Técnicas y Modelos Algorítmicos

Ingeniería de la Salud, 2021

Relación de ejercicios (2021) ----------

Los ejercicios serán evaluados a partir de un Informe de trabajo que debe contener una breve explicación de cómo se ha desarrollado / organizado el código (a nivel funcional, no son importantes los detalles); y un equivalente a una guía de usuario (*user guide*) o a un manual de ejecución (*guided exercises*) que indique ejemplos de entrada y resultados

Los resultados se deben comentar en lo relativo a resultados comparativos y/o evolución parametrizada de las ejecuciones. Es decir, no es importante indicar que una gráfica "crece con la longitud"; sino, "a que se debe que la gráfica crezca con la longitud"; o como evoluciona el tiempo de ejecución en función del tamaño de los datos.

Es conveniente entregar el código en un fichero aparte; separado del Informe

Identifique todos sus trabajo con su APELLIDOS (en caso de grupos, separe los apellidos con el signo +) y una breve descripción.

1.- (**obligatorio**) Cálculo de fragmentos entre codones de parada

EL programa debe ser capaz de:

- Leer una secuencia en formato FASTA desde un fichero cuyo nombre (path) se recibirá como argumento en la invocación del programa
- Deber recorrer los 6 diferentes marcos de lectura buscando codones de parada (TAA, TAG, TGA) y determinando la distancia entre los fragmentos contenidos entre dichos codones
- Producir un histograma que muestre la distribución de las longitudes de los histogramas

2.- (Opcional). Cálculo de las secuencias, reversa y reversa complementaria

- Leer una secuencia en formato FASTA desde un fichero cuyo nombre (path) se recibirá como argumento en la invocación del programa
- Producir las secuencias reversa, y reversa complementaria

3.- (**obligatorio**) Calculo de frecuencias de Kmers

- Leer una secuencia en formato FASTA desde un fichero cuyo nombre (path) se recibirá como argumento en la invocación del programa. En este caso es importante considerar ficheros multi-fasta, es decir, que contengan más de una secuencia. Los resultados se referirán a cada una de las secuencias. Si es oportuno se podrán incluir resultados globales

- El tamaño de la palabra (kmer) se obtendrá desde la línea de comando (argumento). Puede ser único, puede ser un rango; …
- El programa deberá crear un vector (array) o lo que se considere adecuado para contar el número de apariciones de cada kmer en la secuencia. Es posible que contar palabras de un determinado tamaño ya no sea posible con los datos en memoria (limitación en el valor de K), pero debería ser capaz de contar hasta valores de K=7. o más
- (opcional) Implementar el mismo ejercicio sin LIMITAICON en el tamaño de K
- Las frecuencias se indicarán en forma absoluta y relativa
- Adicional: Es posible calcular frecuencias de di-nucleótidos (k=2) para un donjunto de secuencias de distintos organismos (preferible la misma secuencia perteneciente a diferentes organismos); y luego calcular la "distancia" entre los vectores de frecuencias
- La distancia entre vectores de frecuencias de di-nucleótidos es un indicador primero de la separación entre especies
- Un resultado interesante de este ejercicio es hacer las gráficas de tiempo de ejecución en función del tamaño de la secuencia; y parametrizadas por el tamaño de la palabra (valor K)

4.- Ejercicio final (Obligatorio)

Desarrolle un programa para comparar secuencias basado en el uso de semillas (repeticiones de kmers en ambas secuencias). El programa:

- Debe leer dos secuencias de ADN desde disco
- Calcular la composición de kmers (palabra, posición)
- Ordenar los kmer de cada secuencia por la palabra (el kmer) y producir una colección de {kmer, pos1, pos2, pos3… posN}; es decir donde aparece cada kmer en la secuencia (es conveniente incluir el valor del "numero de repeticiones del kmer)
- Obtener las semillas o hits (kmer, posX, posY) donde X e Y son las dos secuencias leídas
- Ordenar las semillas por diagonal y offset (la diagonal es posX – posY) y como offset puede usar posX o posY, pero siempre la misma
- A partir de los hits o semillas, extender los fragmentos. Utilice pesos +1 y -1 para coincidencias y diferencias
- Reporte el conjunto de fragmentos

El profesor: Oswaldo Trelles

Málaga, 20 de Marzo de 2021