CSE305 Project: Parallel sequence alignment

Gleb Pogudin, gleb.pogudin@polytechnique.edu

Sequence alignment is one of the fundamental problems in computational bioinformatics. The problem is, given two sequences of nucleotides (A, C, G, T) or amino acids, compute a "distance" between them. This is a prominent tool when it comes to classifying/clustering sequences and searching for mutations.

Form the computational standpoint, the problem is similar to the classical edit distance problem. The classical sequence alignment algorithms such as Needleman-Wunsh and Smith-Waterman are also based on dynamic programming resulting in time complexity $\mathcal{O}(mn)$, where m and n are the lengths of the sequences to align. If the lengths are or the order of thousands (and with modern sequencing techniques grow further and further), this takes a very long time, so it is natural to try to parallelize the problem. On the other hand, dynamic programming algorithms are often not so easy to parallelize.

The goal of the project is to explore parallel versions of the classical sequence alignment algorithms for CPUs or GPUs.

The goal of the project will be to implement one of the recent parallel algorithms for sequence alignment and test its performance on real-life data. As a starting point, we recommend the survey paper "Parallel Optimal Pairwise Biological Sequence Comparison: Algorithms, Platforms, and Classification" (ACM Computing Surveys, 2016, https://dl.acm.org/doi/10.1145/2893488). You can choose one of the approaches described there (a group of three is strongly encouraged to take two related approaches, e.g. the same basic algorithm but on both CPU and GPU).

One the algorithm is implemented, you should analyze the performance of the implementation depending on the inputs sizes, number of cores, and the similarity of the input sequences. Use one of the publicly available databases such as UniProtKB (https://www.uniprot.org/).