

Proyecto del Módulo de Clustering

Escribir un programa en R que realice un clustering jerárquico con un archivo de secuencias de proteínas en formato FASTA.

Requerimientos:

1. Leer un archivo de secuencias homologas de proteínas en formato fasta.
 - a) No más de 100 para que corra rápido y se pueda interpretar fácilmente el arbol. Asegurense que las proteínas tengan más o menos el mismo tamaño.
 - b) Incluyan proteínas de grupos taxonómicos bien definidos. Modifiquen el identificador de las proteínas para agregar el grupo taxonómico (e.g. NP_41487_entero o NP_41487_gamma). Así, cuando visualizen el arbol in FigTree rápidamente podrán checar la congruencia con la taxonomía.
2. Correr BLASTP de todas las secuencias contra todas las secuencias.
 - a) usar las opciones: `-outfmt 7 -max_hsps 1 -use_sw_tback`
3. Generar una matriz de disimilitud (distancia) con base en los bit scores generados. Es decir, conviertan los bit scores (que son una medida de similitud) en una matriz de disimilitudes que pueda ser pasada al algoritmo de agrupamiento usando la función “as.dist()”:
 - a) Primero normalicen los bit scores (b) o similitudes para que queden en el rango $[0,1]$.
 Ignorando los scores en la diagonal de la matriz de similitud, calcular un bit score normalizado ($B_{i,j}$) por cada par de proteínas i, j dividiendo todos los bit scores ($b_{i,j}$) por el valor del bitscore más alto:

$$B_{i,j} = \frac{b_{i,j}}{\max(b_{x,y} : x,y=1..n)} : i \neq j, x \neq y, \text{ considerando que } B_{i,j} = 1 \text{ cuando } i = j$$
 - b) Par convertir las similitudes $B_{i,j}$ en disimilitudes (o distancias) solo es necesario:

$$d_{i,j} = 1 - B_{i,j}$$
4. Correr clustering jerárquico y correr varios métodos para obtener el número de clusters.
5. Salvar el dendograma como árbol filogenético en formato **Newick** en R.
6. Comparar los árboles obtenidos cuando se aplican los métodos single, average, complete y ward.
7. Obtener el "Agglomerative Coefficient" de todos los árboles.

Entrega de proyecto y evaluación del módulo (pueden formar equipos de hasta cinco personas):

Terminar el programa y entregar un reporte con los resultados de la comparación entre los diferentes métodos. El reporte debe contener lo siguiente:

- a) Una introducción que incluya las consideraciones más importantes que se deben tomar en cuenta en el análisis de clustering.
- b) Incluir imágenes de los diferentes árboles y discutan sus impresiones. ¿Cuál es el árbol más informativo? ¿Cuál es el árbol menos informativo? ¿Cuántos árboles son congruentes con la taxonomía de las proteínas?
- c) ¿Cuál es el árbol con el agglomerative coefficient más alto?

Fecha límite de entrega: viernes 20 de Marzo a la media noche.