GSAlign: An efficient sequence alignment tool for intra-species genomes

Aaloke Mozumdar (2019004)
Aditya Choudhary (2020489)
Ananya Jain (2019408)
Mohit Bansal (2020526)
Nikhil Kumar Gupta (2020530)

Abstract

Personal and comparative genomics are becoming increasingly important in clinical practice and genomics research. Both regions require sequence alignments to discover sequence conservation and variation. Many methods have been developed, some designed to compare small genomes, others not efficient for large-scale genome comparisons. Furthermore, most existing genome comparison tools have not been systematically evaluated for sequence alignment accuracy. A wrong sequence alignment would produce false sequence variants.

Algorithm used - GSAlign

GSAlign is an ultra-fast sequence alignment algorithm for intra-species genome comparison. GSAlign includes three unique features:

- It is the first attempt to use Burrows-Wheeler Transform on genome sequence alignment
- It adopts a divide-and-conquer strategy to separate a query sequence into regions that are easy to align and regions that require gapped alignment.

With all these features, we demonstrate that GSAlign is very efficient and sensitive in finding both the exact matches and differences between two genome sequences and it is much faster than existing state-of-the-art methods.

Dataset Description

- We randomly generate sequence variations with the frequency of 20,000 substitutions (SNVs), 350 small, indels (1~10 bp), 100 large indels (11~20 bp) for every 1M base pairs.
- To increase the genetic distance, we generate different frequencies of SNVs.
- Benchmark datasets labelled with 1X contain around 20,000 SNVs for every 1M base pairs, whereas datasets labelled with 3X (or 5X) contain 60,000 (or 100,000) SNVs per million bases.
- We generate three synthetic datasets with different SNV frequencies using the human genome (GRCh38).
- The synthetic datasets are referred to as simHG-1X, simHG-3X, and simHG-5X, respectively.

Details

EXISTING CODE BASE

https://github.com/hsinnan75/GSAlign
(Algorithm has previously been implemented in C++)

CHANGES

We will be implementing the GSAlign algorithm using Python

Paper Link: https://doi.org/10.1186/s12864-020-6569-1