Explorando la relación entre expresión génica y propiedades topológicas de la red de interacción de *Arabidopsis thaliana*.

Alumna:

Silvia M. Giménez Santamarina

Máster Universitario en Bioinformática y Bioestadística

Tutor UOC: **Brian Jiménez-García**

Tutor externo CSIC: **Guillermo Rodrigo Tárrega**

Fecha de entrega: 26 de diciembre del 2016

A) Creative Commons:

B) GNU Free Documentation License (GNU FDL)

C) Copyright

FICHA DEL TRABAJO FINAL

Título del trabajo:	Explorando la relación entre expresión génica y propiedades topológicas de la red de interacción de *Arabidopsis thaliana*.
Nombre del autor:	Silvia M. Giménez Santamarina
Nombre del consultor:	Brian Jiménez-García
Fecha de entrega (mm/aaaa):	12/2016
Área del Trabajo Final:	TFM “Ad-hoc”. Tema propuesto por el alumno.
Titulación:	*Máster oficial en Bioinformática y Bioestadística.*

Resumen del Trabajo (máximo 250 palabras):

Dentro del campo de la Biología de Sistemas, los modelos de redes de interacción se presentan como una herramienta de gran utilidad para representar sistemas vivos de gran complejidad, ayudando a comprender cómo las moléculas se coordinan e interaccionan para dar respuestas dinámicas adecuadas al entorno que las rodea. La red de interacciones entre proteínas de la planta modelo *Arabidopsis thaliana* L.*,* publicada y validada experimentalmente, presenta las características típicas de las redes jerárquicas, lo que evidencia la existencia de organizaciones internas, como por ejemplo módulos funcionales. El objetivo general de este estudio es relacionar perfiles fenotípicos, en forma de datos transcriptómicos, con la red de interacción de proteínas (PPI) de *A. thaliana*, investigando posibles patrones de relación y organización.

En este estudio se ha determinado la correlación entre perfiles de expresión derivados de perturbaciones transcripcionales de experimentos con micromatrices de RNA en estado estacionario, bajo diversas condiciones de crecimiento, estados de desarrollo, estímulos abióticos y bióticos y una variedad de genotipos mutantes; con la topología que presentan los mismos genes en contexto de la red PPI, como el grado de conexión o coeficiente de agrupamiento. Se han identificado 12 genes que mantienen relaciones de coexpresión y a su vez presentan interacción directa en la red PPI, se han predicho y caracterizado los módulos funcionales que los integra y elementos conectores de gran interés por su papel de relación de los diferentes módulos, actuando como posibles dianas de transmisión de información dentro de la red.

Abstract (in English, 250 words or less):

In the field of Systems Biology, models of interaction networks are a powerful tool to approach living complex systems, also facilitating the understanding of the structure and dynamics of the complex intercellular web of interactions that contribute to the structure and function of a living cell. The protein-protein interaction (PPI) network of *Arabidopsis thaliana* L., previously published and experimentally validated, has the characteristic properties of hierarchical networks. It is shown that PPI networks are dynamically organized into functional modules. The aim of this study is to determine the correlation of transcriptomic profiles with the PPI topology, seeking to demonstrate relational or structural patterns within the network.

Correlation coefficients between expression profiles derived from microarray data evaluated at steady-state RNA expression levels from several growth conditions, developmental stages biotic and abiotic stresses, and a variety of mutant genotypes, and topological properties, such as connectivity degree or clustering coefficient, in PPI context have been evaluated. We identified 12 genes showing relations as coexpression and direct interaction in the network; their functional modules were predicted and characterized. Connector nodes were detected showing an important role as potential targets involved in information transmission through the network.

Palabras clave (entre 4 y 8):

Systems Biology, Computational analysis, Omics data, Biological networks, protein-protein interaction networks, Interactome, *Arabidopsis thaliana*.

Índice

1. Introducción 1
1.1 Contexto y justificación del Trabajo 1
1.2 Objetivos del Trabajo 5
1.3 Enfoque y método seguido 6
1.4 Planificación del Trabajo 7
1.5 Breve sumario de productos obtenidos 10
1.6 Breve descripción de los capítulos de la memoria 11
2. Materiales y métodos 12
2.1 Datos biológicos 12
2.1.1 Datos de transcriptómica 12
2.1.2. El interactoma de *Arabidopsis thaliana* 14
2.2 Softwares para el desarrollo del proyecto 16
2.2.1 R 16
2.2.2 Python y GitHub 16
2.2.3 Cytoscape 18
2.2.4 PANTHER y GeneOntology 18
2.3 Métodos estadísticos 20
3. Resultados y discusión 21
3.1 Análisis del promedio de expresión génica y volatilidad. 21
3.1.1 Descripción y procesamiento de los datos. 21
3.1.2 Propiedades topológicas para los datos de expresión. 23
3.1.3 Propiedades topológicas en relación al coeficiente de volatilidad 25
3.1.4 Discusión 29
3.2 Correlación génica en contexto de la red interactoma (PPI). 31
3.2.1 Descripción y procesamiento de los datos. 31
3.2.2 Discusión. 38
4. Conclusiones 40
4.1 Conclusiones del estudio 40
4.2 Planificación y metodología 41
4.3 Caminos futuros 42

5. Glosario 43
6. Bibliografía 44
7. Anexos 48

Lista de figuras

Figura 1: Planificación. Diagrama de Gantt 9
Figura 2: Resultados del análisis de correlación de la expresión 24
Figura 3: Gráficos de dispersión de los datos 27
Figura 4: Resultados del análisis de correlación de la volatilidad 28
Figura 5: Representación gráfica de módulos funcionales 35
Figura 6: Representación gráfica de módulos funcionales filtrados 36

Lista de tablas

Tabla 1: Genes coexpresados con interacción en la red PPI 31

1. Introducción

1.1 Contexto y justificación del Trabajo

Los mecanismos de regulación que generan respuestas dinámicas por parte de un sistema vivo, es decir, cómo éste responde y se adapta a los diferentes estímulos que recibe del entorno que lo rodea, aún son altamente desconocidos a nivel molecular.

El desafío de la biología molecular en la actualidad es comprender la estructura y la dinámica de la compleja red de interacciones intracelulares que contribuyen en la función y la estructura interna de una célula viva. Los mapas de interacciones génicas y proteicas pueden revelar aspectos físicos y funcionales, aún desconocidos, de un sistema vivo [1-2]. Así pues, la comprensión de los sistemas complejos asciende hacia un enfoque de la Biología de Sistemas, en la cual, en lugar de analizar los elementos del sistema individualmente, intenta integrar todos los elementos que intervienen en el mismo como conjunto, estudiando también la manera en que se relacionan y organizan [3], tanto a nivel de interacciones directas, como mecanismos de regulación de las estructuras internas.

Paralelamente, el rápido desarrollo de las tecnologías aplicadas al estudio de la biología, tales como la secuenciación masiva o técnicas ómicas de alto rendimiento; y la biología computacional, ha supuesto un gran desafío en las ciencias ómicas, ya que la generación masiva de datos supone un problema a la hora de almacenar, analizar e interpretar una gran cantidad de datos en su contexto biológico.

La Biología de Sistemas, aplica numerosas herramientas de diferentes campos de conocimiento. Por ejemplo, la Biología Computacional aplica modelos y métodos matemáticos que permiten diseñar y desarrollar modelos o aproximaciones interactivas que simulan sistemas vivos de gran complejidad, como puede ser una célula. Concretamente, la teoría de redes o teoría de grafos, es una herramienta de gran utilidad en la simulación de las interacciones moleculares de los procesos biológicos [4]. De esta manera, la aplicación de una metodología novedosa a los problemas biológicos permite una comprensión del sistema anteriormente imposible. Las redes biológicas ofrecen una visión detallada del sistema vivo y los elementos que intervienen en él, y cómo estos se relacionan. A nivel molecular se utilizan para representar interacciones entre genes,

proteínas, metabolitos, etc. Las relaciones son representadas a través de gráficos donde miles de nodos están conectados con miles de vértices. Algunas de las redes biológicas que más se están estudiando hoy en día son las redes de interacción proteína-proteína (PPI), las redes de regulación (GRNs) y las redes metabólicas y bioquímicas [5].

De las de redes de proteínas PPI se puede extraer información de cómo las proteínas se coordinan para llevar a cabo funciones y realizar los procesos biológicos que ocurren a nivel intracelular, tales como mecanismos de transducción de señales [6,7], modificaciones post-traduccionales y procesos de desarrollo [8,9]. Ofrecen también, la posibilidad de desarrollar herramientas para la predicción de proteínas con función desconocida hasta el momento, ya que, la función de un gen puede ser definida estudiando sus genes relacionados, a nivel evolutivo, de similitud de secuencias o interacciones moleculares [3].

Sin embargo, las propiedades de las redes biológicas, como robustez y plasticidad, [10-16] se presentan como cuestiones centrales, ya que su comprensión puede generar grandes impactos en el futuro tanto en la medicina, en ciencia básica o cualquier otra rama científica de investigación [3].

Por otro lado, la tecnología de micromatrices permite medir los valores de expresión de miles de genes en unas condiciones determinadas a las que está sometido un organismo. Esta tecnología es muy útil para identificar si los perfiles de expresión difieren de manera significativa entre varios grupos de estudio. Concretamente, las micromatrices de proteínas son útiles para identificar y cuantificar proteínas, péptidos, componentes de bajo peso molecular y oligosacáridos o DNA, además de ser utilizados para estudiar interacciones entre proteínas [17].

Los meta-análisis de las colecciones de datos de micromatrices se pueden utilizar para construir redes, aportando información sobre funciones biológicas e interacciones entre los elementos que constituyen las células, órganos y organismos [18]. Existe una extensa cantidad de estudios realizados con micromatrices. Los resultados de muchos de ellos se almacenan en bases de datos y repositorios.

En el presente Trabajo Final de máster, se ha llevado a cabo un proyecto de investigación donde se estudia la red de interacciones moleculares proteína-proteína (PPI) en relación a datos de transcriptómica a nivel celular de un organismo modelo vegetal, con el fin de profundizar en la organización interna de la red PPI y el modo de regulación de módulos funcionales que desencadenan respuestas dinámicas.

Sobre el organismo modelo *Arabidopsis thaliana*, muy utilizado en estudios de función de proteínas, interacciones y estudios mutacionales [19], existe una gran cantidad de información biológica contenida en bases de datos tales como TAIR [20], AtPID [21], AtPIN [22] o en toda la literatura científica. El análisis de los perfiles transcriptómicos mediante la tecnología micromatrices de oligos de *A. thaliana,* muestra un gran potencial para identificar respuestas dinámicas y los mecanismos de regulación de respuestas.

La red PPI de *A. thaliana* utilizada en el estudio ha sido construida mediante la identificación de alto rendimiento de proteínas en levadura [23, 24], además de la integración de todas las interacciones conocidas y contenidas en la base de datos BioGRID [25]. A partir de estos mismos datos se deducen relaciones de coexpresión entre genes, donde se pueden identificar patrones de expresión. Todo ello complementa la información previa conocida, para entender cómo este organismo ha evolucionado y cómo responde a los diferentes ambientes en los que habita.

El estudio de red consiste en un análisis de la topología de la red interactoma PPI para un organismo modelo vegetal, algunas de las propiedades estudiadas son la centralidad de un cierto nodo, el grado de conectividad y el coeficiente de agrupamiento. A través de métodos probabilísticos y métodos de estimación de correlaciones se relacionan datos de naturaleza ómica con la topología de la red PPI. Los resultados han sido sometidos a análisis estadísticos para comprobar su significancia.

Posteriormente, se contextualizan los datos de los perfiles de expresión génica y coexpresión en la red interactoma, localizando los nodos que representan los genes de interés, con la finalidad de complementar la información, tanto estructural como biológica que ofrece la red. Se desea probar la robustez genética de la red global y de sus estructuras de organización interna, aplicando los recientes conceptos publicados sobre la dinámica de las redes de interacción.

La motivación de este trabajo reside en el interés que despierta la Biología de Sistemas en la alumna. La investigación e innovación en este campo presenta una nueva perspectiva en el estudio de la biología, el cual se centra en la comprensión del sistema vivo integrando todos los elementos que intervienen en él. Es una ciencia multidisciplinar que integra herramientas de diferentes campos del conocimiento. Este hecho, induce al trabajo entre profesionales de diferentes disciplinas, los cuales pueden tener diferentes puntos de vista sobre el estudio. El diálogo entre ellos y la conjunción de las herramientas y enfoques que aportan cada uno conlleva un gran avance en el entendimiento de la vida.

Las aplicaciones que se desarrollan del estudio de redes biológicas, también abarcan una gran diversidad de campos, principalmente en biotecnología y biomedicina. Algunos ejemplos son desde la ciencia básica aportando conocimiento; en la medicina presenta gran utilidad en el desarrollo de terapias génicas o medicina personalizada; en agricultura permite la manipulación de caracteres de interés; en el campo de la virología se desarrollan estrategias antivirales, etc.

He elegido el Instituto de Biología Molecular y Celular de Plantas (IBMCP), centro mixto del Consejo Superior de Investigaciones Científicas (CSIC) y de la Universidad Politécnica de Valencia (UPV) para realizar el TFM por el prestigio de esta institución. En el IBMCP se llevan a cabo numerosas investigaciones en el campo de la Biología Computacional y Biología de Sistemas de plantas y de la interacción planta-patógeno. Por ello es un centro adecuado para continuar con el aprendizaje en estos campos relacionados íntimamente con el máster cursado de la Universitat Oberta de Catalunya.

1.2 Objetivos del Trabajo

El objetivo general del presente Trabajo Final de Máster es el estudio de las bases moleculares que permiten a los organismos complejos tener respuestas dinámicas ante perturbaciones ambientales. En particular, el estudio de la red de interacciones entre las proteínas de una célula de un organismo modelo vegetal, *A. thaliana*, con el fin de entender su organización para dar respuestas dinámicas.

Para alcanzar dicho objetivo se proponen unos objetivos específicos en los que se basan las tareas para el desarrollo del proyecto:

- Desarrollo de scripts propios para el análisis de datos.
- Contextualización e interpretación de datos ómicos en un modelo de la red de interacción de proteínas.
- Caracterización del interactoma de *A. thaliana in silico*.
- Estudio de la robustez genética de la red del interactoma objeto y de las organizaciones internas.
- Estudio de las formas modulares de organización de la red.
- Estudio de los mecanismos de activación de grupos de proteínas que interaccionan entre sí.

1.3 Enfoque y método seguido

El enfoque del estudio se concentra en la realización del análisis *in silico* de datos transcriptómicos, perfiles de expresión génica de un organismo modelo vegetal en diferentes condiciones, en relación a la red de interacciones proteína-proteína (PPI) del mismo organismo. Se han planteado hipótesis de correlación entre parámetros y se han realizado pruebas de significancia con métodos estadísticos no paramétricos adecuados a las características de los datos analizados. En base a los estudios más recientes y revisiones de la divulgación científica, se plantean los siguientes análisis: procesamiento de ficheros computacionalmente, análisis de correlación entre expresión y volatilidad génica con las propiedades topologicas de la red PPI o interactoma de *A. thaliana*, identificación de relaciones de coexpresión en contexto de la red, y análisis modular de genes coexpresados.

El objeto de partida son ficheros extraídos y procesados de diferentes bases de datos: i) dos ficheros que recogen la información de la red de *A. thaliana* hasta la fecha conocida [23-24], a partir de los cuales se representa la estructura de la red y se estudian sus propiedades topológicas; ii) un fichero que contiene datos de perfiles de expresión génica de experimentos de micromatrices de oligos, de *A. thaliana* en diversas condiciones [18]; y iii) datos de coexpresión entre genes del mismo organismo obtenidos a partir del análisis de experimentos de micromatrices [18].

Procesamiento de datos utilizando la línea de comandos de la terminal de Mac del sistema operativo OS X, que utiliza el intérprete de comandos - bash; y scripts creados con lenguaje Python [26]. Análisis estadísticos con R [27]. Se han utilizado Python y R [26, 27] debido a su facilidad en el lenguaje de programación, su versatilidad para abordar múltiples problemas, y la gran disponibilidad de librerías ya creadas. Además de que son los dos softwares principales que se han impartido en el máster de Bioinformática y Bioestadística de la UOC, por lo que el nivel de conocimiento adquirido por la alumna ha permitido su utilización. Utilizando teoría de redes, se desea predecir y simular la estructura de la red de interacción. Para la representación gráfica de la red se utilizará el software Cytoscape (versión 3.4.0) [28].

1.4 Planificación del Trabajo

En primer lugar, los ficheros de partida que contienen toda la información biológica a analizar, son procesados mediante scripts de Python propios y en algunos casos, mediante la línea de comandos de la terminal. Dichos scripts se encuentran disponibles para su consulta en GitHub [29] (https://github.com/silsgs/masterthesis).

A continuación, se realiza un análisis de expresión génica sobre los genes caracterizados en la red interactoma, contrastando sus niveles de expresión con las propiedades topológicas que caracterizan la red PPI utilizada, las cuales son el grado de conectividad de los genes y el promedio de caminos cortos que conectan dos puntos de la red.

Se realizó a continuación un análisis de las relaciones de coexpresión en la red, donde se han localizado aquellos genes de los que se conoce que mantienen una relación de coexpresión y a su vez interaccionan directamente en la red.

Finalmente, el estudio concluirá con un análisis de la modularidad de la red interactoma. En base a las interacciones proteína-proteína, se localizan los genes con mayor y menor expresión génica, los genes que muestran coexpresión e interacción directa en la red, y módulos funcionales, en base a las teorías dinámicas de regulación de las redes de tipo PPI.

Cada fase del estudio finaliza con la elaboración de un corto informe a modo resumen, donde se recoge la metodología utilizada, conclusiones y dificultades encontradas durante su realización.

Las tareas para cada parte del proyecto se pueden resumir en los siguientes puntos:

- Creación de scripts en código Python [26].
- Utilización de comandos en la terminal para el procesamiento de ficheros.
- Importación de datos y análisis estadísticos en R [27].
- Creación de representaciones gráficas tanto de redes de interacción como de los análisis estadísticos realizados.
- Análisis funcional con la herramienta PANTHER [30] y categorías Gene Ontology [31].
- Conclusión de los análisis elaborando un informe resumen de todas las tareas realizadas y los problemas que hayan podido ocurrir, así como las conclusiones extraídas en cada parte del proyecto.

El preludio del Trabajo Final consiste en la elaboración de dos informes a modo resumen de propuesta y planificación del trabajo que se va a realizar.

Durante el desarrollo, se elaboraron 2 informes de seguimiento que constan como las Pruebas de Evaluación Continua 2 y 3, PEC2 y PEC3, las cuales se incluyen como anexo, para la asignatura del Trabajo Final de Máster. En ellas se relata de manera objetiva el desarrollo del trabajo y los avances hasta cada una de las fechas de entrega. Se documentan también los retrasos en el programa o imprevistos; las tareas realizadas y no realizadas; y una crítica sobre el tiempo planificado para cada tarea. De esta manera el tutor realiza una revisión de seguimiento de las competencias específicas y transversales que el alumno debe ir adquiriendo durante la realización del trabajo final.

Finalmente, el trabajo concluye con la presentación de la memoria final y su defensa ante un tribunal de expertos a través de un vídeo.

A continuación, se muestra un esquema de la planificación de las tareas que se han desarrollado en éste trabajo final de master, determinando el tiempo de dedicación con las fechas de inicio y fin. Se incluyen también las tareas de seguimiento propuestas por la Universitat Oberta de Catalunya.

Septiembre 2016:

- ✓ Propuesta TFM “Ad-hoc” y matriculación.
- ✓ Análisis de expresión génica: 15/09/16 – 03/10/2016

Octubre 2016:

- ✓ PEC1 – Plan de Trabajo: 03/10/2016 – 10/10/2016
- ✓ Análisis de coexpresión: 04/10/16 – 20/10/16
- ✓ Inicio del análisis de redes: 04/10/16
- ✓ PEC2 – Desarrollo del trabajo Fase I: Informe del seguimiento del proyecto, especificando tareas previstas realizadas, y no realizadas. 10/10/2016 – 31/10/16

Noviembre 2016:

- ✓ PEC3 – Desarrollo del trabajo Fase II: informe de seguimiento periódico del desarrollo del trabajo realizado hasta el momento. 31/10/16 – 28/11/16
- ✓ Finalización del análisis de redes: 22/11/16

Diciembre 2016 – enero 2017:

- ✓ Memoria del trabajo final: elaboración de la memoria del trabajo final, máximo 90 páginas. 28/11/16 – 26/12/16
- ✓ ENTREGA MEMORIA TRABAJO FINAL: 26/12/2016
- ✓ Presentación del trabajo e informe de autoevaluación: 26/12/2016 – 02/01/2017
- ✓ Defensa pública ante tribunal: 10/01/2017 – 17/01/2017

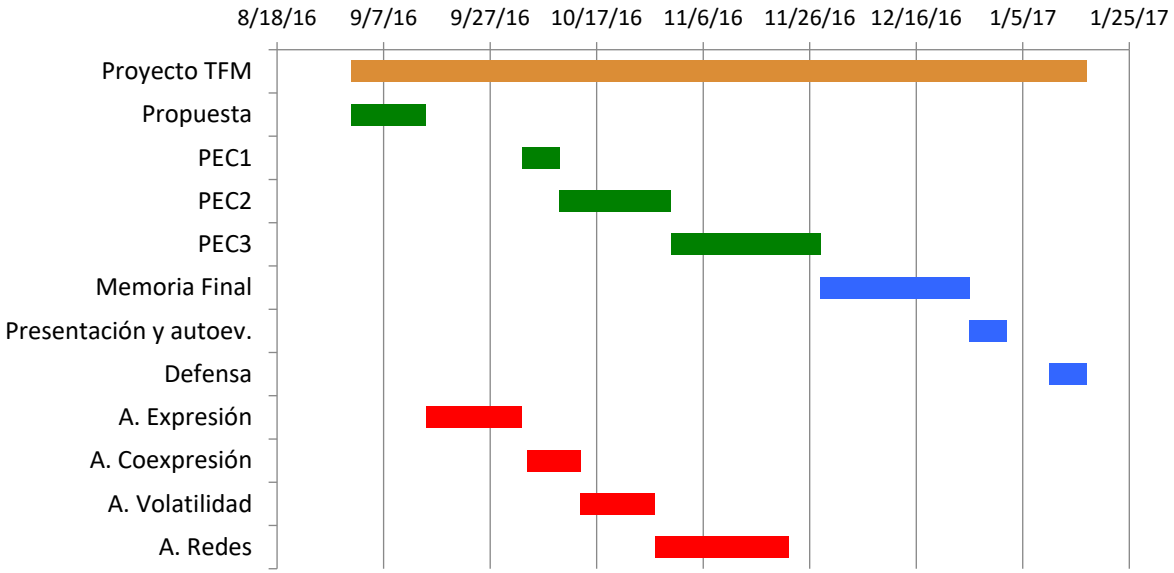

Figura 1: Diagrama de Gantt que ilustra la planificación de las tareas del proyecto relacionado con el tiempo de dedicación previsto para cada tarea.

1.5 Breve sumario de productos obtenidos

Documentos oficiales para la Universitat Oberta de Catalunya:

- Propuesta de Trabajo Final de Máster “Ad-hoc”.
- Pruebas de evaluación continua 1, Planificación del trabajo; 2, Informe de seguimiento Fase I; y 3, Informe de seguimiento Fase II.
- Memoria Final.
- Informe de Autoevaluación.

Resultados del estudio:

- Repositorio online de scripts en Python, en GitHub, (https://github.com/silsgs/masterthesis).
- Scripts en código Python para el procesamiento de ficheros.
- Scripts en RMarkdown para el análisis de resultados.
- Investigación sobre redes biológicas PPI, a publicar en el futuro en revistas especializadas.

1.6 Breve descripción de los capítulos de la memoria

El proyecto se detalla en los siguientes capítulos:

En el primer capítulo Introducción se ha expuesto la problemática que se desea investigar y su contexto, y se han introducido los aspectos más relevantes para facilitar la comprensión del contenido.

En el segundo capítulo, Materiales y Métodos, se detallan los materiales utilizados en los análisis, desde la descripción de los datos de partida, los softwares utilizados, y métodos estadísticos aplicados.

A continuación en el tercer capítulo, Resultados, se documentan en una primera sección los resultados obtenidos del análisis de los perfiles de expresión génica; y en una segunda sección los resultados para el análisis de coexpresión y modularidad de los genes coexpresados. En este capítulo se narra todo el proceso hasta la obtención de los resultados, apoyado por gráficos explicativos y tablas. Estos gráficos están asociados a un pie de figura que revela los detalles más importantes de la misma. Además, una tabla que recoge información biológica acerca de los resultados obtenidos en el análisis de coexpresión. Los resultados son discutidos realizando una crítica objetiva y una comparación con otros estudios publicados.

Finalmente, en el capítulo cuarto, se detallan las conclusiones obtenidas del estudio, una crítica sobre la planificación del proyecto, y una sección de aspectos futuros en relación al estudio realizado.

2. Materiales y métodos

2.1 Datos biológicos

En este apartado se explican los datos de partida del estudio, las variables en las que está contenida la información biológica de interés y su origen.

El proyecto comienza con datos biológicos sobre *A. thaliana* de dos tipos diferentes. Unos datos contienen información sobre perfiles de expresión de *A. thaliana* sometida a diferentes condiciones [18]; y otros que contienen la información para la construcción y análisis topológico de la red PPI de *A. thaliana* [23-25].

2.1.1 Datos de transcriptómica

Los datos que se analizan tiene su origen en los análisis realizados en un estudio anterior [18] en el que se recopilaron datos de perfiles de expresión transcripcional de micromatrices de oligos despositados en la base de datos TAIR (versión 7) [32]. Se calcularon valores promedio de expresión de genes de *A. thaliana* en diferentes condiciones, con la finalidad de inferir la red de regulación transcripcional para dicho organismo.

Los perfiles de expresión de mRNA en estado estacionario derivados de perturbaciones transcripcionales fueron obtenidos de TAIR (versión 7) [32], bajo el patrón de búsqueda “transcriptional factor”. Estos datos contienen perfiles de expresión de *A. thaliana* bajo diversas condiciones experimentales: diferentes condiciones de crecimiento, estados de desarrollo, condiciones en las que se sometió el organismo a estímulos que generan estrés biótico o abiótico y una variedad de genotipos mutantes. Se encontraron 1.187 factores de transcripción en las anotaciones funcionales del genoma de *A. thaliana* disponible en TAIR (versión 7) [32]. El conjunto de datos contiene datos de expresión pre-procesados de 1.436 experimentos de hibridación en los que se utilizaron 22.810 sondas marcadas en Affimetrix’s GeneChip Arabidopsis ATH1 Genome Array [33]. En el estudio de Carrera, *et al.* [18]; se consideraron 22.094 genes. Los arrays se obtuvieron de la casa NASCArrays [34] y

AtGenExpress [35]. La normalización de los datos se realizó utilizando el método robusto del promedio muti-array [36].

Así el presente estudio comienza con tres ficheros que contienen la información biológica en forma de datos de naturaleza transcriptómica, resultado de la metodología explicada. El primero contiene datos de expresión para los genes de *A. thaliana*, y los dos restantes contienen información sobre relaciones de coexpresión entre los genes del mismo organismo. Estos datos se sometieron a una fase de procesamiento que trata las líneas con elementos repetidos, o filas que asocia más de dos genes. Las tareas de procesamiento de ficheros se explican en el próximo apartado.

Los datos de expresión contienen los siguientes valores: i) un valor promedio de los niveles de expresión génica para 23.140 genes de *A. thaliana*; y ii) un segundo valor que corresponde a la desviación estándar del cálculo del promedio, también denominado coeficiente de variación.

A partir de los datos de expresión génica se calculó el coeficiente de volatilidad mediante un script en código Python. El coeficiente de volatilidad equivale al coeficiente de variación de la expresión génica; es decir, corresponde a la relación entre la desviación típica de la expresión y su promedio obtenido de las diferentes condiciones de estudio. A partir de este dato, se puede conocer más detalles sobre los niveles de expresión de los genes, así como si un gen es muy volátil, puede presentar picos de expresión en determinadas situaciones, o de lo contrario, si es poco volátil significa que sus niveles de expresión fluctúan muy poco.

El coeficiente de volatilidad se calculó de la siguiente manera:

$$v_i = \frac{\sigma_i}{\mu_i}$$

Donde v_i es el coeficiente de volatilidad de un gen o coeficiente de variación; σ_i corresponde a la desviación típica y μ_i es el promedio de expresión génica. Los valores obtenidos del coeficiente de volatilidad se almacenan en una nueva variable junto con los datos del promedio y de la desviación en un fichero independiente.

En base a los resultados obtenidos en el estudio de Carrera *et al.* [18], se obtuvieron relaciones de coexpresión entre genes de *A. thaliana.* La coexpresión entre genes se mide correlacionando sus perfiles de expresión en unas condiciones dadas [37], de modo que se entiende que existe coexpresión entre dos genes cuando, al amentar o disminuir los niveles de expresión del primer gen, el segundo gen también ve alterados sus niveles de expresión. Se recogieron en un fichero 722.659 parejas de genes de *A. thaliana* que simbolizan correlación entre ellos.

2.1.2. El interactoma de *Arabidopsis thaliana*

Las técnicas a gran escala y de alto rendimiento pueden detectar proteínas que interactúan dentro de un organismo. Entre ellos, los métodos más utilizados para la detección de interacciones proteína-proteína son los ensayos de tracción ("pull down assays") [38], la purificación mediante afinidad ("tandem affinity purification", TAP) [39], los ensayos de doble híbrido en levaduras ("yeast two-hybrid", Y2H) [40], espectrometría de masas [41], micromatrices [42] y exhibición en bacteriófagos ("phage display") [43, 3]. A partir de estas técnicas se calcularon los datos, y se infirió la red biológica de interacción tipo PPI o interactoma, que refleja las interacciones entre proteínas detectadas experimentalmente.

El interactoma o red PPI de *A. thaliana* utilizada en este estudio ha sido construida mediante la identificación de interacciones binarias entre proteínas de alto rendimiento en levadura [23, 24], además se han integrado las interacciones conocidas y contenidas en la base de datos BioGRID [25]. El interactoma cubre alrededor de 8.000 proteínas, y tiene cerca de 22.000 interacciones no redundantes, todas ellas validadas con pruebas experimentales. El interactoma utilizado en este trabajo se proporciona en los materiales anexos. La red se organiza internamente como una red libre de escalas, indirecta y multi-vértices. Para comprender la organización interna de las redes biológicas, se incluyen las siguientes informaciones generales propias de la teoría de grafos.

La mayoría de las redes biológicas son libres de escalas, lo que significa que la distribución del grado se aproxima a una ley universal que sigue la expresión matemática $P(k) \sim k^{-\gamma}$, donde γ es el exponente del grado de conexiones [1]. Se dice que un gráfico G es indirecto cuando puede definirse como un par (V, E) donde V es un conjunto de vértices que representan los nodos y E es un conjunto de aristas que representan las conexiones entre los nodos sin dirección específica [3]. Una red presenta multi-aristas cuando dos o más aristas acaban en un mismo nodo. Son multi-vértices las redes en las que se pueden enlazar dos elementos a través de más de una conexión, como ocurre en la mayoría de las redes biológicas [3].

La representación de la red PPI de *A. thaliana* se realiza mediante la herramienta Cytoscape. Para ello, es necesario un fichero de entrada que contenga las interacciones directas que se dan en el interactoma a modo de lista de parejas de genes. El formato de éste fichero es sencillo ya que se compone de dos columnas con un primer gen que interacciona con el

segundo. Posteriormente, una vez importado el fichero en Cytoscape, el software interpretará las interacciones.

Cuando se habla de la topología de la red, se hace referencia a las propiedades estructurales o topológicas de la red global y de sus elementos. Estas propiedades ofrecen información de cómo están conectados los nodos de la red, su localización, o información sobre los elementos que los rodean. Algunas de las más destacadas en el estudio de redes biológicas son:

Conectividad o grado de conectividad (*k*): es la propiedad más característica de un nodo, indica cuantas aristas relacionan un nodo con otros. Esta propiedad presenta un único valor para redes indirectas [3].

Coeficiente de agrupación (*C*): en la teoría de redes, *C* corresponde a la medida del grado en que los nodos de un gráfico tienden a agruparse [44]. En muchas redes, si un nodo A está conectado con B, y B conecta con C, es muy probable que A conecte también con C.; esta relación es la que cuantifica el valor del coeficiente de agrupamiento. Los estudios sobre la modularidad de las redes muestran que los módulos funcionales tienden a estar formados por nodos con valores altos de C, es decir que estos nodos conectan entre ellos formando triángulos [44].

Camino más corto y media de longitud de caminos (*l*): la distancia en las redes es medida con la longitud de un camino, lo que proporciona información de cuantos nodos se debe atravesar para llegar de un nodo cualquiera a otro nodo aleatorio de la red [1, 3].

Centralidad intermedia o “betweenness centrality”: es un indicador de la centralidad de un nodo en la red. Es el número de todos los trayectos más cortos de todos los vértices a todos los demás que pasan por ese nodo. Un nodo con alta centralidad tiene una gran influencia en la transmisión de información a través de la red. La centralidad intermedia de un borde, es la suma de la fracción de caminos más cortos de todos los pares que pasan [44].

2.2 Softwares para el desarrollo del proyecto

2.2.1 R

Para el análisis de los datos se utilizó el software libre R [27], a través de la interfaz RStudio [45], para relacionar y analizar estadísticamente las variables objeto de cada parte del estudio, y evaluar las hipótesis que se plantean. El motivo por el que se ha escogido dicho software es por la amplia variedad de modelos estadísticos y técnicas gráficas que proporciona. R es un conjunto integrado de instalaciones de software para la manipulación de datos, cálculo y visualización gráfica [27]. Además, permite crear paquetes de extensión por parte de los usuarios creando nuevas herramientas muy útiles para el análisis de datos. La interfaz RStudio [45], es un entorno de desarrollo integrado de R, que facilita la utilización y comprensión del código [45]. Presenta diferentes áreas dentro de la ventana de trabajo donde se pueden visualizar tablas de datos, las variables definidas por el usuario, consola de comandos, visualización de gráficos, y la herramienta ayuda que imprime el manual de las funciones integradas en R y en los paquetes de extensión cargados.

Dentro de todos los paquetes de extensión que ofrece R, destacamos el paquete “ggplot2”, el cual se ha utilizado para generar los gráficos que aparecen en este estudio. “ggplot2” produce estadísticas o gráficos de datos a través de una gramática basada en el libro “The grammar of graphics” [46], compuesta por un conjunto de componentes independientes que se pueden combinar de muchas formas diferentes [47].

2.2.2 Python y GitHub

Se utilizó el lenguaje de programación Python [26] para el procesamiento de ficheros que contienen los datos biológicos que se han analizado y han sido explicados al inicio de esta sección.

La herramienta Python [26] se presenta como un lenguaje de programación para la integración y desarrollo de softwares. Este lenguaje nace como solución al problema de interoperabilidad de la gran variedad de métodos computacionales aplicados a la biología, para simular, analizar y comprender las complejas propiedades de los sistemas vivos [48]. Este lenguaje de programación pretende facilitar el uso de métodos computacionales, y se ha establecido como uno de los más populares en

la computación científica [49]. Es un lenguaje interpretado, interactivo y orientado a objetos, que ofrece una sintaxis sencilla e intuitiva, basado en códigos de alto rendimiento como Fortran, C o C++ [50]. Provee una gran cantidad de estructuras de almacenamiento de datos y la posibilidad de crear funciones propias [48]. Con funciones pre-programadas que facilitan su comprensión y aplicación. Además, dispone de la posibilidad de habilitar paquetes de extensión que proveen funciones orientadas al tipo de aplicación deseada. En este caso, no se ha utilizado ningún paquete de extensión adicional, ya que, para el procesamiento de ficheros, con las funciones implementadas por defecto, y las estructuras de datos disponibles es suficiente. La creación de nuevos scripts donde el input se declara como una variable general, permite la reproducibilidad de los datos, y la fácil adaptación de los scripts para modificaciones en los análisis. Los scripts son aplicados sobre los ficheros de entrada a través de la terminal del sistema operativo OS X.

Los scripts elaborados a lo largo del desarrollo del proyecto se han depositado públicamente en un repositorio online de libre acceso llamado GitHub. Esta plataforma de desarrollo colaborativo permite alojar proyectos utilizando un sistema de control de versiones.

Este repositorio se creó con la finalidad de una supervisión eficaz de los scripts que se elaboraron para el procesamiento de ficheros. Link de acceso: https://github.com/silsgs/masterthesis

Las principales tareas de procesamiento de ficheros que realizan los scripts son:

Ordenación de los datos que contienen más de un elemento en un mismo campo. Por ejemplo, en la variable ‘id’ que contiene los identificadores de los genes, pueden aparecer más de un gen separados por el carácter ‘;’ en el mismo campo. Esta tarea se ha realizado sobre todos los ficheros de entrada a excepción de los que contienen la información del interactoma de *A. thaliana*.

Funciones de cálculo de coeficientes, en cuanto al coeficiente de volatilidad; y valores promedio para los campos de expresión y desviación típica. Para los genes repetidos en el fichero que contiene los valores promedio de expresión y desviación típica, se tomaron los valores para los genes coincidentes y se calculó la media para ambas variables. Para el mismo fichero, se calculó el coeficiente de volatilidad según la formula anteriormente explicada, para cada entrada o gen del fichero.

Asociación de valores cuyo origen son diferentes fuentes o ficheros de entrada. Los valores de las propiedades topológicas se encuentran contenidos en los ficheros obtenidos de la base de datos TAIR [32], a través de un script, se mapean los genes integrados en el interactoma que

aparecen también en los datos de expresión. En un nuevo fichero se combinan los valores de expresión y volatilidad con los valores de las propiedades topológicas de los mismos genes. Aquellos que no se encuentran en el interactoma se asignan el código ‘n/a’ para las variables de la topología de la red.

Creación de subconjuntos de datos aleatorios. En el análisis de coexpresión, se creó un nuevo fichero con asociaciones de genes aleatorios a partir del conjunto de datos contenido en el fichero de correlación génica entre una pareja de genes.

2.2.3 Cytoscape

Cytoscape es una plataforma de software de código abierto para visualizar redes de cualquier naturaleza o aplicación. Presenta gran utilidad en la inferencia de redes biológicas, para representar interacciones moleculares y/o rutas bioquímicas, además de poder integrar a estas redes anotaciones sobre perfiles de expresión génica y otros datos de interés [28].

Se utilizó este software para las representaciones del interactoma de la red PPI de *A. thaliana*, utilizando los datos extraídos de la base de datos TAIR [32]. Todas las manipulaciones de la red, y selección de nodos se realizan a través de las herramientas que ofrece Cytoscape. Concretamente se utilizó la versión 3.4.0 [28].

2.2.4 PANTHER y GeneOntology

La herramienta PANTHER (Protein ANalysis THrough Evolutionary Relationships) Classification System [30], es un sistema de fácil comprensión que combina herramientas de análisis funcional de genes, de ontología, de rutas metabólicas que permiten analizar datos de secuenciación, proteómicos, o experimentos de perfiles de expresión génica a gran escala, de genomas completos [51].

La versión 11 de PANTHER [30] dispone, a día de hoy, información completa sobre la evolución y la función de genes codificantes de proteínas de 104 genomas completos, entre ellos el de *A. thaliana* (Actualización de noviembre del 2016) [52].

Esta herramienta analiza funcionalmente listas de genes en base a los términos ontológicos anotados en la base de datos Gene Ontology [31, 53]. Gene Ontology es una base de datos que recoge todo el conocimiento

computacional de las funciones de los genes y sus productos. Se organiza en códigos para diversas categorías, procesos biológicos, funciones moleculares y componente celular en el que interviene.

PANTHER [30] se empleó en el análisis funcional de los genes localizados en el interactoma que presentan una correlación de expresión, y se analizaron las listas de los genes involucrados en los módulos analizados en la última parte del estudio.

2.3 Métodos estadísticos

La mayoría de los datos de micromatrices no siguen una distribución normal y con frecuencia hay ruido presente en la muestra [54]. Resulta difícil aplicar un test que elimine el ruido tan bien como si se aplicara un test paramétrico en una distribución normal. Por lo que en estos datos lo más adecuado es aplicar pruebas de tipo no paramétrico. [13]. Se entiende como ruido las perturbaciones implícitas en el método experimental, así como la redundancia de datos.

Teniendo esto en cuenta, y para identificar si existe una relación lineal entre los datos de expresión génica y las propiedades topológicas de genes de *A. thaliana* en contexto de red, se emplearon medidas de correlación. Dos variables se correlacionan positivamente si, al aumentar los valores de la primera variable, los niveles de la segunda aumentan; y la correlación es negativa si, en caso contrario, con el aumento de la primera, la segunda disminuye. Para comprobar si existe correlación entre las variables mencionadas, se utilizó el coeficiente de correlación de Pearson, ya que sirve para comparar matrices numéricas. El comando aplicado en la herramienta de análisis estadístico R es "cor()"[1], especificando en una de sus opciones el test de Pearson.

Debido a la influencia del ruido en los resultados, se realizan pruebas con subgrupos del conjunto global de los datos. Para comprobar la significancia de las comparaciones entre los subgrupos, se aplica la prueba U de Mann-Whitney o test Mann-Whitney-Wilcoxon para comparar dos muestras independientes. Este test requiere que los grupos de datos sean simétricos. Esta es una prueba no paramétrica aplicada para evaluar la significancia entre la correlación de dos variables que en primera instancia se presentan como independientes. Se ha aplicado en R a través del comando "wilcox.test()"[2] del paquete "stats".

Otra función que se ha querido destacar del proceso de análisis es "sample"[3], a través de la cual se crea un bucle que genera un grupo de 150 muestras aleatorias del conjunto de datos que cumple la condición de volatilidad baja-media en la segunda parte del análisis de expresión génica. Más adelante se explican los detalles de su aplicación. Esta función está incluida en las dependencias predeterminadas de R.

[1] Página del manual: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/cor.html
[2] Página del manual: https://stat.ethz.ch/R-manual/R-devel/library/stats/html/wilcox.test.html
[3] Página manual: https://stat.ethz.ch/R-manual/R-devel/library/base/html/sample.html

3. Resultados y discusión

3.1 Análisis del promedio de expresión génica y volatilidad.

El objetivo de los análisis en esta primera parte del proyecto es conocer si existe una correlación positiva o negativa entre los niveles de expresión génica de un conjunto de genes dado, y las características que muestran estos mismos genes en el contexto de la red Interactoma PPI.

El método a seguir son dos análisis *in silico*, cuyo objeto de estudio son datos de transcriptómica que recogen información de los perfiles de expresión génica de un conjunto de genes del organismo modelo *A. thaliana* en diversas condiciones.

En el primer análisis se tratan datos de expresión en forma de valores promedio de expresión génica del conjunto de genes de *A. thaliana*; en el segundo análisis se tratan datos de volatilidad, parámetro entendido como coeficiente de variación de los niveles expresión de los genes en diferentes situaciones.

Dentro de la topología de la red, se destacan entre otras, las propiedades de grado de conectividad (*k*), promedio de caminos cortos (*l*), y el coeficiente de agrupamiento (*C*), debido a que son aquellas propiedades que muestran una cierta relación con los datos biológicos comparados y de las que se obtuvo una relación significante.

3.1.1 Descripción y procesamiento de los datos.

Se realizaron los estudios de correlación de manera independiente para el promedio de expresión y el coeficiente de volatilidad. Sin embargo, la primera fase del análisis es común en cuanto al procesamiento de los ficheros que contienen la información de interés. En ambos análisis se siguió el proceso que a continuación se explica.

En primer lugar, se aplicaron los scripts en código Python diseñados para el procesamiento de los ficheros iniciales.

Para la expresión génica se obtuvo un fichero final que recoge todos los datos biológicos de interés, esto es, un fichero que asocia los genes de *A. thaliana* con diferentes variables, que son: i) el promedio de expresión

génica, ii) la desviación estándar de dicho promedio y iii) el valor de grado de conectividad para cada gen, extraído de los ficheros que contienen la información de las propiedades topológicas del interactoma de *A. thaliana*. Se almacenaron los datos en un fichero separado por tabulaciones que se puede comprender como tabla de contenido.

El mismo ejercicio fue aplicado con un nuevo script, para el valor promedio de caminos cortos. De esta manera se obtuvo un nuevo fichero de texto tabulado que contiene la tabla con los genes de *A. thaliana*, i) el valor promedio de la expresión génica, ii) la desviación estándar del promedio y iii) el valor de los caminos cortos en contexto de la red PPI.

En cuanto a los datos de volatilidad, se aplicaron los scripts correspondientes al procesamiento del mismo fichero que contiene los datos promedio de expresión y desviación típica; y se calculó el coeficiente de volatilidad como se ha explicado en la sección de Materiales y Métodos.

Finalmente se obtuvo un fichero con los genes de *A. thaliana* y diferentes variables para la información biológica: i) valor promedio de expresión, ii) el coeficiente de volatilidad obtenido, iii) el valor de cada gen en cuanto al grado de conectividad en la red PPI, (k); iv) el valor del promedio de los caminos cortos, (l); y v) el valor para la propiedad grado de agrupamiento, (C).

Debido a la adquisición de nuevos conocimientos de programación en Python por parte de la alumna, el diseño de los scripts experimenta una optimización en las diferentes etapas del estudio. Primeramente, se dividen todas las funciones a realizar en scripts sencillos y cortos, y posteriormente acaban siendo scripts un poco más complejos y que realizan varias funciones a la vez.

Cabe destacar, que los datos de expresión génica abarcan el genoma completo del organismo objeto del estudio, mientras que el interactoma cubre únicamente a día de hoy, un 25% de la totalidad del genoma. En las variables correspondientes a la topología, se declaró la cadena de texto ‘n/a’ para aquellos genes no integrados en el interactoma PPI a día de hoy.

Se importaron de los ficheros obtenidos al software R, donde el primer paso fue filtrar aquellos genes que no tenían un valor en las variables de las propiedades topológicas.

3.1.2 Propiedades topológicas para los datos de expresión.

Se desea comprobar si existe correlación significativa entre las propiedades topológicas de grado de conexión y caminos cortos con el promedio de expresión. Después de la exploración previa de los datos, se categorizó la variable del promedio de expresión en genes con alto, medio y bajo nivel de expresión según los valores de los cuartiles de la distribución. Se decidió determinar el valor de corte en los valores de los cuartiles debido a que la media y la mediana de la distribución total tienen un valor muy cercano al del primer y tercer cuartil. Esto quiere decir que el 50% de la distribución presenta un rango de valores de expresión pequeño comparado al 25% de los datos situados en los extremos de la distribución. Por lo tanto, es muy difícil determinar diferencias significantes con los elementos localizados en la categoría media.

Los valores de conectividad para los genes localizados dentro de las categorías de baja y alta expresión se sometieron al test Wilcoxon-Mann-Witney o test de la U, el cual con un p-valor de 1.88e-06 apoya la hipótesis de que puede existir una correlación significativa entre el valor de conectividad y el promedio de la expresión génica.

De la misma manera, se sometieron al mismo test estadístico los datos para el promedio de caminos cortos de los genes con un promedio de expresión alto y bajo acorde al sesgo de los datos anteriormente explicado. Con un p-valor menor a 2.2e-16, se demuestra que existe una correlación significativa entre el valor de caminos cortos y el nivel de expresión en cuanto al promedio de esta.

Esto quiere decir que los genes más conectados presentan mayor promedio del nivel de expresión, a su vez, presentan la propiedad de llegar de manera más rápida a cualquier otro punto de la red; mientras que los genes menos conectados tienen un valor promedio de expresión más bajo, y los caminos cortos para llegar a cualquier otro punto de la red tiene un valor promedio más alto.

Finalmente se localizaron los genes contenidos en las categorías del estudio para alta y baja expresión en la red global del interactoma. Se construyó la red global PPI o interactoma del organismo *A. thaliana* mediante la herramienta Cytoscape. Con la herramienta de selección de nodos, e importando un fichero con la lista de los genes para cada categoría se identificaron estos genes en la red.

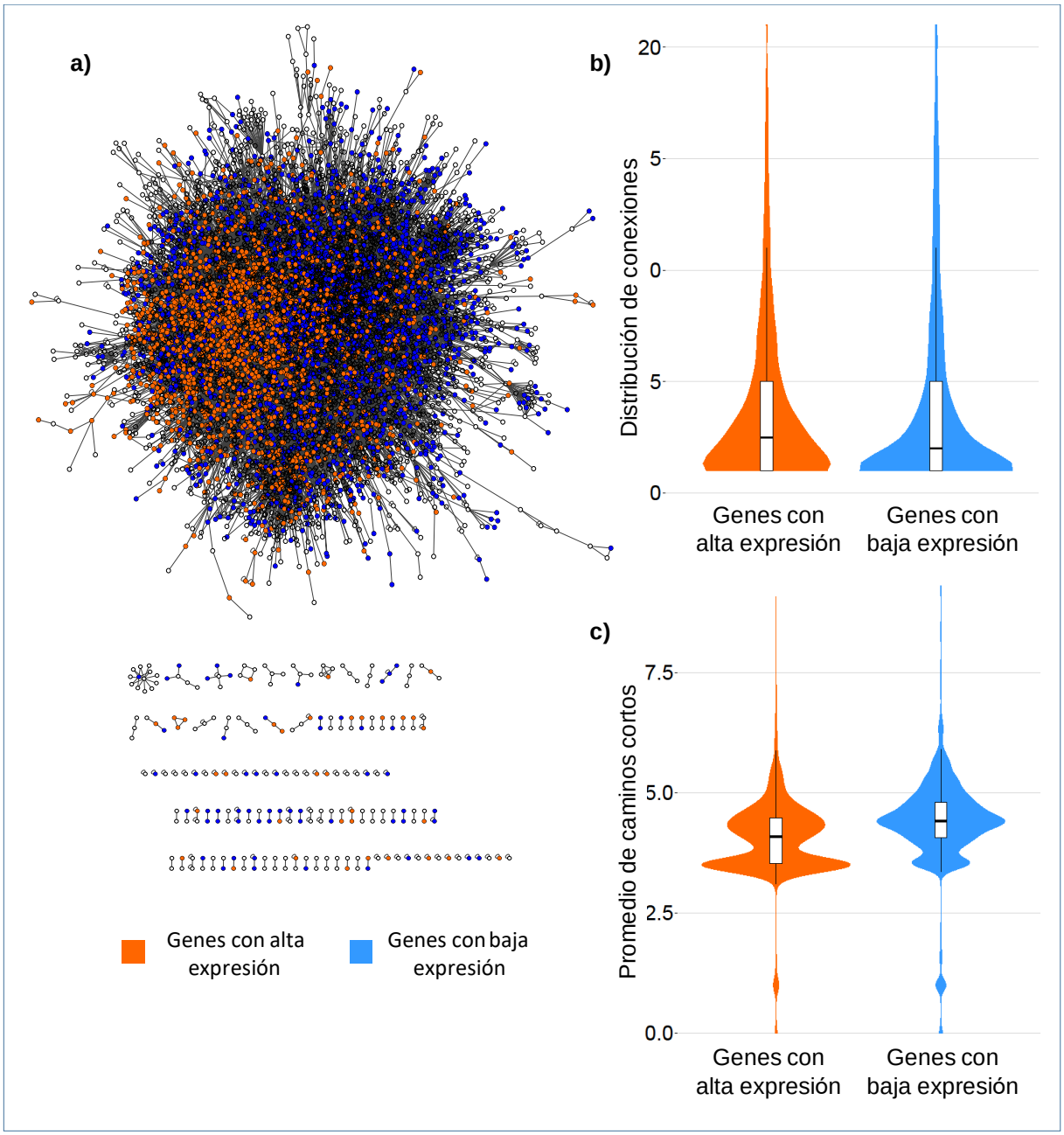

Figura 2: resultados para el análisis de correlación entre el promedio de expresión con las propiedades topológicas, grado de conectividad y promedio de caminos cortos. a) Representación gráfica de la red interactoma PPI global de *A. thaliana* junto con los módulos externos, y localización de los genes para las categorías de genes con un promedio de expresión alto, en naranja; y bajo promedio de expresión, en azul. b) Violin plot combinado con boxplot superpuesto generado para la comparativa de poblaciones con alta y baja expresión para el grado de conectividad; correlación significante con un p = 1.88e-6; c) Violin plot con boxplot superpuesto generado para la comparativa de poblaciones con alta y baja expresión para el promedio de caminos cortos; correlación significante con un p < 2.2e-16.

3.1.3 Propiedades topológicas en relación al coeficiente de volatilidad.

En este apartado se investigan la correlación entre la volatilidad y la topología de los genes en cuanto al grado de conectividad y coeficiente de agrupamiento.

Al igual que en el caso del estudio de la expresión, el análisis comienza con la importación del fichero resultante de la etapa de procesamiento de los datos, a R. Se obtuvo una tabla que relaciona los valores de volatilidad, y en este caso tres propiedades topológicas, que son: grado de conectividad (*k*), promedio de caminos cortos (*l*) y el coeficiente de agrupamiento (*C*).

En la exploración previa de los datos se observó una tendencia lineal en la distribución de los datos para el grado de conectividad y la volatilidad; también se observó una cierta correlación lineal en el gráfico de dispersión para el coeficiente de agrupamiento y la volatilidad.

Se sometieron los datos para ambos contrastes a la prueba de correlación de Pearson. Los resultados fueron los siguientes: i) entre el coeficiente volatilidad y el grado de conectividad la correlación fue r = -0.003709, ii) entre el coeficiente de volatilidad y el promedio de caminos cortos la correlación dio el valor r = 0.007743, iii) y entre el coeficiente de volatilidad y el coeficiente agrupamiento el valor de la correlación fue r = -0.005635. A pesar de que los valores del test de correlación no indican una correlación importante, con motivo de la tendencia lineal que se observa en el gráfico 3c, donde se observa una tendencia de disminución de la volatilidad a medida que el grado de conectividad aumenta, se plantearon diversas estrategias estadísticas para comprobar la existencia de la correlación entre las variables realizando tareas de filtrado de los parámetros.

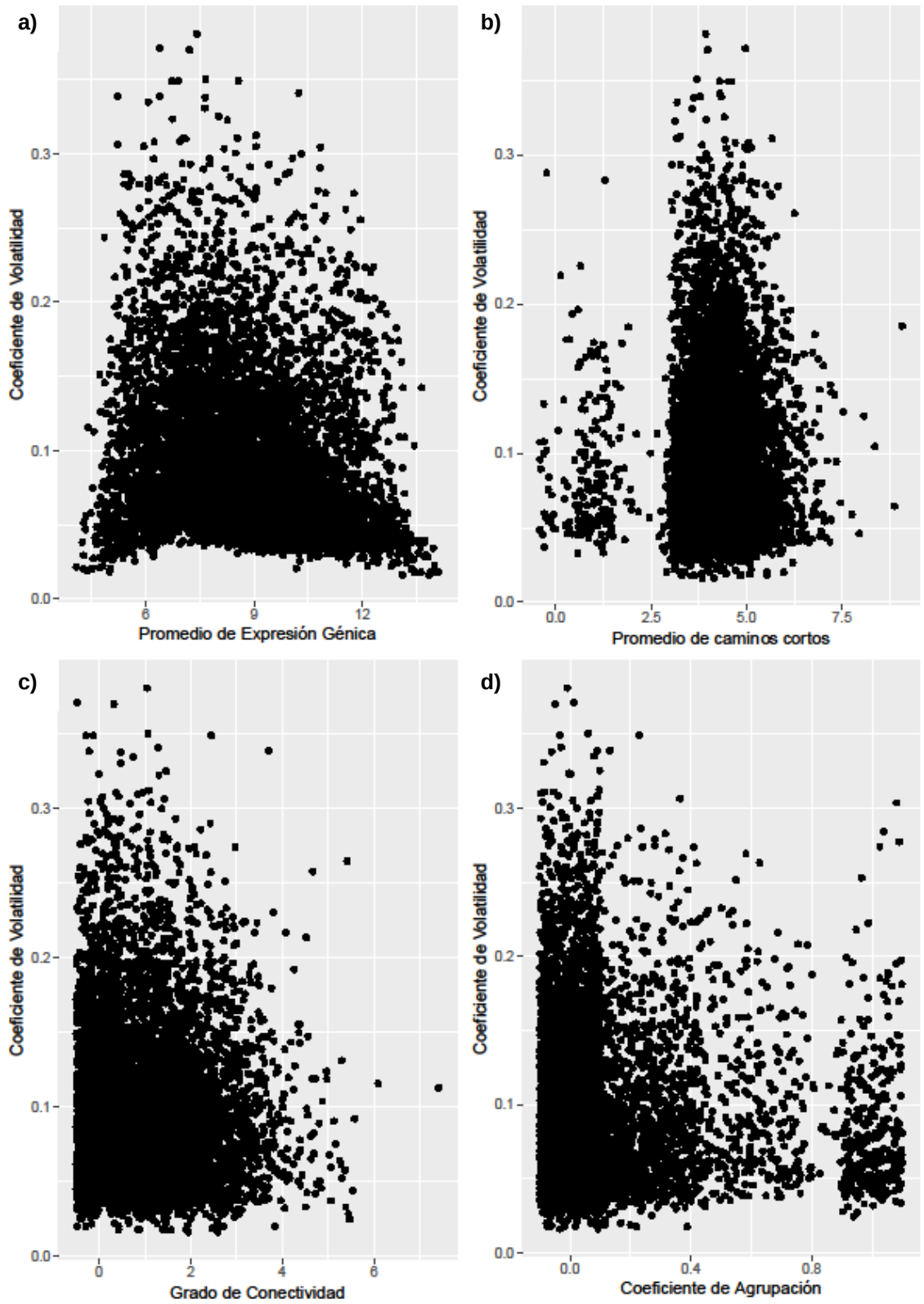

Figura 3: gráficos de dispersión que contrastan los valores para el coeficiente de volatilidad contra: a) grado de conectividad; b) promedio de caminos cortos; c) coeficiente de agrupamiento; y por último d) con la expresión media.

En la figura 3, el primer gráfico de dispersión donde se representan los datos para el coeficiente de volatilidad y el grado de conectividad se observa que existe una cierta relación donde, a mayor conectividad la probabilidad de que el coeficiente de volatilidad tenga un rango amplio disminuye. Así la estrategia que se planteó fue caracterizar los genes altamente conectados, contrastando los datos de volatilidad para los genes poco conectados con los que tienen muchas conexiones. Se categorizó la variable coeficiente de volatilidad, para genes muy volátiles y genes con baja y media volatilidad. De esta última categoría se seleccionaron genes de manera aleatoria creando un subgrupo con un tamaño muestral al de los genes con muy alta volatilidad.

Los grupos para muy alta volatilidad y baja-media volatilidad se sometieron a la prueba de Mann-Witney, que con un p-valor de 0.00053, se acepta que existe una asociación negativa entre la conectividad y la volatilidad. De manera que, los genes más conectados muestran unos niveles de expresión menos fluctuantes.

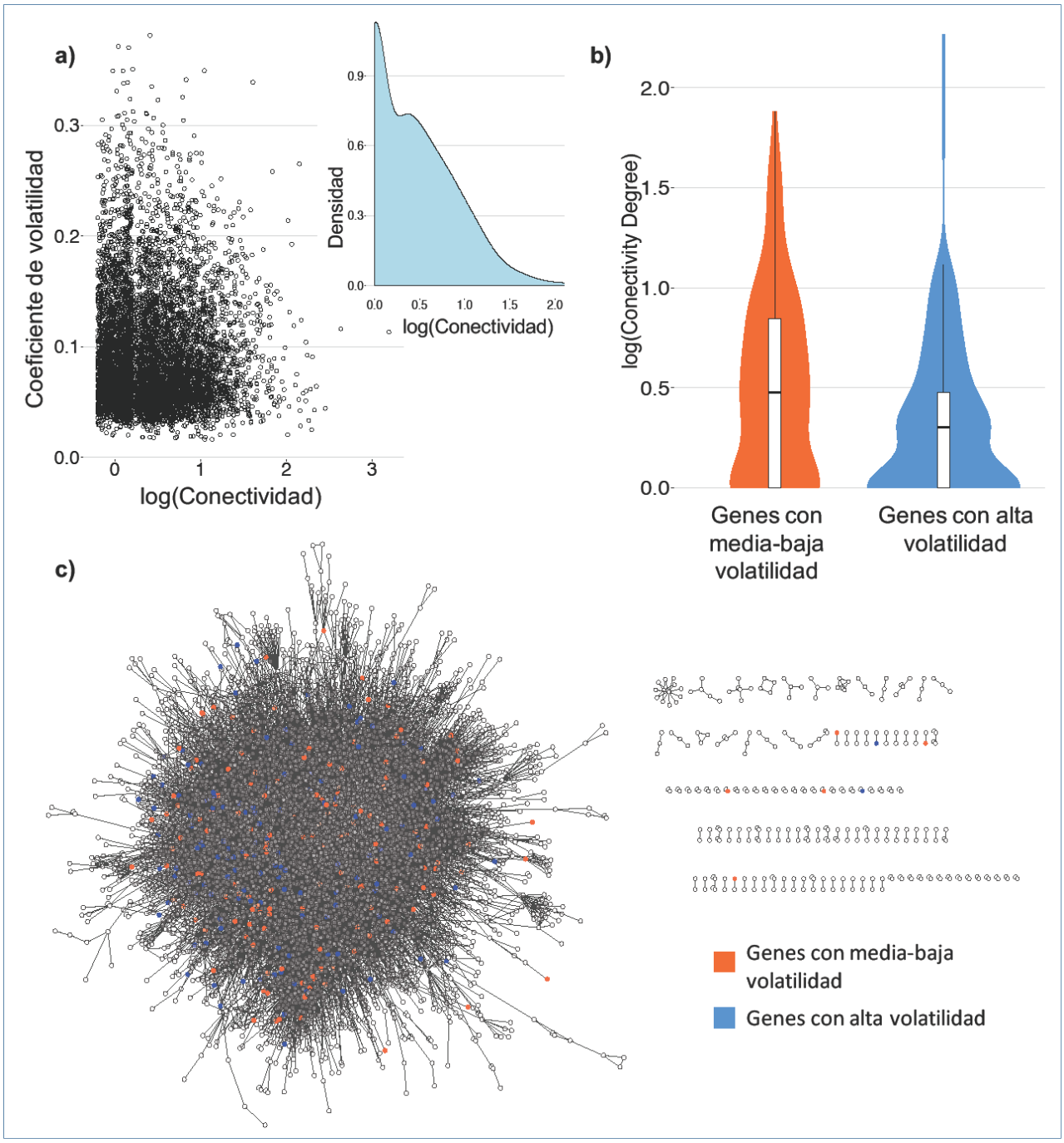

Figura 4: resultados para el análisis de correlación entre el coeficiente de volatilidad con la propiedad topológica grado de conectividad. a) Diagramas de los datos analizados; diagrama de dispersión para los datos de conectividad y volatilidad; curva de densidad de los datos de conectividad, se observa que el 50% de los datos se localizan en valores bajos de conectividad. b) Violin plot combinado con boxplot superpuesto generado para la comparativa de poblaciones de genes con un nivel medio-bajo y alto del coeficiente de volatilidad; correlación significante con un p = 0.00053; c) Localización en la red de los nodos que representan los genes para las categorías del estudio, en azul los genes con media-baja volatilidad, y en naranja aquellos genes con alta volatilidad.

3.1.4 Discusión

Como se ha comentado anteriormente, los datos transcriptómicos analizados en este estudio tienen su origen en un estudio anterior, en el que se realizó el ensamblaje de los datos de micromatrices de oligos depositados en TAIR [32] de los perfiles de transcripción en múltiples condiciones, para calcular un valor promedio de expresión y obtener vectores de correlación entre genes [18]. Sin embargo, la finalidad del estudio de Carrera *et al.* [18] es inferir la red de regulación transcripcional (GRNs) a través de métodos de ingeniería reversa. Es decir, deducir la estructura de la red GRNs desde el producto final que son los niveles de expresión génica en diferentes condiciones, ascendiendo al origen de las regulaciones que dan paso a que los genes se expresen y los organismos respondan dinámicamente. Concluyeron que la red GRNs de *A. thaliana* es libre de escalas y con agrupaciones internas, ambas características son propias de las redes jerárquicas [18].

Por contraste, en este estudio se ha trabajado con la red interactoma PPI, ya publicada y validada experimentalmente en la base de datos TAIR [32], en la cual se han contextualizado datos de perfiles de expresión. Se conoce que la red PPI también es sin escalas, modulada y jerárquica. Se han validado correlaciones entre la expresión génica y la volatilidad con las características topológicas que muestran en la red PPI. La integración de información de redes de interacción de proteínas a gran escala con pantallas fenotípicas podría cambiar fundamentalmente el proceso de descubrimiento y diseño de fármacos. Es posible que tales técnicas puedan generar en el futuro un gran número de nuevos puntos de intervención terapéutica que lleven a nuevos tratamientos de enfermedades [12].

Recordando la filosofía de la Biología de Sistemas de estudiar el organismo como un todo, ambos estudios se complementan para alcanzar el mismo objetivo, que es comprender la dinámica molecular de un sistema vivo. Ambas redes están íntimamente relacionadas, ya que la regulación génica da paso a la síntesis de proteínas. Una de las trayectorias futuras en las investigaciones de redes biológicas, incluyendo las redes metabólicas o bioquímicas, podrían ser los estudios multi-dimensionales donde se relacionen las diferentes redes biológicas como diferentes dimensiones conectadas que generan las respuestas dinámicas de los organismos.

La robustez biológica o genética, es la persistencia de un cierto rasgo o característica en un sistema bajo condiciones de perturbación o incertidumbre.

Por otro lado, en el estudio de Maslov y Sneppen [55]; se compararon dos redes biológicas, de regulación e interacciones de proteínas. Se cuantificaron las correlaciones entre la conectividad de los nodos y se compararon con un modelo de red nula, donde las interacciones fueron aleatoriamente determinadas. Se obtuvo que tanto en las redes PPI como en las GRNs, los nodos altamente conectados son sistemáticamente suprimidos, mientras que aquellos con un nivel de conexión entre alto y bajo se ven favorecidos [55]. Se ha publicado extensamente que las redes biológicas son robustas contra las perturbaciones tales como mutaciones. Por el contrario, también se ha sabido que las redes biológicas son a menudo frágiles frente a mutaciones inesperadas [56]. En el presente estudio se ha obtenido que los genes más conectados también presentan mayores niveles de expresión promedio, pero estos niveles fluctúan poco, ya que la correlación con la volatilidad es negativa. Podría pensarse que esto favorece la robustez de las estructuras internas que conectan.

La robustez facilita la evolución y los rasgos robustos son a menudo seleccionados por la evolución. Este proceso mutuamente beneficioso es posible gracias a características arquitectónicas específicas observadas en sistemas robustos. Pero hay compromisos entre la robustez, la fragilidad, el rendimiento y las demandas de recursos, que explican el comportamiento del sistema, incluyendo los patrones de fracaso. El conocimiento de las propiedades inherentes de los sistemas robustos nos proporcionará una mejor comprensión de las enfermedades complejas y un principio guía para el diseño de la terapia [57].

3.2 Correlación génica en contexto de la red interactoma (PPI).

Los datos de correlación génica de *A. thaliana*, fueron analizados con el objetivo de identificar aquellos genes que ven alterados sus niveles de expresión cuando otro gen aumenta o disminuye sus niveles de expresión, que, además, presenten una interacción directa en contexto de la red interactoma PPI.

Para ello se desarrollaron scripts en código Python que contrasta los ficheros que asocian dos genes en base a la correlación génica por cada línea del fichero, e identifica si estos dos genes aparecen también asociados en el fichero que recoge las interacciones que constituyen el interactoma PPI de *A. thaliana*.

3.2.1 Descripción y procesamiento de los datos.

Después del procesamiento de ficheros con scripts de Python, se obtuvo un nuevo fichero conteniendo los elementos coincidentes en las correlaciones e interacciones. Únicamente 6 parejas de genes cumplen las condiciones de correlación e interacción.

Tabla 1: tabla de los genes coexpresados que a su vez interaccionan directamente dentro de la red PPI de *A. thaliana*, con sus nombres y descripción de la función molecular que llevan a cabo en el organismo, TAIR [32].

	ID	Nombre	Proceso biológico	Función molecular	Conexiones	Coeficiente de agrupación
1	AT1G78380	GST8	cellular response to water deprivation, glutathione metabolic process, oxidation-reduction process, response to cadmium ion, response to	Actividad glutatión transferasa.	6	0.2

			oxidative stress, toxin catabolic process			
	AT2G29490	GST19	glutathione metabolic process, response to toxic substance, toxin catabolic process	Unión a glutatión.	6	0.267
2	AT3G16580		F-box and associated interaction domains-containing protein	Función desconocida.	10	0.022
	AT3G60010	SK13	protein ubiquitination, ubiquitin-dependent protein catabolic process	Actividad ubiquitina transferasa, unión a proteínas.	81	0.003
3	AT1G22300	GF14	brassinosteroid mediated signaling pathway, response to abscisic acid	Unión a ATP, unión a aminoácidos fosforilados.	80	0.023
	AT1G35160	GF14 PHI	brassinosteroid mediated signaling pathway	protein phosphorylated amino acid binding. Unión de dominio específico.	12	0.197
4	AT3G58780	AGL1, SHP1	MAPK cascade, carpel development, plant ovule development, positive regulation of transcription from RNA	Unión a DNA y proteínas. Actividad como factores de transcripción. RNA polymerase II regulatory region	24	0.333

			polymerase II promoter, transcription, DNA-templated	sequence-specific DNA binding.		
	AT4G09960	AGL11, STK	MAPK cascade, carpel development, plant ovule development, positive regulation of transcription from RNA polymerase II promoter, regulation of cell wall organization or biogenesis, regulation of double fertilization forming a zygote and endosperm, regulation of transcription, DNA-templated, seed coat development, seed development, transcription, DNA-templated		13	0.709
5	AT3G13540	MYB5	mucilage biosynthetic process involved in seed coat development, regulation of gene expression, regulation of transcription from RNA polymerase II	Unión a DNA. Factor de transcripción de la RNA polimerasa II.	3	0

			promoter, regulation of transcription, DNA-templated, seed coat development, seed germination, transcription, DNA-templated, trichome differentiation, trichome morphogenesis			
	AT4G09820	ATTT8, TT8	positive regulation of anthocyanin biosynthetic process	Unión a DNA. Factor de transcripción.	14	0.015
6	AT5G46630	AP2M	endocytosis, intracellular protein transport	Unión a lípidos.	161	5.57E-04
	AT4G23460	--	intracellular protein transport, vesicle-mediated transport	Transporte intracelular de proteínas, mediado por vesícula.	4	0.167

Inicialmente se planteó caracterizar topológicamente los genes coexpresados en la red, respondiendo a la cuestión de si pudiera existir alguna correlación positiva o negativa significante entre los parámetros de coexpresión génica y las características topológicas de los nodos que representan estos genes en la red PPI.

Sin embargo, debido a los pocos resultados favorables obtenidos que cumplan ambas condiciones de correlación e interacción, se plantearon nuevos caminos de estudio.

Las redes PPI se organizan internamente en módulos dinámicos. La modularidad de las redes relaciona genes cuyas proteínas interaccionan físicamente entre sí cuando son producidas. Estos módulos se organizan dinámicamente a través de mecanismos de regulación, y se han descrito

en la divulgación científica con una dicotomía clara [58], las interacciones tipo “fiesta”, activación de un conjunto de nodos interaccionados entre sí; y tipo “cita”, constituida por una pareja de nodos activada en un momento dado [58].

Se investiga si esta interacción tipo cita puede dar paso a la conexión entre módulos tipo fiesta a través del estudio de la modularidad de los genes.

Para ello se han identificado los interactores directos de los genes diana, con interacción directa en la red y coexpresión en los datos de transcriptómica; se seleccionaron en Cytoscape sus interactores, mediante selección de vértices y nodos adyacentes. Se obtuvo una representación gráfica de grupos conectados entre sí que a su vez presentan conexión directa entre estos grupos.

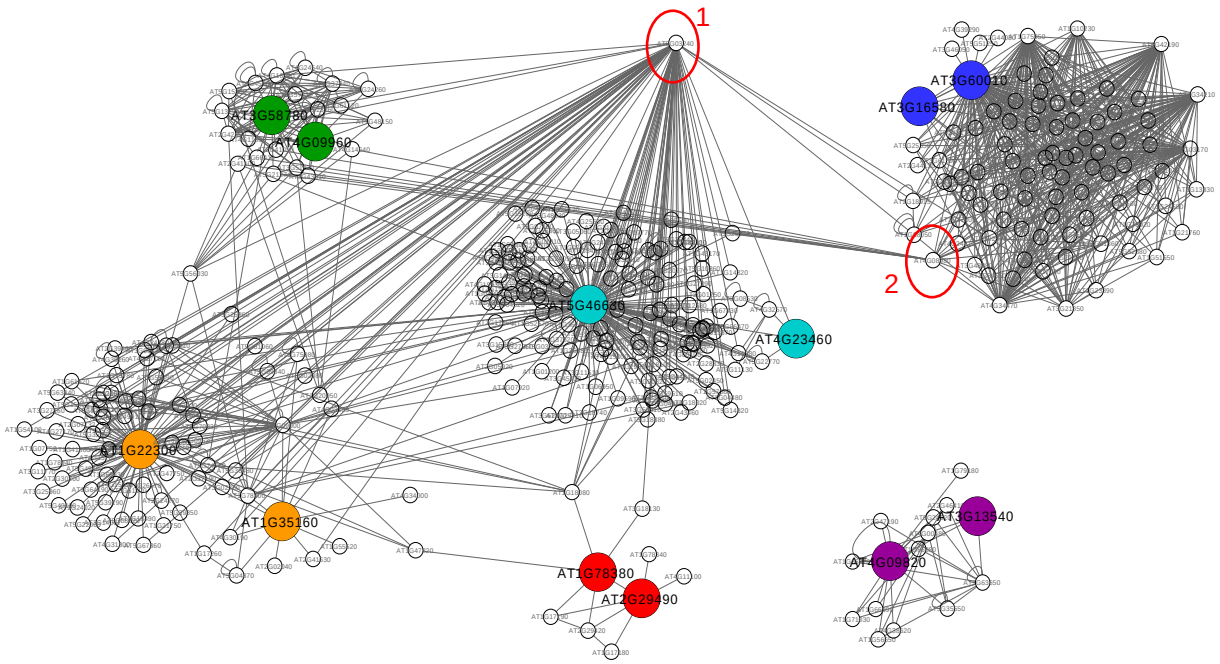

Figura 5: representación gráfica de los módulos formados por los nodos diana, que muestran interacción directa y tienen una relación de coexpresión génica, elaborado en Cytoscape. La primera pareja de genes en la tabla corresponde a aquellos en color rojo, la segunda en azul, la tercera en color naranja, la cuarta en verde, la pareja cinco en color morado, y la última en azul claro. Están representadas todas las interacciones tipo 1 sin ningún parámetro de filtrado. Los nodos remarcados en rojo corresponden a los genes UBQ3 (1) y FBW2 (2).

Se observan módulos claramente definidos, algunos albergan mayor cantidad de nodos, y otros están formados por un número menor. Se observa también la existencia de nodos conectores, que, con un gran número de conexiones relacionan hasta cinco de los seis módulos identificados. Estos nodos conectores podrían tener un papel importante en la activación y desactivación funcional de las estructuras funcionales.

Se decidió realizar un filtraje más específico en cuanto a la selección de nodos o genes, que presenten un valor del coeficiente de agrupamiento alto. Esto significa, nodos que presenten interacciones, que éstas a su vez, interaccionan entre ellas mismas, formando triángulos. Este parámetro es el que sugiere de manera más fiable la existencia de módulos funcionales en base a las interacciones directas entre nodos.

Así, se elaboró un script en Python que extrae los interactores de los genes diana, devuelve los valores de la topología de los mismos, y filtra los interactores por un valor de corte para el parámetro coeficiente de agrupamiento, *C*, y devuelve la lista de estos interactores que cumplen la condición. Estas interacciones son clasificadas como interacciones de tipo 1, es decir, interactúan directamente con los nodos diana.

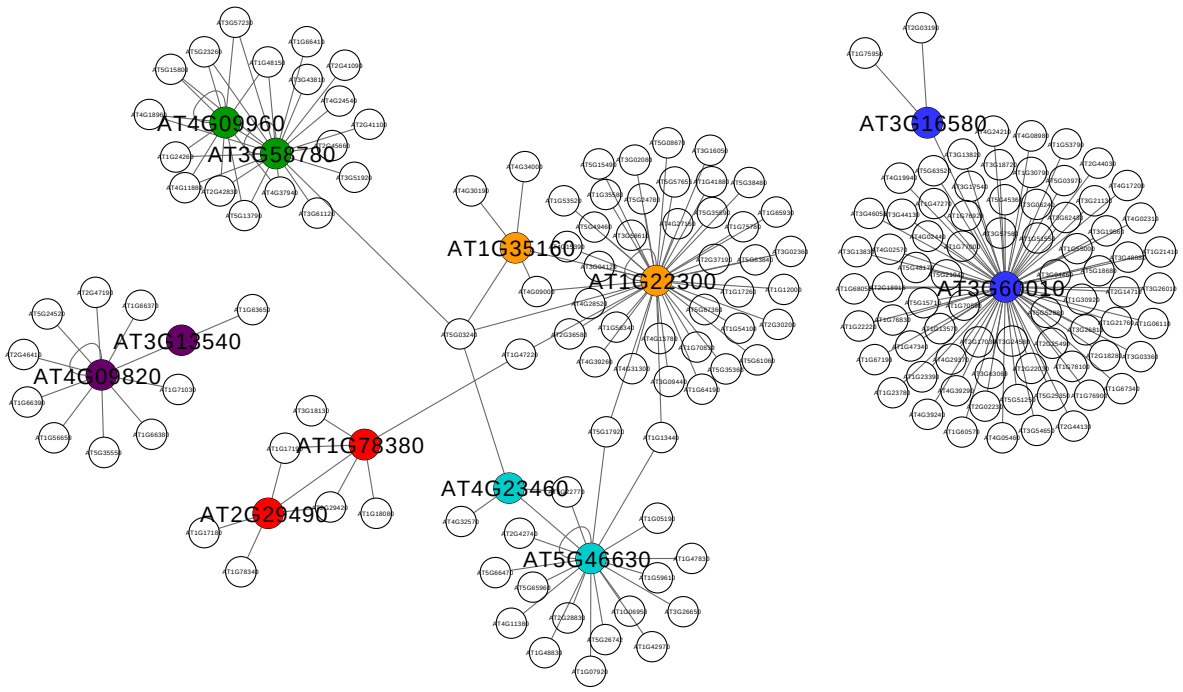

Figura 6: representación gráfica de los nodos diana, genes en los que se ha encontrado una interacción directa dentro de la red PPI, y a su vez muestran coexpresión en los datos de transcriptómica analizados. Las interacciones representadas son de tipo 1.

Comparando las figuras 5 y 6, se observa la existencia de módulos claramente diferenciados. Sin embargo, en la figura 6, tras realizar el filtrado por el coeficiente de agrupamiento, se han perdido numerosos nodos, entre ellos los genes conectores que relacionaban los módulos.

Posteriormente se realizó un análisis funcional utilizando la herramienta PANTHER [30], la cual analiza listas de genes en base a las categorías Gene Ontology [32] de los mismos. El objetivo es conocer si los genes realizan la misma función biológica o al menos relacionada. Esto indica que pueden tratarse de módulos funcionales que actúan conjuntamente.

En el primer módulo, marcados los genes diana en rojo, engloba nueve genes donde predominan las proteínas transferasas de unión a glutatión, las funciones representadas son acción catalítica (GO: 0005488) y de unión (GO:0003824).

El segundo módulo, representando con los genes coexpresados en azul, se presenta como un grupo de 77 proteínas F-Box con función desconocida. Las funciones moleculares representadas son de unión (GO:0005488), actividad catalítica (GO:0003824) y actividad de moléculas estructurales (GO:0005198).

El tercer grupo modular 46 genes, cuyos genes coepresados se presentan en naranja, engloba una gran variedad de tipos de proteínas, donde las funciones moleculares representadas son de unión (GO:0005488), actividad catalítica (GO:0003824), actividad de proteína estructural (GO:0005198), actividad receptora (GO:0004872) y de transporte (GO:0005215).

En el cuarto módulo, donde los genes diana se colorean en verde, es un conjunto que recoge 22 proteínas de tipo agamous y calmodulinas. Las funciones moleculares representadas únicamente son de unión (GO:0005488). Las proteínas de tipo agamous son propias de *A. thaliana*, de unión a DNA que contiene el dominio MADS-box; y las calmodulinas es una familia de proteínas conocidas como uno de los reguladores en la transducción de la señal de calcio en la célula.

El quinto subconjunto está formado por 12 factores de transcripción, de unión a DNA, MYB114, MYB90, TT2, TTG1 entre otros. La función biológica representada es la de unión (GO:0005488), en este caso a DNA.

Finalmente el sexto conjunto está compuesto por 25 proteínas de diversos tipos, incluye tres factores de elongación, dominios U-box, etc. Y las funciones moleculares representadas son de unión (GO:0005488), actividad catalítica (GO:0003824), actividad de proteína estructural (GO:0005198), y actividad de regulación traduccional (GO:0045182).

3.2.2 Discusión.

Se han querido destacar algunos elementos de las representaciones modulares por las características de interconexión modular que muestran algunos elementos.

El gen UBQ3 se ha destacado por la característica de relacionar cinco de los seis módulos analizados, y la cantidad de interacciones que presenta. Este gen codifica para la ubiquitina, la cual se liga a proteínas destinadas a la degradación. Se conoce que sus niveles de expresión son modulados por UV-B y tratamientos de luz/oscuridad.

Al mismo tiempo, otro gen llama la atención por similares razones, es el gen FBW2. Es un gen de tipo F-Box con función molecular desconocida, sin embargo, está catalogado recientemente como un nuevo regulador negativo de los niveles de proteína AGO1, implicada en silenciamiento génico y respuesta antiviral; y puede desempeñar un papel importante en la señalización de ácido abscísico (ABA) y/o respuesta.

Una gran dificultad que obstaculizó los análisis modulares de la red global es que los datos de expresión no contienen medidas de parámetros temporales que relacionen los perfiles de expresión con las relaciones dinámicas según diferentes condiciones. Después de las observaciones realizadas, puede intuirse la funcionalidad de las organizaciones internas o módulos en base al estudio de los genes involucrados como indicadores de las condiciones biológicas.

Se ha publicado un estudio muy recientemente, noviembre 2016, realizado por Yan, *et al.* [59], donde se propone un nuevo método para detectar correlaciones dinámicas como solución a la problemática de detectar módulos y definir relaciones dinámicas entre ellas. Se ha denominado el método como LANDD (Asociación Líquida para la Detección de Dinámica de Red), y se centra en el estudio modular en base a subredes que tienden a incluir genes funcionalmente relacionados, ya que el comportamiento de los genes en una subred es mucho más fiable en comparación al uso de genes únicos como indicadores. [59] Han validado el método utilizando datos de expresión génica real y la red PPI humana.

En el estudio de Chang *et al.* [58] se discute que los análisis sugieren que la organización modular de la red es mucho más compleja que las teorías de la dicotomía de interacciones tipo “fiesta” y “cita”. Por lo que los estudios en las redes de interacción, han de centrarse en caracterizar los parámetros que pueden definir la presencia de módulos funcionales y posteriormente analizarlos. Así como la integración de parámetros

temporales de los perfiles de expresión, puede ayudar en la comprensión de la regulación de las estructuras modulares de la red, identificando bajo qué condiciones concretas se activan los módulos o desactivan, o cómo transmiten la información los nodos conectores de módulos, etc.

4. Conclusiones

4.1 Conclusiones del estudio

- Se ha determinado la correlación positiva entre el promedio de expresión génica y el grado de conectividad en la red, de manera que, los genes más conectados presentan unos niveles mayores en el promedio de expresión génica.
- Se ha determinado la correlación negativa entre el promedio de expresión génica y el promedio de caminos cortos en la red, donde los genes más expresados son capaces de conectar con cualquier otro gen arbitrario de la red de manera más rápida, es decir, atravesando un menor número de nodos en la red.
- Se ha determinado la correlación negativa entre el coeficiente de volatilidad y el grado de conexión. Los genes más conectados presentan unos niveles de expresión menos fluctuantes.
- Se han identificado 6 parejas de genes que tienen una relación de coexpresión entre sí, y a su vez, presentan interacción directa entre los nodos que los representan en la red PPI.
- Se han identificado módulos claramente diferenciados a partir de los genes coexpresados que muestran interacción directa en la red.
- Se ha demostrado mediante el análisis funcional que los genes que componen los módulos, presentan funciones íntimamente relacionadas.
- Finalmente, se han identificado los genes UBQ y FBW2 como elementos conectores de gran interés en la red. Son genes relevantes que conectan diferentes módulos funcionales y presentan un gran número de interacciones que podrían caracterizarse como reguladores de las estructuras modulares internas de la red.

4.2 Planificación y metodología

En líneas generales, se han logrado todos los objetivos propuestos inicialmente en la planificación del estudio. Sin embargo, el análisis modular se planteó en un primer momento con mayor ambición, deseando analizar la red PPI global e identificar módulos funcionales de toda la red. Debido a contingencias imprevistas externas a la alumna, se ha limitado el objeto de estudio a los genes identificados que se coexpresan, relacionándose los resultados obtenidos con las teorías de dicotomía de clasificación de interacciones tipo “cita” y “fiesta”. En un futuro se procederá al análisis modular de la red PPI de *A. thaliana* global.

La planificación se vio alterada poco después de realizar la propuesta de Trabajo Final de Máster, ya que, analizando los datos del promedio de expresión, surgió la idea de analizar la volatilidad en base a la desviación típica del promedio. De esta manera, se integró un nuevo análisis con el que no se contaba en un primer momento, este es el análisis de volatilidad.

4.3 Caminos futuros

En cuanto al desarrollo futuro de esta investigación, se pretende actualizar los datos transcriptómicos analizados incorporando las investigaciones experimentales más recientes. Como se ha explicado, provienen de un estudio anterior realizado en el año 2009 [18], por lo que se prevé que la base de datos TAIR [32] haya sido actualizada y extendida con los experimentos de micromatrices realizados y publicados posteriores al año 2009. Se repetirá de nuevo la metodología aplicada y se compararán los resultados. Se desea investigar en la predicción del módulo formado por las proteínas F-Box.

Se integrará un análisis modular con la herramienta LANDD publicada muy recientemente [58] para la detección de módulos en la red PPI de *A. thaliana,* y en caso de no estar satisfechos con los resultados, se procederá a diseñar e implementar un pipeline computacional para realizar análisis modulares, estudiando profundamente los parámetros que caracterizan los módulos funcionales.

5. Glosario

Ordenado alfabéticamente:

ABA: Ácido abscísico.

AGO1: Proteína argonauta 1.

BioGRID: Database of Protein, Chemical, and Genetic Interactions.

C: Coeficiente de agrupación.

DNA: Ácido desoxirribonucleico.

FBW2: F-BOX WITH WD-40 2

GRNs: Redes de regulación transcripcional

GO: Gene Ontology.

k: Grado de conectividad.

l: Promedio de caminos cortos.

LANDD: Asociación Líquida para la Detección de Dinámica de Red.

mRNA: Ácido ribonucleico mensajero.

p: P-valor.

PANTHER: Protein ANalysis THrough Evolutionary Relationships.

PPI: Interacción proteína-proteína.

r: Coeficiente de correlación.

TAIR: The Arabidopsis Information Resource.

TAP: Tandem Affinity Purification.

UBQ3: Poliubiquina 3.

UV-B: Radiación ultravioleta B onda media.

v_i: Coeficiente de volatilidad.

YAP: Yeast Two-Hybrid.

6. Bibliografía

1. Barabási AL, Oltvar ZN. Network biology: understanding the cell's functional organization. Nature Reviews Genetics, 101-113, 5, 2004.
2. Ideker T, Krogan NJ. Differential network biology. Molecular Systems Biology, 565, 8, 2012.
3. Pavlopoulos GA; Secrier M, Moschopoulos C, Soldatos TG, Kossida J, *et al*. Using graph theory to analyze biological networks. BioData Mining, 10, 4, 2011. doi:10.1186/1756-0381-4-10
4. Pellegrini M, Haynor D, Johnson JM. Protein interaction networks. Expert Review of Proteomics, 1, 2, 2004.
5. Kitano H. Systems Biology: A Brief Overview. Science, 1662-1664, 295, 2002.
6. Liu Y, Zhao H. A computational approach for ordering signal transduction pathway components from genomics and proteomics Data. BMC Bioinformatics, 158-163, 5, 2004.
7. Schulze WX, Deng L, Mann M. Phosphotyrosine interactome of the ErbB-receptor kinase family. Molecular Systems Biology, E4100011–E410001, 1, 2005.
8. Sakakibara H, Takei K, Hirose N. Interactions between nitrogen and cytokinin in the regulation of metabolism and development. Trends Plant Science, 11, 440-448, 2006.
9. Cui J, Li P, Li G, Xu F, Zhao C, *et al*. AtPID: Arabidopsis thaliana protein interactome database—an integrative platform for plant systems biology. Nucleic Acids Research. D999-D1008, 36, 2007.
10. Kitano H. Biological Robustness. Nature Reviews Genetics, 826-837, 5, 2004.
11. Carlson JM; Doyle J. Complexity and robustness. Proceedings of the National Academy of Sciences, 2538-2545, 99 Suppl 1, 2002.
12. Agrawal AA. Phenotypic plasticity in the interactions and evolution of species. Science, 321-326, 294, 2001.
13. Demongeot J, Ben Amor H, Elena A, Gillois P, Noual M, *et al*. Robustness in regulatory interaction networks. A generic approach with applications at different levels: physiologic, metabolic and genetic. International Journal of Molecular Sciences, 4437–4473, 10, 2009.
14. Rizk A, Batt G, Fages F, Soliman S. A General Computational Method for Robustness Analysis with Applications to Synthetic Gene Networks. Bioinformatics, 169-178, 25, 2009.
15. Whitacre JM. Biological robustness: paradigms, mechanisms, and systems principles. Frontiers in Genetics, 67, 3, 2012.

16. Kitano H. Cancer as a robust system: implications for anticancer therapy. Nature Reviews Cancer, 227–235, 4, 2004.
17. Mathur, SK. Statistical Bioinformatics with R. Capítulo 2. Academic Press. London, UK, 2010.
18. Carrera J, Rodrigo R, Jaramillo A, Elena S. Reverse-engineering the Arabidopsis thaliana transcriptional network under changing environmental conditions. Genome Biology, R96, 10, 2009.
19. Somerville C, Koornneef M. A fortunate choice: the history of Arabidopsis as a model plant. Nature Reviews Genetics, 883–889, 3, 2002.
20. Poole RL. The TAIR database. Methods in Molecular Biology, 179–212, 406, 2007.
21. Cui J, Li P, Li G, Xu F, Zhao C, *et al*. AtPID: Arabidopsis thaliana protein interactome database—an integrative platform for plant systems biology. Nucleic acids research, D999-D1008, 36(suppl 1), 2008.
22. Brandão MM, Dantas LL, Silva-Filho MC. AtPIN: Arabidopsis thaliana Protein Interaction Network. BMC Bioinformatics, 454, 10, 2009.
23. Arabidopsis_Interactome_Mapping_Consortium. Evidence for network evolution in an Arabidopsis interactome map. Science, 601-607, 333, 2011.
24. Mukhtar MS, Carvunis AR, Dreze M, Epple P, Steinbrenner J, *et al*. Independently evolved virulence effectors converge onto hubs in a plant immune system network. Science, 596–601, 333. 2011.
25. Chatr-Aryamontri A, Breitkreutz BJ, Oughtred R, Boucher L, Heinicke S, Chen D, *et al*. The BioGRID interaction update database: 2015 updated. Nucleic Acids Res, D470–8, 43, 2015.
26. Python Software Foundation. https://www.python.org/
27. R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.
28. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, *et al*. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome research, 2498-2504, 13(11), 2003.
29. GitHub, Inc. https://github.com/
30. PANTHER Classification System. http://www.pantherdb.org/
31. Gene Ontology Consortium. http://www.geneontology.org/
32. TAIR. http://www.arabidopsis.org/
33. ATH1 Genome Array. http://www.affymetrix.com/products_services/arrays/specific/arab.affx
34. NASCArrays. http://affymetrix.arabidopsis.info/narrays/experimentbrowse.pl
35. AtGenExpress. http://www.arabidopsis.org/info/expression/ATGenExpress.jsp
36. Irizarray RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis K, *et al*. Exploration, normalization, and summaries of high density oligonucleotide array probe level data. Biostatistics, 249-264, 4, 2003.

37. Lynch, M; Walsh, B. Encyclopedia of Systems Biology. 1998.
38. Vikis HG, Guan KL. Glutathione-S-transferase-fusion based assays for studying protein-protein interactions. Methods in Molecular Biology, 175-186, 261, 2004.
39. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, *et al*. The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods, 218-229, 24 (3), 2001.
40. Ito T, Chiba T, Ozawa R, Yoshida M, Hattori M, *et al*. A comprehensive two-hybrid analysis to explore the yeast protein interactome. Proceedings of the National Academy of Sciences, 4569-4574, 98 (8), 2001.
41. Gavin AC, Bosche M, Krause R, Grandi P, Marzioch M, *et al*. Functional organization of the yeast proteome by systematic analysis of protein complexes. Nature, 141-147, 415 (6868), 2002.
42. Stoll D, Templin MF, Bachmann J, Joos TO. Protein microarrays: applications and future challenges. Current Opinion in Drug Discovery and Development, 239-252, 8(2), 2005.
43. Willats WG. Phage display: practicalities and prospects. Plant Molecular Biology, 837-854, 50(6), 2002.
44. Dubitzky, W; Wolkenhauer, O; Yokota, H; Cho, K. H. Encyclopedia of Systems Biology Encyclopedia of systems biology. Springer Publishing Company, Incorporated. 2013.
45. RStudio Team (2015). RStudio: Integrated Development for R. RStudio, Inc., Boston, MA URL. http://www.rstudio.com/
46. Wilkinson, L. The Grammar of Graphics. 2005.
47. Wickham H. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2009.
48. Sanner MF. Python: a programming language for software integration and development. Journal of Molecular Graphics and Modelling, 57-61, 17(1), 1999.
50. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, *et al*. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 2825-2830, 12, 2011.
51. Mi H, Muruganujan A, Casagrande JT, Thomas PD. Large-scale gene function analysis with the PANTHER classification system. Nature Protocols, 1551-1566, 8, 2013.
52. Mi H, Huang X, Muruganujan A, Tang H, Mills C, *et al*. PANTHER version 11: expanded annotation data from Gene Ontology and Reactome pathways, and data analysis tool enhancements. Nucleic Acids Research, doi: 10.1093/nar/gkw1138, 2016.
53. Gene Ontology Consortium. Expansion of the Gene Ontology knowledgebase and resources. Nucleic Acids Research, doi: 10.1093/nar/gkw1108, 2016.
54. Hunter L, Taylor R, Leach SM, Simon R. GEST: a gene expression search tool based on a novel Bayesian similarity metric. Bioinformatics, S115-S122,

17(suppl 1), 2001.
55. Maslov S, Sneppen K. Specificity and Stability in Topology of Protein Networks. Science, 910-913, 296, 2002.
56. Kwon YK, Cho KH. Quantitative analysis of robustness and fragility in biological networks based on feedback dynamics. Bioinformatics, 987-994, 24(7), 2008.
57. Kitano H. Biological robustness. Nature Reviews Genetics, 826-837, 5, 2004.
58. Chang X, Xu T, Li Y, Wang K. Dynamic modular architecture of protein-protein interaction networks beyond the dichotomy of ‘date’and ‘party’hubs. Scientific reports, 1691, 3, 2013.
59. Yan Y, Qiu S, Jin Z, Gong S, Bai Y, *et al*. Detecting subnetwork-level dynamic correlations. Bioinformatics, doi: 10.1093/bioinformatics/btw616, 2016.

7. Anexos

- Scripts de código Python.
- Ficheros filtrados post-procesamiento.
- Listas de genes resultado de análisis funcional con PANTHER sobre los genes que forman módulos funcionales.
- Pruebas de Evaluación Continua 1, 2 y 3.
- Informes tipo resumen de los análisis en formato RMarkdown.