ANÁLISIS DE DATOS DE EXPRESIÓN GENÉTICA

Beatriz Pontes, Domingo S. Rodríguez-Baena, Norberto Díaz-Díaz
Dto. Lenguajes y Sistemas Informáticos, Universidad de Sevilla, {bepontes,dsavio,ndiaz}@lsi.us.es

Raúl Giráldez
Escuela Politécnica Superior, Universidad Pablo de Olavide, rgirroj@upo.es

Resumen

El análisis de datos de expresión genética es una de las tareas fundamentales dentro de la Bioinformática. Para llevar a cabo este estudio se hace necesaria la aplicación de técnicas de Minería de Datos. Las técnicas de Clustering han probado ser de gran utilidad a la hora de descubrir grupos de genes que intervienen en una misma función celular o que están regulados de la misma manera. Recientemente, el Biclustering ha sido propuesto como método para descubrir patrones de comportamiento específico en los que el valor de expresión de un subgrupo de genes evoluciona de la misma forma a lo largo de un subgrupo de condiciones de laboratorio. En este artículo se revisan las distintas técnicas usadas en el análisis de datos de expresión genética, estudiándose en profundidad los métodos basados en Biclustering, además de discutir los diferentes métodos de validación para evaluar el modelo generado por las distintas propuestas.

Palabras clave: Bioinformática, Expresión Genética, Microarray, Clustering, Biclustering.

1. INTRODUCCIÓN

La Bioinformática es una disciplina reciente cuya aparición es propiciada por la revolución llevada a cabo en los campos de la biología molecular y la informática en el último siglo [21]. Podemos definirla como el uso de bases de datos y algoritmos computacionales para analizar proteínas, genes y la completa colección de ADN que forma un organismo (genoma), generando herramientas software que ayuden a solucionar problemas biológicos relacionados con la estructura y función de las macromoléculas, procesos bioquímicos, enfermedades, evolución, etc [35].

Uno de los principales objetivos de la Bioinformática se centra en el genoma, el transcriptoma (ARN generado a partir del ADN) y el proteoma (secuencias de proteínas). Estos millones de secuencias moleculares representan una gran oportunidad para aplicar técnicas de extracción de conocimiento cuyo objetivo final sea determinar la funcionalidad de ciertas células y las relaciones gen-proteína. La disciplina más importante en esta perspectiva es la *Extracción y Análisis de Secuencias de ADN, ARN y Proteínas* [32].

Los organismos individuales representan la segunda perspectiva de la Bioinformática. Cada organismo cambia a través de sus diferentes estados de desarrollo en diferentes regiones del cuerpo. Los genes son regulados dinámicamente en respuesta al paso del tiempo, la región y el estado fisiológico. De esta forma, la expresión genética varía en estados de enfermedad o en respuesta a una gran variedad de señales. Muchas de las herramientas generadas por la Bioinformática son aplicadas al *Análisis de bases de datos de expresión genética* [32] a partir de diferentes tejidos o distintas condiciones experimentales.

Actualmente, las bases de datos biológicas albergan secuencias de ADN de aproximadamente 100000 organismos. La Bioinformática ayuda en este nivel a estudiar las similitudes existentes entre los seres vivos a nivel molecular y en la comparación de genomas. Ramas de la Bioinformática como el *Estudio y Medición de la Biodiversidad o la Comparación de Genomas* [32] intentan aplicar la tecnología computacional para recopilar información sobre las distintas especies y contrastarla.

A lo largo de este artículo se van a tratar técnicas relacionadas con el *Análisis de datos de Expresión Genética*, haciendo especial énfasis en las técnicas de biclustering, y analizando algunos de los métodos de validación.

A continuación exponemos la organización del presente artículo. En la sección 2 se introduce el concepto de expresión genética y cuáles son las alternativas para su estudio. El Clustering como técnica de extracción de conocimiento es analizado en la sección 3. El apartado 4 analiza

las características de las técnicas de Biclustering, mientras que en la sección 5 se comentan algunas técnicas de evaluación de resultados. Por último, en el apartado 6 se resumen las principales conclusiones.

2. EXPRESIÓN GENÉTICA

La expresión de un gen se obtiene cuando una porción de ADN se transcribe para crear una molécula de ARN como primer paso en la síntesis de proteínas. Del numeroso grupo de genes que se encuentran en el núcleo y que codifican proteínas, sólo un subconjunto de ellos se expresará en un momento determinado. Esta expresión selectiva viene regulada por distintos aspectos: tipo de célula, fase de desarrollo del ser vivo, estimulo interno o externo, enfermedades, etc.

La comparación entre distintos perfiles de expresión genética es una herramienta básica para poder responder a un gran número de cuestiones biológicas. Cabe destacar la identificación de los genes de un organismo que se activan durante un ciclo celular o una producción de proteínas para el exterior, así como el estudio del efecto de enfermedades en las expresiones genéticas de roedores, primates y humanos.

Para poder llevar a cabo estos objetivos se han utilizado distintas tecnologías con el fin de obtener, almacenar y analizar esta información. La tecnología basada en Microarrays ha supuesto en los últimos años una gran revolución, ya que permite el almacenamiento en un soporte sólido de la respuesta simultanea de miles de genes frente a una serie de condiciones de laboratorio: diferentes muestras de tejidos en distintas condiciones, estados de temperatura y humedad determinados, procesos celulares completos, etc.

Esta tecnología surgió del trabajo pionero de un grupo de científicos de Stanford y el NIH, entre otros [14], y se ha convertido en una potente técnica para la medición de datos de expresión genética. El Microarray, también llamado DNA chip o Biochip, es un soporte sólido construido normalmente en cristal o en membrana de nylon. Son diversas las técnicas usadas para construir estos chips biológicos integrados: Fotolitografía, robots piezoeléctricos, uso de haces de fibra óptica, etc.

El resultado final del proceso de creación de un Microarray [33] es una matriz sobre un Biochip en la que las filas se corresponden con genes y las condiciones experimentales se sitúan en las columnas. El color de cada celda simboliza el grado de expresión genética de dicho gen frente a una determinada condición experimental. Los niveles de expresión presentes en las celdas serán posteriormente cuantificados para poder ser tratados de manera automática [5], utilizándose para ello representaciones mediante números reales de dichos niveles de expresión. Estos datos serán normalizados y transformados antes de trabajar con ellos para disminuir las variaciones y hacer los cálculos posteriores más sencillos.

Así pues, las bases de datos de expresión obtenidas mediante esta tecnología representan un gran potencial para la extracción de conocimiento útil: qué genes se expresan de forma desmesurada, positiva o negativamente, o qué patrones de comportamiento común existen entre las distintas agrupaciones de genes en la matriz. Como aspecto negativo destacar la enorme complejidad y coste que supone la obtención de biochips, hasta tal punto que algunos científicos lo consideran prohibitivo.

La aplicación de técnicas basadas en Microarrays constituye un campo muy amplio, en el que cabe citar el estudio de la expresión genética ante sustancias tóxicas, el diagnóstico de enfermedades, el estudio de enfermedades genéticas complejas, detectar polimorfismos y mutaciones, el análisis del comportamiento de la célula ante fármacos, etc.

3. CLUSTERING

La predicción y la descripción son dos de los objetivos principales que persigue la Minería de Datos. Dentro del ámbito de la descripción, las técnicas de clasificación nos ayudan a organizar un conjunto de datos mediante la asignación de clases. Una de las técnicas de clasificación no supervisada más importante es el Clustering, cuyo objetivo es el de formar grupos o clases de datos, llamados clusters, de tal forma que los datos de un mismo grupo comparten una serie de características y similitudes mientras que los datos de grupos distintos tienen mayores diferencias [25].

Las técnicas de Clustering han probado ser de gran utilidad a la hora de comprender la funcionalidad de los genes, su regulación, los procesos celulares y los distintos subtipos de células existentes [17]. Una de las mayores tareas en el análisis de datos de expresión genética es la de descubrir grupos de genes que intervienen en una misma función celular o que están regulados de la misma manera, promoviendo incluso la comprensión de la funcionalidad de ciertos genes de los cuales no existía conocimiento previo [37]. La búsqueda de secuencias comunes de ADN en las regiones organizadoras de los genes que se encuentran en

un mismo cluster permiten la identificación de los elementos reguladores dentro de dicho cluster [10, 37]. Finalmente, la aplicación de técnicas de Clustering sobre las condiciones experimentales puede revelar subtipos de células que serían difíciles de descubrir utilizando los tradicionales métodos morfológicos [2, 20].

Normalmente, un Microarray consta de entre 10^3 y 10^4 genes, y se espera que este número llegue al orden de 10^6. Sin embargo, el número de condiciones experimentales es generalmente menor que 100. Esta configuración de los atributos y ejemplos hace que sea significativo aplicar las técnicas de Clustering tanto a genes como a condiciones. Mediante las técnicas de Clustering basadas en genes, éstos son agrupados en clusters dependiendo de la evolución de sus valores de expresión a lo largo de todas las condiciones experimentales [7, 17]. En el caso contrario, tenemos las técnicas de Clustering basadas en condiciones en las que éstas se agrupan en función de sus características teniendo en cuenta todos los genes. Cada cluster se corresponde en este caso con un determinado fenotipo macroscópico, es decir, un tipo de cáncer, un tipo de tejido o un síndrome clínico determinado [20].

Algunos algoritmos de Clustering convencionales, como el *K-means* [4, 30] o las técnicas jerárquicas [13, 28], pueden ser usados indistintamente para agrupar genes o condiciones. Pero en la mayoría de los casos esta distinción implica cambios en las tareas a realizar para la agrupación de objetos similares. También se han desarrollado técnicas nuevas enfocadas a este tipo tan específico de bases de datos, las más importantes basadas en teoría de grafos, consiguiendose muy buenos resultados [7, 34].

4. BICLUSTERING

La aplicación de algoritmos de Clustering sobre datos de expresión genética presenta una importante dificultad. El conocimiento actual sobre los procesos celulares nos hace esperar que un subconjunto de genes sea co-regulado y co-expresado sólo bajo ciertas condiciones experimentales, mientras que bajo otras condiciones estos genes pueden comportarse de forma independiente. El descubrimiento de estos patrones locales de comportamiento puede ser la clave para descubrir pathways genéticos (relaciones entre los distintos componentes que intervienen en un proceso bioquímico) que de otra manera serían difíciles de conocer.

Cuando se aplica un algoritmo de Clustering, cada gen en un cluster de genes es definido usando para ello todas las condiciones experimentales. De forma similar, cada condición en un cluster de condiciones se caracteriza mediante el análisis de todos los genes de la matriz. Sin embargo, las técnicas de Biclustering pueden realizar el agrupamiento en las dos dimensiones de forma simultánea, generando así un modelo local. Es decir, cada gen es seleccionado utilizando para ello sólo un subconjunto de condiciones y en la elección de cada condición de un bicluster sólo intervienen un subgrupo de genes. Por lo tanto, el objetivo principal de las técnicas de Biclustering consiste en identificar subgrupos de genes y subgrupos de condiciones aplicando Clustering sobre las filas y columnas de forma simultánea.

Estos métodos son ideales cuando en los datos se dan algunas situaciones tales como que sólo un pequeño grupo de genes participe en un proceso celular de interés, que una interesante actividad celular sólo se produzca bajo un subconjunto de condiciones experimentales, o que un sólo gen pueda participar en múltiples pathways que no se den en todas las condiciones experimentales.

Otra de las diferencias entre ambas técnicas reside en que los clusters forman grupos disjuntos tanto de filas como de columnas. En el caso de los biclusters, las restricciones son mucho menores ya que se tratan de submatrices que no tienen que ser ni exclusivas ni exhaustivas, es decir, un gen o una condición podría pertenecer a un bicluster, a más de uno o a ninguno, tal y como se muestra en la figura 1. Así, la falta de reglas estructurales en los biclusters permite una gran libertad a los distintos métodos. Por este motivo, las técnicas de Biclustering tienen que garantizar con mayor énfasis que los resultados obtenidos posean validez biológica.

4.1. DEFINICIÓN FORMAL DE BICLUSTER

El objeto de estudio de las técnicas de Biclustering sobre datos de expresión genética será una matriz $\mathcal{M}$ con unas dimensiones dadas (N filas o genes y M columnas o condiciones), en la que cada elemento m_{ij} será, generalmente, un número real que representa el nivel de expresión del gen i bajo la condición experimental j.

Es posible, por tanto, definir la matriz del Microarray $\mathcal{M}$ como un conjunto de filas $X = \{x_1, ..., x_n\}$ y un conjunto de columnas $Y = \{y_1, ..., y_m\}$. Se define un bicluster $\mathcal{B} = (\mathcal{I}, \mathcal{J})$ a partir de una Matriz $\mathcal{M}$ como un subconjunto de filas y un subconjunto de columnas donde $I = \{i_1, ..., i_k\}$ es un subconjunto de genes ($I \subseteq X$ y $k \leq n$) y $J = \{j_1, ..., j_s\}$ es un subconjunto de condiciones experimentales ($J \subseteq Y$ y $s \leq m$).

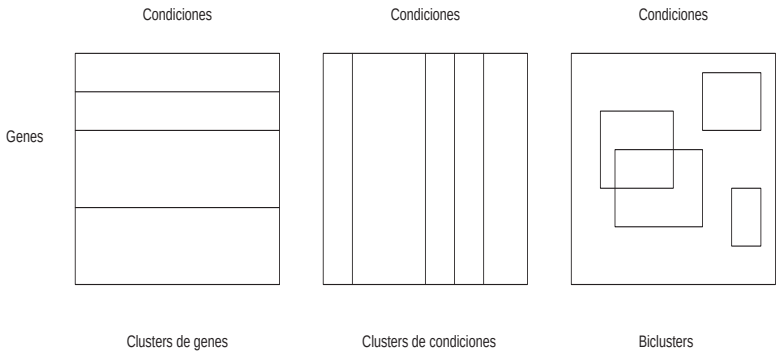

Figura 1: Diferencias estructurales entre clusters y biclusters.

A partir de las formulaciones anteriores, el problema del Biclustering se define como sigue: dada una matriz $\mathcal{M}$ se desea identificar un grupo de biclusters B_l contenidos en ella, de tal forma que cada bicluster B_l satisfaga ciertas características de homogeneidad y con el mayor tamaño posible. Las características de homogeneidad son muy variadas y de ellas depende en gran medida el algoritmo de Biclustering que se utilice.

Debido a la falta de restricciones estructurales comentadas en la anterior sección, el problema de obtener todos los posibles biclusters de una matriz supone una complejidad NP-completa. En su forma más simple, la matriz $\mathcal{M}$ es una matriz binaria en la que cada elemento m_{ij} tiene el valor 0 o 1. En este caso particular, un bicluster corresponde a un subgrafo en el correspondiente grafo bipartito. Encontrar un bicluster de tamaño máximo es, por lo tanto, equivalente a encontrar un subgrafo con el máximo número de aristas en un grafo bipartito, problema que tiene una complejidad NP-completa [31]. En casos más complejos, en los que los valores numéricos de la matriz de partida son tomados en cuenta para determinar la calidad de un bicluster, la complejidad no es menor que la comentada anteriormente. Es por ello que la gran mayoría de algoritmos utilizan técnicas heurísticas para encontrar biclusters, en muchos casos precedidas de un preprocesamiento de los datos con el fin de hacer más evidentes los patrones de interés.

4.2. ALGORITMOS DE BICLUSTERING

En la literatura existe una gran diversidad de técnicas de Biclustering propuestas. Todos ellas comparten el mismo objetivo, agrupar sub-conjuntos de genes que evolucionen de forma similar bajo sub-conjuntos de condiciones experimentales, aunque éste se intenta alcanzar de maneras muy dispares. El motivo de esta variedad es la falta de restricciones estructurales que tienen los biclusters. Este hecho, unido a los escasos conocimientos que se tiene de las reglas y pautas que rigen el comportamiento de los genes, hace que sea muy complejo definir cuál es el objetivo final de un algoritmo de Biclustering.

Debido a la gran variedad de algoritmos propuestos es necesario hacer una clasificación de los mismos basada en el tipo de método utilizado para encontrar submatrices en una matriz de datos.

4.2.1. Combinación de Clustering sobre filas y columnas

Los algoritmos basados en esta técnica consisten en aplicar las originales técnicas de Clustering a cada una de las dos dimensiones: filas y columnas, para después combinar los resultados. Hay que tener en cuenta que al ser una adaptación de las técnicas ya existentes de Clustering, estos métodos heredan tanto sus ventajas como sus limitaciones.

Getz et al. [18] fueron de los primeros autores en aplicar este tipo de técnicas. El objetivo principal que persigue su algoritmo CTWC (Coupled Two-Way Clustering) es el de identificar parejas de pequeños subconjuntos de atributos y objetos, pudiendo asociar ambos tanto a filas como a columnas. CTWC presenta un proceso iterativo en el que los resultados de aplicar técnicas de Clustering sobre una matriz de entrada se combinan para crear submatrices (biclusters), que serán las entradas de la siguiente iteración. El algoritmo propuesto establece un marco genérico en el que poder utilizar cualquier tipo de técnica de Clustering para obtener submatrices significativas de datos. En su artículo, Getz et al. utilizan una técnica jerárquica de Clustering denominada SPC (Super-paramagnetic clustering

of data) [9], que consta de un parámetro que controla la progresiva división de los clusters y la estabilidad de los mismos. La posibilidad del uso de distintos algoritmos de Clustering proporciona mucha libertad a la hora de aplicar el método, aunque los resultados pueden variar mucho dependiendo de la técnica utilizada.

4.2.2. Búsqueda voraz iterativa

A estas clases de algoritmos (Greddy Algorithms) corresponden aquellos métodos que intentan obtener biclusters de la manera más rápida y simple posible. Para ello, aplican una meta-heurística basada en la maximización de criterios locales. Es decir, intentan simplificar la búsqueda tomando decisiones locales con la esperanza de encontrar la solución óptima global. Por ello se pierde calidad de resultados pero se gana sencillez y velocidad a la hora de resolver los problemas.

Dentro de los algoritmos de Biclustering que utilizan búsqueda voraz se encuentra uno de los más conocidos por haber sido el primer trabajo que utilizó el nombre de bicluster aplicado a datos de expresión genética. Cheng y Church [12] establecieron un método que basaba la búsqueda del resultado final en un problema de optimización. Parten de la suposición de que para que un subgrupo de genes y condiciones sea un bicluster, sus valores han de evolucionar al unísono, y esta característica está representada por un valor estadístico: el residuo cuadrado medio (*Mean Squared Residue* (MSR)). El algoritmo toma como entrada la matriz original de datos y un valor umbral de residuo. Su objetivo es encontrar un sólo bicluster en cada ejecución y para ello lleva a cabo dos fases principales. En la primera de ellas, el algoritmo elimina filas y columnas de la matriz original, buscando disminuir el residuo. En la segunda fase, filas y columnas son añadidas a la submatriz obtenida en la fase anterior, teniendo en cuenta no superar el umbral fijado para el residuo. El proceso finaliza cuando un posible aumento del tamaño de la submatriz hace que el valor de su residuo supere el umbral fijado en los parámetros de entrada. Así pues, el resultado final es la submatriz máxima que cumple la condición de no sobrepasar el valor umbral fijado para el MSR.

La propuesta de Cheng y Church ha sido de gran relevancia al ser la primera que introducía el concepto de bicluster en datos de tipo biológico, y que usaba un algoritmo original para su obtención. Ha sido fuente de múltiples estudios y muchos trabajos han sido generados con objetivo de mejorar sus debilidades. Recientes estudios sobre el residuo elaborado por otros autores han mostrado que dicha medida no es adecuada para cuantificar la calidad de los biclusters encontrados [1]. Otros algoritmos de biclustering basados en búsqueda voraz iterativa son FLOC [39] o OPSM [6].

4.2.3. Búsqueda exhaustiva

Algunos autores prefieren basar su búsqueda de biclusters en una enumeración exhaustiva de todas las posibles sub-matrices existentes en la matriz de datos. Sin embargo, este tipo de métodos conllevan un gran coste computacional. Para solucionar este problema, se añaden restricciones para reducir el espacio de búsqueda en base a diversas variables: tamaño, número de condiciones, número de genes, residuo, etc.

Tanay et al. [36] utilizan una combinación de teoría de grafos y de análisis estadístico en su trabajo, SAMBA (Statistical-Algorithmic Method for Bicluster Analysis). En este trabajo se busca un tipo especial de coherencia entre los datos de un bicluster basada en la activación (aumento de la cantidad relativa de ARNm o aumento del valor de expresión) de los genes a lo largo de una serie de condiciones. Para ello se hace uso de la representación en forma de grafo bipartito de los datos de entrada. Los dos conjuntos de vértices estarán formados por los genes y las condiciones. Si una arista une un gen con una condición, significará que ese gen ha respondido de forma positiva a esa condición, estableciendose además un peso a cada una de las aristas. SAMBA comienza con una discretización de los datos, y posteriormente se considerará que un gen está activo si su valor normalizado es superior a 1 y no activo si este valor es inferior a -1. Para tener en cuenta la evolución de los valores de expresión se le añade a cada arista, que representa una activación, un signo que indicaría un aumento o disminución de la expresión genética. Por lo tanto, tendríamos un grafo con tres tipos de relaciones binarias: actividad positiva, actividad negativa y no actividad. De esta forma se podrá buscar un sub-grafo en el que los genes tengan la misma tendencia o incluso la tendencia opuesta. Posteriormente, se realiza una búsqueda de los k subgrafos más pesados en el grafo. El algoritmo aplica iterativamente la mejor modificación para cada bicluster hasta que no sea posible mejorar su peso, filtrando también aquellos sub-grafos que son similares.

4.2.4. Identificación de parámetros de distribución

Este tipo de investigaciones intenta aproximar el concepto de bicluster a partir de un modelo estadístico. El objetivo más importante será el encontrar los parámetros de distribución necesarios para generar las submatrices que se ajusten a dicho modelo estadístico. Son técnicas con un gran contenido matemático y se distinguen por las distintas caracterizaciones que llevan a cabo de los biclusters. El álgebra lineal y la teoría matricial son de las herramientas matemáticas más utilizadas dentro de este contexto.

Lazzeroni y Owen [29] introdujeron el concepto de *plaid models*. Su idea principal se basa en considerar a la matriz de genes y condiciones como una superposición de capas, siendo cada una de ellas un subconjunto de filas y columnas con unos valores de expresión concretos. Estos valores son representados en la matriz de entrada a partir de una determinada coloración en función del nivel de expresión, formándose así una matriz coloreada. El orden de las filas y columnas de una matriz como la descrita anteriormente puede ser arbitrario. Es habitual considerar una reordenación para agrupar filas y columnas similares y así conseguir una matriz formada por bloques, cada uno de ellos formados por valores de un color similar. Así, la matriz de entrada se considera como una suma de capas, siendo cada capa considerada como un bicluster.

4.2.5. Búsqueda estocástica

Se define proceso estocástico como aquel proceso aleatorio que evoluciona con el tiempo. De esta manera, e imitando procesos existentes en la naturaleza, estas técnicas se basan en dichos procesos para encontrar biclusters válidos en una matriz de entrada. La búsqueda estocástica es además una manera de superar los problemas de localidad de ciertos algoritmos voraces, como el utilizado por Cheng y Church [12].

Como ejemplo de aplicación de este tipo de técnicas caben destacar los trabajos de Bryan et al. [11], basados en enfriamiento simulado, así como de Aguilar y Divina [16], que proponen un algoritmo evolutivo [19] de búsqueda de biclusters llamado SEBI, cuyo objetivo es el de encontrar biclusters de tamaño máximo, con un valor del residuo de Cheng y Church [12] inferior a un umbral dado, con un valor elevado de varianza de fila y con un bajo nivel de solapamiento entre las submatrices encontradas. El algoritmo SEBI es un proceso iterativo cuyo número de repeticiones viene dado por una condición de parada. En cada iteración se genera un bicluster máximo cuyo valor de residuo es menor que un umbral introducido como parámetro de entrada. Este bicluster es almacenado en una lista de resultados y el proceso comienza de nuevo. Inicialmente, la población consistirá en biclusters conteniendo un sólo elemento de la matriz de entrada, teniendo además la propiedad de poseer un valor nulo de residuo. Los individuos de la población evolucionarán en un proceso de refinamiento hasta llegar a convertirse en las soluciones óptimas. Al final de cada generación, un determinado número de individuos de la población es seleccionado de forma aleatoria y los mejores de entre ellos serán elegidos como padres, que posteriormente serán cruzados y mutados para dar lugar a la siguiente generación. Para determinar la calidad de un individuo se utiliza una función de ajuste que es combinación de varios criterios: residuo, volumen, varianza de fila y nivel de solapamiento con otros biclusters. Al finalizar el proceso evolutivo, si el mejor individuo de la nueva generación es considerado como un bicluster válido, es decir, con un valor de residuo menor que un umbral, unos límites mínimos de tamaño establecidos y con un nivel de solapamiento aceptable, será devuelto como resultado por el algoritmo SEBI.

5. EVALUACIÓN DE RESULTADOS

Una vez que se han obtenido determinados resultados según las técnicas de Clustering o Biclustering, el siguiente paso es determinar si éstos son validos para los datos procesados, y poder así discriminar entre los distintos métodos e incluso poder elegir la mejor técnica entre las validadas correctamente.

Las técnicas de validación en Bioinformática pueden clasificarse en validación estadística, basadas en estudios sobre los datos y validación biológica, basadas en la comparación del conocimiento obtenido con algún otro extraído de forma biológica experimental. El segundo grupo es un caso particular de evaluación basada en conocimiento previo, que aporta una estimación del modelo extraído desde el punto de vista biológico. Las técnicas pertenecientes al primer grupo aportan una valoración desde un punto de vista estadístico.

Las técnicas de validación estadística pueden ser dividas en medidas de validación externas o internas [22]. Estos dos grupos de técnicas difieren fundamentalmente en sus enfoques, y encuentran aplicación en distintos marcos experimentales. Las medidas de validación externa comprenden a todos los métodos que evalúan los resultados de un cluster basándose en el conocimiento

de las etiquetas correctas de la clase [24, 38]. Evidentemente, ésta es de gran utilidad al permitir una evaluación y comparación totalmente objetiva de los algoritmos de clustering o biclustering sobre un conjunto de datos, para los cuales las etiquetas de la clase corresponden con la estructura verdadera de cluster. En el caso donde no se disponga esta clase, o la clase sea dudosa, se deberá utilizar una medida de validación interna [23, 8]. Las técnicas de validación interna no usan un conocimiento adicional, sino que sólo se basan en la información intrínseca de los datos [15, 40].

Los métodos que consisten en la comparación de los resultados generados con la información biológica real pueden ser divididos en dos fases. Una primera, donde es necesario obtener y procesar los datos biológicos que nos servirán como conocimiento real. Y otra segunda, en donde se compara la solución obtenida y la solución real. El conocimiento biológico recogido hasta el momento ayuda a la interpretación biológica de cualquier resultado generado por un aprendizaje automático. Las nociones de biología actuales puede estar recogidas de dos formas; en ontologías [3, 27] o en bases de datos biomédicas [26].

6. CONCLUSIONES

La Bioinformática es una nueva disciplina que surge con la necesidad de computar grandes cantidades de datos biológicos en un tiempo razonable. El análisis de datos de expresión genética se ha convertido en uno de sus objetivos principales. Para llevar a cabo dicho objetivo se hace uso de técnicas relacionadas con la minería de datos. Dentro de la minería se pueden encontrar multitud de métodos, tanto de extracción de conocimiento como de evaluación de los resultados obtenidos. Sin embargo, la gran mayoría de esas técnicas no pueden ser aplicables a la Bioinformática, debido al conjunto de datos tan particular que ésta maneja (Microarrays). Por ello, los trabajos publicados en los últimos años tratan de resolver esta problemática desde el paradigma de la biología. En este trabajo se han presentado las distintas áreas que abarca esta nueva disciplina, así como las diferentes técnicas de análisis de datos de tipo biológicos más utilizadas en la actualidad.

Referencias

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

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

[3] M. Ashburner. et al. Gene ontology: tool for the unification of biology. The Gene Ontology. *Nature Genetics*, 25:25–29, 2000.

[4] G. Ball and D. Hall. A clustering technique for summarizing multivariate data. *Behaviorial Sciences*, 12(2):153–155, 1967.

[5] P. Barrero. Aplicaciones de la técnica de microarrays en ciencias biomédicas: presente y futuro. *Química viva*, 3(4):1–10, 2005.

[6] A. Bellaachia and D. Portnoy. et al. E-CAST: A Data Mining Algorithm for Gene Expression Data. 2002.

[7] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3-4):281–297, 1999.

[8] J. Bezdek and N. Pal. Some new indexes of cluster validity. *IEEE Trans. Syst. Man Cybernet.*, 28:301–315, 1998.

[9] M. Blatt, S. Wiseman, and E. Domany. Super-paramagnetic clustering of data. *Physical Review Letters*, 76:3251–3254, 1996.

[10] A. Brazma and J. Vilo. Minireview: Gene expression data analysis. *Federation of European Biochemical societies*, 480:17–24, 2000.

[11] K. Bryan, P. Cunningham, and N. Bolshakova. Biclustering of expression data using simulated annealing. In *18th IEEE Symposium on Computer-Based Medical Systems*, pages 383–388, Dublin, Ireland, 2005.

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

[13] R. M. Cormarck. A review of classificiation (with discussion). *J. royal Statustical Society, Series A*, 134:321–367, 1971.

[14] J. DeRisi, L. Penland, P. Brown, M. Bittner, P. Meltzer, M. Ray, Y. Chen, Y. Su, and J. Trent. Use of a cdna microarray to analyse gene expression patterns in human cancer. *Nature. Genetics*, 14:457–460, 1996.

[15] C. Ding and C. He. K-nearest neighbor consistency in data clustering: incorporating local information into global optimization. In

Proceedings of the 2004 ACM Symposium on Applied Computing, ACM Press, pages 584–589, New York, 2004.

[16] F. Divina and J. S. Aguilar-Ruiz. Evolutionary biclustering of microarray data. *Lecture Notes on Computer Science*, 3449:1–10, 2005.

[17] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc.Natl.Acad.Sci.*, 95(25):14863–14868, 1998.

[18] G. Getz, E. Levine, and E. Domany. Coupled two-way clustering analysis of gene microarray data. In *NaturalAcademy of Sciences*, pages 12079–12084, USA, 2000.

[19] D. E. Goldberg. *Genetic Algorithms in Search, Optimization & Machine Learning.* Addison Wesley, 1989.

[20] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gassenbeek, J. Mesirov, H. Coller, M. Loh, J. Downing, M. Caligiuri, D. Bloomfield, and E. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[21] J. Hagen. The origin of bioinformatics. *Nat Rev Genet.*, 1(3):231–236, 2001.

[22] M. Halkidi. et al. On clustering validation techniques. *IJ. Intell. Inform. Syst*, 17:107–145, 2001.

[23] J. Handl, J. Knowles, and D. B. Kell. Computational cluster validation in postgenomic data analysis. *Bioinformatics*, 21(15):3201–3212, 2005.

[24] A. Hubert. Comparing partitions. *J. Classif.*, 2:193–198, 1985.

[25] D. Jiang, C. Tang, and A. Zhang. Cluster analysis for gene expression data: A survey. *IEEE Transactions on Knowledge and Data Engineering*, 16(11):1370–1386, 2004.

[26] M. Kanehisa and S. Goto. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res.*, 28:27–30, 2000.

[27] P. Khatri and S. Draghici. Ontological analysis of gene expression data: current tools, limitations, and open problems. *Bioinformatics, In Press*, 2005.

[28] G. Lance and W. Williams. A general theory of classification sorting strategies.1.hierarchical systems. *The computer journal*, 9:373–380, 1967.

[29] L. Lazzeroni and A. Owen. Plaid models for gene expression data. In *Technical report, Stanford University*, 2000.

[30] E. Lehmann and H. DÁbrera. *Nonparametrics: Statistical Methods Based on Ranks.* Prentice Hall, 1998.

[31] R. Peeters. The maximum edge biclique problem is np-complete. *Discrete Applied*, 131:651–654, 2003.

[32] J. Pevsner. *Bioinformatics and Functional Genomics.* Wiley-Liss, 2003.

[33] D. Shalon, S. Smith, and P. Brown. A dna microarray system for analyzing complex dna samples using two-color fluorescent probe hybridization. *Genome Res*, 6:639–645, 1996.

[34] R. Sharan and R. Shamir. Click: A clustering algorithm with applications to gene expression analysis. In *In Proceedings of the eighth International Conference on Intelligent Systems for Molecular Biology (ISMB), AAAI Press*, pages 307–316, 2000.

[35] F. M. Sánchez, G. L. Campos, and N. I. de Andrés. *Impacto de la bioinformática en las ciencias biomédicas. Servicios de Salud: ¿estrategias o tecnologías?* Editorial Médica Panamericana, 1999.

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

[37] S. Tavazoie, D. Hughes, M. Campbell, R. Cho, and G. Church. Systematic determination of genetic network architecture. *Nature Genetics*, pages 281–285, 1999.

[38] C. van Rijsbergen. Information retrieval. *2nd. edn. Butterworths*, 1979.

[39] J. Yang and W. Wang. Enhanced biclustering on expression data. In *3rd IEEE Conference on Bioinformatics and Bioengineering*, pages 321–327, 2003.

[40] K. Y. Yeung. et al. Validating clustering for gene expression data. *Bioinformatics*, 17:309–318, 2001.