Práctica PRO2

Primavera 2020

1. Introducción

En biología evolutiva, antropología, lingüística y otras muchas disciplinas científicas uno de los problemas principales que se abordan es la construcción de *árboles filogenéticos*, diagramas que representan de manera esquemática las relaciones evolutivas entre un conjunto de N entidades (en biología evolutiva suele hablarse de *especies*, aunque con frecuencia no se trate de especies sino de familias u órdenes), como en los ejemplos de las figuras 1 y 2. El árbol filogenético se construye en base a las similitudes y diferencias en las características físicas o genéticas de las N entidades.

En un árbol filogenético *enraizado* cada nodo con descendientes representa al ancestro común más reciente de sus descendientes, y es usual que la longitud de las aristas/ramas del árbol sea proporcional al tiempo transcurrido entre las entidades representadas. Los nodos internos son entidades hipotéticas, ya que no pueden ser directamente observadas, sólo las N entidades de las hojas del árbol son las que conocemos.

La construcción del árbol filogenético que mejor explica los datos observados, optimizando un cierto criterio, es un problema computacionalmente difícil, en el sentido de que el coste de los cálculos necesarios crece exponencialmente con N. Además otros muchos problemas complican la obtención de árboles filogenéticos que reflejen la historia evolutiva con precisión: datos sobre las entidades inexactos, hibridaciones, evolución convergente, ...

Por esta razón se han desarrollado numerosos métodos aproximados que generan árboles filogenéticos de manera eficiente y que se aproximan muy bien al árbol óptimo en la mayoría de casos, pero no siempre. En esta práctica nuestro objetivo será construir un programa y los módulos necesarios para construir el árbol filogenético para un conjunto de N especies utilizando uno de esos métodos aproximados: el método conocido como WPGMA (*weighted pair group with arithmetic mean*).

2. El método WPGMA

En esta subsección describimos cómo funciona la construcción de un árbol filogenético mediante el método WPGMA. Partiremos de la base de que para cualesquiera dos especies (o entidades) i y j, disponemos de su *distancia* $\delta(i, j)$.

En la sección 4 damos detalles de cómo calcularemos las distancias entre especies en nuestra práctica. Pero por el momento demos por sentado que disponemos de dicha información: dado un conjunto de N especies de las cuales queremos construir su árbol filogenético, supondremos que ya tenemos calculada la distancia entre cualquier par de ellas.

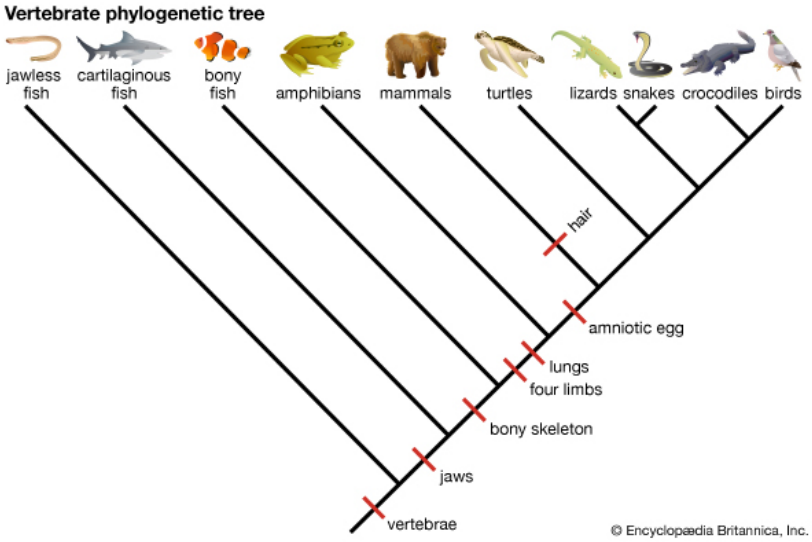

Figura 1: Un ejemplo de árbol filogenético para los vertebrados.

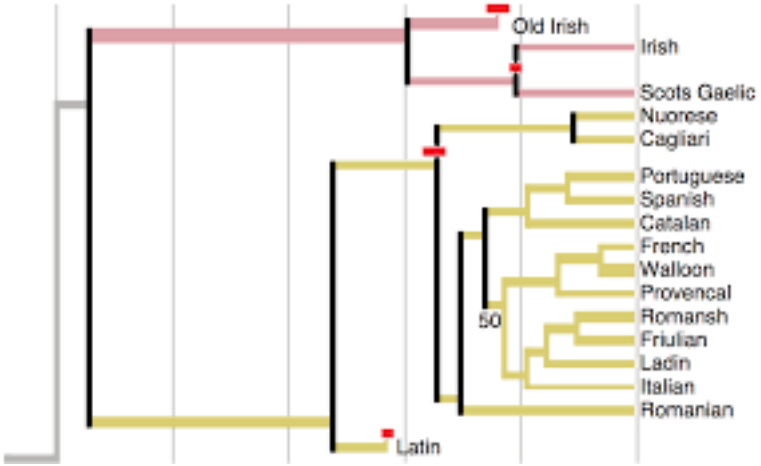

Figura 2: Un ejemplo de árbol filogenético para varios lenguajes indo-europeos.

El algoritmo WPGMA construye un árbol binario que refleja la hipotética evolución de las N especies en un proceso iterativo en el cual se van creando *clústers* o grupos de especies: cada subárbol representa un clúster, dados dos clústers los podemos unir para formar uno mayor que reúne a las especies de ambos. El nodo raíz de cada subárbol representa al ancestro común más reciente de todas las especies del clúster. La construcción del árbol se fundamenta en el principio de que todas las distancias entre la raíz de un subárbol y sus hojas son idénticas, pues la longitud/distancia de cada rama es proporcional al tiempo transcurrido y daremos por sentado que todas las N especies de las que queremos construir el árbol existen actualmente.

Por ejemplo, en la figura 1, el nodo raíz del subárbol que contiene a lagartos, serpientes, cocodrilos y aves representaría al antecesor común de todos ellos (los diápsidos). El antecesor de éste es el nodo raíz del clúster que también contiene a las tortugas (anápsidos), de manera que representa al antecesor común de todos los saurópsidos (básicamente todos los reptiles y aves, extintos o no). En la otra figura, el árbol filogenético agrupa en un clúster a todas las lenguas derivadas del latín, del cual se "desgaja" en primer lugar un grupo que conduce a lenguas habladas en Cerdeña (Nuorese, Cagliari), dialectos del sardo. Del resto del clúster se separa más tarde el rumano, y más adelante dos clústers, uno con lenguas habladas en la Península Ibérica, y el resto en Italia (italiano, friulano, ladino), Francia (francés), Suiza (romanche), ...

El algoritmo WPGMA comienza con un conjunto S de N clústers: cada clúster contiene una sola especie y le corresponde un árbol consistente en una hoja (una raíz que representa al clúster sin descendientes). La distancia entre dos clústers $A = \{a\}$ y $B = \{b\}$, cada uno de los cuales contiene una sola especie, viene dada por la distancia entre las especies correspondientes:

$$\Delta(A, B) = \delta(a, b).$$

Después, en cada iteración del WPGMA hay tres pasos principales:

1. Identificar los dos clústers A y B a menor distancia en S y formar un nuevo clúster $C = A \cup B$.

   El árbol asociado al clúster C consta de una raíz que representa al clúster C en su totalidad y subárboles izquierdo y derecho que corresponden a los clústers A y B. Más adelante veremos cómo asignar identificadores a los clústers, de manera que si el identificador de A es menor que el identificador de B entonces el subárbol izquierdo será A y el subárbol derecho será B.

   También usaremos el orden entre identificadores para resolver los empates entre distancias. Si se plantea la situación, elegiremos el par de clústers A y B que den lugar a un nuevo clúster C con identificador más pequeño.

2. Calcular la distancia entre el nuevo clúster $C = A \cup B$ y los restantes clústers en S. Para cualquier clúster D en S la distancia entre C y D es:

$$\Delta(C, D) = \frac{\Delta(A, D) + \Delta(B, D)}{2}$$

3. Añadir el clúster C a S y eliminar A y B de S.

El algoritmo concluye cuando el conjunto de clústers se haya reducido a un único elemento. Entonces habrá generado el árbol filogenético del conjunto de especies dado inicialmente.

Veamos un ejemplo. Supongamos que hay $N = 5$ entidades o especies, a, b, c, d y e. Inicialmente tendremos 5 clústers, cada uno con una sola especie. En la tablas del ejemplo usaremos, por comodidad, los identificadores (a, b, ...), para los clústers formados por una sola especie, en vez de escribir $\{a\}$, $\{b\}$, ... que sería lo más preciso. Como ya hemos dicho la distancia entre dos clústers A y B, cada uno de los cuales contiene una sola especie es la distancia entre las dos especies.

Supongamos que tenemos la siguiente matriz de distancias:

	a	b	c	d	e
a	0	17	21	31	23
b	17	0	30	34	21
c	21	30	0	28	39
d	31	34	28	0	43
e	23	21	39	43	0

En la primera iteración los clústers $\{a\}$ y $\{b\}$ forman el par con distancia mínima y se unen en un nuevo clúster $\{a, b\}$; la tabla de distancias queda actualizada así:

	$\{a, b\}$	c	d	e
$\{a, b\}$	0	25.5	32.5	22
c	25.5	0	28	39
d	32.5	28	0	43
e	22	39	43	0

En la segunda iteración los clústers a unir son el $\{a, b\}$ y $\{e\}$, dando lugar a un nuevo clúster $\{\{a, b\}, e\}$. La matriz de distancias actualizada es

	$\{\{a, b\}, e\}$	c	d
$\{\{a, b\}, e\}$	0	32.25	37.75
c	32.25	0	28
d	37.75	28	0

En la tercera iteración se agrupan los clústers $\{c\}$ y $\{d\}$ en uno nuevo $\{c, d\}$; tras la actualización las distancias resultantes son:

	$\{\{a, b\}, e\}$	$\{c, d\}$
$\{\{a, b\}, e\}$	0	35
$\{c, d\}$	35	0

En el paso final se unen los dos últimos clústers remanentes $\{\{a, b\}, e\}$ y $\{c, d\}$. en un clúster final $\{\{\{a, b\}, e\}, \{c, d\}\}$.

La figura 3 muestra el árbol filogenético construído por el algoritmo WPGMA. En dicha figura se han tenido en cuenta los convenios sobre identificadores de clústers que se describen en la siguiente sección.

3. Clústers

De la descripción anterior se desprende que un clúster o bien contiene un único elemento o *especie*, o bien está formado por un par de clústers: en definitiva, un clúster es asimilable a

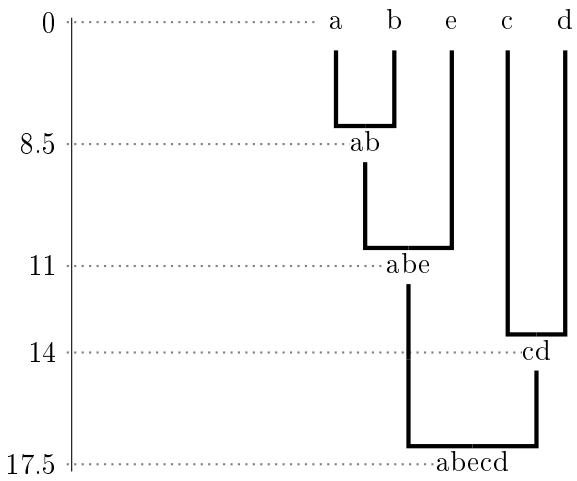

Figura 3: Árbol filogenético construído por el algoritmo WPGMA.

un árbol binario, en donde tenemos especies en las hojas. Nuestro sistema deberá ser capaz de guardar y actualizar dinámicamente una "tabla de distancias" entre clústers. Las "filas" y "columnas" de dicha tabla vienen indexadas por clústers, y se ha de poder añadir y eliminar eficientemente la información de la distancia entre un par de clústers dados.

Por otro lado se ha de poder dotar de contenido inicial a una de tales "tablas de distancias" a partir de un conjunto de N especies, considerando cada una de las N especies como consituyentes de los correspondientes N clústers, cada uno de ellos con un único elemento.

Aunque el algoritmo WPGMA de construcción del árbol filogenético no prescribe quién es el "hijo izquierdo" y quién el "hijo derecho" (es irrelevante) en un clúster, en esta práctica adoptaremos los siguientes convenios:

1. Un clúster con una única especie tiene como identificador el mismo que el de la especie.
2. Dados dos clústers con identificadores α y β tales que $\alpha < \beta$ en orden lexicográfico, su combinación es un clúster que tiene como hijo izquierdo al clúster identificado por α, como hijo derecho al clúster identificado por β y su identificador es $\gamma = \alpha \cdot \beta$, la concatenación de los identificadores de sus hijos.

En los ejemplos de la sección anterior no hemos empleado los identificadores de los clústers para faciltar la explicación[1], pero veamos ahora, con el ejemplo que hemos desarrollado previamente, cómo se asignan identificadores a los clústers a medida que avanza el algoritmo WPGMA siguiendo los convenios recién descritos. Inicialmente tenemos los clústers $\{a\}$, ..., $\{e\}$ con identificadores `"a"`, `"b"`, ..., `"e"`. Después se combinan los clústers con identificadores `"a"` y `"b"` para formar el clúster $\{a, b\}$ cuyo identificador es `"ab"`. En la segunda iteración combinamos los clústers $\{a, b\}$ y $\{e\}$, y al clúster resultante $\{\{a, b\}, e\}$ se le asigna el identificador `"abe"`.

Nótese que al finalizar el proceso solo quedará un clúster que contiene a todas las especies y su identificador será una concatenación de los identificadores de todas las especies, independendientemente de cúal es el conjunto de especies de partida, pero el orden será naturalmente

[1] En la figura 3 sí se han reflejado los convenios sobre los identificadores de los clústers que acabamos de describir en esta sección.

dependiente de cúales son las distancias entre ellas y cúal es la estructura arborescente resultante. En nuestro ejemplo el identificador del clúster final (= árbol filogenético) es `"abecd"`. Fijaos por otro lado que el identificador no nos permite reconstruir de manera única la estructura arborescente.

4. Especies y genes

En las secciones precedentes no hemos entrado en detalle de qué es exactamente una entidad o especie y hemos dado por sentado que, dadas dos especies i y j podemos calcular la distancia entre ambas $\delta(i, j)$.

Tanto la definición de lo que es una entidad como el cálculo de distancias variarán sustancialmente en función de la aplicación que estemos considerando—mientras que el proceso de construcción WPGMA se fundamenta en la abstracción de especie y distancia, y es completamente independiente de lo que realmente sean las especies y de cómo se calculan las distancias. Vuestra práctica, y en concreto vuestro diseño modular, debe reflejar esta circunstancia.

En nuestra práctica simplificaremos en extremo lo que habitualmente nos encontramos en la aplicaciones con relación a la información disponible de cada especie y del cálculo de la distancia entre dos especies. Habitualmente para cada especie tendríamos un conjunto amplio de rasgos o de genes secuenciados, pero nosotros consideraremos que cada especie consiste en un identificador (una *string*) y un *gen* (otra string).

Cada gen consiste en una (larga) secuencia de símbolos, tomados de entre los únicos cuatro posibles: A, C, T o G[2]. Así por ejemplo un gen podría ser `AACTTGCGAGCTACAACCTGGGATTA`.

Solo falta, pues, determinar cómo calcular la distancia entre dos especies. Existen muy diversas nociones de distancia entre dos secuencias de símbolos (es la manera en que representamos los genes en esta práctica), pero muchas de ellas requieren un cómputo muy costoso ya que las secuencias implicadas pueden ser muy largas (p.e., el gen DMD contiene 2.4 millones de nucleótidos —pares de bases en la doble hélice de DNA— de manera que la secuencia tiene longitud igual a 2.4 millones de símbolos!).

En esta práctica vamos a utilizar un procedimiento más simple para estimar la distancia entre dos secuencias g_1 y g_2. Dada una secuencia cualquiera g y un valor $k \geq 1$ diremos que $\text{kmer}(g, k)$ es el conjunto de las subsecuencias contiguas de g de longitud k (los k-meros de g), es decir, si n es la longitud de g entonces

$$\text{kmer}(g, k) = \{g[i..i + k - 1] \,|\, 0 \leq i < n + 1 - k\},$$

donde $g[i..j]$ denota la subsecuencia de g entre los símbolos i-ésimo y j-ésimo de g, que asumimos se indexan de 0 a $n - 1$. Por ejemplo, si $g = \texttt{ACATTATCATGC}$ y $k = 3$ entonces

$$\text{kmer}(g, 3) = \{\texttt{ACA}, \texttt{ATC}, \texttt{ATG}, \texttt{ATT}, \texttt{CAT}^2, \texttt{TAT}, \texttt{TCA}, \texttt{TGC}, \texttt{TTA}\},$$

donde hemos escrito los 3-meros de g en orden lexicográfico creciente, y los superíndices indican la multiplicidad—por ejemplo, el 3-mero `CAT` aparece dos veces en g. La distancia entre g_1 y g_2 se define entonces como

$$\delta_k(g_1, g_2) = \left(1 - \frac{\#(\text{kmer}(g_1, k) \cap \text{kmer}(g_2, k))}{\#(\text{kmer}(g_1, k) \cup \text{kmer}(g_2, k))}\right) \times 100,$$

[2]Un gen es un segmento en una secuencia de núcleotidos en una cadena de ADN. En una cadena de ADN solo existen cuatro tipos de nucleótidos: adenina (A), citosina (C), guanina (G) y tiamina (T).

donde $\#A$ denota la cardinalidad del multiconjunto A. Recordemos que si x aparece i veces en A ($x^i \in A$) y j veces en B ($x^j \in B$) entonces $A \cap B$ contiene mín(i, j) apariciones de x, y que $A \cup B$ contiene máx(i, j) apariciones de x.

El valor de k es un parámetro que se fijará globalmente al principio de la ejecución del programa y servirá para el cálculo de todas las distancias entre pares de especies. Es apropiado pensar que es un valor asociado a la tabla de distancias de un conjunto de especies y consecuentemente es un valor conocido incluso antes de fijar el o los conjuntos de especies de los que pretendemos construir su árbol filogenético.

Aunque vuestra práctica debe funcionar correctamente con cualquier valor de $k > 0$, los mejores resultados en situaciones reales se obtienen con valores de k entre 5 y 15 (nosotros vamos a usar genes "cortos" en los juegos de pruebas, de manera que un valor de k en el rango bajo será más apropiado). El conjunto kmer(g, k) ~~tiene como mucho $|g|$~~ $k + 1$ elementos distintos y por lo tanto un valor de k muy alto seria poco útil pues habría muy pocos elementos y todos ellos tendrían multiplicidad muy baja, por regla general. En el extremo opuesto un valor muy bajo de k tampoco sería útil, pues en kmer(g, k) acabaríamos teniendo todos los k-meros posibles (hay 4^k posibles) y la estimación de la distancia entre dos especies sería entonces muy burda.

5. Funcionalidades

5.1. Decisiones sobre los datos

1. Todos los identificadores son strings constituídos exclusivamente por letras minúsculas y mayúsculas (de la A a la Z, sin caracteres especiales, sin vocales acentuadas, ...), dígitos y por el carácter de subrayado '_'.

2. Ningún identificador de especie es un prefijo propio de ningún otro identificador de especie. Ello garantiza que los identificadores de clústers que se crean concatenando identificadores reflejan cúales son las especies que forman parte del clúster (no su estructura de árbol). También garantiza la validez del desempate entre clústers que hemos definido.

3. No habrá especies que compartan un mismo gen. Sí que podrá haber especies a distancia 0 (es decir, con el mismo conjunto de k-meros), pero de cara al algoritmo WPGMA no representan un caso especial.

4. Todas las distancias, entre especies o entre clústers, son números reales (`double`) que se imprimen con cuatro dígitos decimales de precisión.

5. La entrada es sintácticamente correcta; p.e., si una operación dice que tiene dos argumentos que son identificadores, la entrada contendrá exactamente eso, dos strings válidos como identificadores.

5.2. Comandos del programa principal

Vuestro programa principal comenzará leyendo el valor k que se usará globalmente para el cálculo de todas las distancias entre especies. Se empieza con un conjunto de especies vacío y un conjunto de clústers también vacío. En lo que resta nos referiremos a ellos como **el** conjunto de especies y **el** conjunto de clústers.

Una vez creados los conjuntos de especies y de clústers se entra en un bucle en el que procesa una opción en cada iteración. Las opciones son las siguientes:

1. `crea_especie`: Crea una especie con el identificador y gen (dos strings) dados. Escribe un mensaje de error si ya existe una especie con el mismo identificador. La especie creada, si no hay error, se agrega al conjunto de especies.

2. `obtener_gen`: Dado un identificador de especie, imprime el gen asociado a la especie. Escribe un mensaje de error si no existe una especie con el identificador dado.

3. `distancia`: Dados dos identificadores de especies, imprime la distancia[3] entre las dos especies. Se escribe un mensaje de error si alguna de las dos especies cuyos identificadores se dan no existen.

4. `elimina_especie`: Dado el identificador de una especie e la elimina del conjunto de especies. Escribe un mensaje de error si la especie con el identificador dado no existe.

5. `existe_especie`: Dado el identificador de una especie e imprime una indicación de si dicha especie existe (es decir, es parte del conjunto de especies).

6. `lee_cjt_especies`: Lee del canal estándar de entrada un entero $n \geq 0$ y a continuación una secuencia de n especies (pares identificador-gen). Las n especies dadas tienen identificadores distintos entre sí. Los contenidos previos del conjunto de especies se descartan —las especies dejan de existir— y las n especies leídas se agregan al conjunto de especies.

7. `imprime_cjt_especies`: Imprime en el canal estándar de salida el conjunto de especies.

8. `tabla_distancias`: Imprime la tabla de distancias entre cada par de especies del conjunto de especies.

9. `inicializa_clusters`: Inicializa el conjunto de clústers con el conjunto de especies en el estado en el que esté en ese momento, e imprime los clústers resultantes, así como la tabla de distancias entre clústers. Al imprimir la tabla de distancias se usarán los *identificadores* de los clústers para indexar filas y columnas.

10. `ejecuta_paso_wpgma`: ejecuta un paso del algoritmo WPGMA (fusiona los dos clústers a menor distancia en uno nuevo) e imprime la tabla de distancias entre clústers resultante. Al imprimir la tabla de distancias se usarán los *identificadores* de los clústers para indexar filas y columnas.

11. `imprime_cluster`: dado un identificador α, imprime el clúster (su "estructura arborescente") con el identificador dado, o un error si no existe un clúster con dicho identificador en el conjunto de clústers.

12. `imprime_arbol_filogenetico`: imprime el árbol filogenético para el conjunto de especies actual; dicho árbol es el clúster que agrupa todas las especies, resultante de aplicar el algoritmo WPGMA. El contenido del conjunto de clústers previo se descarta y se reinicializa con el conjunto de especies en el estado en el que esté en ese momento, para

[3] Todas las distancias, sean entre especies o entre clústers, se imprimirán con cuatro dígitos decimales de precisión.

a continuación aplicar el algoritmo. El conjunto de clústers final es el que queda después de aplicar el algoritmo.

Se imprimirá la estructura arborescente del clúster con los identificadores de los clústers (raíces de los subárboles) y la distancia entre cada clúster y sus hojas descendientes (véase la figura 3; dichas distancias son los números a la izquierda, y se pueden calcular fácilmente a partir de la distancia entre los clústers cuya combinación da origen a cada clúster). El formato preciso en el que se ha de imprimir el árbol se mostrará en los juegos de pruebas públicos.

13. `fin`: finaliza la ejecución del programa.

Importante: el conjunto de clústers solamente registra las altas y bajas del conjunto de especies cuando se reinicializa mediante `inicializa_clusters` o `imprime_arbol_filogenetico`. Si, tras una aplicación de cualquiera de estas dos operaciones, el conjunto de especies sufre altas o bajas, estas no afectan al conjunto de clústers hasta la siguiente reinicialización.