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

Chi-Man Liu1†, Thomas Wong1†, Edward Wu1, Ruibang Luo1, Siu-Ming Yiu1, Yingrui Li2, Bingqiang Wang2, Chang Yu2, Xiaowen Chu2, Kaiyong Zhao3, Ruiqiang Li4†, Tak-Wah Lam1∥

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

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.


Contact: linuqiang@gmail.com, taml@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. 2008], SOAP2 [Li et al. 2008, 2009], Stampy [Li and Homer 2010], BWA [Li and Durbin 2009], Bowtie [Langmead et al. 2009], and Durbin [2009]. Refer to [Blom et al. 2011] and [Li and Homer 2011b] 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 1 G 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 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. 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]. SOAP3 requires a CUDA-enabled GPU with at least 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.

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

∥To whom correspondence should be addressed.

© The Author 2012. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com
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**


