Extracción de conocimiento en repositorios biológicos estructurados

Dra. Rocío C. Romero Zaliz

Departamento de Computación

Facultad de Ciencias Exactas y Naturales

Universidad de Buenos Aires

Argentina

Ciudad Universitaria - Pabellón I

rromero@dc.uba.ar

Resumen

Los avances en la biología molecular y el desarrollo de nuevas técnicas computacionales permiten la investigación exhaustiva de procesos moleculares de gran complejidad que ocurren en los sistemas biológicos. El desarrollo continuo de grandes y sofisticados repositorios de información y conocimiento facilita el acceso a una vasta cantidad de datos. Paradójicamente, el uso de estas bases de datos se ve parcialmente limitado por la dificultad de buscar en ellas en términos que

satisfagan las necesidades y expectativas de los usuarios. Por ejemplo, resulta muy difícil identificar las características relevantes que describen un sistema en una base de datos altamente interrelacionada. Estos repositorios son estructurados, en el sentido que describen a los objetos que almacenan por las relaciones entre sus características y no solamente por las características en sí mismas. La extracción de conocimiento a partir de conjuntos de datos es un proceso básico en la minería de datos (*data mining*). En particular, las técnicas de agrupamiento o *clustering* permiten extraer conocimiento que se encuentra oculto a simple vista. Estas técnicas no necesitan la supervisión de un experto, en contraposición a los métodos de clasificación más utilizados, como ser las redes neuronales no supervisadas o los árboles de decisión. La utilización de técnicas clásicas para inferir patrones o perfiles similares deben ser adaptados de forma tal que puedan ser aplicados a repositorios de datos biológicos estructurados, para poder así optimizar el nuevo conocimiento adquirido. En este trabajo se introducirán los problemas y alcances de la minería de datos en bioinformática, específicamente en lo que concierne a las técnicas de agrupamiento, y se presentará brevemente a modo de ejemplo la aplicación de clustering al problema de la respuesta inflamatoria en humanos utilizando información ontológica.

Knowledge extraction in structural biological repositories. Advances in molecular biology and new computational techniques are enabling researchers to systematically investigate the complex molecu-

lar process underlying biological systems. The continued development of large, sophisticated repositories of knowledge and information has facilitated the accessibility to vast amounts of biological data (e.g., cis-regulatory features, metabolic pathways, regulatory networks). However, paradoxically, the usefulness of these databases is partially limited by the inability to search those databases in terms that match the needs and experience of their users. For example, researchers usually get lost when they try to identify the distinguishing features that describe their target systems in highly interconnected databases. These repositories are structural in the sense that they describe objects in terms of their characteristics, but also in terms of the relationships between these characteristics. Database knowledge extraction is a basic topic in the *data mining* field. In particular, *clustering* techniques allow to uncover hidden information inside large repositories in an unsupervised fashion. This is in contrast to other classical machine learning approaches, like neural networks or decision trees, that need supervised information provided by an expert that is not always available. In spite of the recent renewed interest in knowledge-discovery techniques, there is a dearth of data analysis methods intended to facilitate understanding of the represented objects and related systems using structural data. Therefore, customization of classical methods are necessary to achieve the best possible results in the search of useful information in structural biological databases. In this work we introduce the problems and advances of data mining in bioinformatics, in particular refering to

clustering techniques. As an example, we breifly present the aplication of clustering to the mmuno-inflammatory response problem in humans using ontology information.

Keywords: Data-mining, Bioinformatics, Data bases, Clustering.

1. Introducción

Durante las últimas décadas, se ha estado acumulando conocimiento, proveniente de diferentes áreas dentro de la biología, en repositorios de datos digitales. Al estar almacenada de esta manera, la información puede ser estudiada y compartida más fácilmente por distintos expertos. A pesar de esto, el aumento constante del tamaño de estos repositorios hace casi imposible, para un ser humano, extraer información útil de los mismos. Por esta razón, se han desarrollado diversas técnicas de minería de datos o *data mining* para poder revelar información oculta en grandes colecciones de datos (Hand et al., 2001; Hernández et al., 2004) y así poder ayudar a los usuarios. Estas técnicas funcionan correctamente con representaciones de datos en forma atributo-valor, es decir, datos no estructurados. Sin embargo, muchos proyectos de adquisición de datos actuales acumulan información estructurada que describe no sólo los objetos de la base de datos, sino también las relaciones que existen entre ellos. Estos conjuntos de datos son estructurados en el sentido de que los objetos que almacenan están descritos por las relaciones

entre las características y no solamente por las características en sí mismas. Debido a ello, existe la necesidad de crear técnicas que permitan analizar y descubrir conceptos definidos mediante subestructuras en repositorios de datos estructurados.

2. Bases de datos

Existen diferentes clases de bases de datos –relacionales, orientadas a objetos, etc.– utilizadas para almacenar información biológica. Al contrario de lo que normalmente se piensa, GenBank (Bension et al., 2005) no es una base de datos. GenBank es un "archivo", el cual acumula información en forma plana, es decir, en archivos de texto convencionales. Por otro lado, Entrez (Schuler et al., 1996) es una base de datos relacional que integra varias fuentes de información heterogénea, entre las cuales se encuentra la propia GenBank y PubMed (PubMed Web Page,). La diferencia entre un archivo y una base de datos radica en su estructura. Las bases de datos, al contener la información organizada, permiten realizar búsquedas más rápidas, valiéndose de uno o varios sistemas de indexación según la consulta o *query* que se esté pidiendo. Esta ventaja resulta vital en repositorios de gran volumen, como los utilizados en las ciencias naturales, incluyendo los de biología molecular.

Si bien existen diversas técnicas para organizar los datos de forma tal que ocupen un menor espacio, que se pueda encontrar un dato de manera más rápida, o bien, que se puedan llegar a realizar consultas más complejas; exis-

ten en muchos casos relaciones no sólo entre los objetos acumulados en estas bases de datos, sino también entre sus atributos. A esta clase de repositorios se los llama *estructurados*.

Bases de datos estructuradas

Las bases de datos estructuradas almacenan objetos, los cuales están descriptos en base a diversas características relacionadas. Un ejemplo común son las *ontologías*, tales como Gene Ontology (GO) (Consortium, 2000). Las ontologías biológicas (Schulze-Kremer, 2002) son un intento por definir, no solo un vocabulario común, mediante la especificación de términos no ambiguos, sino también las relaciones existentes entre los distintos términos. Una ontología se puede llegar a ver como un *grafo* en donde sus nodos son los términos aceptados en el vocabulario y sus ejes constituyen las relaciones entre estos términos. Los ejes pueden ser dirigidos o no, dependiendo de la relación que los une.

La búsqueda en bases de datos estructuradas es más compleja que en otras bases de datos, ya que debe tener en cuenta su estructura para poder así extraer toda la información posible. Por ejemplo, supongamos que se desea buscar aquellos genes de un organismo particular que participan en el proceso biológica de la *fotosíntesis* utilizando la información de GO. Sin tener en cuenta la estructura de la ontología, una herramienta computacional simple buscaría todos aquellos genes que tengan anotados el término `GO:0015979` (*photosynthesis*), el cual se encuentra a nivel 3 dentro de la ontología de *pro-*

ceso biológico de GO. Sin embargo, los genes normalmente estarán anotados con el término de proceso biológico más específico del que se tenga noción, el cual puede encontrarse a niveles más bajos en la jerarquía. Estos términos pueden tener como antecesor al término correspondiente a fotosíntesis, con lo cual heredan, por las relaciones que lo unen (relación "is_a" o "part_of" (Smith et al., 2005)), la participación en ese proceso. Por lo tanto, teniendo en cuenta las relaciones del grafo se puede llegar a aprovechar al máximo el poder de información que provee la ontología.

No solo las ontologías constituyen una forma de información estructurada, también existen otras fuentes de información que contienen relaciones entre los atributos. Por ejemplo, la posición relativa entre distintos componentes de una misma secuencia, tales como la distancia entre un promotor y un sitio de *binding* (Alberts et al., 2003), o bien, en forma genérica, la distancia entre un patrón y otro patrón, la cual puede resultar crucial para que un sistema biológico funcione de una determinada manera.

3. Extracción de conocimiento en bases de datos

Las técnicas de extracción de conocimiento en bases de datos o KDD (*Knowledge Discovery in Databases*) son parte del proceso de data mining. La minería de datos es aplicada a una amplia gama de problemas, ya sea para predecir, caracterizar o clasificar datos. En el ámbito de la biología, la minería

de datos es una herramienta fundamental ya que mediante su uso se puede llegar a analizar una gran cantidad de información, extrayendo patrones o similitudes entre sus componentes. Una de las técnicas más extendidas en data mining es el agrupamiento de datos o *clustering*.

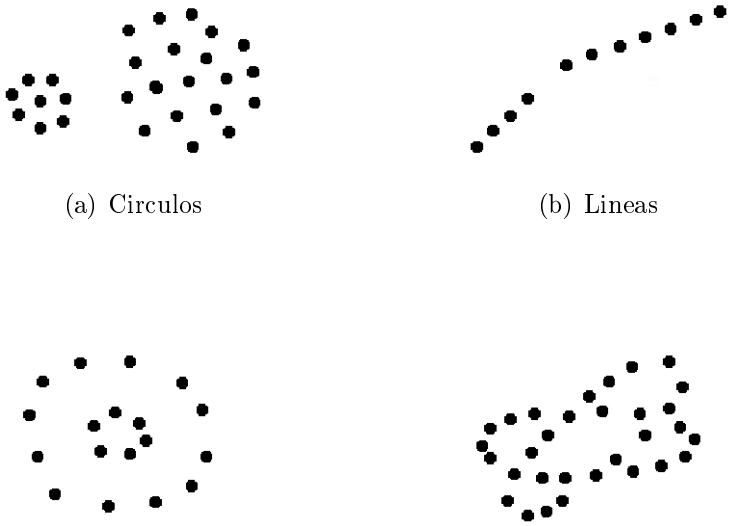
Clustering

El objetivo del agrupamiento de datos es la clasificación de objetos de acuerdo a similitudes entre ellos, para luego organizar estos datos en grupos. Las técnicas de clustering están incluidas entre los métodos de aprendizaje *no supervisado*, debido a que no utilizan conocimiento de identificadores de clases. La mayoría de los algoritmos de clustering tampoco se basan en asunciones comunes, como si lo hacen los métodos estadísticos convencionales, tales como la distribución estadística subyacente de los datos, y por ello son útiles en situaciones donde existe poco conocimiento sobre los mismos. El potencial de los algoritmos de clustering para revelar las estructuras subyacentes de los datos puede ser explotado no sólo para la clasificación y reconocimiento de patrones sino también para la reducción de la complejidad en modelado y optimización.

Las técnicas de clustering pueden aplicarse a datos que sean cuantitativos (numéricos), cualitativos (categóricos) o una mezcla de ambos. Los datos son típicamente observaciones de algún proceso físico.

Pueden formularse varias definiciones de cluster, pero en general se puede aceptar la visión de que un cluster es un grupo de objetos, los cuales son más

similares entre sí que los miembros de otros clusters. El término "similaridad" debería ser entendido como la similaridad matemática, medida formalmente. En espacios métricos, la similaridad está definida comúnmente como una *norma* de distancia. La distancia puede medirse entre los vectores de datos en sí, o como la distancia de un vector de datos a algún objeto prototípico del cluster. Los prototipos no son usualmente conocidos de antemano y los algoritmos de clustering los buscan simultáneamente con la partición de los datos. Los prototipos pueden ser vectores de igual dimensión que la de los objetos de datos, pero se pueden también definir como objetos geométricos de "alto nivel", tales como subespacios lineales y no-lineales, o funciones.



(a) Circulos (b) Lineas

(c) Circunferencias (d) Elipses cruzados

Figura 1: Diferentes formas de los clusters

Los datos pueden revelar clusters de diferentes formas geométricas, tama-

ños y densidades, como se puede ver en la Figura 1. Mientras que los clusters (a) son esféricos, los clusters (b), (c) y (d) pueden caracterizarse como subespacios lineales y no-lineales del espacio de datos. El rendimiento de la mayoría de los algoritmos de clustering está influenciado no solo por la forma geométrica y densidad de los clusters individuales, sino también por las relaciones espaciales y distancias entre ellos. Los clusters pueden estar bien separados, conectados en forma continua, o solapados entre ellos. La separación de los clusters está influenciada por el factor de escala y la normalización de los datos.

Clustering conceptual

El agrupamiento conceptual o *conceptual clustering*, es similar al considerado en el análisis de clusters tradicional, pero está definido de una manera diferente. Dados un conjunto de descripciones en base a atributos de ciertas entidades, un lenguaje de descripción para caracterizar clases de estas entidades y un criterio de calidad de clasificación; el problema consiste en particionar las entidades en clases de tal manera que se maximice el criterio de calidad de clasificación y, simultáneamente, en determinar descripciones generales de estas clases en el lenguaje de descripción dado. Por ello, un método de clustering conceptual busca no sólo una clasificación de las entidades, sino también una descripción simbólica de las clases propuestas. Un aspecto importante que distingue al clustering conceptual es que, a diferencia del análisis de clusters clásico, las propiedades de las descripciones de clases se

toman en consideración en el proceso de determinación de las clases.

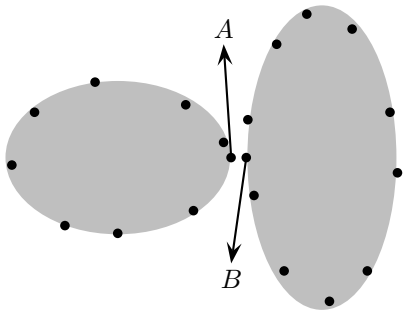


Figura 2: Diferencia entre cercanía y cohesión conceptual

Por tanto, dos objetos pueden ser similares, es decir, cercanos de acuerdo a una cierta medida de distancia (o similaridad), pero pueden tener una baja cohesividad conceptual, y viceversa. Un ejemplo de la primera situación se puede ver en la Figura 2. Los puntos negros A y B son "cercanos" entre sí y, por ello, serían ubicados en el mismo cluster por cualquier técnica basada únicamente en las distancias entre los puntos. Sin embargo, estos puntos tienen una cohesividad conceptual pequeña debido a que pertenecen a configuraciones que representan diferentes conceptos. Si se dispone de un lenguaje de descripción apropiado, un método de clustering conceptual permitirá agrupar los puntos de la Figura 2 en dos "elipses", como lo haría la mayoría de las personas.

Un criterio de calidad de clasificación utilizado en clustering conceptual puede involucrar una variedad de factores, tales como el *ajuste* de la des-

cripción de un cluster a los datos, la *simplicidad* de una descripción u otras propiedades de las entidades o conceptos que los describen.

4. Extracción de conocimiento en bases de datos biológicas estructuradas basada en clustering conceptual

A partir de lo visto en las secciones anteriores, se plantea la posibilidad de utilizar las técnicas de clustering conceptual para extraer información relevante de bases de datos biológicas estructuradas. Existen varias ventajas de utilizar esta clase de técnicas para resolver el problema en cuestión: (1) la propiedad del clustering conceptual de brindar una explicación para cada conjunto de observaciones agrupadas en cada cluster, lo que la convierte en una técnica de *caja blanca* (Babuska, 1998); (2) la facilidad para un usuario, sin mucho conocimientos en técnicas de agrupamiento, de interpretar cada cluster escogiendo de manera apropiada el lenguaje; y (3) la capacidad de trabajar con la estructura existente en la bases de datos, incorporándola al lenguaje seleccionado.

Una vez que se define el lenguaje que se utilizará para realizar el clustering conceptual, es necesario explorar el espacio de posibles conceptos de cluster, es decir, las descripciones que definen cada grupo. Cuanto más descriptivo sea el lenguaje, en base a la variedad de términos y relaciones entre ellos,

más descriptivos serán los clusters. Esta exploración del espacio de conceptos puede realizarse mediante un algoritmo exhaustivo, si el espacio es reducido, o mediante alguna técnica heurística. La búsqueda de estos conceptos debe estar guiada por algún criterio o varios criterios. Como ya se ha mencionado en la sección anterior, un criterio de calidad de clasificación puede involucrar una variedad de factores, algunos de los más utilizados son: el ajuste de la descripción de un cluster a los datos, la simplicidad de una descripción, la distancia entre los clusters identificados, etc. Estos criterios de optimización, a menudo deben ser adaptado al problema en particular que se quiere resolver.

4.1. Optimización multiobjetivo

Muchos problemas reales se caracterizan por la existencia de múltiples medidas de optimización, las cuales deberían ser optimizadas, o al menos ser satisfechas, simultáneamente. Como el propio nombre sugiere, el problema de la optimización multiobjetivo consiste en el proceso de optimización simultánea de más de una función objetivo (Cohon, 1978).

La falta de metodologías para resolver este tipo de problemas llevó a que, en un principio, se resolvieran como problemas mono-objetivo. Sin embargo, no es correcto tomar esta determinación ya que existen diferencias entre los principios en que se basan los algoritmos que tratan un solo objetivo y los que trabajan con varios. De esta forma, al trabajar con problemas mono-objetivo nos enfrentamos con la búsqueda de una solución que optimice esa única función objetivo, tarea distinta a la que se nos plantea al trabajar

con problemas multiobjetivo. En este último caso, no pretendemos encontrar una solución óptima que se corresponda a cada una de las funciones objetivo, sino varias soluciones que satisfagan todos los objetivos a la vez de la mejor manera posible. Como en un problema de optimización con un solo objetivo, también suele existir un número de restricciones que debe satisfacer cualquier solución factible.

Una dificultad común en la optimización multiobjetivo es la aparición de un conflicto entre objetivos (Hans, 1988), es decir, el hecho de que ninguna de las soluciones factibles sea óptima simultáneamente para todos los objetivos. En este caso, la solución matemática más adecuada es quedarse con aquellas soluciones que ofrezcan el menor conflicto posible entre objetivos. Estas soluciones pueden verse como puntos en el espacio de búsqueda que están colocados de forma óptima a partir de los óptimos individuales de cada objetivo, aunque puede que dichas soluciones no satisfagan las preferencias del experto que quiera establecer algunas prioridades asociadas a los objetivos.

Para encontrar tales puntos, todas las técnicas clásicas reducen el vector objetivo a un escalar, es decir, a un único objetivo. En estos casos, en realidad, se trabaja con un problema sustituto buscando una solución sujeta a las restricciones especificadas. Sin embargo, las técnicas clásicas más comunes para afrontar problemas con múltiples objetivos presentan diversos inconvenientes (Deb, 2001).

Debido a ello, se han desarrollado diferentes clases de algoritmos de optimización multiobjetivo (Deb, 2001). La mayoría de estos algoritmos usan el

concepto de *dominancia* en su búsqueda del óptimo. La noción de dominancia entre dos posible soluciones se define intuitivamente como: *"una solución 'x' domina a otra 'y' si 'x' es mejor a 'y' en todos los objetivos a optimizar"*. En otras palabras, si existe al menos un objetivo de 'x' que no supera a 'y' y, a la vez, 'y' no domina a 'x', entonces ambas soluciones se consideran no dominadas ya que no es posible decidir si una es mejor a la otra.

Mediante técnicas que permitan encontrar el conjunto de soluciones no dominadas, es posible brindar diversas soluciones para un mismo problema. Diferentes usuarios preferirán unas soluciones de este conjunto por sobre otras. La posibilidad de obtener varias posibles soluciones es muy útil cuando se trabaja con objetivos contrapuestos ya que proveen de distintas visiones sobre el mismo problema, permitiendo que luego cada especialista extraiga la información que le sea de provecho.

4.2. Ejemplo: aplicación al problema de respuesta inflamatoria en humanos utilizando información ontológica

La técnica de clustering conceptual mediante optimización multiobjetivo en bases de datos estructuradas se ha aplicado al agrupamiento de genes basándose en el lenguaje de términos y relaciones provistos por Gene Ontology (Consortium, 2000). Para ello contamos con un conjunto de genes que provienen de un experimento (Calvano et al., 2005). Para cada uno de ellos

contamos con su descripción en términos de GO. La pregunta que nos hacemos es: ¿en qué se parecen estos genes? Podemos agruparlos comprobando su similitud definiendo conceptos en base a los términos de GO, aprovechando la estructura jerárquica de esta base de información, subiendo o bajando en la ontología para así conseguir conceptos más generales o más específicos respectivamente. A su vez, es de especial interés poder tener en cuenta las tres ontologías definidas en GO de forma simultánea, de esta manera pudiendo generar conceptos más descriptivos que al utilizarlas en forma aislada.

Asimismo, el uso de esta forma de agrupamiento nos permite conseguir clusters que estén definidos por atributos (términos) localmente óptimos, en contraposición con una preselección de características globalmente relevantes (Mitchell, 1997). La exploración del espacio de conceptos permite incorporar o eliminar términos dinámicamente y así conseguir esta selección de atributos especializada para cada cluster.

En esta aplicación, se quieren obtener clusters que sean, a la vez, descriptivos y que representen la mayor cantidad de genes posibles. Dado que estas dos características son contradictorias, ya que a mayor expresividad del cluster menor cantidad de genes representados y vice versa, se ha optado por utilizar una técnica de optimización multiobjetivo. Para un conjunto elevado de genes es necesario el uso de alguna técnica heurística, en nuestro caso utilizamos un algoritmo genético multiobjetivo (Deb, 2001). Entonces, los criterios de optimización utilizados son la *especificidad*, calculada en base al nivel dentro de la jerarquía de GO para cada término y la similitud con cada

una de las observaciones que representa, y la *sensibilidad*, calculada como la cantidad de observaciones que representa el cluster en cuestión.

Como resultado de nuestro experimento, logramos obtener un conjunto de clusters, cada uno de ellos definidos por un concepto en base a los términos de GO (Romero-Zaliz, 2005). En la Tabla 1 se muestran como ejemplo algunos de ellos. Como puede observarse, se obtienen clusters definidos por conceptos muy diferentes, que involucran a términos en distintas ontologías, inclusive a más de un término de una misma ontología. Es necesario aclarar que no está permitido tener, en un concepto de un cluster, más de un término que pertenezca a una misma rama de una ontología, ya que esto brindaría información redundante. Solo el término más específico, es decir el que se encuentre más abajo en la jerarquía, se mantendrá en esos casos.

Una vez obtenidos los clusters y sus conceptos asociados, resulta sencillo utilizar esta información para poder realizar una clasificación de cada gen. Incluso es posible inferir conocimiento nuevo sobre otros genes, utilizando para ello los conceptos de los clusters como modelos. Asimismo se puede realizar una búsqueda de otras instancias de estos modelos en grandes repositorios de datos. De esta manera conociendo información sobre los genes es posible que se pueda asociar ese conocimiento a otros genes que comparten el mismo modelo. Por ejemplo, teniendo información sobre la expresión en el tiempo de un conjunto de genes agrupados conjuntamente, es posible inferir la expresión de otro gen que sea luego clasificado dentaré de este grupo, del cual se desconoce a priori información de su expresión (más información puede

Proceso biológico	**Función molecular**	**Componente celular**
GO:0008151 cell growth and/or maintenance (level: 4)	GO:0005488 binding (level: 2)	GO:0016020 membrane (level: 3) GO:0000267 cell fraction (level: 3)
GO:0006915 apoptosis (level: 6)		GO:0005887 integral to plasma membrane (level: 4)
		GO:0005575 cellular_component (level: 1)

Cuadro 1: Algunos ejemplos de conceptos que definen clusters.

consultarse en (Romero-Zaliz, 2005)).

5. Discusión

Las tareas más simples de la *bioinformática* conciernen a la creación y mantenimiento de bases de datos de información biológica. Secuencias nucleotídicas (y las secuencias proteicas que derivan de las mismas) componen la mayoría de la información que está almacenada en estos repositorios. Mientras que el almacenamiento y organización de millones de nucleótidos está muy lejos de ser una tarea trivial, el diseño de una base de datos y el desarrollo de una interfaz con la cual los investigadores puedan tanto acceder a la información existente como agregar nuevas instancias, es simplemente el

comienzo.

Sin embargo, para que los investigadores puedan beneficiarse de esta información, es necesario cumplir con dos requisitos: (1) tener acceso inmediato al conjunto de secuencias coleccionadas y (2) tener una forma de extraer de este conjunto solamente aquellas secuencias que interesen al investigador. La simple colección, de forma manual, de toda la información necesaria para un proyecto dado a partir de un artículo de revista publicado puede convertirse rápidamente en una tarea epopéyica. Luego de obtener estos datos, es necesario organizarlos y analizarlos.

Es por ello que es necesario el desarrollo de técnicas que permitan extraer información relevante de estos repositorios de datos. En particular, en bases de datos más complejas como son las bases de datos estructuradas que requieren una adaptación de las herramientas de extracción de conocimiento existentes para poder hacer un buen uso de toda la información contenida en esos repositorios.

El eje central de la propuesta que se presenta en este texto se basa en el modelado de una base de datos estructurada y en la utilización de la optimización multiobjetivo para conseguir distintas soluciones que permitan al usuario seleccionar aquellas de su preferencia y retroalimentar a la metodología haciendo uso de esta información. En concreto, hemos considerado el uso de los algoritmos genéticos multiobjetivo para llevar a cabo el clustering conceptual de la información de entrada para así extraer similitudes entre los genes de la base de datos. A esta propuesta la hemos aplicado a una base de

datos de genes, la cual se conocen las anotaciones de cada uno según las ontologías definidas en el proyecto GO, obteniéndose clusters definidos a partir de los términos este proyecto que resultan a la vez precisos e interpretables.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P.: 2003, *Biología molecular de la célula. Cuarta Edición*, Omega

Babuska, R.: 1998, *Fuzzy Modeling for Control*, Kluwer Academic Publishers, Norwell, MA, USA

Bension, D., Karsch-Mizrachi, I., Lipman, D., Ostell, J., and Wheeler, D.: 2005, *Nucleic Acids Research* **34**, 16, Database issue

Calvano, S., Xiao, W., Richards, D., Felciano, R., Baker, H., Cho, R., Chen, R., Brownstein, B., Cobb, J., S.K., T., Miller-Graziano, C., Moldawer, L., Mindrinos, M., Davis, R., Tompkins, R., Lowry, S., ProgramInflamm, L. S. C. R., and to Injury, H. R.: 2005, *Nature* **437(7061)**, 1032

Cohon, J.: 1978, *Multiobjective Programming and Planning*, Academic Press, New York

Consortium, T. G. O.: 2000, *Nature Genet.* **25**, 25

Deb, K.: 2001, *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, Inc.

Hand, D., Mannila, H., and Smyth, P.: 2001, *Principles of Data Mining*, MIT Press, HAN d3 01:1 1.Ex

Hans, A.: 1988, *Multicriteria Optimization in Engineering and Sciences* **19**, 309

Hernández, J., Ramírez, M., and Ferri, C.: 2004, *Introducción a la Minería de Datos (in spanish)*, Pearson Prentice Hall

Mitchell, T.: 1997, *Machine Learning*, McGraw-Hill, New York

PubMed Web Page, http://www.pubmed.gov

Romero-Zaliz, R.: 2005, *Ph.D. thesis*, Universidad de Granada

Schuler, G., Epstein, J., Ohkawa, H., and Kans, J.: 1996, *Methods Enzymol* **266**, 141

Schulze-Kremer, S.: 2002, *In Silico Biol.* **2(3)**, 179

Smith, B., Ceusters, W., Klagges, B., Köhler, J., Kumar, A., Lomax, J., Mungall, C., Neuhaus, F., Rector, A., and Rosse, C.: 2005, *Genome Biology* **6(5)**, R45