Running on HPC Galaxy-based workflows for predictive biomarkers from RNA-Seq clinical data

Calogero Zarbo, Marco Chierici, Cesare Furlanello
Fondazione Bruno Kessler, Trento, Italy

GCC 2013

1. Introduction

We present a Galaxy-based framework for clinical diagnostic on big datasets of RNA deep-sequencing (RNA-Seq) data. The framework implements the Data Analysis Plan (DAP) of the FDA SEQC project for Array and RNA-Seq gene expression analysis pipelines: the goal of the DAP is to provide a principled sequence of machine learning methods for predictive biomarker identification.

Our solution extends functions from the paramiko v1.7.5 module to transport the Galaxy workflow processes through a virtual bash shell, by an SSH data stream connection, on a high performance computing (HPC) system, e.g. a Linux cluster with the SGE queue system. The goal is to achieve parallelization with one workflow, keeping the same flexibility of a direct interaction with the SGE. The workflow is being run on the FBK KORE HPC Facility, a Linux cluster consisting of ~1000 cores, and tested on several SEQC datasets.

2. Modules

- Extended - Paramiko v1.7.5

```
>>> from paramiko import SSHClient
>>> class CloudSSHClient(SSHClient):
        def mkdir (params)
        def put_dir_recursvely (params)
```

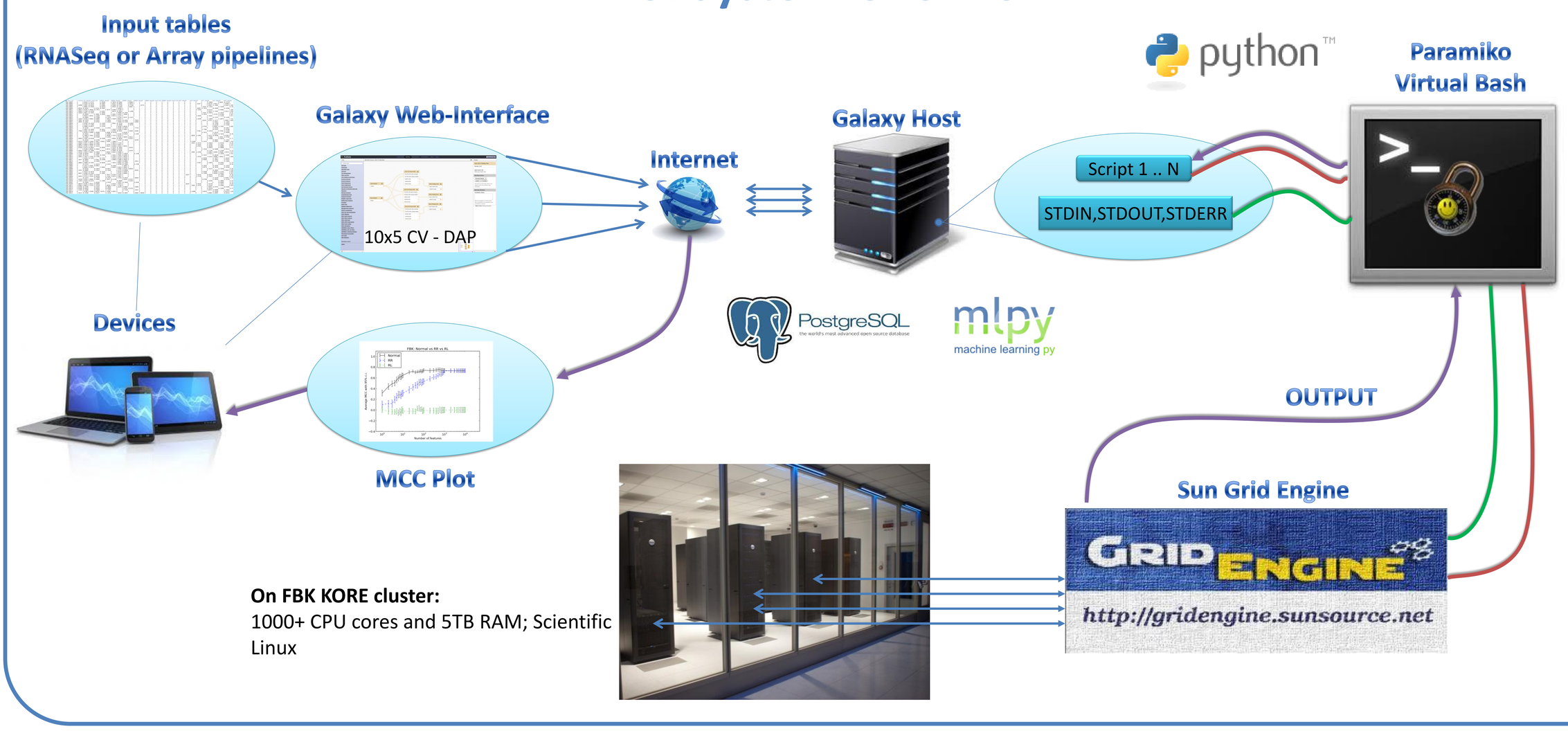
- Sun Grid Engine queue system

```
>>>def qsub(max_mem, queue,
    job_name, script_name, script_param):
        .....
```

- Processes' status control and standard communication streams

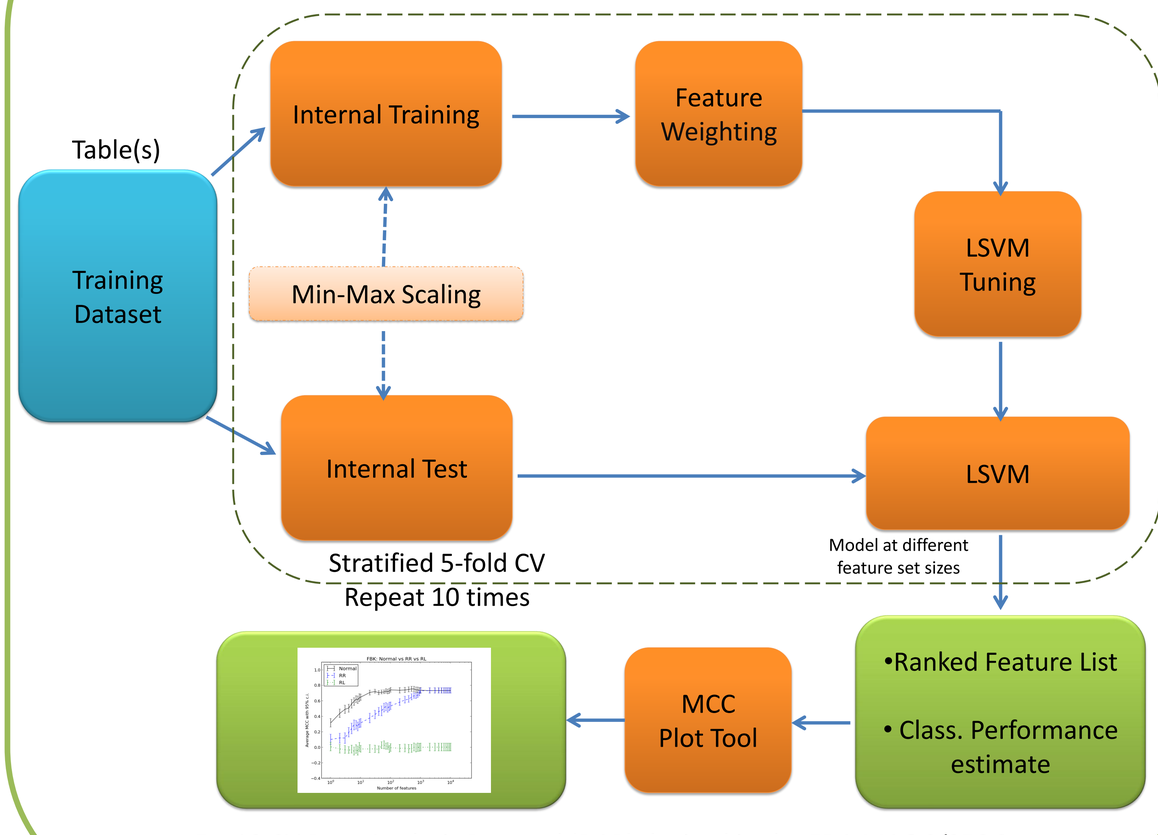
```
>>> client = CloudSSHClient.CoudSSHClient()
>>> client.load_system_host_keys()
>>> client.connect("hpc_system")
>>> sys.stdin, sys.stdout, sys.stderr = client.exec_command(
.....       qsub (8 , bld.q, "job1", "10x5CV.py", [p1, p2]))
```

3. System Overview



On FBK KORE cluster:
1000+ CPU cores and 5TB RAM; Scientific Linux

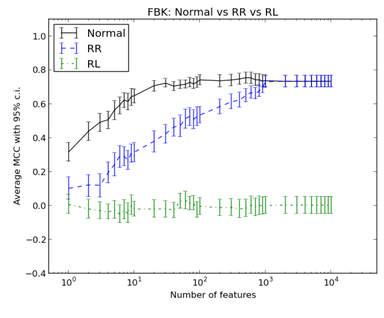
4. Data Analysis Plan



5 x 10 CV Data Analysis Protocol (DAP) devised in the FDA MAQC/SEQC Initiative: machine learning tools from the MLPY library.(4)

5. Experiment Details

- **Datasets**:
  - MicroArray liver rat samples (Aceview): 54 samples x 24 583 features
  - Next Generation Sequencing liver rat samples (Affymetrix): 54 samples x 31 100 features
  - Next Generation Sequencing neuroblastoma human samples: 250 samples x 262 982 features
- **High Performance Computing facility:**
  - FBK KORE HPC Facility, a Linux cluster with SGE queue system and 90 nodes (~1000 cores, 5TB RAM).



REFERENCES

1. Paramiko v1.7.5 http://www.lag.net/paramiko/legacy.html
2. Galaxy Workflow Modeller http://galaxyproject.org/
3. Sun Grid Engine http://gridengine.sunsource.net/
4. MLPY - Machine Learning in Python http://mlpy.sourceforge.net/
5. PostgreSQL http://www.postgresql.org/
6. The MicroArray Quality Control (MAQC) Consortium. The MAQC-II Project: A comprehensive study of common practices for the development and validation of microarray-based predictive models. Nature Biotechnology, 28(8):827-838, 2010