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

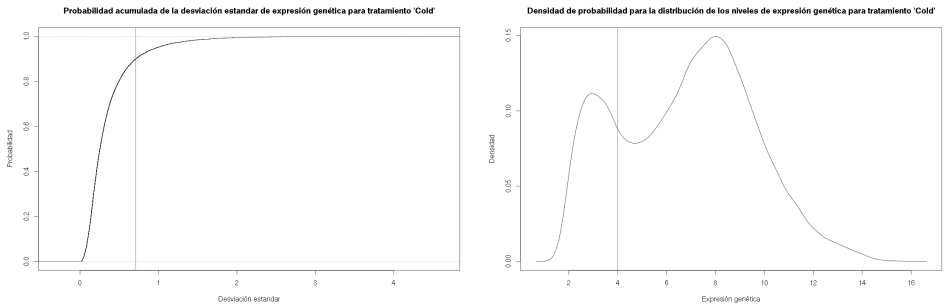
El conjunto de datos Wiegel utilizado consta de los niveles de expresión de 22810 sondas que se mapean a 20149 genes a lo largo de 11 tratamientos diferentes y con entre 4 y 9 muestreos en dos réplicas. Para poder manejar esta cantidad de información es necesario realizar un filtrado (una selección) 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". La tabla 1.1 muestra los filtros aplicados y la cantidad de genes finales por tratamiento. Una vez aplicados los filtros 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 de estandarización de genes para que cada gen tenga media cero y varianza unitaria implica realizar la transformación:

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

Con x_i cada observación del gen x a lo largo del tiempo para un determinado tratamiento. Una vez realizado el filtrado y estandarizado procedimos a agrupar los datos mediante los diferentes métodos mencionados en el capítulo 3.



(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

El método de agrupamiento k-means hace uso de la distancia euclidia para minimizar la suma de los cuadrados. Si los datos están estandarizados y centrados, es posible relacionar la distancia euclidia d con el coeficiente de correlación mediante la fórmula:

$$d(\vec{x}, \vec{y}) = \sqrt{2(d-1)(1 - r(\vec{x}, \vec{y}))} \tag{1.2}$$

y por lo tanto, para datos estandarizados, la distancia euclidia se comportará de forma similar a la distancia de correlación y podremos utilizar el método k-means. revisar esta frase y que quiero decir

Para decidir el k a utilizar en el método, se realizó un barrido variando k entre $k = 2$ y $k = 30$ con pasos de 1. Al tratarse de un método heurístico, no existe garantía de convergencia al óptimo global y el resultado del mismo puede entonces depender de los grupos iniciales. Por lo tanto, para cada k, se realizaron cien agrupamientos y se midieron los índices de validación internos Calinski-Harabasz y Dunn en cada uno, definidos respectivamente como:

$$CH_k = \frac{SS_B}{SS_W} \frac{n-k}{n-1} \tag{1.3}$$

con SS_B el promedio de la varianza entre grupos, SS_W el promedio de la varianza intra grupos, k es el número de grupos y n el número de observaciones y:

$$DI = \frac{min\delta}{\text{máx}\,\Delta} \tag{1.4}$$

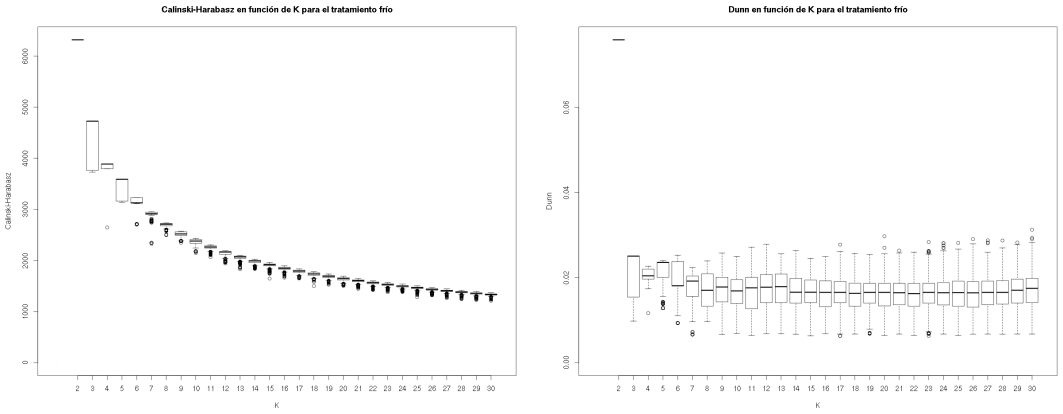
Tratamiento	σ	A	Cantidad de genes
Control	0.37	4	1885
Frío	0.71	3	1955
Osmótico	0.71	3	1923
Sal	0.88	3	1927
Sequía	0.54	4	1870
Genotóxico	0.46	3	1899
Oxidativo	0.41	3	1880
UV-B	0.51	4	1872
Heridas	0.41	4	1877
Calor	0.75	2	1960
Calor y recuperación	0.65	2	1944

Cuadro 1.1: Cantidad de genes y filtros utilizados por tratamiento.

con δ la menor de las de distancias entre grupos y Δ la mayor de las distancias intra grupos.
Grupos bien definidos tendrán distancias grandes entre ellos comparados con las distancias intra grupos, por lo que a mayor CH o DI, mejor definidos estarán los grupos. Las figuras 1.2a y 1.2b muestran un gráfico de caja (o boxplot en inglés), para el índice CH y Dunn respectivamente para cada uno de los k en el barrido. Un boxplot consiste en una caja con una linea horizontal que indica el segundo cuartil, es decir, la mediana del conjunto de datos, y dos lineas verticales llamadas bigotes (o whiskers en inglés) que se extiende una desde el primer cuartil hasta el valor más pequeño del conjunto (con excepción de puntos aislados) y la otra desde el tercer cuartil hasta el valor más grande. Los puntos aislados se grafican de forma separada en el gráfico. Se observa que la cantidad de grupos que maximiza estos índices es 2. Se realizó entonces un agrupamiento con $k = 2$, obteniéndose los perfiles que muestra la figura 1.3, con una correlación media de $\rho = 0{,}74$ para el primero y de $\rho = 0{,}79$ para el segundo, con aproximadamente el 50 % de los genes en cada grupo. Estas estructuras tan grandes son de difícil interpretación biológica, ya que si bien las respuestas de expresión dentro de cada grupo son similares, existe mucha heterogeneidad en las funciones biológicas de los genes que los componen. El método k-means está entonces trabajando a una escala que no permite extraer información biológica de los grupos. Será necesario entonces aumentar la granularidad mediante otros métodos de agrupamiento.

1.3.3. Agrupamiento con corte de árbol dinámico

Utilizando el método de corte de árbol dinámico se realizó un agrupamiento para cada tratamiento, utilizando alternativamente los parámetros $deepSplit = 1$ (menor



(a) Índice CH de particiones realizadas con k-means para k entre 2 y 30.

(b) Índice Dunn de particiones realizadas con k-means para k entre 2 y 30.

Figura 1.2: Índices de validación interna para particiones realizadas con k-means

granularidad) y $deepSplit = 4$ (mayor granularidad). Las figuras 1.4a y 1.4b muestran algunos de los perfiles obtenidos con cada parámetro respectivamente para el tratamiento "Frío".
En general, para todos los tratamientos, los grupos obtenidos por este método tienen mayor correlación media (ρ)que los obtenidos por el método k-means, obteniéndose una mayor cantidad de grupos con el $deepsplit = 4$ que con $deepsplit = 1$.

Para cada parámetro, cada tratamiento y cada grupo, se realizó un control nulo consistente en tomar la misma cantidad de genes presentes en el grupo, pero de forma aleatoria, del conjunto de genes que formaban el tratamiento, y medir su correlación media. Esto se realizó 1000 veces para cada grupo. Las figuras **??** y **??** muestran la correlación media por tamaño de grupo y el control nulo para $deepsplit = 1$ y $deepsplit = 4$ respectivamente. Los grupos fueron agrupados por tamaño de a 10 genes, donde los colores más claros indican mayor cantidad de grupos que los oscuros. El gráfico tiene además la media, en negro, y el segundo y tercer cuartil, en gris, para la distribución del control nulo.

Se observa que la correlación media de los grupos es en todos los casos superior a la del control nulo. Esto muestra que existe estructura en los grupos hallados para ambos parámetros.
Por otro lado, en las figuras **??** y **??** se observa que $deepsplit = 1$ llega a tener grupos de mayor tamaño que $deepsplit = 4$. Esto es esperable ya que cada parámetro aumenta o disminuye la granularidad del método. Estos grupos de $deepsplit = 1$ comparativamente grandes tienen alta correlación. Sin embargo hay una menor correlación en los grupos pequeños para $deesplit = 1$ que para $deesplit = 1$. Una posible explicación para esto es

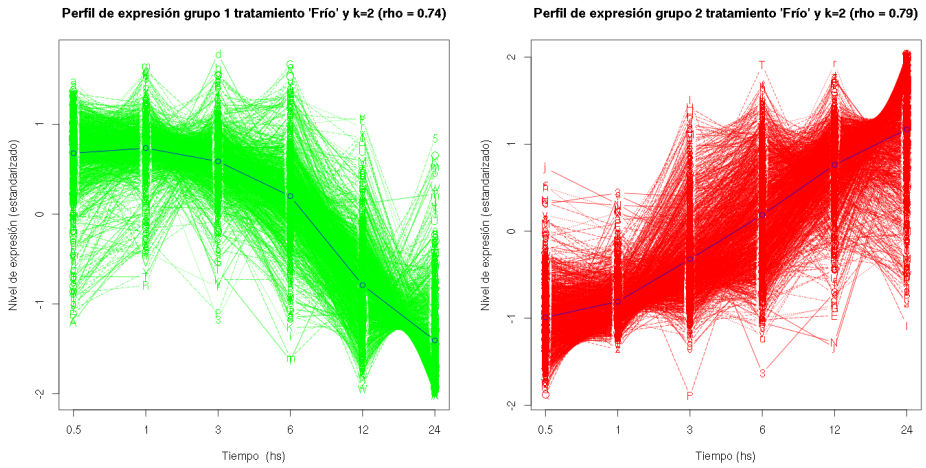
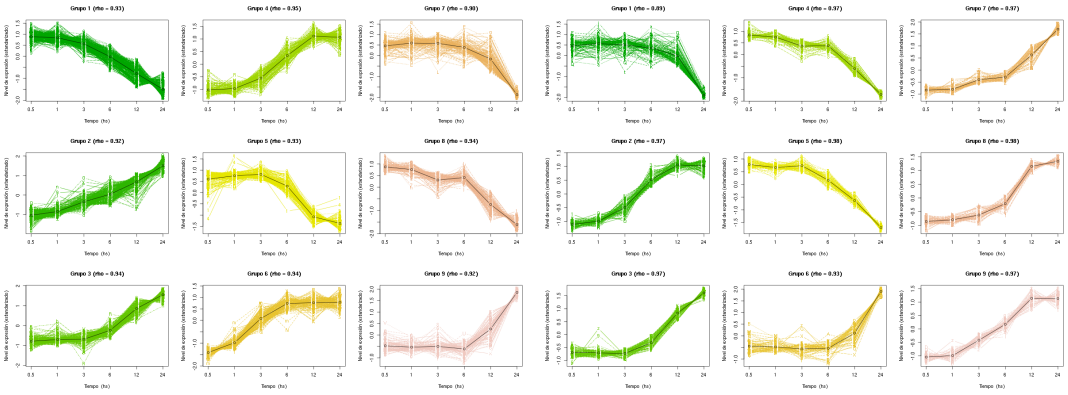


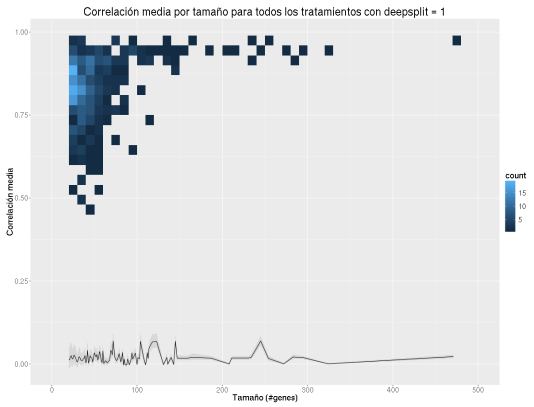
Figura 1.3: Perfiles de expresión génica obtenidos con el método k-means (k=2) para el tratamiento 'Frío'. En azul, el valor medio de cada grupo.



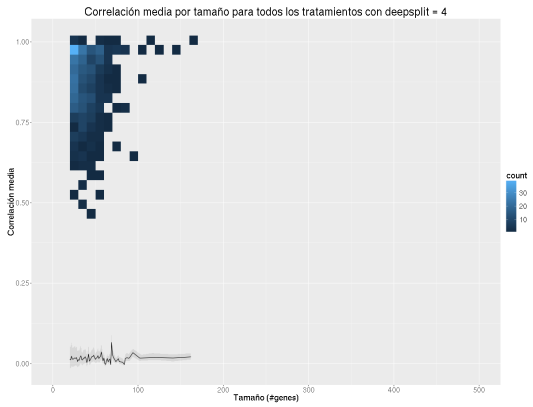
(a) Perfiles obtenidos con *deepsplit* = 1. (b) Perfiles obtenidos con *deepsplit* = 4.

Figura 1.4: Perfiles de expresión génica obtenidos con el método corte de árbol dinámico para *deepsplit* = 1 y *deepsplit* = 4 para el tratamiento 'Frío'. En negro, el valor medio de cada grupo.

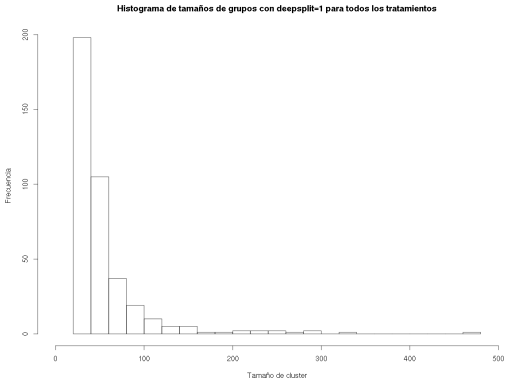
que para que exista un grupo grande, es necesario que el mismo tenga alta correlación. De lo contrario, el método buscará partirlo en grupos más chicos hasta maximizar la correlación de cada grupo.
Otra forma de visualizar esta diferencia en los tamaños de los grupos que obtiene cada método es mediante la función de distribución acumulada empírica que se observa en la figura **??**. En la misma se observa que corte de árbol dinámico con *deepsplit* = 4 produce la mayor cantidad de grupos con los menores tamaños, seguida por la misma técnica pero con *deepsplit* = 1 y finalmente por k-means con solamente dos grupos muy



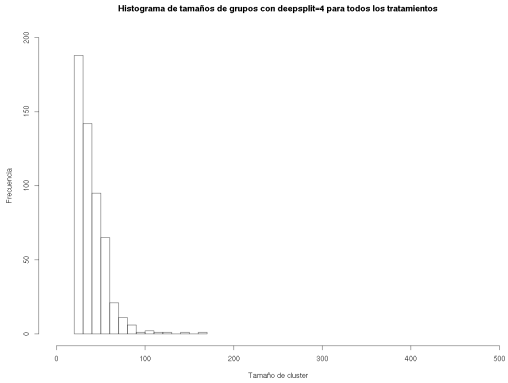
(a) Correlación media por tamaño de grupo para los grupos obtenidos con $deepsplit = 1$.



(b) Correlación media por tamaño de grupo para los grupos obtenidos con $deepsplit = 4$.



(c) Histograma por tamaño de grupo para los grupos obtenidos con $deepsplit = 1$.



(d) Histograma por tamaño de grupo para los grupos obtenidos con $deepsplit = 4$.

Figura 1.5: Correlación media por tamaño de grupo para los grupos obtenidos por corte de árbol dinámico con $deepsplit = 1$, $deepsplit = 4$ y control nulo para todos los tratamientos y sus respectivos histogramas

masivos.

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

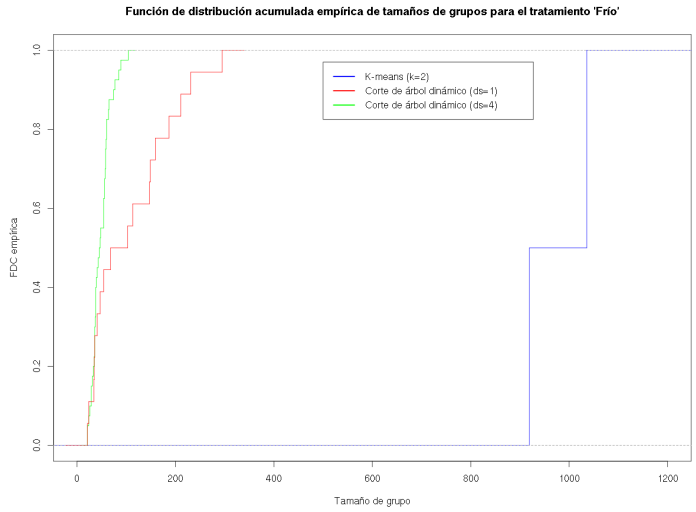


Figura 1.6: Función de distribución acumulada empírica para los métodos k-means (k=2) y corte de árbol dinámico (ds=1 y ds=4) para el tratamiento 'Frío'.

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. 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