GateKeeper: A New Hardware Architecture for Accelerating Pre-Alignment in DNA Short Read Mapping

Mohammed Alser[1], Hasan Hassan[2], Hongyi Xin[3], Oğuz Ergin[2], Onur Mutlu[4], and Can Alkan[1]

[1]Department of Computer Engineering, Bilkent University, 06800 Bilkent, Ankara, Turkey.
[2]TOBB University of Economics and Technology, Sogutozu Cad. 43, Sogutozu, Ankara, Turkey.
[3]Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213, USA.
[4]Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland.

Motivation:

Until today, it remains challenging to sequence the entire DNA molecule as a whole. In the era of high throughput DNA sequencing (HTS) technologies, genomes are sequenced relatively quickly but result in an excessive number of small DNA segments (called short reads and are about 75-300 basepairs long). Resulting reads do not have any information about which part of genome they come from; hence the biggest challenge in genome analysis is to determine the origin of each of the billions of short reads within a reference genome to construct the donor's complete genome. Identifying the potential origin of each read, called *alignment*, typically performed using quadratic-time dynamic programming algorithms. These optimal alignment algorithms are unavoidable and essential for providing accurate information about the quality of the alignment. In recent works [1-4], researchers observed that the majority of candidate locations in the reference genome do not align with a given read due to high dissimilarity. Calculating the alignment of such incorrect candidate locations wastes the execution time and incur significant computational burden. Therefore, it is crucial to develop a fast and effective heuristic method that can detect incorrect candidate locations and eliminate them before invoking computationally costly alignment algorithms.

Results:

We propose GateKeeper, a new hardware accelerator that functions as a *pre-alignment* step that quickly filters out most incorrect candidate locations. GateKeeper is the first design to accelerate pre-alignment using Field-Programmable Gate Arrays (FPGAs), which can perform pre-alignment much faster than software. When implemented on a single FPGA chip, GateKeeper maintains high accuracy (on average >96%) while providing, on average, 90-fold and 130-fold speedup over the state-of-the-art software pre-alignment techniques, Adjacency Filter and Shifted Hamming Distance (SHD), respectively. The addition of GateKeeper as a pre-alignment step can reduce the verification time of the mrFAST mapper by a factor of 10.

Availability:

GateKeeper is open-source and freely available online at https://github.com/BilkentCompGen/GateKeeper.

References:

1. Alser, M., et al., *GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping.* Bioinformatics, 2017. **33**(21): p. 3355-3363.
2. Xin, H., et al., *Shifted Hamming Distance: A Fast and Accurate SIMD-Friendly Filter to Accelerate Alignment Verification in Read Mapping.* Bioinformatics, 2015. **31**(10): p. 1553-1560.
3. Xin, H., et al., *Accelerating read mapping with FastHASH.* BMC genomics, 2013. **14**(Suppl 1): p. S13.
4. Kim, J., et al., *Genome Read In-Memory (GRIM) Filter: Fast Location Filtering in DNA Read Mapping using Emerging Memory Technologies*, Tech. rep. 2017. url: https://people. inf. ethz. ch/omutlu/pub/GRIM-genome-read-in-memoryfilter_psb17-poster. pdf (cit. on p. 29).