UNIVERSIDAD NACIONAL DE COLOMBIA

Identificación de relaciones entre genes utilizando técnicas de inteligencia computacional

Liliana Marcela Olarte Mesa

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

Identificación de relaciones entre genes utilizando técnicas de inteligencia computacional

Liliana Marcela Olarte Mesa

Tesis de investigación presentada como requisito parcial para optar al título de:

Magister en Ingeniería de Sistemas y Computación

Director:

Luis Fernando Niño Vásquez, Ph. D.

Codirectora:

Liliana López Kleine, Ph. D.

Grupo de Investigación:

Laboratorio de Investigación en Sistemas Inteligentes (LISI)

Universidad Nacional de Colombia

Facultad de Ingeniería

Departamento de Ingeniería de Sistemas e Industrial

Bogotá, Colombia

2014

A mis padres y hermanos.

Declaración de autoría

Durante el desarrollo de la presente tesis, se presentaron los siguientes trabajos en eventos académicos internacionales:

Olarte, L. M.; López-Kleine, L.; Niño, L. F. *Strategy for the detection of functional clusters of genes using data mining techniques*. ISCB Latin America 2012 Conference on Bioinformatics. Chile. 2012.

Olarte, L. M.; López-Kleine, L.; Niño, L. F. *Identification and analysis of genes clusters in biological data*. IEEE International Conference on Bioinformatics and Biomedicine (BIBM12). Estados Unidos. 2012.

Agradecimientos

Primero que todo, agradezco a Dios por darme la oportunidad de realizar mis estudios de posgrado y por poder terminar este trabajo de investigación. A la vez, agradezco a mis directores de tesis, el profesor Luis Fernando Niño y la profesora Liliana López Kleine, por su apoyo, dedicación, enseñanza y orientación durante todo el desarrollo de este trabajo. A mi asesor Daniel Restrepo Montoya por ser una guía y apoyo en el entendimiento de los conceptos relacionados con este trabajo. También agradezco a mis padres, hermanas y abuela, porque siempre han estado presentes en mi vida y me han apoyado para alcanzar mis proyectos. A Juan Guillermo Carvajal Patiño, por su cooperación. Finalmente, agradezco a mis compañeros del grupo de investigación LISI de la Universidad Nacional de Colombia.

Resumen

En este trabajo se propone una metodología general para la identificación de relaciones entre genes a partir de datos de expresión obtenidos mediante dos técnicas diferentes: microarreglos de ADN y secuenciación directa del ARN mensajero (RNA_Seq), e integrando datos de categorías biológicas con las que están asociados los genes. La metodología propuesta contempla diversas fases como selección de genes, agrupamiento, análisis de los grupos, construcción de redes de interacción entre genes y comparación biológica de los resultados. En cada una de las fases de la metodología se aplican técnicas de inteligencia computacional conformadas por teorías y algoritmos de minería de datos y aprendizaje de máquina. Para llevar a cabo cada una de estas fases se emplearon datos de expresión y categóricos de la planta *Arabidopsis thaliana.* Los resultados obtenidos reflejaron que la metodología propuesta permite la integración de datos de diferente naturaleza aportando más información al caso de estudio y adicionalmente obtener relaciones entre genes.

Palabras clave: Gen, biología de sistemas, expresión de genes, redes biológicas.

Abstract

In this work, a general methodology for the identification of relationships between genes from expression data using two different techniques (DNA microarrays and RNA_Seq) is proposed. This technique is based on integrating data from biological categories associated to the genes. The proposed methodology comprises several stages such as gene selection, gene clustering, group analysis, building of interaction networks between genes, and biological comparison of the results. In each phase of the methodology, some computer intelligence techniques, based on data mining and machine learning theories and algorithms, were applied. To carry out each phase, expression and category data from the plant *Arabidopsis thaliana* were used. The results showed that the proposed methodology allows the integration of different kinds of data contributing more information to the case study and obtaining gene-gene relationships.

Keywords: Gene, systems biology, gene expression, biology networks.

Contenido

Lista de figuras

Lista de tablas

Lista de símbolos y abreviaturas

Abreviaturas

Abreviatura	Término
ACP	Análisis de componentes principales
ADN	Ácido desoxirribonucleico
ADN_C	Ácido desoxirribonucleico complementario
AGRIS	*Arabidopsis* Gene Regulatory Information Server
ARN	Ácido ribonucleico
AtTFDB	*Arabidopsis thaliana* Transcription Factor Data Base
bp	Pares de bases
GO	Gene Ontology
KACP	*Kernel* análisis de componentes principales
KEGG	Kyoto Encyclopedia of Genes and Genomes
NCBI	National Center for Biotechnology Information

Introducción

En el último siglo, ha aumentado continuamente la cantidad de información biológica, debido al incremento en el número de experimentos y proyectos en este campo, haciendo que se dificulten los procesos de manejo y análisis de la misma [1]. Por lo tanto, han surgido diversas áreas de investigación que permiten el procesamiento y análisis de la información, como la Bioinformática, la Biología Computacional y la Biología de Sistemas. La Bioinformática consiste en la aplicación y desarrollo de métodos computacionales para comprender y organizar la información biológica [2]. Por su parte, la Biología Computacional permite descubrir y comprender conocimiento a partir del análisis de información acerca de sistemas biológicos mediante el uso de la computación [3]. Por otro lado, la Biología de Sistemas es un campo de investigación que trata de analizar sistemas biológicos a gran escala [4]. Teniendo presente lo expuesto, estas áreas ahora son consideradas un elemento importante de la investigación biológica contemporánea [3].

Dentro de estas áreas, se estudian diversos problemas biológicos que buscan ser solucionados a través del uso de diversas herramientas o modelos computacionales. Entre estos se halla el problema de la identificación de interacciones y/o relaciones entre genes, el cual es el fundamento del presente trabajo de investigación. Las relaciones entre genes tienen que ver con el control de los procesos celulares, debido a que en el proceso de control celular están involucradas la regulación de los genes y las interacciones entre ADN, ARN, proteínas y pequeñas moléculas. La regulación de la actividad de los genes contempla la activación o la inhibición de los mismos en un momento determinado. Por medio de esta regulación se puede adquirir un mejor entendimiento acerca de los procesos llevados a cabo a nivel molecular [5].

Además, la identificación de las interacciones entre genes se considera importante en la medida en que permite deducir las propiedades reflejadas en un sistema vivo y las propiedades de las proteínas que codifican los genes, puesto que son especificadas por

los mismos genes [3]. En este sentido, una vía para observar las propiedades de un sistema vivo a partir de sus genes y proteínas es estudiando la regulación de la expresión de dichos genes. Para ello, es relevante tener en cuenta el dogma central de la biología, en donde se puede apreciar la relación entre el ADN y las proteínas y la descripción de cómo el ADN se transcribe a ARN mensajero y éste se traduce a proteína [6].

El proceso de transcripción es crucial en la regulación de los genes, debido a que finalmente puede indicar muchas otras cascadas de eventos biológicos y relaciones entre ellos. El estudio de los niveles de ARN en una célula puede brindar información útil para la comprensión de una amplia variedad de sistemas biológicos [7]. Además, en el proceso de transcripción algunos genes pueden expresarse o no de acuerdo a diversos factores internos y externos al organismo, dando lugar a diversos perfiles de expresión, los cuales proporcionan una imagen general de la función celular.

El análisis de perfiles de expresión de genes posibilita el hallazgo de la similitud de expresión entre los mismos genes, los cuales pueden reflejar diferentes relaciones entre ellos: metabólicas, de señalización, de regulación, etc. Los genes fuertemente correlacionados tienen una mayor probabilidad de expresarse de manera similar y de compartir las mismas funciones o mecanismos de regulación [8]. Los perfiles de expresión génica permiten la comparación de perfiles de genes en tejidos y células (normales y patológicos). Por otra parte, permiten establecer relaciones entre genes (por ejemplo, agrupamiento de genes, patrones de expresión de coincidencia temporal), comprender los mecanismos de la enfermedad a nivel molecular, y definir y validar nuevas drogas [9].

Para el análisis de datos de expresión es necesario tener en cuenta la forma en la que se va a cuantificar el nivel de expresión de los genes y la forma en la que van a ser analizados. De este modo, para la cuantificación del nivel de expresión en los genes bajo ciertas condiciones particulares, existen técnicas que pueden basarse en visualización, hibridación o en secuenciación [10]. Todas ofrecen información acerca del perfil de expresión de múltiples genes a la vez, mientras que, en técnicas anteriores —por ejemplo *Northern blot analysis* [11], *Serial analysis of gene expression* (SAGE) [12] y *Differential display* [13]—, no era posible tener una medida de varios genes al mismo tiempo. En el caso del análisis de datos de expresión, existen diversas formas; por ejemplo, el análisis de la expresión diferencial, la construcción y análisis de grupos de genes a partir de su

expresión, y la construcción de redes de interacción entre genes para visualizar de forma gráfica las relaciones entre estos.

Por lo tanto, para el empleo y análisis de datos de expresión en este trabajo se han escogido los dos tipos de técnicas mencionadas: hibridación utilizando microarreglos y secuenciación utilizando RNA_Seq. Por otra parte, este trabajo se enfoca en dos tipos de análisis de datos de expresión: construcción de grupos de genes y redes de interacción entre genes. Se emplean métodos computacionales enmarcados dentro de las áreas de minería de datos y del aprendizaje maquinal para procesar, analizar, interpretar e integrar expresión génica con categorías funcionales atribuidas a los genes como funciones biológicas, rutas biológicas y factores de transcripción. Esto con el fin de aportar conocimiento biológico previo para fortalecer la identificación de relaciones entre genes y brindar un marco de análisis completo, descartando análisis de únicamente datos de expresión y obteniendo datos compuestos de naturaleza heterogénea.

De esta forma, se integró la información de expresión con las categorías de funciones y rutas biológicas empleando técnicas de aprendizaje de máquina. Con esto se aplicaron posteriormente algunos métodos de minería de datos para crear grupos de genes, los cuales fueron analizados y relacionados con los factores de transcripción. Se identificaron así los factores de transcripción más representativos en cada grupo, que a la vez estaban relacionados entre sí. Además, se emplearon otros métodos computacionales para la identificación de las relaciones entre genes y la construcción de redes de interacción entre los mismos, seleccionando aquellas relaciones que fueran más significativas y que representaran un grafo no aleatorio.

Este documento se encuentra organizado como se describe a continuación. En el capítulo uno, se presentan los conceptos fundamentales relacionados con la investigación realizada. En el capítulo dos, se presentan las técnicas y recursos computacionales previamente publicados que fueron aplicados para resolver el problema planteado. Posteriormente, en el capítulo tres, se describen los datos de expresión y categóricos utilizados; también se presenta la metodología que se propone como base para la identificación de relaciones entre genes. En el capítulo cuatro, se presentan los resultados obtenidos y el análisis de los mismos después de aplicar la metodología propuesta. Por

último, se encuentran las conclusiones y trabajo futuro que puede permitir complementar la investigación realizada.

1. Conceptos fundamentales

Con el fin de tener una mayor comprensión acerca del contenido de este documento, en este capítulo se describen de forma breve algunos de los conceptos biológicos fundamentales y los métodos computacionales empleados en esta investigación.

1.1 Conceptos biológicos

En esta sección, se presentan los conceptos biológicos que servirán como base para el entendimiento del problema biológico de *la identificación de relaciones entre genes*, el cual se pretende resolver en este trabajo.

1.1.1 Gen

Según [14] un gen es una unidad de información genética: “es una secuencia de ADN cromosómico necesaria para la elaboración de un producto funcional, sea un polipéptido o una molécula de ARN funcional”.

1.1.2 Genoma

El genoma es la suma total de la información genética correspondiente a cada especie. Es posible estudiar todo el genoma de forma completa y no solamente cada gen por separado. De esta forma, es posible analizar la expresión génica, las variaciones de los genes y las interacciones entre ellos y el ambiente [14].

1.1.3 Transcriptoma

El transcriptoma de un organismo es el conjunto completo de transcripciones en una célula. El entendimiento del transcriptoma es esencial para la interpretación de los elementos funcionales del genoma y para revelar los componentes moleculares de células y tejidos.

También es útil para comprender el desarrollo del organismo y las enfermedades con las que se encuentra asociado [15].

Existen diversos métodos para realizar mediciones cuantitativas sobre un transcriptoma; entre los más representativos se encuentran los métodos basados en hibridación y los basados en secuenciación.

Entre los métodos basados en hibridación se encuentran los microarreglos, los cuales representan el genoma de alta densidad y permiten el mapeo de las regiones transcritas a una resolución muy alta, a partir de varios pares de bases, aproximadamente 100 bp [16] [17]. Los enfoques basados en hibridación son de alto rendimiento y relativamente económicos, pero presentan algunas limitaciones relacionadas con la confianza en la información presente acerca de la secuencia del genoma, los altos niveles de fondo debido a la hibridación cruzada [18] y un rango dinámico limitado de detección debido al fondo y la saturación de las señales. Además, el proceso de análisis y comparación de los niveles de expresión a través de diversos experimentos puede ser difícil y necesitar métodos de normalización complicados.

Por otro lado, entre los métodos basados en secuenciación se halla el RNA_Seq (*RNA sequencing*), el cual utiliza tecnologías de secuenciación profunda. En general, una población de ARN (total o fraccionada) es convertida en una biblioteca de fragmentos de ADNc con adaptadores adjuntos a uno o ambos extremos. Cada molécula, con o sin amplificación, es luego secuenciada con alto rendimiento para obtener secuencias cortas de uno solo o de ambos extremos. Las secuencias suelen ser de 30 a 400 bp, dependiendo de la tecnología de secuenciación usada [15].

1.1.4 Red biológica

Las redes biológicas se describen como representaciones de sistemas biológicos en forma de grafos, en las que los nodos simbolizan entes biológicos (por ejemplo, moléculas); y las aristas, las interacciones entre ellas [19]. Estas redes relacionan genes, productos de genes, proteínas, familias de proteínas, etc., que interactúan de forma coordinada en un proceso biológico específico [20].

Existen tres tipos principales de redes biológicas que se encuentran relacionadas con el

presente estudio: redes de regulación genética, de transducción de señales y metabólicas.

- Redes de regulación genética

Están conformadas por ARN, proteínas, ADN, entre otras moléculas que se regulan entre sí a través de diversos mecanismos. Al regularse, se establecen cuáles genes se expresan y cuáles no, en respuesta a factores ambientales de la célula [20]. La regulación de genes tiene lugar gracias a la participación de factores de transcripción. Éstos son proteínas que se unen a los sitios de regulación de los genes y pueden activarlos o inhibirlos. Dentro de este tipo de red, se encuentran: red de regulación de genes, red de co-expresión y red de regulación transcripcional.

Una *red de regulación de genes* se puede representar como un grafo mixto compuesto por conexiones dirigidas y no dirigidas como se muestra en la ecuación (1.1). Las conexiones dirigidas representan relaciones causales entre las actividades de los genes y las conexiones no dirigidas representan asociaciones dinámicas entre las actividades de los genes debido a variables ocultas (metabolitos, proteínas, etc.). Una red de regulación de genes describe la comunicación entre los genes y la regulación celular completa presentando las relaciones entre las actividades de los genes. Entre algunos ejemplos de relaciones de genes pueden estar los casos cuando un gen A codifica un factor de transcripción que regula a un gen B y cuando una proteína A podría modificar la tasa de degradación de ARN del gen A [21].

$$G \coloneqq (V, U, D) \tag{1.1}$$

donde G corresponde a un grafo, V a los nodos (genes), U a las conexiones no dirigidas y D a las conexiones dirigidas.

La *red de co-expresión* se infiere a partir de datos de expresión, al igual que la red de regulación de genes, basada en perfiles de expresión similares. Esta red contiene flechas no dirigidas, las cuales representan las asociaciones significativas de las actividades de los genes determinadas por su medida de expresión. Además, las actividades de los genes pueden estar correlacionadas por efectos directos y efectos indirectos, teniendo en cuenta que la correlación no implica causalidad y sí transitividad [21].

La *red de regulación transcripcional* presenta un enfoque mecanicista contemplado dentro de la biología molecular, relacionado con la regulación de los genes a través de transcripción. Las redes de regulación transcripcional son inferidas directamente a partir de resultados experimentales y predominantemente de datos de chip de ADN. Las conexiones entre los genes son únicamente dirigidas y corresponden a la unión física del producto proteínico del gen origen en la región promotora del gen destino, teniendo en cuenta que todos los genes origen codifican factores de transcripción [21].

- Redes de transducción de señales

Las redes de transducción de señales representan un conjunto de pasos encadenados para permitir que una célula pase una señal o estímulo en otro. Están conformadas por biomoléculas que tienen diferentes tipos de interacciones. Además, incluyen procesos relacionados con la bioquímica de una célula [6] y exploran la actividad de los genes y relaciones causa-efecto entre genes y proteínas bajo diferentes condiciones ambientales.

- Redes metabólicas

Estas redes establecen la base para la acumulación de biomoléculas en organismos vivos e integran la transferencia de información, producción de energía, especificación célula-destino, generación de masa y reacciones celulares [22].

1.2 Métodos computacionales

En esta sección, se presentan los conceptos y métodos computacionales fundamentales relacionados con la presente investigación.

1.2.1 Medida de distancia

Las medidas de distancia determinan qué tan cercanos son dos objetos. Son definidas en [23] como funciones que toman dos puntos en el espacio y calculan un número real que satisface los siguientes axiomas:

1. No existen distancias negativas entre dos puntos.
2. Las distancias entre dos puntos son positivas, excepto para la distancia desde un punto hacia sí mismo.
3. La distancia es simétrica.
4. Desigualdad triangular.

Existen diversas medidas de distancia para determinar la similitud entre objetos; por ejemplo, se encuentran la distancia euclidiana y la Minkowski.

La distancia euclidiana es la distancia más conocida y se encuentra dentro de un espacio euclidiano n-dimensional, donde los puntos son vectores de n números reales. Es referenciada como distancia de norma-L_2 [23] y representada de acuerdo a la ecuación (1.2), en donde se ve que esta distancia es la raíz cuadrada de la suma de los cuadrados de las distancias entre los puntos en cada dimensión.

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1.2)$$

La *distancia Minkowski* es la generalización de la distancia euclidiana. Aquí se define el parámetro con el cual se representa el denominado orden, por lo que la distancia euclidiana es un caso particular de la distancia de Minkowski en el cual el grado es 2 [REF14]. Es referenciada como distancia de norma-L_r [23]. En la ecuación (1.3) se representa esta distancia.

$$d([x_1, x_2, \dots, x_n], [y_1, y_2, \dots, y_n]) = \left(\sum_{i=1}^{n} |x_i - y_i|^r\right)^{1/r} \quad (1.3)$$

donde r es una constante.

1.2.2 Agrupamiento

El término agrupamiento es definido en [24] como el proceso de examinar una colección de datos y agruparlos de acuerdo a alguna forma de comparación entre los elementos (por ejemplo, una medida de distancia). En este proceso se busca que los datos en el mismo grupo tengan características similares.

Los algoritmos de agrupamiento pertenecen al grupo de técnicas de clasificación no supervisada y están diseñados para agrupar datos en un conjunto de categorías o grupos

en donde se encuentran juntos los datos que tengan características o patrones similares[23].

Existe una gran cantidad de algoritmos de agrupamiento, por lo que a continuación se presentan solo algunos que posteriormente serán empleados en el desarrollo de esta investigación.

- *K-means*

Algoritmo no jerárquico, es el más conocido con base en una medida de distancia. Este algoritmo divide los datos de observación en *k* grupos (con *k* definido previamente), ubicando cada objeto en un grupo con base en la distancia al centroide de dicho grupo. A continuación se describen los pasos que sigue [23].

1. Selecciona *k* objetos de forma aleatoria o con base en conocimiento a priori.
2. Asigna los *k* objetos seleccionados como centroides de los grupos.
3. Ubica cada objeto del conjunto de datos al grupo con el que tenga el centroide más cercano.
4. Recalcula los centroides de cada grupo.
5. Distribuye todos los objetos según el centroide más cercano.
6. Repite los pasos 4 y 5 hasta que no haya cambios en los grupos.

- *Fuzzy k-means*

Fuzzy k-means es la versión difusa del algoritmo *k-means* que no emplea actualizaciones incrementales de los centroides de los grupos. También es conocido como *c-means* [25].

Este algoritmo lleva a cabo en los siguientes pasos:

1. Seleccionar una seudopartición (partición no excluyente de los genes) difusa inicial.
2. Calcular el centroide de cada grupo empleando la seudopartición difusa.
3. Recalcular la seudopartición difusa.
4. Repetir los pasos 2 al 3 hasta que los centroides en cada grupo no cambien.

Con este algoritmo se construyen grupos difusos, cuyos componentes pueden pertenecer a más de un grupo, en donde en cada uno tiene un grado de membresía en el rango de 0 a 1.

1.3 Métodos de kernel

Los métodos de *kernel* son considerados herramientas para obtener relaciones no lineales en los datos, a través de un mapeo no lineal embebido en un espacio vectorial llamado *espacio de características*. Este mapeo se realiza por medio de una función denominada *función kernel.* Esta función estará relacionada directamente con el tipo de dato analizado y el dominio de conocimiento involucrado [4]. Los métodos de *kernel* son cada día más utilizados para resolver diversos problemas en múltiples áreas del conocimiento. Además, proporcionan las herramientas necesarias para procesar, analizar y comparar muchos tipos de datos de diferente naturaleza [4].

Estos métodos tienen como base los *kernels*, los cuales inducen una medida de similitud que surge a partir de una representación de los patrones inmersos en los datos analizados [26]. A la vez, son considerados productos punto en un espacio de características.

Para establecer la medida de similitud, se emplea una función *kernel*, la cual es simétrica y se representa en la ecuación (1.4).

$$\begin{aligned} K\colon\ & \chi \times \chi \ \rightarrow \ \mathbb{R} \\ & (x, x') \ \mapsto \ k\,(x, x') \end{aligned} \tag{1.4}$$

Existe una variedad de funciones *kernel* previamente establecidas. A continuación, se describen algunas de estas funciones y, entre ellas, las que se usaron en este trabajo.

- *Kernel* gaussiano

El *kernel gaussiano* es una función decreciente de la distancia euclidiana. Es representado en la ecuación (1.5) [4].

$$k_G(x, x') = exp\left(-\frac{\delta(x,x')^2}{2\sigma^2}\right) \quad (1.5)$$

donde σ es un parámetro y δ es la distancia euclidiana.

- *Kernel* polinomial

El *kernel polinomial* es un *kernel* vectorial de grado d mayor a cero. Es representado en la ecuación (1.6).

$$k_{Poly}(x, x') = (x^T\ x' + \ c)^d \quad (1.6)$$

donde c es una constante.

Este *kernel* corresponde al espacio de características conformado por todos los productos internos de más de d variables. Cuando el valor de la constante c es cero, el *kernel* representará el espacio de características conformado por todos los productos internos de exactamente d variables [4].

- *Kernel* coseno

El *kernel coseno* normaliza el *kernel* lineal a través de la norma de dos vectores factoriales totales. Su poder radica en que la normalización que realiza no se puede obtener con métodos lineales clásicos [27]. Es representado en la ecuación (1.7).

$$k(w_1, w_2) = \frac{<w_1, w_2>}{||w_1||\ ||w_2||} \quad (1.7)$$

donde w_1 y w_2 son vectores.

1.4 Bases de datos biológicas

Las bases de datos biológicas son colecciones de datos biológicos organizados y actualizados, los cuales pueden ser consultados fácilmente por los usuarios. Estas bases se clasifican en primarias, secundarias y compuestas.

Las bases de datos biológicas primarias contienen la información de estructuras o secuencias. Las secundarias contienen información relacionada con las bases de datos primarias; por ejemplo, secuencias conservadas y residuos de sitios activos. Las bases

compuestas contienen información de las bases de datos primarias, la cual puede ser consultada a través de diferentes criterios de búsqueda; es decir, se puede buscar específicamente por un gen, proteína, función, etc. [28].

A continuación, se presentan algunos ejemplos de bases de datos biológicas empleadas en el desarrollo de este trabajo.

1.4.1 NCBI

El NCBI (*National Center for Biotechnology Information*, Centro Nacional de Información Biotecnológica de Estados Unidos) tiene una metabase de datos biológica que proporciona información biomédica y genómica, y almacena otras bases de datos de nucleótidos y proteínas [28]. Se puede consultar gratuitamente en la dirección web *http://www.ncbi.nlm.nih.gov*.

1.4.2 KEGG

KEGG (*Kyoto Encyclopedia of Genes and Genomes*, Enciclopedia de Genes y Genomas de Kioto) es una metabase de datos. Fue iniciada por el proyecto japonés del genoma humano. Es considerada una red de bases de datos y servicios computacionales. Permite realizar investigaciones en genómica y áreas relacionadas. En KEGG se puede encontrar información acerca de genes, proteínas, funciones de los genes, redes biológicas, diagramas de las redes, entre otras [29]. Se puede acceder a través de la dirección web *http://www.genome.jp/kegg.*

1.4.3 GO

GO (*Gene Ontology*) es una base de datos que ha buscado estandarizar la representación de los genes y los productos de los mismos. Para ello, ha creado un vocabulario controlado y dinámico que puede ser aplicado a todos los organismos eucariotas [30]. A la vez, contiene la anotación de los productos de los genes y herramientas para la búsqueda y procesamiento de los datos. Además, contiene tres ontologías independientes: procesos biológicos, funciones moleculares y componentes celulares, las cuales presentan conexiones entre cada una [31]. Se puede acceder a esta herramienta en la dirección web *http://www.geneontology.org*.

1.4.4 STRING

STRING es una base de datos y recurso web que contiene información acerca de interacciones directas e indirectas entre proteínas. Al 16 de enero de 2014 contenía 5.214.234 proteínas de 1133 organismos. La información de las interacciones es derivada de diversas fuentes como análisis de información genómica, co-expresión, información reportada en la literatura y en otras bases de datos. Por otro lado, ofrece un visualizador gráfico de cada una de las redes de interacción entre los genes/proteínas construidas [32]. Este recurso se puede consultar en la dirección web *http://string-db.org*.

2. Trabajo previo

Con el fin de resolver el problema del análisis de datos de expresión y la identificación de relaciones entre genes a partir de este tipo de datos, se han empleado y desarrollado diversas técnicas y/o métodos. Desde el punto de vista computacional, a continuación se presentará una revisión de algunos de estos métodos y se describirá la forma en la que han sido aplicados sobre datos de expresión, tanto para el análisis como para la identificación de relaciones y validación de las mismas.

2.1 Métodos empleados para encontrar relaciones entre genes

2.1.1 Biología molecular

La biología molecular es “el estudio de la estructura, función y composición de las moléculas biológicamente importantes” [1]. Está caracterizada por generar, gracias a sus técnicas de alto rendimiento, un gran conjunto de volúmenes de datos con diferentes tipos como secuencias, estructuras, interacciones, localización, expresión, etc. Estos datos se han obtenido a través de diversas tecnologías como secuenciación e hibridación.

Un ejemplo de aplicación de esta área se encuentra en la obtención de datos de expresión de genes (cantidad de ARN mensajero). Este tipo de datos se representan como una matriz que permite visualizar en cada fila a los genes y su nivel de expresión para cada una de las condiciones experimentales analizadas. De este modo, existen tecnologías con base en hibridación y secuenciación que permiten obtener este tipo de datos. Con base en hibridación se encuentran los microarreglos, los cuales son placas que contienen pequeños pozos en donde se pueden evaluar grandes cantidades de genes en corto tiempo [33]. Esta es una tecnología que ha tenido gran auge, debido a que permite obtener

expresión de miles de genes al mismo tiempo. Además, ha generado el interés en el estudio y desarrollo de herramientas para el análisis de perfiles de transcripción de gran escala. Se han diseñado, con base en lo anterior, nuevas formas de realizar el diagnóstico y el tratamiento de pacientes con alguna enfermedad [24]. Otra tecnología empleada para obtener datos de expresión con mayor precisión y con base en secuenciación es conocida como RNA_Seq. Esta técnica se ha desarrollado y utilizado para la generación de perfiles de transcripción y estudio de expresión de los genes, la cual *"utiliza tecnologías de secuenciación profunda y proporciona una medición mucho más precisa de los niveles de las transcripciones y sus isoformas*" [15].

2.1.2 Minería de datos

La minería de datos es definida como el proceso de descubrir modelos a partir de datos [23] y utiliza métodos de inteligencia computacional para extraer conocimiento. En el proceso de minería de datos, se encuentran diferentes métodos para analizar la información contenida en datos provenientes de técnicas mencionadas en la sección anterior, como los microarreglos y RNA_Seq. En particular, se pueden emplear para el descubrimiento de patrones y para técnicas de reducción de dimensión, como análisis de componentes principales (ACP) y agrupamiento.

- Agrupamiento

Los algoritmos de agrupamiento incluyen métodos no supervisados para la organización de datos multivariados en grupos con patrones similares. Es decir, identifican diferencias o características similares entre los elementos de un mismo conjunto de datos, para luego dividirlo en grupos de acuerdo a las diferencias o relaciones encontradas [24] [34].

La aplicación de algoritmos de agrupamiento sobre datos de expresión de genes se han aplicado en diversos tipos de análisis, tales como asociación e identificación de funciones de genes anteriormente desconocidas e identificación de nuevos genes asociados a enfermedades [24].

A partir de la identificación de patrones, es decir, después del proceso de encontrar y caracterizar relaciones generales en un conjunto de datos de expresión y clases relacionadas con éstos, se pueden inferir redes de regulación génica. Según Donna K.

Slonim [24], si se busca específicamente información sobre las interacciones de genes indicados por expresión de datos se pueden sugerir nuevas redes y asociaciones.

Higgs B. W., Elashoff M., Richman S. y Barci B. [35] utilizaron microarreglos y agrupamiento enfocados hacia el estudio de enfermedades cerebrales humanas. Para esto, realizaron la comparación de diferentes plataformas de microarreglos e identificaron en éstas diferencias en sensibilidad y escalabilidad. Además, identificaron en cada enfermedad (esquizofrenia, desorden bipolar y depresión) los mecanismos biológicos asociados a cada una.

Entre otros estudios relacionados se encuentran [36] [37] [38] [39], en los que se realizó agrupamiento con base en datos de expresión de genes de organismos como la levadura (ampliamente estudiado) y se usó diversas medidas de distancia y algoritmos para este fin. Además, al emplear técnicas de agrupamiento sobre estos datos, han podido descubrir patrones interesantes en los genes estudiados como, por ejemplo, porcentajes significativos de motivos comunes en las secuencias de los genes.

- Análisis de componentes principales (ACP)

El análisis de componentes principales (ACP) es una forma de identificar patrones en los datos y, con ello, las semejanzas o diferencias entre los elementos de cada conjunto de datos. Para esto realiza una transformación lineal de los datos a través de la reducción de dimensiones sin perder demasiada información. Igualmente, realiza diversos pasos aritméticos como calcular la media de los datos, la matriz de covarianza y los vectores y valores propios de esa matriz; a partir de estos pasos, llega a la construcción de los datos finales conformados por un vector de características representativas de la información original [40].

El ACP ha sido empleado sobre datos de expresión de genes como una herramienta para extraer la información representativa de los conjuntos de datos, reduciendo las dimensiones que pueden no aportar información útil al estudio. También ha sido aplicado antes de emplear algún algoritmo de agrupamiento sobre los datos, como en el caso presentado en [41]. Allí se encuentra un ejemplo de este tipo de aplicaciones de ACP en el que fue utilizado para encontrar patrones correlacionados e interdependientes en la expresión de los genes que conforman rutas metabólicas humanas y que, al ser combinado

con algoritmos de agrupamiento, permitió identificar los genes que están relacionados con algunos tipos de tumores.

2.1.3 Aprendizaje de máquina

El aprendizaje de máquina se basa en conceptos y métodos de muchas áreas como estadística, inteligencia artificial, filosofía, teoría de la información, biología y la teoría de control. Su objetivo es desarrollar algoritmos o técnicas para extraer conocimiento a partir de datos [42].

Dentro de las técnicas enmarcadas dentro del aprendizaje de máquina se encuentran los métodos de *kernel* [4] [26], los cuales han sido empleados sobre datos biológicos en diversos problemas como identificación de la estructura de las proteínas, predicción de la función de los genes, rol de los genes en enfermedades, etc. [43] [26]

Adicionalmente, han sido empleados para la integración de datos biológicos heterogéneos [44] [45] para realizar el análisis de los datos. Por ejemplo en [44], emplean métodos de *kernel* para realizar la integración de información química y genómica relacionada con las proteínas humanas; allí, identifican las redes de interacción entre las proteínas que específicamente estaban relacionadas con receptores nucleares, canales iónicos, GPCR (*G-protein-coupled receptors*) y enzimas, y construyen la red de interacción entre dichas proteínas.

3. Materiales y metodología

En este capítulo se describen los conjuntos de datos utilizados en esta investigación y la metodología propuesta para la identificación y análisis de relaciones entre genes.

3.1 Materiales

En esta sección se detallan cada uno de los conjuntos de datos empleados, tanto datos de expresión como categóricos, que fueron utilizados durante el desarrollo del presente trabajo.

3.1.1 Conjuntos de datos

Los datos de expresión de genes se obtienen con técnicas de biología molecular como microarreglos o RNA_Seq, entre otras, a través del estudio de los genes del organismo de interés en diversas condiciones ambientales o estados de desarrollo, y pueden ser tomados sobre diversos tejidos del organismo. Los datos de expresión génica se representan como una matriz, en donde cada fila indica el nombre del gen y cuyos valores de expresión para cada uno de los experimentos o condiciones están representados en cada una de las columnas. La cantidad de expresión es una variable continua que indica la cantidad de ARN mensajero producido por cada gen en una condición dada.

Para la experimentación de este trabajo se utilizaron datos de expresión génica obtenidos con experimentos sobre la planta *Arabidopsis thaliana* empleando dos técnicas diferentes: microarreglos y RNA_Seq. Además, se trabajó con conjuntos de datos categóricos que también describían a cada uno de los genes contemplando factores de transcripción, rutas biológicas y funciones de los genes.

- Microarreglos

Los datos de microarreglos se encontraron en la base de datos GEO-datasets del NCBI (*National Center for Biotechnology Information*). Los datos están relacionados con experimentos de patogenicidad (resistencia de la planta a patógenos).

Estos datos están conformados a su vez por dos conjuntos de datos diferentes. El primer conjunto es denominado microarreglos_1, consiste de 278 experimentos y 22.171 genes, en donde los 278 experimentos contienen experimentos originales y réplicas de los mismos. En el anexo A se encuentran los identificadores de acceso de NCBI.

A continuación, se describen algunos de estos experimentos:

- Estudio de los cambios de la transcripción en *Arabidopsis thaliana* hacia la resistencia a la penetración del hongo biotrópico *Blumeriagraminis* (BGH); se compararon muestras de *Arabidopsis thaliana* de tipo silvestre sin inocular e inoculadas con un alelo mutante ATAF1, comprometido en la resistencia a la penetración. El experimento incluyó el muestreo de 8 rosetas para cada muestra replicada en plantas de 6 semanas de edad, 12 horas después de la inoculación.
- Estudio de la resistencia de la planta a la sequía a través de la expresión de los genes NFYA5, los cuales son regulados con estrés hídrico, no sólo a nivel transcripcional sino también a nivel postranscripcional.
- Estudio de la quinasa LysM como receptor, mediando la percepción de quitina y resistencia a los hongos en la planta.
- Estudio del impacto de los efectores de tipo III en las respuestas de defensa de la planta, centrándose en cómo las plantas responden a los patógenos bacterianos, cambios de la planta con los fitopatógenos en terobacterias *P. Syringae.*
- Estudio de la respuesta sistémica a la infección bacteriana observando la respuesta local de la planta a cambios bióticos en redes de defensa basal a través del reconocimiento y respuesta a patógenos conservados asociados a patrones moleculares (PAMPs, por sus iniciales en inglés).

El segundo conjunto de datos es denominado microarreglos_2 y consiste de 24 experimentos y 22.810 genes, en donde los 24 experimentos contienen experimentos

originales y réplicas de los mismos. Este conjunto de datos está relacionado específicamente con la respuesta del mutante pen3 a la invasión del hongo *hordei*. En el anexo A se encuentran los identificadores de acceso de NCBI.

- RNA_Seq

Los datos de RNA_Seq están relacionados con experimentos de patogenicidad realizados en la planta *Arabidopsis thaliana* y consisten de 12 experimentos y 23.517 genes.Estos datos fueron obtenidos a partir de secuenciación de librerías de ADNc con la tecnología Illumina del Departamento de Genética Molecular de la Universidad Estatal de Ohio. Para construir las librerías se extrajeron y secuenciaron muestras de ARN de dos tipos de plantas, con igual número de tratamientos, cada uno a tres tiempos. Se tenían plantas silvestres de *Arabidopsis thaliana* Col-0 y plantas mutantes que carecen del gen de resistencia RPS4. Las plantas fueron infectadas con la bacteria Pph o inyectadas con búfer salino como control negativo. Se recolectó tejido a 6 h, 12 h y 24 h postinoculación.

Las lecturas resultantes de la secuenciación fueron mapeadas sobre 33.518 genes de *Arabidopsis thaliana* y los valores fueron normalizados utilizando la medida RPKM (*reads per kilobase per million reads*) [46] utilizando los programas de *software* R-seq y seq-map [47].

- Factores de transcripción (FT)

La información proporcionada por los factores de transcripción asociados al organismo de estudio se utilizan en el análisis de datos de expresión teniendo en cuenta que el proceso de regulación de la expresión génica está relacionado con los factores de transcripción, debido a que estos se unen a la región regulatoria de los genes produciendo su activación o inhibición [48]. Además, es de interés estudiar los genes que compartan el mismo o los mismos factores de transcripción porque pueden estar asociados a los mismos procesos biológicos.

Los datos de factores de transcripción se obtuvieron en la base de datos pública *AtTFDB* del Arabidopsis Gene Regulatory Information Server (AGRIS), de la Universidad Estatal de Ohio [49]. Estos datos corresponden a los factores de transcripción de los genes de la planta

Arabidopsis thaliana. Para la identificación de estos factores, realizaron la combinación de BLAST y búsqueda de motivos con base en información disponible en la literatura entre factores de transcripción conocidos o motivos conservados entre factores de transcripción de una familia [3].

- Rutas biológicas (KOF)

La información de rutas biológicas se considera importante en el análisis de datos de expresión debido a la relación que existe entre gen, ruta biológica y expresión génica. Esta relación existe en la medida en que los genes que están presentes en diversas rutas biológicas pueden presentar patrones de co-expresión y estar influenciados por diferentes mecanismos de regulación [50]. En consecuencia, se espera que los genes que pertenecen a las mismas rutas biológicas se expresen de forma similar y que los genes que no compartan las mismas rutas tengan perfiles de expresión diferente.

En la base de datos KEGG (*Kyoto Encyclopedia of Genes and Genomes*) [51] [52] se obtuvo información acerca de las rutas biológicas en las que están involucrados los genes de la planta *Arabidopsis thaliana* y con esta información se construyeron los datos denominados KOF, los cuales consisten en la asociación de cada gen a las rutas de KEGG. Con esto, se obtuvo la asociación de los genes a una o varias de las 953 rutas encontradas.

- Funciones biológicas (GOF)

Los productos de los genes pueden estar involucrados en diferentes categorías biológicas como procesos biológicos, funciones moleculares o tener asociado un componente celular específico; cada una de estas categorías están consolidadas en la base de datos de GO (*Gene Ontology*) y se denominan *ontologías* [31] [30]. Estas ontologías son relevantes al momento de analizar datos de expresión génica, porque representan información complementaria al ser atributos de los genes o de los productos de los genes, lo cual permite realizar un mejor análisis integrando y/o comparando los datos tanto de la expresión como de la(s) categoría(s) de un gen.

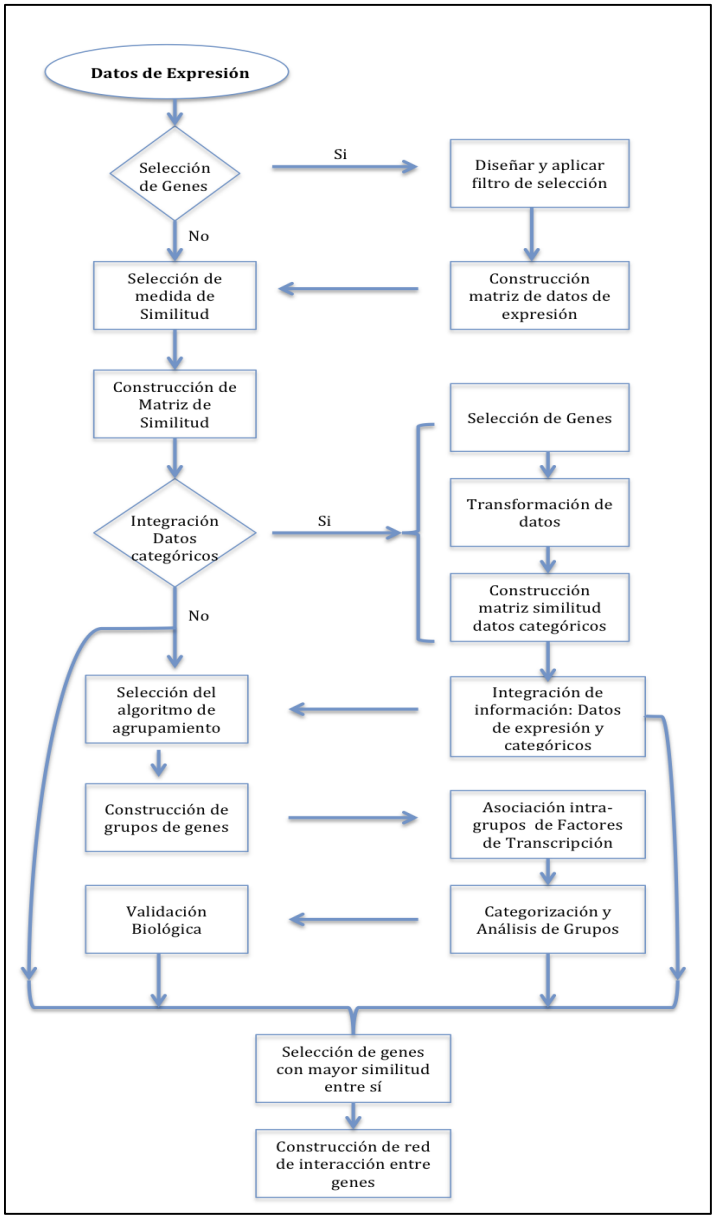
De esta forma, en este estudio se obtuvo información acerca de las ontologías en las que están involucrados los genes de la planta *Arabidopsis thaliana.* Con esta información se

construyeron los datos denominados GOF (Gene Ontology Functions), los cuales consisten en la asociación de cada gen a algunas de las 741 categorías de GO encontradas.

3.2 Metodología

La metodología propuesta y seguida en esta investigación se presenta en la Figura 3-1.

Figura 3-1: Metodología propuesta



3.2.1 Selección de genes

Para comenzar a trabajar con los conjuntos de datos de expresión de microarreglos y RNA_Seq se analizó el número de genes en cada uno y la expresión de los mismos. En caso de que se desee realizar un proceso de selección de información, es importante filtrar aquellos genes cuyo comportamiento sea más significativo para el análisis que se quiera hacero de acuerdo a las limitaciones de recursos computacionales que se tengan. Por lo tanto, se propone emplear un filtro que consiste en seleccionar genes con base en el umbral de expresión. De esta forma, se seleccionan los genes que tengan la mayor (o menor) expresión en cada conjunto de datos de acuerdo al criterio de selección. En este estudio, se seleccionaron aquellos genes cuyo promedio de valor de expresión estuviera por encima del valor del tercer cuartil de expresión media de todos los genes en cada uno de los datos de microarreglos y de RNA_Seq.

3.2.2 Normalización de datos

Un tema importante en el análisis de datos de expresión es la normalización de los mismos. Este proceso es útil para poder ajustar el valor de expresión de los genes, debido a que existe la posibilidad de un desequilibrio en la intensidad de las muestras de ARN, lo cual no hace comparables los datos. Este desequilibrio es producido por diversos motivos técnicos, como diferencias en el ajuste de la tensión de desequilibrio PMT, cantidad total de ARN disponible en cada muestra, etc. [53]

Por consiguiente, para llevar a cabo el proceso de normalización sobre los datos de microarreglos y de RNA_Seq, en este estudio se emplea el método propuesto por *Wolfgang Huber*, el cual consiste en una variación del método del estimador de máxima probabilidad, realizando una transformación sobre la medida de los niveles de expresión y una diferencia estadística, en donde la varianza es aproximadamente constante a lo largo de todo el rango de los valores de expresión [7] [8].

Específicamente, para la experimentación de este paso, se trabajó con la librería *vsn* de *R Project,* la cual implementa el método de Huber, siguiendo los comandos:

```
library(Biobase)
library(vsn)
matExp<-as.matrix(exp) #matExp=matriz de expresión
expNorm<- justvsn(matExp) #normalización
meanSdPlot(expNorm) #graficar la normalización
```

3.2.3 Construcción de la matriz de similitud

- Selección de medida de similitud

Existen diversas medidas de similitud que, aplicadas sobre datos de expresión génica, sirven para determinar la relación de semejanza entre dos o más perfiles de expresión de los genes. Esta similitud puede implicar correlaciones positivas, positivas y negativas e información mutua. Cada una de estas medidas ofrece información acerca de la regulación de los genes, por lo que no es una medida fácil de seleccionar y es por esto que no existe en la actualidad una única y mejor medida de similitud para datos de expresión génica [54]. Debido a esto, se deja libre la selección siempre y cuando se tenga en cuenta el proceso de normalización de los datos que sea necesario para aplicar la medida seleccionada. Por ejemplo, en [55] emplean la distancia euclidiana como medida de comparación entre los perfiles de expresión de los genes, planteando que se debe realizar la normalización con respecto al máximo nivel de expresión de cada gen teniendo en cuenta también el valor mínimo; otra forma es con respecto al valor de la media o la desviación estándar de cada perfil. A la vez, también indican que si los perfiles de expresión tienen datos relativos se deben normalizar con base en el logaritmo de los valores de expresión relativos.

En este trabajo se propone el empleo de métodos de *kernel* para la construcción de la matriz de similitud. De este modo, se selecciona una función de *kernel* como medida de similitud entre los genes. Los métodos de *kernel* son técnicas enmarcadas dentro del área de *aprendizaje de máquina, "están basados en diferentes representaciones del conjunto de datos estudiado como una matriz de similitud entre sus elementos"* [4] y son útiles para procesar, analizar y comparar diversos tipos de datos [9]. De esta forma, a través de la selección de una función *kernel* se mapean los genes en un espacio vectorial R^n (espacio de Hilbert) denominado espacio de características [9] [56], en donde se miden distancias

entre ellos, calculando la similitud entre cada par de genes a través del cálculo del producto punto entre cada vector proyectado.

- Construcción matriz de similitud

Los métodos de *kernel* son empleados en este trabajo para la construcción de la matriz de similitud a partir de los conjuntos de datos de expresión. De esta forma, se seleccionó una función *kernel* para cada conjunto y a partir de esta función se construyó la matriz *kernel* de tamaño $n \times n$, con n como el número de genes, la cual representa la similitud entre cada gen de los datos de expresión.

De este modo, se empleó un *kernel gaussiano* representado en la ecuación (1.5) sobre los datos de microarreglos y de RNA_Seq. Para la selección del parámetro *sigma* (σ) se generaron diferentes valores aleatorios entre 0 y 1 y se se siguió un procedimiento heurístico, el cuál consistió en:

a. Definir los parámetros para el *kernel*
b. Construir la matriz *kernel*
c. Emplear un *kernel* PCA sobre la matriz construida
d. Graficar la proyección de los dos primeros componentes principales
e. Analizar la dispersión de los datos en la gráfica
f. Incrementar el parámetro y repetir el procedimiento (en caso de querer mejorar la dispersión de los datos)

Específicamente, para la construcción de las matrices *kernel* empleando el *kernel* gaussiano, se trabajó con la librería *kernlab* de *R Project,* ejecutando los siguientes comandos con cada uno de los conjuntos de microarreglos y de RNA_Seq:

```
library(kernlab)
matExp<-as.matrix(expDataSet) #matriz con los datos de expresión
# kernel gaussiano. Sig= parámetro sigma
rbfkernel <- rbfdot(sigma = sig)
matKernel<-kernelMatrix(rbfkernel, matExp) #matriz kernel
```

Adicionalmente, se construyeron matrices *kernel* de los datos de expresión empleando el *kernel coseno* representado en la ecuación (1.7), con el objetivo de tener dos medidas basadas en métodos de *kernel* aplicadas a datos biológicos y que serían aplicables como medidas de similitud entre dos genes.

Para poder emplear el *kernel* coseno sobre los datos de expresión, se desarrolló un algoritmo en R que implementa la ecuación (1.7), el cual puede ser ejecutado con cualquier conjunto de datos de expresión.

- Análisis de componentes principales

Al construir matrices *kernel*, si el número de dimensiones de la matriz original aumenta, es decir, si el número de genes con el que se esté trabajando se incrementa, el costo de su procesamiento también lo hará. Por ello, es necesario utilizar técnicas que permitan reducir el número de dimensiones de estas matrices antes de realizar los análisis.

Una técnica para realizar esta reducción es conocida como el análisis de componentes principales (ACP), el cual realiza una transformación lineal para diagonalizar y estimar la matriz de covarianza de los datos y proporciona un conjunto de ejes ortogonales llamados componentes principales, permitiendo describir la mayor parte de los datos con sólo los primeros ejes, aquellos que poseen la varianza más alta en el nuevo espacio [57].

Una extensión del ACP es denominada *Kernel* ACP (KACP), la cual se empleó en este trabajo para obtener los componentes principales de cada matriz y reducir el número de dimensiones de las matrices *kernel* construidas. El KACP transforma los datos realizando un mapeo no lineal en un espacio de mayor número de dimensiones (espacio de Hilbert) que no está relacionado linealmente con el espacio de entrada [57]. El KACP es empleado en vez del ACP clásico con el objetivo de obtener los componentes principales de una forma no lineal, encontrando información que con el ACP clásico podría no hallarse.

En particular, se empleó en este trabajo la librería *kernlab* de *R Project* para la extracción de los componentes principales con el método KACP. Se ejecutaron los siguientes comandos con cada una de las matrices *kernel* de los datos de expresión (microarreglos y RNA_seq):

```
library(kernlab)
kpc<-kpca(matKernel, features=2) #KPCA sobre la matriz kernel
```

3.2.4 Datos categóricos

Con el fin de obtener mayor información acerca de los genes estudiados y aumentar los criterios de selección de similitud entre ellos, se considera importante la inclusión de otros datos relacionados con los genes. En este estudio se utilizaron, además de los datos de expresión con datos de categorías de los genes, datos de rutas biológicas obtenidas de la base de datos KEGG (datos KOF), funciones biológicas obtenidas de la base de datos *GO* (datos GOF) y factores de transcripción obtenidas de la base de datos *AtTFDB* (datos FT).

Para emplear los datos categóricos como información adicional a los datos de expresión es necesario realizar una selección de los genes y transformación de la información contenida en ellos.

- Selección de genes comunes a todos los tipos de datos

Con el objetivo de realizar la asociación de la información proporcionada por los datos de expresión y categóricos, los genes en cada conjunto categórico —datos KOF, datos GOF y datos FT— deben corresponder a los genes de los datos de expresión. De esta forma, se identificaron los genes presentes en los datos de expresión luego, se buscaron estos genes en los datos categóricos, seleccionándolos y descartando los demás genes contenidos en los datos categóricos.

- Transformación de datos

Los datos categóricos KOF y GOF están compuestos por una lista de genes y una categoría asociada a cada uno de estos genes. Es decir, se tiene una relación uno a uno, por lo que se busca obtener mayor información a partir de estos datos realizando una transformación a los mismos, en la que se aumenta el espacio de categorías relacionadas con los genes, ampliándose las variables del conjunto del datos.

De este modo, se desarrolló un programa que realiza el proceso de la transformación de los datos, el cual consiste en identificar todas las categorías inmersas en los datos y crear

una columna por cada una. Posteriormente, se analiza cada fila y se le asigna el valor de uno (1) a la categoría que esté relacionada con el gen y cero (0) en caso contrario.

3.2.5 Construcción matriz de similitud datos categóricos

- Selección de la medida de similitud

Para encontrar la similitud entre los genes de los conjuntos de datos categóricos, se emplea, de la misma forma que para los datos de expresión, métodos de *kernel* para la construcción de la matriz de similitud.

- Construcción matriz de similitud

Para la construcción de la matriz *kernel* (matriz de similitud) en los conjuntos de datos categóricos, se seleccionó y empleó una función *kernel*: *kernel polinomial* definido en la ecuación (1.6) sobre los datos KOF, GOF y FT, y se seleccionó el valor del parámetro *grado del polinomio* de forma similar al valor sigma en el *kernel* gaussiano sobre los datos de expresión, generando valores aleatorios mayores a 1 y escogiendo aquel valor que otorgara la mejor distribución en los datos en el diagrama 2D.

De forma específica, para la construcción de las matrices *kernel* de los datos categóricos, se trabajó con la librería *kernlab* de *R Project,* ejecutando los siguientes comandos con cada uno de los conjuntos de datos KOF, GOF y FT:

```
library(kernlab)
matCateg<-as.matrix(categDataSet)#matriz de datos categóricos
#kernel polinomial. d=parámetro grado del polinomio:
polyKernel<-polydot(degree = d, scale = 1, offset = 0)
matKernel<-kernelMatrix(polyKernel,matCateg) #matriz kernel
```

- Análisis de componentes principales

Al igual que con las matrices *kernel* de datos de expresión se emplea el *Kernel* ACP (KACP) para obtener la información más representativa de cada matriz de datos categóricos y reducir el número de dimensiones de las matrices para realizar los análisis respectivos. Específicamente, se empleó la librería *kernlab* de *R Project,* ejecutando los

siguientes comandos, con cada una de las matrices kernel de los datos categóricos (KOF, GOF y FT):

```
library(kernlab)

datosKpca<-kpca(matKernel, features=2) #KACP sobre la matriz kernel
```

3.2.6 Integración de información – datos de expresión y categóricos

En algunas ocasiones el conocimiento proporcionado por los datos de expresión génica no aporta en su totalidad información significativa al estudio. Es decir, no ofrece la información necesaria para elucidar los patrones y características propias de los genes estudiados. Por consiguiente, se hace ineludible incluir en el análisis información de categorías a las que pertenecen o están relacionados los genes, teniendo en cuenta que el conocimiento a priori acerca de los genes puede proporcionar un peso mayor a la similitud proporcionada por solamente los datos de expresión [9]. Para ello, se creó una matriz *kernel* que incorporara información de expresión y de las categorías de los genes.

Por consiguiente, se desarrolló un algoritmo que realiza la combinación de *kernels* creando matrices *kernel* múltiples, las cuales permiten integrar datos heterogéneos. Una matriz *kernel* múltiple (k) es construida a partir de la suma de matrices *kernel* básicas ($k_1, k_2 \ldots, k_i$) de acuerdo a la ecuación (3.1). Esta matriz resultante es una matriz *kernel* definida por el producto interno de los espacios de características de los *kernels* básicos [9].

$$k = \sum_{i=1}^{c} k_i \tag{3.1}$$

con k_i como cada una de las matrices kernel.

En consecuencia, para la construcción de las matrices *kernel* múltiples en este trabajo, se toman como punto de partida las matrices *kernel* construidas para los datos de expresión y categóricos. Luego, se aplica el algoritmo desarrollado en donde la información de estas matrices se usa para llevar los espacios de cada una al mismo sistema de coordenadas y son combinadas asignándoles diferentes pesos (valores entre 0.1 y 0.9) a cada una al momento de la fusión de acuerdo a la ecuación (3.2), generando diversas matrices *kernel* con la información tanto de datos categóricos como de expresión.

$$k = \sum_{i=1}^{c} \mu_i k_i \quad (3.2)$$

con μ_i siendo el peso de cada *kernel* y k_i es cada *kernel*.

Para la asignación de los pesos cuando se integran las matrices kernel, estos se fueron variando dentro del algoritmo, construyendo la matriz *kernel* y analizando esta matriz resultante de acuerdo con:

a. La diagonal, la cual debe ser 1 o muy cercana a 1, y
b. La proyección de los dos primeros componentes principales calculados con el *kernel* PCA.

De esta forma, se escogen las mejores matrices para realizar los siguientes pasos de la metodología. Así mismo, se aprecia que el criterio de selección de los pesos es heurístico y se emplean métodos de *kernel* descriptivos, debido a que no se tiene un estándar pre-establecido lo suficientemente confiable con el cual contrastar.

3.2.7 Selección del algoritmo de agrupamiento

Los métodos de agrupamiento permiten encontrar relaciones entre las muestras de un conjunto de datos, realizando clasificación automática de las mismas en un número de grupos de acuerdo a una medida de similitud [54].

Estos métodos son aplicados a los conjuntos de datos estudiados en este trabajo con el objetivo de encontrar genes similares o con perfiles de expresión similar, a través de la formación de grupos. De esta forma, se pueden identificar genes con características similares dentro del mismo grupo y que tengan diferencias con genes de otros grupos.

- Selección del algoritmo de agrupamiento

Existen numerosos algoritmos de agrupamiento cuyo resultado en algunos casos estará relacionado con la medida de similitud seleccionada [54]. Se debe seleccionar una medida de similitud y luego ser utilizada en algún algoritmo de agrupamiento que la requiera, mientras que, para construir grupos de genes, es recomendable emplear más de un algoritmo de agrupamiento para obtener variedad en los grupos de genes y poder comparar y validar los patrones encontrados por cada algoritmo.

De esta forma, es útil emplear algoritmos que busquen relaciones lineales y no lineales en los datos de expresión génica. Además, es importante tener en cuenta para la construcción de grupos de genes, que un gen puede pertenecer a más de un grupo y no sólo a uno, influenciando la forma general de varios grupos [58].

En consecuencia, en este trabajo se seleccionaron los algoritmos *k-means y fuzzy k-means* para construir los grupos. Se seleccionó el algoritmo *k-means* por ser muy conocido, sencillo y de fácil implementación [33]. Además, ha sido ampliamente utilizado sobre datos de expresión. El algoritmo *fuzzy k-means* fue escogido dada su propiedad de asignación de un gen a varios grupos, lo que permite obtener grupos más diversos y no limitar la participación de un gen a un único grupo, generando pérdida de características o comportamientos particulares de los grupos a los que podría pertenecer y que podrían ser de interés para el caso de estudio.

3.2.8 Construcción de grupos de genes

Para la construcción de los grupos de genes se utilizaron los algoritmos seleccionados en el paso anterior: *k-means* y *fuzzy k-means* sobre el *kernel* ACP de cada una de las matrices *kernel* construidas de los datos de expresión y de los datos categóricos.

Cada algoritmo se ejecutó para diferentes valores de *k* según la heurística definida en la ecuación (3.3) para obtener un valor aproximado al número adecuado de grupos que se deberían formar para cada tipo de dato.

$$k \approx \sqrt{n/2} \tag{3.3}$$

donde n es el número de elementos del conjunto de datos.

Para construir los grupos, se trabajó con *R Project* ejecutando los siguientes comandos:

```
matrizDatos<- as.matrix(rotated(datosKpca))
#k-means:
classKM<-kmeans(matrizDatos,n_g,n_iter)

#fuzzy k-means:
classKM<-cmeans(matrizDatos, n_g,n_iter)
#n_g=número de grupos, n_iter= número de iteraciones
```

Adicionalmente, luego de ejecutar el algoritmo *fuzzy k-means* es recomendable seleccionar la participación de los genes en los grupos formados utilizando una métrica que consta en la determinación de un umbral. Para este umbral se propone calcular la mediana de los valores de membresía de un gen en cada uno de los grupos y, a partir de este valor, seleccionar aquellos grupos en donde el valor de membresía se encuentre por encima o sea igual al valor de la mediana. De esta forma, se asociarán a los grupos los genes con los valores de membresía más altos.

3.2.9 Asociación intra-grupos de factores de transcripción

Con el objetivo de buscar patrones en los grupos resultantes del paso anterior, es importante relacionarlos con otros datos biológicos para poder realizar un análisis más detallado de los mismos. Se propone utilizar datos de los factores de transcripción (FT) de los genes del organismo bajo estudio para buscar relaciones de regulación e identificar aquellos genes que se comportan de forma similar y, además, respondan al mismo factor de transcripción. En este paso, no es necesario que en los datos FT cada gen tenga asociado un factor de transcripción debido a que surgirán algunos genes con factores de transcripción desconocidos que estarán asociados a un grupo de genes con comportamiento de expresión similar y que podrían estar relacionados con los mismos sitios de regulación. En este sentido, en cada grupo formado se identifican los genes que lo integran, se busca y adiciona la información de los factores de transcripción a los que se encuentra asociado cada gen. Así, en cada grupo no sólo se va a tener información de los genes que lo componen sino también de los factores de transcripción de cada gen o de algunos de estos genes que se encuentren en el mismo grupo.

3.2.10 Categorización y análisis de grupos

Después de asociar los factores de transcripción a los genes de los grupos formados, se busca caracterizar o categorizar cada uno de los grupos. Específicamente, un grupo será definido por el número de genes, los genes, los factores de transcripción de dichos genes y los factores de transcripción más representativos en aquellos grupos. Un factor de transcripción es representativo si tiene el valor de frecuencia más alto en el grupo, es decir, que presenta el número mayor de genes asociados a él en el grupo.

Al realizar la asociación de los datos FT a los grupos, se identifican los factores de transcripción relevantes en cada grupo y se podrán elucidar aquellos que se encuentren relacionados con el caso de estudio. Además, se contará con la categorización de los grupos en la medida en la que cada uno estará asociado a unos factores de transcripción característicos y se establecerá cuál es el factor más significativo para el grupo y se identificará cuáles son los factores representativos con mayor frecuencia en todos los grupos.

3.2.11 Comparación datos biológicos

Al obtener diferentes grupos de genes y de factores de transcripción asociados a estos con cada uno de los tipos de datos, es necesario realizar una etapa de comparación con información biológica para determinar la precisión de la formación de los grupos, de los factores de transcripción que resultaron ser los más representativos en los grupos y encontrar procesos de regulación involucrados con los genes estudiados.

De este modo, se analizaron los factores de transcripción que tenían el mayor número de genes asociados en cada grupo y aquellos que presentaban la mayor frecuencia en todos los grupos, es decir, que surgieron como representativos en la mayoría de los grupos. En el análisis realizado se buscaron las características biológicas de cada uno de los factores representativos y se comprobó su relación con la respuesta de los genes a las condiciones experimentales de los datos de microarreglos y de RNA_Seq. En otras palabras, se hallaron las asociaciones de regulación entre los genes y sus correspondientes factores de transcripción.

En este sentido, teniendo en cuenta que la regulación de los genes no necesariamente está influenciada por un único factor de transcripción, se investigaron en las bases de datos biológicas y en la literatura las relaciones entre los factores de transcripción seleccionados en el paso anterior.

3.2.12 Selección de genes con mayor similitud entre sí

A partir de la construcción de las matrices *kernel* para cada tipo de dato, se encontraron relaciones entre la mayoría de los genes, motivo por el cual se buscó seleccionar las relaciones más fuertes entre los genes.

Para realizar esta selección se propone emplear diferentes métodos que permitan establecer las conexiones más fuertes entre los genes y eliminar aquellas conexiones que por su valor tan pequeño puedan ser descartadas y no ser relevantes en el estudio. De esta forma, en este trabajo se seleccionaron dos métodos: algoritmo de los k vecinos más y la determinación de un umbral a partir de la ley de transitividad empleando el coeficiente de agrupamiento.

El algoritmo de los *k* vecinos más cercanos es un clasificador basado en aprendizaje por analogía, comparando una tupla desconocida con tuplas de entrenamiento similares a esta [59]. En el presente trabajo se empleó este algoritmo no como clasificador sino como medida de selección de los *k* genes más similares a un gen particular. Así, se escogieron para cada gen aquellos genes (vecinos) cuya similitud tuviera los valores más altos. Para ello, se ordenaron los valores de cada fila de las matrices *kernel* de mayor a menor y posteriormente se seleccionaron los primeros k elementos de cada una. De este modo, se filtraron las relaciones entre genes, resultando cada gen relacionado con los *k* genes más similares. Posteriormente, se volvieron a construir las matrices manteniendo el valor original si corresponde a una relación de un par de genes con mayor similitud y asignando el valor de cero en caso contrario.

El método de la selección del umbral con base en la teoría de la transitividad hace referencia a la medida de la transitividad de un grafo conocida como *coeficiente de agrupamiento* [60] [61] [62]. De esta forma, el coeficiente de agrupamiento es calculado para cada gen de forma separada y para todos los genes que pertenezcan al conjunto de datos y estén relacionados con otros genes. Este coeficiente se representa en la ecuación (3.4).

$$C_i = \frac{E_i}{k_i(k_i-1)/2} \tag{3.4}$$

donde k_i es el número de vecinos del gen i y E_i es el número de conexiones presentes entre los genes conectados a este.

De esta forma, se calcula el coeficiente de agrupamiento para toda la matriz de similitud y se relaciona el promedio de los valores del coeficiente de agrupamiento versus el umbral de similitud. Con esto, se puede encontrar cuál es el valor de umbral que debe ser seleccionado. Este umbral debe tener las siguientes características [60]*:*

a) Ser tan bajo que permita conservar un número suficiente de conexiones.
b) Ser consistente con la propiedad de transitividad de relaciones lineales.
c) Ser tan alto como para que la probabilidad de que las relaciones lineales implícitas que ocurren por azar sea baja.

En consecuencia, para la selección del umbral se implementó computacionalmente y siguió el siguiente algoritmo para cada valor del umbral definido en el rango de 0 a 1 y para cada una de las redes construidas:

a) Calcular el coeficiente de agrupamiento de cada gen de la red
b) Calcular el coeficiente de agrupamiento de toda la red
c) Generar una red aleatoria con el mismo número de genes
d) Calcular el coeficiente de agrupamiento de cada gen de la red aleatoria
e) Calcular el coeficiente de agrupamiento de la red aleatoria
f) Calcular la diferencia entre los coeficientes de agrupamiento de la red construida y la red aleatoria
g) Graficar la diferencia calculada en (f) e identificar el máximo local entre los vecinos más cercanos. Este valor será el umbral que se debe seleccionar.

De este modo, se empleó este algoritmo sobre las matrices de *kernel* construidas, se determinó el umbral para cada matriz y se eliminaron todas las conexiones que se encontraron por debajo del umbral establecido.

3.2.13 Construcción de red de interacción entre genes

Las redes de interacción entre genes permiten visualizar las relaciones entre genes de una forma gráfica y comprender las interacciones de un gen con los demás.

Para la construcción de las redes de interacción entre los genes estudiados, se emplearon

las matrices *kernel* construidas en los pasos anteriores y previamente aplicado el filtro de las conexiones sobre estas. Después de esto, se verificó la simetría de las matrices. En caso de no ser así, es necesario realizar algún procedimiento para volver simétricas las matrices. Además, se requiere verificar y anular el valor de las relaciones de un gen consigo mismo; es decir, si el valor de la relación entre un gen con el mismo es mayor a cero, se cambia por cero. De este modo, no se tendrán bucles dentro de la red que se genere y se graficarán únicamente las relaciones de un gen con los demás genes presentes.

Por último, se debe trabajar con alguna herramienta de *software* que permita construir grafos. En este trabajo, se emplea la librería *igraph* contenida en *R Project,* la cual crea las redes de genes a partir de las matrices simétricas. En estas redes, cada nodo identifica a un gen y cada interacción refleja las relaciones entre los genes.

En *R Project* se ejecutaron los siguientes comandos para la creación y visualización de la red:

```
library(igraph)

grafoGenes<-graph.adjacency(matKernel,mode="undirected",weighted=TRUE)

plot(grafoGenes)
```

4. Aplicación de la metodología propuesta a datos reales

En esta sección se presentarán los resultados de aplicar la metodología propuesta en el capítulo anterior a un conjunto de datos reales.

4.1 Datos de expresión

Los datos de expresión empleados en este trabajo fueron datos de microarreglos y de RNA_Seq para el mismo ente biológico, la planta *Arabidopsis thaliana.*

Después de realizar el proceso de selección de genes y escoger aquellos genes con mayor expresión, se reestructuró la conformación de cada conjunto de datos en el número de genes y el número de experimentos final (ver Tabla 4-1).

Tabla 4-1: Datos de expresión

Conjunto de datos	Número de genes	Número de experimentos
Microarreglos_1	5529	278
RNA_Seq	2095	12
Microarreglos _2	5590	24

Luego de tener los nuevos conjuntos de datos, se construyeron las matrices de similitud (matrices *kernel*), empleando el *kernel coseno* y el *kernel gaussiano* con el valor para el parámetro sigma de 0.000001 para microarreglos y de 0.0001 para RNA_Seq, de acuerdo a la mejor separación y distribución de los datos en el plano 2D (ver sección 3.2.3).

4.2 Datos categóricos

Para la experimentación con datos categóricos asociados a la planta *Arabidopsis thaliana,* se emplearon los datos relacionados con factores de transcripción, rutas KEGG y categorías GO asociados a los genes de la planta. Con el objetivo de vincular esta información con los datos de expresión se filtraron los genes y se construyeron conjuntos de datos categóricos asociados con los genes incluidos en los datos de microarreglos y en los datos de RNA_Seq. De esta forma, se conformaron los conjuntos de datos categóricos (ver Tabla 4-2).

Tabla 4-2: Datos categóricos

Tipo de dato	Número de categorías
Factores de transcripción	85
Rutas KEGG	953
Funciones GO	741

Posteriormente, se construyeron las matrices de similitud (matrices *kernel*) de los datos de rutas de KEGG y GO, empleando un *kernel* polinomial con el valor para el grado del polinomio de 4 para datos KEGG y de 3 para datos GO, de acuerdo a la mejor distribución de los datos en el diagrama 2D.

4.3 Datos fusionados

A partir de las matrices *kernel* construidas para los datos de expresión y para los datos categóricos, se construyeron las matrices *kernel* fusionadas, las cuales consistieron en la integración de cada matriz *kernel* expresión con cada matriz *kernel* categórica. De esta forma, se obtuvieron las matrices *kernel* de los datos fusionados (ver Tabla 4-3).

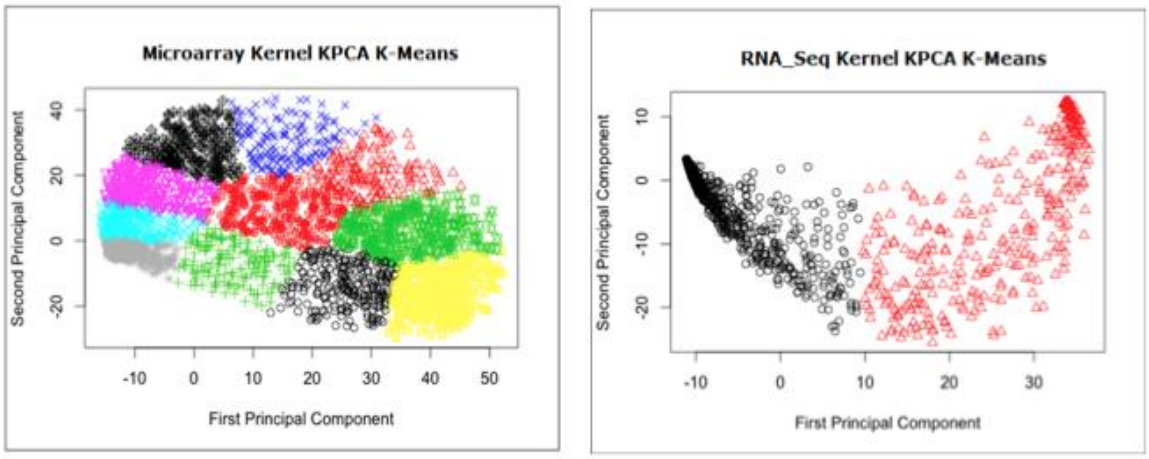
Tabla 4-3: Datos fusionados

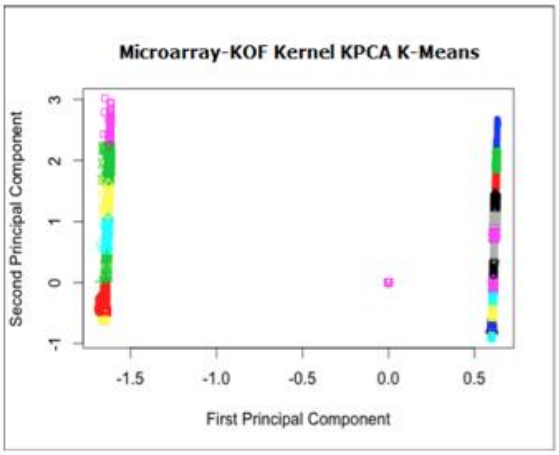
Matriz	Datos fusionados	Dimensiones
MicroKOF	Microarreglos, rutas KEGG	5529 x 5529
MicroGOF	Microarreglos, categorías GO	5529 x 5529
RNAKOF	RNA_Seq, rutas KEGG	2095 x 2095
RNAGOF	RNA_Seq, categorías GO	2095 x 2095

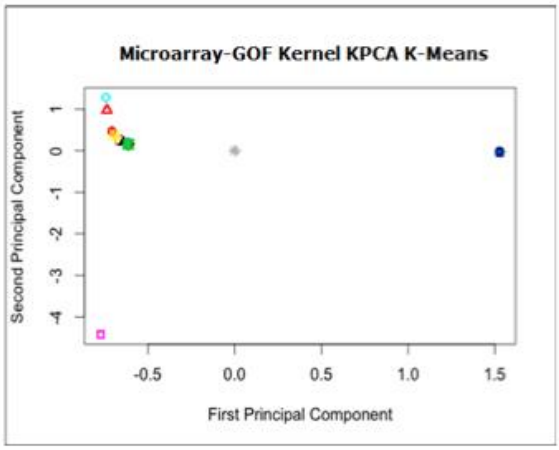
4.4 Construcción grupos de genes

Para construir grupos de genes, se emplearon los algoritmos de agrupamiento *k-means* y *fuzzy k-means* aplicados sobre el *kernel* ACP de las matrices *kernel* de los datos de expresión, categóricos y fusionados. Además, se realizó la visualización de la proyección de los grupos construidos sobre los datos obtenidos después de aplicar el KACP sobre las matrices de similitud de cada uno de estos conjuntos de datos (ver Figura 4-1).

Figura 4-1: Representación visual de *k-means* sobre el KACP de las matrices de similitud







En la Tabla 4-4 se muestra el número de grupos obtenido con cada conjunto de datos, el grupo con el mayor número de genes y el grupo con el menor número de genes. El número de grupos se obtuvo teniendo en cuenta una regla heurística representada en la Ecuación 3.3. En los grupos con un número alto de genes se observa que estos últimos presentan una mayor relación entre sí, reflejándose muchos patrones compartidos por varios genes y no diversos grupos pequeños compartiendo propiedades muy particulares. Es decir, los genes agrupados reflejan patrones de similitud a nivel de expresión teniendo en cuenta que son los genes con mayor expresión los que se han empleado para el análisis de los datos de microarreglos y de RNA_Seq. Además, los genes pertenecientes a un mismo grupo también reflejan comportamientos de regulación similares visualizados más adelante con la asociación de factores de transcripción sobre cada uno de los grupos formados.

4.5 Categorización de los grupos

En el proceso de análisis de los grupos obtenidos se busca encontrar las características propias de cada uno. Para ello, de acuerdo a la metodología propuesta, se relacionaron los genes pertenecientes a cada grupo con los datos de factores de transcripción buscando obtener patrones de regulación significativos y factores representativos en cada grupo.

En consecuencia, se presentan los factores de transcripción más representativos encontrados en los grupos resultantes y la cantidad de asociaciones a estos (ver Tablas 4-5 y 4-6).

Como se puede apreciar en las Tablas 4-5 y 4-6, los promotores más representativos que se encontraron con los algoritmos *k-means* y *fuzzy k-means* son similares, pero existe una

pequeña variación en el orden en el que se encuentran y la cantidad de asociaciones a cada uno. Sin embargo, con el algoritmo *fuzzy k-means* fue posible encontrar un mayor número de asociaciones en cada grupo y en cada uno de los tipos de datos.

El número de genes que conforma cada grupo y el número de asociaciones a los factores de transcripción en los grupos construidos con el algoritmo fuzzy k-means es mayor que los resultados obtenidos con los grupos construidos con el algoritmo k-means, debido a la propiedad que presentan los grupos difusos y que se encuentra relacionada con que un gen puede pertenecer a más de un grupo a la vez y no se limita su participación a un solo grupo como ocurre con el algoritmo k-means. Por lo tanto, al momento de calcular el número total de asociaciones a los factores de transcripción (ver Tabla 4-7), este valor aumenta considerablemente en este tipo de grupos difusos. Esto se debe a que, si un gen está asociado con un factor y este gen pertenece a diferentes grupos, en cada grupo en el que se encuentre va a ser tenido en cuenta para incrementar el número de asociaciones a dicho factor.

Tabla 4-4: Grupos construidos con cada algoritmo de agrupamiento sobre cada tipo de dato

Método	Número de grupos	Grupo mayor	Grupo menor
K-means microarreglos	11	3017	160
K-means RNA_Seq	2	1597	498
K-means KOF	21	4684	6
K-means GOF	19	1646	52
K-means MicroKOF	24	1493	64
K-means MicroGOF	22	875	20
Fuzzy K-means microarreglos	11	5149	828
Fuzzy K-means RNA_Seq	2	1605	490
Fuzzy K-means KOF	21	5071	19
Fuzzy K-means GOF	19	3406	246
Fuzzy K-means MicroKOF	24	5119	456
Fuzzy K-means MicroGOF	22	5510	246

Tabla 4-5: Factores de transcripción más representativos encontrados con el algoritmo *k-means*

FACTOR DE TRANSCRIPCIÓN	CANTIDAD DE ASOCIACIONES
GATA [LRE	44779
RAV1-A	42374
MYB4	37972
W-box	35392
LFY	35265
DPBF1&2	34570
Tbox	31548
Ibox	27496
ATB2	27187
BoxII	26771
ARF	25187
AtMYC2 BS in RD22	25360
ARF1	25194
SORLIP2	24903
Bellringer	23409

Tabla 4-6: Factores de transcripción más representativos encontrados con el algoritmo *fuzzy k-means*

FACTOR DE TRANSCRIPCIÓN	CANTIDAD DE ASOCIACIONES
GATA [LRE	362208
RAV1-A	352224
MYB4	315786
W-box	293417
DPBF1&2	286973
LFY	285676
Tbox	262511
ATB2	223066
Ibox	222749
BoxII	222708
ARF	209750
ARF1	209164
SORLIP2	206270
AtMYC2 BS in RD22	205485
Bellringer	188959

4.6 Análisis de los grupos construidos

El número de grupos que se formó para cada tipo de dato se obtuvo usando el criterio heurístico descrito en la Ecuación 3.3 y los métodos de agrupación empleados fueron no supervisados, debido a que no se tiene un estándar previamente establecido y que sea completamente confiable con el cual comparar los grupos construidos. Además, se hubiera podido hacer un análisis supervisado para un subgrupo de genes bien conocido, pero es difícil clasificar un número tan grande de genes como los trabajados en cada conjunto de datos cuando se tiene información reportada sobre tan pocos.

De esta forma, se realizó el análisis funcional de los grupos a través de la asociación con información de factores de transcripción de los genes que conformaban cada grupo. Después de obtener las relaciones intra-grupos de factores de transcripción y encontrar los factores más representativos en los grupos, se realizó un proceso de comparación de los resultados obtenidos con información biológica reportada en las bases de datos biológicas y en la literatura.

De esta forma, de los factores de transcripción representativos se pueden resaltar MYB4 y DPBF1&2, por encontrarse dentro de las cinco promotores con más genes asociados tanto con el algoritmo *k-means* como con el algoritmo *fuzzy k-means*; además, por sus características biológicas, estando relacionados con la respuesta a las condiciones experimentales de los datos de microarreglos y de RNA_Seq.

Adicionalmente, se encontró la relación biológica entre los factores de transcripción representativos en los grupos, es decir, se pudo ver cómo están relacionados entre sí, no sólo con los genes sino la relación que existe entre ellos mismos. De este modo, surgieron la importancia y la validez de los factores de transcripción encontrados y su función directa en la regulación de los genes de la planta estudiada, las condiciones ambientales a las cuales esta se encontraba y, por último, las condiciones experimentales registradas en los datos de expresión. Para enfatizar este punto, se citan las siguientes relaciones:

Partiendo de los dos factores más representativos, MYB4 y DPBF1&2, se puede decir que el factor de transcripción DPBF1&2 también es conocido como DC3, y es un factor de respuesta al ácido abscísico y al estrés. Se encuentra involucrado en rutas de señalización reguladas por este ácido, regulación positiva de transcripción y respuesta a carencia de agua [63] [64].

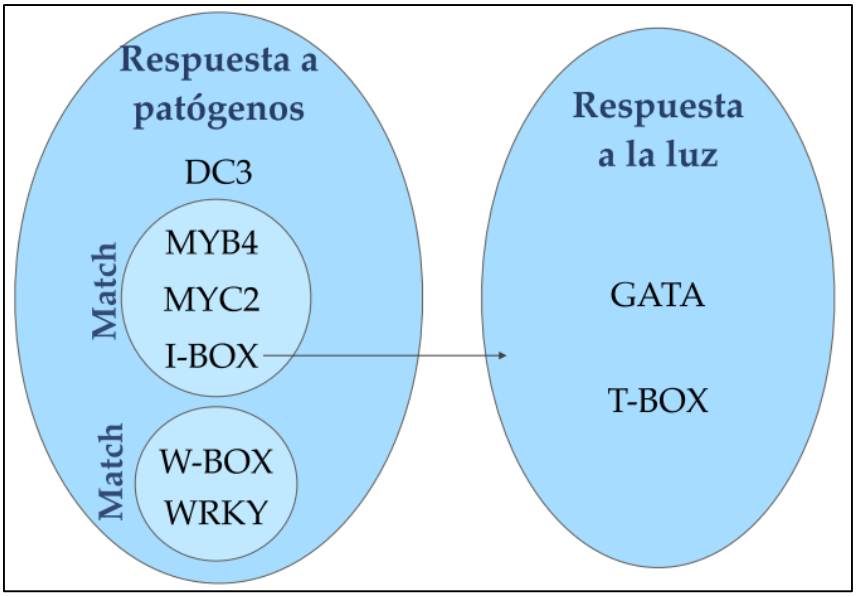
El factor de transcripción MYB4 está presente como respuesta al estrés ambiental en la planta. Este promotor pertenece a la familia de proteínas MYB, la cual es la familia putativa de factores de transcripción para el promotor I-Box; dado esto, el promotor IBox presenta la mejor asociación con las proteínas pertenecientes a esta familia; además, también está relacionado con respuesta a estrés y está involucrado en la fotosíntesis [65]. El factor de transcripción ATMYC2 en cooperación con la proteína MYB2 participa en la regulación de los genes mediados por ácido abscísico en condiciones de estrés hídrico [66].

El factor de transcripción W-Box también presenta una función putativa en respuesta a estrés ambiental. El factor de transcripción W-Box fue reconocido específicamente por el ácido salicílico (SA) inducido por proteínas de unión a ADN WRKY, un grupo de sitios de unión WRKY actúa como elementos reguladores negativos para la expresión inducible de resistencia a las enfermedades, ya que tienen la propiedad de ser reguladores de la inmunidad transcripcional [67].

También se encuentran los factores de transcripción GATA y T-Box, los cuales están relacionados con la luz; el factor GATA se encuentra involucrado con la regulación de genes que responden a la luz y las mutaciones en el factor de transcripción T-box han resultado en reducciones de luz activadas por la transcripción de los genes [68].

De forma general, se pueden apreciar todas las relaciones dilucidadas entre los factores de transcripción de la planta *Arabidopsis thaliana* encontrados como significativos (Figura 4-2).

Figura 4-2: Relaciones encontradas entre los factores de transcripción más representativos



4.7 Construcción de redes de genes

Para la construcción de las redes de interacción entre genes e identificar otras relaciones asociadas a los genes, se emplearon las matrices *kernel* construidas previamente y, en el caso de los datos de expresión, se escogió y empleó el *kernel* coseno como medida de similitud. Además, sobre estas matrices se aplicaron los algoritmos escogidos para realizar el filtro de las conexiones más fuertes entre cada par de genes.

Sin embargo, entre los algoritmos *k vecinos más cercanos* y el *umbral basado en el coeficiente de agrupamiento*, se escogió como principal algoritmo de poda de conexiones el método de selección del umbral con base en el coeficiente de agrupamiento, debido a su soporte matemático-estadístico y a sus bases teóricas relacionadas con la validación de que las redes obtenidas sean realmente redes biológicas y no redes aleatorias [60] [61] [62]. Con este filtro realizado, se construyeron las redes de genes, empleando (como se mencionó en el capítulo anterior) la librería *igraph* del programa R.

Para el cálculo de este umbral en cada red construida se empleó el algoritmo implementado previamente para este caso (ver sección 3.2.12) y se aplicó sobre cada una las redes. En la Figura 4-3, se presenta el resultado de la selección del umbral. Específicamente, allí se visualiza la diferencia entre los valores del coeficiente de agrupamiento de la red construida con los datos de microarreglos y los valores del coeficiente de agrupamiento de la red construida de forma aleatoria, para este caso, el valor del umbral fue de 0.6.

Un ejemplo de las redes construidas se presenta de forma muy general con la visualización de la red de genes construida a partir de los datos de microarreglos en la Figura 4-4.

Con las redes construidas se realizó el análisis de cada una a través del cálculo de determinadas métricas topológicas como *el diámetro*, el cual representa el camino mínimo entre los dos genes más alejados, y *el componente conexo*, que indica todas las subredes que tienen todos los genes conectados entre sí [69]. Además, se calculó una propiedad representativa de las redes biológicas como *la transitividad,* que nos permite identificar si todos los genes adyacentes a cada gen de la red están conectados entre sí [60] (ver Tabla 4-7).

Para verificar las relaciones encontradas entre los genes en las redes, se realizó la búsqueda de conexiones reportadas en la base de datos *String.* Esta base de datos contiene información previamente reportada en bases de datos biológicas y en la Literatura, contemplando información de experimentos de alto rendimiento, co-expresión, conocimiento previo y contexto genómico [32]. Las relaciones encontradas fueron entonces comparadas con las conexiones obtenidas. A partir de esta comparación, se determinó cuales conexiones entre genes formadas con la metodología propuesta estaban también reportadas; estas conexiones se visualizan en el Anexo B.

Figura 4-3: Selección del umbral con base en el coeficiente de agrupamiento para los datos de microarreglos

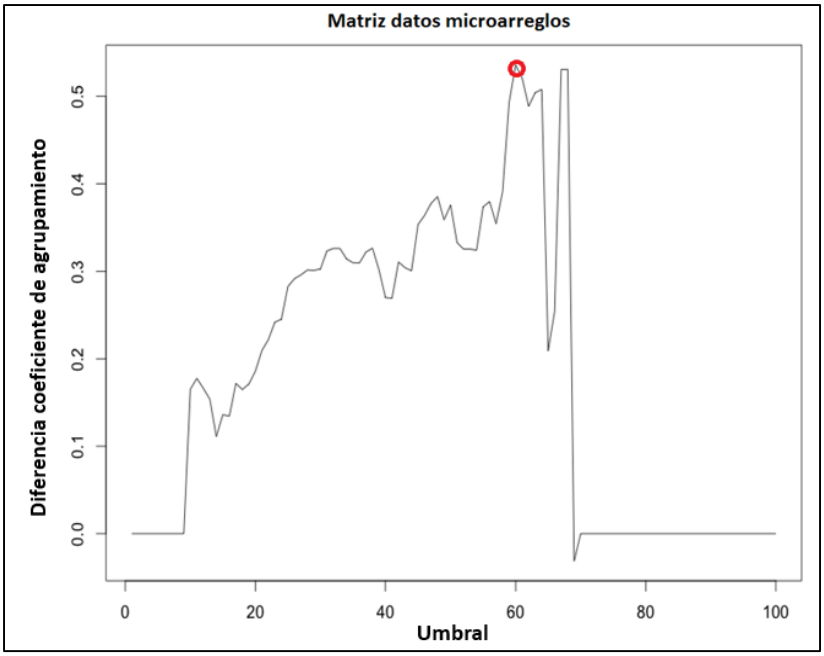


Figura 4-4: Esquema representativo de la red de interacción entre genes a partir de los datos de microarreglos

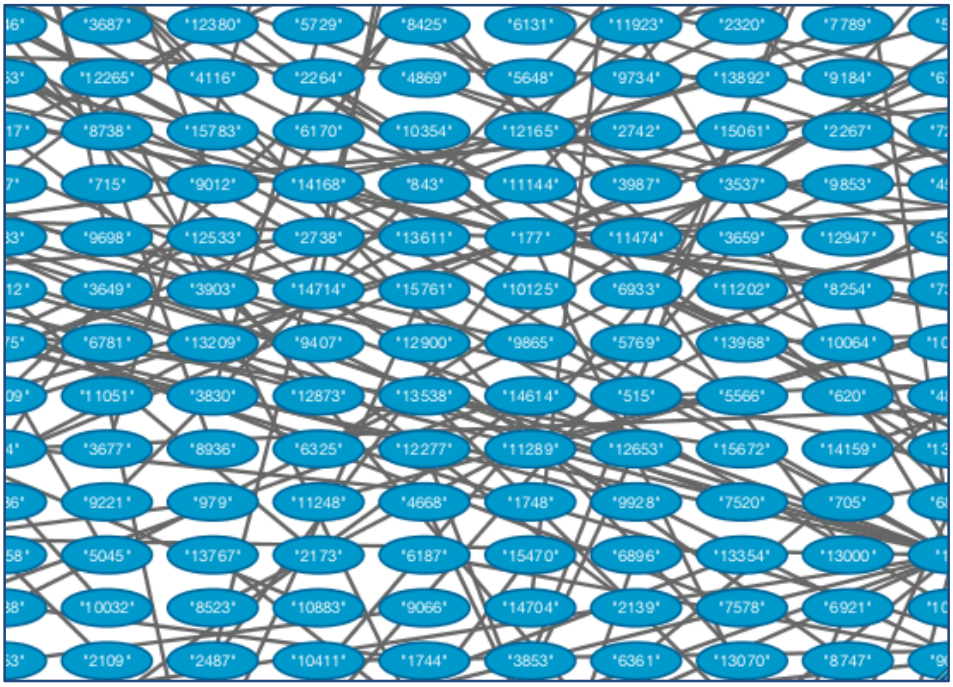


Tabla 4-7: Descripción de las características de las redes construidas

Tipo de dato	Algoritmo de poda	Umbral del algoritmo	Número de genes	Número de conexiones	Componente conexo	Diámetro	Transitividad
Microarreglos_2	Coef. de agrupamiento	0.99	5590	5712613	10	11.253	0.9345
RNA_Seq	Coef. de agrupamiento	0.99	2095	317	187	8.9355	0.5945
Microarreglos_2-KOF	Coef. de agrupamiento	0.92	2858	3359457	2	1.9449	0.9430

Por otro lado, como proceso de verificación de las redes de interacción entre genes construidas, se desarrolló en el programa R el procedimiento para construir redes teniendo como medida de similitud entre los genes la *correlación de Pearson*, la cual ha sido empleada en varios casos de estudio y análisis de datos de expresión.

En la Tabla 4-8 se visualizan las características que posee la red construida con los datos de microarreglos y esta medida de similitud. De este modo, se puede apreciar que, partiendo del mismo conjunto de datos (microarreglos_2) pero aplicando dos medidas de similitud diferentes —el *kernel* coseno y la correlación de Pearson—, se obtuvieron a su vez dos redes diferentes. La red construida con el *kernel* coseno será denominada redCoseno (sus características se visualizan en la primera fila de la tabla 4-7) y la construida con la correlación de Pearson será denominada redPearson (ver Tabla 4-8). A partir de lo explicado, se analiza en primer lugar el número de conexiones: este número es mayor en la redCoseno, lo que indica que por medio del *kernel* coseno se encuentran más genes conectados unos a otros. En segundo lugar, se analiza el componente conexo, el cual es más grande en la redPearson, por lo que se infiere que con la correlación de Pearson se formaron diversos grupos con poca cantidad de genes y conexiones, mientras que con el *kernel* coseno se obtuvieron pocos grupos con un número mayor de genes interconectados entre sí. Esto también se puede relacionar con el diámetro de las redes, el cual refleja que los genes en la redPearson están más alejados entre sí, a diferencia de la redCoseno, donde las distancias entre los genes son menores. Por último, el valor de transitividad resume lo expuesto anteriormente al ser mayor en la redCoseno, indicando

que a diferencia de la redPearson los genes tienen un número mayor de conexiones entre ellos y entre los genes conectados a cada uno de los genes de la red.

Tabla 4-8: Descripción de las características de la red construida empleando la correlación de Pearson

Tipo de dato	Algoritmo de poda	Umbral del algoritmo	Número de genes	Número de conexiones	Componente conexo	Diámetro	Transitividad
Microarreglos_2	Coef. de agrupamiento	0.95	5590	1738276	1885	17.842	0.7437

4.8 Verificación de las redes construidas con información biológica reportada previamente

A partir de las redes construidas empleando el *kernel* coseno, se realizó un proceso de verificación biológica de las redes con base en información reportada en la literatura y en bases de datos. Este proceso consiste en la caracterización de los genes presentes y conectados en las redes por medio de la asociación con atributos biológicos que han sido reportados experimentalmente para cada gen como, por ejemplo, funciones biológicas, rutas metabólicas y procesos biológicos [70].

En este sentido, se seleccionó un grupo de genes específico para realizar el proceso de asociación de categorías biológicas. Para la construcción de este grupo, se identificaron los genes con mayor expresión en cada conjunto de datos, que estuvieran conectados con otros en las redes formadas y cuyo nivel de conexión estuviera por encima del umbral predefinido. Con base en esto, se realizó la búsqueda de las categorías biológicas de cada uno de estos genes en la base de datos Gene Ontology. Específicamente se buscó por procesos biológicos involucrados con este grupo de genes y se establecieron los procesos que eran comunes entre ellos y que estaban sobrerrepresentados en las redes (ver Tabla 4-9). Para la identificación de estos procesos se empleó el *plug-in Bingo* del programa *Cytoscape* [71].

De este mismo modo, se realizó la búsqueda de información en la misma Gene Ontology para identificar las funciones moleculares sobrerrepresentadas en las redes, teniendo como base el grupo de genes previamente establecido (ver Tabla 4-10).

Adicionalmente, se pudieron establecer y visualizar las relaciones existentes entre los procesos biológicos representativos resultantes y, a la vez, cómo están relacionados con otros procesos en la planta (ver Figura 4-4), es decir, cada uno de estos procesos está relacionado funcionalmente en la medida que a nivel de cada organismo se llevan a cabo procesos que pueden afectar o depender de otro proceso particular [72].

De igual forma, se encontraron las relaciones entre las funciones moleculares atribuidas a los genes y que se establecieron como representativas en las redes. Se aprecia la relación tanto entre las funciones sobrerrepresentadas como la relación con otras funciones biológicas de la planta, las cuales también interactúan a nivel molecular para el funcionamiento del organismo (ver Figura 4-5). Para visualizar gráficamente las relaciones entre los procesos biológicos y entre las funciones moleculares se empleó el programa *Cytoscape* [71].

Tabla 4-9: Procesos biológicos sobrerrepresentados en las redes

Proceso biológico GO
Respuesta a estímulos
Respuesta a estímulos químicos
Respuesta estímulo abiótico
Respuesta a estrés
Respuesta a hongo
Respuesta estímulo biótico
Respuesta a otro organismo
Proceso biosintético celular
Fotosíntesis
Respuesta a estímulos de temperatura
Respuesta a bacteria
Respuesta a estrés osmótico

Figura 4-5: Interconectividad entre los procesos biológicos sobrerrepresentados en las redes construidas

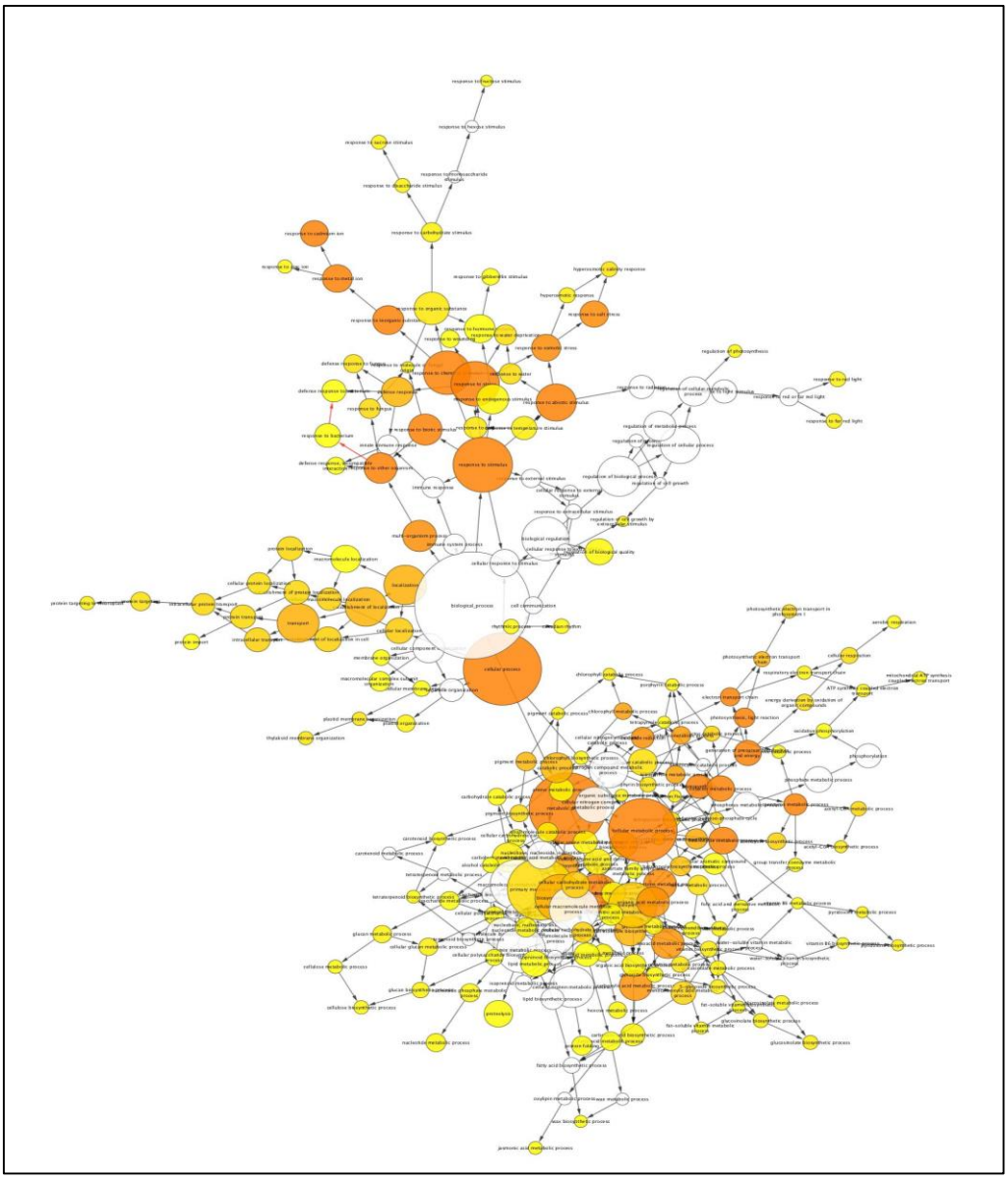
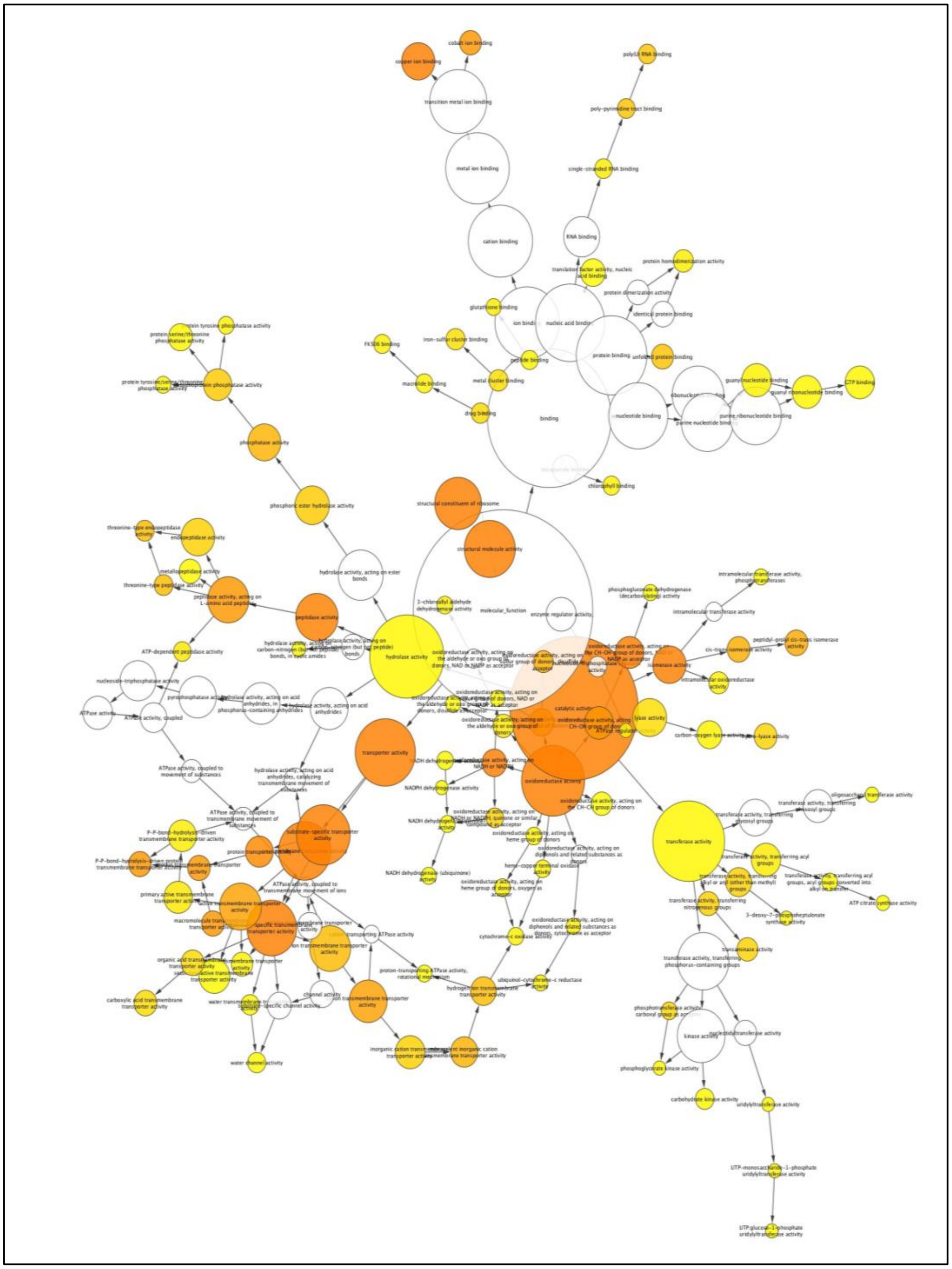


Tabla 4-10: Funciones moleculares sobrerrepresentadas en las redes

Función molecular GO
Actividad del factor de traducción, unión del ácido nucleico
Actividad de la oxidoreductasa, actuando en el aldehído o grupo oxo de donantes, disulfuro como aceptador
Actividad ATPasa transportadora de protones, mecanismo giratorio
Actividad de la NADH deshidrogenasa (ubiquinona)
Actividad de la NADH deshidrogenasa (quinona)
Actividad de la transferasa oligosacarina
Actividad de la hidrolasa, actuando en uniones de carbón-nitrógeno (pero no péptidos), en amidas cíclicas
Actividad de citrato sintasa de ATP
Actividad de óxidoreductasa, actuando en el grupo de donantes CH-CH
Actividad del fosfogluconato deshidrogenasa (descarboxilante)
Actividad de transferasa intramolecular, fosfotransferasas
Actividad de transportador de [proteínas] transmembranas secundarias activas
Actividad del canal de agua [celular]
Actividad del transportador de [proteína] transmembrana de agua
Actividad de transferasa, transfiriendo grupos acilos, grupos acilos convertidos en alquilo(s) en la transferencia
Actividad de la uridiltransferasa
Actividad de fosfoglicerado kinasa
Actividad reguladora de ATPasa

Figura 4-6: Interconectividad entre las funciones moleculares sobrerrepresentadas en las redes construidas



4.9 Análisis de las redes de genes construidas

Con base en los datos de expresión y categóricos se construyeron las redes de interacción entre genes empleando la metodología propuesta en el presente trabajo, tanto para los datos de expresión de forma separada como para los datos fusionados (expresión y categóricos).

Se pudo establecer que las redes de genes construidas presentan cualidades básicas de una red biológica, enmarcadas dentro de la teoría de la transitividad [73] en donde se resaltan parámetros como el coeficiente de agrupamiento y el diámetro de la red. El coeficiente de agrupamiento es una propiedad de las redes biológicas y representa la probabilidad de que los genes adyacentes a un gen particular estén conectados a la vez entre ellos. De esta forma, se precisa que si este valor es “uno” para un gen significa que todos los genes conectados a este gen también están conectados entre sí, y si este valor es “cero” significa que no hay conexión entre los genes que se conectan a este. En las redes construidas, el valor del coeficiente de agrupamiento dio muy cercano a 1, lo que indica que los genes en la red sí se encuentran conectados unos a otros y que entre los vecinos de cada gen también hay conexión. De acuerdo con lo anterior, y revisando los resultados obtenidos en la Tabla 4-7, se puede decir que las redes obtenidas con la metodología seguida si corresponden a redes biológicas y no a redes aleatorias dadas las propiedades que presenta. En especial, teniendo en cuenta la conexión existente entre los genes vecinos a cada uno de los genes que conforman las redes y a la no formación de islas de genes (pequeños subgrupos de genes), lo que no se esperaría que se obtuviera en una red aleatoria, en la cual los genes vecinos a un gen no necesariamente están conectados entre sí.

Con relación al valor del diámetro en las redes, se encontró que este valor es mayor cuando hay un número más grande de genes, pero también disminuye directamente al momento de construir las redes con datos fusionados (ver Tabla 4-7), lo que significa que al fusionar los datos se forma una mayor conexión entre los genes.

A partir de las redes de interacción entre genes construidas, se pudo ver cómo los genes que resultaron conectados en las redes también están conectados de alguna forma en actividades biológicas reales de la planta, como es el caso de los procesos biológicos y funciones moleculares encontradas como sobrerrepresentadas en las redes. De esta

forma, se puede apreciar que los genes conectados en las redes están relacionados funcionalmente y hacen parte de los mismos procesos o funciones biológicas.

Los procesos biológicos sobrerrepresentados están relacionados a su vez con el tipo de condiciones biológico-experimentales de los conjuntos de datos de microarreglos y de RNA_Seq, es decir, se ve reflejado cómo ante condiciones de resistencia a patógenos se activan diversos procesos de forma representativa: por ejemplo, respuesta a estrés, respuesta a otro organismo, respuesta a bacteria, etc.

Con relación a las funciones moleculares encontradas como sobrerrepresentadas en las redes, se visualiza que algunas corresponden al metabolismo basal como la fotosíntesis y otros a procesos específicos de inmunidad que se activan cuando la planta está sometida a algún patógeno (respuesta a una bacteria o a estrés).

Cabe resaltar que el análisis de verificación de las conexiones en las redes, realizado con el procedimiento de contraste con información biológica reportada tanto en *Gene Ontology* como en la base de datos *String,* fue útil para dar soporte a la metodología seguida y a las conexiones encontradas, verificando que tuvieran un sentido biológico. Pero, vale anotar, que no se validaron en un sentido estricto, debido a que no se confirmaron a través de experimentos biológicos.

5. Conclusiones y recomendaciones

En esta sección se presentan las conclusiones del trabajo de investigación realizado y las recomendaciones para futuras investigaciones relacionadas con la identificación de las relaciones entre genes.

5.1 Conclusiones

El objetivo del presente trabajo fue plantear una metodología para identificar las relaciones entre genes, problema que se ha venido trabajando desde hace varios años pero que aún sigue siendo de gran envergadura para la comunidad científica de la actualidad. El interés en este problema radica en la importancia que presenta al considerar las relaciones entre genes como un mecanismo para el entendimiento de procesos biológicos en los seres vivos. Sin embargo, aún no existe una solución definitiva a este problema, dada la complejidad que tiene la recolección de información necesaria de datos biológicos, su procesamiento y respectivo análisis.

Con la metodología propuesta se busca la identificación de relaciones entre genes y que pueda ser aplicada sobre diferentes tipos de datos de expresión, como es el caso de microarreglos de ADN y RNA_Seq. Aunque los microarreglos están basados en hibridación y RNA_Seq en secuenciación, permiten obtener medidas de los niveles de transcripción de miles de genes simultáneamente, se puede aplicar la metodología sobre datos de estos tipos y realizar un análisis de los genes en unas condiciones particulares. Se escogieron datos de microarreglos y de RNA_Seq del mismo ente biológico y que estuvieran relacionados con experimentos de resistencia a infección de patógenos. De esta forma, se pudieron encontrar relaciones entre genes, patrones de regulación y funcionalidades comunes entre ellos, a partir de los dos tipos de datos.

Además, la metodología presenta la funcionalidad para combinar datos heterogéneos que incluyen datos categóricos y datos de expresión, permitiendo descubrir patrones no lineales que podrían no ser descubiertos con el uso de métodos lineales.

También la metodología propuesta permite obtener grupos de genes independientemente del tipo de dato con el que construyan (datos de expresión, datos combinados entre expresión y categóricos). En particular, permite identificar relaciones entre genes a nivel de co-expresión y regulación de los mismos. Además, por medio de la asociación y categorización de los grupos con datos de factores de transcripción reportados en la base de datos *AtTFDB,* fue posible determinar que los genes dentro de un mismo grupo compartían patrones de regulación similares. Para el caso de los datos de estudio fue posible encontrar, a partir de la comparación con información biológica previamente reportada en la literatura acerca de los factores de transcripción más representativos en los grupos la evidencia que dichos factores estaban relacionados con estrés, respuesta al ácido abscísico y respuesta a la luz, lo que a su vez está relacionado directamente con las condiciones experimentales de los datos de expresión, los cuales contemplaban la infección por patógenos en *Arabidopsis thaliana.* De este modo, los grupos obtenidos pueden ser usados para predecir factores de transcripción relacionados con los genes o, al menos, para predecir los patrones de regulación comunes y, por lo tanto, predecir indirectamente las funciones de los genes.

De manera similar, a partir de las redes construidas es posible identificar los genes que están relacionados, comparten alguna propiedad biológica o participan al tiempo en algún mecanismo funcional, tales como procesos biológicos y funciones moleculares. Esto fue evidenciado al realizar el proceso de verificación de las relaciones con información biológica reportada previamente en base de datos como Gene Ontology. Al mismo tiempo, al identificar las funciones y procesos biológicos sobrerrepresentados en las redes, se encontraron asociaciones con los datos de expresión y las condiciones de patogenicidad y ambientales; por ejemplo, los procesos de respuesta a estrés, de respuesta a hongos y las funciones moleculares de regulación y transporte de proteínas.

Al comparar los resultados obtenidos con los métodos de *kernel* y los obtenidos con una técnica tradicional como la correlación de Pearson, los primeros brindaron mejores resultados al formar redes de interacción con un número mayor de conexiones entre los

genes y con un grado mayor de conexiones entre los genes adyacentes a cada gen de la red.

Vale destacar que la construcción global de una red de interacción entre genes de un ser vivo es una tarea difícil, debido a que los datos obtenidos de manera experimental solo aportan cierta información acerca del mismo y no sobre todos los mecanismos biológicos llevados a cabo en su interior.

5.2 Recomendaciones

Como trabajo de investigación futuro, se plantea la exploración y experimentación con otros datos de expresión tanto de microarreglos como de RNA_Seq para realizar una validación adicional de la metodología propuesta.

Por otro lado, se sugiere realizar experimentación con otras funciones *kernel* para comparar las relaciones encontradas y las redes formadas, y determinar posibles nuevas relaciones.

Adicionalmente, se sugiere aplicar otros métodos para realizar el filtro de las conexiones encontradas entre los genes en el proceso de construcción de las redes de interacción entre genes. También se recomienda usar otras herramientas para la visualización de las redes construidas.

A. Anexo: Experimentos conjuntos de datos de microarreglos (NCBI)

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM322545
GSM322546
GSM322547
GSM322548
GSM322549
GSM322550
GSM322551
GSM322552
GSM322553
GSM322554
GSM322555
GSM322556
GSM304028
GSM304029
GSM304031
GSM304032
GSM206274
GSM206275
GSM206276
GSM206277
GSM206278
GSM206279
GSM206280
GSM206281
GSM206282

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM206283
GSM206284
GSM206285
GSM260880
GSM260881
GSM260882
GSM260883
GSM189096
GSM189097
GSM189099
GSM189100
GSM189101
GSM189102
GSM189103
GSM189104
GSM189105
GSM189106
GSM189107
GSM189108
GSM189109
GSM189110
GSM189111
GSM189112
GSM189114
GSM189116
GSM189117
GSM189118
GSM189119
GSM189120
GSM189121
GSM189122

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM189123
GSM189124
GSM189163
GSM189164
GSM189165
GSM189170
GSM189171
GSM189172
GSM189173
GSM189174
GSM189175
GSM189176
GSM189177
GSM189178
GSM142829
GSM142830
GSM142831
GSM142832
GSM142833
GSM142834
GSM142835
GSM142836
GSM142837
GSM142838
GSM142839
GSM142840
GSM142841
GSM142842
GSM142843
GSM142844
GSM142845
GSM142846
GSM142847

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM142848
GSM142849
GSM142850
GSM142851
GSM142852
GSM142853
GSM142854
GSM142855
GSM157373
GSM157374
GSM157375
GSM157376
GSM157377
GSM157378
GSM157379
GSM157380
GSM157381
GSM134430
GSM134431
GSM134432
GSM134433
GSM134434
GSM134435
GSM134436
GSM134437
GSM134438
GSM134439
GSM134440
GSM134441
GSM134442
GSM134443
GSM134444
GSM134445

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM134446
GSM134447
GSM134394
GSM134395
GSM134396
GSM134397
GSM134398
GSM134399
GSM134400
GSM134401
GSM134402
GSM134403
GSM134404
GSM134405
GSM134406
GSM134407
GSM134408
GSM134409
GSM134410
GSM134411
GSM134376
GSM134377
GSM134378
GSM134379
GSM134380
GSM134381
GSM134382
GSM134383
GSM134384
GSM134385
GSM134386
GSM134387

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM134388
GSM134389
GSM134390
GSM134391
GSM134392
GSM134393
GSM134448
GSM134449
GSM134450
GSM134451
GSM134452
GSM134453
GSM134454
GSM134455
GSM134456
GSM134457
GSM134458
GSM134459
GSM134460
GSM134461
GSM134462
GSM134463
GSM134464
GSM134465
GSM134358
GSM134359
GSM134360
GSM134361
GSM134362
GSM134363
GSM134364
GSM134365
GSM134366

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM134367
GSM134368
GSM134369
GSM134370
GSM134371
GSM134372
GSM134373
GSM134374
GSM134375
GSM134340
GSM134341
GSM134342
GSM134343
GSM134344
GSM134345
GSM134346
GSM134347
GSM134348
GSM134349
GSM134350
GSM134351
GSM134352
GSM134353
GSM134354
GSM134355
GSM134356
GSM134357
GSM133896
GSM133897
GSM133898
GSM133899
GSM133900

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM133901
GSM133902
GSM133903
GSM133904
GSM133905
GSM133906
GSM133907
GSM133782
GSM133783
GSM133784
GSM133785
GSM133786
GSM133787
GSM90867
GSM90868
GSM90869
GSM90870
GSM90871
GSM90872
GSM70995
GSM70996
GSM70997
GSM70998
GSM70999
GSM71000
GSM71001
GSM71002
GSM71003
GSM71004
GSM71005
GSM71006
GSM71007

Experimentos Microarreglos_1 (Geo DataSets - NCBI)
GSM71008
GSM71009
GSM71010
GSM71011
GSM71012
GSM71013
GSM71014
GSM71015
GSM71016
GSM71017
GSM71018
GSM6227
GSM6544
GSM6571
GSM6572
GSM6573
GSM6574
GSM6575
GSM6576
GSM6577
GSM6578
GSM6579
GSM6580
GSM6581
GSM6582
GSM6583
GSM6584

Experimentos Microarreglos_2 (Geo DataSets - NCBI)
GSM71002
GSM71003
GSM71004
GSM71005
GSM70998
GSM70999
GSM71000
GSM71001
GSM70995
GSM70996
GSM70997
GSM71017
GSM71013
GSM71014
GSM71015
GSM71016
GSM71010
GSM71011
GSM71012
GSM71018
GSM71006
GSM71007
GSM71008
GSM71009

B. Anexo: Conexiones reportadas en la base de datos String

Conexiones reportadas STRING_DB	
Q8VZ23	Q9FJC6
Q9FJC6	Q9LU69
Q9FJC6	Q9FJM2
Q9FJC6	Q9FGF0
Q9FJC6	Q9FN15
O82234	Q9FJC6
Q8VY52	Q9FJC6
Q9S720	Q9FJC6
Q9SJ89	Q9FJC6
Q9LT18	Q9FJC6
Q9SR41	Q9FJC6
Q9SMV8	Q9FJC6
Q9SFB3	Q9FJC6
Q9XIA4	Q9FJC6
Q9SGU7	Q9FJC6
Q9ZVZ9	Q9FJC6
Q9LM71	Q9FJC6
Q94F10	Q9FJC6
Q9FWS4	Q9FJC6
Q9S9K3	Q9FJC6
Q9M9H4	Q9FJC6
Q9LYI9	Q9FJC6
Q94K68	Q9FJC6
Q9LKU2	Q9FJC6

Conexiones reportadas STRING_DB	
Q9FI13	Q9FJC6
Q9FKB7	Q9FJC6
Q9LU63	Q9FJC6
Q9LK47	Q9FJC6
Q9LK00	Q9FJC6
Q9LSE4	Q9FJC6
Q9T016	Q9FJC6
Q9SUL0	Q9FJC6
Q9SEL7	Q9FJC6
Q9LZ14	Q9FJC6
Q9FNM5	Q9FJC6
O64750	Q9FJC6
Q9C5M1	Q9FJC6
Q9LUB2	Q9FJC6
Q94JY0	Q9FJC6
Q9FFW9	Q9FJC6
Q9LVV6	Q9FJC6
Q9LVV5	Q9FJC6
Q94AU3	Q9FJC6
Q94K51	Q9FJC6
Q9SXP7	Q9FJC6
Q9C7I0	Q9FJC6
Q9LW20	Q9FJC6
Q9SYX1	Q9FJC6

Bibliografía

[1] J. D. Watson, “The Human Genome Project: Past, Present and Future.” .

[2] N. M. Luscombe, D. Greenbaum, and M. Gerstein, “Review What is bioinformatics ? An introduction and overview,” *Gene Expr.*, pp. 83–100.

[3] A. Yilmaz, M. K. Mejia-Guerra, K. Kurz, X. Liang, L. Welch, and E. Grotewold, “AGRIS: the Arabidopsis Gene Regulatory Information Server, an update,” *Nucleic Acids Res.*, vol. 39, no. suppl 1, pp. D1118–D1122, Jan. 2011.

[4] J. Vert and B. Sch, “1 A primer on kernel methods,” no. 1992, 2004.

[5] G. Joshi-Tope, M. Gillespie, I. Vastrik, P. D’Eustachio, E. Schmidt, B. de Bono, B. Jassal, G. R. Gopinath, G. R. Wu, L. Matthews, S. Lewis, E. Birney, and L. Stein, “Reactome: a knowledgebase of biological pathways.,” *Nucleic Acids Res.*, vol. 33, no. Database issue, pp. D428–32, Jan. 2005.

[6] H. Toyoshiba, T. Yamanaka, H. Sone, F. M. Parham, N. J. Walker, J. Martinez, and C. J. Portier, “Gene Interaction Network Suggests Dioxin Induces A Significant Linkage Between Ah-Receptor and Retinoic Acid Receptor Beta,” *Environ. Health Perspect.*, vol. 1121, no. 1217, 2004.

[7] W. Huber, A. Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, “Variance stabilization applied to microarray data calibration andto the quantification of differential expression,” *Bioinformatics*, vol. 18, no. 1997, pp. S96–S104, 2002.

[8] W. Huber, A. Heydebreck, H. Sultmann, A. Poustka, and M. Vingron, “Statistical Applications in Genetics and Molecular Biology Parameter estimation for the calibration and variance stabilization of microarray data Parameter estimation for the calibration and,” *Bioinformatics*, vol. 2, no. 1, 2003.

[9] B. Schölkopf, K. Tsuda, and J.-P. Vert, *Kernel Methods in Computational Biology*. The MIT Press, 2004.

[10] R. A. Shimkets, *Gene Expression Profiling Methods and Protocols*, vol. 258. 2004.

[11] J. C. Alwine, D. J. Kemp, and G. R. Stark, “Method for detection of specific RNAs in agarose gels by transfer to diazobenzyloxymethyl-paper and hybridization with DNA probes.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 74, no. 12, pp. 5350–4, Dec. 1977.

[12] J. Aach, W. Rindone, and G. M. Church, “Systematic management and analysis of yeast gene expression data.,” *Genome Res.*, vol. 10, no. 4, pp. 431–45, Apr. 2000.

[13] P. Liang and A. B. Pardee, “Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction,” *Science (80-.).*, vol. 257, pp. 967–971, 1992.

[14] R. L. Nussbaum, R. R. McInnes, and H. F. Willard, *Thompson & Thompson. Genética en medicina*. 2007, p. 600.

[15] Z. Wang, M. Gerstein, and M. Snyder, “RNA-Seq: a revolutionary tool for transcriptomics,” *Nature Reviews Genetics*, 2009.

[16] L. David, W. Huber, M. Granovskaia, J. Toedling, C. J. Palm, L. Bofkin, T. Jones, R. W. Davis, and L. M. Steinmetz, “A high-resolution map of transcription in the yeast genome.,” *Proc. Natl. Acad. Sci. U. S. A.*, vol. 103, no. 14, pp. 5320–5, Apr. 2006.

[17] K. Yamada, J. Lim, J. M. Dale, H. Chen, P. Shinn, C. J. Palm, A. M. Southwick, H. C. Wu, C. Kim, M. Nguyen, P. Pham, R. Cheuk, G. Karlin-Newmann, S. X. Liu, B. Lam, H. Sakano, T. Wu, G. Yu, M. Miranda, H. L. Quach, M. Tripp, C. H. Chang, J. M. Lee, M. Toriumi, M. M. H. Chan, C. C. Tang, C. S. Onodera, J. M. Deng, K. Akiyama, Y. Ansari, T. Arakawa, J. Banh, F. Banno, L. Bowser, S. Brooks, P. Carninci, Q. Chao, N. Choy, A. Enju, A. D. Goldsmith, M. Gurjal, N. F. Hansen, Y. Hayashizaki, C. Johnson-Hopson, V. W. Hsuan, K. Iida, M. Karnes, S. Khan, E. Koesema, J. Ishida, P. X. Jiang, T. Jones, J. Kawai, A. Kamiya, C. Meyers, M. Nakajima, M. Narusaka, M. Seki, T. Sakurai, M. Satou, R. Tamse, M. Vaysberg, E. K. Wallender, C. Wong, Y. Yamamura, S. Yuan, K. Shinozaki, R. W. Davis, A. Theologis, and J. R. Ecker, *Empirical analysis of transcriptional activity in the Arabidopsis genome.*, vol. 302, no. 5646. 2003, pp. 842–6.

[18] M. J. Okoniewski and C. J. Miller, “Hybridization interactions between probesets in short oligo microarrays lead to spurious correlations.,” *BMC Bioinformatics*, vol. 7, no. Mm, p. 276, Jan. 2006.

[19] U. Alon, “Biological networks: the tinkerer as an engineer.,” *Science*, vol. 301, no. 5641, pp. 1866–7, Sep. 2003.

[20] F. Masulli and S. Mitra, “Natural computing methods in bioinformatics: A survey,” *Inf. Fusion*, vol. 10, no. 3, pp. 211–216, Jul. 2009.

[21] S. Das, D. Caragea, S. M. Welch, and W. H. Hsu, *Handbook of Research on Computational Methodologies in Gene Regulatory Networks*. IGI Global, 2009.

[22] S. Mitra, R. Das, H. Banka, and S. Mukhopadhyay, “Gene interaction – An evolutionary biclustering approach,” *Inf. Fusion*, vol. 10, no. 3, pp. 242–249, Jul. 2009.

[23] A. Rajaraman and J. D. Ullman, *Mining of Massive Datasets*. 2011, p. 340.

[24] D. K. Slonim, “From patterns to pathways: gene expression data analysis comes of age.,” *Nat. Genet.*, vol. 32 Suppl, no. december, pp. 502–8, Dec. 2002.

[25] T. S. Gardner, D. di Bernardo, D. Lorenz, and J. J. Collins, “Inferring genetic networks and identifying compound mode of action via expression profiling.,” *Science*, vol. 301, no. 5629, pp. 102–5, Jul. 2003.

[26] B. Scholkopf and A. J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. Cambridge, MA, USA: MIT Press, 2001.

[27] N. Dehak, R. Dehak, K. Patrick, N. Brummer, P. Ouellet, and P. Dumouchel, *Support Vector Machines versus Fast Scoring in the Low Dimensional Total Variability Space for Speaker Verification*. 2009.

[28] M. M. Babu, “Biological Databases and Protein Sequence Analysis,” 1986.

[29] H. Ogata, S. Goto, K. Sato, W. Fujibuchi, H. Bono, and M. Kanehisa, “KEGG: Kyoto Encyclopedia of Genes and Genomes,” *Nucleic Acids Res*, vol. 27, no. 1, pp. 29–34+.

[30] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock, “Gene Ontology: tool for the unification of biology,” *Nat Genet*, vol. 25, no. 1, pp. 25–29, May 2000.

[31] T. Consortium Gene Ontolgy, M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, G. M. Rubin, and G. Sherlock, “Gene Ontology : tool for the unification of biology,” vol. 25, no. 1, pp. 25–29, 2011.

[32] D. Szklarczyk, A. Franceschini, M. Kuhn, M. Simonovic, A. Roth, P. Minguez, T. Doerks, M. Stark, J. Muller, P. Bork, L. J. Jensen, and C. von Mering, “The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored.,” *Nucleic Acids Res.*, vol. 39, no. Database issue, pp. D561–8, Jan. 2011.

[33] K.-J. Kim and S.-B. Cho, “Ensemble classifiers based on correlation analysis for DNA microarray classification,” *Neurocomputing*, vol. 70, no. 1–3, pp. 187–199, Dec. 2006.

[34] A. K. Jain and R. C. Dubes, *Algorithms for clustering data*. 1988.

[35] B. W. Higgs, M. Elashoff, S. Richman, and B. Barci, “An online database for brain disease research.,” *BMC Genomics*, vol. 7, p. 70, Jan. 2006.

[36] D. Jiang, “Cluster Analysis for Gene Expression Data : A Survey,” *Technology*, pp. 1–40, 2004.

[37] S. Datta and S. Datta, “Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.,” *BMC Bioinformatics*, vol. 7, p. 397, Jan. 2006.

[38] I. Priness, O. Maimon, and I. Ben-Gal, “Evaluation of gene-expression clustering via mutual information distance measure,” *BMC Bioinformatics*, vol. 8, no. 1. p. 111, Mar-2007.

[39] R. Birnie, S. D. Bryce, C. Roome, V. Dussupt, A. Droop, S. H. Lang, P. a Berry, C. F. Hyde, J. L. Lewis, M. J. Stower, N. J. Maitland, and A. T. Collins, “Gene expression profiling of human prostate cancer stem cells reveals a pro-inflammatory phenotype and the importance of extracellular matrix interactions.,” *Genome Biol.*, vol. 9, no. 5, p. R83, Jan. 2008.

[40] L. I. Smith, *A tutorial on Principal Components Analysis*. 2002, p. 27.

[41] A. Manuscript, “Heterogeneity of tumor-induced gene expression changes in the human metabolic network,” *Nat. Biotechnol.*, vol. 31, no. 6, pp. 522–529, 2013.

[42] T. M. Mitchell, *Machine Learning*. 1997, p. 419.

[43] B. E. Boser, T. B. Laboratories, I. M. Guyon, and V. N. Vapnik, “A Training Algorithm for Optimal Margin Classi ers,” *Proc. 5th ACM Workshop on Computational Learning Theory*, pp. 44 –152, 1992.

[44] Y. Yamanishi, M. Araki, A. Gutteridge, W. Honda, and M. Kanehisa, “Prediction of drug-target interaction networks from the integration of chemical and genomic spaces.,” *Bioinformatics*, vol. 24, no. 13, pp. i232–40, Jul. 2008.

[45] K. Bleakley and Y. Yamanishi, “Supervised prediction of drug-target interactions using bipartite local models.,” *Bioinformatics*, vol. 25, no. 18, pp. 2397–403, Sep. 2009.

[46] A. Mortazavi, B. A. Williams, K. Mccue, L. Schaeffer, and B. Wold, “Mapping and quantifying mammalian transcriptomes by RNA-Seq,” *Nat. Methods*, vol. 5, no. 7, pp. 1–8, 2008.

[47] H. Jiang and W. H. Wong, “SeqMap: mapping massive amount of oligonucleotides to the genome.,” *Bioinformatics*, vol. 24, no. 20, pp. 2395–6, Oct. 2008.

[48] a. E. Kel, “MATCHTM: a tool for searching transcription factor binding sites in DNA sequences,” *Nucleic Acids Res.*, vol. 31, no. 13, pp. 3576–3579, Jul. 2003.

[49] S. K. Palaniswamy, S. James, H. Sun, R. S. Lamb, R. V Davuluri, and E. Grotewold, “AGRIS and AtRegNet . A Platform to Link cis-Regulatory Elements and Transcription Factors into Regulatory Networks 1 [W][OA],” *Bioinformatics*, vol. 140, no. March, pp. 818–829, 2006.

[50] J. Ruan, A. K. Dean, and W. Zhang, “A general co-expression network-based approach to gene expression analysis: comparison and applications.,” *BMC Syst. Biol.*, vol. 4, p. 8, Jan. 2010.

[51] M. Kanehisa, M. Araki, S. Goto, M. Hattori, M. Hirakawa, M. Itoh, T. Katayama, S. Kawashima, S. Okuda, T. Tokimatsu, and Y. Yamanishi, “KEGG for linking genomes to life and the environment.,” *Nucleic Acids Res.*, vol. 36, no. Database issue, pp. D480–4, Jan. 2008.

[52] M. Kanehisa, S. Goto, Y. Sato, M. Furumichi, and M. Tanabe, “KEGG for integration and interpretation of large-scale molecular data sets.,” *Nucleic Acids Res.*, vol. 40, no. Database issue, pp. D109–D114, Jan. 2012.

[53] R. M. Smon, E. L. Korn, L. M. McShane, M. D. Radmacher, G. W. Wright, and Y. Zhao, *Design and Analysis of DNA microarrays*. 2003.

[54] J. Vilo, a Brazma, I. Jonassen, a Robinson, and E. Ukkonen, “Mining for putative regulatory elements in the yeast genome using gene expression data.,” *Proc. Int. Conf. Intell. Syst. Mol. Biol.*, vol. 8, pp. 384–94, Jan. 2000.

[55] P. D’haeseleer, S. Liang, and R. Somogyi, “Genetic network inference: from co-expression clustering to reverse engineering.,” *Bioinformatics*, vol. 16, no. 8, pp. 707–26, Aug. 2000.

[56] H. J. Bierens, “Introduction to Hilbert Spaces,” pp. 1–18, 2007.

[57] I. Fasel, “Kernel PCA Scholkopf , Smola and Muller: Nonlinear Component Analysis as a Kernel Eigenvalue Problem,” 2001.

[58] D. D.´le´and P. Kastner, “Fuzzy C-means method for clustering microarray data,” *Bioinformatics*, vol. 19, pp. 973–980, 2003.

[59] J. Han and M. Kamber, *Data Mining Concepts and Techniques*. 2006, p. 772.

[60] A. Gupta, C. D. Maranas, and R. Albert, “Elucidation of directionality for co-expressed genes: predicting intra-operon termination sites.,” *Bioinformatics*, vol. 22, no. 2, pp. 209–14, Jan. 2006.

[61] L. L. Elo, H. Järvenpää, M. Oresic, R. Lahesmaa, and T. Aittokallio, “Systematic construction of gene coexpression networks with applications to human T helper cell differentiation process.,” *Bioinformatics*, vol. 23, no. 16, pp. 2096–103, Aug. 2007.

[62] F. Luo, Y. Yang, J. Zhong, H. Gao, L. Khan, D. K. Thompson, and J. Zhou, “Constructing gene co-expression networks and predicting functions of unknown genes by random matrix theory.,” *BMC Bioinformatics*, vol. 8, p. 299, Jan. 2007.

[63] A. Abdeen, J. Schnell, and B. Miki, “Transcriptome analysis reveals absence of unintended effects in drought-tolerant transgenic plants overexpressing the transcription factor ABF3,” *BMC Genomics*, vol. 11, no. 1, p. 69+, Jan. 2010.

[64] T. Yoshida, Y. Fujita, H. Sayama, S. Kidokoro, K. Maruyama, J. Mizoi, K. Shinozaki, and K. Yamaguchi-Shinozaki, “AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation,” *Plant J.*, vol. 61, no. 4, pp. 672–685, Nov. 2010.

[65] K. Vandepoele, M. Quimbaya, T. Casneuf, L. De Veylder, and Y. Van de Peer, “Unraveling transcriptional control in Arabidopsis using cis-regulatory elements and coexpression networks.,” *Plant Physiol.*, vol. 150, no. 2, pp. 535–46, Jun. 2009.

[66] H. Abe, K. Yamaguchishinozaki, T. Urao, T. Iwasaki, D. Hosokawa, and K. Shinozaki, “Role of Arabidopsis MYC and MYB homologs in drought- and abscisic acid-regulated gene expression,” *Plant Cell*, vol. 9, no. 10, pp. 1859–1868, 1997.

[67] K. Heidrich, K. Tsuda, S. Blanvillain-Baufumé, L. Wirthmueller, J. Bautor, and J. E. Parker, “Arabidopsis TNL-WRKY domain receptor RRS1 contributes to temperature-conditioned RPS4 auto-immunity.,” *Front. Plant Sci.*, vol. 4, no. October, p. 403, Jan. 2013.

[68] G. R. Teakle, I. W. Manfield, J. F. Graham, and P. M. Gilmartin, “Arabidopsis thaliana; GATA factors: organisation, expression and DNA-binding characteristics,” *Plant Mol. Biol.*, vol. 50, no. 1, pp. 43–56, 2002.

[69] D. S. Lee and J. L. Kalb, “Network Topology Analysis,” no. January, 2008.

[70] A. Alexeyenko, W. Lee, M. Pernemalm, J. Guegan, P. Dessen, V. Lazar, J. Lehtiö, and Y. Pawitan, “Network enrichment analysis: extension of gene-set enrichment analysis to gene networks.,” *BMC Bioinformatics*, vol. 13, p. 226, Jan. 2012.

[71] P. Shannon, A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, “Cytoscape : A Software Environment for Integrated Models of Biomolecular Interaction Networks,” pp. 2498–2504, 2003.

[72] M. Chagoyen and F. Pazos, “Quantifying the biological significance of gene ontology biological processes--implications for the analysis of systems-wide data.,” *Bioinformatics*, vol. 26, no. 3, pp. 378–84, Feb. 2010.

[73] T. Schank and D. Wagner, “Approximating Clustering Coefficient and Transitivity Basic Definitions,” vol. 9, no. 2, pp. 265–275, 2005.