TEXAS TECH UNIVERSITY
Edward E. Whitacre Jr.
College of Engineering
Computer Science

Data-Intensive Scalable Computing Laboratory (DISCL)

Technical Report

Department of Computer Science

Texas Tech University

Distributed Framework and Algorithm Design for Large-scale Biological Sequence Alignment

Frank Conlon, Jiang Zhou, Shengping Yang, Yong Chen

frank.conlon@ttu.edu, jiang.zhou@ttu.edu, shengping.yang@ttuhsc.edu, yong.chen@ttu.edu

June 2017

Technical Report № TTU/DISCL-2017-02

http://discl.cs.ttu.edu

Distributed Framework and Algorithm Design for Large-scale Biological Sequence Alignment

Frank Conlon, Jiang Zhou, Shengping Yang*, Yong Chen

Department of Computer Science, Texas Tech University, Lubbock, TX
*Texas Tech University Health Sciences Center, Lubbock, TX

frank.conlon@ttu.edu, jiang.zhou@ttu.edu, shengping.yang@ttuhsc.edu, yong.chen@ttu.edu

Abstract

One of the most pressing issues in modern genome analysis is the issue of how to quickly and efficiently align short reads against a whole sequenced genome. Short reads are small portions of a genome that have been sequenced. These short reads are faster and much less expensive to produce than an entire genome sequence. Once a series of these short reads has been produced though they must be aligned against a base genome that has been completely sequenced in order to analyze what they mean. Improving this alignment process is a constant an ongoing effort.

Many improvements have been made to the process of aligning short reads over the years. One of the fastest and most popular methods for aligning short reads is to perform a Burrows-Wheeler transformation on the completely sequenced genome that the short reads will be aligned against to make reduce the Big O time complexity of the alignment process. This algorithm provides fast and inexpensive alignment, but it still does not take advantage of modern improvements in distributed computing. Attempts have been made to remedy this, one of which will be discussed here, but no single method has stood out among the rest.

One tool that is available to researchers to perform genome alignment is a piece of software called Bowtie. Bowtie is a tool that implements the previously mentioned Burrows-Wheeler algorithm. Like many other aligners available, this tool is not optimized for parallel computing on a distributed memory system. There is a tool available, PMAP, that can be used to execute Bowtie in a parallel manner on a distributed system. However, this tool is inefficient and has to be run as a middleware layer between Bowtie and the system. This report will discuss how PMAP parallelizes Bowtie, some of its drawbacks, and a distributed framework and parallel algorithms for directly parallelizing Bowtie.

Keywords: Burrows-Wheeler algorithm, parallel, genome, PMAP