Abstract

Genome sequence matching is used to reveal biological information hidden in the DNA sequences and genome sequences. The main objective is to find whether the given sequence is like other sequence or not. To find the similarity between the diseases and intensity of the disease DNA sequences are matched. There is large number of sequences and the database is still growing. Given a genome sequence and to find matching sequences from the complete database is a big challenge. The genome sequence matching algorithms are also computation intensive like BLAST; which performs large number of string matching operations. So to handle this genome sequence matching algorithms and to store data which is Big data; Hadoop is used. Hadoop is a parallel processing Big data framework. The genome sequence database can be stored on Hadoop distributed filesystem. And then can be efficiently processed using Map/Reduce. The data is distributed in the form of blocks and for every block an instance of mapper is mapped to process the block and then output of all the mappers is combined by reducer. This Map/Reduce process has inter-node parallelism. To further speedup the process and to efficiently utilize the resources like Central processing unit and Graphical processing unit,
a parallel processing framework called OpenCL is used. In this work OpenCL is integrated with Hadoop using an API called APARAPI. In addition to inter-node parallelism, intra-node parallelism is also provided and Map/reduce is accelerated for BLAST algorithm which is termed as HCLBLAST. The HCLBLAST is compared with HBLAST and BLAST algorithm for different datasets. It is found that HCLBLAST outperforms in all cases.

References


Index Terms

Computer Science

Information Sciences

Keywords

DNA, HCLBLAST, BLAST, NGS