UNIVERSIDAD DE BUENOS AIRES
FACULTAD DE CIENCIAS EXACTAS Y NATURALES

Desarrollo de una metodología de extracción de conocimientos a partir de datos de micromatrices de DNA basada en ontologías genéticas

TESIS PRESENTADA PARA OPTAR AL TITULO DE
MAGISTER EN EXPLOTACIÓN DE DATOS Y
DESCUBRIMIENTO DEL CONOCIMIENTO

AUTOR: LIC. EN SIST. ARMANDO TAIÉ
DIRECTOR : DRA. ANA SILVIA HAEDO
CO-DIRECTORES : DRA. NORMA PANIEGO
DR. MARCELO SORIA

BUENOS AIRES, DICIEMBRE DE 2008

A Mónica y Macarena
La más vívida expresión del amor de Dios

Agradecimientos

A Dios origen y fin de toda mi vida.

A Mónica y Macarena, por el amor y la paciencia. Por acompañarme en estos años de esfuerzo, cediéndome muchas horas de familia para permitirme seguir mi ideal

A mi familia, mis padres Fernando e Isabel y mi hermana María de los Milagros, por el apoyo incondicional y aliento permanente.

Al INTA por el apoyo económico recibido para concretar esta Maestría.

A los Directivos del INTA en las personas del Ing. Hugo García, Ing. Hugo Roig, Ing. Alfredo Marín y el Dr. Ramón Vasquez. Por permitirme reencontrarme con mi futuro.

A los Directivos y Profesores de la Maestría, por el esfuerzo y dedicación. En especial al Profesor Gustavo Denicolay.

A los Directores de mi Tesis Dra. Ana Haedo, Dra. Norma Paniego y Dr. Marcelo Soria. Por la excelencia de sus conocimiento, la calidad y calidez de sus personas.

A mis "hermanos de la vida", mis compañeros de la Maestría, en especial aquellos con los que compartimos largas horas de trabajos prácticos y presentaciones: Claudio, Diego, Fernando, Ariel, Gustavo, Pablo, Andrés, Guillermo y todos aquellos que me fui encontrando en este camino. A Paula Fernandez y Marisa Farber por la el apoyo.

A Mónica Cuschnir, Mercedes y Nicolás por su ayuda desde la secretaría de la Maestría. A los conservadores de los laboratorios Guillermo, Fernando, Mateo, por el soporte técnico que me brindaron. A Diego el administrador del servidor de la maestría.

A mis Compañeros y Amigos del Proyecto Arroz del INTA EEA Corrientes, Tano, Miguel, Juan, Rita, María Inés, al bibliotecario Toti y el personal de campo.

A la Secretaria de Dirección Ana María. A la Administración del INTA, Jorge, Alberto, María Paz, Gabriel, Fabiana, Gisela, Franco, Fernando, Daniel y Sol.

A Mis suegros Pepe y Florita. A Juan y su familia y Jorge.

A los integrantes del Movimiento Familiar Cristiano, "Poroto" y Alba, Kent y Estela, Juanchi y Patricia, Ariel y Alicia, Oscar y Laura.

Al Sacerdote Pablo Sanchez.

Al personal de APINTA, Luís, Tito, Mabel, Ana, Pablo y a todas las personas que conocí allí, Roberto y Sra., Esteban y Sra., Hugo y Sra., y mi amiguita Araceli.

Y dejo para lo último y entre los más importantes. A mis queridos Tío Guido y Tía Nelly. Creyeron en mí desde siempre.

Y a todos los que me ayudaron y que estuvieron y estarán en mi corazón.

Abreviaturas

ADN	ácido desoxirribonucleico
ADNc	ácido desoxirribonucleico complementario
ARN	ácido ribonucleico
mARN	ácido ribonucleico mensajero
tARN	ácido ribonucleico de transferencia
A	Adenina
C	Citosina
G	Guanina
T	Timina
U	Uracilo
NCBI	Nacional Center of Biotechnology Information
EMBL	European Molecular Biology Laboratory
EBI	European Bioinformatics Institute
DDBJ	DNA Data Bank of Japan
CBI-DDBJ	Center for Information Biology and DNA Data Bank of Japan
Cy-3	Cianina-3
Cy-5	Cianina-5
DE	diferencialmente expresado
PM	Perfect Match probe (sonda de apareamiento perfecto)
MM	Mismatch probe (sonda de apareamiento imperfecto)
TIGR	The Institute for Genomic Research
MGED	The Microarray Gene Expression Data Society
MIAME	Minimal Information about a Microarray Experiment
MAGE-OM	Microarray Gene Expresion – Object Model
MAGE-ML	Microarray Gene Expresion – Implementación XML
MAGE-stk	Microarray Gene Expresion – Software toolkit
GO	Gene Ontology Consortium
DAG	Direct Acyclic Graph
XML	eXtensible Markup Language
HTML	Hyper Text Markup Language
GEO	Gene Expression Omnibus
GPLxxx	Formato de Número de Acceso GEO de una plataforma
GSExxx	Formato de Número de Acceso GEO de una serie
GSMxxx	Formato de Número de Acceso GEO de una muestra
GEPAS	Gene Expression Profile Analysis Suite
ABA	ácido abscísico
SOM	Self-organizing maps
PAM	Partioning Around Medoids
CLARA	Clustering LARge Aplications
KDD	Knowledge Discovery Database

INDICE

RESUMEN

Los experimentos de micromatrices de DNA permiten obtener información sobre la expresión conjunta de cientos o miles de genes, lo que ha producido un importante incremento en el volumen de datos disponibles en el área de las ciencias biológicas. Sin embargo, esta disponibilidad de información no ha implicado un aumento proporcional en el avance del conocimiento relacionado. La minería de datos (data-mining) surge como una tecnología emergente que sirve de soporte para el descubrimiento de conocimiento, que se revela a partir de patrones observables en datos estructurados o asociaciones que usualmente eran desconocidas. El presente trabajo consiste en desarrollar una metodología de análisis de datos que permita descubrir conocimientos biológicamente relevantes, partiendo de datos de micromatrices de arroz almacenados en repositorios públicos, enriqueciendo esta información mediante la asociación con los términos de la Ontología de Genes (Gene Ontology, GO). La GO propone establecer descripciones coherentes de los genes a partir del desarrollo de vocabularios controlados y proporciona tres redes estructuradas de términos controlados para describir los atributos de los genes que pueden ser aplicados a cualquier organismo.

La metodología desarrollada se basa en la aplicación de paquetes de software de código abierto para el análisis de datos, como el lenguaje R, que provee un entorno de procesamiento estadístico y gráfico. R posee una instalación base y módulos que se agregan según el tipo de análisis que se realice. Entre ellos se encuentra el módulo Bioconductor que permite el análisis de datos bioinformáticos. Este tipo de iniciativas de código abierto y libre, facilitan la comunicación entre los usuarios creando comunidades que se van fortaleciendo y enriqueciendo a través de los conocimientos compartidos. Se utilizó un paquete especial del Bioconductor para consultar y rescatar información de la Base de Datos de la GO (GO.db). Estas aplicaciones, asociadas al administrador de Base de Datos MySQL, fueron usadas en el desarrollo de una pipeline para implementar los procedimientos de extracción del conocimiento propuestos en esta tesis. Se utilizaron como modelo, los datos crudos obtenidos de estudios independientes sobre perfiles de expresión de genes de arroz inducidos ante estreses abióticos.

Palabras claves: Micromatrices, Affymetrix, Agrupamiento de Genes, Ontología, Explotación de Datos, Descubrimiento del Conocimiento, Arroz, *Oryza Sativa*.

ABSTRACT

DNA microarray technology allows scientists to study the expression of thousands of genes simultaneously; however the increase of biological data has not implied a proportional growth of related knowledge. Knowledge discovery from the amount of data collected depends on the development and appropriate use of data mining and statistical tools. This work involves the application of techniques for extracting knowledge implicit previously unknown and potentially useful from the biological information obtained from gene expression studies using microarrays.

Three sets of experimental DNA microarray data from selected Oryza sativa abiotic stress experiments were analyzed using a pipeline based on MySQL database and R/Bioconductor routines. A secondary refinement process using the GO annotations was introduced to enrich the level of biological information included in the clusters. The result was a high-level biological significance categorization of microarray data based on GO resources.

Key word: Microarray, Affymetrix, Gene clustering, Gene Ontology, Data mining, Knowledge Discovery, Rice, Oryza Sativa.

I. INTRODUCCIÓN

I.1. Gen y Genoma

Todos los individuos de una misma especie son básicamente iguales y al mismo tiempo poseen rasgos que los distinguen físicamente dentro de la misma especie. Estas semejanzas y diferencias provienen de las instrucciones genéticas, contenidas en los genes, y que determinan las características particulares, las que son heredadas de padres a hijos **[1].**

Podemos definir al gen como la unidad física y funcional de herencia, que lleva información de una generación a la siguiente.

Un gen está conformado por una secuencia específica de ácido desoxirribonucleico (ADN) que contiene la información necesaria para sintetizar una molécula de ácido ribonucleico (ARN) que a su vez se puede traducir en una proteína.

La mayoría de nuestro ADN corresponde a:

- Genes que codifican para todas las proteínas y enzimas que necesita el organismo
- ADN con función estructural
- ADN repetitivo (secuencias cortas o largas que se repiten cientos o miles de veces dispersas a lo largo del genoma) ó
- Reliquias que han quedado de la evolución o genes que han dejado de ser funcionales en la actualidad por alguna razón (pseudogenes)

Toda esta colección de secuencias constituye el genoma de un organismo. De manera que la definición más simple de **genoma nuclear es todo el ADN presente dentro del núcleo de la célula de una especie [2].** Dicho de otra manera, un genoma es la totalidad del ADN de un organismo vivo. Es el conjunto completo de instrucciones genéticas para la construcción, el funcionamiento y el mantenimiento de dicho organismo **[1].**

I.2. Los ácidos nucleótidos, polinucleótidos, oligonucléotidos, ADN y ARN.

El ADN es el conjunto de instrucciones para determinar las características de un individuo.

Las células están formadas por diferentes tipos de moléculas, por ejemplo: agua, minerales, proteínas, azúcares, grasas y ADN. De ellas, las proteínas son particularmente

importantes, ya que son los componentes fundamentales del cuerpo que determinan como se organizan y actúan todas las moléculas. Por lo tanto, las proteínas juegan un papel esencial tanto en nuestro aspecto como en la manera en que crecemos.

El ADN actúa como código molecular para la creación de estas proteínas. La secuencia de ADN de cada gen proporciona instrucciones para hacer una proteína. Es decir que en su conjunto, todo el ADN del genoma puede interpretarse como el plano de un individuo y las proteínas son utilizadas para la construcción y el funcionamiento del mismo **[1].**

La estructura básica de los ácidos nucleicos, ADN y ARN, son los nucleótidos. Podemos clasificar a los ácidos nucleicos en dos tipos.

ADN formada por una doble cadena de desoxinucleótidos.

ARN formada por una simple cadena de ribonucleótidos **[3].**

Los nucleótidos se construyen

- con la base nitrogenada
  - adenina (A) en ADN y ARN
  - citosina (C) en ADN y ARN
  - guanina (G) en ADN y ARN
  - timina (T) en ADN
  - uracilo (U) en ARN
- un azúcar (del tipo desoxirribosa o ribosa) y
- un fosfato **[2].**

Los **polinucleótidos** se forman uniendo nucleótidos en forma de cadena, tanto en el ADN como en el ARN.

Los **oligonucleótidos** son secuencias de ADN o ARN entre 20 bases (ARN) o pares de bases (ADN) y 70 nucleótidos **[3].**

El **ADN** está formado por una doble cadena en forma una doble hélice que va rotando hacia la derecha y hace una vuelta completa cada 10 nucleótidos en el tipo de conformación más usual. En el tipo de conformación más usual estas dos cadenas se enfrentan entre sí apareando las bases nitrogenadas de los nucleótidos según las reglas que propusieron Watson y Crik (1953) **[27].**

(A) adenina - timina (T)
(C) citosina - guanina (G) **[2].**

Esta escalera helicoidal tiene el esqueleto de azúcar (desoxirribosa) y fosfato por fuera y las bases apareadas hacia adentro. En la Figura 1 observamos la doble cadena de ADN **[3].**

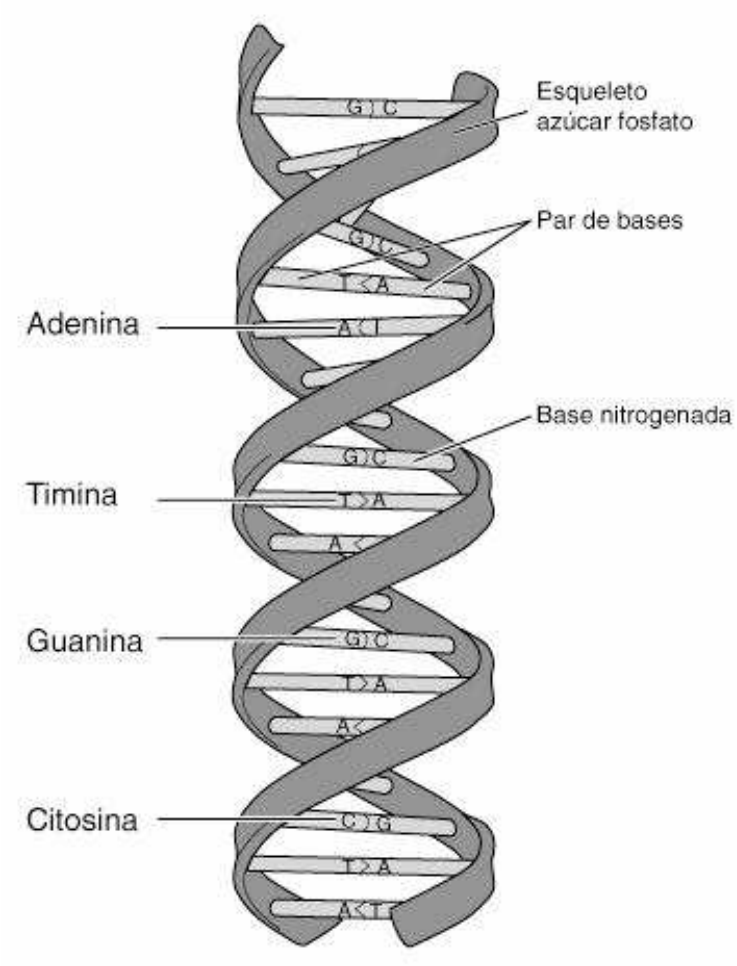

Figura 1: Doble cadena ADN [3].

El **ARN** es un polinucleótido de simple cadena con las mismas bases que las del ADN, salvo la timina (T) que es reemplazada por el uracilo (U) y el azúcar es ribosa **[3].**

I.3. Dogma central de la biología molecular

Todas nuestras células contienen la misma información genética. Entonces ¿qué es lo que hace que, por ejemplo, las células de la piel sean diferentes de las del hígado?

Estas diferencias resultan del hecho que **diferentes genes se expresan en diferentes niveles en los diferentes tejidos, órganos e incluso en distintos momentos del desarrollo.**

Entonces, ¿**qué significa que un gen se exprese**?

Para responder a esta pregunta debemos recurrir al Dogma Central de la Biología Molecular, postulado por Francis Crick en 1958 **[27]**, establece lo siguiente.

> *Una porción del ADN del cromosoma se copia (transcripción) a una cadena simple de mARN (ARN mensajero) que sale del núcleo llevando consigo la información necesaria para codificar (traducción) una proteína.*

La Figura 2 muestra en forma esquemática el Dogma Central de la Biología Molecular.

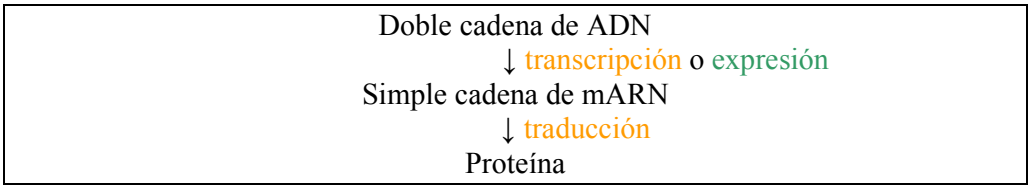

Figura 2: Dogma Central de la Biología Molecular [3].

Esto explica cómo fluye la información desde el ADN hasta las proteínas. El ADN se encuentra aislado en un sector especializado de la célula, el núcleo, el cual está rodeado por una membrana tapizada por pequeños poros.

Las proteínas, en cambio, son fabricadas en el citoplasma de las células, o sea que se encuentran en compartimentos físicamente separados.

Debido a esta separación, se necesita que algún tipo de mensajero molecular transcriba la información y la lleve desde el núcleo al citoplasma celular, que es donde se "fabrican" las proteínas. La molécula que cumple la función de trascribir la información de un lugar a otro es el ARN mensajero (mARN).

Una vez en el citoplasma, falta una traducción (cambiar el lenguaje) para generar proteínas con la información que vino desde el ADN.

De esta forma, la información fluye desde el ADN hacia el mensajero y desde el mensajero hacia la fábrica de proteínas. El ADN permanece resguardado en el núcleo y se autoduplica para perpetuar la información en las sucesivas generaciones. A partir de una única porción de ADN se pueden generar cientos de copias del mensajero para generar cientos de proteínas iguales y así amplificar la información **[2].** La figura 3, publicada por *National Institutes of Health***[1],** esquematiza el proceso de síntesis de una proteína**.**

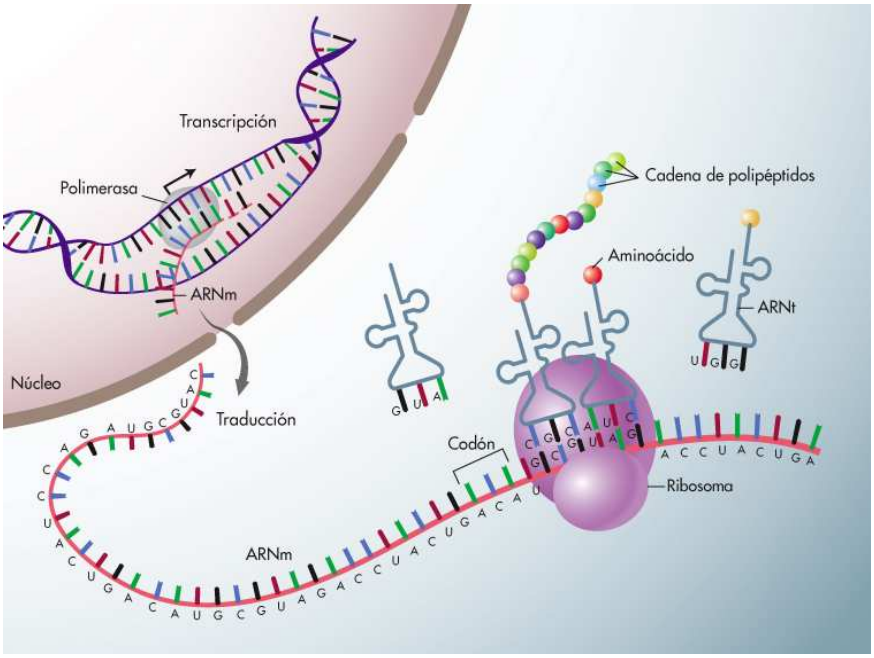

Figura 3: Cómo se crea una proteína.

Cada base de nucleótido (A, C, T, G) en una hebra de la cadena de ADN en doble hélice, tiene unida una base complementaria situada en la otra hebra. La adenina (A) siempre se une con complementaria, la timina (T). La citosina (C) siempre se une a la guanina. Cuando se utiliza la información de un gen para hacer una proteína, primero se "transcribe" (se copia) a una molécula de ARN mensajero (mARN). Las hebras complementarias de ADN se "separan" para dejar expuesto el gen codificado, y un mecanismo molecular conocido como polimerasa crea una hebra complementaria de mARN. Las moléculas de mARN salen del núcleo de la célula y se mueven a un ribosoma, donde los codones que forman el código genético especifican los aminoácidos particulares que son necesarios para formar la proteína dada. El mARN asociado con un ribosoma requiere un aminoácido particular, según lo determina un "código genético". Cada aminoácido es llevado al ribosoma por otro tipo especial de ARN denominado ARN de transferencia (tARN). Estos tARN son específicos para el aminoácido particular que transportan, y reconocen los codones a lo largo del mARN. Al ser llevado cada aminoácido al ribosoma por el tARN, y ser añadido a una cadena creciente de polipéptidos, el ribosoma avanza a lo largo de la cadena de mARN hasta llegar al siguiente codón, y así sucesivamente hasta completar toda la secuencia. La cadena completa de polipéptidos puede entonces doblarse y ensamblarse para obtener una proteína funcional [1].

Dentro de cada porción de DNA, que llamaremos gen, hay segmentos conocidos que tienen un papel activo en el proceso de codificación (exones, es la parte del mARN que sale del núcleo luego de la transcripción) y también hay otros segmentos que no codifican (intrones), son partes del mensajero inmaduro que se "editan" antes de que éste salga del núcleo y esté listo para la traducción.
El mARN que sale del núcleo, denominado mARN maduro, sólo tiene los exones y en general es más corto que la porción de ADN que lo codificó.
Para contestar la pregunta del comienzo de este apartado:

- *Cualquier secuencia (cadena genómica o gen) que es copiada a un mensajero, y que eventualmente se traducirá a proteína, se dice que esta* ***expresada.***

- *El* ***nivel de expresión*** *de un gen es la cantidad de copias de mARN transcriptos presentes en la célula en un determinado momento* **[3].**

I.4. La bioinformática y la extracción de conocimiento de datos biológicos.

La bioinformática es el área de las ciencias de la informática que aborda la búsqueda de soluciones a los problemas biológicos con herramientas computacionales. Tanto el análisis genómico como sus disciplinas relacionadas pueden ser abordados desde esta nueva perspectiva a partir de la enorme disponibilidad de secuencias moleculares acumuladas. Existen bases de datos internacionales donde se encuentran secuencias como GenBank, que está disponible a través de los servidores del NCBI (National Center of Biotechnology Information, Centro Nacional para la Información Biotecnológica) en Estados Unidos; en Europa se encuentra la base de datos EMBL (European Molecular Biology Laboratory) que administra el EBI (European Bioinformatics Institute) y en Japón la base de datos DDBJ (DNA Data Bank of Japan) localizada en el CBI-DDBJ (Center for Information Biology and DNA Data Bank of Japan). Estas tres entidades sincronizan los datos frecuentemente entre ellas. Esto, junto con el desarrollo de las tecnologías informáticas y la llegada de internet han transformado la estructura de almacenamiento y acceso a los datos y han permitido el desarrollo de aspectos fundamentales para el avance de los conocimientos sobre los sistemas biológicos.

Según la definición del NCBI, bioinformática es el campo de la ciencia en la que la biología, la computación y la tecnología de la información se fusionan para formar una sola disciplina. El objetivo final del campo es permitir el descubrimiento de nuevas ideas biológicas, así como a crear una perspectiva global unificando los principios que en biología se pueden discernir" **[4].**

La situación actual sobre los métodos de extracción de conocimientos de datos biológicos son expuestos por *Larrañaga et al* **[5]** en cuya Introducción expresa:

"El crecimiento exponencial de la cantidad de datos biológicos disponible implica dos problemas:

1. *El eficiente almacenamiento y manejo de la información y*
2. *La extracción de información útil de esos datos.*

El segundo problema es uno de los principales desafíos en biología computacional, que requiere el desarrollo de herramientas y métodos capaces de transformar la heterogeneidad de datos en ***conocimiento biológico sobre el mecanismo subyacente****. Estas herramientas y métodos pueden permitirnos ir más allá de la mera descripción de los datos y proveer de conocimiento en forma de modelos. Por la abstracción simplificada que constituye un modelo, nos permitirá obtener una predicción del sistema.*

Esta revisión abarca las diferentes técnicas utilizadas por la bioinformática como: clasificación supervisada, agrupamiento, modelos probabilísticos, gráficos y optimización. Dichas técnicas se aplican en distintos dominios biológicos para la extracción de conocimiento. Uno de estos dominios corresponde a las micromatrices, que demandan de los métodos computacionales el manejo de grandes cantidades de datos y experimentos complejos, que implican dos problemas diferentes:

1. ***Preprocesamiento de los datos.*** *Consiste en modificar los mismos para ser utilizados por los algoritmos de aprendizaje automático.*
2. ***Análisis de los datos.*** *La elección del método depende del objetivo que se tiene. La aplicación más típica es la identificación de patrones de expresión, clasificación e inducción a redes genéticas* **[5].**

I.5. Las micromatrices

El desarrollo reciente de las tecnologías asociadas al desarrollo y evaluación de micromatrices de ADN (1990) hizo posible obtener medidas cuantitativas de la expresión de genes de un experimento biológico. La posibilidad de interrogar de manera concertada y para una condición en particular un importante número de genes facilita la visión global de la expresión de genes y se puede usar en muchas aplicaciones, como:

- Buscar agrupaciones de genes, identificando aquellos que se expresan de manera similar en distintas condiciones experimentales.
- Caracterizar diferencias celulares entre diferentes tipos de tejidos (células normales y cancerosas, ó diferentes respuestas a tratamientos, ó células control y tratadas con una droga en particular)
- Evaluar muestras tomadas en series temporales durante un proceso biológico.

Por lo tanto, identificar los genes diferencialmente expresados es una de las aplicaciones más importantes e inmediatas del análisis de datos de micromatrices **[6].**

I.5.1. ¿Qué es una micromatriz ?

Es un soporte sólido, generalmente de vidrio o silicio, al que se le han adherido, mediante un robot, en forma ordenada sondas (probes) con diferentes cadenas conocidas de material genético (ADN, cADN, oligos) (cubriendo parte o toda la secuencia de un genoma – transcriptoma de un organismo), en forma de matriz de miles puntos (10.000 – 40.000) equiespaciados. Cada secuencia se asocia con un único gen. Cada punto, o sonda (*probe*), contiene millones de secuencias clonadas "idénticas" **[7].**

I.5.2. ¿Cómo se construye una micromatriz?

El proceso de construcción de una micromatriz puede clasificarse en dos grandes grupos:

- Síntesis y pegado: sondas sintetizados aparte mediante PCR – *polymerase chain reaction*-, utilizando BACs –*bacterial artificial chromosomes*-, y luego fijados al soporte sólido (*spotted DNA microarray*) por impresión de contacto.
- Síntesis *in situ*: cada sonda se construye base por base directamente sobre la micromatriz mediante la utilización de diferentes procesos, que varían con los

fabricantes, por ejemplo, fotolitografía. De este procedimiento resultan cadenas cortas de oligonucleótidos (20 a 70 bases) **[3].**

En las Figuras 4 y 5 se observa un ejemplo de micromatriz fabricada por la empresa Affymetrix (izquierda). El tamaño real se puede apreciar en comparación con el tamaño de la mano (derecha).

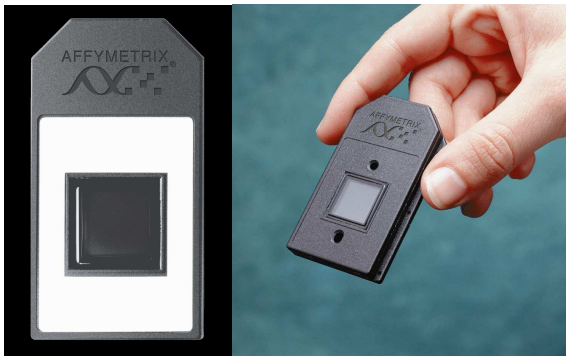

Figura 4: Imagen de un chip conteniendo una micromatriz y su tamaño real en relación con el tamaño de una mano (Imagen Cortesía de Affymetrix).

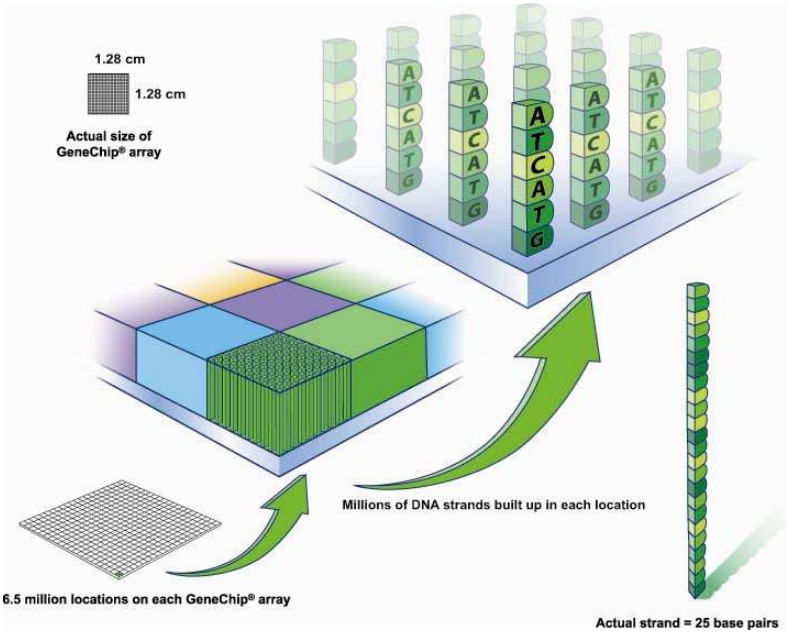

Figura 5: Describe de manera esquematizada y general cómo esta construida una matriz de ADN de la firma Affymetrix. En la parte superior izquierda se observan las medidas de 1,28 cm de lado de la matriz de sondas. A la derecha abajo, esquematiza una de las cadenas de 25 pares que representa un sector específico de un gen particular. En el centro se observan ampliaciones sucesivas de la micromatriz (Imagen Cortesía de Affymetrix).

I.5.3. ¿Para qué sirve una micromatriz?

El uso de micromatrices de ADN en experimentos en el área de genómica funcional permite medir simultáneamente la actividad concertada de cientos o miles de genes e inferir la interacción entre los mismos. Una aplicación típica de las micromatrices radica en la utilización de las mismas para la identificación de genes expresados diferencialmente **[7].**

I.5.4. ¿Cómo actúan las sondas de una micromatriz?

En un experimento con micromatrices el mARN maduro (Figura 6) de uno o más tejidos se extrae para hibridarlo con el material que se encuentra previamente depositado sobre la micromatriz. La micromatriz actúa como un detector de la cantidad de ARN mensajero presente en el tejido **[3].**

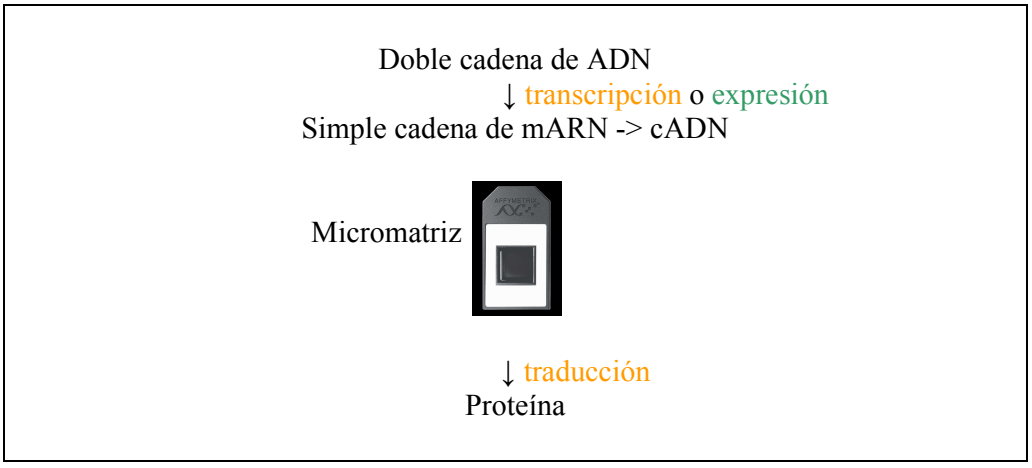

Figura 6: Se muestra en el flujo de información propuesto por el Dogma Central de la Biología Molecular, dónde se ubica el mARN en el proceso de expresión de las regiones codificantes del genoma y que proceso se analiza con una micromatriz [3].

El principio biológico de la ***complementariedad*** por el cual estas sondas aparean bajo condiciones específicas es el mismo que determina que el ADN en las células tenga una estructura de doble cadena. Establece que las secuencias de ADN o de ARN que tienen bases complementarias tienden a pegarse.

Por ejemplo, las dos secuencias que siguen son complementarias

```
… A A A G C T A G T C G A T G C T A G …
    | | | | | | | | | | | | | | | | |
… T T T C G A T C A G C T A C G A T C …
```

Para cada gen o conjunto de genes en particular que interese ser estudiado en un determinado tejido u órgano (*target, blanco, objetivo*) se puede construir una sonda o *probe* utilizando el principio de complementariedad. La posición de la sonda en la micromatriz nos referencia un gen específico.

Cada sonda de la micromatriz actúa a modo de tubo de ensayo. Al poner en contacto bajo condiciones controladas la sonda incógnita presente en una muestra correspondiente con las sondas identificadas e impresas en la matriz, aquellas cadenas que sean complementarias se pegan o hibridizan por el principio antedicho, formando una doble cadena **[3].**

I.5.5. ¿Cómo es la metodología del uso de las micromatrices?

Para determinar la expresión global de genes por medio de micromatrices se requiere la obtención de ARN mensajero (mARN) de distintas fuentes, por ejemplo, las células y/o tejidos a estudiar (células/tejidos con una característica especial o que han sido estimuladas) y las células/tejidos control (células normales o que no han sido estimuladas). Seguidamente, a partir del mARN se sintetiza ADN complementario (cADN) por medio de una reacción de transcripción inversa en la que simultáneamente se agregan nucleótidos marcados con fluorocromos tales como Cyanina-3 (Cy-3) y Cyanina-5 (Cy-5). Dependiendo del diseño experimental, se puede utilizar la Cy-3 para la muestra control y Cy-5 para la muestra problema.

Cy-3 emite una fluorescencia verde, mientras que la Cy-5 emite una fluorescencia roja. Esta marcación diferencial permite no sólo localizar las señales fluorescentes en la micromatriz facilitando la identificación de genes, sino que también permite la cuantificación de las señales correspondientes a cada uno de ellos como una medida directa del grado de expresión del mARN correspondiente.

El proceso incluye una fase donde se prepara una mezcla de reacción que incluye el cADN marcado (al que se lo denomina blanco) que se pone en contacto con la micromatriz bajo condiciones controladas para favorecer la hibridación con las sondas

presentes en ella (Figura 7). Esta explicación es para micromatrices de dos canales y la figura del ejemplo es de un solo canal. **[4].**

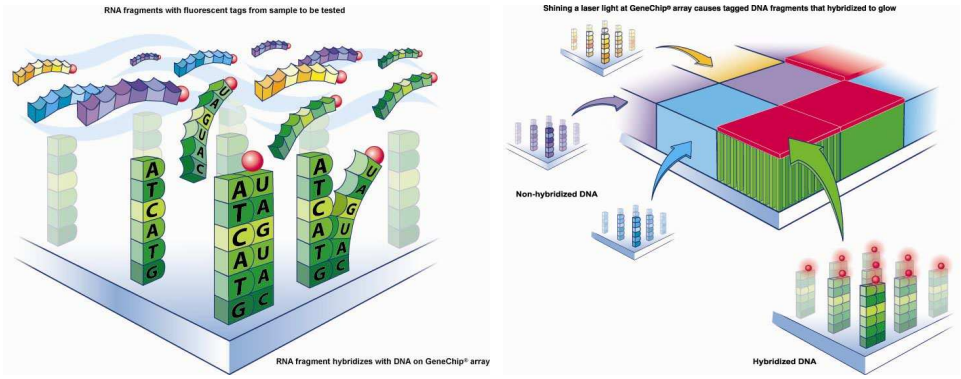

Figura 7: Esquema del proceso de hibridación de las sondas incógnitas marcadas con un florocromo con las moléculas pegadas a la superficie de la micromatriz (Imagen Cortesía de Affymetrix)

Luego de la reacción de hibridación, las micromatrices se analizan con un fluorómetro para cuantificar la intensidad de la fluorescencia en cada punto de la matriz que reaccionó positivamente con una sonda presente en la muestra. La intensidad medida en cada punto de la matriz es la resultante de la relación Cy3/Cy5, (verde y rojo), la cual se expresa como intensidad de fluorescencia en escala logarítmica y en números absolutos. Los colores son una representación de las ondas emitidas, no son reales:

- Si ésta es 1 (color amarillo) significa que el gen se expresa de manera similar en ambas muestras (control y problema).
- Si es menor que 1 (color verde) significa que el gen se expresó más en las células control.
- Si la relación es mayor que 1 (color rojo) significa que el gen se sobreexpresó en la muestra problema **[4].**

En el ejemplo de la Figura 8 (izquierda) se esquematiza un ejemplo de análisis con micromatrices de dos tejidos, uno normal y otro tumoral, a los cuales se etiquetó con colores verde y rojo respectivamente. Una vez hibridados, la fluorescencia se interpreta de la siguiente manera: las sondas expresadas sólo en el control se ven como puntos verdes; las sondas coexpresadas en el control y en la muestra como puntos amarillos y los expresados en la muestra, como puntos rojos.

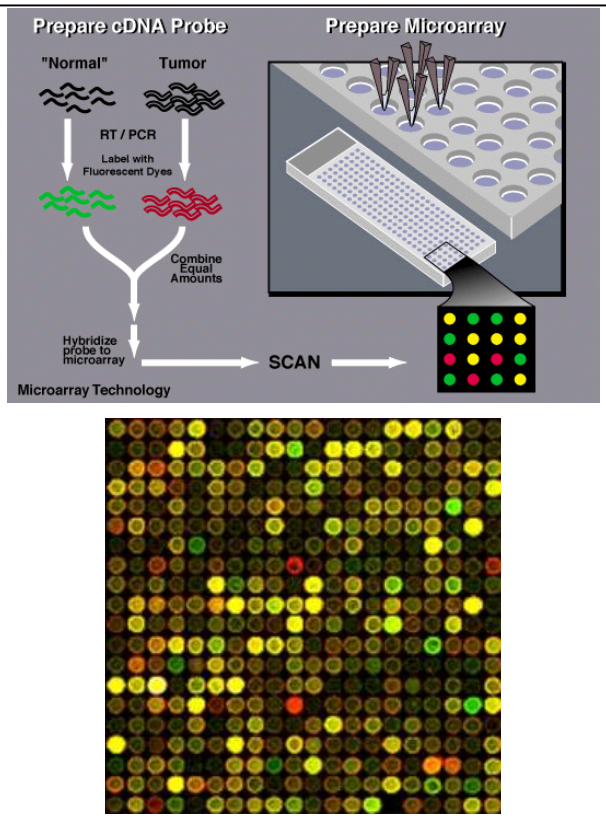

Figura 8: Esquema general de un proceso de hibridación y lectura de una micromatriz de ADN y vista real de una imagen escaneada (Imagen Cortesía de National Human Genome Research Institute).

I.5.6. Tecnología Affymetrix.

Los datos que se procesarán fueron obtenidos utilizando una matriz comercial de la firma Affymetrix específico para el genoma del arroz GeneChip® Rice Genome Array **[8].**

El GeneChip® Rice Genome Array para el análisis de expresión de genes incluye 51279 sondas correspondientes a secuencias expresadas y prediciones de secuencias que se trancriben provenientes de dos cultivares de arroz. Están representados en la micromatriz 48564 transcriptos de *Oryza Sativa* variedad japónica y 1260 transcriptos del cultivar índica. Este diseño único se creó dentro del Programa de Consorcios Affymetrix GeneChip® y proporciona a los científicos una única micromatriz que se puede utilizar para el estudio de arroz. Las secuencias para el diseño de sondas representadas en las micromatrices comerciales se obtuvieron de las bases de datos generales de secuencias de ESTs depositados en GenBank NCBI y de las bases de datos secundarias ”Gene Indices” del The Institute of Genomic Research (TIGR, actualmente Craig Venter Institute **[36]**) y

del proyecto de secuenciamento "International Rice Genome" **[37]**. Las micromatrices fueron diseñados utilizando UniGene NCBI Build #52, (7 de mayo de 2004), incorporación de genes de GenBank y el conjunto de datos TIGR Os1 v2.

I.5.7. Algunas características distintivas de la tecnología de Affymetrix

El diseño de micromatrices de Affymetrix está compuesto por un conjunto de sondas, cada una de 25 bases de largo para cada transcripto. Estas sondas se agrupan, a su vez en dos subconjuntos, uno es el *Perfect Match Probe*; PM, y el otro, *Miss Match Probe*; MM. La diferencia entre ambos es que las sondas PM son perfectamente complementarias a las secuencias de los genes de arroz y las sondas MM tienen un "mismatch", o mal apareamiento, que consiste en la transición de una base, por ejemplo A→ G, C → T en la posición central de la sonda. Los MM se usan para controlar la calidad de las hibridaciones de los PM .

Esta tecnología prevé la hibridación con una sola marca fluorescente, y utiliza un solo canal, por lo que las imágenes escaneadas con un equipo específico de la marca son imágenes en blanco y negro **[3].**

I.5.8. ¿ Cúales son los análisis de los datos generados por las micromatrices ?

La Figura 9 esquematiza los pasos de análisis de la micromatriz a partir de la obtención de la imagen digital de la intensidad de fluorescencia capturada en el mismo para el caso de las tecnologías basadas en el uso simultáneo de sondas de dos colores (dos canales) micromatriz **[7].**

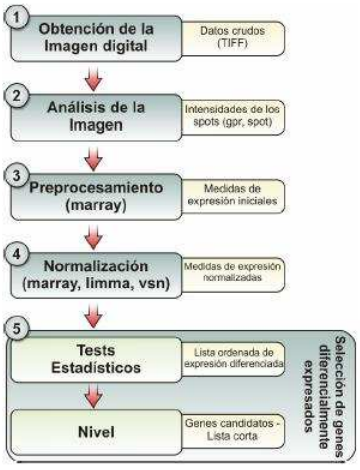

Figura 9: Esquema general de los pasos de análisis de datos provenientes de micromatrices de dos canales

I.5.8.1 Obtención de la imagen digital

Para captar la emisión de las micromatrices hibridadas de acuerdo a cualquiera de las tecnologías mencionadas que hacen uso de sondas fluorescentes, se usan escáneres de diferentes tecnologías. Estos actúan excitando cada grupo fluorescente unido al blanco mediante una luz monocromática producida por un láser o luz blanca y colectando la luz de emisión (fluorescencia) convirtiendo la corriente de fotones en valores digitales que pueden ser almacenados en una computadora como un archivo de imagen. A cada fluorocromo corresponde una longitud de onda de excitación y una longitud de onda de emisión diferente.

Para un experimento típico de micromatrices, se generan dos imágenes, una para cada emisión fluorescente, que se componen al final por superposición en una única imagen. Estas imágenes constituyen los datos crudos del experimento a partir de los cuales se extrae una medida relativa de la abundancia de transcriptos presentes en la muestra incógnita referida a las sondas representadas en la micromatriz **[7].**

I.5.8.2 Análisis de la imagen

El primer paso en el análisis de la imagen obtenida es medir el ruido de fondo inespecífico y las señales específicas de cada punto de la micromatriz para los canales rojo y verde. Una vez obtenida la imagen el análisis de la misma puede separarse en tres etapas como se observa en la figura 10.

Figura 10: Esquema de las etapas del análisis de imágenes

- **Grillado o direccionamiento:** Es el proceso de asignación de coordenadas para identificar la localización de cada punto que representa una sonda particular sobre la matriz. Si bien los puntos en cada bloque se encuentran equiespaciados, pueden existir pequeñas variaciones durante la impresión de la matriz, las cuales pueden causar distorsiones significativas en la imagen escaneada.

- **Segmentación de la imagen:** Es el proceso de clasificación de los pixeles como señal (foreground) y fondo (background).
- **Extracción de la intensidad:** Se calculan las intensidades de fluorescencia para los canales rojo y verde correspondiente a cada punto sobre la micromatriz y estima medidas de calidad de cada uno de los puntos **[7].**

I.5.8.3 Preprocesamiento

El preprocesamiento de los datos escaneados en el punto anterior consiste de estadísticas descriptivas (n, promedio, desvío estándar, valores máximos y mínimos) y representaciones gráficas que permiten visualizar errores sistemáticos vinculados con factores como las diferencias en la eficiencia de la incorporación de fluorocromos en las cantidades de mARN entre muestras, en los parámetros de escaneado, entre otros **[7].**

I.5.8.4 Normalización

El propósito de la normalización es solucionar los errores sistemáticos detectados en el preprocesamiento. Una de las variaciones más frecuentes en la tecnología de dos canales es la diferencia de intensidad del canal rojo en relación a las intensidades medidas en el canal verde. Usualmente estas intensidades suelen no ser constantes tanto entre puntos de un mismo arreglo como entre arreglos En la figura 11 observamos las dos etapas del proceso de normalización **[7].**

Figura 11: Esquema de las etapas de Normalización

I.5.8.5 Selección de genes diferencialmente expresados

El último paso del análisis es identificar los genes diferencialmente expresados (DE). Los pasos de identificación de los mismos pueden separarse en dos etapas.

- seleccionar un estadístico que permita clasificar los genes en función de los valores medidos de expresión. Algunos de estos estadísticos se observan en la Figura 12

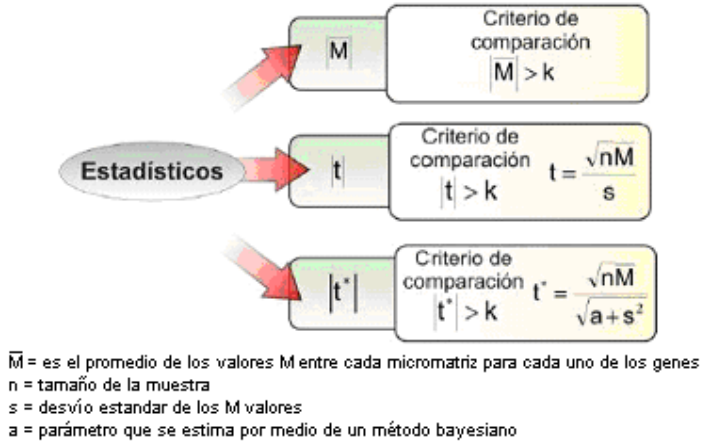

Figura 12: Distintos estadísticos utilizados para identificar los genes diferencialmente expresados

- determinar un valor crítico para el ordenamiento anterior de los genes por encima del cual cualquier valor resulta significativo y ubica el gen en cuestión dentro del grupo de genes DE **[7].**

I.5.9. Estándares de calidad en los ensayos con micromatrices de ADN. La *Microarray Gene Expresión Data Society* (MGED) y los Estándares

La diversidad de plataformas de micromatices sobre las que se realizan ensayos biológicos ha demandado el establecimiento de estándares que faciliten las comparaciones entre sistemas. En Noviembre de 1999 fue fundada la Sociedad de Datos de Expresión de Genes de Micromatriz -*Microarray Gene Expresión Data Society*- (MGED) con la intención de establecer estándares para la anotación de datos de micromatrices y permitir la creación de bases de datos públicas de de expresión concertada de genes. En esta sociedad están representadas muchas de las más importantes instituciones relacionadas con el desarrollo y aplicación de micromatrices; institutos de investigación, universidades, organizaciones comerciales y revistas científicas.

El consorcio MGED se reúne anualmente para discutir y desarrollar su trabajo en cuatro grupos de interés:

- MIAME: Minimal Information About a Microarray Experiment (*Desarrollado en extensión en el punto 1.5.10*)
- Ontologías: Determinar ontologías para describir experimentos de micromatrices usando vocabularios controlados para describir genes, muestras y datos.

- MAGE: Formular modelos objeto (MAGE-OM), lenguaje de intercambio (MAGE-ML) y módulos de software (MAGE-stk) para implementación del sofware de micromatrices.
- Transformaciones: Determina recomendaciones describiendo métodos de transformación, normalización y estandarización de datos de micromatrices

I.5.10. MIAME – Minimal Information About a Microarray Experiment.

Determina la información mínima requerida para almacenar un experimento de micromatrices con el objeto de describir y compartir el experimento en la comunidad científica **[9].**

Abarca dos áreas:

- Descripción del diseño del arreglo: incluye detalles de diseño de la micromatriz, incluyendo factores físicos (tamaño y material), factores químicos (tipos de attachment) y factores lógicos (secuencias) de la micromatriz.
- Descripción del experimento: incluye información detallada del diseño experimental; Las muestras *(Sample)* usadas, el método de extracción, preparación y etiquetado de las mismas; los procedimientos de hibridación y parámetros de lectura; los datos medidos y las especificaciones de procesamiento de los datos obtenidos.

I.6. Ontologías

Las bases de datos bioinformáticas, como GenBank, almacenan para cada gen depositado la secuencia de nucleótidos, uno o más identificadores, las coordenadas que indican dónde se encuentran puntos de interés sobre esa secuencia y una descripción de la función de ese gen: para qué proteína codifica, cuál es la función de esa proteína, en qué vía metabólica interviene, etc. Esta descripción se llama anotación y a medida que las bases de datos crecen se transforma en una valiosa fuente de información sobre la cual realizar búsquedas, comparaciones y agrupamientos. Un problema de las anotaciones tradicionales es que se realizan con un lenguaje libre: un operador humano escribe la anotación, que debe ser precisa e informativa, pero no sigue reglas en cuanto a su construcción, lo que dificulta la minería. Esta situación surgió en otros campos también, y la respuesta fue generar vocabularios controlados para realizar descripciones donde los términos de estos vocabularios están relacionados entre sí de manera jerárquica, yendo desde términos muy generales a muy específicos **[9]**.

Un tipo de ontología muy anterior a las realizadas para el descubrimiento del conocimiento son las taxonomías biológicas. En ellas los seres vivos se agrupan en categorías muy generales o reinos, que se dividen en categorías cada vez más específicas, hasta llegar a especies, incluso a veces, puede detallarse aún más hasta niveles más específicos como subespecie, cepa, cultivar, etc.

El objetivo de las ontologías es dar un marco para una representación formal de un sujeto que incluye el vocabulario o nombres para referirnos a una materia determinada y cómo estos términos están relacionados entre sí. El uso de ontologías asociadas al análisis de micromatrices permite establecer un marco conceptual a partir del cual se pueden diseñar bases de datos de expresión facilitando la consulta de las mismas evitando ambigüedades. Dentro de las ontologías aplicables al proceso de análisis de micromatrices se encuentran las establecidas por el consorcio Gene Ontologies (GO) **[17]** para la anotación de productos de genes provistos.

El uso extensivo de la tecnología de micromatrices y la necesidad de compartir los datos obtenidos a partir de estos experimentos entre quienes trabajan con un sistema biológico similar ha promovido el establecimiento de un protocolo que define la mínima información que debe reportarse sobre este tipo de experimentos para asegurar la interpretación y reproducción precisa de los mismos. Este estándar se denomina MIAME (*Minimum Information About a Microarray Experiment*; www.mged.org/miame). Asociado a este estándar, surgieron ontologías específicas, como MGED *Ontology* (MO) desarrollada por el grupo de trabajo de ontologías de MGED, las cuales proveen los términos para anotar todos los aspectos relacionados a los experimentos con micromatrices según MIAME. El MAGE-OM (*Microarray Gene Expression Object Model*) y el MAGE-ML (*MicroArray Gene Expresión Markup Language*) definen un formato sintáctico no ambiguo para el intercambio de datos de acuerdo a los lineamientos de MIAME **[38].**

I.6.1. Ontologías Genéticas

En 1999 se creó el Gene Ontology (GO) Consortium para definir un vocabulario controlado aplicable a la anotación de genomas que permita la descripción precisa de los productos de genes, facilite la comparación entre especies y simplifique las consultas de bases de datos.

La GO describe tres tipos de ontologías independientes:

- **Función Molecular**: relativo a la actividad o función bioquímica que cumple el producto de un gen. Ejemplos: factor de trascripción, fosforilasa, etc.
- **Proceso Biológico:** relativo a los que son llevados a cabo por conjuntos ordenados de funciones moleculares definidos en un sentido, por ejemplo, "mitosis", "metabolismo de purinas", etc.
- **Componente Celular:** relativo a estructuras subcelulares, localizaciones, complejos macromoleculares. Ejemplos: núcleo, telómero, mitocondria, etc.

Cualquier gen puede ser mapeado en estas ontologías. O dicho de otra forma, el producto de un gen individual tiene una función molecular, es parte de algún proceso biológico y ocurre en algún componente celular.
Cada término GO tiene varios campos asociados:

- GO ID: Identificador numérico único.
- Sinónimo: Término alternativo para algunos términos.
- Ultima modificación: Día y hora en el que el GO ID fue modificado
- Padres: Clase más general a la cual pertenece el término.
- Hijos: Clase más específica derivada del término.

En la ontología clásica, cada término tiene un solo padre. Sin embargo, debido a la complejidad de la información biológica, en la GO cada término puede tener más de un padre. Algunas veces es útil pensar la relación de los términos como árbol; más precisamente, los términos se organizan en lo que se conoce como "**Direct Acyclic Graph" (DAG,** Gráfico Acíclico Dirigido). La Figura 13 muestra un ejemplo de DAG, por ejemplo el término "F" tiene 2 padres "B" y "C"

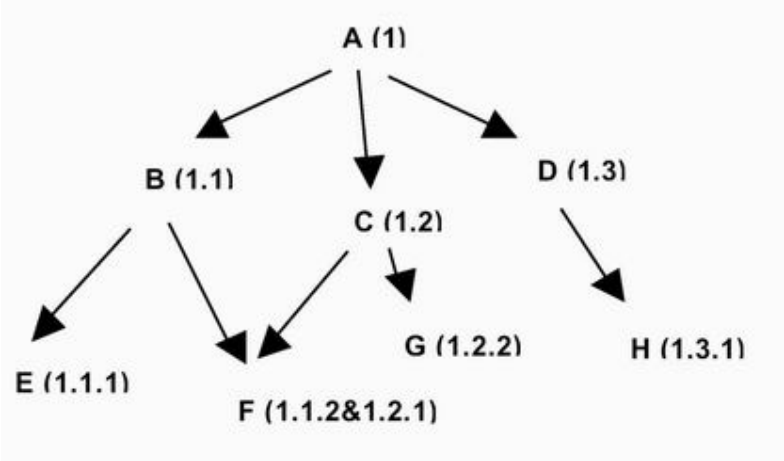

Figura 13: Ejemplo de gráfico DAG

- **Ontología MGED**

El grupo de trabajo de Ontologías del MGED ha elaborado Ontologías para Micromatrices (MO) para estandarizar las anotaciones de los experimentos que usan micromatrices las cuales incluyen tres categorías

- **Clases:** referidas a las categorías de información, por ejemplo edad, sexo o protocolo.
- **Propiedades:** referida a los atributos o propiedades de las clases, por ejemplo una Clase *Protocolo* tiene como Popiedad *tiene.cita* que toma el valor *ReferenciaBibliográfica.*
- **Individuos:** refiere a los valores reales o instancias de las clases, por ejemplo la Clase *Género* tiene como Individuo *macho, hembra, hermafrodita, sexo-mixto, desconocido* como instancias de la clase los cuales pueden ser usados para describir el género.

I.6.2. MAGE – Microarray Gene Experiment [9]

Microarray y Gene Expression (MAGE) facilita la implementación técnica de MIAME e incluye.

- **MAGE – OM**: Un modelo de objeto que representa una estructura compleja de información, tal como se presenta la información contenida en un experimento con micromatrices.

- **MAGE – ML**: es la representación del MAGE-OM. ML está basado en el lenguaje de intercambio de datos XML, el cual permite al usuario codificar la información a partir del uso de etiquetas y términos que facilitan el intercambio de información estructurada entre diferentes plataformas. Los datos almacenados com MAGE-ML pueden ser transferidos entre diferentes aplicaciones y bases de datos, incluidas aquellas referidas al almacenamiento público de datos de micromatrices como GEO o ArrayExpress (descriptas en el punto I.4.13 de esta tesis)
- **MAGE – stk** (MAGE Software ToolKit): es una aplicación que facilita la adopción de MAGE, provee una interfase para convertir MAGE–OM en MAGE–ML utilizando distintos lenguajes **[39]**.

I.7. Bases de datos públicas para experimentos con micromatrices

El trabajo de estandarización de la información de experimentos con micromatrices va aparejado con la creación de bases de datos públicas capaces de recibir, almacenar y distribuir los datos obtenidos experimentales. Existen diferentes bases de datos de expresión génica en general, entre las más consultadas se encuentran Gene Expresión Omnibus (GEO; NCBI; **[10]**) y ArrayExpress (EBI; **[41]**). Por otra parte, se está desarrollando una nueva línea de trabajo que está dirigiendo sus pasos hacia la creación de bases de datos exclusivas para micromatrices.

Los datos usados para el desarrollo de esta tesis fueron extraídos de GEO. Esta base de datos almacena y distribuye libremente datos de micromatrices depositados allí por distintos miembros de la comunidad científica siguiendo los lineamientos de MIAME. La arquitectura de GEO consta de cuatro secciones, pero a continuación se describen sólo tres **[11] [40]:**

- **Plataforma:** Definen una lista de características de la micromatriz (ej:cDNA, sondas de oligonucleotido). A cada registro de plataforma se le asigna un único y estable Número de Acceso GEO, con el formato (GPLxxx). Una plataforma puede referenciar varias Muestras y Serie.
- **Muestra:** describe el material biológico y las condiciones experimentales bajo las cuales se llevó a cabo el experimento y todos los datos obtenidos de él. A cada registro Sample es asignado a un único y estable Número de Acceso GEO (GSMxxx)

- **Series:** define un grupo de Muestras relacionadas y consideradas parte de un mismo experimento. A cada registro de la serie se le asigna un único y estable Número de Acceso GEO (GSExxx)

Para facilitar los procesos de minería de datos, GEO organiza en colecciones de conjunto de datos todas aquellas muestras ("Samples") experimentales relacionados y procesados de manera similar.

I.8. Explotación de datos y descubrimiento del conocimiento en bioinformática

El proceso de descubrimiento del conocimiento (KDD – Knowledge Discovery in Database) tiene como objetivo el descubrimiento de información implícita, no trivial, previamente desconocida y potencialmente útil de una gran base de datos. En la Figura 14 vemos el proceso de Descubrimiento del Conocimiento.

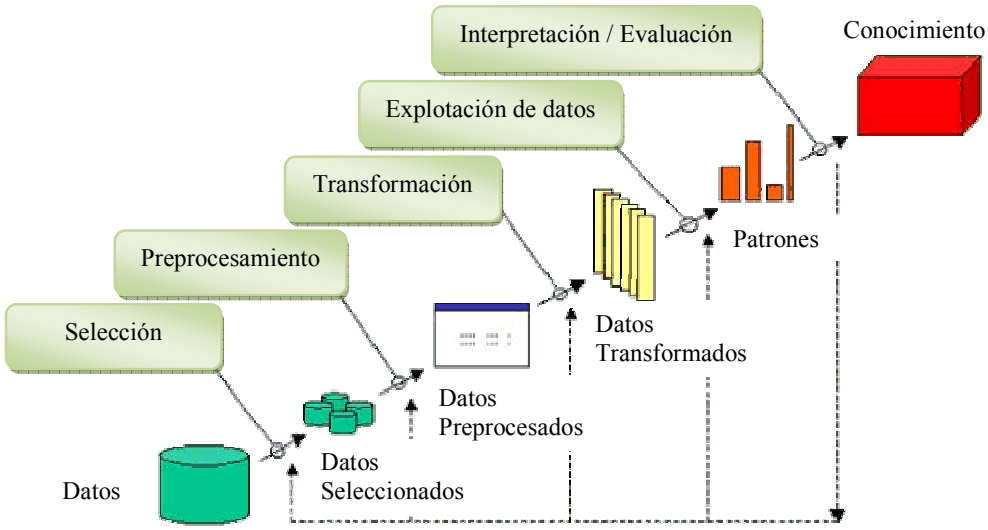

Figura 14: Proceso de Descubrimiento del Conocimiento

El proceso de KDD comienza comprendiendo el dominio de la aplicación, entendiendo cuales sus los conocimientos relevantes y los objetivos de los usuarios. La **selección** es el proceso por el cual se obtiene un conjunto de datos, focalizando en un subconjunto de registros o de variables. En el **preprocesamiento** se limpian los datos, eliminando los datos erróneos, detectando y validando los valores extremos, recolectando información necesaria para modelar, determinando estrategias para el manejo de datos perdidos, etc.

La **transformación de los datos** se enfoca en reducir la dimensionalidad o utilizar métodos de trasformación que reducen el número de variables a considerar. Luego se ejecuta la tarea de explotación de datos dependiendo del objetivo del proceso de KDD se elige el método más adecuado (clasificación, regresión, agrupamiento, etc.) y se define en consecuencia el algoritmo y los parámetros para la búsqueda de patrones presentes en los datos. Como resultado de este procedimiento se buscan patrones de interés en una forma de representación particular o un conjunto de representaciones como ser, reglas de clasificación o árboles, regresión, agrupamientos, matrices con pesos sinápticos de las redes neuronales, etc. Estos patrones obtenidos son interpretados y evaluados buscando consolidar el conocimiento descubierto. Como se observa en las flechas punteadas, el proceso es iterativo en cualquiera de sus etapas.

El proceso de preparación de los datos representa la actividad que mayor tiempo demanda, se estima en alrededor de dos tercios del tiempo total del desarrollo previsto. La aplicación de los métodos de extracción del conocimiento, si bien demanda tiempo de cómputo, se facilita porque los comandos pueden ser escritos en pequeños programas que automaticen el proceso de explotación de datos.

Para el presente trabajo de tesis se utilizaron algunos métodos de agrupamientos los que son explicados en el siguiente punto.

1.8.1. Métodos de agrupamientos

Para el procesamiento de los datos de expresión de los genes pueden utilizar distintos algoritmos de agrupamientos, a saber:

- Jerárquico
- K-medias
- Partición alrededor de medioides (PAM – Partitioning Around Medoids),
- Agrupamiento en grandes aplicaciones (CLARA – Clustering LARge Applications)
- Mapas Auto Organizados (SOM – Self-Organizing Map)

1.8.1.1. Agrupamiento Jerárquico

El propósito del análisis de conglomerados, es agrupar las observaciones de forma que los datos sean muy homogéneos dentro de los grupos (mínima varianza) y que estos grupos sean lo más heterogéneo posible entre ellos (máxima varianza). De este modo obtenemos una clasificación de los datos multivariada con la que podemos comprender mejor los

mismos y la población de la que proceden. En el caso del método jerárquico al comienzo cada gen es un cluster y en sucesivas iteraciones se van uniendo, según la distancia entre conglomerados que se haya determinado **[33].**

En primer lugar se calcula la distancia entre los genes. Existen muchas métricas para realizar este cálculo, en este trabajo se utilizaron a modo exploratorio las siguientes **[34]:**

Manhattan $D_1 = \sum_{i=1}^{M} | x_i - y_i |$

Euclidea $D_2 = \sqrt{(\sum_{i=1}^{M} (x_i - y_i)^2)}$

Canberra $D_{Can}(x,y) = \sum_{i=1}^{M} \frac{| x_i - y_i |}{| x_i + y_i |}$

Donde x_i e y_i representan la expresión de cualquiera de los genes.

La distancia Manhattan es la distancia a lo largo de los ejes de coordenadas en valor absoluto, y la distancia Euclidea se deduce del teorema de Pitágoras, ambas varían frente a cambios de escala de las variables. La distancia Canberra no varía frente a cambios de escala, es decir, el peso que se atribuye a la diferencia entre individuos es mayor cuanto menor es la dispersión en esa variable.

El paso siguiente es calcular las distancias entre los conglomerados, utilizando la matriz de distancia generada en el paso anterior. Existen también varios métodos para este propósito, los más comunes son los siguientes:

Ligamento simple $\boldsymbol{\delta(S,T) = \min_{\{x \in S, y \in T\}} d(x,y)}$,

Ligamento completo $\boldsymbol{\delta(S,T) = \max_{\{x \in S, y \in T\}} d(x,y)}$,

Ligamento promedio $\boldsymbol{\delta(S,T)} = \frac{1}{| S || T |} \sum_{x \in S, y \in T} d(x,y)$,

Ward $\boldsymbol{\delta(S,T)} = \sum_{x \in S} d^2(x, \bar{x}) + \sum_{y \in T} d^2(y, \bar{y})$.

Donde *S, T* son dos cualesquiera de los conglomerados.

El método comienza suponiendo que cada gen compone un grupo, en cada paso se unen aquellos grupos con menores distancias entre sí. Para determinar la distancia resultante entre grupos de genes se utilizan los criterios de ligamiento listados más arriba. El ligamento simple utiliza la menor distancia entre las observaciones de cada grupo; el ligamento completo, la mayor distancia entre las observaciones de cada grupo; el ligamiento promedio, el promedio de las distancias de las observaciones en cada grupo y el ligamiento de Ward, junta el par de grupos que produce la varianza más pequeña entre los grupos unidos. El agrupamiento final resultante se grafican en forma de dendograma. Este gráfico permite determinar distintos niveles de corte, en cada uno de los cuales se determina un número distinto de grupos formados. Es decir, a priori, uno no determina la cantidad de grupos, sino que en base a los gráficos y a distintas métricas al finalizar el método define la cantidad adecuada de grupos.

1.8.1.2. Agrupamiento K-medias

Es un método de particionamiento en un número especificado de K conglomerados.

En el caso de K-medias se eligen arbitrariamente K centroides: c_1, c_2, c_3,..... ,c_k.

Luego iterativamente se reasignan las observaciones a los conglomerados cuyo centroide este más cercano, es decir:

$$\mathbf{C(i) = argmin_{1\leq k\leq K}||x_i - c_k||} \ .$$

Donde **argmin** se define como el agrupamiento óptimo donde cada observación está asignada a su centroide más próximo, es decir minimizando las distancias.

Se recalculan los centroides basados en los elementos que están contenidos en el cluster, tendiendo a minimizar la suma de cuadrados dentro del cluster.

$$WSS = \sum_{k=1}^{K} \sum_{C(i)=k} || \mathbf{x}_i - \mathbf{c}_{k.} ||^2$$

Esto se realiza hasta que se cumple algún criterio de parada, por ejemplo que la suma de los cuadrados dentro de los conglomerados sea la más pequeña ó que el método cumpla una determinada cantidad de iteraciones **[34].**

1.8.1.3. Agrupamiento PAM

La partición alrededor de medioide es un método similar a K-medias, pero el centro de cada conglomerado es un medioide, es decir aquel elemento que mejor representa al grupo al que pertenece.

Para un número especificado de K agrupamientos, el procedimiento PAM está basado en la búsqueda iterativa de los K Medioides. Siendo

$$\mathbf{M = (m_1, m_2, m_3, ..., m_k)\ ,}$$

los k medioides de los k conglomerados en los que se agruparán todas las observaciones a clasificar. Para encontrar M hay que minimizar la suma de las distancias de las observaciones a su medioide más cercano

$$\mathbf{M = argmin}_{\mathbf{M}} \sum \mathbf{min}_{\mathbf{k}}\ \mathbf{d(x_i, m_k)}\ .$$

Esto es, se eligen k elementos como medioides, luego se les asigna las instancias restantes al medioide más cercano. A continuación se selecciona al azar una instancia no-medioide y se calcula el costo total de cambiar de medioide, si se mejora la calidad se cambia el medioide. Esto se realiza iterativamente hasta que no haya cambios.

1.8.1.4. Agrupamiento CLARA

La diferencia entre los algoritmos de PAM y CLARA es que este último aplica muestreos: solamente se elige una pequeña porción del total de datos. La idea es que si la muestra se seleccionó al azar, es seguro que representa correctamente el conjunto total de datos y por lo tanto, los objetos representativos (medioides) elegidos, serán similares a los que se hubieran elegido del conjunto total de datos. CLARA extrae muestras múltiples y busca agrupamientos donde las distancias sean mínimas, entre las distintas instancias y sus medioides a los que fueron asignados. Al realizar muestreos, este metodo permite trabajar con conjunto de datos más grandes que PAM, que busca los mejores medioides de K entre un conjunto total de datos, mientras que CLARA busca los mejores medioides de K entre la muestra seleccionada del conjunto total de datos.

1.8.1.5. Agrupamiento SOM

Los mapas autoorganizados (SOM) son algoritmos de particionamiento que se basan en el hecho de que los clusters pueden ser representados en una estructura regular de dimensión baja, tal como un vector o una matriz de dos dimensiones. Las más usados son los SOM en una y dos dimensiones.

Los agrupamientos que son cercanos entre sí aparecen en celdas adyacentes de la grilla. Es decir, SOM mapea el espacio de entrada de las muestras (variables que describen a la

instancia) en un espacio de menor dimensión en el cual la medida de semejanza entre las muestras está dada por la relación de cercanía de los vecinos.

Cada uno de los K clusters se representa por un objeto prototipo M_i , i = 1,2,…,K .

En este método el orden de entrada en el procesamiento de las distintas observaciones influye en la asignación a un grupo. Esto se debe a que a medida que se van procesando las distintas instancias, se modifican los vectores prototipos que contienen los pesos sinápticos y que determinan el grupo al que será asignado.

Un SOM es entrenado como una red neuronal, es decir por cada una de las instancias que se procesan, se modifican los vectores prototipos que determinan la posición y el agrupamiento. Iterando este proceso se obtienen una matriz donde se almacenan los pesos sinápticos que constituyen el conocimiento adquirido y determina los conglomerados posteriores. Se construye un gráfico en donde las observaciones similares se muestran cercanas entre sí.

El algoritmo general de SOM es:

Paso1- Seleccionar los números de filas (q_1) y columnas (q_2) en el grillado. Luego habrá $K=q_1q_2$ clusters.

Paso2- Inicializar el tamaño del parámetro de actualización $\boldsymbol{\alpha}$ (la tasa de aprendizaje en términos de redes neuronales) y el radio del grillado de aprendizaje $\boldsymbol{r}$.

Paso3- Inicializar los vectores prototipos M_j, j Є (1,…,q_1) x (1,….,q_2) mediante la elección aleatoria de K observaciones.

Paso4- Para cada observación x_i del conjunto de datos hacer los siguiente:

Identificar el vector índice j' del prototipo M_j más cercano a x_i.

Identificar un conjunto S de prototipos vecinos de M_j'. Es decir,

$S=\{j: \text{distancia } (j ,j')<r\}$.

La distancia puede ser Euclideana o cualquier otra.

Actualizar cada elemento de S moviendo el correspondiente prototipo hacia x_i:

$M_j \leftarrow M_j + \alpha (x_i-M_j)$ para todo j Є S

Paso5- Disminuir el tamaño de α y r en una cantidad predeterminada y continuar hasta alcanzar convergencia.

Se debe definir la topología de la red que define la estructura de las conexiones laterales entre neuronas, determinando la vecindad sobre la que va a tener influencia la neurona ganadora. Las topologías usuales son rectangular (**rect**) o hexagonal (**hexa**).

Otro parámetro que caracteriza la vecindad es la forma de la función por la que se cambian los valores de los vectores que hay en ella. Una forma tipo gaussiana (**gaussian**) hará que el cambio de valores disminuya con la distancia, mientras que una forma tipo burbuja (**bubble**) cambiará de la misma forma todos los vectores que pertenezcan a la vecindad.

El **radio del grillado (r)** establece el radio a partir del cual se actualizan los elementos. Este radio disminuirá a medida que avance las iteraciones de aprendizaje.

La **tasa de aprendizaje (α)** determina cómo se ajustan los valores vecinos e irá disminuyendo durante el aprendizaje.

Tanto el radio como la tasa de aprendizaje permiten en un primer momento del entrenamiento realizar un ajuste grueso, y a medida de que avancen las iteraciones el ajuste es más fino hasta llegar a la convergencia del método.

1.8.2. Enriquecimiento de conglomerados utilizando Ontologías Genéticas

Una vez analizados y clasificados los datos obtenidos de experimentos con micromatrices de ADN, el paso siguiente consiste en extraer la información biológica subyacente a los resultados experimentales. Para realizar esta aproximación se han desarrollado numerosas herramientas algunas de las cuales han sido presentadas en el punto 1.8.1, cuyo objetivo es identificar estructuras o subclases de los objetos en las bases de datos e inferir conclusiones posibles en relación al experimento analizado. La utilización de ontologías para procesos biológicos, componentes y funciones moleculares, facilita y enriquece la extracción de conocimiento. Esto se debe a que al establecer un vocabulario controlado en lugar del lenguaje natural se elimina el uso de palabras alternativas para definir una función, proceso o componente determinado. Otra ventaja del uso de estas ontologías es que sus términos están organizados de acuerdo a una estructura jerárquica, lo cual permite distinguir distintos grados de especialización o generalización de funciones, procesos o componentes cuando se analiza la información para descubrir nuevos conocimientos.

Para la utilización de esas ontologías en los procesos de minería de datos se han desarrollado diferentes programas, algunos de los cuales son herramientas que están disponibles utilizando navegadores de la web. Mayoritariamente estas herramientas están orientadas al procesamiento de datos de micromatrices provenientes de ensayos en humanos o animales. Un ejemplo de ello es GEPAS **[14],** que es una plataforma web que permite procesar los datos provenientes de micromatrices incluyendo métodos para la

normalización, determinar conglomerados (clustering), seleccionar genes, predictores y realizar anotación funcional de experimentos.

En el área de biotecnología vegetal existen algunos antecedentes de extracción del conocimiento a partir de datos obtenidos del análisis de micromatrices utilizando estas ontologías. A partir de los cuales se han identificado grupos de genes con funciones biológicamente relevantes para diferentes escenarios. Un ejemplo es el estudio integrado de un conjunto de experimentos orientados al estudio de respuestas a estreses abióticos en *Arabidopsis thaliana*, en los que se enfrentaron plantas a diferentes concentraciones de ácido abscisico (ABA), una hormona involucrada en las respuestas al estrés **[15]**. Los autores de este trabajo desarrollan una herramienta computacional para extraer conocimiento a partir de múltiples clusters de genes diferencialmente expresados bajo las condiciones de estudio determinando la relevancia biológica de los mismos haciendo uso de la información obtenida a partir del análisis de los términos GO asociados a los miembros de cada cluster. Para ello desarrollan un algoritmo que rescata los términos GO que están relacionado directamente con el número de acceso o identificador de genes presentes en cada agrupamiento. Luego caracteriza cada grupo a partir de los términos GO más frecuente dentro del mismo.

En este trabajo se propone la aplicación de un algoritmo que está basado en mapas autoorganizados, para un propósito similar de extracción del conocimiento a partir de datos públicos para ensayos de micromatrices **[16].** Este método es utilizado para agrupar genes acordes a su perfil de expresión y su anotación en la GO. Se usa también con múltiples conjuntos de datos para proveer de una mejor separación y definición biológica de los conglomerados de genes.

Existen muchas herramientas que aprovechan las anotaciones GO para complementar el análisis de experimentos de micromatrices. Entre otras funciones, estas herramientas permiten la navegación y consultas dentro de las ontologías, visualizando las anotaciones **[17].** Además existe una versión reducida de la GO denominada Slims, que incluye un subconjunto de términos GO que aportan información general del contenido de la ontología sin entrar en detalles específicos de los términos **[18].**

I.9. El Arroz – Situación Mundial y Regional.

En su recopilación de datos sobre la producción del arroz, Kraemer **[12]** expresa que es una de las cultivos más importantes del mundo, cultivadas en todos los continentes,

sumando alrededor de 150 millones de hectáreas en el año 2004, con una producción mundial de 636,7 millones de toneladas. Es una de las principales fuentes de alimentos, siendo el responsable de la alimentación de más de la mitad de la población mundial, contribuyendo con el 23 % de todas las calorías, casi lo mismo que el aporte del maíz y el trigo juntos, transformándose en la principal fuente mundial de consumo directo de calorías.

La producción de arroz en Argentina se ve reflejada en los siguientes números **[13]:**

- Superficie sembrada: 137.000 hectáreas.
- Principales provincias productoras: **Corrientes (51%)** y Entre Ríos (38%) .
- Producción anual: 730.000 toneladas.
- Destino de la producción: 45 % se exporta.
- Participación en el comercio mundial de arroz: 1%.
- Cantidad de molinos arroceros en actividad: 87.
- Consumo per cápita: 6,5kg anuales.
- Brasil es el principal destino externo del arroz argentino; representa el 45% de las exportaciones del cereal.

II. OBJETIVOS, DESARROLLO Y APORTES ORIGINALES DE LA TESIS

El objetivo de esta tesis es explorar una metodología de análisis de datos que permita descubrir conocimientos biológicamente relevantes, partiendo de datos de micromatrices de arroz almacenados en repositorios públicos. El trabajo consiste en la aplicación de técnicas para la extracción del conocimiento implícito, previamente desconocido y potencialmente útil, utilizando el volumen cada vez mayor de datos de expresión de genes de arroz (*Oryza sativa*) existente en bases de datos públicas, enriqueciendo esta información mediante la asociación con los términos de la Gene Ontology.

Se analizan datos de tres experimentos de estrés abiótico con micromatrices para estudiar la expresión génica en arroz realizados la plataforma comercial desarrollada por la empresa Affimetrix **[8]**, indexada en la base de datos de expresión GEO del NCBI con el código GPL2025 **[19]**. Se seleccionaron estos tres experimentos para analizarlos en forma exploratoria debido a que compartían la característica de que todos ellos fueron sometidos a estrés salino, de esta manera que se espera un comportamiento de la expresión genética sea parecido. La metodología desarrollada se basa en la aplicación de paquetes de software de código abierto para el análisis de datos, como es el lenguaje R **[20]**, que provee un entorno de procesamiento estadístico y gráfico. R posee una instalación base y módulos que se agregan según el tipo de análisis que se realice. Entre ellos se encuentra el módulo Bioconductor **[21]** que permite el análisis de datos bioinformáticos. Este tipo de iniciativas de código abierto y libre, facilitan la comunicación entre los usuarios creando comunidades que se van fortaleciendo y enriqueciendo a través de los conocimientos compartidos. Se utiliza el paquete GO.db de Bioconductor **[22]** el cual pone a disposición comandos para consultar y rescatar información de la base de datos de la GO. Estas aplicaciones, asociadas al administrador de base de datos MySQL **[23]**, se aplican al desarrollo de una *pipeline* para extracción del conocimiento.

La naturaleza del aporte de esta tesis es el diseño e implementación de una metodología que integra análisis de datos de expresión génica, métodos estadísticos y el uso de ontologías para reconocer agrupaciones de genes y enriquecerlas con información relativa a la función, localización y procesos en los que intervienen las proteínas codificadas por aquellos genes.

Dentro de los aspectos originales que aporta este trabajo de tesis podemos mencionar

- La utilización de experimentos conducidos por diferentes grupos de investigación que comparten una plataforma común y comercial desarrollada para conducir estudios de expresión génica en arroz. El conjunto de datos seleccionado incluye 47 muestras ("*Samples"*, ver 1.7*)* que representan tres experimentos independientes, todos focalizados al estudio de las respuestas inducidas en plantas de arroz sometidas a diferentes factores de estrés abiótico. No se conocen reportes previos que describan la exploración de datos obtenidos a partir de diferentes perspectivas biológicas, con el propósito de descubrir nuevos conocimientos implícitos. Se espera que la aproximación propuesta sobre la reutilización de estos datos compartidos dentro de un *dataset* aporte experiencia y medios para extraer conocimientos que individualmente o en forma conjunta hasta ahora no se han descubierto.
- La utilización, en los procesos de extracción del conocimiento a partir de bases de datos, de la ontología génica tanto de manera transversal, a través del uso de sus diferentes categorías -Proceso Biológico, Componente Celular y Función Molecular-, como a nivel vertical, recuperando los términos GO asociados a genes en los diferentes niveles jerárquicos propuestos con la opción de seleccionar el nivel de profundidad informativo óptimo entre la generalidad y especificidad de términos GO.

III. MATERIALES Y MÉTODOS

III.1. Micromatrices analizadas

En primer lugar se seleccionó un conjunto de datos experimentales obtenidos a partir de la interrogación de una plataforma de micromatriz desarrollada para arroz por la empresa Affymetrix por los siguientes motivos: 1) existen muchos trabajos en los repositorios públicos realizados sobre esta plataforma; 2) el diseño de la micromatriz está muy bien documentado en el sitio web de la empresa; 3) existen bibliotecas específicas desarrolladas en lenguaje R dentro del paquete Bioconductor que facilitan el procesamiento de datos proveniente de esta plataforma; 4) la selección de experimentos realizados sobre la misma plataforma reduce la variabilidad entre las muestras de los experimentos y entre los experimentos. La información detallada que describe los datos de la Plataforma Affymetrix y los enlaces desde donde se pueden descargar archivos (en formato txt y/o xml) con datos de la plataforma, se puede encontrar en los siguientes sitios:

NCBI : http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GPL2025
Affymetrix: http://www.affymetrix.com/support/technical/byproduct.affx?product=rice

Se eligieron tres experimentos del repositorio del NCBI cuya característica en común es que en todos ellos evaluaban el nivel de expresión de genes de plantas de arroz sometidas a estrés salino o hídrico. Esta selección permitía el análisis comparativo de la coexpresión genética, para cada experimento en forma individual o para los tres experimentos en conjunto. En consecuencia, es posible encontrar genes que mantuvieron perfiles de expresión similares en los tres experimentos, y otros que se expresaron solo en algunos de ellos, debido posiblemente a diferencias entre los cultivares de arroz usados, o las condiciones específicas de estrés de cada tratamiento.

Los experimentos seleccionados fueron:

GSE3035: Se evaluaron las líneas de arroz FL478 e IR29, ambas índica, sometidas a estrés salino. Posee 11 muestras.
url: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc= GSE3035.

GSE4438: Analiza la respuesta a estrés salino en la etapa de reproducción temprana (inicio de la panoja), se evaluaron las siguientes líneas: m103 (japónica sensitiva), agami (japónica tolerante), ir29 (índica sensitiva) y ir63731 (índica tolerante). Posee 24 muestras.
url: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4438.

GSE6901: Se evalúan respuestas al estrés por salinidad, sequía y frío, se evaluó la línea ir64 (índica). Posee 12 muestras.
url: http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6901.

El número total de datos de expresión genética extraídos por muestra es de **57.381**, en total se analizaron 47 muestras distribuidas en 3 experimentos cada uno de los cuales incluye 11, 24 y 12 muestras respectivamente (Tabla 1). La cantidad de datos de expresión génica procesados es de $57.381 \times 47 = 2.696.907$.

Tabla 1: Micromatrices de expresión génica analizados. Los números de acceso provistos corresponden a la base de datos GEO (NCBI). La cantidad de muestras (47) corresponde con la cantidad de micromatrices analizadas. La columna Etiqueta de la muestra indica si trata de un control o describe un tratamiento aplicado.

Plataforma N° Acceso	Serie N° Acceso	Muestra N° Acceso	Etiqueta de la muestra	Referencia
GPL2025	GSE3053:	GSM67052	FL478 control replicate 1	**[24] [25]**
		GSM67053	FL478 control replicate 2	
		GSM67054	FL478 control replicate 3	
		GSM67055	FL478 salt stressed replicate 1	
		GSM67056	FL478 salt stressed replicate 2	
		GSM67057	FL478 salt stressed replicate 3	
		GSM67058	IR29 control replicate 1	
		GSM67059	IR29 control replicate 2	
		GSM67060	IR29 salt stressed replicate 1	
		GSM67061	IR29 salt stressed replicate 2	
		GSM67062	IR29 salt stressed replicate 3	
GPL2025	GSE4438	GSM99858	TCHW P1_m103 control, biological rep1	**[26] [27]**
		GSM99859	TCHW P2_m103 control, biological rep2	
		GSM99860	TCHW P3_m103 control, biological rep3	
		GSM99861	TCHW P4_m103 salt stress, biological rep1	
		GSM99862	TCHW P5_m103 salt stress, biological rep2	

		GSM99863	TCHW P6_m103 salt stress, biological rep3	
		GSM99864	TCHW P7_ir29 control, biological rep1	
		GSM99865	TCHW P8_ir29 control, biological rep2	
		GSM99866	TCHW P9_ir29 control, biological rep3	
		GSM99867	TCHW P10_ir29 salt stress, biological rep1	
		GSM99868	TCHW P11_ir29 salt stress, biological rep2	
		GSM99869	TCHW P12_ir29 salt stress, biological rep3	
		GSM99870	TCHW P13_agami control, biological rep1	
		GSM99871	TCHW P14_agami control, biological rep2	
		GSM99872	TCHW P15_agami control, biological rep3	
		GSM99873	TCHW P16_agami salt stress, biological rep1	
		GSM99874	TCHW P17_agami salt stress, biological rep2	
		GSM99875	TCHW P18_agami salt stress, biological rep3	
		GSM99877	TCHW P19_ir29 control, biological rep1	
		GSM99878	TCHW P20_ir29 control, biological rep2	
		GSM99879	TCHW P21_ir29 control, biological rep3	
		GSM99880	TCHW P22_ir29 salt stress, biological rep1	
		GSM99881	TCHW P23_ir29 salt stress, biological rep2	
		GSM99882	TCHW P24_ir29 salt stress, biological rep3	
GPL2025	GSE6901	GSM159259	7-day-old Seedling, biological rep 1	**[28]** **[29]**
		GSM159260	7-day-old Seedling, biological rep 2	
		GSM159261	7-day-old Seedling, biological rep 3	
		GSM159262	Drought stress, biological rep 1	
		GSM159263	Drought stress, biological rep 2	
		GSM159264	Drought stress, biological rep 3	
		GSM159265	Salt stress, biological rep 1	
		GSM159266	Salt stress, biological rep 2	
		GSM159267	Salt stress, biological rep 3	
		GSM159268	Cold stress, biological rep 1	
		GSM159269	Cold stress, biological rep 2	
		GSM159270	Cold stress, biological rep 3	

De todas las variables que conforman el archivo de la plataforma descargada, tanto de la NCBI como de Affymetrix, sólo se utilizan las variables:

ID: Identificador del Probe Set de Affymetrix

Gene Ontology Biological Process: 0, 1 ó varios identificadores de Proceso Biológico de la GO.

Gene Ontology Cellular Component: 0, 1 ó varios identificadores de Componente Celular de la GO.

Gene Ontology Molecular Function: 0, 1 ó varios identificadores de Función Molecular de la GO.

Esta plataforma de "GeneChip Rice Genome Array" de Affymetrix permite interrogar 57381 sondas o transcriptos de manera concertada.

Se modificó la organización de los datos descargados, para estructurarlos de manera de facilitar su carga en la tabla **affychip** de la base de datos, como se describe más adelante en el punto **III.2.3.** Una vez reformateados los datos se grabaron en formato CSV para facilitar la carga desde el MySQL a la tabla **affychip**.

Los datos de expresión, identificador del Prob Set (**id_Ref**) y valor de expresión (**Value**), se formatearon para acomodarlos al diseño de la tabla **Sample** de la base de datos (Ver **III.2.1.).** Una vez realizado este proceso se grabó en formato CSV para permitir ser cargados desde el MySQL a la tabla mencionada. Esta tabla contiene entonces el identificador Affymetrix de cada Gen y su valor de expresión. Para identificarlo unívocamente se agregan variables que identifican el experimento y la muestra a la que pertenece.

Debido a que cada experimento estaba compuesto por distintos tratamientos, se creó una tabla *exper*, que registra códigos que permiten rescatar cualquier subconjunto de datos. El diseño de esta tabla es descripto en el punto **III.2.2.** Esta tabla fue generada en MS-Excel y se grabó en formato CSV para ser cargada en el gestor de base de datos.

III.2. Preparación de los datos y armado de una base de datos relacional

Los datos crudos disponibles en la base de datos GEO para cada uno de estos experimentos fueron organizados en diferentes tablas para poder operar con ellos dentro de una base de datos relacional que permitiera ejecutar consultas variadas. En total se diseñaron tres tablas que se describen a continuación:

1) **Tabla *sample.*** Incluye el total de los datos 2.696.907 informados en los tres experimentos y consta de 4 campos:

id_Affy : Identificador del gen según nomenclatura de Affymetrix

id_Exp : Identificador de la *serie* a la que pertenece (ej: GSE3053)

id_Sample: Identificador del *sample* pertenece (ej: GSM67052). Este campo junto con id_Exp identifica unívocamente el valor de expresión medido para un gen determinado en una condición particular.

Expresion: El valor de expresión determinado para cada gen presente en la micromatriz.

2) Tabla exper. Incluye el total de micromatrices (47) analizadas en los tres experimentos y consta de 9 campos:

id_Exp : Identificador de la *serie* a la que pertenece

id_Sample: Identificador de la muestra (*sample)*.

exp: El valor de este campo es 1 si la micromatriz pertenece al grupo de los tratamientos experimentales y 0 en caso contrario.

ctrl: El valor de este campo es 1 si la micromatriz pertenece al grupo de los tratamientos control y 0 en caso contrario.

sal: El valor de este campo es 1 si la micromatriz pertenece al grupo de las micromatrices sometidos a estrés salínico y 0 en caso contrario.

seq: El valor de este campo es 1 si la micromatriz pertenece al grupo de las micromatrices sometidos a estrés por sequía y 0 en caso contrario.

frio: El valor de este campo es 1 si la micromatriz pertenece al grupo de las micromatrices sometidos a estrés por frío y 0 en caso contrario.

estad: El valor de este campo es 1 si la micromatriz pertenece al grupo de las micromatrices que midieron la expresión en plantas y 0 en caso de que lo hayan hecho en plántulas.

sample: El valor de este campo es numérico y se repite cuando son micromatrices pertenecientes al mismo grupo de repetición o de control.

3) Tabla affychip.

Incluye las anotaciones GO asignadas por Affymetrix para cada gen representado en la matriz. Un gen puede tener anotación en una, dos o en las tres categorías que abarca la ontología génica (GO), proceso biológico (PB), función molecular (FM) y componente celular (CC). A su vez cada gen puede estar asociado a uno o más términos GO dentro de la misma categoria. El número total de registros en esta tabla es de 20.346 y corresponde a la anotación de 4476 genes. La misma consta de 2 campos:

id_Affy : Identificador del gen según la nomenclatura asignada por Affymetrix

acc: Identificador del término GO para PB, FM y/o CC.

Esta tabla permite relacionar el identificador de Affymetrix con el identificador de términos GO y enlazar los datos de expresión de los experimentos con la información de la ontología respectiva, ya sea PB, CC o FM.

Las tablas descriptas fueron integradas a la estructura de una base de datos MySQL disponible en el sitio Gene Ontology Consortium **[30]** que almacena las anotaciones de genes y productos de genes según la ontología GO (Figura 15). La misma se actualiza regularmente y puede ser consultada de manera remota o descargada para ser ejecutada de manera local.

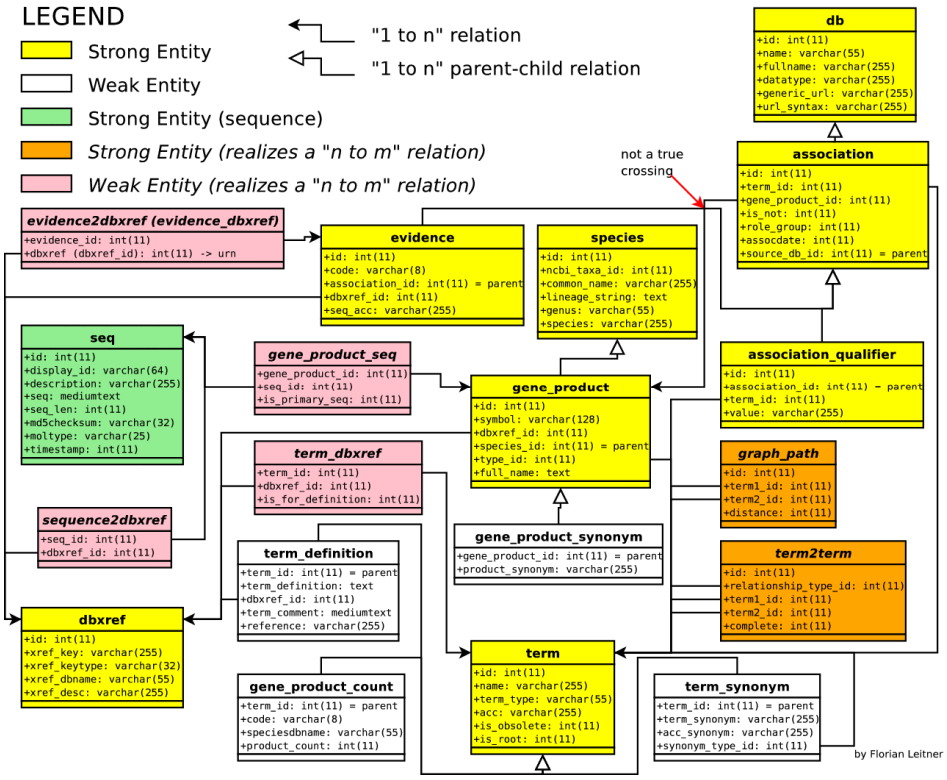

Figura 15: Estructura de la base de datos de términos GO, visto en este diagrama de entidad–relación. (http://www.geneontology.org/images/go-database-ER-diagram.png)

La base de datos de las ontologías fue descargada del sitio de la Gene Ontology: http://www.geneontology.org/GO.downloads.database.shtml.

Se instaló el MySQL (**url**: http://dev.mysql.com/downloads/) en el servidor que dispone la Maestría, cuya configuración es la siguiente: Server HP ML350, 4 GB Ram, 250 GB de almacenamiento y un micro de 1,8 GHz (un núcleo). El sistema operativo instalado es Windows Server 2003. Esto le permite soportar grandes cargas de procesamiento.

Los comandos para la creación y carga en MySQL de la base de datos GO son:

1. Desde dentro del programa MySQL se crea la base de datos

create database GO;

2 .Desde la línea de comando del sistema operativo se ejecuta el comando de carga.

>mysql --user= ..... -- pasword= ..... GO < go_200806

Donde se utiliza el nombre del usuario y la clave para cargar el achivo descargado de la Gene Ontology (**go_200806**) a la base de datos creada en el paso anterior (**GO**).

3. Para preservar todos los datos, se generaron copias de seguridad con el comando ejecutado desde la línea de comando del sistema operativo

> **mysqldump --opt --user= ..... -- pasword= ..... GO > GO_010608**

El nombre del backup sirve para registrar la fecha en que se realizó. En el ejemplo, 1 de junio de 2008. El comando para reinstalar esta copia de seguridad es el mismo con el que se cargó la base de datos GO en el paso anterior

4. Se crean a continuación las tablas Affymetrix, Samples y Exper dentro de la base de datos GO y se cargan los datos que fueron reformateados anteriormente. El comando de MySQL **create table** se utiliza para crear las tablas y **load data local infile** se usa para cargar las mismas.

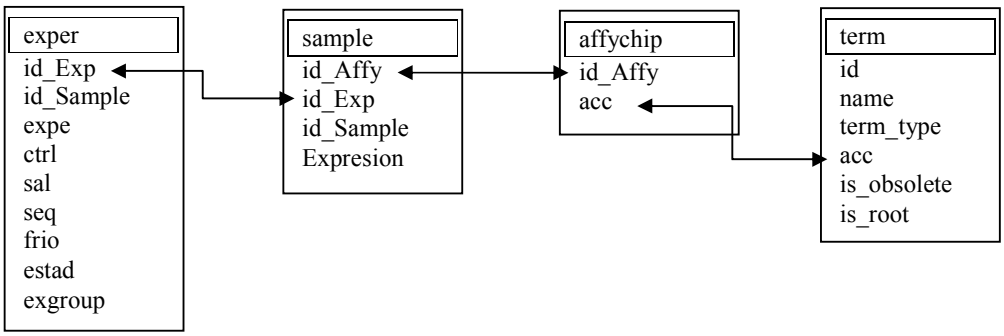

Figura 16: Estructura de la base de datos de términos GO, (diagrama de entidad–relación) con las cuatro tablas utilizadas para las consultas.

Como resultado de estas operaciones se crean las tablas **Samples** (2.696.907 registros), **Affychip** (20.346 registros) y la tabla **Exper** (47 registros).

El enlace de las tablas correspondientes a los datos de las micromatrices y las de la ontología, se hizo a través de la tabla **term** la cual contiene una variable denominada *term_type* que determina si lo que vamos a extraer son datos de PB, FM o CC. La Figura 16 muestra la integración de las tablas a la base de datos lo cual facilita la realización de una variedad importante de consultas.

Para armar las consultas se usaron herramientas provistas por R. En primer lugar se descargó el software estadístico R (http://www.r-project.org/) y se instaló en el servidor de la maestría. Para poder conectar a R con MySQL se utilizó la librería RMySQL que permite la realizar consultas desde R al gestor de bases MySQL.

Para conectar R con MySQL se utilizan los siguientes comandos.

```
library(MySQL)
m <- MySQL()
con <- dbConnect(m, host="localhost", user="......", password = "....." , dbname = "go")
```

Luego de esta conexión, es posible automatizar consultas mediante la implementación de instrucciones sencillas, como la que se describe a continuación.

```
select sample.id_Affy, exper.id_Sample, sample.Expresion, exper.exgroup from
        ((exper join sample use index (samp_idx) using (id_Sample))
                join affychip use index (affyc_idx, affych_idx) using (id_Affy))
                        join term use index (term_idx, termt_idx) using (acc)
```

En el **Anexo A.1**. se exponen en forma completa las instrucciones de conexión a la base de datos, la construcción de la consulta y la ejecución de la misma. Allí se indica como determinar el subconjunto de datos a rescatar con las consulta, usando el comando **paste** que permite unir las variables que definen el criterio de búsqueda.

Para realizar una consulta sobre la nbase de datos MySQL desde R se debe ejecutar el comando SQL dentro de una sentencia **dbGetQuery** como se observa a continuación

```
matri <- dbGetQuery(con, arma_con)
```

Se diseñaron las consultas de MySQL teniendo en cuenta que se necesitaba utilizar cuatro tablas de la base de datos GO, Exper, Sample, Affychip y Term, utilizando la

opción **left join** para enlazar dos tablas y el resultado de estas a su vez con una tercera tabla y estas con una cuarta.

Se construye la consulta utilizando las distintas combinaciones provistas por la tabla Exper para variar la búsqueda y rescatar el subconjunto de registros que necesitamos para procesar. A esta consulta se pueden agregar distintos criterios de selección mediante la cláusula "where", por ejemplo con los siguientes campos:

- term.term_type selecciona si es Proceso Biológico, Componente Celular o Función Molecular.
- exper.id_Exp selecciona con cuál de los tres experimentos se va a trabajar, GSE3053, GSE4438 o GSE6901.
- exper.ctrl selecciona si se trabaja con el control o no.
- exper.sal selecciona si trabaja con micromatrices tratadas o no con estrés salino
- exper.seq selecciona si trabaja con micromatrices tratadas o no con estrés por sequía.
- exper.frío selecciona si se trabaja con micromatrices tratadas o no con estrés por frío.
- exper.estad selecciona si se trabaja con micromatrices analizadas con material biológico en el estadío de plántulas o plantas.

A fin de optimizar el tiempo de respuesta de las consultas, se agregaron índices a las tablas con el comando **alter table**.

Se indexó la tabla **Sample** por la variable **id_Affy** generando el índice **samp_idx.**

Se indexó la tabla **Affychyp** por la variable **id_Affy** generando el índice **affyc_idx** y por la variable **acc** generando el índice **affych_idx.**

Por último se indexó la tabla **term** por la variable **term_type** generando el índice **termt_idx** y por la variable **acc** generando el índice **term_idx.**

Se utilizaron estos índices dentro de las consultas y los tiempos de respuestas se redujeron sustancialmente.

III.3. Preprocesamiento de Datos

III.3.1. Promediar las muestras de repeticiones

Los datos correpondientes a réplicas de tratamientos y controles dentro de cada experimento se promediaron usando la función **aggregate.data.frame** de R. Esta función

se aplica a datos almacenados en "dataframes" de R, y permite pasar como una lista de parámetros las variables que se usan para agrupar. En el ejemplo que sigue la variable **Expresión** es promediada agrupándola por las variables **id_Affy** y **exgroup.**

```
matri1 <- aggregate.data.frame(matri$Expresion, list(matri$id_Affy, matri$exgroup), FUN=mean)
```

III.3.2. Reformatear los datos en matriz.

Una vez que se obtienen medias agregadas de los datos, éstas se deben reorganizar de manera que en las filas se encuentren lo genes (referidos con el identificador de Affymetrix) y en cada columna queden los valores de los tratamientos y controles. Esta reorganización se realiza con el comando **reshape** de R.

```
matrire <- reshape(matri1, timevar="Group.2", idvar=c("Group.1"), direction="wide")
```

El siguiente paso del análisis es analizar la dispersión de los datos dentro de un experimento. Esto se debería realizar graficando todos las combinaciones de muestras disponibles tomándolas de a dos. En algunos casos esto representa una gran cantidad de gráficos y por ello diferentes autores recomiendan calcular un vector cuyos elementos son las medianas entre tratamientos para cada gen. Luego se realizan los gráficos de dispersión de cada uno de las muestras contra el vector de medianas.

Se calcula la muestra mediana iterando en R de la siguiente manera.

```
for (i in 1:dim_matrire_rec[1])
        {
                med_matrire_rec[i,1] <- median(as.numeric(matrire_rec[i,]))
        }
```

III.3.3. Normalización de Datos

Los datos crudos se normalizaron para poder comparar los valores dentro y entre experimentos **[31]**.

La normalización se realiza a tres niveles

- Normalización dentro de la Muestra

  Se observa que la distribución de los datos es altamente asimétrica. La mayor frecuencia de ellos se presentan en valores bajos de expresión y esta frecuencia va disminuyendo en la medida que aumentan los valores de expresión. Para lograr una mayor simetría en la distribución, la transformación recomendada en la

bibliografía del proveedor de la micromatriz indica aplicar el logaritmo en base 2 para distribuir la frecuencia más uniformemente **[32]**. Esta recomendación también es indicada por Stekel **[31]**. La instrucción en R es:

```
matrire2 <- log(matrire_rec,2)
```

Luego, los valores extremos, aquellos que se ubican por encima del percentil 98 y por debajo del percentil 2 se reemplazan por el valor de los percentiles respectivos. La iteración en R es la siguiente.

```
for (i in 1:dim_mat2[2])
{
        matri_cuar <- quantile(matrire2[,i], c(0.02 , 0.98))
        matri_min <- ifelse(matrire2[,i] < matri_cuar[1], matri_cuar[1], matrire2[,i])
        matri_max <- ifelse(matri_min > matri_cuar[2], matri_cuar[2], matri_min)
        if (i == 1) {matri_fin <- matri_max} else {matri_fin <- cbind(matri_fin, matri_max)
}
```

- Normalización entre muestras y entre experimentos

  La normalización entre muestras y entre experimentos se debe realizar porque los promedios y varianza son distintos en las distintas muestras. En el caso de los experimentos, existen otros factores que aumentan la varianza, generados por el equipamiento con el que se trabaja, el operador o los procedimientos que pueden influir en la medición. Estas variaciones pueden controlarse o aislarse en el laboratorio realizando pruebas preliminares. Cuando se comparan experimentos realizados en diferentes laboratorios a partir de los datos depositados es imposible estimar el efecto individual de cada uno de esos factores. En nuestro caso normalizamos los datos provenientes de distintos tratamientos según lo recomendado por Stekel **[31**, capítulo 5**]**.

  En R esta operación se realiza con el comando ***scale***.

```
matrire2_es <- scale(matri_fin)
```

III.4. Análisis de Datos con R

Los datos ya normalizados en sus tres niveles, es decir, dentro de cada muestra, entre muestra y entre experimentos, se analizaron utilizando diferentes técnicas de agrupamiento (cluster). Estos procedimientos permiten reconocer qué genes están coexpresándose; es decir permite encontrar grupos de genes que comparten patrones de comportamiento común tanto de expresión, o de no expresión. Los métodos utilizados se listan a continuación.

- Jerárquico,
- K-medias,
- Partición alrededor de medioides (PAM – Partitioning Around Medoids),
- Agrupamiento en grandes aplicaciones (CLARA – Clustering LARge Applications) y
- Mapas Auto Organizados (SOM – Self-Organizing Map)

III.4.1. Agrupamiento Jerárquico

De acuerdo a lo expuesto en el **punto 1.8.1.1**, para agrupar los genes, primero se debe calcular la matriz de distancias. A modo exploratorio en el ejemplo que se muestra a continuación, se utiliza la instrucción ***dist***, con la distancia ***manhattan.***

```
mat10 <- dist(matrire2_es, method="manhattan")
```

Luego de obtener la matriz de distancias de los niveles de expresión entre genes, estas se procesan estas para generar los distintos agrupamientos, calculando la distancia entre los agrupamientos y dentro los mismos de acuerdo a lo expresado en **1.8.1.1**.

Las siguientes instrucciónes de R son usadas para:

- generar el dendograma de agrupamientos: ***hclust*** ; en este ejemplo la distancia entre grupos es ***ligamiento completo***

  ```
  mat11 <- hclust(mat10, method = "complete")
  ```

- visualizar el dendograma: ***plot***

  ```
  plot(mat11)
  ```

- visualizar la secuencia de aglomeración de los genes: ***$merge***

  ```
  mat11$merge
  ```

- visualizar los grupos en el dendograma: ***print(rect,hclust()).*** En este ejemplo, el número de grupos es 10.

  ```
  print(rect.hclust(mat11,k=10))
  ```

- identificar a qué grupo pertenece cada gen: ***cutree.*** En este ejemplo, el número de grupos es 10.

  ```
  cutree(mat11,k=10)
  ```

- adjuntar al *data frame* de datos, el vector que identifica a que cluster pertenece cada gen.

  matrire2a=data.frame(matrire_rec_affy, matrire2_es,cutree(mat11,k=10))

III.4.2. Agrupamiento por K-medias

La instrucción para agrupar los datos por K-medias, expuesta en el **punto 1.8.1.2.**, es ***kmeans***, en este caso y a modo de ejemplo el número solicitado de grupos es 8.

km <- kmeans(matrire2_es, 8)

Para adjuntar al *data frame* de datos el vector que identifica a qué grupo pertenece cada gen, se utiliza ***$cluster***.

matrire2=data.frame(matrire, km$cluster)

III.4.3. Partición alrededor de medioides (PAM–Partitioning Around Medoids)

El software R no cuenta dentro de sus procedimientos "base" al PAM. Por ello es necesario descargar de R la librería ***cluster***. Luego se la llama mediante el comando ***library*** **[34] .**

library(cluster)

La instrucción para procesar los datos con PAM, expuesto en el **punto 1.8.1.3**., es ***pam***. En este ejemplo se busca agrupar los genes en 8 conglomerados.

pamk <- pam(matrire2_es, 8)

Para adjuntar al *data frame* de datos el vector qué identifica a que grupo pertenece cada gen, se utiliza ***$clustering***.

matrire2=data.frame(matrire, pamk$clustering)

III.4.4. Agrupamiento para grandes aplicaciones (CLARA – Clustering LARge Applications)

Al igual que el procedimiento PAM, CLARA necesita descargar del sitio de R la biblioteca ***cluster***. Luego se la llama con el comando ***library*** **[35]**.

library(cluster)

La instrucción para particionar los datos por medio de CLARA, expuesto en el **punto 1.8.1.4**, es utilizar la instrucción ***clara,*** en este ejemplo buscando la formación de 8 grupos**.**

clarax <- clara(matrire2_es, 8)

Para adjuntar al *data frame* de datos el vector que identifica a que cluster pertenece cada gen, se utiliza el comando ***$clustering***.

```
matrire2=data.frame(matrire, clarax$clustering)
```

III.4.5. Mapas Autoorganizados (SOM – Self-Organizing Map)

Se utilizó la librería *som* de R para construir mapas autoorganizados. La cual, una vez instalada, se invoca con el comando **[34]**:

```
library(som)
```

La instrucción para crear mapas autoorganizados, expuesto en el **punto 1.8.1.5**., es ***som***. Algunos de los parámetros que observamos en este ejemplo son: ***xdim***: tamaño de la dimensión x; ***ydim***: tamaño de la dimensión y; ***topol***: topología de la red usada; ***neigh***: función de actualización de la vecindad.

```
y.som <- som(matrire2_es, xdim = xd, ydim = yd, topol = topo, neigh = neig)
```

La instrucción para mapear la red SOM es ***plot***:

```
plot(y.som)
```

Para adjuntar al *data frame* de datos el vector que identifica a que cluster pertenece cada gen, se utiliza la variable ***$visual*** de la salida de som.

```
matrire2=data.frame(matrire, y.som$visual)
```

III.5. Enriquecimiento de los clusters usando términos GO con R/Bioconductor

El resultado de aplicar los métodos de agrupamiento descritos en la sección anterior es la identificación de los grupos de genes que comparten patrones de expresión similares. El paso siguiente es determinar los términos de la ontología GO más representativos de cada uno de esos grupos.

Este proceso consta de tres pasos:

- Rescate de términos GO para las tres jerarquías: proceso biológico, componente celular y función molecular.
- Rescate de los ancestros de esos términos GO.
- Rescate de la profundidad dentro del árbol jerárquico de las notaciones GO.

III.5.1. Rescatar términos GO

Para cada uno de los genes que componen cada uno de los cluster generados, se busca en la ontología génica la o las notaciones correspondientes a cualquiera de las categorías (PB, CC y/o FM). Las instrucciones de armado de la consulta y la extracción de la información de la Base de Datos Gene Ontology ampliada con las tablas de experimentos se encuentran en el **Anexo A.4.1**.

III.5.2. Rescatar ancestros de términos GO

A partir de los términos GO asociados a los genes y rescatados en el paso anterior, se recuperan todos aquellos términos GO que son sus ancestros y que van desde el más específico (que es el que se obtuvo en el paso anterior) hasta el más genérico dentro de cada categoría, que será la raíz de cada una de las tres ontologías, PB, CC y/o FM.

Este proceso se realizó con el paquete GO.db de R, el cual permite rescatar los ancestros de cada término GO. En el **Anexo A.4.2.** se describe como se construye y ejecuta la búsqueda.

III.5.3. Rescatar profundidad jerárquica de los ancestros de términos GO

Una vez recuperados los ancestros, debemos identificar a qué profundidad de la escala jerárquica se encuentran cada uno de estos términos. Este paso es necesario para la caracterización final de cada uno de los cluster. En el **Anexo A.4.3** se describe como se construye la búsqueda y el rescate de los niveles de anotación.

IV. RESULTADOS Y DISCUSIÓN

Estrés hídrico y salino en arroz. Selección de experimentos

El estrés salino, provocado por altas concentraciones de sales en el suelo, y el estrés hídrico, que ocurre en situaciones de sequía, son dos condiciones de estrés que afectan negativamente el desarrollo de las plantas de arroz cultivadas. A veces, ambas situaciones ocurren simultáneamente, por ejemplo, en una situación de sequía progresiva se concentran las sales del suelo en el contenido cada vez menor de agua disponible. Ambos tipos de estrés inducen en las plantas respuestas similares que incluyen la acumulación de solutos y la modificación de componente de la pared celular entre otros para evitar la deshidratación, síntesis de proteínas protectoras e inducción de mecanismos de tolerancia a deshidratación ocasionada por cualquiera de los estreses mencionados, evitando o reparando el daño celular ocasionado (Verslues y col, 2006) **[42]**. La activación de estas respuestas requiere de una compleja red de señalizaciones y transducciones de señales, algunas de ellas comunes a varios estreses y otras específicas de un tipo de estrés en particular como aquellas asociadas con las secuencias DREB/DBF (Chinnusamy y col, 2003; Zhu, 2001) **[43] [44]**. Las plantas en el campo no se encuentran sometidas a un solo tipo de estrés sino a una suma de ellos (bióticos y abióticos) lo que causa que la respuesta molecular a ellos sea de tipo compleja (Agarwal y col, 2006; Fujita y col, 2006) **[45] [46]**. Existen diferentes estrategias experimentales para dilucidar los mecanismos de tolerancia osmótica, una de las más frecuentes en los últimos años es el empleo de micromatrices para estudiar la expresión génica.

Para realizar agrupamientos de genes que comparten perfiles de expresión y enriquecer los aglomerados con información ontológica se seleccionaron tres experimentos de expresión génica en arroz, en los que se trabajó con las micromatrices de Affymetrix para esta especie. Los experimentos seleccionados, el detalle de los tratamientos, los números de acceso en la base de datos GEO y los url correspondientes se describen en el punto III.1 y en la Tabla 1 de Materiales y Métodos.

Almacenamiento y pre-procesamiento de los datos

Los datos de expresión génica y las jerarquías ontológicas se almacenaron en una base de datos relacional construida con el gestor de base de datos de fuente abierta MySQL. Se agregaron identificadores uniformes para los tratamientos realizados en cada experimento, con el objeto de facilitar las comparaciones entre experimentos.

IV.1. Normalización de Datos - Visualización (Ver Anexo A.2 – Modulo 2)

Los datos de expresión genética dentro de cada muestra presentan una distribución altamente asimétrica. Por otra parte, las distribuciones que provienen de distintos experimentos y muestras presentan variaciones entre ellas. Estas variaciones deben ser corregidas antes de aplicar los métodos de agrupamiento. El apartado **III.3.3.** de Materiales y Métodos detalla los procedimientos para realizar las normalizaciones a los tres niveles, dentro de la muestra, entre muestras y entre experimentos.

Para ilustrar el procedimiento se presentará a modo de ejemplo en el punto IV.1.1., el procesamiento de los datos del Experimento GSE3053 analizando para este caso aquellos referidos a los Procesos Biológicos.

Las Figuras 17, 18 y 19 muestran los datos originales y las Figuras 20, 21 y 22 presentan los datos procesados. Para la normalización se promediaron las repeticiones de las muestras (**ver III.3.1**), los datos se reorganizaron en un formato matricial (**ver III.3.2**), y luego se calcularon las muestras medianas (**ver III.3.2**) que permite generar los gráficos de dispersión.

IV.1.1. Datos sin normalizar dentro de la muestra - Gráficos

En la Figura 17 se grafican los datos de una muestra de ejemplo (eje vertical) versus la muestra mediana calculada (eje horizontal). Se observa una alta asimetría, presentándose una mayor frecuencia en los valores bajos de expresión, disminuyendo esta frecuencia en la medida de que los valores de expresión aumentan.

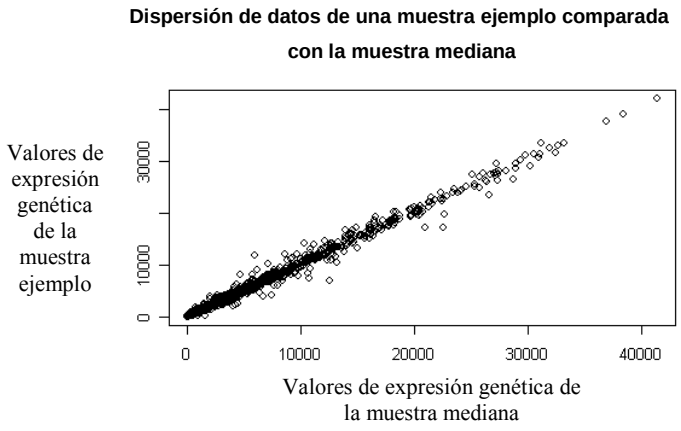

Figura 17: Ejemplo de dispersión de frecuencias según nivel de expresión genética entre una muestra ejemplo versus la muestras mediana.

En el gráfico 18, se presentan los mismos datos de la muestra ejemplo en formato de histograma. También se puede ver la asimetría de las frecuencias entre en los valores bajos de expresión y los valores altos de expresión.

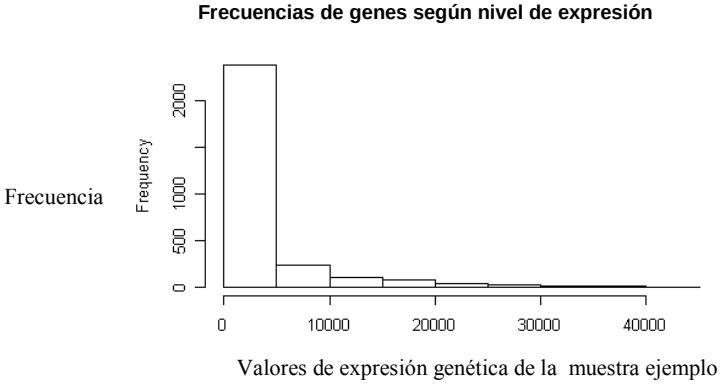

Figura 18: Ejemplo de dispersión de frecuencias dentro de la muestra

Si bien se grafica una muestra ejemplo, esta asimetría se presentan en todas las muestras analizadas en esta tesis. Si no se corrigiera este problema se presentaría una fuerte distorsión cuando se procesan los agrupamientos. En la Figura 19 se muestra un análisis de agrupamiento jerárquico, con datos de la muestra sin normalizar.

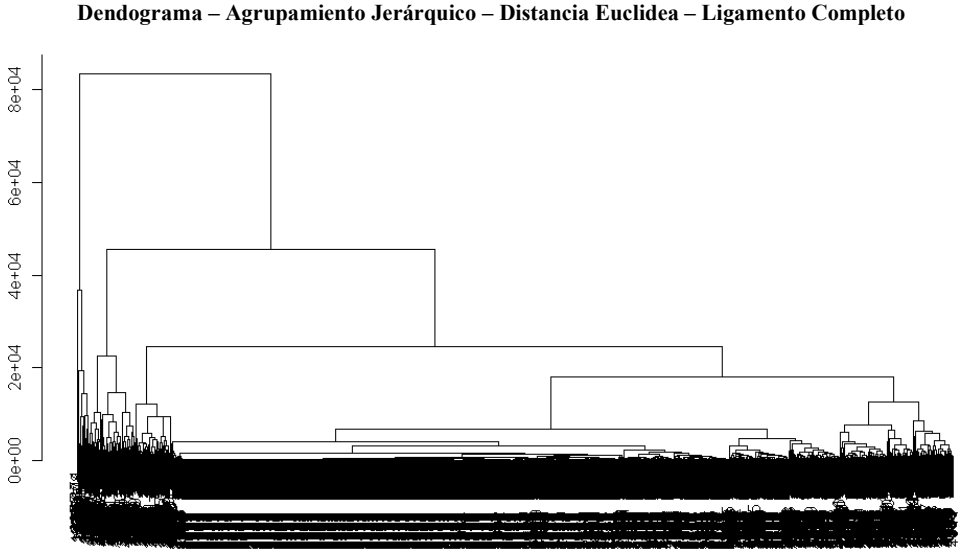

Figura 19: Distorsión hacia la izquierda del dendograma jerárquico, debido a los datos sin normalizar.

IV.1.2. Datos normalizados dentro de la muestra - Gráficos

Para lograr una distribución de valores más homogénea y con ello un mejor agrupamiento de los genes, los datos se transforman calculando el logaritmo en base 2 a cada uno de los valores (**ver III.3.3**). Por otra parte, los valores mayores al percentil 98 y menores al percentil 2 de esta distribución, son considerados valores extremos **[31]** y se asume que están seriamente afectados por errores de procedimiento en la generación de la micromatriz y/o en la lectura de los datos. Por ello se procedió a reemplazar los valores de expresión genética inferiores al percentil 2, con el valor del percentil 2; de la misma forma se reescribieron los valores superiores al percentil 98, con el valor del percentil 98.
Se puede observar el efecto de estas correcciones en las Figuras 20, 21 y 22 que corresponden a los mismos tratamientos mostrados en las figuras 17, 18 y 19, pero con datos logarítmicos y ajustes de los valores extremos. Se observa que las distribuciones de frecuencias obtenidas fueron más homogéneas a cualquier nivel de expresión genética de la muestra mediana (eje horizontal), y de las muestra ejemplo (eje vertical).

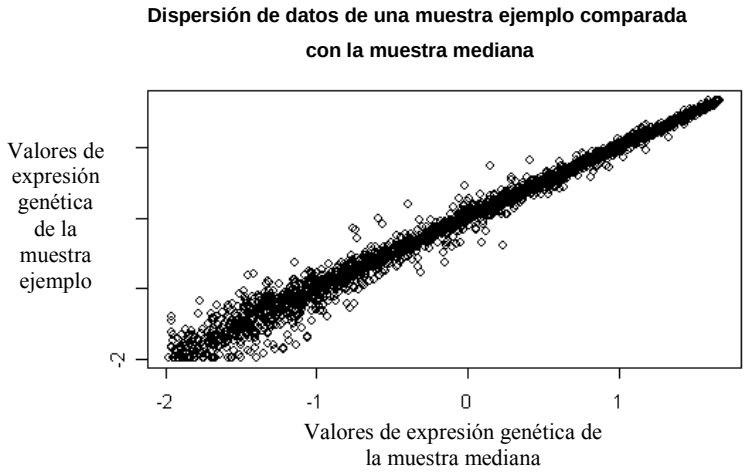

Figura 20: Ejemplo de dispersión de frecuencias según nivel de expresión genética entre una muestra ejemplo versus la muestras mediana.

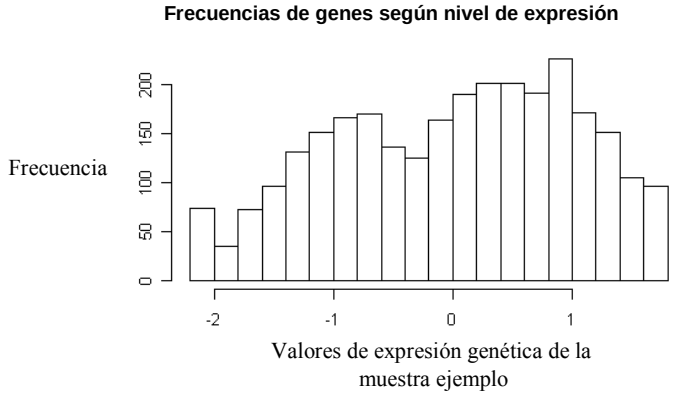

Figura 21: Ejemplo de dispersión de frecuencias dentro de la muestra

La Figura 22 muestra el mismo conjunto de datos procesados en la Figura 19, pero con las correcciones realizadas dentro de la muestra, observándose un dendograma más homogéneo.

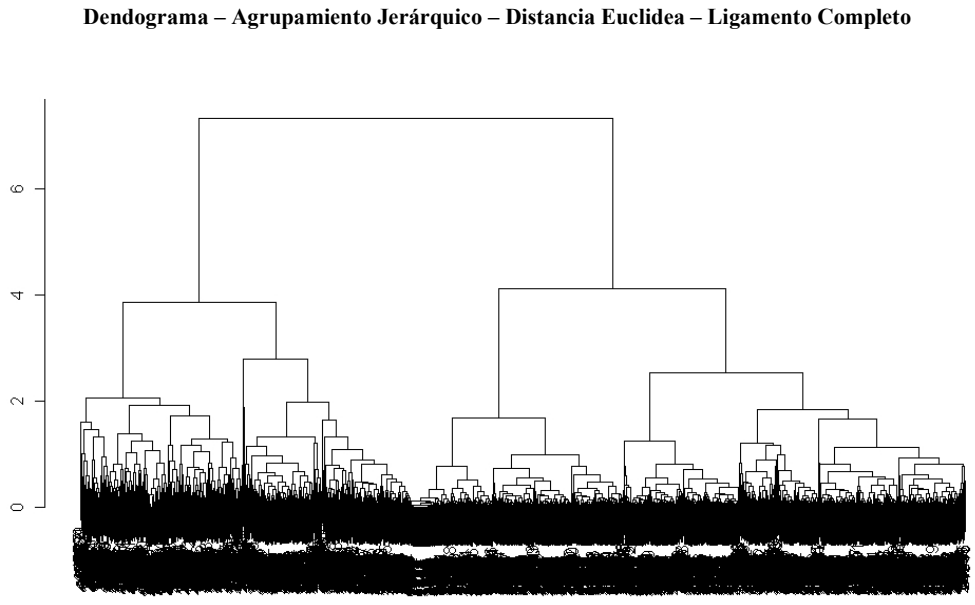

Figura 22: Dendograma jerárquico sin distorsión, debido a la normalización de los datos.

IV.1.3. Datos sin normalizar entre muestras y entre experimentos - Gráficos

Los datos provenientes de diferentes muestras y diferentes experimentos presentan promedios y varianzas distintas (**ver III.3.3. – segundo punto**). En la Figura 23 se presenta un gráfico, donde se aprecian pequeñas diferencias entre las cuatro muestras del experimento GSE3053. El origen probable de estas variaciones son pequeñas diferencias en los procedimientos de hibridación y lectura de los resultados, que ocurren a pesar de trabajar con una plataforma estandarizada de trabajo, como es el sistema Affymetrix.

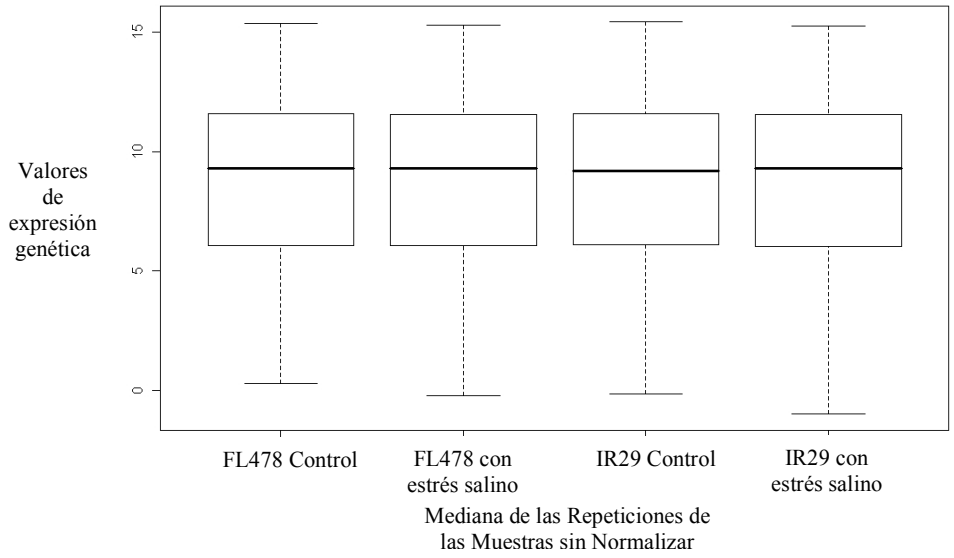

Figura 23: Gráfico de las cuatro muestras del Experimento 3053 – Proceso Biológico no escalados.

IV.1.4. Datos normalizados entre muestras y entre experimentos - Gráficos

Para lograr unificar estos datos provenientes de distintas muestras y distintos experimentos, se realiza un escalamiento de los datos restando a cada uno el promedio de su muestra y dividiendo por el desvío estándar de la misma muestra. Esto permite realizar comparaciones entre las distintas muestras y los distintos experimentos (**ver III.3.3. – segundo punto**). En la Figura 24 se observan los mismos datos de las muestras de la Figura 23 ya escalados.

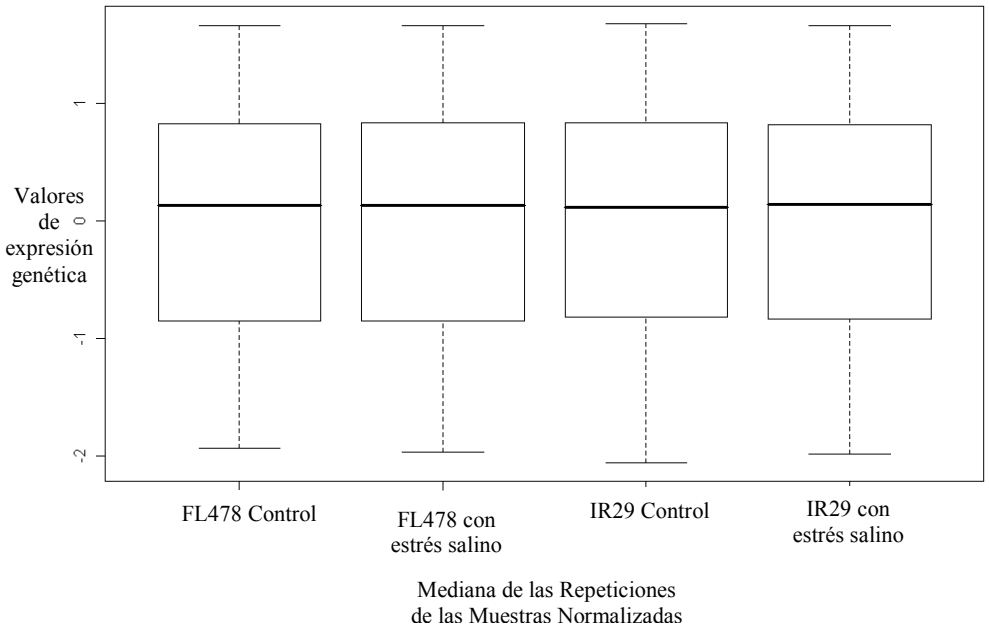

Figura 23: Gráfico de las cuatro muestras del Experimento 3053 – Proceso Biológico escalados.

IV.2. Agrupamiento – Visualización (Ver Anexo A.3 – Modulo 3)

Para la construcción de los agrupamientos se utilizaron los métodos descriptos en la sección **III.4.** con algunos parámetros por defecto. Se determinó como criterio de selección del mejor método de conglomerados aquel que permitiera una distribución homogénea de genes para cada profundidad de los términos GO en los gráficos DAE.

Es decir, el criterio adoptado para seleccionar el método y los parámetros respectivos se determinarán en el paso siguiente, cuando se enriquezcan los conglomerados generados, con los términos GO.

Para ajustar el proceso se continuó trabajando con el experimento GSE3053, y con aquellos genes que tuvieran al menos una anotación GO para procesos biológicos. Estos datos fueron preprocesados como se explica en los puntos anteriores y se los agrupó según los métodos y configuraciones que se ven en la sección **III.4.** Los que se utilizaron se resumen en la Tabla 2.

Tabla 2: Métodos de agrupamientos y configuraciones con los que se procesaron las muestras de ejemplo.

Métodos de cluster	Distancias	
Jerárquico	Manhattan	Ligamento completo
		Ward
	Euclidea	Ligamento completo
		Ward
K-medias	500 iteraciones	
	1000 iteraciones	
PAM	Euclidean	
	Manhattan	
CLARA	Euclidea	
	Manhattan	
SOM	Hexagonal	Gaussian
		Bubble
	Rectangular	Gaussian
		Bubble

Se realizaron agrupamientos iniciales en los que se fijó el número de conglomerados a analizar en nueve, para todos los métodos aplicados. Esta cantidad de grupos *a priori* es reducida, dado que nos sitúa probablemente en el paso posterior de enriquecimiento en donde todos los conglomerados quedan caracterizados por unos pocos términos GO muy generales. Sin embargo, este bajo número facilitó la exploración y el ajuste del método. El análisis se repitió luego, ampliándose el número de conglomerados a 16 y 25 grupos.

Como resultados de esta exploración inicial se obtuvieron 14 salidas que se encuentran recopiladas en el archivo **Gráficos de salida 9 grupos.xls** en el CD anexo que acompaña esta tesis.

En el Gráfico 24 se muestran a modo de ejemplo los dendogramas del agrupamiento jerárquico con 9 grupos, para las distintas combinaciones de distancia referidas en la Tabla 2.

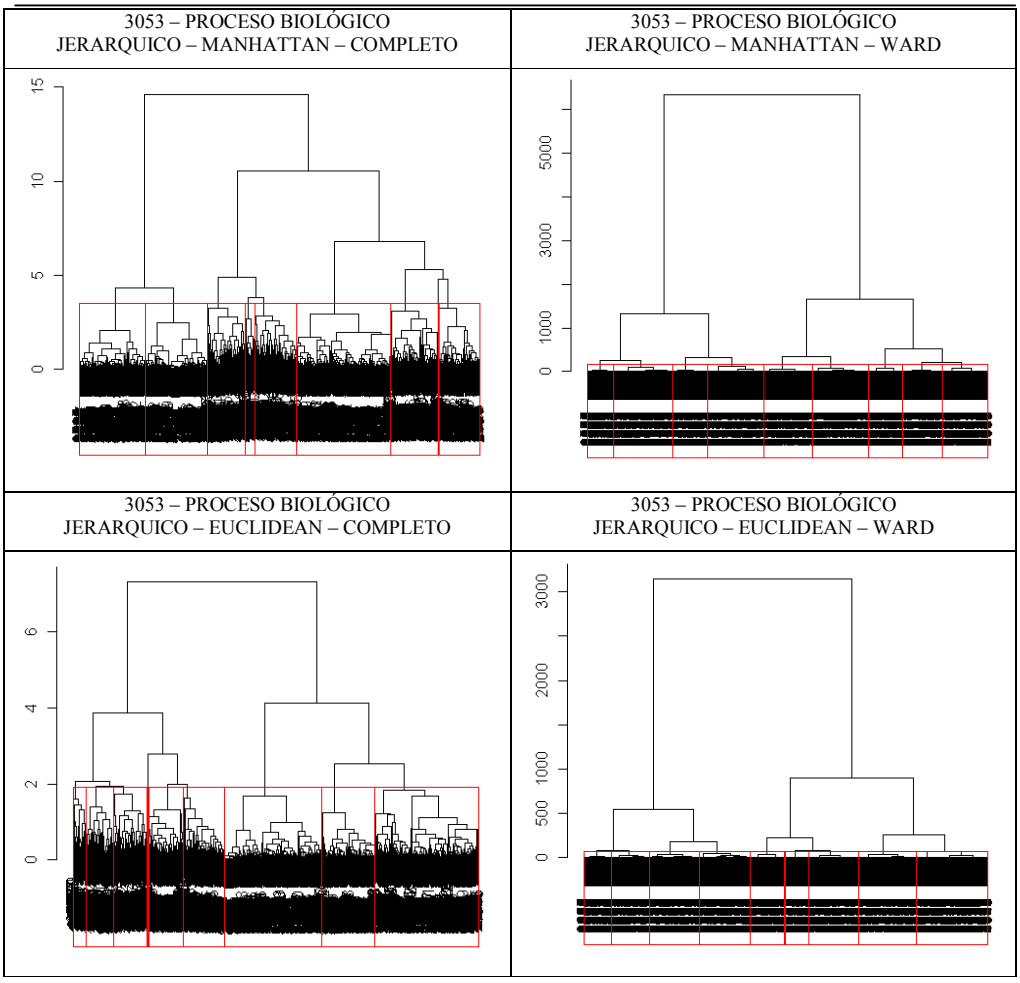

Figura 24: Gráfico de las cuatro dendogramas de cluster jerárquico para el Experimento 3053 – Proceso Biológico y 9 agrupamientos, con las configuraciones dadas en la Tabla 2.

Se puede observar que la distancia entre agrupamientos Ward, distribuye *a priori* en forma más pareja que la distancia entre agrupamientos Completo.

En la Figura 25 se muestran a manera de ejemplos, los gráficos de salida del agrupamiento por método SOM, con las configuraciones establecidas en la Tabla 2

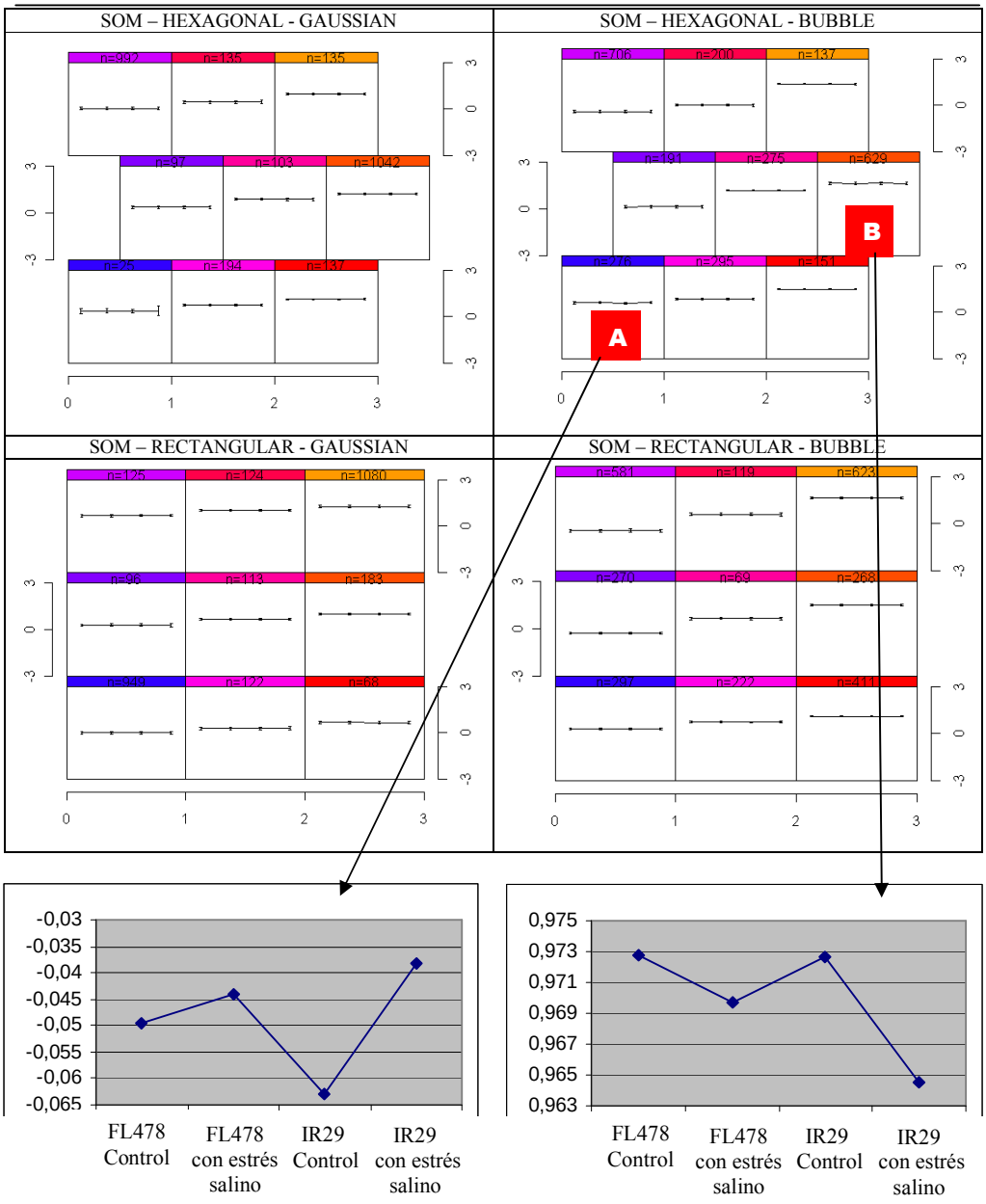

Figura 25: Gráfico de las cuatro salidas de cluster SOM para el experimento GSE3053 y genes con anotación GO para proceso biológico. Se calcularon nueve conglomerados, con las configuraciones dadas en la Tabla 2. Se amplían dos sectores de agrupamientos indicados como A y B para una mejor visualización de distintos promedios entre muestras proveniente de tratamiento y control.

Se observa que, excepto para la configuración Rectangular-*Bubble*, con el método SOM se obtienen dos grupos con un elevado número de miembros en relación a los restantes grupos. Pero se indicó anteriormente, la decisión sobre cuál método es el más adecuado se realizará teniendo en cuenta la homogeneidad de frecuencias en cada una de las profundidades de niveles de términos GO.

Estos gráficos, al ser tan pequeños, no permiten apreciar diferencias en las líneas que representan el promedio de los valores de expresión en cada grupo, tanto para los 2 tratamientos y los 2 controles. Por ello en el mismo gráfico se ampliaron dos agrupamientos (A y B) para una mejor visualización. Sabiendo que en ambos gráficos 1 y 3 corresponden a las muestras control y, 2 y 4 a las muestras con tratamientos, se observa en el gráfico A una mayor sobreexpresión de los genes pertenecientes a los tratamientos frente a los de control y en B se visualiza el caso contrario.

IV.3. Enriquecimiento términos GO (Ver Anexo A.4 – Modulo 4)

A continuación se realizó el enriquecimiento de términos GO de cada uno de los conglomerados obtenidos en los 14 análisis de la sección anterior. (ver archivo **Gráficos de salida 9 grupos.xls** en el CD que acompaña esta tesis)

Cada una de estas salidas fue enriquecida a partir de la caracterización de los grupos en base a los términos GO más representados, lo cual se realizó en tres pasos:

1. **Búsqueda del término GO que corresponde a cada uno de los genes (IV.3.1)**
2. **Recuperar los ancestros dentro del gráfico DAG de GO (IV.3.2)**
3. **Determinación de la profundidad de cada término GO (IV.3.3)**

IV.3.1 Búsqueda del término GO que corresponde a cada uno de los genes.

Para asignar los términos GO correspondientes a cada uno de los accesos interrogados en la micromatriz, se realiza una consulta iterativa a la base de datos GO, utilizando como argumento de búsqueda la identificación de cada uno de los genes y rescatando para este ejemplo, el término GO para proceso biológico asignado a cada gen. Esta consulta utiliza dos tablas de la base de datos GO, Affychip y Term. En el **Anexo A.4.1** se observa el procedimiento con los comandos respectivos de consulta. Es pertinente aclarar que para cada gen se pueden encontrar uno o varios términos GO para procesos biológicos asociados

La salida de este proceso son el o los términos GO asignados a cada uno de los genes y el identificador de los conglomerados que los contienen.

IV.3.2. Recuperar los ancestros dentro del gráfico DAG de GO

Por cada uno de los términos GO encontrados en el paso anterior se buscaron todos los ancestros, partiendo del término GO originalmente asignado al gen en la notación de Affymetrix hasta el término GO raíz de la jerarquía inclusive, en este ejemplo, proceso biológico. Para realizar esto se utilizó la librería **GO.db** de **Bioconductor.** En el **Anexo A.4.2**, podemos ver el procedimiento con los comandos respectivos de la consulta.

Los resultados obtenidos de esta consulta se agregaron como registros, tantos como niveles tenga hasta la raíz de la ontología establecida, incluyendo en cada registro la identificación del gen y el grupo al que pertenecía.

IV.3.3. Determinación de la profundidad de cada término GO

Una vez recuperados los términos GO ancestros, se determinó su profundidad dentro del gráfico DAG de la ontología, es decir, la profundidad dentro de la jerarquía ontológica. Este proceso se describe en el **Anexo A.4.3.** utiliza además la tabla **graph_path** que contiene todas las distancias a través de los caminos del gráfico DAG, entre dos cualesquiera de sus anotaciones GO.

El diseño de la base de datos GO permite conocer la profundidad, o dicho de otra forma la distancia, entre cualquier término GO y la raíz de la jerarquía. El proceso de búsqueda de profundidades se realizó iterativamente para cada término GO asociado a los genes y para sus ancestros. La salida es un vector que se adjuntó a la matriz resultante del paso anterior.

IV.4. Descubrimiento del conocimiento en los conglomerados enriquecidos con términos GO

Hasta este punto se encontraron los agrupamientos, los términos GO asociados a cada gen dentro de cada grupo, los ancestros de estos términos y las profundidades dentro de la jerarquía de anotación GO.

El siguiente punto fue determinar la frecuencia de los términos GO, tanto los asignados a genes como sus términos ancestros, a cada nivel de profundidad en la jerarquía y para cada agrupamiento. Estas distribuciones de frecuencia asisten al experto del dominio al determinar cuan genérica o cuan específica es la información GO con la que desea trabajar. Aquellos términos que están muy en la superficie (profundidad entre los niveles

1 a 4) son notaciones GO muy genéricas. A partir del nivel 4 en adelante, la s especificidades de estas notaciones representan un conocimiento más sustancial para el investigador. Más allá del nivel 6, las anotaciones no son tan útiles debido a la escasa frecuencia y la excesiva especificidad de la notación. Para conducir estudios comparativos, es importante definir un nivel de anotación lo más específico y frecuente posible. Por ello, se muestra como ejemplo la tabla 3, que ilustra la frecuencia de términos GO a las distintas profundidades para un determinado grupo, el porcentaje de cada uno de las profundidades y el porcentaje acumulado.

Tabla 3: Ejemplo de frecuencias de términos GO, porcentaje y porcentaje acumulado.

Prof.	0	1	2	3	4	5	6	7	8	9	10
N	546	1119	1496	953	733	485	305	56	12	1	1
%	10	20	25	17	13	8	5	2	0	0	0
% Acumulado	10	30	55	72	85	93	98	100	100	100	100

La Figura 26 muestra una representación grafica de la Tabla 3.

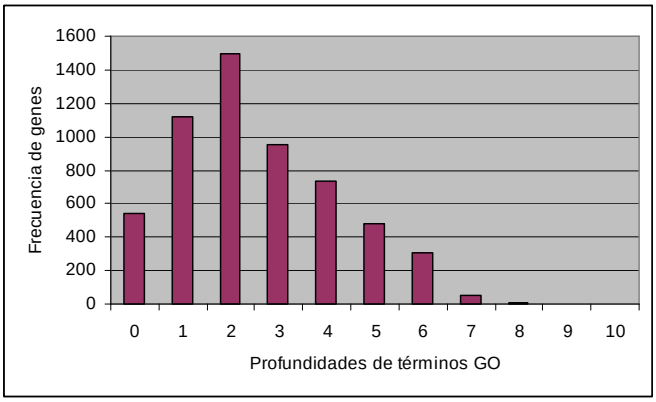

Figura 26: Ejemplo de distribución de la frecuencia de genes dentro de un grupo en las distintas profundidades del árbol jerárquico de términos GO

A partir de la información de las frecuencias de términos GO, la profundidad de las anotaciones y el conocimiento previo sobre el sistema biológico que se está estudiando se puede establecer el nivel de profundidad óptimo para caracterizar al grupo asociado, buscando un equilibrio en los niveles de las anotaciones, es decir, que estas no sean muy genéricas ni muy específicas.

Se construyeron tablas a partir de las salidas como la que se obtuvieron en el punto **IV.3.3,** en las que se calculó para cada conglomerado la frecuencia absoluta de términos GO a cada nivel de profundidad. La Tabla 4 muestra las frecuencias de términos GO que se rescataron para cada nivel de profundidad de los genes agrupados con el método de K-medias con 500 iteraciones para 9 conglomerados. Al enriquecer con los términos GO se obtuvieron profundidades hasta el nivel 10.

Tabla 4: Frecuencias de términos GO, por grupo y por profundidad para el Experimento GSE3053, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones.

		Profundidad											
		0	1	2	3	4	5	6	7	8	9	10	
Grupo	1	527	760	1073	916	577	494	226	97	29	1	0	
	2	808	1096	1430	930	704	524	297	53	11	1	1	
	3	1127	1556	2081	1569	1247	851	497	134	39	2	1	
	4	938	1312	1680	1165	883	626	371	107	28	3	2	
	5	766	1159	1543	1302	844	672	378	145	38	1	0	
	6	1020	1444	1936	1508	1100	832	509	147	55	9	2	
	7	962	1278	1715	1172	918	687	397	113	30	3	1	Sumatoria de los rangos
	8	1098	1549	2096	1595	1280	885	538	150	48	4	0	
	9	846	1150	1508	1054	835	588	357	113	39	8	6	
Rango		600	796	1023	679	703	391	312	97	44	8	6	4659

En la última fila se indica el rango de frecuencias en cada una de las profundidades. En la celda inferior derecha, la sumatoria de los rangos. Este valor de la sumatoria expresa cuan homogénea es la distribución de los distintos grupos a través de las distintas profundidades. Esto lo vemos en la Figura 27 donde se presentan gráficamente los datos de la Tabla 4.

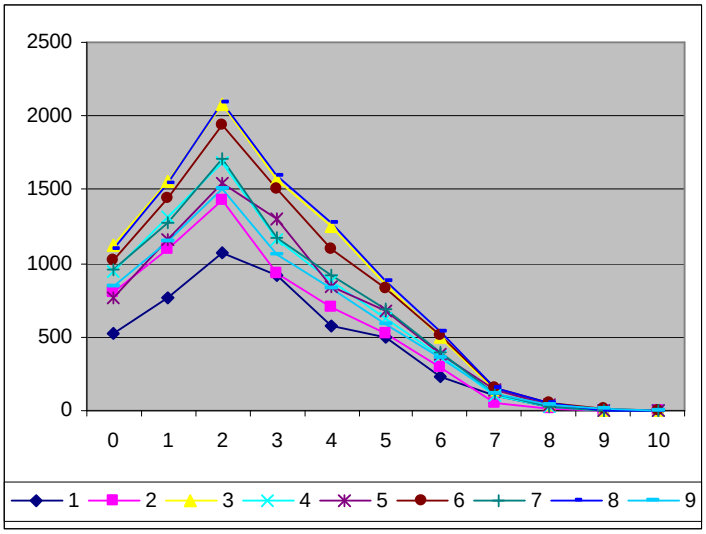

Figura 27: Distribución de frecuencias de cada uno de los agrupamientos a través de las distintas profundidades para el Experimento GSE3053, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones.

Para seleccionar qué método ajustaba mejor se calculó la sumatoria de los rangos de frecuencias de cada uno de los 14 métodos. La Tabla 5 muestra los resultados obtenidos

Tabla 5: Resultados obtenidos de la sumatoria de rangos para cada uno de los métodos.

Método	Sumatoria de los rangos
Kmedias 500 iteraciones	4659
PAM – Manhattan	4863
PAM – Euclidean	4941
Jerárquico – Manhattan – Ward	4972
Kmedias 1000 iteraciones	5823
CLARA – Euclidean	7134
Jerárquico – Euclidean – Ward	7499
CLARA – Manhatan	8248
SOM –Hexagono-Bubble	12290
SOM –Rectangular-Bubble	13180
Jerárquico – Manhattan – Ligamento Completo	15741
Jerárquico – Euclidean – Ligamento Completo	16831
SOM –Rectangular-Gaussian	24398
SOM –Hexagono-Gaussian	24873

En el archivo **3053CPBTOTAL.xls** que se encuentra en el CD anexo que acompaña esta tesis, podemos encontrar las distintas frecuencias de anotaciones por método de agrupamiento, por grupo dentro de método y por profundidad dentro de cada grupo.

El menor valor de la sumatoria indica cuál es el método de agrupamiento que presentó la menor variabilidad entre los 14 métodos ensayados, Kmedias con 500 iteraciones. Se procedió a aplicar este método al mismo conjunto de datos GSE3053 para Procesos Biológicos, aumentando la cantidad de grupos a 25. Se buscó con esto tratar de evitar que los genes con frecuencias menores de anotación dentro de cada conglomerado, queden solapados por anotaciones que poseen mayores frecuencias.

En el archivo **15K05PBprof-clust.xls** (ver CD) se pueden ver las distintas anotaciones o términos GO y sus frecuencias absolutas, porcentuales y porcentuales acumuladas para las profundidades 4, 5 y 6, y dentro de cada una de estas, para cada agrupamiento. Allí debemos ver aquellas anotaciones GO que tengan una frecuencia porcentual mayor al 5%. En referencia a esto, el comentario del profesional con conocimiento de dominio es el siguiente:

"*En el nivel 4 y 5 los términos GO más frecuentes –regulación de la trascripción, regulación de la traducción y metabolismo del ARN-, son comunes a varios grupos. Para nivel 6 las anotaciones se hacen más variadas. No conviene analizar los términos GO con frecuencias relativas menores a 4 o 5%. Para frecuencias más bajas, los términos son más abundantes y es más difícil establecer un punto de corte".*

"Las anotaciones de nivel 4 más frecuentes son bastante generales y poco informativa, convendría trabajar con las anotaciones de nivel 5 y 6. El análisis de ambos niveles da una información biológica parecida, lo cuál está bien, porque se espera que los agrupamientos de términos GO sean robustos e insensibles al nivel. En todo caso esperamos que a mayor profundidad, mayor detalle, pero referido al mismo proceso, función o compartimiento."

Nuevamente, el número de grupos es arbitrario y se lo determina luego de iterar el proceso en busca una clasificación subjetivamente buena que satisfagan dos criterios: la primera es que, al aumentar el número de grupos, se descubran genes que por su poca frecuencia de anotación GO quedaban enmascarados por genes con mucha frecuencia de anotación, cuando se definían pocos grupos; la segunda razón es subjetiva y tiene que ver con el descubrimiento de genes no triviales, previamente desconocidos y potencialmente

útiles, por su "importancia biológica" y con poca frecuencia en un grupo. Esto es, genes que *a priori* no se esperaría que se estén coexpresando con aquellos de mayor frecuencia dentro del conglomerado. Al momento de realizarse esta tesis no se tiene conocimiento de que exista un índice que permita medir este concepto subjetivo de "importancia biológica".

En la Tabla 6 podemos ver los resultados obtenidos de las frecuencias de anotación para el conjunto de datos de GSE3053, para la categoría Proceso Biológico, cuando se realizaron agrupamientos utilizando Kmedias 500 iteraciones sobre 25 grupos.

Tabla 6: Frecuencias de términos GO, por grupo y por profundidad para el Experimento GSE3053, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones y 25 agrupamientos.

		Profundidad											
Cluster		0	1	2	3	4	5	6	7	8	9	10	
	1	479	710	958	819	517	434	241	100	29	1	0	
	2	279	378	499	334	255	186	103	17	2	0	0	
	3	205	257	367	242	206	141	81	22	6	0	0	
	4	156	224	288	223	174	116	74	24	9	1	0	
	5	281	390	511	350	281	187	111	32	14	4	3	
	6	407	577	750	580	455	313	167	37	12	0	0	
	7	83	117	151	117	76	54	33	8	1	0	0	
	8	513	674	924	631	506	353	220	66	22	3	1	
	9	652	952	1264	984	719	556	355	111	45	9	2	
	10	593	873	1169	950	659	503	285	95	27	1	0	
	11	336	478	647	510	416	281	171	51	12	0	0	
	12	382	496	660	416	319	253	138	28	7	0	0	
	13	183	251	331	210	159	117	71	17	3	0	0	
	14	565	778	1089	845	657	461	283	70	22	1	0	
	15	382	484	679	452	365	264	139	31	9	1	1	
	16	332	484	601	438	335	252	157	55	17	2	0	
	17	194	270	366	270	196	133	73	30	10	3	3	
	18	409	593	833	711	440	369	156	69	17	0	0	
	19	134	190	239	176	128	98	51	19	2	0	0	
	20	74	105	130	93	77	60	42	18	4	0	0	
	21	21	31	39	23	21	14	6	2	0	0	0	
	22	283	398	483	342	243	175	106	32	10	1	0	
	23	563	804	1059	823	649	445	262	72	23	2	0	Sumatoria de los rangos
	24	270	366	467	306	254	189	117	27	8	2	2	
	25	316	424	558	366	281	205	128	26	6	1	1	
Rango		631	921	1225	961	698	542	349	109	45	9	3	5493

En la Figura 28 se grafica la Tabla 6.

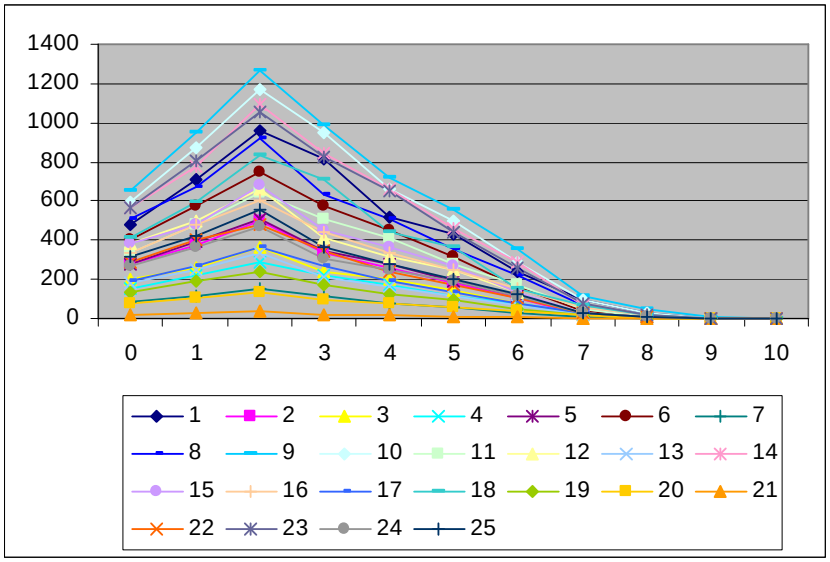

Figura 28: Distribución de frecuencias de cada uno de los agrupamientos para las distintas profundidades para el Experimento GSE3053, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones y 25 agrupamientos

Una vez realizado el agrupamiento y la caracterización de cada uno de los 25 conglomerados, obtenidos a partir de los datos del experimento GSE3053 para Proceso Biológicos, se procesaron los datos de los 3 experimentos en forma conjunta (GSE3053, GSE4438 y GSE6901) también para Proceso Biológico, buscando agrupar y caracterizar 25 conglomerados. De esta manera es posible la búsqueda y comparación de grupos buscando encontrar cualquiera de estas situaciones:

a) Conglomerados que habiéndose formado en el experimento individual, no se encuentren en el análisis conjunto. Esto sugeriría que la expresión de estos genes en los otros dos experimentos no está coordinada e influyen separando al grupo en el análisis conjunto.

b) En el caso opuesto, la aparición de agrupamientos en el análisis conjunto que no estén presente en alguno de los análisis individuales, supondría que existe un patrón de comportamiento en los otros dos experimentos que permite agrupar a dichos genes.

c) La tercera opción sería encontrar grupos semejantes en ambos análisis. Esto permitiría inferir que el conglomerado es insensible a las variaciones producidas en cada uno de los experimentos.

El resultado del análisis de los tres experimentos para Proceso Biológico, utilizando K-medias con 500 iteraciones en busca de 25 grupos, se pueden observar en la Tabla 7.

Tabla 7: Frecuencias de términos GO, por grupo y por profundidad para TODOS los Experimentos, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones y 25 agrupamientos.

		Profundidad											
		0	1	2	3	4	5	6	7	8	9	10	
Cluster	1	379	525	687	467	380	263	152	41	14	2	2	
	2	169	239	304	232	190	129	82	23	10	1	0	
	3	522	722	966	744	595	433	276	66	27	3	0	
	4	159	229	265	198	155	129	88	30	4	0	0	
	5	484	742	978	768	505	394	235	81	24	3	2	
	6	371	512	687	455	319	259	143	33	8	1	0	
	7	419	588	751	528	406	295	164	51	15	2	1	
	8	112	155	191	153	101	74	45	17	2	0	0	
	9	144	193	238	188	112	73	37	9	3	0	0	
	10	467	658	877	673	578	366	214	43	3	0	0	
	11	298	402	530	401	319	216	129	27	9	0	0	
	12	344	476	661	437	320	243	134	37	11	1	1	
	13	298	381	512	358	273	189	94	21	4	0	0	
	14	309	458	635	570	356	320	176	79	25	1	0	
	15	227	297	401	271	218	150	98	29	5	0	0	
	16	150	205	262	196	151	95	63	23	11	3	3	
	17	357	488	690	482	348	265	150	38	14	2	0	
	18	471	701	961	799	540	420	241	88	26	1	0	
	19	200	274	312	208	158	109	71	18	4	0	0	
	20	487	655	907	692	571	396	229	65	24	0	0	
	21	353	500	664	489	355	254	145	44	12	3	1	
	22	255	336	466	356	290	221	142	39	14	3	0	
	23	444	616	804	576	462	331	202	63	24	6	3	Sumatoria de los rangos
	24	342	513	712	593	381	312	130	54	12	0	0	
	25	331	439	601	377	305	223	130	40	12	0	0	
Rango		410	587	787	646	494	360	239	79	25	6	3	3636

En la Figura 29 se grafica la Tabla 7.

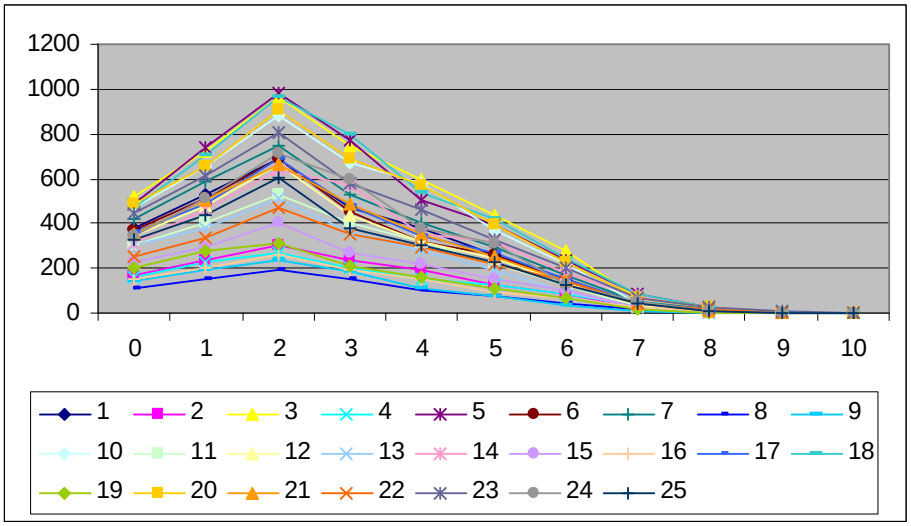

Figura 28: Distribución de frecuencias de cada uno de los agrupamientos para las distintas profundidades para TODOS los Experimentos, Proceso Biológico, agrupados por el método K-medias con 500 iteraciones y 25 agrupamientos

En el archivo **16K05PBprof-clust.xls** (incluido en el CD) se muestran los términos GO encontrados en los niveles de profundidad 4, 5 y 6, divididos por sus grupos. Cada notación GO tiene su frecuencia absoluta, la porcentual y porcentual acumulada. En el análisis se consideraron notaciones con frecuencias porcentuales mayor al 5 %, dentro de cada nivel y de cada agrupamiento.

Con respecto a la comparación entre los resultados obtenidos de los análisis de un experimento, el GSE3053 (archivo anexo 15K05PBprof-clust.xls) y el obtenido de los 3 experimentos en forma conjunta GSE3053, GSE4438 y GSE6901 (archivo anexo 16K05PBprof-clust.xls), el comentario del profesional con conocimiento de dominio es el siguiente:

"El punto de corte para comparar las anotaciones dentro de cada agrupamiento se estableció para términos con una abundancia relativa mayor o igual a 3.5% , con una profundidad 6 dentro de los niveles de ontología".

"Los términos GO asignados en los diferentes clusters son bastante parecidos entre los dos análisis. La mayoría de los clusters tienen asignados términos GO que corresponden a regulación de la transcripción y a modificaciones postraduccionales, sobre todo

fosforilaciones. Este segundo mecanismo es una forma de controlar la actividad de las proteínas ya sintetizadas, el agregado de un grupo fosfato en un lugar específico de la proteína puede cambiar dramáticamente su actividad. Es decir, que la mayoría de los clusters tienen asignados genes que intervienen en la regulación de los procesos celulares vía control de la transcripción o por control directo de las proteínas ya sintetizadas".

"Diferentes clusters, para los dos análisis de agrupamientos, contienen términos GO asociados al metabolismo del glucano, que ya fue informado como un factor asociado a las respuestas a estrés salino. Por ejemplo, en la variedad IR-28 de arroz, sensible a sal, al enfrentarla a un estrés salino se observó una disminución en la síntesis de almidón, un polisacárido de la familia de los glucanos".

"También se observan en los dos análisis, numerosos clusters con términos GO relacionados con el metabolismos de los aminoácidos y de los hidratos de carbono. Existen trabajos previos que indican que en varias especies vegetales la respuesta al estrés salino incluye un aumento de los pooles intracelulares de aminoácidos y algunos azúcares".

"Para los dos análisis se observa al menos un cluster que incluye un término GO referido al metabolismo de las hormonas vegetales, específicamente auxina. Además se encontraron otros clusters con anotaciones GO relacionadas con otro proceso de control y regulación, la muerte celular programada, que pertenece al grupo de respuestas que se induce en algunas especies frente al estrés salino".

"También se observan numerosos clusters que incluyen términos GO que se refieren a vías secretorias de proteínas y al tráfico intracelular de las mismas".

V. CONCLUSIONES

- Este trabajo abordó todos los pasos de un proceso de Descubrimiento del Conocimiento (KDD – Knowledge Discovery in Databases);

- La **preparación de los datos**, desde la búsqueda bibliográfica, la selección del conjunto de datos de prueba que reunieran las características adecuadas para encarar este trabajo, el análisis de la estructura de los datos y de los protocolos de almacenamiento de los datos en repositorios públicos, hasta la readecuación de las variables, incluidos los tratamientos de normalización de datos para permitir el procesamiento de los mismos.
- La utilización de los métodos de **explotación de los datos**, especificamente distintos métodos de agrupamiento con variadas configuraciones.
- Finalmente **la visualización de esta información** por medio de la interpretación de patrones de agrupamiento enriquecidos mediante el uso de ontologías, el establecimiento de un nivel de profundidad adecuado para analizar los conglomerados en relación a los términos GO asignados a cada gen, favoreciendo la extracción del conocimiento, especialmente para los casos de genes con frecuencia minoritaria dentro del conglomerado.

- Los pasos que insumieron la mayor parte del tiempo empleado para realizar el procesamiento completo de los datos fueron los siguientes en orden de importancia: la etapa inicial de entender la estructura de los datos a analizar y transformar o preprocesar los datos para luego aplicar el correspondiente algoritmo de minería de datos; la etapa de caracterizar cada acceso de la tabla por niveles de la jerarquía ontológica dado la gran cantidad de datos asociados por gen que presenta la Base de Datos GO; la etapa de procesamiento de los algoritmos de minería de datos, que demandó básicamente tiempo de computación.

- Es fundamental el conocimiento del dominio para definir y determinar la importancia biológica de los resultados. En este sentido se estableció como criterio para evaluar los métodos de agrupamiento aplicados la capacidad de incluir un gen con funciones no habituales en un agrupamiento de funcionalidad homogénea.

- Se determinó el uso de la sumatoria de los rangos como criterio para seleccionar que método formaba clusters más homogéneos en relación a la profundidad de los términos GO asociados.

- De todos los métodos de agrupamiento aplicados Kmedias con 500 iteraciones resulto ser el óptimo.

- Se alcanzó el objetivo principal propuesto en este trabajo, la exploración conjunta de datos obtenidos a partir de experimentos independientes, con el propósito de descubrir patrones de expresión que aporten nuevos conocimientos implícitos. Se refinó esta caracterización mediante el uso de jerarquías ontológicas GO, permitiendo completar un análisis predictivo sobre datos complejos que constituye el punto de partida para el planteo de hipótesis de futuros trabajos.

- Cabe destacar que el modelo propuesto se enmarca en áreas de actualidad e importancia en las ciencias de la vida. En un contexto donde existe una gran cantidad de datos disponibles y una fuerte necesidad de herramientas de análisis de datos complejos es imprescindible la convergencia de diversas áreas de conocimiento que den lugar al diseño e implementación de sistemas informáticos que soporten la integración de bases de datos heterogéneas con la información genómica y biológica en general.

VI. Trabajos a Futuro

Durante el desarrollo de este trabajo surgieron nuevas ideas y preguntas relacionadas, que aunque quedan fuera del alcance de esta tesis, sirven tanto para profundizar esta aproximación como para abrir nuevas líneas de investigación como las siguientes:

- Explorar la factibilidad de la aplicación de esta técnica para procesar todos los genes incluidos en la matriz de expresión, tanto los que tienen términos GO como los que no, como alternativa para asignar una posible anotación funcional a los genes aun no clasificados por GO.

- Construir un índice para evaluar el funcionamiento de los métodos de agrupamiento teniendo en cuenta la calidad y cantidad de términos GO no triviales, previamente desconocidos y potencialmente útil presenten en cada agrupamiento. Dado que este paso requiere de un conocimiento profundo del dominio en el que se esta trabajando, el índice debería "aprender" cuáles son los términos importantes o que características tienen los mismos para que en una aproximación heurística pudiera resolver el problema.

- Aplicar el modelo de análisis de datos desarrollado en este trabajo de tesis a distintos proyectos de investigación.

BIBLIOGRAFIA

[1] National Institutes of Health. National Human Genome Research Institute. 2003. "Del mapa a usted". Publication of NIH Num. 03-5377S. Maryland, Estados Unidos.

[2] Vazquez, Martin. 2006. "La intimidad de las moléculas de la vida. De los genes a las proteínas". Colección Ciencia Joven. Eudeba 108 Páginas. Buenos Aires. Argentina.

[3] Kelmansky, Diana. 2006. Apuntes de la Materia: "Análisis Exploratorio y Confirmatorio de Datos de Experimentos de Micromatriz". Departamento de Matemática - Instituto de Cálculo. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Buenos Aires. Argentina.

[4] Fernandes, Paula del Carmen. (Tesista). Heinz, Ruth A. (Directora). 2006. "Análisis Genómico de Girasol: Desarrollo de ESTs y de una plataforma bioinformática para estudios de expresión de genes candidatos en respuestas a estreses biológicos". Tesis Doctoral en Ciencias Biológicas. Universidad de Buenos Aires. Instituto de Biotecnología CICVyA INTA – Castelar. 2006. Páginas 35-52. Buenos Aires. Argentina.

[5] Larrañaga, Pedro. Calvo, Borja. Santana, Roberto. Bielza, Concha. Galdiano, Josu. Inza, Iñaki. Lozano, José A. Armañanzas, Rubén. Santafe, Guzman. Pérez, Aritz and Robles, Victor. 2005. "Machine learning in bioinformatics". Briefings in Bioinformatics. Vol. 7. Nº I. Pag. 86-112. Oxford, Inglaterra.

[6] Aas, Kjersti. 2001. "Microarray Data Mining: A Survey". Norsk Regnesentral / Norwegian Computing Center. 35 paginas. Oslo. Noruega.

[7] Bello, Ricardo; Colombini, Mauro; Takeda, Eugenia. 2006. "Análisis de Datos de Micromatriz de dos canales. Un acercamiento Teórico – Práctico", Apuntes de la Materia: Análisis Exploratorio y Confirmatorio de Datos de Experimentos de Micromatriz. Departamento de Matemática - Instituto de Cálculo. Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. Buenos Aires. Argentina.

[8] GeneChip® Rice Genome Array
http://www.affymetrix.com/products/arrays/specific/rice.affx

[9] Stekel, Dov. 2003. "Data Standards, Storage and Sharing". Microarray Bioinformatics. Cambridge University Press. Capitulo 11. Páginas 231 – 257.

[10] National Center for Biotechnology Information – NCBI
http://www.ncbi.nlm.nih.gov

[11] GEO Overview – NCBI
http://www.ncbi.nlm.nih.gov/geo/info/overview.html

[12] Kraemer, Alejandro Fausto. (Tesista). Marchesan, Enio. (Orientador). 2008. "Residual da mistura formulada dos herbicidas imazethapyr e imazapic em áreas de arroz sob diferentes manejos de solo". Tesis Maestre em Agronomia. Universidad Federal de Santa María. 63 Páginas. Santa María. Río Grande do Sul. Brasil.

[13] Asociación Argentina de Consorcios Regionales de Experimentación Agrícola.
http://www.aacrea.org.ar/economia/articulos/pdf/et03-01-04agroalimentos.pdf

[14] Gepas; http://gepas.bioinfo.cipf.es Departamento de Bioinformática. Centro de Investigación Príncipe Felipe. Valencia. España

[15] Kankainen, Matti; Brader, Günter; Törönen, Petri; Palva, E. Tapio and Holm, Liisa. 2006. "Identifying functional gene sets from hierarchically clustered expression data: map of abiotic stress regulated genes in *Arabidopsis thaliana*". Nucleic Acids Research, 2006, Vol. 34, No. 18 e124. 9 Páginas. Oxford, Inglaterra.

[16] Brameier, Markus; Wiuf, Carsten. 2006. "Co-clustering and visualization of gene expression data and gene ontology terms for Saccharomyces cerevisiae using self-organizing maps". Science Direct – Journal of Biomedical Informatic 40 (2007) pag 160-173.

[17] Gene Ontology Tools
http://www.geneontology.org/GO.tools.shtml

[18] GO Slim and Subset Guide
http://www.geneontology.org/GO.slims.shtml

[19] NCBI Plataforma GPL2025 – Affymetrix Gene Chip Rice Genome Array
http://www.ncbi.nlm.nih.gov/sites/entrez?db=gds&term=GPL2025%5Baccession%5D&cmd=search

[20] Software R
http://www.r-project.org/

[21] Modulo Bioconductor
http://www.bioconductor.org/

[22] Paquete GO.db
http://bioconductor.org/packages/2.2/data/annotation/html/GO.db.html

[23] Administrador de Base de Datos MySQL
http://www.mysql.com/

[24] GSE3053 record: Rice salt expression – NCBI
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE3053

[25] Walia H, Wilson C, Condamine P, Liu X et al. Comparative transcriptional profiling of two contrasting rice genotypes under salinity stress during the vegetative growth stage. Plant Physiol 2005 Oct;139(2):822-35.

[26] GSE4438 record: Expression data from rice under salinity stress – NCBI
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE4438

[27] Watson J, Crick F. A structure for Desoxyribose Nucleic Acid. Nature 1953. Vol 1, Num 171. Pag 737.
http://academy.d20.co.edu/kadets/lundberg/dnapaper.html

[28] GSE6901 record: Expression data for stress treatment in rice seedlings – NCBI
http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE6901

[29] Jain M, Nijhawan A, Arora R, Agarwal P et al. F-box proteins in rice. Genome-wide analysis, classification, temporal and spatial gene expression during panicle and seed development, and regulation by light and abiotic stress. Plant Physiol 2007 Apr;143(4):1467-83.

[30] GO Database Downloads – Gene Ontology
http://www.geneontology.org/GO.downloads.database.shtml

[31] Stekel, Dov. Microarray Bioinformatics. Cambridge Univesity Press. 2003, Cambridge, United Kingdom

[32] Affymetrix. Statistical Algorithms Description Document. 2002.Santa Clara. California. EEUU

[33] Curso de R: Capitulo 10: Análisis de conglomerados (cluster) I
http://es.geocities.com/r_vaquerizo/Manual_R10.htm

[34] Clustering (Clasificacion No supervisada) I – Edgar Acuna – Departamento de Matematicas – Universidad de Puerto Rico – Mayaguez – Puerto Rico.
math.uprm.edu/~edgar/marrayclass12.ppt

[35] Revisión de técnicas de agrupamiento de minería de datos espaciales en un SIG
http://www.monografias.com/trabajos27/datamining/datamining.shtml

[36] TIGER – The Institute of Genomic Research – actualmente Craig Venter Institute.
http://www.tigr.org/db.shtml

[37] IRGSP – The International Rice Genome Sequencing Project.
http://rgp.dna.affrc.go.jp/IRGSP/

[38] Whetzel, P.L, et al. The MGED Ontology: a resource for semantics-based description of microarray experiments. Bioinformatics Advance Access. Bioinformatics 2006 22: 866-873. January 21, 2006.
http://bioinformatics.oxfordjournals.org/content/vol22/issue7/

[39] Paul T Spellman, Paul T, et al. Design and implementation of microarray gene expression markup language (MAGE-ML). Genome Biology. 2002.
http://genomebiology.com/2002/3/9/research/0046

[40] Barrett, T., et al.; NCBI GEO: mining tens of millions of expression profiles—database and tools update.
http://nar.oxfordjournals.org/cgi/content/abstract/gkl887v1

[41] ArrayExpress – EBI
http://www.ebi.ac.uk/microarray-as/aer/#ae-main[0]

[42] Verlues PE, Agarwal M, Katiyar-Agarwal S, Zhu J, Zhu K: Methods and concepts in quantifying resistance to drought, salt and freezing, abiotic stresses that affect plant water status (2006). Plant J. 45(4):523-39

[43] Zhu, J.K: Plant salt tolerance. Trends in Plant Science (2001) (a), 6, 2, 66-71.

[44] Chinnusamy V, Ohta M, Kanrar S, Lee BH, Hong X, Agarwal M, Zhu JK: ICE1: a regulator of cold-induced transcriptome and freezing tolerance in Arabidopsis (2003). Genes y Dev. 17(8):1043-54.

[45] Agarwal, P.K, Agarwal, P., Reddy, M.K. and Sopory, S.: Role of DREB transcription factors in abiotic and biotic stress tolerance in plants. Plant Cell Rep. (2006), DOI 10.1007/s00299-006-0204-8

[46] Fujita M., Fujita Y., Noutoshi Y., Takahashi F., Narusaka Y., Yamaguchi-Shinozaki K. and Shinozaki K.: Crosstalk between abiotic and biotic stress responses: a current view from the points of convergence in the stress signaling network (2006). Current Opinion in Plant Biology, 9:436-442.

ANEXO A

A.1. Consulta de muestras a analizar

```
# MODULO: 01-CONSULTA
# Establece enlace con la base de datos GO en MySQL y realiza una consulta segun los parámetros
# solicitados al comienzo.
# Luego reordena la anterior salida en formato matricial
# ------------------------------------------------------------------------------------------------------------------
# carga el data set go
rm(list=ls())
# carga la libreria necesaria
library(RMySQL)
# ------------------------------------------------------------------------------------------
# Para realizar la consulta deseada se debe definir los siguientes parámetros
# ------------------------------------------------------------------------------------------
# ¿QUÉ TIPO DE TERMINO GO VA A PROCESAR?
# ------------------------------
# | OPCION | TIPO DE TERMINO GO |
# ------------------------------
#    1     PROCESO BIOLOGICO
#    2     COMPONENTE CELULAR
#    3     FUNCION MOLECULAR
# ------------------------------
tg <- 1
# ------------------------------------------------------------------------------------------
# ¿QUÉ EXPERIMENTO O EXPERIMENTOS VA A PROCESAR?
# ----------------------
# | OPCION | EXPERIMENTO|
# ----------------------
#    1      GSE3053
#    2      GSE4438
#    3      GSE6901
#    4      TODOS
# ----------------------
ex <- 4
# ------------------------------------------------------------------------------------------
# ¿TRABAJARÁ CON LOS EXPERIMENTOS, LOS CONTROLES O TODOS ?
# --------------------------------
# | OPCION | EXPERIMENTO/CONTROL |
```

```
# ---------------------------------
#    0     EXPERIMENTO
#    1     CONTROL
#    2     TODOS
# ---------------------------------
ct <- 2
# ------------------------------------------------------------------------------------
# ¿TRABAJARA CON LOS EXPERIMENTOS SOMETIDOS A STRESS SALÍNICO, O LOS
#  RESTANTES O TODOS ?
# ----------------------------
# | OPCION | STRESS SALINICO |
# ----------------------------
#    0     NO
#    1     SI
#    2     TODOS
# ----------------------------
sa <- 2
# ------------------------------------------------------------------------------------
# ¿TRABAJARÁ CON LOS EXPERIMENTOS SOMETIDOS A ESTRÉS HÍDRICO, O LOS
# RESTANTES O TODOS ?
# ----------------------------
# | OPCIÓN | STRESS HIDRICO |
# ----------------------------
#    0     NO
#    1     SI
#    2     TODOS
# ----------------------------
se <- 2
# ------------------------------------------------------------------------------------
# ¿TRABAJARÁ CON LOS EXPERIMENTOS SOMETIDOS A ESTRÉS DE FRIO, O LOS
# RESTANTES O TODOS ?
# ----------------------------
# | OPCION | ESTRÉS DE FRÍO |
# ----------------------------
#    0     NO
#    1     SI
#    2     TODOS
# ----------------------------
fr <- 2
# ------------------------------------------------------------------------------------
```

```
# ¿TRABAJARA CON LOS EXPERIMENTOS CON ESTADIO DE PLANTAS DESARROLLADAS,
# O PLÁNTULAS O TODOS ?
# ------------------------------
# | OPCION | PLANTA / PLANTULA |
# ------------------------------
#    0     PLANTULA
#    1     PLANTA
#    2     TODOS
# -------------------------
es <- 2
# ------------------------------------------------------------------------------------------------------------------
m <- MySQL()
con <- dbConnect(m, host="localhost", user="root", password = "dm" , dbname = "go")
# arma la consulta
# exper.exgroup
arma_con <- "select sample.id_Affy, exper.id_Sample, sample.Expresion, exper.exgroup from((exper
join sample use index (samp_idx) using (id_Sample)) join affychip use index (affyc_idx, affych_idx)
using (id_Affy)) join term use index (term_idx, termt_idx) using (acc) "

if (tg==1) arma_con <- paste(arma_con,"where term.term_type = 'biological_process' ") else
if (tg==2) arma_con <- paste(arma_con,"where term.term_type = 'cellular_component' ") else
if (tg==3) arma_con <- paste(arma_con,"where term.term_type = 'molecular_function' ")

if (ex==1) arma_con <- paste(arma_con,"and exper.id_Exp = 'GSE3053' ") else
if (ex==2) arma_con <- paste(arma_con,"and exper.id_Exp = 'GSE4438' ") else
if (ex==3) arma_con <- paste(arma_con,"and exper.id_Exp = 'GSE6901' ") else
if (ex==4) arma_con <- arma_con

if (ct==0) arma_con <- paste(arma_con,"and exper.ctrl = 0 ") else
if (ct==1) arma_con <- paste(arma_con,"and exper.ctrl = 1 ") else
if (ct==2) arma_con <- arma_con

if (sa==0) arma_con <- paste(arma_con,"and exper.sal = 0 ") else
if (sa==1) arma_con <- paste(arma_con,"and exper.sal = 1 ") else
if (sa==2) arma_con <- arma_con

if (se==0) arma_con <- paste(arma_con,"and exper.seq = 0 ") else
if (se==1) arma_con <- paste(arma_con,"and exper.seq = 1 ") else
if (se==2) arma_con <- arma_con
```

```
if (fr==0) arma_con <- paste(arma_con,"and exper.frio = 0 ") else
if (fr==1) arma_con <- paste(arma_con,"and exper.frio = 1 ") else
if (fr==2) arma_con <- arma_con

if (es==0) arma_con <- paste(arma_con,"and exper.estad = 0 ") else
if (es==1) arma_con <- paste(arma_con,"and exper.estad = 1 ") else
if (es==2) arma_con <- arma_con

arma_con

# realiza la consulta
matri <- dbGetQuery(con, arma_con)
matri1 <- aggregate.data.frame(matri$Expresion, list(matri$id_Affy, matri$exgroup), FUN=mean)
matrire <- reshape(matri1, timevar="Group.2", idvar=c("Group.1"), direction="wide")
```

A.2. Normalizar dentro de muestras, entre muestras y entre experimentos Gráficos

```
# ------------------------------------------------------------------------------------------------------------
# MODULO: 02-NORMALIZACIÓN
# Normaliza los valores de la matriz obtenida en tres niveles. Dentro de la matriz, entre matrices y
entre experimentos.
# ------------------------------------------------------------------------------------------------------------
# ||| DENTRO DEL ARRAY ||| Transforma la matriz de datos a logaritmo (base 2)
#                    Plot e histogramas
# ------------------------------------------------------------------------------------------------------------
# Calcula la dimensión de la matriz, extrae la columna Id_Affy y valores de Experimentos por
# separado
# Calcula la dimensión de la matriz de valores
#-------------------------------------------------------------------------------------------------
dim_matrire <- dim(matrire)
matrire_rec_affy <- matrire[1:dim_matrire[1],1]
matrire_rec <- matrire[1:dim_matrire[1],2:dim_matrire[2]]
dim_matrire_rec <- dim(matrire_rec)
#-------------------------------------------------------------------------------------------------
# Grafica los experimentos contra la mediana de los experimentos
#-------------------------------------------------------------------------------------------------
med_matrire_rec <- array(0, c(dim_matrire_rec[1],1))
for (i in 1:dim_matrire_rec[1])
        {
                med_matrire_rec[i,1] <- median(as.numeric(matrire_rec[i,]))
        }
if ((trunc(dim_matrire_rec[2]/2)) == (dim_matrire_rec[2]/2))
                {fil_graf <- dim_matrire_rec[2]/2} else
                {fil_graf <- (trunc(dim_matrire_rec[2]/2)+1)}
par(mfrow = c(fil_graf , 2))
for (i in 1:dim_matrire_rec[2])
        {
                plot(med_matrire_rec[,1], matrire_rec[,i], type="p")
        }
#-------------------------------------------------------------------------------------------------
# Grafica el histograma de los experimentos
#-------------------------------------------------------------------------------------------------
```

```
if ((trunc(dim_matrire_rec[2]/2)) == (dim_matrire_rec[2]/2))
                    {fil_graf <- dim_matrire_rec[2]/2} else
                    {fil_graf <- (trunc(dim_matrire_rec[2]/2)+1)}
par(mfrow = c(fil_graf , 2))
for (i in 1:dim_matrire_rec[2])
        {
                    hist(matrire_rec[,i],nclass=30)
        }
#-------------------------------------------------------------------------------------------
# Calcula el logaritmo de los experimentos
#-------------------------------------------------------------------------------------------
matrire2 <- log(matrire_rec,2)
dim_mat2 <- dim(matrire2)
# ------------------------------------------------------------------------------------------------------------------------
# ||| ENTRE ARRAY Y ENTRE EXPERIMENTOS |||
# Centrar y escalar la matriz de datos logaritmo (base 2) - BoxPlot
#                         Reescribe los menores a 2 % y mayores al 98%
# ------------------------------------------------------------------------------------------------------------------------
matri_fin <- array(0, c(dim_mat2[1],1))
for (i in 1:dim_mat2[2])
        {
                    matri_cuar <- quantile(matrire2[,i], c(0.02 , 0.98))
                    matri_min <- ifelse(matrire2[,i] < matri_cuar[1], matri_cuar[1], matrire2[,i])
                    matri_max <- ifelse(matri_min > matri_cuar[2], matri_cuar[2], matri_min)
                    if (i == 1) {matri_fin <- matri_max} else {matri_fin <- cbind(matri_fin, matri_max)}
    }
dimnames(matri_fin)[[2]] <- dimnames(matrire2)[[2]]
matrire2_es <- scale(matri_fin)
dim_matrire2_es <- dim(matrire2_es)
#-------------------------------------------------------------------------------------------
# Grafica los experimentos contra la mediana de los experimentos CORREGIDOS
#-------------------------------------------------------------------------------------------
med_matrire2_es <- array(0, c(dim_matrire2_es[1],1))
for (i in 1:dim_matrire2_es[1])
        {
                    med_matrire2_es[i,1] <- median(as.numeric(matrire2_es[i,]))
        }
if ((trunc(dim_matrire2_es[2]/2)) == (dim_matrire2_es[2]/2))
                    {fil_graf <- dim_matrire2_es[2]/2} else
                    {fil_graf <- (trunc(dim_matrire2_es[2]/2)+1)}
```

```
par(mfrow = c(fil_graf , 2))
for (i in 1:dim_matrire2_es[2])
        {
                plot(med_matrire2_es[,1], matrire2_es[,i], type="p")
        }
#-------------------------------------------------------------------------------------------
# Grafica el histograma de los experimentos CORREGIDOS
#-------------------------------------------------------------------------------------------
if ((trunc(dim_matrire2_es[2]/2)) == (dim_matrire2_es[2]/2))
                {fil_graf <- dim_matrire2_es[2]/2} else
                {fil_graf <- (trunc(dim_matrire2_es[2]/2)+1)}
par(mfrow = c(fil_graf , 2))
for (i in 1:dim_matrire2_es[2])
        {
                hist(matrire2_es[,i],nclass=15)
        }
```

A.3. Métodos de Agrupamiento

```
# ------------------------------------------------------------------------------------------------------------
# CLUSTER JERÁRQUICO
# ------------------------------------------------------------------------------------------------------------
-------------------
MÉTODO DE DISTANCIA : euclidean - maximum - manhattan - canberra - binary - minkowski
-------------------
mat10 <- dist(matrire2_es, method="euclidean")
-----------------
MÉTODO DE CLUSTER : ward - single - complete - average - mcquitty - median - centroid
-----------------
mat11 <- hclust(mat10, method = "ward")
*** plotea el dendograma
plot(mat11)
*** historia de la aglomeración
mat11$merge
*** recorta el dendograma visualmente
rect.hclust(mat11,k=9)
*** plotea el recorte del dendograma
print(rect.hclust(mat11,k=10))
*** agrupa los individuos en 2
cutree(mat11,k=10)
*** agrupa los individuos en 2 y 3
cutree(mat11,k=2:3)
*** adjunta al data frame los agrupamiento de individuos en 2 y 4
matrire2a=data.frame(matrire_rec_affy, matrire2_es,cutree(mat11,k=9))
# ------------------------------------------------------------------------------------------------------------
# CLUSTER K MEDIAS
# ------------------------------------------------------------------------------------------------------------
# con 500 y 1000 iteraciones (entre 1000 y 2000 iteraciones cambian de grupo peron mantienen los
# patrones de agrupamiento)
centr <- 9
km <- kmeans(matrire2_es, iter.max = 1000, centr)
matrire2a=data.frame(matrire_rec_affy, matrire2_es, km$cluster)
# ------------------------------------------------------------------------------------------------------------
# CLUSTER PAM
# ------------------------------------------------------------------------------------------------------------
library(cluster)
# MÉTODO DE DISTANCIA : euclidean  - manhattan
```

```
medoid <- 9
pamy <- pam(matrire2_es, medoid, diss = FALSE, metric = "manhattan" )
matrire2a=data.frame(matrire_rec_affy, matrire2_es, pamy$clustering)
# ------------------------------------------------------------------------------------------------------------------
# CLUSTER Clara (clustering large applications: PAM method for large data sets)
# ------------------------------------------------------------------------------------------------------------------
library(cluster)
# MÉTODO DE DISTANCIA: euclidean  - manhattan
clmed <- 9
clarax <- clara(matrire2_es, clmed, metric = "manhattan", stand = FALSE)
matrire2a=data.frame(matrire_rec_affy, matrire2_es, clarax$clustering)
# ------------------------------------------------------------------------------------------------------------------
# SOM
# ------------------------------------------------------------------------------------------------------------------
library(som)
xd <- 3
yd <- 3
----------
TOPOLOGÍA : "hexa" - "rect"
----------
topo <- "rect"
---------------------
FUNCIÓN DE VECINDAD: "bubble" - "gaussian"
---------------------
nei = "bubble"
y.som <- som(matrire2_es, xdim = xd, ydim = yd, topol = topo, neigh = nei)
plot(y.som)
matrire2a=data.frame(matrire_rec_affy, matrire2_es, y.som$visual)
dim_matrire2a <- dim(matrire2a)
for (i in 1:dim_matrire2a[1])
{
if (matrire2a[i,(dim_matrire2a[2]-2)] == 0 & matrire2a[i,(dim_matrire2a[2]-1)] == 0)
        {matrire2a[i,dim_matrire2a[2]] = 1}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 0 & matrire2a[i,(dim_matrire2a[2]-1)] == 1)
        {matrire2a[i,dim_matrire2a[2]] = 2}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 0 & matrire2a[i,(dim_matrire2a[2]-1)] == 2)
        {matrire2a[i,dim_matrire2a[2]] = 3}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 1 & matrire2a[i,(dim_matrire2a[2]-1)] == 0)
        {matrire2a[i,dim_matrire2a[2]] = 4}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 1 & matrire2a[i,(dim_matrire2a[2]-1)] == 1)
        {matrire2a[i,dim_matrire2a[2]] = 5}
```

```
if (matrire2a[i,(dim_matrire2a[2]-2)] == 1 & matrire2a[i,(dim_matrire2a[2]-1)] == 2)
        {matrire2a[i,dim_matrire2a[2]] = 6}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 2 & matrire2a[i,(dim_matrire2a[2]-1)] == 0)
        {matrire2a[i,dim_matrire2a[2]] = 7}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 2 & matrire2a[i,(dim_matrire2a[2]-1)] == 1)
        {matrire2a[i,dim_matrire2a[2]] = 8}
if (matrire2a[i,(dim_matrire2a[2]-2)] == 2 & matrire2a[i,(dim_matrire2a[2]-1)] == 2)
        {matrire2a[i,dim_matrire2a[2]] = 9}
}
for (i in 1:dim_matrire2a[1])
{
matrire2a[i,(dim_matrire2a[2]-2)] = matrire2a [i,(dim_matrire2a[2])]
}
matrire2a <- matrire2a[,1:(dim_matrire2a[2]-2)]
```

A.4.1. Rescatar términos GO

```
# Carga de la librería GO.db
library(GO.db)
# ------------------------------------------------------------------------
# RENOMBRA ultima columna de matrire2a
# ------------------------------------------------------------------------
dim_matrire2a <- dim(matrire2a)
dimnames(matrire2a)[[2]][dim_matrire2a[2]] <- "cluster"
# ------------------------------------------------------------------------
# Rescata todos los términos GO para cada uno de los ID_Affy del cluster solicitado
# ------------------------------------------------------------------------
GO_acu <- array(0, c(1,3))
cc = 0
for (i in 1:dim_matrire2a[1])
#for (i in 1:100)
{
cc = cc + 1
arma_c1 <- "select id_Affy, acc from affychip join term use index (term_idx, termt_idx) using (acc) "

if (tg==1) arma_c1 <- paste(arma_c1,"where term.term_type = 'biological_process' ") else
if (tg==2) arma_c1 <- paste(arma_c1,"where term.term_type = 'cellular_component' ") else
if (tg==3) arma_c1 <- paste(arma_c1,"where term.term_type = 'molecular_function' ")

arma_c1 <- paste(arma_c1,"and affychip.id_Affy = '",matrire2a[i,1],"'", sep = "")

cons_GO <- dbGetQuery(con, arma_c1)
dim_cons_GO <- dim(cons_GO)

cons_G <- array(matrire2a[i,dim_matrire2a[2]], c(dim_cons_GO[1],1))
cons_GO <- cbind(cons_GO, cons_G)

if (i == 1) {GO_acu <- cons_GO} else {GO_acu <- rbind(GO_acu, cons_GO)}
}
```

A.4.2. Rescatar ancestros de términos GO

```
dim_GO_acu <- dim(GO_acu)
GO_anc <- array(0, c(1,4))
GO_ancb <- array(0, c(1,4))
k = 1
for (i in 1:dim_GO_acu[1])
        {
                comp1 <- paste("GO:",substr(GO_acu[i,2],4,10), sep ="")
                if (tg==1) comp2 <- GOBPANCESTOR else
                if (tg==2) comp2 <- GOCCANCESTOR else
                if (tg==3) comp2 <- GOMFANCESTOR
                aux_GO <- get(comp1,comp2)
                dim_aux_GO <- length(aux_GO)
                for (j in 2:dim_aux_GO)
                        {
                        if(k == 1)
                                {
                                GO_anc[k,1] <- GO_acu[i,1]
                                GO_anc[k,2] <- GO_acu[i,2]
                                GO_anc[k,3] <- GO_acu[i,3]
                                GO_anc[k,4] <- aux_GO[j]
                                }
                        if(k >= 2)
                                {
                                GO_ancb[1,1] <- GO_acu[i,1]
                                GO_ancb[1,2] <- GO_acu[i,2]
                                GO_ancb[1,3] <- GO_acu[i,3]
                                GO_ancb[1,4] <- aux_GO[j]
                                GO_anc <- rbind(GO_anc, GO_ancb)
                                }
                        k = k + 1
                        }
                GO_ancb[1,1] <- GO_acu[i,1]
                GO_ancb[1,2] <- GO_acu[i,2]
                GO_ancb[1,3] <- GO_acu[i,3]
                GO_ancb[1,4] <- GO_acu[i,2]
                GO_anc <- rbind(GO_anc, GO_ancb)
        }
```

A.4.3. Rescatar la profundidad jerárquica de los ancestros del términos GO

```
dim_GO_anc <- dim(GO_anc)
GO_anc_DF <- as.data.frame(GO_anc)
GO_prof <- array(0, c(dim_GO_anc[1],4))

if (tg==1) cod <- 5655 else
if (tg==2) cod <- 3596 else
if (tg==3) cod <- 1963

cce = 0

for (i in 1:dim_GO_anc[1])
        {
        prof1 <- "select term.acc, graph_path.term1_id, term.name, graph_path.term2_id, graph_path.distance from term use index (term_idx) left join graph_path ON term.id = graph_path.term2_id where acc = "
        profGO <- paste(prof1,"'",GO_anc_DF[i,4],"'"," and term1_id = ", cod, sep ="")
        GO_prof_aux <- dbGetQuery(con,profGO)
        a = dim(as.matrix(GO_prof_aux))
        if (a[1] != 0)
        {
        GO_prof[i,1] <- GO_prof_aux[1,2]
        GO_prof[i,2] <- GO_prof_aux[1,3]
        GO_prof[i,3] <- GO_prof_aux[1,4]
        GO_prof[i,4] <- GO_prof_aux[1,5]
        }
        else
        {
        cce = cce + 1
        }
        }
GO_prof_DF <- as.data.frame(GO_prof)
FINAL <- cbind(GO_anc_DF, GO_prof_DF)
TOTAL <- table(FINAL[,3], FINAL[,8])
table(FINAL[,5])
write.table(FINAL,file="E:/Tesis Maestria/ataie/tot53pbmacom0901.dat")
abc <- read.table("E:/Tesis Maestria/ataie/tot53pbmacom0901.dat")
```

ANEXO B

Glosario

Background	Señal de fondo
Foreground	Señal específica
Direct Acyclic Graph. (DAG).	Gráfico Acíclico Dirigido
Clustering	Conglomerados , agrupaciones
Self-Organizing Maps (SOM)	Mapas autoorganizados
Dataset	Conjunto de datos
Pipeline	Flujo de procesos
Sample	Muestra
Data frame	Marco de datos