Summary

During the past years, DNA sequencers have been constantly improved in performance and operating costs, generating a genomic data deluge. This situation has fostered the general improvement and parallelisation of alignment algorithms, taking profit of different high performance environments.

In bioinformatics, the term *alignment* refers to the comparison of two potentially dissimilar reads of DNA, RNA or proteins. This comparison is made in terms of the relationships between its nucleotides: matches, mismatches, insertions and deletions. When aligning short reads, the more concrete term *sequence mapping* is employed. Several algorithms for the inexact mapping of short biological sequences are presented in this thesis, along with its parallelisation in environments like GPGPU, distributed memory and shared memory.

Currently, inexact mapping methods consist on a combination of seeding techniques followed by local alignment techniques. On the one hand, seeding algorithms are usually based on backward search methods, using the Burrows-Wheeler Transform, the Ferragina and Manzini Index and Suffix Arrays to locate the alignment candidate areas of a read. On the other hand, local alignment algorithms generate matrices of weights using dynamic programming, obtaining the best scoring alignment among the candidate areas.

First of all, the thesis presents an study of the backward search methods. Concretely, we describe the relationships between the Burrows-Wheeler Transform and the Suffix Array of a reference text. We also describe the mechanics of the FM-Index that enable backward search.

Secondly, two backward search algorithms using the FM-Index have been parallelised using GPGPUs in this thesis. The first one covers exact mapping on GPUs. It can be used to accelerate seeding techniques. The second one is an hybrid CPU-GPU implementation, which performs inexact mapping with one error and returns the pair-ends of a read. Both approaches outperform existing implementations.

Thirdly, an inexact mapping algorithm supporting any number of differences has been implemented. Such algorithm combines backward search with search tree exploration techniques, implementing pruning strategies specifically suited for genomic data. This new approach constitutes the most significant contribution of this thesis, achieving higher sensitivity and a 7x speed-up over similar algorithms. In addition, an out-of-core index has been implemented for employing this approach with large genomes on systems without expensive primary memory configurations.

Finally, the thesis formulates a formal specification of the local alignment problem using list homomorphisms and semiring structures. Along with this specification we employ the Generate, Test and Aggregate (GTA) framework, which allows to automatically derive pure functional algorithms. This implementation is excellent for distributed memory environments and, also, reaches a superlinear speed-up in shared memory multi-core machines.

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, memoria distribuida o 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.

En primer lugar, la tesis presenta un estudio los métodos de búsqueda hacia atrás. Concretamente, describimos la relación entre la transformada de Burrows-Wheeler y las matrices de sufijos de un texto de referencia. También describimos las mecánicas del FM-Index que permiten realizar búsqueda hacia atrás.

En segundo lugar, 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.

En tercer lugar, 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. Además, se ha implementado un índice out-of-core para trabajar con genomas grandes en sistemas con configuraciones de memoria primaria limitadas.

Finalmente, la tesis formula una especificación formal del problema del alineamiento local usando listas homomórficas y semianillos. Junto a esta especificación, empleamos el marco de trabajo Generar, Testear y Agregar, el cual permite derivar automáticamente algoritmos puramente funcionales. Esta implementación es excelente para entornos de memoria distribuida, alcanzando un speed-up superlineal en maquinas multiprocesador de memoria compartida.

Resum

Durant els últims anys, els seqüenciadors d'ADN han estat millorats en velocitat i costos de funcionament, generant una allau de dades genòmiques. Això ha fomentat la millora i paral·lelització dels algorismes d'alineament, buscant aprofitar els diferents entorns de computació d'alt rendiment.

En bioinformàtica, el terme *alineament* es defineix com la comparació de dues lectures d'ADN, ARN o proteïnes potencialment diferents. Aquesta comparació es fa d'acord amb les relacions entre els seus nucleòtids: encerts, errors, insercions i esborrats. Més específicament, quan es comparen seqüències curtes s'empra el terme *mapatge de seqüència*.

En aquesta tesi es descriuen diversos algorismes per al mapatge inexacte de seqüències biològiques curtes, amb la seua paral·lelització en entorns com GPGPU, memòria distribuïda o compartida.

Actualment, els mètodes de mapatge inexacte consisteixen en una combinació de tècniques de cerca de llavors seguides de tècniques d'alineament local. D'una banda, els algorismes de cerca de llavors solen basar-se en tècniques de recerca cap enrere, utilitzant la transformada de Burrows-Wheeler, l'índex de Ferragina i Manzini i matrius de sufixos per localitzar les àrees on podria alinear-se una lectura. D'altra banda, els algorismes d'alineament local generen matrius de pesos usant programació dinàmica, obtenint així l'alineament millor puntuat d'entre totes les àrees destacades.

En primer lloc, la tesi presenta un estudi dels mètodes de recerca cap enrere. Concretament, descrivim la relació entre la transformada de Burrows-Wheeler i les matrius de sufixos d'un text de referència. També descrivim les mecàniques del FM-Index que permeten realitzar recerca cap enrere.

En segon lloc, dos algorismes de recerca cap enrere que usen el FM-Index s'hi han paral·lelitzat en GPGPUs. El primer permet mapatge exacte en GPUs i pot usar-se per accelerar les tècniques de cerca de llavors. El segon és una implementació CPU-GPU híbrida que permet mapeig inexacte amb un error i retorna els parells finals d'una lectura. Els dos superen les implementacions existents.

En tercer lloc, s'ha implementat un algorisme de mapatge inexacte que permet qualsevol nombre de diferències. L'algorisme combina recerca cap enrere amb tècniques d'exploració d'arbres de cerca, implementant estratègies de poda específiques per a dades genòmiques. Aquest nou mètode és la contribució més significativa de la tesi, aconseguint major sensibilitat i un speed-up de 7x respecte a algorismes similars. A més, s'ha implementat un índex out-of-core per treballar amb genomes grans en sistemes amb configuracions de memòria primària limitades.

Finalment, la tesi formula una especificació formal del problema de l'alineament local usant llistes homomórfiques i semianells. A més d'aquesta especificació, fem servir el marc de treball Generar, Testejar i Afegir, el qual permet derivar automàticament algorismes purament funcionals. Aquesta implementació és excel·lent per a entorns de memòria distribuïda, assolint un speed-up superlineal en màquines multiprocessador de memòria compartida.