology & Bioinformatics*, volume 222, pages 145–153, 2013.

[57] R.M. Luque-Baena, D. Urda, M. Gonzalo Claros, L. Franco, and J.M. Jerez. Robust gene signatures from microarray data using genetic algorithms enriched with biological pathway keywords. *Journal of Biomedical Informatics*, 49(0):32 – 44, 2014.

[58] Sara C. Madeira and Arlindo L. Oliveira. A polynomial time biclustering algorithm for finding approximate expression patterns in gene expression time series. *Algorithms for Molecular Biology*, 4, 2009.

[59] Sara C. Madeira, Miguel C. Teixeira, Isabel Sá-Correia, and Arlindo L. Oliveira. Identification of regulatory modules in time series gene expression data using a linear time biclustering algorithm. *IEEE/ACM Trans. Comput. Biology Bioinform.*, 7(1):153–165, 2010.

[60] S.C. Madeira and A.L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE Transactions on Computational Biology and Bioinformatics*, 1(1):24–45, 2004.

[61] R. Marti and M. Laguna. *Scatter Search. Methodology and Implementation in C.* Kluwer Academic Publishers, Boston, 2003.

[62] Ignacio Medina, JosÃ© Carbonell, Luis Pulido, Sara C. Madeira, Stefan Goetz, Ana Conesa, JoaquÃn TÃ¡rraga, Alberto Pascual-Montano, Ruben Nogales-Cadenas, Javier Santoyo, Francisco GarcÃa, Martina MarbÃ , David Montaner, and JoaquÃn Dopazo. Babelomics: an integrative platform for the analysis of transcriptomics, proteomics and genomic data with advanced functional profiling. *Nucleic Acids Research*, 38(suppl 2):W210–W213, 2010.

[63] B. Mirkin. *Mathematical Classification and Clustering.* Academic Press: Boston-Dordrecht, 1996.

[64] D. Mishra and A.K. Rath. Cpb: A model for biclustering. In *Proceedings of International Conference on Information Management and Engineering, 2009 - ICIME '09.*, pages 629–632, 2009.

[65] Meeta Mistry and Paul Pavlidis. Gene ontology term overlap as a measure of gene functional similarity. *BMC Bioinformatics*, 9(1):327, 2008.

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

[67] J. Morgan and J. Sonquistz. Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, 58(302):415–434, 1963.

[68] T.M. Murali and S. Kasif. Extracting conserved gene expression motifs from gene expression data. In *Proceedings of Pacific Symposium on Biocomputing*, pages 77–88, 2003.

[69] Juan A. Nepomuceno, Alicia Troncoso, and Jesus Aguilar-Ruiz. Biclustering of gene expression data by correlation-based scatter search. *BioData Mining*, 4(1):3, 2011.

[70] Juan A. Nepomuceno, Alicia Troncoso, and Jesús S. Aguilar-Ruiz. Evolutionary metaheuristic for biclustering based on linear correlations among genes. In *SAC 2010: Proceedings of the 2010 ACM Symposium*

on Applied Computing (SAC), Sierre, Switzerland, March 22-26, 2010, pages 1143–1147, 2010.

[71] Haiyuan; Paccanaro Alberto Nepusz, Tamas; Yu. Detecting overlapping protein complexes in protein-protein interaction networks. *Nature Methods*, 9:471, 2012.

[72] Catia Pesquita, Daniel Faria, Hugo Bastos, Antonio Ferreira, Andre Falcao, and Francisco Couto. Metrics for go based protein semantic similarity: a systematic evaluation. *BMC Bioinformatics*, 9(Suppl 5):S4, 2008.

[73] Gianvito Pio, Michelangelo Ceci, Domenica D'Elia, Corrado Loglisci, and Donato Malerba. A novel biclustering algorithm for the discovery of meaningful biological correlations between micrornas and their target genes. *BMC Bioinformatics*, 14(Suppl 7):S8, 2013.

[74] B. Pontes, F. Divina, R. Giraldez, and J.S. Aguilar-Ruiz. Virtual error: A new measure for evolutionary biclustering. *Lecture Notes in Computer Science*, 4447:217–226, 2007.

[75] B. Pontes, R. Giraldez, and J.S. Aguilar-Ruiz. Shifting patterns discovery in microarrays with evolutionary algorithms. *10th International conference on Knowledge-based & Intelligent Information & Engineering Systems (KES-06)*, 3102:1264–1271, 2006.

[76] Beatriz Pontes, Raúl Giráldez, and Jesús S Aguilar-Ruiz. Measuring the quality of shifting and scaling patterns in biclusters. In Tjeerd Dijkstra, Evgeni Tsivtsivadze, Elena Marchiori, and Tom Heskes, editors, *Pattern Recognition in Bioinformatics*, volume 6282 of *Lecture Notes in Computer Science*, pages 242–252. Springer Berlin / Heidelberg, 2010.

[77] A. Prelic, S. Bleuler, P. Zimmermann, A. Wille, P. Buhlmann, W. Gruissem, L. Hennig, L. Thiele, and E. Zitzler. A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, 22(9):1122–1129, 2006.

[78] Monika Ray, Jianhua Ruan, and Weixiong Zhang. Variations in the transcriptome of alzheimer's disease reveal molecular networks involved in cardiovascular diseases. *Genome Biology*, 9(10):R148, 2008.

[79] David Reiss, Nitin Baliga, and Richard Bonneau. Integrated biclustering of heterogeneous genome-wide datasets for the inference of global regulatory networks. *BMC Bioinformatics*, 7(1):280, 2006.

[80] Kahn Rhrissorrakrai and Kristin Gunsalus. Mine: Module identification in networks. *BMC Bioinformatics*, 12(1):192, 2011.

[81] Domingo S. Rodriguez-Baena, Antonio J. Perez-Pulido, and Jesus S. Aguilar-Ruiz. A biclustering algorithm for extracting bit–patterns from binary datasets. *Bioinformatics*, 27(19):2738–45, 2011.

[82] Rodrigo Santamaría, Luis Quintales, and Roberto Therón. Methods to bicluster validation and comparison in microarray data. In *Proceedings of Intelligent Data Engineering and Automated Learning - IDEAL 2007*, volume 4881 of *Lecture Notes in Computer Science*, pages 780–789. Springer Berlin / Heidelberg, 2007.

[83] Vemparala et al. Subbarayan. Inverse relationship between 15-lipoxygenase-2 and ppar-γ gene expression in normal epithelia compared with tumor epithelia. *Neoplasia*, 7(3):280–293, March 2005.

[84] Fran Supek, Matko Bosnjak, Nives Skunca, and Tomislav Smuc. Revigo summarizes and visualizes long lists of gene ontology terms. *PLoS ONE*, 6(7):e21800, 07 2011.

[85] A. Tanay, R. Sharan, and R. Shamir. Discovering statistically significant biclusters in gene expression data. *Bioinformatics*, 18(90001):136–144, 2002.

[86] A. Tanay, R. Sharan, and R. Shamir. Biclustering algorithms: A survey. *Handbook of Computational Molecular Biology*, 9:26–1, 2005.

[87] Marie Verbanck, Sebastien Le, and Jerome Pages. A new unsupervised gene clustering algorithm based on the integration of biological knowledge into expression data. *BMC Bioinformatics*, 14(1):42, 2013.

[88] Hong Yan Wen-Hui Yang, Dao-Qing Dai. Finding correlated biclusters from gene expression data. *IEEE Transactions on Knowledge and Data Engineering.*, IEEE computer Society Digital Library. IEEE Computer Society:568–584, 2010.

[89] Scott Williams and Jason Moore. Big data analysis on autopilot? *BioData Mining*, 6(1):22, 2013.

[90] Christof Winter, Glen Kristiansen, Stephan Kersting, Janine Roy, Daniela Aust, Thomas KnÃ¶sel, Petra RÃ$\frac{1}{4}$mmele, Beatrix Jahnke, Vera Hentrich, Felix RÃ$\frac{1}{4}$ckert, Marco Niedergethmann, Wilko Weichert, Marcus Bahra, Hans J. Schlitt, Utz Settmacher, Helmut Friess, Markus

BÃ¼chler, Hans-Detlev Saeger, Michael Schroeder, Christian Pilarsky, and Robert GrÃ¼tzmann. Google goes cancer: Improving outcome prediction for cancer patients by network-based ranking of marker genes. *PLoS Comput Biol*, 8(5):e1002511, 05 2012.

[91] J. Yang, H. Wang, W. Wang, and P. Yu. Enhanced biclustering on expression data. In *Proceedings of 3th IEEE Simposium on Bioinformatics and Bioengeneering*, pages 321–327, 2003.

[92] Taegyun Yun and Gwan-Su Yi. Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion. *BMC Genomics*, 14(1):144, 2013.

[93] Taegyun Yun and Gwan-Su Yi. Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion. *BMC Genomics*, 14:144, 2013.

[94] Tao Zeng and Jinyan Li. Maximization of negative correlations in time-course gene expression data for enhancing understanding of molecular pathways. *Nucleic Acids Research*, 38(1):e1, 2010.