Onto-CC: un servidor web para la anotación automática de genes utilizando *gene ontology*

R. Romero-Zaliz, C. del Val, I. Zwir

Resumen— **El vocabulario de *Gene Ontology (GO)* ha sido ampliamente utilizado para analizar las funciones de genes co-expresados. Sin embargo, a pesar de su uso tan extendido en biología y medicina, sigue habiendo altos niveles de incertidumbre sobre cual sub-ontología debe ser utilizada (i.e., proceso biológico, función molecular o componente celular), y a que nivel de especificidad. Por otra parte, la base de datos de GO puede contener información incompleta debido a errores en las anotaciones, o al ser altamente influenciada por el conocimiento disponible sobre una rama específica en una sub-ontología. A pesar de estas desventajas, existe una tendencia a ignorar estos problemas e incluso a utilizar los términos GO para realizar búsquedas de perfiles de expresión de genes (i.e., expresión + GO) en lugar de, simplemente, considerarlos como una fuente independiente de validación (i.e., expresión vs. GO). Por lo tanto, la incertidumbre se propaga y se generan análisis sesgados sobre posibles hipótesis que agrupan a conjuntos de genes.**

En este trabajo presentamos Onto-CC, un método automático especialmente adecuado para la explicación/validación independiente de hipótesis que agrupan a conjuntos de genes (e.g. genes co-expresados) basándose en agrupamientos (clusters) de términos GO (i.e., expresión vs. GO). Onto-CC reduce la incertidumbre de las consultas identificando los clusters conceptuales óptimos que combinan términos de diversas sub-ontologías simultáneamente, al igual que términos definidos en distintos niveles de especificidad en la jerarquía de GO. Para ello, se utiliza la metodología de EMO-CC para encontrar clusters conceptuales en bases de datos estructurales (e.g., grafos dirigidos acíclicos, como GO). Esta metodología permite considerar los distintos clusters óptimos como posibles hipótesis paralelas, guiados por técnicas de optimización multiobjetivo/multimodal. Por lo tanto, se podrán generar explicaciones alternativas y, a la vez, óptimas de las consultas, lo cual puede proporcionar nuevo conocimiento para un problema dado.

Existen dos versiones de Onto-CC para uso publico: *Ready2GO*, una versión de EMO-CC precalculada para varios genomas y una versión de *Onto-CC avanzada* para su uso con archivos de anotación propios (`http://gps-tools2.wustl.edu/onto-cc/index.html`).

Palabras clave— **Agrupamiento de datos, Algoritmos Evolutivos MultiObjetivo, Expresión genética.**

I. Introducción

Las técnicas experimentales de procesamiento de alto-rendimiento, tales como los *microarrays* [1], producen grandes cantidades de datos y de conocimiento sobre los niveles de la expresión de un conjunto de genes. El análisis de esta información suele ser con frecuencia una lista de genes con niveles de expresión significativamente diferenciados, la cual puede llegar a contener hasta cientos de genes. Esta información es de poco uso si no es posible interpretar los resultados en un contexto biológico [2]. Para solventar este problema, el *Gene Ontology Consortium* [3] proporciona descripciones de productos de genes. Este conocimiento biológico se organiza como vocabularios jerárquicos, estructurados y controlados llamados *Gene Ontologies (GOs)* [3], los cuales describen productos de genes en términos de sus procesos biológicos (PB), funciones moleculares (FM) y componentes celulares (CC). Hoy en día, el Gene Ontology Consortium proporciona anotaciones utilizando estos términos GO para diversos organismos [3].

En los últimos años se han estado desarrollando varias herramientas para identificar grupos de términos GO que puedan explicar conjuntos de genes co-expresados provenientes de experimentos de microarrays [4]. Estas herramientas suelen buscar términos GO sobre-representados que describan a un grupo de genes usando diversas ticónicas estadísticas, tales como el test de Fisher (FatiGO [5]), χ^2 o la distribución binomial (Onto-Express [6]), o calculando z-scores mediante la distribución hiper-geométrica (MAPPFinder [7]).

Uno de los principales problemas al identificar clusters biológicos significativos en la base de datos de GO es que la calidad de las anotaciones está basada en el conocimiento disponible. Por ejemplo, algunos procesos biológicos son estudiados más detalladamente que otros, así generando ramas largas con términos GO muy específicos mientras que otras ramas siguen estando muy poco estudiadas. Para tratar con esta incertidumbre, la mayor parte de las herramientas actualmente disponibles piden al usuario que seleccione un nivel de especificidad (e.g., nivel 3) para los términos recuperados, a menudo restringiendo los términos GO encontrados (e.g., todos los procesos biológicos) a los mismos niveles recuperando entonces no sólo información limitada sino también información muy general o demasiado específica. Por otra parte, la mayor parte de los métodos de agrupamiento (clustering) disponibles buscan en cada en sub-ontología en forma independiente y, por ello, perdiendo relaciones relevantes entre términos de diversas sub-ontologías.

La desventaja crucial compartida por estos métodos es que su algoritmo de clustering subyacente no ha sido diseñado para trabajar con información jerárquica [8], restringiendo su capacidad de buscar relaciones complejas subyacentes contenidas en el gráfico dirigido acíclico (DAG) de la base de datos de GO. Una base de datos estructural se puede

Departamento de Ciencias de la Computación e Inteligencia Artificial, Universidad de Granada. E-mail: {rocio,delval,igor}@decsai.ugr.es

ver como grafo que contiene nodos, que representan objetos; y las relaciones entre estos objetos representadas por ejes. Luego, una subestructura corresponde a un sub-grafo del DAG de GO [9]. Las técnicas de clustering conceptual ya han sido aplicadas con éxito a bases de datos estructurales al buscar a través de un espacio predefinido de hipótesis potenciales (i.e., subestructuras) aquellas que mejor se ajusten a los ejemplos del entrenamiento [9], [10]. Sin embargo, la búsqueda de clusters conceptuales en una estructura con forma de grafo como el DAG de GO, daría lugar a la generación de muchas subestructuras pequeñas, dado que es más fácil modelar subconjuntos de datos más pequeños que grupos más grandes y representativos [11].

La utilidad de los sistemas existentes de inferencia de perfiles funcionales se ve afectada por el sesgo en la base de datos de GO y por las restricciones impuestas por los métodos de clustering. Por lo tanto, para extraer conceptos mejor definidos, Onto-CC utiliza una metodología inspirada en técnicas de clustering conceptual [12], permitiendo obtener clusters óptimos basados en su especificidad, diversidad y en el número de productos de genes recuperados. Éstos son criterios en conflicto que pueden ser estudiados como un problema de optimización. El desafío es evitar el sesgo potencial causado al realizar sumas pesadas de los objetivos [11], que siempre culmina por una convergencia a un conjunto de soluciones que corresponden a regiones limitadas del espacio de búsqueda (i.e., el DAG de GO). Este problema resulta significativo dado que la minera de datos clásica, particularmente en biología computacional, tiende a acentuar los patrones mas frecuentes [8] que, en general, encubren mas que revelan conocimiento nuevo y útil sobre el problema [13], [14].

II. Metodología

El servidor Onto-CC [15] realiza búsquedas de explicaciones/validaciones funcionales que estén potencialmente relacionadas, en un grupo de genes provisto por el usuario (e.g., genes co-expresados). El conjunto de genes dado como entrada es estadísticamente comparado con un conjunto de clusters obtenidos de GO precalculados independientemente, para el organismo seleccionado. Estos clusters comparten un conjunto de características (i.e., términos de GO) organizados jerárquicamente en distintos niveles de especificidad en la base de datos estructural (i.e., DAG de GO). De hecho, Onto-CC considera las tres sub-ontologías de GO simultáneamente. Los grupos que resultan de las relaciones anteriores (i.e., los clusters conceptuales) deben ser óptimos, evitando redundancia, pero permitiendo descripciones de genes desde diferentes puntos de vista. Es decir un gen puede pertenecer a diversos clusters conceptuales caracterizados por diversos conjuntos de características [16]. Resumiendo, esta herramienta web permite que los usuarios validen sus hipótesis sobre un conjunto de productos de genes, estableciendo relaciones entre ellos y los clusters de términos GO, los cuales han sido identificados por un algoritmo inspirado en el clustering conceptual. Onto-CC no sólo recupera clusters de genes, sino que también realiza una selección de características diferenciada para cada cluster [17].

Los clusters precalculados (i.e., subestructuras) se obtienen siguiendo estos pasos: (i) Dado un archivo de anotación de GO para un genoma específico, el algoritmo crea aleatoriamente las subestructuras potenciales que contienen distintas características (i.e., términos GO) definidas en varios niveles de especificidad y en distintas sub-ontologías de GO. Onto-CC no selecciona a priori un nivel de especificidad en el DAG de GO como lo hacen la mayor parte de las herramientas actuales (e.g., nivel 3), busca subestructuras en diversos niveles de especificidad a través del espacio del DAG usando un algoritmo evolutivo (AE) [18]. (ii) Las subestructuras iniciales evolucionan dirigidas por un sistema de optimización multiobjetivo/multimodal basado en dos objetivos: el grado de similitud entre los términos contenidos en la subestructura y los términos GO que caracterizan un subconjunto de producto de genes (i.e., especificidad) y el número de productos de genes descritos por la subestructura (i.e., soporte). Éstas son objetivos contradictorios, dado que cuando la especificidad aumenta, el soporte generalmente disminuye, y viceversa. Particularmente, el objetivo es seleccionar las subestructuras que satisfacen una solución de compromiso entre especificidad y soporte. (iii) Los clusters finales se alcanzan cuando se llega al número máximo de evaluaciones de la función de coste o a un número máximo de generaciones determinado. Los resultados obtenidos son clusters destacados de genes/términos GO no-dominados, es decir no son peores que cualquier otra solución final en ambos objetivos. Estos grupos consisten en todas las variaciones óptimas posibles de términos de GO definidos en diversos niveles de especificidad, sub-ontologías y productos de genes.

El servidor Onto-CC provee dos servicios: Ready2GO y la versión avanzada de Onto-CC. El servicio Ready2GO es una versión precalculada de los clusters conceptuales para más de 30 genomas distintos anotados por el GO Consortium. La versión avanzada de Onto-CC está pensada para usuarios que trabajan con genomas no completamente descritos, anotaciones de genomas propios o genomas aun no anotados por el GO Consortium. En este caso, los clusters conceptuales serán calculados en linea basándose en los archivos de anotación proporcionados por el usuario.

III. Interfaz web

El servidor web está implementado usando scripts cgi que comunican con varios scripts de Perl y el ejecutable unix de EMO-CC. Cada una de las versiones

de Onto-CC, Ready2GO y avanzado, tiene un tutorial disponible junto con un conjunto de archivos de prueba. Los tutoriales explican qué parámetros pueden ser ajustados y entre qué rangos pueden ser modificados. Los ajustes por defecto son recomendados para nuevos usuarios. Los tutoriales en linea cubren los siguientes temas: organismo, anotación, entrada, valor de umbral, parámetros de EMO-CC, salidas adicionales, resultados por correo electrónico y en HTML. Los resultados se proporcionan en HTML para una inspección visual y pueden ser también recibidos por correo electrónico. En caso de error, se exhibe un mensaje en pantalla.

A. Bases de datos

Las bases de datos estándar de proteínas son utilizas para consultar términos GO mediante listas de números de acceso proporcionadas por el usuario. Los números de acceso que pueden ser utilizados son: UniProt (número de acceso o identificación) [19], RefSeq [20], Ensembl [21], Vega [22], GI [23], nombre del gen, Dictybase [24], CGD [25], Flybase [26], GeneDB [27], TIGR [28], MGD [29], RGD [30], SGD [31], PseudoCAP [32], TAIR [33], Wormbase [34], ZFIN [35] y/o PDB [36]. Para cada organismo existe una relación entre todas estas bases de datos y el proyecto GO calculada fuera de linea. Los archivos de anotación y los clusters son actualizados cada 6 meses. Para ello, aprovechamos la ventaja de técnicas evolutivas, como la del algoritmo propuesto para permitir que se actualicen los clusters realizando unas pocas generaciones usando los clusters anteriores como semilla [37]. Este aprendizaje incremental [38] acelera y reduce la complejidad de cómputo del proceso de actualización, dejando el recalculo completo para casos extremos e inusuales (manuscrito en preparación).

B. Ready2GO

B.1 Entrada

El fichero de entrada es una lista de números de acceso que pertenecen a uno de los organismos para los cuales el proyecto GO proporciona una anotación [3] (Figura 1 (a)). Este fichero de entrada consiste en identificadores, uno por línea, de cualquiera de las bases de datos mencionadas anteriormente.

B.2 Parámetros

Existen dos parámetros a especificar: organismo y *p-value*. El organismo puede ser cualquiera de los genomas anotados por el GO Consortium, listado en el menú; el menú incluye eucariotas, microorganismos y multiespecies (Figura 1 (a)). El segundo parámetro es el p-value [39] y representa la probabilidad de observar al azar una intersección específica entre los productos de genes dados por el usuario y los productos de genes que pertenecen a los clusters calculados de antemano. El p-value puede tomar valores entre 0 y 1, donde valores más bajos representan mayor confiabilidad.

B.3 Salida

Los resultados se muestran como una tabla HTML que contiene cada uno de los clusters encontrados en ningún orden particular. Los clusters se pueden ordenar por el número de productos de genes o por p-value, usando los botones que se encuentran sobre la tabla (Figura 1, (b)). La tabla contiene los siguientes campos: número de identificación del cluster (i.e., columna Cluster ID), términos GO y su descripción correspondientes a la sub-ontología de PB (i.e., columnas Biological Process y BP Description, respectivamente), términos GO y su descripción correspondientes a la sub-ontología de FM (i.e., columna Molecular Function y MF Description, respectivamente), términos GO y su descripción correspondientes a la sub-ontología de CC (i.e., columnas Cellular Component y CC Description, respectivamente), la lista de números de acceso que pertenecen al cluster (i.e., columna ACC) y el p-value entre el conjunto de los números de acceso dados y el cluster (i.e., columna P-value). Además de la versión en HTML el archivo de salida se puede descargar como un archivo de texto separado por comas (.csv, conveniente para MS Excel) y como texto separado por tabulaciones (.txt).

C. Versión avanzada

La versión avanzada de Onto-CC permite obtener un conjunto de descripciones de términos GO para una lista de números de acceso a partir de la información de anotación de términos GO provista por el usuario. En este caso las subestructuras (i.e., clusters conceptuales) serán calculadas en linea basándose en los archivos de anotación proporcionados por el usuario. Para poder llevar a cabo este objetivo son necesarios dos pasos: *Paso 1*, creación de los clusters conceptuales propios y *Paso 2*, creación de una descripción usando términos GO para una lista de números de acceso usando los clusters conceptuales calculados anteriormente.

C.1 Paso 1

El fichero de entrada para este paso es un archivo de anotación GO. Este archivo describe la relación entre un producto de un gen/una proteína y los términos GO. El archivo de anotación contiene una descripción por línea, donde el identificador del gen/de la proteína se separa de su descripción GO por una coma. Cada identificador puede tener múltiples términos GO, que son separados por puntos y comas y pueden pertenecer a cualquiera de las subontologías en el proyecto GO.

El EMO-CC es un AE con varios objetivos. Un AE utiliza mecanismos inspirados en la evolución biológica para optimizar las soluciones a un problema, como ser la reproducción, la mutación, la recom-

Ready2GO version: *(Tutorial)*

Organism [?]

Eukaryote	Microorganism	Multispecies
Arabidopsis thaliana (TAIR/TIGR)	Anaplasma phagocytophilum HZ (TIGR)	Protein Data Bank (EBI)
Bos taurus (EBI)	Bacillus anthracis Ames (TIGR)	UniProt (EBI)
Caenorhabditis elegans (WormBase)	Carboxydothermus hydrogenoformans Z-2901 (TIGR)	
Candida albicans (CGD)	Campylobacter jejuni RM1221 (TIGR)	
Danio rerio (ZFIN)	Coxiella burnetii RSA 493 (TIGR)	
Dictyostelium discoideum (DictyBase)	Dehalococcoides ethenogenes 195 (TIGR)	
Drosophila melanogaster (FlyBase)	Ehrlichia chaffeensis Arkansas (TIGR)	
Gallus gallus (EBI)	Geobacter sulfurreducens PCA (TIGR)	
Homo sapiens (EBI)	Listeria monocytogenes 4b F2365 (TIGR)	
Leishmania major (Sanger GeneDB)	Methylococcus capsulatus Bath (TIGR)	
Mus musculus (MGI)	Neorickettsia sennetsu Miyayama (TIGR)	
Oryza sativa (Gramene)	Pseudomonas aeruginosa PAO1 (PseudoCAP)	
Plasmodium falciparum (Sanger GeneDB)	Pseudomonas syringae DC3000 (TIGR)	
Rattus norvegicus (RGD)	Shewanella oneidensis MR-1 (TIGR)	
Saccharomyces cerevisiae (SGD)	Silicibacter pomeroyi DSS-3 (TIGR)	
Schizosaccharomyces pombe (Sanger GeneDB)	Vibrio cholerae (TIGR)	
Trypanosoma brucei (Sanger GeneDB)		
Trypanosoma brucei chr 2 (TIGR)		

Input [?] Examinar...

Threshold [?] 0.05 (default = 0.05, min = 0, max = 1)

Additional Outputs [?] Excel (.csv) Tab delimited (.txt)

E-mail results [?] (optional)

Submit Reset Test Set & Configuration

(a)Captura de pantalla del formulario de entrada. Se pueden seleccionar de entre varios genomas y bases de datos multiespecies.

Ready2GO version

Calculating clusters for input file *testset.txt* on organism *Oryza sativa* with threshold *0.5*...

.....done

Download .csv file

Download .txt file

See Results unordered | See Results ordered by number of genes | See Results ordered by PI

Cluster id	Biological Process (BP)	Molecular Function (MF)	Cellular Component (CC)	BP Description	MF Description
1	GO:0006075	GO:0003843	GO:0005575	beta-1 ;3 glucan biosynthesis	1 ;3-beta-glucan synthas
2	GO:0040029	GO:0003674	GO:0005575	regulation of gene expression ; epigenetic	molecular_functic

Back

(b)Captura de pantalla de la salida de resultados. Además de la tabla HTML, se pueden descargar dos archivos con los resultados: .csv (versión separada por comas, conveniente para MS Excel) y .txt (separado por tabulaciones).

Fig. 1. Interfaz de Ready2GO.

binación, la selección natural o la supervivencia del más apto. Se pueden modificar varios parámetros en un AE, pero solamente dos están disponibles para el usuario: el tamaño de la población y el número de evaluaciones. Los cambios en estos parámetros modifican el funcionamiento del algoritmo y tienen un efecto en el número y la calidad de los clusters encontrados. Los AEs trabajan sobre una población de representaciones abstractas, llamadas cromosomas, de soluciones candidatas, llamados individuos, a un problema de la optimización. Tamaños más grandes de población darán lugar a un funcionamiento más lento, pero a mejores resultados. Aumentando el tamaño de la población genera más espacio disponible para guardar mas soluciones, por lo tanto, promoviendo la evolución a mejores áreas y la diversidad. A medida que el tamaño de la población aumenta, el número de evaluaciones realizadas debe también aumentar. El tamaño de la población puede ser modificado por el usuario en el rango [10-1000] con un valor prefijado de 200. Este valor es apropiado para una lista de aproximadamente 2000 IDs anotados. Generalmente, la población inicial de candidatos consiste en soluciones generadas aleatoriamente. Durante cada generación sucesiva, se selecciona una porción de la población existente para producir una nueva generación. Una función de coste se utiliza para dirigir la búsqueda y se aplica a las soluciones candidatas y a cualquier descendiente subsecuente para calcular el costo de una solución de tal manera de poder comparar un cromosoma particular contra el resto de cromosomas. Cada uno de estas evaluaciones de la función de coste se puede utilizar para determinar cuando detener la ejecución de un AE. El usuario puede especificar el número máximo de evaluaciones para el AE, donde los valores se extienden desde 100 a 99.999 y el valor prefijado es de 20.000. Como una regla básica, se puede calcular el número de evaluaciones como un múltiplo del tamaño de la población. Este número será aproximadamente el número de generaciones a realizar.

La tabla HTML de salida muestra cada uno de los clusters encontrados en ningún orden particular. La tabla es muy similar a la tabla de salida de Ready2GO pero sin la columna del PI y con la adición de columnas de especificidad y soporte. Los valores de especificidad se encuentra en el rango [0-1] con 1 como el mejor caso en el cual todos los productos de genes, descrito por el cluster, comparten los mismos términos GO. Los valores de la segunda función objetivo, el soporte, se encuentran en el rango [0-1] con 1 como el mejor caso en el cual el cluster describe todos los productos de genes en el fichero de entrada. Al igual que en el caso anterior, el archivo de salida se puede transferir como una versión separada por comas (.csv, conveniente para MS Excel) y como texto separado por tabulaciones (.txt), además de la versión HTML.

C.2 Paso 2

Este paso necesita dos entradas: (1) el archivo de clustering obtenido del paso 1 y (2) el fichero de entrada que contiene los IDs que el usuario quiere analizar. La salida es igual que la explicada para Ready2GO.

IV. Implementación

El script que convierte de un número de acceso de una base de datos en otra esta escrito en Perl usando los módulos de BioPerl y accediendo a varios servicios web (e.g., biomaRt, pagina web del genoma del organismo y el servicio web de UniProt ID mapping). Onto-CC fue desarrollado en Eiffel v6.0 (Eiffel es un lenguaje de programación con estándar ISO, orientado a objetos y basado en el paradigma del diseño por contrato).

Para la versión de Ready2GO los tiempos de ejecución varían dependiendo de la cantidad de números de acceso de entrada combinados con el tamaño de la anotación del genoma. El archivo de prueba con los valores prefijados tarda aproximadamente un minuto en un ordenador de 64-bits con un procesador de 2 GHz. Para la versión avanzada, el paso 1 tarda varios minutos para un archivo estándar. Esta consumición del tiempo no sólo depende de la cantidad de números de acceso y de las anotaciones de la entrada, sino también del tamaño de la población del AE y del número de evaluaciones a realizarse. Recomendamos guardar los resultados obtenidos del paso 1 para reutilizarlos en el paso 2 sin tener que reconstruir el clustering conceptual. El archivo de prueba con los valores prefijados tarda menos de 15 segundos para el paso 1 y menos de 5 segundos para el paso 2 utilizando el mismo ordenador especificado anteriormente.

V. Conclusiones

El vocabulario de GO ha sido explorado extensivamente para analizar las funciones de genes co-expresados [5], [6]. Sin embargo, y a pesar de su uso extendido, todavía hay controversia sobre su utilidad para validar hipótesis de agrupaciones de genes. Hemos propuesto Onto-CC como un método automático especialmente adecuado para la explicación y validación independiente de las hipótesis que agrupan a conjuntos de genes (e.g., genes co-expresados) basadas en clustering de términos GO (i.e., expresión vs. GO), en lugar del uso de los términos GO para conducir búsquedas de perfiles de expresión de genes (i.e., expresión + GO) [5]. El método de clustering usado en nuestro propuesta es suficientemente robusto para reproducir resultados, independientemente de los niveles de especificidad de la anotación del organismo utilizado. Los experimentos del funcionamiento del algoritmo en la base de datos de GO con distintas complejidades ha mostrado una distribución similar de soluciones. La complejidad reducida de una base de datos aumenta el número de solucio-

nes altamente específicas con un soporte bajo, que indica la presencia de clusters solapados (i.e., clustering difuso) causados por un conjunto de términos más generales y condensados. Aunque para distintas ejecuciones sobre diferentes bases de datos de GO con distintas complejidades para el mismo organismo se obtienen pequeñas diferencias en la especificidad de los clusters, la mayoría de los clusters mejor evaluados son reconocidos en la versión completa y en la reducida, caracterizando los mismos genes.

Onto-CC reduce la incertidumbre de las consultas identificando los clusters conceptuales óptimos que combinan términos de diversas sub-ontologías simultáneamente y, a la vez, usando términos definidos en diversos niveles de especificidad en la jerarquía de GO. De hecho, un gen puede pertenecer a más de un cluster [40], así proporcionando explicaciones alternativas que pueden generar nuevo conocimiento para un problema dado.

El punto mas destacado de la metodología propuesta es su flexibilidad para integrar diversas fuentes de conocimiento basándose en pruebas estadísticas [12], lo cual facilita el uso de Onto-CC conjuntamente con otras fuentes de anotación independientes tales como IPA [41]. La validación de la metodología usada por Onto-CC, al igual que su funcionamiento en comparación con otras herramientas utilizados para bases de datos con términos GO ya ha sido publicada [12]. El desarrollo del servidor presentado aquí ha sido diseñado pensando en el usuario final desde el comienzo. Su funcionalidad está siendo continuamente actualizada y extendida en respuesta a las peticiones y a las sugerencias que emergen de nuestros usuarios.

Agradecimientos

Este trabajo ha sido subvencionado en parte por el Ministerio de Ciencia y Tecnología de España bajo el proyecto TIN-2006-12879 y en parte por la Conserjería de Innovación, Investigación y Ciencia de la Junta de Andalucía bajo el proyecto TIC-02788. Coral del Val también pertenece al "Programa de Retorno de Investigadores"de la Junta de Andalucía, e Igor Zwir es también un investigador senior subvencionado por el Howard Hughes Medical Institute.

Referencias

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

[2] T. Beissbarth, "Interpreting experimental results using gene ontologies," *Methods Enzymol*, vol. 411, pp. 340–352, 2006.

[3] The Gene Ontology Consortium, "Gene ontology: tool for the unification of biology," *Nature Genet.*, vol. 25, pp. 25–29, 2000.

[4] P. Khatri and S. Draghici, "Ontological analysis of gene expression data: current tools, limitations, and open problems," *Bioinformatics*, vol. 21, pp. 3587–3595, 2005.

[5] F. Al-Shahrour, P. Minguez, J. TÃ¡rraga, I. Medina, E. Alloza, D. Montaner, and J. Dopazo, "Fatigo+: a functional profiling tool for genomic data. integration of functional annotation, regulatory motifs and interaction data with microarray experiments.," *Nucleic Acids Res.*, vol. 35, pp. W91–W96, 2007.

[6] S. Draghici, P. Khatri, P. Bhavsar, A. Shah, S. Krawetz, and M.A. Tainsky, "Onto-tools, the toolkit of the modern biologist: onto-express, onto-compare, onto-design and onto-translate," *Nucleic Acids Res.*, vol. 31, pp. 3775–3781, 2003.

[7] S.W. Doniger, N. Salomonis, K.D. Dahlquist, K. Vranizan, S.C. Lawlor, and B.R. Conklin, "Mappfinder: using gene ontology and genmapp to create a global gene-expression profile from microarray data," *Genome Biology*, vol. 4, no. I, pp. Electronic publication, 2004.

[8] I. Zwir, H. Huang, and E.A. Groisman, "Analysis of differentially-regulated genes within a regulatory network by gps genome navigation," *Bioinformatics*, vol. 21, pp. 4073–4083, 2005.

[9] I. Jonyer, D. J. Cook, and L. B. Holder, "Discovery and evaluation of graph-based hierarchical conceptual clusters," *Journal of Machine Learning Research*, vol. 2, pp. 19–43, 2001.

[10] T. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.

[11] E. Ruspini and I. Zwir, "Automated generation of qualitative representations of complex object by hybrid soft-computingmethods," in *Pattern Recognition: From Classical to Modern Approaches*, S. Pal and A. Pal, Eds., Singapore, 2001, pp. 453–474, World Scientific Company.

[12] R.C. Romero-Zaliz, C. Rubio-Escudero, J.P. Cobb, F. Herrera, O. Cordón, and I. Zwir, "A multi-objective evolutionary conceptual clustering methodology for gene annotation within structural databases: a case of study on the gene ontology database," *IEEE Transactions on Evolutionary Compututation*, 2008, in press.

[13] L. McCue, W. Thompson, C. Carmack, M. P. Ryan, J. S. Liu, V. Derbyshire, and C. E. Lawrence, "Phylogenetic footprinting of transcription factor binding sites in proteobacterial genomes," *Nucleic Acids Res*, vol. 29, pp. 774–82, 2001.

[14] A. Martinez-Antonio and J. Collado-Vides, "Identifying global regulators in transcriptional regulatory networks in bacteria," *Curr Opin Microbiol*, vol. 6, pp. 482–9, 2003.

[15] R. Romero-Zaliz, C. del Val, J.P. Cobb, and I. Zwir, "Onto-cc: a web server for identifying gene ontology conceptual clusters," *Nucleic Acids Res*, 2008.

[16] D. Cook, L. Holder, S. Su, R. Maglothin, and I. Jonyer, "Structural mining of molecular biology data," *IEEE Engineering in Medicine and Biology, special issue on Advances in Genomics*, vol. 4, no. 20, pp. 67–74, 2001.

[17] R. Kohavi and G.H. John, "Wrappers for feature subset selection," *Artificial Intelligence*, vol. 97, pp. 273–324, 1997.

[18] D. Goldberg, *Genetic Algorithms in Search Optimization and Machine Learning*, Addison-Wesley, 1989.

[19] The UniProt Consortium, "The universal protein resource (uniprot)," *Nucleic Acids Res.*, vol. 35, pp. D193–D197, 2007.

[20] K.D. Pruitt, T. Tatusova, and D.R. Maglott, "Ncbi reference sequence (refseq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Res.*, vol. 35, pp. D61–D65, 2007.

[21] T.P.J. Hubbard, B.L. Aken, K. Beal, B. Ballester, M. Caccamo, Y. Chen, L. Clarke, G. Coates, F. Cunningham, T. Cutts, T. Down, S. C. Dyer, S. Fitzgerald, J. Fernandez-Banet, S. Graf, S. Haider, M. Hammond, J. Herrero, R. Holland, K. Howe, N. Johnson, A. Kahari, D. Keefe, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, C. Melsopp, K. Megy, P. Meidl, B. Ouverdin, A. Parker, A. Prlic, S. Rice, D. Rios, M. Schuster, I. Sealy, J. Severin, G. Slater, D. Smedley, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, M. Wood, T. Cox, V. Curwen, R. Durbin, X.M. Fernandez-Suarez, P. Flicek, A. Kasprzyk, G. Proctor, S. Searle, J. Smith, A. Ureta-Vidal, , and E. Birney, "Ensembl," *Nucleic Acids Res.*, vol. 35, pp. D610–D617, 2007.

[22] J.L. Ashurst, C.K. Chen, J.G. Gilbert, K. Jekosch, S. Keenan, P. Meidl, S.M. Searle, J. Stalker, R. Storey, S. Trevanion, L. Wilming, and T.S. Hubbard, "The vertebrate genome annotation (vega) database," *Nucleic Acids Res.*, vol. 33, pp. D459–D465, 2005.

[23] D.A. Benson, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, and D.L. Wheeler, "Genbank," *Nucleic Acids Res.*, vol. 35, pp. D21–D25, 2007.

[24] R.L. Chisholm, P. Gaudet, R.M. Just, K.E. Pilcher, P. Fey, S.N. Merchant, and W.A. Kibbe, "dictybase, the model organism database for dictyostelium discoideum," *Nucleic Acids Res.*, vol. 34, pp. D423–D427, 2006.

[25] M.B. Arnaud, M.C. Costanzo, M.S. Skrzypek, G. Binkley, C. Lane, S.R. Miyasato, and G. Sherlock, "The candida genome database (cgd), a community resource for candida albicans gene and protein information," *Nucleic Acids Res.*, vol. 33, pp. D358–D363, 2005.

[26] M.A. Crosby, J.L. Goodman, V.B. Strelets, P. Zhang, W.M. Gelbart, and The FlyBase Consortium, "Flybase: genomes by the dozen," *Nucleic Acids Res.*, vol. 35, pp. D486–D491, 2007.

[27] C. Hertz-Fowler, C.S. Peacock, V. Wood, M. Aslett, A. Kerhornou, P. Mooney, A. Tivey, M. Berriman, N. Hall, K. Rutherford, J. Parkhill, A.C. Ivens, M.A. Rajandream, and B.K. Barrell, "Genedb: a resource for prokaryotic and eukaryotic organisms," *Nucleic Acids Res.*, vol. 32, pp. D339–D343, 2005.

[28] J.D. Peterson, L.A. Umayam, T.M. Dickinson, E.K. Hickey, and O. White, "The comprehensive microbial resource," *Nucleic Acids Research*, vol. 29, pp. 123–125, 2001.

[29] J.T. Eppig, C.J. Bult, J.A. Kadin, J.E. Richardson, J.A. Blake, and Members of the Mouse Genome Database Group, "The mouse genome database (mgd): from genes to mice-a community resource for mouse biology," *Nucleic Acids Res.*, vol. 33, pp. D471–D475, 2005.

[30] S.N. Twigger, M. Shimoyama, S. Bromberg, A.E. Kwitek, H.J. Jacob, and the RGD Team, "The rat genome database, update 2007 - easing the path from disease to data and back again," *Nucleic Acids Res.*, vol. 35, pp. D658–D662, 2007.

[31] J.M. Cherry, C. Ball, S. Weng, G. Juvik, R. Schmidt, C. Adler, B. Dunn, S. Dwight, L. Riles, R.K. Mortimer, and D. Botstein, "Genetic and physical maps of saccharomyces cerevisiae," *Nature*, vol. 387 (Suppl. 6632), pp. 67–73, 1997.

[32] G.L. Winsor, R. Lo, S.J. Sui, K.S. Ung, S. Huang, D. Cheng, W.K. Ching, R.E. Hancock, and F.S. Brinkman, "Pseudomonas aeruginosa genome database and pseudocap: facilitating community-based, continually updated, genome annotation," *Nucleic Acids Res.*, vol. 33, pp. D338–D343, 2005.

[33] S.Y. Rhee, W. Beavis, T.Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L.A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D.C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon, and P. Zhang, "The arabidopsis information resource (tair): a model organism database providing a centralized, curated gateway to arabidopsis biology, research materials and community," *Nucleic Acids Res.*, vol. 31, pp. 224–228, 2003.

[34] T. Bieri, I. Antoshechkin, C. Bastiani, D. Blasiar, P. Canaran, J.C. Chan, W.J. Chen, P. Davis, T.J. Fiedler, L. Girard, M. Han, T.W. Harris, R. Kishore, R. Lee, S. McKay, H.M. Müller, C. Nakamura, A. Petcherski, A. Rangarajan, A. Rogers, G. Schindelman, E.M. Schwarz, W. Spooner, M.A. Tuli, K. Van Auken, D. Wang, X. Wang, G. Williams, R. Durbin, L.D. Stein, P.W. Sternberg, and J. Spieth, "Wormbase: new content and better access," *Nucleic Acids Res.*, vol. 35, pp. D506–D510, 2007.

[35] J. Sprague, D. Clements, T. Conlin, P. Edwards, K. Frazer, K. Schaper, E. Segerdell, P. Song, B. Sprunger, and M. Westerfield, "The zebrafish information network (zfin): the zebrafish model organism database," *Nucleic Acids Res.*, vol. 31, pp. 241–243, 2003.

[36] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, and P.E. Bourne, "The protein data bank," *Nucleic Acids Research*, vol. 28, pp. 325–242, 2000.

[37] G. Sheng-Uei and Z. Fangming, "An incremental approach to genetic-algorithms-based classification," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 35, no. 2, pp. 227–239, 2005.

[38] C. Giraud-Carrier, "A note on the utility of incremental learning," *AI Commun.*, vol. 13, pp. 215–223, 2000.

[39] S. Tavazoie, J. Hughes, M. Campbell, R. Cho, and G. Church, "Systematic determination of genetic network architecture," *Nature Genet.*, vol. 22, no. 3, pp. 281–285, 1999.

[40] A.P. Gasch and M.B. Eisen, "Exploring the conditional coregulation of yeast gene expression through fuzzy k-means clustering," *Genome Biology*, vol. 3, pp. RESEARCH0059, 2002.

[41] Ingenuity Systems, "Ingenuity pathways analysis," .