Inferencia de Redes de Asociación de Genes Guiada por Similitud Semántica

José Luis Galván-Rojas, Isabel A Nepomuceno-Chamorro, Juan A. Nepomuceno, and José C. Riquelme-Santos

Dpto. Lenguajes y Sistemas Informáticos,
Universidad de Sevilla, Spain
josgalroj@alum.us.es
inepomuceno@us.es

Resumen En este trabajo se propone el uso de conocimiento a priori como heurística en métodos de inferencia de redes de genes a partir de datos de expresión obtenidos con tecnología de Microarray. Utilizamos Gene Ontology [15] como fuente de conocimiento a priori. Este repositorio se nutre de la información de anotaciones de relaciones en el material genético basadas en evidencias científicas. En este trabajo se propone el uso de medidas de similitud semántica, de manera más concreta la medida SimGIC en un método de inferencia basado en regresión. La propuesta se compara frente al mismo método sin integración de información y frente a otros métodos clásicos obteniendo mejoras y resultados comparables en otros casos.

Keywords: Redes de Asociación de Genes, Ontología, Similitud Semántica de Genes

1. Introducción

En la actualidad, la producción de información susceptible de analizar crece exponencialmente. Gracias a los avances tecnológicos, disponemos de medios suficientes para poder manejar esta información. Por otra parte, las ciencias Biológicas y Biomédicas no han sido ajenas a estos avances y gracias a tecnologías como la de Microarray[1] podemos analizar el nivel de expresión de miles de genes en simultáneo y así obtener conclusiones sobre información biológica, tarea que hubiera sido imposible sin estos avances científicos.

Este crecimiento exponencial de información biológica hace necesario establecer un vocabulario controlado para la información genética y nacen proyectos como Gene Ontology (GO) [15] que definen una estructura de información biológica en forma de ontologías que están en constante actualización y que se nutren de anotaciones de los productos genéticos. Anotaciones realizadas por

[1] Microarray: Chip que permite el análisis de fragmentos de ADN. Se conforman en forma de matriz y permiten obtener un nivel de expresión de ciertos genes bajo un conjunto de condiciones específicas.

los investigadores y basadas en evidencias que son codificadas manteniendo una traza con la información de la procedencia y fundamentos de dichas anotaciones. Esta información puede ser utilizada para caracterizar la inferencia de redes de asociación de genes mediante la aplicación de medidas de similitud semántica. A su vez estas medidas de similitud semántica pueden servir para la reducción del espacio de búsqueda en algoritmos que tienen como objetivo la obtención de redes de asociación de genes.

Las redes de asociación de genes son una forma de representación de la interacción que se produce entre pares de genes, ya que estos no son elementos disociados en su ciclo de vida, sino que colaboran de forma conjunta formando interacciones denominadas rutas bioquímicas o pathways [1]. Estas interacciones o asociaciones contienen gran cantidad de información biológica y pueden ser inferidas a partir de datos de microarray utilizando técnicas de Minería de Datos.

Existen diferentes técnicas para la extracción de estas redes de asociación pero en su mayoría exploran espacios demasiado amplios, por lo que sin la aplicación de heurísticas, la precisión en la obtención de resultados con alto interés biológico desciende.

En este trabajo, implementamos una metodología para la obtención de Redes de asociación de genes aplicando una heurística basada en medidas de similitud semántica entre genes denominada simGIC [2] sobre el algoritmo RegNet [3]. Este algoritmo a diferencia de los modelos basados en correlación que analizan la similitud bajo un conjunto completo de condiciones se basa en dividir el espacio de búsqueda de manera iterativa. Junto con la aplicación de simGIC, el método dirige en la búsqueda de soluciones descartando aquellas relaciones que no superan un umbral de similitud semántica considerado óptimo.

2. Trabajos Relacionados

Diferenciaremos dos tipos de trabajos relacionados. Por un lado abordaremos los trabajos relacionados con la inferencia de redes de asociación de genes y por otro lado abordaremos los trabajos relacionados con las medidas de similitud semántica entre genes.

2.1. Inferencia de Redes de Asociación de Genes

Los métodos utilizados en la actualidad para la inferencia de redes de asociación de genes son diversos. Destacamos los siguientes:

Métodos basados en similitud: Estos métodos buscan pares de genes que tienen un nivel de expresión similar bajo un conjunto de condiciones experimentales. También denominados redes de co-expresión o redes de dependencia y se basan en la utilización de técnicas estadísticas para calcular la similitud bajo el conjunto de condiciones, ejemplo de estas técnicas son las basadas en medidas de correlación, correlación parcial o información mútua [4,5,6]. Otros métodos

se basan en algoritmos de Clustering como por ejemplo [7]. Estas propuestas realizan agrupaciones, clasificando las variables en grupos según una medida de distancia. Estos métodos están motivados por una simple idea válida en genómica funcional, se dice que siguen una heurística denominada *guilt-by-association* [4] basadas en el supuesto de que genes con un mismo perfil de expresión siguen el mismo régimen regulador y pertenecen a un mismo proceso biológico.

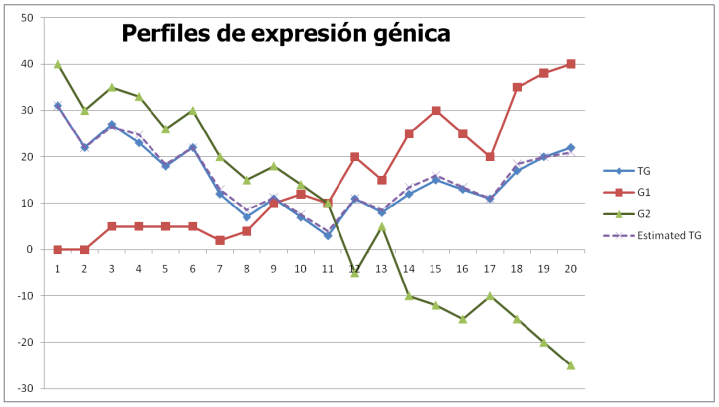
Métodos basados en Redes Bayesianas: Son métodos probabilistas y entre ellos destaca el trabajo realizado por Friedman et al. [10]. Este se basa en el uso del algoritmo *Sparse Candidate* que selecciona de forma iterativa un conjunto de candidatos influyentes en cada gen con el objetivo de maximizar la función de evaluación. Otros ejemplos de métodos bayesianos son [12] y [11], el último utiliza un algoritmo de EM y árboles de regresión para construir grafos a maximizar con puntuación bayesiana.

Métodos basados en árboles: Uno de los trabajos de referencia para este tipo de metodologías es el trabajo realizado por Soinov et al. [8] basado en árboles de decisión. En este trabajo se aplica el algoritmo C4.5 [9] desarollado por Ross Quinlan para la construcción de árboles de decisión. El algoritmo C4.5 se basa en el concepto de entropía de información y cómo esta medida se ajusta entre diferentes particiones del espacio de búsqueda.

También destaca en este terreno la metodología RegNet [3]. RegNet se basa en árboles modelo o de regresión, es decir, en la detección de similitud lineal entre pares de genes sobre un subconjunto de condiciones y no sobre un conjunto completo a diferencia de otros métodos. En la Figura 1 de [13] se puede observar que el gen objetivo denominado TG (target gene) y el gen denominado G1 no están correlacionados, pero tienen una dependencia lineal para el subconjunto de muestras o condiciones que cumplen que el valor de expresión del G1 es mayor a 10. Este método supone la existencia de dependencias lineales entre un gen objetivo y un subconjunto de genes dividiendo el espacio de búsqueda en subespacios más pequeños. Este método trata cada gen del conjunto de datos de Microarray como gen objetivo de manera iterativa.

RegNet consta de una serie de fases. La primera de ellas consiste en la construcción de árboles basados en el algoritmo M5', para ello recorre de forma iterativa cada gen, estableciendo en cada iteración un gen objetivo para la construcción del árbol en el que los nodos contienen un modelo lineal que aproxima el valor del gen objetivo para un subespacio de condiciones o muestras. El conjunto de árboles M5' generados lo denominaremos bosque. El segundo paso consiste en la poda del bosque y extracción de dependencias. Para la poda se establece como valor umbral el error relativo, de manera que aquel árbol que no supere dicho umbral se descarta. En este segundo paso también se produce la extracción de dependencias determinando como hipótesis de asociación la relación entre el gen objetivo y cada uno de los genes involucrados en los modelos lineales. El tercer paso consiste en la construcción de un grafo de asociación a partir de las hipóte-

TG	G_1	G_2
31	0	40
22	0	30
27	5	35
23	5	33
18	5	26
22	5	30
12	2	20
7	4	15
11	10	18
7	12	14
3	10	10
11	20	-5
8	15	5
12	25	-10
15	30	-12
13	25	-15
11	20	-10
17	35	-15
20	38	-20
22	40	-25



Corr(TG,G1) = -0.09

Corr(TG,G2) = 0.35

IF G1<=10 AND G2>10 then Estimado-TG = 0.9 * G2 -5

IF G1>10 then Estimado-TG = 0.5 * G1 + 1

Figura 1. *En la figura se muestra un hipotético ejemplo donde la correlación entre el gen objetivo y otros dos genes es débil pero sin embargo podemos observar una dependencia local fuerte.* [13]

sis aplicando el método estadístico para la eliminación de Falsos descubrimientos de Benjamini-Yekutieli [14].

2.2. Medidas de similitud semántica

Gene Ontology (GO) es una ontología en donde se anotan los genes según su funcionalidad biológica. Esta ontología constituye un repositorio público de información muy utilizado en el campo de la Bioinformática y estructura su información en tres ramas [15] que contienen términos que describen los productos genéticos y sus asociaciones. Estas ramas son diferentes características de la biología celular:

Cellular Component: Se refiere al espacio celular donde se encuentra el producto genético.

Biological Process: Se corresponde con transformaciones químicas o físicas realizados por uno o más conjuntos organizados.

Molecular Function: Son actividades que ocurren a nivel molecular como pueden ser la actividad catalítica o la actividad de unión.

Las ontologías en GO [15] se organizan en forma de Grafo Acíclico Dirigido (GAD) donde los términos son vértices y las relaciones entre términos se corresponden con las aristas. Las asociaciones se basan en evidencias que son

tipificadas y que permiten identificar el tipo de interacción que se produce entre los términos.

Las medidas de similitud semántica nos permiten obtener valores numéricos de la cercanía semántica entre términos de una ontología. Existen diversas métricas para el cálculo de esta similitud entre términos basadas en ontologías como GO y que podemos clasificar en tres categorías: métodos basados en las relaciones de la ontología y sus tipos; métodos basados en los nodos y su información; y métodos híbridos que se basan tanto en las relaciones como en los nodos de los GAD. Tal como muestra Pesquita et al. [2], los métodos híbridos obtienen buenos resultados siendo la medida simGIC la que mejores resultados obtiene de forma general.

La medida simGIC se calcula como sigue:

$$simGIC(A,B) = \frac{\sum_{t \in \{GO(A) \cap GO(B))\}} IC(t)}{\sum_{t \in \{GO(A) \cup GO(B))\}} IC(t)} \quad (1)$$

siendo A y B los genes para los que se calcula la medida de similitud, IC[2] es el contenido de información, GO(A) son los términos GO asociados al gen A, GO(B) son los términos GO asociados al gen B, $t \in \{GO(A) \cap GO(B))\}$ y $t \in \{GO(A) \cup GO(B))\}$ son los términos resultado de la intersección o la unión de los términos GO del gen A y del gen B.

La herramienta *GOssTo* [17] (the Gene Ontology Semantic Similarity Tool) nos permite calcular diferentes medidas de similitud entre conjunto de genes a partir de GO. GossTo ofrece 6 medidas de similitud semántica entre ellas simGIC. Esta herramienta puede ser utilizada desde línea de comandos o también puede integrarse con otras herramientas o programas ya que puede ser utilizada fácilmente a través de la API que proporciona. Una de las ventajas de GOssTo es la usabilidad del software ya que se puede integrar en ella nuevas medidas de similitud utilizando la API que proporcionan.

3. Metodología

Proponemos la integración de información a priori en un método para la inferencia de redes de asociación de genes. Ampliamos la metodología RegNet incorporando en el proceso información a priori, basada en la medida de similitud semántica simGIC. Integramos la herramienta GossTo en la metodología RegNet.

El algoritmo propuesto se divide en 4 etapas diferencias:

Etapa 1. Construcción y poda del Bosque M5': En esta etapa se itera sobre cada gen del conjunto de entrada. En cada iteración, un solo gen se establece como

[2] IC (Information Content): El contenido de información es una medida de cuan específico e informativo es un término determinado. El IC de un término c se cuantifica como la probabilidad logarítmica negativa $-\log p(c)$, dónde $p(c)$ es la probabilidad de ocurrencia de c en un corpus especifico tal como la base de datos del conocimiento Uniprot siendo estimado normalmente por frecuencia de anotación. [18]

atributo objetivo del conjunto y se construye un árbol modelo mediante el algoritmo M5' propuesto por Witten, I. y Frank, E. (2005) [19] implementado por herramientas como la librería Weka que utilizamos de soporte para la aplicación del algoritmo M5'. Cada árbol generado se evalua determinando si el error relativo es superior a un valor umbral ϵ en cuyo caso el árbol se poda o elimina y no pertenece al bosque de resultados.

Etapa 2. Extracción de dependencias: Durante esta etapa se generan las relaciones de dependencia entre pares de genes, para ello el algoritmo recorre el bosque de árboles M5'. Para cada árbol se extrae de los nodos hoja el modelo lineal generado con la asunción de que existen diversos genes asociados con una función biológica que influenciará las relaciones de co-expresión de un gen [13]. Los modelos lineales generados siguen el siguiente esquema:

$$ML : g_x = \sum_i \lambda_i g_{y_i} \tag{2}$$

Etapa 3. Cálculo de la Similitud Semántica: Esta etapa constituye la fase de integración de información a priori basado en el cálculo de la similitud semántica simGIC [2]. En esta etapa se hace uso de GossTo [17] para generar un valor de similitud entre cada asociación de par de genes, posteriormente se comprueba si el valor determinando de similitud semántica obtenido para una relación está por encima de un valor umbral ϵ en cuyo caso se afianza la relación entre el par de genes propuesto ya que hay evidencia de cierta relación biológica. En caso de que la relación no supere el umbral ϵ se descarta ya que se considera que no hay evidencia biológica de dicha asociación.

Etapa 4. Control de falsos descubrimientos: La última etapa consiste en la aplicación de un test estadístico que permite identificar la proporción de falsos descubrimiento para controlar el número de errores tipo I entre todos los descubrimientos realizados. El procedimiento estadístico llevado a cabo es el propuesto por Benjamini y Yekutieli, (2001) [14]. Este procedimiento consiste en un test estadístico que controla el ratio de falsos descubrimientos (FDR) para hipótesis $H_0^1, H_0^2, ..., H_0^m$. Sean $p_1, p_2, ..., p_m$ los p-values asociados a las m hipótesis nulas. Sean $p_{(1)} \leq p_{(2)} \leq ... \leq p_{(m)}$ la lista de p-values ordenados de manera creciente. El procedimiento define un valor k que se utiliza para rechazar las hipótesis $H_0^1, H_0^2, ..., H_0^k$ siguiendo la siguiente ecuación para el cálculo de k [13].

$$k = max\{i : p_{(i)} \frac{m}{i} \sum_{k=1}^{m} \frac{1}{k} \leq \alpha\} \tag{3}$$

La hipótesis no será rechazada si no existe un i que cumpla con la ecuación. Una relación entre par de genes identificada como una arista entre ellos no será válida si no siguen una relación monotómica significativa y dado el subespacio identificado por el nodo hoja del árbol modelo propuesto. Para testear esta monotonía se hace uso del estadístico *Tau de Kendall* que mide el valor de

significatividad a partir de un nº de instancias n y un valor del estadístico τ correspondiente al *Tau de Kendall*.

$$z = \frac{3\tau\sqrt{n(n-1)}}{\sqrt{2(2n+5)}} \tag{4}$$

El rechazo de la hipótesis nula al nivel de significancia α significa que la relación es válida y la arista como relación entre pares de genes pertenecerá al grafo resultante.

4. Resultados

Se han realizado diferentes estudios para comparar los resultados de la metodología propuesta con otros trabajos publicados en el campo de la inferencia de redes. En los siguientes apartados describiremos: las medidas de evaluación utilizadas en la comparativa; el diseño de la experimentación y los resultados de la comparativa.

4.1. Definición de medidas utilizadas para el análisis de rendimiento

Utilizamos como marco de comparación el establecido en el trabajo [22] y las medidas de evaluación que describimos a continuación:

Definición 1 *Exactitud de la Red: Se corresponde con la proporción de verdaderos positivos (TP y TN) sobre el número total de casos de ejemplo.*

$$Exactitud = \frac{TP + TN}{TP + FP + TN + FN} \tag{5}$$

donde:

- *TP: Es el número de asociaciones de par de genes obtenidos por el algoritmo y que también aparecen en la red utilizada como test.*
- *FP: Es el número de asociaciones de par de genes obtenidas por el algoritmo y no aparecen en la red utilizada como test.*
- *TN: Es el número de asociaciones de par de genes no obtenidas por el algoritmo y que no aparecen en la red utilizada como test.*
- *FN: Es el número de asociaciones de par de genes no obtenidas por el algoritmo y que aparecen en la red utilizada como test.*

Definición 2 *Precisión de la Red: Se define como la proporción de verdaderos positivos sobre todos los positivos (TP y FP) obtenidos por el algoritmo.*

$$Precisión = \frac{TP}{TP + FP} \tag{6}$$

Definición 3 *Sensibilidad de la Red: Se corresponde con la proporción de casos clasificados correctamente y se representa como la proporción de verdaderos positivos sobre los correctamente identificados.*

$$Sensibilidad = \frac{TP}{TP + FN} \tag{7}$$

Definición 4 *Especificidad de la Red: Se corresponde con la proporción de verdaderos negativos sobre los correctamente identificados.*

$$Especificidad = \frac{TN}{TN + FP} \tag{8}$$

Definición 5 *Número de aristas obtenidas: Se corresponde con el número de aristas obtenidas como salida por el algoritmo en cuestión, de las cuales una proporción habrá sido correctamente identificada.*

4.2. Conjunto de datos y diseño experimental

Se ha utilizado como conjunto de datos de entrada la matriz de expresión obtenida con tecnología de microarray de *Spellman [20] y Cho [21]* para el ciclo celular de la levadura. De este conjunto de datos a modo de ejemplo se han seleccionado un subconjunto de 20 genes bajo 24 condiciones experimentales considerando estos un conjunto de genes bien descritos, que codifican proteínas importantes para la regulación del ciclo celular de manera análoga al trabajo de Soinov [8].

Las redes de referencias o reales para el cálculo de las medidas de rendimiento de las redes resultados son YeastNet [25], GO [26] y Co-citacion [27]. Se utilizan estas redes como test de rendimiento ciego para comparar las redes generadas por la metodología contra las redes reales y así calcular los valores de exactitud, precisión, especificidad y sensibilidad. Para la medida de sensibilidad, la metodología propuesta y otras que se utilizarán en la comparativa obtendrán valores muy bajos ya que estas redes de referencia son reales y de gran tamaño en comparación con las redes que se obtendrán a partir del dataset de entrada.

En la Tabla 1 se señalan los valores de entrada para el cálculo de similitud semántica.

Cuadro 1. Parámetros de entrada de GossTo

Ontología GO:	Biological Process
Evidencias GO:	EXP, IDA, IPI, IMP, IGI, IEP, TAS, IC, ISS, ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD, RCA, NAS, ND, IEA
Relaciones GO:	is_a, part_of, regulates, positively_regulates, negatively_regulates, has_part
Fecha Obo File:	25/06/2013
Fecha Goa File:	27/04/2015

En la Tabla 2 se observan las diferentes configuraciones utilizadas para analizar el rendimiento de la metodología propuesta. Se han variado los parámetros de la siguiente forma: el valor umbral del error relativo del árbol modelo se ha variado de 10 en 10 valores de 0 a 100; el valor umbral de similitud se ha variado de una unidad en una unidad de 0.10 a 0.30; el nivel α de significancia se ha fijado en 0.05; finalmente, en las columnas de exactitud, precisión, especificidad y sensibilidad se muestran los valores en media de estas medidas. Se ha decidido fijar el valor umbral de similitud entre 0.1 y 0.3 calculando la media y moda del valor de similitud entre todos los genes del conjunto de datos y calculando la distribución de frecuencias por debajo de 0.3 se encuentra más del 80 % de las parejas de genes. Se ha tomado como red real o de referencia YeastNet y se observa en la tabla que los valores en media para los umbrales de similitud son similares, luego se puede fijar el umbral de similitud entre 0.1 y 0.3 ya que no se observa variación en los resultados.

Cuadro 2. Resultados Promedio de Experimentación

ID	θ	σ	α	Exactitud	Precisión	Sensibilidad	Especificidad
1	[0-100]	[0.10-0.15]	0.05	59.30 %	44.24 %	5.35 %	95.26 %
2	[0-100]	[0.16-0.20]	0.05	59.21 %	43.08 %	5.10 %	95.29 %
3	[0-100]	[0.21-0.25]	0.05	59.24 %	43.06 %	5.10 %	95.33 %
4	[0-100]	[0.26-0.30]	0.05	59.34 %	43.90 %	4.84 %	95.68 %

4.3. Resultados y comparativa

Los resultados de la metodología APrioriRegNet frente a RegNet se muestran en la primera y segunda columna de la Tabla 3. Se observa que la integración de información mejora los resultados en el caso de utilizar como redes reales YeastNet y Cocitation. En el caso de la red real GO se obtienen resultados comparables y algo menores que en el caso de utilizar la metodología sin integración de información a priori.

Además comparamos la aproximación de integración de información en la metodología de inferencia de redes frente a los resultados de otras aproximaciones a la inferencia de redes de manera análoga al marco de comparativa establecido en [22]. Utilizamos como aproximaciones de referencia: el método GarNet basado en reglas de asociación [22]; el método de optimización GRNCOP [23]; el método basado en árboles de decisión [8]; y el método basado en lógica de primer orden [24]. Los resultados se observan de la tercera a la séptima columna de la Tabla 3 y puede observarse que la aproximación utilizando información a priori mejora los resultados frente al resto de aproximaciones en el caso de tomar YeasteNet como red real, mientras que en el caso de tomarse como redes reales Co-Citation o GO los resultados son comparables al resto de aproximaciones.

Cuadro 3. Comparativa de algoritmos de Inferencia de Redes de Asociación de Genes

		APR*	RegNet	GarNet1	Garnet2	GRNCOP2	Soinov et al.	BLS**
YeastNet	Precisión	66,67 %	100,00 %	100,00 %	93,75 %	93,33 %	50,00 %	88,89 %
	Sensibilidad	5,26 %	7,14 %	20,40 %	15,31 %	14,29 %	3,06 %	8,19 %
	Especificidad	98,25 %	100,00 %	100,00 %	98,91 %	98,91 %	96,74 %	98,91 %
	Exactitud	**61,05 %**	**52,11 %**	**58,94 %**	**55,79 %**	**55,27 %**	**48,41 %**	**52,09 %**
Co-citacion	Precisión	100,00 %	100,00 %	95,00 %	93,75 %	93,33 %	50,00 %	88,89 %
	Sensibilidad	7,23 %	8,13 %	22,89 %	18,07 %	16,87 %	3,61 %	9,64 %
	Especificidad	100,00 %	100,00 %	99,07 %	99,07 %	99,07 %	97,20 %	99,07 %
	Exactitud	**59,47 %**	**58,42 %**	**65,79 %**	**63,68 %**	**63,16 %**	**56,29 %**	**60,00 %**
GO	Precisión	60,00 %	71,43 %	70,00 %	75,00 %	73,33 %	50,00 %	55,56 %
	Sensibilidad	3,49 %	5,81 %	16,28 %	13,95 %	12,79 %	3,49 %	5,81 %
	Especificidad	98,08 %	98,08 %	94,23 %	96,16 %	96,15 %	97,12 %	96,15 %
	Exactitud	**55,26 %**	**56,32 %**	**58,96 %**	**58,95 %**	**58,42 %**	**54,75 %**	**55,24 %**

$APR^* = AprioriRegnet$
$BLS^{**} = Bulashevska$

5. Conclusiones y Trabajos Futuros

En este trabajo se presenta la integración de información a priori en una metodología de inferencia de redes. Se ha integrado en ésta el cálculo de similitud semántica entre genes a partir de Gene Ontology, de manera más concreta utilizando la medida SimGIC. Se ha establecido dos comparativas, la primera frente a la metodología sin integración de información y la segunda frente a otros métodos de inferencia de redes. En esta comparativa se puede observar que el algoritmo propuesto mejora la exactitud de RegNet para dos de las redes de test, en el caso de YeastNet desciende su precisión. Este descenso posiblemente se deba a la ausencia de anotaciones de evidencia en GO que hace que la etapa de cálculo de similitud semántica se realice la poda de relaciones sin suficiente información biológica. Finalmente, en comparación con el resto de métodos se observa una mejora del valor de exactitud para la red de test YeastNet, destacando también que la comparación con la red de test Co-citacion Garnet1 y Garnet2 destaca en el valor de exactitud aunque reduciendo la precisión.

Concluimos tras el análisis de los resultados experimentales que el método propuesto es una solución válida para la inferencia de redes de asociación genética. Además, observamos que mediante la incorporación de una heurística con información a priori se puede guiar al algoritmo hacia la construcción de una red de asociación donde las relaciones identificadas disponen de una evidencia científica válida.

En trabajos futuros se extenderá el conjunto de entrada para aplicar la propuesta sobre un conjunto de datos con mayor información, que permita identificar un mayor número de relaciones. Además, se trabajará en la construcción de un plugin para la herramienta Cytoscape que permita integrar la ejecución de la propuesta dentro de esta herramienta de forma que la red de asociación se pueda obtener y analizar de forma visual.

Referencias

1. Zhou, X., Kao, M.C., Wong, W.H.: From the Cover: Transitive functional annotation by shortest-path analysis of gene expression data. Proceedings of the National Academy of Sciences 99(20), 12783–12788 (2002).
2. Pesquita,C., Faria, D., Bastos, H., Ferreira, A., O Falcão, A., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics 29(9), S4 (2008).
3. Nepomuceno-Chamorro, I.A., Aguilar-Ruiz, J.S., Riquelme, J.C.: Inferring gene regression networks with model trees. BMC Bioinformatics 11, 517–617 (2010).
4. Markowetz, F., Spang, R.: Inferring cellular networks-a review. BMC Bioinformatics 8, S5 (2007).
5. Fitch, A., Jones, M.: Shortest path analysis using partial correlations for classifying gene functions from gene expression data. Bioinformatics 25, 42–47 (2009).
6. Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Favera, R., Califano, A.: ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. BMC Bioinformatics 7(Suppl 1), S7 (2006).
7. Eisen, M.B., Spellman, P.T., Brown, P.O., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proceedings of the National Academy of Sciences of the United States of America 95, 14863-14868 (1998).
8. Soinov, L., Krestyaninova, M., Brazma, A.: Towards reconstruction of gene networks from expression data by supervised learning. Genome Biol 4, R6 (2003).
9. Quinlan, J.R.: C4.5: Programs for Machine Learning. Morgan Kaufmann, San Francisco, CA, USA, (1993).
10. Friedman, N., Linial, M., Nachman, I., Peter, D.: Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology 7, 601–620 (2000).
11. Segal, E., Shapira, M., Regev, A., Pe'er, D., Botstein, D., Koller, D., Friedman, N.: Module networks: identifying regulatory modules and their condition-specific regulators from gene expression data. Nature Genet 34, 166-176 (2003).
12. Steele, E., Tucker, A., 't Hoen, P.A., Schuemie, M.J.: Literature-based priors for gene regulatory networks. Bioinformatics 25(14), 1768–1774 (2009).
13. Nepomuceno Chamorro. I.A: Reconocimiento de Redes de Genes Mediante Regresión. Editorial Fenix, (2011).
14. Benjamini, Y., Yekutieli, D.: The control of the false discovery rate in multiple testing under dependency. Ann. Statist 29(4), 1165–1188 (2001).
15. Gene Ontology Consortium: Ontology Documentation.http://geneontology.org/page/documentation
16. Pesquita Catia, Faria Daniel, Falcão André O., Lord Phillip, Couto Francisco M. (2013): Section of the GO graph showing the three aspects (molecular function, biological process, and cellular component) and some of their descendant terms. Fig1.tif. PLOS Computational Biology. 10.1371/journal.pcbi.1000443.g001.
17. Caniza, H., Romero, A. E., Heron, S., Yang, H., Devoto, A., Frasca, M., Mesiti, M., Valentini, G., Paccanaro, A. : GOssTo: a user-friendly stand-alone and web tool for calculating semantic similarities on the Gene Ontology. Bioinformatics 30, 2235–2236 (2014).
18. Pesquita, C., Faria, D., Falcão, A.O., Lord, P., Couto, F.M.: Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5(7),: e1000443 (2009).
19. Witten, I.H., Frank, E., Hall, M.A.: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann.

20. Spellman, P., Sherlock, G., Zhang, M., Iyer,V., Anders, K., Eisen, M., Brown, P., Botstein, D., Futcher, B.: Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. Mol BiolCell 9, 3273-–3297 (1998).
21. Cho, R., Campbell, M., Winzeler, E., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T., Gabrielian, A., Landsman, D., Lockhart, D., Davis, R.: A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 2, 65-–73 (1998).
22. Martínez-Ballesteros, M., Nepomuceno-Chamorro, I.A., Riquelme, J.C.: Discovering gene association networks by multi-objective evolutionary quantitative association rules. Journal of Computer and System Sciences 80, 118–136 (2014).
23. Ponzoni, I., Azuaje, F.A., Glass, J.: Inferring adaptive regulation threshold and association rules from gene expression data through combinatorial optimization learning. IEEEACM Transactions on computational biology and bioinformatics 4, 624–634 (2007).
24. Bulashevska, S., Eils, R.: Inferring genetic regulatory logic from expression data. Bioinformatics 21, 2706–2713 (2005).
25. Lee, I., Li, Z., Marcotte, E.: An improved, bias-reduced probabilistic functional gene network of baker's yeast, Saccharomyces cerevisiae. PLoS ONE 2, e988 (2007).
26. Dwight, S., Harris, M., Dolinski, K., Ball, C., Binkley, G., Christie, K., Issel-Tarver, L.F., Schroeder, G., Sherlock, A., Sethuraman, S., Weng, D., Botstein, J.: Saccharomyces genome database (sgd) provides secondary gene annotation using the gene ontology (go). Nucleic Acids Res 30, 69-–72 (2002).
27. Lee, I., Date, S., Adai, A., Marcotte, E.: A Probabilistic Functional Network of Yeast Genes: Science 306, 1555–1558 (2004).
28. Nepomuceno, J.A., Troncoso, A., Nepomuceno-Chamorro, I.A., Aguilar-Ruiz, J.S.: Integrating biological knowledge based on functional annotations for biclustering of gene expression data. Comput Methods Programs Biomed. 119(3), 163–80 (2015).