Procesamiento de señales en grafos para medir distancias en la ontología de genes

Tiago López[1], Leandro E. Di Persia[2], y Diego H. Milone[2].
{tlopez,ldipersia,dmilone}@sinc.unl.edu.ar

[1] Universidad Nacional del Litoral, Facultad de Ingeniería y Ciencias Hídricas, Ingeniería en Informática, (3000) Santa Fe, Argentina, Tel: +54 (0342) 4571110
[2] sinc(i), Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, FICH/UNL-CONICET, (3000) Santa Fe, Argentina, Tel: +54 (0342) 4575233/34/39/44/45 ext 195, Fax: +54 (342) 4575224

Resumen Las medidas de similitud semántica en base a ontologías son útiles en muchas aplicaciones, por ejemplo para la inferencia de funciones de genes o proteínas. Las medidas tradicionales para esta aplicación se basan en la frecuencia de anotación de términos, y no cumplen con ciertas propiedades básicas cuando se las utiliza como distancias. En este trabajo se propone definir nuevas medidas que incorporen la estructura del grafo de la ontología de genes, aplicando la teoría de procesamiento de señales en grafos. Los genes representados como señales en el grafo, definiendo caminos que recorren el grafo desde la raíz hasta los términos anotados. El desempeño de las medidas propuestas se evalúa en primer lugar calculando la distancia entre genes anotados en una sub-ontología sencilla, con distintos conjuntos de genes. Luego se evalúa el desempeño en la aplicación de interés, prediciendo funciones de genes mediante un enfoque Bayesiano. Las medidas propuestas se ajustan mejor a las nociones intuitivas de distancia y obtienen un mejor desempeño en la inferencia.

Palabras clave: Ontología de genes · distancias · señales en grafos · Laplaciano · transformada de Fourier en grafos.

1. Introducción

Una ontología es una estructura que se puede representar como un grafo, especificando en la misma conceptos, objetos y otras entidades de un área de interés [1]. En los nodos se definen los conceptos y las aristas indican las relaciones entre conceptos. La ontología de los genes (GO, del inglés *Gene Ontology*) [2], es una de las ontologías más relevantes, ya que en ella se pueden expresar todas las funciones posibles de los genes. La GO tiene numerosos usos, entre ellos la predicción de módulos funcionales, la representación de vías biológicas y la inferencia de funciones desconocidas [3].

En la inferencia es necesario contar con medidas de similitud semántica entre genes anotados en la GO. En este trabajo analizamos las medidas más usadas, Resnik, Lin Relevance, coseno y exactitud. Estas medidas las denominamos medidas clásicas y tienen algunos inconvenientes que vamos a presentar para cada una de ellas. En general, podemos mencionar que el problema de estas medidas

es que no incorporan en forma explícita información de la estructura del grafo, lo cual puede limitar fuertemente sus desempeños en las aplicaciones.

En este trabajo analizamos formas alternativas de caracterizar la similitud entre genes, que tengan en cuenta la estructura del grafo para atacar las limitaciones de las medidas clásicas. Para ello definimos un conjunto de nociones que dejan sentado qué se busca en una medida de similitud y proponemos nuevas medidas a partir del procesamiento de señales en grafos sobre la estructura de la GO. Luego, aplicamos estas medidas en la medición de distancias con distintos conjuntos de genes, analizando y comparando los resultados entre todas las medidas, y en la inferencia de funciones de genes, mediante un método Bayesiano, para un conjunto de genes de la levadura.

2. Medidas de similitud entre genes

Una ontología se puede representar como un grafo en donde los nodos son las funciones de los genes y las aristas son las relaciones entre esas funciones. Así, es posible representar un gen como un conjunto de nodos y sus conexiones. Se puede determinar el parecido de los elementos de la ontología a través de las medidas de similitud. Estas medidas se utilizan para medir el parecido entre términos o genes, y se pueden utilizar para inferir nuevos términos en los genes.

Las medidas de similitud se pueden clasificar de acuerdo a si miden nodos o aristas. En este trabajo analizamos las medidas sobre nodos, debido a que los nodos contienen información de las anotaciones de los genes. En las medidas sobre nodos se suele utilizar el contenido de información (IC, del inglés *Information Content*), que es una medida de la especificidad de un término y se define como $IC(c) = -\log p(c)$, donde c es el término del que se quiere conocer el contenido de información y $p(c)$ es la probabilidad de ocurrencia de c para un determinado organismo, que se estima con la frecuencia de anotación [4].

Medidas de similitud entre términos de la ontología: las medidas de similitud semántica más importantes son las de Resnik [5], Lin [6] y Relevance [7]. La similitud de Resnik se obtiene a partir del ancestro común con mayor contenido de información, al que también se lo llama ancestro común más significativo (MICA, por sus siglas en inglés). La medida de Resnik se define como $s_{Res}(c_1, c_2) = IC(c_{MICA})$, donde c_1 y c_2 son los términos de la ontología a los que se les calcula la similitud y c_{MICA} es el ancestro común más significativo para los términos anteriores. Al usar esta medida puede ocurrir que dos pares de genes diferentes con diferencias importantes en sus funciones pero con el mismo MICA, tengan la misma similitud.

La similitud de Lin agrega a la medida de Resnik el contenido de información de los términos a comparar $s_{Lin}(c_1, c_2) = \frac{2 \cdot IC(c_{MICA})}{IC(c_1)+IC(c_2)}$, lo que soluciona el problema de la medida anterior pero no aporta información sobre la profundidad a la que se encuentran los términos dentro del grafo.

La similitud de Relevance añade información de la distribución dentro del grafo, utilizando la probabilidad de ocurrencia del MICA, esto es $s_{Rel}(c_1, c_2) = s_{Lin}(c_1, c_2) \cdot (1 - p(c_{MICA}))$. Pero al estar basada en el MICA, en el caso de dos

pares de genes diferentes con diferencias importantes en sus funciones pero con el mismo MICA, les asigna la misma profundidad.

Medidas de similitud entre genes completos: los genes anotados en la ontología contienen varios términos, por lo que no se puede medir la similitud de forma directa con las medidas anteriores. Una alternativa es obtener la similitud entre cada par de etiquetas de términos (tomando una de cada gen a comparar) y luego, a partir de estos resultados parciales, obtener la similitud global entre los genes. Para ello se puede utilizar el mínimo, máximo, promedio global o el promedio de las mejores coincidencias (BMA, del inglés *Best-Match Average*) [8].

La similitud entre anotaciones de genes también se puede estimar con otras medidas globales sobre los términos sin considerar la estructura de la ontología. Las más usadas son la similitud coseno y exactitud. Estas medidas tienen la ventaja de que se pueden calcular de forma directa sobre el gen completo.

En este caso necesitamos representar los genes como vectores binarios $\boldsymbol{g} \in \{0,1\}^N$, que tiene tantos elementos como etiquetas hay en el grafo, indicando con 1 a las etiquetas que pertenecen al gen y con 0 a las que no le pertenecen. El orden de los términos en el vector es arbitrario, en tanto se utilice el mismo orden para todos los genes. De esta forma, la similitud coseno se obtiene simplemente como el coseno del ángulo entre vectores $s_{cos}(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a} \cdot \boldsymbol{b}}{\|\boldsymbol{a}\| \|\boldsymbol{b}\|}$, donde $\boldsymbol{a}$ y $\boldsymbol{b}$ son los vectores binarios que representan a los términos de los genes A y B, respectivamente. El problema de esta medida es que no contiene información de la estructura del grafo.

Para definir la medida basada en la exactitud suponemos que contamos con un gen y buscamos medir el parecido de un segundo gen con respecto al primero. Denominamos a los términos anotados (valores 1) en los genes como "positivos" (P) y a los no anotados (valores 0) como "negativos" (N). Luego, de la comparación término a término surgen los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN), y falsos negativos (FN).

Conociendo estos valores para un par de genes A y B, representados en forma binaria como $\boldsymbol{a}$ y $\boldsymbol{b}$, se define la similitud basada en exactitud como $s_{exa}(\boldsymbol{a}, \boldsymbol{b}) = \frac{TP+TN}{M}$, donde $M = TP + FP + TN + FN$. Además, se puede utilizar como medida de similitud ya normalizada dado que vale 1 si los genes son iguales y 0 si sus términos son todos diferentes. La medida es muy sencilla pero tampoco contiene información sobre la estructura de la ontología, considerando por ejemplo la profundidad de los términos dentro del grafo.

3. Medidas propuestas

Las medidas de similitud presentadas anteriormente son muy utilizadas, a pesar de esto hay aspectos de las mismas que son indeseables. A su vez, no cumplen con ciertas intuiciones básicas que tendría que tener una medida de similitud.

En ciertas aplicaciones que utilizan genes anotados en la GO se requieren de medidas de distancia en lugar de similitud. Estas se pueden obtener restando a 1 la similitud, siempre que esta última tenga valores entre 0 y 1. De aquí en

adelante vamos a referirnos a las medidas de distancia, en lugar de las medidas de similitud, con lo cual son aún más relevantes las críticas antes mencionadas en relación con las medidas de similitud existentes.

Para definir nuevas medidas de distancia que caractericen mejor a los genes incorporando la estructura de la GO, necesitamos tener en cuenta las propiedades o características que serían deseable que tengan. En este trabajo consideramos las siguientes: i) La distancia entre un gen y sí mismo debe ser cero. ii) La distancia entre genes con etiquetas cercanas a la raíz debería ser menor que la de genes con etiquetas de mayor profundidad. Esto se debe a que las etiquetas a mayor profundidad son más específicas. iii) Los términos que no son iguales en los genes aportan un incremento al valor de la distancia, mientras que los términos iguales no modifican ese valor. Por ejemplo, si se compara un gen con sí mismo, la distancia debe ser cero. Si se agrega un término extra a uno de ellos y se mide, la distancia debería incrementarse. iv) Las distancias deben estar entre $[0, 1]$.

Las medidas descritas en la sección anterior no cumplen con estas nociones. En algunos casos la distancia de un gen con sí mismo resulta distinta de cero. Además, hay casos en que la distancia entre dos genes con cierto número de diferencias, pero muy lejanos a la raíz, es menor que la de dos genes con similar número de diferencias pero cercanos a la raíz.

Como las medidas clásicas no se ajustan a estas nociones vamos a proponer nuevas medidas basadas en la estructura del grafo. En términos generales, se busca obtener una matriz de transformación D, construida a partir del grafo, y utilizarla para transformar los genes $\boldsymbol{g}$, esto es $\boldsymbol{g}' = D\boldsymbol{g}$, donde $\boldsymbol{g}$ es la representación binaria de los genes descripta anteriormente. Luego, la distancia se podría calcular como la norma euclídea de la diferencia entre genes transformados $d(\boldsymbol{g}'_1, \boldsymbol{g}'_2) = \|\boldsymbol{g}'_1 - \boldsymbol{g}'_2\|$. La distancia se normaliza dividiendo por la mayor de las distancias obtenidas. Entonces, la distancia ajustada al rango $[0, 1]$ es $d' = \frac{d}{d_{max}}$.

La construcción de la matriz D corresponde a la información del grafo que consideramos relevante en la definición de la medida de distancia. En nuestro caso la matriz D puede contener información directa de la estructura del grafo o información frecuencial del grafo. En las siguientes secciones se presenta la definición de las matrices para las medidas propuestas y su aplicación en la medida de distancia.

Grafos y genes: para el desarrollo de las medidas se requiere de un grafo definido numéricamente, con nodos definidos como enteros entre 0 y $N-1$, donde N es la cantidad de nodos del grafo (etiquetas de función). Para cada subgrafo de la GO, se genera una matriz de transformación D en donde cada etiqueta de la GO tiene asociado un índice con valores entre 0 y $N-1$. El orden de las etiquetas en las matrices D está dado por el orden de aparición de las mismas en el archivo que define la GO.

Los genes se anotan en la GO asignando un conjunto de etiquetas, asociadas a las funciones de los genes. Los términos de las GO tienen padres (a excepción del nodo raíz), y esa relación indica que el término hijo representa una función

de mayor especificidad. Al buscar los padres en forma recursiva hasta la raíz se obtienen los ancestros del gen. Las etiquetas de los ancestros se pueden incorporar al gen, ya que esas etiquetas también lo caracterizan. En este trabajo nos interesa esta representación de los genes, ya que contiene mayor información de la estructura del grafo y es un procedimiento estándar en las evaluaciones de los algoritmos de predicción [9].

En el análisis de las medidas de similitud se utiliza la representación de las etiquetas del gen como un vector binario, explicada anteriormente.

Laplaciano: la GO contiene información frecuencial que se puede obtener utilizando la Transformada de Fourier para grafos. Esto nos permite transformar los genes y medir sus distancias en el espacio de las frecuencias. La definición de la Transformada de Fourier para grafos utiliza los autovalores y autovectores del laplaciano del grafo. Éste se define a partir de la matriz de grado Ω y la matriz de adyacencia A del grafo. El laplaciano queda definido como $L = \Omega - A$. Tiene asociado un conjunto autovectores ortonormales $\{\boldsymbol{u}_l\}_{l=0,1,...,N-1}$ y autovalores $\{\lambda_l\}_{l=0,1,...,N-1}$, que satisfacen la ecuación $L\boldsymbol{u}_l = \lambda_l \boldsymbol{u}_l$, para $l = 0, 1, ..., N-1$.

La Transformada de Fourier para grafos se define como $\hat{f}(\lambda_\ell) = \langle f, \boldsymbol{u}_\ell \rangle = \sum_{i=1}^{N} f(i)\boldsymbol{u}_\ell^*(i)$, donde λ_l son los autovalores del laplaciano del grafo, $f \in \mathbb{R}^N$ es una función definida sobre los vértices del grafo y $\boldsymbol{u}_l$ son los autovectores del laplaciano del grafo [10]. En nuestro caso las funciones definidas sobre el grafo son los genes $\boldsymbol{g}$.

Los autovectores nos definen la base de Fourier para realizar la transformación. Se seleccionará un subconjunto de autovectores como filas de la matriz D con la cual transformar los genes, debido al costo computacional que implica obtener los autovectores. La selección de los autovectores no es arbitraria, se comprobó en experimentos preliminares que los autovectores asociados a los autovalores más grandes dan mejores resultados. A esta medida la llamaremos "medida del Laplaciano" y la expresamos en notación matemática como $d_{eig}(\cdot, \cdot)$.

Caminos: en esta medida buscamos realizar una transformación de los genes, considerando la estructura de la GO para medir distancias entre los genes transformados. La estructura del grafo se incluye en la matriz de transformación con los caminos dentro del grafo como filas. Las filas de esta matriz representarán un diccionario, probablemente redundante, para representar a los vectores binarios de etiquetas de cada gen.

Los términos de la GO que no tienen términos hijos se denominan "términos hoja". De esta forma un camino a un término hoja es el camino entre el mismo término y la raíz, y se representa como la lista de etiquetas que forman ese camino en un vector binario. Si se obtienen todos los caminos a los términos hojas (incluyendo todos los posibles caminos entre un nodo hoja y el raiz) entonces tenemos una representación de la estructura del grafo. Con los caminos a las hojas se construye una matriz D. A esta medida la llamaremos "medida de todos los caminos" y la expresamos en notación matemática como $d_{path}(\cdot, \cdot)$.

Hojas: esta medida de distancia es similar a la previamente presentada, con una ligera variación. En este caso en lugar de tomar todos los caminos posibles a una hoja dada, se toma un único camino por hoja que contiene todas las etiquetas

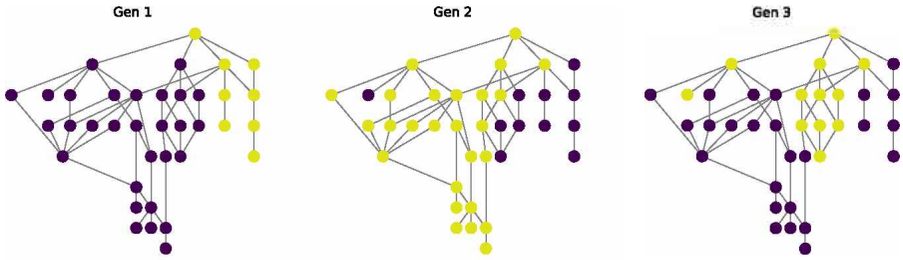

Figura 1. Genes de la levadura.

que se encuentran entre el término hoja y la raíz. De esta forma contamos con tantos elementos como términos hoja tenga el grafo. La matriz D se construye a partir de dichos elementos como filas. A esta medida la llamaremos "medida de los caminos a las hojas" y la expresamos en notación matemática como $d_{leaf}(\cdot,\cdot)$.

4. Aplicación de las medidas en una sub-ontología

En esta sección comparamos los resultados de las medidas de distancia propuestas con respecto a las medidas de distancia clásicas, sobre un subgrafo de la ontología y utilizando genes reales. La sub ontología que se usará como caso de estudio se presenta en la Figura del Anexo I. Ésta proviene del subgrafo de procesos biológicos, fue descargada[3] en formato *obo* y su fecha de publicación es 09/12/2019. La numeración de los términos en el grafo es la que se utiliza para armar el grafo y los genes, y además sirve para referirnos a los mismos de una forma más sencilla que por sus identificadores GO.

Para este análisis se seleccionaron 3 genes de levadura, obtenidos del archivo de anotación[4] del organismo con fecha de publicación 17/12/2019. Los genes son los siguientes: YDR106W, YDR263C y YER042W. Por simplicidad los vamos a nombrar Gen 1, Gen 2 y Gen 3, en el mismo orden en que se listaron.

A las anotaciones de cada gen se agregan los términos de sus ancestros hasta la raíz, para contar con mayor información de cada uno, como se explicó anteriormente. En la Figura 1 se presentan los genes propagados en el grafo, los nodos en color amarillo corresponden a los términos anotados en los genes y los nodos en color violeta a los términos que no están anotados.

Previo a obtener las distancias analizamos los resultados que en principio esperamos obtener de acuerdo a nuestras nociones sobre las medidas de distancia y la importancia biológica de las anotaciones: i) la menor distancia corresponde a los genes 1 y 3, que si bien comparten pocos términos, también tienen menor cantidad de términos en diferencia, ii) le sigue la distancia entre los genes 2 y 3, dado que comparten varios términos, iii) y finalmente, la distancia entre 1 y 2 es la mayor, porque tiene mayor cantidad de términos que difieren.

En la Tabla 1 se presentan los resultados de las distancias con las medidas de todos los caminos, los caminos a las hojas, Laplaciano, Resnik, Lin, Relevance,

[3] http://data.bioontology.org/ontologies/GO/submissions/1779/download?apikey=8b5b7825-538d-40e0-9e9e-5ab9274a9aeb

[4] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/YEAST/goa_yeast.gaf.98.gz

	Gen 1 vs Gen 2	Gen 1 vs Gen 3	Gen 2 vs Gen 3
$d_{path}(\cdot,\cdot)$	**1.0000**	**0.3042**	**0.8705**
$d_{leaf}(\cdot,\cdot)$	**1.0000**	**0.2622**	**0.8684**
$d_{eig}(\cdot,\cdot)$	**1.0000**	**0.2738**	**0.8714**
Coseno	0.8613	0.7959	0.6037
Exactitud	**0.8108**	**0.4324**	**0.6486**
Resnik promedio	0.9868	0.9963	0.9887
Lin promedio	0.9717	0.9931	0.9724
Relevance promedio	0.9967	0.9992	0.9911

Tabla 1. Distancias entre genes de la levadura.

coseno y exactitud. Para la medida del laplaciano se usan 10 autovectores, de los 36 que se obtienen del grafo. Las medidas de Resnik, Lin y Relevance[5] se aplican con el promedio [11]. En la primera columna se listan las medidas de distancia, en las siguientes columnas se presentan las distancias entre los genes; en la segunda columna entre el Gen 1 y el Gen 2, en la tercera columna entre el Gen 1 y el Gen 3, y en la cuarta columna entre el Gen 2 y el Gen 3.

De la Tabla 1 se puede observar que la medida de todos los caminos brinda los resultados esperados, marcados en negrita, la distancia entre los genes 1 y 3 es la menor, seguida por la distancia entre los genes 2 y 3, y finalmente la distancia entre los genes 1 y 2 es la mayor. Aplicando el mismo análisis a la medida de los caminos a las hojas podemos observar que también presenta los resultados esperados. Estas medidas contienen información de los caminos del grafo en la matriz de transformación D, lo que permite que se representen de forma correcta los genes en el espacio transformado y las distancias den los resultados esperados.

La medida del laplaciano también obtiene los resultados esperados (marcados en negrita). En este caso los genes son bien representados en el espacio reducido de las frecuencias del grafo. La medida de exactitud cumple con nuestras expectativas, ya que el caso los genes cuentan con una cantidad considerable de términos en diferencia y esto permite determinar fácilmente las distancias por la cantidad de términos en común. Por otro lado, la medida coseno no brinda los resultados esperados, la distancia entre los genes 2 y 3 es menor que la distancia entre los genes 1 y 3. Esto puede deberse a que en el espacio de los vectores que representan los genes el Gen 1 se encuentra más cerca del Gen 2 que el Gen 3, además, no se tiene en cuenta la estructura de la ontología.

La medida de Resnik, calculada con el promedio de las distancias, no cumple con las nociones, la distancia entre los genes 1 y 2 es la menor, seguida por la distancia entre los genes 2 y 3 y por último la distancia entre los genes 1 y 3. En las medidas de Lin y Relevance, la distancia entre los genes 1 y 3 resulta ser mayor, lo que no coincide con el resultado esperado.

En el Anexo II se presentan ejemplos de aplicación de las medidas sobre genes inventados. Los ejemplos de esta sección y el Anexo II nos permitieron ver en qué casos las diferentes medidas de distancia tienen resultados aceptables y cuáles fallan de acuerdo a nuestra percepción de la magnitud relativa de las distancias. Las medidas basadas en similitud, Resnik, Lin y Relevance, en ningún

[5] https://github.com/tanghaibao/goatools

caso cumplen con las relaciones básicas esperadas, como vimos al analizar las distancias del conjunto de 3 genes de la levadura; a diferencia de las medidas propuestas que si lo hacen. La medida de coseno no diferencia entre términos en distintas ramas, a diferencia de las medidas propuestas, y puede dar magnitudes diferentes a las esperadas. La medida de exactitud no distingue entre términos a diferentes niveles en el grafo, lo que si realiza la medida basada en el Laplaciano. Las medidas de todos los caminos y los caminos a las hojas presentan buenos resultados, y aunque suelen no distinguir términos de una misma rama, en la mayoría de los casos concuerdan con los resultados esperados. Las medidas propuestas en este trabajo permiten diferenciar correctamente a los genes definidos en la ontología y a distinguir términos que se encuentran en distintos caminos de la misma.

5. Predicción de etiquetas por inferencia bayesiana

Método de inferencia: definimos $G = \{\boldsymbol{g}_j\}$, $j = 1, \ldots, m$ como el conjunto de genes con etiquetas conocidas para una sub ontología de la GO específica. $L_j = \{\ell_{jk}\}$ es el conjunto de todas las etiquetas del gen $\boldsymbol{g}_j$, con $L = \cup L_j$ como el conjunto de todas las etiquetas asociadas a G. Si $\boldsymbol{g}_i$ es un gen sin etiquetar, usando inferencia Bayesiana se puede estimar la probabilidad de una etiqueta $\ell \in L$ le pertenezca como $p(\ell|\boldsymbol{g}_i) \propto p(\ell) \cdot p(\boldsymbol{g}_i|\ell) = \frac{1}{C} \sum_{\boldsymbol{g}_j} I(\ell, \boldsymbol{g}_j) \cdot \sum_{\boldsymbol{g}_j/\ell \in L_j} S(\boldsymbol{g}_i, \boldsymbol{g}_j)^\gamma$, donde $I(\ell, \boldsymbol{g}_j)$ es una función que indica con un valor 1 si el gen $\boldsymbol{g}_j$ fue etiquetado con la etiqueta ℓ y 0 en otro caso. $S(\boldsymbol{g}_i, \boldsymbol{g}_j)$ es una medida de similitud sobre los genes $\boldsymbol{g}_i$ y $\boldsymbol{g}_j$, y C es una constante de normalización para las probabilidades. El exponente γ permite asignar mayor importancia a los términos de los genes más cercanos. Esto es particularmente importante cuando hay un gran número de genes involucrados. En este trabajo usamos medida de similitud $S(\boldsymbol{g}_i, \boldsymbol{g}_j) = \frac{2}{1+d(\boldsymbol{g}_i, \boldsymbol{g}_j)} - 1$, que toma valores en el intervalo $[0, 1]$. Luego L es ordenada de forma descendente por $p(\ell|\boldsymbol{g}_i)$, y las etiquetas con la probabilidad más alta se asignan a $\boldsymbol{g}_i$ hasta un valor máximo de probabilidad acumulada μ.

Los resultados de la inferencia se miden con la métrica F_1. Para definirla retomamos los conceptos de verdaderos positivos, falsos positivos, verdaderos negativos y falsos negativos definidos en secciones anteriores. Primero se definen la sensibilidad $s = \frac{TP}{TP+FN}$ y precisión $p = \frac{TP}{TP+TN}$, y luego el F_1 se define como $F_1 = 2\frac{sp}{s+p}$.

Resultados de la inferencia: aplicamos el método de inferencia presentado a un conjunto de genes de la levadura [12], de los cuales consideramos únicamente los genes que no tienen valores faltantes, pertenecen a la categoría de procesos biológicos de la GO y que no tienen código de evidencia ND (del inglés *non biological data available*). De esta forma, se usaron sólo etiquetas de función que han sido validadas experimentalmente. Pasamos del conjunto original de 2467 genes a 587 genes. Los ancestros de los genes se obtuvieron del archivo de anotación[6] del organismo con fecha de publicación 17/12/2019. Las

[6] ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/old/YEAST/goa_yeast.gaf.98.gz

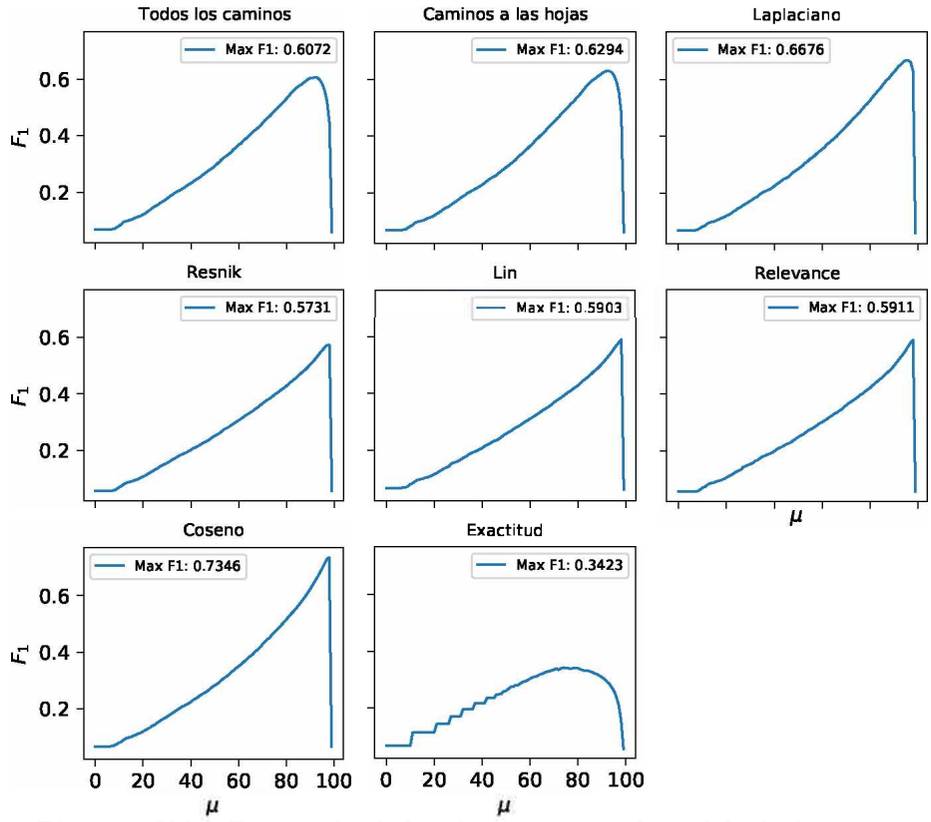

Figura 2. Valor F_1 promedio de la inferencia para cada medida de distancia.

medidas utilizadas son Resnik, Lin, Relevance (aplicando el promedio global entre etiquetas), coseno, exactitud, todos los caminos, caminos a las hojas y el Laplaciano.

El proceso de inferencia se aplica con un *leave-1-out* en donde se infieren funciones a cada uno de los genes, asumiendo que no conocemos las etiquetas de ese gen, pero sí las del resto del conjunto. De la misma forma, se supone que conocemos las distancias semánticas entre todos los genes.

El criterio de corte a utilizar en la probabilidad acumulada también es importante. En estos experimentos aplicamos la inferencia para 100 criterios de corte espaciados entre 0 y 1. En cada caso calculamos el valor F_1 del resultado de la inferencia. El mejor método será el que pueda alcanzar el mayor valor de F_1, independientemente del umbral de corte al que lo alcance [9]. En la Figura 2 se muestra, para cada una de las medidas de distancia, el promedio de los valores F_1 para todos los criterios de corte. El eje de las abscisas indica el umbral de probabilidad acumulada y el eje de las ordenadas el valor F_1 promedio para dicho umbral. La leyenda en las figuras muestra el valor F_1 máximo para todos los umbrales. Se puede observar que los mayores promedios de F_1 se obtienen para las medidas de coseno, Laplaciano, caminos a las hojas y todos los caminos, en ese orden. Estas medidas no solo obtienen mejores resultados en la inferencia

sino que permiten comparar genes completos, con todas sus anotaciones, en lugar de tener que medir término a término y aplicar un método aproximado de integración.

6. Conclusiones y trabajo a futuro

En este trabajo se propusieron medidas de distancia que consideran la estructura del grafo de la ontología de genes. Éstas se ajustan mejor a las nociones de distancia entre genes, y obtienen mejores resultados que las medidas clásicas. La medida basada en el Laplaciano obtiene buenos resultados en todos los casos analizados. En la inferencia de funciones de genes las medidas propuestas obtienen un buen desempeño, superando a las medidas clásicas de Resnik, Lin, Relevance y exactitud. De las propuestas destaca la medida basada en el Laplaciano que obtiene el mejor desempeño, aunque no supera en desempeño a la distancia coseno para esta tarea. La medida basada en el Laplaciano obtiene resultados que nos impulsa a explorar más su definición, analizando en particular la selección de los autovalores, con una estrategia que maximice el desempeño en la inferencia de funciones sin aumentar significativamente el el costo computacional.

Referencias

1. Gruber, T.: A translation approach to portable ontology specifications. Knowledge Acquisition, 5, 199-220 (1993)
2. Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., Sherlock, G.: Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nature genetics, 25(1), 25–29 (2000)
3. Leale, G., Baya, A. E., Milone, D. H., Granitto, P. M., Stegmayer, G.: Inferring Unknown Biological Function by Integration of GO Annotations and Gene Expression Data. IEEE/ACM transactions on computational biology and bioinformatics, 15(1), 168–180 (2018)
4. Pesquita, C., Faria, D., Falcão, A. O., Lord, P., Couto, F. M.: Semantic Similarity in Biomedical Ontologies. PLoS Comput Biol 5(7): e1000443 (2009)
5. Resnik, P.: Using Information Content to Evaluate Semantic Similarity in a Taxonomy. ArXiv, abs/cmp-lg/9511007 (1995)
6. Lin, D.: An Information-Theoretic Definition of Similarity. ICML (1998)
7. Schlicker, A., Domingues, F., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on Gene Ontology. BMC Bioinformatics, 7, 302 - 302 (2006)
8. Pesquita, C., Faria, D., Bastos, H. P., Ferreira, A. E., Falcão, A., Couto, F.: Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics, 9, S4 - S4 (2008)
9. Zhou, N., Jiang, Y., Bergquist, T. R., Lee, A. J., Kacsoh, B. Z., Crocker, A. W., Lewis, K. A., Georghiou, G., Nguyen, H. N., Hamid, M. N., Davis, L., Dogan, T., Atalay, V., Rifaioglu, A. S., Dalkıran, A., Cetin Atalay, R., Zhang, C., Hurto, R. L., Freddolino, P. L., Zhang, Y., ... Friedberg, I.: The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens. Genome biology, 20(1), 244 (2019)

10. Shuman, D., Narang, S., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains. IEEE Signal Processing Magazine, 30, 83-98 (2013)
11. Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., Tang, H.: GOATOOLS: A Python library for Gene Ontology analyses. Scientific reports, 8(1), 10872 (2018)
12. Eisen, M., Spellman, P., Brown, P., Botstein, D.: Cluster analysis and display of genome-wide expression patterns. Proc. Nat. Acad. Sciences, 95(5), 14863–14868 (1998)

Anexo I: Sub-ontología de la GO

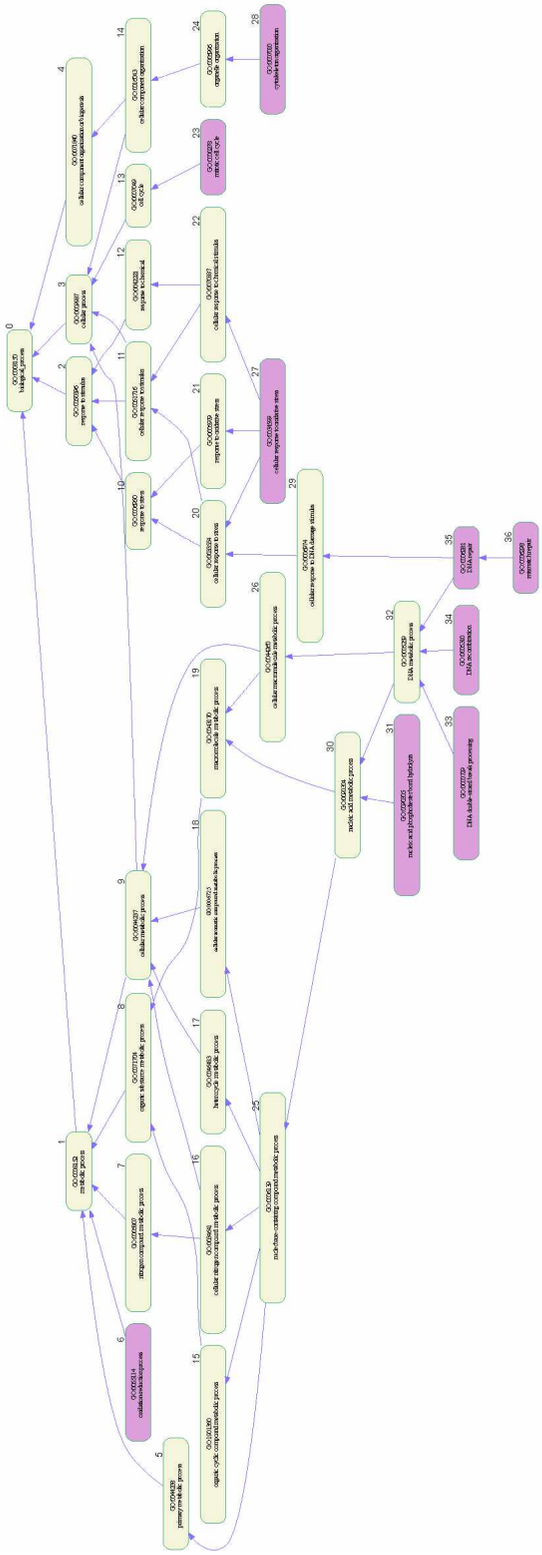

Figura 3. Subgrafo de la ontología de procesos biológicos.

Anexo II: Aplicación en genes simulados

En este anexo vamos a analizar el comportamiento de todas las medidas presentadas, aplicadas sobre genes simulados. El primer caso se presenta en la Figura 4. En estos genes se generó una diferencia de un término en distintas ramas, del Gen 1 con el Gen 2 y del Gen 1 con el Gen 3. A partir de los resultados podremos determinar si la medida solo tiene en cuenta los términos del gen más allá de su ubicación en la ontología, en cuyo caso las distancias entre los genes mencionados resultaría igual.

En este experimento esperamos que la distancia entre los genes 1 y 2 sea la menor, dado que el término en el que difieren es más cercano a la raíz. Le sigue la distancia entre los genes 1 y 3, que tienen un término de diferencia. Y finalmente, la distancia entre 2 y 3 que difieren en dos términos.

Podemos observar en la Tabla 2 que las medidas propuestas cumplen con nuestras expectativas (distancias marcadas en negrita), dado que la distancia entre los genes 1 y 2 es menor que la distancia entre los genes 1 y 3. Sin embargo, las medidas de coseno y exactitud no cumplen con las expectativas, ya que la distancia entre los genes 1 y 2 es igual a la distancia entre los genes 1 y 3. La medida de exactitud falla debido a que solo tiene en cuenta los términos de los genes, entonces como en ambas comparaciones teníamos una etiqueta de diferencia y la misma cantidad de etiquetas en común, la medida da el mismo resultado. En la medida coseno ocurre que los genes 2 y 3 tienen el mismo ángulo con respecto al Gen 1.

Otro caso que tenemos que probar es cuando los términos que difieren entre gen y gen están en una misma rama. Esto nos puede indicar si las medidas reconocen los distintos términos en una misma rama. Para ello se definen los genes de la Figura 5, donde se puede observar que el Gen 1 tiene términos en dos ramas completas del grafo y el Gen 2 es similar al anterior pero eliminamos un término de la rama más larga. En cambio el Gen 3 es similar al Gen 1 pero eliminamos los dos últimos términos de la rama más larga.

En este caso esperamos que la distancia entre los genes 2 y 3 sea la menor, dado que la cantidad de términos en común es menor que en los otros casos. Utilizando el mismo razonamiento le sigue la distancia entre los genes 1 y 2. Finalmente, la distancia entre los genes 1 y 3 es la mayor, porque tiene mayor cantidad de términos que difieren.

En la Tabla 3 se presentan los resultados de las medidas de distancia para este experimento. Las medidas de todos los caminos, los caminos a las hojas y exactitud no logran distinguir las distancias entre los genes 1 y 2, y las distancias entre 2 y 3. En las medidas de los caminos esto ocurre porque los genes se encuentran en un mismo camino y la medida no puede distinguir entre distintos términos en un mismo camino. En el caso de exactitud se debe a que solo tiene en cuenta la cantidad de términos en diferencia, que es la misma en los dos casos. Las medidas del Laplaciano, en negrita, y coseno distinguen entre las distancias mencionadas, pero la relación entre las distancias no es la esperada, ya que la distancia entre los genes 2 y 3 debería ser la menor. En la medida del

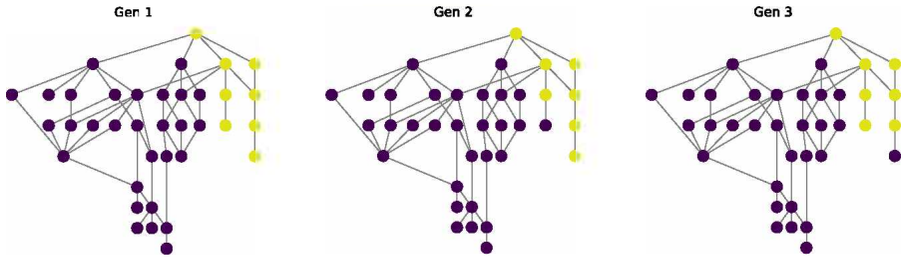

Figura 4. Genes inventados con términos que varían en ramas distintas.

	Gen 1 vs Gen 2	Gen 1 vs Gen 3	Gen 2 vs Gen 3
$d_{path}(\cdot,\cdot)$	**0.5774**	**0.8165**	**1.0000**
$d_{leaf}(\cdot,\cdot)$	**0.5000**	**0.8660**	**1.0000**
$d_{eig}(\cdot,\cdot)$	**0.5054**	**0.8700**	**1.0000**
Coseno	0.0646	0.0646	0.1429
Exactitud	0.0270	0.0270	0.0541

Tabla 2. Distancias entre genes inventados con términos que varían en ramas distintas.

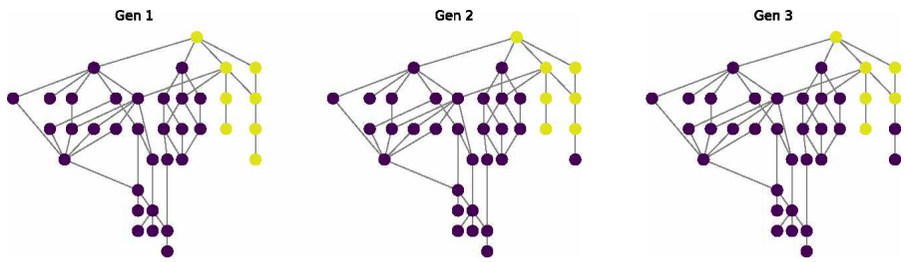

Figura 5. Genes inventados con términos modificados en una misma rama.

	Gen 1 vs Gen 2	Gen 1 vs Gen 3	Gen 2 vs Gen 3
$d_{path}(\cdot,\cdot)$	0.5000	1.0000	0.5000
$d_{leaf}(\cdot,\cdot)$	0.5000	1.0000	0.5000
$d_{eig}(\cdot,\cdot)$	**0.4983**	**1.0000**	**0.5476**
Coseno	0.0646	0.1340	0.0742
Exactitud	0.0270	0.0541	0.0270

Tabla 3. Distancias entre genes con términos modificados en una misma rama.

Laplaciano puede estar ocurriendo que en el espacio de las frecuencias el término en diferencia entre los genes 1 y 2 tenga menos peso que el término en diferencia de los genes 2 y 3, y que por ello la distancia entre los primeros resulte menor. En la medida coseno la diferencia se puede deber a que el ángulo entre los vectores que representan a los genes sea mayor cuando esperamos que sea menor.