Extending BioVLab Cloud Workbench to a TeraGrid Gateway

Suresh Marru, Patanachai Tangchaisin, Marlon Pierce, Heejoon Chae, Kenneth Nephew, Sun Kim
Indiana University
www.collab-ogce.org
Outline

• Introduce OGCE
• Introduce BioVLAB
• Bio Informatics workflows on EC2
• mCpG computational needs – TeraGrid
• Experiences with EC2 to TeraGrid transition
• Lessons Learned
Gateway Computing Environments Workshop 2011

• Gateway(Grid) Workshop Series at Supercomputing Conferences since 2005.

• Format: presentation of peer-reviewed papers.
  – Papers published in ACM or IEEE digital proceedings.

• Next one will be in Nov 2011 at SC11. See http://www.collab-ogce.org/gce11.
# Open Gateway Computing Environments Collaboration

<table>
<thead>
<tr>
<th>Institution</th>
<th>People</th>
</tr>
</thead>
<tbody>
<tr>
<td>Indiana University</td>
<td>Marlon Pierce, Suresh Marru, Raminder Singh, Lahiru Gunathilake, Gerald Guo, Xuan Wu</td>
</tr>
<tr>
<td>NCSA/UIUC</td>
<td>Sudhakar Pamidighantam, Shaowen Wang, Yan Liu, Liang Yu, Ye Fan</td>
</tr>
<tr>
<td>Purdue University</td>
<td>Carol Song, Lan Zhao, David Braun, Shawn Wu</td>
</tr>
<tr>
<td>UTHSCSA</td>
<td>Emre Brookes, Borries Demeler, Dan Zollars, Gary Gorbet</td>
</tr>
</tbody>
</table>
OGCE Funds Software Lifecycle

1. Next Generation Requirements
2. Gateway Software Development
3. Software Testing
4. Deployment
5. Integration Testing
# OGCE: Science Gateway Suite

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>OGCE Gadget Container</td>
<td>An OpenSocial and Google gadget-compatible Web container for running Web gadgets.</td>
</tr>
<tr>
<td>GFAC</td>
<td>A Web service for generating, securely invoking, and managing the lifecycle of scientific applications on Grids and Clouds.</td>
</tr>
<tr>
<td>Workflow Tools</td>
<td>Composer (XBaya), enactment engines, event system, and service registry to support scientific workflows on Grids and Clouds.</td>
</tr>
</tbody>
</table>
OGCE Gateway Suite
What Are Ingredients for Sustainable Software (Besides Funding)?

- Committed user base
- Open source or free licensing
  - Apache or GPL license families
- Code should use supporting tools from SourceForge, GitHub, Google Code, etc
  - Version control software, trouble tickets, mailing lists, IRC, etc.
- Project management philosophy
  - Meritocracy, benevolent dictator, etc.
- Reproducible build and test
- Open community
  - Overlooked but key
  - Process for attracting and rewarding new developers
  - Encourage and reward projects for diversifying developer base
  - Encourage cross-project collaboration
Apache Rave is a joint submission of OneHippo, MITRE, SURFNet, and IU.

- http://incubator.apache.org/rave/
- Unusual submission

It’s a generalization of the OGCE Gadget Container
Champion: Ate Douma
Mentors:
- Sylvain Wallez
- Upayavira
- Ross Gardler
- Hadrian Zbarcea
Apache Airavata consists of the workflow tools developed by the OGCE
- http://incubator.apache.org/airavata/
- More traditional submission

- Champion: Ross Gardler
- Mentors:
  - Alek Slominski
  - Ate Douma
  - Sanjiva Weerawarna
  - Paul Fremantle
  - Chris A. Mattmann
More Information

• OGCE Web Site: http://www.collab-ogce.org
• News Feed/Blog: http://collab-ogce.blogspot.com
• Contact us:
  – ogce-discuss@googlegroups.com
  – http://groups.google.com/group/ogce-discuss/
BioVLAB

- BioVLAB- A Reconfigurable Cloud Computing Environment for Bioinformatics.
- Collaboration between OGCE and Prof. Sun Kim’s group, School of Informatics and Computing, Indiana University.
- Used for executing Bioinformatics applications in Cloud since late 2006.
- BioVLAB-Microarray published in eScience 2008
- BioVLAB-MMIA, BioVLAB-mCpG – 2011
Data-driven Science

• A large volume of data
  – Biological databases are growing at a very fast exponential rate, which is greater than the rate of improvement of microprocessors. (BMC Bioinformatics 2008, 9(Suppl 2):S10 doi:10.1186/1471-2105-9-S2-S10)

• Data processing and analysis are main problem

• Example
  – Next generation sequencing data
    ➔ String pattern matching
      • Mapping billions of short reads (strings) to human reference genome
  – Gene expression data
    ➔ Graph traversal
      • Nodes: 2M probes, edges: relationship between probes
Cloud as Computational Platform

- Software (SaaS), Platform (PaaS) and Infrastructure (IaaS) as a service
- **Characteristics**
  - On-demand
  - Pay as you go billing model
  - Rapid elasticity
  - Easy network access
  - Location independent

- **Spectrum of Clouds**
  - Amazon EC2 (instruction set VM), Microsoft Azure (Bytecode VM), Google AppEngine (Framework VM)

http://edgewatertech.files.wordpress.com/2009/03/cloud-computing-diagram.jpg
BioVLAB Architecture
How to use BioVLAB

1. Get an Amazon account
2. Download workflow
3. Provide Amazon EC2 Credentials
4. Run the workflow

Usability!
BioVLAB Application Development Procedure

User

• Develop a command line app.

Admin

• Install the app. in Amazon EC2
• Let the app. store any output to Amazon S3 / Microsoft Application-Based Storage
• Make a virtual machine image
• Register the app. by using Gfac

User

• Instantiate EC2 and run the app. by using XBaya
Executing EC2 jobs through GFAC
BioVLAB-Microarray

- Analysis of high throughput microarray experiment
- Multiple tasks in a single batch
- Output of a task can be plugged into another task
- Repeat the same set of tasks with small changes of parameters
• MicroRNAs (miRNAs)
  • small (19-22 nucleotide) non-protein-coding RNA molecules
  • regulate the expression of specific gene products
  • effect translational blockade or message degradation
• MMIA: microRNA and mRNA integrated analysis

BioVLAB-MMIA
• Computation in the Cloud
• MMIA expertise in workflow
OSU-IU Center for Cancer Systems Biology

The mCpG-SNP-EXPRESS package

Integrated analysis of DNA methylation, SNP, and gene expression data is a complicated data mining process. We have developed a package called, mCpG-SNP-EXPRESS, that helps perform the integrated analysis.
Workflows of mCpG-SNP-EXPRESS

Pairwise phenotype analysis

Multiple phenotype analysis

0. gene set
1. Pdiffexpr
2. Pdiffmethyl
3. PdiffSNP
4. Infodiff expr
5. Infodiff methylSNP
6. Infodiff SNP
7. Infodiff methyl
8. gene set Pdiffexpr
9. gene set Pdiffmethyl
10. gene set PdiffSNP
11. gene set Infodiff expr
12. gene set Infodiff methylSNP
13. gene set Infodiff SNP
14. gene set Infodiff methyl
15. Common Pairwise
16. Common Infocombined
17. Common Infomultiple
18. gene set pairwise
19. gene set infocombined
20. gene set infomultiple
21. Methylated region Summary
22. Visualization (Integrative Genomic Viewer, pathway-express)
23. Profiles per gene (SNP, methyl)
24. SNP, Methylation Heatmap
Explaining steps in the workflow

**Differentially expressed genes** are calculated for pairwise and multiple phenotypes.

- **Pdiffexpr** for pairwise phenotypes (task 1 in the workflow below) is calculated by first using bioconductor to perform RMA pre-processing on the raw CEL files. This performs a background adjustment and quantile normalization on the probe measurements. The fold change is then calculated among each subtype. If there are multiple CEL files belonging to the same subtype then the mean expression level is used. A fold change value below -0.5 is then considered to be down regulated. With respect to a phenotype p1 (average of multiple cell lines of p1) using expression levels:

  - **Infodiffexpr** for multiple phenotypes (task 4 in the workflow below) is calculated by computing a normalized entropy considering average expression levels as gene expression probabilities.

**Differentially methylated genes** are calculated for pairwise and multiple phenotypes.

- **Pdiffmethyl** for pairwise phenotypes (task 2 in the workflow below) is calculated by the presence and absence of hypermethylated peaks determined by MACS peaks.

- **Infodiffmethyl** for multiple phenotypes (task 7 in the workflow below) is calculated by computing a normalized entropy considering methylation levels measured by the number of mapped reads in a bin (by default 1kb) as methylation probabilities.

**Differentially mutated (SNP) genes** are calculated for pairwise and multiple phenotypes.

- **PdiffSNP** for pairwise phenotypes (task 3 in the workflow below) is calculated by computing Pearson's correlation using vectors of the number of predicted SNPs using VCFTools for for pairwise phenotypes.

- **InfodiffSNP** for multiple phenotypes (task 6 in the workflow below) is calculated by computing a normalized entropy considering the number of predicted SNPs in a bin (by default 1kb) as SNP probabilities.

**Differentially methylated and mutated (SNP) genes** (task 5 in the workflow below) are calculated for multiple phenotypes by computing either a ratio of normalized entropies for methylation (task 7) and SNP (task 6) or a relative entropy of methylation and SNP.
Differentially methylated, SNP, and expressed gene sets

To guide the data mining process, we need to narrow down data to potential candidate regions. The package is designed to take a gene set as input, thus the analysis can be performed iteratively when a new gene set of interest is identified.
BioVLAB-mCpG
BioVLAB Summary

• **Usability (Reconfigurable environments)**
  – As an adoption of the SaaS model of Cloud Computing for BioVLAB, end-users only need to launch the pre-composed BioVLAB workflows. With XBaya, users can easily customize it by modifying just a few components and input parameters.

• **Flexibility (Full privileges)**
  – As a way of the IaaS model, BioVLAB workflow developers can have flexibility for handling computing resources and implementing applications with Amazon Cloud. They can choose specific systems resources to satisfy their needs with a fully controlled access power.

• **Reducing processing time & Cost effective**
  – Users can have number of servers, and control their usage time as they want. That reduces researching cost and initial time to construct physical infrastructure for research.
Switch between Cloud & Grid

- Users
- OGCE based Gateway
- Clouds
- TeraGrid
GFac Current & Future Features

Color Coding:
- **Blue** Existing Features
- **Red** Planned/Requested Features

- Input Handlers
- Registry Interface
- Scheduling Interface
- Monitoring Interface
- Output Handlers
- Fault Tolerance
- Data Management Abstraction
- Job Management Abstraction
- Auditing
- Checkpoint Support
- Unicore
- Condor

Globus
Campus Resources
Amazon Eucalyptus

XSUL/Apache Axis2
TeraGrid
Indiana University
Pervasive Technology Institute
OGCE Layered Workflow Architecture: Derived from LEAD Workflow System
Putting It All Together
Gateways and Clouds

• Amazon EC2 and S3 provide infrastructure for **gateways** and **development efforts**.
  
• OGCE software lets you expose your applications on VMs as network accessible services.
  – And control VM lifecycles.

• Same software will allow you to move to the TeraGrid if you need larger resources.
  – Or you can use Amazon as a “spill-over” service if you need urgent resources.

• This can be done with OGCE off the shelf.
  – BioVLAB will illustrate these issues.
Demos Next