Universidad de Concepción
Departamento de Ing. Informática y Cs. de la Computación

Proyecto 1: Tópicos en Manejo de Grandes Vols. de Datos
Cecilia Hernández

Sketches para estimación de similitud entre genomas

El proyecto lo pueden realizar en grupos de 2 estudiantes.
Fecha de entrega: viernes 27 de Septiembre, 11:59hrs.

En esta tarea se implementará y analizará algoritmos de streaming y sketches para la estimación de similitud entre genomas usando sketches de cardinalidad.

En particular se pide que implemente los algoritmos en C/C++ usando alguna función hash disponible en smhasher. El software usa cmake para compilar, de manera que si no lo tiene debe instalarlo antes de compilar. Para bajar el código y compilar en linux proceder de la siguiente manera:

```
$ git clone https://github.com/aappleby/smhasher.git
$ cd smhasher
$ cd src
$ cmake .
$ make
```

Para el desarrollo del proyecto se pide lo siguiente:

1. Implementar los Hyperloglog. (1.5 puntos)

   Para la implementación puede usar la función de gcc *__builtin_clz*, que cuenta el número de leading ceros. Esto le sirve para extraer del valor de hash el valor que debe ingresar en el sketch. A tal valor le debe sumar 1 para determinar la posición del primer 1 (ver algoritmo en clases para mas detalle).

2. Implementar la similitud de Jaccard, es decir $J(A,B) = \frac{|A \cap B|}{|A \cup B|}$ usando Hyperloglog. En este parte debe implementar dos alternativas: (2.5 puntos)

   Su labor es estimar Jaccard usando el principio de inclusión/exclusión. Luego, se puede definir el coeficiente o similitud de Jaccard entre dos conjuntos como sigue:

$$J(A,B) = \frac{|A| + |B| - |A \cup B|}{|A \cup B|}$$

Para ambas implementaciones considere los siguientes valores de $k = 20$, $k = 25$ y $w = 50$, $w = 100$. Además considere un error estándar de estimación para Hyperloglog de aproximadamente de 0.01 (1 %), o 0.02.

a) Alternativa 1: Usando k-mers. Esta alternativa consiste en representar cada genoma como un multiconjunto de k-mers y estimar la cardinalidad del conjunto de k-mers. Los k-mers son substrings de largo k consecutivos en una secuencia de nucleótidos de ADN.

b) Alternativa 2: Usando minimizers. Esta alternativa es igual a la alternativa anterior donde en lugar de usar k-mers debe usar minimizers. Un minimizer es un k-mer seleccionado en una región de un genoma y se computan usando un esquema de ventana deslizante. Luego, se computan en base a dos parámetros:

- k, largo de k-mer.
- w, largo de la ventana deslizante. Esta ventana contiene un número de k-mers. Luego, $w > k$, donde el número de k-mers en la ventana esta dado por $w - k + 1$.

Considere la siguiente manera para computar minimizers:

- Definir variable i, para indicar la ventana actual de procesamiento, donde el largo de la ventana está dado por w. La ventana actual indica la región actual del genoma en procesamiento. Al inicio $i = 1$ y luego las posiciones del genoma en la primera región están en el rango $[1, w]$.
- Construir todos los k-mers dentro de la ventana.
- Definir como minimizer en la ventana al k-mer menor lexicográficamente.
- Desplazar la ventana en 1. Para ello debe incrementar la variable i, y deslizar ventana en 1, lo que define rango de posiciones en el genoma de $[2, w + 1]$.
- Procesamiento continua procesando ventanas hasta completar de procesar el genoma.

Ejemplo de cómputo de minimizers usando un $k = 3$ y $w = 6$.

```
G = ACGTGACCG
Primera ventana: ACGTGA
k-mers: ACG, CGT, GTG, TGA
Minimizer: ACG

Segunda ventana: CGTGAC
k-mers: CGT, GTG, TGA, GAC
Minimizer: GAC

Tercera ventana: GTGACC
k-mers: GTG, TGA, GAC, ACC
Minimizer: ACC

Cuarta ventana: TGACCG
k-mers: TGA, GAC, ACC, CCG
Minimizer: ACC
```

3. Debe procesar al menos 5 genomas de los que están disponibles aquí.

4. Realizar una evaluación experimental usando las medidas de error definidas por Error Relativo Medio (ERM) y Error Absoluto Medio (EAM) como sigue : (2 puntos)

$$ERM = \frac{1}{n}\sum_{i=1}^{n}\frac{|\hat{J}_i - J_i|}{J_i}$$

$$EAM = \frac{1}{n}\sum_{i=1}^{n}|\hat{J}_i - J_i|$$

   donde $\hat{J}_i$ corresponde al Jaccard estimado entre un par de genomas y J_i es el Jaccard real.

5. El desarollo del proyecto debe incluir un informe con lo realizado. La calificación de su proyecto incluirá el funcionamiento, calidad de código e informe. (0.5 puntos)