biolab

A Grid Approach for Large Data Processing in Biomedicine

Fato Marco, Papadimitropoulos Adam, ***Porro Ivan****, Scaglione Silvia, Schenone Andrea, Torterolo Livia, Viti Federica*

Department of Communication, Computer and System Science (DIST)
University of Genoa, Italy
pivan@dist.unige.it

What

- A grid platform built with scalability and interoperability in mind
- It will allow large scale (“thousands”) bioinformatics data analysis and will have a simple web-based interface
- Implemented for microarray (first case study) but designed to be flexible enough to handle different data types

What not

- It’s NOT a Grid-based database
- It’s NOT (yet!) a tool to access everything everywhere
- It’s NOT a closed solution (in terms of licenses and applications)
- It’s NOT designed to be a vertical application

Why

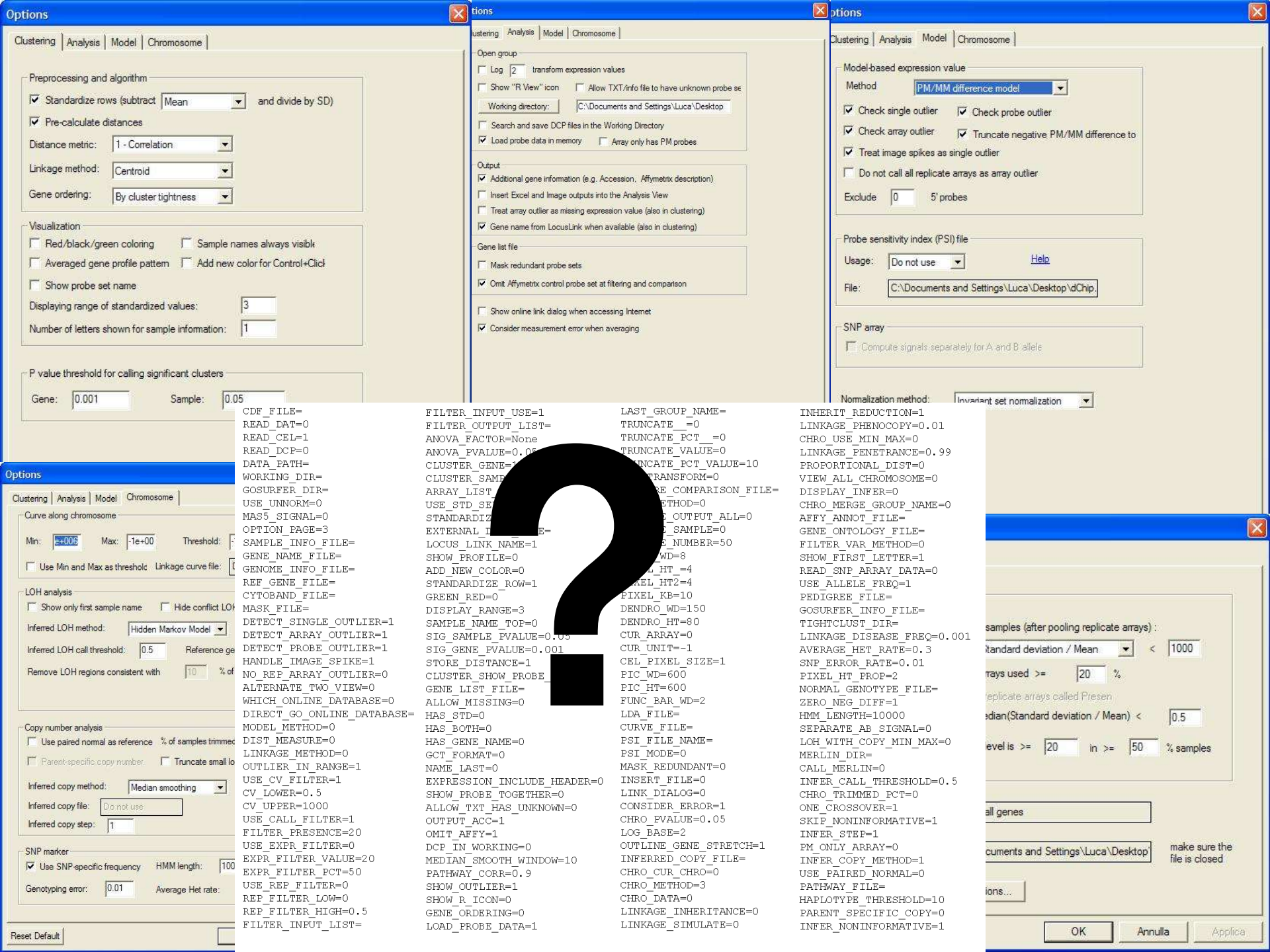
- Microarray have intrinsics costs (chip, raw materials, manpower), about 500€ each chip
  - laboratories usually work on small experiment (10 – 30 uA)
  - the better statistical analysis you want, the larger the dataset must be (N° samples << N° variables)
  - client side software tools interfere with data share
- Bio"everything" data is growing in size and speed of acquisition
  - Data mining is difficult
  - Data storage is challenging (hirarchical vs. on-line tradeoff)

Activities

- Identify free open source tools for microarray analysis → **dChip**
- Identify a network infrastructure to provide secure and reliable storage → **gLite Grid**
- Design a modular system, open to third party modules (algorithms, workflows) → **W3C standards**
- Implement a user frienldy web interface prototype
  - Available: **Genius**
  - Custom: **PHP + Apache, J2EE, portlet**

dChip porting

- dChip is a Windows GUI application for microarray data analysis
- One of available options, choosed as a case study
- **Q:** Is it possible to run dChip command line on a Linux environment and, maybe, over The Grid?
- **A:** Yes. Steps:
  - Linux port
  - Investigate parameters and configuration
  - Standardize I/O files (also intermediate ones)
  - Separate the program into standalone modules
  - Make them parallel


Options
Clustering
Analysis
Model
Chromosome
Preprocessing and algorithm
Standardize rows (subtract
Mean
and divide by SD)
Pre-calculate distances
Distance metric:
1 - Correlation
Linkage method:
Centroid
Gene ordering:
By cluster tightness
Visualization
Red/black/green coloring
Sample names always visibl
Averaged gene profile pattern
Add new color for Control+Click
Show probe set name
Displaying range of standardized values:
3
Number of letters shown for sample information:
1
P value threshold for calling significant clusters
Gene:
0.001
Sample:
0.05
Open group
Log
2
transform expression values
Show "R View" icon
Allow TXT/info file to have unknown probe se
Working directory:
C:\Documents and Settings\Luca\Desktop
Search and save DCP files in the Working Directory
Load probe data in memory
Array only has PM probes
Output
Additional gene information (e.g. Accession, Affymetrix description)
Insert Excel and Image outputs into the Analysis View
Treat array outlier as missing expression value (also in clustering)
Gene name from LocusLink when available (also in clustering)
Gene list file
Mask redundant probe sets
Omit Affymetrix control probe set at filtering and comparison
Show online link dialog when accessing Internet
Consider measurement error when averaging
Model-based expression value
Method
PM/MM difference model
Check single outlier
Check probe outlier
Check array outlier
Truncate negative PM/MM difference to
Treat image spikes as single outlier
Do not call all replicate arrays as array outlier
Exclude
0
5' probes
Probe sensitivity index (PSI) file
Usage:
Do not use
Help
File:
C:\Documents and Settings\Luca\Desktop\dChip
SNP array
Compute signals separately for A and B allele
Normalization method:
Invariant set normalization
Curve along chromosome
Min:
Max:
-1e+00
Threshold:
Use Min and Max as threshold
Linkage curve file:
LOH analysis
Show only first sample name
Hide conflict LOH
Inferred LOH method:
Hidden Markov Model
Inferred LOH call threshold:
0.5
Reference ge
Remove LOH regions consistent with
% of
Copy number analysis
Use paired normal as reference
% of samples trimmed
Parent-specific copy number
Truncate small lo
Inferred copy method:
Median smoothing
Inferred copy file:
Do not use
Inferred copy step:
1
SNP marker
Use SNP-specific frequency
HMM length:
Genotyping error:
0.01
Average Het rate:
Reset Default
CDF_FILE=
READ_DAT=0
READ_CEL=1
READ_DCP=0
DATA_PATH=
WORKING_DIR=
GOSURFER_DIR=
USE_UNNORM=0
MAS5_SIGNAL=0
OPTION_PAGE=3
SAMPLE_INFO_FILE=
GENE_NAME_FILE=
GENOME_INFO_FILE=
REF_GENE_FILE=
CYTOBAND_FILE=
MASK_FILE=
DETECT_SINGLE_OUTLIER=1
DETECT_ARRAY_OUTLIER=1
DETECT_PROBE_OUTLIER=1
HANDLE_IMAGE_SPIKE=1
NO_REP_ARRAY_OUTLIER=0
ALTERNATE_TWO_VIEW=0
WHICH_ONLINE_DATABASE=0
DIRECT_GO_ONLINE_DATABASE=
MODEL_METHOD=0
DIST_MEASURE=0
LINKAGE_METHOD=0
OUTLIER_IN_RANGE=1
USE_CV_FILTER=1
CV_LOWER=0.5
CV_UPPER=1000
USE_CALL_FILTER=1
FILTER_PRESENCE=20
USE_EXPR_FILTER=0
EXPR_FILTER_VALUE=20
EXPR_FILTER_PCT=50
USE_REP_FILTER=0
REP_FILTER_LOW=0
REP_FILTER_HIGH=0.5
FILTER_INPUT_LIST=
FILTER_INPUT_USE=1
FILTER_OUTPUT_LIST=
ANOVA_FACTOR=None
CLUSTER_GENE=1
ARRAY_LIST_
USE_STD_SE
LOCUS_LINK_NAME=1
SHOW_PROFILE=0
ADD_NEW_COLOR=0
STANDARDIZE_ROW=1
GREEN_RED=0
DISPLAY_RANGE=3
SAMPLE_NAME_TOP=0
SIG_GENE_PVALUE=0.001
STORE_DISTANCE=1
GENE_LIST_FILE=
ALLOW_MISSING=0
HAS_STD=0
HAS_BOTH=0
HAS_GENE_NAME=0
GCT_FORMAT=0
NAME_LAST=0
EXPRESSION_INCLUDE_HEADER=0
SHOW_PROBE_TOGETHER=0
ALLOW_TXT_HAS_UNKNOWN=0
OUTPUT_ACC=1
OMIT_AFFY=1
DCP_IN_WORKING=0
MEDIAN_SMOOTH_WINDOW=10
PATHWAY_CORR=0.9
SHOW_OUTLIER=1
SHOW_R_ICON=0
GENE_ORDERING=0
LOAD_PROBE_DATA=1
LAST_GROUP_NAME=
TRUNCATE__=0
TRUNCATE_PCT__=0
TRUNCATE_VALUE=0
TRANSFORM=0
ETHOD=0
_SAMPLE=0
_NUMBER=50
WD=8
_HT_=4
XEL_HT2=4
PIXEL_KB=10
DENDRO_WD=150
DENDRO_HT=80
CUR_ARRAY=0
CUR_UNIT=-1
CEL_PIXEL_SIZE=1
PIC_WD=600
PIC_HT=600
FUNC_BAR_WD=2
LDA_FILE=
CURVE_FILE=
PSI_FILE_NAME=
PSI_MODE=0
MASK_REDUNDANT=0
INSERT_FILE=0
LINK_DIALOG=0
CONSIDER_ERROR=1
CHRO_PVALUE=0.05
LOG_BASE=2
OUTLINE_GENE_STRETCH=1
INFERRED_COPY_FILE=
CHRO_CUR_CHRO=0
CHRO_METHOD=3
CHRO_DATA=0
LINKAGE_INHERITANCE=0
LINKAGE_SIMULATE=0
INHERIT_REDUCTION=1
LINKAGE_PHENOCOPY=0.01
CHRO_USE_MIN_MAX=0
LINKAGE_PENETRANCE=0.99
PROPORTIONAL_DIST=0
VIEW_ALL_CHROMOSOME=0
DISPLAY_INFER=0
CHRO_MERGE_GROUP_NAME=0
AFFY_ANNOT_FILE=
GENE_ONTOLOGY_FILE=
FILTER_VAR_METHOD=0
SHOW_FIRST_LETTER=1
READ_SNP_ARRAY_DATA=0
USE_ALLELE_FREQ=1
PEDIGREE_FILE=
GOSURFER_INFO_FILE=
TIGHTCLUST_DIR=
LINKAGE_DISEASE_FREQ=0.001
AVERAGE_HET_RATE=0.3
SNP_ERROR_RATE=0.01
PIXEL_HT_PROP=2
NORMAL_GENOTYPE_FILE=
ZERO_NEG_DIFF=1
HMM_LENGTH=10000
SEPARATE_AB_SIGNAL=0
LOH_WITH_COPY_MIN_MAX=0
MERLIN_DIR=
CALL_MERLIN=0
INFER_CALL_THRESHOLD=0.5
CHRO_TRIMMED_PCT=0
ONE_CROSSOVER=1
SKIP_NONINFORMATIVE=1
INFER_STEP=1
PM_ONLY_ARRAY=0
INFER_COPY_METHOD=1
USE_PAIRED_NORMAL=0
PATHWAY_FILE=
HAPLOTYPE_THRESHOLD=10
PARENT_SPECIFIC_COPY=0
INFER_NONINFORMATIVE=1
samples (after pooling replicate arrays) :
tandard deviation / Mean
<
1000
rays used >=
20
%
eplicate arrays called Presen
edian(Standard deviation / Mean) <
0.5
evel is >=
20
in >=
50
% samples
all genes
cuments and Settings\Luca\Desktop
make sure the file is closed
ions...
OK
Annulla
Applica

Computing & Data Grid: gLite

- Secure and distributed data storage
- Access to several thousands CPU
- LCG (LHC Computing Grid) is the middleware born to support forthcoming (2007) High Energy Physics experiments
- gLite is its evolution (deployment in production on April 2006)

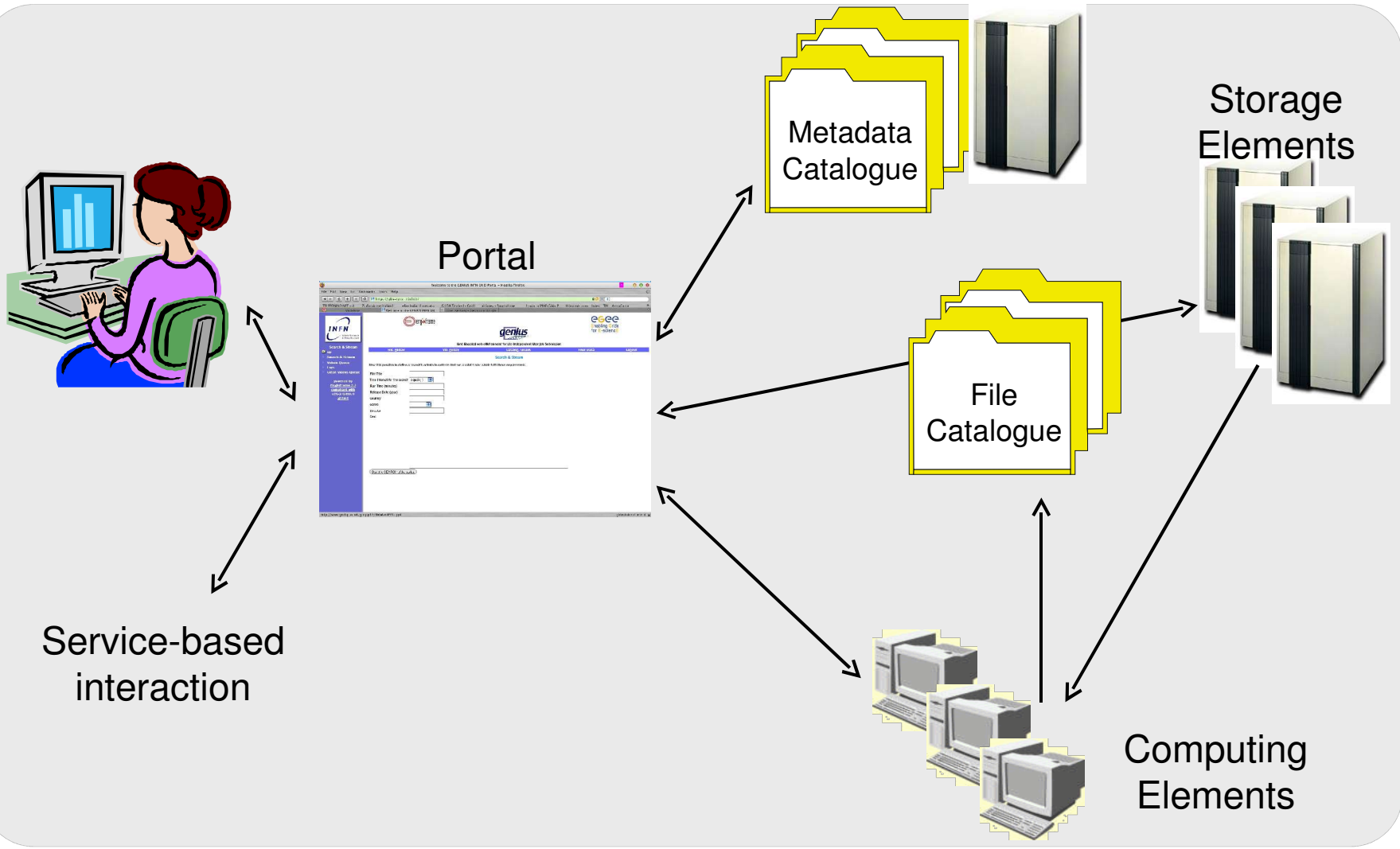
Metadata: MIAME & AMGA

- Metadata are relationally structured information associated to experiment data
- MIAME (Minimum Information About Microarray Experiment) is a MGED de-facto standard
- We defined a custom ontology for stem cells experiments to maintain homogeneity in collected data among different laboratories
- AMGA (ARDA Metadata Grid Application) is the Metadata catalogue of gLite 1.5, with replication and federation support

Interface prototype: Genius

- Provide a open source grid-portal based on Nice Srl product Enginframe
- Features:
  - Grid Security Infrastructure compliant
  - Modular and flexible: it can handle almost every programming and scripting language for underlying logic
- It could publish its interface components also as web services (WSDL)

System Architecture

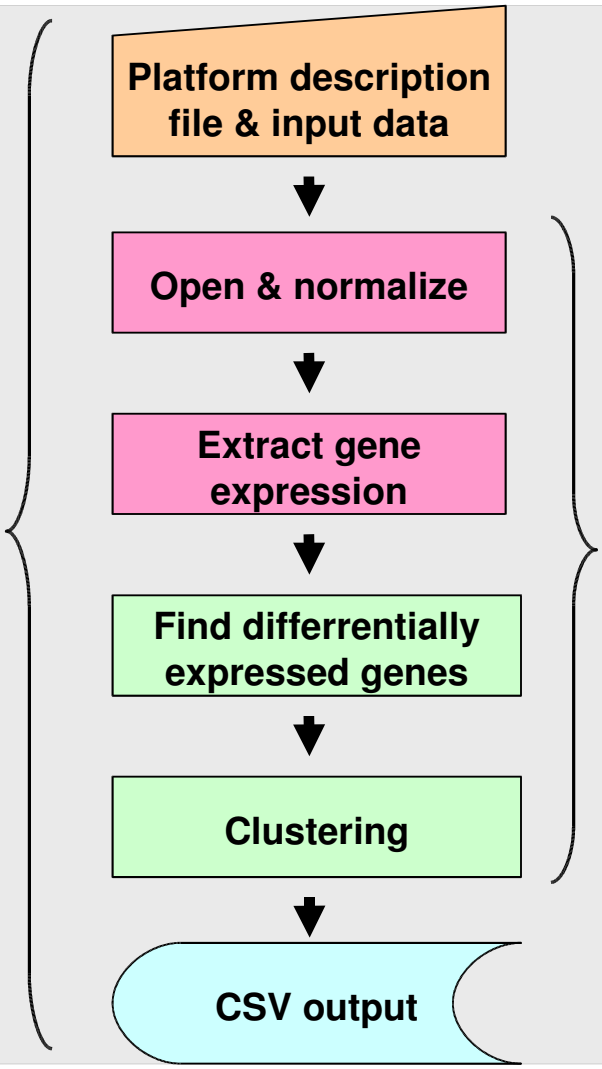


MPI-dChip available

- dChip has been succesfully ported to Grid and parallelized with MPI (mpich2)
- Two version available (Linux):
  - standard file I/O for MPI clusters (RHEL3/4)
  - grid GFAL I/O for MPI grid job (gLite 3.0)
  - mix available
- Source code is GNU/GPL
- Availability: on request [LITBIO portal? Sourceforge?]

MPI-dChip structure

Original dChip pipeline

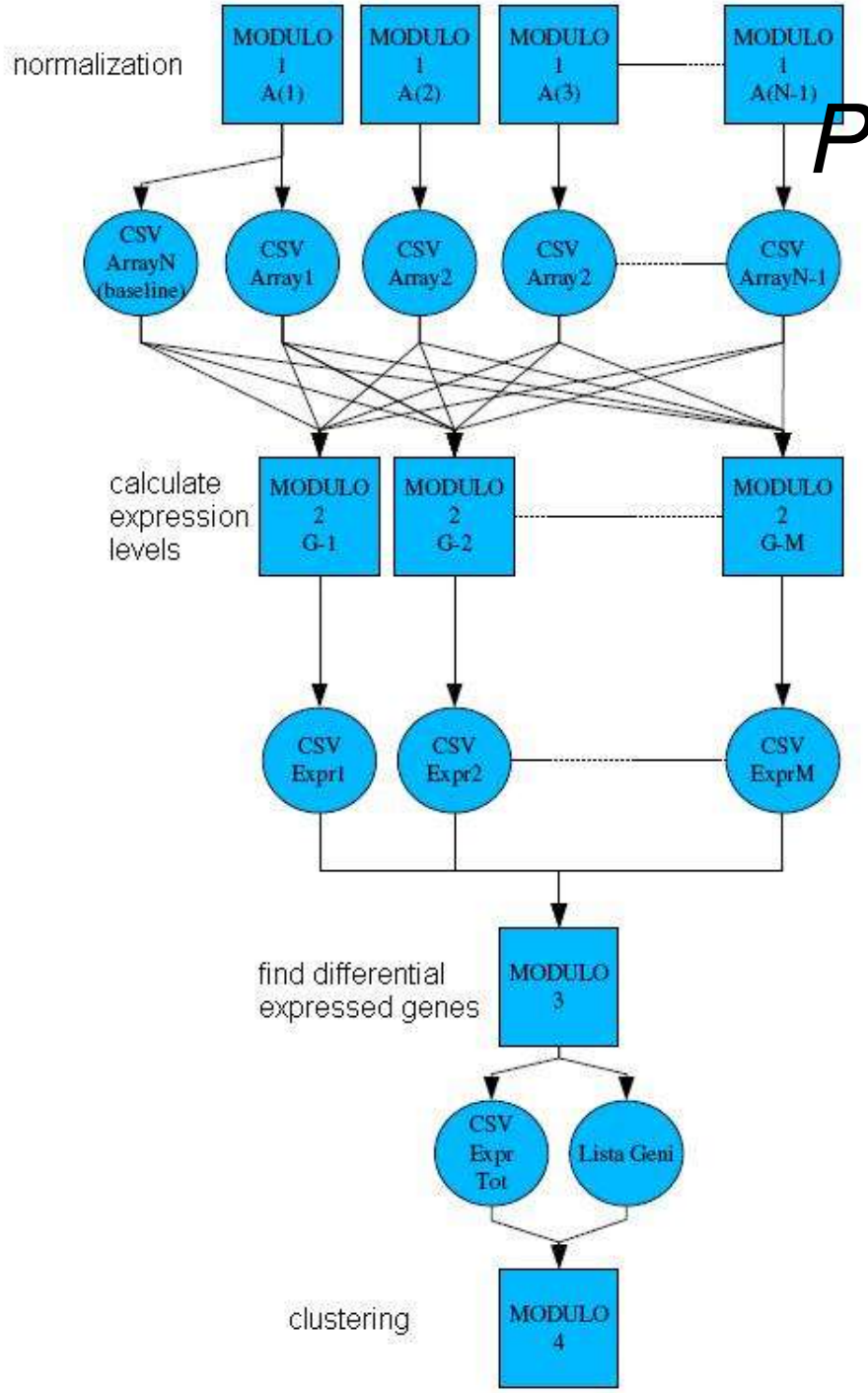


- Execution on the Grid or on a cluster
- All intermediate files are Comma Separated Value
- All jobs can be executed sequentially with a shell script or as standalone jobs

parallel

not parallel

Parallelization schema



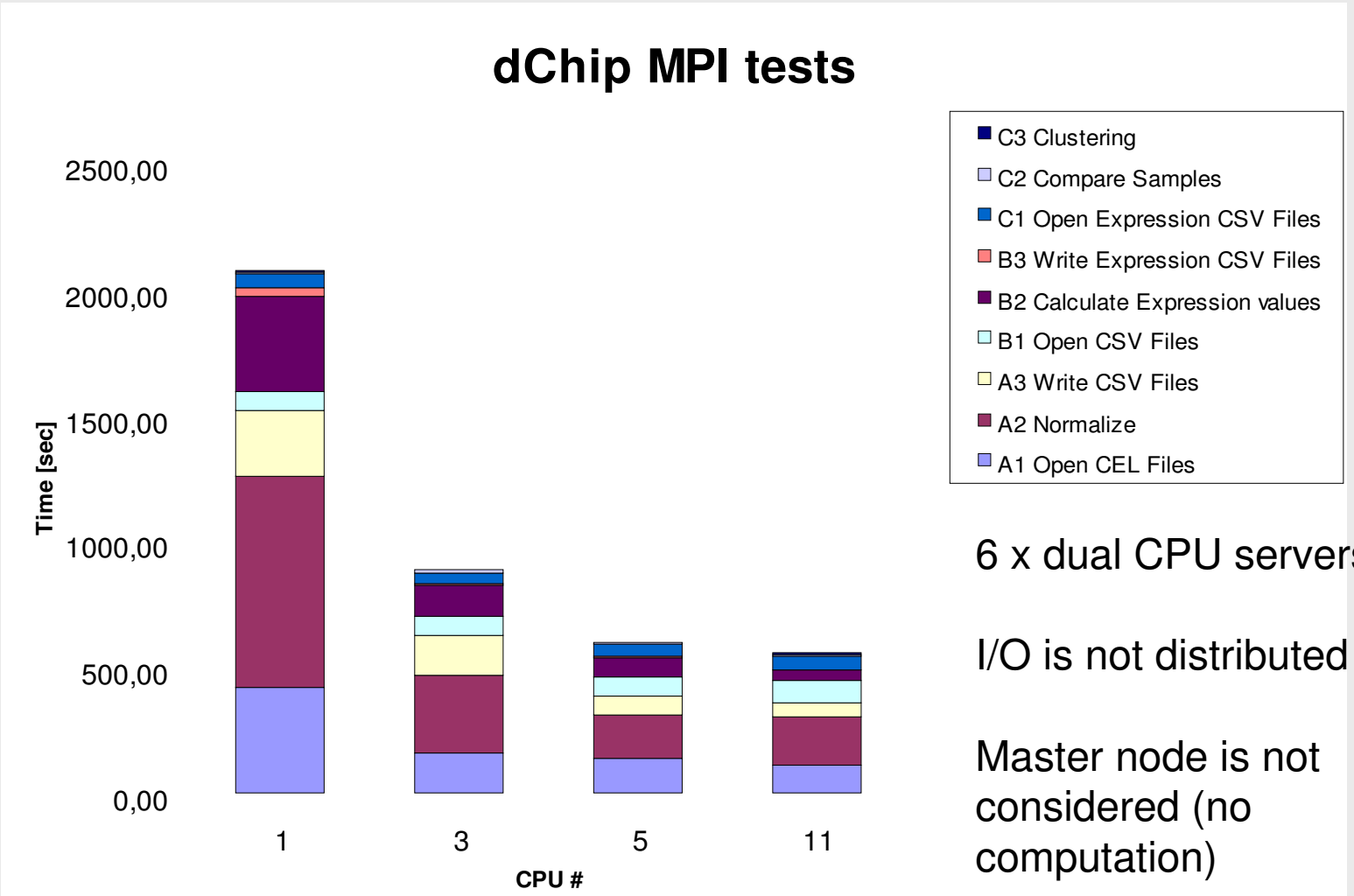
e.g.: we have 10 job and 200 files, each with 20.000 genes

Each normalize job takes 200/10 files and the base file and produce 200 normalized input files

Each model (PM or PM/MM) job takes 20.000/10 genes and produce 10 expression level files

Differential expressed genes calculation and clustering are not parallel. Results are saved to 2 CSV files

Results



6 x dual CPU servers

I/O is not distributed

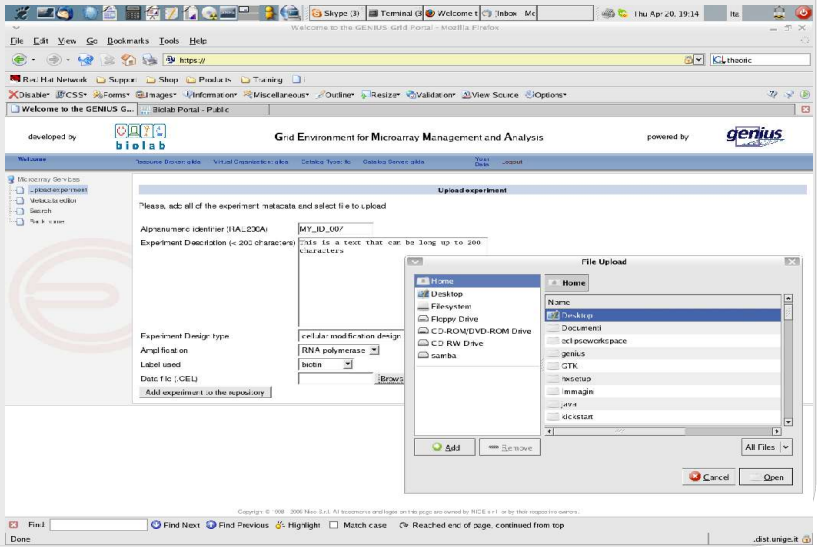
Master node is not considered (no computation)

Results /2

We succesfully installed and customized Genius Grid portal

We developed microarray data management service prototype:

- Annotatated experiment upload
- Search
- Edit



Open issues & future work

- Genius 3.1 / Enginframe under extensive testing
  - Standard development is easy
  - Low level customization and software maintenance could be difficult
- Workflow enaction with Moteur engine partially successful
  - Moteur limitations for data and services
  - Genius integration has been undertaken
- Webservice available only for Genius interface services
- Integrate other tools
  - ArrayExpress Import/Export
  - R/Bioconductor based tools

Acknowledgments

- This platform is currently part of the Italian FIRB project LITBIO (Laboratory for Interdisciplinary Technologies in BIOinformatics), http://www.litbio.org

- Thanks to Ulrich Pfeffer, IST Genoa for providing data for our tests and to Luca Corradi, thesis student, for coding MPI dChip

- Questions?