Abstract

Genomic sequence alignment of varied species is one of the most sort of applications in bioinformatics. In future bioinformatics technologies are expected to produce genomic data of terabyte. Bioinformatics computation require super computer for sequence alignment computation which involves huge cost. Parallelization technique is a way forward in computing sequence alignment with limited cost and time. Cloud computing and MapReduce framework play an important role in bioinformatics intensive application to achieve parallelization since it provides a consistent performance over time and it also provides good fault tolerant mechanism. The existing gene sequencing methodologies are designed based on Hadoop-MapReduce framework which adopts a serial execution strategy which is an area of concern. This work introduces a Smith-Waterman Alignment on the Bulk synchronous Parallel Map Reduce (SW-BSPMR) cloud platform for bioinformatics gene sequence alignment. This work adopts a widely accepted and accurate SW algorithm for sequence alignment and parallel synchronous scheduler methodology of map and reduce framework process is considered. A customized MapReduce based on Microsoft Azure cloud platform is developed to overcome the
An Efficient Bulk Synchronous Parallelized Scheduler for Bioinformatics Application on Public Cloud

issue in Hadoop-MapReduce framework. The experimental study presented in this work proves that the SW-BSPMR can accurately and effectively align bioinformatics genomic sequences of various read length.

References


36. Michael C. Schatz: CloudBurst: highly sensitive read mapping with MapReduce.
37. Tung Nguyen, Weisong Shi and Douglas Ruden: CloudAligner: A fast and full-featured MapReduce based tool for sequence mapping.
38. LI Xubin, JIANG Wenrui, JIANG Yi and ZOU Quan: Hadoop Applications in Bioinformatics.
39. Xiao-liang Yang, Yu-long Liu, Chun-feng Yuan, Yi-hua Huang: Parallelization of BLAST with MapReduce for Long Sequence Alignment

Index Terms
Computer Science Distributed Computing

Keywords
Bioinformatics, genomic sequence, Scheduler, parallelization, hadoop, Microsoft Azure