GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies

Jeremie Kim[1], Damla Senol[1], Hongyi Xin[2], Donghyuk Lee[1,3], Mohammed Alser[4], Hasan Hassan[5], Oguz Ergin[5], Can Alkan[4] and Onur Mutlu[6,1]

[1] ECE Department, Carnegie Mellon University
[2] CS Department, Carnegie Mellon University
[3] NVIDIA Research
[4] CE Department, Bilkent University
[5] CE Department, TOBB University of Economics and Technology
[6] CS Department, ETH Zürich

Seed location filtering is critical in DNA read mapping, a process where billions of DNA fragments (reads) sampled from a donor are mapped onto a reference genome in order to identify the genomic variants of the donor. State-of-the-art read mappers determine the original location of a read sequence within a reference genome in 3 generalized steps. A read mapper 1) quickly generates possible mapping locations for seeds (i.e., smaller segments) within a read, 2) extracts the reference sequence at each of the mapping locations, and 3) determines the similarity score between the read and its associated reference sequences with a computationally-expensive algorithm (i.e., sequence alignment). With the similarity scores across all possible locations, the read mapper can determine the original location of the read sequence. The differences between the read sequence and the matching reference sequence indicate the genomic variants of the donor, which can be further analyzed for preventative care or diagnosis.

A seed location filter (e.g., FastHASH [2], SHD [3], GateKeeper [4]) comes into play before sequence alignment (step 3) and reduces the number of unnecessary alignments. A seed location filter efficiently determines whether a candidate mapping location would result in an incorrect mapping before performing the computationally-expensive sequence alignment step for that location. In the ideal case, a seed location filter would discard *all* poorly matching locations prior to alignment such that there is no wasted computation on unnecessary alignments.

We propose a novel seed location filtering algorithm, GRIM-Filter, optimized to exploit 3D-stacked memory systems that integrate computation within a logic layer stacked under memory layers, to perform processing-in-memory (PIM). GRIM-Filter quickly filters seed locations by 1) introducing a new representation of coarse-grained segments of the reference genome, and 2) using massively-parallel in-memory operations to identify read presence within each coarse-grained segment. Our evaluations show that for a sequence alignment error tolerance of 0.05, GRIM-Filter 1) reduces the false negative rate of filtering by 5.59x--6.41x, compared to the best previous seed location filtering algorithm, and 2) provides an end-to-end read mapper speedup of 1.81x--3.65x, compared to a state-of-the-art read mapper employing the best previous seed location filtering algorithm [2].

This work will appear at the 16th Asia Pacific Bioinformatics Conference in January 2018 [1]. The preliminary version of the full article is at https://arxiv.org/pdf/1711.01177.pdf.

[1] Kim, Jeremie S, et al. "GRIM-Filter: Fast Seed Location Filtering in DNA Read Mapping Using Processing-in-Memory Technologies." *to appear in BMC Genomics* (2018).
[2] Xin, Hongyi, et al. “Accelerating read mapping with FastHASH.” *BMC Genomics* (2013).
[3] Xin, Hongyi, et al. “Shifted Hamming distance: a fast and accurate SIMD-friendly filter to accelerate alignment verification in read mapping.” *Bioinformatics* (2015).
[4] Alser, Mohammed, et al. "GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping." *Bioinformatics* (2017).