Resumen

Durante los últimos años, los secuenciadores de ADN han sido mejorados en velocidad y costes de funcionamiento, generando una avalancha de datos genómicos. Esto ha fomentado la mejora y paralelización de los algoritmos de alineamiento, buscando aprovechar los distintos entornos de computación de alto rendimiento.

En bioinformática, el término *alineamiento* se define como la comparación de dos lecturas de ADN, ARN o proteínas potencialmente diferentes. Esta comparación se hace en base a las relaciones entre sus nucleótidos: aciertos, fallos, inserciones y borrados. Más específicamente, cuando se comparan secuencias cortas se emplea el término *mapeo de secuencia*. En esta tesis se describen varios algoritmos para el mapeo inexacto de secuencias biológicas cortas, junto con su paralelización en entornos como GPGPU o memoria compartida.

Actualmente, los métodos de mapeo inexacto consisten en una combinación de técnicas de semilleo seguidas de técnicas de alineamiento local. Por un lado, los algoritmos de semilleo suelen basarse en técnicas de búsqueda hacia atrás, utilizando la transformada de Burrows-Wheeler, el índice de Ferragina y Manzini y matrices de sufijos para localizar las áreas donde podría alinearse una lectura. Por otro lado, los algoritmos de alineamiento local generan matrices de pesos usando programación dinámica, obteniendo así el alineamiento mejor puntuado de entre todas las áreas destacadas.

La tesis se enfoca en el estudio de los métodos de búsqueda hacia atrás. Concretamente, describimos las relaciones entre la transformada de Burrows-Wheeler, las matrices de sufijos y el FM-Index de un texto de referencia.

Dos algoritmos de búsqueda hacia atrás que usan el FM-Index se han paralelizado en GPGPUs. El primero permite mapeo exacto en GPUs y puede usarse para acelerar las técnicas de semilleo. El segundo es una implementación CPU-GPU híbrida, la cual permite mapeo inexacto con un error y devuelve los pares finales de una lectura. Los dos superan a las implementaciones existentes.

Además, se ha implementado un algoritmo de mapeo inexacto que permite cualquier número de diferencias. Dicho algoritmo combina búsqueda hacia atrás con técnicas de exploración de árboles de búsqueda, implementando estrategias de poda específicas para datos genómicos. Este nuevo método constituye la contribución más significativa de la tesis, alcanzando mayor sensibilidad y un speed-up de 7x respecto a algoritmos similares.

Finalmente, durante la estancia en Japón el algoritmo ha sido modificado para trabajar con un índice out-of-core. Dicho índice permite usar el algoritmo de mapeo inexacto con genomas grandes en sistemas con configuraciones de memoria primaria limitadas.