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 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.

This thesis focuses in backward search methods. Concretely, we describe the relationships between the Burrows-Wheeler Transform, the Suffix Array and the FM-Index of a reference text.

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.

Also, 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. This speed-up has been achieved without employing parallelism techniques.

Finally, during the internship in Japan the algorithm has been modified to support an out-of-core index. This index allows to use the inexact mapping algorithm with large genomes on systems without expensive primary memory configurations.