Tarea 3 - Alineamiento de lecturas

Algoritmos en Biología Computacional

BCOM 4006

Laura Valentina Acosta Corredor 201911225

Juanita Puentes Mozo 201814823

Jhon Stewar Rayo Mosquera 201720166

Septiembre 26, 2023

Universidad de los Andes

September 29, 2023

Problem 1

Alineamiento de lecturas en el cluster.

Solution.

El objetivo de este punto del taller fue alinear lecturas empleando los comandos y modulos del cluster. En primer lugar, se alinearon las lecturas al genoma de referencia haciendo uso de la herramienta bowtie2. Luego, se construyo un índice FM para el genoma de referencia utilizando el comando bowtie2-build. La operación principal de bowtie2 es el alineamiento de lecturas cortas al genoma de referencia. Para este punto llevamos a cabo el alineamiento de cuatro muestras diferentes: *Seg5*, *PoolEthProd*, *PoolEthTol* y *Unselected*. Para cada una de estas muestras se presentan los resultados en la siguiente ruta:

/hpcfs/home/cursos/bcom4006/estudiantes/estudiante20/Tarea3/

A partir de cada uno de estos alineamientos se obtuvieron estadisticas del alineamiento como el número de fragmentos que aparecen en los archivos fastq. Tambien se incluyen la cantidad que fueron alineados como pares consistentes, tanto de forma única como múltiples veces, la cantidad que fueron alineados como pares de manera inconsistente con el experimento (muy cerca, muy lejos o con orientación incorrecta) y de los que no fueron alineados como pares, las lecturas que se pudieron alinear como lecturas sencillas. La Tabla **??** resume las estadisticas de alineamiento para todas las muestras.

	Seg5	***PoolEthTol***	***PoolEthProd***	***Unselected***
Paired	1000000 (100.00%)	1000000 (100.00%)	1000000 (100.00%)	1000000 (100.00%)
Aligned concordantly 0 times	25092 (2.51%)	26099 (2.61%)	26362 (2.64%)	36868 (3.69%)
Aligned concordantly exactly 1 time	875441 (87.54%)	856045 (85.60%)	857282 (85.73%)	851137 (85.11%)
Aligned concordantly more than 1 time	99467 (9.95%)	117856 (11.79%)	116356 (11.64%)	111995 (11.20%)
Pairs aligned concordantly 0 times	25092	26099	26362	36868
Aligned discordantly 1 time	3064 (12.21%)	2117 (8.11%)	2755 (10.45%)	2648 (7.18%)
Pairs aligned 0 times concordantly or discordantly	22028	23982	23607	34220
Mates make up the pairs	44056	47964	47214	68440
Aligned 0 times	26135 (59.32%)	37453 (78.09%)	35584 (75.37%)	57386 (83.85%)
Aligned exactly 1 time	8923 (20.25%)	5539 (11.55%)	6102 (12.92%)	6013 (8.79%)
Aligned more than 1 time	8998 (20.42%)	4972 (10.37%)	5528 (11.71%)	5041 (7.37%)
Overall alignment rate	98.69%	98.13%	98.22%	97.13%

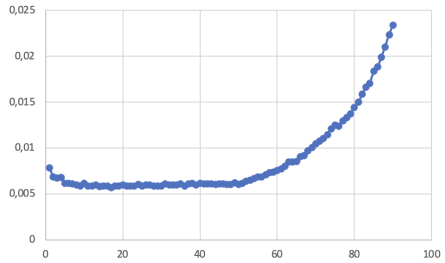
Table 1: **Estadisticas de alineamiento.**

Posteriormente, se calcularon las estadísticas de calidad, cubrimiento y tamaño de fragmento para cada una de las muestras. A partir de ello, al dividir el número de diferencias con el genoma de referencia considerando solo alineamientos únicos entre el total de alineamientos únicos, se obtuvo la tasa de error asociada. Las tasas de error promedio para cada muestra se presentan en la Tabla **??**. No obstante, esta es una sobreestimación porque algunas de las diferencias que se contabilizan en las columnas 2 (Número de diferencias con el genoma de referencia considerando todos los alineamientos) y 3 (Número de diferencias con el genoma de referencia considerando solo alineamientos únicos) son debidas a variación real.

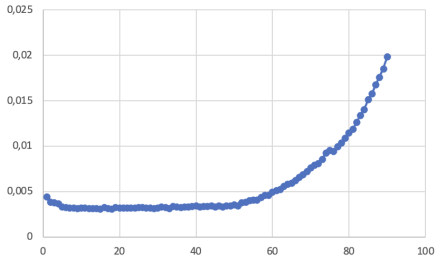
	Seg5	*PoolEthTol*	*PoolEthProd*	*Unselected*
Tasa de error promedio	0.0086	0.0057	0.0058	0.0059

Table 2: **Tasas de error de alineamiento para las diferentes muestras.**

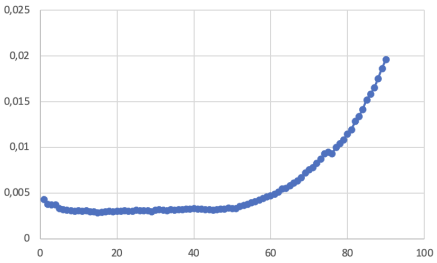
Luego, se graficó la curva que representa la tasa de error promedio para cada muestra (Figura **??**). La tasa de error máxima para las muestras Seg5, PoolEthTol, PoolEthProd y Unselected fueron 0.0235, 0.0196, 0.0199 y 0.0209 respectivamente. En todos los casos, en la base en la que ocurre esta tasa máxima es la que se encuentra en el extremo 3', es decir, la base número 90.



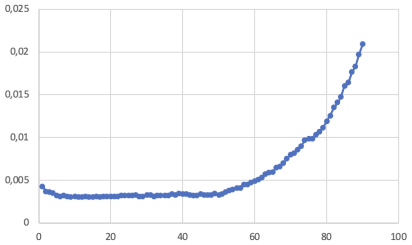
(a) Muestra: Seg5



(b) Muestra: PoolEthProd



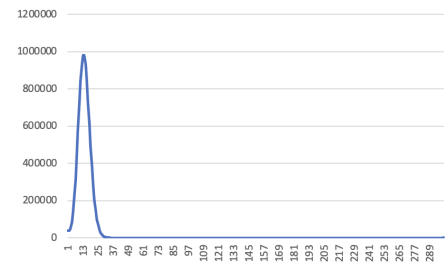
(c) Muestra: PoolEthTol



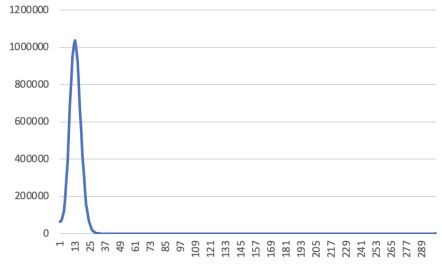
(d) Muestra: Unselected

Figure 1: **Distribución de la tasa de error promedio para las diferentes muestras.**

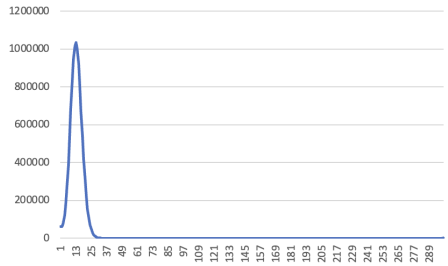
Finalmente, se calcularon las estadísticas de cubrimiento. La Figura a continuación muestra las gráficas de profundidad que se obtienen y la moda de la distribución para cada muestra. En todos los casos se presenta una distribución con moda cercana a 13. Estas distribuciones se calcularon empleando los datos de la columna 3.



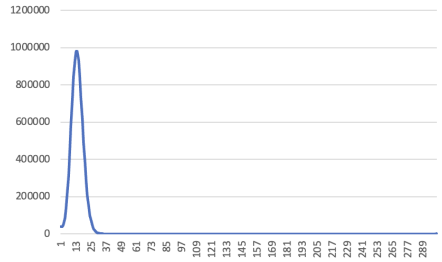
(a) Muestra: Seg5



(b) Muestra: PoolEthProd



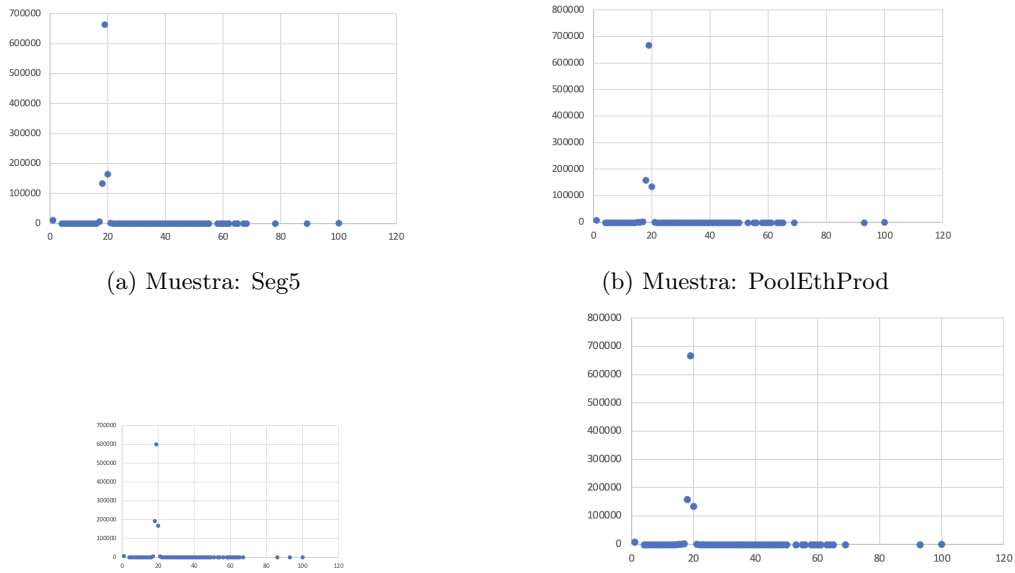
(c) Muestra: PoolEthTol



(d) Muestra: Unselected

Figure 2: **Distribución de la tasa de error promedio para las diferentes muestras.**

Finalmente, se calcularon las distribuciones de tamaño de fragmento para cada muestra con el objetivo de determinar qué tamaño se escogeria como mínimo y como máximo si fuera necesario ejecutar de nuevo el alineamiento. En las gráficas se observan algunos outliers, no obstante, la mayoria de datos se encuentran entre 0 y 500. Por lo anterior, se deberia seleccioanr 500 como el tamaño maximo. A continuación se presentan las graficas de distribución:



(a) Muestra: Seg5

(b) Muestra: PoolEthProd

(c) Muestra: PoolEthTol

(d) Muestra: Unselected

Figure 3: **Distribuciones de tamaño de fragmento para cada muestra .**

Problem 2
Visualización de alineamientos.

Solution. Después de la generación de los archivos .bam y .bai obtenidos a partir del genoma de levadura, pasamos a visualizarlos por medio de la herramienta IGV, en busca de regiones de interés. La visualización de lecturas alineadas por pares se ve como e la Figura **??**

Figure 4: Visualización de alineamientos por pares en IGV

¿Qué significan los diferentes colores asignados por IGV a los alineamientos?

De acuerdo con la documentación de IGV, los colores que se asignan a segmentos de los archivos ingresados pueden seguir tres códigos principales: según el tamaño del insert, según la orientación por pares, y según el bisulfite mode (que tiene seis modos posibles). En este caso, como es de nuestro interés identificar SNPs y otras modificaciones que pueden incluir inserciones o deleciones, el código de color a seguir es el de tamaño del insert. Para visualizar esto dentro de IGV, especificamos que la opción de *Color Alignments* este marcada en *insert size and pair orientation.*La asociación de colores se da como se indica a continuación:

Insertions: Las inserciones dentro de IGV se identifican cuando las lecturas, respecto al genoma de referencia, tienen un tamaño esperado mayor que el inferido sobre la secuencia de referencia. Estas secciones son marcadas en color **azul**.

Deletions: Las deleciones dentro de IGV se identifican cuando las lecturas, respecto al genoma de referencia, tienen un tamaño esperado menor que el inferido sobre la secuencia de referencia. Estas secciones son marcadas en color **rojo**.

Inter-chromosomal rearrangements: Las asociaciones de lecturas pareadas a su pertenencia a un cromosoma en particular se marcan según códigos de colores, teniendo cada cromosoma un color diferente asociado.

¿Cómo se ve una posición con un SNP? Determine 3 posiciones en las que visualmente parece haber un SNP e indique para cada muestra si el genotipo sería homocigoto o heterocigoto?

Un SNP puede definirse como un cambio en uno o pocos nucleotidos dentro de una secuencia de ADN, respecto a su genoma de referencia. Teniendo en cuenta el código de colores ya descrito, los SNPs dentro de IGV pueden observarse al revisar regiones coloreadas en azul o rojo sobre la curva de profundidad en dichas posiciones.

Al revisar a lo largo del alineamiento para la muestra Seg5, generada con bowtie2, se identificaron regiones de este tipo en múltiples posiciones y en distintos cromosomas. Se tomo la precaución de iniciar por colorear y organizar por pares las lecturas forward y reverse obtenidas de los mapeos, para poder descartar la posibilidad de que los cambios en 1 nt no fueran SNPs sino falsos positivos. Al observar que las mutaciones se presentan en ambas hebras, se logra comprobar que se trata de un SNP.

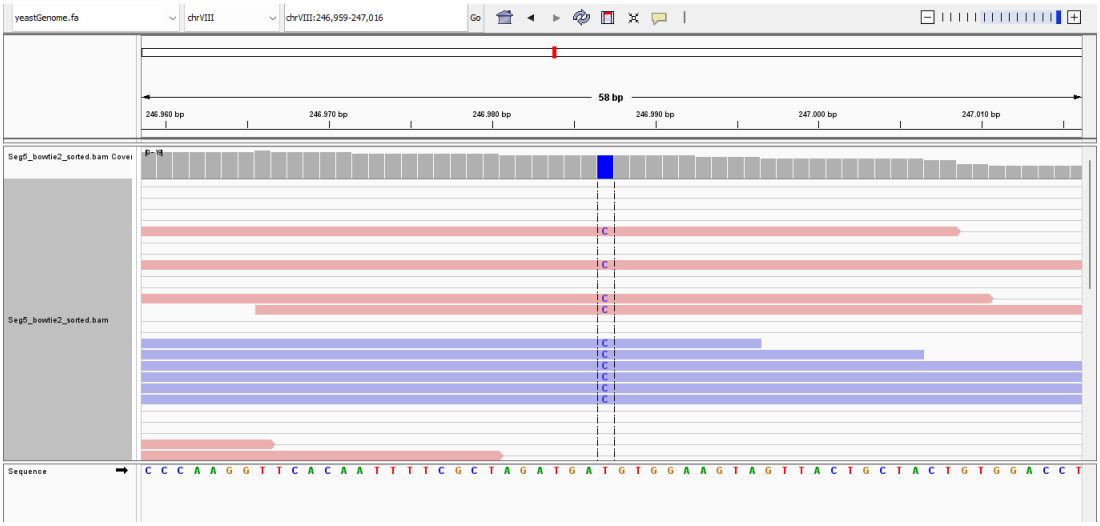


Figure 5: SNP 1 - Inserción de citosina en Ch VIII

La Figura **??** muestra la identificación de un posible SNP que corresponde a una inserción de una citosina en la posición 246987 del genoma. Este SNP es de caracter homocigoto, puesto que solo presenta una misma base añadida en todas las lecturas respecto a la secuencia de referencia.

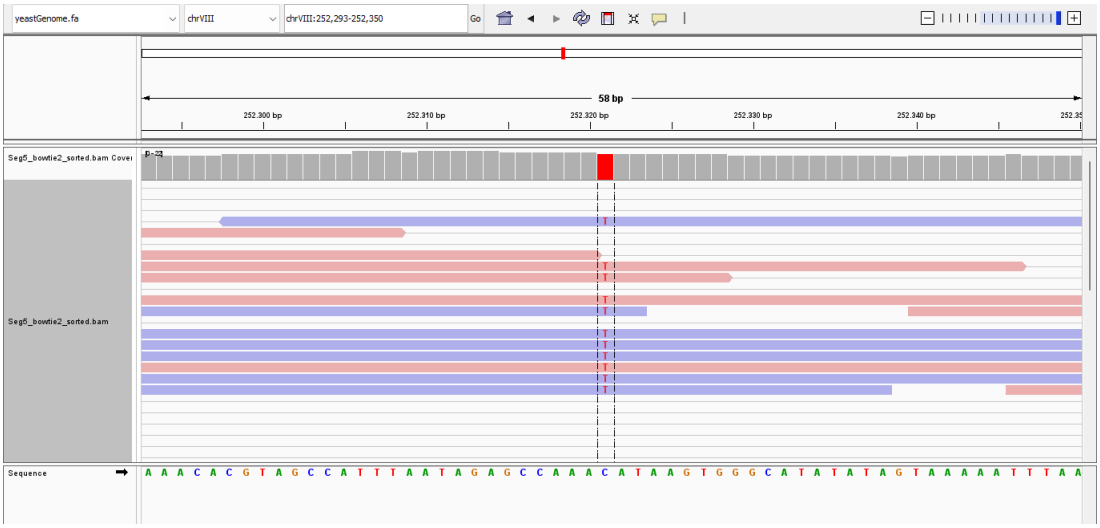


Figure 6: SNP 2 - Deleción de timina en Ch VIII

La Figura **??** muestra la identificación de un posible SNP que corresponde a una deleción de una timina en la posición 246987 del genoma. Este SNP es también de caracter homocigoto, puesto que solo presenta una misma base añadida en todas las lecturas respecto a la secuencia de referencia.

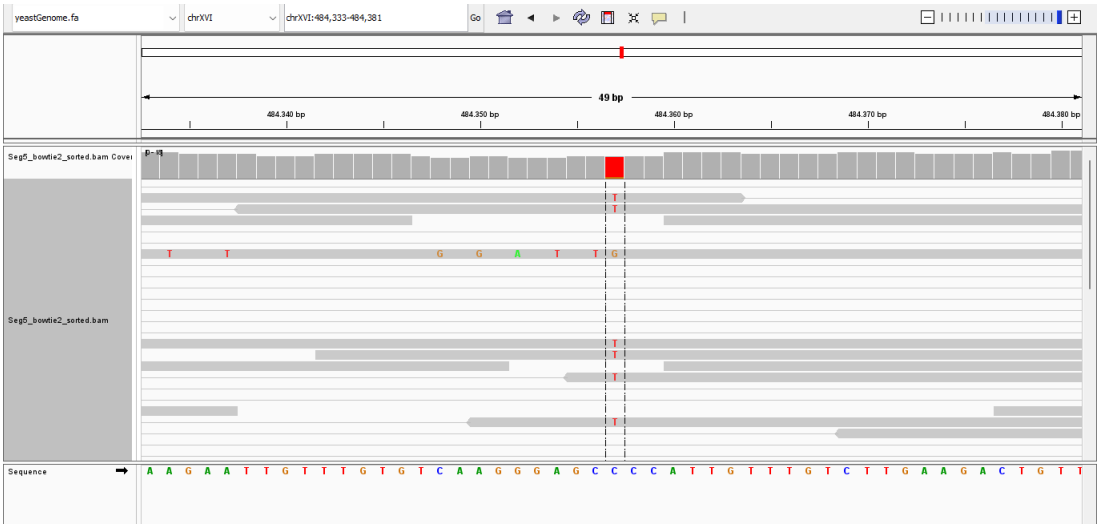


Figure 7: Deleción de timina con variante guanina en Ch XVI

Por otro lado, al revisar otros cromosomas mapeados encontramos un tercer caso presentado en el cromosoma XVI. La Figura **??** muestra la identificación de un posible SNP que corresponde a una deleción de una timina en la posición 246987 del genoma. Este SNP, sin embargo, difiere de los dos anteriores casos presentados, puesto que una de las lecturas contiene una guanina en lugar de una timina en esta posición. Por tal motivo, asociamos este SNP a una mutación de carácter heterocigoto (con dos posibles estados del mismo gen en esa posición).

Mismas preguntas del punto anterior para 2 borrados y para 2 inserciones pequeñas.

Pese a que no logramos identificar regiones de este estilo dentro de las lecturas mapeadas al genoma de referencia (dado que esperabamos encontrar estas regiones reflejadas en la curva de profundidad), si podemos establecer que estas deben seguir un mismo patrón de color y de cantidad de alelos. La Figura **??** muestra regiones subsecuentes de nucleótidos que no corresponden a los presentados en las posiciones de la secuencia original, y que pueden corresponder a inserciones o deleciones de regiones pequeñas pese a no verse reflejadas en múltiples lecturas. La homocigocidad o heterocigocidad de las mismas esta dada bajo los criterios expresados para los SNPs individuales previamente descritos.

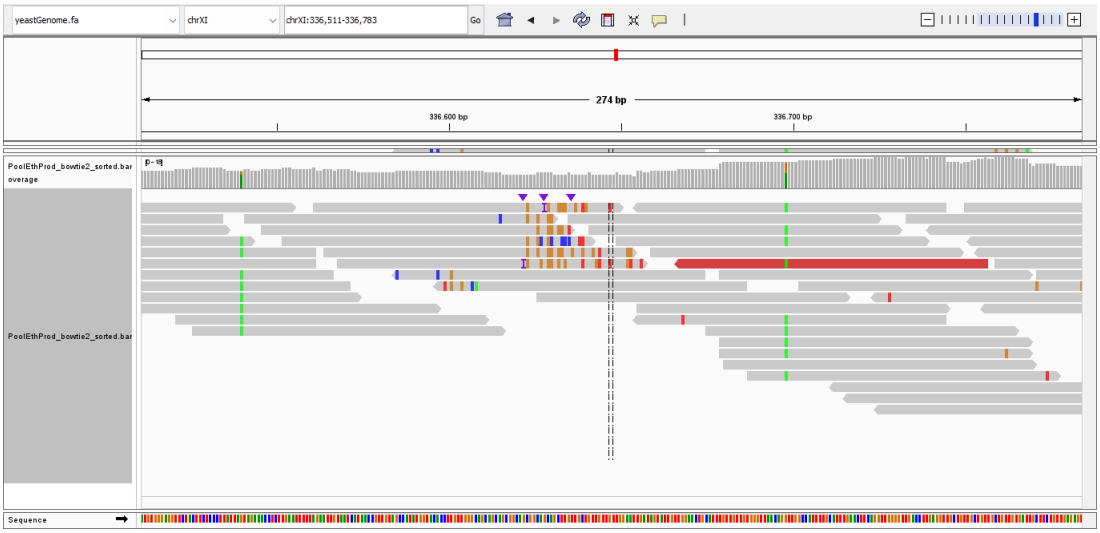


Figure 8: Posibles inserciones o deleciones en región del Ch XI

.

Problem 3
Alineamiento de lecturas largas.

Solution.

Por medio del cluster se realizó nuevamente mapeo con bowtie2, y adicionalmente mapeo con minimap2; esta vez de lecturas contra un genoma de referencia de coronavirus. Los resultados estan almacenados en la carpeta *estudiante20* de la carpeta estudiantes ubicada en bcom4006. Se registró la tasa de mapeo en ambos casos, siendo esta más directamente obtenida dado el funcionamiento de bowtie2. La tasa de alineamiento para bowtie2 fue de 27.8%, con 231 de 831 lecturas alineadas al menos 1 vez.

Problem 4
Desarrollo de índices FM.

Solution.

Para la implementación del índice FM, se diseño un programa orientado a objetos en Java, y se utilizó el genoma del CoronaVirus disponible en la página de NCBI: https://www.ncbi.nlm.nih.gov/nuccore/1798174254 para verificar la correcta funcionalidad.

Arreglo de sufijos. Primero, el algoritmo crea todos los sufijos de la secuencia con su posición de inicio. Luego, se ordenan los sufijos lexicográficamente, y se mapea a un arreglo de números

usando las posiciones donde aparece cada sufijo en la secuencia, siendo esto lo que se conoce como *Arreglo de sufijos*. Esta implementación para construir la estructura tiene una complejidad de tiempo de $O(n^2 lgn)$, por lo que requiere ordenar los n sufijos de tamaño n.

Utilizando el archivo $coronavirus.fasta$ y el archivo $coronavirus\_10\_20.fastq$ que contiene 10 lecturas de tamaño 20, se probo la implementación y se obtuvo el siguiente resultado.

Read:TACATTCTTTAAGAGTTTGT found at position 3740
Read:GTCTTTTGTACTGTAAATGC found at position 17301
Read:AATGTCTGGTAAAGGCCAAC found at position 28971
Read:TAATAGATACTTAGCTCTTT found at position 9876
Read:CTTAAGGGTGTAGAAGCTGT found at position 5548
Read:GGAACCTGAGTTTTATGAGG found at position 16187
Read:GCAATATGGCAGTTTTTGTA found at position 23823
Read:AATTCCCTCGAGGACAAGGC found at position 28466
Read:CTAGAGTATTAGGTTTGAAA found at position 6605
Read:TAATTAGAGGTGATGAAGTC found at position 22763

Al validar con el archivo del genoma, pudimos corroborar la correctitud del programa.

Notar que para este caso solo se reportaría el indice donde empieza la primera ocurrencia encontrada, pudiendo haber más. Ahora, buscar una lectura en la secuencia tiene un complejidad de $O(klgn)$ donde k corresponde al tamaño de las lecturas.

Indice FM. Con esta estructura, queremos poder encontrar ocurrencias de lecturas mucho más rápido que usando un *Arreglo de sufijos*.

En particular la complejidad de construir el índice está mayormente determinada por la construcción del arreglo de sufijos, pues este se utiliza para calcular la transformada de *Burrows-Wheeler*.

Por su parte, encontrar ocurrencias de lecturas en la secuencia tendría una complejidad de $O(k)$ asumiendo un tamaño del alfabeto constante.

Pudimos, igualmente, validar la correctitud del programa al comparar los resultados arrojados por el índice FM con los resultados obtenidos previamente usando la misma entrada de datos.

Problem 5

Bono. Modificar la construcción del arreglo de sufijos para generar el arreglo ordenado de posiciones sin tener que calcular explicitamente los sufijos

Solution. Almacenar explicitamente los sufijos de la secuencia requiere $O(n^2)$ espacio, pues almacenamos n sufijos de tamaño $O(n)$.

Es posible realizar esta construcción sin necesidad de usar espacio adicional utilizando un comparador que determine para dos sufijos S y T, cual es menor lexicográficamente mirando únicamente los cáracteres de la secuencia dada. Ver método *initV2* en la clase *SuffixArray.java*