Application of GPGPU for Acceleration of Short DNA Sequence Alignment in Unipro UGENE Project

Konstantin Okonechinkov*, German Grekhov, Konstantin Stepanyuk, Mikhail Fursov*
Unipro Center Of Information Technologies
{kokonech, ggrekhov, const24, mfursov} @unipro.ru

Address: Unipro, 6/1 Lavrentiev Avenue, 630090, Novosibirsk, Russia
Web site: http//ugene.unipro.ru
* Corresponding authors

Abstract

A dramatic increase of available sequencing datasets has resulted in the need of fast sequence alignment methods. Plenty of novel methods were proposed to perform the fast alignment of NGS data and some of them appeared to be rather effective, however a relatively small number of existing alignment tools use Graphic Processing Units (GPUs) to speed up the alignment procedure. Unfortunately these tools are available only as source code packages and have very limited support for existing biological data formats.

In this report we describe UGENE Genome Aligner, an efficient GPU-accelerated tool for fast short reads alignment. This tool is created as an extension to Unipro UGENE bioinformatics toolkit, a popular open-source software package for molecular biologists. The performance benchmarks of the Genome Aligner demonstrated a 2x speedup in comparison with existing tools. The relative performance of the GPU-accelerated Genome Aligner was measured by comparison with the CPU version of the algorithm. In result the GPU-optimized search procedure showed ~6x speedup over the CPU version.

The GPU-accelerated Genome Aligner is fully integrated into UGENE framework and freely available for download. The integration of the Genome Aligner allows it to be easily included into any research pipeline provided by UGENE and makes it available "out-of-the-box" for end-user without any additional efforts.

Introduction

A recent dramatic increase of available sequencing datasets catalyzed by improvements in high-throughput sequencing technologies has resulted in the need of fast sequence alignment methods. Moreover these methods should allow the rapid processing of billions of short DNA sequences (short reads). The requirement of speed has become especially relevant in the context of whole human genome resequencing.

Plenty of novel methods were proposed to perform the fast alignment of NGS data and some of them appeared to be rather effective [2, 3, 4]. Since the alignment of short reads implies simple data-parallelism many of available methods make use of multi-core and multiprocessor systems to work faster [2, 5]. A relatively small number of existing alignment tools use Graphic Processing Units (GPUs) to speed up the alignment procedure [6, 7]. Unfortunately these tools are available only as source code packages and have very limited support for existing biological data formats, which makes it difficult to use them in every-day research without additional efforts.

In this paper we describe UGENE Genome Aligner, an efficient GPU-accelerated tool for fast short reads alignment. This tool is created as an extension to Unipro UGENE bioinformatics toolkit. Unipro UGENE [1] is an open-source software package for molecular biologists; its main goal is to integrate popular bioinformatics tools and algorithms within a single flexible user interface. UGENE includes multiple bioinformatics algorithms and supports a great majority of biological data formats. Some of the included algorithms are optimized for a multi-core environment and have GPU implementations. One great UGENE advantage is that the included GPU-optimizations are available "out-of-the box" for the end-user and works on any system with compatible hardware.

UGENE Genome Aligner Description

UGENE Genome Aligner uses a simple but efficient method for short reads alignment to a reference sequence based on the application of suffix arrays [8]. The algorithm starts with building a suffix array for the reference sequence. Similar to most of the existing aligners the suffix array is used to find the seed portion of a short read. To optimize the searching, each suffix array entry is encoded into a 64-bit integer value. The encoding uses 2 bits to represent each DNA nucleotide and allows searching for seeds up to 31 base pair size. In the context of existing NGS instruments this limitation doesn't have much impact because of the typically small size of short reads. During the alignment procedure each seed is encoded using the above-mentioned technique and then a standard binary search procedure is applied to find the seed in the suffix array. After the seed is found a refining comparison of a short read and corresponding reference sequence region might be performed based on algorithm settings. The UGENE Genome Aligner allows up to 3 mismatches in result alignment and can report either all available short read alignments or only the best alignment in terms of mismatch count.

The proposed algorithm has several minor drawbacks such as limited ability to work with non-standard nucleotide codes and small allowed substitution percentage. However despite its pitfalls the Genome Aligner almost achieved the same efficiency as modern alignment tools. We tested the accuracy and performance of the Genome Aligner against Bowtie (Table 1). In this test we aligned approximately 2 million randomly selected short reads with size of 76 bp (experiment ERR008366 from NCBI Short Reads Archive[12]) to the mouse genome chromosome 1 (GenbankID:AC157543.8). As a result Bowtie was faster than the Genome Aligner in the cases of 0 and 1 permitted mismatch, but a performance boost was demonstrated by the Genome Aligner in the case of 2 and 3 mismatches. Both tools demonstrated equal accuracy.

Number of allowed mismatches	Bowtie time, sec	Bowtie, number of reads aligned in percents	Genome Aligner time, sec	Genome Aligner, number of reads aligned in percents
0	4	4.75	15	4.75
1	6	7.04	18	7.04
2	38	8.55	30	8.55
3	135	9.74	60	9.74

Table 1 Performance comparison of Bowtie and UGENE Genome Aligner. Test machine configuration: Intel Q9550 2.7 GHz, 8 Gb Ram. Bowtie is launched with the following parameters: *-p 4 --norc -v X,* the Genome Aligner had the following parameters: *--best --n-mis=X.*

Parallel Binary Search on GPU

The general purpose application of GPUs proved to be successful in multiple areas. An increasing availability of powerful GPUs such as NVIDIA Tesla allows for a typical PC workstation to solve complex computational problems; therefore the utilization of GPU in UGENE Genome Aligner was a task of high importance for us. We started analyzing the possibility to use GPU for searching the seed portions of short reads in the suffix array. The search procedure typically takes from 0.3 to 0.6 of the entire alignment time, the result being a significant increase in performance.

The suffix array in our case is represented as a sorted array of integer values therefore the search task can be described as a problem of locating a position of a specific item in a sorted array. Since we are searching for multiple independent items it is possible to solve this problem by applying binary search algorithm in parallel. There were some recent attempts to implement parallel binary search on GPU which resulted in significant performance increase. For example, the so called "p-ary" search, discussed in the paper by T. Kaldewey and others [9], achieved 8x speedup over the CPU binary search.

Taking into account existing investigations we implemented parallel binary search procedure using both NVIDIA CUDA toolkit [10] and OpenCL [11]. Performance tests showed the expected 8x speedup over CPU. It is worth mentioning that our implementation includes a number of optimizations, which are targeted on minimizing global GPU memory access.

Results

The relative performance of the GPU-accelerated Genome Aligner was measured by comparison with the CPU version of the algorithm. We measured the runtime of search in the suffix array procedure and total application runtime separately. Our test stand was based on Intel Core2Quad Q9550 processor, 8 GB RAM with NVIDIA Tesla 1060 on board. The CPU-version of the algorithm made use of multiple processor cores implementing several binary searches in parallel similar to the GPU-version.

In our performance tests we used synthetic reads generated from the complete genomes of *Anopheles gambiae* (GenbankID: NC_004818.2) and Mouse Genome Chromosome 1 and 2 (AC157543.8). For each reference sequence we simulated 2000000-length sets of short reads. Each set included only reads with the size of 35 bp, 50 bp and 100 bp correspondingly (Table 2).

Species	Общее время сборки для CPU, сек	CPU-time (search in suffix array), сек	Overall alignment GPU-time, sec	GPU- time (search in suffix array), sec
Anopheles gambiae	112	45	85	8
Mus Musculus Chromosome 1	164	50	103	8
Mus Musculus Chromosome 2	181	68	110	13

Table 2 Performance Comparison of CPU and GPU implementations of the Genome Aligner. Test machine configuration: Intel Q9550 2.7 GHz, 8 Gb Ram, NVidia Tesla C1060. The Genome Aligner had the following parameters: *--best --n-mis=3.*

In each performance test the GPU-optimized search procedure showed a significant speedup over the CPU version. However in general the GPU-optimized Genome Aligner was only ~1.5x faster than the CPU version of the algorithm. The whole alignment procedure has a number of other steps other than the search procedure: building a suffix array, performing a refining comparison, working with I/O, and each of these steps can require a significant amount of time. We strongly believe that there is potential for further performance improvement. First, our implementation of parallel binary search on GPU can be further optimized in order to assure that global GPU memory accesses are coalesced etc. Second, we are planning to carefully analyze and rework the algorithm in order to port the refining comparison step and building a suffix array step to GPU and also minimize the I/O impact.

The current implementation of the GPU-accelerated Genome Aligner is fully integrated into UGENE and freely available for download. Even in its current implementation the GPU-optimized Genome Aligner is very efficient when performing alignment of large NGS datasets. The integration of the Genome Aligner allows it to be easily included into any research pipeline provided by UGENE framework and makes it available for end-user without any additional efforts.

References

1. **Unipro UGENE** (http://ugene.unipro.ru)
2. Langmead B, Trapnell C, Pop M, Salzberg SL. ”Ultrafast and memory-efficient alignment of short DNA sequences to the human genome.” Genome Biol 2009 10:R25.
3. Homer N, Merriman B, Nelson SF. “BFAST: an alignment tool for large scale genome resequencing.” PLoS One. 2009 Nov 11;4(11):e7767.
4. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, Wang J. “SOAP2: an improved ultrafast tool for short read alignment.”, Bioinformatics. 2009 Aug 1;25(15):1966-7. Epub 2009 Jun 3.
5. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. “Searching for SNPs with cloud computing.”, Genome Biol. 2009;10(11):R134. Epub 2009 Nov 20.
6. Schatz M, Trapnell C, Delcher A, Varshney A “High-throughput sequence alignment using Graphics Processing Units”, BMC Bioinformatics 2007, 8:474
7. Gharaibeh A., Ripeanu M. "Size Matters: Space/Time Tradeoffs to Improve GPGPU Applications Performance, , IEEE/ACM International Conference for High Performance Computing, Networking, Storage, and Analysis (SC 2010), New Orleans, LA, November 2010.
8. Gusfield D. “Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology”, Cambridge University Press, 1997
9. Kaldewey T., Hagen J, Di Blas A., Sedlar E. “Parallel Search On Video Cards”, Oracle Server Technologies - Special Projects
10. NVIDIA CUDA toolkit, http://www.nvidia.com/object/cuda_home_new.html
11. OpenCL, http://www.khronos.org/opencl/
12. NCBI Short Reads archive, http://www.ncbi.nlm.nih.gov/sra