Accelerating DNA Long Read Mapping with Emerging Technologies

Roman Kaplan
Andrew and Erna Viterbi Faculty of Electrical Engineering
Technion, Israel Institute of Technology
Haifa, Israel
romankap@gmail.com

Poster Summary

Background and Motivation

Constructing human DNA sequence in real time is paramount to development of precision medicine [1] and on-site pathogen detection of disease outbreaks [2]. Single-molecule, real-time (SMRT) sequencing from Pacific Biosciensces [3] (PacBio) and Oxford Nanopore Technologies [4] (ONT) are new technologies that can produce long reads within minutes, potentially enabling real time genomic analysis. However, compared with the high-accuracy short reads from 2nd generation sequencers, SMRT sequencing poses new challenges. First, long reads contain many thousands of base pairs (bps). Second, long reads tend to exhibit about 15-20% insertion, deletion and substitution errors [3][4].

To construct a complete host sequence, in case a reference sequence exists (from a previously sequenced organism), long reads are mapped to high-similarity locations of the reference sequence. Determining the optimal mapping location of every read onto the reference sequence requires a computationally intensive local alignment procedure (e.g., Smith-Waterman [5]). Read mappers (e.g., minimap [6], minimap2 [7]) find regions of high similarity (overlaps) between reads or between a read and a reference sequence. Once a mapping exists, the alignment can be performed on a specific region of the reference, reducing its duration and resource requirements [8]. Therefore, read mapping can be viewed as a pre-alignment step that reduces the problem size for aligners by narrowing the regions to ones with potentially high-scoring alignment.

Existing pre-alignment hardware solutions [9],[10] target short reads (up to several hundreds bps). Such reads contain a small number of errors (less than 5%) and have a different error profile than that of PacBio or ONT long reads [3],[4]. However, current solutions [9] have high false positive rates when the number of errors is higher than 5%. Thus, the current solutions for short reads are not applicable for long reads.

Approximate computing techniques are known to trade accuracy for speed or energy efficiency. In case of long reads, multiple errors are a natural part of the sequencing output. Therefore, long read DNA mapping inherently tolerates the imprecision.

With the end of Dennard scaling and the slowdown of Moore's law, novel hardware solutions for data intensive problems are researched. Emerging technologies such as resistive memories enable new architectures with better performance and energy efficiency. Resistive approximate Hamming distance solutions exist [11]. However, these do not provide the parallelism required to support a high throughput applications such as DNA read mapping.

RASSA: Resistive Approximate Similarity Search Accelerator

This work presents RASSA, a Resistive Approximate Similarity Search Accelerator architecture for long read DNA mapping. RASSA is a massively parallel in-memory processor, facilitating simultaneous compare and mapping of a long read onto a reference sequence. The key performance breakthrough of RASSA is achieved by applying the similarity search in parallel to the entire reference.

RASSA employs resistive elements, memristors, serving at the same time as single bit storage elements and comparators. It allows storing and in-situ processing of large datasets. RASSA enables comparing a key pattern with the entire dataset in parallel. Every number of mismatches (of the key pattern vs. each data element that is in each memory row) causes a specific voltage drop, allowing quantifying the number of mismatching locations (called a *mismatch score*). Additional evaluation transistors translate mismatch scores into voltage levels, which are converted to digital values using analog to digital converters. The mismatch score is compared with a predefined threshold value to indicate the locations which have the desired degree of similarity with the compared pattern.

Long DNA reads are divided to fixed-size chunks that serve as key patterns to compare against the entire RASSA array. All locations with a mismatch score below the predefined threshold are marked by RASSA and indicate a valid mapping location.

RASSA performance is compared with two existing solutions. The first solution, minimap2 [7], a state-of-art read mapper that uses a SIMD extensions and multi-threading. Minimap2 was executed on a high-end server with 16-core Intel Xeon E5-2650 and 64GB of RAM. RASSA achieves 16-77× speedup over minimap2, with comparable accuracy.

The second solution compared with RASSA is GateKeeper [9], a pre-alignment accelerator that counts mismatches between short reads and a reference sequence, then filters-out reads with mismatch score above a certain threshold. GateKeeper's FPGA implementation throughput is compared with RASSA. RASSA is

found to outperform GateKeeper by more than 2 orders of magnitude.

To summarize, this work makes the following contributions:

1. RASSA, an in-memory processing resistive approximate similarity search accelerator, is introduced. The parallel processing architecture is presented bottom-up, from the memristor-based bitcell to base pair encoding and up to a complete RASSA system;
2. RASSA based implementation of long read mapping is developed;
3. Comparative analysis of RASSA's mapping accuracy, execution time and throughput with two existing solutions is conducted.

The full paper can be found at: https://arxiv.org/abs/1809.01127.

REFERENCES

[1] Jameson, J.L. and Longo, D.L. "Precision medicine—personalized, problematic, and promising." Obstetrical & gynecological survey, vol. 70, no. 10, pp. 612-614, 2015.

[2] Quick, J., Loman, N.J., Duraffour, S., et al, "Real-time, portable genome sequencing for Ebola surveillance." Nature, 530(7589), pp. 228-232.

[3] Rhoads, Anthony, and Kin Fai Auc. "PacBio sequencing and its applications." Genomics, proteomics & bioinformatics 13.5, pp. 278-289, 2015.

[4] Laver, T., Harrison, J., O'neill, P. A., Moore, K., Farbos, A., Paszkiewicz, K., & Studholme, D. J. "Assessing the performance of the oxford nanopore technologies minion." Biomolecular detection and quantification, vol. 3, pp. 1-8, 2015.

[5] Kaplan, R., Yavits, L., Ginosar, R. and Weiser, U. "A Resistive CAM Processing-in-Storage Architecture for DNA Sequence Alignment." IEEE Micro, vol. 37, no. 4, pp. 20-28, 2017.

[6] Li, Heng. "Minimap and miniasm: fast mapping and de novo assembly for noisy long sequences." Bioinformatics 32.14, pp. 2103-2110, 2016.

[7] Li, H., Minimap2: pairwise alignment for nucleotide sequences. Bioinformatics, vol. 1, p. 7, 2018.

[8] Berlin, K., Koren, S., Chin, C. S., Drake, J. P., Landolin, J. M., & Phillippy, A. M. "Assembling large genomes with single-molecule sequencing and locality-sensitive hashing." Nature biotechnology, 33(6), p. 623, 2015.

[9] Alser, M., Hassan, H., Xin, H., Ergin, O., Mutlu, O. and Alkan, C. "GateKeeper: a new hardware architecture for accelerating pre-alignment in DNA short read mapping." Bioinformatics, vol. 33, no. 21, pp. 3355-3363, 2017.

[10] Khatamifard, S. K., Chowdhury, Z., Pande, N., Razaviyayn, M., Kim, C., & Karpuzcu, U. R. "A Non-volatile Near-Memory Read Mapping Accelerator." arXiv preprint arXiv:1709.02381.

[11] Imani, M., Rahimi, A., Kong, D., Rosing, T., & Rabaey, J. M. "Exploring hyperdimensional associative memory". In IEEE International Symposium on High Performance Computer Architecture (HPCA), pp. 445-456, 2017.