Using Bioinformatics Applications on the Cloud

Hyungro Lee
School of Informatics and Computing, Indiana University
815 E 10th St.
Bloomington, IN 47408
lee212@indiana.edu

ABSTRACT

Dealing with large genomic data on a limited computing resource has been an inevitable challenge in life science. Bioinformatics applications have required high performance computation capabilities for next-generation sequencing (NGS) data and the human genome sequencing data with single nucleotide polymorphisms (SNPs). From 2008, Cloud computing platforms have been widely adopted to deal with the large data sets with parallel processing tools. MapReduce parallel programming framework is dominantly used due to its fast and efficient performance for data processing on cloud clusters. This study introduces various research projects regarding to reducing a data analysis time and improving usability with their approaches. Hadoop implementations and workflow toolkits are focused on address parallel data processing tools and easy-to-use environments.

Keywords

Cloud Computing, Workflows, Bioinformatics

1. INTRODUCTION

Bioinformatics tools become much easier to use even though its complexity for the distributed computations and data analyses gets higher. Researchers have been working on dealing with large sequencing data to discover new findings without computation time lag. Cloud computing which is an on-demand and pay-as-go model, has been used to provide dynamic computing resources without commitment or upfront costs for building physical systems. From a small research laboratory to a large institute, cloud computing are broadly used for big data analysis with parallel processing tools such as Apache Hadoop. MapReduce, which is a programming model consisting of map() and reduce() functions, is implemented in Apache Hadoop for processing large data sets in parallel. Bioinformatics software widely adopted cloud computing with Hadoop implementation to manage large genomic data and to perform data analysis. Several studies have shown that this combination satisfies the needs for efficient and fast computation in genomic research.

One of the challenges in genomic research is to understand and share analysis processes with others. Scientific workflow system such as Galaxy offers simple web-based workflow toolkits and scalable computing environments to resolve this challenge. CloudMan, for example, runs Galaxy workflow system on Amazon EC2 to perform data-intensive computational analysis on the cloud. In the following section, scientific workflows, parallel applications, i.e. Hadoop MapReduce framework, and related work are arranged in chronological order.

2. PARALLEL APPLICATIONS

Bioinformatics researchers have been confronted with big data and computational challenges and many studies have been conducted with parallel applications such as Message Passing Interface (MPI) and Hadoop on cloud computing to speed up computational time on data processing and analysis. Hadoop provides parallel processing across multiple nodes with its file system, which is Hadoop Distributed File Systems (HDFS) for processing large datasets with MapReduce paradigm.

2.1 Cloud Computing Clusters

Cloud computing providers have accommodated High Performance Computing (HPC) in the cloud with highly scalable computing resources. HPC in the cloud is not as fast as traditional HPC or Grid from top research institutions but it supports relatively powerful computing components in terms of leveraging time and cost. MapReduce/Hadoop gives an enough boost to cloud computing clusters for most embarrassingly parallel problems (EPP). Such data-parallel programming models are useful to analyze or solve the problems in clusters of commodity machines ,i.e. cloud computing clusters.

2.2 MapReduce and Cloud Computing

We have identified that Hadoop (MapReduce implementation) is broadly used to utilize multiple compute nodes in cloud computing by map and reduce functions. For those simple tasks such as sequence search, read alignment and image recognition, cloud computing and mapreduce parallel programming is a good combination for data analysis in genomic research. MapReduce splits large inputs into small pieces for independent sub-processes and combines completed results from the sub-processes with Hadoop Distributed File System (HDFS). Table 1 shows related re-

searches in chronological order with the changes of technical tools.

2.3 mpiBLAST

In 2003, mpiBLAST was introduced as a parallelization version of BLAST to improve performance of the sequence searching. MPI was used to implement database segmentation and mpiBLAST splits the databases into small chunks on a shared storage so that each worker node performs blast search with the piece of database segmentation. As a result, mpiBLAST shows that the single worker needs 22.4 hours whereas 128 workers complete the search within 8 minutes. Beowulf cluster was used in the test to take advantage of a low-cost and efficient Linux cluster.

2.4 CloudBLAST

CloudBLAST [9] is a software which combines virtual compute resources, virtual network technologies and parallel technique for Bioinformatics applications. Apache Hadoop (MapReduce implementation) performs data analysis with NCBI BLAST in parallel and ViNe, which is a virtual network architecture for grid computing, enhances network connectivity with better performance in packet transfer and latency. With this implementation and configuration, CloudBLAST shows better performance than mpiBLAST on the Xen VMs from two regions: University of Florida and University of Chicago.

2.5 CloudBurst

CloudBurst [11] is a read mapping algorithm using MapReduce for mapping single end next generation sequence (NGS) data. A short-read mapping software, RMAP performs similar data analysis but with much slow performance than MapReduce parallel programming framework. CloudBursts shows 30 times faster than RMAP. The seed-and-extend algorithm used in both RMAP and CloudBursts is suitable for parallel processing in MapReduce to expedite the mapping process.

2.6 Crossbow

In 2009, Langmead et al [5] developed a software tool that runs the aligner Bowtie with the SNP caller SOAPsnp on Amazon hadoop clustser. CloudBurst [11], which is a short read-mapping algorithm for mapping next-generation sequence (NGS) data with MapReduce on the cloud, previously showed that Hadoop was useful to improve computational processes by performing data analysis in parallel with virtual compute nodes. Crossbow uses two programs for human genome alignment and single nucleotide polymorphism (SNP) detection. Bowtie aligns short DNA sequences (reads) to the human genome, and SOAPsnp is a Bayeian based SNP-detection program for multiple whole-human datasets. Crossbow showed that long running sequential data processing (over 1,000 hours) can be completed within three hours using on-demand cloud computing resources (i.e. Amazon EC2) and a MapReduce parallel software framework. Hadoop has been applied to Bowtie and SOAPsnp to accelerate computation performance and Amazon EC2 has been adopted to facilitate reproducibility and cooperation of Crossbow with experiment data. 40 EC2 Extra Large High CPU Instances were used to provide 320 cores in total for analyzing 38-fold coverage of the human genome in the Amazon cloud.

2.7 SparkSeq

SparkSeq [13] is a library for next generation sequencing (NGS) data with MapReduce framework and Apache Spark on the cloud. Apache Spark, which is a cluster computing framework using memory, expedites analysis of sequencing alignment files by caching the datasets in memory. SparkSeq performs fast and efficient in-memory computations on the cloud by Apache Spark and Resilient Distributed Datasets (RDDs) which is a distributed memory abstraction. Memory-based Apache Spark showed better performance than disk-based architecture such as Apache Mahout for iterative machine learning algorithms or low-latency applications.

2.8 SeqPig

SeqPig [12] is a set of scripts that uses Apache Pig, which is a programming tool generating MapReduce programs automatically, for large sequencing data sets. SeqPig scripts ease a way of data manipulation, analysis and access with Hadoop and the scripts have been tested on Amazon Elastic MapReduce (EMR) with amazon storage service (S3) to perform data processing in parallel on the cloud. SeqPig runs with Hadoop-BAM to read an input BAM file which is a binary version of a SAM file. Hadoop-BAM provides an easy access to BAM data using the Hadoop MapReduce framework with the Picard SAM JDK, and Samtools-like command line tools.

2.9 AzureBlast

AzureBlast [8] is a parallel BLAST (Basic Local Alignment Search Tool) engine on Windows Azure cloud platform. BLAST is one of the popular applications in bioinformatics to find regions of local similarity between sequences. AzureBlast runs BLAST on multiple instances of Azure Cloud by the query segmentation data parallel pattern. Other Cloud-enabled BLAST implementations such as CloudBLAST [9] use Hadoop MapReduce runtime instead. In AzureBlast, the input sequences will be divided into multiple partitions per each worker node and be merged together once the computation finished.

2.10 Rainbow

Rainbow [14] is an enhancement of Crossbow, which detects single nucleotide polymorphisms (SNPs) in whole-genome sequencing (WGS) data on Amazon hadoop cluster. Rainbow claims four improvements: 1) support for BAM input files (previously FASTQ input files only used), 2) data preprocessor to deal with large FASTQ files quickly by splitting them into small pieces 3) monitoring cluster nodes, and 4) SOAPsnp aggregator to support genome-wide association studies (GWAS) by merging multiple SNPs outputs to a single chromosome-based genotype file. In their test run, the WGS data for 44 subjects were analyzed with Rainbow on Amazon Cloud with about 8 TB input files.

2.11 BioPig

BioPig [10] is a sequence analysis tool with Hadoop and Apache Pig for large-scale sequencing data. Pig Latin, the programming language used in the Pig program, is easy to write and understand so that implementing MapReduce functions is less complicated. Compared to SQL, Pig Latin has some benefits to load user data at any point.

Name	Year	Description	Application tools
CloudBLAST	2008	Combining MapReduce and Virtualization on Distributed Resources for Bioinformatics Applications	Hadoop, ViNe, BLAST
CloudBurst	2009	highly sensitive read mapping with MapReduce	MapReduce, Amazon EC2
Crossbow	2009	Searching for SNPs with cloud computing	Hadoop, bowtie, SOAPsnp, Amazon EC2
Myrna	2010	Cloud-scale RNA-sequencing differentialexpression analysis	Hadoop, Amazon EMR, HapMap
Galaxy	2010	Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences	Python, web server, SQL database
Galaxy CloudMan	2010	delivering cloud compute clusters	Amazon EC2, Bio-Linux, Galaxy
AzureBlast	2010	A Case Study of Developing Science Applications on the Cloud	Azure, BLAST
CloudAligner	2011	A fast and full-featured MapReduce based tool for sequence mapping	CloudBurst, MapReduce, Amazon EMR
CloVR	2011	virtual machine for automated and portable sequence analysis from the desktop using cloud computing	VM, VirtualBox, VMWare
Cloud BioLinux	2012	pre-configured and on-demand bioinformatics computing for the genomic community	VM, Amazon EC2, Eucalyptus, VirtualBox
FX	2012	an RNA-Seq analysis tool on the cloud	Hadoop, Amazon EC2
Rainbow	2013	Tool for large-scale whole-genome sequencing data analysis using cloud computing	Crossbow, bowtie, SOAPsnp, Picard, Perl, MapReduce
BioPig	2013	a Hadoop-based analytic toolkit for large-scale sequence data	Hadoop, Apache Pig
SeqPig	2014	simple and scalable scripting for large sequencing data sets in Hadoop	Hadoop, Apache Pig
SparkSeq	2014	fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision	Apache Spark, Scala, samtools

Table 1: Cloud-enabled bioinformatics platforms

3. SCIENTIFIC WORKFLOWS

In life sciences, data analysis and process are important subjects since accessing resources in distributed systems is getting complicated due to rapid changes of the compute platforms and heterogeneity. Scientific workflow tools such as Taverna, Kepler, Galaxy, or XBaya (now Airavata) have developed methods for collecting distributed data, pipelines, or protocols to support bioinformatics applications. Workflow has been widely used to improve coordination between biologists and bioinformaticians since it offers better interoperability between bioinformatics applications, but without explaining how it is implemented in detail.

Workflow is a depiction of tasks, processes, flows, and data in distributed environments. The combination of processes and flows describes an entire application with individual tasks in a visual form. The set of methods and technologies in the workflow application supports a business process by controlling the workflow components with distributed information so that the workflow system can orchestrate a task, a process, a flow, and data with resources. Coordination between others by sharing 'process knowledge' is a key feature supported in the workflow [6]. The specification languages (e.g. Apache ODE, BPEL, and SCUFL) help transfer composed workflows and pipelines between people. The key concepts in workflow are the processes, matching human resources to tasks, matching information resources to tasks and process managements. When the workflow performs the key concepts need to ensure three things. The content of process, the owner of the process, and the act of the process should be indentified to explain the analysis tasks performed in the workflow [6].

3.1 Galaxy

Galaxy [3] is a web-based scientific workflow system for genomic research. Bioinformatics tools such as BLAST, HMMER tools, etc. are easily accesible through its repository in Galaxy workflow system. Reproducible Research System (RRS), which is a concept to support reproducible computational experiments, is applied to Galaxy system to enable reproducibility by recording and repeating computational analyses of user workflows. To support distributed workloads in cluster, Galaxy works with portable batch system (PBS), or Sun Grid Engine (SGE) clusters. CloudMan [1] is a cloud resource management system for Galaxy workflow system that enables Galaxy users to utilize dynamic computational infrastructure with their data sets. Amazon EC2 and NERC Bio-Linux [4] are included in the CloudMan to provide a suite of biological tools on the cloud.

3.2 CloudMan

CloudMan is a deployment toolkit of Galaxy workflow system on the Amazon EC2. Galaxy CloudMan Console manages compute cluster on the cloud to increase or decrease the cluster size. Sun Grid Engine(SGE) is mainly used to control cluster jobs on the cloud with Galaxy CloudMan. Amazon Elastic Block Storage (EBS), which is an external and permanent data volume, is used to keep the data after shutting down cloud instances.

3.3 CloudBioLinux

Northwest Environmental Business Council (NEBC), J. Craig Venter Institute (JCVI), Harvard School of Public Health and others offer cloud infrastructure to enable genome analysis on cloud computing platforms with a BioLinux [2] bioinformatics suite. More than 135 bioinformatics software for sequence alignment, clustering, assembly, display, editing, and phylogeny can be accessed via CloudBioLinux on Ama-

zon EC2, Eucalyptus and VirtualBox. This rich set of bioinformatics packages reduces effort to prepare and configure data analysis environments. The latest version of BioLinux version 8 runs on Ubuntu14.04 with up-to-date versions of many packages including R, QIIME, Mothur, Jalview, Artemis, BLAST and Bowtie-Bio. Galaxy workflow, which is one of the main components in CloudBioLinux, is easily deployable using CloudBioLinux as well as CloudMan.

3.4 Galaxy Workflow with Globus toolkits

Bo Liu et al [7] extends Galaxy workflow with Globus toolkits and Chef orchestration for large-scale next-generation sequencing (NGS) analyses. Globus Transfer, which is a file transfer service using GridFTP, is included in this workflow platform to support fast and reliable file transfer for large datasets such as NGS data. Globus Transfer outperforms at least six times faster in transferring data than HTTP and FTP. Globus Provision (GP) automates creating and scaling of the workflow platform in the Amazon cluster with Chef orchestration toolkit. Initial configurations and installations are written in the scripts. Prerequisite packages are installed by running the scripts as well as configuring the clusters. GP consists of the scripts, which are called recipes in Chef, to enable automated provision on cloud-based platforms like Amazon EC2. Two workflows were introduced with this system to demonstate scalability and automation. CRData workflow showed the easy use of BioConductor R scripts with a Galaxy workflow and RNA-Seq workflow showed the integration of CummeRbund packages.

4. CONCLUSIONS

These days, individual research laboratory is able to generate terabytes of data (or even larger), which is no suprises to new sequencing technologies in genomic research. High performance computation environments keep improving on processing large-scale data at low cost. The combination of MapReduce and cloud computing facilitates fast and efficient parallel processing on the virtual environment for terabyte-scale data analysis in bioinformatics, if the analysis consists of embarrassingly parallel problems. MapReduce framework is suitable for the simple and dividable tasks such as read alignment, sequence search and image recognition. Easy-to-use methods and user-friendly cloud platforms have been provided to researchers so that they can easily have access to the cloud with their large data sets uploaded on the cloud in a secure manner. Scientific workflow may focus on improving data transfer and handling tasks regarding these usability problems. More challenges are expected to deal with data storage and analysis since it grows at unprecendented scales.

5. REFERENCES

[1] E. Afgan, D. Baker, N. Coraor, B. Chapman, A. Nekrutenko, and J. Taylor. Galaxy cloudman: delivering cloud compute clusters. *BMC bioinformatics*, 11(Suppl 12):S4, 2010.

[2] D. Field, B. Tiwari, T. Booth, S. Houten, D. Swan, N. Bertrand, and M. Thurston. Open software for biologists: from famine to feast. *Nature biotechnology*, 24(7):801–804, 2006.

[3] J. Goecks, A. Nekrutenko, J. Taylor, et al. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86, 2010.

[4] K. Krampis, T. Booth, B. Chapman, B. Tiwari, M. Bicak, D. Field, and K. E. Nelson. Cloud biolinux: pre-configured and on-demand bioinformatics computing for the genomics community. *BMC bioinformatics*, 13(1):42, 2012.

[5] B. Langmead, M. C. Schatz, J. Lin, M. Pop, and S. L. Salzberg. Searching for snps with cloud computing. *Genome Biol*, 10(11):R134, 2009.

[6] P. Lawrence, editor. *Workflow Handbook 1997*. John Wiley & Sons, Inc., New York, NY, USA, 1997.

[7] B. Liu, R. K. Madduri, B. Sotomayor, K. Chard, L. Lacinski, U. J. Dave, J. Li, C. Liu, and I. T. Foster. Cloud-based bioinformatics workflow platform for large-scale next-generation sequencing analyses. *Journal of biomedical informatics*, 2014.

[8] W. Lu, J. Jackson, and R. Barga. Azureblast: a case study of developing science applications on the cloud. In *Proceedings of the 19th ACM International Symposium on High Performance Distributed Computing*, pages 413–420. ACM, 2010.

[9] A. Matsunaga, M. Tsugawa, and J. Fortes. Cloudblast: Combining mapreduce and virtualization on distributed resources for bioinformatics applications. In *eScience, 2008. eScience'08. IEEE Fourth International Conference on*, pages 222–229. IEEE, 2008.

[10] H. Nordberg, K. Bhatia, K. Wang, and Z. Wang. Biopig: a hadoop-based analytic toolkit for large-scale sequence data. *Bioinformatics*, 29(23):3014–3019, 2013.

[11] M. C. Schatz. Cloudburst: highly sensitive read mapping with mapreduce. *Bioinformatics*, 25(11):1363–1369, 2009.

[12] A. Schumacher, L. Pireddu, M. Niemenmaa, A. Kallio, E. Korpelainen, G. Zanetti, and K. Heljanko. Seqpig: simple and scalable scripting for large sequencing data sets in hadoop. *Bioinformatics*, 30(1):119–120, 2014.

[13] M. S. Wiewiórka, A. Messina, A. Pacholewska, S. Maffioletti, P. Gawrysiak, and M. J. Okoniewski. Sparkseq: fast, scalable, cloud-ready tool for the interactive genomic data analysis with nucleotide precision. *Bioinformatics*, page btu343, 2014.

[14] S. Zhao, K. Prenger, L. Smith, T. Messina, H. Fan, E. Jaeger, and S. Stephens. Rainbow: a tool for large-scale whole-genome sequencing data analysis using cloud computing. *BMC genomics*, 14(1):425, 2013.