

Análisis y Detección de Correlaciones en Relevamientos Transcripcionales de Gran Escala

Andrés Rabinovich

Director: Dr. Ariel Chernomoretz

Departamento de Física
Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires

Marzo 2016.

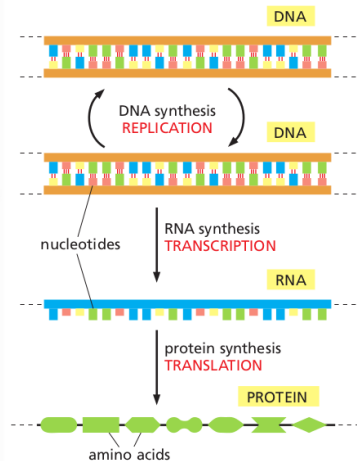
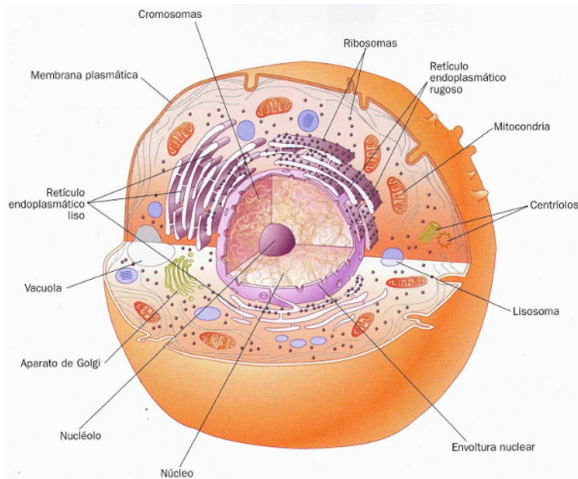


Contenido

- 1 Introducción
 - Relevamientos transcripcionales de gran escala
 - Detección de correlaciones
- 2 Análisis de relevamientos transcripcionales
 - Medidas de similaridad y distancia
 - Métodos de agrupamiento utilizados
 - Caracterización de particiones
- 3 Congruencia biológica
 - Ontología génica (GO)
 - Cuantificando la congruencia biológica
- 4 Coherencia entre métricas
 - Métrica en GO
 - KTA global
 - Modulación de heterogeneidades transcripcionales con GO
- 5 Conclusiones

Transcripción y traducción (dogma central de la biología molecular)

Células, ADN, ARNm, proteínas y otras yerbas...



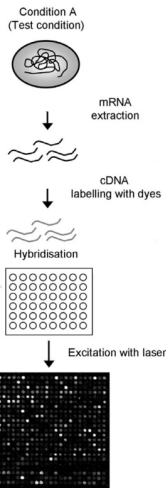
Cambios transcripcionales en respuesta a estrés abiótico en *A. thaliana*

A. Thaliana [1]

Gen	1 hs.	6 hs.	24 hs.
AT4G09000	0.2	0.8	0.4
AT1G78300	1.2	1.2	0.8
AT5G38480	0.5	0.7	1.2
-	-	-	-
-	-	-	-
AT1G35160	1.1	0.9	0.6

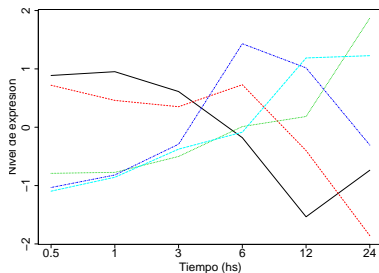
Micromatriz de ADN

[1] arabidopsis.org: AtGenExpress.



Datos de estrés abiótico:

- **10 tratamientos + control:** frío, calor, osmótico, salinidad, sequía, genotoxicidad, oxidación, UV, herida, recuperación.
- ≈ 22000 genes.
- ≈ 6000 genes se **movieron** en algún tratamiento.
- **Entre 4 y 8 mediciones** temporales por gen y por tratamiento.

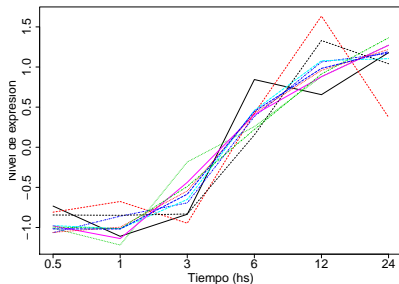


Detección de correlaciones

Queremos inferir estrategias del organismo frente a los tratamientos.

Lo vamos a hacer usando **métodos de agrupamiento o “clustering”** para encontrar relaciones y estructura en esta gran cantidad de datos.

- Son métodos no supervisados.
- Consisten en agrupar elementos **“similares entre sí”**.
- Permiten el **descubrimiento de patrones** en los datos.
- Posibilitan obtener **conclusiones** sobre los datos.

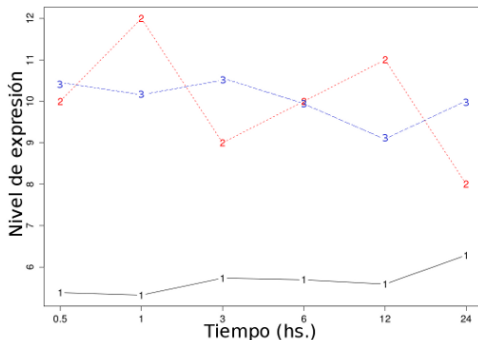


Usamos el coeficiente de correlación de Pearson

Distancia basada en el
 coeficiente de correlación de
 Pearson

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

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



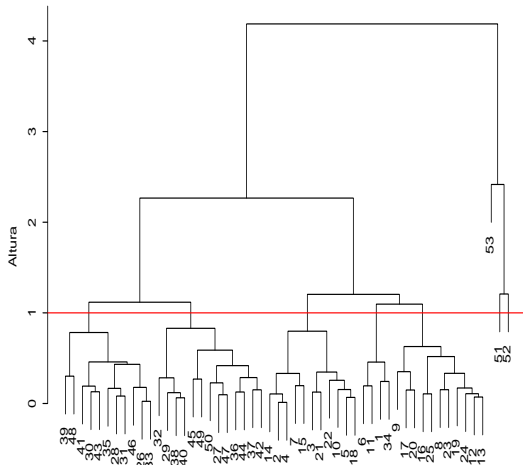
Métodos de agrupamiento

Método k-means

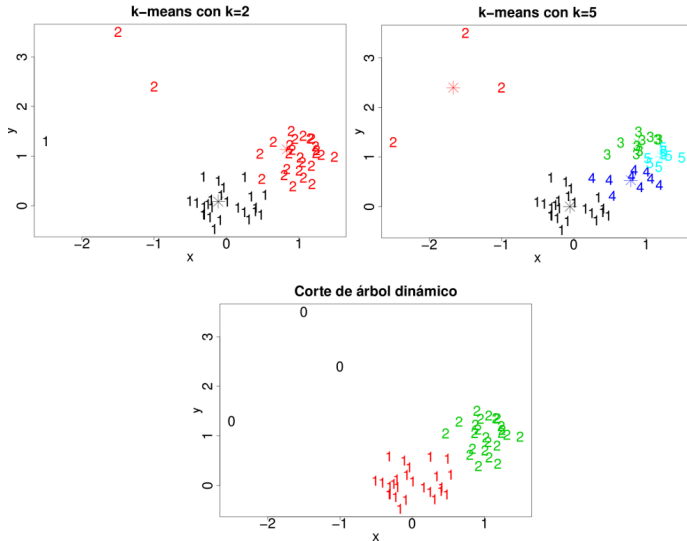
- Busca estructuras compactas.
- Muy rápida ejecución.
- La cantidad k de grupos debe ser fijada a priori.
- Existen figuras de mérito para decidir el k óptimo.

Método corte de árbol dinámico

- Agrupamiento jerárquico.
- Representación mediante **dendrograma**.
- Ajuste por *deepsplit* (usamos DS1 y DS4).

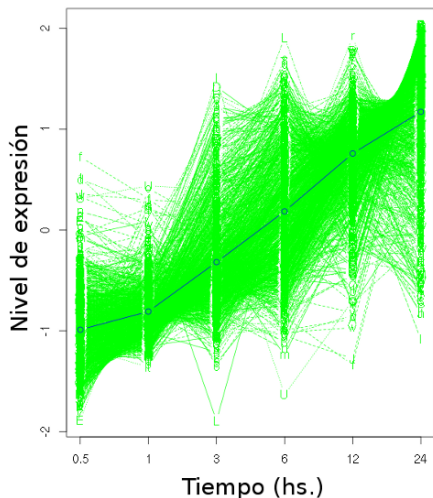


Ejemplo de agrupamientos

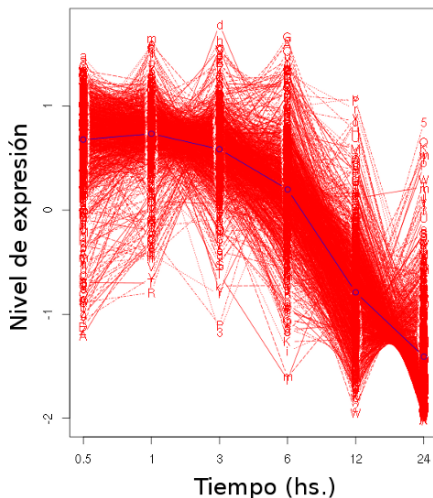


Grupos con k-means

Perfil de expresión grupo 1 tratamiento 'Frio' y k=2 ($\rho = 0.74$)

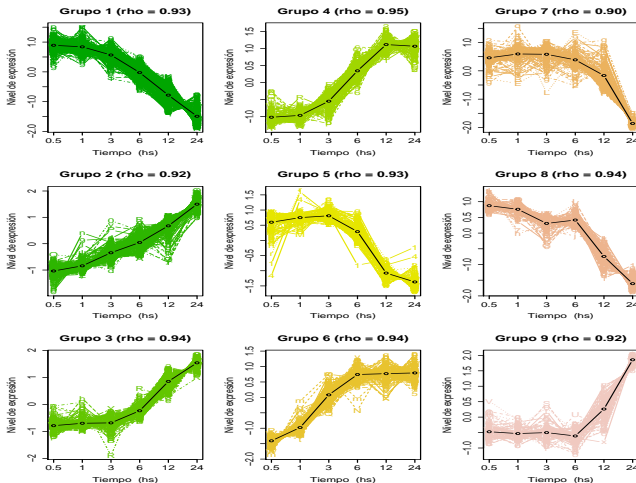


Perfil de expresión grupo 2 tratamiento 'Frio' y k=2 ($\rho = 0.79$)



Grupos con corte de árbol dinámico

A modo de ejemplo, los nueve perfiles más grandes de una partición de tratamiento “Frío” y DS1.

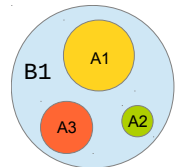


El problema de la escala

Granularidad y resolución de los métodos

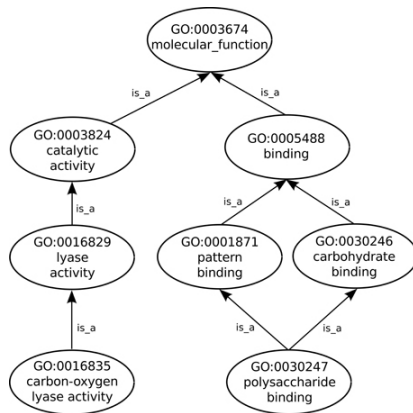
- Una partición A es **más fina** que una partición B si cada grupo de A está contenido en un grupo de B .
- Tenemos **tres formas de realizar particiones** de nuestros datos.
- DS4 genera particiones más finas que DS1 y este a su vez que k-means.
- Tenemos distintas maneras de encontrar estructura en nuestros datos y las **distintas heterogeneidades aparecerán a distintas escalas**.

Vamos a ver si existe una **escala óptima en un sentido biológico** a la que trabajar con éste conjunto de datos y para eso vamos a utilizar un espacio de conocimiento biológico.



Ontología génica (GO)

- Provee un vocabulario controlado de términos.
- Permite comparar y clasificar entidades biológicas.
- Tres ontologías: procesos biológicos (BP), componentes celulares (CC) y funciones moleculares (MF).
- Estructura de grafo acíclico dirigido (DAG).
- Cada nodo representa un término que describe alguna función.
- Los nodos se unen entre si por medio de relaciones “es un” o “es parte de”.



Un gen descrito por un término está “anotado” en ese término.

Observables

Buscamos cuantificar la congruencia biológica de las particiones halladas

Densidad de interacción:

$$ID(GO_j) = \frac{NE(GO_j)}{N(GO_j)} \quad (1)$$

Con $NE(GO_j)$ la cantidad de pares de genes anotados en GO_j que se encuentran juntos en un mismo grupo transcripcional C_x y $N(GO_j)$ la cantidad de pares de genes anotados en GO_j .

Índice de homogeneidad biológica:

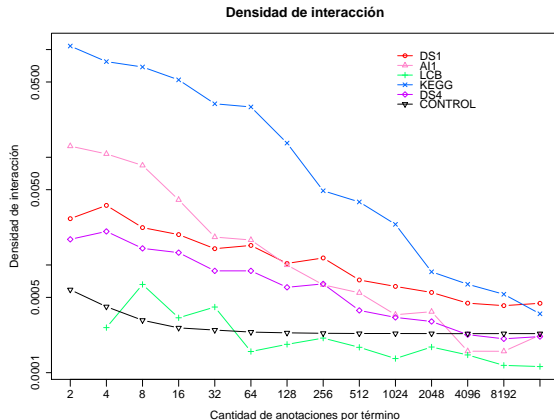
$$BHI_j = \frac{1}{n_j(n_j - 1)} \sum_{x \neq y \in D_j} I(C(x) = C(y))$$

Con n_j la cantidad de genes anotados en el grupo D_j .

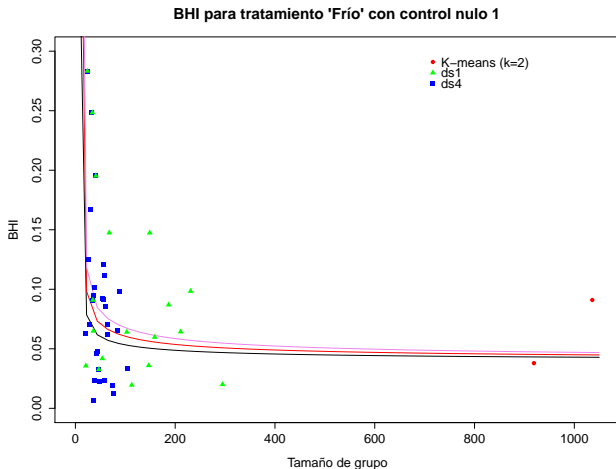
La función indicadora $I(C(x) = C(y))$ toma el valor 1 si hay al menos una clase en donde ambos genes estén anotados, y 0 en caso contrario.

Densidad de interacción

- 1 Términos mas específicos presentan mayor ID en una relación decreciente.
- 2 DS1 presenta mayor congruencia biológica que DS4. Indicio acerca de la escala apropiada.
- 3 Ambos presentan mayor congruencia biológica que el control nulo.
- 4 Los agrupamientos inducidos por otra información presentan mayor congruencia que los inducidos por expresión.



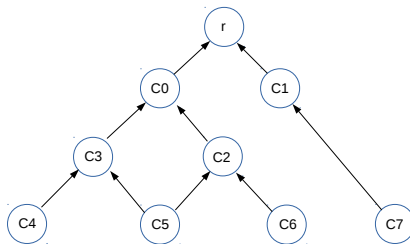
Índice de homogeneidad biológica



Grupos altamente coherentes pero de baja calidad de BHI. No tienen soporte biológico.

Similaridad entre genes en GO

Podemos definir similitudes entre genes en el espacio GO



Utilizando la similitud entre términos:

$$Sim_{res}(c_i, c_j) = \max_{c \in S(c_i, c_j)} (-\log_2[P(c)]) = IC(MICA[c_i, c_j]) \quad (2)$$

KTA global

La noción de similaridad de a pares en cada espacio está dada en términos de una función k llamada kernel tal que

$$K = K_{ij} = k(x_i, x_j) \quad (3)$$

El KTA de un kernel k_1 con respecto a un kernel k_2 del conjunto C

cuantifica la similaridad entre dos espacios y se define como:

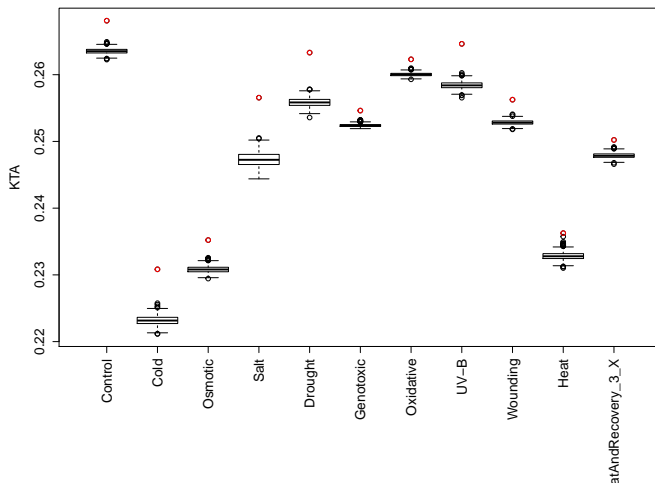
$$\hat{A}(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \quad (4)$$

con $\langle K_1, K_2 \rangle_F = \sum_{i,j=1}^m K_1(x_i, x_j) K_2(x_i, x_j)$ es el producto interno de Frobenius.

Intuitivamente, si $\langle K_1, K_2 \rangle$ es grande, ambos kernels son coherentes.

KTA global

KTA global entre expresión y ontología BP con control nulo

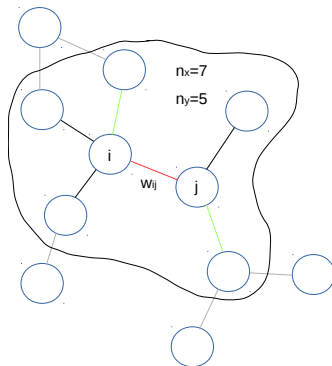


Red 30 primeros vecinos mutuos - vecindades locales

Queremos detectar zonas de alta coherencia.

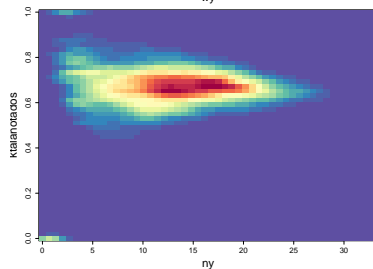
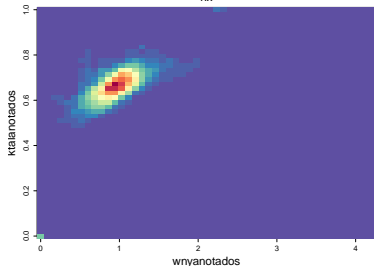
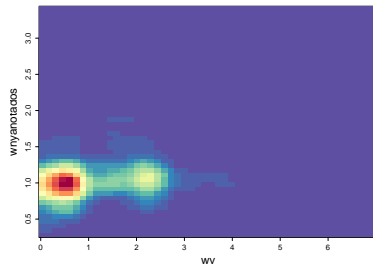
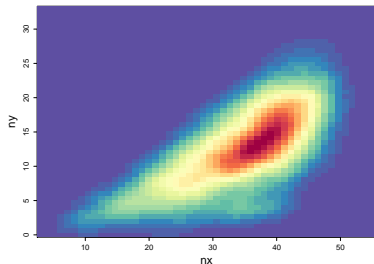
Generamos una red de 30 primeros vecinos mutuos y vamos a ver arista por arista, una localidad definida por los primeros vecinos:

- n_x nodos.
- n_y nodos anotados.
- w_{ij} distancia entre nodos i y j en GO.
- w_{ny} similaridades promedio en una vecindad en GO.
- $w_{nyanotados}$ similaridades promedio en una vecindad en GO con nodos anotados.
- $ktal_{anotados}$ KTA en la vecindad de los nodos anotados.



A modo de ejemplo, la red para tratamiento “Frío” consta de 1951 nodos y 18436 aristas.

Caracterización de vecindades locales tratamiento “Frío”



Métrica mixta

Dada una arista, el peso de una arista y el promedio de pesos, tenemos una manera de cuantificar si una vecindad es o no biologicamente coherente.

Vamos a usar esto para encontrar grupos transcripcionales teniendo en cuenta las coherencias biológicas locales modificando los pesos:

$$w_{ij} = \text{simcor}_{ij}^{\beta * \text{stress}_{ij}} \quad (5)$$

Donde:

$$\text{stress}_{ij} = \frac{KTA_{fondo}}{KTA_{ij}} \quad (6)$$

Típicamente el *stress* oscila entre 0,8 y 1,2.

β es un parámetro que va a acentuar las heterogeneidades para poder detectar subgrupos.

Métrica mixta y métodos heurísticos

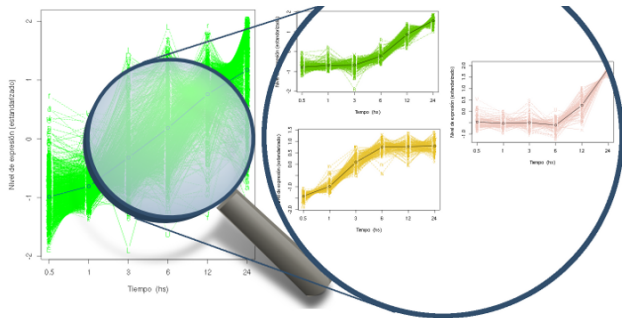
Buscamos subestructura en los grupos a partir de la métrica mixta

Heurísticas con métrica mixta:

- lkta.dtc
- lkta.cnm
- lkta.infomap

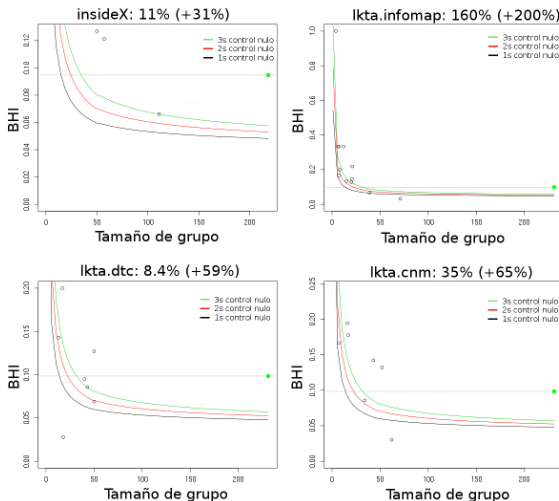
Comparación con métrica transcripcional:

- InsideX



Coherencia biológica medida con BHI - Ejemplo

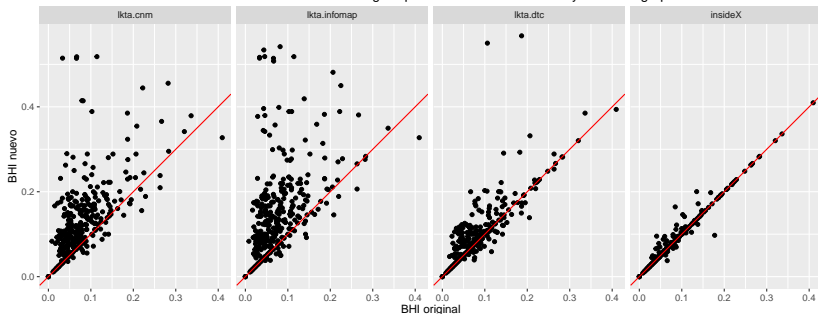
Subestructura en grupo 2 de tratamiento “Frío”



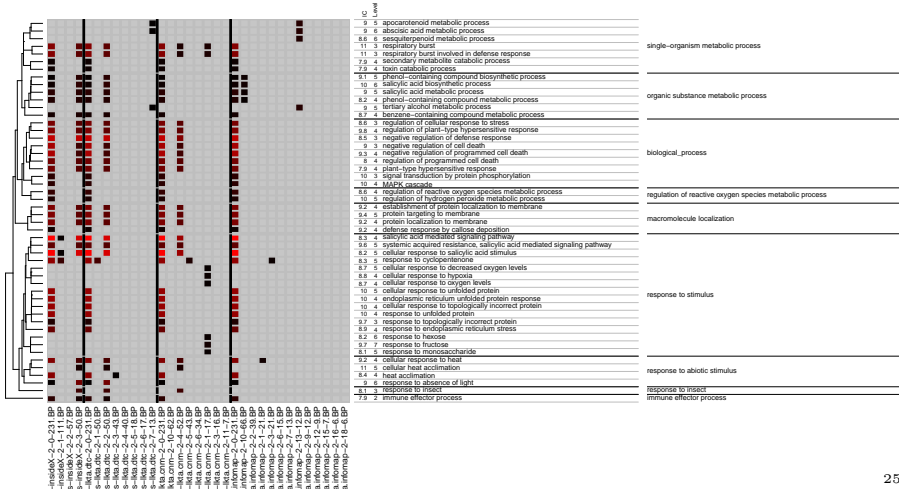
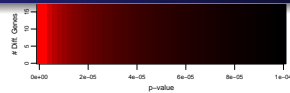
Coherencia biológica medida con BHI

Caracterizamos los nuevos subgrupos hallados

BHI nuevo en función de BHI original para todos los tratamientos y todos los grupos



Coherencia biológica a partir de sobrerrepresentación (test de Fisher)



Conclusiones

- Diferentes técnicas de agrupamiento nos permitieron obtener grupos correlacionados en el espacio de expresión.
- Cada método obtiene descripciones a diferente resolución.
- Buscamos analizar estas descripciones en función de la interpretabilidad biológica.
- Presentamos los observables BHI y KTA para cuantificar la coherencia entre los espacios y corroboramos que lo detectado en el espacio transcripcional es en general coherente con el conocimiento biológico.
- Introdujimos una versión local de KTA que nos permitió definir una métrica mixta.
- Presentamos heurísticas para identificar subestructuras transcripcionales con alta interpretabilidad y coherencia biológica.

Muchas gracias

¡Muchas gracias por su atención!

¿Preguntas o comentarios?

Agradecimientos

