Capítulo 1

Análisis de conjunto de datos transcripcionales Wiegel

En este capítulo analizaremos el conjunto de datos transcripcionales Wiegel & Lohmann para la planta Arabidopsis thaliana presentados en la sección **??**, utilizando para ello los métodos de agrupamiento k-means (sección **??**) y corte de árbol dinámico híbrido (sección **??**) introducidos en el capítulo **??** para obtener grupos en el espacio de expresión.
Una vez obtenidos los grupos en el espacio de expresión, utilizaremos los índices BHI e Interacting Densities para cuantificar el grado de coherencia entre estas estructuras y los conocimientos (entendidos como nociones de similitud) en el espacio GO.
Luego, analizaremos la coherencia de los resultados obtenidos en el espacio de expresión con la de resultados obtenidos en otros espacios de conocimiento, como GO (sección **??**), PIN (sección **??**) y KEGG (sección **??**), esperando que estos conocimientos sean diferentes pero no ortogonales, utilizando para ello el índice KTA.

1.1. Descripción del dataset

esto esta en sec:wiegel habra que profundizar mas?

1.2. Métricas transcripcionales

esto esta en el capitulo 3, o la idea es poner otra cosa?

1.3. Agrupamiento

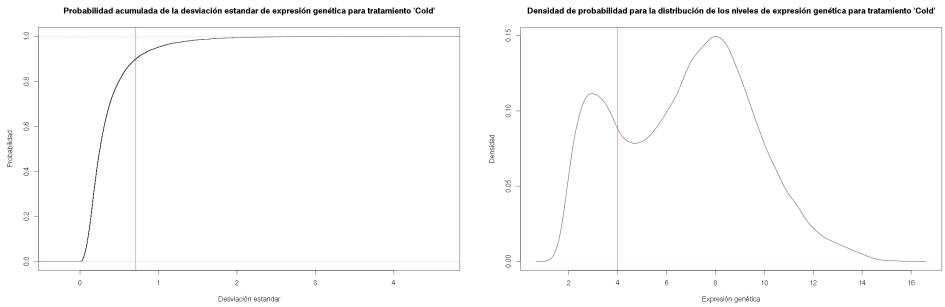
1.3.1. Proceso de filtrado y estandarización de datos

El conjunto de datos Wiegel utilizado consta de los niveles de expresión de 22810 sondas que se mapean a 20149 genes a lo largo de 11 tratamientos diferentes y con entre 4 y 9 muestreos en dos réplicas. Para poder manejar esta cantidad de información es necesario realizar un filtrado (una selección) previo de los datos que permita quedarse únicamente con aquellos genes que se expresaron o inhibieron, ya que serán estos los genes que estarán siendo regulados en función del tratamiento y por lo tanto los de interés.

Para ello, se aplicaron dos tipos de filtros por tratamiento, por desviación estandar y por de tipo "$KsobreA$". Para el primero, se calculó la desviación estandar por gen a lo largo de todo el tratamiento y se decidió tomar los genes cuya desviación estandar se encontrara en el cuantil 0.9, es decir, utilizar el 10 % de los genes con mayor desviación estandar, considerando estos como los que formaron parte de la respuesta biológica al tratamiento. La figura 1.1a muestra la distribución de probabilidad acumulada (empírica) de la desviación estandar para los genes del tratamiento "Cold".

Una vez aplicado este filtro por desviación estandar, se aplicó un filtro de tipo "$KsobreA$", que toma unicamente con aquellos genes que tengan al menos K datos por encima del valor A. En nuestro caso, decidimos utilizar como valor de K, la mitad de las mediciones que tuviera el tratamiento. Si el tratamiento tenía mediciones cada 0 minutos, 30 minutos, 1 hora, 3 horas, 6 horas, 12 horas y 24, es decir, 6 mediciones en total, se tomó $K = 3$. Para A, se decidió utilizar una medida usual de $A = 4$, ya que valores de señal menores a 4 no se distinguen del ruido paper sobre esto? cuales son las unidades de estos datos? son en escala logaritmica?. La figura 1.1b muestra la distribución de probabilidad para los niveles de expresión para el tratamiento "Cold". Una vez aplicados los fitros y obtenido los genes de mayor variabilidad en su expresión, se estandarizaron los datos obtenidos para poner a todos los genes en igualdad de condiciones y pesarlos de la misma forma en el agrupamiento. Un procedimiento normal para estandarización de genes implica realizar la transformación:

$$\tilde{x_i} = \frac{x_i - \bar{x}}{\sigma_x} \tag{1.1}$$



(a) Distribución de probabilidad acumulada de la desviación estandar para los genes del tratamiento *Cold*. Todos los genes con desviación estandar menor que la indicada por la recta vertical roja son descartados.

(b) distribución de probabilidad para los niveles de expresión para el tratamiento *Cold*. La recta vertical roja muestra el valor a partir del cual se hace un corte.

Figura 1.1: Funciones de distribución de probabilidad para perfiles de expresión

1.3.2. Agrupamiento con k-means

1.3.3. Agrupamiento con dynamic tree cut

1.3.4. Análisis de los métodos y problemas de escala de resolución

1.4. Coherencia entre la métrica transcripcional y otros espacios de conocimiento

idea esperamos que los conocimientos (entendidos como nociones de similitud) de los distintos espacios sean diferentes pero no ortogonales...cuantificacion...veamos que estructuras son en cierto grado coherentes

1.4.1. Interacting densities

genex1 /genex4 VS BPa/BPb/CC PINinfomap / KEGGinfomap/LCI para referencia

1.4.2. Test de fisher

genex1 /genex4 VS BPa/BPb/CC

1.4.3. KTA y zKTA

Global KTA Genex por tratamiento + PIN + KEGG + LCI / GOBPa, GOBPb, GOCC zKTA: por tratamiento Gx/GOBPa, Gx/GOBPb, Gx/GOCC, Gx/PIN, Gx/LCI, Gx/Kegg

Bibliografía

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

[2] Wikipedia.org. *Functional genomics*. Accedido: 2016-01-13.
URL https://en.wikipedia.org/wiki/Functional-genomics

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

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

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

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

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

[8] 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.

[9] 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

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

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

[12] 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.

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

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

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

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

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

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

[19] 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`

[20] 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`

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

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

[23] 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.

[24] 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.

[25] 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`

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

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

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

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

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

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

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

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

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

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

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

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

[41] 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.

[42] 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`

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