ABSTRACT

AUTOMATING THE PARALLEL ALIGNMENT OF DNA USING KUBERNETES

By Aideen Fay

Supervised by Jeremy Jones

Trinity College Dublin | School of Computer Science and Statistics

B.A.I. Computer Engineering

Read alignment is the first stage in the DNA analysis pipeline responsible for processing the ever-increasing amount of genomic data produced by next generation sequencing technologies. Currently available alignment tools are not capable of processing the available genomic data quickly and cheaply enough, some requiring days to complete a full alignment[1].

BWBBLE is a read alignment program that uses a collection of genomes to reduce the bias and improve accuracy when performing read alignment. The benefits BWBBLE offers are dwarfed by the fact that it is 100 times slower than other read alignment programs which use a single reference genome[2]. This is compounded by the fact that read alignment is already the bottleneck stage of DNA analysis.

To minimize the time taken to perform short-read alignments with BWBBLE, various parallelized execution models such as AWS-BWBBLE[3] and SparkBWBBLE[4] have been developed. While AWS-BWBBLE succeeded in demonstrating the potential for a linear speedup by distributing reads amongst several cloud based virtual machines, it introduced significant cost and infrastructure management overhead. As a result of this overhead, adoption has been limited and the practical speedups achieved remain out of reach of DNA researchers.

This project successfully demonstrated that a linear speedup in BWBBLE is possible by parallelizing the read alignment stage using Kubernetes jobs. We also demonstrated that it was possible to achieve this cheaply and with minimal operational overhead using a Kubernetes Controller. In doing so, we have made it possible for more researchers to easily leverage BWBBLE for their work.

---

[1] Arram, J., Kaplan, T., Luk, W. & Jiang, P., 2017. Leveraging FPGAs for Accelerating Short Read Alignment. *IEEE/ACM Transactions on Computational Biology and Bioinformatic,* May.14(3).

[2] Huang, L., Popic, V. & Batzoglou, S., 2013. Short read alignment with populations of genomes. *Bioinformatics,* July, 29(13), p. i361–i370.

[3] McGinley, K., 2019. *Parallel DNA Read Alignment Using the Amazon Cloud, MSc* Computer Science, Trinity College Dublin.

[4] Stratford, B., 2018. Cloud based high-speed parallel DNA read alignment using the Burrows-Wheeler Transform, B.A. Computer Science, Trinity College Dublin.