Algoritmo para eliminar MEMs contenidos en otros MEMs

Julio César García Vizcaíno

27 de septiembre de 2012

1. Introducción

Un MEM es una coincidencia exacta máxima y es utilizado en el alineamiento de genomas. El MEM se define como una cadena de caracteres, de cierta longitud mínima, que se repite una o mas veces dentro de una secuencia de referencia y una secuencia de consulta y que no puede extenderse hacia la izquierda o la derecha sin incurrir en una discrepancia.
El alineamiento de genomas involucra la búsqueda de MEMs y debido a la longitud de los genomas de referencia y de consulta la tarea computacional de encontrar todos los MEMs tiene un tiempo de ejecución casi óptimo, donde el tiempo óptimo sería proporcional a la suma de los tamaños de los genomas de referencia y consulta y el número de MEMs encontrados. Sin embargo, si la longitud mínima de MEM a buscar es muy pequeño, entonces la cantidad de MEMs puede ser muy grande.
Para reducir el tiempo de ejecución se puede utilizar técnicas de paralelismo. Una opción es el paralelismo de datos, esta técnica consiste en hacer una distribución de los datos de entrada, en este caso el genoma de referencia y de consulta entre las diferentes instancias de procesamiento de cómputo (threads, procesos, nodos de cómputo, etc.) y en cada unidad de procesamiento realizar la búsqueda de MEMs de manera local en el segmento de datos asignado y en la fase posterior unir la lista de MEMs de cada segmento para tener la lista completa de MEMs. Los problemas que surgen de esta técnica son:

1. Uso eficiente de los recursos de cómputo.

2. Distribución de datos.

3. Evitar:

   - Redundancia de datos.
   - Pérdida de datos.

2. Búsqueda de MEMs

La primera fase para el alineamiento de grandes genomas es la búsqueda de coincidencias exactas entre dos secuencias de ADN, la secuencia de referencia R y de consulta Q. Una coincidencia exacta máxima es uno de los mecanismos utilizados en esta fase. El resultado de buscar todos los MEMs ente R y Q es una lista donde cada elemento está compuesto por la posición de inicio del MEM en la referencia, la posición de inicio del MEM en la consulta y la longitud del MEM.

2.1. Ejecución serial

Encontrar el conjunto de MEMs entre una secuencia de referencia de longitud n y de consulta de longitud m, de una longitud mínima de MEM l tiene una complejidad computacional $O(m + k)$ donde k es el número de MEMs encontrados para la coincidencia de longitud m.

2.2. Ejecución paralela

Realizar la búsqueda de MEMs de forma paralela implica definir como se realizará la ejecución paralela. Se parte del principio de utilizar el paralelismo a nivel de datos, es decir, dividir los datos de entrada y realizar la búsqueda en cada uno de estos bloques y al finalizar obtener la lista global de MEMs. Sin embargo esto no es tan sencillo de llevar a cabo.
Esta técnica se divide en las siguientes 3 fases:

1. Dividir la secuencia de consulta entre el número de instancias de procesamiento de cómputo disponibles.

2. Ejecución en paralelo para cada uno de los bloques de la secuencia de consulta.

3. Obtener la lista de MEMs definitivos a partir del conjunto de lista de MEMs local de cada uno de los bloques de la secuencia de consulta.

La división del genoma de consulta se realiza utilizando el paradigma del paralelismo a nivel de datos que consiste en la generación de bloques de una secuencia con un tamaño fijo y un solapamiento fijo. Un problema surge al hacer la división del genoma de consulta, esto es al hacer la división se tiene que tomar en cuenta si la partición se hace dentro de una subcadena que contiene múltiples repeticiones de la letra "n", que biológicamente puede significar cualquier nucleótido: (a,c,g,t). De tal manera que la subcadena entera de repeticiones de "n" debe estar contenida en un solo bloque al momento de hacer la división.
Los dos parámetros que definen como se realizará la generación de los bloques para el procesamiento en paralelo son:

- Tamaño del bloque: la longitud del bloque está dado por el número de instancias de procesamiento de cómputo disponibles, sin embargo el tamaño del bloque no siempre será igual si el bloque se corta en una subcadena que contiene la repetición de "n".

- Longitud del solapamiento: aunque se podría hacer una división sin considerar una longitud de solapamiento, es

Conjunto de MEMs parciales.