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

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.

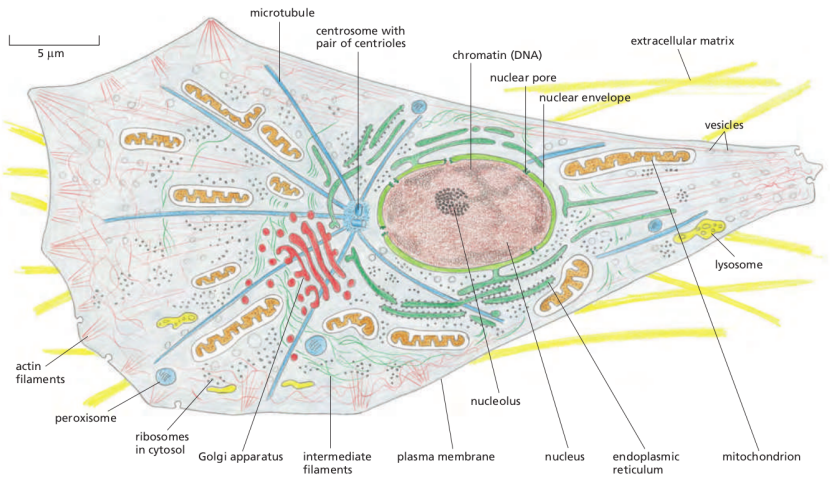
A modo de ejemplo

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

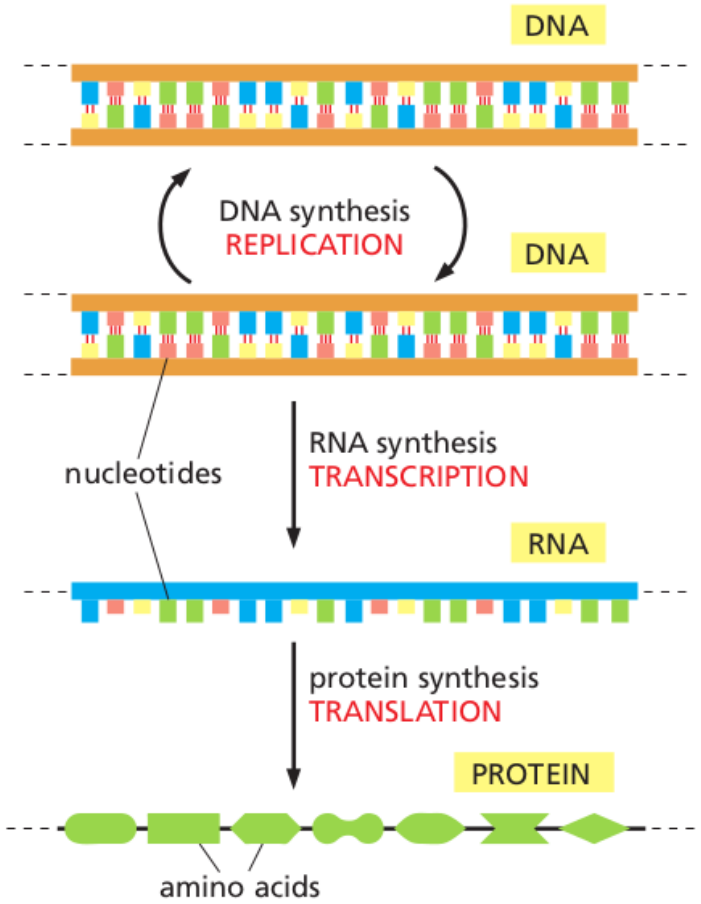
Agrupado por módulo: $\{-5\}$, $\{-3, 3\}$ y $\{-2, 2\}$

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

Transcripción y traducción (dogma central de la biología molecular)



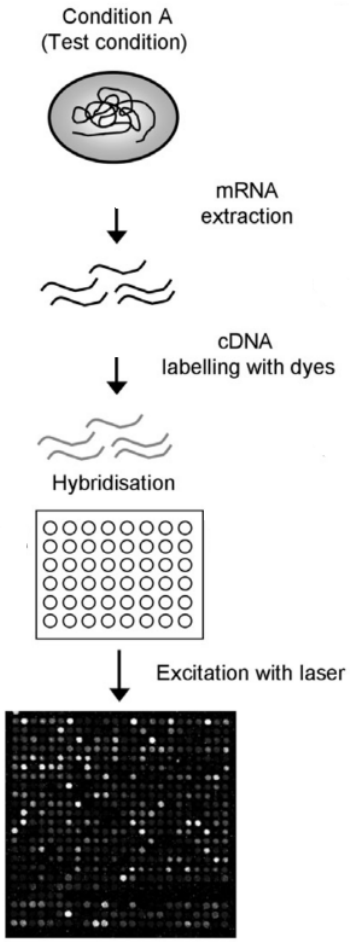
(a) Célula eucariota



(b) Dogma central de la biología molecular

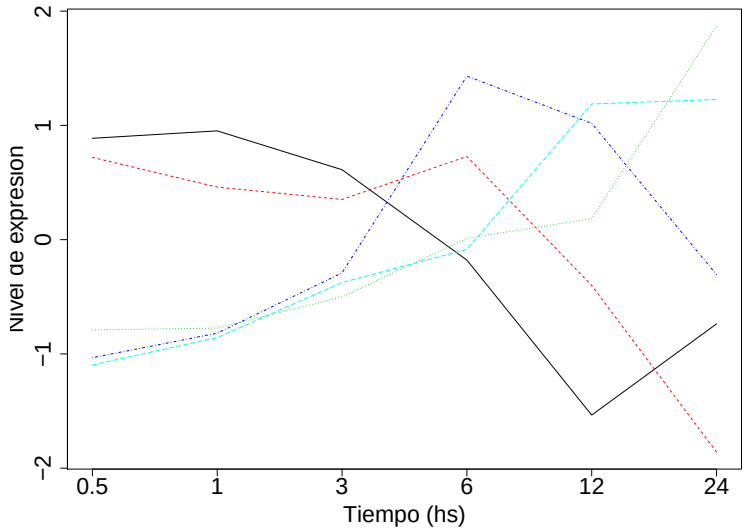
Cambios transcripcionales en respuesta a estrés abiótico en *A. thaliana*

A. Thaliana



Datos de estrés abiótico:

- 11 tratamientos
- $\approx$ 22000 genes
- entre 4 y 8 mediciones temporales por gen y por tratamiento

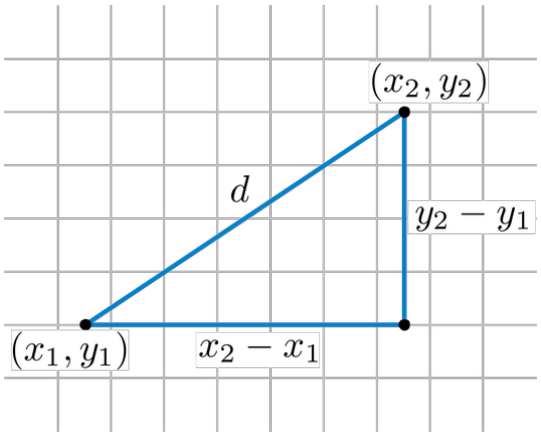


Medidas de similaridad y distancia

Necesitamos definir que significa que dos datos sean "similares"

Distancia euclidiana en espacio de alta dimensionalidad:

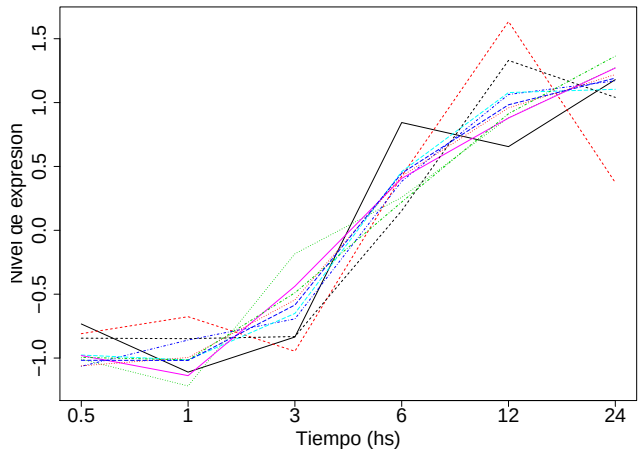
$$d_{euc}(\vec{x},\vec{y}) = [\sum_{i=1}^{n}(x_i - y_i)^2]^{\frac{1}{2}} \quad (1)$$



Distancia basada en el coeficiente de correlación de Pearson:

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

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



Tipos de agrupamiento

Métodos k-means, corte de árbol dinámico

Caracterización de particiones

El problema de la escala

Ontología génica (GO)

Densidades de interacción

Indice de homogeneidad biológica

Coherencia entre métrica transcripcional y espacio GO

KTA global

KTA local para modulación de heterogeneidades transcripcionales

Métrica mixta

Método heurístico

Interpretación biológica

Conclusiones y perspectivas