Capítulo 1

Congruencia biológica de las particiones halladas

Esperamos que los conocimientos (entendidos como nociones de similitud) de los distintos espacios (el de expresión y el biológico) sean diferentes pero no ortogonales. Por lo tanto, una vez detectadas las estructuras en distintas resoluciones en el espacio de expresión, nos interesará cuantificar la congruencia biológica de las mismas. Para ello haremos uso de varios índices, BHI, BHI_{IC} IC, BHI_{Resnik} y zBHI que servirán como criterios biológicos de validación externos.

1.1. Índice de homogeneidad biológica

El índice de homogeneidad biológica (o BHI por sus siglas en inglés) de una partición, introducido por Datta [**?**] es un observable que cuantifica el grado en que una partición presenta grupos biológicamente homogéneos, reportando, para cada grupo, la máxima proporción de pares de genes agrupados que comparten una misma clase funcional de Ontología Génica. Consideremos dos genes x e y que pertenecen a un mismo grupo D de una partición dada, con un total de k grupos, y sean $C(x)$ y $C(y)$ los conjuntos de todas las clases funcionales que tienen anotados a los genes x e y respectivamente. Sea además 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. Entonces, el índice de homogeneidad biológica queda definido como:

$$BHI = \frac{1}{k} \sum_{j=1}^{k} \frac{1}{n_j(n_j - 1)} \sum_{x \neq y \in D_j} I(C(x) = C(y)) \tag{1.1}$$

con n_j la cantidad de genes anotados en el grupo D_j.
Los valores de BHI calculados para cada uno de los grupos del tratamiento "frío" en las particiones k-means (puntos rojos), *deepsplit* = 1 (triángulos verdes) y *deepsplit* = 4

(cuadrados azules) se presentan en la figura 1.1 junto con un control nulo consistente en 1000 reasignaciones aleatorias de las etiquetas de cada partición. Los grupos fueron ordenados según su masa de forma creciente.
Se observa que de los dos grupos de kmeans, solo uno presenta un BHI superior al tercer cuartil para el control nulo, mientras que para *deepsplit* = 1, solo el 30 % de los grupos lo superan. Finalmente, para la partición *deepsplit* = 4, el BHI del 40 % de

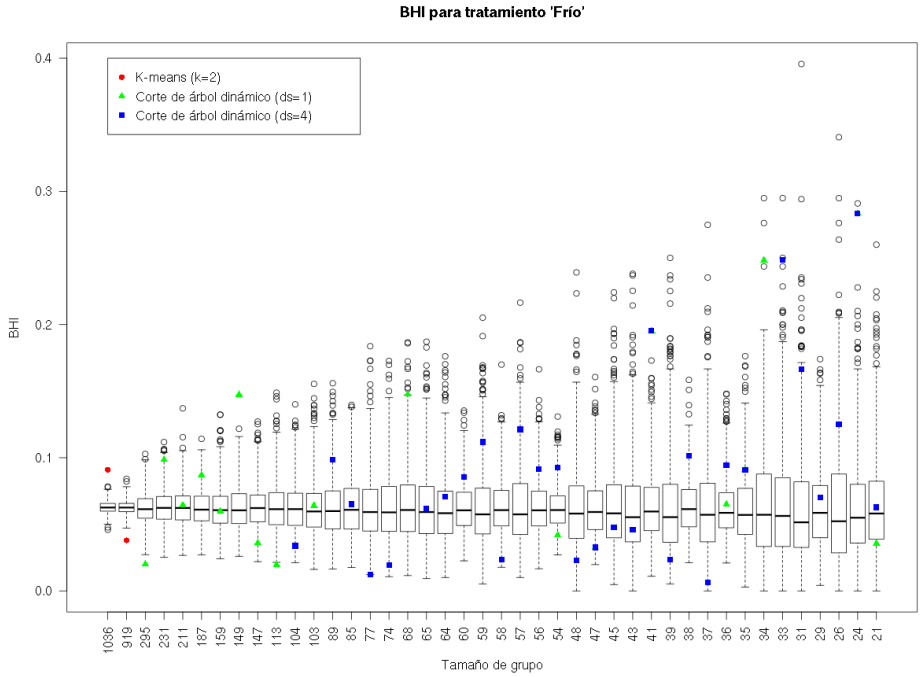

Figura 1.1: Índice de Homogeneidad Biológica, BHI, para cada uno de los grupos del tratamiento 'Frío' obtenidos con kmeans, *deepsplit* = 1 y *deepsplit* = 4.

los grupos se encuentra por sobre el tercer cuartil del control nulo. Esto sugiere que si bien el aumentar la granularidad de la partición con el método corte de árbol dinámico resulta en un aumento de la consistencia biológica global de las estructuras observadas, esto no implica que las resoluciones utilizadas sean las óptimas.

1.2. Modificaciones al Índice de homogeneidad biológica

Presentaremos a continuación dos variantes del BHI que modificarán la función indicadora para hacer uso de la similaridad semántica y del contenido de información génico.

El índice de homogeneidad biológica con contenido de información (BHI_{IC}) para un grupo se define como:

$$BHI_{IC} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in D_j} I(C(x)=C(y))IC(Gx) \tag{1.2}$$

donde el $IC(Gx)$ es el contenido de información del gen x, definido como el máximo de los contenidos de información de los conceptos en los que el gen x se encuentra anotado. Este índice permite pesar la homogeneidad biológica de un grupo la especificidad de los genes que lo componen.
Por otro lado, el índice de homogeneidad biológica Resnik para un grupo, BHI_{Resnik} queda definido como:

$$BHI_{Resnik} = \frac{1}{k}\sum_{j=1}^{k}\frac{1}{n_j(n_j-1)}\sum_{x\neq y\in D_j} I(C(x)=C(y))Sim_{rcmax}(C(x),C(y)) \tag{1.3}$$

donde $Sim_{rcmax}(C(x),C(y))$ es como fuera definida en **??**.
Este índice pesa la homogeneidad biológica de un grupo por la similaridad semántica de los genes que lo componen. Notar que en BHI_{IC} el peso viene dado por la especificidad de cada gen individual que compone una partición, mientras que en BHI_{Resnik} el peso está dado por la similaridad semántica entre pares de genes.
Finalmente, el índice de homogeneidad biológica estandarizado para un grupo, zBHI, se define como:

$$zBHI = \frac{BHI - < BHI_r >}{s(BHI_r)} \tag{1.4}$$

donde BHI_r es el conjunto de valores del BHI del grupo para las 1000 particiones aleatorias.
Para caracterizar el comportamiento de cada uno de estos índices se midieron los mismos para el tratamiento "Frío" y se calculó su correlación de a pares de índices. La figura 1.2 muestra las distribuciones y correlaciones de a pares para estos índices. Se encuentra que los índices modificados tienen una alta correlación entre ellos y con BHI, y por lo tanto, no aportan más información que la que se obtiene a través del índice original. Por ser el más sencillo de calcular, es el que utilizaremos como criterio de validación externa de la calidad de una partición.

1.3. Discusión

Ojo que esto quedo de la discusion del capitulo anterior En el presente análisis de estructura de los grupos obtenidos por medio de los métodos k-means, *deepsplit* = 1 y *deepsplit* = 4, encontramos que todos los métodos producen particiones altamente

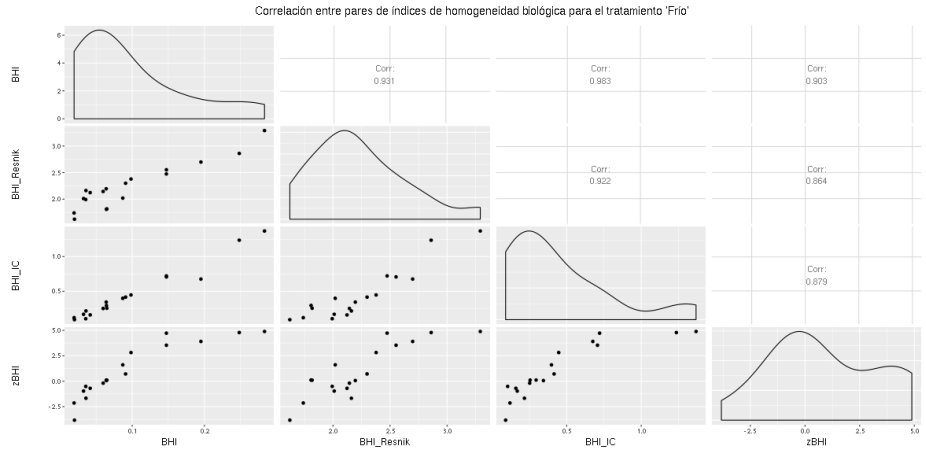

Figura 1.2: Correlación de a pares para los distintos índices de homogeneidad biológica presentados para cada uno de los grupos del tratamiento 'Frío' obtenidos con $ds = 1$. Se observa que todos los índices tienen una alta correlación entre si.

coherentes, con el método k-means generando las particiones más gruesas y los métodos subsiguientes, refinamientos de las mismas. La alta coherencia detectada es indicativo de que cada método logra hallar estructuras en el espacio de expresión génica, aunque no siempre es factible realizar una interpretación biológica de las estructuras encontradas. Sobre todo en los grupos encontrados con k-means, que solamente toman en cuenta la expresión o inhibición de los genes. Sin embargo, el análisis de BHI indica que el aumentar la resolución con corte de árbol dinámico tampoco consigue encontrar la escala óptima en el análisis.
En el capitulo siguiente introduciremos algunas herramientas que buscarán cuantificar la homogeneidad biológica de particiones para encontrar dicha escala.

Bibliografía