

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**
 - Detección de correlaciones
 - Relevamientos transcripcionales de gran escala
- 2 **Análisis de relevamientos transcripcionales**
 - Medidas de similaridad y distancia
 - Métodos de agrupamiento utilizados
 - 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 y perspectivas**

Detección de correlaciones

Queremos encontrar relaciones entre grandes cantidades de datos.

Lo vamos a hacer usando métodos de agrupamiento o “clustering”.

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

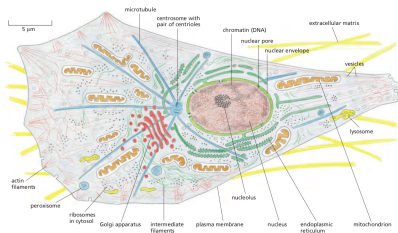
A modo de ejemplo

El conjunto: $\{-5, -3, -2, 2, 3\}$

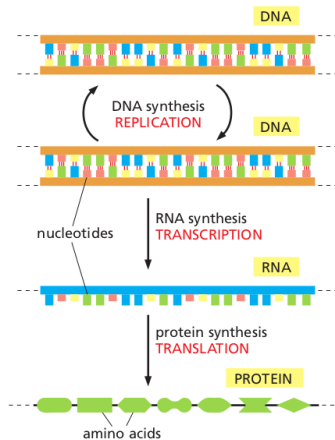
Agrupado por módulo: $\{-5\}$, $\{-3, 3\}$ y $\{-2, 2\}$

Agrupado por signo: $\{-5, -3, -2\}$ y $\{2, 3\}$

Transcripción y traducción (dogma central de la biología molecular)



(a) Célula eucariota

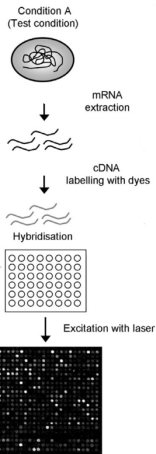


(b) Dogma central de la biología molecular

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

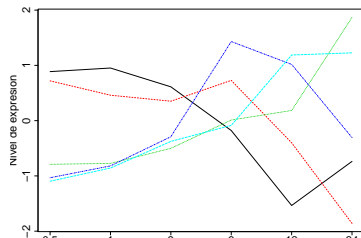


A. Thaliana



Datos de estrés abiótico:

- 11 tratamientos
- ≈ 22000 genes
- entre 4 y 8 mediciones temporales por gen y por tratamiento

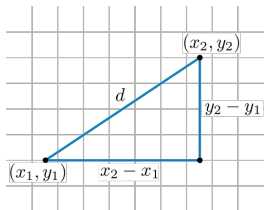


Medidas de similaridad y distancia

Necesitamos definir que significa que dos datos sean “similares”

Distancia euclidiana en espacio de alta dimensionalidad:

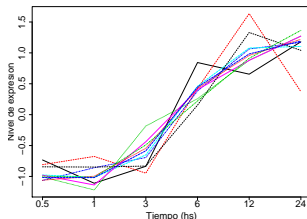
$$d_{euc}(\vec{x}, \vec{y}) = \left[\sum_{i=1}^n (x_i - y_i)^2 \right]^{\frac{1}{2}} \quad (1)$$



Distancia basada en el coeficiente de correlación de Pearson:

$$r(\vec{x}, \vec{y}) = \frac{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{s_x s_y} \quad (2)$$

$$d_{ccp}(\vec{x}, \vec{y}) = 1 - r(\vec{x}, \vec{y}) \quad (3)$$



Método de agrupamiento k-means

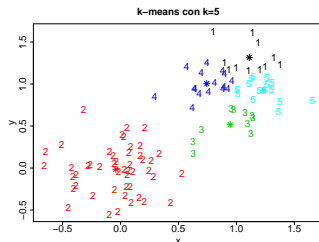
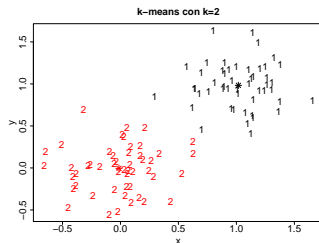
- Agrupamiento no jerárquico.
- Cada observación pertenece al grupo con la media más cercana.
- La cantidad k de grupos debe ser fijada a priori.
- Utiliza la distancia euclidiana.

Para datos estandarizados:

$$\tilde{x}_i = \frac{x_i - \bar{x}}{s_x} \quad (4)$$

la distancia euclidiana se relaciona con la correlación como:

$$d(\vec{x}, \vec{y}) = \sqrt{2(n-1)(1-r(\vec{x}, \vec{y}))} \quad (5)$$

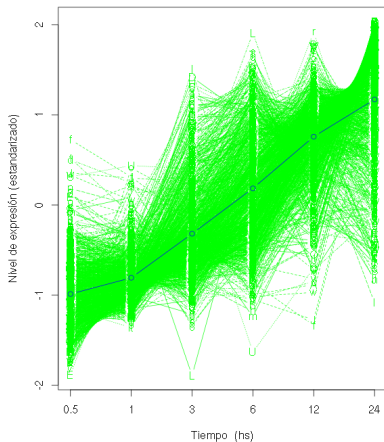


- Agrupamiento jerárquico.
- Utiliza la distancia de correlación.
- Se puede “sintonizar” la resolución del método.
- DS1 particiones gruesas, con pocos grupos bien definidos.
- DS4 particiones finas, con muchos grupos más dispersos.

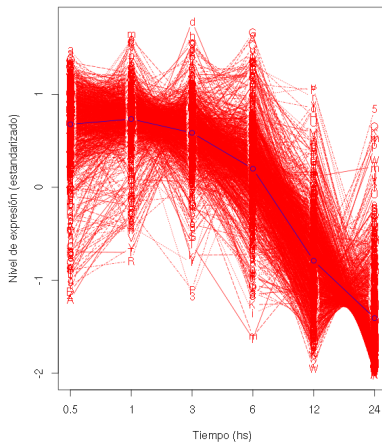


Perfiles tratamiento "Frío" con k-means

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

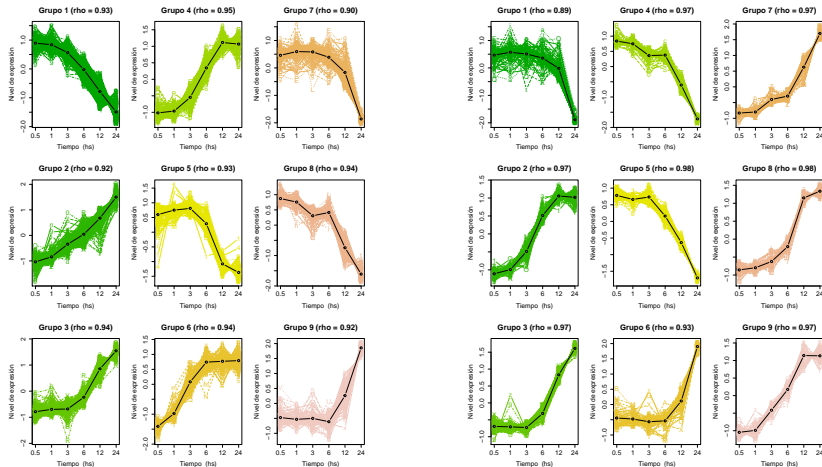


Perfil de expresión grupo 2 tratamiento 'Frío' y k=2 ($\rho = 0.79$)



Perfiles tratamiento “Frío” con corte de árbol dinámico

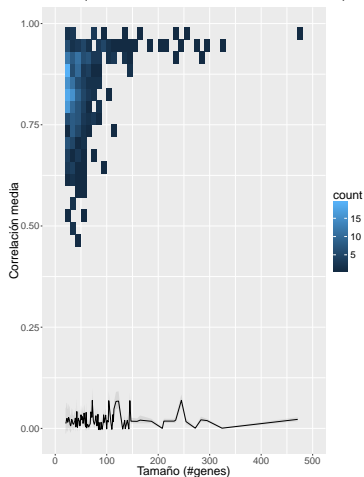
A modo de ejemplo, los nueve perfiles más grandes de cada partición
DS1 (particiones más gruesas) DS4 (particiones más finas)



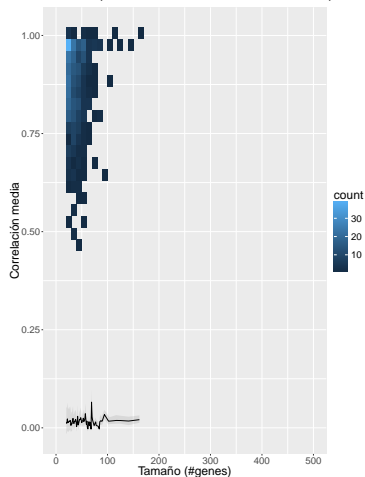
Caracterización de particiones corte de árbol dinámico

Correlación media por tamaño de grupo

DS1 (particiones más gruesas)

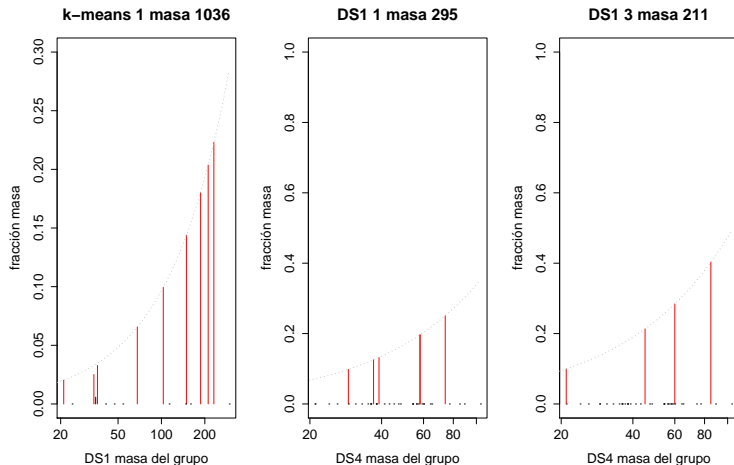


DS4 (particiones más finas)



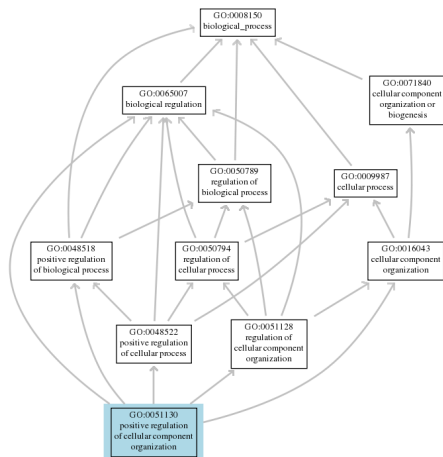
Caracterización de granularidad de las particiones halladas

Fracción de grupos en una partición más fina dentro de grupos de una partición más gruesa (tratamiento “Frío”)



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 sí 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:

$$ID(GO_j) = \frac{NE(GO_j)}{N(GO_j)} \quad (6)$$

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:

$$BHI_j = \frac{1}{n_j(n_j - 1)} \sum_{x \neq y \in D_j} I(C(x) = C(y)) \quad (7)$$

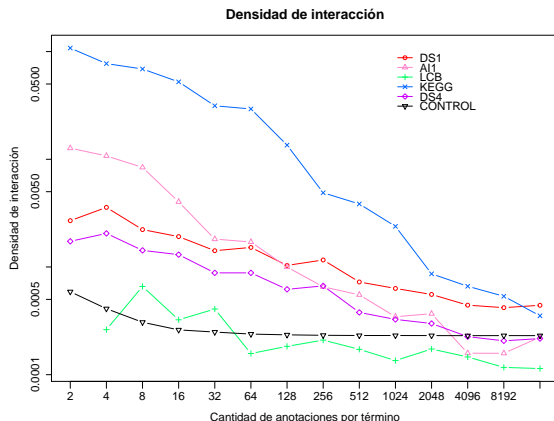
Con n_j la cantidad de genes anotados

en el grupo D_j .

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

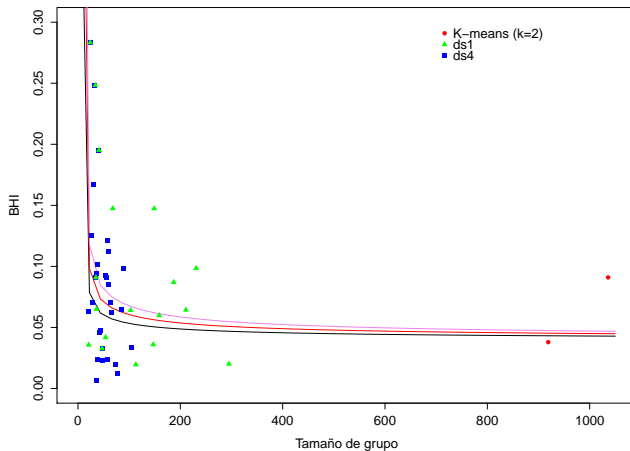
Densidad de interacción

- 1 Términos mas específicos presentan mayor ID.
- 2 Agrupamientos hasta 100 genes correlacionan con la información biológica embebida en la ontología.
- 3 Las estructuras en KEGG presentan mayor congruencia con las ontologías, seguidas por AI1 y expresión.
- 4 *ds1* presenta mayor congruencia biológica que *ds4*. Indicio acerca de la escala de granularidad apropiada.



Indice de homogeneidad biológica

BHI para tratamiento 'Frío' con control nulo 1



Particiones altamente coherentes pero de baja calidad de BHI.

Similaridad entre genes en GO

Definimos la similaridad entre genes en el espacio GO como:

$$Sim_{rcmax}(GO(g_1), GO(g_2)) = \max\left\{\frac{1}{N} \sum_i \max_{1 \leq j \leq M} S_{ij}, \frac{1}{M} \sum_j \max_{1 \leq i \leq N} S_{ij}\right\} \quad (8)$$

Donde:

$$S(g_1, g_2)_{ij} = Sim_{res}(GO(g_1^i), GO(g_2^j)), \forall i \in \{1, \dots, N\} y \forall j \in \{1, \dots, M\} \quad (9)$$

con:

$$Sim_{res}(c_i, c_j) = \max_{c \in S(c_i, c_j)} (-\log_2[P(c)]) = IC(MICA[c_i, c_j]) \quad (10)$$

la similaridad entre términos.

KTA global

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

$$K = K_{ij} = k(x_i, x_j) \quad (11)$$

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

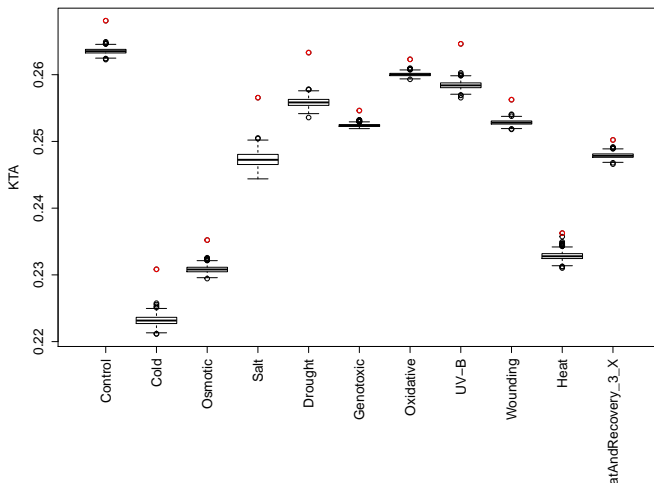
cuantifica la similaridad entre dos espacios y se define como:

$$\hat{A}(C, k_1, k_2) = \frac{\langle K_1, K_2 \rangle_F}{\sqrt{\langle K_1, K_1 \rangle_F \langle K_2, K_2 \rangle_F}} \quad (12)$$

con $\langle K_1, K_1 \rangle_F = \sum_{i,j=1}^m K1(x_i, x_j)K2(x_i, x_j)$ es el producto interno de Frobenius.

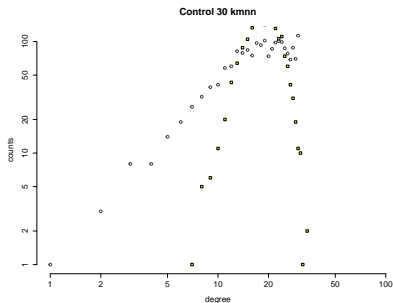
KTA global

KTA global entre expresión y ontología BPB con control nulo



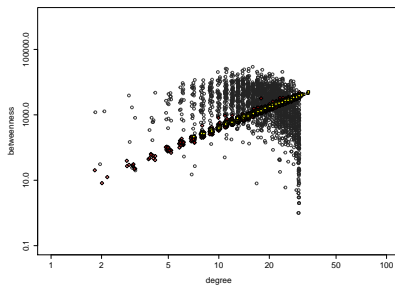
Red 30 primeros vecinos mutuos

Distribución de grado



Red y modelo nulo Erdős-Renyi

Betweenness



Red, modelo nulo Erdős-Renyi y
modelo configuracional

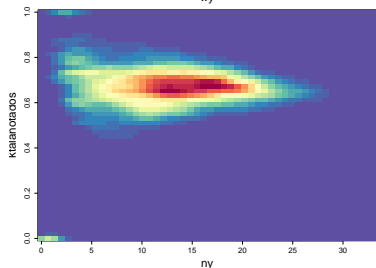
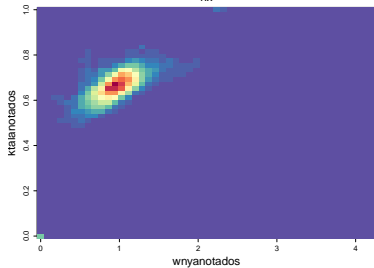
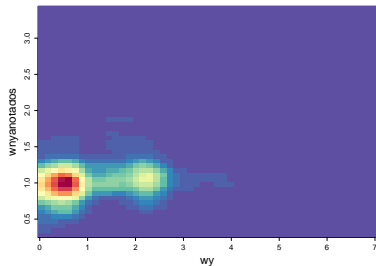
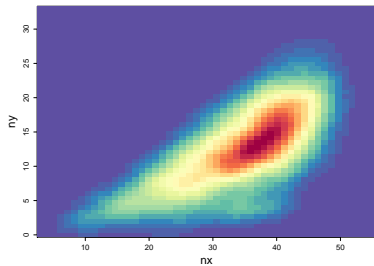
Red 30 primeros vecinos mutuos - vecindades locales

Vecindad de una arista a primeros vecinos:

GRAFICO DE VECINDAD

- n_x nodos.
- n_y nodos anotados.
- wyn promedio de pesos de aristas en GO.
- $wyn_{anotados}$ promedio de pesos de aristas en GO con nodos anotados.

Caracterización de vecindades locales tratamiento “Frío”



Métrica mixta

Modificaremos los pesos de las aristas utilizando la información de coherencia biológica presente en KTA:

$$w_{ij} = \text{simcor}_{ij}^{\text{stress}} \quad (13)$$

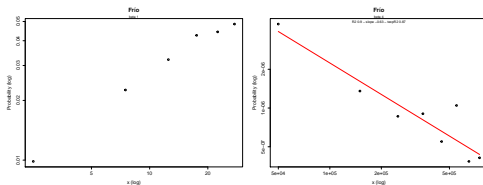
Donde:

$$\text{stress} = \frac{KTA_{\text{fondo}}}{KTA_{ij}} \quad (14)$$

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

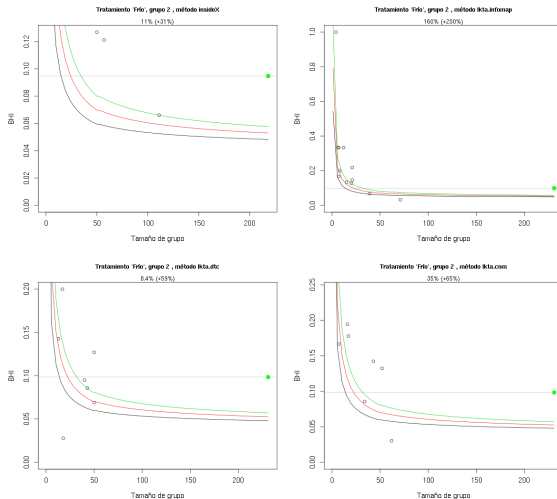
Buscamos un parámetro β tal que la distribución de *strenght* de la red siga una ley de potencias:

$$w'_{ij} = w_{ij}^{\beta} \quad (15)$$



Métodos heurísticos

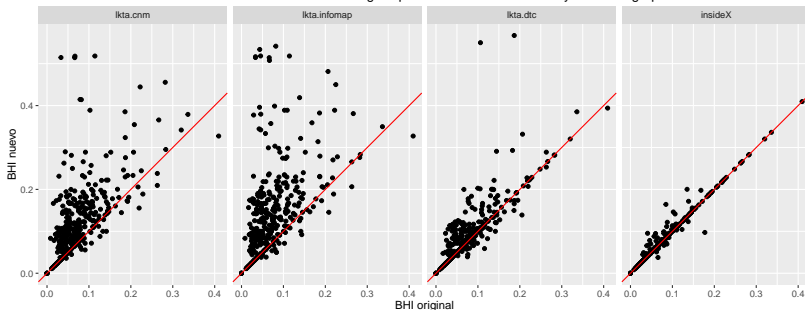
Buscamos subestructura en los grupos a partir de la métrica mixta

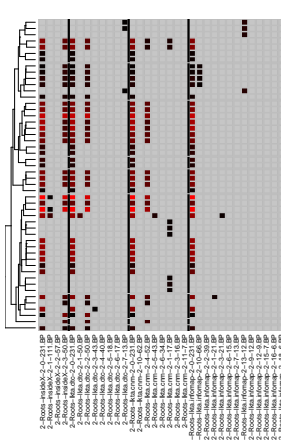


Métodos heurísticos - caracterización de particiones

Caracterizamos los nuevos subgrupos hallados

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





8	1	5	1	
9	1	5	1	
10	1	5	1	
11	1	5	1	
12	1	5	1	
13	1	5	1	
14	1	5	1	
15	1	5	1	
16	1	5	1	
17	1	5	1	
18	1	5	1	
19	1	5	1	
20	1	5	1	
21	1	5	1	
22	1	5	1	
23	1	5	1	
24	1	5	1	
25	1	5	1	
26	1	5	1	
27	1	5	1	
28	1	5	1	
29	1	5	1	
30	1	5	1	
31	1	5	1	
32	1	5	1	
33	1	5	1	
34	1	5	1	
35	1	5	1	
36	1	5	1	
37	1	5	1	
38	1	5	1	
39	1	5	1	
40	1	5	1	
41	1	5	1	
42	1	5	1	
43	1	5	1	
44	1	5	1	
45	1	5	1	
46	1	5	1	
47	1	5	1	
48	1	5	1	
49	1	5	1	
50	1	5	1	
51	1	5	1	
52	1	5	1	
53	1	5	1	
54	1	5	1	
55	1	5	1	
56	1	5	1	
57	1	5	1	
58	1	5	1	
59	1	5	1	
60	1	5	1	
61	1	5	1	
62	1	5	1	
63	1	5	1	
64	1	5	1	
65	1	5	1	
66	1	5	1	
67	1	5	1	
68	1	5	1	
69	1	5	1	
70	1	5	1	
71	1	5	1	
72	1	5	1	
73	1	5	1	
74	1	5	1	
75	1	5	1	
76	1	5	1	
77	1	5	1	
78	1	5	1	
79	1	5	1	
80	1	5	1	
81	1	5	1	
82	1	5	1	
83	1	5	1	
84	1	5	1	
85	1	5	1	
86	1	5	1	
87	1	5	1	
88	1	5	1	
89	1	5	1	
90	1	5	1	
91	1	5	1	
92	1	5	1	
93	1	5	1	
94	1	5	1	
95	1	5	1	
96	1	5	1	
97	1	5	1	
98	1	5	1	
99	1	5	1	
100	1	5	1	

Conclusiones y perspectivas

- Mediante técnicas de agrupamiento de datos fue posible encontrar grupos de genes con perfiles de expresión altamente correlacionados.
- Distintos métodos darán distintas particiones en función de la resolución que logran.
- Mediante una métrica mixta fue posible encontrar particiones con alta homogeneidad biológica y con alta correlación transcripcional.
- Utilizamos la ontología GO para dar una interpretación biológica a los grupos obtenidos y encontramos que en general, la granularidad óptima de los grupos fue de ≈ 50 genes.
- Estas técnicas podrían funcionar como punto de partida para inferir funciones biológicas de genes de los que se tiene poco conocimiento.
- Sería interesante en un futuro agregar la información contenida en otros espacios de conocimiento biológico, como ser vías metabólicas o redes de interacción de proteínas.

Agradecimientos

¡Muchas gracias!
FOTO DEL GRUPO