Proyecto del Módulo de Clustering

Escribir un programa en Python o Perl que internamente llame a R y 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 rapidamente podrán checar la congruenca con la taxonomía.

2. Correr BLASTP de todas las secuencias contra todas las secuencias.

a) usar: -outfmt 7 -max_hsps 1 -use_sw_tback

3. Generar una matriz de disimilitud (distancia) con base en los bit scores generados.

4. Normalizar las disimilitudes (*d*) para que queden en el rango [0,1].

Ignorando las comparaciones en la diagonal de la matriz de distancia, Calcular un bit score normalizado ($B_{i,j}$) por cada par the proteins i, j dividiendo todos los bit scores ($b_{i,j}$) por el valor del bitscore más alto:

$$B_{i,j} = \frac{b_{i,j}}{\max\left(b_{x,y} : x,y=1..n\right)} : i \neq j, x \neq y, \text{ considerando que } B_{i,j} = 1 : i = j$$

En este caso $B_{i,j}$ es una similitud. Para obtener la disimitud solo es necesario:

$$d_{i,j} = 1 - B_{i,j}$$

5. Correr clustering jerárquico y correr varios métodos para obtener el número de clusters.

6. Salvar el dendograma como árbol filogenético en formato **Newick** en R.

7. Comparar los árboles obtenidos cuando se aplican los métodos single, average, complete y ward.

8. 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? ¿Cuantos árboles son congruentes con la taxonomía de las proteínas?

c) ¿Cuál es el árbol con el agglomerative coefficient más alto?

Fecha de entrega: Domingo 29 de Marzo a la media noche.