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 o memòria 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.

La tesi s'enfoca en l'estudi dels mètodes de recerca cap enrere. Concretament, descrivim la relació entre la transformada de Burrows-Wheeler, les matrius de sufixos i el FM-Index d'un text de referència.

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.

A més, 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.

Finalment, durant l'estada al Japó l'algorisme ha estat modificat per treballar amb un índex out-of-core. Aquest índex permet usar l'algorisme de mapeig inexacte amb genomes grans en sistemes amb configuracions de memòria primària limitades.