DISTANCIAS SEMÁNTICAS PARA INFERENCIA EN LA ONTOLOGÍA DE GENES

Tiago López[†], Diego Milone[†] y Leandro Di Persia[†]

[†]sinc(i), *Instituto de Investigación en Señales, Sistemas e Inteligencia Computacional, FICH/UNL-CONICET, Santa Fe, Argentina.* {*tlopez,dmilone,ldipersia*}*@sinc.unl.edu.ar*

Resumen: Las medidas de similaridad semántica en la ontología de genes se utilizan para la inferencia de nuevas funciones de genes o proteínas desconocidas. Las medidas clásicas están basadas en la frecuencia de anotación de términos de función y son inconsistentes con ciertas propiedades básicas cuando son usadas para calcular distancias. En este trabajo proponemos nuevas medidas que incorporan la estructura del grafo de la ontología, a partir de la transformada de Fourier de los genes representados como señales en el grafo y los caminos entre términos. Evaluamos el desempeño de las medidas calculando la distancia entre genes en una sub-ontología sencilla y prediciendo funciones de genes con un enfoque Bayesiano. Las medidas propuestas son consistentes y obtienen mejores resultados en la inferencia.

Palabras clave: *Señales en grafos, distancias, Laplaciano, transformada de Fourier, Ontología de genes.*
2000 AMS Subject Classification: 92C55

1. INTRODUCCIÓN

Una ontología es una estructura en donde se especifican de forma explícita conceptos, objetos y otras entidades que existen en un área de interés, junto con las relaciones que los unen [1]. Las ontologías se pueden representar como grafos en donde los conceptos se ubican en los nodos y las relaciones en las aristas. En particular, la ontología de genes (GO) [2] representa las distintas funciones de los genes y, entre otras aplicaciones, se pueden mencionar la predicción de módulos funcionales, la representación de vías biológicas y la inferencia de funciones desconocidas [3].

En la mayoría de los casos es necesario contar con medidas de similaridad semántica entre genes anotados en la GO. Las medidas más usadas son las de Resnik [4], Lin [5] y Relevance [6]. La primera tiene en cuenta el contenido de información del ancestro común de menor probabilidad de ocurrencia, lo que puede generar que dos pares de genes diferentes, con diferencias importantes en sus funciones pero con el mismo ancestro común, tengan la misma similaridad. La similaridad de Lin aporta a la anterior la probabilidad de ocurrencia de los términos cuya similaridad se calcula, y esto soluciona el problema antes mencionado pero no aporta información sobre la profundidad a la que se encuentran los términos dentro del grafo. La similaridad Relevance intenta solucionar el punto anterior agregando a la similaridad de Lin el producto con la probabilidad de ocurrencia del ancestro común de mayor contenido de información.

En este trabajo nos proponemos analizar formas alternativas de caracterizar la similaridad entre genes, que tengan en cuenta la estructura del grafo para atacar las limitaciones de las medidas clásicas. Para esto definimos un conjunto de nociones que dejen sentado qué se busca en una medida de similaridad y proponemos nuevas medidas a partir del procesamiento de señales en grafos sobre la estructura de la GO.

2. MEDIDAS PROPUESTAS

Las medidas de similaridad semántica en la GO utilizan el contenido de información (IC) como medida de la especificidad de un término. 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 [7]. Luego, para calcular la similaridad entre genes, cada uno representado como un conjunto de etiquetas de términos, se obtiene la similaridad entre cada par de etiquetas de los genes a comparar y a partir de estas medidas parciales se estima la similaridad global con el mínimo, máximo, promedio global o el promedio de las mejores coincidencias [7].

En general es conveniente realizar estas comparaciones con medidas de distancia, por lo que en adelante solo nos referiremos a las últimas. Para definir nuevas medidas de distancia que caractericen mejor a los genes incorporando la estructura de la GO, necesitamos tener en cuenta ciertas propiedades básicas deseables: 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 distancia entre genes con etiquetas con la misma ubicación relativa pero a mayor profundidad en el grafo (por proveer información más específica); iii) los términos que difieren en los genes aportan un incremento al valor de la distancia, mientras que los términos iguales no deberían modificar ese valor; iv) las distancias deben estar en $[0, 1]$.

Para este trabajo codificamos los genes como vectores binarios $\boldsymbol{g} \in \{0, 1\}^N$, donde cada componente del vector está asociada a uno de los N posibles términos de la GO. Nuestra propuesta es obtener una matriz de transformación D, construida a partir del grafo, y utilizarla para transformar los genes $\boldsymbol{g}' = D\boldsymbol{g}$. Luego, la distancia se podría calcular como la norma euclídea de la diferencia entre genes transformados $d(\boldsymbol{g}'_1, \boldsymbol{g}'_2) = \|\boldsymbol{g}'_1 - \boldsymbol{g}'_2\|$. Como primera propuesta las distancias se normalizarán simplemente dividiendo por el máximo. La construcción de la matriz D depende de la información del grafo que consideramos relevante en la definición de la medida de distancia. En los siguientes párrafos se presentan las tres medidas propuestas.

Laplaciano: la transformada de Fourier del grafo se define a partir de los autovectores y autovalores del Laplaciano del grafo. Sean Ω la matriz de grado y A la matriz de adyacencia del grafo, su Laplaciano se define como $L = \Omega - A$. La operación análoga a la Transformada de Fourier en 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, $\boldsymbol{u}_\ell$ son los autovectores, y $f \in \mathbb{R}^N$ es una función definida sobre sus vértices [8]. En nuestro caso las señales definidas sobre el grafo son los genes $\boldsymbol{g}$. Como no buscamos realizar un cambio de base, seleccionamos el subconjunto de P autovectores asociados a los autovalores más grandes para formar la matriz D_e. A partir de la matriz D_e se obtienen las distancias con el procedimiento explicado anteriormente. A esta medida la expresamos como $d_e(\cdot, \cdot)$.

Caminos: en esta medida buscamos que la estructura de la GO esté presente al medir distancias entre los genes transformados, usando los caminos desde la raíz a cada hoja como filas de la matriz D_p. En este caso se considera un camino a la hoja como la secuencia de términos entre un término hoja y la raíz. Si tomamos todos los caminos posibles a los términos hojas entonces tenemos una representación de la estructura del grafo. La distancia se calcula con el procedimiento presentado anteriormente. A esta medida la expresamos como $d_p(\cdot, \cdot)$.

Hojas: la tercer medida de distancia que proponemos es similar a la previamente presentada, con una ligera variación. En este caso en lugar de tomar todos los caminos a las hojas, se toma un único camino por hoja, que contiene todas las etiquetas que se encuentran entre el término hoja y la raíz. De esta forma contamos con tantos elementos en el diccionario como términos hoja tenga el grafo. Esta medida la expresamos como $d_l(\cdot, \cdot)$.

3. APLICACIÓN DE LAS MEDIDAS DE DISTANCIA

Comparamos las medidas en un caso sencillo, utilizando un subgrafo de la GO. La subontología proviene del subgrafo de procesos biológicos[1] y para este análisis se seleccionaron 3 genes de levadura[2]: YDR106W, YDR263C y YER042W. En la Figura 1 se presentan los términos de estos genes en el grafo de la subontología. Previo a obtener las distancias analizamos los resultados que 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 pero tienen diferencias a mayor profundidad; iii) y finalmente, la distancia entre 1 y 2 debería ser la mayor, porque tiene mayor cantidad de términos que difieren.

De la Tabla 1 se puede observar que las tres medidas propuestas brindan los resultados esperados. Sin embargo, la medida de Resnik[3] [9] no cumple con lo esperado ya que 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. Ninguna de estas medidas cumple con la propiedad básica de que la distancia entre un gen y sí mismo debe ser cero.

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

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

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

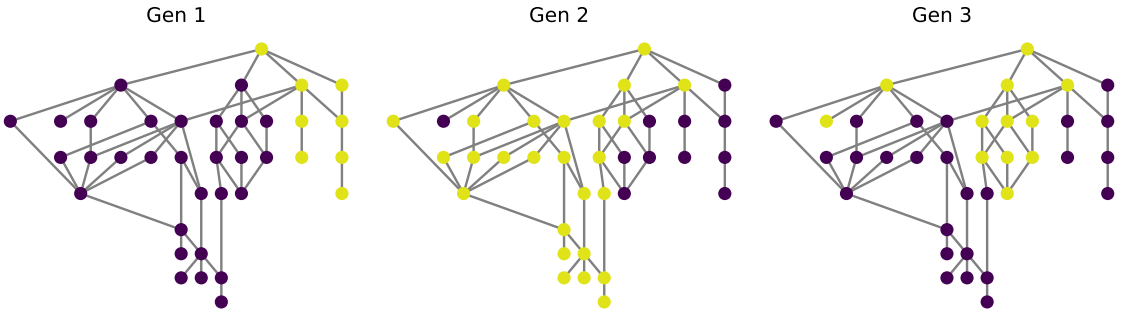

Figura 1: Genes de prueba para evaluar las medidas de distancia, en amarillo los términos que posee cada gen.

	$d_p(\cdot,\cdot)$	$d_l(\cdot,\cdot)$	$d_e(\cdot,\cdot)$	Resnik	Lin	Relevance
Gen 1 vs Gen 2	**1.0000**	**1.0000**	**1.0000**	0.9868	0.9717	0.9967
Gen 1 vs Gen 3	**0.3042**	**0.2622**	**0.2738**	0.9963	0.9931	0.9992
Gen 2 vs Gen 3	**0.8705**	**0.8684**	**0.8714**	0.9887	0.9724	0.9911

Tabla 1: Resultados de las distancias para el conjunto de tres genes.

En un caso real de inferencia de funciones, se busca determinar si una etiqueta pertenece a un gen (previamente sin etiquetar). Mediante un enfoque Bayesiano es posible calcular la probabilidad de que dicha etiqueta pertenezca al gen. Sin embargo, como no se puede medir la distancia de genes totalmente desconocidos, asumimos que éstas se obtuvieron en base a otra información de esos genes [10]. Definimos $G = \{\boldsymbol{g}_j\}$, $j = 1, \ldots, m$ como el conjunto de genes con etiquetas conocidas para una sub-ontología específica de la GO. $L_j = \{\ell_{jk}\}$ 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 una inferencia Bayesiana se puede estimar la probabilidad de que una etiqueta $\ell \in L$ le pertenezca como $p(\ell|\boldsymbol{g}_i) = p(\ell) \cdot p(\boldsymbol{g}_i|\ell) = \frac{1}{C} \sum_{\boldsymbol{g}_j} I(\ell, \boldsymbol{g}_j) \cdot \sum_{\boldsymbol{g}_j/\ell \in L_j} S(\boldsymbol{g}_i, \boldsymbol{g}_j)^\gamma$, donde $I(\ell, \boldsymbol{g}_j)$ es una función que indica con 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 similaridad sobre los genes $\boldsymbol{g}_i$ y $\boldsymbol{g}_j$, y C es una constante de normalización. 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 la siguiente medida de similaridad $S(\boldsymbol{g}_i, \boldsymbol{g}_j) = \frac{2}{1+d(\boldsymbol{g}_i, \boldsymbol{g}_j)} - 1$, para las distancias propuestas. Luego, L es ordenado 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.

Aplicamos el método de inferencia a un conjunto de 587 genes de la levadura [11, 3]. Los términos de los genes se obtuvieron del archivo de anotación del organismo[4]. Las distancias con las medidas clásicas se integraron con el promedio global. Para medir las tasas de acierto en la inferencia se utilizó un esquema de validación *dejar-1-afuera*, en donde aplicamos el método de inferencia descrito a cada uno de los genes, asumiendo que conocemos las etiquetas del resto del conjunto. Para cuantificar el desempeño se usó la métrica de clasificación $F_1 = 2\frac{sp}{s+p}$, donde $s = \frac{TP}{TP+FN}$ es la sensibilidad y $p = \frac{TP}{TP+TN}$ es la precisión, en término de los verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN). Para evaluar la capacidad de cada distancia semántica en todo el rango de inferencia utilizamos 100 criterios de corte μ, equiespaciados entre 0 y 1. En cada caso calculamos el valor F_1 del resultado de la inferencia.

En la Figura 2 se presenta, para cada una de las distancias semánticas, el promedio de los valores F_1 sobre todos los genes para cada uno de los criterios de corte μ y su máximo. Considerando los valores máximos de F_1 promedio, las medidas se pueden ordenar de mayor a menor según su desempeño: Laplaciano, Caminos a las hojas, Todos los caminos, Relevance, Lin y Resnik. Claramente las medidas propuestas tienen mejores resultados en la inferencia. Además, hay que destacar que las medidas propuestas permiten comparar los genes completos, con todas sus anotaciones, en lugar de tener que medir término a término y luego aplicar algún método aproximado de integración.

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

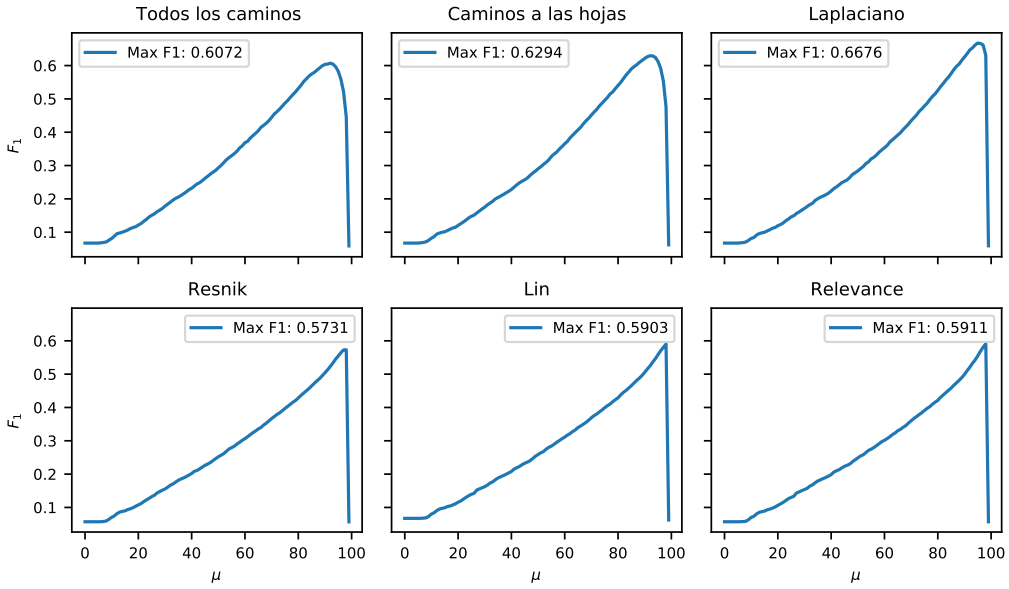

Figura 2: F_1 promedio sobre todos los genes de levadura para cada una de las medidas de distancia.

4. Conclusiones y trabajo a futuro

Se propusieron nuevas medidas que consideran la estructura del grafo en la ontología de genes. Estas medidas se ajustan mejor a las nociones intuitivas para las relaciones de similaridad entre genes, son consistentes y permiten obtener un mejor desempeño en la inferencia de funciones de genes. El mejor desempeño de la medida basada en el Laplaciano nos impulsa a explorar más su definición, en particular en la selección de los autovectores, y tratando de mantenerlos a un mínimo para reducir el costo computacional.

Referencias

[1] T. Gruber, *A translation approach to portable ontology specifications*, Knowledge Acquisition, Volume 5, Issue 2, 1993.

[2] M. Ashburner, C.A. Ball, J.A. Blake, D. Botstein, H. Butler, J.M. Cherry, A.P. Davis, K. Dolinski, S.S. Dwight, J.T. Eppig, M.A. Harris, D.P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J.C. Matese, J.E. Richardson, M. Ringwald, G.M. Rubin, and G. Sherlock, *Gene ontology: tool for the unification of biology.* The Gene Ontology Consortium, Nature genetics, 25(1), 2000, 25–29.

[3] G. Leale, A.E. Bayá, D. Milone, P. Granitto, and G. Stegmayer, *Inferring Unknown Biological Function by Integration of GO Annotations and Gene Expression Data*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, 15, 2018, 168-180.

[4] P. Resnik, *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*, Volume 1 (IJCAI'95). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 448–453.

[5] D. Lin, *An Information-Theoretic Definition of Similarity*, In Proceedings of the Fifteenth International Conference on Machine Learning (ICML '98). Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 296–304.

[6] A. Schlicker, F. Domingues, J. Rahnenführer, and T. Lengauer, *A new measure for functional similarity of gene products based on Gene Ontology*, BMC Bioinformatics, 7, 2006, 302 - 302.

[7] C. Pesquita, D. Faria, A.O. Falcão, P. Lord, and F.M. Couto, *Semantic Similarity in Biomedical Ontologies*, PLoS Comput Biol, 2009, 5(7): e1000443.

[8] D. Shuman, S. Narang, P. Frossard, A. Ortega, and P. Vandergheynst, *The emerging field of signal processing on graphs: Extending high-dimensional data analysis to networks and other irregular domains*, IEEE Signal Processing Magazine, 30, 2013, 83-98.

[9] D.V. Klopfenstein, L. Zhang, B.S. Pedersen, F. Ramírez, A. Warwick Vesztrocy, A. Naldi, C.J. Mungall, J.M. Yunes, O. Botvinnik, M. Weigel, W. Dampier, C. Dessimoz, P. Flick, and H. Tang, *GOATOOLS: A Python library for Gene Ontology analyses*, Scientific reports, 8(1), 2018, 10872.

[10] L. Di Persia, D.H. Milone, and G. Stegmayer, *Annotation pipeline for inferring gene functions integrating GO annotations and expression data*, A2B2C 10th Meeting, Mendoza, Argentina, 2019. http://sinc.unl.edu.ar/sinc-publications/2019/DMS19

[11] M. Eisen, P. Spellman, P. Brown, and D. Botstein, *Cluster analysis and display of genome-wide expression patterns*, Proc. Nat. Acad. Sciences, 95(5), 1998, 14863–14868.