A User Story

Globus and Galaxy in the Project
Mr.SymBioMath

http://www.mrsymbiobiomath.eu/
Agenda

- Mr.SymBioMath – an Overview
- Galaxy and Workflows
- Galaxy and Globus
- Demo
- Conclusion
Mr. SymBioMath – an Overview
Mr. SymBioMath

- Industry Academia Partnership and Pathways Project
- Funded in the 7th Framework Program under grant agreement number 324554
- Mr. SymBioMath
  - Mr → MapReduce
  - Sym → Symbolic Computing
  - Bio → Bio*
  - Math → Mathematics
Mr. SymBioMath

- Big-Data and computationally intensive scientific applications
- Cloud and high performance computing
- Visualization
- Application domain: bioinformatics and biomedicine

Enhance industry-academia cooperation in terms of research training, career development and knowledge sharing
Mr. SymBioMath –
Objectives and Challenges

• Research on Big-Data problems and computational intensive applications in HPC
  - New software libraries for out-of-core management of Big-Data
  - Reduce the computational space and memory demand
Mr. SymBioMath – Objectives and Challenges

• New data processing applications in comparative genomics & biomedicine:
  
  – Modelling evolutionary events (SNP, mutations, translocations, deletions, duplications, inversions, etc.)
  – Understand evolution and extrapolate from the divergence between species (phylogenetic trees)
  – Inter-genome distances to estimated divergence
  – Gene-gene interactions studies correlated with phenotype data
Mr. SymBioMath – Objectives and Challenges

• Implement and packing commercial applications prototypes:
  - Interfacing, visualisation, integration of external data sources and functional annotations.
  - Adverse reactions to treatment in allergic patients (biological markers)
  - Multi & meta-genome analysis; biomedical applications (GWAS)
## The Consortium

<table>
<thead>
<tr>
<th>Nº</th>
<th>Partner</th>
<th>Acronym</th>
<th>Country</th>
<th>Scientist in charge</th>
<th>Main expertise</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>University of Malaga</td>
<td>UMA</td>
<td>ES</td>
<td>Oswaldo Trelles</td>
<td>HPC, Bioinformatics, Comparative genomics applications</td>
</tr>
<tr>
<td>2</td>
<td>RISC Software GmbH</td>
<td>RISC</td>
<td>AT</td>
<td>Michael Krieger</td>
<td>Cloud and infrastructure provider, Tree comparison algorithms.</td>
</tr>
<tr>
<td>3</td>
<td>Johannes Kepler University</td>
<td>JKU</td>
<td>AT</td>
<td>Sepp Hochreiter</td>
<td>Machine learning (biomed apps.: Genotype-Phenotype correlations)</td>
</tr>
<tr>
<td>4</td>
<td>Integromics, S.L.</td>
<td>ITG</td>
<td>ES</td>
<td>Eduardo Gonzalez-Couto</td>
<td>Visual Analytics WFs, platform integration and commercial packing</td>
</tr>
<tr>
<td>5</td>
<td>Hospital Carlos Haya</td>
<td>HCH</td>
<td>ES</td>
<td>Miguel Blanca</td>
<td>Clinical data on allergies, medical data analysis</td>
</tr>
<tr>
<td>6</td>
<td>Leibniz-Rechenzentrum</td>
<td>LRZ</td>
<td>DE</td>
<td>Christoph Anthes</td>
<td>Advanced 3D and VR visualization</td>
</tr>
<tr>
<td>7</td>
<td>National Lab for Scientific Computation</td>
<td>LNCC</td>
<td>BR</td>
<td>Ana Ribeiro de Vasconcellos</td>
<td>Genomics: sequence data provider, final user tester</td>
</tr>
<tr>
<td>8</td>
<td>Institute of Computer Graphics JKU</td>
<td>ICG</td>
<td>AT</td>
<td>Marc Streit</td>
<td>Visual analytics, bioinformatics</td>
</tr>
<tr>
<td>9</td>
<td>IBM Austria</td>
<td>IBM</td>
<td>AT</td>
<td>Olaf Hahn</td>
<td>Big Data Analytics</td>
</tr>
</tbody>
</table>
Galaxy and Workflows
Galaxy

- Galaxy is an open, web-based platform for data intensive biomedical research. (see http://galaxyproject.org/)

- Galaxy is developed by the Nekrutenko lab in the Center for Comparative Genomics and Bioinformatics at Penn State and the Taylor lab at Johns Hopkins University, along with contributions from the community.

- For Mr.SymBioMath:
  - Workflow engine
  - Web-based user interface
  - Flexible Execution of Scripts and Executables
  - Standard tool which is often used within Bioinformatics / Biomedicine
Customisation
Example: Biomedicine

• Background
  - NSAIDs hypersensitivity
  - Genome wide association study (GWAS) on drug hypersensitivity (HCH-FIMABIS/IBIMA)

• One of the performed analyses on this dataset:
  - CLOSER: GWAS raw-data to processed genotype file (vcf) (RISC/UMA/HCH/JKU)
Experimental Design

Patients

Patient DNA

Non-patients

Non-patient DNA

Disease-specific SNPs

Non-disease SNPs

Compare differences to discover SNPs associated with diseases.
CLOSER

- **Output: VCF**
- **Standard format to report genomic variants**
- **Describes variant calls and meta data of analysis**

Takes a long time
Embarrassingly parallel
Perfect for the CLOUD
CLOSER

- Workflow is used to produce a VCF file from the raw probe intensity measurements

- Resulting VCF can be used for downstream analysis

- Pipeline is publicly available
- Presented at international conference
- Published as proceedings paper
Example: Comparative Genomics

SeqX
- words
- Sort
- W2HD

SeqY
- words
- Sort
- W2HD

Hits sort frags
Input data for example workflows

• CLOSER (1 workflow)
  - CEL-files (roughly 70 MB)
  - models (> 100 MB)
  - CDF (> 100 MB)
  - probes (KB size)
  - HG – fasta format (~3 GB)
  - pickle file (> 100 MB)

• Comparative Genomics
  - 2 Genome files (KB - GB)
Galaxy and Globus
Challenges

• What do users want?
  – short execution time
    • dynamic provisioning (auto-scaling) of resources
  – efficient data transfer
    • Is HTTP or FTP enough?
  – secure and large (almost unlimited) data storage
  – single authentication / single password
Challenges

**What do administrators want?**

- ease of administration
  - (almost) no new software to develop
  - use of standard-features
- „self-containing“ services
  - clear security and system borders
  - Separate environments for services – also necessary for scalability
- security updates from original software providers
  - use of standard software
System Architecture

- GridFTP
- MyProxy
- LDAP
- Database
- Web-server
- NFS
- PBS
- Worker
- Worker
...the „devil“ is in the details

• Usernames in Galaxy are e-mail addresses
  – Linux and usernames with special characters???

• Galaxy has to access FTP-user home directories as the process user (for example: „galaxy“)

• Home directories of GridFTP users should be only accessible by owner
Access rights manipulation

- GridFTP

- Webserver

- Galaxy
Demo
Demo
Demo
Demo

Galaxy / Mr.SymbioMath

Files_2_FTP (version 0.0.1)

Folder To Create on FTP server:

This folder will hold the transferred data

Data Files to downloads

Add new Data Files to download

Execute

What it does

Copies selected datasets to your FTP account on this galaxy instance. You can then use an FTP client to download the files. The files will be placed in a subfolder with the specified name. If no name is specified, a folder with the current data will be used. Input

Any dataset from the history

Output

An html file, with an overview of the copied files.

This tool was installed from a Tool Shed, you may be able to find additional information by following this link:

http://toolshed.g2.bx.psu.edu/view/geert-vandeweyer/files_to_ftp

You are over your disk quota. Tool execution is on hold until your disk usage drops below your allocated quota.

This history is empty. You can load your own data or get data from an external source.
Conclusion

● Possible to rebuild „similar“ installation to Globus Genomics on-site.

● Possible to use standard software and standard installations.

● TBD:
  - Auto-Deployment of whole setup (with HEAT for OpenStack)
  - Auto-Scaling of PBS workers
Thank you!

Questions?