Biology and Bioinformatics

- Overview
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Issues/Gaps
- Crossbow project: SNPs searching with cloud computing 1, 2
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Cloud computing paradigms for pleasingly parallel biomedical applications 3
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Cloud technologies for bioinformatics applications 4
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Hybrid cloud and cluster computing paradigms for life science applications 5
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Cloud-scale RNA-sequencing differential expression analysis with Myrna 6
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- AzureBlast 7
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Current applications of Hadoop/MapReduce/HBase framework in bioinformatics 8
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- The case for cloud computing in genome informatics 9
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- High-Performance cloud computing 10
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
  - Issues/Gaps
- Cloud and heterogeneous computing for the big data problems in biology 11
  - Summary
  - Workflow
  - Data
  - Cloud platform
  - Cloud performance
Overview

Summary

- There are a lot of practice of implementing biology/bioinformatics applications on cloud computing resources. This is because:
  - Many bioinformatic applications such as DNA sequencing require processing of large data throughput
  - There are a lot of existing tools (e.g., open source tools like Hadoop implementation of MapReduce) which can be easily run in the cloud
  - Many opensource projects can be easily implemented in cloud such as Myrna, CloudBLAST (Langmead et al. 2010), and Galaxy (Afgan et al. 2010).

- There are some review papers in this area (Stein 2010, Schatz et al. 2010)

Workflow

- MapReduce framework are frequently used.
- Other MapReduce extensions are also used such as Microsoft Dryad (Qiu et al. 2009 and Lu et al. 2010) and Twister (Qiu et al. 2010).

Data

The data throughput (such as DNA sequences) in biology/bioinformatic applications are usually very big.

Examples:

- A human sample comprising 2.7 billion reads were genotyped by Crossbow in about 4 hours including data uploading time in Amazon Cloud and the cost is about $85 (Schatz et al. 2010 and Langmead et al. 2009).
- 1 PB of data can be traversed on a 1,000 node instance on Amazon EC2 within ~350 mintues and cost $2,040 (Schadt et al. 2011).

Cloud platform

- Amazon Cloud service is the most popular cloud platform (Schatz et al. 2010, Langmead et al. 2009, Gunarathne et al. 2010, Qiu et al. 2010, Langmead et al. 2010, Vecchiola et al. 2009, Schadt et al. 2011, Nguyen et al. 2011, Afgan et al. 2010) in the biology/bioinformatics area since most bioinformatics use linux-based system and technologies, and it is convenient to implement those technologies on the Amazon cloud due to the IaaS property of Amazon AWS.
- Windows Azure is also used (Qiu et al. 2009, Qiu et al. 2010, Lu et al. 2010). Dryad is a Azure-based implementation of MapReduce framework.
- Applications on other cloud services are also investigated such as FutureGrid (Qiu et al. 2010) and Magellan (Taylor et al. 2010).

Issues/Gaps

- Data transfer (Schatz et al. 2010)
  The input data (which is usually with large size) must be deposited in a cloud resource to run a cloud program over the data set. So the
compatibility between data-generation and transfer speeds achievable must be assessed.
Option 1: High-speed Internet. Such as Internet2 and JANET.
Option 2: Ship physical hard drives to the cloud vendor (http://aws.amazon.com/importexport).

- Data security and privacy (Schatz et al. 2010)
  Policy on security for data storing and processing in the cloud is still under development. Users need to determine whether cloud computing is compatible with any privacy or security requirements associated with their institutions.
- Applications redesigning (Schatz et al. 2010)
- Usability (Schatz et al. 2010)

Crossbow project: SNPs searching with cloud computing $^{1,2}$

Summary

Crossbow is a Hadoop-based software tool combining short read aligner Bowtie with the accuracy of the SNP caller SOAPsnp to perform alignment and SNP detection for multiple whole-human datasets per day.
In this work, Crossbow analyzes data comprising 38-fold coverage of a Han Chinese male genome in 4 hours 30 minutes (including transfer time) using a 320-core cluster rented from Amazon EC2.
The Crossbow project uses cloud computing (with MapReduce and Hadoop) to efficiently parallelize existing sequence alignment and genotyping algorithms.
By taking advantage of commodity processors available via cloud computing services, Crossbow condenses over 1,000 hours of computation into a few hours without requiring the user to own or operate a computer cluster.
Also, running on standard software (Hadoop) and hardware (EC2 Instances) makes it easier for results reproduction and customized analysis with Crossbow.

Workflow

- "Map-shuffle-scan" framework

Data

- 38-fold coverage of a Han Chinese male genome
- Data: 2.66 billion reads (~85 Gb)

Cloud platform

- Cloud platform: Amazon EC2/S3

Cloud performance

Computation was performed both locally and in Amazon EC2 cluster.

- Local cluster
  - Hadoop 0.20 cluster with 10 worker nodes
  - Each node: 4-core 3.2 GHz Intel Xeon (40-core total)
  - 64-bit Redhat Enterprise Linux Server 5.3
  - Each node: 4 GB memory and 366 GB local storage available for the HDFS
  - Connection: gigabit ethernet
  - Performance: Requires about 1 day of wall clock time to run

- Amazon EC2
  - Amazon EC2 service on clusters of 40 nodes
  - Each node: EC2 Extra Large High CPU Instance (High-CPU XL)
  - Each node: a virtualized 64-bit computer with 7 GB of memory and the equivalent of 8 processor cores clocked at approximately 2.5 to 2.8 Ghz.
  - Cost: $0.68 per node per hour (2009) (Actually the price is the same as 2011, see current Amazon price)
  - Overall performance in Amazon EC2: Discovered 3.7M SNPs in one human genome for ~$100 in an afternoon. Accuracy validated at >99% (See the table in the figure below)

Issues/Gaps

Cloud computing paradigms for pleasingly parallel biomedical applications $^{3}$

Summary
Two parallel biomedical applications: assembly of genome fragments and dimension reduction in the analysis of chemical structures are presented in the paper which are implemented utilizing cloud infrastructure service based utility computing models of Amazon AWS and Windows Azure. Apache Hadoop and Microsoft DryadLINQ are used as data processing framework.

**Workflow**
- Cap3: assembles DNA sequences by aligning and merging sequence fragments to construct whole genome sequence
- GTM (Generative Topographic Mapping) and MDS (Multidimensional Scaling): dimension reduction algorithms

**Data**
- Cap3: Assembling 200 FASTA files (458 reads per file)
- GTM interpolation: Processing 26 million Pubchem data points

**Cloud platform**
- Amazon AWS (EC2, S3 and SQS)
- Azure
- Apache Hadoop is used
- DryadLINQ is used (Dryad is a framework developed by Microsoft as a general-purpose distributed execution engine)

**Cloud performance**
- Cap3: Process 4096 FASTA files (~1GB)
  - EC2: 58 minutes, total cost $11.19
  - Azure: 59 minutes, total cost $15.77
  - Local cluster using Hadoop: 10.9 minutes, total cost: $650,000
- GTM Interpolation (More memory-intensive than Cap3)
  - When input data is larger, Hadoop & DryadLINQ have an advantage of data locality based scheduling over EC2
- MDS interpolation can only be implemented using Azure since it requires .net framework

**Issues/Gaps**

NA

Cloud technologies for bioinformatics applications

**Summary**
Microsoft Dryad (an implementation of extended MapReduce from MS) and Azure are applied to three bioinformatics applications

**Workflow**
- PhyloD
- EST
- Alu clustering

**Data**

**Cloud platform**
- Microsoft Azure are used
- Microsoft Dryad
- Hadoop is used to compare with Dryad.

**Cloud performance**
- The flexibility of clouds and MapReduce makes them the preferred approaches over traditional MPI approaches.

**Issues/Gaps**
- The experiment is carried on Windows platform. Most of other bioinformatics applications used Linux based systems and technologies.

Hybrid cloud and cluster computing paradigms for life science applications

**Summary**
A hybrid cloud and cluster computing paradigms is designed for life science applications. Non iterative cases are tested with Amazon, Azure and FutureGrid. Twister Iterative MapReduce is benchmarked against basic MapReduce (Hadoop) and MPI in information retrieval and life sciences applications.

**Workflow**
- Commercial clouds support "massively parallel" or "many tasks" applications which are loosely coupled.
- Hybrid cloud-cluster architecture can be used to link MPI and MapReduce components.

**Data**

**Cloud platform**
- Amazon
- Azure
- FutureGrid
- Twister is developed. Twister is a extension of MapReduce that can deal with iterative structure.

**Cloud performance**
- Amazon EC2

**Issues/Gaps**

**Cloud-scale RNA-sequencing differential expression analysis with Myrna**

**Summary**
- Myrna is a cloud-computing pipeline for calculating differential gene expression in large RNA-Seq data sets.
- Cloud computing is good for algorithm which can be made to run efficiently on many loosely coupled processors.

**Workflow**

Myrna workflow:
- Preprocess
- Align - using Bowtie
- Overlap
- Normalize
- Statistical analysis
- Summarize
- Postprocess

**Data**
- HapMap RNA-Seq dataset: 1.1 billion 35-bp unpaired reads

**Cloud platform**
- Amazon EC2
- Amazon S3: Store and preprocess the input data

**Cloud performance**
- Transfer cost: $11: $6.40 in cluster rental fees and $4.30 in data transfer fees
- Transfer time: 43GB from public HTTP server located at the University of Chicago to an S3 repository located in US in 1 hr 15 min (82 Mb/s)*
- Calculation: 1.1 billion RNA-seq reads in less than 2 hours of wall clock time for about $66 (320 processor cores used)

**Issues/Gaps**
- Cloud data transfers are inconvenient and sometimes too slow
- Privacy concerns (internal review board requirements)
AzureBlast

Summary

- BLAST algorithm is implemented on Windows Azure cloud platform.
- BLAST is a popular life science algorithm used to discover the similarities between the two bio-sequences.
- The implementation: AzureBlast is a parallel BLAST engine running on the Windows Azure cloud.
- No high-level programming models or runtimes such as MapReduce are used.

Workflow

BLAST

Data

- NR database: non-redundant protein sequence database of 10,427,007 sequences (3,558,078,962 total letters, about 10 GB)

Cloud platform

- Windows Azure
- PaaS in contrast to IaaS (Amazon AWS)
- Software can be implemented as a Service application. (In contrast, Amazon EC2 provides a host for virtual machines)
- Current data center network architectures are optimized for high scalability at low cost, and are not optimized for low latency communication

Cloud performance

- With Extra large Azure instance (CPU: 8 x 1.6 GHz, Memory: 14 GB, Storage: 2,040 GB)
  - Performance: ~50 seq/min with 300 sequences
  - Cost: 3000 seq$/ with 300 sequence ($0.1/seq)
  - Scalability:
- Read and Write throughput of Blob Storage: >200 MB/sec reading, 100 MB/sec writing (64 instances)

Issues/Gaps

NA

Current applications of Hadoop/MapReduce/HBase framework in bioinformatics

Summary

- Overview of the Hadoop/MapReduce/HBase framework
- Use them in cloud via data upload to cloud vendors who have implemented Hadoop/HBase

Workflow

Data

Cloud platform

Opensource projects built on top of Hadoop:

- Hive: a data warehouse framework at Facebook, designed for batch processing not online transaction processing
- Pig: for batch processing
- Mahout and other expansions to Hadoop
- Cascading
- HBase:
  - Apache Hadoop-based project
  - Modeled on Google's BigTable database
  - distributed, fault-tolerant scalable database
  - built on top of the HFDS file system

Cloud performance

Issues/Gaps

- DOE is exploring scientific cloud computing in the Magellan project
The case for cloud computing in genome informatics

Summary
This is a review paper.

Workflow
Data
Cloud platform
Cloud performance
Issues/Gaps

High-Performance cloud computing

Summary
- An enterprise cloud computing solution, Aneka is described.
- Case study of Aneka for classification of gene expression data and the execution of fMRI brain imaging workflow

Workflow
- Classification of Gene Expression Data
  - CoXCS
  - fMRI imaging
    - Only the spatial normalization is modeled

Data
- Gene expression data classification: BRCA and Prostate dataset
- fMRI: 40 brain images of 20GB

Cloud platform
- Aneka cloud on top of Amazon EC2 infrastructure
  - PaaS
  - Support for Compute
  - User access interface: Web APIs, Custom GUI

Cloud performance
Issues/Gaps

Cloud and heterogeneous computing for the big data problems in biology

Summary
1 PB of data can be traversed on a 1,000 node instance on Amazon EC2 with ~350 minutes and cost $2,040.

Workflow
Data
Cloud platform
Cloud performance
Issues/Gaps
CloudBlast

Summary
- CloudBlast is an implementation combining MapReduce and Virtualization on Distributed Resources for Bioinformatics applications.

Workflow

Data

Cloud platform

Cloud performance

Issues/Gaps

CloudAligner

Summary
- CloudAligner is a Hadoop MapReduce based tool for sequence mapping.
- CloudAligner can be implemented on cloud services such as Amazon EC2

Workflow

1. Upload CloudAligner to Amazon S3
2. Create job flows in Amazon Elastic MapReduce to execute it.

Data
NA

Cloud platform
Amazon AWS

Cloud performance
- CloudAligner is faster than CloudBurst in the cloud.

Issues/Gaps

Galaxy Cloudman

Summary
- Introduced Galaxy CloudMan, an integrated solution that leverages existing tools and package on cloud resources.
- Interactions with Galaxy CloudMan and the cloud cluster management is performed through a web based user interface, so no computational expertise is needed.

Workflow

1. Use the AWS Management Console to start a master EC2 instance
2. Use the CloudMan web console on the master instance to manage the cluster size

Data
NA

Cloud platform
- Amazon EC2
Cloud performance

Issues/Gaps

Galaxy Project

Summary

- Galaxy for processing of greatly variable amounts of data over time.
- Galaxy instantiated on cloud computing infrastructure such as Amazon EC2

Workflow

- Start an EC2 instance
- Use Galaxy CloudMan web interface on the started EC2 instance to manage the compute cluster
- Use Galaxy on the cloud as a personal instance

Data

Cloud platform

- Amazon EC2

Cloud performance

- pay-as-you-go
- Scalable computing resource

Issues/Gaps

References

General Review Papers:
1. Stein L.D. *Genome Biol.* (2010)

Hadoop Review Papers:

Individual Stories:
4. Qiu, J. *et al.* *BMC Bioinformatics* 11 2010
12. Galaxy and Cloud