Obtención de redes de asociación directa con aplicación a datos ómicos

Marina Ainciburu
Plan de Estudios del Estudiante
Estadística y bioinformática

Nombre Consultor/a – Esteban Vegas
Nombre Profesor/a responsable de la asignatura – Alex Sánchez

Fecha Entrega

05 de junio de 2018

FICHA DEL TRABAJO FINAL

Título del trabajo:	*Obtención de redes de asociación directa con aplicación a datos ómicos*
Nombre del autor:	*Marina Ainciburu Fernández*
Nombre del consultor/a:	*Esteban Vegas Lozano*
Nombre del PRA:	*Alexandre Sánchez Plá*
Fecha de entrega (mm/aaaa):	03/2018
Titulación::	*Máster en bioinformática y bioestadística*
Área del Trabajo Final:	*Estadística y bioinformática*
Idioma del trabajo:	*Español*
Palabras clave	*Red de asociación directa, Gaussian Graphical Models (GGM), …*

Resumen del Trabajo (máximo 250 palabras): *Con la finalidad, contexto de aplicación, metodología, resultados i conclusiones del trabajo.*

Los microRNA (miRNA) son importantes reguladores génicos, que inhiben la traducción y/o promueven la degradación de sus RNA diana. Participan en procesos esenciales y juegan un papel en muchas patologías. Por ello, establecer las redes de regulación miRNA – genes puede ser importante. Existen diversos métodos computacionales de predicción de dianas. En este trabajo, exploramos métodos para inferir asociaciones directas a partir de datos de expresión de experimentos de microarrays. Para ello, utilizamos algoritmos que estiman correlaciones parciales y permiten crear Gaussian Graphical Models (GGM) a partir de datos de grandes dimensiones, con mayor número de variables que muestras.

Evaluamos el rendimiento de algoritmos de este tipo implementados en paquetes de R aplicándolos a datos de diversas características: datos pequeños con n > p, datos normales multivariantes simulados, datos genómicos simulados y datos reales de expresión de mRNA y miRNA.

Los resultados muestran la utilidad de las redes de asociación creadas por GGM en datos con pocas variables y muchas muestras. Sin embargo, el fallo de estos métodos a la hora de inferir asociaciones se hace evidente conforme aumenta el número de variables.

Abstract (in English, 250 words or less):

microRNA (miRNA) are key genic regulators that inhibit translation and/or promote degradation of their target RNA. They take part in essential biological processes and play a role in various diseases. Thus, stablishing the miRNA – genes regulation network can be important. Different computational methods have been developed to predict miRNA targets. In this study, we explore methods to infer direct associations from expression data coming from microarray experiments. In order to do this, we use algorithms that estimate partial correlations and allow us to create Gaussian Graphical Models (GGM) from high-dimensional data, where there are more variables than samples.

We evaluate the performance of algorithms already implemented in R packages, by applying them to different kinds of data: small data, with $n > p$, simulated multivariate normal data, simulated genomic data and real mRNA and miRNA expression data.

Results show how association networks created by GGM are useful in conditions with little variable and lots of samples. However, we become aware of the failure of these methods to infer associations as the number of variables grows bigger.

Índice

1. Introducción

1.1 Contexto y justificación del Trabajo

Descripción general

Este trabajo se centra en la creación de redes de asociación directa a partir de resultados de microarray de expresión de mRNA y miRNA.

Los microarrays son un tipo de experimentos ómicos que permiten estudiar de forma simultánea la expresión del conjunto del genoma en unas determinadas condiciones. Los resultados obtenidos permiten construir redes con las que visualizar las asociaciones entre los diferentes elementos estudiados. En este caso, se pretende estudiar las asociaciones entre mRNAs y microRNAs, ya que los segundos regulan de forma específica los primeros, inhibiendo la traducción de sus mRNA diana.

Por otra parte, se quiere representar solamente las asociaciones directas entre estos dos tipos de moléculas, excluyendo relaciones más distantes. Para ello se utilizarán algoritmos que permiten aplicar Gaussian Graphical Models (GGM) a datos en los que existen más variables que muestras. Esto es lo que ocurre generalmente en los experimentos ómicos, donde se estudia gran cantidad de genes, proteínas, metabolitos… en pocos indivíduos.

Justificación del TFM

Los experimentos ómicos proporcionan herramientas muy útiles en el estudio de enfermedades complejas, en las que una gran cantidad de elementos pueden tener un papel relevante.

A su vez, los miRNA tienen gran importancia en enfermedades como el cáncer, ya que cambios en su papel regulador pueden ser una causa de la sobrexpresión o silenciamiento de genes involucrados en la enfermedad. Experimentos a pequeña escala, dirigidos a estudiar asociaciones mRNA – miRNA individuales, tienen un coste muy alto. Por ello, el desarrollo de métodos computacionales para predecir asociaciones puede ser de gran ayuda.

En conjunto, la interpretación de los resultados de microarrays mediante redes de asociaciones directas puede resultar muy útil, ya que se crean gráfico que se acercan a la representación de relaciones causales. Esto facilita la identificación de asociaciones individuales miRNA – mRNA potencialmente relevantes.

1.2 Objetivos del Trabajo

Objetivos generales

I. Aplicar GGM a datos de dimensiones pequeñas, con n > p

II. Generar datos simulados de distribución normal multivariante para omparar diversos algoritmos para la aplicación de GGM a datos con más variables que muestras

III. Generar datos simulados de niveles de expresión de mRNA y miRNA con asociaciones conocidas, sobre los cuales aplicar GGM.

IV. Aplicar técnicas GGM a la búsqueda de asociaciones mRNA-miRNA.

V. Establecer un sistema de evaluación de los resultados, mediante su comparación con información sobre asociaciones mRNA – miRNA ya validadas.

Objetivos específicos

I. Aplicar GGM a datos de dimensiones pequeñas, con n > p
   a. Crear redes con correlaciones, correlaciones parciales y GGM
   b. Comparar resultados

II. Datos normales simulados
   a. Generar datos simulados
   b. Comparar diversos algoritmos GGM ya implementados en paquetes de R
   c. Ajustar parámetros de penalización de los algoritmos

III. Datos genómicos simulados
   a. Generar datos simulados de niveles de expresión de mRNA y miRNA con asociaciones conocidas
   b. Aplicar algoritmos GGM para recuperar las asociaciones

IV. Aplicar técnicas GGM a la búsqueda de asociaciones mRNA-miRNA.
   a. Aplicar los métodos con mejor resultado para buscar asociaciones mRNA - miRNA
   b. Aplicar los métodos implementados en el paquete *miRLab*

V. Establecer sistema de evaluación
   a. Validar las asociaciones miRNA – mRNA con los métodos GGM aplicados a datos reales, contrastándolas con información de bases de datos
   b. Calcular parámetros de evaluación del rendimiento para comparar métodos

1.3 Enfoque y método seguido

Las redes de asociación génica se pueden construir de diversas formas. En este trabajo se ha escogido representar asociaciones directas mediante GGM de acuerdo con las razones expuestas en la justificación del TFM.

Por otro lado, la aplicación de GGM a datos ómicos se puede realizar siguiendo varias estrategias, que se pueden resumir en tres grupos.

1. Análisis clásico de GGM. Para ello, se deben escoger muy pocos genes para analizar o hacer grandes clusters de genes para conseguir n > p

2. Utilizar correlaciones parciales de orden limitado. Se crea un modelo intermedio entre GGM y uno basado en correlaciones.

3. Crear GGM regularizados. Se basa en encontrar estimadores para las matrices de covarianza y de precisión (inversa de la matriz de covarianza). Existen diversos métodos para estimar estos valores.

El tercer método es el más popular de los tres y es el que se utilizará en este trabajo. Para empezar, se compararán varios algoritmos que realizan la estimación de valores mediante técnicas distintas, utilizando datos simulados (datos normales y datos genómicos) con asociaciones conocidas. Los métodos que ofrezcan mejores resultados se usarán para la búsqueda de asociaciones mRNA – miRNA, a partir de datos de niveles de expresión. Para evaluar el rendimiento de los algoritmos aplicados a estos datos, se implementará un sistema de comparación con información de bases de datos con asociaciones probadas. Por último, se utilizará el paquete *miRLab*, que implementa métodos de inferencia de asociaciones miRNA – gen, basándose en el mismo concepto que este trabajo.

En cuanto al software utilizado, se ha decidido trabajar con R ya que dispone de todas las herramientas necesarias para llevar a cabo este trabajo. Por un lado, permite la importación de los datos y el trabajo con éstos en forma de matrices. También existen paquetes en los que se implementan los algoritmos y métodos estadísticos que se utilizarán. Además, ofrece diversas opciones de representación gráfica de los resultados.

1.4 Planificación del Trabajo

1. Aplicar GGM a datos con n > p
    a. Conseguir un set de datos apropiado
    b. Crear redes de correlaciones, correlaciones parciales y GGM
    c. Comparar resultados de la red de correlaciones parciales y GGM

2. Comparar algoritmos GGM con datos normales simulados
    a. Aplicar diversos algoritmos ya implementados en paquetes de R como *GGMselect* y *huge* con datos simulados
    b. Probar parámetros de penalización
    c. Comparar el rendimiento de cada método

3. Comparar algoritmos GGM con datos genómicos simulados
    a. Generar niveles de expresión de mRNAs´y miRNAs con asociaciones conocidas
    b. Aplicar algoritmos GGM
    c. Comparar resultados

4. Aplicar técnicas GGM a la búsqueda de asociaciones mRNA - miRNA
    a. Obtener fuentes de datos de experimentos de microarrays de mRNA y miRNA
    b. Realizar una exploración y preprocesado de los datos
    c. Crear modelo GGM con el método escogido en las simulaciones
    d. Aplicar métodos del paquete miRLab a los mismos datos

5. Establecer sistema de evaluación
    a. Obtener información sobre asociaciones validadas
    b. Contar las asociaciones validadas de las generadas por el modelo GGM
    c. Calcular parámetros de evaluación

1.5 Breve sumario de productos obtenidos

Además de la memoria, durante la realización de este estudio se han recopilado datos y se ha generado código R que se adjunta como material suplementario. Por un lado, se incluyen dos datasets para la validación de asociaciones y dos archivos con datos a partir de los cuales generar redes de asociación. Por otro lado, se ha creado un informe dinámico en formato rmd con el código R escrito para lograr los resultados de cada uno de los subapartados.

1.6 Breve descripción de los otros capítulos de la memoria

En el capítulo siguiente se desarrolla una introducción teórica más detallada, en la que se habla del proceso de regulación de mRNA por parte de miRNA, así como de las técnicas para su estudio y de las estrategias para establecer redes de asociación. También se comentan diferentes estrategias para crear redes de asociaciones y se dan más detalles sobre los GGM.

A continuación, se describen los materiales y métodos. Este capítulo se centra en explicar la procedencia y naturaleza de los datos usados y las bases de los algoritmos aplicados.

El apartado de resultados se ordena siguiendo los objetivos del trabajo y recoge figuras y tablas que resumen los resultados del trabajo, junto con pequeñas descripciones.

En la discusión se realiza un análisis global e integrativo de los resultados obtenidos y se valoran los problemas encontrados y las posibles estrategias alternativas. Finalmente, se termina con una pequeña conclusión.

2. Contexto teórico

2.1 Regulación génica por miRNA

Un microRNA (miRNA) es un RNA pequeño, de unos 22 nucleótidos, que no codifica para una proteína y que lleva a cabo un silenciamiento específico post-transcripcional de genes. Su actividad resulta tanto en la inhibición de la traducción como en la degradación del RNA diana. Los miRNA son una de las moléculas reguladoras más abundantes en organismos pluricelulares. Hay evidencias de su papel esencial en procesos como el desarrollo, la diferenciación y la proliferación celular, así como en varias enfermedades como el cáncer o la diabetes[1].

Se estima que existen entre 200 y 255 miRNA distintos en el genoma humano[2]. Estos se encuentran a lo largo de todo el genoma, tanto en regiones alejadas de genes codificantes como en intrones de otros genes. El segundo caso puede dar lugar a la expresión coordinada de un mRNA y un miRNA regulador. Existen también clusters de miRNA que se expresan conjuntamente. Por otro lado, la expresión varía en función del tipo celular y el estado de desarrollo, por lo que se pueden establecer diferentes patrones de expresión de miRNA[3].

La transcripción de miRNA da lugar a moléculas precursoras (pri-miRNA), de unos 80nt, con cap-5' y una cola de poli A-3'. Los extremos de estas moléculas son complementarios, por lo que se pliegan sobre sí mismas en forma de horquilla, con un bucle en el centro (en su parte no complementaria). Tras una primera escisión en el núcleo se generan los pre-miRNA, de 65nt. Estos salen al citoplasma, donde son reconocidos por la RNAsa Dicer, que corta el bucle del pre-miRNA, dando lugar a un RNA dúplex, formado por el miRNA y una cadena complementaria. El miRNA dúplex se incorpora al complejo silenciador inducido por RNA (RISC) y en este proceso la cadena complementaria se degrada, quedando el miRNA simple. El miRNA es la guía para encontrar un mRNA diana, ya que éste es complementario, de forma parcial, al extremo 3' no traducible del segundo. De esta forma, el complejo RISC puede bloquear la síntesis de proteína o degradar de manera específica los mRNA diana[1].

Cada vez hay más evidencias de la importancia de estos RNA en algunas enfermedades. Se han identificado miRNA cuyo silenciamiento o sobrexpresión da lugar a la aparición o el rescate de fenotipos patológicos[4,5]. Por otro lado, se han estudiado los patrones de expresión de miRNA en diversos tipos de cáncer y se han encontrado similitudes en función de la respuesta al tratamiento o el tipo de tumor[6]. Todo esto lleva a considerar el potencial de los miRNA como herramientas de diagnóstico y pronóstico, así como futuras estrategias de terapia.

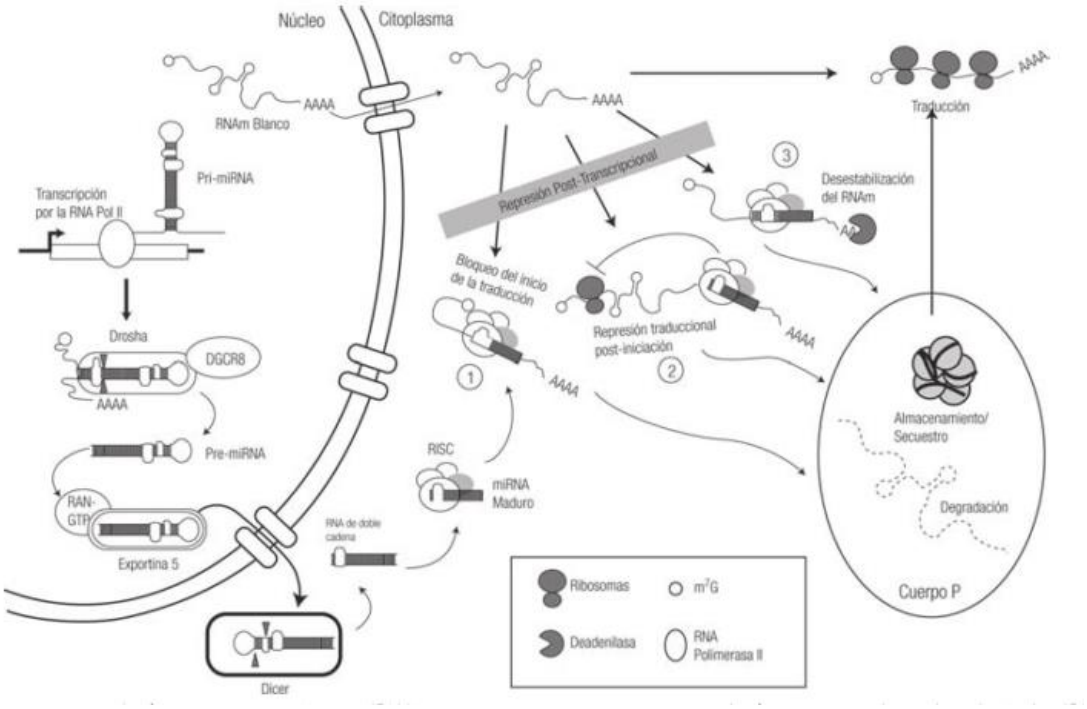

Fig. 1: ruta de síntesis, procesado y mecanismos de regulación génica de los miRNA. Fuente, Lugo Trampe, 2009[1]

Predicción de mRNA diana

La imperfección de la complementariedad de bases hace que un miRNA pueda contar con varias dianas. A su vez, un mismo mRNA puede ser el blanco de distintos miRNA, ya sea por medio de un solo lugar de unión o de varios. Esto da lugar a una complejidad enorme de interacciones[7].

Para estudiar estas interacciones se utilizan métodos biológicos y computacionales. Por un lado, los métodos biológicos consisten en experimentos a pequeña escala, generalmente enfocados a estudiar un solo miRNA, con un coste muy alto. Por ello, se están desarrollando numerosos métodos computacionales para la predicción de dianas a gran escala.

Los algoritmos para la identificación de dianas se basan principalmente en la combinación de cuatro criterios[7].

1. Complementariedad de bases
2. Accesibilidad del lugar de unión del mRNA en función de su estructura secundaria y energía necesaria para exponerlo
3. Estabilidad de la unión
4. Conservación de las dianas entre especies

En general, se ha visto que todos los algoritmos tienen una tasa de falsos positivos y falsos negativos muy alta[8]. Además, cambios pequeños en los criterios de predicción dan lugar a resultados muy divergentes[9].

Como consecuencia, han ido apareciendo métodos computacionales alternativos, basados en identificación de relaciones a partir de perfiles de expresión de miRNA y mRNA. La suposición sobre la que trabajan es que si un miRNA actúa como regulador sobre un mRNA, los niveles de éste deben cambiar y esto debería quedar reflejado en forma de correlación en los niveles de expresión de ambos[10]. A continuación, se explican más detalladamente estrategias para identificar estas asociaciones y generar redes con ellas.

2.2 Identificación de asociaciones y creación de redes

La estrategia convencional a la hora de buscar asociaciones es basarse en medidas de correlación tradicionales, como el coeficiente de Pearson o el de Spearman, calculados de forma individual para cada par de mRNA – miRNA. Al trabajar con grandes matrices de expresión genómica, esto da lugar a la identificación de asociaciones tanto directas cómo indirectas[10]. En este caso, las primeras son las interacciones miRNA – mRNA diana, mientras que las segundas son relaciones más distantes, fruto de la compleja red de regulación génica.

Es posible suprimir las asociaciones indirectas mediante el uso de correlaciones parciales. Este parámetro indica la correlación entre dos variables condicionada a la correlación con el resto de variables. Su cálculo se basa en realizar una regresión de cada una de las dos variables sobre el resto de variables y obtener el coeficiente de Pearson de los residuos de las dos regresiones. De esta forma, se elimina el efecto lineal del resto de variables sobre el par de variables estudiado[11].

Normalmente, la matriz de correlaciones parciales se obtiene a partir de la inversa de la matriz de covarianzas: la matriz de concentración o precisión.

La matriz de correlaciones parciales permite crear **Gaussian Graphical Models (GMM)**; gráficos no dirigidos, en los que las aristas representan dependencia condicional entre dos variables. En otras palabras, la presencia de una arista indica que dos variables están correlacionadas una vez se ha eliminado el efecto del resto de variables (dependencia condicional). Por otro lado, la ausencia de arista se corresponde con una correlación parcial igual a 0 e implica independencia condicional, lo cual no quiere decir que las variables puedan estar relacionadas de forma indirecta. De esta forma, los GGM proporcionan una buena representación de la dependencia, pero no de la independencia, al contrario que los gráficos basados en correlaciones, en que la ausencia de aristas indica la independencia marginal de dos variables[12].

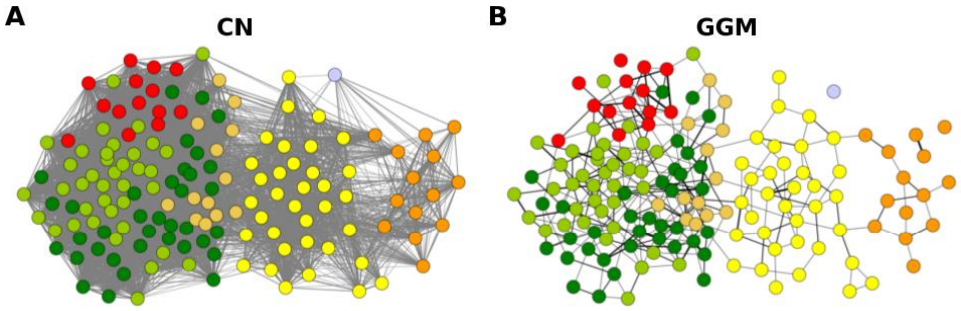

Fig. 2: ejemplos de redes de asociación de rutas metabólicas creadas a partir de correlaciones (A) y de correlaciones parciales (B). Los colores representan clases de metabolitos. Fuente, Krumsiek *et al,* 2011[11].

Para decidir si existe una asociación entre dos nodos, debe calcularse la significación de cada elemento de la matriz de correlaciones parciales. Las conexiones establecidas se pueden resumir en una matriz de adyacencia; una matriz p x p simétrica de unos y ceros, en que los unos indican la presencia de conexión entre dos variables.

Si se quiere construir un gráfico bipartito, con dos tipos de variables, la matriz de adyacencia se puede sustituir por una matriz de incidencia, también con unos y ceros, pero en este caso con las variables de un tipo situadas en las filas y las del otro tipo, en las columnas.

Los GGM clásicos se crean asumiendo normalidad de los datos y a partir de datos con mayor número de muestras que variables ($n > p$). Sin embargo, cuando se parte de la situación contraria ($n < p$), la matriz de covarianza es singular, no invertible. Por ello, la matriz de precisión no se puede calcular [13]. Existen diversas estrategias para afrontar este problema:

1. Llevar los datos a condiciones clásicas $n > p$, ya sea por selección de variables, por su agrupación en clusters…

2. Usar correlaciones parciales de orden limitado. Una correlación parcial de orden n es aquella que se calcula incluyendo n variables independientes en la regresión. La correlación parcial de orden completo tiene en cuenta el efecto de todas las variables. La correlación parcial de orden 0 es el coeficiente clásico de correlación. De esta forma, el uso de correlaciones parciales de orden limitado da lugar a gráficos a medio camino entre un GGM y una red de correlaciones incondicionales.

3. Aplicar GGM regularizados. Los métodos englobados en esta categoría buscan estimaciones para la matriz de precisión, ya sea a partir de las observaciones o estimando también la matriz de covarianzas. Es común incorporar un parámetro de penalización para estimar la matriz de precisión

3. Materiales y métodos

3.1 Datos clínicos

Se ha realizado una prueba inicial de la creación de redes de asociación utilizando un dataset con más muestras que variables. Se trata de un dataset de 403 individuos con diabetes en los que se miden 19 variables[14]. Hemos eliminado variables categóricas como *gender* y *location*, así como dos variables en que prácticamente todos los valores eran desconocidos. También hemos eliminado individuos con datos desconocidos, tras lo cual nos hemos quedado con un dataset con 375 muestras y 13 variables.

Con estos datos hemos creado redes de asociación para visualizar las relaciones entre las 13 variables. Por un lado, hemos creado un gráfico a partir de los coeficientes de correlación de Pearson (calculados con el paquete *Hmisc*). Al tener mayor número de muestras que variables, también se han podido calcular las correlaciones parciales directamente a partir de la matriz de precisión (paquete *ppcor*). En ambos casos, se ha estimado la significación de cada parámetro mediante un ajuste de los p valores por false Discovery rate (fdr) y se ha indicado un nivel de significación de 0.05. Por último, se ha creado otro gráfico utilizando algoritmos GGM que estiman la matriz de precisión y las correlaciones parciales. En este apartado, sólo se han utilizado los métodos del paquete *GGMselect*, explicados más adelante (3.3).

3.2 Simulación de datos

Datos normales multivariantes

En el paquete *Huge*, la función *huge.generator()* permite generar datos que sigan una distribución normal multivariante y con una estructura gráfica conocida. Se debe especificar el número de muestras (n) y variables (d), así como la estructura deseada para el gráfico (random, hub, cluster…). En función de la estructura, se pueden añadir parámetros adicionales. En nuestro caso, hemos utilizado diferentes combinaciones de n y d y una estructura con 3 clusters (graph = "cluster", g = 3).

La función crea en primer lugar la estructura del gráfico, representada mediante una matriz de adyacencia. A partir de aquí, se generan las matrices de covarianza y precisión y las observaciones.

El conocimiento de la matriz de adyacencia nos permite evaluar el rendimiento de diferentes métodos GGM.

En sucesivas simulaciones, se han generado sets de datos normales multivariantes (con *mvrnorm*) con las mismas medias y matriz de covarianzas que los datos de *huge.generator*. En cada simulación se ha generado un modelo GGM y se ha comparado su matriz de adyacencia con la matriz conocida. Finalmente, se han calculado los valores medios de los parámetros de evaluación para el conjunto de las simulaciones.

El mismo mecanismo se ha llevado a cabo para analizar los resultados en función del número de muestras y variables, el rendimiento de algoritmos implementados en diferentes paquetes de R (explicados en 3.3) y los cambios provocados por los parámetros de penalización.

Datos genómicos

En primer lugar, se ha utilizado el paquete *madsim* para simular resultados de microarrays de expresión génica con dos condiciones. La función *madsim* permite especificar el número de genes y muestras de cada condición, el porcentaje de genes con expresión diferencial, la desviación estándar, el rango de los niveles de expresión y más parámetros.

Después se ha creado una matriz de incidencia en que, de forma aleatoria, se han establecido asociaciones entre genes y un conjunto de miRNAs.

Una vez conocidos los genes con que interacciona cada miRNA, se ha pasado a generar niveles de expresión de miRNAs de forma que reflejen estas interacciones. Para ello, se han creado ecuaciones lineales con pendiente negativa y un error aleatorio de distribución normal. El miRNA es la variable dependiente y todos los mRNAs con que esté asociado son las variables independientes. Así, las ecuaciones son de la forma:

$$miRNA = \beta_0 - \beta_1 Gen1 - \ldots - \beta_x Genx + \varepsilon$$

β_0 se genera a partir de la suma de los valores máximos de expresión de los genes incluidos en la ecuación. El resto de parámetros β son un número aleatorio entre 0.6 y 0.9. El término de error sigue una distribución normal $\varepsilon \sim N(0, 0.05)$.

Se han intentado controlar los valores de los coeficientes de correlación para que los datos creados se ajusten lo mejor posible a las asociaciones.

3.3 Algoritmos GGM

A continuación, se detallan los métodos utilizados para crear los GGM. Se trata de algoritmos basados en la estimación de la matriz de precisión, que incluyen una función de penalización para que las estimaciones se acerquen más a 0 y se obtengan redes más dispersas.

Están implementados en diferentes paquetes de R.

Paquete *GGMselect*

En el paquete GGMselect[15] se incluyen 3 algoritmos GGM:

- C01: basado en el proceso de estimación propuesto por Wille y Buhlmann[16]
- LA: basado en el proceso de estimación propuesto por Meinshausen y Buhlmann[17]
- EW: es una version modificada de LA, basada en el adaptative lasso de Zou[18]

También es posible usar combinaciones de 2 o 3 métodos, de forma que se elegirá automáticamente el gráfico óptimo de entre todos los generados.

Se aplican mediante la función selectFasta(), en la que se introduce la matriz de observaciones y el método que se va a usar (C01, LA, EW o alguna combinación de los 3). Parámetros opcionales importantes son dmax: número máximo de aristas por nodo y k: el valor para la función de penalización que, por defecto es K = 2.5 y suele oscilar entre 1 y 3.

Paquete *Huge*

En el paquete Huge se implementan otras 3 variantes de algoritmos GGM:

- mb: estimación propuesta por Meinshausen-Buhlmann[17]
- glasso: graphical lasso[19]
- ct: método basado en crear umbrales en la matriz de correlación convencional[20]

Se aplican mediante la función huge(), que acepta como input la matriz de observaciones n x p y el método deseado. Además, se incluyen varios parámetros para ajustar el valor de penalización lambda. Los autores recomiendan dejar que el programa calcule los valores de lambda automáticamente, pero, si se desea, se puede introducir un vector con valores entre 0 y 1 o indicar el nº de valores que se quieren utilizar (nlambda) o el valor mínimos (lambda.min.ratio).

Paquete *GeneNet*

En *GeneNet*[21] se implementa un método para obtener estimadores por contracción de las correlaciones parciales. La función ggm.estimate.pcor() calcula estos estimadores a partir de la matriz de observaciones. Con la matriz de estimaciones, network.test.edges() calcula la significación de cada coeficiente, con un pvalor y la probabilidad de que exista asociación. Finalmente, extract.network() permite seleccionar las asociaciones que cumplan un cierto valor de significación.

3.4 Datos genómicos reales

Los datos descritos a continuación, tanto los de expresión como los de validación, se encuentran disponibles en https://sourceforge.net/projects/mirlab/files/?source=navbar

Datos de expresión

Se ha utilizado un dataset en el que se comparan líneas celulares tumorales de tipo epitelial y mesenquimal, creado para estudiar el proceso de transición epitelio mesenquimal (EMT) que tiene lugar durante el desarrollo embrionario, la cicatrización y la progresión tumoral.

En el conjunto de datos se han integrado niveles de expresión de mRNA y miRNA, procedentes de estudios diferentes, Shankavaram *et al*[22] y Sokilde *et al*[23], respectivamente.

En ambos casos, el análisis se realizó sobre las 60 líneas del National Cancer Institute (NCI-60).

La integración de los datos se llevó a cabo en un tercer estudio[24]. Los autores realizaron un análisis de expresión diferencial entre 11 muestras epiteliales y 36 mesenquimales, mediante el paquete *limma* de Bioconductor. Se encontraron 35 sondas de miRNA y 1154 sondas de mRNA con expresión diferencial (p valor ajustado por BH < 0.05).

El archivo de que disponemos contiene los niveles de expresión normalizados de los 35 miRNAs y 1154 mRNAs con expresión diferencial en las 11 muestras epiteliales y 36 muestras mesenquimales.

En este apartado se ha utilizado solamente el método con el que se han obtenido mejores resultados en los apartados de simulaciones.

Datos de validación

El número de dianas de miRNA experimentalmente validadas es muy limitado, por lo que no es posible evaluar y comparar los métodos utilizados de forma completa. Aquí utilizamos dos datasets diferentes con información validada.

Por un lado, Le T.D. *et al*[24] crearon una base de datos con interacciones miRNA – gen a partir de la información disponible en 4 bases de datos: Tarbase v6.0, miRecords v2013, miRWalk v2.0 y miRTarBase v4.5. En estos repositorios se recogen interacciones verificadas en la bibliografía y revisadas de forma manual. Una vez eliminados los duplicados, se consiguió un dataset con 62.858 interacciones únicas.

Por otro lado, Yue Li *et al*[25] recogieron resultados de diversos experimentos de perturbación de miRNA por transfección. En estos experimentos, se transfectan células con un miRNA, causando su sobrexpresión y se compara la expresión génica con la de células control. Los autores recogieron resultados de experimentos en 77 tejidos y líneas celulares humanas y sobre 113 miRNA distintos. La información se resume en una matriz con filas de genes, columnas de miRNAs y los valores de expresión diferencial entre condiciones en forma de log2 fold change (LFC). Consideramos que un gen es diana de un miRNA si $|LFC| > 1$.

Para facilitar la evaluación de resultados, hemos resumido estos dos datasets de validación en uno, seleccionando solamente miRNA y genes presentes en nuestros datos de expresión. Eliminando duplicados, han quedado 424 asociaciones validadas que se pueden encontrar en el dataset de EMT.

3.5 Paquete miRLab

miRLab es un paquete de Bioconductor creado por Le T.D. *et al*[26] para detectar asociaciones entre miRNA y mRNA a partir de datos de niveles de expresión. Implementa diversos métodos para encontrar asociaciones, así como funciones para seleccionar los resultados más significativos y validarlos enfrentándolos a datos conocidos. Los creadores del paquete

proponen aplicar varios métodos de inferencia de asociaciones sobre los mismos datos e integrar los resultados mediante "ensemble methods", para lo cual también se han creado funciones.

El paquete permite importar datos en formato *csv* y aplicar sus métodos de inferencia y validación. Así, hemos utilizado los mismos datos de expresión y validación – que, de hecho, han sido implementados por los mismos autores – y hemos comparado el rendimiento de los métodos de *miRLab* con el de nuestros métodos.

De entre los métodos de inferencia disponibles, hemos aplicado por un lado la regresión Lasso y, por otro, la combinación de la correlación de Pearson, Regresión Lasso y el método de inferencia causal IDA.

3.6 Evaluación de resultados

El rendimiento de los métodos GGM se ha evaluado mediante el contaje de verdaderos y falsos positivos y negativos, así como el cálculo de las tasas de sensibilidad y especificidad. También se indica el número total de asociaciones que crea cada método.

4. Resultados

4.1 Datos clínicos

En una primera prueba del funcionamiento de los modelos GGM, hemos utilizado un dataset pequeño, con muchas más muestras (375) que variables (13). Se han creado redes de asociación para visualizar las relaciones entre las 13 variables. Los coeficientes de correlación dan lugar a una red concentrada, mientras que los otros dos métodos resultan en redes más dispersas. Además, en estas dos redes se observa cierta lógica en las interacciones, por ejemplo, en la relación de las variables colesterol, hdl y ratio (colesterol/hdl) (fig. 3).

Por otro lado, la evaluación de la red GGM contrastándola con la red de correlaciones parciales revela que, en el mejor de los resultados (C01.LA.EW), sólo aparece una interacción falsa (tabla 1), que resulta ser entre *glyhb* y *age*.

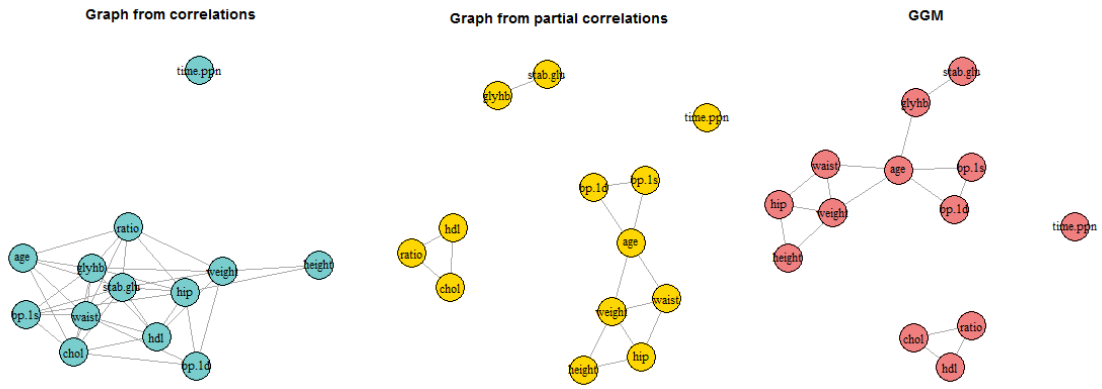

Fig. 3: redes de asociación, de izquierda a derecha, a partir de coeficientes de correlación de Pearson, correlaciones parciales y algoritmos GGM del paquete *GGMselect*

Method	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
C01	13	10	3	61	4	71.429	95.312
LA	15	13	2	62	1	92.857	96.875
EW	15	14	1	63	0	100.000	98.438
C01.LA.EW	15	14	1	63	0	100.000	98.438

Tabla 1: evaluación del rendimiento de los algoritmos de *GGMselect* en datos con más muestras que variables

Los resultados muestran un alto rendimiento de los modelos GGM en una situación de n > p.

4.2 Simulación de datos normales multivariantes

Con *huge.generator()* se han generado datos de distribución normal multivariante estructurados en 3 clusters, con matrices de covarianza y adyacencia conocidas, de la forma de la figura 4.

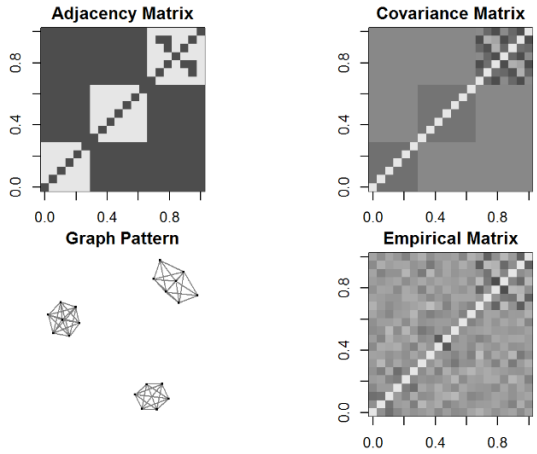

Fig. 4: estructura de los datos normales multivariantes simulados con *huge.generator*

El análisis de resultados en función del número de muestras y variables se ha llevado a cabo con la combinación de los métodos C01.LA del paquete *GGMselect*. En la tabla 2 se puede observar que los resultados han sido negativos. En las condiciones especificadas, prácticamente no se crean interacciones significativas, con lo que las tasas de sensibilidad son despreciables y la especificidad permanece muy alta. Solamente cuando se alcanzan 200 muestras aparecen algunas interacciones. En cualquier caso, esta no es la situación que nos interesa, ya que buscamos crear modelos a partir de datos con n < p.

	10	50	100	200
50	0.5 (0 , 99.954)	0.5 (0.333 , 100)	3 (1.361 , 99.907)	14.5 (8.497 , 99.86)
200	1 (0 , 99.994)	0.5 (0.025 , 100)	1 (0.051 , 100)	11 (0.451 , 99.989)
500	1 (0 , 99.999)	0 (0 , 100)	1 (0.008 , 100)	2.5 (0.02 , 100)

Tabla 2: evaluación de los métodos GGM C01.LA del paquete *GGMselect*. En las filas aparece el nº de variables y en las columnas, el número de muestras. En cada celda se recoge el nº medio de asociaciones creadas por cada modelo y, entre paréntesis, los porcentajes de sensibilidad y especificidad.

Al analizar por separado los distintos métodos de *GGMselect*, con n = 20 y d = 500, obtenemos los mismos resultados (Tabla 3). El método EW ha sido descartado debido a su enorme tiempo de ejecución.

	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
C01	0	0	0	112194	12556	0	100
LA	0	0	0	112194	12556	0	100
C01.LA	0	0	0	112194	12556	0	100

Tabla 3: comparación de métodos de *GGMselect* con n = 20, d = 500

El parámetro K es el que determina el valor de la función de penalización. Cambiando su valor predeterminado (K = 2.5), obtenemos mayor número de asociaciones al utilizar valores más pequeños, que implican menor penalización. Sin embargo, la sensibilidad prácticamente no aumenta, indicando que muchas de las asociaciones creadas no se corresponden con las asociaciones de referencia.

	1	1.5	2	2.5	3
C01	21 (0.024 , 99.984)	1.5 (0.012 , 100)	1 (0 , 99.999)	0 (0 , 100)	0 (0 , 100)
LA	16 (0.02 , 99.988)	2 (0.016 , 100)	1 (0 , 99.999)	0 (0 , 100)	0 (0 , 100)
C01.LA	16 (0.02 , 99.988)	2 (0.016 , 100)	1 (0 , 99.999)	0 (0 , 100)	0 (0 , 100)

Tabla 4: comparación de resultados con los métodos de *GGMselect,* n = 20, d = 500, en función del parámetro de penalización K

A continuación se han evaluado los 3 métodos del paquete *Huge*, con diferentes valores para el parámetro de penalización, que en este caso es lambda. El algoritmo calcula automáticamente una serie de valores descendentes, de 1 a 0, para lambda. En este caso se han calculado 5 valores. Los resultados muestran un incremento en el número medio de asociaciones conforme la penalización disminuye. Por otro lado, el método glasso es el que proporciona mejores resultado en términos de sensibilidad, aunque genera más falsos positivos, dando lugar a una menor especificidad. En todo caso, el número de asociaciones correctas que se crea sigue siendo muy bajo y a su vez aparecen gran cantidad de asociaciones que no se corresponden con la matriz de referencia.

	l1	l2	l3	l4	l5
mb	0 (0 , 100)	1366.5 (1.574 , 98.958)	3453 (3.82 , 97.35)	5008.5 (5.298 , 96.129)	5971 (6.22 , 95.374)
glasso	0 (0 , 100)	4116 (4.674 , 96.855)	10236 (10.521 , 92.054)	12658 (12.54 , 90.122)	13884 (13.537 , 89.141)
ct	1 (0.008 , 100)	1561 (1.94 , 98.826)	3120 (3.716 , 97.635)	4679 (5.274 , 96.42)	6238 (6.916 , 95.214)

B

	l1	l2	l3	l4	l5
mb	0.807	0.454	0.255	0.144	0.081
glasso	0.807	0.454	0.255	0.144	0.081
ct	0.807	0.555	0.506	0.475	0.449

Tabla 4: (A) comparación de métodos de *Huge* con n = 20, d = 500 y diferentes valores de lambda. En cada celda se recoge el nº medio de asociaciones creadas por cada modelo y, entre paréntesis, la sensibilidad y especificidad. (B) Valores de lambda utilizados en cada método.

Por último, el algoritmo implementado en *GeneNet* no ha conseguido crear ninguna asociación significativa.

4.3 Simulación de datos genómicos

Se han simulado resultados de microarrays con 100 genes, 2 condiciones con 10 muestras cada una y el 100% de los genes diferencialmente expresados, utilizando el paquete *madsim*. El análisis exploratorio de los datos ha revelado que éstos siguen una distribución típica de niveles de intensidad normalizados y que existe una diferenciación clara entre las dos condiciones (Fig. 5). Además, un análisis de expresión diferencial con *limma* ha demostrado que, efectivamente, todos los genes simulados tienen diferencias de expresión significativas entre condiciones.

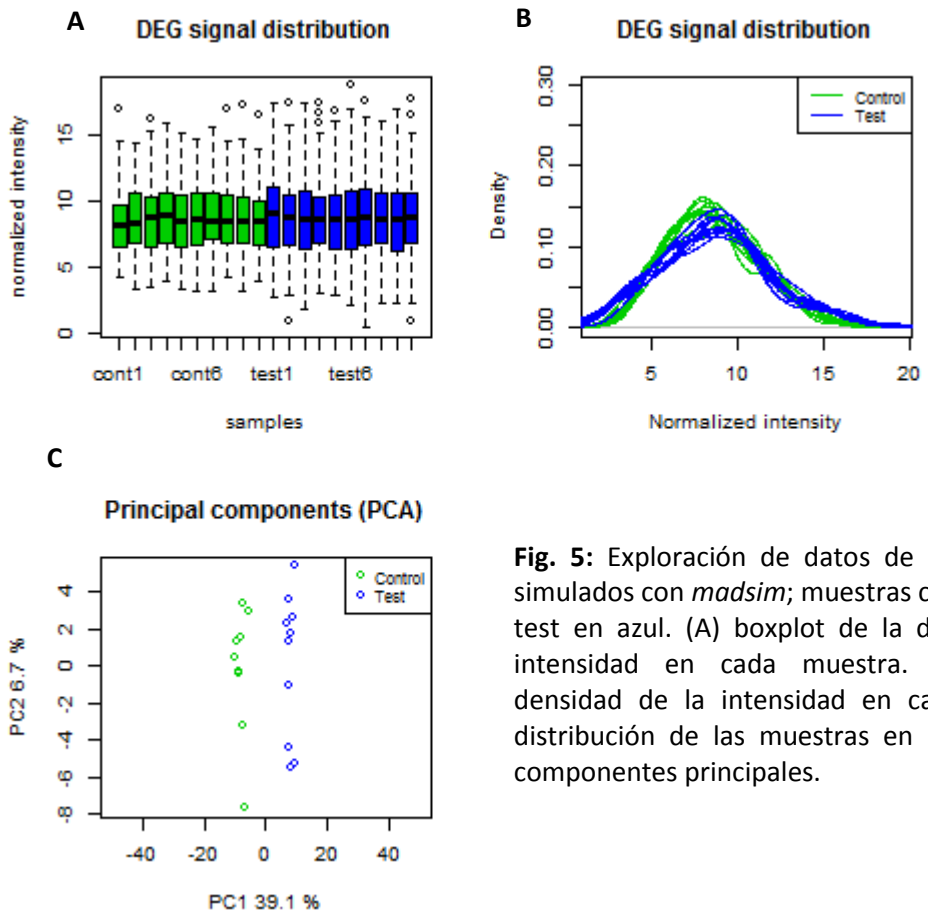

Fig. 5: Exploración de datos de expresión génica simulados con *madsim*; muestras control en verde y test en azul. (A) boxplot de la distribución de la intensidad en cada muestra. (B) Gráfico de densidad de la intensidad en cada muestra. (C) distribución de las muestras en las dos primeras componentes principales.

A continuación, se han generado las asociaciones y valores de expresión de miRNA siguiendo el procedimiento descrito en materiales y métodos. En los casos en que un miRNA se asocia con un único gen, los coeficientes de correlación resultan muy altos (alrededor de -0.9). Sin embargo, cuando se crean asociaciones con múltiples genes, los valores de correlación que se consiguen son más bajos. Esto hace más complicado que las correlaciones reflejen de forma clara las asociaciones. Se ha intentado controlar éste problema poniendo restricciones en las correlaciones que se aceptan. Por ejemplo, si un miRNA no está asociado con ningún gen, no se permite que ninguna de sus correlaciones sea $r < -0.3$. Estas restricciones generan gran cantidad de iteraciones y hacen que el procedimiento no sea eficiente. Por ello, sólo se han conseguido generar datasets pequeños (hasta 30 miRNA) y basados en redes dispersas (con un 5% de asociaciones).

Primero se ha generado un dataset de expresión con 10 miRNA y 20 genes (seleccionados de los 100 genes simulados inicialmente) y 20 muestras, con un 5% de conexiones (10 conexiones). El gráfico de asociaciones conocidas crea una red dispersa, bien replicada por el paquete *GGMselect*. El método glasso crea más asociaciones de las existentes, mientras que *GeneNet* también genera una red dispersa, aunque con varias conexiones incorrectas (Fig. 6). La tabla 5 refleja estos resultados en forma cuantitativa, mostrando la mayor sensibilidad del método glasso, junto con su especificidad decreciente conforme disminuye lambda. Por su parte, el método de *GGMselect* sólo ha generado dos falsos negativos, mientras que todas las asociaciones detectadas han sido correctas.

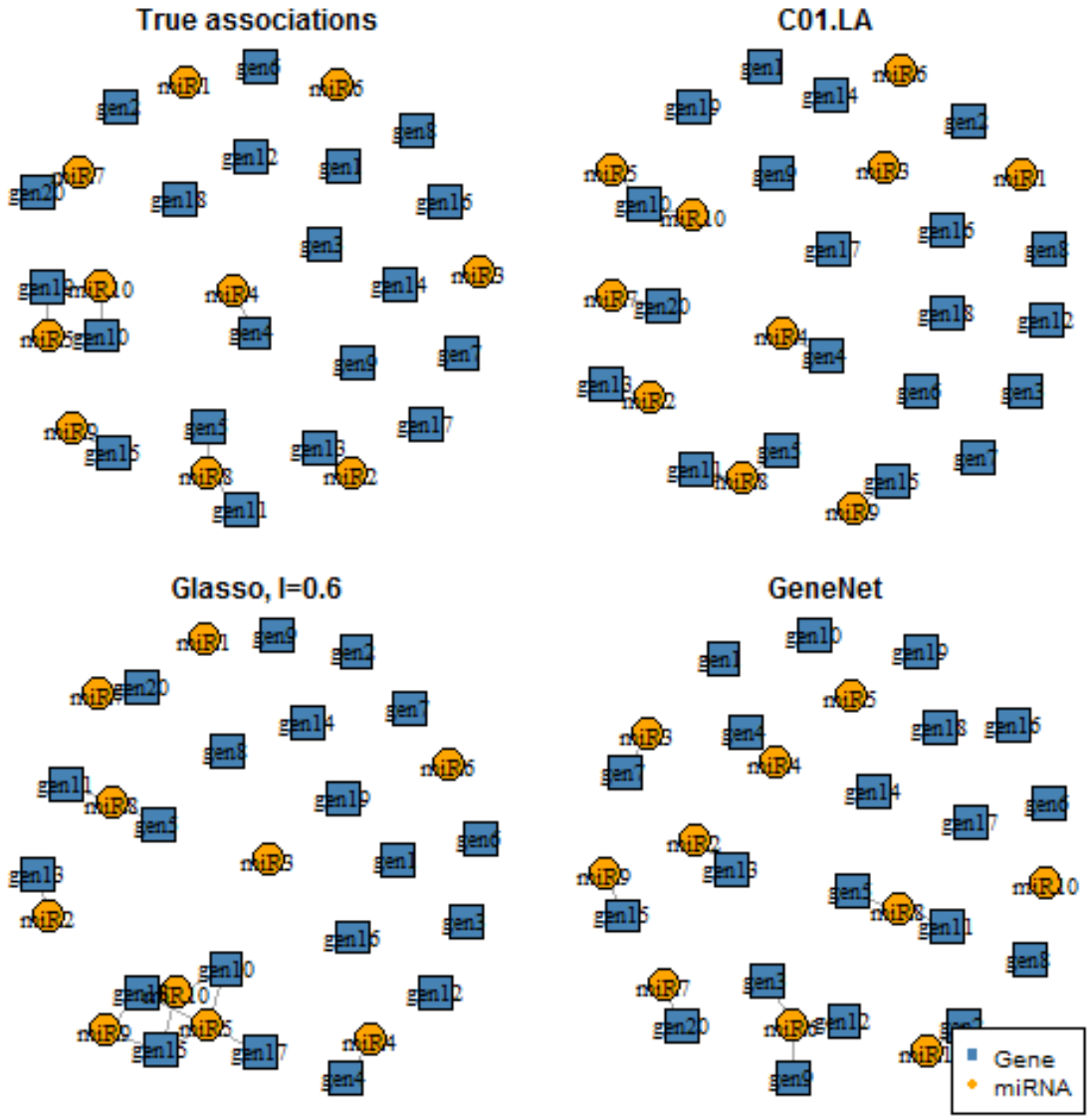

Fig. 6: representación gráfica de las asociaciones entre 10 miRNA y 20 genes simulados a partir de sus niveles de expresión. De izda. a dcha. y de arriba abajo, las redes se han creado a partir de la matriz de asociaciones conocidas, con la combinación de los métodos C01.LA del paquete *GGMselect*, con el método glasso del paquete *Huge* (lambda = 0.68) y con el paquete *GeneNet*.

	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
C01.LA	8	8	0	90	2	80	100.000
Glasso, l = 0.68	14	8	6	84	2	80	93.333
Glasso, l = 0.46	45	9	36	54	1	90	60.000
Glasso, l = 0.32	59	10	49	41	0	100	45.556
Glasso, l = 0.22	74	10	64	26	0	100	28.889
Glasso, l = 0.15	87	10	77	13	0	100	14.444
Glasso, l = 0.1	96	10	86	4	0	100	4.444
GeneNet	11	6	5	85	4	60	94.444

Tabla 5: evaluación del rendimiento de diversos métodos GGM para reconstruir la estructura de un gráfico con 20 genes, 10 miRNAs y 10 asociaciones entre ellos.

A continuación se ha generado un dataset más grande, también con 20 muestras y ahora con 30 miRNA y 50 genes y un 5% de conexiones (75 conexiones). En este caso, se observan muchas más dificultades a la hora de reconstruir la red de asociaciones por medio de GGM. El gráfico de asociaciones verdaderas muestra una red con muchos elementos interconectados y asociaciones de muchos miRNAs con múltiples genes. Por su parte, los modelos GGM han creado redes mucho más dispersas, principalmente con asociaciones únicas miRNA – gen (Fig. 7). La red de *GeneNet* no se ha representado ya que sólo ha generado una asociación, que ha resultado incorrecta. En la tabla 6 se puede observar la alta especificidad que conserva el método C01.LA, que sin embargo no consigue generar suficientes conexiones; parece que sólo considera significativas correlaciones muy altas. Por su parte, glasso consigue mayor sensibilidad, a costa de generar también gran cantidad de falsos positivos. La especificidad se mantiene alta, pero hay que tener en cuenta que existe una cantidad mucho mayor de nodos no conectados que de nodos conectados, con lo que puede haber muchos falsos positivos sin que disminuya mucho la especificidad. En concreto, en glasso vemos que con lambda = 0.46 o 0.32, se crean más del doble de conexiones incorrectas que correctas y sin embargo, los valores de sensibilidad y especificidad son de los más equilibrados.

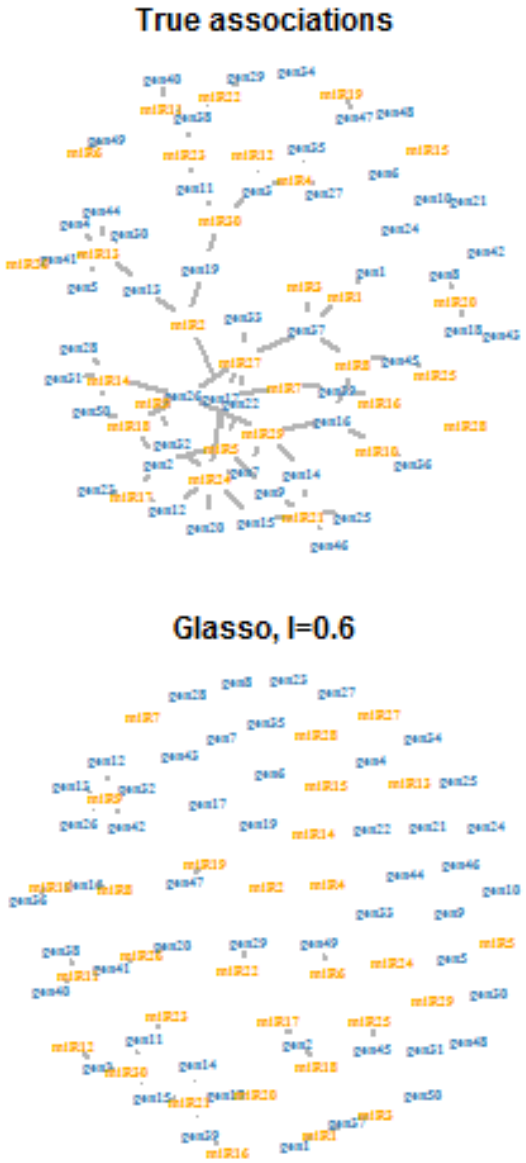

Fig. 7: representación gráfica de las asociaciones entre 30 miRNA y 50 genes simulados a partir de sus niveles de expresión. De izda. a dcha. y de arriba abajo, las redes se han creado a partir de la matriz de asociaciones conocidas, con la combinación de los métodos C01.LA del paquete *GGMselect* y con el método glasso del paquete *Huge* (lambda = 0.68). Genes en azul y miRNA en amarillo

	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
C01.LA	21	21	0	825	54	28.000	100.000
Glasso, l = 0.68	33	25	8	817	50	33.333	99.030
Glasso, l = 0.46	182	48	134	691	27	64.000	83.758
Glasso, l = 0.32	255	57	198	627	18	76.000	76.000
Glasso, l = 0.22	318	63	255	570	12	84.000	69.091
Glasso, l = 0.15	401	66	335	490	9	88.000	59.394
Glasso, l = 0.1	492	71	421	404	4	94.667	48.970
GeneNet	1	0	1	824	75	0.000	99.879

Tabla 6: evaluación del rendimiento de diversos métodos GGM para reconstruir la estructura de un gráfico con 50 genes, 30 miRNAs y 75 asociaciones entre ellos.

4.4 Inferencia de asociaciones miRNA – mRNA a partir de datos reales de expresión

Para este apartado se ha utilizado una matriz de expresión normalizada con los niveles de 1154 mRNA y 35 miRNA con expresión diferencial entre líneas celulares tumorales de tipo epitelial y mesenquimal (11 y 36 muestras, respectivamente). Un primer análisis exploratorio ha revelado la adecuada distribución de los datos y la clara diferenciación entre los dos grupos (Fig. 8).

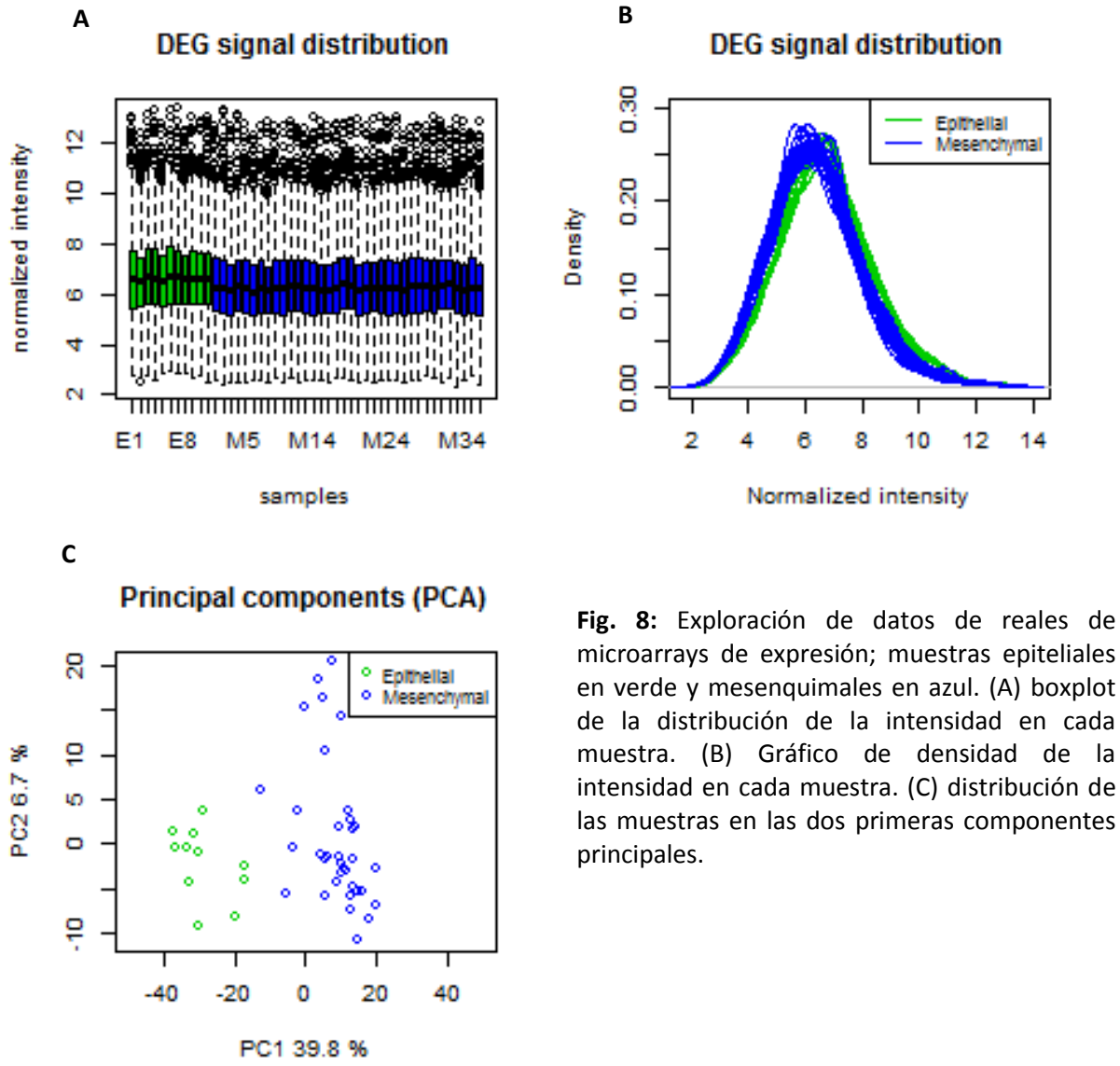

Fig. 8: Exploración de datos de reales de microarrays de expresión; muestras epiteliales en verde y mesenquimales en azul. (A) boxplot de la distribución de la intensidad en cada muestra. (B) Gráfico de densidad de la intensidad en cada muestra. (C) distribución de las muestras en las dos primeras componentes principales.

Tras los resultados obtenidos en los apartados de simulaciones, se ha decidido utilizar el método *glasso* de *huge*, con el objetivo de obtener un número relativamente alto de asociaciones y comprobar en qué medida es posible validarlas. Además, en pruebas preliminares se han aplicado los métodos de *GGMselect* y prácticamente no se han obtenido resultados.

Primero se han creado modelos a partir de las 1189 variables y 47 muestras. Una vez más, se han utilizado diferentes valores para lambda y se ha comprobado la gran variabilidad de resultados en función de estos. En ningún caso se ha conseguido detectar un gran número de asociaciones validadas y se ha generado gran cantidad de falsos positivos. Hay que tener en cuenta que la validación en este apartado es limitada, ya que procede de la recopilación de datos sobre dianas conocidas de miRNAs y resultados experimentales.

Lambda	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
0.771	183	9	174	39792	415	2.123	99.565
0.597	871	17	854	39112	407	4.009	97.863
0.462	1061	17	1044	38922	407	4.009	97.388
0.358	861	16	845	39121	408	3.774	97.886
0.277	795	15	780	39186	409	3.538	98.048
0.214	1006	13	993	38973	411	3.066	97.515
0.166	1302	15	1287	38679	409	3.538	96.780
0.129	1639	18	1621	38345	406	4.245	95.944
0.100	1980	22	1958	38008	402	5.189	95.101

Tabla 7: evaluación de resultados de modelos GGM creados por glasso y distintos valores de lambda.

Por otro lado, se ha analizado si una reducción del número de variables ayuda a la mejora de los resultados. Se han filtrado miRNA y mRNA por sus coeficientes de correlación de Pearson con el resto de variables, eliminando todos aquellos sin al menos una correlación negativa $r < -0.1$. Este filtraje ha dado lugar a la conservación de los 35 miRNA y a la reducción del número de mRNA a 326. De esta forma, se ha obtenido un dataset de 361 variables y 47 muestras. Además, el dataset de validación también se ha reducido para incluir sólo las asociaciones presentes en la nueva matriz de expresión, que han resultado ser 187.

El número de asociaciones validadas detectadas ha disminuido ligeramente, pero al tener muchas menos asociaciones validadas, han disminuido en mayor proporción los falsos negativos. Con ello ha aumentado la sensibilidad, pero se ha obtenido una especificidad menor.

Lambda	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
0.771	14	3	11	11212	184	1.604	99.902
0.597	238	12	226	10997	175	6.417	97.986
0.462	624	15	609	10614	172	8.021	94.574
0.358	565	12	553	10670	175	6.417	95.073
0.277	501	10	491	10732	177	5.348	95.625
0.214	540	9	531	10692	178	4.813	95.269
0.166	694	12	682	10541	175	6.417	93.923
0.129	931	16	915	10308	171	8.556	91.847
0.100	1176	21	1155	10068	166	11.230	89.709

Tabla 8: evaluación de resultados de modelos GGM creados por glasso y distintos valores de lambda, tras filtrar variables por sus coeficientes de correlación.

4.5 Paquete miRLab

Utilizando los métodos implementados en el paquete miRLab, se han vuelto a inferir asociaciones miRNA – gen en los mismos datos del apartado anterior. La validación también se ha llevado a cabo con los mismos datos que en el caso anterior.

El método de evaluación de resultados disponible en *miRLab* evalúa las X asociaciones con más puntuación encontradas para cada miRNA e informa de cuántas se han encontrado en cada uno de los dos datasets de validación (Fig. 8). Por ejemplo, en el top100 (35x100 = 3500 asociaciones), 67 de ellas están validadas al aplicar lasso y 60 al utilizar el método combinado. Con los los dos métodos se obtienen resultados similares, algo mejores con lasso, a pesar de que la combinación Pearson + IDA + Lasso es el método recomendado por los autores. Esto nos indica la variación de los resultados en función del tipo de datos con que se trabaja. Las conexiones validadas que se detectan aumentan conforme se evalúan más asociaciones, lo que conlleva un incremento de sensibilidad, a costa de un gran número de falsos positivos (Tabla 9). Se necesita evaluar 17.000 asociaciones para conseguir una sensibilidad del 73 %, pero esto da lugar a una cantidad enorme de falsos positivos, lo que conduce a un valor predictivo ínfimo: 312/(312 + 17.188) = 1,78%.

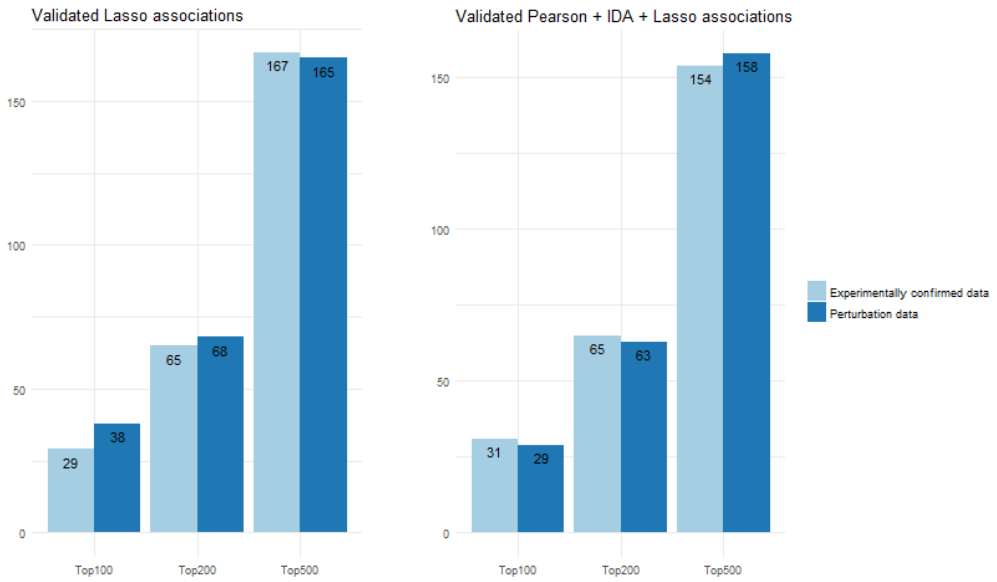

Fig. 8: nº de resultados validados el top 100, 200 y 500 de asociaciones de cada miRNA en los modelos creados con Lasso (izda.) y Pearson + Lasso + IDA (dcha.)

	Associations	True Positives	False Positives	True Negatives	False Negatives	Sensitivity	Specificity
Top10	350	7	343	39623	417	1.651	99.142
Top30	1050	19	1031	38935	405	4.481	97.420
Top100	3500	60	3440	36526	364	14.151	91.393
Top200	7000	128	6872	33094	296	30.189	82.805
Top500	17500	312	17188	22778	112	73.585	56.993

Tabla 9: evaluación de resultados de la inferencia de asociaciones mediante la combinación de correlaciones de Pearson, Lasso e IDA

5 Discusión

El objetivo de este trabajo era la creación de redes de asociaciones directas con GGM y su aplicación en datos ómicos, para descubrir interacciones entre miRNA y sus mRNA diana. Estudios previos habían explorado la aplicación de este tipo de métodos con el mismo objetivo[10,26]. Aquí, nos hemos centrado en la utilización de algoritmos GGM que estiman la matriz de precisión y correlaciones parciales y generan matrices de adyacencia con la estructura de la red de asociaciones.

En una prueba preliminar, se han utilizado datos con n > p y se ha observado la diferencia en los gráficos surgidos a partir de correlaciones clásicas y de correlaciones parciales, siendo los segundos más dispersos y con una estructura dividida en clusters que pueden ser interpretados. Además, se han obtenido buenos resultados con la aplicación de GGM; se ha conseguido reproducir la red de correlaciones parciales casi por completo. Sería interesante explorar las posibilidades de estos métodos en datos de altas dimensiones, pero que conservan un mayor número de muestras que de variables. Estudios de metabolómica, por ejemplo, han obtenido resultados positivos aplicando GGM en este tipo de condiciones[11].

El aumento del número de variables hasta n < p ha resultado en una caída enorme del rendimiento. Para empezar, los métodos de *GGMselect* prácticamente no han encontrado asociaciones significativas en datos normales simulados con n = 20 y d = 500 (tampoco en simulaciones con menos variables y más muestras, como muestra la tabla 2). La disminución de la penalización no ha mejorado mucho los resultados. Por otro lado, *GGMselect* ha vuelto a dar buenos resultados – además de con los datos clínicos – al reconstruir la red de asociaciones de datos genómicos simulados, con pocas variables. Estos resultados parecen indicar que se trata de métodos muy conservadores, que sólo detectan asociaciones en casos en que las estimaciones de correlaciones parciales sean muy significativas.

Por otra parte, de los métodos implementados por el paquete *huge*, el graphical lasso ha destacado como el más interesante. Aún y todo, en ningún momento ha llegado a detectar una cantidad alta de asociaciones validadas y los resultados han mostrado gran variabilidad en función del parámetro de penalización, lambda. Hemos observado que el número de falsos positivos aumenta progresivamente con la disminución de lambda. Al aplicar glasso a datos reales de microarrays, se ha conseguido una sensibilidad muy baja al contrastar los resultados con información de bases de datos. El filtraje de variables por coeficientes de correlación ha mejorado algo el conjunto de los resultados, dando una mayor sensibilidad y menos falsos positivos.

El filtraje se ha realizado bajo la suposición de que asociaciones miRNA – mRNA deben corresponderse con correlaciones negativas, debido a la regulación negativa que los primeros ejercen sobre los segundos. Por ello, se han seleccionado elementos con r < -0,1. Sin embargo, al crear una nueva matriz de expresión tras el filtraje, el número de asociaciones validadas que se podían encontrar en los datos ha disminuido de 424 a 187. Este hecho sugiere que hay gran cantidad de asociaciones validadas que no se reflejan en valores negativos de correlación. Por un lado, hay que tener en cuenta que en diferentes condiciones, tipos celulares, tejidos… se evidencian diferentes asociaciones y se crean redes

de regulación distintas. Además, la variabilidad de los datos también complica el establecimiento de correlaciones claras.

El uso del paquete *miRLab*, creado por Le T.D. *et al*[26], con los mismos objetivos que este trabajo, ha generado resultados similares a los obtenidos con glasso sobre los mismos datos. Los métodos de *miRLab* calculan matrices numéricas (de correlaciones, coeficientes de regresión…) con las que se crean rankings de asociaciones. Los métodos de este trabajo, por su parte, se limitan a construir la matriz de adyacencia, a partir del cálculo de la significación de cada coeficiente, de forma que no es posible ordenar las asociaciones, sino sólo establecer si existen o no. De esta manera, en *miRLab* se pueden considerar todas las asociaciones que se deseen, ordenándolas de mayor a menor puntuación. Sin embargo, la evaluación de los resultados indica que no existe una mayor concentración de asociaciones validadas entre las de mayor puntuación, sino que estas se distribuyen a lo largo de toda la matriz, lo cual pone en cuestión la idoneidad de los resultados.

En conjunto, en este estudio se ha observado que los algoritmos GGM no están preparados para el análisis de datos de altas dimensiones y pocas muestras. Para que su aplicación sea útil, es necesario tender hacia condiciones n > p, ya sea utilizando un mayor número de muestras o mediante la selección de variables.

6. Conclusiones

La aplicación de Gaussian Graphical Models resulta útil en situaciones en que las muestras superan a las variables, ya que revelan redes de asociaciones directas, cercanas a gráficos causales, que ayudan a la interpretación de los datos. Sin embargo, hemos comprobado que los métodos implementados para aplicar estos modelos a situaciones con gran número de variables tienen un rendimiento muy bajo. Los algoritmos GGM que hemos probado aún no están preparados para su aplicación a datos ómicos.

7. Glosario

Datos ómicos: información obtenida del uso de técnicas que estudian de forma simultánea el conjunto de un tipo de moléculas biológicas, como la proteómica, genómica o metabolómica.

microRNA: RNA pequeño, de unos 22 nucleótidos, que no codifica para una proteína y que lleva a cabo un silenciamiento específico post-transcripcional de genes.

mRNA diana: RNA mensajero con una secuencia de unión para uno o varios miRNA, que es sujeto de degradación o inhibición traduccional por parte de éstos.

microarray de expresión génica: herramienta para el estudio de la expresión génica mediante la cuantificación de RNA marcado, unido por hibridación a un soporte sólido.

correlación parcial: correlación entre dos variables calculada teniendo en cuenta la correlación con el resto de variables.

Gaussian Graphical model: gráfico de asociaciones no dirigido, en el que las aristas representan dependencia condicional entre dos variables.

Dependencia condicional: interacción entre dos nodos de una red condicionada a la correlación con el resto de nodos

Matriz de adyacencia: matriz p x p simétrica de unos y ceros, en que los unos indican la presencia de conexión entre dos variables.

8. Bibliografía

1. Ángel Lugo-Trampe, K. del C. T.-M. MicroRNAs: reguladores clave de la expresión génica. *Med. Univ.* **11,** 187–192 (2009).
2. Lim, L. P., Glasner, M. E., Yekta, S., Burge, C. B. & Bartel, D. P. Vertebrate MicroRNA Genes. *Science (80-.).* **299,** 1540–1540 (2003).
3. Bartel, D. P. MicroRNAs: Genomics, Biogenesis, Mechanism, and Function. *Cell* **116,** 281–297 (2004).
4. Sárközy, M., Kahán, Z. & Csont, T. A myriad of roles of miR-25 in health and disease. *Oncotarget* **9,** 21580–21612 (2018).
5. Li, B., Fan, J. & Chen, N. A Novel Regulator of Type II Diabetes: MicroRNA-143. *Trends Endocrinol. Metab.* **29,** 380–388 (2018).
6. Ju, X. *et al.* DIFFERENTIAL microRNA EXPRESSION IN CHILDHOOD B-CELL PRECURSOR ACUTE LYMPHOBLASTIC LEUKEMIA. *Pediatr. Hematol. Oncol.* **26,** 1–10 (2009).
7. Liu, B., Li, J. & Cairns, M. J. Identifying miRNAs, targets and functions. *Brief. Bioinform.* **15,** 1–19 (2014).
8. Rajewsky, N. microRNA target predictions in animals. *Nat. Genet.* **38,** S8–S13 (2006).
9. John, B. *et al.* Human MicroRNA Targets. *PLoS Biol.* **2,** e363 (2004).
10. Lee, M. & Lee, H. J. DMirNet: Inferring direct microRNA-mRNA association networks. *BMC Syst. Biol.* **10,** (2016).
11. Krumsiek, J., Suhre, K., Illig, T., Adamski, J. & Theis, F. J. Gaussian graphical modeling reconstructs pathway reactions from high-throughput metabolomics data. *BMC Syst. Biol.* **5,** 21 (2011).
12. UB. *Graphical Gaussian Models (GGM) for high-dimensional data Nutrimetabolomics studies*.
13. Huisman, S. Penalised graphical models for combined mRNA and miRNA data. (2013).
14. Diabetes data. Available at: http://staff.pubhealth.ku.dk/~tag/Teaching/share/data/Diabetes.html. (Accessed: 4th June 2018)
15. Bouvier, A., Giraud, C., Huet, S. & Verzelen, N. GGMselect: R package for estimating Gaussian graphical models.
16. Wille, A. & Bühlmann, P. Low-Order Conditional Independence Graphs for Inferring Genetic Networks. *Stat. Appl. Genet. Mol. Biol.* **5,** Article1 (2006).
17. Meinshausen, N., Bühlmann, P. & Zürich, E. HIGH-DIMENSIONAL GRAPHS AND VARIABLE SELECTION WITH THE LASSO. *Ann. Stat.* **34,** 1436–1462 (2006).
18. Zou, H. The Adaptive Lasso and Its Oracle Properties. doi:10.1198/016214506000000735
19. Friedman, J., Hastie, T. & Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. (2007).
20. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9,** 559 (2008).
21. Schwender, H. *et al.* Reverse Engineering Genetic Networks using the GeneNet Package. *J. Am. Stat. Assoc. Bioinforma. Comput. Biol-ogy Solut. Using R Bioconductor* **65,** 1151–1160 (2006).
22. Shankavaram, U. T. *et al.* Transcript and protein expression profiles of the NCI-

60 cancer cell panel: an integromic microarray study. *Mol. Cancer Ther.* **6,** 820–832 (2007).
23. Sokilde, R. *et al.* Global microRNA Analysis of the NCI-60 Cancer Cell Panel. *Mol. Cancer Ther.* **10,** 375–384 (2011).
24. Le, T. D., Zhang, J., Liu, L. & Li, J. Ensemble Methods for MiRNA Target Prediction from Expression Data. *PLoS One* **10,** e0131627 (2015).
25. Li, Y., Goldenberg, A., Wong, K.-C. & Zhang, Z. A probabilistic approach to explore human miRNA targetome by integrating miRNA-overexpression data and sequence information. *Bioinformatics* **30,** 621–628 (2014).
26. Le, T. D., Zhang, J., Liu, L., Liu, H. & Li, J. MiRLAB: An R based dry lab for exploring miRNA-mRNA regulatory relationships. *PLoS One* **10,** 1–15 (2015).

9. Material suplementario

ggm_datos_pequeños.zip

- **Diabetes.csv**: datos clínicos
- **ggm_datos_pequeños.rmd**: código para aplicar GGM en datos clínicos

ggm_datos_normales_simulados.rmd: código para aplicar GGM a datos normales simulados.

ggm_datos_omicos_simulados: código para generar datos ómicos simulados y aplicar GGM

validacion.zip

- **Groundtruth-all.csv**: dataset de validación a partir de bases de datos
- **logFCimputed.rda*:*** dataset de validación con logFC obtenidos en experimentos de perturbación por transfección

ggm_datos_reales.zip

- **EMT_MCC_BRS_DEG_S1.xlxs:** matriz de expresión de mRNA y miRNA en muestras tumorales epiteliales y mesenquimales
- **ggm_datos_reales.rmd:** código para aplicar ggm a datos de expresión de microarrays reales