BIOINFORMATICS APPLICATIONS NOTE

Vol. 28 no. 6 2012, pages 878–879
doi:10.1093/bioinformatics/bts061

Sequence analysis

Advance Access publication January 28, 2012

SOAP3: ultra-fast GPU-based parallel alignment tool for short reads

Chi-Man Liu[1,†], Thomas Wong[1,†], Edward Wu[1], Ruibang Luo[1], Siu-Ming Yiu[1], Yingrui Li[2], Bingqiang Wang[2], Chang Yu[2], Xiaowen Chu[3], Kaiyong Zhao[3], Ruiqiang Li[4,*], Tak-Wah Lam[1,*]

[1]Department of Computer Science, The University of Hong Kong, Hong Kong, [2]BGI Shenzhen, China, [3]Department of Computer Science, Hong Kong Baptist University, Hong Kong and [4]Peking-Tsinghua Center for Life Sciences, Biodynamics Optical Imaging Center and School of Life Sciences, Peking University, Beijing, China

Associate Editor: John Quackenbush

ABSTRACT

Summary: SOAP3 is the first short read alignment tool that leverages the multi-processors in a graphic processing unit (GPU) to achieve a drastic improvement in speed. We adapted the compressed full-text index (BWT) used by SOAP2 in view of the advantages and disadvantages of GPU. When tested with millions of Illumina Hiseq 2000 length-100 bp reads, SOAP3 takes < 30 s to align a million read pairs onto the human reference genome and is at least 7.5 and 20 times faster than BWA and Bowtie, respectively. For aligning reads with up to four mismatches, SOAP3 aligns slightly more reads than BWA and Bowtie; this is because SOAP3, unlike BWA and Bowtie, is not heuristic-based and always reports all answers.

Availability: SOAP3 is available at: http://www.cs.hku.hk/2bwt-tools/soap3; http://soap.genomics.org.cn/soap3.html.

Contact: liruiqiang@gmail.com, twlam@cs.hku.hk

Received on November 17, 2011; revised on January 23, 2012; accepted on January 25, 2012

Recent sequencing technologies are able to generate a large volume of reads in a fast and low-cost manner. A single sequencer (e.g. Illumina HiSeq 2000) can generate 600 million pair-end reads of length 100 in 10 days. Large genome centers can afford to have tens to over a hundred of sequencers, providing a cost-effective high-throughput platform for generating sufficient reads for many exciting biological applications (e.g. mapping DNA–protein interactions, whole-transcriptome sequencing and whole genome expression profiling). Most of these applications require the mapping of the reads onto a reference genome as the first step, followed by various downstream analyses. Thus, an extremely fast alignment tool is needed. As the reads are longer, we need alignment that can allow three or more mismatches.

There are quite a number of software tools for aligning short reads onto a reference genome. The most popular ones are MAQ (Li *et al.*, 2008a), SOAP2 (Li *et al.*, 2008b, 2009), Stampy (Lunter and Goodson, 2011), Bowtie (Langmead *et al.*, 2009), BWA (Li and Durbin, 2009). Refer to (Blom *et al.*, 2011; Li and Homer, 2010) for two recent surveys about these tools. Using the human genome as the reference, aligning 70 million read pairs (equivalent to the throughput of one lane of the Illumina HiSeq 2000) with at most four mismatches takes > 3.5 h using the fastest existing aligner. To align 1G read pairs (100 Gb sequences, about 30× coverage for a human genome), it takes > 2 days to complete the alignment step. To further reduce the time substantially using a single CPU seems to be very difficult.

In this note, we present the first short read alignment tool SOAP3 that leverages the multi-processors in a graphic processing unit (GPU) to achieve a drastic improvement in speed. We developed a GPU version of the compressed full-text indexing data structure used by SOAP2, which is based on BWT. The novelty of SOAP3 stems from two aspects. BWT is a sophisticated compressed index. Pattern searching using BWT requires many random memory access. A direct implementation of the CPU version on GPU would induce heavy memory contention among different threads and degrade the overall performance drastically. We solved the problem by redesigning the data structure to reduce memory accesses as much as possible, while retaining the efficiency of the index. The other difficulty is that GPU works in a single-instruction multiple-thread (SIMT) mode. Processors in the same unit [called streaming multiprocessor (SM)] must execute the same instruction. Too many diverging branches in the execution path would force some of the processors to idle. However, how many diverging branches a pattern may introduce cannot be determined until runtime. We derive a useful parameter to determine in runtime whether a pattern would introduce too many branches (called *hard* patterns). We stop the execution of hard patterns, group them and re-do the alignment of them in another round to reduce the idle time of processors.

The current version of SOAP3 can support alignment with up to four mismatches. We evaluated the performance of SOAP3 on two real datasets with human reference genome build 37.1 as the reference, and compared it to BWA and Bowtie. The evaluation was conducted on a computer with a 3.07 GHz quad-core CPU and 24 G memory. SOAP3 is supported by a NVIDIA GTX 580 GPU card with 3G memory.[1] We have chosen two datasets with different quality. The first one contains 70.7 M read pairs, sequenced from YH1 Cell-line DNA using Illumina HiSeq 2000 (Wang *et al.*, 2008;

*To whom correspondence should be addressed.

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

[1]SOAP3 requires a CUDA-enabled GPU with at most 2.5 GB memory for indexing a human genome. SOAP3 has also been tested with NVIDIA Tesla C2070 and M2050; see SOAP3 website for a list of GPUs supported.

Table 1. Running time and percentage of aligned reads of SOAP3, BWA and Bowtie on two paired-end datasets with 70.7 M (HiSeq 2000) and 25.3 M (Genome Analyzer II) read pairs

Dataset	Four mismatches		Three mismatches		Two mismatches		One mismatch		Zero mismatch	
	Time (s)	%	Time (s)	%	Time (s)	%	Time (s)	%	Time (s)	%
70.7 M read pairs										
SOAP3	1839	81.46	1019	79.43	695	76.48	521	70.94	452	53.11
BWA	13756	81.03	10590	79.19	8920	76.38	6064	70.90	5272	53.11
Bowtie	not supported		29178	79.42	2082	76.47	1570	70.93	1216	53.10
25.3 M read pairs										
SOAP3	735	86.58	453	85.34	356	83.46	291	79.17	266	60.30
BWA	4700	86.17	3803	85.12	3298	83.38	2352	79.17	1800	60.30
Bowtie	not supported		9486	85.33	1431	83.45	617	79.17	484	60.13

Each time reported below includes index loading time, read loading time (259 s for 70.7 M; 107 s for 25.3 M) and the alignment time. For BWA, we opt for faster speed by disabling the gapped alignment.

Table 2. Breakdown of running time and percentage of aligned reads of SOAP3 and SOAP2 on the above two datasets

	SOAP3		SOAP2	
	Time (s)	%	Time (s)	%
Index loading	133		40	
HiSeq dataset: 70.7 M				
Read loading	259		259	
Four-mismatch alignment	1447	81.46	12206	77.25
Three-mismatch alignment	626	79.43	12198	76.65
Two-mismatch alignment	303	76.48	4370	76.48
G. A. II dataset: 25.3 M				
Read loading	107		107	
Four-mismatch alignment	294	86.58	3495	84.13
Three-mismatch alignment	212	85.34	3453	83.63
Two-mismatch alignment	356	83.46	1465	83.46

http://yh.genomics.org.cn). The second dataset contains 25.3 M read pairs, sequenced by Illumina Genome Analyzer II (NCBI SRA, SRR211279). Both datasets have read length 100. Table 1 shows the results on finding a best alignment with different number of mismatches allowed. The time reported in each case includes the loading time of the index and reads, and the alignment time.

As shown in Table 1, SOAP3 is much faster than BWA and Bowtie. When aligning with up to three or four mismatches, the HiSeq 2000 dataset and Genome Analyzer II dataset reveal that SOAP3 is at least 7.5 times and 6.4 times times faster than BWA, respectively; and Bowtie is the slowest in this setting. It is worth-mentioning that SOAP3 favors large dataset as it takes longer time to load the index. For aligning with zero to two mismatches, Bowtie is faster than BWA, and SOAP3 is 2–6 times faster than Bowtie. Notice that when aligning with zero or one mismatch, SOAP3's running time is dominated by the loading time of the reads (e.g. 259 s for 70.7 M); thus the speed-up gained by SOAP3 does not look as significant as it should be. BWA and Bowtie are heuristics based, while SOAP3 always reports all alignments, and our experiments reveal that SOAP3 aligns slightly more reads than BWA and Bowtie.

We have further conducted a more detailed comparison of SOAP3 with its predecessor SOAP2. SOAP3 has a larger index and requires more time to load the index. Yet SOAP3 can often align up to 10 times faster than SOAP2 (see Table 2). Thus, the effect of larger index loading time becomes less significant when aligning multi-millions of reads. SOAP2 finds all alignments for up to two mismatches, but it becomes heuristics-based for aligning with three or four mismatches and aligns fewer reads than SOAP3. Similar to SOAP2, the output formats of SOAP3 include text and SAM/BAM format.

We are in the process of enhancing SOAP3 with GPU-based dynamic programming so as to report alignments with indels and gaps; the preliminary results show that the percentage of aligned reads could be improved by 5–8%.

ACKNOWLEDGEMENTS

We would like to thank the users of SOAP3 (in particular, BGI and NIH) for their support and feedbacks.

Conflict of Interest: none declared.

REFERENCES

Blom,J. *et al.* (2011) Exact and complete short read alignment to microbial genomes using gpu programming. *Bioinformatics*. **27**, 1351–1358.

Langmead,B. *et al.* (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, **10**, R25.

Li,H. and Durbin,R. (2009) Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, **25**, 1754–1760.

Li,H. and Homer,N. (2010) A survey of sequence alignment algorithms for next-generation sequencing. *Brief. bioinform.*, **II**, 473–483.

Li,H. *et al.* (2008a) Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res.*, **18**, 1851–1858.

Li,R. *et al.* (2008b) SOAP: short oligonucleotide alignment program. *Bioinformatics*, **24**, 713–714.

Li,R. *et al.* (2009) Soap2: an improved ultrafast tool for short read alignment. *Bioinformatics*, **25**, 1966–1967.

Lunter,G. and Goodson,M. (2011) Stampy: a statistical algorithm for sensitive and fast mapping of illumina sequence reads. *Genome Res.*, **21**, 936–939.

Wang,J. *et al.* (2008) The diploid genome sequence of an Asian individual. *Nature*, **456**, 60–65.