Tarea 2 - Ensamblaje de genomas

Algoritmos en Biología Computacional

BCOM 4006

Laura Valentina Acosta Corredor 201911225

Juanita Puentes Mozo 201814823

Jhon Stewar Rayo Mosquera 201720166

Septiembre 13, 2023

Universidad de los Andes

September 18, 2023

Problem 1

Probar el programa implementado en la clase "ReadsAnalyzerExample" utilizando el comando "Kmers" y las lecturas dadas. Probar con cada fastq de ejemplo tamaños de k-mer 5, 10, 20, 50 y 75 y registrar la cantidad de k-mers, la abundancia media y el tiempo de ejecución.

Solution.
Al implementar los métodos de la clase **KmersTable**, se calcularon la cantidad de k-mers, la abundancia promedio y el tiempo de ejecución para cada uno de los archivos fastq que presentan diferentes profundidades de lectura (10x, 20x, 50x y 100x). Los resultados de este análisis se presentan en la Tabla 1.

Se observa que, para todos los tamaños de k-mer, la profundidad de lectura está directamente relacionada con la abundancia media. Esto significa que, a medida que aumenta la profundidad de lectura, la abundancia promedio de los k-mers también aumenta. Esto se interpreta considerando que la abundancia media representa cuántas veces, en promedio, se encuentra cada k-mer en el conjunto de datos de secuenciación. Por lo que si la profundidad es mayor, un k-mer tendrá mayor probabilidad de tener un mayor número de ocurrencias.

Por otro lado, el tiempo de ejecución aumenta de manera proporcional a la profundidad de lectura y de manera inversa al tamaño del k-mer. Esto indica que a medida que se aumenta la profundidad de lectura (es decir, se tienen más datos), se requiere más tiempo para analizarlos. Lo anterior se explica teniendo en cuenta que procesar una mayor cantidad de datos generalmente implica un mayor tiempo de cómputo. Por otro lado, el tiempo de ejecución disminuye a medida que se incrementa el tamaño del k-mer, ya que el uso de k-mers más pequeños acelera el proceso de análisis. Esto se debe a que con k-mers más pequeños, hay una mayor diversidad de k-mers diferentes en los datos, lo que ralentiza el proceso de análisis. En cambio, cuando el tamaño del k-mer aumenta, hay menos k-mers diferentes para analizar, lo que puede reducir el tiempo necesario para llevar a cabo el análisis.

Por último, la cantidad de k-mers para un tamaño de k-mer específico aumenta a medida que se incrementa la profundidad de lectura, dado que se generan más datos para cada secuencia. Además, se observa que la cantidad de k-mers en promedio es mayor cuando el tamaño del k-mer es 20. Para un tamaño de k-mer igual a 5, se presenta la menor cantidad de k-mers distintos. Esto se explica dado que, al tener una longitud menor, es mayor la probabilidad de que haya caracteres repetidos en los k-mers, lo que reduce la diversidad de k-mers distintos. Sin embargo, al utilizar un tamaño de k-mer igual a 20, disminuye la probabilidad de que haya k-mers repetidos, lo que resulta en una mayor cantidad de k-mers distintos en promedio.

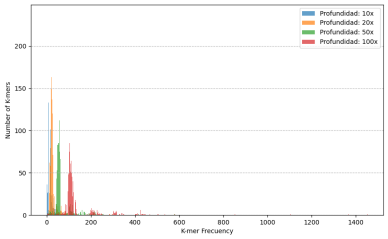
Tamaños de K-mer	10x	20x	50x	100x	K-mers	Abundancia Media	Tiempo de ejecución (ms)
5	✓	-	-	-	2153	13.383	79
	-	✓	-	-	2177	26.458	72
	-	-	✓	-	2175	66.207	85
	-	-	-	✓	2179	132.171	102
10	✓	-	-	-	2703	10.100	74
	-	✓	-	-	2741	19.920	73
	-	-	✓	-	2739	49.836	86
	-	-	-	✓	2743	99.526	107
20	✓	-	-	-	2758	8.811	69
	-	✓	-	-	2796	17.382	83
	-	-	✓	-	2794	43.486	85
	-	-	-	✓	2798	86.848	105
50	✓	-	-	-	2735	5.594	67
	-	✓	-	-	2773	11.035	66
	-	-	✓	-	2771	27.607	81
	-	-	-	✓	2775	55.135	100
75	✓	-	-	-	2549	3.06 0	63
	-	✓	-	-	2737	5.700	67
	-	-	✓	-	2746	14.202	77
	-	-	-	✓	2750	28.364	94

Table 1: **K-mers.** Resultados de cantidad de K-mers, abundancia media y tiempo de ejecución para diferentes tamaños de k-mer y distintas profundidades de lectura.

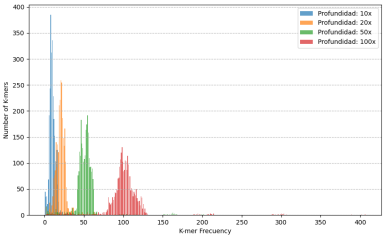
Posteriormente, se graficaron las distribuciones que representan el histograma de número de veces que aparece un k-mero. En esta graficas, el **eje x** hace al conteo de k-mers y el **eje y** representa la cantidad de k-mers. La Figura 1 presenta las diferentes distribuciones de k-mers para distintos tamaños de k-mero. Es importante tener en cuenta que en estas graficas, la altura de la barra representa la frecuencia con la que aparece ese k-mero en la secuencia. Cuanto más alta sea la barra, más frecuente es ese k-mero en la secuencia.

El tamaño del k-mero en una secuencia influye en su distribución de frecuencias. Con k pequeño, la presencia de numerosas combinaciones de k-meros con frecuencias bajas y desiguales da lugar a una distribución puntiaguda que se asemeja a la distribución de Poisson. Esto se debe a que algunos k-meros pueden ser comunes en varias partes de la secuencia, mientras que otros son raros. Por ejemplo, para un tamaño de k-mero igual a 5 (Figura 1a) se observan distribuciones con pendientes elevadas y con muchos outliers. En contraste, a medida que aumenta k (Figura 1d-1e), la probabilidad de que un k-mero específico aparezca disminuye, y la distribución tiende a converger hacia una forma más suave y simétrica, similar a una distribución normal. Esto sucede debido a la disminución de la variabilidad en las frecuencias de k-meros individuales.

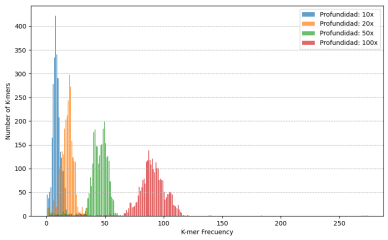
Es importante destacar que a medida que aumenta la profundidad de lectura, se observa un aplanamiento en la curva de las distribuciones y, este comportamiento se observa para todos los tamaños de k-mero (Figura 1). Cuando la profundidad de lectura es baja, las variaciones afectan más y el sesgo influye significativamente en la distribución de k-meros (dado que hay menor cantidad de datos). No obstante, a medida que se generan más datos con una mayor profundidad de lectura, se obtiene una representación más precisa de la distribución real de k-meros en la secuencia. Cuantos mas datos hay, la media tienda a acercarse al valor esperado, reduciendo así las variaciones y permitiendo una comprensión más precisa de la composición de la secuencia.



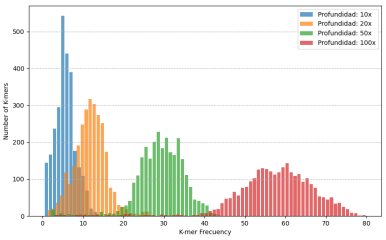
(a) Tamaño K-mer: 5



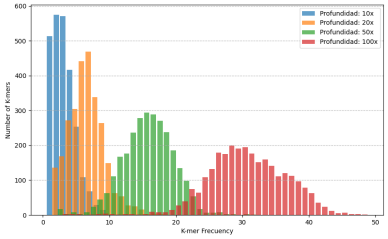
(b) Tamaño K-mer: 10



(c) Tamaño K-mer: 20



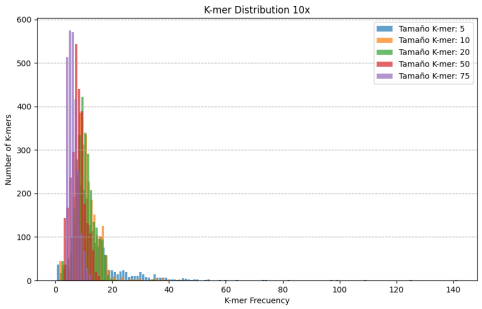
(d) Tamaño K-mer: 50



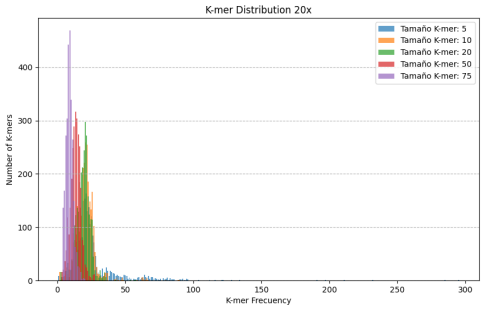
(e) Tamaño K-mer: 75

Figure 1: **Distribución de K-mers.** Distribuciones de K-mers agrupadas por los diferentes tamaños de k-mero (5,10,20,50 y 100) con distinciones para cada una de las profundidades.

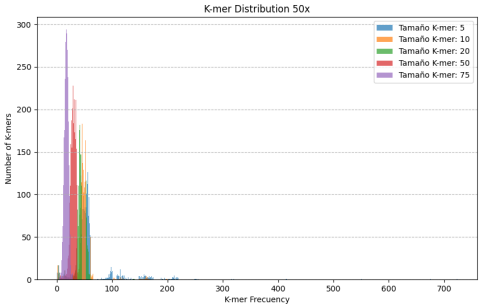
Se realizaron gráficas de distribución agrupadas según la profundidad de lectura con el propósito de evaluar el impacto del tamaño del k-mero en cada caso particular (Figura 2). En términos generales, se observa que a medida que aumenta el tamaño del k-mero, se incrementa la pendiente del pico en la distribución. Esto concuerda con el comportamiento previamente descrito, donde un tamaño de k-mero más pequeño resulta en una mayor diversidad de k-meros distintos y, por lo tanto, una dispersión más amplia a lo largo del eje x. Importante destacar que al aumentar la profundidad de lectura (Figura 2d), se observa una mayor diferenciación entre las curvas correspondientes a distintos tamaños de k-meros. Esto sugiere que una profundidad de lectura más elevada facilita la distinción entre k-meros de diferentes longitudes, haciendo que los k-meros más cortos, como k=5 y k=10, sean más fácilmente identificables debido a la mayor cantidad de datos disponible para cada tamaño de k-mero.



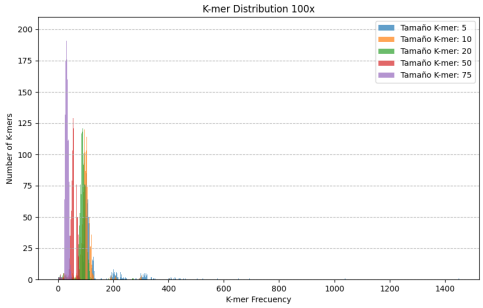
(a) 10x



(b) 20x



(c) 50x



(d) 100x

Figure 2: **Distribución de K-mers.** Distribuciones de K-mers agrupadas por la profundidad de lectura con distinciones para los diferentes tamaños de k-mero (5,10,20,50 y 100).

Problem 2

De la clase "OverlapGraph" implementar los primeros métodos hasta el método "calculateOverlapDistribution". Probar el programa implementado en la clase "ReadsAnalyzerExample" utilizando el comando "Overlap" y las lecturas dadas. Reportar para cada fastq de ejemplo disponible en la carpeta "data" la distribución de abundancias de secuencias, la distribución de sucesores por secuencia y el tiempo de ejecución.

Se implementaron los métodos de la clase **OverlapGraph** con diferentes tamaños de sobrelape minimo para las profundidades de lectura 10x, 20x, 50x y 100x. A partir de esto, se obtuvieron los resultados de tiempo de ensamblaje, tiempo de construcción del grafo, numero de sucesores y numero de secuencias diferentes para cada caso. Los resultados se presentan en la Tabla 2.

Al igual que ocurre con los K-mers, el aumento en la profundidad de lectura implica un incremento en el tiempo requerido tanto para ensamblar como para construir el grafo en la clase OverlapGraph. Esta relación se debe al hecho de que a medida que se aumenta la profundidad de lectura, se generan mayores cantidades de datos. Por esto, aumenta la carga computacional y se necesita más tiempo para ejecutar el algoritmo. Esto sucede dado que se necesita mayor capacidad

de procesamiento para manejar y analizar conjuntos de datos más grandes.

En los resultados de la Tabla 2, se observa un patrón distintivo: a medida que tanto el tamaño mínimo de solapamiento como la profundidad de lectura aumentan, se incrementan la cantidad de sucesores. Esto ocurre dado que al aumentar el tamaño mínimo, se permite una mayor superposición entre estos fragmentos, lo que resulta en la generación de sucesores más largos y, por lo tanto, en un mayor número de ellos. Una mayor profundidad de lectura implica que se están secuenciando más copias de la misma región, lo que aumenta la probabilidad ensamblar secuencias sucesoras adicionales que estén presentes en el genoma.

Finalmente, a pesar de que el número de sucesores aumenta, el numero de secuencias diferentes permanece constante para las diferentes profundidades de lectura. Este fenomeno puede ser el resultado de secuencias sucesoras más largas estén formadas a partir de las mismas secuencias originales. Es decir, los sucesores adicionales pueden ser redundantes o derivarse de regiones que ya se habían incluido como secuencias diferentes.

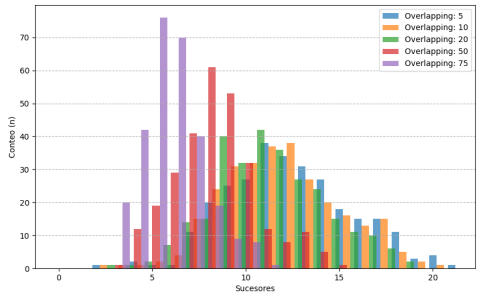
Overlapping	10x	20x	50x	100x	Sucesores Media	Abundancia Media	Tiempo ensamblaje (ms)	Tiempo construcción (ms)	Numero de secuencias diferentes
5	✓	-	-	-	12.995	95.000	3	380	285
	-	✓	-	-	15.486	135.500	5	935	542
	-	-	✓	-	16.083	231.600	16	3849	1158
	-	-	-	✓	19.042	261.143	41	8347	1828
10	✓	-	-	-	13.571	95.000	4	368	285
	-	✓	-	-	16.938	135.500	4	1096	542
	-	-	✓	-	18.677	231.600	15	3982	1158
	-	-	-	✓	22.024	261.143	30	9365	1828
20	✓	-	-	-	15.000	95.000	3	385	285
	-	✓	-	-	18.067	135.500	7	1305	542
	-	-	✓	-	24.125	231.600	13	3816	1158
	-	-	-	✓	27.697	261.143	25	9466	1828
50	✓	-	-	-	19.000	95.000	2	378	285
	-	✓	-	-	24.636	135.500	2	1036	542
	-	-	✓	-	34.059	231.600	9	3752	1158
	-	-	-	✓	40.622	261.143	22	9677	1828
75	✓	-	-	-	25.909	95.000	1	367	285
	-	✓	-	-	38.714	135.500	2	1100	542
	-	-	✓	-	50.348	231.600	6	1158	1158
	-	-	-	✓	70.308	261.143	13	9183	1828

Table 2: **Overlapping.** Resultados del numero de secuencias diferentes, abundancia media, sucesores promedio y tiempo de ejecución para diferentes tamaños de solapamiento y distintas profundidades de lectura.

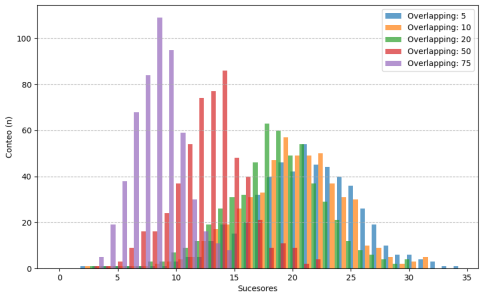
Posteriormente, se gráfico la distribución de los sucesores para cada una de las profundidades de lectura haciendo distinción entre los diferentes tamaños mínimos de solapamiento (Figura 3). Es notable que cuanto mayor sea la profundidad de lectura, hay mayor distinción entre las distribuciones de los diferentes tamaños de solapamiento. Por ejemplo, para 10X es evidente que las distribuciones para todos los tamaños de solapamiento se ubican, en su mayoría, entre valores de 5 a 20. Además, en este caso, las frecuencias de cada sucesor son generalmente menores a 40. En contraste, para una profundidad de 100x, se distinguen claramente las diferentes distribuciones marcadas por el tamaño mínimo de solapamiento. Cuando aumenta la profundidad de lectura, se secuencian más copias de la misma región del genoma, lo que aumenta la precisión y la cantidad de datos disponibles. Como resultado, se vuelven más evidentes las diferencias en las secuencias y en la distribución de sucesores asociados a diferentes tamaños mínimos de solapamiento.

Los sucesores con mayores tamaño de solapamiento se ubican mas a la izquierda (respecto al eje x) en la distribución. Cuanto menor es el tamaño de solapamiento, la distribución se ubica

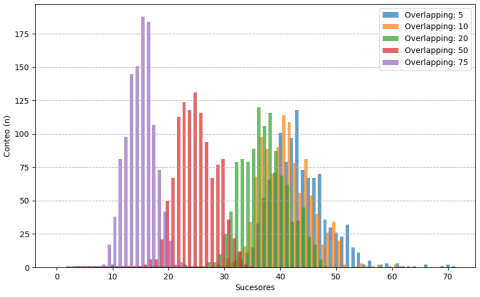
mas a la derecha. Además, cuento menor es el tamaño mínimo de solapamiento, la distribución se asemeja a una normal. Mientras que para tamaños mínimos de solapamiento mas grandes, la distribución tiene un pico mas grande. A medida que aumenta la profundidad de lectura, se detectan más fragmentos superpuestos, lo que permite una mayor variabilidad en la longitud de los sucesores. Los sucesores con tamaños de solapamiento más grandes se encuentran a la izquierda en la distribución porque representan fragmentos más largos que comparten más secuencia en común. Los sucesores con tamaños de solapamiento más pequeños se ubican más a la derecha porque representan regiones genómicas que se superponen en menor medida. Además, las distribuciones se asemejan a una distribución normal para tamaños mínimos de solapamiento más bajos, ya que hay una mayor probabilidad de que las secuencias se superpongan en múltiples puntos, lo que resulta en sucesores más diversos. Para tamaños mínimos de solapamiento más grandes, la distribución tiene un pico más pronunciado, lo que sugiere que hay un número limitado de sucesores predominantes.



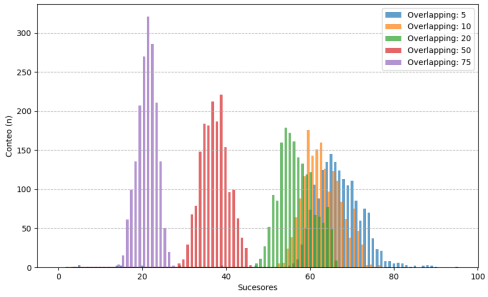
(a) 10x



(b) 20x



(c) 50x



(d) 100x

Figure 3: Distribución de Overlapping. Distribuciones agrupadas por la profundidad de lectura con distinciones para los diferentes tamaños minimos de solapamiento (5, 10, 20, 50 y 75)

Finalmente, se grafico la distribución de abundancia para diferentes profundidades de lectura (Figura 4). De igual manera como se observo para casos anteriores, a medida que aumenta la profundidad de lecturas se incrementan la cantidad de datos. Y, por ende, la abundancia se incrementa conforme la profundidad de lectura lo hace.

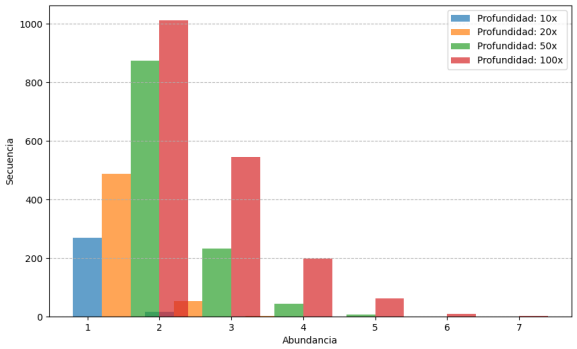


Figure 4: **Abundancia.** Distribución de abundancias para diferentes profundidades de lectura.

Problem 3

Implementar los métodos faltantes de la clase "OverlapGraph". Correr el programa para verificar si el programa trata de ensamblar una secuencia. Reportar la secuencia que se genera y el tiempo de ejecución para cada fastq de ejemplo. Reportar para cada fastq entre qué tamaños de sobrelape se puede ensamblar la secuencia.

Luego de implementar los métodos faltantes de la clase **OverlapGraph**, se ensamblaron las secuencias para cada fastq de ejemplo. En la sección de Anexos se presentan cada una de las secuencias ensambladas para cada profundidad de lectura. De cada uno de los ensamblajes es notorio que a medida que aumenta la profundidad de lectura, hay una refinación significativa en los detalles del párrafo ensamblado. Además, es evidente que el significado semántico y la coherencia entre las oraciones se fortalecen a medida que la profundidad de lectura se incrementa. Esto se debe se debe a la disponibilidad de una mayor cantidad de datos para cada fragmento, lo que permite una reconstrucción más precisa del texto original.

Para ilustrar esto con un ejemplo, consideremos una profundidad de lectura de 10X. En este caso, el texto ensamblado comienza con la frase: **"l pelotón de fusilamiento el coronel"**. Sin embargo, al aumentar la profundidad de lectura, el inicio del texto se transforma en algo más completo, como: **"Muchos años después, frente al pelotón de fusilamiento, el coronel"**. Esta evolución demuestra cómo la información adicional disponible con profundidades de lectura mayores enriquece el texto y mejora su coherencia semántica.

La Tabla 3 muestra los tamaños de sobrelape requeridos para ensamblar cada secuencia junto con el tiempo necesario para ejecutar el ensamblaje total. A medida que la profundidad de lectura aumenta, se observa una reducción en la amplitud del rango de tamaños de solapamiento necesarios. Por ejemplo, en el caso de una profundidad de lectura de 10x, este rango oscila entre 50 y 100. Sin embargo, cuando la profundidad de lectura alcanza los 100x, este rango se estrecha aún más, situándose entre 92 y 100. Esto sucede dado que a medida que la profundidad de lectura aumenta, se reduce la incertidumbre con respecto a la ubicación precisa de los solapamientos entre

los fragmentos de secuencia. En profundidades de lectura más bajas, donde la cantidad de datos es limitada, es necesario considerar un rango más amplio de tamaños de solapamiento debido a la mayor probabilidad de errores y variabilidad en la secuenciación. Por otro lado, en profundidades de lectura más altas, la precisión en la determinación del solapamiento se incrementa, lo que se traduce en un rango más estrecho de tamaños de solapamiento requeridos.

Finalmente, es importante destacar que el tiempo necesario para llevar a cabo el proceso de ensamblaje está directamente relacionado con la profundidad de lectura. A medida que la profundidad aumenta, se requiere analizar una mayor cantidad de datos, lo que prolonga el tiempo necesario para completar la ejecución completa del algoritmo.

Profundidad de Lectura	Tamaños de sobrelape	Tiempo de ensamblaje (ms)
10x	50-100	3
20x	70-100	5
50x	87-100	16
100x	92-100	41

Table 3: **Tamaños de sobrelape.** Tamaños de sobrelape con los que se puede ensamblar cada secuencia con diferentes profundidades de lectura y que tiempo de ejecución es requerido.

Problem 4

Completar el script "SimpleReadsSimulator" para generar lecturas aleatorias de una secuencia de por lo menos 10Kbp en formato fasta. Probar los comandos "Overlap" y "Kmers" de "ReadsAnalyzerExample" con tamaños de secuencia 50, 100, 200, 500 y con profundidad promedio de 5X, 10X, 20X, 50X y 100X. Reportar tiempos de ejecución de los algoritmos en cada caso. Determinar la cantidad de lecturas máxima que puede procesar cada comando con un máximo de memoria de 4Gb.

Solution.

Para la ejecución del simulador de errores mediante la generación de lecturas aleatorias, se hizo uso del archivo .fasta correspondiente al genoma completo de *Homo sapiens isolate NS07 mitochondrion*, con número de acceso GU170821.1. Esta secuencia corresponde a 16569 bp de ADN circular mitocondrial de *Homo sapiens*. El algoritmo se planteo de tal manera que se pudieran recibir valores de tamaño de lectura y de profundidad variables, bajo los cuales se realizan los fragmentos de lectura de salida. De esta manera, para un tamaño de lectura de 50bp y una profundidad de 5X, se obtienen un total de 5 lecturas con sitio de inicio aleatorio, y una longitud de 50bp. Las tablas a continuación ilustran los resultados obtenidos para las distintas combinaciones de tamaño de lectura y profundidad promedio en términos del tiempo de ejecución en ms, para los métodos de sobrelapamiento y k-meros respectivamente.

Los tiempos de ensamblaje para el método de sobrelapamiento (overlapping) se presentan coherentes para todos los tamaños de lectura tomando cada una de las profundidades planteadas (5X, 10X, 20X, 50X y 100X), con excepción de la profundidad de 20X para un tamaño de lectura

de 50bp. Este comportamiento puede asociarse a factores de manipulación referentes a un uso secundario de la memoria RAM, o a un aumento en la complejidad en el ensamblaje debido a las lecturas aleatorias generadas. Este aumento también fue consistente con el de la generación de la gráfica de sobrelapamiento, bajo la cual se realiza el proceso de ensamblaje. Las diferencias de magnitud en el tiempo de ejecución total según el tamaño de la lectura corresponden a lo esperado, siendo estos proporcionales a este factor, y a la cantidad total de lecturas evaluadas. Se utilizó un valor de sobrelapamiento mínimo de 5 bases, teniendo en cuenta que para el texto ensamblado en puntos anteriores (28 letras del alfabeto, alrededor de 800 combinaciones diferentes), el valor sugerido fue 50 para una profundidad de 10X. Este parámetro fue disminuido entonces, partiendo de la cantidad de combinaciones posibles para solo 4 letras (cantidad de nucleótidos diferentes).

Tamaño de la lectura	5x	10x	20x	50x	100x	Tiempo gràfica sobrelapamiento (ms)	Tiempo de ensamblaje (ms)
50	✓					95	0
		✓				88	0
			✓			60	4
				✓		103	1
					✓	157	4
100	✓					48	0
		✓				64	0
			✓			85	0
				✓		94	2
					✓	184	4
200	✓					83	0
		✓				77	0
			✓			92	0
				✓		139	2
					✓	341	4
500	✓					56	1
		✓				77	4
			✓			112	4
				✓		305	4
					✓	796	13

Table 4: **Overlapping.** Resultados de tiempo de generación de gráfica de sobrelapamiento y tiempo de ejecución para tamaños de secuencia de 50, 100, 200, 500 y con profundidad promedio de 5X, 10X, 20X, 50X y 100X, generados mediante la clase *SimpleReadsSimulator*, a partir del gen *Homo sapiens isolate NS07 mitochondrion* (HS-MIT). Overlap minimo = 5.

Además de estas experimentaciones, se realizaron pruebas de la cantidad máxima de secuencias que podían ser procesadas con una capacidad de alrededor de 10 GB, dado que los comandos de limitación de memoria para la ejecución de la clase no pudieron ser implementados. La Tabla 5 muestra el tiempo de ejecución del algoritmo para diferentes número de secuencias, en términos del tiempo para generar la gráfica de sobrelapamiento y el tiempo para realizar el ensamblaje, en ms. Se observa que el tiempo crece de manera proporcional a la cantidad de secuencias ingresadas, siendo de mayor magnitud para la construcción de la gráfica en todos los casos. Se partió del valor mayor de secuencias con las que se realizaron las experimentaciones anteriores (500bp), y se aumentó de manera progresiva para observar un cambio gradual en los tiempos. Evidenciamos que, nuevamente, el comportamiento del tiempo es exponencial, tomando un total de 0.2 min para procesar 500 lecturas y 40 min para el procesamiento de 7500 secuencias. Para el caso de 10000 secuencias, se estimó que el tiempo total de procesamiento sería de aproximadamente 60 min, de acuerdo a todos los comportamientos anteriores. Sin embargo, el programa no generó un output ni tampoco se detuvo al cabo de 75 min, por lo que se asumió que el máximo de secuencias a procesar

con 10 GB de memoria está en el rango entre 7500-10000.

Número de secuencias	Tiempo gráfica de sobrelapamiento (ms)	Tiempo de ensamblaje (ms)
500	13348	51
800	32392	81
1000	51682	138
2000	217179	239
3000	473815	604
5000	1202053	2447
7500	2371913	2860
10000	-	-

Table 5: **Overlapping.** Cantidad de lecturas máximas para procesamiento con memoria de 9,88 GB

Se condujo un análisis similar para el caso de los K-meros, esta vez registrando el tiempo de ensamblaje en ms y el número de k-meros generados en cada caso, para un tamaño de k-mero de 75. Una vez más, este número creció de forma proporcional al aumento en el número de secuencias y la profundidad, aunque los tiempos de ensamblaje no presentaron este mismo comportamiento. La variabilidad del tiempo se atribuye al factor aleatorio bajo el cual se generan las secuencias a ensamblar, puesto que esto remueve la continuidad que se esperaría en la ejecución. Al utilizar un tamaño de k-mero (k) de esta magnitud, se obtiene que para un tamaño de lectura de 50bp - menor al k) - no se generan k-meros debido a que no hay una longitud suficiente para generarlos. Para todos los otros grupos de tamaño de lectura, se generan tantos k-meros de longitud 75 como es posible dentro de esa extensión de secuencia.

Tamaño de la lectura	5x	10x	20x	50x	100x	Número de Kmers	Tiempo de ensamblaje (ms)
50	✓					0	55
		✓				0	57
			✓			0	72
				✓		0	59
					✓	0	94
100	✓					130	50
		✓				260	54
			✓			520	53
				✓		1214	50
					✓	2414	57
200	✓					630	71
		✓				1260	52
			✓			2411	77
				✓		5405	84
					✓	8532	95
500	✓					2130	64
		✓				3976	65
			✓			6508	83
				✓		12178	84
					✓	14269	112

Table 6: **Kmers.** Resultados de cantidad de Kmers y tiempo de ejecución para tamaños de secuencia de 50, 100, 200, 500 y con profundidad promedio de 5X, 10X, 20X, 50X y 100X, generados mediante la clase *SimpleReadsSimulator*, a partir del gen *Homo sapiens isolate NS07 mitochondrion* (HS-MIT). Tamaño del Kmer = 75.

Al igual que pare el método de sobrelapamiento, se realizó un proceso iterativo de procesamiento de secuencias de distintas longitudes, con el fin de evaluar la cantidad de secuencias bajo la cual el programa era incapaz de continuar dada su memoria disponible. Se inició también por un total de 500 secuencias a procesar, lo que requirió de 0.003 min para ejecutarse; y se evaluaron un total de hasta 20000 secuencias, lo cuál tomó 0.04 min de procesamiento (Tabla 7). Con esto se pude evidenciar que el método de k-meros toma menor cantidad de tiempo general para la ejecución completa que involucra el procesamiento de las secuencias de entrada respecto al método de sobrelapamiento, y requiere de menos memoria para procesar la misma cantidad de secuencias. Se estima que para una memoria de 4 GB, el procesamiento de las 20000 debería demorar aproximadamente 1 min.

Número de secuencias	Número de Kmers	Tiempo de ensamblaje (ms)
500	16453	192
800	16483	191
1000	16883	208
2000	16464	349
3000	8869	487
5000	16491	675
7500	16486	1115
10000	16490	1267
20000	16493	2278

Table 7: **Kmers.** Cantidad de lecturas máximas para procesamiento con memoria de 9,88 GB

Problem 5

Mejorar el script reads simulator incluyendo un parámetro nuevo que permita simular una tasa de error aleatoria. Simular datos con un tamaño de lectura de 50bp y profundidad promedio de 20X. Visualizar y comparar la distribución de abundancia de k-mers en cada caso.

Solution.

Para este caso, se realizaron dos funciones auxiliares dentro del *SimpleReadsSimulator*, para introducir el valor de tasa de error escogido dentro de la generación de secuencias aleatorias. Una de las funciones se encarga de escoger aleatoriamente una base entre 3 restantes a la base analizada, con el fin de generar una lectura errónea en esa posición. Esta función solo se aplica al validar que el valor de error aleatorio que se genera por posición, sea efectivamente menor al valor de la tasa de error ingresada por parámetro. Se realizaron pruebas para valores de tasa de error de 0.01, 0.1, 0.25, 0.5, 0.75 y 1.0, siendo el valor más pequeño aquel reportado como asociado a algunas tecnologías de secuenciación existentes. La Tabla 8 muestra el comportamiento presentado por el método de sobrelapamiento, cuyo tiempo de ensamblaje se mantiene para todas las tasas, y el tiempo de generación de las gráficas aumenta conforme la tasa va incrementando. Este comportamiento puede ser explicado por el criterio de aleatoriedad, y la probabilidad de que un valor aleatorio sea

menor a la tasa indicada. Entre mayor es el valor de la tasa, aumenta la probabilidad de caer sobre este nuevo rango, por lo que obtener lecturas que no coincidan en al menos 5bp resulta más sencillo.

Tasa de error	Tiempo gráfica de sobrelapamiento (ms)	Tiempo de ensamblaje (ms)
0.01	80	1
0.1	67	1
0.25	64	1
0.5	71	1
0.75	104	1
1.0	147	1

Table 8: **Overlapping.** Efecto sobre tiempo de ejecución de la tasa de error definida.

Se realizó un análisis similar para el método de k-meros, bajo el cual solo se obtuvo una diferencia en el tiempo del ensamblaje, y se obtuvo un valor fijo de 520 k-meros en todos los casos. El tiempo de ensamblaje tampoco se comportó de acuerdo al aumento de la tasa, probablemente debido a la cantidad de secuencias únicas generadas por el valor aleatorio agregado.

Tasa de error	Número de Kmers	Tiempo de ensamblaje (ms)
0.01	520	58
0.1	520	65
0.25	520	54
0.5	520	68
0.75	520	73
1.0	520	58

Table 9: **Kmers.** Efecto sobre tiempo de ejecución de la tasa de error definida.

Con el fin de corroborar la distribución de K-meros, se generó la gráfica ilustrada en la Figura 5. Esta muestra como solo se genera 1 - máximo 2 - k-meros por cada una de las tasas de error, lo cual es consistente con la introducción de las. Es probable que se estén generando errores cada vez más frecuentemente, lo que incrementa la probabilidad de generar secuencias similares a partir de secuencias inicialmente diferentes, en especial cuando se trata de un conjunto tan pequeño de elementos (4 nts). La Figura muestra como, para una tasa de error de 0.01, se alcanzan a formar 2 k-meros a partir de las secuencias aleatorias de 50bp, mientras que al aumentar la tasa, las secuencias se agrupan y convergen a un solo grupo.

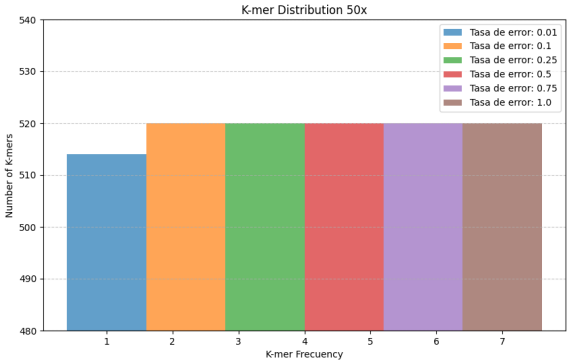


Figure 5: **Abundancia de K-mers.** Distribución de K-mers para diferentes tasas de error, y un valor de 50x de profundidad.

References

1 Anexos

A continuación se presentan los ensamblajes obtenidos para diferentes profundidades de lectura.

Profundidad 10X

l peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y cañabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que señalarlas con el dedo. Todos los años por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentrañar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en

aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empeño en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo

Profundidad 20X

chos años despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y cañabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que señalarlas con el dedo. Todos los años por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentrañar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empeño en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura

encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo de mujer.

Profundidad 50X

os años despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y cañabrava construidas a la orilla de un rio de aguas diafanas que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que señalarlas con el dedo. Todos los años por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentrañar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empeño en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo de mujer.

Profundidad 100X

Muchos años despues frente al peloton de fusilamiento el coronel Aureliano Buendia habia de recordar aquella tarde remota en que su padre lo llevo a conocer el hielo. Macondo era entonces una aldea de veinte casas de barro y cañabrava construidas a la orilla de un rio de aguas diafanas

que se precipitaban por un lecho de piedras pulidas blancas y enormes como huevos prehistoricos. El mundo era tan reciente que muchas cosas carecian de nombre y para mencionarlas habia que señalarlas con el dedo. Todos los años por el mes de marzo una familia de gitanos desarrapados plantaba su carpa cerca de la aldea y con un grande alboroto de pitos y timbales daban a conocer los nuevos inventos. Primero llevaron el iman. Un gitano corpulento de barba montaraz y manos de gorrion que se presento con el nombre de Melquiades hizo una truculenta demostracion publica de lo que el mismo llamaba la octava maravilla de los sabios alquimistas de Macedonia. Fue de casa en casa arrastrando dos lingotes metalicos y todo el mundo se espanto al ver que los calderos las pailas las tenazas y los anafes se caian de su sitio y las maderas crujian por la desesperacion de los clavos y los tornillos tratando de desenclavarse y aun los objetos perdidos desde hacia mucho tiempo aparecian por donde mas se les habia buscado y se arrastraban en desbandada turbulenta detras de los fierros magicos de Melquiades. Las cosas tienen vida propia - pregonaba el gitano con aspero acento - todo es cuestion de despertarles el anima. Jose Arcadio Buendia cuya desaforada imaginacion iba siempre mas lejos que el ingenio de la naturaleza y aun mas alla del milagro y la magia penso que era posible servirse de aquella invencion inutil para desentrañar el oro de la tierra. Melquiades - que era un hombre honrado - le previno: Para eso no sirve. Pero Jose Arcadio Buendia no creia en aquel tiempo en la honradez de los gitanos asi que cambio su mulo y una partida de chivos por los dos lingotes imantados. Ursula Iguaran su mujer que contaba con aquellos animales para ensanchar el desmedrado patrimonio domestico no consiguio disuadirlo. Muy pronto ha de sobrarnos oro para empedrar la casa replico su marido. Durante varios meses se empeño en demostrar el acierto de sus conjeturas. Exploro palmo a palmo la region inclusive el fondo del rio arrastrando los dos lingotes de hierro y recitando en voz alta el conjuro de Melquiades. Lo unico que logro desenterrar fue una armadura del siglo XV con todas sus partes soldadas por un cascote de oxido cuyo interior tenia la resonancia hueca de un enorme calabazo lleno de piedras. Cuando Jose Arcadio Buendia y los cuatro hombres de su expedicion lograron desarticular la armadura encontraron dentro un esqueleto calcificado que llevaba colgado en el cuello un relicario de cobre con un rizo de mujer.