Universidad de Concepción
Dirección de Postgrado
Facultad de Ingeniería - Programa de Magíster en Ciencias de la
Ingeniería con mención en Ingeniería Eléctrica

Acelerador hardware para búsqueda de motivos emergentes en streams de secuencias de ADN

Tesis para optar al grado de Magíster en Ciencias de la Ingeniería con mención en Ingeniería Eléctrica

ANTONIO SEBASTIÁN SAAVEDRA MONDACA
CONCEPCIÓN-CHILE
2018

Profesor Guía: Dr. Miguel Figueroa T.
Profesor Co-guía: Dra. Cecilia Hernández R.
Comisión 1: Dr. Mario Medina C.
Comisión 2: Dr. Gonzalo Carvajal B.
Dpto. de Ingeniería Eléctrica, Facultad de Ingeniería
Universidad de Concepción

Resumen

El descubrimiento de motivos en cadenas de ADN se define como la búsqueda de secuencias cortas de elementos compartidos en un conjunto largo de bases de nucleótidos que poseen una función biológica común. El descubrimiento de motivos entre los sitios de unión de los factores de transcripción, debido a la importancia de su función regulatoria en la expresión genética, resulta un problema de relevancia biológica. Este tipo de problemas presenta una alta complejidad computacional, especialmente debido a las dificultad de trabajar con bases de datos masivas. Las soluciones existentes en este tipo de problema se enfocan, por lo general, a plataformas en grandes clusters de alto costo, elevados tiempos de ejecución y consumo de potencia.

En este trabajo se desarrolla un acelerador hardware reconfigurable para la búsqueda de motivos emergentes en secuencias de ADN. Los motivos emergentes se definen como aquellos que cumplen requisitos establecidos de frecuencia dentro de la secuencias analizadas. Su búsqueda representa un problema biológicamente relevante que presenta altos requisitos de memoria y costos computacionales. La plataforma se propone en base a algoritmos capaces de resolver el problema de la búsqueda de elementos más frecuentes dentro de un stream de datos. Estos algoritmos utilizan estructuras de datos conocidas como sketches para realizar una aproximación al proceso de conteo para determinar los elementos más frecuentes. A diferencia de un conteo tradicional, la utilización de sketches permite resolver, a través de procesos probabilísticos, en espacio sublineal, la estimación de la frecuencia de cada elemento del stream.

Se implementaron en software los algoritmos CountSketch, Countmin, y Countmin-CU. Utilizando bases de datos biológicas públicas, se analizaron las dimensiones requeridas para operar con buena precisión y sensibilidad. El algoritmo Countmin-CU es capaz de encontrar los motivos emergentes de largos entre 10 y 20 utilizando arreglos de 65 mil contadores. El conteo tradicional requeriría sobre 100 mil millones. Se diseñó una arquitectura hardware dedicada que permite utilizar un FPGA como acelerador en un contexto de computación heterogénea. El algoritmo de streaming logra un balance adecuado entre el cómputo y los accesos requeridos a memoria permitiendo explotar el paralelismo fino de este tipo de plataforma. De esta manera, la lógica programable del FPGA con un diseño especializado nos permite reducir los costos de tiempo y el consumo de potencia de la solución. Este modelo de computación acelerada por hardware, con el FPGA nos permite trabajando con un reloj de 300MHz y consumiendo 3 Watts de potencia, nos permite alcanzar una aceleración de hasta 290 veces sobre la versión en software.