Proyecto 2 Sistemas Operativos

Cecilia Hernández

April 19, 2016

Fecha inicio: Martes, 19 de Abril, 2016.
Fecha entrega: Lunes, 3 de Mayo, 2016.

1. **Descripción del problema** El ácido desoxirribonucleico, comúnmente conocido como ADN, contiene instrucciones genéticas usadas en el desarrollo y funcionamiento de todos los organismos vivos y es responsable de su transmisión hereditaria. El ADN contiene la información necesaria para construir otros componentes de las células, como las proteínas y las moléculas de ARN. El ADN consiste de una cadena de compuestos químicos llamados nucleótidos o bases que son Adenina (A), Timina (T), Citocina (C) y Guanina (G). La disposición secuencial de estos cuatro nucleótidos a la largo de la cadena es la que codifica la información genética, y los segmentos de ADN que llevan información genética se llaman genes.

   Para resolver este problema le entregaremos algunos ejemplos de archivos de entrada que podrá procesar usando su aplicación desde línea de comando. El formato del archivo de entrada contiene 4 líneas de texto por secuencia de ADN asociada a una lectura. Cada una de estas líneas tiene el siguiente aspecto:

```
@SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
TTGCCTGCCTATCATTTTAGTGCCTGTGAGGTGGAGATGTGAGGATCAGT
+SRR566546.970 HWUSI-EAS1673_11067_FC7070M:4:1:2299:1109 length=50
hhhhhhhhhhghhghhhhhfhhhhhfffffe‘ee[‘X]b[d[ed‘[Y[^Y
```

   - Tipo 1. La primera línea empieza con un “@” seguido con el nombre de la lectura.
   - Tipo 2. La segunda línea corresponde a la secuencia de nucleótidos. Debe verificar la correctitud de cada nucleótico, porque pueden haber distintos a A,T,G, y C.
   - Tipo 3. La tercera línea empieza on un “+” normalmente contiene lo mismo que la primera línea, pero también puede sólo contener el “+”.
   - Tipo 4. La cuarta línea contiene información de calidad de la lectura, de manera que tiene el mismo largo que la línea de secuencia. Esta línea reporta para cada nucleótido un caracter que indica la confiabilidad de la lectura del nucleótido. Los valores de caracteres (correlativos en ASCII) asociados a la calidad son los siguientes, donde la calidad aumenta de izquierda a derecha:

donde "!" corresponde a la menor calidad y "~" corresponde a la mayor calidad. La codificación de calidad Q está dada por la fórmula $Q = ASCII - 33$, donde 33 corresponde al caracter ASCII de "!", de manera que $0 \leq Q \leq 93$ para cualquier caracter. Esta medida de calidad está asociada a la probabilidad de error en la lectura mediante la siguiente fórmula $Q = -10 \log_{10} P_e$, donde P_e es la probabilidad de error.

Su trabajo consiste en escribir un programa en C en paralelo usando pthreads o Java Threads que permita ingresar por teclado el largo máximo de la secuencia (línea 2) y la probabilidad de error (P_e), junto con un conjunto de lecturas y realice las operaciones que se dan a continuación. Debe considerar que P_e es la probabilidad de error máxima para considerar un nucleótido como válido. Su programa deberá imprimir las secuencias válidas y confiables en el archivo de salida Salida_Secuencias.txt y la siguiente informaci´n en el archivo de Salida_Stats.txt.

(a) El P_e ingresado

(b) El número total de nucleótidos en las secuencias de ADN de todas las lecturas.

(c) El número total de nucleótidos confiables (dado el P_e y nucleótidos válidos) en todas las secuencias de ADN.

(d) El número total de todos los errores de todas las lecturas en la secuencia de ADN (caracteres que no coinciden con A, C, G o T)

(e) El número total de todos los nucleótidos válidos, pero erróneos determinado por P_e

(f) El contenido de Guanina y Citocina en datos de ADN (GC content). Existen varias razones por las cuales los biólogos están interesados en el contenido de GC, tales como:

- Contenido de GC puede identificar tipos de genes en la secuencia de ADN. Genes tienden a tener mas alto contenido de GC que otras secciones de ADN, así como largas regiones de codificación tienen mayor contenido de GC.
- Regiones de ADN con alto contenido de GC requieren mayores temperaturas para ciertas reacciones químicas como duplicación de ADN.
- El contenido de GC tambin puede ser usado para determinar la clasificación de especies.

El contenido de GC se define como el porcentaje $100 \times \frac{G+C}{A+T+G+C}$

(g) El contenido de AT, el cual se define como $100 \times \frac{A+T}{A+T+G+C}$

(h) La razón AT/GC el cual se define como $\frac{A+T}{G+C}$

(i) Clasificación segn GC. El contenido de GC puede ser utilizado para clasificar organismos. Para ello considere la siguiente clasificación:

- Si el contenido de GC está sobre el 60%, se considera que el organismo tiene un alto contenido de GC
- Si el contenido de GC está bajo el 40%, se considera que el organismo tiene un bajo contenido de GC
- Si no se cumple lo anterior se considera que el organismo tiene un contenido moderado

Algunos ejemplos de wikipedia (GC content *https://en.wikipedia.org/wiki/GC-content*) se pueden clasificar algunos organismos como los siguientes:

- En Streptomyces coelicolor A3(2), el contenido GC es 72%
- El contenido GC de Yeast (Saccharomyces cerevisiae) es 38%
- En Thale Cress (Arabidopsis thaliana), el contenido GC es 36%
- En Plasmodium falciparum, el contenido GC es aproximadamente 20%

2. Requerimientos

Su labor es resolver este problema usando la ejecución de hebras en paralelo y sincronización. Para resolver este problema debe:

- Paralelizar el proceso de validación de secuencias, es decir extraer nucleótidos inválidos y nucleótidos con confiabilidad en la lectura superior al error Pe. La paralelización debe realizarla mediante un número de hebras (N) ingresada como parámetro de entrada al programa.
- Usar un buffer intermedio de tamaño limitado (M registros) que le permita ir almacenando los registros que contengan las secuencias válidas generadas por las hebras. Cada vez que este buffer se llene todas las hebras deben esperar a que la hebra que escribió ese registro vacie el buffer en un archivo de salida. Una vez hecho esto el buffer queda disponible a que todas las hebras puedan seguir leyendo su partición en el archivo de entrada para continuar con su procesamiento. M debe ser ingresado como parámetro a su aplicación.

3. Ejemplos

A continuación se proporciona un ejemplo de ejecución.

```
Contenido A0.fastq

@HWI-ST141_0363:2:1101:1175:2080#ATCACG/1
NCCGTATCC
+HWI-ST141_0363:2:1101:1175:2080#ATCACG/1
!!!!Z~~~~
@HWI-ST141_0363:2:1101:1141:2203#ATCACG/1
GTTATTCTT
+HWI-ST141_0363:2:1101:1141:2203#ATCACG/1
!!!!!!!~~

Ejecucion.

proyecto2 A0.fastq 8 2 1 Salida_Secuencias.txt Salida_Stats.txt
(proyecto2 archivo.fastq Num_lineas_archivo.fastq N M Salida_Secuencias.txt Salida_Stats.txt)
```

```
El contenido del archivo Salida_Secuencias.txt es:

TATCC
TT

El contenido del archivo Salida_Stats.txt es:

Pe = 0.001000
largo sec total = 18
largo sec con filtro Pe = 7
nn = 1
ne = 10
Contenido GC = 28.57 %
Contenido AT = 71.43 %
Razon AT GC = 2.50
GC bajo
```

4. Datos

   Los archivos de entrada están en formato fastq (ver wikipidia para mayor información). Fastq es un archivo de texto (en ASCII) que almacena la secuencia biológica con su respectiva calidad de lectura. Actualmente es el formato por defecto para almacenar la salida de instrumentos de secuenciación (analizadores de genomas) de alto rendimiento.

   Los archivos de prueba estarán disponibles en infoda. Para el desarrollo de su proyecto, es recomendable usar archivos de prueba cortos y ficticios, que ustedes manipulen para verificar su programa antes de ejecutar con los archivos de prueba mas largos que les daremos.

5. Evaluación

   La evaluación del poyecto será en 3 partes cuyos porcentajes estan estipulados a continuación:

   - Correctitud de la solución: cumpliendo todas las exigencias pedidas. Se probarán varios casos de entrada los cuales determinarán en parte su nota. Tambin se analizará el código. (35 %)
   - Informe escrito: el proyecto debe estar documentado en un informe de no más de 4 páginas, en las cuales expresen en palabras, pseudocódigo o diagrama de flujo lo que hace el algoritmo que implementaron en su solución.(30 %)
   - Interrogación grupal: habrá también una interrogación grupal. El objetivo de esta interrogación es que quede clara la autoría de las soluciones expuestas. Cada integrante del grupo debe conocer el funcionamiento de las soluciones expuestas y de los programas.(35 %)

6. **Términos y condiciones**

   Si se descubren prácticas deshonestas de cualquier tipo, el grupo y todos los involucrados, serán calificados con la nota mínima.