
Mapreduce and Hadoop Algorithms in Bioinformatics Papers

Abhishek Tiwari 

Citation: *A. Tiwari*, "Mapreduce and Hadoop Algorithms in Bioinformatics Papers", Abhishek Tiwari, 2012. [doi:10.59350/d8s8t-w3818](https://doi.org/10.59350/d8s8t-w3818)

Published on: August 09, 2012

Solely inspired by Atbrox's list of [academic papers for Mapreduce & Hadoop Algorithms](#). Unlike computer science where applications of Mapreduce/Hadoop are very much diversified, most of published implementations in bioinformatics are still focused on the analysis and/or assembly of biological sequences. As usual this list will be updated time to time. If you find that any important paper that is missing from the list then please drop a comment at end of the post.

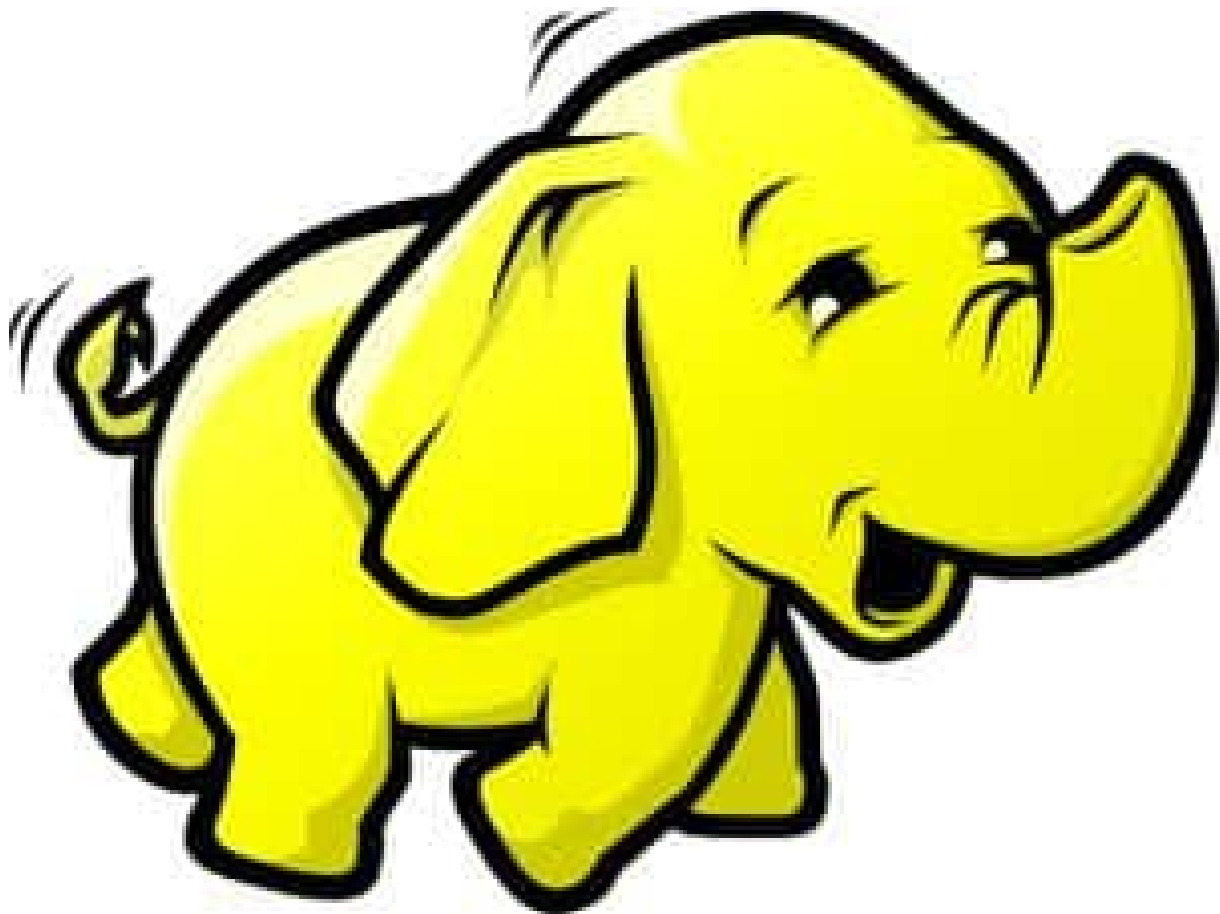


Figure 1: Hadoop Logo

Review articles

1. [An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics](#)

Paper describes the concepts behind Hadoop and the associated HBase project, and current bioinformatics software that employ Hadoop.

Sequence analysis/assembly

2. [PeakRanger: A cloud-enabled peak caller for ChIP-seq data](#)

PeakRanger paper describes a Hadoop version with supports for splitting the job by chromosomes to take advantage of the chromosome-level independence (CLI) of ChIP-seq data sets. In the CLI case, “map-then-reduce” becomes “split-by-chromosome-then-call-peaks” where chromosomes are used as keys.

1. [Quake: quality-aware detection and correction of sequencing errors](#)

Hadoop cluster was used for Counting k-mers and also to sum together the partial counts computed on individual machines using an extension of the MapReduce word counting algorithm.

2. [Biomedical Case Studies in Data Intensive Computing](#)

Study illustrates two use case, one the analysis of gene sequence data (35339 Alu sequences) and other a study of medical information (over 100,000 patient records), and compares the performance of MapReduce computing model with MPI.

3. [Cloud-scale RNA-sequencing differential expression analysis with Myrna](#)

Myrna is a cloud-computing pipeline for calculating differential gene expression in large RNA-Seq datasets. Myrna is designed with a parallel Hadoop/MapReduce model in mind. Myrna can be run on the cloud using Amazon Elastic MapReduce, on any Hadoop cluster, or on a single computer (without requiring Hadoop).

4. [Cloud computing for comparative genomics](#)

Describes a typical comparative genomics algorithm, the reciprocal smallest distance algorithm (RSD), to run within Amazon’s Elastic Computing Cloud (EC2).

5. [BlastReduce: High Performance Short Read Mapping with MapReduce](#)

Describes a parallel read mapping algorithm optimized for aligning next-generation sequence data to reference genomes.

6. [Biodoop: Bioinformatics on Hadoop](#)

Describes Hadoop implementation to three algorithms: BLAST, GSEA and GRAMMAR.

7. [CloudBurst: highly sensitive read mapping with MapReduce](#)

Describes algorithm [CloudBurst](#), a new highly scalable read-mapping algorithm optimized for next-generation sequence data.

8. [CloudBLAST: Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications](#)

Describes an implementation which integrates Hadoop, Virtual Workspaces, and ViNe as the MapReduce, virtual machine and virtual network technologies, respectively, to deploy the commonly used bioinformatics tool NCBI BLAST on a WAN-based test bed.

9. [Searching for SNPs with cloud computing](#)

Describes [Crossbow](#), a cloud-computing software tool that executes in parallel using Hadoop and combines the aligner Bowtie and the SNP caller SOAPsnp.

10. [The Genome Analysis Toolkit: a MapReduce framework for analyzing next generation DNA sequencing data](#)

Describes [Genome Analysis Toolkit \(GATK\)](#), a structured programming framework designed to ease the development of efficient and robust analysis tools for next-generation DNA sequencers using the functional programming philosophy of MapReduce.

11. [Cloud Technologies for Bioinformatics Applications](#)

Describes [Dryad](#) (Microsoft's implementation of MapReduce) and Azure with application in EST sequence assembly, identification of HLA-associated viral evolution, and a pairwise Alu gene alignment. Dryad combines the MapReduce programming style with dataflow graphs to solve the computation tasks.

12. [A novel approach to multiple sequence alignment using hadoop data grids](#)

Describes multiple sequence alignment method with improved computation time and accuracy using Hadoop framework.

13. [Parallelizing bioinformatics applications with MapReduce](#)

Describes Hadoop based implementation of BLAST and GSEA (Gene Set Enrichment Analysis) algorithms.

14. [Data Intensive Computing for Bioinformatics](#)

Describes wide range of topics using Microsoft's MapReduce framework Dryad including iterative MapReduce programming model to analyse the metagenomics data . (Highly recommended)

Phylogenetic

15. [MrsRF: an efficient MapReduce algorithm for analyzing large collections of evolutionary trees](#)

Describes MapReduce framework for designing phylogenetic applications.

Workflow

16. [Kepler + Hadoop : A General Architecture Facilitating Data-Intensive Applications in Scientific Workflow Systems](#)

Describes integration of Hadoop with Kepler workflow system which enables users to compose and execute MapReduce applications.

17. [Data Parallelism in Bioinformatics Workflows Using Hydra](#)

Describes MapReduce like middleware Hydra to support data parallelism and parameter sweep parallelism in bioinformatics workflows.

Pattern finding/motif detection

18. [MapReduce-Based Pattern Finding Algorithm Applied in Motif Detection for Prescription Compatibility Network](#)

Describes a MapReduce-based pattern finding algorithm (MRPF) for analyzing the complex network.