Parallelization of Mapping Algorithms for Next Generation Sequencing Applications

Doruk Bozdağ[1], Catalin C. Barbacioru[2], and Umit Catalyurek[1]

[1] The Ohio State University, Dept. of Biomedical Informatics, {bozdagd,umit}@bmi.osu.edu
[2] Applied Biosystems, catalin@appliedbiosystems.com

With the advent of next-generation high throughput sequencing instruments, large volumes of short sequence data are generated at an unprecedented rate. Processing and analyzing these massive data requires overcoming several challenges. A particular challenge addressed in this abstract is the mapping of short sequences (reads) to a reference genome which is a significantly time consuming combinatorial problem in many applications, including whole-genome resequencing, targeted sequencing, transcriptome/small RNA, DNA methylation and ChiP sequencing. This computationally intensive process takes time on the order of days using existing sequential techniques on large scale datasets. In this work, we describe new parallelization methods to speedup short sequence mapping and to reduce the execution time from the order of days under just a few hours for such large datasets [1]. These methods are designed to optimize the distribution of data to enhance the parallel performance and can be used to parallelize most mapping algorithms that utilize hashing or indexing techniques.

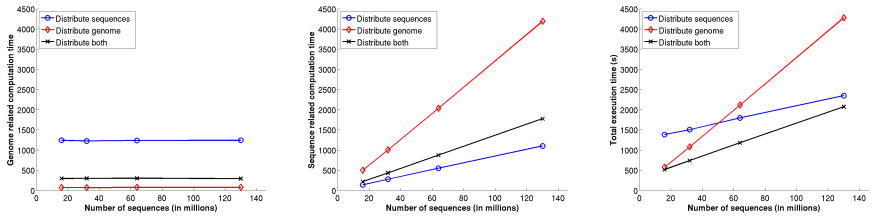

Fig. 1. Comparison of three parallelization methods for mapping short sequences to a 800Mbp genome using parallelized MapReads [2] program on a 16-node cluster.

Each parallelization approach has different advantages in terms of scalability. We investigated theoretical cost models of six different methods and also compared them through experiments on real datasets. In Figure 1, comparison of three of the proposed methods are given for varying number of short sequences. As demonstrated in this example, efficient distribution of sequence and genome data helps improving the parallel execution time. Furthermore, choosing the best parallelization method based on the given data sizes are made possible by the proposed cost models. To the best of our knowledge this is the first study on parallelization of short sequence mapping problem.

References

1. D. Bozdag, C. C. Barbacioru, U. Catalyurek. Parallel Short Sequence Mapping for High Throughput Genome Sequencing *23rd International Parallel and Distributed Processing Symposium*, to appear (2009).
2. http://solidsoftwaretools.com/gf/project/mapreads/.

Dagstuhl Seminar Proceedings 09061
Combinatorial Scientific Computing
http://drops.dagstuhl.de/opus/volltexte/2009/2129