Optimization of Rabin Karp Pattern Matching Algorithm Based on Parallel Computing Techniques for DNA Sequence Analysis

M.G.M. Anjalee, W.P.U. Fernando and D.R.V.L.B. Thambawita

Department of Computer Science and Technology, Uva Wellassa University, Badulla, Sri Lanka

String matching algorithms are used to discover the occurrences of a defined pattern in a given text or a pool of strings which is widely used in detecting plagiarism, spam filtering and most importantly in computational biology including DNA sequencing. The existence and the intensity of a muted sequence in DNA caused for various diseases can be identified using Rabin Karp string matching algorithm. The main contribution of the study is to bring an efficient version of Rabin Karp algorithm by minimizing the spurious hits while using both Central Processing Unit (CPU) parallel techniques and General Purpose Graphics Processing Unit (GPGPU) parallel techniques specifically for DNA sequence analysis. The improved Rabin Karp is implemented using C language with POSIX Threads library, OpenMP and MPI and using Compute Unified Device Architecture (CUDA). When accelerating computations based on GPU, a special consideration has given to global memory, shared memory and texture memory, the types of memories with particular importance offered in CUDA architecture. By experimental studies, we investigated a new method to eliminate brute force matching and the GPU optimization is presented with stencil method ensuring efficiency in terms of memory overhead due to redundant data access in the serial CPU implementation. We have compared these parallel implementations for evaluating the effect of varying number of threads per block as well as varying DNA file sizes. The results obtained in this study present that the proposed implementation provides acceleration surpassing 36x speedup for string size 2^{20} characters compared to a sequential (CPU) implementation. Eventually, using the empirical results, we could conclude that the improved CUDA C implementation of shared memory version can achieve 35 times of performance than serial implementation for a large pool of DNA data in string matching.

Keywords: Rabin karp, GPU, CUDA, Pthread, String matching