Semantic Mining in Clusters from Signaling Pathways Networks

Rangel, C., and Altamiranda, J.
Centro de Estudios en Microcomputación y Sistemas Distribuidos (CEMISID)
Universidad de Los Andes
Mérida, Venezuela
{carlosran| altamira}@ula.ve

Aguilar, J.
Centro de Estudios en Microcomputación y Sistemas Distribuidos (CEMISID)
Universidad de Los Andes, Mérida, Venezuela.
Prometeo Researcher
Universidad Técnica Particular Loja, Ecuador.
aguilar@ula.ve

***Abstract*— This paper describes how to semantically enrich clusters from signaling pathways networks. The study is divided into two phases, the first is the detection of clusters in signaling pathways networks, after getting these clusters, they are passed to an extraction process of centrality within each one, so the second phase can enrich them semantically. The centrality chosen for the case study is the measure of closeness to other nodes, and it is who is enriched semantically in each cluster. The selected case study is the signaling pathway of TGF-β, and the central nodes found were enriched with the Gene Ontology.**

Keywords—Bioinformatics, clustering, semantic enrichment, TGF-β, Ontology Mining

I. INTRODUCCIÓN

Los conocimientos biológicos han inspirado el surgimiento de proyectos como la Ontología de Genes (GO), que permite realizar anotaciones a decenas de miles de genes de varias especies de otros proyectos o estudios. Esta ontología proporciona un conocimiento considerable, que le permite a los biólogos entender el comportamiento de un gen específico, o el producto génico en un sistema biológico. Estas anotaciones en los genes previstas por el proyecto GO describen la función de un solo gen o de grupos pequeños de genes, pero los biólogos están más interesados en el análisis de grandes listas de genes.

Por otro lado, dentro del área de las redes biológicas es interesante estudiar las interacciones moleculares. Los genes desempeñan sus funciones específicas a partir de sus interacciones temporales, y pueden cambiar de función mediante la interacción con diferentes vecinos [1]. Esto implica que el análisis funcional de la lista de genes, sin tener en cuenta las interacciones, no es óptimo. Por lo tanto, surge la necesidad de anotar funciones teniendo en cuenta al mismo tiempo las moléculas y sus interacciones [2], es decir, para anotar una función biológica se deben considerar las redes biomoleculares o redes biológicas [3,4]. Una red biológica se define como un conjunto de nodos y enlaces. Por lo general, los nodos representan genes o sus productos, y si dos nodos tienen algún tipo de interacción, habrá un enlace entre ellos. Actualmente, muchas redes biológicas han sido ampliamente estudiadas, tales como las redes de interacción de proteínas [5], redes reguladoras de genes [6] y las redes metabólicas [7].

Estudios recientes revelan que las redes biológicas son dinámicas, recableándose para responder a diferentes respuestas externas, con la aparición o desaparición de enlaces en el tiempo. Un ejemplo de la dinámica de una red transcripcional son las redes de regulación de levaduras, y un ejemplo de red dinámica de interacción proteína-proteína son las redes de interacción de proteínas de tejido (ver [7]). Estos ejemplos muestran una misma lista de genes, con diferentes formas de interacciones en distintas condiciones, lo que conlleva a diversos significados o funciones biológicas. El análisis funcional de las redes biológicas, teniendo en cuenta tanto los genes como sus interacciones, supera la capacidad de las herramientas de análisis actual, que consideran sólo los genes individualmente.

En ese sentido, es particularmente deseable determinar *clusters* densos en cuanto a cantidad de nodos. Este problema aparece en el contexto de un gran número de aplicaciones vinculadas a la partición de grafos y al problema de corte mínimo. La determinación de regiones densas en un grafo es un problema crítico desde la perspectiva de diferentes aplicaciones, por ejemplo en las redes sociales y en minería web [16]. Un número importante de técnicas han sido diseñadas en la literatura para la agrupación de grafos densos [17, 18, 19].

Las redes de genes en personas sanas tienen la misma lista de genes, pero las conexiones son diferentes, y por lo tanto, tienen distintos fenotipos. En esta situación, los métodos actuales claramente no pueden decir la diferencia porque la información de enlaces no se considera. Así, hay una gran necesidad por desarrollar nuevos métodos de análisis sobre la función de las redes biológicas, que exploten plenamente la información topológica de la red.

Por otro lado, una red de vías de señales, o *signaling pathway,* es el conjunto de reacciones implicadas en la

978-1-4673-9143-6/15/$31.00

reacción de una célula a un estímulo externo. En ese conjunto de reacciones se pueden detectar subconjuntos, para lo cual se necesita usar una técnica de *clustering*. Los *clusters* no dan mucha información, pero al identificar las funciones biológicas que identifican cada *cluster* se pueden definir familias, diferenciándose cada una del resto.

En este trabajo se propone la detección de grupos de genes, tomando en cuenta la estructura topológica de la red de relaciones dada por estímulos externos, que es conocida como redes de vías de señales, o *signaling pathway* networks. Después de detectar los grupos, el trabajo enriquece los grupos usando GO y técnicas de Minería Ontológica, enriqueciendo no solo un gen con GO, sino un grupo de genes.

El estudio se realizó en la red TGF-β *signaling pathway,* ya que se poseen suficientes datos de alta calidad para explotarlos. TGF-β es una proteína que controla la proliferación celular y la diferenciación, que además está notablemente implicada en la inmunidad y el cáncer. Es interesante realizar el estudio en redes como TGF-β *signaling pathway,* ya que permitirá detectar funciones biológicas propias a la proliferación celular de ciertas células, en específico células cancerígenas.

Este artículo consta de 5 secciones, como primer punto la introducción, en la segunda sección se resume el estado del arte de trabajos relacionados, y la tercera sección presenta las bases teóricas para el entendimiento de la propuesta. En la cuarta sección se presenta nuestra propuesta de Minería Semántica para clusters en Signaling Pathways, seguido en la quinta sección con un caso de estudio. Por último, se presentan algunas conclusiones.

II. Estado del Arte

Algunos trabajos relacionados al área de enriquecimiento semántico de redes de genes son descritos a continuación.

A. *Análisis estadístico y visualización de perfiles funcionales de los genes y grupos de genes*

En últimos años se han diseñado técnicas experimentales de alto rendimiento, como los microarrays, ARN-Seq y espectrometría de masas, que pueden detectar moléculas celulares a nivel de sistemas. Este tipo de análisis genera enormes cantidades de datos, que deben ser objeto de una interpretación biológica. Un enfoque comúnmente utilizado es a través de la agrupación de diferentes genes en base a sus similitudes [20].

Por otro lado, para buscar funciones compartidas (similitud funcional) entre los genes, una forma común es incorporar conocimiento biológico, usando bases de conocimiento como Gene Ontología (GO) y Kyoto Enciclopedia de genes y genomas (KEGG), para la identificación de temas biológicos predominantes en una colección de genes.

Después de la agrupación, los investigadores no sólo quieren determinar si hay un tema común en un grupo de genes, también quieren comparar los temas biológicos entre grupos de genes. Este paso para elegir grupos de interés es manual, seguido del enriquecimiento y análisis de cada conglomerado seleccionado, lo cual normalmente es lento y tedioso. Para llenar este vacío [20] diseñaron clusterProfiler, una herramienta para comparar y visualizar los perfiles funcionales entre grupos de genes.

B. *Comparación de redes de proteínas en el cáncer colorrectal (CRC), bajo un modelo experimental enriquecido semánticamente.*

El objetivo de [21] es el desarrollo de un método que detecte y muestre diferencias entre varias interacciones de proteína-proteína (PPI). El propósito de este método es ayudar a los investigadores en el análisis de las interacciones moleculares que podrían ser comunes o distintas en diferentes manifestaciones de la CRC. Esto podría conducir al descubrimiento de nuevos bio-marcadores predictivos.

El método descrito en [21] integra estas redes en una red principal de proteínas, llamada red de conocimiento. Esta red se monta a partir de un conjunto de bases de datos de proteínas disponibles públicamente, y se enriquece a través de aplicaciones de Minería. Esto se lleva a cabo usando el identificador de proteínas Uniprot, combinándolo con pesos de enlaces utilizando una función de probabilidad de combinación. Posteriormente, las proteínas de ambas redes integradas se clasifican utilizando el análisis de centralidad. Mediante la comparación de las listas resultantes de las proteínas, las regiones de interés que contiene las principales similitudes entre filas de proteínas se encuentran. Estas regiones de interés se clasifican y se visualizan para permitir a los investigadores una fácil orientación y nuevas pistas sobre la mecánica de las enfermedades.

La idea detrás de este procedimiento es que las enfermedades con fenotipos similares son propensos a ser la consecuencia de mutaciones en genes idénticos o funcionalmente relacionados [22]. Encontrar regiones similares en las redes de proteínas de las líneas celulares de CRC, por tanto, podría arrojar algo de luces en los mecanismos moleculares de la enfermedad. Esto podría revelar marcadores potenciales o dianas terapéuticas de la enfermedad. Dado que los trastornos complejos como el cáncer no se pueden describir suficientemente como una lista de genes involucrados, un enfoque basado en red parece prometedor para identificar marcadores potenciales de subredes [23]

Por otro lado entre los recursos para realizar enriquecimiento semántico de redes signaling pathway se pueden mencionar:

NOA

El análisis de la Ontología Genes (GO) se ha convertido en una herramienta popular e importante en el estudio de la bioinformática. Actualmente se lleva a cabo principalmente en el gen individual, o una lista de genes. Sin embargo, análisis recientes a la red molecular revela que la misma lista

de genes con diferentes interacciones puede realizar diferentes funciones [8]. Por lo tanto, es necesario considerar las interacciones moleculares para anotar correctamente y específicamente las redes biológicas. En este caso, se propone un nuevo método de análisis de ontologías de redes (NOA), para llevar a cabo el análisis de ontologías de genes enriquecidos en las redes biológicas. Específicamente, NOA define primero una ontología de enlace que asigna funciones a las interacciones basadas en las anotaciones de los genes conocidos, a través de la optimización de dos índices, "cobertura" y "diversidad" [8]. Entonces, NOA genera dos conjuntos de referencia alternativos para clasificar estadísticamente los términos funcionales enriquecidos para una red biológica dada. Al comparar NOA con los métodos de análisis de enriquecimiento tradicionales en varias redes biológicas, se puede encontrar que: (i) NOA puede capturar el cambio de funciones no sólo en la transcripción dinámica de redes de regulación, sino también en volver a cablear las redes de interacción de proteínas, mientras que los métodos tradicionales no pueden, y (ii) NOA puede encontrar las funciones más relevantes y específicas que los métodos tradicionales de diferentes tipos de redes estáticas. Un servidor web de libre acceso para el NOA se ha desarrollado en http://www.aporc.org/noa/ [8].

MEDLINE

El reconocimiento automático de las relaciones entre un término específico de la enfermedad y sus genes relevantes, o términos de proteínas, es una práctica importante de la bioinformática. Teniendo en cuenta la utilidad de los resultados de este enfoque, se ha identificado el cáncer de próstata y los términos de genes con las etiquetas de identificación de bases de datos públicas biomédicas. Por otra parte, teniendo en cuenta que los expertos en genética usan estos resultados, ellos lo clasificaron basado en seis temas, que pueden ser utilizados para analizar el tipo de cáncer de próstata, los genes y sus relaciones [9].

Los Métodos que se utilizaron son un reconocedor de entidad en base a una entropía máxima, y un reconocedor de relación aplicado a un enfoque basado en el corpus. Se recogen los resúmenes relacionados con el cáncer de próstata a partir de MEDLINE, y se construye un corpus anotado de genes y el cáncer de próstata, con las relaciones basadas en los seis temas. Fue usado para entrenar al reconocedor de entidad mencionado, y para crear la relación máxima basada entropía. Los resultados de este trabajo, en relación al reconocimiento, alcanzaron un 92,1% de precisión para las relaciones (un incremento del 11,0% de la obtenida en un experimento de línea de base). Para todos los temas, la precisión fue de entre 67,6 y 88,1%. En conclusión, [9] reveló que un sistema de reconocimiento cuidadosamente diseñado usando el reconocimiento de entidades, puede mejorar el rendimiento del reconocimiento de las relaciones. En cuanto a la clasificación, el reconocimiento se puede abordar de manera efectiva a través de un enfoque basado en el corpus, mediante una anotación manual y técnicas de aprendizaje automático.

CePa

CePa es un paquete de R con el objetivo de encontrar *pathways* importantes a través de la información de topología de red [10]. El paquete tiene varias ventajas en comparación con las herramientas de enriquecimiento de trayectorias. En primer lugar, el nodo de *pathway* en lugar de definir solo el gen, este es tomado como la unidad básica en el análisis de redes para satisfacer el hecho de que los genes forman parte de sistemas complejos para mantener las funciones normales. En segundo lugar, múltiples centralidades de red se aplican simultáneamente para medir la importancia de los nodos basada en diferentes aspectos, para hacer una vista completa en el sistema biológico. Cepa extiende los métodos de enriquecimiento, para incluir tanto procedimientos de análisis de sobre-representación como de análisis gen-set [10]. CePa se ha evaluado con un alto rendimiento en los datos del mundo real, y se le puede dar más información directamente relacionada con los problemas biológicos actuales. Esta herramienta se encuentra disponible en la red de Archivo R Integral (CRAN): http://cran.r-project.org/web/packages/CePa/

Cytoscape y PSICQUIC

El estudio de la totalidad del interactome (las interacciones proteína-proteína que tienen lugar en una célula) ha experimentado un enorme crecimiento en los últimos años. Representaciones de redes biológicas y sus análisis, se han convertido en una herramienta cotidiana para muchos biólogos y para la bioinformática, ya que los gráficos de interacción nos permiten mapear y caracterizar las vías de señalización y predecir la función de proteínas desconocidas [11]. Sin embargo, dado el tamaño y la complejidad de los conjuntos de datos del interactome, extraer información significativa de las redes de interacción puede ser una tarea desalentadora. Haciendo uso de la herramienta de código abierto Cytoscape, y de otros recursos como PSICQUIC, se puede acceder a varios repositorios de interacción de proteínas al mismo tiempo, el plugin clusterMaker encuentra grupos topológicos dentro de la red resultante, y el plugin bingo realiza el enriquecimiento con GO, de los grupos que se encuentran con clusterMaker [11].

A la luz de los trabajos anteriores, nuestra propuesta se diferencia en que se basa en la estructura del grafo generado del *signaling pathway,* y se analiza usando técnicas de Análisis de Redes Sociales (SNA), específicamente técnicas basadas en la teoría de grafos para la detección de *clusters* (en SNA conocidos como comunidades)), y técnicas de Minería Ontológica para el enriquecimiento de los *clusters.*

III. MARCO TEÓRICO

A. Signaling Pathway

En algunos casos, la activación del receptor provocada por la unión a un receptor de ligando se acopla directamente a la respuesta de la célula al ligando. Por ejemplo, el

neurotransmisor GABA puede activar un receptor de la superficie celular que es parte de un canal iónico. La unión a un receptor GABA A en una neurona GABA abre un canal de ion cloruro selectivo que es parte del receptor. La activación del receptor GABA permite que los iones cloruro negativamente-harged moverse dentro de la neurona, lo que inhibe la capacidad de la neurona para producir potenciales de acción. Sin embargo, para muchos receptores de la superficie celular, las interacciones ligando-receptor no están directamente vinculadas a la respuesta de la célula. El receptor activado debe primero interactuar con otras proteínas dentro de la célula antes de que se produzca el efecto fisiológico final de ligado en el comportamiento de la célula. A menudo, el comportamiento de una cadena de varias proteínas celulares que interactúan se altera después de la activación del receptor. El conjunto de cambios celulares inducidos por la activación del receptor se llama un mecanismo de transducción de señal o vía. [12]

En el caso de la señalización de Notch mediada, el mecanismo de transducción de la señal puede ser relativamente simple. Como se muestra en la Figura 1, la activación de Notch puede causar que la proteína Notch sea alterada por una proteasa. Parte de la proteína Notch se libera de la membrana de la superficie celular y toma parte en la regulación génica. Investigación sobre la señalización celular implica estudiar la dinámica espacial y temporal de ambos receptores y los componentes de las vías de señalización que se activan por los receptores en diversos tipos de células. [12]

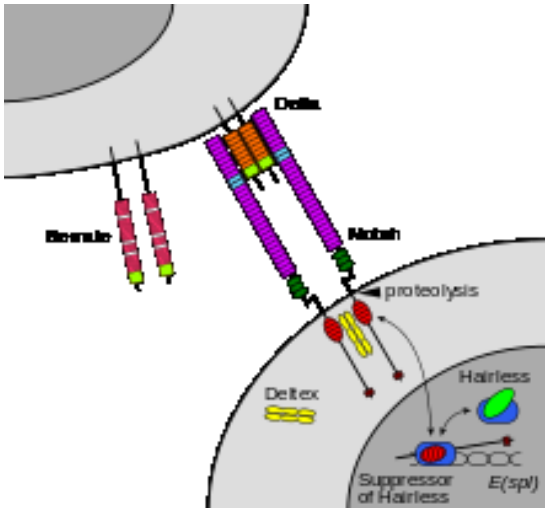

Fig 1. Activación de Notch

Muchos factores de crecimiento se unen a receptores en la superficie celular y estimulan a las células para el progreso a través del ciclo celular y la división. Varios de estos receptores son quinasas y fosforila otras proteínas cuando se une a un ligando. Esta fosforilación puede generar un sitio de unión para una proteína diferente, y por lo tanto, inducir la interacción proteína-proteína. Por ejemplo, una de las vías de transducción de señales que se activan se llama la vía activada por mitógenos de la proteína quinasa (MAPK). El componente de la transducción de señales etiquetadas como "MAPK" en la vía se llamaba originalmente "ERK," por lo que la vía se llama la vía MAPK / ERK. La proteína MAPK es una enzima, una proteína quinasa que puede unir fosfato a proteínas diana, tales como el factor de transcripción MYC, y por tanto, alterar la transcripción de genes, y en última instancia, la progresión del ciclo celular. Muchas proteínas celulares se activan corriente abajo de los receptores de factores de crecimiento (tales como EGFR) que inician esta vía de transducción de señal. [12]

B. Minería Semántica (SM) y Minería Ontológica (OM)

Uno de los desafíos de la Minería de Datos (DM por sus siglas en inglés *Data Mining*) ha sido incorporar conocimiento de un dominio desde los datos.

La minería semántica se encarga de extraer conocimiento semántico desde diferentes fuentes semánticas, como lo son páginas web, contenido sin estructura en la web, contenido estructurado en la web, grafos anotados, ontologías, entre otros. La Minería Semántica se divide en tres grandes grupos, Minería de datos semántica, Minería web semántica y Minería ontológica, este último es el de mayor interés para este trabajo y se describe a continuación.

La extracción de patrones de comportamiento, de conocimiento, entre otras características, usando las técnicas de DM, con la finalidad de construir o enriquecer ontologías, es conocida como Minería Ontológica (OM). Actualmente, con el gran crecimiento en las cantidades de ontologías disponibles, es necesaria el área de OM para explorar técnicas que puedan extraer conocimiento global de un conjunto de ontologías. Algunas de las técnicas que se han venido desarrollando son de enlazado, mezcla, o alineamiento entre varias ontologías.

En particular, en este trabajo nos interesa caracterizar los patrones de agrupamiento dentro de las ontologías, viéndolas como grafos, con el fin de crear un patrón de conocimiento que sea particular a cada grupo. Los algoritmos de minería para grafos son usados para extraer patrones, tendencias, clases y grupos en los grafos. En algunos casos, pueden necesitar ser aplicados a grandes colecciones de grafos. Algunos métodos de minería para grafos se encuentran en [16].

C. Conceptos de la teoria de grafos

A continuación presentamos los concepto de interés en esta área para este trabajo.

Modularidad

La modularidad es una medida de la estructura de las redes o grafos. Fue diseñada para medir la fuerza de la división de una red en módulos (también llamados grupos comunidades). Las redes con alta modularidad tienen conexiones sólidas entre los nodos dentro de los módulos, pero escasas conexiones entre los nodos en diferentes módulos. La modularidad se utiliza a menudo en los métodos de optimización para la detección de la estructura comunitaria en las redes.

Centralidad

En teoría de grafos y análisis de redes sociales la centralidad se refiere a una medida posible de un vértice o nodo en dicho grafo, que determina su importancia relativa dentro de éste [13]. Poder reconocer la centralidad de un nodo puede ayudar a determinar, por ejemplo, el impacto de un gen involucrado en un conjunto de reacciones en una red *signaling pathway*. Algunas métricas de centralidad que podemos mencionar son las siguientes:

La centralidad de grado (degree centrality en inglés) es la primera y más simple de las medidas de centralidad. Corresponde al número de enlaces que posee un nodo con los demás [14]. Esta se puede dividir en centralidad de grado de entrada o centralidad de grado de salida, para grafos dirigidos.

Centralidad de Cercanía (Closeness centrality en inglés), esta medida de cercanía, es la más conocida y utilizada de las medidas radiales de longitud. Se basa en calcular la suma, o el promedio, de las distancias más cortas desde un nodo hacia todos los demás [14].

La intermediación (betweenness centrality en inglés) es una medida que cuantifica la frecuencia o el número de veces que un nodo actúa como un puente a lo largo del camino más corto entre dos nodos [14]. Es de suma importancia al estudiar nodos críticos para la propagación de enfermedades o de opiniones en SNA.

La centralidad de vector propio (eigenvector centrality en inglés) mide la influencia de un nodo en una red, y corresponde al principal vector propio de la matriz de adyacencia del grafo analizado [14].

PageRank es un algoritmo utilizado para asignar de forma numérica la relevancia de los documentos (o páginas web) indexados por un motor de búsqueda, este algoritmo se ha extrapolado al análisis de redes o grafos en general. PageRank es utilizado por Google para ayudar a determinar la importancia o relevancia de una página. Google interpreta un enlace de una página A a una página B como un voto de la página A, para la página B. Los votos emitidos por las páginas consideradas "importantes", es decir con un PageRank elevado, valen más, y ayudan a hacer a otras páginas "importantes".

D. Clustering Jerárquico

En minería de datos, el agrupamiento jerárquico es un método de análisis de grupos el cual busca construir una jerarquía de grupos. Estrategias para agrupamiento jerárquico generalmente caen en dos tipos:

Aglomerativas: Este es un acercamiento ascendente, cada observación comienza en su propio grupo, y los pares de grupos son mezclados mientras uno sube en la jerarquía.

Divisivas: Este es un acercamiento descendente, todas las observaciones comienzan en un grupo, y se realizan divisiones mientras uno baja en la jerarquía.

En los métodos de clustering jerárquico los nodos no se particionan en clusters inmediatamente, primero se realizan particiones sucesivas seguido de la agregación o agrupamiento. El clustering jerárquico produce taxones o clusters de diferentes niveles, estructurados de forma ordenada, estableciendo una jerarquía.

Para poder establecer la clasificación jerárquica se realiza una serie de particiones del conjunto de nodos total:

$$W = \{ i1 , i2 , ...,iN \}$$

Donde i1, iN son los identificadores de los clusters, en un principio cada identificador es asignado a cada uno de los nodos, sucesivamente estos se van agrupando a otros (alglomerativo), hasta el punto que se desee, ya sea una cantidad de nodos por clusters, o una cantidad máxima de clusters.

La representación de la jerarquía de clusters obtenida suele llevarse a cabo por medio de un diagrama en forma de árbol invertido llamado dendograma, en el que las sucesivas fusiones de las ramas a los distintos niveles nos informan de las sucesivas fusiones de los grupos en grupos de superior nivel (mayor tamaño, menor homogeneidad).

Para efectos de este trabajo, utilizaremos métodos aglomerativos. En general, las mezclas y divisiones son determinadas de forma golosa. Los resultados del agrupamiento jerárquico son usualmente presentados en un dendrograma, como se observa en la fig. 2.

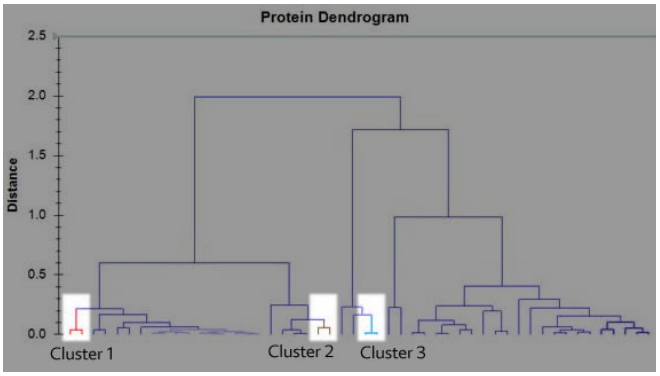

Fig 2. Clustering Jerárquico

El nivel de agrupamiento para cada fusión viene dado por un indicador llamado "valor cofenético", que debe ser proporcional a la distancia o disimilaridad considerada en la fusión (distancia de agrupamiento).

Una vez completamente definida la distancia para nodos, la clasificación jerárquica se puede llevar a cabo mediante el siguiente macro-algoritmo:

El macro algoritmo de clustering Jerárquico

0. Inicio
1. Formar la partición inicial, Considerando cada individuo como un cluster:
   P = { i1},{ i2 },...{ iN }
2. Repetir
   2.1. Determinar los dos clusters más próximos (de menor distancia) ii ,ij , y agruparlos en uno solo.
   2.2. Formar la partición:
   P = { i1},{ i2 },...{ ii u ij },...,{ iN }
3. hasta obtener la partición final Pr= {W}
4. Fin

El marco algoritmo, asigna un cluster a cada individuo o nodo, esto es la partición inicial P (paso 1); seguidamente se van agrupando los nodos que estén más cercanos entre sí, usando técnicas de distancia entre nodos, como la euclidiana en el caso de que se tengan nodos con características numéricas (paso 2.1), en caso contrario se deben usar técnicas otras técnicas de cercanía (como se describe más adelante, la maximización de la modularidad); la nueva partición es formada con los nodos más cercanos entre sí, agrupados en un mismo cluster (paso 2.2); los pasos 2.1 y 2.2 se repiten hasta que las condiciones deseadas se cumplan (número máximo de clusters o número de nodos por cluster) .

IV. PROPUESTA (SEMIC)

A. Aspecto filosófico

Aquí vamos a definir como son usados los conceptos de la teoría de redes en nuestro trabajo:

Modularidad

La modularidad es calculada y optimizada a traves del *Método de Louvain* para la detección de comunidades o clusters, es un método para extraer las comunidades de grandes redes creadas [15].

La modularidad es un valor de escala entre -1 y 1 que mide la densidad de enlaces interiores en las comunidades a los enlaces de las comunidades externas. La optimización de este valor teóricamente resulta en la mejor agrupación posible de los nodos de una red dada, sin embargo, ir a través de todas las posibles iteraciones de los nodos en grupos no es práctico. El método de detección de comunidades de Louvain, empieza primero con pequeñas comunidades, que se encuentran mediante la optimización de la modularidad de forma local en todos los nodos, entonces cada pequeña comunidad se agrupa en un solo nodo, y el primer paso se repite hasta converger.

Centralidad

En este estudio, las medidas de centralidad ayudan a identificar los nodos más significativos dentro de un *cluster*, estos nodos se enriquecerán semánticamente en GO, extrapolando la información semántica de los otros nodos de cada *cluster.* Las centralidades usadas para la detección de estos nodos más significativos son: *Degree Centrality, Closeness Centrality*, *Betweenness Centrality*, y *PageRank.*

B. Macroalgoritmo

A continuación se presenta el macroalgopritmo que permite detectar los *clusters* dentro de una red *signaling pathway*, y enriquecerlos con GO.

El macro algoritmo de la propuesta

0. Inicio
1. Recibir como entrada una ontologia de *signaling pathway*
2. La ontología es llevada a un formato de red (las proteínas serán tratados como nodos y las reacciones como relaciones)
3. Calcular la modularidad para cada nodo en la red
4. Calcular un dendograma, usando la modularidad
5. Realizar el cluster jerárquico, usando el dendograma
6. Calcular los centroides de cada cluster, usando técnicas de centralidad de redes.
7. Enriquecer cada centroide semánticamente con

GO
8. Retornar los clusters con el contenido semántico de sus centroides
9. Fin

El marco algoritmo es descrito a continuación: lo primero a realizar en la propuesta después de recibir la entrada (paso 1) es llevar la red de *signaling pathway,* la cual es recibida en formato OWL (Ontology Web Language), a un formato de red tradicional para poder ser analizada por la herramienta de análisis de redes sociales Gephi (paso 2). Entre los formatos de red que dicha herramienta permite se encuentran: NET, DOT y CSV; seguidamente se pasa al cálculo de la modularidad de todos los nodos (paso 3), esto se hace con la herramienta Gephi, que permite la maximización de la modularidad a través del método de Louvain, y así calcular el dendograma (hecho por la misma herramienta en el proceso de detección de comunidades), tal como se muestra en la figura 2 (paso 4).

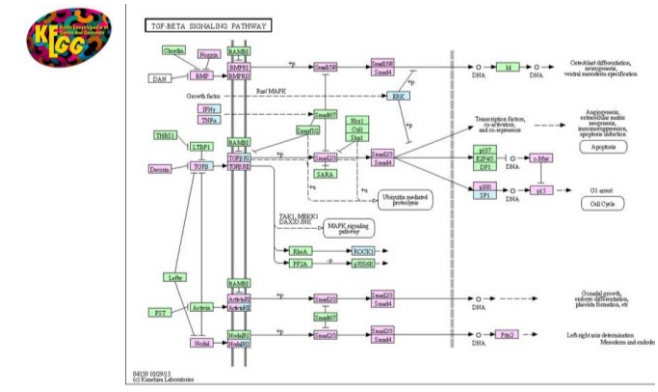

Fig 3. KEGG TGF-β

Los clústeres de cada comunidad serán detectados en el paso 4 (pasos 5) usando la herramienta Gephi. Esto se hace para una red *signaling pathway* como por ejemplo la enciclopedia de Genes y Genomas TGF-β, que es mostrada en la figura 3. Ese paso dará como resultado lo observado en la figura 4, donde hipotéticamente se encuentran los clusters representados por los nodos que se encuentran encerrados en cada circunferencia (esto es de manera ilustrativa, lo que la propuesta logra hacer en la detección de clusters con el método de detección de comunidades de SNA), cada uno de estos grupos de genes pasa al siguiente paso de detección de nodos centrales.

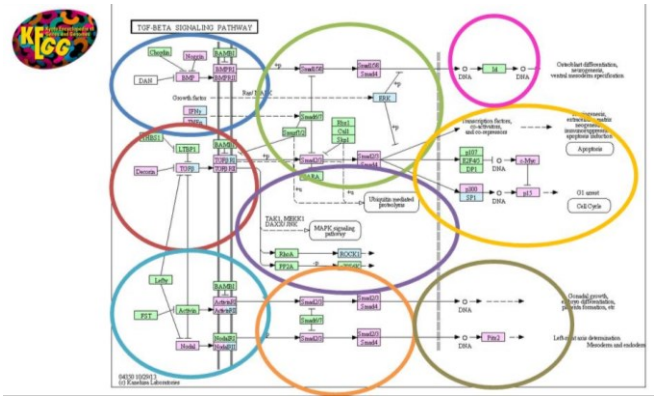

Fig 4. KEGG. Clusters example in TGF-β

A continuación se extraen los centroides (paso 6) de cada cluster. Para este caso los individuos dentro del clúster no poseen características numéricas (recordando que la red se está tomando como nodos y las reacciones entre ellos), razón por la cual se utiliza clustering jerárquico. Los centroides se tomarán para efectos de este estudio como equivalentes a los nodos más centrales, tomando en cuenta las medidas de centralidad: *Degree Centrality, Closeness Centrality*, *Betweenness Centrality*, y *PageRank*. Esto es, cada cluster tendrá no sólo un centroide, sino que dicho centroide será representado por las características semánticas provenientes de enriquecer los nodos altamente centrales de cada cluster.

Seguidamente, al tener los nodos centrales, estos pasan a un enriquecimiento semántico (paso 7), esto se realiza usando la base de conocimiento Gene Ontology (GO). En la figura 5 se ilustra este proceso con los clusters detectados en la figura 4. Para ello se usa una herramienta para extraer conocimiento de GO. La herramienta usada está dentro de "AmiGO 2", que es un proyecto de GO, el cual es un sistema en la web oficial de GO para buscar y navegar por la base de datos de la ontología de genes. PANTHER es una herramientas dentro de AmiGO para el análisis de proteínas a través de relaciones evolutivos (Protein Analysis Through Evolutionary Relationships). El sistema de clasificación usa una gran base de datos biológica de las familias de genes/proteínas y sus subfamilias funcionalmente relacionados, que se pueden utilizar para clasificar e identificar la función de los productos génicos. Las proteínas son clasificadas de acuerdo con la familia (y subfamilia), la función molecular, y su proceso biológico. Esta herramienta (PANTHER) recibe el identificador de un gen, y devuelve el contenido semántico de dicho gen. Para este trabajo, los identificadores que se pasan a enriquecer en PANTHER son los de los nodos altamente centrales de cada cluster, y la información semántica devuelta será extrapolada al cluster.

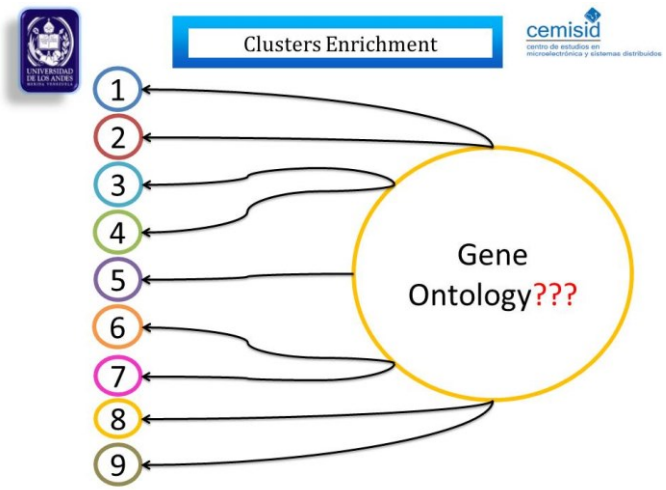

Fig 5. Clusters central nodes

V. CASO DE ESTUDIO TGF-B

A continuación se presenta en detalle el experimento con el *signaling pathway* TGF-β.

Como se ha mencionado anteriormente, se escogió TGF-β porque es una proteína relacionada que controla la proliferación celular implicada en el cáncer. Este estudio permitirá a biólogos detectar funciones biológicas propias a la proliferación celular cancerígena. La red usada se muestra en la fig 6, ya llevada al formato de red que admite Gephi. Dicha red posee 1534 nodos o genes, y 3029 relaciones o reacciones entre ellos.

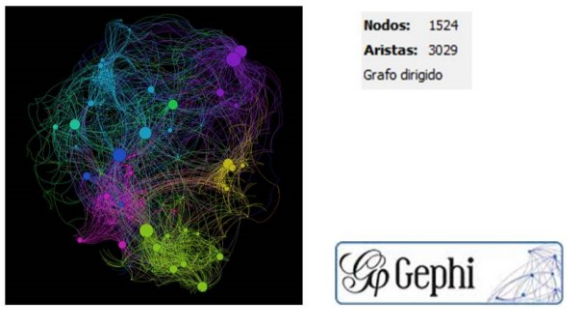

Fig 6. Gephi Clusters

Al ejecutar los algoritmos de cálculo de Modularidad y optimización de la misma en la herramienta Gephi, se detectaron 16 comunidades, que para este trabajo son 16 clusters de genes.

Al realizar el cálculo de centralidades para todos genes de la red, la centralidad de cercanía (*Closeness Centrality*), resultó altamente interesante, ya que para este tipo de redes se va a detectar que nodo crítico en la red, que pueda estar causando una enfermedad, en este caso cancerígenas. Estos nodos con centralidad de cercanía alta, propagaran más rápido enfermedades, o son los causantes de desencadenar un conjunto de reacciones que llevan a estas enfermedades; los nodos más centrales se muestran en la figura 7, estos para visualizarlos mejor se muestran con un mayor tamaño a los que tienen menor centralidad de cercanía (el tamaño es proporcional a la medida de centralidad).

Para mejor visualización de los nodos centrales, y mayor entendimiento para los biólogos, se realizan filtros, y así sólo mostrar lo más interesante; la figura 8 muestra la misma red de la figura 7, pero usando un filtro que sólo permite nodos altamente centrales, dichos nodos son genes potencialmente críticos en el desarrollo de enfermedades cancerígenas.

Agregando un segundo filtro que sólo permita genes con un alto grado de entrada y otro filtro que permita sólo genes con alto grado de salida, la red resultante es ilustrada en la figura 9, estos genes ya vienen del primer filtro de alta centralidad de cercanía, al agregar este nuevo filtro de alto grado de entrada y salida, se ilustran los genes que propagaría más rápidamente una enfermedad, ya que vienen de ser nodos críticos que están más cerca de los otros genes en la red, y además poseen la mayor cantidad de reacciones hacia otros nodos, es decir, propagarían más rápido la enfermedad en cuestión.

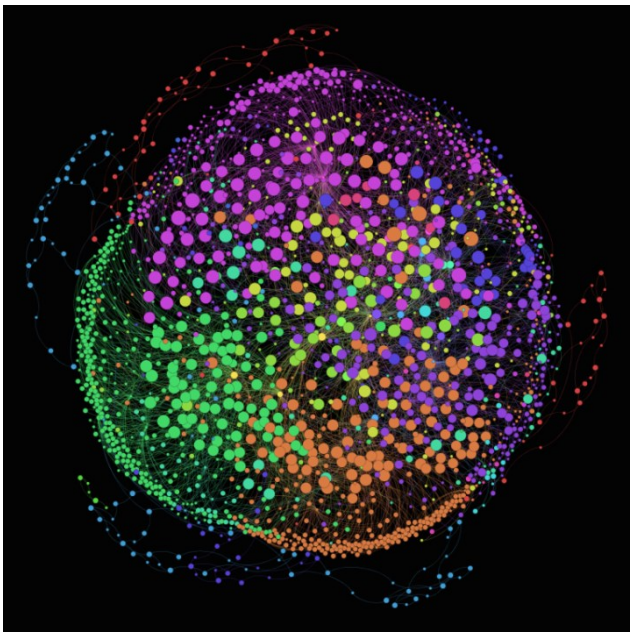

Fig 7. Vista de la red con nodos centrales agrandados

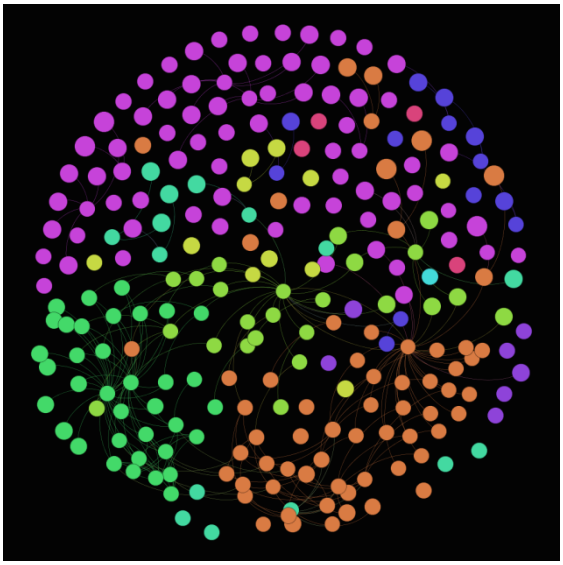

Fig 8. Vista más cercana de la red, usando un filtro para *Closeness Centrality*

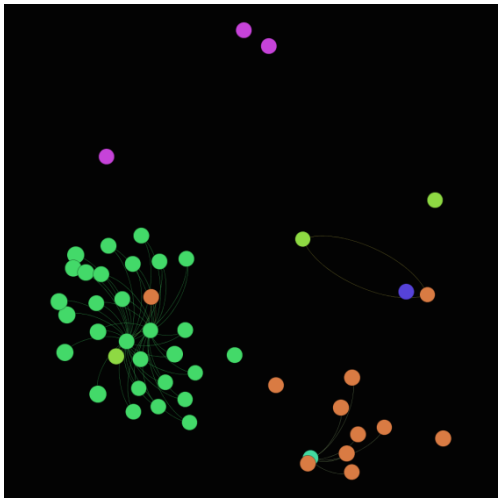

Fig 9. Vista más cercana de la red, usando dos filtros nuevos de alto grado de entrada y grado de salida

En la tabla I vemos la salida que nos proporciona la herramienta gephi, donde Label es el identificador del gen, Grado el grado de entrada y salida del nodo o gen en la red, *Closeness Centrality* el valor de centralidad de cercanía de los genes, Modularity Class es el número de la clase que se le da en la detección de comunidades a cada comunidad (número para identificar el cluster).

La tabla I es una versión reducida de la tabla real, donde sólo se muestran 5 de los nodos más centrales, pertenecientes a dos clusters diferentes; los genes _:A615, _:A617 y _:A091 pertenecen al cluster 7 (Modularity Class 7), y son los genes con mayor centralidad de cercanía y mayor grado de toda la red y por lo tanto de dicho cluster. Por otro lado, los genes _:A092 y _:A664 pertenecen al cluster 4 (Modularity Class 4), siendo los nodos más centrales (usando C*loseness Centrality* y el grado) del cluster 4.

El siguiente paso es el enriquecimiento semántico de los datos de la tabla I, que como ya se mencionó es realizado con la herramienta PANTHER, ofrecida por los mismos desarrolladores de GO. La lista de identificadores de los genes se le da como entrada a PANTHER, y una salida que da la herramienta se puede observar en la figura 10, donde ya todos los términos están referenciados a un concepto en GO.

TABLA I. SALIDA GEPHI

Label	Grado	Closeness Centrality	Modularity Class
_:A615	4	5.69672131	7
_:A617	4	5.69672131	7
_:A1091	4	5.69672131	7
_:A1092	4	5.69672131	4
_:A664	4	5.68032787	4

Term	Background frequency	Sample frequency	Expected	+/-	P-value
transforming growth factor beta receptor signaling pathway (GO:0007179)	134	14	3.134e-01	+	1.810e-17
transmembrane receptor protein serine/threonine kinase signaling pathway (GO:0007178)	213	15	4.982e-01	+	2.548e-16
cellular response to transforming growth factor beta stimulus (GO:0071560)	168	14	3.930e-01	+	4.065e-16
response to transforming growth factor beta (GO:0071559)	168	14	3.930e-01	+	4.065e-16
regulation of cellular response to growth factor stimulus (GO:0090287)	173	13	4.047e-01	+	2.807e-14
negative regulation of cellular response to growth factor stimulus (GO:0090288)	102	11	2.386e-01	+	1.488e-13
regulation of transforming growth factor beta receptor signaling pathway (GO:0017015)	103	11	2.409e-01	+	1.654e-13
response to endogenous stimulus (GO:0009719)	1289	23	3.015e+00	+	3.420e-13
cellular response to growth factor stimulus (GO:0071363)	565	17	1.322e+00	+	1.077e-12
cellular response to endogenous stimulus (GO:0071495)	927	20	2.168e+00	+	1.269e-12
enzyme linked receptor protein signaling pathway (GO:0007167)	811	19	1.897e+00	+	1.675e-12
response to growth factor (GO:0070848)	582	17	1.361e+00	+	1.738e-12

Fig 10. Nodos con contenido semantico en PANTHER

V. CONCLUSIONES

En este trabajo se propuso el uso de técnicas de clustering orientadas en un principio para el Análisis de Redes Sociales (SNA), para detectar comunidades o grupos en redes de *signaling pathway*. Como principal aporte, con respecto a las demás técnicas de análisis de *signaling pathway* es que no se usa una técnica de clustering tradicional, sino que son de otro ámbito (SNA). De esta manera, se pudo usar simplemente la idea de modularidad, ya que no es necesario estudiar las características de los nodos para ir creando los grupos.

Dentro del SNA existen métricas de centralidad, las cuales son diferentes de la de centroides de los clusters, estas métricas permitieron identificar nodos centrales dentro de los grupos, sin necesidad de nuevo de hacer un estudio de las características de los nodos, solo estudiando sus estructuras y conectividad. El resultado del SNA son los datos de la red separados por grupos, también llamados comunidades. En particular, cada nodo contiene su respectivo valor de centralidad para las diferentes métricas.

Por otro lado, para el enriquecimiento semántico en específico se realizó una búsqueda en Gene Ontology (GO) usando el motor de enriquecimiento PANTHER, para enriquecer semánticamente los nodos más centrales de cada grupo. Esto aporta mucha información de valor para los biólogos.

Particularmente, queda como trabajo futuro una aplicación integrada que use todas estas herramientas, y dé cómo salida los datos que se logran enriquecer con GO.

AGRADECIMIENTO

Al Proyecto CDCHTA I – 1407 – 14 – 02 – B de la Universidad de Los Andes por su apoyo financiero. Dr. Aguilar ha sido parcialmente financiado por el Proyecto Prometeo del Ministerio de Educación Superior, Ciencia, Tecnología e Innovación de la República del Ecuador.

REFERENCES

[1] Kitano,H. (2002) Systems biology: a brief overview. Science, 295, 1662–1664.

[2] Barabasi,A.B. and Oltvai,Z.N. (2004) Network biology: understanding the cell's functional organization. Nat. Rev., 5, 101–113.

[3] Chen,L., Wang,R.S. and Zhang,X.S. (2009) Biomolecular Networks: Methods and Applications in Systems Biology. John Wiley & Sons, Hoboken, NJ.

[4] Chen,L., Wang,R.Q. and Aihara,K. (2010) Modeling Biomolecular Networks in Cells: Structures and Dynamics. Springer, London.

[5] Stelzl,U., Worm,U., Lalowski,M., Haenig,C., Brembeck,F.H., Goehler,H., Stroedicke,M., Zenkner,M., Schoenherr,A., Koeppen,S. et al. (2005) Human protein–protein interaction network: a resource for annotating the proteome. Cell, 122, 957–968.

[6] Hasty,J., McMillen,D., Isaacs,F. and Collins,J.J. (2001) Computational studies of gene regulatory networks: in numero molecular biology. Nat. Rev. Genet., 2, 268–279.

[7] Ravasz,E., Someraz,A.L., Mongru,D.A., Oltvai,Z.N. and Barabasi,A.L. (2002) Hierarchical organization of modularity in metabolic networks. Science, 297, 1551–1555.

[8] Wang, J., Huang, Q., Liu, Z., Wang, Y., Wu, L., Chen, L., and Zhang1, X. (2011) NOA: a novel Network Ontology Analysis method. Nucleic Acids Research, 2011, Vol. 39, No. 13 e87 doi:10.1093/nar/gkr251

[9] Chun, H., Tsuruoka, Y., Kim, J., Shiba R., Nagata, N., Hishiki, T., and Tsujii, J. Automatic recognition of topic-classified relations between prostate cancer and genes using MEDLINE abstracts. BMC Bioinformatics, BioMed Central. BMC Bioinformatics 2006, 7(Suppl 3):S4 doi:10.1186/1471-2105-7-S3-S4

[10] Gu, Z., Wang, J. (2009). CePa: an R package for finding significant pathways weighted by multiple network centralities. Vol. 29 no. 5 2013, pages 658–660 BIOINFORMATICS APPLICATIONS NOTE doi:10.1093/bioinformatics/btt008

[11] Porras, P. (2013). Network generation and analysis through Cytoscape and PSICQUIC. EMBL-EBI V6. Wellcome Trust Genome Campus Hinxton Cambridge CB10 1SD, U.K.

[12] Bettembourg, C., Diot, C., Dameron, O. (2014) Semantic particularity measure for functional characterization of gene sets using gene ontology. PLoS One. 2014 Jan 28;9(1):e86525. doi: 10.1371/journal.pone.0086525. eCollection 2014.

[13] Borgatti, S. (2005). Centrality and network flow. Social Networks 27: 55–71.

[14] Sun, J., Tang, J. (2011). A survey of models and algorithms for social influence analysis. En Charu C. Aggarwal. Social network data analytics (Nueva York: Springer): 177–214

[15] Blondel, V., Guillaume, J., Lambiotte, R., Lefebvre, E. (2008). Fast unfolding of communities in large networks. Journal reference: J. Stat. Mech. (2008) P10008 DOI: 10.1088/1742-5468/2008/10/P10008

[16] Aggarwal, C., and Wang, H. (2010). Managing and Mining Graph Data. Advances in Database Systems, Springer.

[17] Agrawal, R., and Srikant, R. (1994). Fast algorithms for mining association rules in large databases, VLDB Conference, 1994.

[18] Agrawal, S., Chaudhuri, S., and Das, G. (2002). A system for keywordbased search over relational databases. ICDE Conference, 2002. DBXplorer.

[19] Bhagat, S., Cormode, G., and Rozenbaum, I. (2007). Applying link-based classification to label blogs. WebKDD/SNA-KDD, pages 97–117, 2007.

[20] Yu, G., Wang, LG., Han, Y., and He, QY. (2012). ClusterProfiler: an R package for comparing biological themes among gene clusters. Journal of Integrative Biology 2012, 16(5):284-287. http://dx.doi.org/10.1089/omi.2011.0118

[21] Bux, M., Leser, U., and Philippe, T. (2012). Diploma Thesis Exposé: Comparing semantically enriched experimental protein networks in colorectal cancer. Humboldt Universität zu Berlin.

[22] Baudot, A., Gomez-Lopez, G., and Valencia, A. (2009). Translational disease interpretation with molecular networks. Genome Biol, 10(6):221, 2009. URL http://dx.doi.org/10.1186/gb-2009-10-6-221.

[23] Chuang, HY., Lee, E., Liu, YT., Lee, D., and Ideker, T. (2007). Network-based classication of breast cancer metastasis. Mol Syst Biol, 3:140, 2007. URL http://dx.doi.org/10.1038/msb4100180.