Capítulo 1

Coherencia entre métrica transcripcional y espacio de conocimiento GO

En los capítulos precedentes cuantificamos por medio de diversos índices la congruencia biológica de los grupos encontrados en el espacio de expresión génica. En este capítulo buscaremos cuantificar la coherencia entre los espacios de expresión génica y de conocimiento biológico desde una óptica diferente: desde la métrica en lugar de desde las agrupaciones.

1.1. Alineamiento de núcleo-objetivo

Una matriz de núcleo o matriz de Gram o matriz de kernel K puede ser pensada informalmente como una matriz semidefinida positiva de similaridad de a pares entre puntos de un conjunto de datos. Para un conjunto de datos $\{x_1, ..., x_m\}$ esta noción de similaridad esta dada en términos de una función k llamada kernel tal que:

$$K = K_{ij} = k(x_i, x_j) \tag{1.1}$$

con $i, j = \{1, ..., m\}$ y $k : \mathbb{R}^m \times \mathbb{R}^m \Rightarrow \mathbb{R}$. Una función $k(x, y)$ es un kernel si y solo si para cualquier conjunto finito de datos $C = \{x_1, ..., x_m\}$ y para cualquier conjunto $\{a_1, ..., a_m\} \in \mathbb{R}^m$ se tiene que:

$$\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0 \tag{1.2}$$

Se puede demostrar que esto implica que K debe ser semidefinida positiva (SDP), es decir, $K = \sum_i \lambda_i v_i v_i'$, con $\lambda_i \geq 0$ los autovalores de la matriz K y v_i sus autovectores.
Intuitivamente, un kernel es una transformación que mapea pares de puntos en un espacio de alta dimensionalidad a un indice de similaridad entre los mismos mediante el uso de un producto interno.
Existen multiplicidad de kernels disponibles y para cada aplicación será necesario encontrar el adecuado.
Es de esperar que si es posible extraer información biológica del espacio de expresión genética,

entonces dos puntos que son similares (en algún sentido a definir por el kernel elegido) en el espacio de expresión, también lo sean en el espacio GO (nuevamente, en algún sentido a definir por el kernel elegido). Para cada espacio habrá que definir un kernel adecuado.
Una forma de cuantificar la similaridad entre estos dos espacios es mediante una cantidad conocida como alineamiento núcleo-objetivo o KTA. El KTA de un kernel k_1 con respecto a un kernel k_2 del conjunto C esta definido como:

$$\hat{A}(S, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \tag{1.3}$$

Donde $\langle K_1, K_1 \rangle_F = \sum_{i,j=1}^{m} K1(x_i, x_j)K2(x_i, x_j)$ es el producto de interno de Frobenius entre matrices y K_i son las matrices de kernels simétricas y semidefinida positivas de los espacios a comparar. Este índice tiene un rango entre $[0, 1]$. [49]
Es posible extender este concepto a matrices simétricas indefinidas (no SDP) S mediante diversas técnicas que consisten en transformar S para obtener una S' SDP. Esto es relevante para nosotros, porque matrices de similaridad basadas en similaridad semántica no suelen ser semidefinidas positiva. La que utilizaremos en este trabajo se conoce como *corrimiento del espectro*. Si S es simétrica entonces admite una descomposición en autovalores y autovectores tal que $S = U\Lambda U^T$ con U una matriz ortogonal y Λ una matriz diagonal de autovalores reales, es decir, $\Lambda = diag(\lambda_1, ..., \lambda_m)$. Entonces, el corrimiento del espectro consiste en correr todo el espectro de S por el mínimo necesario:

$$S_{corrida} = U(\Lambda + |\min \lambda_{min}(S), 0|I)U^T \tag{1.4}$$

Decidimos utilizar este método porque el mismo solo aumenta las autosimilaridades, sin modificar la similaridad entre dos puntos distintos, preservando la estructura de grupo al agrupar datos no necesariamente métricos. [50]
Notar que esta medida es una medida global, ya que toma en cuenta todas las similaridades para calcular KTA.

1.2. Espacio de expresión y GO

Para cuantificar la coherencia métrica entre el espacio de expresión de cada tratamiento y las ontologías GOBPA, GOBPB y GOCC, utilizamos como kernel de espacio de expresión, K_x, la similaridad derivada de la correlación:

$$k_x(g_i, g_j) = (\frac{correlacion(g_i, g_j) + 1}{2}) \tag{1.5}$$

con g_i y g_j genes pertenecientes al tratamiento en cuestión.
Para el kernel del espacio de ontologías, utilizamos la similaridad definida en la ecuación **??** y transformamos la matriz en SDP por medio de 1.4. La matriz se construyó tomando en cuenta todos los genes del tratamiento. Si un gen del tratamiento no se encontraba anotado en la ontología, se lo anotaba al nodo raíz y por lo tanto su similaridad con el resto de los genes era cero. Se calculó entonces para cada tratamiento y cada ontología, el KTA y se

construyó además un control nulo de tipo 2, realizando 1000 reordenamientos aleatorios de las etiquetas de la matriz K_x.
Las figuras 1.1, 1.2 y 1.3 presentan en un boxplot, para cada tratamiento y cada ontología, el KTA del control nulo, junto con un punto rojo para el KTA de expresión.

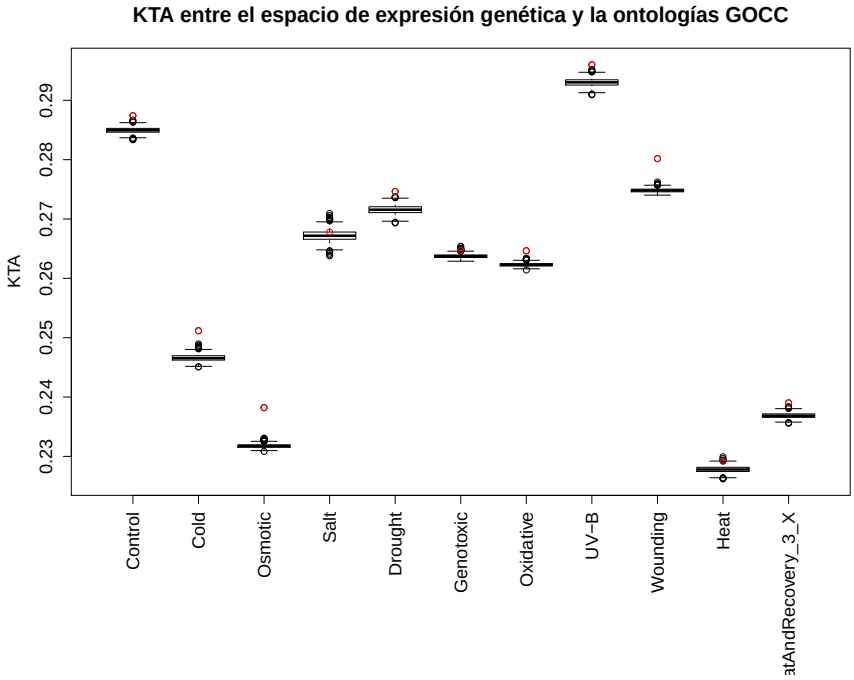


Figura 1.1: KTA para distintos tratamientos entre espacio de expresión y ontología CC.

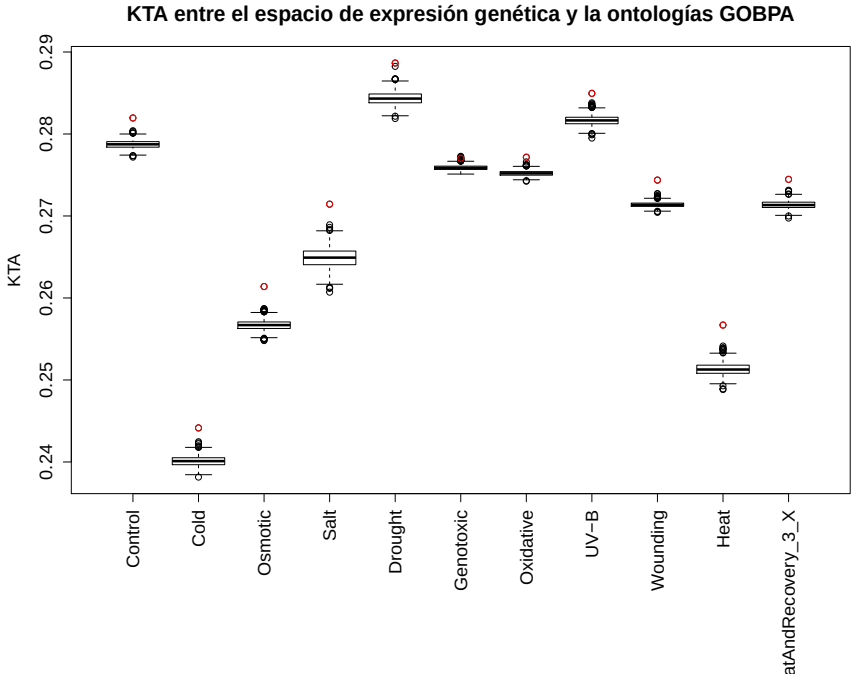


Figura 1.2: KTA para distintos tratamientos entre espacio de expresión y ontología BPA.

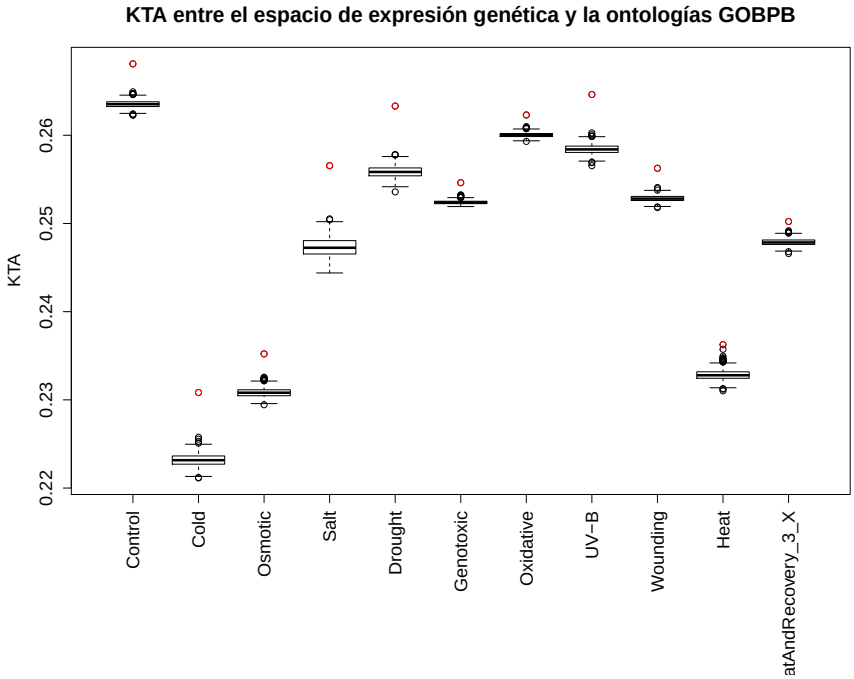


Figura 1.3: KTA para distintos tratamientos entre espacio de expresión y ontología BPB.

En todos los casos encontramos que el KTA de expresión supera todos los valores del KTA de control nulo, lo que soporta la idea de que existe un grado de coherencia no trivial, de orden global, entre la métrica de expresión transcripcional y la métrica del espacio de conocimiento biológico, inferido a partir de la estructura de GO.

1.3. Alineamiento de núcleo-objetivo local

En la sección anterior presentamos una forma de cuantificar globalmente la coherencia entre la métrica transcripcional y el espacio GO mediante el índice KTA. Es posible redefinir este índice para obtener una medida de alineamiento de estos dos espacios pero de forma local, en vecindades transcripcionales. La idea de esta nueva medida apunta a cuantificar e identificar regiones acotadas del espacio transcripcional (i.e. grupos de genes) donde exista un alineamiento espacio-transcripcional vs. espacio-GO diferente a la media. Este índice nos será de gran utilidad en un análisis posterior para caracterizar la congruencia biológica en grupos de genes de similar comportamiento transcripcional y para definir una métrica mixta, que permita definir agrupamientos de genes que presente alto grado de coherencia no solo a nivel transcripcional, sino también en el espacio de conocimiento biológico.
Para definir vecindades transcripcionales consideramos redes de k-primeros-vecinos-mutuos, donde cada nodo es un gen y cada arista tiene un peso w_{ij} entre $[0, 1]$ dado por la similaridad de correlación entre los dos genes g_i y g_j que son unidos por esa arista. Construimos una red $k = 30$, que llamamos $30kmnn$, y estudiamos su topología [1].

[1]Exploramos diferentes alternativas en este punto que incluye el uso de redes de k primeros vecinos en lugar de k primeros vecinos mutuos, y de un número variable de primeros vecinos. Encontramos que la red $30kmnn$ presentada en el trabajo era un buen compromiso entre cobertura y tiempo de computo.

1.3.1. Vecindades transcripcionales

Como primer paso para entender alguna de las particularidades y limitaciones de nuestra caracterización de las vecindades transcripcionales estudiaremos a la red $30kmnn$ mediante dos observables topológicos, la distribución de grado y la intermediación central o *betweenness centrality*.
Además, compararemos la red con dos modelos de redes aleatorias, el modelo de Erdös Renyi, que consiste en construir una red aleatoria con N nodos, conectando pares elegidos al azar y omitiendo múltiples conexiones entre dos mismos nodos hasta alcanzar una cantidad total de K aristas impuestas a priori [51], y el modelo configuracional, que genera una red aleatoria a partir de un recableado de conexiones que mantiene fijo el grado de cada nodo, es decir, no altera la distribución de grado $P(k)$ de la red original [52].

Distribución de grado

El grado k_i de un nodo de la red es la cantidad de primeros vecinos que tiene el nodo.
La distribución de grado $P(k)$ es entonces la probabilidad de un que nodo i tomado al azar tenga grado k.
La figura 1.4a muestra la distribución de grado en escala logarítmica de la red de 30 primeros vecinos mutuos y un modelo nulo de Erdös-Rényi con la misma cantidad de nodos y aristas para el tratamiento frió (1951 nodos y 18436 aristas). Se observa que el grado máximo que alcanzan estas distribuciones está relacionado directamente con el k utilizado, ya que a lo sumo un nodo tendrá k primeros vecinos (k aristas). El hecho de tener un gran número de nodos con k máximo sugiere que ya alcanzamos la "saturación" en una representación basada en interacciones de vecinos mutuos. Vemos de la figura que la distribución obtenida se aleja de la Poissoniana correspondiente al modelo nulo Erdös-Rényi.

Intermediación central o betweenness centrality

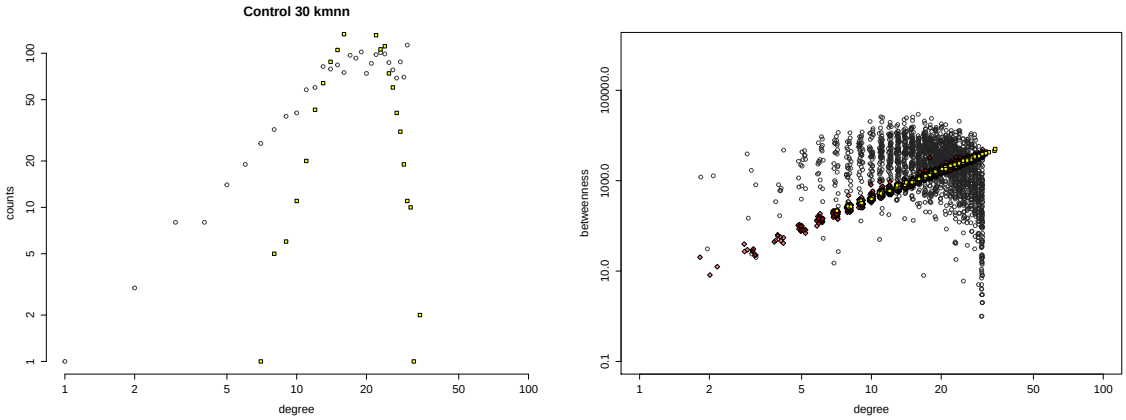
La longitud de un camino entre dos nodos se define como la cantidad de aristas que se recorren para llegar de un nodo al otro. El camino (o caminos) más corto es aquél camino cuya longitud es la menor entre todos los caminos. La longitud de un camino más corto se conoce como distancia geodésica.
El betweenness centrality de un nodo i es igual a la cantidad de caminos más cortos desde todos los nodos a todos los otros nodos que pasan por el nodo i. Es una medida de la influencia del nodo i en la red, ya que un nodo con alto betweenness centrality recibirá una gran parte de la carga de la red, suponiendo que la carga se distribuye a través de los caminos más cortos.
Muchas redes reales presentan algunos nodos de alta conectividad, llamados conectores o *hubs*, por donde pasa la mayor parte de la carga de la red.
La figura 1.4b presenta en escala logarítmica la relación entre el betweenness y el grado k de los nodos de la red $30kmnn$, en círculos grises, la de Erdös-Rényi, en triángulos amarillos y la configuracional, en rombos rojos.

En todos los casos se observa una correlación positiva entre el betweenness y el grado de la red, pero cabe destacar que para un grado fijo k, la red $30kmnn$ presenta una mayor dispersión en el betweenness que las otras, sobre todo en los grados más altos, y con mayor betweenness en los grados intermedios. Esto implica que existen en esta red *hubs* de tamaño intermedio que se conectan a otros *hubs*, lo que da un indicio de la existencia de una estructura modular.
Estas diferencias en la distribución de grado y betweenness centrality entre la red $30kmnn$ y las redes aleatorias serían una evidencia a priori de patrones de conectividad no triviales y estructura modular, es decir, de existencia de grupos de genes dentro de la red altamente relacionados entre si. Como factor extra para elegir trabajar con la red de $30kmnn$, por ser una red con relativamente pocos nodos y aristas, es posible realizar sobre la misma todos los cálculos de KTA local en un tiempo computacional razonable.

1.3.2. KTA local red 30kmnn

Para cada arista de la red $30kmnn$ de cada tratamiento, encontramos su vecindario local a primeros vecinos (los primeros vecinos de los nodos unidos por la arista) y construimos la matriz de similaridad de correlación reducida, consistente en la similaridad de correlación entre esos nodos y sus primeros vecinos. Generamos además la matriz de similaridad semántica reducida usando únicamente los genes anteriores, para cada una de las ontologías GOBPA, GOBPB y GOCC.
Aquellos genes que no estaban anotados en las ontologías, fueron anotados en la raíz de cada una respectivamente, por lo que los genes en la vecindad de un gen i en el espacio de expresión no necesariamente son vecinos del mismo gen en el espacio de ontología.
Observamos que la cantidad de nodos vecinos anotados en la ontología, ny, depende linealmente de la cantidad de nodos vecinos en la red, nx, como muestra la figura 1.5a. Por otro lado, podemos ver que el promedio de los pesos en una vecindad de una arista anotados en la ontología, $wynanotados$, presenta una gran dispersión en su relación con los pesos de las aristas, como se observa en la figura 1.5b. Esto implica que el comportamiento en una vecindad no puede ser predicho a partir del conocimiento de un edge, sino que es necesario tomar toda la vecindad en su conjunto. A partir de estas relaciones, introduciremos el índice KTA local como el índice KTA aplicado a las matrices reducidas de ambos espacios, previamente transformadas en SDP. Para caracterizar su comportamiento nos interesa observar primero que el comportamiento de este índice depende linealmente de $wynanotados$, como muestra la figura 1.5c, siendo entonces coherente con el promedio de las aristas de su vecindad, lo que da cuenta que la información contenida en la ontología GO está relacionada con el índice KTA local. Por otro lado, de la figura 1.5d podemos verificar que no existe una dependencia fierte del KTA local con la cantidad de vecinos anotados $nyanotados$. Por lo tanto, el KTA local presenta valores más altos para vecinos con alta coherencia biológica, independientemente de la cantidad de vecinos que se tome en cuenta.

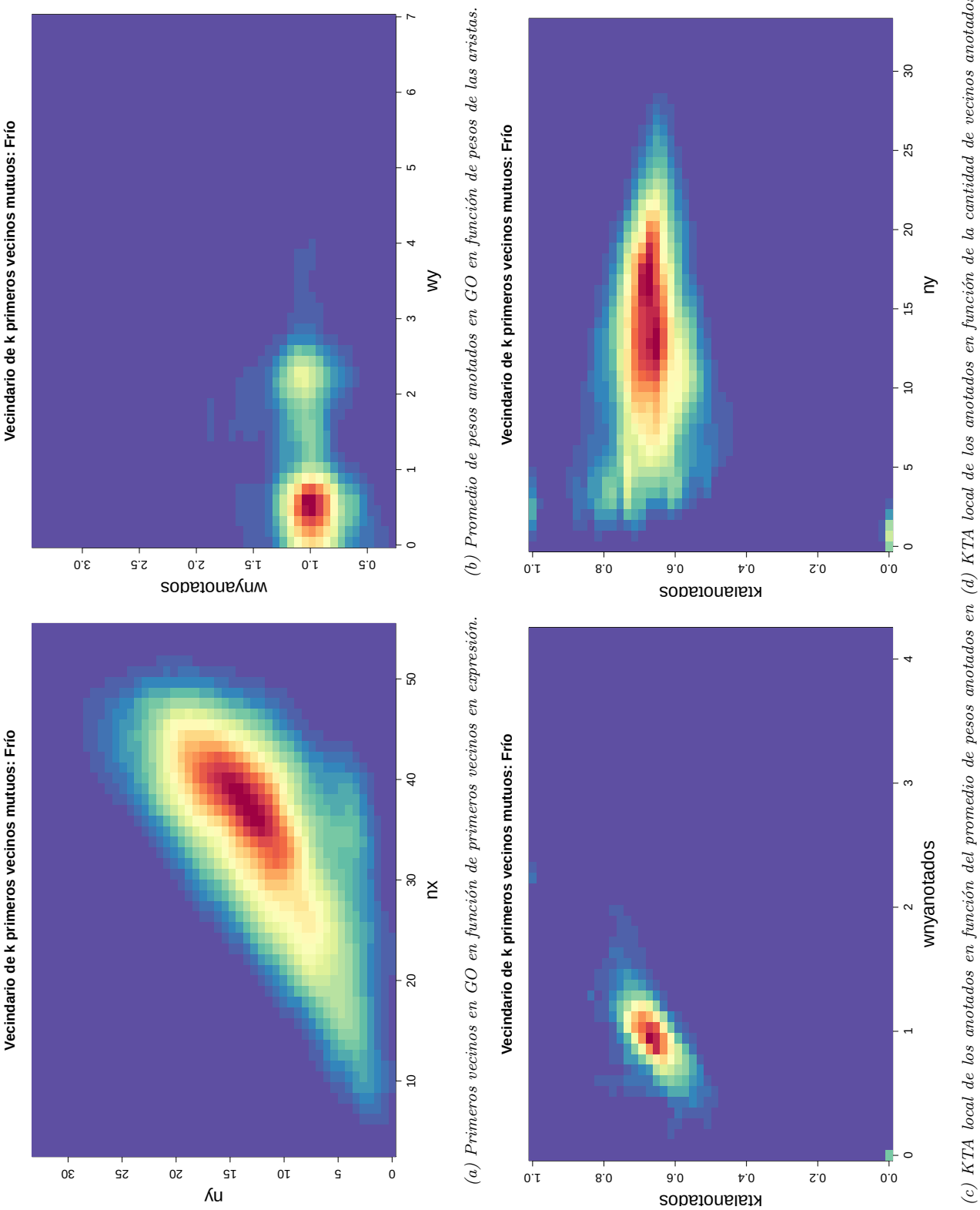
En el capitulo siguiente utilizaremos los resultados obtenidos en este capítulo para desarrollar una variante al procedimiento córte de árbol dinámico que integre la información contenida en el espacio de expresión con la información contenida en el espacio de conocimientos GO para favorecer la búsqueda de estructuras biológicamente coherentes.



(a) Distribución de grado para los nodos de la red 30*kmnn para el tratamiento 'Frío' (círculos vacíos) y para una red de Erdös-Rény (cuadrados amarillos) con idéntica cantidad de nodos y aristas.*

(b) Distribución de betweenness centrality en función del grado de la red para los nodos de las redes 30*kmnn para el tratamiento 'Frío', Erdös-Rényi y configuracional.*

Figura 1.4: Distribución de grado y de betweenness centrality para los nodos de la red 30*kmnn y modelos nulos*



(a) Primeros vecinos en GO en función de primeros vecinos en expresión.

(b) Promedio de pesos anotados en GO en función de pesos de las aristas.

(c) KTA local de los anotados en función del promedio de pesos anotados en GO.

(d) KTA local de los anotados en función de la cantidad de vecinos anotados en GO.

Figura 1.5: Caracterización de KTA local para tratamiento 'Frío'.

Bibliografía

[1] Nature.com. *Functional genomics*. Accedido: 2016-01-13.
URL `http://www.nature.com/subjects/functional-genomics`

[2] Arabidopsis-Interactome-Mapping-Consortium. *Evidence for Network Evolution in an Arabidopsis Interactome Map*. Annual review of plant biology **10** (2013) 161.

[3] M. Kanehisa. *Yeast Biochemical Pathways. KEGG: Kyoto encyclopedia of genes and genomes*. Nucleic Acids Res **28** (2000) 27.
URL `http://pathway.yeastgenome.org/biocyc/`

[4] arabidopsis.org. *org.At.tair.db*. `https://www.arabidopsis.org/biocyc/`.

[5] E. Domany. *Cluster Analysis of Gene Expression Data 1* **110** (2003) 1117.

[6] B. Alberts. *Molecular Biology of The Cell*, volume 6 (2015).

[7] B. Bose. *In Vitro Differention of Pluripotent Stem Cells into Functional B Islets Under 2D and 3D Culture Conditions and In Vivo Preclinical Validation of 3D Islets*. Methods in Molecular Biology (2016) 257.

[8] M. Babu. *An Introduction to Microarray Data Analysis*. Computational Genomics: Theory and Application (2004) 225.
URL `http://www.mrc-lmb.cam.ac.uk/genomes/madanm/microarray/chapter-final.pdf`

[9] A. Schulze. *Navigating gene expression using microarrays: a technology review*. Nature cell biology **3** (2001) E190.

[10] arabidopsis.org. *Microarray data from AtGenExpress*. `https://www.arabidopsis.org/portals/expression/microarray/ATGenExpress.jsp`.

[11] J. Kilian *et al. The AtGenExpress global stress expression data set: Protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses*. Plant Journal **50** (2007) 347.

[12] A. Brückner *et al. Yeast two-hybrid, a powerful tool for systems biology*. International Journal of Molecular Sciences **10** (2009) 2763.

[13] M. E. Cusick *et al. NIH Public Access*. Nature Methods **6** (2009) 39.

[14] G. Sales *et al. graphite: GRAPH Interaction from pathway Topological Environment* (2015). R package version 1.16.0.

[15] E. Segal *et al. Discovering molecular pathways from protein interaction and gene expression data*. Bioinformatics **19** (2003).

[16] J. Pandey *et al. Functional coherence in domain interaction networks*. Bioinformatics **24** (2008) 28.

[17] P. Resnik. *Using Information Content to Evaluate Semantic Similarity in a Taxonomy*. roceedings of the 14th international joint conference on Artificial intelligence - Volume 1 - IJCAI'95 **1** (1995) 6.
URL http://arxiv.org/abs/cmp-lg/9511007

[18] C. Pesquita *et al. Semantic similarity in biomedical ontologies*. PLoS Computational Biology **5** (2009).

[19] A. Berenstein. *Análisis de redes complejas en sistemas biomoleculares* (2014).

[20] Ashburner. *Gene ontology: tool for the unification of biology*. Nat Genet **25** (2000).

[21] G. Gan *et al. Data Clustering: Theory, Algorithms, and Applications*, volume 20 (2007).

[22] M. Halkidi *et al. On clustering validation techniques*. Journal of Intelligent Information Systems **17** (2001) 107.

[23] E. Domany. *Superparamagnetic clustering of data—the definitive solution of an ill-posed problem*. Physica A: Statistical Mechanics and its Applications **263** (1999) 158.
URL http://www.sciencedirect.com/science/article/pii/S0378437198004944

[24] S. Chen *et al. On the similarity metric and the distance metric*. Theoretical Computer Science **410** (2009) 2365.
URL http://dx.doi.org/10.1016/j.tcs.2009.02.023

[25] L. W. Kheng. *Image Registration* (2010).

[26] P. D'haeseleer. *How does gene expression clustering work?* Nat Biotech **24** (2005).

[27] C. Hennig. *How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification*. Journal of the Royal Statistical Society. Series C: Applied Statistics **62** (2013) 309.

[28] M. Eisen. *Cluster analysis and display of genome-wide expression patterns*. Proceedings of the National Academy of Sciences of the United States of America **95** (1998) 14863.

[29] H. K. Lee *et al. Coexpression Analysis of Human Genes Across Many Microarray Data Sets* (2004) 1085.

[30] J. Sevilla. *Correlation between gene expression and go semantic similarity*. In: IEEE/ACMTransactions on Computational Biology and Bioinformatics (2005).

[31] P. Lord. *Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation.* Bioinformatics (2003).

[32] D. Lin. *An Information-Theoretic Definition of Similarity.* In: Proc. of the 15th Internatio- nal Conference on Machine Learning (1998) 296.

[33] J. Jiang. *Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy.* Proceedings of International Conference Research on Computational Linguistics (1997) 19.

[34] J. Hartigan. *A K-Means Clustering Algorithm.* Journal of the Royal Statistical Society **28** (1979) 100.

[35] J. Kogan. *Introduction to Clustering Large and High-Dimensional Data* (2006).

[36] H. S. Park. *A simple and fast algorithm for K-medoids clustering.* Expert Systems with Applications **36** (2009) 3336.

[37] L. Ibrahim. *Using Modified Partitioning Around Medoids Clustering Technique in Mobile Network Planning* **9** (2012) 299.

[38] J. Stephen. *Hierarchical clustering schemes.* Psychometrika (1967).

[39] C. Shalizi. *Distances between Clustering , Hierarchical Clustering.* Data Mining (2009) 36.

[40] P. Langfelder *et al. Defining clusters from a hierarchical cluster tree: The Dynamic Tree Cut package for R.* Bioinformatics **24** (2008) 719.

[41] P. Langfelder *et al. Dynamic Tree Cut : in-depth description , tests and applications* (2007) 1.

[42] S. Horvath. *The Generalized Topological Overlap Matrix For Detecting Modules in Gene Networks.* bioinformatics (2007).

[43] A. Barabási. *Network biology: understanding the cell's functional organization.* Nat. Rev.Genet. (2004) 101.

[44] L. Hartwell *et al. From molecular to modular cell biology.* Nature (1999) 47.

[45] M. Rosvall. *Maps of random walks on complex networks reveal community structure.* Proceedings of the National Academy of Sciences of the United States of America **105** (2008) 1118.

[46] a. Clauset *et al. Finding community structure in very large networks.* Phys. Rev. E **70** (2004) 66111.
URL `http://prola.aps.org/abstract/PRE/v70/i6/e066111`

[47] J. Dutkowski *et al. A gene ontology inferred from molecular networks.* **31** (2013) 38.
URL `http://www.nature.com/nbt/journal/v31/n1/abs/nbt.2463.html\nhttp://www.pubmedc`

[48] S. Datta. *Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes.* BMC bioinformatics **7** (2006) 397.
URL `http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1590054{&}tool=pmcer`

[49] N. Cristianini *et al. On kernel target alignment.* Studies in Fuzziness and Soft Computing **194** (2006) 205.

[50] Y. Chen *et al. Learning kernels from indefinite similarities.* Proceedings of the 26th Annual International Conference on Machine Learning - ICML 2009 (2009) 1.

[51] P. Erdos *et al. On random graphs.* Publicationes Mathematicae (1959) 290.

[52] M. Molloy *et al. A critical-point form random graphs with a given degree sequence.* Random structures and Algorithms (1995) 161.

[53] A. J. Berenstein. *Técnicas de Mecánica estadística para la detección de correlaciones en perfiles de expresión génica.* Tésis de Licenciatura en Ciencias Físicas (2010).

[54] S. Horvat. *A general framework for weighted gene co-expression network analysis.* Stat. Appl. Genet. Mol. Biol. (2005).