**UNIVERSIDAD DE CHILE**
FACULTAD DE CIENCIAS

# REDES DE COEXPRESIÓN EN LA BÚSQUEDA DE NUEVOS GENES INVOLUCRADOS EN EL METABOLISMO DE LA PARED CELULAR

Tesis entregada a la
Universidad de Chile
en cumplimiento parcial de los requisitos
para optar al grado de

Doctor en Ciencias con Mención en Biología Molecular, Celular y Neurociencias

por

**Alexander Vergara Robles**



Septiembre, 2009
Santiago - Chile

Director de Tesis: Dr. Ariel Orellana López

Co-Director de Tesis: Dr. Rodrigo Gutiérrez Ilabaca

**UNIVERSIDAD DE CHILE**
**FACULTAD DE CIENCIAS**

# INFORME DE APROBACION

# TESIS DE DOCTORADO

**Se informa a la Escuela de Postgrado de la Facultad de Ciencias que la Tesis de Doctorado presentada por el candidato.**

**Alexander Vergara Robles**

**Ha sido aprobada por la comisión de Evaluación de la tesis como requisito para optar al grado de Doctor en Ciencias con mención en Biología Molecular Celular y Neurociencias, en el examen de Defensa Privada de Tesis rendido el día 9 de Julio del año 2009.**

**Director de Tesis:**
**Prof. Ariel Orellana López**

**Co-Director de Tesis**
**Prof. Rodrigo Gutiérrez Ilabaca**

**Comisión de Evaluación de la Tesis**

**Prof. Marco Tulio Nuñez (Presidente)**

**Prof. Juan Carlos Letelier**

**Prof. Mauricio González**

**Prof. Michael Handford**

# DEDICATORIA

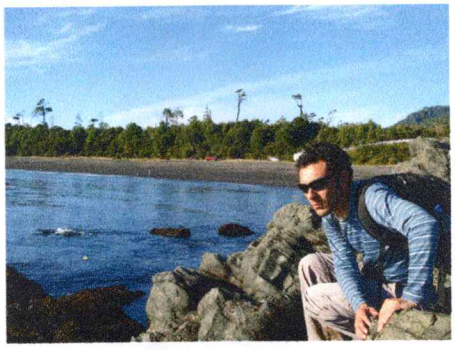
Dedico este trabajo con mucho cariño y amor a toda mi familia y de forma especial a mi amada Vany.

TRINITY: You came here because you wanted to know the answer to a hacker's question.

NEO: The Matrix. What is the Matrix?

TRINITY: Twelve years ago I met a man, a great man, who said that no one could be told the answer to that question. That they had to see it, to believe it. He told me that no one should look for the answer unless they have to because once you see it, everything changes. Your life and the world you live in will never be the same. It's as if you wake up one morning and the sky is falling.....
The truth is out there, Neo. It's looking for you and it will find you, if you want it to....

**THE MATRIX.**



Playa Santa Barbara -Chaitén. Febrero 2008.

Alexander Vergara nació el 6 de abril de 1979 en Santiago de Chile. Estudió Ingeniería en Biotecnología Molecular en la Facultad de Ciencias de la Universidad de Chile, carrera en la que mostró tempranamente su interés por la ciencia fundamental, siendo alumno en pasantía en el antiguo "laboratorio de membranas" ya en su primer año de estudios universitarios. En este laboratorio ubicado en "las barracas", trabajó en el joven grupo del Dr. Ariel Orellana y también con el grupo del Dr. Tulio Nuñez. Allí, conoció el ambiente de un laboratorio de investigación y lo emocionante de realizar experimentos. Posteriormente, Alexander realizó su tesis de pregrado en el laboratorio del Dr. Orellana, donde trabajó en ciencia fundamental con mucho entusiasmo. El año 2004, ingresó al exigente programa de doctorado de Biología Molecular Celular y Neurociencias de la misma facultad donde se formó como estudiante de pregrado. Hoy, luego de un largo camino esta feliz de terminar su tesis. La primera tesis en "Systems Biology" de este programa de doctorado.

## AGRADECIMIENTOS

Quiero agradecer profundamente a Ariel Orellana, mi tutor de tesis, por permitir que cambiara mi primer proyecto de doctorado, por la libertad y el apoyo que me ha entregado. Muchas gracias Ariel por tus críticas y guía durante todos estos años. También quiero agradecer a Rodrigo Gutiérrez, mi co-tutor de tesis, por todo lo que me ha enseñado y guiado durante este trabajo. Gracias Rodrigo por aceptar guiar este trabajo, por los trucos y la disciplina.

Agradezco a todos los miembros del *Centro de Biotecnología Vegetal* de la Universidad Andrés Bello y a los miembros del *"Plant Systems Biology Lab"* de la Pontificia Universidad Católica de Chile, por todos sus comentarios y correcciones en todas las presentaciones que realicé de este trabajo en las reuniones de grupo.

Agradezco a mi familia y a Maribel Donoso por todo el apoyo que me han entregado en los momentos difíciles. Muchas gracias por la paciencia y comprensión.

Por último, agradezco a todos los amigos de club social y deportivo OPA. Con las conversaciones científicas que tuvimos, siempre pude explorar nuevas ideas y encontrar apoyo cuando todo parecía completamente perdido.

## FINANCIAMIENTO

* Millennium Nucleus in Plant Cell Biotechnology

* Millennium Nucleus for Plant Functional Genomics

* Beca de Postgrado CONICYT, 2004, #21040486. PROGRAMA NACIONAL DE BECAS DE POSTGRADO.

* Beca pasantías en el extranjero CONICYT, 2005.

* Beca asistencia a cursos cortos en el extranjero CONICYT, 2007.

# ÍNDICE

# ÍNDICE DE TABLAS

# ÍNDICE DE FIGURAS

## ABREVIATURAS

*CC* : Coeficiente de Clustering

<*CC*> : Coeficiente de Clustering promedio

CPAN : Comprehensive Perl Archive Network

DAG: Directed Acyclic Graph

GB: Gigabyte

GHz: Gigahertz

GO: Gene Ontology

*K* : Grado

<*k*> : Grado promedio

Kb : Kilo bases

Mb : Mega bases

RMA : Robust Multiarray Analysis

TAIR : The Arabidopsis Information Resource

TF : Factor de transcripción

TNA : Transportador de nucleótidos azúcar

## RESUMEN

La matriz extracelular vegetal o pared celular, es una red compleja constituida por una red entrecruzada de glicoproteínas y polisacáridos. Esta estructura participa activamente en la comunicación célula-célula, ciclo celular, morfogénesis y desarrollo. Además, constituye una barrera de protección frente al ataque de patógenos. Pese que conocemos bastante de la estructura y composición de la pared celular, desconocemos la identidad de la mayoría de las proteínas involucradas en su metabolismo. De hecho, en *Arabidopsis thaliana,* el primer organismo vegetal con su genoma completamente secuenciado, hasta la fecha solamente alrededor de un 50% de los genes de este organismo poseen algún proceso biológico asignado.

El objetivo de este trabajo es encontrar nuevos genes involucrados en el metabolismo de la pared celular, poniendo especial énfasis en el análisis de aquellos genes implicados en el metabolismo de hemicelulosas y pectinas, polímeros de azúcar de la pared celular que son sintetizados en la ruta secretoria.

En este trabajo, hemos realizado un análisis global de expresión génica en *Arabidopsis thaliana*. Utilizando 1.701 microarreglos públicos ATH1 Affymetrix™, construimos redes de correlación de expresión, donde los nodos representan los genes y los arcos entre éstos la similitud entre sus perfiles de expresión. Análisis de topología indican que las redes analizadas poseen

propiedades distintas a una red aleatoria, con características de red *"pequeño mundo"*, donde la mayoría de los nodos de la red pueden ser conectados entre sí con unos pocos arcos. Además, la distribución de frecuencia del número de vecinos que posee cada gen en nuestras redes de coexpresión es "scale free", donde hay muy pocos nodos altamente conectados. Sorprendentemente, un número importante de los "hubs" o nodos altamente conectados por coexpresión no tienen función molecular conocida. Además, en estos genes hemos encontrado un enriquecimiento negativo en mutantes insercionales de T-DNA que poseen ambas copias de los genes alteradas, lo que sugiere que no es posible obtener plantas viables de dichas líneas insercionales homocigotas, por la letalidad que produce dicha mutación en estos genes, los cuales participan en muchos procesos biológicos.

Con el objetivo de proponer nuevos genes involucrados en el metabolismo de la pared celular, desarrollamos un algoritmo que en base a las características del vecindario de un gen en la red de coexpresión, es capaz de asignarle una función en cierto proceso biológico en particular. Para ello, utilizamos las anotaciones de procesos biológicos (BP) de Gene Ontology (GO). Utilizando un método de asignación de función que cuenta las anotaciones presentes en el vecindario de un gen y las pondera según sus distancias, podemos propagar la anotación de los genes con el proceso biológico GO "Organización de la pared celular y su biogénesis" a 43 nuevos genes que previamente no tenían esta etiqueta asignada. Estas predicciones, tienen una certeza mínima esperada de ser correctas de aproximadamente el

48.92%, según análisis de validación cruzada que realizamos en este trabajo. De los 43 genes, 7 no poseen ninguna descripción en sus anotaciones que los relacione a algún proceso biológico o función molecular ("unknown proteins"), lo que los hace muy interesantes por tratarse de genes poco estudiados y de los que aún sabemos muy poco. Estos genes no son implicados en el metabolismo de la pared celular por análisis de identidad de secuencia, el cual puede ser considerado un método paralelo de asignación de función génica.

Hoy, nos encontramos en la era post-genómica. Tenemos las secuencias de los genes y nos enfrentamos ante el gran desafío de dilucidar la función de cada uno de los genes codificantes para proteínas. El procedimiento bioinformático seguido en este trabajo, ha sido desarrollado para encontrar nuevos genes involucrados en el metabolismo de la pared celular. Sin embargo, puede además ser utilizado para propagar anotaciones de muchos otros procesos biológicos y puede ser considerado como un "screening *in silico*", que propone nuevos genes involucrados en cierto proceso biológico en particular, dando además una estimación de la certeza de las predicciones, las cuales fueron calculadas utilizando los genes de anotación conocida como grupo de entrenamiento. Esto, permite evaluar las predicciones según la oportunidad que tengan éstas de ser verificadas experimentalmente y trabajar con un número de genes más reducido de los que podrían ser seleccionados para estudios de genómica funcional en base a análisis de identidad de secuencias, tomando una aproximación centrada en el proceso biológico en sí mismo y no únicamente en las funciones moleculares posiblemente involucradas.

## ABSTRACT

The plant extracellular matrix or cell wall, is a complex network of glycoproteins and polysaccharides that is involved in essential plant cell processes, such as cell-cell interaction, cellular cycle, morphogenesis, development and pathogen defense mechanisms. These cell wall polysaccharides can be broadly classified as cellulose, hemicelluloses and pectins. While cellulose biosynthesis take place at the plasma membrane, the biosynthesis of hemicelluloses and pectins occurs inside the Golgi lumen by the action of glycosyltransferases that incorporate sugars into acceptors.

The *Arabidopsis thaliana* genome was the first to be sequenced and is the best characterized among plants. However, around 50% of the genes still have not a biological process assigned and their annotation is a key challenge in functional genomics. The objective of this work is to identify new candidate genes involved in cell wall metabolism, with a special focus on genes playing a role in the metabolism of hemicellulose and pectin.

The DNA microarray technology is currently the most widely used approach for monitoring genome-wide gene expression changes in model organisms like Arabidopsis. In this work, using publicly available microarray data we performed a global pair-wise linear correlation analysis of expression profiles using 1.701 quality-screened *Affymetrix*™ ATH1 microarray chips and constructed gene co-expression networks. Topology analysis of these networks

shows that they are fundamentally non-random, with "scale-free" degree distribution and "small world" characteristics. These networks show many genes with a low number of connections and a few highly connected genes called "hubs". Hubs in scale-free networks are essential for the function of the network. However, we found that many Arabidopsis co-expression hubs are genes with unknown molecular function, suggesting that we know little of these potentially key genes. In addition, there is a negative enrichment in homozygous T-DNA mutant plants on these genes, suggesting a lethal gene-deletion phenotype on these essential genes and their involvement on key biological processes.

Today, the recent availability of gene expression profiles across many microarrays, allow us to propose gene annotations by different methods, which can be evaluated using the annotations recall of a training set, like an information retrieval index to optimize it. In order to propose new cell wall genes involved, we tested three neighbourhood based methods using Gene Ontology (GO) terms to spread the annotations in the networks. We obtained the best recall of annotations using a counting based method that consider the correlation values between the gene query and its neighbours. Using this method, in this work we proposed 43 new genes involved in the cell wall metabolism, spreading the GO term *GO:0007047 "cell wall organization and biogenesis"* in the co-expression networks, with a precision of 48.92% in the cross validation analysis. Of the proposed genes, 7 still not have a biological process or molecular functions assigned in its annotations ("unknown proteins"), and are very interesting if the focus is to find new cell wall metabolism genes.

These genes could not be related to cell wall metabolism by only using sequence analysis, a gene function prediction method that can consider a parallel method.

Today is the post-genomic era. We have genes sequences and the next step is to know all the genes functions. The bioinformatic approach presented in this work, using massive expression data, was developed in order to find new cell wall metabolism genes. However, it can be considered like a "screening in sillico", that propose the involvement of genes in a given biological process, providing information on the certainty of the recall values obtained in the cross validation at the selected settings, using genes with a biological process already assigned as a training set. This approach, with a focused biological process, increases the chance to confirm the prediction since it produces a low number of proposed genes in comparison to an approach that only consider sequence comparison analysis.

# INTRODUCCIÓN

## La pared celular vegetal

Todas las células vegetales poseen una matriz extracelular denominada pared celular, la cual es esencial para muchos procesos fisiológicos y de desarrollo de la planta. Esta estructura, entrega resistencia mecánica a los tejidos vegetales y regula el crecimiento, volumen y morfología celular. Además, está involucrada en fenómenos de comunicación célula-célula y constituye una barrera de protección frente al ataque de patógenos (Carpita y col. 1993). La pared celular es una estructura compleja formada por polímeros de azúcar que junto a glicoproteínas conforman una entrecruzada red. Si bien la composición de esta matriz extracelular varía según los distintos tejidos, se ha clasificado en dos tipos: la pared celular primaria y secundaria. La pared celular primaria se forma durante el crecimiento de todas las células vegetales, es una estructura no especializada compuesta principalmente de hemicelulosas y pectinas. La pared secundaria se forma una vez que el crecimiento celular ha cesado y sólo la poseen algunos tipos celulares. Se caracteriza por poseer un mayor contenido de celulosa, menor cantidad de pectinas, hemicelulosas y por poseer lignina (Carpita y col. 1993; Somerville y col. 2004). Las hemicelulosas son polímeros de azúcares neutros ramificados que interactúan con celulosa, siendo el xiloglucano el más abundante de estos polímeros en *Arabidopsis thaliana*. El

xiloglucano está constituido por una cadena lineal de glucosas β 1-4 sustituidas por residuos de α 1-6 xilosa. Estos residuos de xilosa pueden estar sustituidos en posiciones específicas por residuos de β 1-2 galactosa, los que a su vez pueden ser sustituidos por residuos de α 1-2 fucosa (Scheible y col. 2004). Las pectinas en cambio, se caracterizan por ser polímeros constituidos principalmente por ácidos urónicos. El Homogalacturonano (HG) es el más simple de ellos y está constituido por una cadena no ramificada de α 1-4 ácido galacturónico. Ramnogalacturonano I y II (RG-I ; RG-II) poseen estructuras más complejas y ramificadas, incluyendo en su composición además de ácido galacturónico, ramnosa, galactosa y arabinosa (Somerville y col. 2004).

Se estima que en *Arabidopsis thaliana* varios cientos de proteínas están involucradas en el metabolismo biosintético y en la remodelación dinámica de la pared celular, la mayoría de las cuales son codificadas por familias génicas con significativas similitudes estructurales, pero con distintos perfiles de expresión entre sus miembros (Somerville y col. 2004; Imoto y col. 2005). Debido a esto, una abstracción a nivel molecular del metabolismo de la pared celular necesariamente implica, analizar la acción de todos los distintos genes que codifican para las proteínas que participan en este proceso biológico del cual aún conocemos muy poco. Hasta la fecha, todas de las proteínas involucradas en el metabolismo de la pared celular, han sido descubiertas mediante el análisis de plantas obtenidas por genética reversa o mutagénesis, que muestran alteraciones en la composición de azúcares de la pared celular.

Así, basándose en análisis de identidad de secuencias con estos genes, se han definido diversas familias génicas cuyo supuesto papel biológico está asociado al metabolismo de la pared celular (Girke y col. 2004). Sin embargo, ¿Están realmente todos los miembros de estas familias génicas asociados al mismo proceso biológico? Debemos mencionar que el denominado "papel biológico" de una proteína no radica en su secuencia u estructura propiamente tal, sino en las interacciones que ésta tiene con otras macromoléculas en un proceso dinámico. Además, la secuencia de un gen puede dar pistas de la función molecular de su producto proteico, pero si se estudia un proceso metabólico en particular, es necesario identificar los genes involucrados en este proceso metabólico, utilizando una aproximación que considere las interacciones funcionales entre las proteínas. Este tipo de análisis de interacciones, a diferencia de los análisis de secuencia, se enfoca en el proceso en que está implicado un producto génico más que en su función molecular, pues un proceso metabólico puede utilizar muchas funciones moleculares distintas en sus distintas reacciones. De hecho, en el metabolismo de la pared celular participan distintos tipos de funciones moleculares. El primer paso en la ruta de biosíntesis de sus polímeros de azúcar, es la conversión de las moléculas asimiladas en la fotosíntesis en moléculas de nucleótidos-azúcar, para lo cual existen intrincadas vías de fosforilación e interconversión metabólica en el citosol (Bonin y col. 1997; Dormann y col. 1998). Las enzimas responsables de la biosíntesis de los elementos no celulósicos de la pared celular, como las hemicelulosas y pectinas propiamente tal, son glicosiltransferasas localizadas

subcelularmente en el aparato de Golgi que se caracterizan por ser proteínas de sólo un dominio transmembrana y poseer su sitio catalítico orientado hacia el lumen de las cisternas del Golgi (Keegstra y col. 2001). Estas enzimas, utilizan como sustrato nucleótidos azúcar que son sintetizados en el citosol y que pueden ser incorporados al lumen del Golgi gracias a la acción de transportadores de nucleótidos azúcar (Bonin y col. 1997; Dormann y col. 1998) **(Figura 1).**

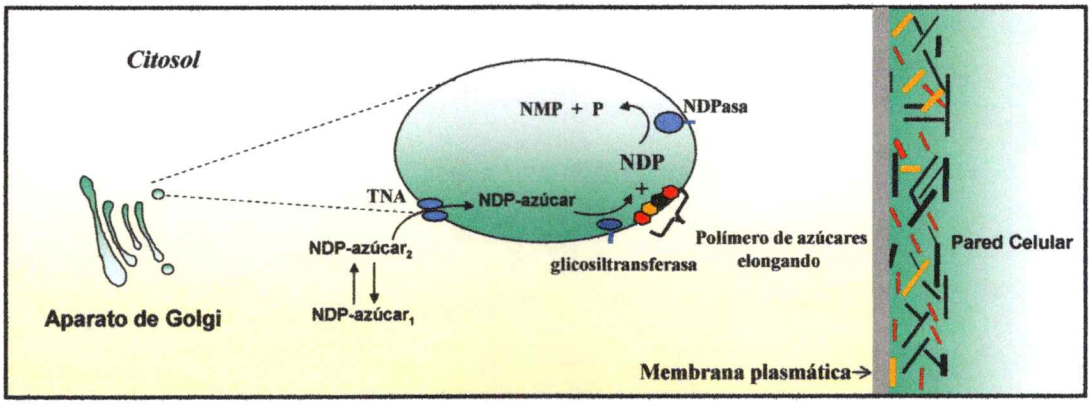



**Figura 1. Modelo propuesto para la biosíntesis de hemicelulosas y pectinas en las células vegetales**. Los nucleótidos azúcar (NDP-azúcar) son metabolitos sintetizados en el citoplasma y transportados hacia el lumen del aparato de Golgi mediante transportadores de nucleótidos azúcar (TNA). Allí, pueden ser utilizados por las glicosiltransferasas, enzimas que se localizan en la membrana de este organelo y que poseen su sitio catalítico orientado hacia el lumen. Estas enzimas, transfieren azúcares a cadenas de polisacáridos en elongación, los que luego son secretados al apoplasto. Los nucleótidos azúcar, son metabolizados en el citosol y algunos pueden ser interconvertidos entre sí por la acción de epimerasas y dehidratasas citosólicas.

## 1. El problema de estudio

### Estudio del metabolismo de hemicelulosas y pectinas

En este proyecto, abordamos el estudio del metabolismo de la pared celular utilizando una aproximación de biología de sistemas, poniendo especial énfasis en el análisis del metabolismo de hemicelulosas y pectinas. Polímeros de azúcar de la pared celular, que a diferencia de la celulosa que es sintetizada en la membrana plasmática, son sintetizados en el aparato de Golgi (Dupree y Sherrier, 1998). ¿Cómo podemos encontrar nuevos genes involucrados en este proceso metabólico? *Arabidopsis thaliana* fue el primer organismo vegetal con su genoma completamente secuenciado y gracias a su ciclo de vida corto y tamaño pequeño es un excelente modelo en genómica funcional. Sin embargo, aún alrededor del 50% de los genes de este organismo no poseen algún proceso biológico asignado.

Los microarreglos de DNA han revolucionado el monitoreo de la expresión génica en *Arabidopsis thaliana* y otros organismos modelo (Sherlock y col. 2001; Craigon y col. 2004; Parkinson y col. 2005). Hoy, la gran cantidad de datos generados con esta tecnología y su disponibilidad pública nos entregan más que nunca antes la posibilidad de analizar la expresión génica desde una perspectiva global. Desde que el proyecto de secuenciación del genoma de *Arabidopsis thaliana* finalizó el año 2000, numerosos tipos de análisis de secuencias y una gran cantidad de datos de expresión utilizando microarreglos ante diversos escenarios fisiológicos han sido reportados y analizados por distintos métodos (Craigon y col. 2004; Steinhauser y col. 2004;

Zimmermann y col. 2004; Obayashi y col. 2007; Horan y col. 2008; Srinivasasainagendra y col. 2008). Justamente el gran desafío de la era post genómica es relacionar los genes con los procesos biológicos en que participan y lograr integrar la numerosa información disponible. Tenemos las secuencias de los genes y mucha información experimental. Sin embargo, aún debemos trabajar en el análisis de esta información para lograr obtener información biológica relevante y comprender la función de los genes.

**Redes de coexpresión**

La construcción de redes y su análisis son una herramienta muy útil para caracterizar los sistemas biológicos, ya que entregan una poderosa herramienta de análisis de la información disponible. Las redes se representan como grafos, donde los nodos (o vértices) son genes, proteínas o metabolitos y los arcos entre éstos representan una relación entre ellos. El número de arcos que posee un nodo con sus nodos vecinos inmediatos en la red, corresponde a su grado (Barabasi y col. 2004).

Una propiedad estructural común entre muchas redes analizadas en distintas áreas de la ciencias, incluyendo las redes biológicas, es que todas ellas poseen la llamada distribución de los grados, escala libre o "scale free" (Jeong y col. 2000). Básicamente una red escala libre, es una red en la cual la distribución de frecuencia de los grados *(k)* de los nodos sigue una distribución en la cual la probabilidad de encontrar un nodo con cierto grado *(P(k))*, obedece a una ley de potencia del tipo: $P(k) \sim Ak^{-\gamma}$ , donde *A* es una constante de

normalización y donde el exponente negativo del grado $\gamma$, toma valores entre $1 < \gamma < 3$ según los datos que han sido reportados. Como una consecuencia de esta topología, las redes escala libre se caracterizan por presentar los llamados "hubs" o nodos altamente conectados, que poseen grados muy altos en comparación con el resto de los nodos de la red. Además, se ha propuesto que los altos valores de grado de los "hubs" y la formación de redes escala libre, puede ser explicada por la propiedad de unión preferencial (*"preferential attachment"*), pues en modelos de redes que crecen por la adición de nuevos nodos, éstos se conectan de forma preferencial a los más conectados. Los nodos que de esta forma adquieren un grado elevado, son pocos, pero resultan ser esenciales para la estructura de la red debido a su altísima conectividad con los otros nodos (Albert y col. 2000).

Analizando datos de transcriptomas e interactomas, se ha demostrado que la mayoría de los genes que son co-regulados transcripcionalmente codifican para proteínas que interactúan funcionalmente, formando parte de una misma vía o proceso metabólico (Ge y col. 2001; Obayashi y col. 2007; Obayashi y col. 2008). Así, el análisis de correlaciones de expresión y la construcción de redes transcripcionales se han transformado en una poderosa herramienta para ayudar a asignar función y a encontrar nuevas proteínas involucradas en cierto proceso biológico de forma independiente de los análisis de secuencia.

En este trabajo, con el objetivo de encontrar nuevos genes involucrados en el metabolismo de la pared celular, hemos confeccionado redes de

coexpresión en *Arabidopsis thaliana* utilizando datos públicos de microarreglos. En nuestras redes los nodos representan a los genes y los arcos entre éstos el nivel de correlación entre sus perfiles de expresión ante distintos escenarios fisiológicos de crecimiento (ver métodos).

Prácticamente es una generalidad, que las redes existentes en el mundo real, es decir, aquellas redes no aleatorias obtenidas al analizar con teoría de grafos distintos sistemas complejos, son robustas a la eliminación de un nodo elegido al azar. Por el contrario, si la eliminación del nodo es dirigida a un "hub", estas redes son muy sensibles a dicha alteración (Barabasi y col. 1999; Albert y col. 2000; Jeong y col. 2001). En nuestro caso, nos interesa analizar este tipo de información pues un gen altamente conectado o "hub" en nuestras redes, es un gen que coexpresa con muchos otros genes, lo que indica que su producto proteico participa probablemente en muchos procesos metabólicos. Así, los "hubs" son muy buenos candidatos para analizar el efecto sobre la planta de su mutación o deleción, esperándose un gran efecto fenotípico si dicha alteración es efectiva. En nuestro trabajo, nos interesa identificar aquellos nodos más conectados, por su potencial importancia en el metabolismo vegetal. Además, es posible por medio del grado, ordenar los genes de nuestro interés según su nivel de conexión con los otros genes y obtener así, dependiendo del tipo de red analizada, potenciales genes regulatorios o genes muy importantes en el metabolismo de la pared celular. Sabemos que hay genes involucrados en el metabolismo de la pared celular que no han sido descubiertos, pues la complejidad de los enlaces de los polímeros de azúcar de la pared celular

requiere un gran repertorio enzimático (Somerville y col. 2004). Debido a esto, proponer nuevos genes involucrados en la biosíntesis de la matriz extracelular vegetal es un gran aporte para la biología vegetal, pues entrega nuevos potenciales participantes en este proceso biológico esencial para la planta. Además, permite postular cuáles de ellos pueden originar un fenotipo más notorio al ser eliminados, basándonos simplemente en su grado en las redes de coexpresión. Luego, podemos caracterizar plantas con alteraciones en las secuencias de estos genes, pues al trabajar con *Arabidopsis thaliana* es posible obtener líneas de estas plantas que posean un elemento insercional en su genoma en la región codificante para un gen de interés y analizar estas líneas fenotípicamente. Bajo este escenario experimental, resulta muy útil reducir las listas de genes de interés, ya sea utilizando un método más exacto para asignar un papel a un gen en cierto proceso biológico ó métodos paralelos, para así lograr efectivamente reducir el número de líneas de plantas a caracterizar. Por ejemplo, hoy en *Arabidopsis thaliana*, en base a análisis de identidad de secuencia, se postula que alrededor del 2.5% de los genes que codifican para proteínas, puede estar participando en el metabolismo de la pared celular (Girke y col. 2004; Somerville y col. 2004). Sin embargo, ¿Realmente todos estos genes participan en el metabolismo de la pared celular? ¿Cómo podemos encontrar nuevos genes involucrados en este proceso metabólico, iluminando funcionalmente a aquellos genes de *Arabidopsis thaliana* que aún no poseen un papel en algún proceso biológico asignado y que quizás juega un papel fundamental en este metabolismo?

## 2. Hipótesis

Utilizando análisis global de datos de expresión génica en *Arabidopsis thaliana*, es posible identificar nuevos genes involucrados en el metabolismo de la pared celular vegetal.

## 3. Objetivo General

Identificar mediante el uso de datos públicos de microarreglos nuevos genes involucrados en el metabolismo de la pared celular.

## 4. Objetivos específicos

1.- Construir un modelo de redes de coexpresión basado en datos públicos de microarreglos.

2.- Realizar un análisis de vecindad en las redes de coexpresión para identificar nuevos genes involucrados en el metabolismo de la pared celular.

3.- Confeccionar un modelo que represente las redes regulatorias involucradas en el metabolismo de la pared celular.

# MATERIALES Y MÉTODOS

## 1. Análisis computacionales

### 1.1 Computadores y programas utilizados:

Para la realización de este trabajo, se utilizó como servidor Web un computador equipado con un procesador Intel Xeon de 3.2 GHz de 64 bits de cuatro núcleos independientes, con 4 GB de memoria RAM, utilizando el sistema operativo Fedora. Además, se utilizó un computador portátil con un procesador de 32 bits de 2 GHz Intel Centrino con 1 GB de memoria RAM, utilizando el sistema operativo Ubuntu. Para el análisis de los datos, se utilizó programas escritos en los lenguajes Perl (http://www.perl.org/), R (http://www.r-project.org/) y comandos de UNIX. El análisis de los datos de redes fue realizado con el programa Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/) utilizando Wine (http://www.winehq.org/).

## 2. Análisis de correlación de perfiles de expresión

### 2.1 Obtención de datos de expresión y selección de microarreglos de calidad

Datos de 1.887 hibridaciones de microarreglos ATH1 Affymetrix™ fueron obtenidos desde NASCArrays (Craigon y col. 2004) a mediados del año 2006 por medio del servicio de suscripción AffyWatch como archivos CEL (http://arabidopsis.info/). Las hibridaciones fueron filtradas para eliminar

aquellos chips que presenten problemas de calidad en sus datos, pues diversas son las posibles fuentes de error que introducen variaciones de origen no biológico en los datos de microarreglos (Wilson y col. 2005). Para identificar potenciales chips de baja calidad y eliminarlos de los análisis posteriores, utilizamos un método que elimina aquellos chips que pueden ser considerados aberrantes, al comparar la distribución de sus señales con la distribuciones observadas en múltiples observaciones tomadas como referencia (Persson y col. 2005). De las 1.887 hibridaciones iniciales obtenidas por nuestro laboratorio desde NASCArrays, 186 fueron seleccionadas aberrantes por este método y fueron eliminadas de los análisis realizados en este trabajo.

2.2 Normalización de datos de expresión

Las 1.701 hibridaciones que pasaron el filtro de calidad fueron normalizadas por el método Robust Multiarray Analysis (RMA) (Irizarry y col. 2003) disponible en los paquetes de análisis para R de Bioconductor (Gentleman y col. 2004).

2.3 Cálculos de correlaciones lineales entre perfiles de expresión

Utilizando todos los 22.810 grupos de sondas presentes en los chips de microarreglos ATH1 Affymetrix™, calculamos las correlaciones lineales entre todos los pares de perfiles de expresión obtenidos en los 1.701 chips filtrados por calidad seleccionados anteriormente. Las correlaciones lineales de Spearman y Pearson, fueron obtenidas con la función *cor* disponible en el lenguaje R (http://www.r-project.org/).

## 3. Construcción de redes de correlación de expresión

### 3.1 Selección de sondas mapeadas a locus

De los 22.810 grupos de sondas presentes en los microarreglos ATH1 Affymetrix™, se eliminaron 1.952 grupos de sondas, ya sea por hibridar inespecíficamente, ser sondas control incluidas en estos chips o por tratarse de casos en que el locus es mapeado por dos o más grupos de sondas. Los 20.858 grupos de sondas remanentes que mapean directamente a 1 locus cada una, fueron analizadas en este trabajo (TAIR 8) (Rhee y col. 2003; Swarbreck y col. 2008).

### 3.2 Visualización de las correlaciones en una red

Las correlaciones entre pares de perfiles de expresión entre todos los grupos de sondas seleccionadas fueron representadas como distancia (distancia = 1- correlación). Las redes fueron visualizadas con el programa para análisis de redes Pajek (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

## 4. Análisis de topología en redes de correlación de expresión

Distintos parámetros topológicos fueron determinados en las redes de coexpresión. El grado *(k)* y los coeficientes de clustering *(CC)* de cada gen de la red fueron calculados con el programa Pajek. El módulo de Perl *graph* fue utilizado para la identificación y análisis del número de componentes conexos en las redes. El módulo se encuentra disponible en CPAN (http://search.cpan.org/~jhi/Graph-0.91/lib/Graph.pod).

## 5. Análisis estadísticos con anotaciones

### 5.1 Manejo de anotaciones de Gene Ontology

Las anotaciones funcionales realizadas por el "Gene Ontology Consortium" (GO) (Ashburner y col. 2000) para cada locus del genoma de Arabidopsis fueron obtenidas desde la página oficial del proyecto (http://www.geneontology.org/). Las anotaciones para cada locus del genoma fueron homogenizadas a un mismo nivel de descripción en GO. El procesamiento y análisis de las anotaciones funcionales se realizó utilizando programas escritos en Perl y utilizando el módulo *GO::OntologyProvider::OntologyParser* disponible libremente en CPAN (http://search.cpan.org/~sherlock/GO-TermFinder-0.83/lib/GO/OntologyProvider/OntologyParser.pm).

### 5.2 Análisis de recuperación de anotaciones con grupos de entrenamiento

Para las validaciones de los distintos métodos y su comparación entre sí, utilizamos todos los genes con un proceso biológico asignado como grupo de entrenamiento, los cuales fueron obtenidos utilizando las anotaciones de Gene Ontology (GO) (Ashburner y col. 2000). Para realizar estos análisis, utilizamos el lenguaje Perl. Los verdaderos positivos (TP) corresponden a aquellos genes en que su etiqueta de anotación predicha concuerda con la que el gen posee a ese nivel en su descripción GO.

### 5.3 Análisis de enriquecimiento en una lista de genes

Los análisis de las anotaciones sobre representadas o enriquecidas en listas de genes en algún término funcional de Gene Ontology (GO), fueron

obtenidos utilizando el módulo de Perl *GO::Term-Finder* que se encuentra disponible en CPAN (http://search.cpan.org/~sherlock/GO-TermFinder-0.83/lib/GO/TermFinder.pm). Este módulo, utilizando la distribución hipergeométrica, permite evaluar si una determinada etiqueta GO se encuentra sobre representada en la lista de genes, al contrastar la frecuencia en que se encuentra en esta lista cada etiqueta GO y contrastando estas frecuencias con las esperadas por simple azar (Boyle y col. 2004).

5.4 Análisis de genes altamente conectados

En este trabajo, se seleccionó al 1% de los genes más conectado de cada red de coexpresión construida. Luego, en estas listas de genes, se realizó análisis de enriquecimiento de anotaciones. Además, en los nodos altamente conectados se analizó la proporción de estos genes que presentan una inserción en su genoma de T-DNA en ambas copias del gen. Para ello, se comparó su frecuencia con la esperada en base a los datos obtenidos hasta la fecha utilizando la distribución hipergeométrica. Los datos del número de genes que poseen una inserción homocigota están disponibles en la página del *Salk Institute* (http://signal.salk.edu/) en el contexto del proyecto NSF 2010 #040126.

## 6. Construcción de red regulatoria

6.1 Obtención de regiones regulatorias y mapeos con factores de transcripción

Desde la base de datos AGRIS (Davuluri y col. 2003; Palaniswamy y col. 2006), es posible obtener los genes anotados como factores de transcripción y acceder al análisis de las regiones río arriba de todos los genes del genoma de

*Arabidopsis thaliana*. Utilizando esta información, se construyó un archivo con los factores de transcripción (TF) y sus potenciales genes blancos en el genoma de Arabidopsis. Debido a que las secuencias reconocidas por estos factores de transcripción son normalmente pequeñas y de baja complejidad, la probabilidad de encontrar una de estas secuencias por azar en el genoma es alta. En este trabajo, analizamos como promotor a las regiones río arriba de cada gen con largo arbitrario de 1.5 Kb. Para definir una interacción posible entre un TF y un gen blanco, sólo consideramos los promotores que poseen un determinado elemento en Cis regulatorio reconocido por TF con una frecuencia alta en comparación con lo esperado por simple azar. Así, definimos como interacción regulatoria, aquellos mapeos entre TF y sus genes blanco (interacción proteína-DNA) que cumplen con la condición de que el gen blanco posee en su región promotora la caja consenso reconocida por el TF, con una frecuencia igual o superior a tres veces la desviación estándar de su frecuencia en el total de los genes del genoma. Este enfoque simple, ha resultado exitoso para la predicción de módulos regulatorios transcripcionales (Gutierrez y col. 2008).

6.2 Obtención de interacciones regulatorias

Para incrementar la probabilidad de que una interacción regulatoria predicha sea realmente efectiva y encontrar nuevos genes de interés como potenciales reguladores maestros de la transcripción de los genes del metabolismo de la pared celular, seleccionamos desde todas las interacciones regulatorias entre un TF y un gen blanco obtenidas en el punto anterior, aquellas que contienen a alguno de los 248 genes involucrados en el

metabolismo de la pared celular. Esta lista de genes contiene 205 genes a los cuales la etiqueta de Gene Ontology *GO:007047* les ha sido asignada en su descripción (05/08/2008) y 43 genes a los cuales en este trabajo les ha sido asignada esta etiqueta utilizando un algoritmo de anotación basado en el vecindario de un gen de simple conteo que considera las correlaciones entre los nodos. Luego, filtramos estas interacciones utilizando los datos de correlación de expresión de Spearman desarrollados en este trabajo, seleccionando sólo aquellas interacciones regulatorias que entre el TF y su gen blanco que posean una correlación de expresión $\geq 0.5$ entre sí. Con estas interacciones, en este trabajo se construyó una red regulatoria y se analizaron los potenciales TF involucrados en el metabolismo de la pared celular por medio de análisis del grado de los nodos de la red.

# RESULTADOS

## 1. Construcción de redes de correlación de expresión

Con el objetivo de construir redes de correlación con datos de expresión que nos permitan extraer información funcional en *Arabidopsis thaliana*, utilizando 1.701 microarreglos públicos filtrados por calidad ATH1 de Affymetrix™, calculamos las correlaciones lineales entre todos los pares de grupos de sondas presentes en estos chips (ver métodos) **(Figura 2-3).** Al analizar la frecuencia de los valores de las correlaciones obtenidas y filtrar aquellas que poseen valores ≥ 0.5, a nivel global observamos que existen más correlaciones positivas que negativas entre los grupos de sondas analizados. **(Figura 4).** Esto indica que en Arabidopsis, es más común observar genes expresándose juntos en las distintas condiciones experimentales analizadas. Los microarreglos analizados en este trabajo pueden ser agrupados en 180 experimentos distintos e incluyen datos de distintos órganos y tejidos, estados de desarrollo y tratamientos con diversos factores experimentales tales como estrés abiótico y biótico (Aceituno y col. 2008). En este trabajo, analizamos aquellas interacciones definidas por la coexpresión de los genes en todo este grupo de experimentos, sin hacer mayor distinción de los grupos de experimentos responsables de la mayor o menor similitud entre dos perfiles de expresión. Además, para la confección de nuestras redes de coexpresión, sólo consideramos las correlaciones positivas, evaluando distintos niveles de corte

en los valores de correlación ≥ 0.5. Al analizar las redes de coexpresión construidas, se observa que a mayor nivel de corte de correlación, menor es el número de arcos de una red de coexpresión, pues no todos los genes poseen correlaciones altas en sus niveles de expresión con otros genes (ver métodos) **(Figura 5).**

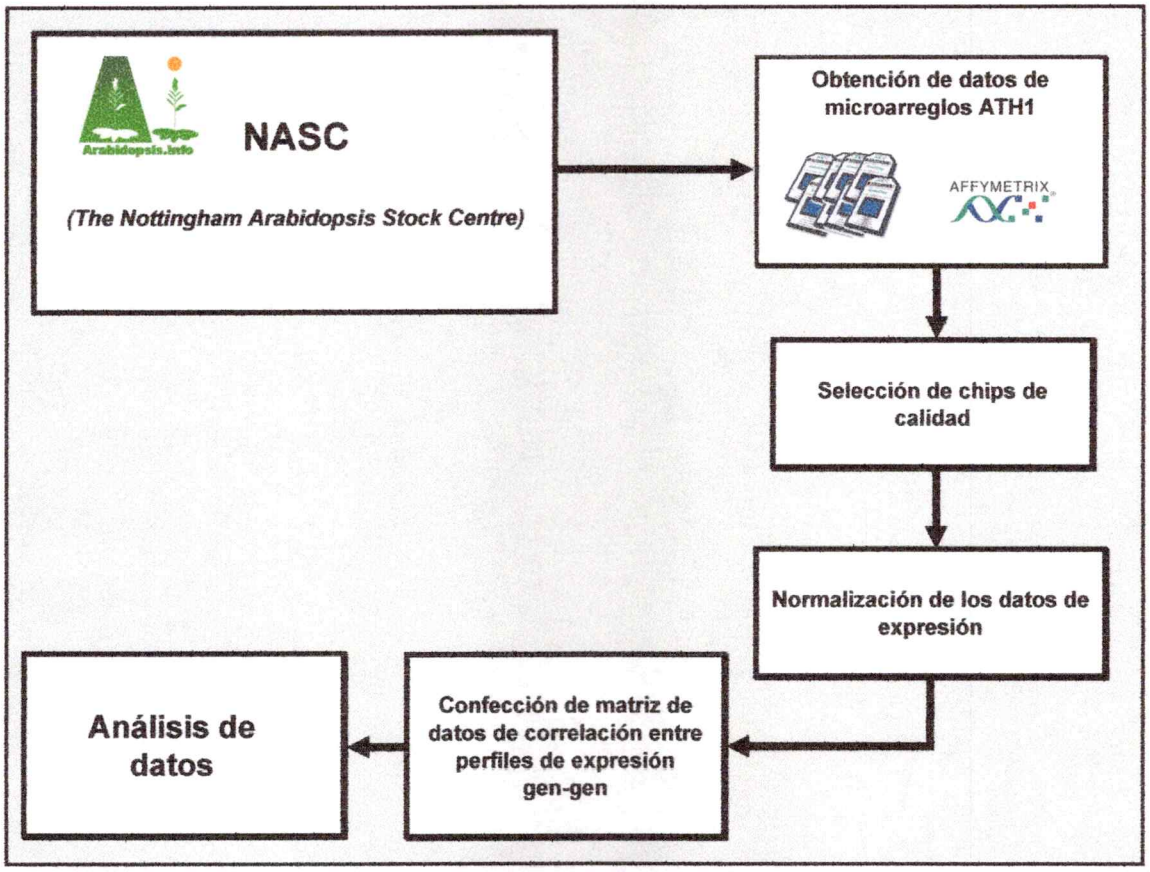



**Figura 2. Esquema que representa los pasos seguidos para la obtención de los datos utilizados en este trabajo.** Desde el "Nottingham Arabidopsis Stock Centre" se obtuvieron datos públicos de microarreglos correspondientes a 1.887 hibridaciones realizadas con los chips ATH1 Affymetrix™. Las hibridaciones fueron filtradas para eliminar aquellos chips que presenten problemas de calidad en sus datos. Las 1.701 hibridaciones remanentes, fueron normalizadas por el método RMA (Irizarry y col. 2003). Las correlaciones lineales entre todos los pares de perfiles de expresión fueron luego organizadas en una matriz, desde la cual es posible obtener los datos de correlación gen-gen entre todos los perfiles de cobertura del chip y realizar los análisis de datos necesarios para la construcción de una red que represente las correlaciones entre pares de perfiles de expresión.

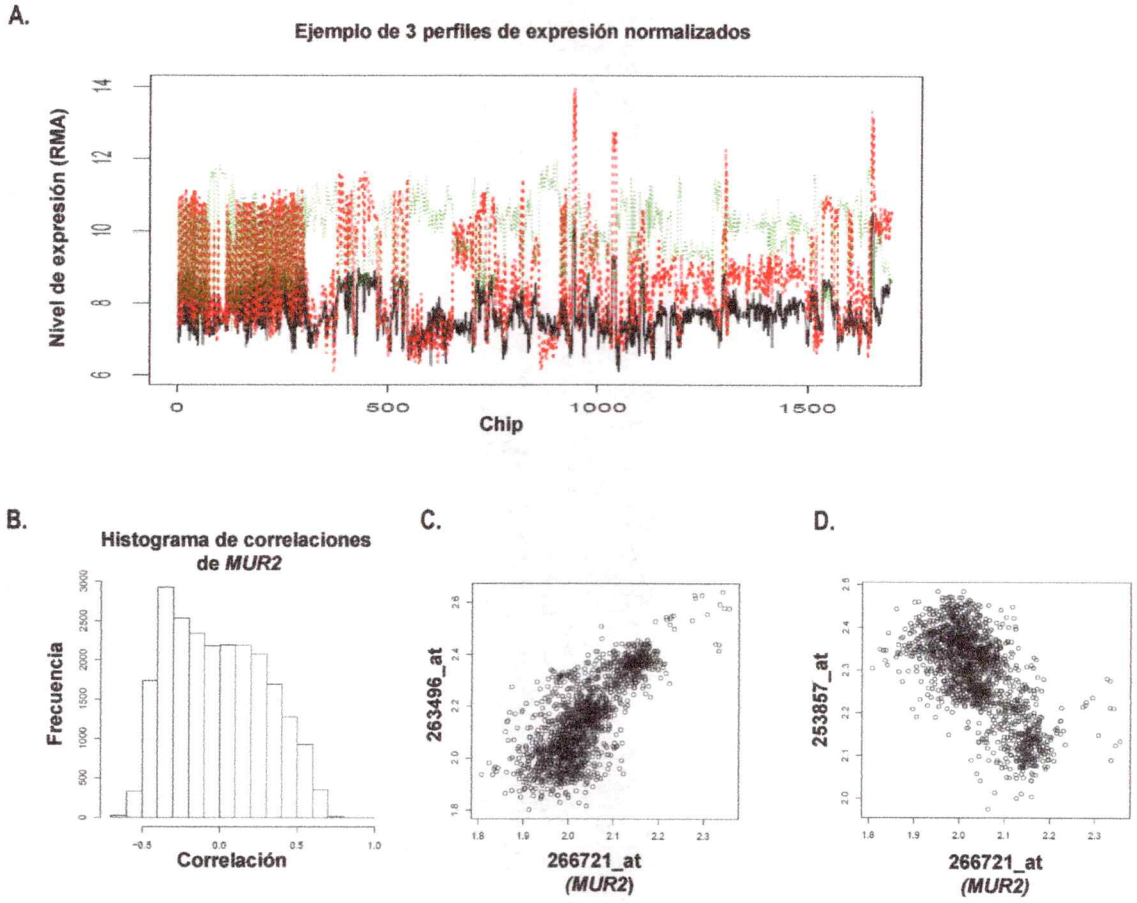



**Figura 3. Ejemplo de correlaciones positivas y negativas entre perfiles de expresión.** (A) Perfil de expresión normalizado por RMA (Irizarry y col. 2003) en los 1.701 microarreglos analizados en este trabajo del gen *MUR2* (perfil en color negro), un gen que codifica una glicosiltransferasa involucrada en el metabolismo del xiloglucano (locus AT2G03220). En rojo graficamos el perfil de expresión del gen con la correlación más alta con *MUR2*. Dicho perfil corresponde al perfil del locus AT2G42570, un gen de función desconocida anotado como *"unknown protein"*. En verde, graficamos el perfil de expresión que presenta la menor correlación con *MUR2*. Este perfil corresponde al del locus AT1G76260, un gen que si bien no posee un proceso biológico asignado, posee asignada la función molecular de unión a nucleótido *"nucleotide binding"*. (B) Histograma que muestra la frecuencia de los valores de correlación entre el perfil de expresión de *MUR2* y los otros 20.857 perfiles de expresión de los grupos de sondas analizados (ver métodos). Nótese que *MUR2* no posee correlaciones ≥ 0.8. (C) Gráfico que muestran el grado de asociación lineal entre el perfil de *MUR2* y AT2G42570, con una correlación de 0.737 (D) Gráfico que muestra el grado de asociación lineal entre el perfil de *MUR2* y AT1G76260, con una correlación de -0.646.

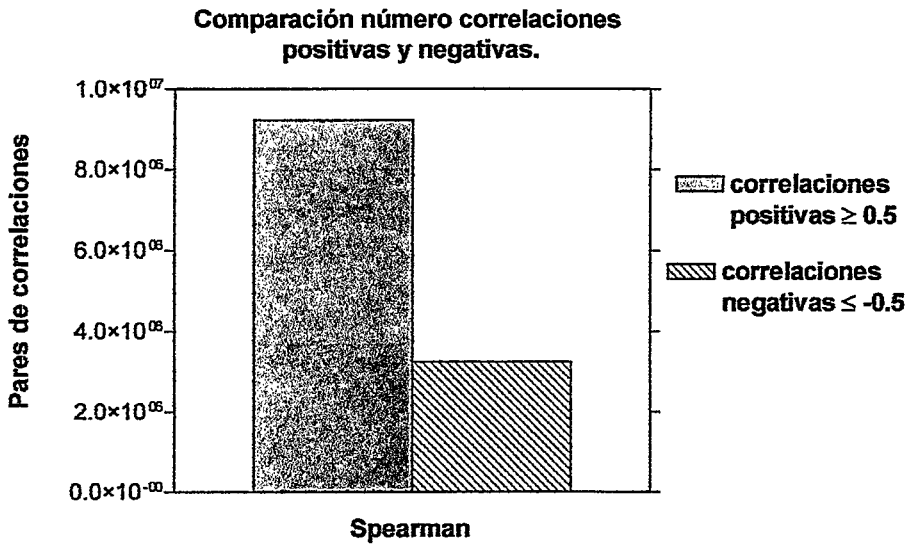
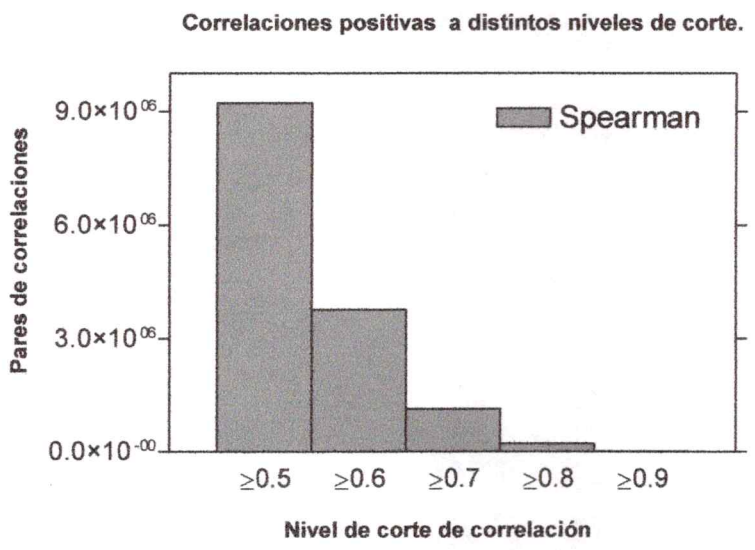



**Figura 4. Existen más correlaciones positivas que negativas.** Al analizar los 260.136.645 pares de correlaciones de Spearman entre todos grupos de sondas presentes en los chips ATH1, encontramos que las correlaciones ≥ 0.5 corresponden a un 4.4% del total de las correlaciones calculadas. Además, encontramos que hay alrededor de tres veces más correlaciones positivas que negativas, si se comparan las correlaciones significativas y con valores mayor o igual que 0.5 o menor o igual que -0.5.

Utilizando los mapeos entre grupos de sondas entregados por TAIR (Rhee y col. 2003; Swarbreck y col. 2008), basados en la última versión del genoma de Arabidopsis a la fecha (TAIR 8), identificamos en estos chips 21.180 grupos de sondas no promiscuas mapeables a algún locus, lo que nos permite alcanzar una cobertura de aproximadamente el 80% de los aproximados 27.000 genes de Arabidopisis descritos hasta la fecha. Con estos grupos de sondas, calculamos las correlaciones de expresión entre sus perfiles de expresión, obteniendo con éstos 224.285.610 pares de correlaciones, las cuales representan el nivel de similitud de los pares de perfiles de expresión analizados. Con estos datos, construimos redes de coexpresión como un grafo no dirigido, donde los nodos representan a los genes y los arcos entre éstos representan los valores de correlación entre pares de perfiles de expresión como distancia **(Figura 5).**

A.




B.

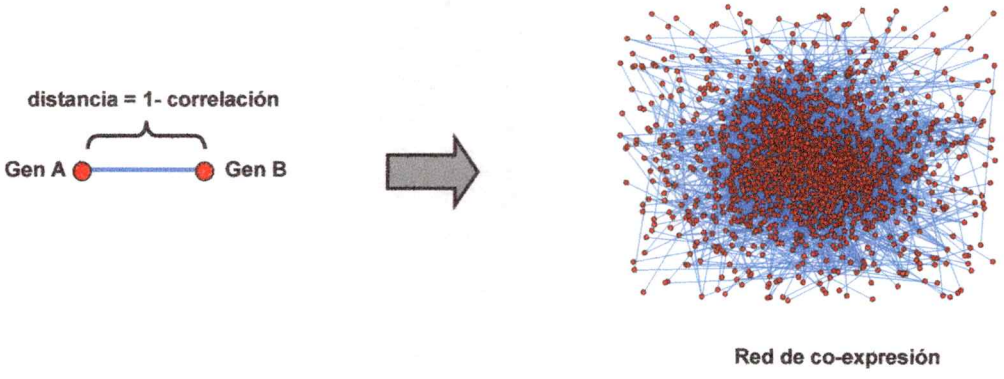



**Figura 5. Con las correlaciones de expresión es posible construir redes de coexpresión.** (A) Comparación del número de pares de correlaciones positivas a distintos niveles de corte de correlación obtenidos. Nótese que a mayor nivel de corte de correlación, menor es el número de arcos de una eventual red de coexpresión. (B) Las correlaciones positivas entre pares de perfiles de expresión fueron expresadas como distancia (distancia=1-valor correlación) y están representadas por los arcos que unen dos nodos (genes) en una red, pues las correlaciones positivas toman valores entre 0 y 1.

## 2. Análisis de topología en redes de coexpresión

Construimos redes de correlación tomando distintos valores umbral en los valores de correlación a partir de valores ≥ 0.5 **(Tabla 1)**. En estas redes, calculamos el grado *(k)* para cada nodo (gen). Este parámetro, mide cuántos vecinos tiene un gen en la red, es decir, con cuántos genes coexpresa. Además, para cada nodo calculamos sus coeficientes de clustering *(CC)*, parámetro que nos dice qué tan agrupados están los vecinos de un gen en la red y que toma valores entre 0 y 1. El *CC* se calcula simplemente tomando el número de arcos entre todos los vecinos de un gen y se divide por el número máximo de posibles arcos que podrían existir entre éstos (Watts y col. 1998; Albert R. 2002). El significado biológico del *CC* en las redes de coexpresión, radica en hecho de que altos valores de este parámetro, nos revelan que los genes se agrupan en módulos entre sí y que están muy conectados unos con otros. Tanto el grado como los coeficientes de clustering, nos permiten obtener información de la estructura de la red. En los últimos años, se ha observado que muchas redes de sistemas complejos poseen la llamada propiedad "small world", pues estas redes se caracterizan por la facilidad con que dos nodos de la red pueden ser unidos entre sí a través de los arcos de la red, gracias al alto grado de agrupamiento entre los nodos. Si dos nodos forman parte de un componente conexo de la red, siempre es posible encontrar un camino para unirlos y siempre hay un camino más corto entre todos los caminos posibles entre dos nodos de la red. Así, como parámetro topológico de las redes suele

utilizarse el diámetro, que se define como el camino más largo entre todos los caminos más cortos que conectan todos los pares de nodos de la red. Las redes "small world" o de pequeño mundo, poseen diámetros muy pequeños y coeficientes de clustering promedio cercanos a ½ (Watts y col. 1998; Amaral y col. 2000). Al analizar la topología de nuestras redes de coexpresión, observamos las redes de coexpresión analizadas en este trabajo poseen altos valores de coeficientes de clustering promedio *(<CC>)*, muy superiores a los obtenidos en redes de igual tamaño y grado promedio *(<k>)*, pero construidas con modelos de redes aleatorias, lo que nos indica que la forma en que se agrupan los nodos en la red no es al azar. Hasta la fecha, muchas redes analizadas en distintos sistemas complejos, poseen sus nodos altamente agrupados entre sí. Además, se caracterizan por poseer pequeños subgrupos de nodos que están muy altamente conectados en comparación con el resto de los nodos de la red. Estos nodos con los mayores valores de grado *(k)* de una red, usualmente reciben el nombre de "hubs". De todos los niveles de corte de correlación analizados, la red que más se aleja en su <CC> de una red aleatoria de tamaño equivalente es la red obtenida al seleccionar aquellos pares de correlaciones ≥ 0.7. El diámetro más pequeño en las redes analizadas fue obtenido al seleccionar aquellos pares de correlaciones ≥ 0.5. Esto implica que como máximo con sólo 8 arcos es posible unir cualquier par de nodos de la red, pues se trata de una red con un alto grado de conexión entre los nodos que la componen **(Tabla 1)**,

| Corte de correlación | Nodos (genes) | Arcos | <k> | <CC> | <CC>/<CC>$_{random}$ | Diámetro |
|---|---|---|---|---|---|---|
| 0.5 | 20.145 | 7.873.880 | 781.720 | 0.603 | 15.533 | 8.0 |
| 0.6 | 17.861 | 3.225.174 | 361.141 | 0.590 | 29.193 | 12.0 |
| 0.7 | 12.248 | 985.891 | 160.988 | 0.566 | 43.14 | 19.0 |
| 0.8 | 5.293 | 186.853 | 70.603 | 0.526 | 40.048 | 26.0 |
| 0.9 | 1.056 | 10.334 | 19.571 | 0.464 | 25.286 | 13.0 |

**Tabla 1. Redes de coexpresión analizadas en este trabajo.** Las correlaciones ≥ 0.5 de Spearman fueron analizadas y se construyó con ellas redes de coexpresión (ver métodos). Los locus mapeados por grupos de sondas no promiscuas en los chips ATH1, fueron utilizados para construir redes de correlación de expresión (ver métodos). En estas redes a los distintos niveles de corte de correlación seleccionados, se analizó el grado promedio (*<k>*) y los coeficientes de clustering promedio (*<CC>*). Además, se comparó el <CC> de cada red con el obtenido desde una red de igual número de nodos y el mismo *<k>*, pero obtenidos desde una red aleatoria construida con el modelo de *Erdos-Renyi* (<CC>$_{random}$).

Los análisis de topología realizados sobre las redes de coexpresión analizadas poseen una distribución de frecuencia de los grados de sus nodos truncada en la zona de los nodos más conectados **(Figura 6-A).** Para obtener el valor del grado de la truncación $K_x$, que justamente define la región en que esta distribución se ajusta a un modelo de decaimiento escala libre, se realizó una regresión con la ecuación de Yule-Simon; $\log y = -\rho \log k - (k/kx)$, donde y corresponde a la frecuencia y $\rho$ al exponente del decaimiento exponencial. Similares distribuciones se observan a los otros niveles de corte de correlación de Spearman $\geq 0.5$ analizados (datos no mostrados). La conclusión de estos datos, es que las redes de coexpresión analizadas, se comportan como una red escala libre en cierto rango de grado, pero presentan una cola en la región de los nodos más conectados, donde la frecuencia de los nodos con cierto número de vecinos en la red, decae aún más rápido con el grado. Esto quiere decir, que si tomamos un nodo al azar, la probabilidad de que este sea un nodo altamente conectado es muy baja y que en la zona de la cola, esta tendencia se hace aún más abrupta. En la zona de los nodos más conectados en la red, hay pocos genes y corresponden a genes que co-expresan con muchos otros genes. Debido a esto, son genes muy interesantes de estudiar, pues se trata de genes que codifican para proteínas que probablemente participan en muchos procesos biológicos o que forman parte de procesos biológicos esenciales.

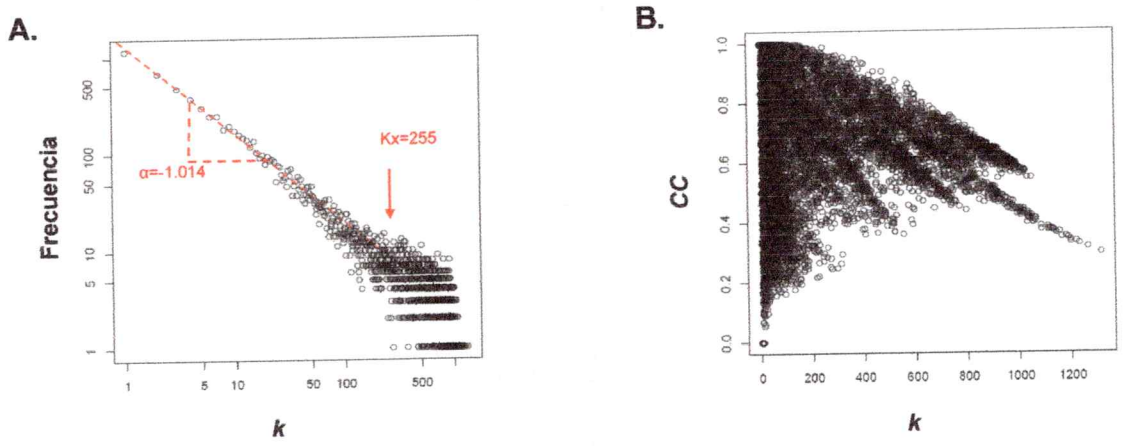



**Figura 6. Análisis de topología en redes de coexpresión de Arabidopsis.** (A) En la figura, se representa como ejemplo la distribución de frecuencia de los grados de los nodos correspondientes a la red de coexpresión construida con correlaciones de Spearman ≥ 0.7 de la tabla 1. El grado *(k)* de cada nodo (gen) se grafica versus su frecuencia en escala log/log. La línea roja es una guía de lo que se espera para un decaimiento exponencial. La distribución obedece a una ley de potencia que se trunca a $k_x$ = 255, pues en esta región del gráfico la frecuencia decae mucho más rápido que lo que lo haría una ley de potencia. α corresponde a la pendiente del decaimiento exponencial en la zona lineal del gráfico, utilizando los datos con grados ≤ $K_x$. (B) La relación entre los coeficientes de clustering *(CC)* y el grado *(k)* de cada gen fue analizada con los datos obtenidos desde red de coexpresión construida con las correlaciones de Spearman ≥ 0.7 de la tabla 1.

Para cada uno de los genes presentes en las redes de coexpresión, examinamos con cuántos genes coexpresa, es decir su grado (*k*), y su coeficiente de clustering. El coeficiente clustering *(CC)*, nos entrega una métrica de qué tan agrupados están entre sí los vecinos de un nodo en la red, tomando valores entre $0 \leq CC \leq 1$. Así, una red con un grado de agrupación por coexpresión que tenga un *CC* promedio *(<CC>)* pequeño, será una red con bajo grado de interacción entre sus genes. En nuestras redes, hay muy pocos genes con muy alto grado en comparación con el resto de los genes. Además, los nodos están muy agrupados entre sí, pues poseen *<CC>* muy superiores a los esperados para una red aleatoria. Sin embargo, ¿Existirá alguna relación entre ambas variables? Al examinar la relación entre los *(CC)* y el grado de cada gen en las redes de coexpresión, al igual que *Bergmann, Ihmels et al.* (Bergmann y col. 2004), observamos a los genes agrupados en "bandas", donde se observa que a medida que aumenta el grado, los coeficientes de clustering también decrecen **(Figura 6-B).** Al encontrar que hay una relación entre *CC y k* evidenciamos que nuestras redes poseen módulos y estructura jerárquica (Newman 2006; Clauset y col. 2008). Esto implica que en la red los nodos se agrupan y que además, estos grupos se subdividen en otros subgrupos. Grupos que en muchas redes analizadas, en distintas disciplinas de la ciencia, han sido coincidentes con unidades funcionales previamente descritas, tales como vías metabólicas en redes metabólicas, nichos ecológicos en redes ecológicas o comunidades en redes sociales (Ravasz y col. 2002; Clauset y col. 2008). ¿Están estos grupos de genes conectados entre sí en las redes? Al analizar

todos los caminos más cortos posibles al recorrer los arcos entre todos los pares de nodos de una red, encontramos que el camino más largo de éstos (diámetro de la red) es bastante pequeño en todas las redes analizadas (tabla 1). Sin embargo, no sabemos qué tantos módulos o grupos no conectados hay en las redes construidas a cada nivel de corte de correlación analizado. Esto es muy importante de analizar, pues nosotros lo que estamos haciendo es propagar las anotaciones de los genes anotados a sus vecinos sin un proceso biológico asignado, lo que claramente está limitado por los posibles agrupamientos funcionales existentes. Debido a esto, un punto importante a considerar en las redes y justamente relacionado con la estructura modular que observamos, es analizar el número de componentes presentes en cada una de estas redes. Dos genes están definidos como parte del mismo componente, si existe un camino a lo largo de distintos arcos de la red que los conecte. Las redes no necesariamente poseen a todos sus nodos conexos entre sí, lo que limita la propagación de las anotaciones. Al analizar el número de componentes presentes en cada uno de los niveles de corte analizados en la tabla 1, observamos que a menor nivel de corte de correlación la mayoría de los genes están agrupados en un solo gran componente **(Figura 7).**

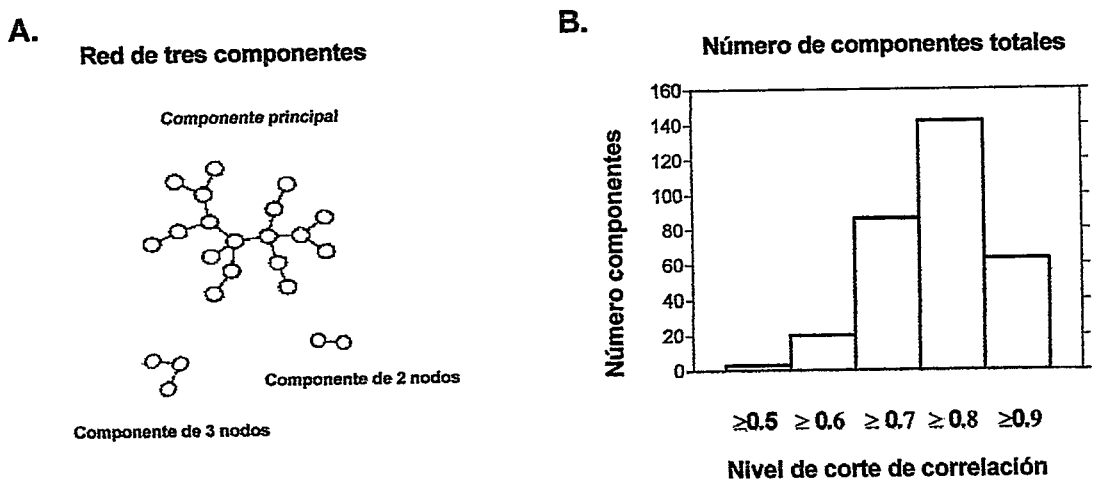



**Figura 7. Análisis de componentes** (A) Dos genes están definidos como parte del mismo componente, si existe un camino a lo largo de distintos arcos de la red que los conecte. Las redes pueden no necesariamente tener a todos sus nodos conexos entre sí. En la figura, se representa una red compuesta de 3 componentes. Un gran componente principal de 18 nodos y 2 componentes más pequeños de 3 y 2 nodos respectivamente. (B) A los distintos niveles de corte de correlación, se obtuvo el número de componentes presentes. Acá el número de componentes totales en cada red de Spearman es graficado a los distintos niveles de corte de correlación analizados en la tabla 1.

¿Qué información biológica podemos obtener de los análisis de topología realizados a las redes de coexpresión? Al analizar funcionalmente los genes más conectados es posible saber si hay algo que los caracteriza. Al analizar sus anotaciones es posible saber si por ejemplo, estos genes están enriquecidos en cierto tipo particular de anotaciones correspondientes a ciertos procesos biológicos en particular o funciones moleculares. Así, con el objetivo de analizar los nodos más conectados de las distintas redes construidas, tomamos el 1% de los nodos más conectados de cada una de las redes, estos nodos los definimos como "hubs" (ver métodos) y analizamos estas listas de genes funcionalmente, utilizando las anotaciones de Gene Ontology (GO). La base de datos de Gene Ontology (http://www.geneontology.org/) (Ashburner y col. 2000), entrega las etiquetas de anotaciones de los genes y las relaciones que las etiquetas poseen entre sí. Estas etiquetas se clasifican en categorías, que dependiendo del aspecto que desean describir, se clasifican en "funciones moleculares", "componentes celulares" y "procesos biológicos". Cuando analizamos en los genes seleccionados como "hubs", las anotaciones que poseen asignadas y sus respectivas frecuencias con respecto a todo el genoma, encontramos que en estos genes, ciertas etiquetas GO están representadas por sobre lo esperado por simple azar **(Figura 8-A).** Además, otro resultado interesante, es el hecho de que estas listas de genes "hubs" correspondientes al 1% más conectado de la red, están enriquecidos en genes no anotados y presentan una menor frecuencia de mutantes de T-DNA homocigotas de lo esperado en base al frecuencia de plantas con inserción

homocigota que han sido identificadas hasta la fecha por proyectos internacionales de obtención de líneas insercionales de T-DNA en Arabidopsis a nivel masivo (ver métodos; proyecto NSF 2010 #040126). La utilización de líneas insercionales de T-DNA como herramienta en genómica funcional en *Arabidospsis thaliana* ha tenido un explosivo crecimiento en los últimos años y es la principal metodología responsable de las anotaciones que relacionan a un gen de esta planta con algún fenotipo obtenido (Meinke y col. 2003; Li y col. 2006) Hoy, gracias a la disponibilidad de esta información, sabemos que de un total de 23.245 genes que poseen una inserción de T-DNA que interrumpa su secuencia, sólo 16.656 la poseen en ambas copias del gen **(Figura 8-B).** Al contrastar esta frecuencia de genes con inserción de T-DNA en ambas copias contra las frecuencias esperadas en las listas de genes más conectados analizadas, observamos que las listas de genes "hubs" presentan un número mucho menor de genes con una línea homocigota de T-DNA que lo esperado por azar, lo que sugiere que estos genes son esenciales y que la eliminación de ambas copias del gen ocasiona la letalidad de la planta.

A.

| Nivel de corte de correlación (cut-off) | N° hubs (1% más conectado) | GO enriquecidos en hubs (p value ≤ 0.05) Procesos biológicos | Función molecular | Anotación gen más conectado |
|---|---|---|---|---|
| ≥ 0.5 | 201 | Ninguno | No anotado<br>Betagalactosidasas | AT4G05230<br>Biological Process: ubiquitin cycle<br>Molecular Function: unknown<br>Cellular Component: unknown |
| ≥ 0.6 | 179 | Ninguno | No anotado | AT4G05230<br>Biological Process: ubiquitin cycle<br>Molecular Function: unknown<br>Cellular Component: unknown |
| ≥ 0.7 | 122 | Ninguno | No anotado | AT4G05230<br>Biological Process: ubiquitin cycle<br>Molecular Function: unknown<br>Cellular Component: unknown |
| ≥ 0.8 | 53 | Respuesta a frío<br>Respuesta a T°<br>Fotosintesis | Unión a drogas | AT1G67700<br>Biological Process: unknown<br>Molecular Function: unknown<br>Cellular Component: chloroplast |
| ≥ 0.9 | 11 | Ninguno | Ninguno | AT1G15980<br>Biological Process: photosynthetic electron transport in photosystem I<br>Molecular Function: none<br>Cellular Component: chloroplast NAD(P)H dehydrogenase complex |

B.

| Nivel de corte de correlación (cut-off) | N° hubs (1% más conectado) | Número de hubs con alguna línea homocigota de T-DNA (observado) | (esperado) | P-value Universo (23306) |
|---|---|---|---|---|
| ≥ 0.5 | 201 | 95 | 144 | $4.1*10^{-13}$ |
| ≥ 0.6 | 179 | 88 | 128 | $2.4*10^{-10}$ |
| ≥ 0.7 | 122 | 61 | 87 | $4.37*10^{-7}$ |
| ≥ 0.8 | 53 | 27 | 38 | 0.0012 |
| ≥ 0.9 | 11 | 7 | 8 | 0.387 |

**Figura 8. Los genes más conectados de redes de coexpresión están enriquecidos en genes de función desconocida y poseen menos mutantes homocigotas de T-DNA de lo esperado.** En base a las frecuencias observadas de cierta clase, es posible calcular la probabilidad de que un muestreo aleatorio sin reposición entregue cierto valor de frecuencia, contrastando con la frecuencia de la clase en un universo utilizando la distribución hipergeométrica. (A) En la tabla, se muestra el número de nodos seleccionados como "hubs" a distintos niveles de corte de correlación y sus anotaciones GO enriquecidas de procesos biológicos y funciones moleculares, seleccionando como enriquecidas aquellas anotaciones con un P-value ≤ 0.05. Además, se representa la anotación del nodo más conectado obtenido a cada uno de los niveles de corte de correlación analizados. (B) Tabla que representa para los "hubs" obtenidos, a distintos niveles de corte de correlación, el número observado y esperado de éstos genes con inserciones de T-DNA homocigotas. Además, se representa el P-value respectivo obtenido al contrastar ambas frecuencias y utilizar la distribución hipergeométrica. En rojo, se representan aquellos P-values ≤ 0.01 obtenidos.

## 3. Predicción de función génica utilizando las redes de coexpresión

Las anotaciones de Gene Ontology (GO) se organizan en un grafo acíclico con dirección (DAG), donde cada etiqueta descriptiva o término GO puede tener términos "padres" y términos "hijos". Esto provoca que no todos los genes necesariamente estén anotados al mismo nivel con respecto al término GO "raíz". Debido a esto, para realizar las evaluación de las predicciones de función debemos previamente homogenizar las anotaciones de todo el genoma hasta un mismo nivel de descripción, para poder así calcular cuántas predicciones son correctas utilizando un grupo de genes de anotación conocida como grupo de entrenamiento (ver métodos). Básicamente, para homogenizar las anotaciones hay que ascender por el DAG de cada anotación hasta la raíz y luego descender hasta el mismo nivel de descripción las anotaciones de todos los genes **(Figura 9).**

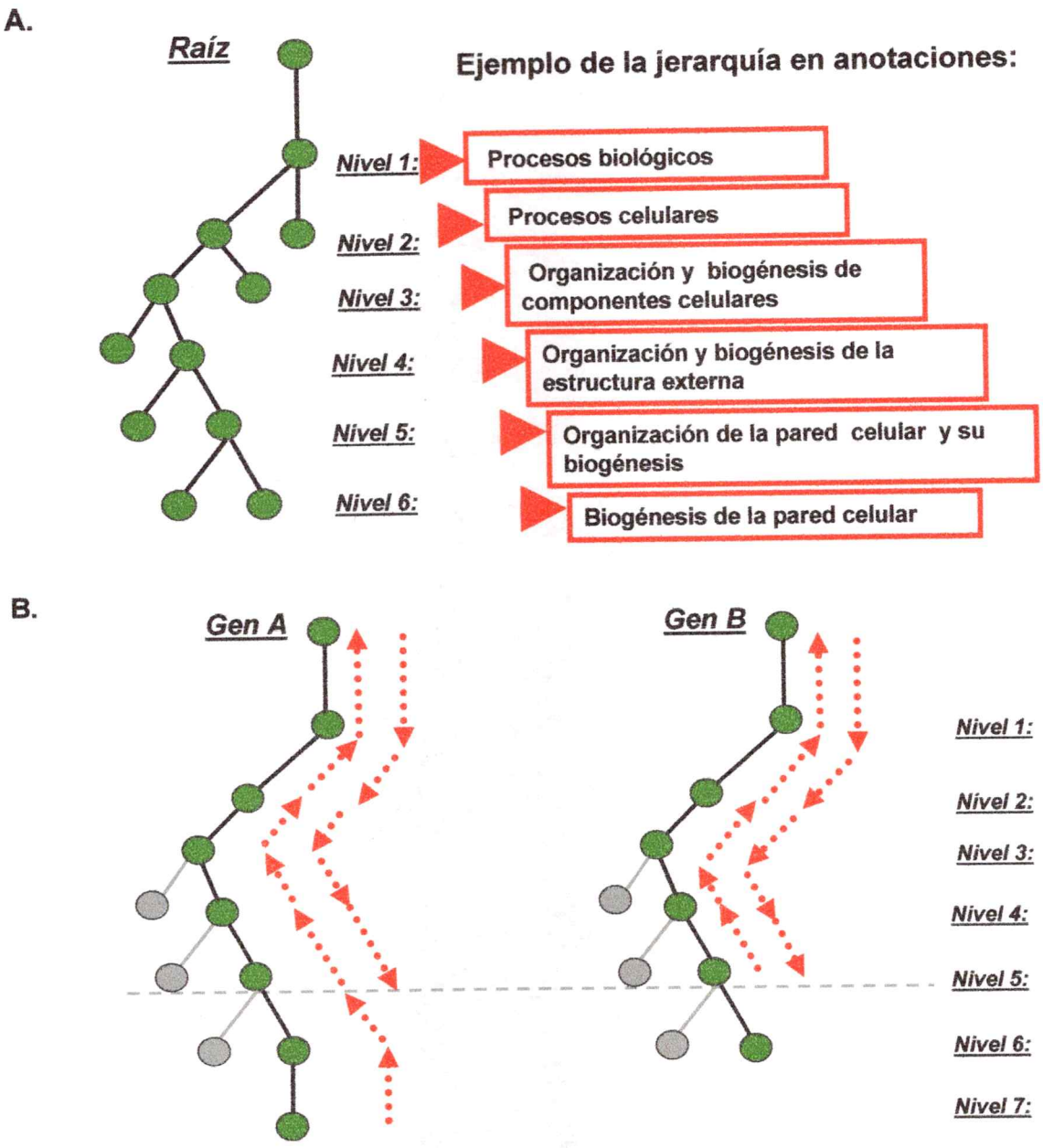



**Figura 9. Las anotaciones de Gene Ontology (GO) deben ser llevadas hasta el mismo nivel de descripción.** (A) Las anotaciones de GO se organizan en un grafo acíclico como el que se presenta acá. Para simplificar la figura se eliminó la dirección de los arcos. Las descripciones o etiquetas se organizan desde la más general a la más específica (B) Debido a que en base a las anotaciones de los vecinos de un gen en la red de coexpresión, analizamos distintos algoritmos para asignar función, es necesario llevar todas las anotaciones hasta un mismo nivel de descripción, homogenización que se esquematiza acá llevando 2 genes hasta nivel 5. Nótese que el gen A originalmente estaba anotado hasta nivel 7 y el gen B hasta nivel 6.

## 3.1 Validación de la metodología

En este trabajo, una vez construidas las redes de coexpresión, analizamos distintos métodos que en base al vecindario de un gen, fueran capaces de asignarle un papel en algún proceso biológico. Dicho de otra forma, lo que hicimos fue propagar las anotaciones de los genes anotados a sus vecinos. Pese a ser Arabidopsis el primer organismo vegetal en ser secuenciado y ser un excelente organismo modelo en genómica funcional, sólo alrededor del 50% de los genes de esta planta poseen un proceso biológico asignado. Debido a esto, proponer algún papel en algún proceso biológico al 50% de los genes remanentes es un gran desafío en genómica funcional.

En este trabajo comparamos el rendimiento de tres algoritmos basados en el análisis del vecindario inmediato de un gen en la red para asignar función. Básicamente, tomando en consideración las anotaciones de los vecinos de un gen, estos métodos proponen alguna anotación para el gen en cuestión a anotar. Sin embargo, el problema es lograr comparar los distintos algoritmos en su poder predictivo. Para solucionar esto, lo que hicimos fue abordar el problema desde una perspectiva utilizada en los algoritmos de clasificación estadística y recuperación de información. Desde el punto de vista de las anotaciones, existen dos tipos de genes: a) aquellos que poseen anotación y b) aquellos que no poseen anotación. Debemos recordar que las anotaciones GO están subcategorizadas en BP, CC, FM, las que corresponden a las anotaciones de los genes en los aspectos de *"Procesos Biológicos"*, *"Componentes Celulares"* y *"Función Molecular"* respectivamente. En este

trabajo, utilizando las anotaciones de BP de GO homogenizadas hasta distintos niveles de descripción, calculamos la capacidad de cada uno de los distintos algoritmos analizados de acertar en las predicciones que entrega. Debemos mencionar que es necesario evaluar distintas variables, tales como el nivel de corte de correlación utilizado para realizar la predicción, la precisión de cada método analizado y el nivel de profundidad de homogenización de las anotaciones de GO. Además, debemos recordar que un gen puede tener 1 o más etiquetas de GO BP asignadas o simplemente no tener ninguna. Debido a ello, para poder comparar los distintos métodos de asignación de función analizados y elegir el mejor, evaluamos cuántos genes tenían una anotación correcta con respecto a su anotación original dentro de un grupo de entrenamiento y por cierto cuántos genes pueden ser anotados bajo esos parámetros como asociados al metabolismo de la pared celular, pues ciertos parámetros pueden proponer muchos nuevos genes anotados en este proceso metabólico. Sin embargo, debemos saber la probabilidad de que dichas predicciones sean correctas. Mientras más genes anotemos, menor será la oportunidad de que todas estas anotaciones propuestas sean verdaderas.

3.2 Métodos de asignación de función analizados

En este trabajo, analizamos la capacidad de distintos algoritmos que en base al vecindario de un gen en la red de coexpresión son capaces de asignarle algún proceso biológico a un gen. Utilizando los genes con alguna etiqueta GO de "proceso biológico" (BP) asignada como grupo de entrenamiento, comparamos los distintos métodos. La idea es encontrar la mejor predicción,

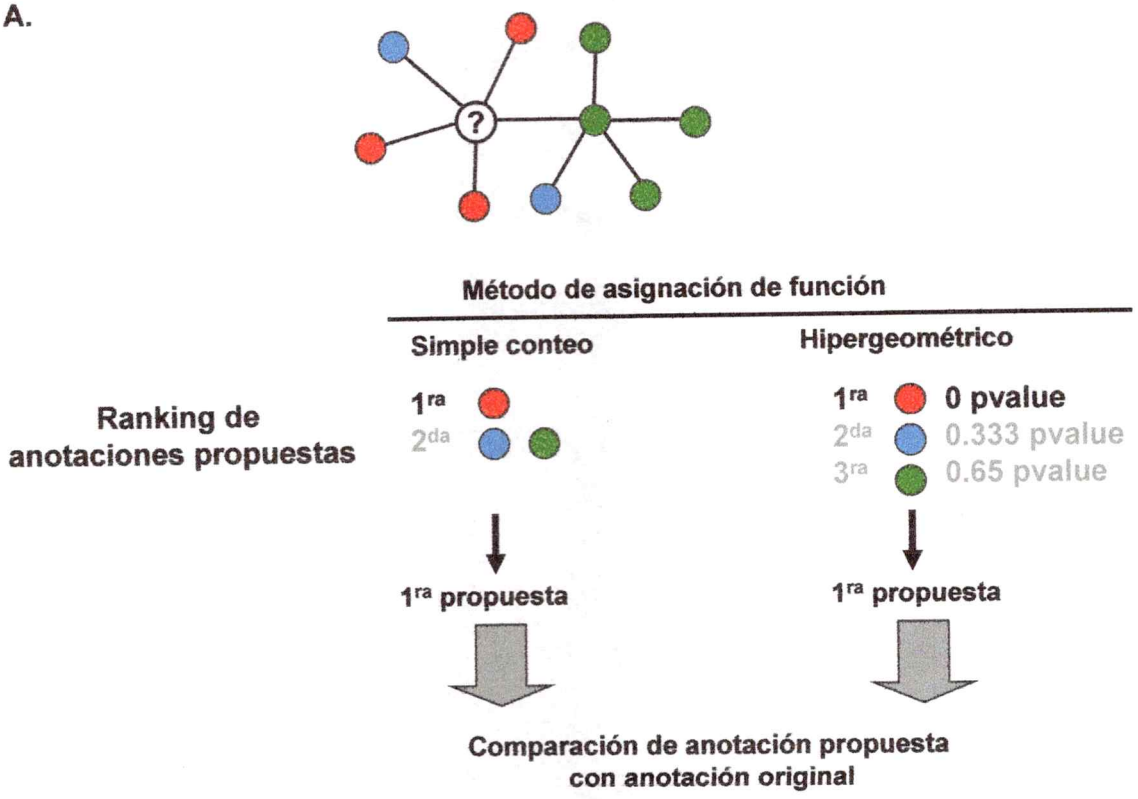
optimizando la recuperación de las anotaciones de los genes del grupo de entrenamiento. Cuando la predicción de función de un gen corresponde a la etiqueta original de GO BP asignada a ese gen, esa prueba es considerada un verdadero positivo (TP). Así, el parámetro a optimizar en los distintos métodos analizados para asignar función a un gen, es el número de TP dividido por el número de genes evaluados (Verdadero Positivos + Falsos Positivos), parámetro que usualmente recibe el nombre de Precisión. En la **Figura 10-A**, podemos ver un ejemplo ilustrativo de una red que posee 10 nodos pintados de 3 colores distintos: rojo, azul y verde. Estos colores representan alguna anotación de GO de algún proceso biológico asignado a ese nodo o gen. Para asignarle alguna función (color) al gen de color no asignado (gen incógnito), una aproximación posible es simplemente contar el número de colores presentes en el vecindario de ese nodo y asignarle el color más abundante. Esta aproximación es una aproximación de simple conteo. En nuestro ejemplo, una aproximación como ésta asigna el color rojo (que representa una anotación) al gen no anotado. Sin embargo, el uso de una aproximación de simple conteo que considera las frecuencias de cada anotación para elegir la más abundante como anotación a asignar tiene un problema, pues quizá hay algunas anotaciones muy abundantes que se propagarán mucho más que otras en la redes. Debido a esto, además del método de simple conteo, decidimos evaluar el rendimiento de un método que usa la distribución hipergeométrica, el cual corrige las frecuencias de las anotaciones en el vecindario de un gen con la abundancia de éstas en todo el resto de los genes y selecciona dentro de

aquellas anotaciones enriquecidas en el vecindario de un gen, aquella que sea más inusual encontrar por simple azar **(Figura 10-A).** En este trabajo, analizamos los vecindarios de cada gen en las redes de coexpresión y con éstos, asignamos anotaciones a los genes, comparando el rendimiento de tres distintos algoritmos de asignación de función: i) un algoritmo de simple conteo, ii) un algoritmo de simple conteo que además considera la distancia con los distintos vecinos de un gen a anotar (conteo-distancia) y iii) un algoritmo hipergeométrico, el cual propone como la anotación a asignar, aquella más sobre representada en comparación con lo esperado por azar. Utilizando los genes con algún proceso biológico asignado como grupo de entrenamiento, evaluamos el rendimiento de cada método ante distintos niveles del GO DAG y distintos niveles de corte de correlación, comparando los valores de precisión entre los distintos métodos **(Figura 11).** A niveles de corte de correlación muy cercanos a 1, las predicciones aumentan su poder predictivo, Sin embargo, menor será el número de genes que podrán propagar sus anotaciones bajo estos niveles de corte que a niveles menos estrictos, pues no necesariamente el perfil de expresión de un gen posee correlaciones tan altas con algún otro perfil. A nivel 5 del GO-DAG, tomando las correlaciones ≥ 0.9, la precisión obtenida es de 74.11% para el método de simple conteo que considera la distancia entre los vecinos para seleccionar la anotación a asignar, 66.35% para método de análisis hipergeométrico y 63.68% para método de simple conteo que no considera la distancia. A altos niveles de corte de correlación y al aumentar la exigencia en la propagación de las anotaciones, es posible

aumentar los valores de precisión. Sin embargo, nótese que este efecto no se observa a bajos niveles de correlación, pues el grado entre los nodos en estas redes es muy alto y el número de componentes no conexos de bajo número de nodos en estas redes es muy bajo **(Figura 11).**

Según los análisis de los valores de precisión, a todos los niveles de corte de correlación analizados el método con la precisión más alta es el de simple conteo que considera la distancia. El siguiente paso es utilizar este algoritmo entonces para encontrar nuevos genes involucrados en el metabolismo de la pared celular.

A.




B.

| | |
|---|---|
| Verdadero positivo **(TP)** | Falso positivo **(FP)** |
| Falso negativo **(FN)** | Verdadero negativo **(TN)** |

**Totales:** **P** **N**

*Precisión* = *TP* / (*TP* + *FP*)

**Figura 10. Esquema que representa cómo trabajan los métodos de asignación de función y su evaluación.** (A) Esquema de una red de 10 nodos, los que representan 10 genes, de los cuales 1 no posee anotación (nodo blanco con signo interrogación) y 9 la poseen, representadas por colores. De los 5 vecinos del gen de anotación incógnita a anotar, 3 son rojos, 1 verde y 1 azul. En la figura, se muestra un ejemplo de cómo un método de simple conteo y un método basado en enriquecimiento hipergeométrico, proponen anotaciones y las ranquean. El método de simple conteo, analiza cada uno de los vecinos de un gen a anotar en la red y la frecuencia de las anotaciones

presentes. Luego, después de ranquear las anotaciones desde la más abundante hasta la menos abundante presentes en el vecindario, propone como anotación a asignar, la mejor ranqueada o más abundante. El método hipergeométrico en cambio, además de analizar la frecuencia de cada anotación en el vecindario del gen a anotar, analiza qué tan frecuentes son cada de una de ellas en todas las anotaciones del genoma y compara qué tan probable es tener ese número de anotaciones por simple azar comparando las frecuencias observada con las esperadas. Así por ejemplo, el gen a anotar en el esquema, pese a poseer igual número de vecinos verdes que azules, al momento de ser ranqueadas las anotaciones propuestas, la anotación representada por el color azul esta más sobre representada que la verde en el vecindario del gen a anotar, pues como hay más nodos verdes en la red, la probabilidad que de los 5 vecinos del gen a anotar 1 sea verde por simple azar es más alta que sea azul, pues hay menos nodos azules que verdes en la red completa. Debido a ello, luego de contrastar las frecuencias observadas con las esperadas de los colores y calcular con estos datos un p value de enriquecimiento, la anotación propuesta es el color rojo, pues es el color con un p value más pequeño y además es ≤ 0.05 (ver métodos). (B) Esquema que representa como fueron evaluadas las anotaciones propuestas al compararlas con las anotaciones originales. Cuando la(s) anotaciones propuestas por algún método corresponde con alguna(s) de la(s) anotaciones originales del gen, esa prueba corresponde a un verdadero positivo (TP).

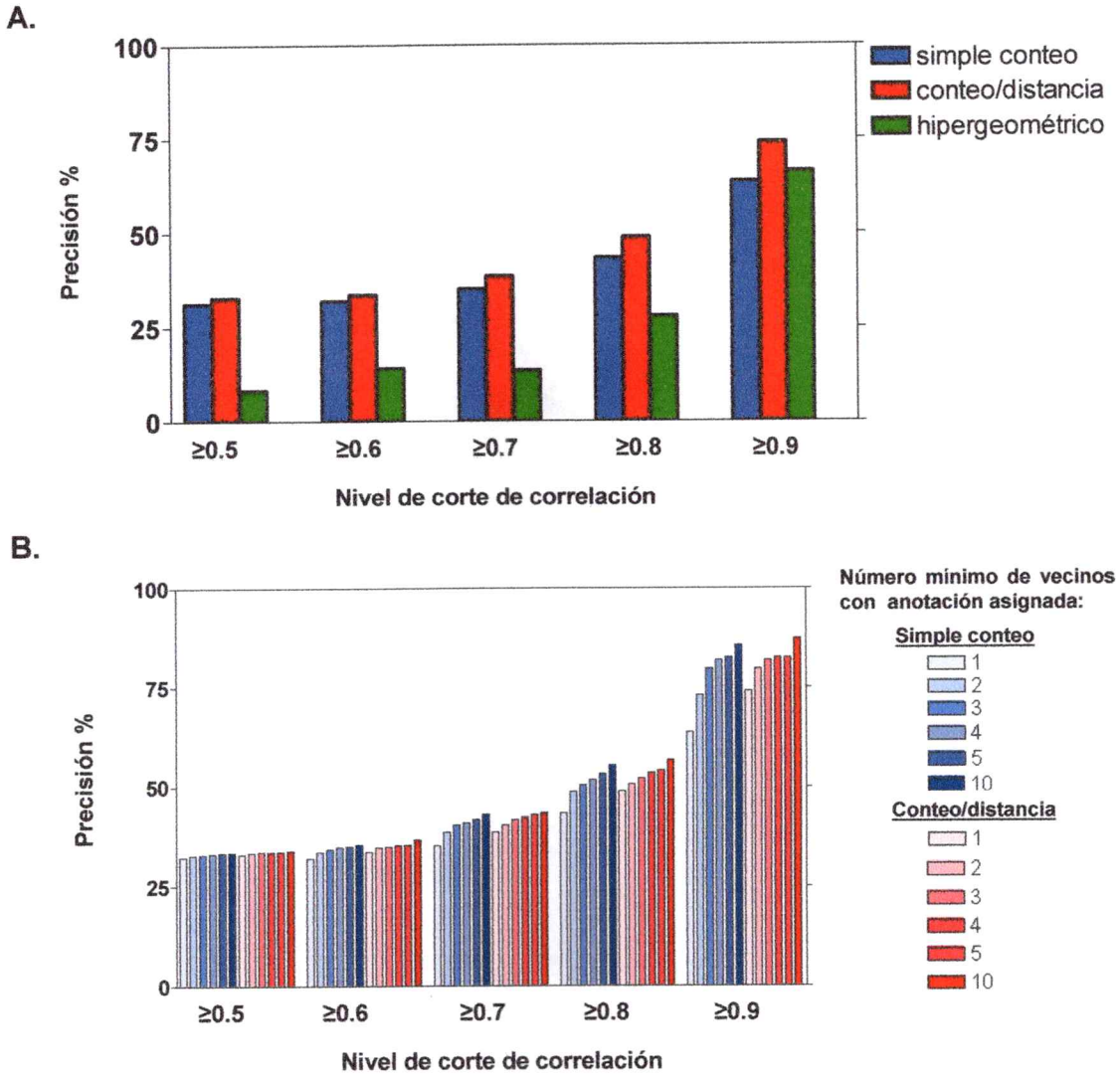



**Figura 11. Comparación del poder predictivo entre distintos métodos de asignación de función.** (A) Utilizando todos los genes con algún proceso biológico asignado como grupo de entrenamiento, calculamos la precisión obtenida con tres distintos métodos: simple conteo, simple conteo que considera las distancias (conteo/distancia) y método hipergeométrico. (B) Al analizar el vecindario de un gen, es posible incrementar la precisión, exigiendo que la anotación asignada a un gen esté presente en al menos un número determinado de vecinos (número mínimo de vecinos con anotación asignada).

## 4. Búsqueda de nuevos genes involucrados en el metabolismo de la pared celular

Con el objetivo de encontrar nuevos genes involucrados en el metabolismo de hemicelulosas y pectinas, en este trabajo hemos confeccionado redes de coexpresión a distintos niveles de corte de correlación. Al analizar estas redes y los genes presentes en éstas, encontramos que las anotaciones de GO correspondientes a los procesos biológicos (BP) están sesgadas. Hay ciertos procesos biológicos asignados a muchos genes y otros asignados a muy pocos (ver figura suplementaria 1). Así, al propagar las anotaciones en las redes de coexpresión por análisis de vecindarios, es necesario tomar en cuenta estos sesgos y evaluar su efecto sobre las posibles predicciones, utilizando por ejemplo un método que justamente ajuste las frecuencias observadas en el vecindario de un gen con respecto a la frecuencia de esa anotación en todo el genoma y comparando su rendimiento predictivo con algún otro método distinto. Con el objeto de encontrar nuevos genes involucrados en el metabolismo de la pared celular, utilizando la información funcional que nos puede entregar la coexpresión, en este trabajo decidimos propagar en las redes de coexpresión la etiqueta de anotación de BP *GO:0007047 "cell wall organization and biogenesis"* , una etiqueta que se encuentra a nivel 5 en el grafo de las anotaciones que se encuentra asignada a 205 genes en el genoma. Sin embargo, debemos recordar que no todos los genes poseen correlaciones necesariamente altas con algún otro gen.

Recordemos por ejemplo, que el perfil de expresión de mur2 no posee correlaciones ≥0.8 con alguno de los otros 20.858 perfiles de expresión de los grupos de sondas analizados (ver Figura 2-B). Así, dependiendo del nivel de corte de correlación que sea elegido, existirá más o menos potenciales genes nuevos a los cuales la etiqueta *GO:0007047 "cell wall organization and biogenesis"* puede ser asignada. A menor nivel de corte de correlación, mayor número de genes podrán ser asociados al metabolismo de la pared celular. Sin embargo, la probabilidad de que las predicciones sean correctas es menor en la medida en que los niveles de corte de las correlaciones son más bajos. ¿Cómo se distribuyen a lo largo de los distintos niveles de corte de correlación los genes asociados al metabolismo de la pared celular? Para contestar esta pregunta, analizamos cuántos de estos genes con la etiqueta *GO:0007047* asignada se encuentran presentes en las distintas redes construidas **(Figura 12).** Sólo aquellos genes que estén presentes en las redes de coexpresión analizadas podrán propagar esta etiqueta a genes no anotados vecinos y nos permitirán encontrar nuevos genes de interés asociados a este metabolismo que previamente no han sido asociados a este proceso. Persson y colaboradores, utilizando un enfoque "guiado por gen" analizaron las correlaciones de expresión de genes que codifican para varias celulosas sintasas y mediante análisis de intersección entre estas listas de genes, proponen nuevos genes involucrados en el metabolismo de la pared celular (Persson y col. 2005). Sin embargo, en dicho trabajo no se realiza un análisis de vecindario, pues no considera todos los vecinos de un gen y la selección de

los genes guía es arbitraria. En este trabajo, explotamos la información de todos los genes vecinos de un gen en la red de coexpresión y tomamos un enfoque "guiado por proceso biológico", centrándonos en la propagación de etiquetas de procesos biológicos de GO en las redes de coexpresión. ¿Cuántos genes es posible asociar al metabolismo de la pared celular por medio de la propagación de anotaciones implementada en este trabajo? Ya vimos que evaluando la precisión obtenida al usar todos los genes con anotación en algún proceso biológico como grupo de entrenamiento, tenemos las mayores probabilidades de que las predicciones sean ciertas utilizando una aproximación que cuenta las anotaciones presentes en el vecindario de un gen y que toma en consideración las correlaciones para proponer una anotación a un gen. Así, utilizando este algoritmo a niveles de corte de correlación ≥0.8, encontramos que podemos anotar 43 nuevos genes involucrados en el metabolismo de la pared celular con una precisión del 43.39% **(Figure 13-14)**. Además, al exigir como mínimo que exista 2 o más vecinos con la anotación de pared celular asignada para propagar las anotaciones, es posible obtener valores de precisión aún más altos, pero por cierto que el número de genes anotados usando estas condiciones más estrictas decae **(Figure 13)**.

A.

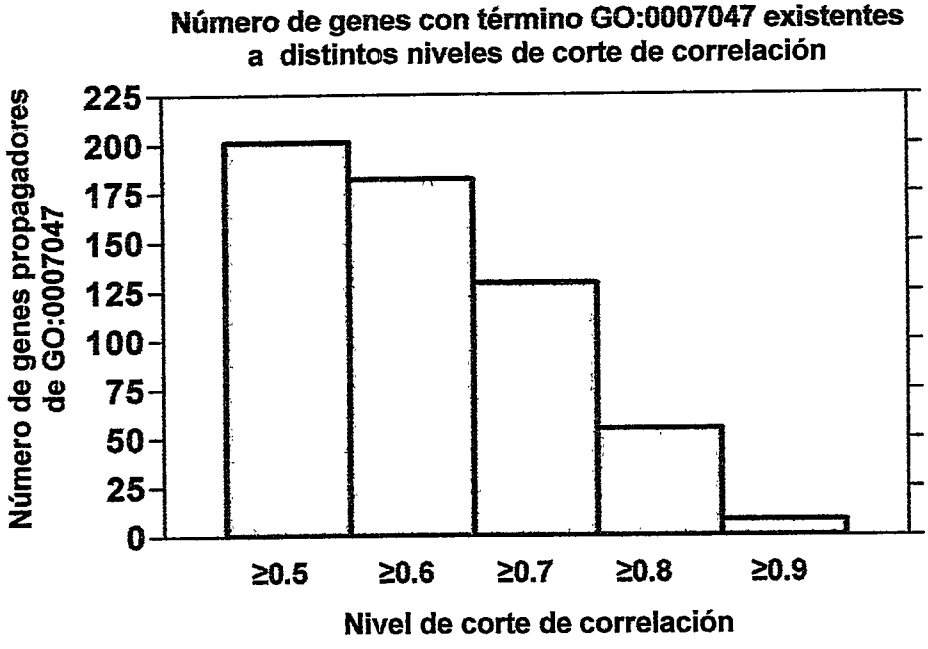



**Figura 12. Distribución de frecuencia del número de genes que poseen asignada la etiqueta *GO:0007047* a distintos niveles de corte de correlación.** (A) Utilizando el método de asignación de función de simple conteo que considera las correlaciones entre los genes como distancia, es posible propagar las anotaciones de los genes anotados a sus genes vecinos utilizando las distintas redes de correlación de la tabla 1. En Arabidopsis hasta la fecha existen 205 genes con la etiqueta *GO:0007047 "cell wall organization and biogenesis"* asignada en su descripción de Gene Ontology, una etiqueta a nivel 5 del DAG. Sin embargo, solo 201 de éstos poseen correlaciones de Spearman en sus perfiles de expresión con algún otro gen ≥ 0.5. Al aumentar el nivel de corte de correlación, el número de genes con esta etiqueta asignada disminuye. Al analizar la distribución de estos 201 genes a distintos niveles de corte de correlación, encontramos que a mayor nivel de corte de correlación, menor es el número de genes asociados al metabolismo de la pared celular que eventualmente pueden propagar sus anotaciones a sus genes vecinos. Pese a que como vimos anteriormente en la figura 10-A, la oportunidad de las predicciones de ser correctas es más alta a mayores niveles de corte de correlación.

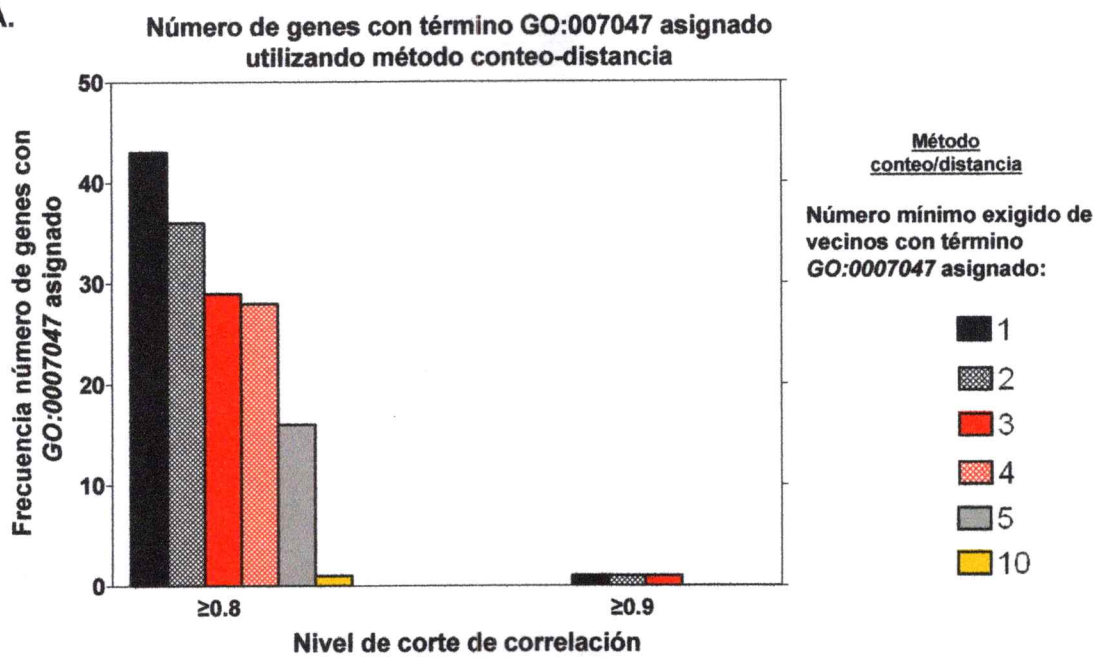



**Figura 13. El número de genes a los cuales la etiqueta *GO:0007047* es asignada decae si se incrementa el número mínimo exigido de vecinos con dicha anotación.** (A) Utilizando el método de asignación de función de simple conteo que considera las correlaciones entre los genes como distancia, se propagó las anotaciones de los genes que poseen la etiqueta *GO:0007047 "cell wall organization and biogenesis"* asignada en su descripción de Gene Ontology, a sus vecinos en la red de coexpresión obtenida seleccionando aquellas correlaciones de Spearman ≥ 0.8 y ≥ 0.9. Como vimos en la figura 11-B, a estos niveles de corte de correlación la oportunidad de que las predicciones sean correctas es más alta a mayores niveles de corte de correlación y en la medida en que el número mínimo exigido de genes que posean esta anotación en el vecindario de un gen es mayor. Así, por ejemplo, al tomar un nivel de corte de correlación Spearman ≥ 0.8 y exigir que como mínimo exista 1 vecino con esta anotación en el vecindario de un gen para que la etiqueta *GO:0007047* le sea asignada, es posible asignar la etiqueta *GO:0007047* a 43 genes que previamente no tenían esta etiqueta asignada, con una precisión de 48.92%. Sin embargo, si se exige que el número mínimo de vecinos con esta anotación en el vecindario de un gen sea 5, la precisón es de 54.18%, pero sólo es posible asignar esta etiqueta a 16 genes. Al tomar un nivel de corte de correlación Spearman ≥ 0.9, la precisión es más alta, alcanzando el 74.11%. Sin embargo, usando estos parámetros es posible asignar la etiqueta *GO:0007047* a un único gen.

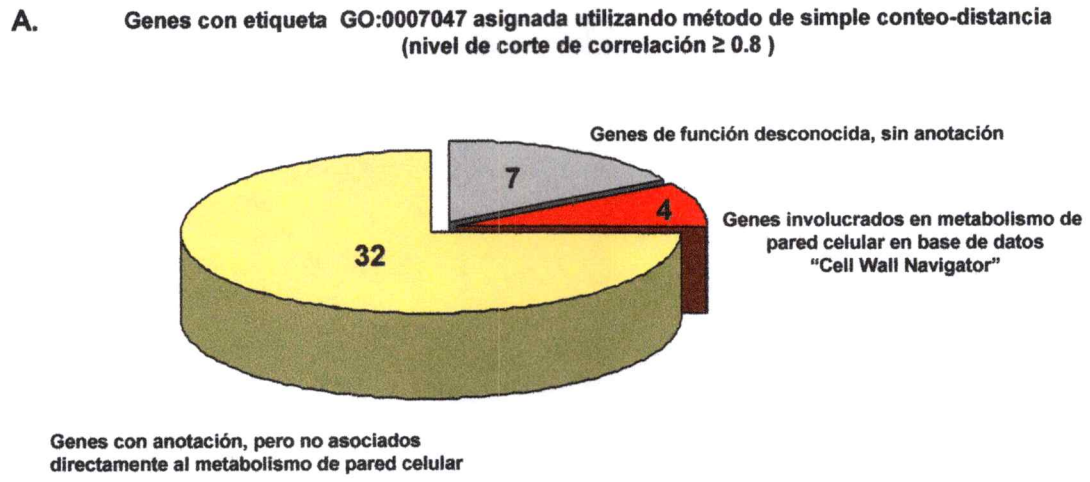



**Figura 14. Descripción de los genes a los cuales la etiqueta GO del metabolismo de la pared celular *GO:0007047* es asignada.** (A) Utilizando un método de simple conteo que considera las correlaciones entre los genes como distancia, tomando un nivel de corte de correlación de Spearman ≥ 0.8 y exigir que como mínimo exista 1 vecino con la anotación a ser asignada en el vecindario de un gen para que la etiqueta sea asignada, es posible asignar la etiqueta *GO:0007047* a 43 genes que previamente no poseían esta etiqueta asignada en su anotación. En la figura, se representa estos genes en un gráfico de torta, donde 7 de estos genes son genes que codifican para proteínas sin ninguna anotación asignada *("unknown proteins")*; 4 corresponden a genes que por análisis de identidad de secuencias han sido previamente implicados en el metabolismo de la pared celular en la base de datos Cell Wall Navigator (Girke y col. 2004) y 32 corresponde a genes que poseen anotación, pero no han sido asociados directamente al metabolismo de la pared celular (ver tabla 2).

| Código | Descripción (TAIR) |
|---|---|
| AT2G34910 | similar to unknown protein [Arabidopsis thaliana] (TAIR:AT1G30850.1) |
| AT3G07900 | similar to unknown protein [Arabidopsis thaliana] (TAIR:AT2G44500.1) |
| AT3G59340 | similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G59310.1) |
| AT3G51540 | similar to unknown protein [Arabidopsis thaliana] (TAIR:AT3G08670.1) |
| AT5G16100 | similar to hypothetical protein [Cleome spinosa] (GB:ABD96916.1) |
| AT5G24313 | unknown protein |
| AT5G14330 | unknown protein |
| AT1G54970 | ATPRP1 (PROLINE-RICH PROTEIN 1); structural constituent of cell wall |
| AT1G48930 | ATGH9C1 (ARABIDOPSIS THALIANA GLYCOSYL HYDROLASE 9C1); hydrolase, hydrolyzing O-glycosyl compounds |
| AT2G20520 | FLA6 (fasciclin-like Arabinogalactan-protein 6) |
| AT4G40090 | AGP3 (Arabinogalactan-protein 3) |
| AT3G62710 | glycosyl hydrolase family 3 protein |
| AT5G22410 | peroxidase, putative |
| AT2G39040 | peroxidase, putative |
| AT3G49960 | peroxidase, putative |
| AT4G26010 | peroxidase, putative |
| AT4G26010 | peroxidase, putative |
| AT5G17820 | peroxidase 57 (PER57) (P57) (PRXR10) |
| AT5G67400 | peroxidase 73 (PER73) (P73) (PRXR11) |
| AT1G30870 | cationic peroxidase, putative |
| AT2G48080 | oxidoreductase, 2OG-Fe(II) oxygenase family protein |
| AT2G42850 | CYP718 (cytochrome P450, family 718); oxygen binding |
| AT2G25160 | CYP82F1 (cytochrome P450, family 82, subfamily F, polypeptide 1); oxygen binding |
| AT5G60020 | LAC17 (laccase 17); copper ion binding / oxidoreductase |
| AT3G05930 | GLP8 (GERMIN-LIKE PROTEIN 8); manganese ion binding / metal ion binding / nutrient reservoir |
| AT3G62020 | GLP10 (GERMIN-LIKE PROTEIN 10); manganese ion binding / metal ion binding / nutrient reservoir |
| AT5G47950 | transferase family protein |
| AT4G25250 | invertase/pectin methylesterase inhibitor family protein |
| AT4G25220 | transporter, putative |
| AT3G07000 | DC1 domain-containing protein |
| AT1G69240 | hydrolase, alpha/beta fold family protein |
| AT2G25240 | serine-type endopeptidase inhibitor |
| AT4G30320 | allergen V5/Tpx-1-related family protein |
| AT4G30420 | nodulin MtN21 family protein |
| AT4G15740 | C2 domain-containing protein |
| AT5G19800 | hydroxyproline-rich glycoprotein family protein |
| AT5G61650 | CYCP4;2 (CYCLIN P4;2); cyclin-dependent protein kinase |
| AT5G10130 | pollen Ole e 1 allergen and extensin family protein |
| AT5G05500 | pollen Ole e 1 allergen and extensin family protein |
| AT4G02270 | pollen Ole e 1 allergen and extensin family protein |
| AT2G33790 | pollen Ole e 1 allergen and extensin family protein |
| AT4G33730 | pathogenesis-related protein, putative |
| AT4G31470 | pathogenesis-related protein, putative |
| AT1G50060 | pathogenesis-related protein, putative |

**Tabla 2. Detalle de los códigos de acceso de los genes con la etiqueta Gene Ontology *GO:0007047 "cell wall organization and biogénesis"* asignada.** Ninguno de los 43 genes con esta etiqueta asignada por el método de simple conteo que considera la distancia entre los nodos, poseían esta etiqueta GO previamente en su anotación. En la tabla se detalla los códigos de acceso y la descripción TAIR de cada uno de estos genes (Rhee y col. 2003; Swarbreck y col. 2008), respetando los colores utilizados en la Figura 14 para clasificarlos. En gris *("unknown proteins")*, en rojo los genes que por análisis de identidad de secuencias han sido previamente implicados en el metabolismo de la pared celular y en amarillo los genes que poseen anotaciones previas, pero que no los relacionan al metabolismo de la pared celular directamente.

## 5. Red regulatoria del metabolismo de la pared celular

Los factores de transcripción (TF), son proteínas que poseen dominios de unión a DNA y que pueden regular la expresión de un gen al reconocer secuencias específicas o cajas regulatorias en la región promotora de un gen blanco, ya sea al facilitar o inhibir el reclutamiento de las proteínas de la maquinaria de transcripción de un gen. Además, los genes que codifican para un factor de transcripción, son a su vez regulados en su expresión. En el área de la llamada genética molecular, un problema central es justamente comprender la regulación de la expresión génica y dilucidar cual(es) son los factores de transcripción que regula(n) la expresión de un gen o grupo de genes de interés, lo que determina que estas interacciones regulatorias entre un TF y su(s) gen(es) blanco(s) sean posibles de estudiar utilizando redes que representen este control transcripcional en sus arcos y donde los nodos representen a los TF y los genes blanco. Estas redes, reciben el nombre genérico de redes regulatorias, aunque existen muchos tipos distintos posibles de estas redes dependiendo de que información se integra y cómo se procede en su construcción. *Arabidopsis thaliana* es un organismo modelo que cuenta con la disponibilidad pública de diversa información útil para análisis de datos de interés en la confección de una red regulatoria, tales como las secuencias río arriba del inicio de la traducción de todos los genes que codifican para proteínas y la asociación entre dominios de unión a DNA presentes en los factores de transcripción (TF) y las cajas regulatorias o elementos en Cis reconocidos por

éstos. Además, es posible analizar en dichas secuencias la frecuencia de cada una de estas cajas, en las regiones promotoras de todos los genes que codifican para proteínas y utilizar esta información para confeccionar redes regulatorias (Obayashi y col. 2007; Gutierrez y col. 2008; Swarbreck y col. 2008).

Con el objetivo de construir una red que represente en sus arcos las interacciones regulatorias entre los genes involucrados en el metabolismo de la pared celular y los factores de transcripción que regulan su expresión, utilizamos los datos de coexpresión y las posibles interacciones regulatorias entre factores de transcripción (TF) y su(s) gene(s) blanco(s) de forma conjunta construyendo redes regulatorias, donde los arcos entre los nodos no poseen dirección (ver métodos) **(Figura 15-16).** Dentro de los mapeos TF-gen blanco, seleccionamos aquellos que incluyeran los 43 genes a los cuales les asignamos la anotación en pared celular **(Figura 14)** y los 205 genes que ya poseen asignada la etiqueta *GO:0007047* correspondiente a este proceso metabólico a nivel 5 del DAG de las anotaciones de Gene Ontology **(Figura 9)** (ver figura suplementaria 1).

Al analizar y representar el grado ($k$) versus los coeficientes de clustering ($CC$) de cada uno de los nodos de la red regulatoria de la pared celular construida, encontramos que 4 de los nodos poseen grados muy altos con respecto a los otros 3.417 existentes en la red y que corresponden a genes que codifican para 4 factores de transcripción **(Figura 16-B)**, lo que indica que estos TF son muy interesantes de estudiar pues destacan en su alto grado en

comparación con los otros 691 TF incluidos de la red regulatoria, pues son genes que poseen interacciones regulatorias con muchos de los 248 genes involucrados en el metabolismo de la pared celular incorporados en este análisis **(Figura 15-16).**

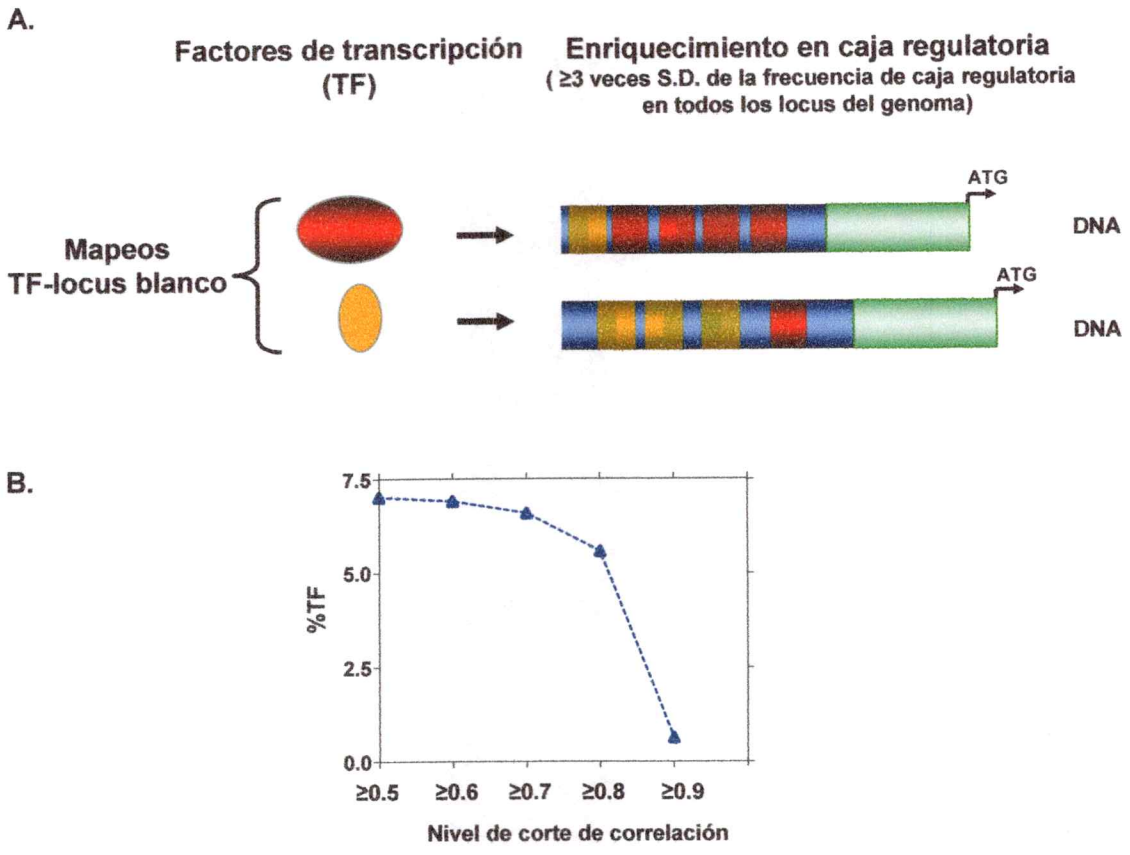



**Figura 15. Esquema que representa la obtención de interacciones regulatorias entre los factores de transcripción y sus genes blancos involucrados en el metabolismo de la pared celular.** (A) Para la construcción de la red regulatoria del metabolismo de la pared celular, utilizamos las interacciones regulatorias entre factores de transcripción (TF) y su(s) gene(s) blanco(s) que poseen un enriquecimiento del elemento en Cis regulatorio que reconoce el TF y luego seleccionamos aquellos mapeos que posean correlaciones ≥ 0.5 entre el TF y su gen blanco mapeado (ver métodos). En la figura, se representa cómo se realizó la selección de los mapeos entre un TF y su gen blanco que evidencian una posible interacción regulatoria. Aquellos genes blanco que poseen el elemento en Cis regulatorio con una frecuencia igual o superior a 3 veces la desviación estándar de la presencia de este elemento regulatorio en todos las regiones río arriba del inicio de la traducción de todos los genes del genoma, fueron mapeados con el TF respectivo. Estos mapeos fueron previamente desarrollados por Gutiérrez y colaboradores, analizando 1.5 Kb río arriba de cada uno de los genes del genoma (Gutierrez y col. 2008). (B) Gráfico de la frecuencia de los factores de transcripción (TF) presentes en las distintas redes de coexpresión de la tabla 1. Nótese que muy pocos factores de transcripción poseen correlaciones de expresión ≥ 0.9.

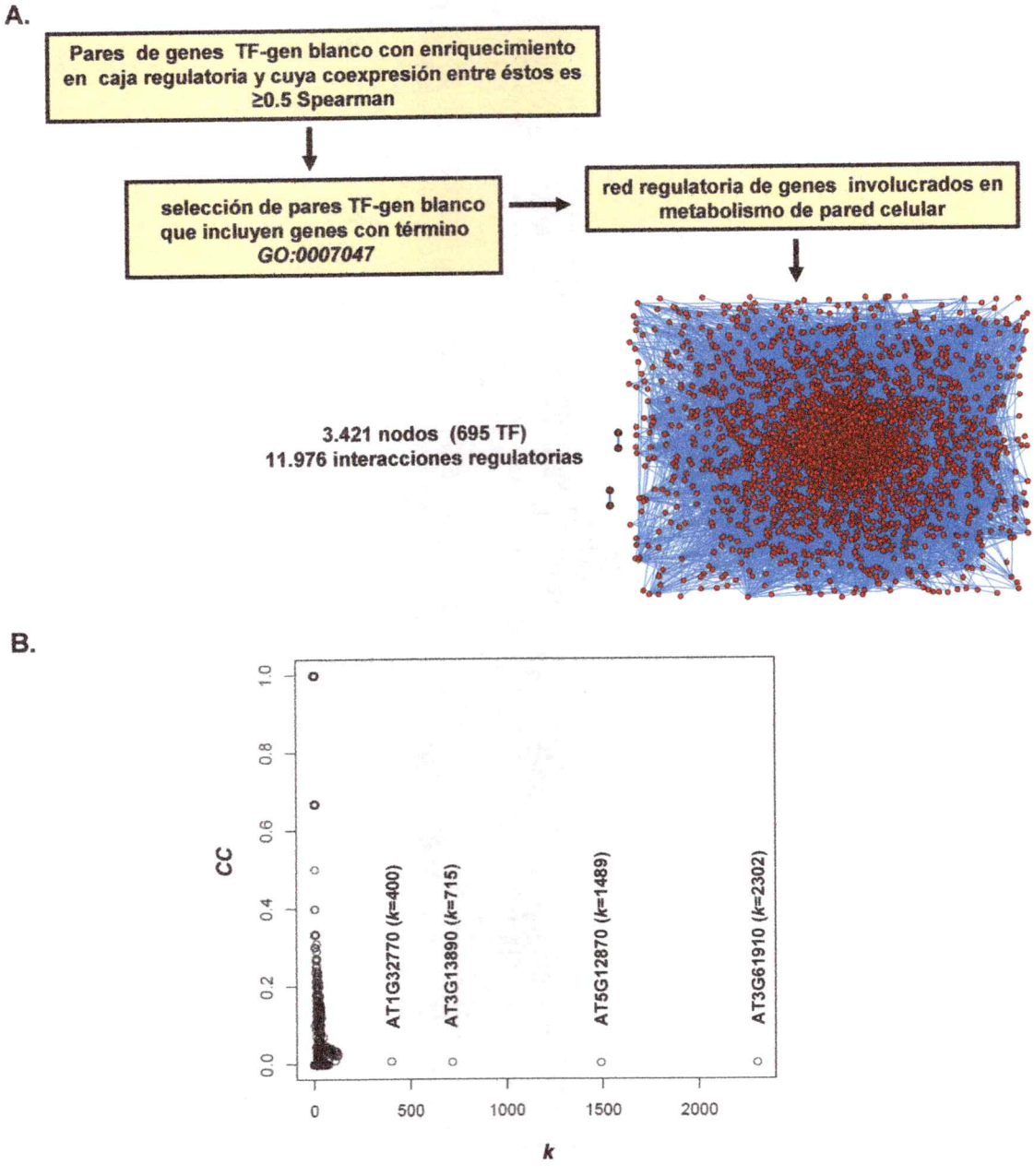



**Figura 16. Análisis de red regulatoria de metabolismo de pared celular.** (A) Diagrama que representa los pasos seguidos para la construcción de una red regulatoria del metabolismo de la pared celular (ver métodos). Al analizar la red, es posible observar que esta posee tres componentes: un gran componente conexo y dos pares de interacciones de dos nodos cada una no conectados a este. (B) Gráfico que representa el grado (*k*) versus los coeficientes de clustering (*CC*) de cada uno de los 3.421 nodos de la red regulatoria de la pared celular. Nótese que hay 4 nodos que poseen grados muy altos con respecto a los otros nodos de la red.

# DISCUSIÓN

Predicción de la función génica en la era post-genómica

Después de que el genoma de 120 Mb de *Arabidopsis thaliana* fue completamente secuenciado el año 2000, la biología vegetal ha puesto un gran esfuerzo en la dilucidación de la función de cada uno de los genes codificantes para proteínas encontrados en estas secuencias. Inicialmente, la anotación de estos genes consistió en la predicción computacional de su localización y de sus elementos estructurales (intrones, exones y putativos elementos regulatorios), junto con la posterior caracterización empírica de sus estructuras por medio de análisis de transcritos de RNA por técnicas masivas que generan muchos datos, tal como los microarreglos. Hoy, tenemos todos estos resultados tanto teóricos como experimentales y el gran desafío es lograr obtener información de la función de estos genes, pues pese a ser Arabidopsis el primer organismo vegetal en tener su genoma secuenciado y ser un excelente organismo modelo en genómica funcional ampliamente utilizado, solo alrededor del 50% de los genes de esta planta poseen un proceso biológico asignado. Debido a esto, proponer algún rol en algún proceso a los genes remanentes es un gran desafío y de mucha utilidad, pues dicha información además de ayudarnos a comprender el metabolismo vegetal y reducir los genes de interés en cierto proceso en particular, podrá ayudar en los trabajos realizados en otros sistemas vegetales por medio de la genómica comparativa, donde el trabajar

con individuos mutantes o con alteraciones puntuales en la secuencia de un gen es muy difícil, ya sea por dificultades técnicas o debido a que el ciclo de vida del organismo en cuestión es demasiado prolongado como para obtener resultados masivos en el corto plazo.

En este trabajo, nuestro desafío ha sido lograr encontrar nuevos genes involucrados en el metabolismo de hemicelulosas y pectinas, polímeros de azúcar de la pared celular sintetizados en la ruta secretoria **(Figura 1).** Para ello, utilizando datos públicos de microarreglos construimos redes de coexpresión y las utilizamos para proponer una lista de genes que tengan una alta oportunidad de efectivamente estar involucrados en estos procesos metabólicos.

¿Cómo es posible asignar algún proceso biológico a un gen?

La transferencia de la información funcional de un gen a otro utilizando su relación biológica recibe el nombre genérico de "construir por asociación". En este trabajo, gracias a la disponibilidad pública de muchos datos de microarreglos existente en Arabidopsis, construimos redes de coexpresión para inferir propiedades funcionales de los genes, pues esperamos que genes que codifiquen para proteínas que participan en el mismo proceso biológico, sean co-expresados si se analiza muchas condiciones distintas de crecimiento y experimentos de expresión. Al utilizar las redes construidas en este trabajo para propagar la información funcional existente en las anotaciones de los genes anotados, podemos anotar los genes no anotados, utilizando un enfoque de análisis global que justamente construye por asociación, pues en base a las

interacciones por coexpresión es posible proponer algún proceso biológico a los genes sin un proceso asignado y nuevos procesos a los genes con anotación funcional existente (ver métodos).

¿Cómo se construyeron las redes de coexpresión analizadas?

En este trabajo, utilizando las correlaciones de expresión de Spearman calculadas entre 21.180 perfiles de expresión normalizados obtenidos desde 1.701 chips de microarreglos de calidad, construimos distintas redes de coexpresión utilizando las correlaciones positivas y tomando distintos niveles de corte en los pares de correlaciones, para justamente considerar esta variable en nuestros análisis realizados **(Tabla 1).** Las correlaciones lineales de los pares de perfiles de expresión utilizadas en la confección de las redes, representan un número bastante bajo con respecto al universo de 260.136.645 correlaciones entre pares de perfiles de expresión calculadas **(Figura 4-5).** Nótese que las correlaciones lineales de Spearman y Pearson poseen una relación lineal entre sí en nuestros datos (datos no mostrados). Sin embargo, en la confección de las redes utilizamos las correlaciones de Spearman, debido a que es un algoritmo menos sensible a valores extremos. Un resultado interesante que encontramos, es que hay alrededor de tres veces más correlaciones positivas que negativas si comparan correlaciones $\geq 0.5$ y las $\leq -0.5$ obtenidas. Lo que indica que los genes, ante las diversas condiciones de crecimiento incluidas en los chips analizados, tienden a expresarse juntos **(Figura 4-B).** Debemos mencionar, que con los chips ATH1 Affymetrix™ utilizados, alcanzamos ~80%

de cobertura del total de los genes codificantes para proteínas existentes en *Arabidopsis thaliana.*

En este trabajo, para anotar los genes por medio de la propagación de anotaciones en las redes de coexpresión, utilizamos las anotaciones de Gene Ontology (GO) de "Procesos Biológicos" existentes en los vecindarios de los genes (ver métodos). Utilizamos estas etiquetas de anotación, pues estamos estudiando el metabolismo de las hemicelulosas y pectinas, procesos biológicos en el que participan distintos tipos de proteínas con funciones moleculares diversas. Debido a esto, necesitamos implementar una metodología que nos permita anotar los genes en relación con los procesos biológicos que poseen asignados, no las anotaciones de "función molecular", las cuales se basan en análisis de identidad de secuencias y no entregan información de los procesos metabólicos en que las proteínas participan. Recordemos que la función metabólica de los productos proteicos codificados por los transcritos que detectamos en los experimentos de microarreglos o cualquier otro, no radica directamente en la función molecular que posean asignada, la cual generalmente es una anotación que poseen asignada debido a una similitud a nivel de identidad de secuencias con alguna otra proteína descrita en algún otro organismo, sino que se basa en las interacciones entre los productos proteicos y otras moléculas, en un proceso dinámico.

¿Qué nos indica el análisis de la topología de nuestras redes de coexpresión?

Previo a la utilización de nuestras redes de coexpresión como herramienta en genómica funcional, es necesario analizar las estructuras de

estas redes por un análisis topológico vía la obtención de distintas variables, pues éste análisis permite analizar las interacciones de coexpresión entre los genes de forma global y obtener importante información de las redes. Al igual que lo observado en redes de interacción proteína-proteína y otras redes de sistemas complejos, las redes de coexpresión analizadas en este trabajo poseen valores de coeficientes de clustering promedio *(<CC>)* muy superiores a los obtenidos en redes de igual tamaño e igual grado promedio *(<k>)*, pero construidas con modelos de redes aleatorias. Es decir, redes en las cuales las interacciones entre los nodos es al azar **(Tabla 1)**. Además, los diámetros de las nuestras redes fueron pequeños. Por ejemplo, en la red obtenida al seleccionar aquellos pares de correlaciones ≥ 0.5 el diámetro obtenido es 8. Esto implica que como máximo con sólo 8 arcos es posible unir cualquier par de nodos de esta red, pues se trata de una red con muchos arcos y que presentan un alto grado de conexión entre los nodos que la componen. Al tener un diámetro pequeño y coeficientes de <CC> muy superiores a los obtenidos en redes aleatorias de igual tamaño, podemos decir que las redes de coexpresión, obedecen al llamado "fenómeno de pequeño mundo", encontrado en redes de diámetros pequeños y coeficientes de clustering promedio cercanos a ½ (Watts y col. 1998; Amaral y col. 2000).

Al analizar las topologías de redes de distintos sistemas complejos se ha observado que no todas las redes son exactamente escala libre, pues algunas poseen una distribución de frecuencia de las conectividades que si bien obedece a una ley de potencia en cierto rango de conectividad, presentan una

"truncación" en la distribución en la zona de los nodos más conectados. Este tipo de distribuciones recibe el nombre de distribuciones de conectividad con "cola exponencial" (Mossa y col. 2002). Los análisis de topología realizados sobre las redes de coexpresión analizadas en este trabajo, revelan que son escala libre, con una distribución de frecuencia de los grados truncada en la zona de los nodos más conectados **(Figura 6-A).**

Un gen altamente conectado o "hub", es un gen que coexpresa con muchos otros genes y que participa en muchos procesos metabólicos. Así, los hubs son muy buenos candidatos para analizar el efecto sobre la planta de su mutación o deleción, esperándose un gran efecto fenotípico si dicha alteración es efectiva. En nuestras redes, sabemos que hay muy pocos genes con un grado muy alto comparados con el resto de los genes. Además, sabemos que los nodos están muy agrupados entre sí, pues poseen <*CC*> muy superiores a los esperados para una red aleatoria. Al examinar la relación entre los *(CC)* y los grados de cada gen en las redes de coexpresión, observamos a los genes agrupados en "bandas", donde se observa que a medida que aumenta el grado, los coeficientes de clustering también decrecen, lo que indica que nuestras redes poseen módulos y estructura jerárquica **(Figura 6-B).** En los distintos niveles de corte de correlación analizados observamos una dependencia entre los *CC* y la *k* **(Figura 6 - Figura suplementaria 2).** Básicamente lo que esto implica, es que en las redes hay nodos divididos en grupos, que además se subdividen en otros subgrupos. Grupos que en muchas redes analizadas en distintas disciplinas de la ciencia, han sido coincidentes con unidades

funcionales previamente descritas, tales como vías metabólicas en redes metabólicas, nichos ecológicos en redes ecológicas o comunidades en redes sociales (Ravasz y col. 2002; Clauset y col. 2008). Sin embargo, ¿Están todos los nodos conexos o conforman módulos funcionales no conectados entre sí? Esto es muy importante de analizar, pues nosotros lo que estamos haciendo es propagar las anotaciones de los genes anotados a sus vecindarios, lo que claramente podría estar relacionado con posibles agrupamientos funcionales existentes entre los genes. Debido a esto, analizar los componentes es un punto importante a considerar en las redes y que puede estar relacionado con la estructura jerárquica y modular observada. Dos genes están definidos como parte del mismo componente, si existe un camino a lo largo de distintos arcos de la red, que los conecte. Al analizar el número de componentes presentes en cada uno de los niveles de corte analizados, observamos que a menor nivel de corte de correlación la mayoría de los genes están agrupados en un solo gran componente. Además, nótese que la red construida con las correlaciones de Spearman $\geq 0.8$ es la que posee mayor número de componentes **(Figura 7).**

¿Que información biológica podemos obtener de este análisis? Al analizar funcionalmente los genes más conectados es posible saber si hay algo que los caracteriza. Al analizar sus anotaciones es posible saber si por ejemplo, estos genes están enriquecidos en cierto tipo particular de anotaciones correspondientes a ciertos procesos biológicos en particular o funciones moleculares determinadas. En las redes escala libre de distintos sistemas complejos, se ha observado que si bien estas redes son robustas a la

eliminación aleatoria de un nodo, son muy vulnerables a la eliminación o ataques selectivos de los nodos más conectados. El número de genes hubs analizado en este trabajo, para cada nivel de corte de correlación es distinto, pues tomamos como hubs al 1% de los genes más conectados de cada red y cada una de éstas, posee distintos números de nodos **(Tabla 1).** Al analizar en las listas de genes hubs de las distintas redes, posibles enriquecimientos de anotaciones en estos grupos de genes, encontramos que los hubs de las redes analizadas, están enriquecidos en genes sin una función molecular (FM) asignada. Muchos de estos genes, no son posibles de anotar por análisis de identidad de secuencias. Sin embargo, son genes muy importantes, pues co-expresan con muchos otros genes **(Figura 8-A).** El gen AT4G05230 es el gen más conectado en las redes utilizando las correlaciones ≥ 0.5, 0.6 y 0.7. Este gen posee el proceso biológico "ubiquitin cycle" asignado, lo que hace bastante sentido, pues se trata de un proceso biológico muy importante para la células eucariontes, pues regula la ubiquitinación y des-ubiquitinación de proteínas. La ubiquitina, es un péptido que puede actuar como señal en el control de la degradación de proteínas, en la vía de ubiquitinación-degradación en el proteosoma. En plantas, este mecanismo de control de la degradación de proteínas, participa en muchos procesos involucrados en el desarrollo y crecimiento, tales como la fotomorfogénesis y respuestas hormonales (Hellmann y col. 2002; Schwechheimer y col. 2004). Por su parte, los genes más conectados en las redes construidas utilizando las correlaciones ≥ 0.8 y 0.9 corresponden a los genes AT1G67700 y AT1G15980 respectivamente.

AT1G67700 no posee ningún proceso biológico asignado y ninguna función molecular asignada, lo que lo hace un gen muy interesante de estudiar más en profundidad. AT1G15980, es un gen que ha sido recientemente involucrado en el transporte de electrones del Fotosistema I del cloroplasto para producir ATP (Takabayashi y col. 2008) **(Figura 8-A).**

La utilización de líneas insercionales de T-DNA como herramienta en genómica funcional en *Arabidopsis thaliana* ha tenido un explosivo crecimiento en los últimos años y es la principal metodología responsable de las anotaciones que relacionan a un gen de esta planta con algún fenotipo obtenido (Meinke y col. 2003; Li y col. 2006). Cuando analizamos en los genes mas conectados la frecuencia de inserciones de T-DNA en ambas copias de estos, encontramos que al contrastar estas frecuencias contra las frecuencias esperadas por simple azar, observamos que las listas de genes "hubs" de nuestras redes, presentan un número mucho menor de genes con una línea homocigota de T-DNA que lo esperado por azar. Esto sugiere, que estos genes son muy importantes y probablemente esenciales, pues la eliminación de ambas copias de estos genes puede estar generando una mayor tasa de letalidad en estas líneas insercionales y sesgando los resultados a la observación de que hay menos líneas insercionales homocigotas en estos genes de lo esperado **(Figura 8-B).**

¿Cómo encontramos nuevos genes involucrados en el metabolismo de la pared celular utilizando nuestras redes de coexpresión?

Se estima que hay muchos genes involucrados en el metabolismo de la pared celular que aún no han sido descubiertos, pues la complejidad de los enlaces de los polímeros de azúcar de la pared celular requiere un gran repertorio enzimático (Somerville y col. 2004). Debido a esto, proponer nuevos genes involucrados en su metabolismo es un gran aporte para la biología vegetal, pues además de entregar nuevos potenciales participantes en este complejo proceso, reduciendo el universo de potenciales blancos, podemos postular cuales de ellos pueden originar un fenotipo más notorio al ser eliminados, basándonos en su grado (*k*) por coexpresión. Debemos recordar, que la simple coexpresión entre dos genes no necesariamente implica que ambos genes participen en el mismo proceso biológico. Simplemente es una métrica de similitud entre 2 perfiles de expresión. Por ejemplo, quizás dos genes que co-expresan tienen significantes valores de coexpresión con muchos otros genes involucrados en otros procesos biológicos, o la correlación ocurre por simple azar. Debido a este punto, en este trabajo analizamos la capacidad de tres distintos algoritmos que en base al vecindario completo de un gen en la red de coexpresión, son capaces de asignarle algún proceso biológico. Sin embargo, ¿Cómo comparamos estos métodos y evaluamos cuál nos entrega las mejores predicciones? Lo que hicimos en este trabajo, fue utilizar los genes con alguna etiqueta GO de "proceso biológico" (BP) asignada como grupo de entrenamiento, seleccionando el método con la mejor recuperación de las anotaciones de los genes del grupo de entrenamiento para realizar nuestras predicciones. Es decir, el método con los valores de precisión más altos **(Figura**

**10).** Analizando los vecindarios de cada gen en las redes de coexpresión, asignamos anotaciones, analizando el rendimiento de tres distintos algoritmos de asignación de función muy simples: i) un algoritmo de simple conteo, ii) un algoritmo de simple conteo que además considera la distancia con los distintos vecinos de un gen a anotar (conteo-distancia) y iii) un algoritmo hipergeométrico, el cual propone como la anotación a asignar, aquella más sobre representada en comparación con lo esperado por azar en el vecindario analizado, ajustando por las frecuencias esperadas por la abundancia de las anotaciones en todos los genes del genoma. A niveles de corte de correlación más cercanos a 1, las predicciones poseen una mayor oportunidad de ser certeras **(Figura 11).** Sin embargo, es importante mencionar que en general, a altos niveles de corte de las correlaciones, menor es el número de genes que podrán propagar sus anotaciones. Por ejemplo, cuando se toman las correlaciones de Spearman ≥ 0.9, donde 1.056 genes están presentes en la red de correlación, es posible anotar pocos genes. Nótese que se observa valores de precisión más altos, cuando tomamos las correlaciones de Spearman ≥ 0.5, donde pese a existir 20.145 genes en la red de coexpresión, la oportunidad de que las predicciones sean las correctas es menor **(Figura 11).** Es posible incrementar los valores obtenidos de precisión, exigiendo que las anotaciones asignadas estén presentes un número mínimo de veces en el vecindario de un gen para que éstas le sean asignadas. Sin embargo, debemos mencionar que el efecto de aumentar esta exigencia, disminuye el número de genes anotados. Además, tomando las correlaciones ≥ 0.5 y ≥ 0.6, no se observa que exista un

efecto de incremento los valores de precisión obtenidos **(Figura 11-B).** Según los análisis de los valores de precisión, en todos los niveles de corte de correlación analizados, el método con la mayor precisión es el de simple conteo que considera las distancias entre los nodos. Debido a esto, fue el método utilizado para encontrar nuevos genes involucrados en el metabolismo de la pared celular. Para propagar una etiqueta de GO que nos permita encontrar genes involucrados en el metabolismo de la pared celular, en este trabajo propagamos la etiqueta Gene Ontology *GO:0007047 "cell wall organization and biogénesis"*, la cual la poseen asignada 205 genes. Esta etiqueta se encuentra a nivel 5 en el grafo acíclico que organiza estas anotaciones. Debemos resaltar, que el hecho de que existan estos genes, no garantiza que propaguen sus anotaciones en las redes, pues dependerá de si coexpresan o no con otros genes y con que niveles de correlación. Por ejemplo, al analizar cómo se distribuyen las coexpresiones de estos genes, observamos que tomando únicamente las correlaciones ≥ 0.8, sólo 55 de los 205 poseen algún vecino a los cuales eventualmente podrán propagar su anotación **(Figura 12).** A este nivel de corte, al propagar las anotaciones de estos genes en las redes, somos capaces de proponer la participación en el metabolismo de la pared celular de 43 nuevos genes, los cuales no poseían previamente esta anotación asignada, con una certeza de que las predicciones sean efectivas estimada del 48.92% **(Figura 13-14).** Debemos resaltar, que debido a que un gen puede participar en 1 o más procesos biológicos, nuestra metodología permite asignar un rol en cierto proceso a genes que previamente no poseen ningún proceso asignado en

anotación o a genes que estaban relacionados a otro proceso biológico. Al cumplir el ambicioso objetivo de encontrar nuevos genes involucrados en el metabolismo la pared celular, encontramos nuevos genes de potencial interés en el estudio del metabolismo de la pared celular, el cual es un proceso metabólico que pese a tener una gran importancia en el desarrollo y una multiplicidad de procesos vegetales, posee pocos genes asignados. De los 43 genes a los cuales les asignamos la anotación GO que los involucra en el metabolismo de la pared celular **(Tabla 2)**, 7 codifican para proteínas sin ninguna anotación asignada, lo que les da un gran potencial, por la novedad que trabajar con ellos implica; 4 corresponden a genes que por análisis de identidad de secuencias han sido previamente implicados en el metabolismo de la pared celular, pues poseen identidad de secuencias con genes ya asociados a este metabolismo en la base de datos "Cell Wall Navigator" (Girke y col. 2004). 32 de los genes, poseen alguna anotación asignada, pues poseen en sus secuencias algún dominio proteico previamente descrito. Sin embargo, esto sólo nos da información de la función molecular de los productos proteicos de estos genes y no permite directamente asociar a los genes con el metabolismo de la pared celular, pues en este proceso biológico participan distintos tipos de funciones moleculares.

Nuestra aproximación "centrada en proceso(s) biológico(s)" es independiente de análisis de secuencias y entrega una ventaja con respecto al conocimiento previo. Además, puede ser utilizada para proponer nuevos genes involucrados en muchos otros procesos biológicos y no tan sólo genes

involucrados en el metabolismo de la pared celular, con conocimiento de las certezas de las predicciones.

Red regulatoria del metabolismo de la pared celular

En este trabajo, con el objetivo de construir una red que represente las interacciones regulatorias entre los genes involucrados en el metabolismo de la pared celular y los potenciales factores de transcripción que regulan su expresión, utilizamos los datos de coexpresión analizados en los puntos anteriores y las potenciales interacciones regulatorias entre factores de transcripción (TF) y su(s) gene(s) blanco(s), como fuente de información conjunta para la construcción de una red regulatoria, donde los arcos (mapeos) entre los nodos no poseen dirección (ver métodos) **(Figura 15-16).** Estos mapeos fueron previamente desarrollados Gutiérrez y colaboradores, analizando 1.5 Kb río arriba de cada uno de los genes del genoma (Gutierrez y col. 2008). Dentro de los mapeos TF-gen blanco, seleccionamos aquellos que incluyeran los 43 genes a los cuales les asignamos la anotación en pared celular y a los 205 genes que ya poseen en sus anotaciones asignada la etiqueta *GO:0007047* **(Figura 9)** (ver métodos).

¿Qué información biológica fue obtenida de la red regulatoria?

Al analizar y representar el grado (*k*) versus los coeficientes de clustering (*CC*) de cada uno de los nodos de la red regulatoria de la pared celular construida, encontramos que 4 de los nodos poseen grados muy altos con respecto a los otros 3.417 existentes en la red y que corresponden a genes que codifican para 4 factores de transcripción **(Figura 16-B).** Esto nos indica que

estos TF son muy interesantes de estudiar pues destacan en su alto grado en comparación con los otros 691 TF incluidos de la red regulatoria, pues son genes que poseen interacciones regulatorias con muchos de los 248 genes involucrados en el metabolismo de la pared celular incorporados en este análisis. Debido a esto, estos genes pueden ser considerados "genes maestros" de la regulación transcripcional del metabolismo de la pared celular. AT3G61910 es el más conectado de estos genes, participando en el 19.22% de los arcos de la red, teniendo interacciones con 2.302 genes. Éste, codifica para el factor de transcripción NST2 y ha sido recientemente involucrado en el engrosamiento de la pared celular y la regulación de genes del metabolismo de lignina (Zhong y col. 2008; Zhou y col. 2009). El locus AT5G12870 MYB46 posee 1489 interacciones regulatorias, lo que abarca el 12.4% de las interacciones totales. Al igual que NST2, MYB46 es un factor de transcripción que previamente ha sido involucrado en la regulación de la expresión de genes involucrados en el metabolismo de la pared celular (Zhong y col. 2007; Zhong y col. 2008). El locus AT3G13890 codifica para el factor de transcripción MYB26 y posee 715 interacciones regulatorias. Se trata de un gen que regula la expresión de genes varios involucrados en el metabolismo de la pared celular secundaria. Plantas mutantes en este gen son estériles pese a poseer polen fértil, pues las anteras no se abren y esto impide que estas estructuras se liberen desde los sacos polínicos (Yang y col. 2007). El locus AT1G32770 posee 400 interacciones regulatorias y codifica para el factor de transcripción

SND1, el cual también ha sido involucrado en la regulación de la expresión de genes de pared celular secundaria (Zhong y col. 2008; Zhou y col. 2009).

Validación experimental: ¿Por qué no se realizó?

El cuarto objetivo específico inicialmente propuesto en este trabajo, fue validar experimentalmente alguna hipótesis generada durante el desarrollo del mismo. Sin embargo, debemos señalar que este objetivo fue planteado previo al desarrollo de los análisis de validación cruzada que desarrollamos en utilizando los genes anotados como grupo de entrenamiento. Estos análisis de recuperación de información, son necesarios si se quiere tener alguna estimación de las certezas de las predicciones realizadas. Además, permiten la optimización de las recuperaciones de las predicciones y analizar en distintos algoritmos de predicción de función, el efecto de distintas variables. En nuestro trabajo, una eventual validación experimental de las predicciones de los procesos biológicos en el cual un gen participa sería por supuesto muy interesante. Sin embargo ¿Qué significa realizar una validación experimental de una predicción de función génica? Dentro de las aproximaciones de validación experimental más utilizadas en genómica funcional al trabajar con *Arabidopsis thaliana* como modelo de estudio, se encuentra el análisis de plantas mutantes, pues en este organismo modelo, se encuentran disponibles semillas de plantas con mutaciones puntuales o con inserciones en sus secuencias, en la mayoría de los genes codificantes para proteínas, las cuales pueden ser solicitadas y ser experimentalmente caracterizadas en el laboratorio (Alonso y col. 2006). Esta aproximación, que se enfoca en un gen y luego intenta por medio de análisis

posteriores asignar función a un gen, recibe el nombre de genética reversa ("reverse genetics"). Debemos resaltar que este tipo de análisis, pese a tomar años, no garantiza que necesariamente se encuentre un fenotipo en las líneas mutantes analizadas que nos permita relacionar al gen mutado con el proceso en que su producto proteico participa y que aumente el impacto de esta caracterización. Varios de los genes cuyos productos proteicos participan en la biosíntesis de polisacáridos de la pared celular se han identificado al analizar plantas mutantes de Arabidopsis. Algunas de estas plantas presentan alteraciones en el crecimiento, en la elongación del hipocotilo, defectos de la morfología de las raíces, fragilidad del hipocotilo, o efectos morfológicos en las células de algunos tejidos (Reiter y col. 1997; Arioli y col. 1998; Taylor y col. 1999; Bouton y col. 2002). Sin embargo, muchas plantas mutantes en genes involucrados en el metabolismo de la pared celular no presentan diferencias morfológicas macroscópicas evidentes en comparación con plantas silvestres, aunque al realizar análisis más exhaustivos, muestran alteraciones en la composición de azúcares de la pared celular (Bouton y col. 2002; Assaad y col. 2004; Usadel y col. 2004). Por ejemplo, la primera celulosa sintasa en ser caracterizada y subsecuentemente clonada en *Arabidopsis thaliana* fue la CesA; plantas mutantes en este gen presentan alteraciones en la composición de celulosa (Arioli y col. 1998; Somerville y col. 2004). Otros buenos ejemplos, son las glicosiltranferasas xiloglucano fucosiltransferasa AtFUT1 (Perrin y col. 1999; Perrin y col. 2003), la xiloglucano xilosiltransferasa XT1 (Faik y col. 2002) y la xiloglucano galactosiltransferasa MUR3 (Vanzin y col. 2002), pues plantas

mutantes en estas proteínas presentan alteraciones en el polímero de la pared celular sintetizado en la ruta secretoria llamado xiloglucano. Debemos considerar la posibilidad que una gran fracción de los genes cuyas plantas mutantes presentan alteraciones en la composición de azúcares de la pared celular, quizás ya han sido descubiertos, pues además de los genes que han sido ya estudiados por genética reversa, el análisis de mutantes vía una aproximación centrada en los análisis de fenotipos conocida como "forward genetics", ha sido previamente desarrollada con el objeto de encontrar genes involucrados en este metabolismo, buscando en poblaciones de plantas mutantes, efectos sobre la composición de azúcares al analizar muestras de pared celular (Reiter y col. 1997).

Cualquier asignación de función realizada con nuestro método de análisis de vecindario podría por cierto ser validada experimentalmente durante los próximos años. En nuestras predicciones hay 43 genes a los cuales les asignamos la etiqueta Gene Ontology *GO:0007047 "cell wall organization and biogénesis"* que podrían ser estudiados con mayor detalle y ser validados funcionalmente en el futuro, con alta oportunidad de entregar información nueva, pues ninguno de estos genes poseen esta etiqueta de proceso biológico asignada en su anotación original **(Figura 14- Tabla 2).** Así, estudiar estos genes durante los próximos años, es el próximo paso lógico y una de las principales proyecciones a largo plazo de este trabajo. Debemos resaltar, que nuestro trabajo es un gran aporte a la genómica funcional, pues puede ser considerado como un "screening *in silico*", que permite identificar potenciales

genes de interés con un foco en los procesos biológicos y permite disminuir el universo de potenciales blancos a un número más reducido y manejable dentro de una investigación centrada en el estudio de algún proceso biológico, dando la base para la validación experimental de algún gen con un rol propuesto en cierto proceso biológico, por medio por ejemplo, del análisis de mutantes o plantas con la expresión de los genes de estudio silenciada.

# CONCLUSIONES

- Al analizar las correlaciones de expresión se observa que hay más correlaciones positivas que negativas, lo que indica que en *Arabidopsis thaliana* los genes ante distintas condiciones de crecimiento tienden a expresarse juntos.

- Al analizar la topología de las redes de coexpresión construidas, se observa que las redes de coexpresión de *Arabidopsis thaliana* son "escala libre" en cierto rango de grados y que son redes de "pequeño mundo", con diámetro pequeño y coeficientes de clustering muy mayores a los observados en redes aleatorias de igual número de nodos y arcos. Además, son redes que presentan jerarquía, donde hay grupos de genes y subgrupos de éstos.

- Los nodos más conectados de las redes de coexpresión ("hubs") son genes que tienden a carecer de funciones moleculares asignadas y a presentar un número muy inferior al esperado de plantas mutantes con ambas copias de sus estos genes alterados, lo que y sugiere que son genes esenciales debido a que participan en muchos procesos metabólicos.

- Al analizar la capacidad de generar predicciones de los procesos biológicos en que participa un gen, utilizando tres algoritmos de análisis de vecindario en las redes de coexpresión y utilizar los genes anotados como grupo de entrenamiento, se observa que el método que genera las predicciones con la mayor oportunidad de ser ciertas es un método de simple conteo que considera las anotaciones presentes en el vecindario de un gen según su frecuencia y la correlaciones de expresión que poseen con el gen a anotar.

- Al propagar las anotaciones de los 205 genes que poseen la etiqueta de Gene Ontology *GO:0007047 "cell wall organization and biogenesis"* asignada en su descripción en las redes de coexpresión, somos capaces de proponer la participación en el metabolismo de la pared celular de 43 nuevos genes.

- Al construir una red regulatoria del metabolismo de la pared celular, encontramos 4 factores de transcripción maestros, que regulan la expresión de muchos genes involucrados en este metabolismo.

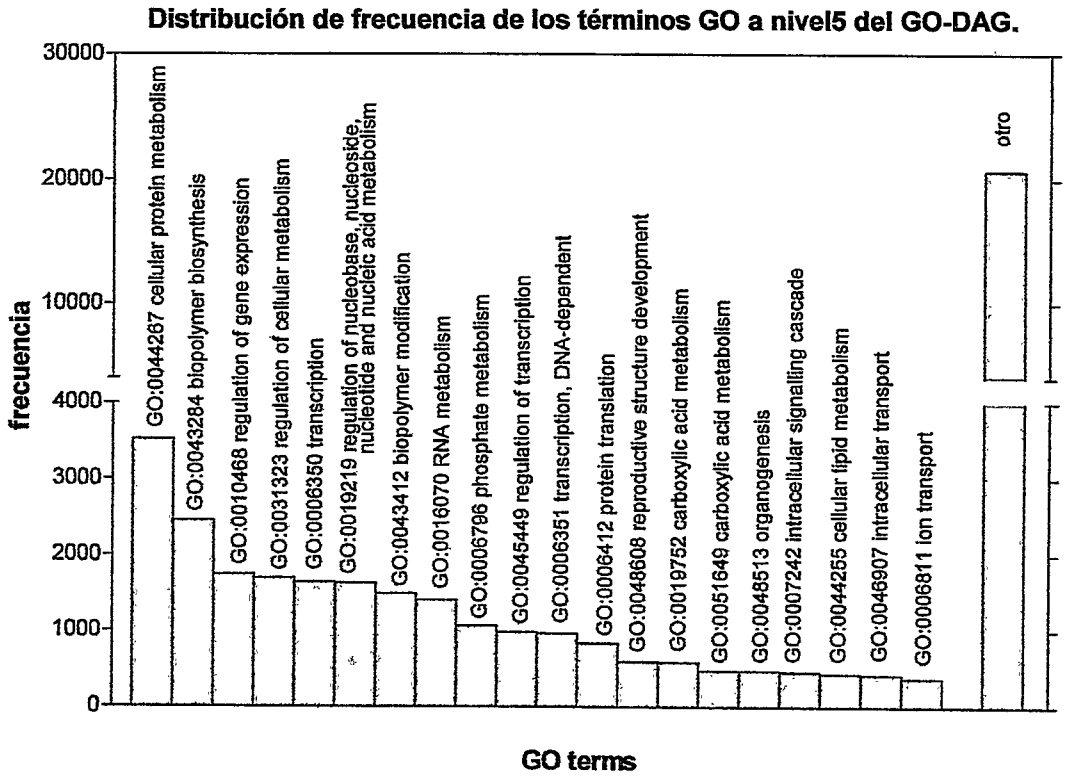



**Figura Suplementaria 1. En las anotaciones de Gene Ontology (GO) hay términos GO muy abundantes.** (A) En la figura, se representa la distribución de frecuencia de los distintos términos GO obtenida al homogenizar todo el genoma hasta a nivel 5 en las descripciones de procesos biológicos en el grafo acíclico (DAG) de sus anotaciones. Por razones de espacio, solo se muestran las descripciones de aquellos 20 términos GO más abundantes, pues los otros términos GO poseen frecuencias ≤ 363 y son representados todos juntos agrupados en la última barra, la cual incluye 125 términos GO.

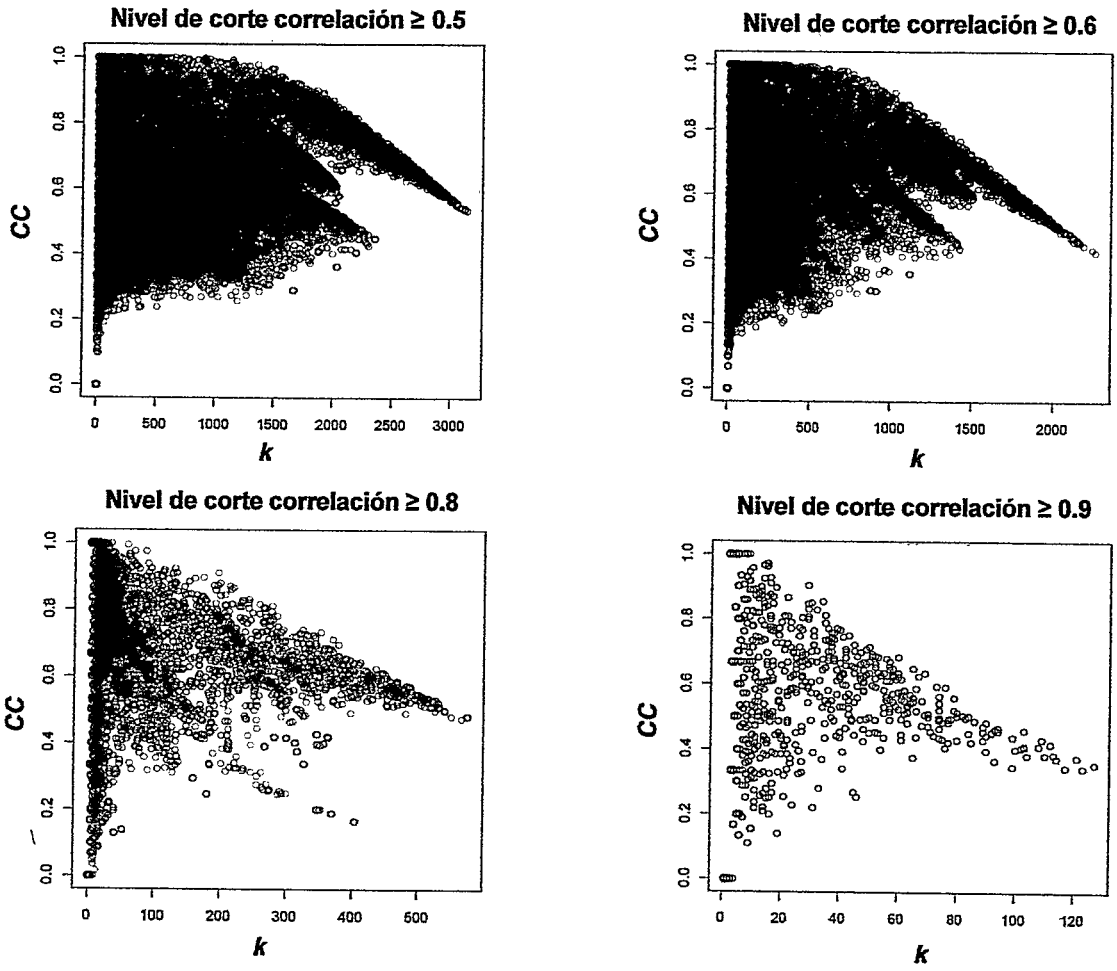



**Figura Suplementaria 2. Relación entre grado y coeficientes de clustering a distintos niveles de corte de correlación.** En la figura, se representa la relación entre el grado (*k*) y coeficientes de clustering (*CC*) de cada nodo, tomando los distintos niveles de corte de correlación de Spearman indicados en la tabla 1. El gráfico correspondiente a nivel de corte de correlaciones ≥0.7 fue presentado previamente en la figura 6-B. En todos los casos analizados se encontró una relación entre ambas variables, observándose un decaimiento de los *CC* con el grado y agrupamientos de los puntos en bandas.

## BIBLIOGRAFÍA


Aceituno, F. F., N. Moseyko, S. Y. Rhee y R. A. Gutierrez (2008). "The rules of gene expression in plants: Organ identity and gene body methylation are key factors for regulation of gene expression in Arabidopsis thaliana." BMC Genomics **9**(1): 438.

Albert, R., H. Jeong y A. L. Barabasi (2000). "Error and attack tolerance of complex networks." Nature **406**(6794): 378-82.

Albert R., B. A.-L. (2002). "Statistical mechanics of complex networks." Rev Mod Phys. **74**: 47–97.

Alonso, J. M. y J. R. Ecker (2006). "Moving forward in reverse: genetic technologies to enable genome-wide phenomic screens in Arabidopsis." Nat Rev Genet **7**(7): 524-36.

Amaral, L. A., A. Scala, M. Barthelemy y H. E. Stanley (2000). "Classes of small-world networks." Proc Natl Acad Sci U S A **97**(21): 11149-52.

Arioli, T., L. Peng, A. S. Betzner, J. Burn, W. Wittke, W. Herth, C. Camilleri, H. Hofte, J. Plazinski, R. Birch, A. Cork, J. Glover, J. Redmond y R. E. Williamson (1998). "Molecular analysis of cellulose biosynthesis in Arabidopsis." Science **279**(5351): 717-20.

Ashburner, M., 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 y G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-9.

Assaad, F. F., J. L. Qiu, H. Youngs, D. Ehrhardt, L. Zimmerli, M. Kalde, G. Wanner, S. C. Peck, H. Edwards, K. Ramonell, C. R. Somerville y H. Thordal-Christensen (2004). "The PEN1 syntaxin defines a novel cellular compartment upon fungal attack and is required for the timely assembly of papillae." Mol Biol Cell **15**(11): 5118-29.

Barabasi, A. L. y R. Albert (1999). "Emergence of scaling in random networks." Science **286**(5439): 509-12.

Barabasi, A. L. y Z. N. Oltvai (2004). "Network biology: understanding the cell's functional organization." Nat Rev Genet **5**(2): 101-13.

Bergmann, S., J. Ihmels y N. Barkai (2004). "Similarities and differences in genome-wide expression data of six organisms." PLoS Biol **2**(1): E9.

Bonin, C. P., I. Potter, G. F. Vanzin y W. D. Reiter (1997). "The MUR1 gene of Arabidopsis thaliana encodes an isoform of GDP-D-mannose-4,6-dehydratase, catalyzing the first step in the de novo synthesis of GDP-L-fucose." Proc Natl Acad Sci U S A **94**(5): 2085-90.

Bouton, S., E. Leboeuf, G. Mouille, M. T. Leydecker, J. Talbotec, F. Granier, M. Lahaye, H. Hofte y H. N. Truong (2002). "QUASIMODO1 encodes a putative

membrane-bound glycosyltransferase required for normal pectin synthesis and cell adhesion in Arabidopsis." Plant Cell **14**(10): 2577-90.

Boyle, E. I., S. Weng, J. Gollub, H. Jin, D. Botstein, J. M. Cherry y G. Sherlock (2004). "GO::TermFinder--open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes." Bioinformatics **20**(18): 3710-5.

Carpita, N. C. y D. M. Gibeaut (1993). "Structural models of primary cell walls in flowering plants: consistency of molecular structure with the physical properties of the walls during growth." Plant J **3**(1): 1-30.

Clauset, A., C. Moore y M. E. Newman (2008). "Hierarchical structure and the prediction of missing links in networks." Nature **453**(7191): 98-101.

Craigon, D. J., N. James, J. Okyere, J. Higgins, J. Jotham y S. May (2004). "NASCArrays: a repository for microarray data generated by NASC's transcriptomics service." Nucleic Acids Res **32**(Database issue): D575-7.

Davuluri, R. V., H. Sun, S. K. Palaniswamy, N. Matthews, C. Molina, M. Kurtz y E. Grotewold (2003). "AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors." BMC Bioinformatics **4**: 25.

Dormann, P. y C. Benning (1998). "The role of UDP-glucose epimerase in carbohydrate metabolism of Arabidopsis." Plant J **13**(5): 641-52.

Faik, A., N. J. Price, N. V. Raikhel y K. Keegstra (2002). "An Arabidopsis gene encoding an alpha-xylosyltransferase involved in xyloglucan biosynthesis." Proc Natl Acad Sci U S A **99**(11): 7797-802.

Ge, H., Z. Liu, G. M. Church y M. Vidal (2001). "Correlation between transcriptome and interactome mapping data from Saccharomyces cerevisiae." Nat Genet **29**(4): 482-6.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. Yang y J. Zhang (2004). "Bioconductor: open software development for computational biology and bioinformatics." Genome Biol **5**(10): R80.

Girke, T., J. Lauricha, H. Tran, K. Keegstra y N. Raikhel (2004). "The Cell Wall Navigator database. A systems-based approach to organism-unrestricted mining of protein families involved in cell wall metabolism." Plant Physiol **136**(2): 3003-8; discussion 3001.

Gutierrez, R. A., T. L. Stokes, K. Thum, X. Xu, M. Obertello, M. S. Katari, M. Tanurdzic, A. Dean, D. C. Nero, C. R. McClung y G. M. Coruzzi (2008). "Systems approach identifies an organic nitrogen-responsive gene network that is regulated by the master clock control gene CCA1." Proc Natl Acad Sci U S A **105**(12): 4939-44.

Hellmann, H. y M. Estelle (2002). "Plant development: regulation by protein degradation." Science **297**(5582): 793-7.

Horan, K., C. Jang, J. Bailey-Serres, R. Mittler, C. Shelton, J. F. Harper, J. K. Zhu, J. C. Cushman, M. Gollery y T. Girke (2008). "Annotating genes of known and

unknown function by large-scale coexpression analysis." Plant Physiol **147**(1): 41-57.

Imoto, K., R. Yokoyama y K. Nishitani (2005). "Comprehensive approach to genes involved in cell wall modifications in Arabidopsis thaliana." Plant Mol Biol **58**(2): 177-92.

Irizarry, R. A., B. Hobbs, F. Collin, Y. D. Beazer-Barclay, K. J. Antonellis, U. Scherf y T. P. Speed (2003). "Exploration, normalization, and summaries of high density oligonucleotide array probe level data." Biostatistics **4**(2): 249-64.

Jeong, H., S. P. Mason, A. L. Barabasi y Z. N. Oltvai (2001). "Lethality and centrality in protein networks." Nature **411**(6833): 41-2.

Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai y A. L. Barabasi (2000). "The large-scale organization of metabolic networks." Nature **407**(6804): 651-4.

Keegstra, K. y N. Raikhel (2001). "Plant glycosyltransferases." Curr Opin Plant Biol **4**(3): 219-24.

Li, Y., M. G. Rosso, B. Ulker y B. Weisshaar (2006). "Analysis of T-DNA insertion site distribution patterns in Arabidopsis thaliana reveals special features of genes without insertions." Genomics **87**(5): 645-52.

Meinke, D. W., L. K. Meinke, T. C. Showalter, A. M. Schissel, L. A. Mueller y I. Tzafrir (2003). "A sequence-based map of Arabidopsis genes with mutant phenotypes." Plant Physiol **131**(2): 409-18.

Mossa, S., M. Barthelemy, H. Eugene Stanley y L. A. Nunes Amaral (2002). "Truncation of power law behavior in "scale-free" network models due to information filtering." Phys Rev Lett **88**(13): 138701.

Newman, M. E. (2006). "Modularity and community structure in networks." Proc Natl Acad Sci U S A **103**(23): 8577-82.

Obayashi, T., S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta y K. Kinoshita (2008). "COXPRESdb: a database of coexpressed gene networks in mammals." Nucleic Acids Res **36**(Database issue): D77-82.

Obayashi, T., K. Kinoshita, K. Nakai, M. Shibaoka, S. Hayashi, M. Saeki, D. Shibata, K. Saito y H. Ohta (2007). "ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis." Nucleic Acids Res **35**(Database issue): D863-9.

Palaniswamy, S. K., S. James, H. Sun, R. S. Lamb, R. V. Davuluri y E. Grotewold (2006). "AGRIS and AtRegNet. a platform to link cis-regulatory elements and transcription factors into regulatory networks." Plant Physiol **140**(3): 818-29.

Parkinson, H., U. Sarkans, M. Shojatalab, N. Abeygunawardena, S. Contrino, R. Coulson, A. Farne, G. G. Lara, E. Holloway, M. Kapushesky, P. Lilja, G. Mukherjee, A. Oezcimen, T. Rayner, P. Rocca-Serra, A. Sharma, S. Sansone y A. Brazma (2005). "ArrayExpress--a public repository for microarray gene expression data at the EBI." Nucleic Acids Res **33**(Database issue): D553-5.

Perrin, R. M., A. E. DeRocher, M. Bar-Peled, W. Zeng, L. Norambuena, A. Orellana, N. V. Raikhel y K. Keegstra (1999). "Xyloglucan fucosyltransferase, an enzyme involved in plant cell wall biosynthesis." Science **284**(5422): 1976-9.

Perrin, R. M., Z. Jia, T. A. Wagner, M. A. O'Neill, R. Sarria, W. S. York, N. V. Raikhel y K. Keegstra (2003). "Analysis of xyloglucan fucosylation in Arabidopsis." Plant Physiol **132**(2): 768-78.

Persson, S., H. Wei, J. Milne, G. P. Page y C. R. Somerville (2005). "Identification of genes required for cellulose synthesis by regression analysis of public microarray data sets." Proc Natl Acad Sci U S A **102**(24): 8633-8.

Ravasz, E., A. L. Somera, D. A. Mongru, Z. N. Oltvai y A. L. Barabasi (2002). "Hierarchical organization of modularity in metabolic networks." Science **297**(5586): 1551-5.

Reiter, W. D., C. Chapple y C. R. Somerville (1997). "Mutants of Arabidopsis thaliana with altered cell wall polysaccharide composition." Plant J **12**(2): 335-45.

Rhee, S. Y., W. Beavis, T. Z. Berardini, G. Chen, D. Dixon, A. Doyle, M. Garcia-Hernandez, E. Huala, G. Lander, M. Montoya, N. Miller, L. A. Mueller, S. Mundodi, L. Reiser, J. Tacklind, D. C. Weems, Y. Wu, I. Xu, D. Yoo, J. Yoon y P. Zhang (2003). "The Arabidopsis Information Resource (TAIR): a model organism database providing a centralized, curated gateway to Arabidopsis biology, research materials and community." Nucleic Acids Res **31**(1): 224-8.

Scheible, W. R. y M. Pauly (2004). "Glycosyltransferases and cell wall biosynthesis: novel players and insights." Curr Opin Plant Biol **7**(3): 285-95.

Schwechheimer, C. y K. Schwager (2004). "Regulated proteolysis and plant development." Plant Cell Rep **23**(6): 353-64.

Sherlock, G., T. Hernandez-Boussard, A. Kasarskis, G. Binkley, J. C. Matese, S. S. Dwight, M. Kaloper, S. Weng, H. Jin, C. A. Ball, M. B. Eisen, P. T. Spellman, P. O. Brown, D. Botstein y J. M. Cherry (2001). "The Stanford Microarray Database." Nucleic Acids Res **29**(1): 152-5.

Somerville, C., S. Bauer, G. Brininstool, M. Facette, T. Hamann, J. Milne, E. Osborne, A. Paredez, S. Persson, T. Raab, S. Vorwerk y H. Youngs (2004). "Toward a systems approach to understanding plant cell walls." Science **306**(5705): 2206-11.

Srinivasasainagendra, V., G. P. Page, T. Mehta, I. Coulibaly y A. E. Loraine (2008). "CressExpress: a tool for large-scale mining of expression data from Arabidopsis." Plant Physiol **147**(3): 1004-16.

Steinhauser, D., B. Usadel, A. Luedemann, O. Thimm y J. Kopka (2004). "CSB.DB: a comprehensive systems-biology database." Bioinformatics **20**(18): 3647-51.

Swarbreck, D., C. Wilks, P. Lamesch, T. Z. Berardini, M. Garcia-Hernandez, H. Foerster, D. Li, T. Meyer, R. Muller, L. Ploetz, A. Radenbaugh, S. Singh, V. Swing, C. Tissier, P. Zhang y E. Huala (2008). "The Arabidopsis Information Resource (TAIR): gene structure and function annotation." Nucleic Acids Res **36**(Database issue): D1009-14.

Takabayashi, A., N. Ishikawa, T. Obayashi, S. Ishida, J. Obokata, T. Endo y F. Sato (2008). "Three novel subunits of Arabidopsis chloroplastic NAD(P)H dehydrogenase identified by bioinformatic and reverse genetic approaches." Plant J.

Taylor, N. G., W. R. Scheible, S. Cutler, C. R. Somerville y S. R. Turner (1999). "The irregular xylem3 locus of Arabidopsis encodes a cellulose synthase required for secondary cell wall synthesis." Plant Cell **11**(5): 769-80.

Usadel, B., U. Schluter, M. Molhoj, M. Gipmans, R. Verma, J. Kossmann, W. D. Reiter y M. Pauly (2004). "Identification and characterization of a UDP-D-glucuronate 4-epimerase in Arabidopsis." FEBS Lett **569**(1-3): 327-31.

Vanzin, G. F., M. Madson, N. C. Carpita, N. V. Raikhel, K. Keegstra y W. D. Reiter (2002). "The mur2 mutant of Arabidopsis thaliana lacks fucosylated xyloglucan because of a lesion in fucosyltransferase AtFUT1." Proc Natl Acad Sci U S A **99**(5): 3340-5.

Watts, D. J. y S. H. Strogatz (1998). "Collective dynamics of 'small-world' networks." Nature **393**(6684): 440-2.

Wilson, C. L. y C. J. Miller (2005). "Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis." Bioinformatics **21**(18): 3683-5.

Yang, C., Z. Xu, J. Song, K. Conner, G. Vizcay Barrena y Z. A. Wilson (2007). "Arabidopsis MYB26/MALE STERILE35 regulates secondary thickening in the endothecium and is essential for anther dehiscence." Plant Cell **19**(2): 534-48.

Zhong, R., C. Lee, J. Zhou, R. L. McCarthy y Z. H. Ye (2008). "A battery of transcription factors involved in the regulation of secondary cell wall biosynthesis in Arabidopsis." Plant Cell **20**(10): 2763-82.

Zhong, R., E. A. Richardson y Z. H. Ye (2007). "The MYB46 transcription factor is a direct target of SND1 and regulates secondary wall biosynthesis in Arabidopsis." Plant Cell **19**(9): 2776-92.

Zhou, J., C. Lee, R. Zhong y Z. H. Ye (2009). "MYB58 and MYB63 are transcriptional activators of the lignin biosynthetic pathway during secondary cell wall formation in Arabidopsis." Plant Cell **21**(1): 248-66.

Zimmermann, P., M. Hirsch-Hoffmann, L. Hennig y W. Gruissem (2004). "GENEVESTIGATOR. Arabidopsis microarray database and analysis toolbox." Plant Physiol **136**(1): 2621-32.