Similitud de genes

La coordinación de álgebra siguió buscando a las personas que ayudaron a los cachorros durante meses, sin tener mayores resultados. Cuando ya habían perdido sus esperanzas, uno de sus informantes llegó con una pista crucial. Mientras seguía a uno de los cachorros, lo vió juntarse con un informático en el DIINF, donde el sospechoso le entregó un pendrive. Audazmente el informante logró obtener un cabello del sospechoso, entregándolo a la coordinación.

Utilizando herramientas de investigación al más puro estilo *CSI Miami,* lograron identificar los genes de cada informático y sus respectivos procesos biológicos en donde se puede hallar, solo les faltaba encontrar el grupo de individuos que fueran amigos del único sospechoso hasta ahora. Con la ayuda de un científico, dedujeron que mientras menor sea la distancia entre dos genes, mayor será la relación entre los individuos, por lo que de esa manera pueden encontrar a todos los culpables.

En el cuarto laboratorio del curso se le pide elaborar un programa que pueda determinar qué genes tienen comportamientos similares al estar involucrados en procesos biológicos. La estructura con la que se deberá trabajar será la de un grafo acíclico dirigido, en la cual existe un nodo raíz y cada descendiente puede tener uno o más ancestros. Cada vértice del grafo representa un proceso biológico, y cada proceso biológico podrá tener 0 o más genes asociados (Figura 1).

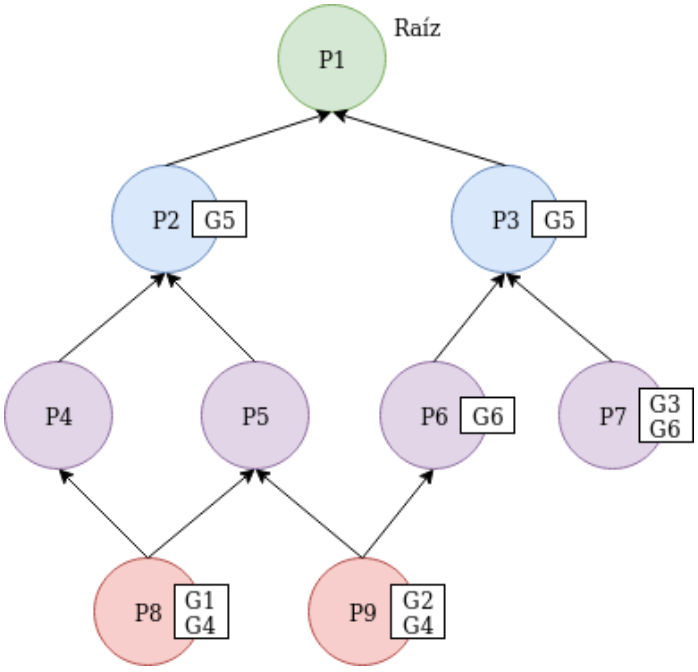

Figura 1: Grafo acíclico dirigido

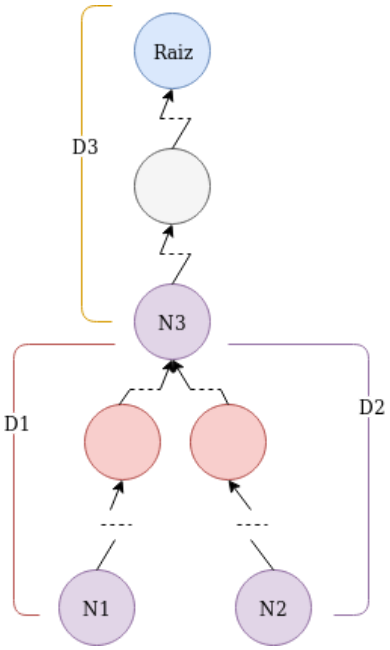

Figura 2: Distancia al ancestro común y distancia a la raíz.

Los profesores de álgebra solicitan ocupar estas medidas de similitud dado que les permitirán de mejor manera identificar a los alumnos involucrados:

Similitud de Wu-Palmer

$$Sim_{wp}(N_A, N_B) = \frac{2 * D3}{D1 + D2 + (2 * D3)}$$

Donde:

- D1: distancia de N1 a N3.
- D2: distancia de N2 a N3.
- D3: distancia de N3 a la raíz.

Por ejemplo para el proceso 9 y el proceso 6 de la figura 1, la similitud sería:

$$Sim_{wp}(P9, P6) = \frac{2 * 2}{1 + 0 + 2 * 2} = 0.800000$$

Similitud de Leacock-Chodorow

$$Sim_{lc}(N_A, N_B) = -\log\left(\frac{D1 + D2 + 1}{2 * D}\right)$$

Donde D es la profundidad del grafo, por ejemplo en la figura 1, D = 4), D1 y D2 son las distancias mostradas en la figura 2.

Por ejemplo para el proceso 9 y el proceso 6, la similitud sería:

$$Sim_{lc}(P9, P7) = -log(\frac{2+1+1}{2*4}) = 0.301030$$

Similitud entre dos genes

La similitud entre dos genes x, y corresponde al promedio de las similitudes entre los pares de procesos biológicos en los que se encuentran involucrados los genes.

$$Simg(g_x, g_y) = \frac{\sum_{i \epsilon x} \sum_{j \epsilon y} Sim(i,j)}{|P_x|\,|P_y|}$$

Figura 3: Similitud entre dos genes

Dónde Sim(i,j) corresponde a la similitud de Wu-Palmer o de Leacock-Chodorow según corresponda entre los procesos involucrados de los genes y $|P_x||P_y|$ corresponde al total de pares de procesos involucrados.

Por ejemplo la similitud de Wu-Palmer entre los genes 4 y 6 de la figura 1 sería la siguiente:

$$Simg_{wp}(G4, G6) = \frac{Sim_{wp}(P8, P6) + Sim_{wp}(P8, P7) + Sim_{wp}(P9, P6) + Sim_{wp}(P9, P7)}{4}$$

$$Simg_{wp}(G4, G6) = \frac{0.000000 + 0.000000 + 0.800000 + 0.400000}{4}$$

$$Simg_{wp}(G4, G6) = 0.300000$$

Y la similitud de Leacock-Chodorow de los genes 4 y 6 sería:

$$Simg_{lc}(G4, G6) = \frac{Sim_{lc}(P8, P6) + Sim_{lc}(P8, P7) + Sim_{lc}(P9, P6) + Sim_{lc}(P9, P7)}{4}$$

$$Simg_{lc}(G4, G6) = \frac{0.124939 + 0.124939 + 0.602060 + 0.301030}{4}$$

$$Simg_{lc}(G4, G6) = 0.288242$$

Entradas

El grafo será entregado en dos archivos, el primero se llama "procesos.in" y contiene los vértices del grafo (los cuales representan los procesos biológicos) de la siguiente forma:

- En la primera línea se indica la cantidad de vértices del grafo
- Luego en las siguientes líneas, el primer elemento corresponde a un vértice del grafo y los siguientes elementos de la línea son los padres de este.

```
9
P1
P2 P1
P3 P1
P4 P2
P5 P2
P6 P3
P7 P3
P8 P4 P5
P9 P5 P6
```

Figura 4: Ejemplo de archivo de entrada 1 para el grafo de la figura 1

El segundo archivo se llama "genes.in" e indica en qué procesos se encuentran involucrados los genes de la siguiente forma:

- En la primera línea se indica la cantidad de genes
- Las siguientes líneas indican el gen y luego los procesos en los que aparece.

```
6
G1 P8
G2 P9
G3 P7
G4 P8 P9
G5 P2 P3
G6 P6 P7
```

Figura 5: Ejemplo de archivo de entrada 2 para el grafo de la figura 1

Luego por consola se deben ingresar dos genes y el programa debe calcular las medidas entregadas previamente, tal como se muestra figura 6.

Salida

Se debe entregar por pantalla los valores de las similitudes de genes descritas anteriormente (Figura 3), como se muestra en la figura 6.

```
franco@franco:~/Escritorio$ ./lab4
Ingrese primer gen: 4
Ingrese segundo gen: 6
Similitud de genes Wu-Palmer: 0.300000
Similitud de genes Leacok-Chorodow: 0.288242
¿Desea ingresar otro par de genes?: No
```

Figura 6: Ejemplo de ejecución.

BONUS (opcional):

La semana del 24 de junio se podrá mostrar en horario de clases un avance de código en que éste cumpla con la funcionalidad de poder leer los archivos de entrada y una de las similitudes entre procesos. Se dará un punto base para quienes presenten el avance.

Observaciones:

- Recuerde realizar todas las validaciones correspondientes a las entradas.

Laboratorio 4
Todas las coordinaciones

FACULTAD DE INGENIERÍA
Departamento de Ingeniería Informática
Análisis de algoritmos y estructuras de datos

UdeSantiago de Chile

Fecha de entrega: **04/07/2019** hasta las **23:30 hrs**.

Instrucciones de entrega:

- Archivo **PDF** con el **informe** (incluye manual de usuario).
- Se debe calcular el **T(n)** y el **O**.
- Código fuente en archivos **.c y .h** (no entregar proyectos de ninguna IDE).
- El código debe permitir ser **compilado en** ambiente **Windows y Linux** por lo que se sugiere usar ANSI C.