



Applicazione delle tecnologie ICT
al settore Life Sciences:
casi di studio e dimostrazioni di utilizzo
delle soluzioni Nice per l'analisi e gestione
dei dati generati dalle analisi biotecnologiche

Livia Torterolo, PhD

livia.torterolo@nice-software.com



Outline

NICE

NICE Products&Solutions

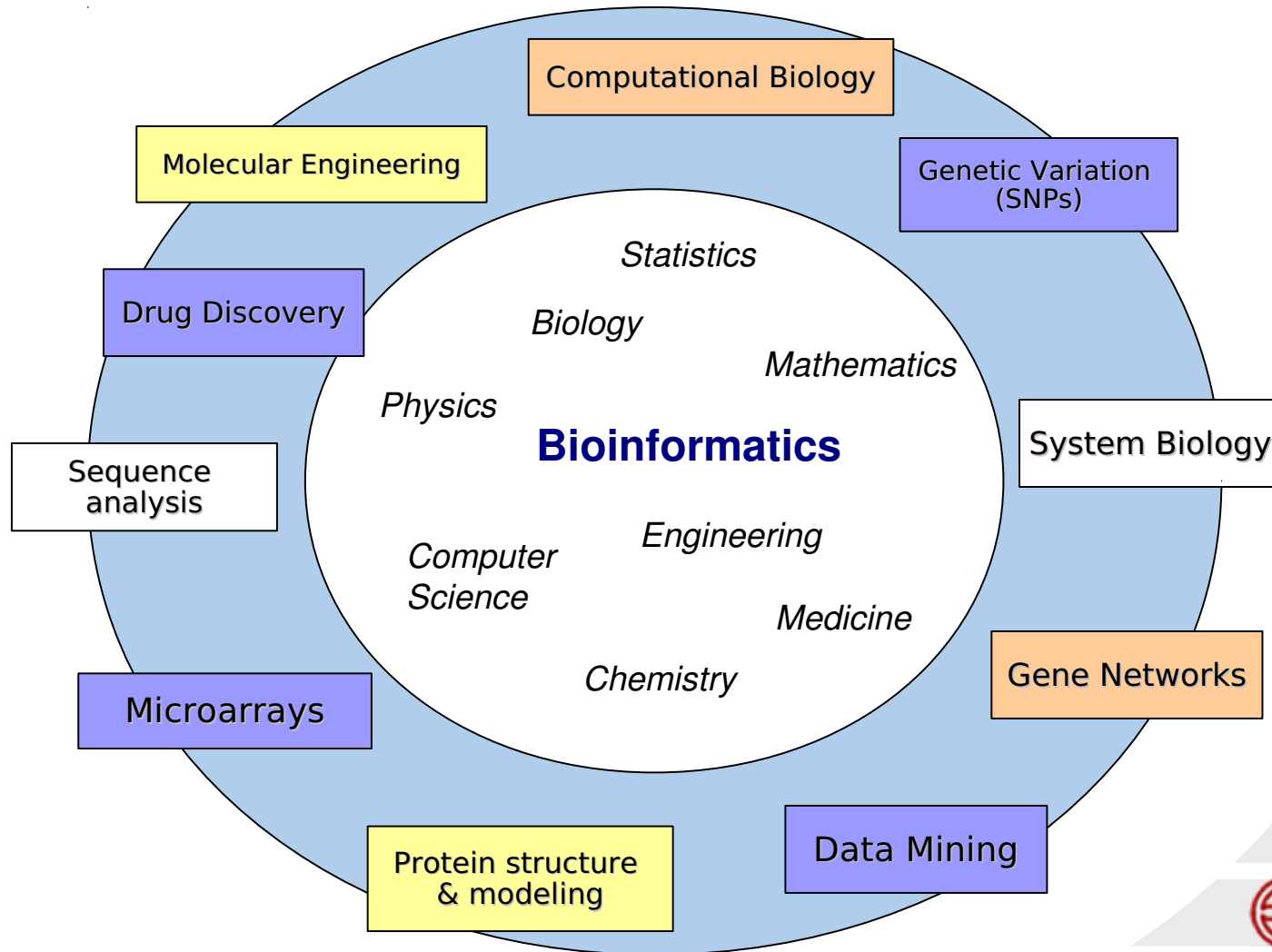
Life Sciences Solutions

Case studies

Demonstrations



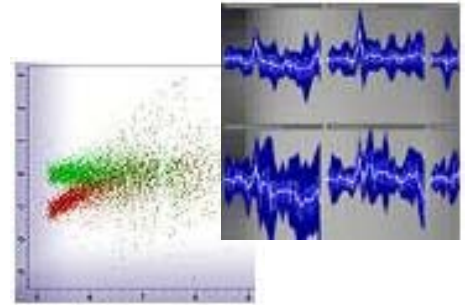
Scenario



Challenges

- **Data Analysis**

- Analyze large datasets
- Automate procedures – reproducibility



- **Data/Metadata Management**

- Archive big amount of data/metadata
- Access and manage distributed data

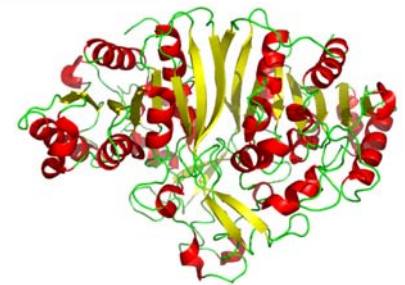


- **Integration**

- Integrate heterogeneous tools and data

- **Remote Visualization**

- Visualize & share data & interactive app remotely

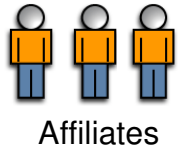
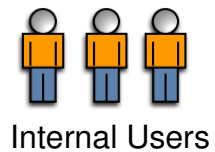


- **Accessibility & Usability**

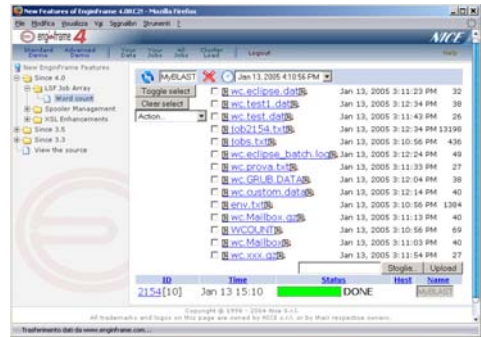
- Technology for everyone



NICE EnginFrame Architecture



Standard Protocols



NICE EnginFrame
Grid Portal / Gateway



Enterprise Portal

Applications
(Interactive & Batch)

License keys

Storage and Data



Grid/Compute/Visualization Farm
(Linux, Windows, ...)



How EnginFrame can be customized for LS...?

... by using/developing Life Sciences plugins!

■ BioInformatics tools (batch)

- R / BioConductor support
- Affymetrix Power Tools (parallel)
- Dchip (Cheng Li Lab) (parallel)
- Plink commands
- Amber, AutoDock, BLAST family, BLAT, Charmm, ClustalW, Glue, Gromacs, MrBayes, Namd, Philyp, RaXML, Vienna RNA, X! Tandem, etc.

■ Molecular modeling (interactive)

- Accelrys DS Visualizer
- MOE
- Schrodinger/Glide

■ Medical Imaging tools (batch&interactive)

- Statistical Parametric Mapping
- SeegViewer (3D Viewer)

■ Matlab (batch&interactive)

■ Workflow tools

- SOAP/web-service enable
- Taverna, Moteur
- A-ware

■ Data/metadata mngt framework

- Irods, gLite

... applications can be provided as services (SaaS)



Outline

NICE

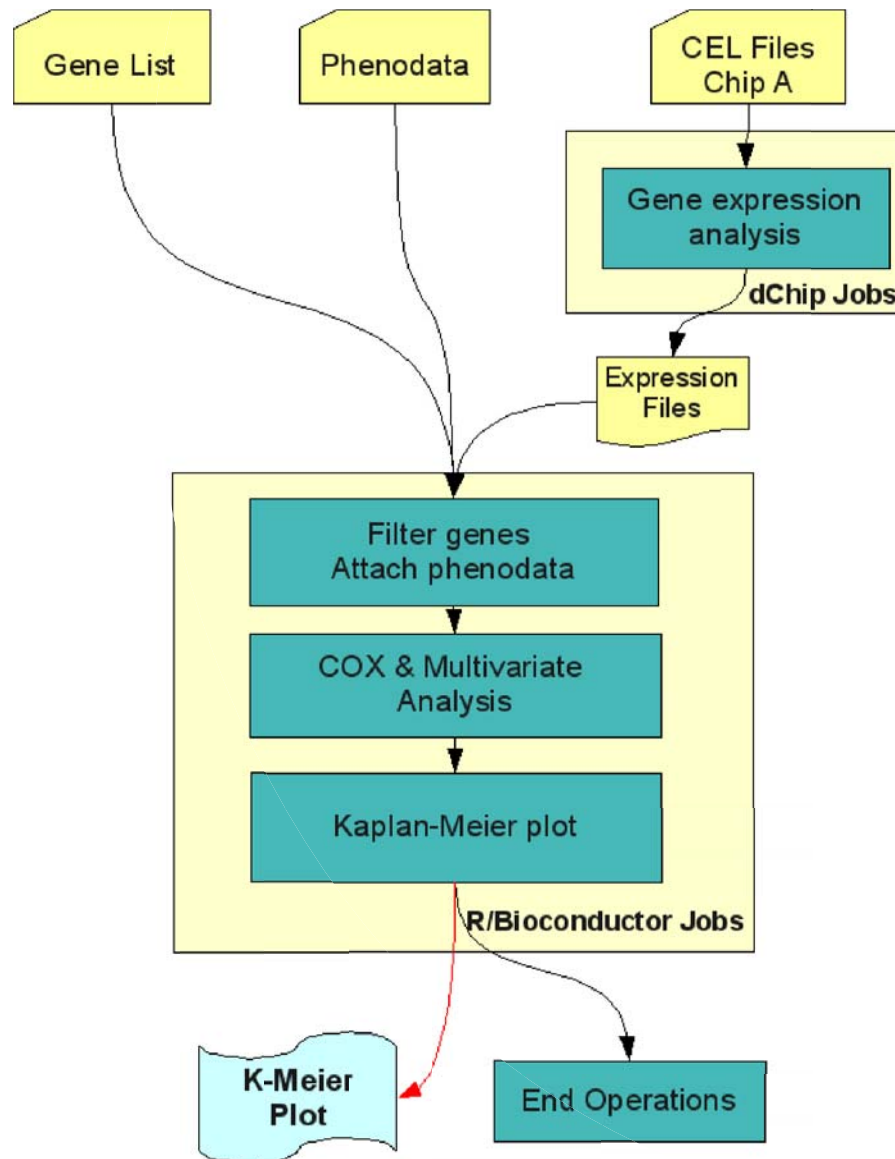
NICE Products&Solutions

Life Sciences Solutions

Case studies

Demonstrations

Case study: Cancer Survival Studies (1)



In collaboration with:



Aim:

Develop a service for automating functional genomics analyses used in breast cancer survival studies and prognosis assessment.

It allows to study the correlation of clinico-pathological and follow-up data with gene expression both for the definition of signatures and to test hypotheses on the effects of over/under expression of single genes.

The service can be used by bio-medical researchers without specific computation skills to validate potential biomarkers or multi-gene classifiers.



Case study: Cancer Survival Studies (2)

Microarray analysis WF LSF

Analysis name: Survival analysis

microarray

DataSet

CDF_file Browse

CEL_file Browse

CEL_file Browse

View XML

Expression values options:

Model-Based: PM only model

Expression: R compatible

Expression file format: R compatible

log2 expressions

filter genes and phenodata options:

Phenodata: Browse...

Gene list: Browse...

Cox and Multivariate options:

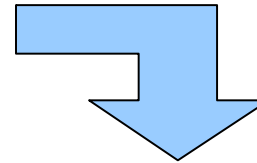
Phenodata 1: SURV_RELAPSE

Phenodata 2: RELAPSE

Perform Multivariate analysis

Kaplan-Meier plot

The analysis is submitted as a "JOB" to the backend infrastructure and its status can be monitored in an easy way through the portal



File Upload

biolab

Places	Name	Modified
Search	Desktop	Monday
Recently Used	Documents	Thursday
biolab	Examples	04/22/2008
Desktop	PDF	Thursday
File System	Public	04/29/2008
18.9 GB Media	R	09/29/2008
	R-2.5.1	09/29/2008
	Documents	
	R-old	05/26/2008

Buttons: Add, Remove, All Files, Cancel, Open

Analysis: Mar 06, 2009 17:19:55

Microarray Analysis_bna_Survival analysis

File	Date	Size
Leve	02 24, 16 19	4096
logs	02 24, 16 19	72
lista_gen2.txt	02 24, 14 58	79
phenodata_stoccorma3.com	02 24, 14 58	228

Buttons: Browse, Upload into Spooler

The user interface allows input data upload (locally or remotely) and parameters selection.

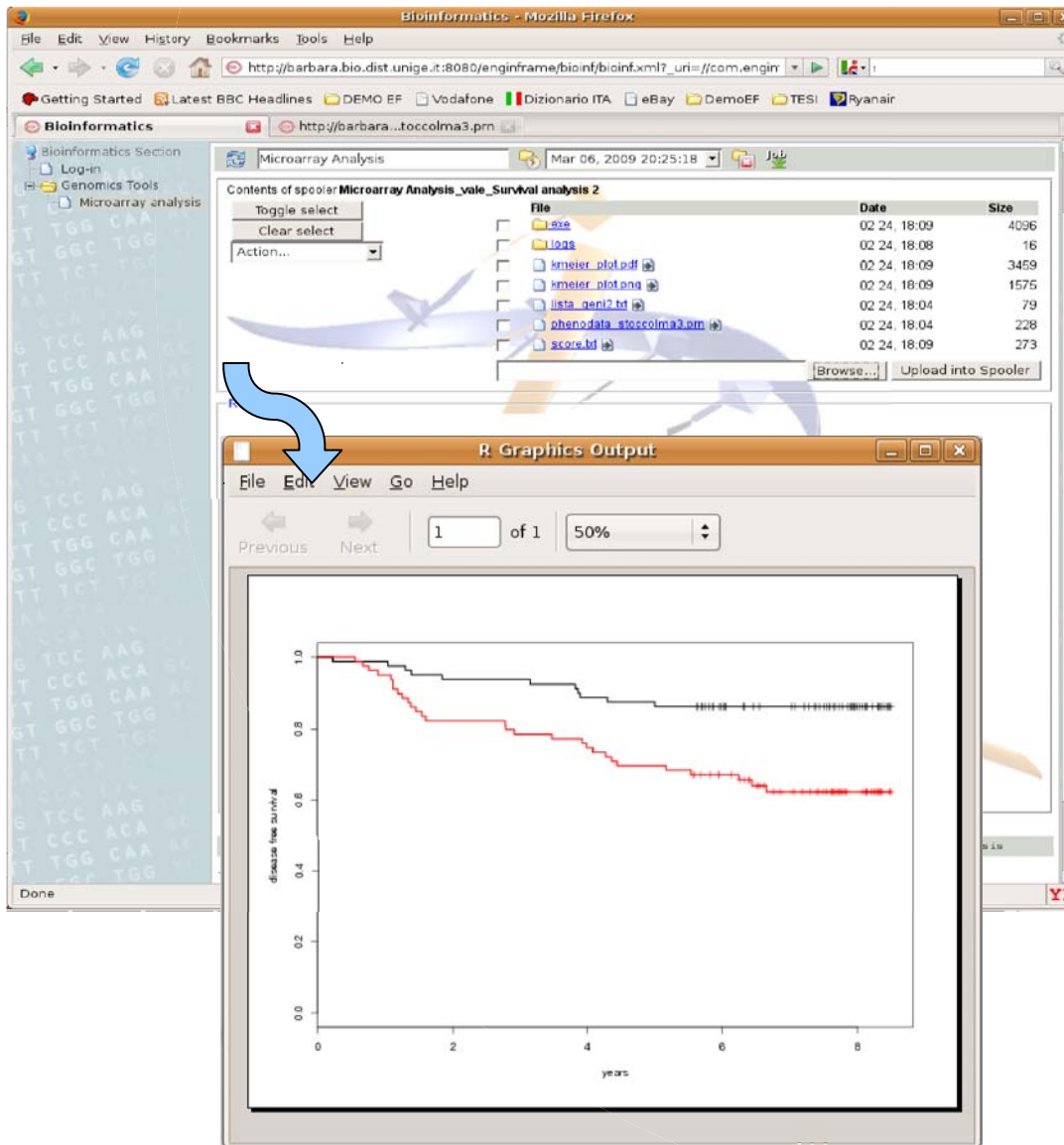
Project	Time	Status	Host	Name
6622	default Feb 24 15:17	DONE	gis-wm01.bio.dist.unige.it	Init
6623 [2]	default Feb 24 15:17	DONE	-	MA_Analysis
6624	default Feb 24 15:17	DONE	gis-wm01.bio.dist.unige.it	Normalize
6625	default Feb 24 15:17	DONE	gis-wm01.bio.dist.unige.it	Filter Genes
6626	default Feb 24 15:17	DONE	gis-wm01.bio.dist.unige.it	COX & Multivariate
6627	default Feb 24 15:17	DONE	gis-wm01.bio.dist.unige.it	Score computation
6628	default Feb 24 15:17	RUN	gis-wm01.bio.dist.unige.it	Kaplan-Meier plot
6629	default Feb 24 15:17	PEND		End operations

Case study: Cancer Survival Studies (3)

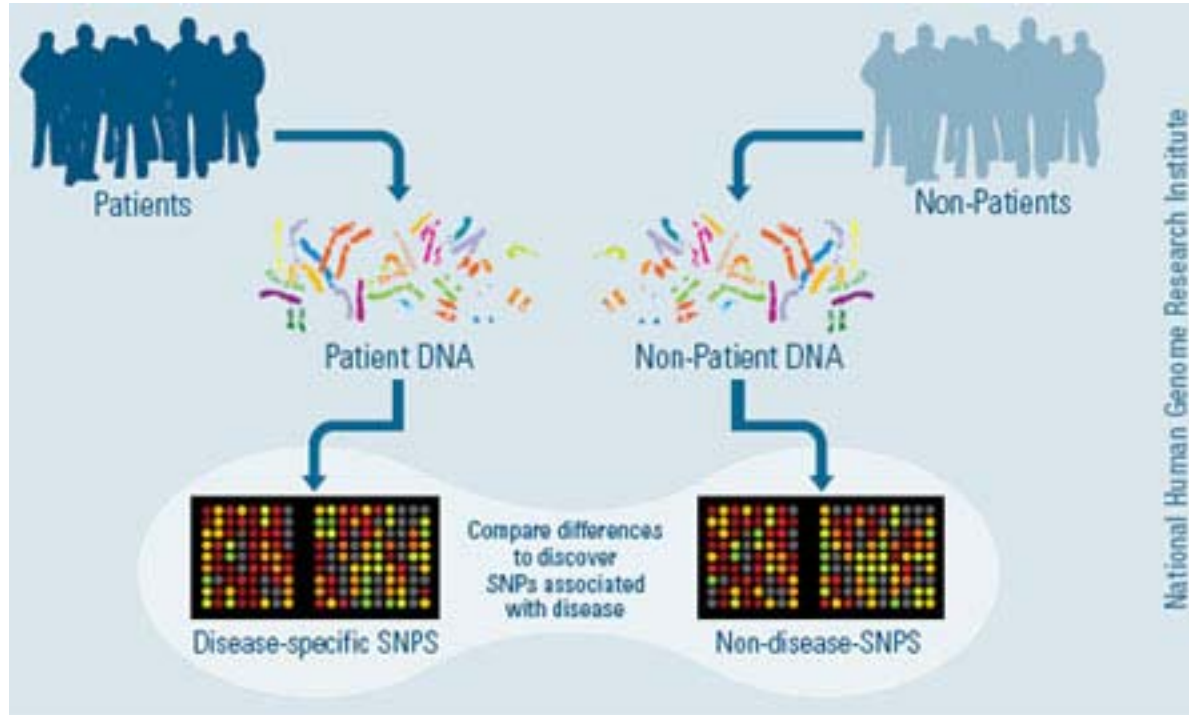
The analysis results (Kaplan-Meier plots) are shown in the portal area. The plot shows the two groups of patients, classified on the basis of low or high risk of disease relapse.

Advantages obtained:

- Integration of different software tools (dChip, R/Bioc) and data (gene expression, phenodata)
- Automation of procedures → time & cost reduction
- Computational power
- Experiment reproducibility
- Usability



Case study: Genome-Wide Association Studies (1)



In collaboration with:



Researchers analyze the DNA of two groups of participants: people with the disease (**Case**) being studied and similar people without the disease (**Control**). Each person's complete set of DNA is placed on tiny chips and scanned on automated laboratory machines, which quickly survey each participant's genome for strategically selected markers of **genetic variation**, which are called single nucleotide polymorphisms, or **SNPs**.



Case study: Genome-Wide Association Studies (2)

Quality Control of Genome-Wide Association Studies



Aim: Define strategies for setting appropriate genotyping rate thresholds for GWAS quality control based on decision theory.

Approach:

Maximize these parameters = maximize the power of the study

100 genotyping rate filters individual/SNP (90%→99%)

Genotyping rate thresholds

Alternatives: different combinations of Ind and SNP genotyping rate
(Campioni>90%,SNPs>90%);
(Campioni>90%,SNPs>91%)...

CRITERIA	ALTERNATIVES			
	CR90,SNP90	CR90,SNP91	...	CR99,SNP99
% of Individuals that pass the genotyping filters	99.45	99.24	...	60.66
% of SNPs that pass the genotyping filters	99.68	99.46	...	87.73
% of SNPs with differences in missing rates c/c (p<0.05)	12.88	12.83	...	93.49
% of SNPs with MAF < 1%	0.11	0.11	...	0.11
% of SNPs deviating from HWE (p<0.001)	0.29	0.27	...	0.11
Heterozygosity Rate Standard Deviation (per individual)	0.01	0.01	...	0.00
Genomic Inflation Factor	1.31	1.31	...	1.18

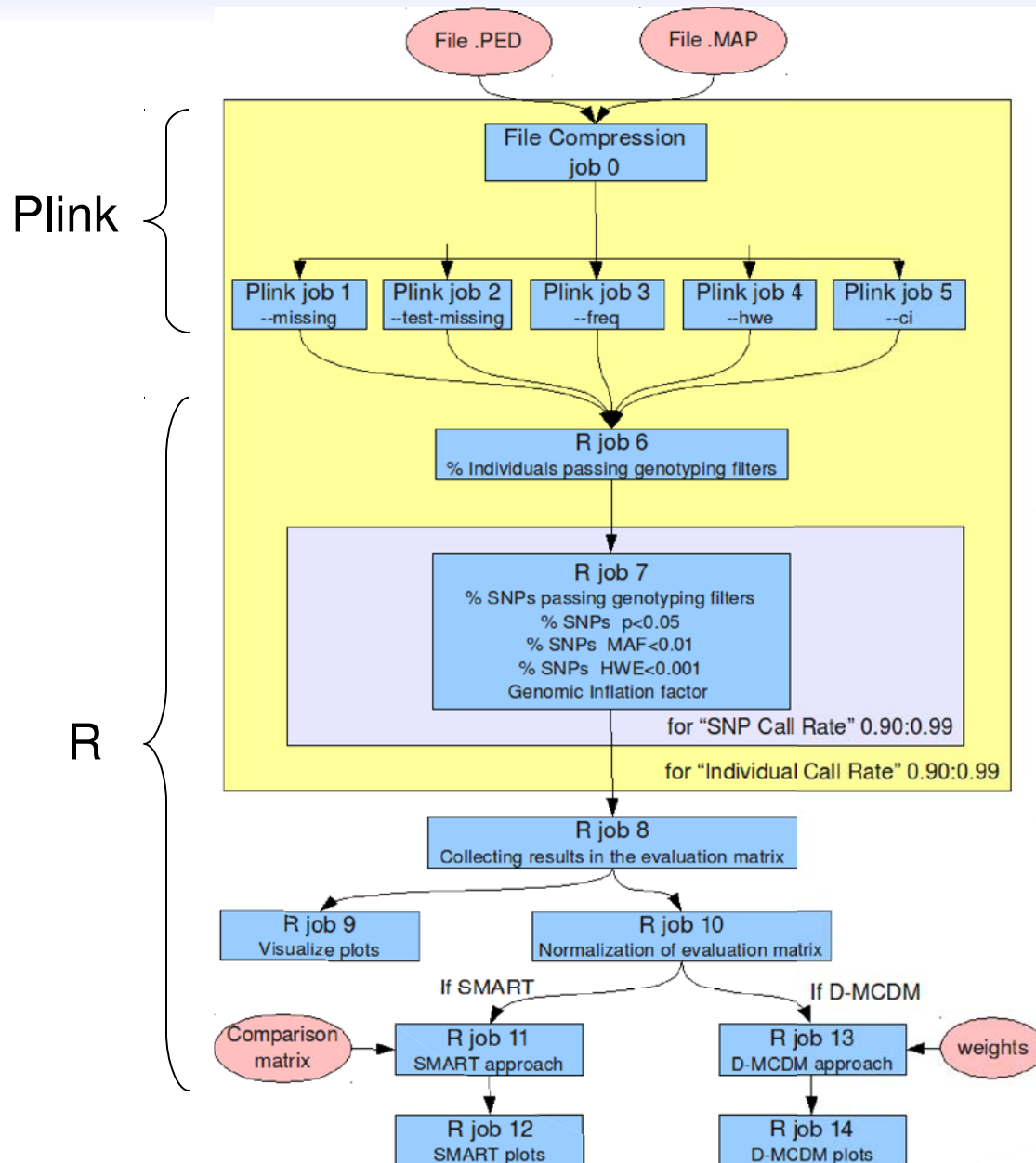
Decision Theory model

Score profile for alternatives

Minimize these parameters = minimize genotyping errors



Case study: Genome-Wide Association Studies (3)



Design the analysis pipeline and study the best computational model to optimize computational effort

... this service submits more than 100 independent jobs, one for any alternative...

100 alternatives

Case study: Genome-Wide Association Studies (4)

Advantages obtained:

- Reduction of computational time → from hours to minutes
- Very large dataset analysis (Giga datasets)
- Integration of different software tools (plink, R/Bioc)
- Automation of procedures → time & cost reduction
- Experiment reproducibility
- Usability

The screenshot displays the Bioinformatics web interface in Mozilla Firefox. The main page is titled "Genotyping QC analysis" and includes a "Service description" and a list of four key benefits: 1. reduce the computational time required for analyses, 2. build automated analysis pipelines that can be shared across the network, 3. run complex application-specific experiments on large data sets, and 4. increase user accessibility by providing user-friendly Web-based interfaces. The interface prompts the user to upload a MAP file and a PED file, and to choose between SMART and D-MCDM score approaches. A table for parameter selection is shown below:

Inds.CR	SNP.CR	p.miss	MAF	HWD	lambda
1	1	5	5	5	5
1	1	5	5	5	5
0.2	0.2	1	1	1	1
0.2	0.2	1	1	1	1
0.2	0.2	1	1	1	1
0.2	0.2	1	1	1	1

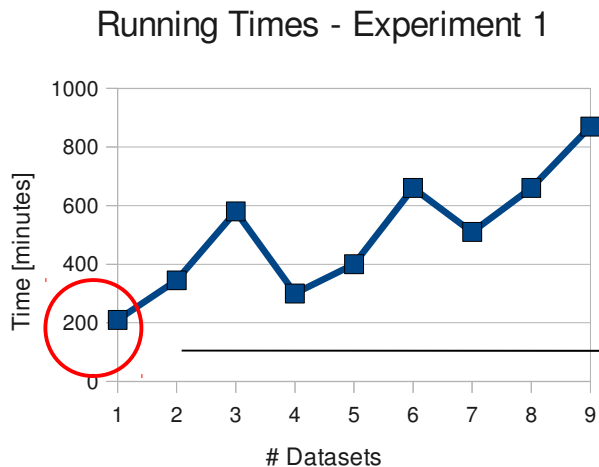
A file browser window shows a directory listing of files for analysis, including various cleaned and processed files. A second screenshot shows the "Contents of spooler" for the analysis, listing files such as "maf_scores_definitive_per_alternatives.col.names.ok_1.png" and "maf_scores_definitive_per_alternatives.col.names.ok_2.png". A large blue arrow points from the parameter table to the analysis results.

The analysis results are displayed in two plots: a "Histogram of MAF" showing the frequency distribution of Minor Allele Frequencies, and a 3D surface plot showing the relationship between SNP, MAF, and Sample. The histogram shows a high frequency of SNPs with low MAF values, while the 3D plot shows a surface representing the MAF values across different samples and SNPs.

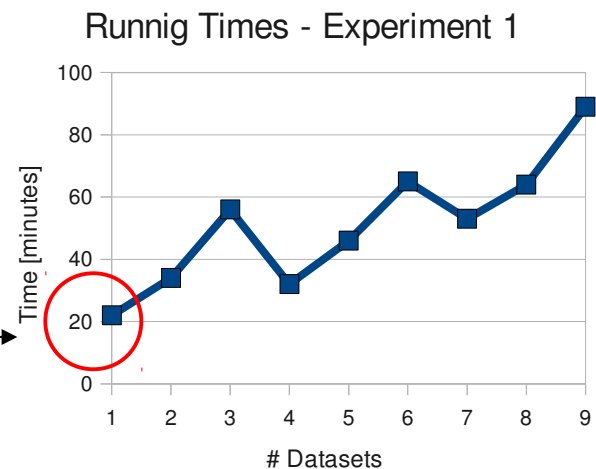
Case study: Genome-Wide Association Studies (5)

Computational tests

Execution on PC/Workstation



Execution on HPC/Grid

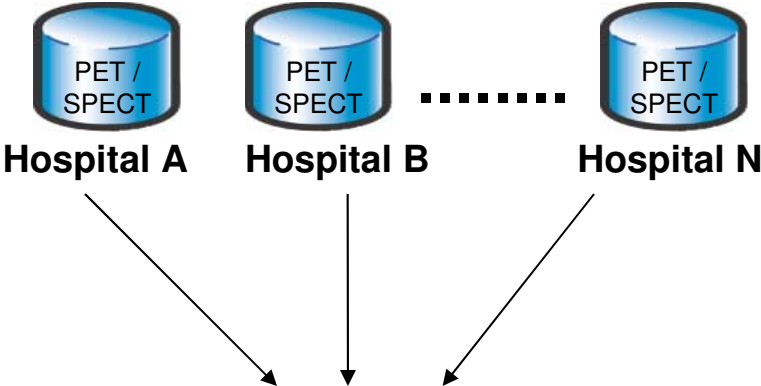


Time reduction!

Dataset 1	cases	controls	SNPs	time PC	time Grid (minuti)
sim-500-samples-300k	500	500	300k	210	22
sim-500-samples-500k	500	500	500k	345	34
sim-500-samples-1000k	500	500	1000k	580	56
sim-1000-samples-300k	1000	1000	300k	300	32
sim-1000-samples-500k	1000	1000	500k	400	46
sim-1000-samples-1000k	1000	1000	1000k	660	65
sim-2000-samples-300k	2000	2000	300k	510	53
sim-2000-samples-500k	2000	2000	500k	660	64
sim-2000-samples-1000k	2000	2000	1000k	870	89

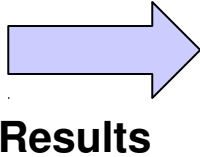
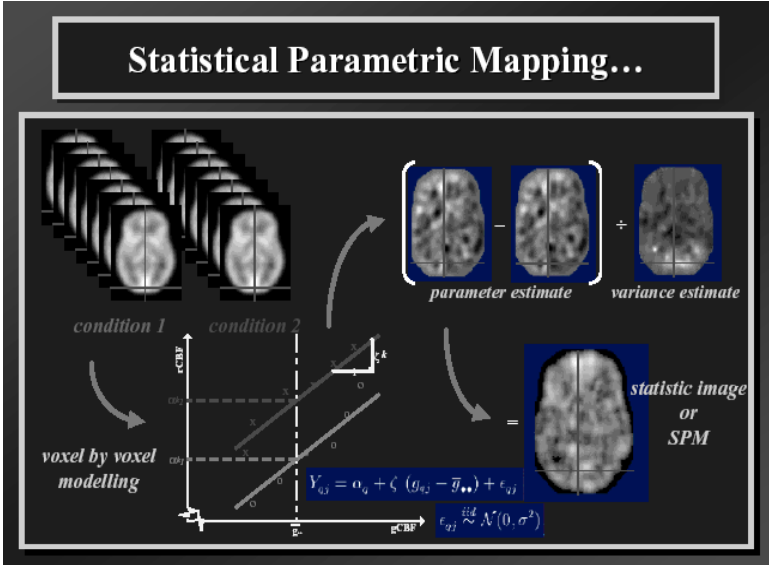


Case study: Data management, Analysis and Visualization in Neuroinformatics (1)

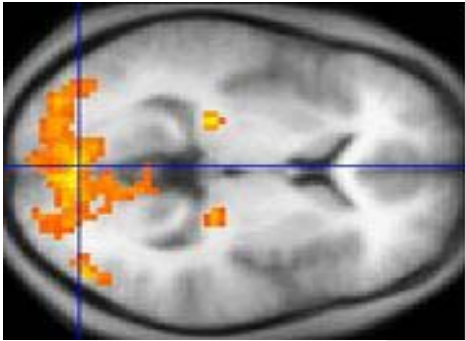


Aim: perform the early diagnosis of Alzheimer through a statistical comparison of PET/SPECT patient image with a database of reference patient images.

Advantages obtained: remote analysis of distributed images



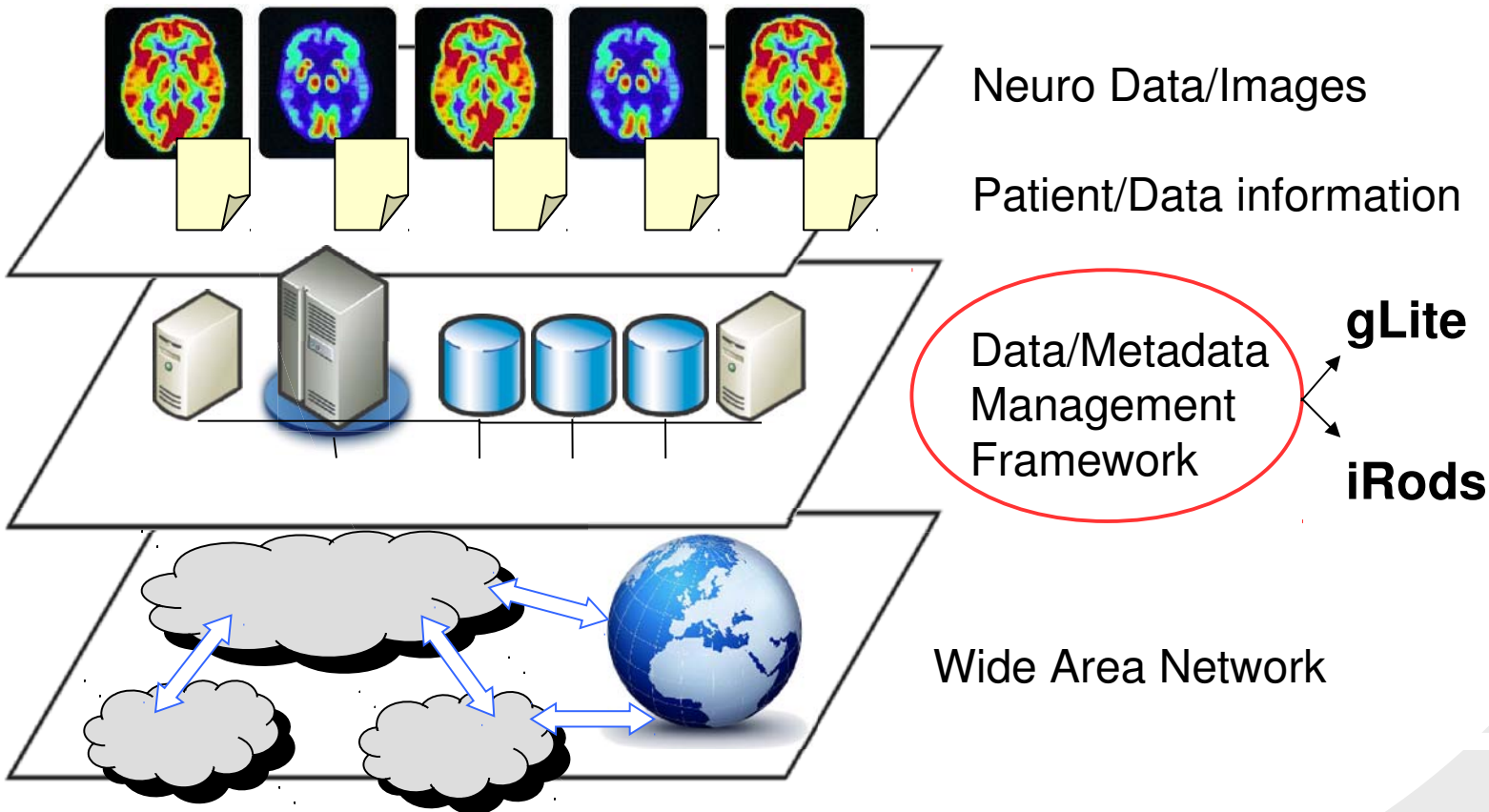
Results



Patient image where suspected brain areas are visualized

Case study: Data management, Analysis and Visualization in Neuroinformatics (2)

How to access to distributed images? How to manage metadata?



Case study: Data management, Analysis and Visualization in Neuroinformatics (3)

gLite implementation



The screenshot shows the neuroinf.it web application interface. The top navigation bar includes 'Home', 'Help', 'My data', and 'My analyses'. The main content area is titled 'New patient' and contains a form for entering patient information. The form fields are: Age (53), Sex (M), Header (16_SPECTm56.hdr), Image (16_SPECTm56.img), and Normalized (YES). An 'Upload patient' button is located below the form. A blue arrow points from the 'Upload patient' button to a second browser window. This second window displays the 'SPM_results' page, which features a 3D brain model with a red region of interest, a 2D axial slice, and a 'Design matrix' plot. The 'Design matrix' plot shows a 3x3 grid of values: 0, 1, 0 in the first row; 0, 1, 0 in the second row; and 0, 0, 0 in the third row. The URL in the address bar is http://marilyn.bio.dist.unige.it:8080/enginframe/spm/SPM.xml?_service=new_pat.

Aim: design, implement, and validate a **service** for the computer-aided extraction of diagnostic markers for Alzheimer's disease and schizophrenia from medical images on a GRID-based e-Infrastructure.

NICE is involved in Front-end development and services provision by integrating EnginFrame with Liferay technology.

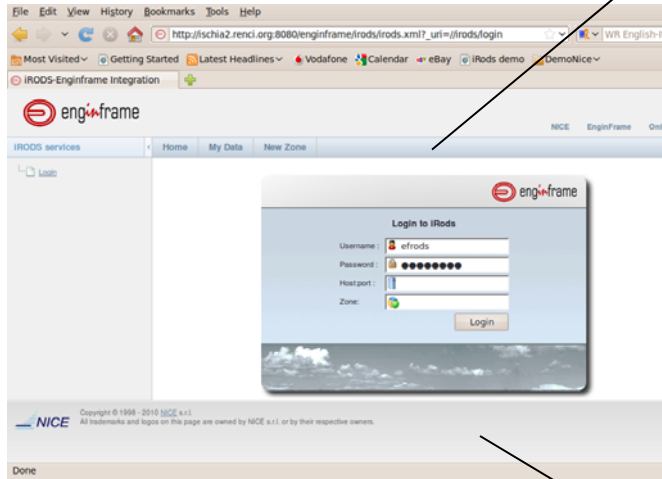
Case study: Data management, Analysis and Visualization in Neuroinformatics (4)



iRODS implementation

Solution for distributed data management and organization of data into sharable logical collections

Replication of data on different physical resources



Visualization & setting of metadata

The screenshot shows the EnginFrame web interface. The main window displays a file list with columns for Name, Size, Resource, and Date Modified. A 'Replicate' dialog box is open, showing a dropdown menu for 'tucasi-unc' and 'Replicate' and 'Cancel' buttons. Below the file list, a 'Metadata for: /TUCASI/home/efrodsadmin/P16_SPECTm56.hdr' dialog box is open, showing a table of metadata attributes and values.

Attribute	Value	Units
<input type="checkbox"/> Age	56	
<input type="checkbox"/> Image format	.hdr	
<input type="checkbox"/> Image type	SPECT	
<input type="checkbox"/> Patient ID	16	
<input type="checkbox"/> Sex	M	

Case study: Data management, Analysis and Visualization in Neuroinformatics (5)

The image displays a web-based visualization portal titled "Visualization Demo Portal - Mozilla". The interface includes a menu bar (File, Edit, View, History, Bookmarks, Tools, Help) and a sidebar with "Interactive Services" such as "My Sessions", "My Data", "My Jobs", and "All Jobs". A table lists active sessions:

Name	Status	Started on
SeegViewer	running	Oct 22 12:20
Google Earth	running	Oct 22 12:45
XTerm	running	Oct 29 10:55
Accelrys DS Visualizer	running	Oct 29 11:00
CEI EnLiten - EF 3D Views	running	Oct 29 11:26

The main content area features a 3D view of a brain model and four 2D views: Axial View (Radiological Conv.), Sagittal View, Coronal View (Radiological Conv.), and Slice View. Below the views are control panels for "Surface 1" (Value: 1, Surface: Red), "Slice Controls" (X: 127, Y: 127, Z: 127, alpha: 0, beta: 0, gamma: 0), and "Axial Controls", "Sagittal Controls", and "Coronal Controls" (Crop +, Box, Crop +).

Advantages obtained: access to interactive 2D/3D applications remotely, in a collaborative environment, through a common web browser



Outline

NICE

NICE Products&Solutions

Life Sciences Solutions

Case studies

Demonstrations

