Long Read Alignment with Parallel MapReduce Cloud Platform (Summary)

The existing sequence aligners mostly support short read genomic sequences and lack support in cloud environments. They exhibit deficiencies in the alignment of long sequence genomic data that are currently generated using NGS technologies. The existing long read aligners that adopt the cloud platform for computation suffer from certain disadvantages. Thus, a cloud infrastructure and the MapReduce framework are combined together as a solution to support long read sequence alignment.

A dual phase execution is used in the MapReduce model; in the first phase, input data is split into fragments (associated with a mapper providing key-value pairs as outputs) whereby reduce workers are provided – key-sorted(value) pairs – to store the results in the Hadoop files system. The workers are usually virtual machines in public cloud environments.

The Burrows-Wheeler Aligner's Smith-Waterman Alignment on Parallel MapReduce (BWASW-PMR) cloud platform for long sequence alignment is implemented to solve the problem of serial execution for the map/reduce phases. The main developments for this long sequence alignment strategy are: the optimisation of SW in the BWA-SW alignment, a custom MapReduce framework to support the required computations, a parallel map and reduce workers execution strategy and a parallel execution of the map and reduce functions at worker nodes.

BWA-SW Alignment relies on the SW algorithm to align the seed matches of sequences using a similarity matrix and backtracking algorithm. The BWA-SW algorithm constructs a full-text index using FM-indexing of the query and reference sequences whereby a prefix directed acyclic word graph and a prefix trie are built respectively with the aid of suffix arrays and their intervals along with a dynamic programming mechanism. To optimize these computations, a reverse post-order traversal scheme is implemented.

The BWASW-PMR uses the MapReduce computation model for cloud computation whereby map and reduce worker nodes are deployed on a cloud cluster consisting of VMs. It considers the genomic sequence alignment in dual phases – map and reduce. In Hadoop, the map phase is executed and then the reduce phase is initiated. A parallel execution strategy of the two phases is considered to overcome Hadoop's disadvantages; i.e. phases are modelled to run in parallel utilizing all programming cores available in the worker VMs.

The optimization of SW is achieved using a wave front parallelization technique. Execution time of the optimised SW is significantly lower than the standard SW. Comparing BWASW-PMR Cloud and Bwasw-Cloud single computing node, the BWASW-PMR aligner showcases a significant speed-up. Comparing BWASW-PMR Cloud and Bwasw-Cloud on Azure, BWASW-PMR exhibits lower makespan time and long sequence alignment is faster.

The results obtained indicate significant improvement and is thus of use to the genomic community to support the required computations for long sequence alignment efficiently. The parallel executions of the map and reduce phases along with SW optimization are the main contributing factors for these experimental results.

Future undertakings include: optimisation of the BWA-SW algorithm as it uses a lot of memory; accelerating the BWA-MEM algorithm of Burrows Wheeler aligner on different platforms.