Next generation read mapping on GPUs

Cole Trapnell

April 24, 2008

Project goal: Extend MUMmerGPU to report inexact alignments k differences between a reference molecule and many short sequencing reads.

Next generation sequencing technologies produce an enormous number of short sequencing reads. These 25-50 base pair (bp) reads are more difficult to use for de novo assembly, but the machines that produce them are very inexpensive to operate. These technologies are thus very attractive for resequencing and comparative genomics projects. A key first step in such a project is to map each read to a closely related *reference* genome. Reads that map to the genome with differences may reveal genomic variation between the reference and the *donor* of the reads. These differences may reveal information about the evolutionary history of both organisms, or in clinical settings, information about the health of the donor.

I recently co-authored MUMmerGPU, a program that reports all exact substring alignments between a reference and a set of short reads. [2] MUMmerGPU uses the grapics processing unit (GPU) in a desktop PC to align reads in parallel, and is thus up to four-fold faster than a high-end workstation on the same workload. In this project, I will implement an inexact alignment procedure to extend the exact substring alignments produced by mummer. The inexact procedure will implement a variant of the Landau-Vishkin algorithm for finding alignments with k differences. [1].

Landau-Vishkin has $O(mn)$ running time, where m is the length of the reference, but in practice, running time is $O(km)$. This is because the core loop in the algorithm is essentially a call to `strcmp`, which makes extremely good use of the cache on a modern microprocessor. In MUMmerGPU, the CPU is mostly idle, as we recently moved nearly all of the computation onto the GPU. Because Landau-Vishkin performs well in the presence of cache, and the GPU has very little cache, the CPU will run inexact alignment procedure. MUMmerGPU uses a streaming model in which batches of queries are aligned to a sliding window of the reference. When all queries have been processed against that window frame, the window slides down to the next section of reference. Since the queries are batched, MUMmerGPU can perform the exact matching on the GPU, and then extend those matches on the CPU while the GPU performs the exact matching for the next batch. Overlapping the two procedures should dramatically shorten the time to perform the full alignment.

MUMmerGPU targets users who wish to align a great many solexa reads, but who cannot afford to buy a machine with an enormous amount of RAM or processing power. As short read data becomes ubiquitous, this is likely to constitute a sizeable number of researchers. They will not be willing to wait for a week to align several million reads to a mammalian-sized genome. For this project, I will extend MUMmerGPU to align 8 million 32bp reads to the human genome with up 2 differences in less than three days.

References

[1] Gad M. Landau and Uzi Vishkin. Fast parallel and serial approximate string matching. *Journal of Algorithms*, 10:157–169, 1989.

[2] Michael C. Schatz, Cole Trapnell, Arthur L. Delcher, and Amitabh Varshney. High-throughput sequence alignment with graphics processing units. *BMC Bioinformatics*, 8(474), 2008.