Nuevo método de cálculo de distancias: distancia coseno

Queremos reemplazar el método de cálculo de distancias entre especies, sustituyendo el índice de Jaccard (es el que hemos estado usando en la práctica) por la distancia coseno.

Puesto que trabajamos con un alfabeto de cuatro letras $\{\texttt{A}, \texttt{C}, \texttt{G}, \texttt{T}\}$ existen 4^k k-meros distintos y los podemos numerar del 0 al $4^k - 1$ en orden lexicográfico. Dado un gen g y su conjunto de k-meros podemos asociarle un vector $v(g) = (v_0, \ldots, v_{d-1})$ de dimensión $d = 4^k$ dónde v_i es el número de veces que el k-mero i-ésimo aparece en g. Por ejemplo si $k = 3$ entonces `AAA` es el 3-mero con índice 0 (es el menor en orden lexicográfico) y `TTT` es el 3-mero con índice 63 ($= 4^3 - 1$, es el mayor en orden lexicográfico). Si un gen contiene `AAA` seis veces y `TTT` no aparece ninguna vez entonces $v(g) = (6, \ldots, 0)$.

Una vez establecido cómo se asocia a cada gen un vector (¡no hemos representar el vector $v(g)$ usando un `vector` de C++: solo necesitamos poder **iterar** sobre la colección de los k-meros de g en orden alfabético!) la distancia entre dos genes g_1 y g_2 se definirá como

$$\delta_k(g_1, g_2) = \left(1 - \frac{1}{\pi} \arccos\left(\frac{v(g_1) \cdot v(g_2)}{\|v(g_1)\| \cdot \|v(g_2)\|}\right)\right) \times 100,$$

donde $v \cdot w$ es el producto escalar de vectores y $\|v\|$ la norma del vector v:

$$v \cdot w = \sum_{0 \le i < d} v_i \cdot w_i,$$
$$\|v\| = \sqrt{\sum_{0 \le i < d} v_i^2}.$$

Por ejemplo, suponed que $k = 2$ y que tenemos dos genes g_1 y g_2 con

$$\text{kmer}(g_1, 2) = \{\texttt{AA}, \texttt{AC}^2, \texttt{CA}, \texttt{CC}, \texttt{CG}^2, \texttt{GC}\}$$
$$\text{kmer}(g_2, 2) = \{\texttt{AC}^2, \texttt{AG}^2, \texttt{CA}^2, \texttt{GA}, \texttt{GC}\}$$

Entonces denotando

$$v = v(g_1) = (1, 2, 0, 0, 1, 1, 2, 0, 0, 1, 0, 0, 0, 0, 0, 0)$$

y

$$w = v(g_2) = (0, 2, 2, 0, 2, 0, 0, 0, 1, 1, 0, 0, 0, 0, 0, 0)$$

tenemos

$$v \cdot w = v_{AA} \cdot w_{AA} + \cdots + v_{TT} \cdot w_{TT} = v_{AC} \cdot w_{AC} + v_{CA} \cdot w_{CA} + v_{GC} \cdot w_{GC} = 2 \times 2 + 1 \times 2 + 1 \times 1 = 7.$$

Por otro lado

$$\|v\| = \sqrt{1^2 + 2^2 + 1^2 + 1^2 + 2^2 + 1^2} = \sqrt{12} = 3{,}4641\ldots$$
$$\|w\| = \sqrt{2^2 + 2^2 + 2^2 + 1^2 + 1^2} = \sqrt{14} = 3{,}7416\ldots$$

Para el cálculo haced un `#include <cmath>` y podréis usar las funciones `double sqrt(double x)` y `double acos(double x)` para la raíz cuadrada y el arco coseno, respectivamente. Si justo antes del `#include <cmath>` escribís la línea `#define _USE_MATH_DEFINES` en vuestro `.cc` entonces también podréis usar la constante `M_1_PI`, una aproximación de mucha precisión al número $1/\pi$.

Nuevo método de clusterización: método UPGMA

En el método UPGMA la construcción de un árbol filogenético se hace exactamente igual que en WPGMA pero para calcular la distancia entre un nuevo clúster $C = A \cup B$ y un clúster D se utiliza la siguiente fórmula

$$\Delta(C, D) = \frac{|A| \cdot \Delta(A, D) + |B| \cdot \Delta(B, D)}{|A| + |B|},$$

donde $|A|$ y $|B|$ son el número de especies (número de hojas) de los clústeres A y B, respectivamente. Observa que para poder implementar este nuevo método te convendrá almacenar para cada clúster el número de especies que agrupa, bien internamente en cada clúster propiamente dicho, bien en una estructura externa al clúster, p.e., agregando información adicional en el conjunto de clústers que el algoritmo va manipulando.

Las distancias $d(C)$ entre cada clúster C y sus hojas descendientes que se ha de imprimir cuando se hace un `#imprimir_cluster` o #imprimir_arbol_filogenetico (los números a la izquierda en la figura 3 del enunciado original de la práctica) se seguirán calculando de la misma forma: si $C = A \cup B$ entonces $d(C) = \Delta(A, B)/2$. Por tanto esta parte de tu práctica no tendrás que cambiarla en absoluto.

Modifica tu práctica para que en el menú principal la operación `ejecuta_paso_clust` sustituya a `ejecuta_paso_wpgma`; `ejecuta_paso_clust` tiene iguales características que la operación a la que sustituye pero aplicando el método UPGMA en vez del método WPGMA.