APPENDIX A

ARTIFACT DESCRIPTION APPENDIX: ACCELERATING DNA LONG READ MAPPING WITH EMERGING TECHNOLOGIES

Abstract

DNA sequencing technologies output only short fragments of a genome, called reads. New single-molecule real-time sequencing technologies can produce long reads, up to tens of thousands base pairs, within minutes. However, these long reads may contain up to 15% errors.

To construct a complete genome from DNA reads, a computationally expensive bioinformatics task, read mapping, is required. Read mapping finds the best-fitting location for each DNA read on a long reference sequence.

The length and error rate of long reads poses a challenge for existing read mapping hardware solutions, designed for short reads with low error rates. This work presents a novel DNA read mapping hardware architecture, RASSA. RASSA is a Resistive Approximate Similarity Search Accelerator that exploits charge distribution and parallel in-memory processing to reflect a mismatch count between DNA sequences. RASSA implementation of long read DNA mapping outperforms the fastest state-of-art long read mapping solution (minimap2) by 16-77× with comparable accuracy.

Description

- **Program**: minimap2. Publicly available at: https://github.com/lh3/minimap2 (Li, Heng. "Minimap2: pairwise alignment for nucleotide sequences." *Bioinformatics* 1 (2018): 7.)
- **Compilation**: only 'make', according to the installation instructions
- **Datasets**:
  - Organisms: Escherichia coli strain K-12 substrain MG1655 (e.coli) and Saccharomyces cerevisiae W303 (yeast)
  - e.coli, reference sequence: https://www.ncbi.nlm.nih.gov/nuccore/CP014225.1
  - e.coli, PacBio: http://wgs-assembler.sourceforge.net/wiki/index.php/Escherichia_coli_K12_MG1655,_using_uncorrected_PacBio_reads,_with_CA8.1
  - e.coli, PacBio CCS: http://files.pacb.com/datasets/secondary-analysis/e-coli-k12-de-novo/1.3.0/Ecoli_MG1655_pacBioToCA.tgz
  - e.coli, ONT: https://s3.climb.ac.uk/nanopore/E_coli_K12_1D_R9.2_SpotON_2.pass.fastas
  - Yeast, reference sequence: https://www.ncbi.nlm.nih.gov/nuccore?term=CM007964:CM007981[PACC]
  - Yeast, PacBio: https://gist.github.com/pb-jchin/6359919
  - Yeast, ONT: https://trace.ncbi.nlm.nih.gov/Traces/sra/?run=ERR789757
- **Run-time environment**: Ubuntu 16.04.4 LTS (GNU/Linux 4.4.0-62-generic x86_64)
- **Hardware**: server with 16-core 2GHz Intel Xeon E5-2650 CPU server, 64GB of RAM, DELL PERC H710 hard drive.
- **Execution:**
  - For PacBio: 'minimap –x map-pb' (parameter settings suitable for PacBio reads)
  - For ONT: 'minimap –x map-ont' (parameter settings suitable for ONT reads)
- **Publicly Available?** Yes

Experiment Workflow and Evaluation

My work proposes a hardware architecture. An in-house behavioral simulator, based on the model of proposed architecture, was used to estimate runtime. Comparisons of the proposal with existing tools were performed to estimate architecture's potential for improvement. The other tool chosen for the comparison is minimap2. Minimap2 is the highest-performing read mapper available that supports long reads. It uses multi-threading and SIMD extensions, fully exploiting currently available hardware resources (Intel 16-core CPU).

Invoking minimap2 on the listed datasets with the parameters listed under 'Execution' executes minimap2 in read mapping mode. The inputs are the reference sequence of the organism and the dataset reads in .fastq format. The output is a .sam file with indicating the mapping location of each read on the reference sequence. Measured time is for the entire execution. An average over 100 executions was used.