Charles Bowen-Rayner

Bioinformatics - sequence alignment

Basic Problems in Bioinformatics Sequence Alignment:

There are two basic problems within this domain. The first is Pairwise Sequence Alignment involves the comparison of two biological sequences, such as DNA, RNA, or proteins, with the aim of identifying regions of similarity between them. These similarities often indicate functional or evolutionary relationships, providing valuable insights into the genetic makeup and evolution of organisms. Secondly, Multiple Sequence Alignment which builds on the first concept by aligning three or more sequences simultaneously, allowing for the identification of conserved regions across a set of related sequences. MSA plays a role in understanding evolutionary relationships among species, as well as in deciphering protein structure and function. Using MSA researchers can identify common patterns and motifs which allows for the inference of evolutionary histories and the prediction of protein functions.

Algorithms Implemented on GPGPUs:

Smith-Waterman Algorithm: A dynamic programming algorithm used for local sequence alignment. It identifies the optimal alignment between two sequences by maximizing a similarity score. The parallelization of this algorithm on GPGPUs accelerates the computation of alignment scores for large sequence datasets.

Needleman-Wunsch Algorithm: A dynamic programming algorithm used for global sequence alignment. The process involves aligning entire sequences to identify the optimal alignment. GPGPU implementations speed up the computation of alignment matrices for large-scale sequence comparisons.

Basic Local Alignment Search Tool (BLAST): BLAST is a heuristic algorithm for comparing a query sequence against a database of sequences to find local sequence similarities. GPU-accelerated versions of BLAST significantly reduce the time required for database searches which improves the efficiency of sequence analysis.

Open Source Software Titles:

CUDA-BLAST: An open-source project that provides a GPU-accelerated implementation of the BLAST algorithm using NVIDIA CUDA technology. It offers significant speedups for sequence database searches, particularly useful for large-scale bioinformatics analyses.

GPU-HMMER: An open-source project that utilizes GPU acceleration for profile Hidden Markov Model (HMM) searches, commonly used for sequence alignment and homology detection. GPU-HMMER speeds up the process of searching biological sequence databases for protein families and domains. A performance of up to 100x faster than a single core of AMD Shanghai 2.3 GHz has been measured for 3 GPU units using this.

Considerations:

1. Despite the massive advantage of GPUs in their ability to handle the data intensive nature of this problem the main challenges are optimizing algorithms and data structures to minimize processing time and memory usage.
2. A study using the Smith-Waterman algorithm for optimal pairwise alignment of a 2 Petacell dataset was able to reduce execution times from 9.5 h on a Kepler GPU to just 2.5 h on a Pascal counterpart, with energy costs cut by 60%. This shows the significant improvement that GPGPUs provide when faced with an extremely data intensive problem.
3. It is very important that the chosen parallelized GPU-accelerated sequence alignment algorithm is able to integrate into the existing workflow pipeline and it does not interfere with the existing process for data pre-processing, visualization and analysis.