Capítulo 1

Métodos de agrupamiento de datos

Un método de agrupamiento de datos o método de "clustering", es un método de clasificación no supervisado que permite la partición de un conjunto de N objetos en K grupos o clases, de tal forma que los objetos miembro de un grupo sean más similares entre si (en algún sentido a definir) que entre los miembros de otros grupos.
Son métodos no supervisados ya que en un proceso de agrupación no existen clases definidas previamente ni ejemplos de que tipo de relaciones se desea encontrar entre los objetos, por lo que el mismo proceso debe generar las clases iniciales a las cuales asignar los objetos en el proceso de clasificación.
Estas técnicas permiten el descubrimiento o identificación de distribuciones y patrones subyacentes en los datos, posibilitando obtener conclusiones sobre los mismos, lo que las hace una de las herramientas más útiles en procesos de minería de datos y aprendizaje automatizado en campos tan diversos como las ciencias sociales, las ciencias médicas y la ingeniería.
Dependiendo de los criterios utilizados para realizar la partición, un proceso de agrupamiento puede resultar en diferentes particiones. Como ejemplo de esto podemos tomar el conjunto de números $\{-5, -3, -2, 2, 3\}$. Si decidimos agruparlos por su módulo, obtendremos los conjuntos $\{-5\}$, $\{-3, 3\}$, $\{-2, 2\}$, mientras que si decidimos agruparlos por positividad o negatividad, obtendremos los conjuntos $\{-5, -3, -2\}$ y $\{2, 3\}$. También podríamos haber optado por agrupar por paridad, si son o no primos, etc. Como se observa de un ejemplo tan sencillo, es de fundamental importancia la elección de las propiedades de los objetos a partir de las cuales realizar el agrupamiento.
En el presente trabajo nos interesará agrupar y caracterizar conjuntos de genes de un organismo modelo, la planta *Arabidopsis thaliana*, en base a sus perfiles de expresión génica a lo largo de diversos tratamientos. [**?**, **?**, **?**]
Discutiremos a continuación diferentes metodologías y criterios de similaridad que pueden ser considerados para ello.

1.1. Similaridad, distancia y disimilaridad

Las distancias y similaridades tienen un rol preponderante en el análisis de agrupamiento de datos y por regla general son conceptos recíprocos.
Una medida de similaridad o coeficiente de similaridad se utiliza para indicar de forma cuantitativa la fuerza de la relación entre dos objetos del conjunto. Los $i = 1, 2, ...N$ objetos de un conjunto E pueden ser definidos en términos de las coordenadas $\vec{X}i$ de sus puntos representativos en un espacio $d - dimensional$. Sean $\vec{x} = \{x_0, x_1, ..., x_d\}$ e $\vec{y} = \{y_0, y_1, ..., y_d\}$ dos puntos $d-dimensionales$. Entonces, el coeficiente de similaridad entre ambos será una función de sus atributos:

$$s(\vec{x}, \vec{y}) = s((x_0, x_1, ..., x_d), (y_0, y_1, ..., y_d)) \tag{1.1}$$

con s una función simétrica, es decir, $s(\vec{x}, \vec{y}) = s(\vec{y}, \vec{x})$. Cuanto mayor es el coeficiente de similaridad, mayor es la similaridad entre ambos.
Por otro lado, las medidas de disimilaridad o de distancia se comportan de forma inversa, a mayor distancia o disimilaridad, más diferentes son dos puntos. Una métrica de distancia es una función $d \in R$ definida sobre un conjunto E que cumple las siguientes propiedades:

1. No-negatividad: $d(\vec{x}, \vec{y}) \geq 0$
2. Reflexividad: $d(\vec{x}, \vec{y}) = 0 \iff \vec{x} = \vec{y}$
3. Conmutatividad: $d(\vec{x}, \vec{y}) = d(\vec{y}, \vec{x})$
4. Desigualdad triangular: $d(\vec{x}, \vec{y}) \leq d(\vec{x}, \vec{z}) + d(\vec{z}, \vec{y})$

con $\vec{x}, \vec{y}, \vec{z}$ objetos arbitrarios del conjunto.

Una medida de disimilaridad es una métrica si cumple con las propiedades antes enunciadas.
Aunque no pareciera existir una definición formal de métrica de similaridad, Chen y colaboradores definen una métrica de similaridad como una función s que cumple:

1. $s(\vec{x}, \vec{y}) = s(\vec{y}, \vec{x})$
2. $s(\vec{x}, \vec{x}) \geq 0$
3. $s(\vec{x}, \vec{x}) \geq s(\vec{x}, \vec{y})$
4. $s(\vec{x}, \vec{x}) = s(\vec{y}, \vec{y}) = s(\vec{x}, \vec{y}) \iff x = y$
5. $s(\vec{x}, \vec{y}) + s(\vec{y}, \vec{z}) \leq s(\vec{x}, \vec{z}) + s(\vec{y}, \vec{y})$

La condición 5 indica que la similaridad entre $\vec{x}$ y $\vec{z}$ a través de $\vec{y}$ no es mayor que la similaridad directa entre $\vec{x}$ y $\vec{z}$ sumada a la autosimilaridad de $\vec{y}$. Esta propiedad es el equivalente de la desigualdad triangular para una distancia métrica.
Si bien es deseable que una similaridad o disimilaridad sea una métrica, existen muchas medidas de similaridad o disimilaridad que dan excelentes resultados en técnicas de agrupamiento de datos sin ser métricas, es decir, sin que necesariamente cumplan la desigualdad triangular o el item 5 de métrica de similaridad. [**?**]
Finalmente, los objetos del conjunto pueden ser especificados por medio de una "matriz de distancia" de $N \times N$ cuyos elementos d_{ij} indican la disimilaridad entre los puntos i y j. [**?**, **?**, **?**, **?**]

1.1.1. Medidas de distancia

El análisis de datos de expresión genética se basa principalmente en la comparación de perfiles de expresión génica. Para poder comprarlos, se requiere una medida que cuantifique cuan similares o disimilares son los objetos considerados. La elección de una medida de distancia será entonces de fundamental importancia para lograr agrupamientos que tengan sentido en el contexto de los datos analizados. En las subsiguientes secciones se listarán las medidas de distancia más comúnmente utilizadas en el agrupamiento de datos (no necesariamente de datos de perfiles de expresión).

Distancia euclidiana

La distancia euclidiana es probablemente la distancia más utilizada en el contexto de datos numéricos. Para dos puntos $\vec{x}$ e $\vec{y}$ en un espacio $d - dimensional$, la distancia euclidiana se define como:

$$d_{euc}(\vec{x}, \vec{y}) = [\sum_{i=1}^{d}(x_i - y_i)^2]^{\frac{1}{2}} = [(\vec{x} - \vec{y})(\vec{x} - \vec{y})^T]^{\frac{1}{2}} \tag{1.2}$$

con x_i e y_i los valores de la $i$esima componente de $\vec{x}$ e $\vec{y}$ respectivamente.

Distancia Manhattan o Taxicab

La distancia Manhattan o taxicab es llamada así por ser la distancia que debería recorrer un taxi en una ciudad para ir de un punto a otro, suponiendo la ciudad como una cuadrícula perfecta. Para dos puntos $\vec{x}$ e $\vec{y}$ en un espacio $d - dimensional$, la distancia Manhattan se define como:

$$d_{man}(\vec{x}, \vec{y}) = \sum_{i=1}^{d} |(\vec{x} - \vec{y})| \tag{1.3}$$

Distancia máxima

Para dos puntos $\vec{x}$ e $\vec{y}$ en un espacio $d-dimensional$, la distancia máxima se define como:

$$d_{max}(\vec{x},\vec{y}) = \max_{1\leq j\leq n} |x_i - y_i| \tag{1.4}$$

Distancia de Minkowsky

Para dos puntos $\vec{x}$ e $\vec{y}$ en un espacio $d-dimensional$, la distancia de Minkowsky se define como:

$$d_{mink}(\vec{x},\vec{y}) = [\sum_{i=1}^{d}(x_i - y_i)^r]^{\frac{1}{r}}, r \geq 1 \tag{1.5}$$

r es el orden de la distancia de Minkowsky. Notar que si tomamos $r = 2, 1, \text{ínf}$ obtenemos la distancia euclidiana, la Manhattan y la máxima, respectivamente.

Coeficiente de correlación de Pearson

Una de las métricas más utilizadas para medir similaridad entre perfiles de expresión, como los presentados en la figura **??**, es el coeficiente de correlación de Pearson [**?**]. El coeficiente de correlación fue desarrollado por Karl Pearson basado en ideas introducidas por Francis Galton alrededor del año 1880.

Para dos puntos $\vec{x}$ e $\vec{y}$ en un espacio $d-dimensional$, representando el perfil de expresión de dos genes a lo largo de un dado tratamiento, el CCP se define como:

$$r(\vec{x},\vec{y}) = \frac{\sum_{i=1}^{d}(x_i - \bar{x})(y_i - \bar{y})}{[\sum_{i=1}^{d}(x_i - \bar{x})^2]^{\frac{1}{2}}[\sum_{i=1}^{d}(y_i - \bar{y})^2]^{\frac{1}{2}}} \tag{1.6}$$

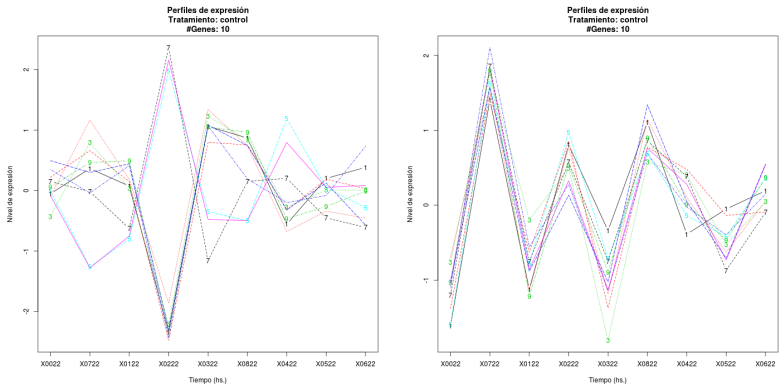
o de forma equivalente:

$$r(\vec{x},\vec{y}) = \frac{\frac{1}{d-1}\sum_{i=1}^{d}(x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \tag{1.7}$$

con s la desviación estandar de la muestra. El centrar alrededor de la media permite comparar la forma de ambos perfiles, en lugar de su magnitud.

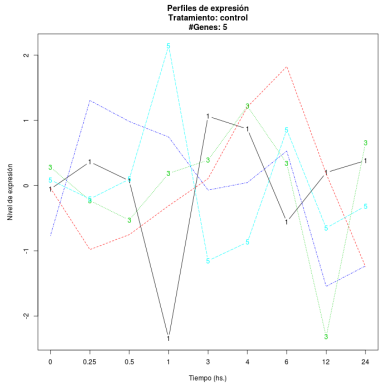
Valores altos de r implican que las fluctuaciones respecto de la media de las respectivas componentes se encuentran "en sincronia". En caso de no estarlo el valor esperado tiende a cero. Finalmente si las fluctuaciones tienden a ocurrir "sincronizadamente" pero en sentidos opuestos $r \Rightarrow -1$.

En nuestro caso, $r = +1$ corresponde a genes que estan siendo coexpresados por la



(a) Perfiles de expresión para el tratamiento *Control* de 10 genes que están co-regulados y anti co-regulados.

(b) Perfiles de expresión para el tratamiento *Control* de 10 genes que están co-regulados.



(c) Perfiles de expresión para el tratamiento *Control* de cinco genes tomados al azar.

Figura 1.1: Distintos grupos de perfiles de expresión

maquinaria celular, mientras que $r = -1$ a genes anti-coexpresados. En el primer caso los perfiles de ambos genes (apropiadamente reescaleados) coinciden perfectamente, mientras que en el segundo son perfectamente opuestos.
La correspondiente medida de distancia puede ser calculada como [**?**]:

$$d_{ccp}(\vec{x}, \vec{y}) = 1 - r(\vec{x}, \vec{y}) \tag{1.8}$$

o alternativamente:

$$d_{ccp}(\vec{x}, \vec{y}) = 1 - |r(\vec{x}, \vec{y})| \tag{1.9}$$

En el caso de la distancia definida en **??**, al tomar el valor absoluto del CCP, genes cuyos perfiles son iguales pero opuestos (están anti-coexpresados) pueden encontrarse

más cerca en el sentido de d_{ccp} que aquellos que son expresados hacia arriba o abajo pero en distintas magnitudes. Por lo tanto, esta distancia permite encontrar grupos de genes que son coexpresados, sin importar en que sentido (Figura **??**) sean coexpresados. En el caso de la distancia definida en **??**, solamente se consideran cercanos aquellos genes cuyos perfiles sean coexpresados o bien hacia arriba o bien hacia abajo (Figura **??**). [**?**, **?**, **?**, **?**]

En el presente trabajo se utilizará como distancia la definida en **??** para encontrar grupos de genes que únicamente se hayan coexpresado o bien hacia arriba o bien hacia abajo. [**?**]

1.1.2. Similaridad semántica

La adopción de ontologías provee los medios para comparar aspectos de entidades que de otra forma no podrían ser comparados. Por ejemplo, si dos productos génicos son anotados dentro del mismo esquema, es posible compararlos mediante el análisis de los términos en los cuales están anotados de forma explícita utilizando medidas de similaridad semántica. Se define una medida de similaridad semántica como una función tal que dados dos términos de la ontología o un conjunto de términos en los que dos genes están anotados, la función devuelve un escalar que refleja la cercanía de sentido entre ellos.
Es posible cuantificar la similaridad semántica en una ontología representada por un grafo como GO, mediante diversas estrategias.

Comparación de términos en GO

Existen esencialmente dos formas distintas de comparar términos en GO: comparación a partir de los arcos del grafo y comparación a partir de los nodos y sus propiedades. En este trabajo estaremos interesados únicamente en comparar términos a partir de sus nodos, ya que son los nodos los que contendrán la información biológica en forma de anotaciones génicas.
Sea C el conjunto de todos los términos de una ontología GO, con un número total $\#C$ de anotaciones. Un término c_i tendrá $\#c_i$ anotaciones, ya sea directamente o por intermedio de cualquiera de sus hijos. La probabilidad de que un gen tomado al azar, sin otro tipo de información, se encuentre anotado al concepto c_i será entonces $P(c_i) = \frac{\#c_i}{\#C}$, con $P : C \Rightarrow [0 : 1]$.
Se define el contenido de información de c_i como $IC = -log_2(P(c_i))$, cantidad en el intervalo $(0, -log_2[\frac{1}{\#C}])$, que indica cuan específico e informativo es un término de la ontología. Para un c_i y c_j tales que $c_i \preceq c_j$, se tiene que $IC(c_i) \geq IC(C_j)$. Cuanto más específico sea un término, es menos probable que un gen dado esté anotado en el mismo, y por lo tanto, su contenido de información es mayor. El nodo raíz de la ontología tiene un contenido de información nulo, ya que es el ancestro de todos los términos de la misma y por lo tanto, saber que un concepto está anotado a la raíz no aporta información.
Si bien el IC puede tener un sesgo, ya que términos en áreas actuales de interés en investigaciones biomédicas van a estar más anotados que otros términos en otras áreas, la utilización del IC sigue teniendo un sentido desde el punto de vista de la probabilidad, porque es mucho más probable (y menos significativo) que dos genes compartan un término frecuentemente usado, que uno no tan frecuente, más allá de si ese término es frecuente porque sea genérico o porque sea un término de interés para la investigación actual.
Es posible definir una medida entre pares de términos utilizando el IC. Existen dos

formas para ello: tomar el ancestro común más informativo (MICA, por sus siglas en inglés), en donde solo el ancestro común con mayor IC es considerado, o tomar el ancestro disjunto común (DCA por sus siglas en inglés), en la cual todos los ancestros comunes disjuntos (ancestros que no tienen ancestros comunes) son considerados.
Una de las medidas de similaridad semántica más comúnmente utilizadas es la medida de similaridad semántica introducida por Resnik en [**?**], que consiste en asignar como la medida de similaridad entre dos términos, el contenido de información del ancestro en común más informativo (el MICA):

$$Sim_{res}(c_i, c_j) = \max_{c \in S(c_i, c_j)} (-log_2[P(c)]) = IC(MICA[c_i, c_j]) \tag{1.10}$$

Con $S(c_i, c_j)$ el conjunto de ancestros comunes de $c_i y c_j$. De esta manera, para cuantificar la información compartida (y estimar entonces su similaridad semántica) se considera el contenido de información de los ancestros en común que dos términos poseen.
A modo de ejemplo, tomemos el DAG de la figura **??**, con 9 términos o conceptos: $C = \{R, c_0, ..., c_7\}$ y con 5 entidades mapeadas (genes anotados): $g_1 = \{5, 6, 2, 0, r\}$, $g_2 = \{5, 4, 2, 3, 0, r\}$, $g_3 = \{7, 1, r\}$, $g_4 = \{4, 3, 0, r\}$ y $g_5 = \{2, 0, r\}$. Podemos calcular la similaridad semántica de Resnik entre los términos c_4 y c_5, por ejemplo, sabiendo que $\#C = 5$ y que el ancestro común más informativo de ambos es c_0, con $\#c_0 = 4$. Se tiene entonces que $Sim_{res}(c_4, c_5) = IC(MICA) = IC(c_0) = -log_2(\frac{\#c_0}{\#C}) = -log_2(\frac{4}{5}) =$ 0,32. Si quisieramos calcular ahora la similaridad semántica Resnik entre c_5 y c_6, obtendríamos $Sim_{res}(c_5, c_6) = IC(c_2) = -log_2(\frac{3}{5}) = 0{,}73$. Por lo tanto, para Resnik, los conceptos c_5 y c_6 son entre sí, más similares que los conceptos c_4 y c_5.
Al considerar solo el IC del MICA, la Sim_{res} no tiene en cuenta la especificidad de los

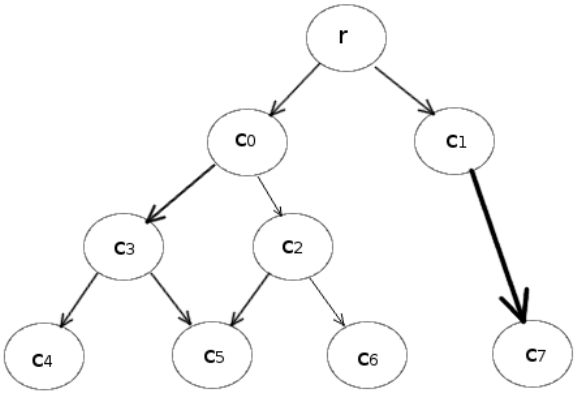


Figura 1.2: DAG con 9 términos o conceptos: $C = \{R, c_0, ..., c_7\}$ y con 5 entidades mapeadas (genes anotados): $g_1 = \{5, 6, 2, 0, r\}$, $g_2 = \{5, 4, 2, 3, 0, r\}$, $g_3 = \{7, 1, r\}$, $g_4 = \{4, 3, 0, r\}$ y $g_5 = \{2, 0, r\}$

términos que compara, es decir, no toma en cuenta la distancia entre los términos y su

MICA. Para tomar en cuenta esta distancia, las medidas de Lin [**?**] y Jiang-Conrath [**?**] relacionan el IC del MICA con el IC de los términos a comparar:

$$Sim_{lin}(c_i, c_j) = \frac{2 \times IC(MICA[c_i, c_j])}{IC(c_i) + IC(c_j)} \tag{1.11}$$

$$Sim_{JC}(c_i, c_j) = 1 - IC(c_i) + IC(c_j) - 2 \times IC(MICA[c_i, c_j]) \tag{1.12}$$

Un inconveniente de estas medidas es que se encuentran desplazadas del grafo, es decir, estas medidas son proporcionales a las diferencias entre los IC de los términos y de sus ancestros comunes, independientemente del valor absoluto de IC del ancestro. Una restricción de todas estas medidas es que solo toman en cuenta el MICA, a pesar de que los términos GO pueden tener varios ancestros disjuntos comunes (DCA). Para evitar esta restricción, [**?**] propuso la aproximación GraSM, que puede ser aplicada a todas las medidas descritas anteriormente, simplemente reemplazando el IC del MICA, por el promedio de los IC de los DCA. Existen más de dos docenas de medidas de similaridad entre términos GO, y no siempre es claro cuál es el mejor para un dado propósito. Sin embargo, generalmente la elección de una medida por defecto es suficiente [**?**]. En este trabajo utilizaremos la Sim_{res}, por tratarse de una medida simple y efectiva.
Una vez establecida una medida de similaridad semantica entre termino GO, existen distintas formas para extender esta idea y definir una similaridad semántica entre genes. Básicamente existen 2 estrategias. La primera se basa en medidas globales (groupwise en inglés), que comparan globalmente los conjuntos de términos en los que dos genes están anotados, $GO(g_1)$ y $GO(g_2)$, por ejemplo, contando cuantos términos comparten: $|GO(g_1) \cap GO(g_2)|$. [**?**]
La segunda estrategia se basa en medidas de a pares (pairwise en inglés), calculando la similaridad semántica término a término de cada uno de los conjuntos $GO(g_1)$ y $GO(g_2)$ y luego aplicando sobre esta similaridad alguna operación para obtener una medida de similaridad entre estos genes.
El primer paso para esto es calcular una matriz de similaridad S de $N \times M$ que contenga la similaridad de a pares, entre todos los pares de términos de estos conjuntos, con $N = |GO(g_1)|$ y $M = |GO(g_2)|$, utilizando alguna de las medidas de similaridad semántica entre términos presentadas anteriormente (Sim_{res}, Sim_{lin}, etc.):

$$S_{ij} = Sim(GO(g_1^i), GO(g_2^j)), \forall i \in \{1, ..., N\} y \forall j \in \{1, ..., M\} \tag{1.13}$$

Notar que esta matriz puede no ser simétrica.
Cada una de las N filas corresponde a la similaridad entre la anotación $i - esima$ del gen 1 y todas las M anotaciones del gen 2 y cada una de las M columnas corresponde a la similaridad entre la anotación $j - esima$ del gen 2 y todas las N anotaciones del gen 1.
A partir de S_{ij} es posible definir tres métodos para obtener una medida de similaridad

entre genes. El primer método, propuesto en [**?**], consiste en tomar como similaridad, la máxima similaridad entre todos los términos:

$$Sim_{max}(GO(g_i), GO(g_j)) = \text{máx}\{S_{ij}\} \tag{1.14}$$

El segúndo método, propuesto en [**?**], consiste en tomar el valor medio de todos los valores de la matriz S_{ij}:

$$Sim_{med}(GO(g_i), GO(g_j)) = \frac{1}{N.M} \sum_{i,j} S_{ij} \tag{1.15}$$

Finalmente, el tercer método, propuesto en [**?**], implica tomar el valor medio de los máximos de cada fila, el valor medio de los máximos de cada columna, y quedarse con el máximo de esos dos valores. Este criterio de similaridad se conoce como *rcmax*:

$$Sim_{rcmax}(GO(g_1), GO(g_2)) = \text{máx}\{\frac{1}{N} \sum_{i} \underset{1 \leq j \leq M}{\text{máx}} S_{ij}, \frac{1}{M} \sum_{j} \underset{1 \leq i \leq N}{\text{máx}} S_{ij}\} \tag{1.16}$$

Como muchos genes están anotados en conceptos muy diversos por participar en procesos biológicos muy distintos, e incluso puede haber genes que no están anotados en ningún concepto, la medida de similaridad Sim_{med} tiende a dar valores más bajos que otros métodos. Por el contrario, la medida Sim_{max} tiende a dar valores más altos, por ser una medida más optimista. En este trabajo utilizaremos el tercer método, Sim_{rcmax}, por ser un compromiso entre ambos casos extremos. [**?**, **?**, **?**, **?**, **?**, **?**, **?**]

1.2. Estrategias de agrupamiento

En la sección anterior abordamos distintas metodologías para cuantificar la noción de similaridad en diversos espacios.
En lo que sigue introduciremos las diferentes estrategias de agrupamiento de datos utilizadas en este trabajo, tanto para agrupamiento de perfiles transcripcionales como de armado de comunidades en las redes presentadas anteriormente.
Es posible distinguir dos tipos de agrupamientos, conocidos como agrupamiento duro (hard clustering en inglés), y agrupamiento difuso (fuzzy clustering en inglés). En el primer caso, el de agrupamiento duro, cada objeto del conjunto de datos es asignado a un y solo un grupo, mientras que en el segundo caso, el de agrupamiento difuso, un elemento del conjunto puede pertenecer a varios grupos, con distinta probabilidad. En este trabajo utilizaremos únicamente métodos de agrupamiento duro.

1.3. Agrupamientos no jerárquicos

Además de la distinción mencionada más arriba, los métodos de agrupamiento pueden dividirse (entre otros) fundamentalmente entre agrupamientos jerárquicos y agrupamientos no jerárquicos. Las dos estrategias de agrupamientos no jerárquicos que se presentan a continuación fueron utilizadas en el desarrollo de este trabajo.

1.3.1. K-means

K-means es un método usual de agrupamiento no jerárquico en donde cada observación pertenece al grupo con la media más cercana a la observación.
El mismo comienza agrupando los objetos de forma arbitraria en K grupos distintos. El número K puede ser elegido de forma aleatoria o estimado mediante algún otro método de agrupamiento jerárquico pero es siempre fijo. Luego, se calcula un promedio de la posición de todas las observaciones de cada grupo, llamado centroide. A continuación, los objetos individuales son redistribuidos de un grupo a otro dependiendo de que centroide esté más cerca de la observación. Este procedimiento de calcular el centroide de cada cluster y re agrupar los objetos más cercanos a los centroides disponibles se repite de manera iterativa una cantidad fija de veces o hasta la convergencia del método (se considera que el método converge cuando una iteración no modifica la iteración anterior).
Más formalmente, sea un conjunto de observaciones $\{\vec{x_1}, ..., \vec{x_n}\}$, k-means construye una partición de las observaciones en k grupos con $k \leq n$ a fin de minimizar una función de costo, como ser la suma de los cuadrados dentro de cada grupo $G = \{g_1, ..., g_k\}$:

$$C = argmin \sum_{i=1}^{k} \sum_{x_j \in g_i} ||x_j - \mu_i||^2 \tag{1.17}$$

Con μ_i el valor medio de los elementos del grupo g_i. La figura **??** muestra un conjunto de observaciones y los grupos que se obtienen fijando $k = 2$ y $k = 5$, junto con sus respectivos centroides. Se observa que dependiendo del k utilizado, el algoritmo encuentra particiones con mayor o menor nivel de *resolución*. Volveremos sobre el tema de la resolución más adelante. [**?**, **?**]

1.3.2. PAM

Si bien k-means es uno de los métodos de partición más utilizados ya que es muy eficiente en términos de tiempo computacional, el mismo es muy sensible a observaciones aisladas. Por esta razón, en algunos métodos se reemplazan los centroides, que son puntos no necesariamente pertenecientes al conjunto de observaciones, por medoides, que son los objetos más centrales dentro del grupo (se reemplaza k-means por k-medoids).

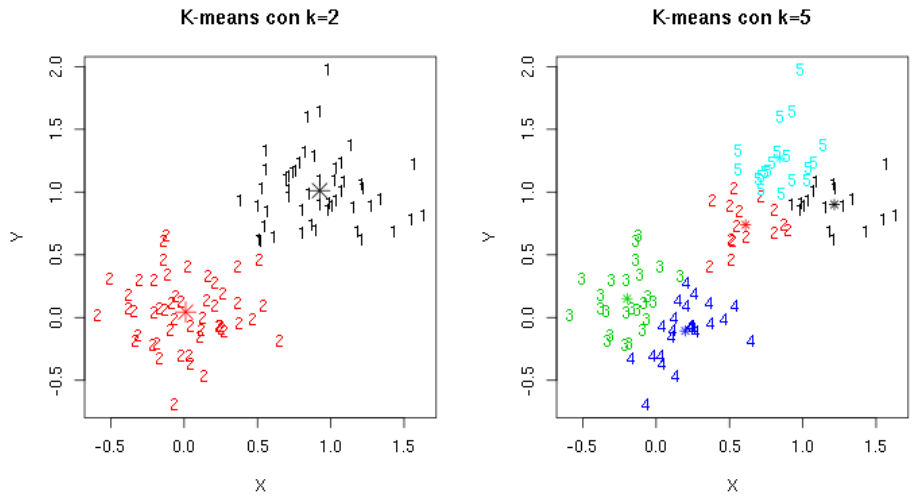


Figura 1.3: Agrupamiento utilizando k-means con $k = 2$ y $k = 5$. mejorar epigrafe

Esto hace que el método sea insensible a observaciones aisladas.
Particionar alrededor de medoides (Partitioning around medoids en inglés) es uno de los métodos más conocidos que hace uso de este concepto, buscando minimizar la función de costo:

$$C = argmin \sum_{i=1}^{k} \sum_{x_j \in g_i} d(x_j, m_i) \tag{1.18}$$

Con m_i el medoide del grupo i y $d(x_j, m_i)$ la distancia entre el objeto x_j del grupo i y el medoide del mismo grupo. [?, ?]

1.4. Agrupamientos jerárquicos

Existen dos acercamientos distintos para realizar un agrupamiento jerárquico: se puede ir "desde abajo hacía arriba", agrupando grupos más chicos en grupos más grandes, lo que se conoce como agrupamiento aglomerativo, ose puede ir "desde arriba hacia abajo", dividiendo grupos más grandes en grupos más chicos, lo que se conoce como agrupamiento divisivo. En este trabajo nos interesará únicamente trabajar con agrupamientos aglomerativos.
Un agrupamiento jerárquico aglomerativo comienza con cada objeto en un grupo separado. Luego, se unen los dos grupos más cercanos de acuerdo a algún criterio definido generando un nuevo grupo a partir de ambos. Al nuevo grupo se le asignará una distancia al resto de los grupos de acuerdo a cierto criterio. Esto se repite hasta que solo quede un único grupo.

Es un tipo de procedimiento determinista y voraz (greedy en inglés), ya que realiza las decisiones tomando en cuenta los óptimos locales en cada etapa, esperando obtener con esto un óptimo global.
Se dice que una partición es más fina (o un refinamiento) de otra partición, si cada grupo de una partición más fina está contenido dentro de un grupo de la partición más gruesa, es decir, cada grupo de la partición más fina es un sub-grupo de un grupo de la partición más gruesa. El agrupamiento jerárquico es un método cuyo resultado es un conjunto de particiones anidadas $P_n, P_{n-1}, ..., P_1$ cada vez más gruesas, donde cada nivel más alto une dos grupos de una partición de un nivel más bajo.
Para poder realizar este procedimiento, es necesario definir cuan cercanos son dos grupos:

1.4.1. Método de Ward

Este método busca unir los grupos de una forma tal que se minimice la pérdida de información asociada a cada unión, usualmente cuantificada como el error de la suma de los cuadrados (ESS). Dado un conjunto de puntos C, el ESS asociado a C queda definido por:

$$ESS(C) = \sum_{\vec{x} \in C} (\vec{x} - \mu(C))(\vec{x} - \mu(C))^T \tag{1.19}$$

con $\mu(C) = \frac{1}{|C|} \sum_{\vec{x} \in C} \vec{x}$, el valor medio de C. Suponiendo que una dada partición está separada en k grupos, $\{C_1, C_2, ..., C_k\}$, entonces se tiene que la pérdida de información de la partición está dada por:

$$ESS = \sum_{i=1}^{k} ESS(C_i) \tag{1.20}$$

En cada etapa de este método, se prueban todas las uniones de grupos posibles de a pares y se realiza aquella unión que minimiza **??**.
En el agrupamiento jerárquico, el ESS comienza en cero, ya que cada punto pertenece a un grupo distinto, y crece a medida que se unen grupos. Al ser un algoritmo voraz, la ESS para un dado número de grupos k no será necesariamente la mínima.

1.4.2. Método de enlace único (o single-link en inglés)

Este método es uno de los métodos más simples para agrupamiento jerárquico. El mismo define la distancia entre dos grupos como la mínima distancia entre entre sus miembros. Sean C_i y C_j dos grupos, entonces la distancia de enlace único se define como:

$$D_{sl}(C_i, C_j) = \min_{\vec{x} \in C_i, \vec{y} \in C_j} d(\vec{x}, \vec{y}) \tag{1.21}$$

con $d(\vec{x}, \vec{y})$ la función de distancia utilizada para calcular la matriz de disimilaridad entre los elementos. El nombre de enlace único hace referencia a que dos grupos están cerca aunque tengan un único par de puntos cerca. Este método permite el manejo de grupos con formas complejas y es invariante ante transformaciones monótonas (como una transformación logarítmica) [**?**].
Este algoritmo solamente considera la separación entre elementos, dejando de lado la compacidad o el balance en los grupos.

1.4.3. Método de enlace completo (o complete-link en inglés)

Este método es similar al método de enlace único, ya que toma la distancia entre dos grupos como el máximo de la distancia entre sus puntos:

$$D_{cl}(C_i, C_j) = \max_{\vec{x} \in C_i, \vec{y} \in C_j} d(\vec{x}, \vec{y}) \tag{1.22}$$

con $d(\vec{x}, \vec{y})$ la función de distancia utilizada para calcular la matriz de disimilaridad entre los elementos.
En este trabajo, utilizaremos el método de enlace completo. [**?**, **?**, **?**]

1.4.4. Representación de un agrupamiento jerárquico - dendrogramas

Un agrupamiento jerárquico puede representarse como un árbol, llamado dendrograma, que permite una rápida interpretación. En un dendrograma, cada nodo está asociado con una altura h, tal que si A y B son dos nodos del dendrograma, h cumple:

$$h(A) \leq h(B) \Leftrightarrow A \subseteq B \tag{1.23}$$

A modo ilustrativo, la figura **??** muestra el agrupamiento jerárquico realizado sobre 10 puntos colocados de forma aleatoria en el plano, agrupados utilizando la distancia euclidiana y mediante los tres métodos vistos anteriormente (Ward, enlace único y enlace completo). De estos gráficos es claro que cada método produce una secuencia diferente de particiones, y dependerá de la aplicación que se requiera, cual de los métodos utilizar.

1.5. Detectando grupos en el agrupamiento jerárquico

El agrupamiento jerárquico organiza los objetos en árboles (dendrogramas) cuyas ramas son los grupos deseados. El proceso de detección de grupos se conoce como corte de árbol, corte de ramas o podado de ramas.

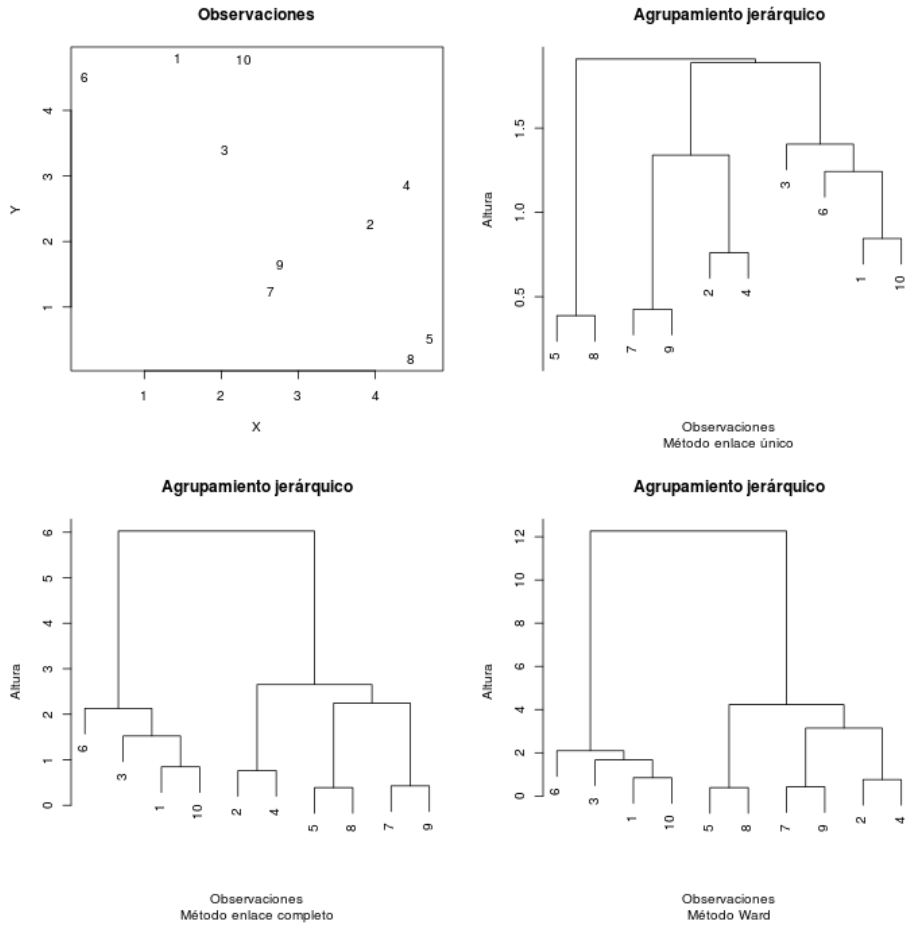


Figura 1.4: Ejemplos de agrupamientos jerárquicos utilizando el mismo conjunto de datos pero distintos métodos de distancia entre grupos.mejorar epigrafe

1.5.1. Corte de árbol estático

El método más sencillo de podado es conocido como corte de árbol estático, y funciona definiendo cada rama contigua debajo de una altura fija de corte, como un grupo separado. La cantidad de grupos obtenidos por éste método depende fuertemente de la altura de corte elegida. La figura **??** muestra dos alturas de corte posibles y los grupos que se obtienen a partir de cada una de ellas. Al cortar el árbol en $h = 3$, se obtienen dos grupos, el grupo g_1, que contiene a las observaciones $\{6, 3, 1, 10\}$ y el grupo g_2 que contiene a las observaciones $\{2, 4, 8, 5, 7, 9\}$, mientras que al cortarlo en $h = 2$, se obtienen cuatro grupos, g'_1 con la observación $\{6\}$, g'_2 con las observaciones $\{3, 1, 10\}$, g'_3 con las observaciones $\{2, 4\}$ y g'_4 con las observaciones $\{5, 8, 7, 9\}$.
A partir de un ejemplo tan sencillo es inmediato notar que el problema del agrupamiento es un problema "mal planteado", es decir, cualquier conjunto de puntos puede ser

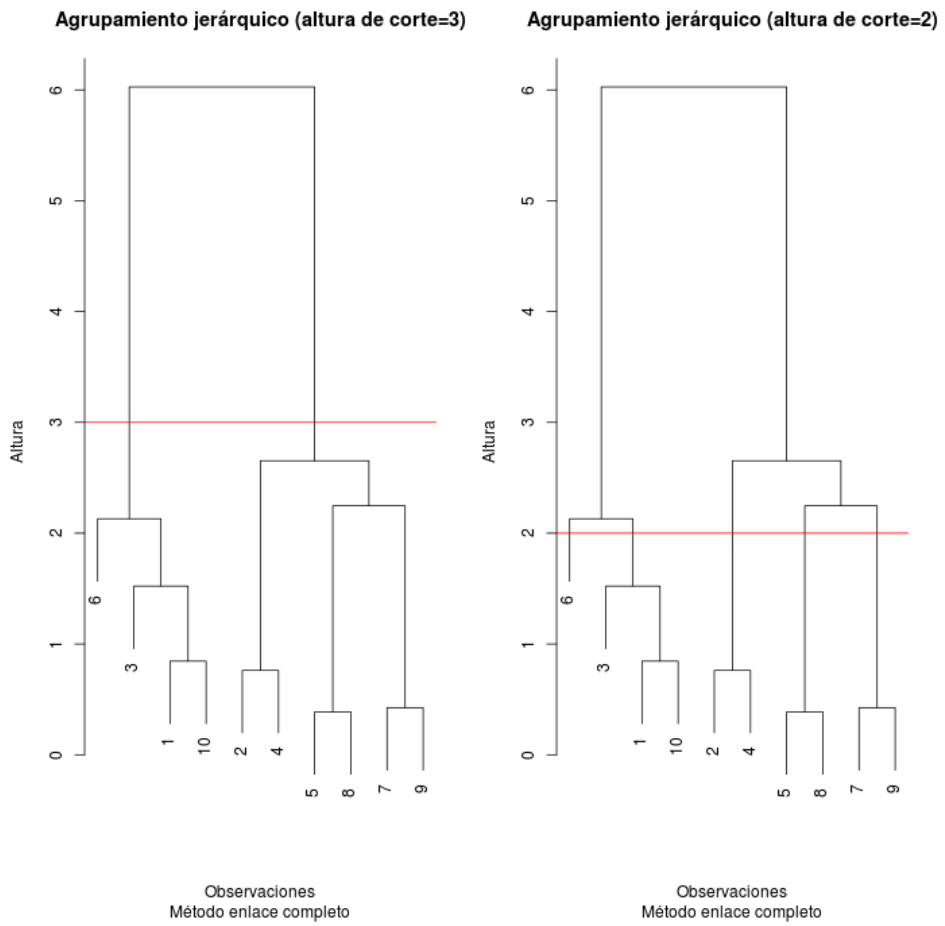


Figura 1.5: Corte de árbol a dos alturas diferentes. mejorar epigrafe

agrupado de maneras drásticamente distintas, sin que exista a priori un único criterio para preferir uno u otro agrupamiento. La fuente de ambigüedades a este respecto más importante, es que la forma en que los datos deberían ser agrupados, depende fuertemente de la *resolución* deseada. Lo que parece una única nube de puntos puede resultar ser, al analizar los datos con mayor resolución, una partición compuesta de muchos grupos. Cada tarea deberá encontrar el nivel adecuado de resolución para obtener la cantidad "correcta" de grupos. [**?**] [**?**]

1.5.2. Corte de árbol dinámico híbrido

Si bien es posible detectar grupos distintos en el dendrograma a partir de una inspección visual, utilizar una técnica de corte de árbol estático de forma programática no siempre logra identificar adecuadamente los grupos, ya que al poseer grupos anidados,

un solo corte a una altura prefijada no será capaz de detectarlos todos. El método de corte de árbol dinámico híbrido ataca este problema analizando la forma de las ramas del dendrograma en lugar de una altura absoluta [**?**]. El mismo construye los grupos de abajo hacia arriba en dos pasos. En el primer paso, se detectan las ramas que satisfacen un criterio específico para ser grupos. Este paso de poda está basado en la información de unión del dendrograma. En el segundo paso, se miden cuán cerca de los grupos detectados en el primer paso están todos los objetos no asignados previamente. Si un objeto está suficientemente cerca de un grupo, es asignado a ese grupo. En este paso, se ignora el dendrograma y se utiliza únicamente la información de disimilaridad. Este paso puede considerarse un método modificado de particionado alrededor de medoides (modified Partitioning Around Medoids o mPAM, en inglés). Por eso el nombre de *híbrido*, al tratarse de una mezcla entre agrupamiento jerárquico y no jerárquico. Los criterios específicos para la detección de grupos se basan en los siguientes cuatro criterios de la forma de las ramas:

1. Un grupo debe tener una cantidad mínima de objetos.
2. Los objetos que están muy lejos del grupo son excluidos del grupo aunque pertenezcan a la misma rama del dendrograma.
3. Cada grupo debe estar separado de su entorno por una brecha o espacio vacío.
4. El núcleo de cada grupo (el conjunto de objetos con menor altura de unión en el grupo) debe estar fuertemente conectado.

O más formalmente, dado un núcleo de un grupo, llamamos d al promedio de las disimilaridades de a pares entre objetos del núcleo, es decir, a su dispersión y definimos la brecha g de un grupo como la diferencia entre d y la altura donde el grupo se une al resto del dendrograma y entonces, una rama se considera un grupo si:

1. Tiene al menos N_0 objetos.
2. Todas las alturas de unión son a lo sumo de h_{max}.
3. La brecha g del grupo es mayor que un g_{min}.
4. La dispersión d del núcleo es a lo sumo d_{max}.

Los parámetros N_0, h_{max}, g_{min} y d_{max} son parámetros ajustables del método. La figura **??** muestra un ejemplo de los parámetros utilizados para definir los grupos en el paso 1.
Para el paso 2, de tipo PAM, los objetos no asignados (o aquellos grupos que no cumplan tener al menos N_0 objetos) son asignados al grupo más cercano si la disimilaridad correspondiente es más pequeña que una disimilaridad máxima definida previamente,

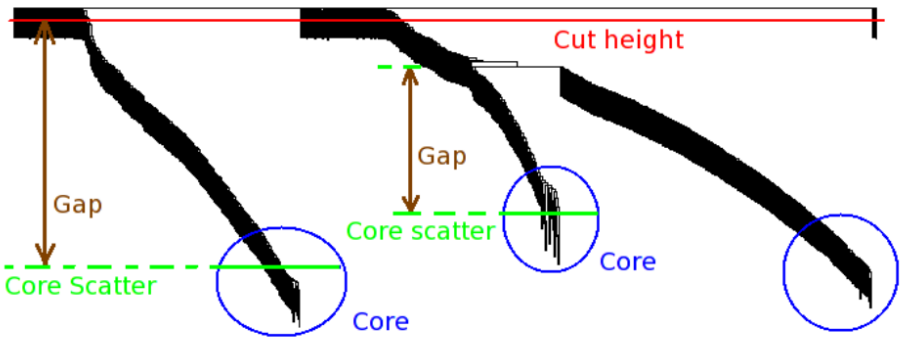


Figura 1.6: Dendrograma simulado con tres ramas con alturas de unión diferentes. La altura de corte corresponde a h_{max} (fuente: [**?**]). poner en castellano

o si es más pequeña que el "radio" del grupo. El "radio" se define como la máxima de las disimilaridades del medoide del grupo al resto de los objetos del mismo.
Es posible controlar la sensibilidad de las divisiones de los grupos mediante el parámetro *deepSplit*, que puede tomar los valores de 1 a 4. Para un $deepSplit = 1$, el método producirá relativamente pocos grupos, de muchos elementos y bien definidos, mientras que para $deepSplit = 4$, el método producirá más grupos pero con una dispersión mayor en el núcleo y separado por brechas más pequeñas.
Para una descripción más detallada del algoritmo, el lector interesado puede referirse a [**?**], [**?**].

1.6. Infomap y CNM

Como se desarrolló en la sección **??**, las redes son construcciones útiles para esquematizar la organización de las interacciones en distintos tipos de sistemas. Sin embargo, por motivos de visualización, solo se pueden representar pequeños sistemas. Las redes reales son usualmente tan grandes que es necesario representarlas mediante algún mecanismo de granularidad más gruesa, es decir, descomponer a la red en módulos que representen varios nodos y arcos. Los módulos son conjuntos de módulos que tienen un alto solapamiento topológico. Este es el objetivo básico de lo que se conoce como *detección de comunidades*.
Una red puede representarse con una matriz de adjacencia $A = [a_{ij}]$ que codifica que pares de nodos están conectados. A es una matriz simétrica donde cada a_{ij} puede tomar un valor entre $[0, 1]$. Para una red no pesada, 0 indica que dos nodos no están conectados, mientras que 1 indica que si lo están. Para una red pesada, el elemento de matriz indica la fuerza de la conexión, tomando un valor entre $[0, 1]$.
A partir de la matriz de adjacencia, es posible construir una matriz de solapamiento

topológico $T = [t_{ij}]$ (TOM por sus siglas en inglés) que es una medida de similaridad para redes biológicas y está definida como:

$$t_{ij} = \begin{cases} \frac{l_{ij}+a_{ij}}{min\{k_i,k_j\}+1-a_{ij}} & i \neq j \\ 1 & i = j \end{cases} \tag{1.24}$$

donde $l_{ij} = \sum_u a_{iu}a_{uj}$, $k_i = \sum_u a_{iu}$ y u es un índice que recorre todos los nodos de la red. El solapamiento topológico de dos nodos refleja su similaridad en términos de los nodos en común que conectan. Básicamente, t_{ij} es un indicador del acuerdo entre el conjunto de nodos vecinos a i con el conjunto de nodos vecinos a j. Utilizando esta similaridad, se obtiene una matriz de disimilaridad $D = 1 - TOM$ y con esto es posible realizar agrupamientos utilizando, entre otras, alguna de las técnicas antes mencionadas. Además de TOM, es posible definir una matriz de solapamiento topológico generalizada de orden m, $T[m] = [t[m]_{ij}]$ (GTOMm), tal que mida el acuerdo entre los nodos que son accesibles por i y por j en m pasos. [**?**]

En este trabajo utilizaremos dos métodos de modularización en redes, Infomap [**?**] y CNM [**?**].

El método o algoritmo Infomap hace uso de criterios de optimización basados en teorías de información, donde los módulos se definen de tal forma que la longitud media de la descripción de un proceso de paseo al azar en el grafo sea mínima, mientras que el desarrollado por Clauset, Newman y Moore que denominaremos CNM, a partir de ciertas heurísticas, busca particiones de la red optimizando directamente una función de calidad Q.

Ambos métodos serán utilizados en este trabajo con el fin de comparar los resultados obtenidos para las comunidades Infomap y CNM con los obtenidos para los métodos de agrupamiento usados. [**?**, **?**]

Capítulo 2

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.

2.1. Descripción del dataset

esto esta en sec:wiegel habra que profundizar mas?

2.2. Métricas transcripcionales

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

2.3. Agrupamiento

2.3.1. Proceso de filtrado

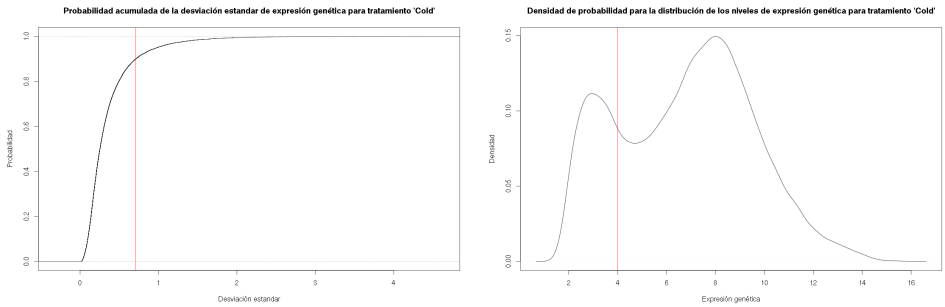
El conjunto de datos Wiegel utilizado consta de los niveles de expresión de 22810 sondas que se mapean a 20149 genes a lo largo de 11 tratamientos diferentes y con entre 4 y 9 muestreos en dos réplicas. Para poder manejar esta cantidad de información es necesario realizar un filtrado (una selección) 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 **??** 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:

$$\tilde{x_i} = \frac{x_i - \bar{x}}{s_x} \tag{2.1}$$

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 2.1: Funciones de distribución de probabilidad para perfiles de expresión

2.3.2. 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:

$$d(\vec{x}, \vec{y}) = \sqrt{2(d-1)(1 - r(\vec{x}, \vec{y}))} \tag{2.2}$$

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:

$$CH_k = \frac{SS_B}{SS_W} \frac{n-k}{n-1} \tag{2.3}$$

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:

$$DI = \frac{min\delta}{\text{máx}\,\Delta} \tag{2.4}$$

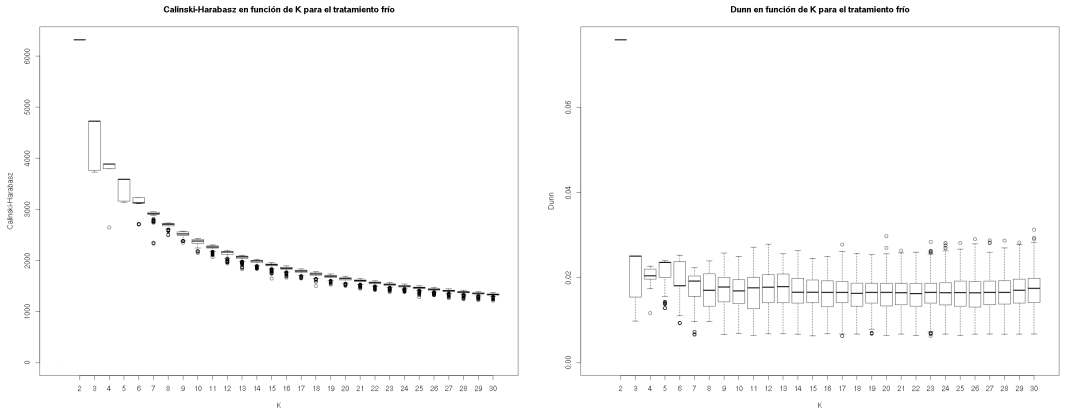
Tratamiento	σ	A	Cantidad de genes
Control	0.37	4	1885
Frío	0.71	3	1955
Osmótico	0.71	3	1923
Sal	0.88	3	1927
Sequía	0.54	4	1870
Genotóxico	0.46	3	1899
Oxidativo	0.41	3	1880
UV-B	0.51	4	1872
Heridas	0.41	4	1877
Calor	0.75	2	1960
Calor y recuperación	0.65	2	1944

Cuadro 2.1: 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.

2.3.3. Agrupamiento con corte de árbol dinámico

Utilizando el método de corte de árbol dinámico se realizó para cada tratamiento ideas: poner aca que se probo con deepsplit 1 y deepsplit 4. Que se logra meter mucho



(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 2.2: Índices de validación interna para particiones realizadas con k-means

mejor, mostrar algunos perfiles para algun tratamiento para ambos y mostrar quizas una tabla con cada tratamiento cuantos clusters y que tamanios. charlar de nuevo sobre la escala.

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

2.4. 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

2.4.1. Interacting densities

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

2.4.2. 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

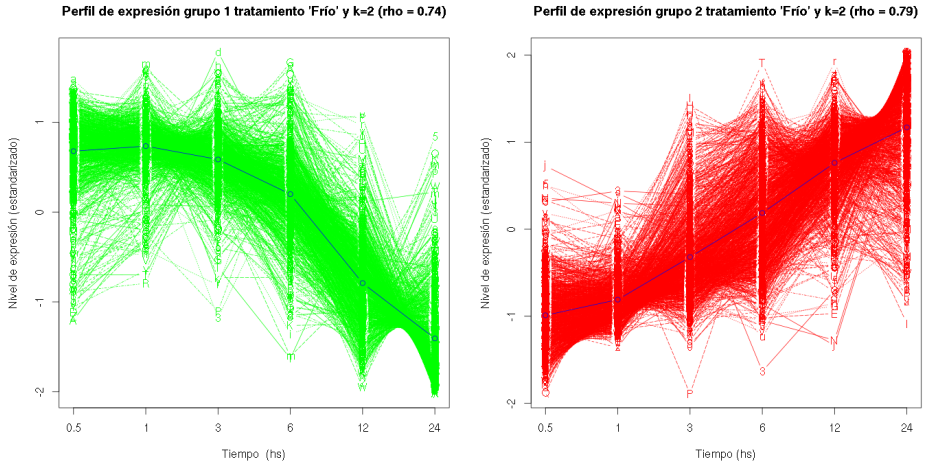


Figura 2.3: Perfiles de expresión génica obtenidos con el método k-means (k=2) para el tratamiento 'Frío'. En azul, el valor medio de cada grupo.

Bibliografía