Accelerating Nucleic Acid Sequencing Data Workflows Using a Rapid Computation of Hamming Distance

This technology is a computational method and software implementation to accelerate the computation of the Hamming distance between two nucleic acid sequences by encoding each sequence as a bitstring.

What is the Problem?

When analyzing high-throughput sequencing data, a computational bottleneck is read alignment in which reads (sequences) are mapped to a common reference genome or transcriptome. Therefore, a common basic operation is the comparison of two character strings. In particular, the Hamming distance between two character strings of equal length is the number of positions at which the corresponding symbols are different.

What is the Solution?

The solution is a computational method and software implementation to accelerate the computation of the Hamming distance between two nucleic acid sequences by encoding each sequence as a bitstring. Bitwise operations are used to generate new comparison strings. Counting the bits in these strings gives us the Hamming distance between the two sequences. The representation of the base nucleotide sequence is based on a 5 letter alphabet “ACGTN”, where “N” means unknown. Our methods used 4 bits to represent all 5 possible sequences. where N is 0000, and A, C, G, T are represented by 1000,0100,0010,0001 respectively, and the method would work on any permutation of the mappings between A,C,G,T and 1000,0100,0010,0001. With this representation, bitwise operations between the binary representations of two letters will give rise to a bitstrings where the number of 1’s in the bitstrings distinguishes between matches, mismatches and partial matches where one letter is unknown. This converts the comparison of two sequences into bitwise operations, followed by bitcounting.

Applications of this invention include the analyses of single cell RNA sequencing data and long-read sequencing Nanopore data. We are currently working on developing a rapid diagnostic assay with efficient software tools that will utilize this invention to reduce the turnaround time for the diagnosis of leukemia patients.

What Differentiates it from Solutions Available Today?

Technology ID

BDP 8704

Category

Software/Bioinformatics
Selection of Available Technologies

Authors

Ling-Hong Hung

Learn more

Existing approaches run into a bottleneck with read alignment. This solution is 70x faster on a CPU than comparison of the letters individually, due to the efficiency and parallelism of the bitwise and bitcounting operations. Larger vector units for bitwise operations that would be available on GPUs, newer CPUs and custom FPGAs should give rise to even faster speeds.

Patent Information:

US20220415444A1