



Enabling Grids for E-scienceE

# BioDCV for functional genomics Status at Spring 2007



*S. Paoli, C. Furlanello, D. Albanese,  
G. Jurman, A. Barla, S. Merler, R. Flor*

<http://mpa.itc.it>

*EGEE Bioinformatics Meeting #4 - Varenna, May 19th 2007*

[www.eu-egee.org](http://www.eu-egee.org)



1. Application for analysis of microarray and proteomics data with Support Vector Machine (SVM) classifiers
2. BioDCV is written in C language, without external libraries or services
3. it uses SQLite database flat files to manage data/results on grid elements
4. Use both non-MPI and LCG2 MPI grid sites
5. With IFOM-FIRC in BICG AIRC project

## Timeline



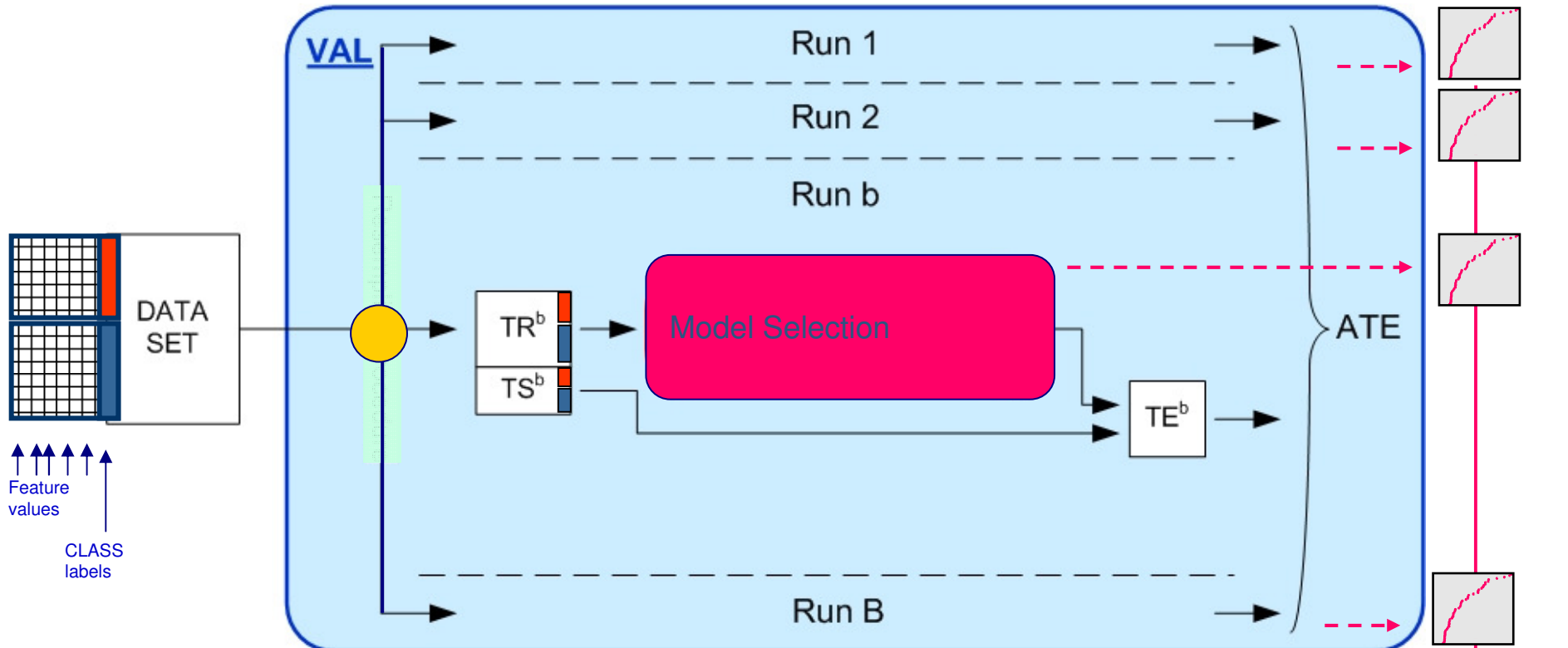
- [Spring 2005]: in production as a grid application
- [Dec. 2005]: 193 CPU days of scalability and footprint tests on microarray data on the INFN grid – Egrid VO
- [Spring 2006]: running in EGEE Biomed VO
- [May 2006]: Maldi-TOF proteomics on Egee Biomed VO
- [Autumn 2006]: Breast cancer microarray data for IFOM
- [Spring 2007]: SELDI-TOF Cromwell simulator

# To Avoid Selection Bias: a complete validation setup\*

- externally, a stratified random partitioning: ●
- internally, a model selection based on a K-fold cross-validation:
- SVM, RFE, HPC on cluster and EGEE Biomed VO grid

From 300.000  
to 2.000.000

SVMs



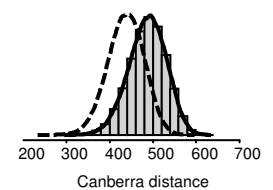
ATE:  
Average Test Error

BioDCV

SET OF LISTS:

- aggregation
- stability

Mutual genelist distance



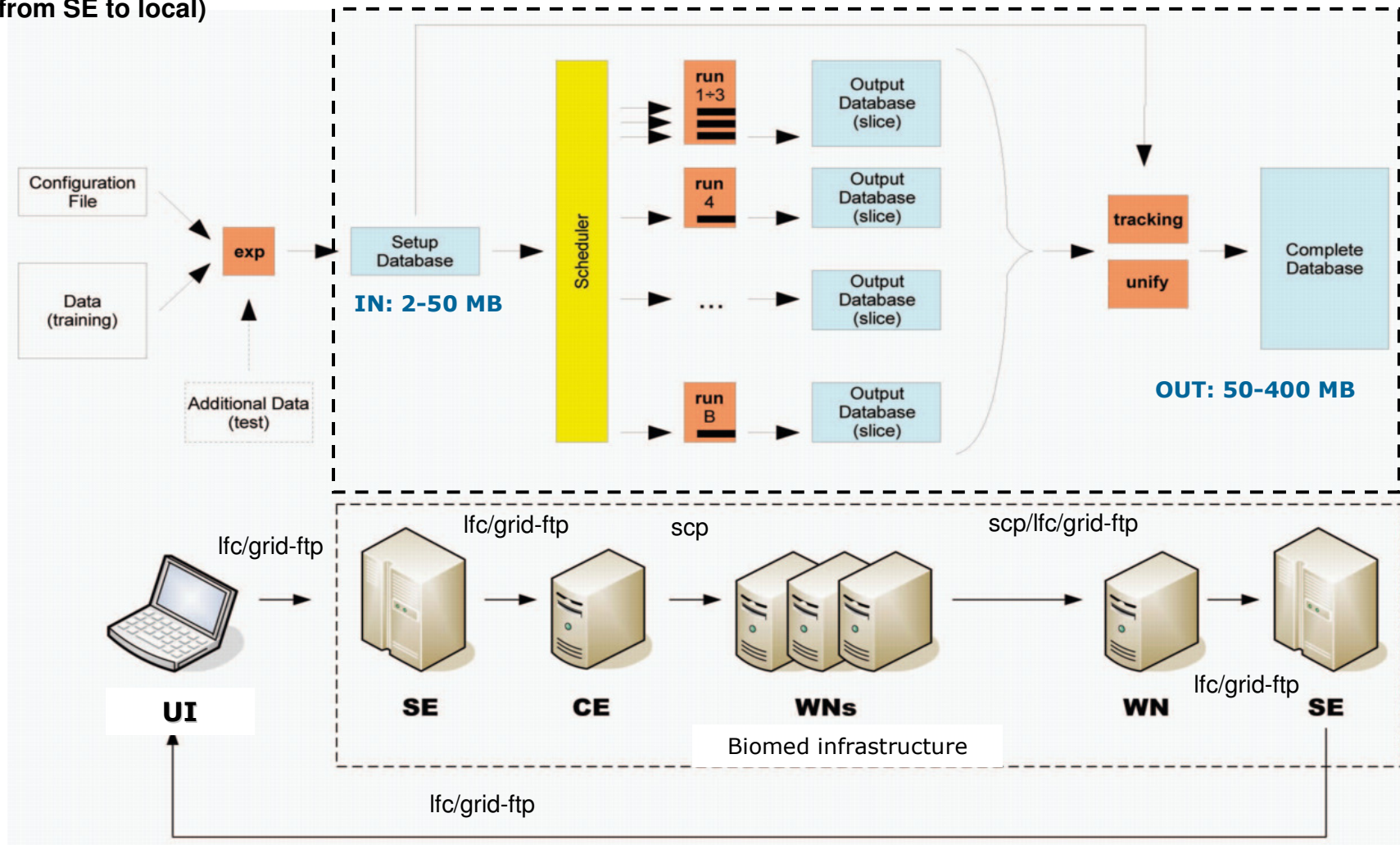
\* Ambroise & McLachlan, 2002, Simon et. al 2003, Furlanello et. al 2003

Project funded by  
AIRC - IFOM

Standard LCG user interface commands are used to transfer

- a. Data + experiment design (setup db)  
**lcg-cp/grid-url-copy db**  
from local to SE
- b. Application  
**edg-job-submit BioDCV.jdl** (jdl file)
- c. Resulting db:  
**lcg-cp/grid-url-copy db**  
(from SE to local)

# Running BioDCV on LCG/gLite middleware

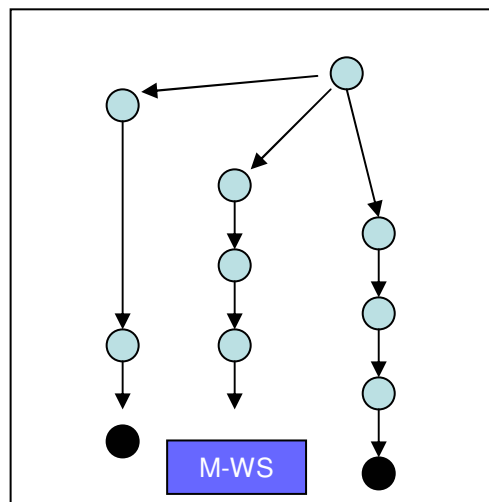


Task:

- Connect BioDCV to other available bioinformatics applications and tools

How:

- Zolera SOAP Infrastructure (ZSI)
  - Python language
  - Apache (mod\_python)
  - HTTP, SOAP and WSDL capabilities
  - Result
    - A first application in M. Cannataro, A. Barla, R. Flor, A. Gallo, G. Jurman, S. Merler, S. Paoli, G. Tradigo, P. Veltri, C. Furlanello “A Grid environment for high-throughput proteomics” IEEE Transactions on Nanobioscience - accepted April 07)

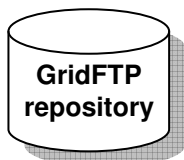


**MS-Analyzer  
Ontology-based  
Workflow Designer**

Repository URL -  
Email -

- Data
- Metadata

- Biomarkers data
- REPORT

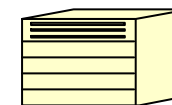


**BioDCV  
Grid-enabled  
Molecular  
Profiling**

- BioDCV Outputs: Visualization of ATE, Sampletracking, HTML publication, Email notification



Biomed VO



Local cluster facility

Egee UI

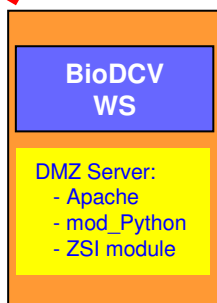
BioDCV - WS - UI.py

Local Cluster front-end

BioDCV - WS - local

Proteomics Data Preparation

BioDCV WS front-end Server



Cannataro et al. "A Grid environment for high-throughput proteomics"  
IEEE Transactions on Nanobioscience

Mozilla Firefox

File Modifica Visualizza Vai Segnalibri Strumenti ?

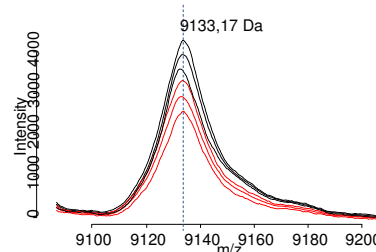
Result

Data:  
<http://poseidon.bioingegneria.unicz.it/~cannataro/dates/2006.08.09/data.zip>  
[Average Test Error](#)  
[Sample tracking curves](#)

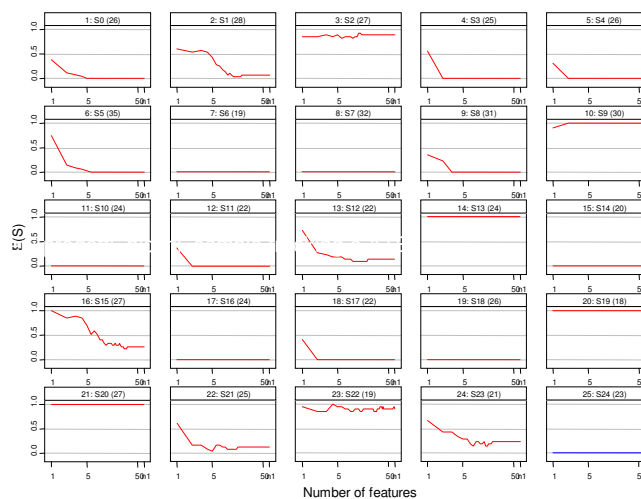
List

#	Peak id	# Exts	Mean	SD
1	29	100	5.3	2
2	49	100	6.3	2.6
3	34	99	2.4	2.4
4	46	99	6.9	4.3
5	42	98	5	4.5
6	55	97	9.8	4.5
7	32	93	11.9	3.1
8	51	91	6.2	4.9
9	41	89	9.8	5.4
10	40	87	8.6	6.5
11	33	87	9.9	3.7
12	35	84	13.4	3.7
13	59	83	14.7	3.1
14	50	79	13.3	3.3
15	43	74	6.6	4.4
16	37	65	12.4	4.6
17	28	64	10.9	5.3
18	30	59	14.4	3.7
19	53	54	14.2	3.3
20	52	53	13.2	4

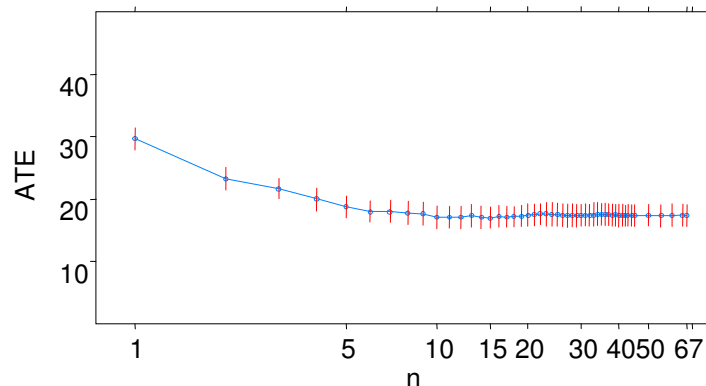
Completato



- D2: mean A
- D2: .95 Student bootstrap CI
- D2: mean B
- D2: .95 Student bootstrap CI



- Error rate (tumour tissue)
- Error rate (non-tumoural tissue)
- No-information error rate



<http://www.r-project.org/>

Spring 2006

## Experiments on Ovarian Cancer (Barla et al, IEEE CBMS 2006)

Data: Nat. Ovarian Cancer Early Detection Program, Northwestern Univ. Hospital MALDI-TOF analyzers from Keck Lab Yale: 93 cancers + 77 controls, ranging in 700-3500 Da for reflectron analyzer and in 3450-28000 Da for the linear one. As in (Wu et al 2005)

- standard analysis (5 datasets)
- random labels analysis (5 datasets)
- A strict deadline for analysis and paper completion ...

## Solution:

- Egee Biomed grid infrastructure
- **20 cpus (Jobs)** per analysis, a total of **100 + 120 jobs**
- The BioDCV jobs were run on 150 Biomed Sites in all Europe
- Failure ~ **2%** . Legenda: a complete BioDCV experiment is splitted in  $N$  subprograms, of which  $n$  fail:  $2\% = 100 * n / N$

## Breast cancer microarray dataset

- **22215** genes and **183** samples  
 4Mega footprint units (footprint = #features x #samples)  
 Original work in (Sotiriou et al, J. Nat Canc Inst 2006)
  - September/October 2006
  - Used: 60 CPUs and about 40 Biomed sites.
  - 20 CPUs x 3 series (alternative machine learning models):  
 RFE-Linear SVM, TR (Terminated Ramp) SVM,  
 Correlation-aware RFE-SVM.
  - Failure: **5** % (3 jobs)
  - Running times (average over 20 runs):
    - Linear SVM ~ 5 hours
    - TR SVM ~ 8 hours
    - Correlation-Aware ~ 15 hours

Spring 2007

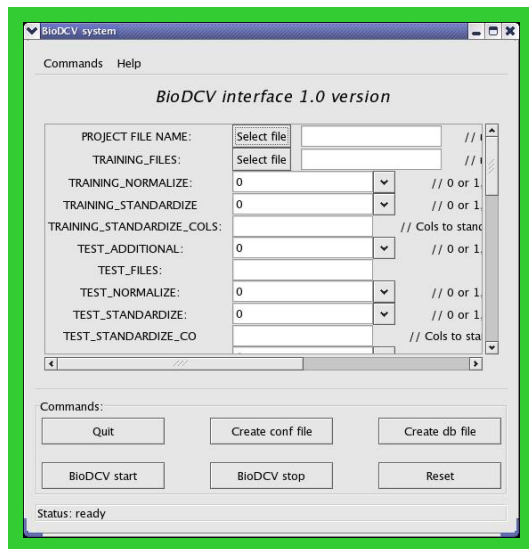
Cromwell: a proteomic MALDI-TOF simulation engine

- 10 replicated experiments:
  - Same m/z from Keck's linear dataset: [3660,26000] Da
  - Synthetic one generated with the Cromwell simulator
  - Multires model:  $R_{tot} = R1 + R2$
  - 100 BioDCV runs: training set: 76 + 76 samples and test set: 100 + 100
- 6 replicated experiment on EGEE Biomed grid
  - $20(\text{jobs}) * 3(R1 + R2 + R_{tot}) * 6 = 360$  jobs -> 360 cpus
  - Each job -> 5 BioDCV runs
  - Failure= 6%
- Results:
  - 1.Valencia Feb07: Proteomics and Pathology, Joint European and American Proteomic Society (multiresolution proteomic profiling)
  - 2.Paper submitted to Journal of Mass Spectrometry
  - 3.Naples April07: BITS2007 Meeting (Biomarker stability of multiresolution proteomic profiling)

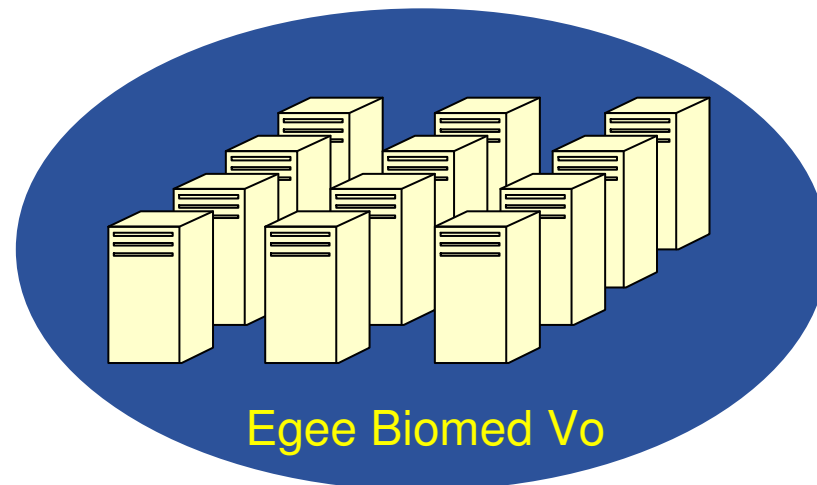
K.R.Coombes et al.  
Understanding the  
characteristics  
of mass spectrometry data,  
Cancer Informatics 2005

- Written in C and Python languages
  - Optimization
    - Internal routines
    - Molecular profiling procedures
  - Allow dealing with integrative biology tasks
  - Sqlite+Pickle(object serialization)
  - NumPy (scientific library)
  - Best integration with TK, PyGTK or wxPython GUI
  - Native linking with Python WS tools (ZSI library )
- R and R+Python scripts
- Install system
- Now: starting tests on the first binaries

## Workstation



- Linux and Windows
- JDL files creation
- Submission to Egee
  - Put data into SE
  - Run BioDCV
  - Control status
  - Get results from grid

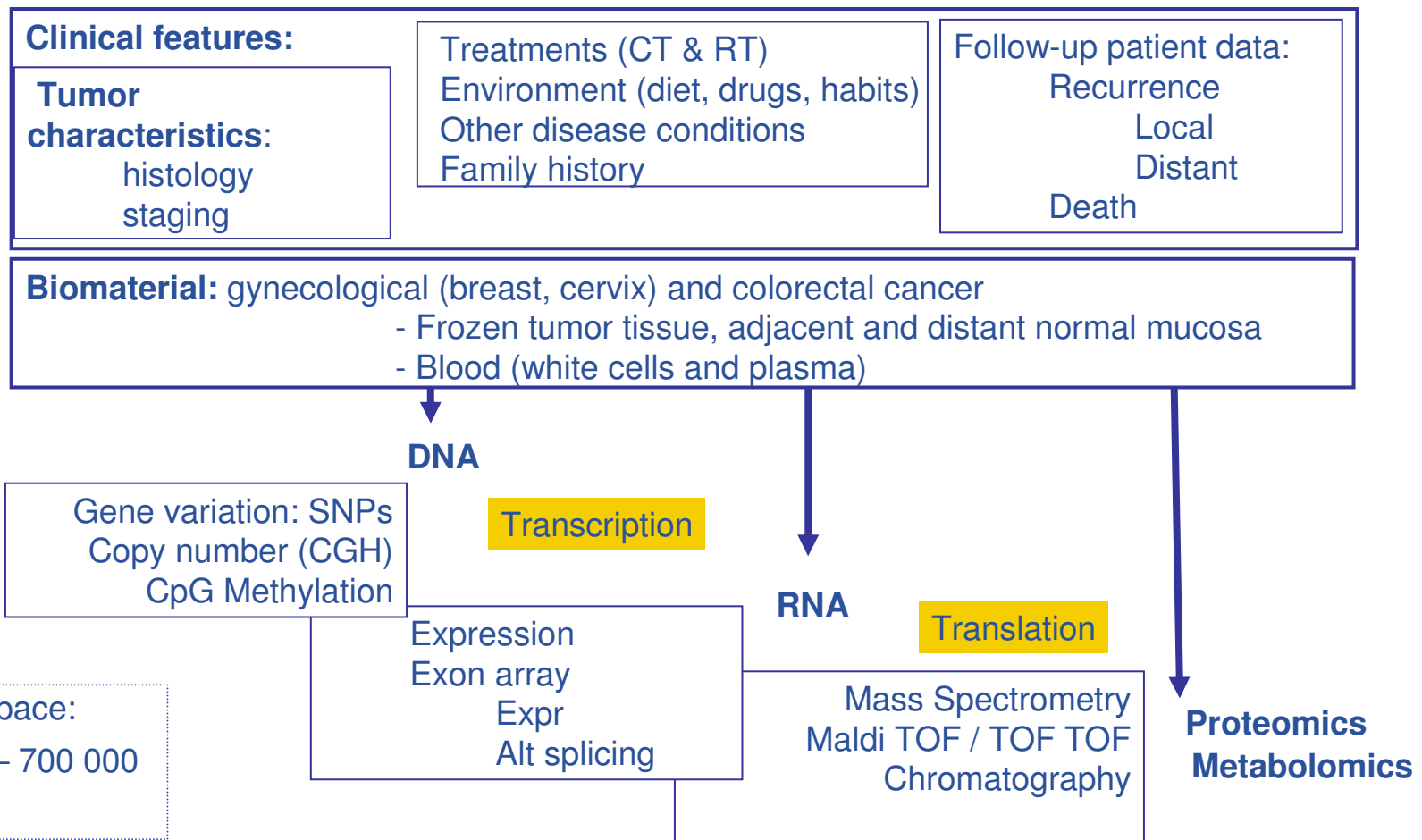


Security Protocol

gLite (Webservices)



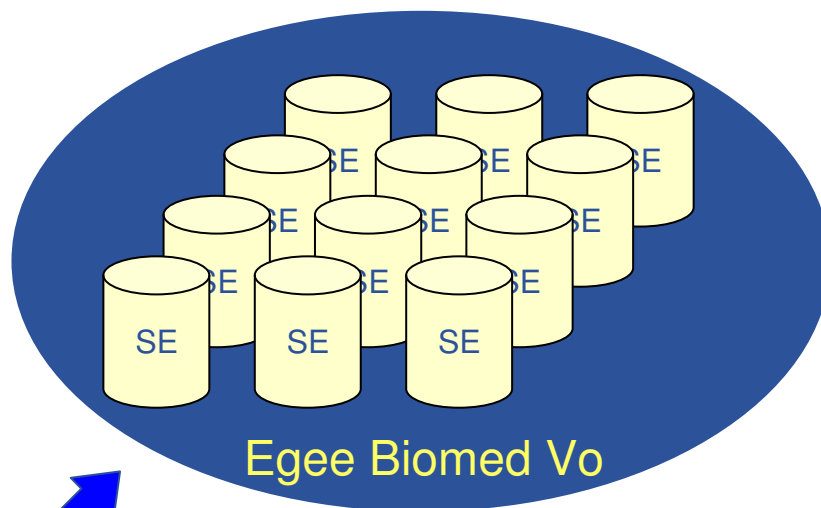
Data from Tumor banks e high-throughput platform in **ONCOPATTERN** (FBK-irst, ICO Barcelona, Katholieke Universiteit Leuven, Bristol Univ., Rotterdam, CCB Mines Paris): **identification of complex pattern in alterations for early diagnosis, prognosis and treatment response.**



Proteomic datasets

Microarray datasets

Clinical data



- Data repository
- Common access
- Control access
- Data encryption
- Backup

Group A



Group B



## BioDCV: summary

- **SUBVersion repository:** <http://biodecv.itc.it>
- **Production version for predictive profiling on high-throughput technologies**
  - Computational procedures for complete validation  
→ Control of Selection Bias / Overfitting
  - Stability analysis of Biomarker Lists
- **NEW: WebServices, GUI interfaces and BioDCV 3.0**
- **Applications in molecular oncology**
  - Microarrays (with IFOM-FIRC, AIRC)
  - Mass spectrometry (with UniCZ)
- **Contacts:** Cesare Furlanello ([furlan@itc.it](mailto:furlan@itc.it))  
Silvano Paoli ([silpaoli@itc.it](mailto:silpaoli@itc.it))