Shifted Hamming Distance (SHD): A Fast and Accurate SIMD-Friendly Filter for Local Alignment in Read Mapping

Hongyi Xin\textsuperscript{1}, John Greth\textsuperscript{1}, John Emmons\textsuperscript{1}, Gennady Pekhimenko\textsuperscript{1}, Carl Kingsford\textsuperscript{1}, Can Alkan\textsuperscript{2}, and Onur Mutlu\textsuperscript{1}

\textsuperscript{1}Departments of Computer Science and Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, PA, USA
\textsuperscript{2}Department of Computer Engineering, Bilkent University, Ankara, Turkey

ABSTRACT
1 Motivation:

Edit distance is used to calculate the identity between to sequences, where some basic “edits” (i.e. insertions, deletions, and substitutions) are allowed. This operation constitutes the primary task performed by many read mappers that compare billions of pairs of short sequences. Although typical dynamic programming algorithms can calculate edit distance between any pairs of strings, the read mappers impose a maximum edit distance threshold for an alignment to be valid. However, the majority of string pairs tested during mapping differ by significantly more errors than the tolerated threshold. Such error-abundant string pairs needlessly waste resources and severely hinder the performance of read mappers. Therefore, it is crucial to develop a fast and accurate filter that can rapidly and efficiently detect error-abundant error pairs and remove them from consideration before more computationally expensive methods are applied.

2 Results:

We present a simple, efficient algorithm, Shifted Hamming Distance (SHD), which utilizes bit-parallelism and SIMD-parallelism to speed up the filtering process. SHD only filters out string pairs that contain more errors than the threshold. It maintains high accuracy with moderate error threshold (up to 5% of the string length) while achieving a 3-fold speedup over Gene Myers’s bitvector algorithm. SHD is also compatible with all mappers that perform sequence alignment for verification.

3 Availability:

The code is available online at:
https://github.com/xhongyi/SHD_code

4 Contact:

hxin@cmu.edu, calkan@cs.bilkent.edu.tr, onur@cmu.edu