Universidad de Santiago de Chile
Departamento de Ingeniería Informática

BIOINFORMÁTICA (2016-2)
Proyecto Final
Herramienta Web de Visualización de Soluciones de Clustering
Profesor: *Mario Inostroza Ponta*
Ayudante: *Jorge Párraga-Álava (jorge.parraga@usach.cl)*

Octubre, 2016

1. Introducción

La visualización del clustering es una de las formas más frecuentes de mostrar datos de alta-dimensionalidad y de evaluar visualmente los resultados obtenidos por los algoritmos de clustering, sobre todo en estudios que buscan similitudes y diferencias en diversos materiales biológicos. Las dos herramientas de visualización más usados en el clustering de genes son *Eisen plot*[1] y *cluster profile plot*[2] gracias a las cuales es posible observar cómo los perfiles de expresión de diferentes clústeres de genes difieren uno de otro mientras que los perfiles de un mismo clúster son más similares. A raíz de ello, este proyecto se centra en crear una herramienta que permita la visualización de clústeres de diversa índole con enfoque en el clustering de genes de datos de expresión.

2. Trabajo a realizar

En grupos de tres estudiantes desarrollar las siguientes actividades:

- Crear una herramienta computacional web que permita la visualización de diversas soluciones de clustering a partir de las particiones o desde matrices de distancia.
- Para el caso de clustering de genes, se deberán mostrar los clústeres en forma de *eisen plot* y *cluster profile plot*, para el efecto la herramienta deberá utilizar la matriz de expresión génica o la matriz de distancia en conjunto con las particiones de los genes. Para cada solución, incluir los n-términos significantes de Gene Ontology (GO) (podrán usar servicios web, o librerías de R como *"RDAVIDWebService"*) para cada clúster con su respectivo *p-value*, en términos de procesos biológicos asociados. Debido a que existen muchas nomenclaturas para identificar genes, es necesario que la herramienta a desarrollar considere la conversión de tales nombres cuando estos no se encuentren en el estándar usado por los términos de GO (R, dispone de librerías como "*AnnotationDbi*" que facilitan esta conversión).
- Cuando no se disponga de la matriz de expresión génica, se utilizarán las matrices de distancias de los elementos a agrupar, por ende la visualización de las soluciones de clustering se realizará en forma de *grafos* donde los nodos corresponden a los elementos que han sido clusterizados y las aristas a la distancia entre ellos. En este caso, cada clúster se visualizará usando un elemento identificatorio (color, forma, tamaño, etc.).
- En todas las formas de visualización, para cada solución de clustering se deberá incluir el valor correspondiente a los índices Silueta, Davies-Bouldin, Dunn, así como la compactacidad (cohesión) y separación de los clústeres, además el número de elementos en cada clúster.
- El formato de las matrices a usar como entradas es el siguiente:

[1] La expresión génica se muestra mediante el uso de heatmaps donde habitualmente los valores coloreados en verde corresponden células sanas, los coloreados en rojo a células enfermas y la ausencia de valores de expresión diferencial se denotan con color negro. Partiendo de ello, en un eisen plot colores similares son agrupados juntos denotando que los perfiles de expresión de los genes de un clúster son similares unos con otros.

[2] Valores normalizados, promedios y desviación estándar de la expresión de los genes, de cada clúster con respecto a las condiciones, puntos de tiempo o muestras.

	Solución 1	…	Solución p
Elemento 1	4	…	1
Elemento 2	4	…	3
Elemento 3	1	…	3
Elemento 4	3	…	4
Elemento 5	2	…	4
…	…	…	…
Elemento n	3	…	2

Tabla 1. Formato de solución de clustering (partición).

	Condición 1	…	Condición m
Gen 1			
…			
Gen n			

Tabla 2. Formato de matriz de expresión génica.

	Gen 1	Gen 2	…	Gen n
Gen 1	0	0.44	…	0.12
Gen 2	0.44	0	…	0.83
…	…	…	…	…
Gen n	0.12	0.83	…	0

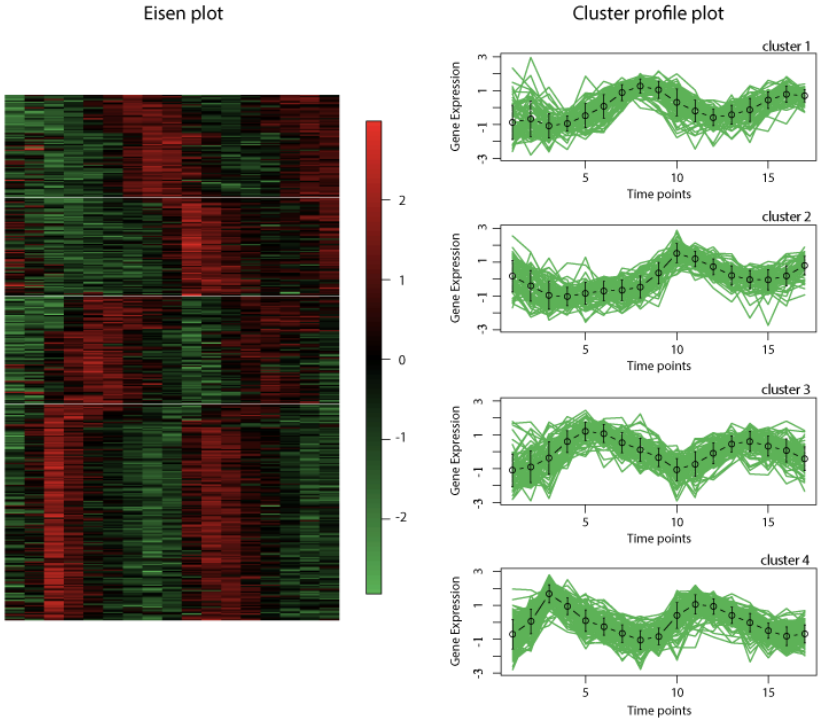
Tabla 3. Formato de matriz de distancia.

3. Bases de datos

Se sugiere utilizar las bases de datos *Arabidopsis Thaliana* (disponibles en UdeSantiagoVirtual) que consta de matriz de expresión, matriz de distancia y soluciones de clustering (particiones). Para comprobar la correcta funcionalidad de la herramienta desarrollada, además pueden obtener soluciones de clustering usando algún algoritmo de clustering (k-means, UPGMA, PAM, etc.).

4. Índices de calidad

- http://wwwae.ciemat.es/~cardenas/docs/lessons/Silueta.pdf
- http://wwwae.ciemat.es/~cardenas/docs/lessons/Davies-Bouldin.pdf
- http://wwwae.ciemat.es/~cardenas/docs/lessons/Dunn.pdf



Ejemplo de visualización mediante eisen y cluster profile plot.

Cluster	Término GO significante	p-value
Cluster 1	Sporulation GO:0043934	4.36e-23
	Sporulation resulting information of a cellular spore GO:0030435	3.17e-20
	Anatomical structure formation involved in morphogenesis GO:0048646	3.62e-43
Cluster 2	Ribosome GO:0030684	2.91e-13
	Ribosome biogenesis GO:0042254	1.45e-18
	Ribonucleoprotein complex biogenesis GO:002261	6.78e-24
Cluster 3	Cytosolic ribosome GO:0022626	2.91e-14
	Cytoplasmic translation GO:0002181	7.14e-34
	Structural constituent of ribosome GO:0003735	6.86e-13
Cluster 4	Glycolysis GO:0006096	1.12e-27
	Glucose catabolicprocess GO:0006007	3.74e-18
	Hexosecatabolicprocess GO:0019320	8.06e-26

Ejemplo de los tres más significantes términos de GO con sus correspondientes p-values.