Enabling User Communities: Experience from porting applications and services in the biomedical area

Ignacio Blanquer

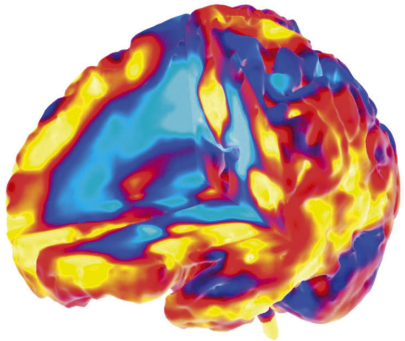
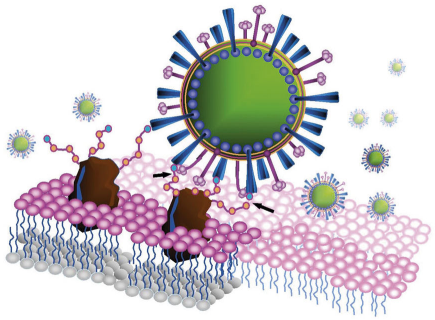
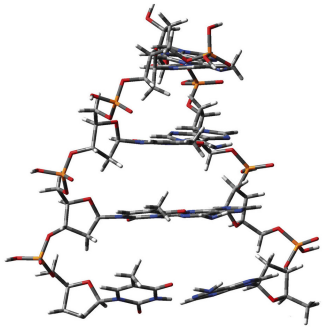
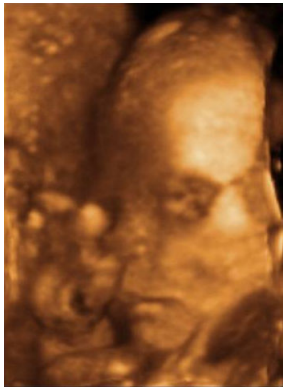
Outline

- The Biomedical community.
- The Biomed VO in EGEE.
- Problems and Needs of the Biomed Community.
- Approaches, Components and Applications.
- Successful Stories.
- Conclusions.

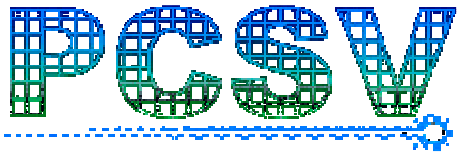
- Many slides have been borrowed from members of the Biomed VO.

The biomedical Community

- Biomedicine integrates many different disciplines related with health, life sciences and biochemistry.
- It comprises the storage, management and processing of data related with the physiology and structure of living beings.
- So the Biomed Community is wide, heterogeneous and has many different challenges.

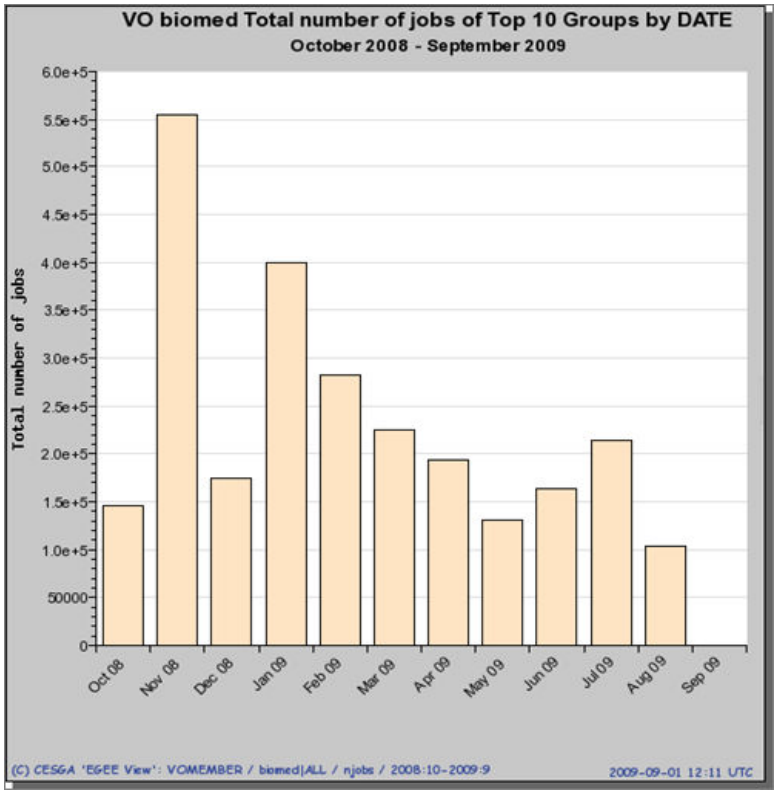





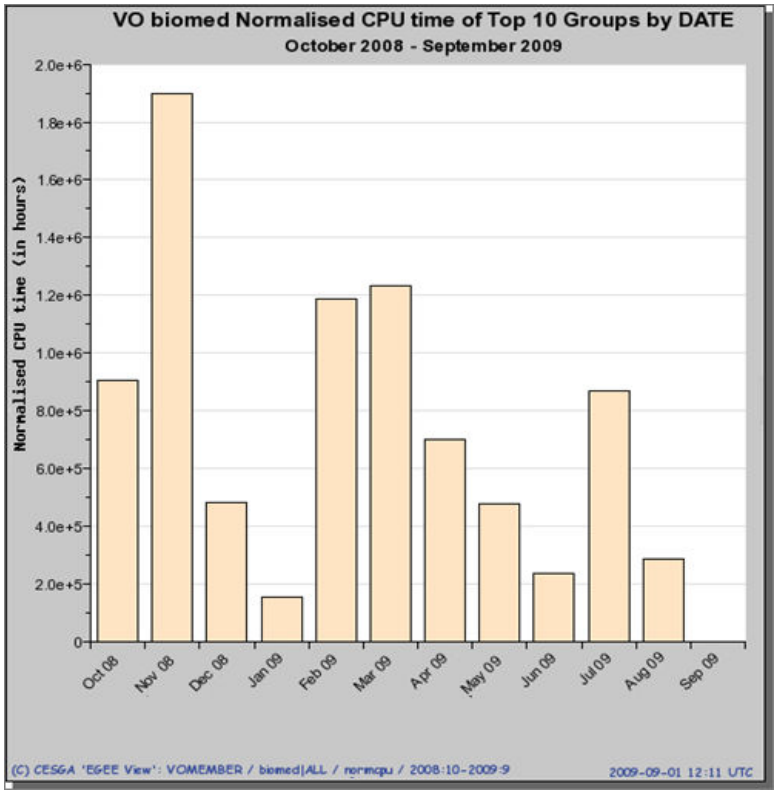
The Absolutely Incomplete EGEE-Related Biomed Community



The Biomed VO in EGEE

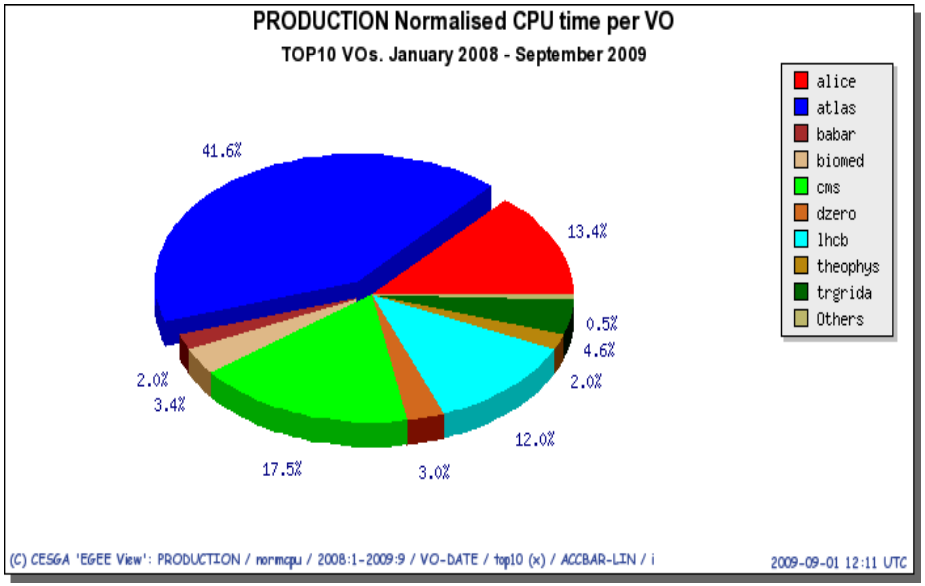
- 253 Registered users.
- Authorised for and important share of the resources available (around 25%).
- 8 Million jobs and 25 million CPU hours since the starting of the accounting.

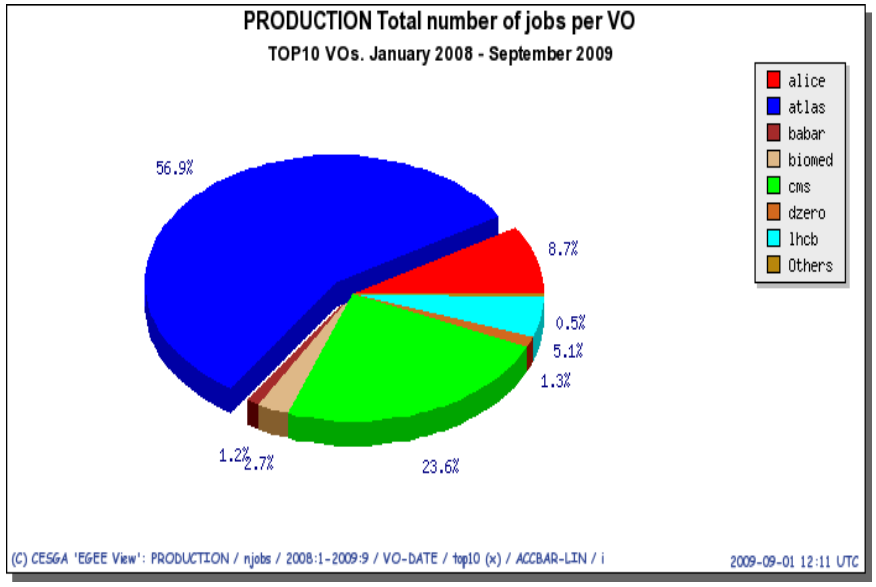




Infrastructure usage by biomed VO

- **Although there is an important difference with HEP VOs, Biomed is the most active VO outside of this topic.**





Problems and Needs of the Biomed Community

Grid Applications in Biomedicine

Genomics / Proteomics
Biomedical Simulation - VPH
Innovative Medicine

Epidemiological studies
Medical Imaging
Clinical Trials

Theraphy Follow-on
Telemedicine

6 ESFRI projects are related to biomedicine and already several ones use Grids

Biomed Communitiy Targets

Medical Imaging

- Medical Imaging federation and processing.
- Oncology, Cardiology, Neurology.

Epidemiology

- Efficiency studies.
- Spreading of diseases.

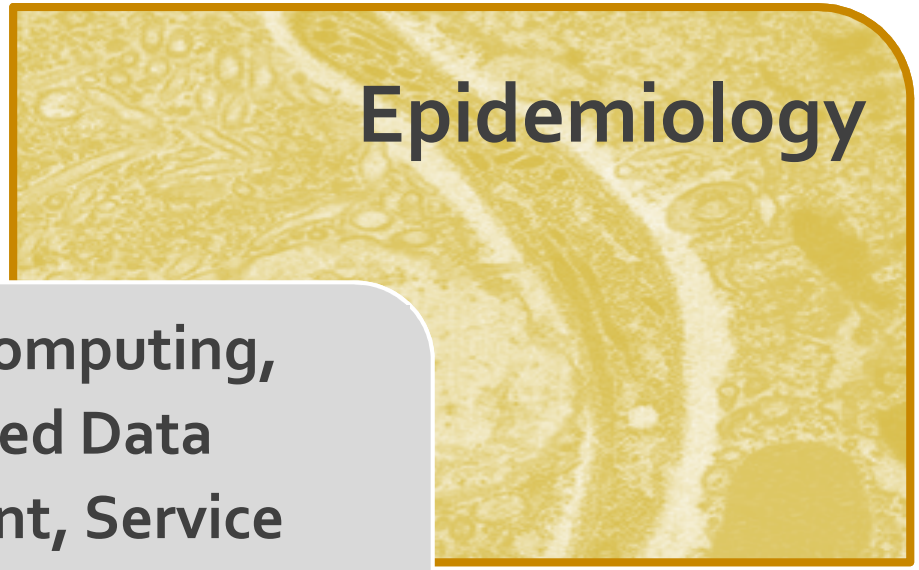
Bioinformatics

- Complex processing on large Bioinformatics Databases.
- Alignment, Phylogenetics, System Biology.

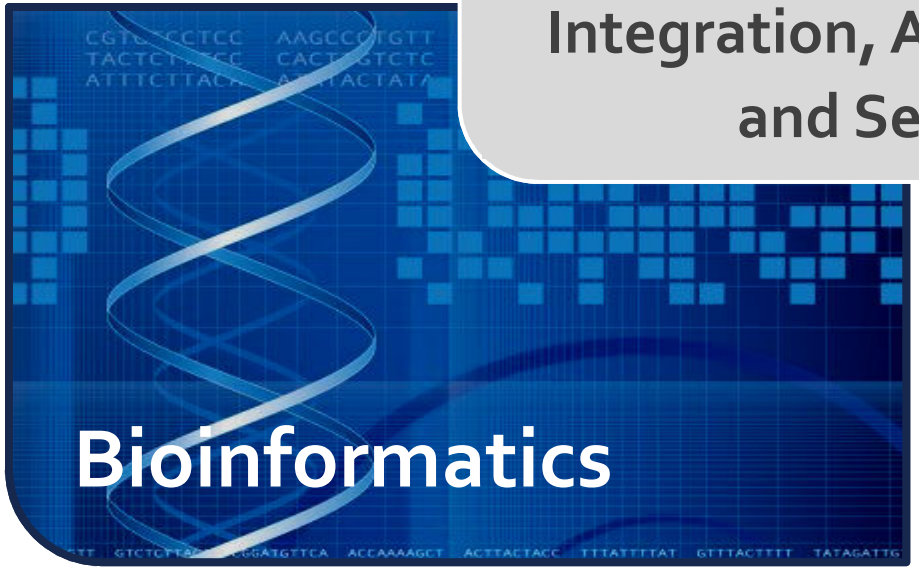
Innovative Medicine

- Drug discovery.
- Resistance to treatment.
- Treatment personalisation.
- Modelling of Diseases and Organs.

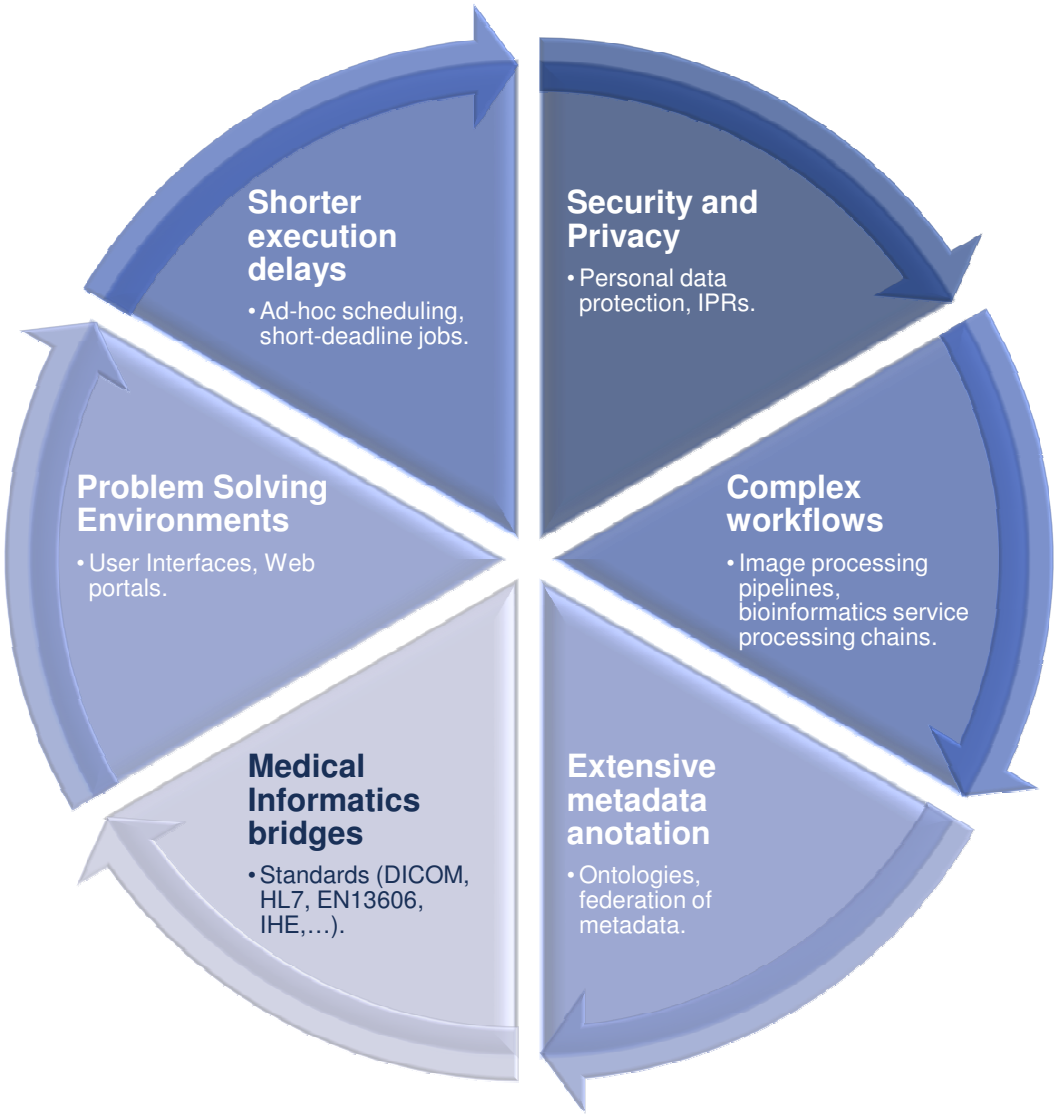
Biomed ICT Needs



Intensive Computing, Distributed Data Management, Service Integration, Authorisation and Security

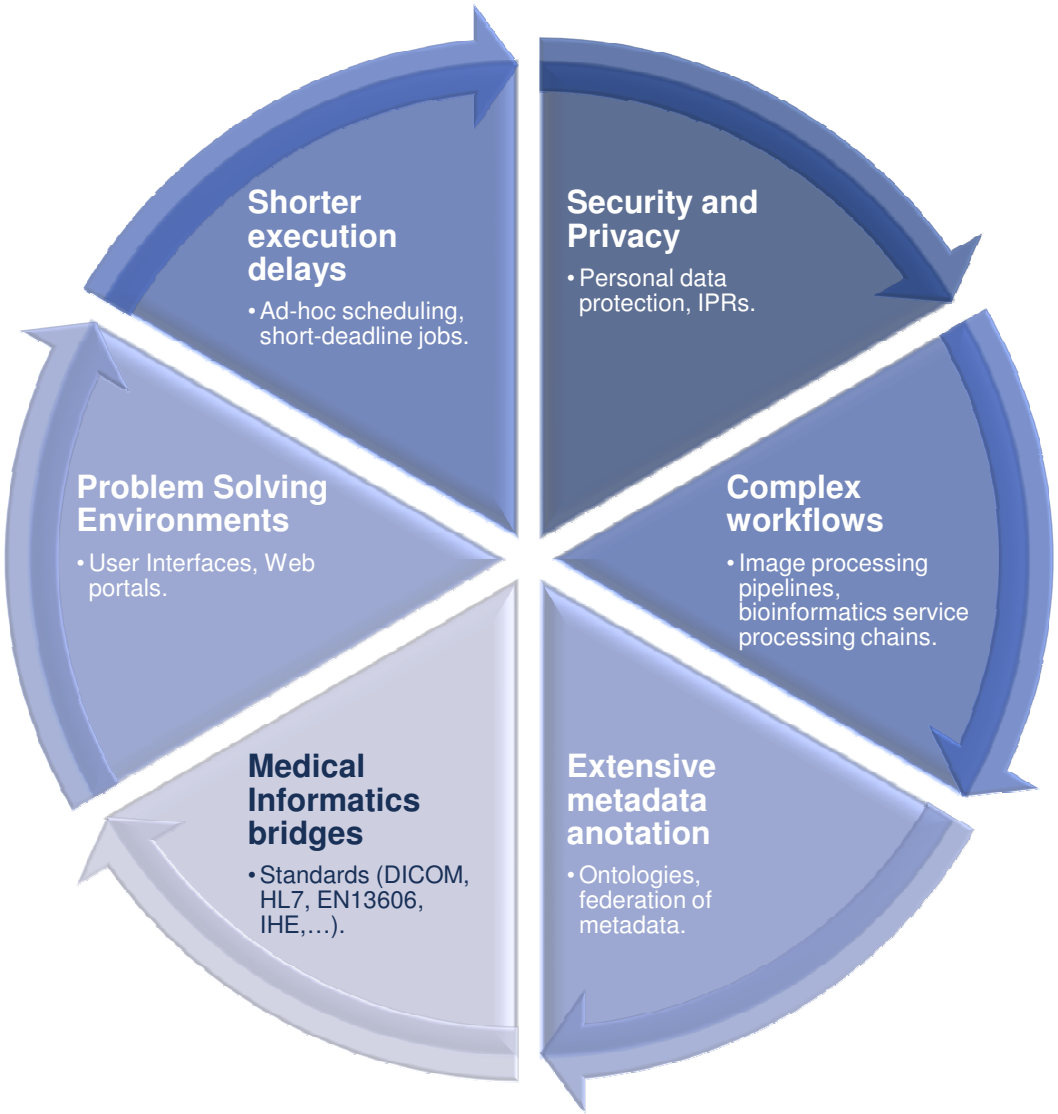


Specific Challenges faced by biomed community



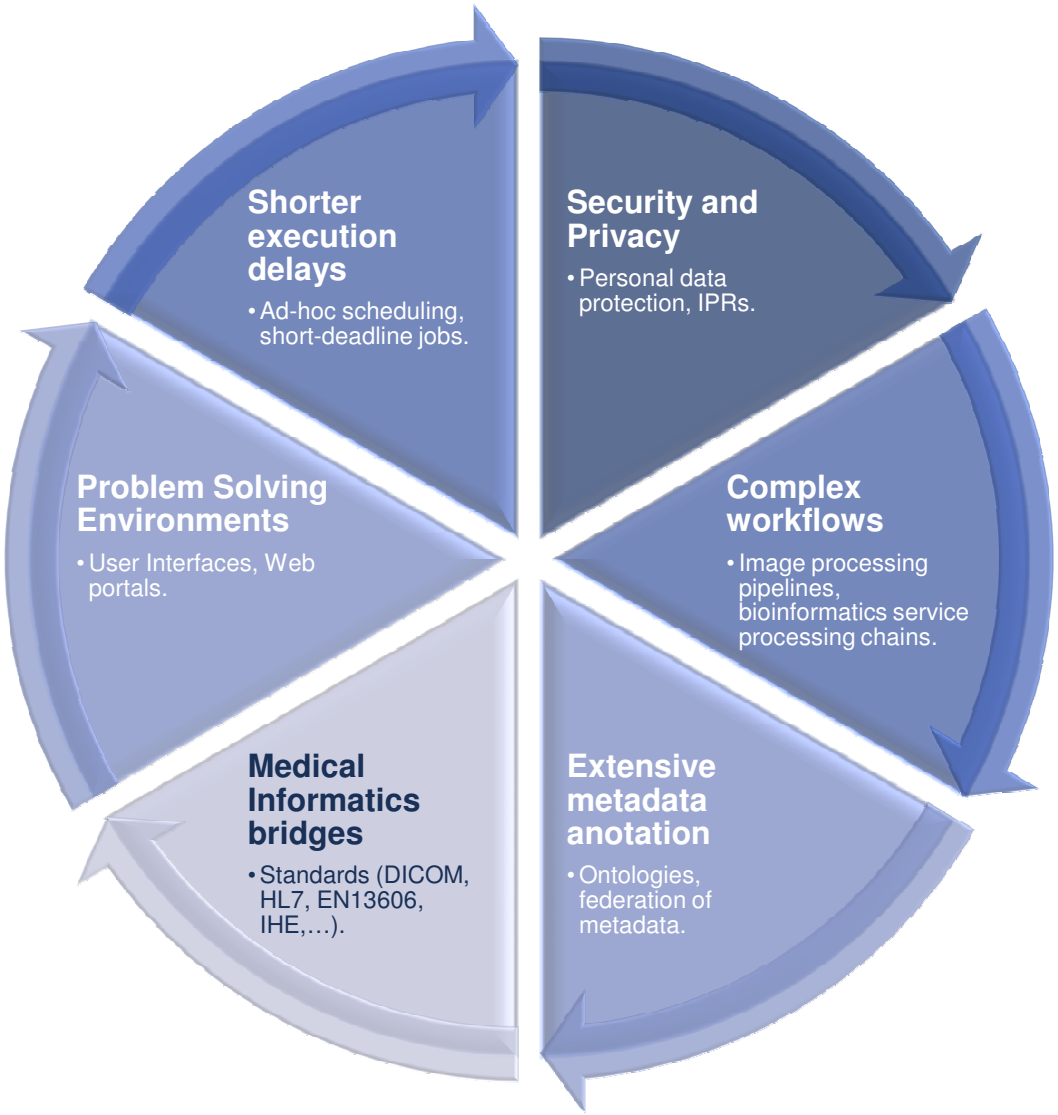
SHARE
Roadmap for European HealthGrid

Specific Challenges faced by biomed community



SHARE
Roadmap for European HealthGrid

Specific Challenges faced by biomed community



S.H.A.R.E
Roadmap for European HealthGrid

Approaches, Components and Applications.

Biomed Activity – Bottom-up Approach

Deployments and Experiments

- Analysis on the accuracy of annotations on gene and protein databases.
- Effect of the mutation of specific genes on cardiology diseases.
- Modelling the spreading of infectious diseases.
- Evaluating the efficacy of radiotherapy treatments.
- Identifying biomarkers for neurodegenerative diseases.
- Selecting new components for the treatment of emerging diseases.

Biomed Activity – Bottom-up Approach

Deployments and Experiments

- The EGEE Grid Infrastructure.
- Availability of large computing data storage and communication resources.
- Enough resources for most of the currently interesting research challenges.

Infrastructure

Biomed Activity – Bottom-up Approach

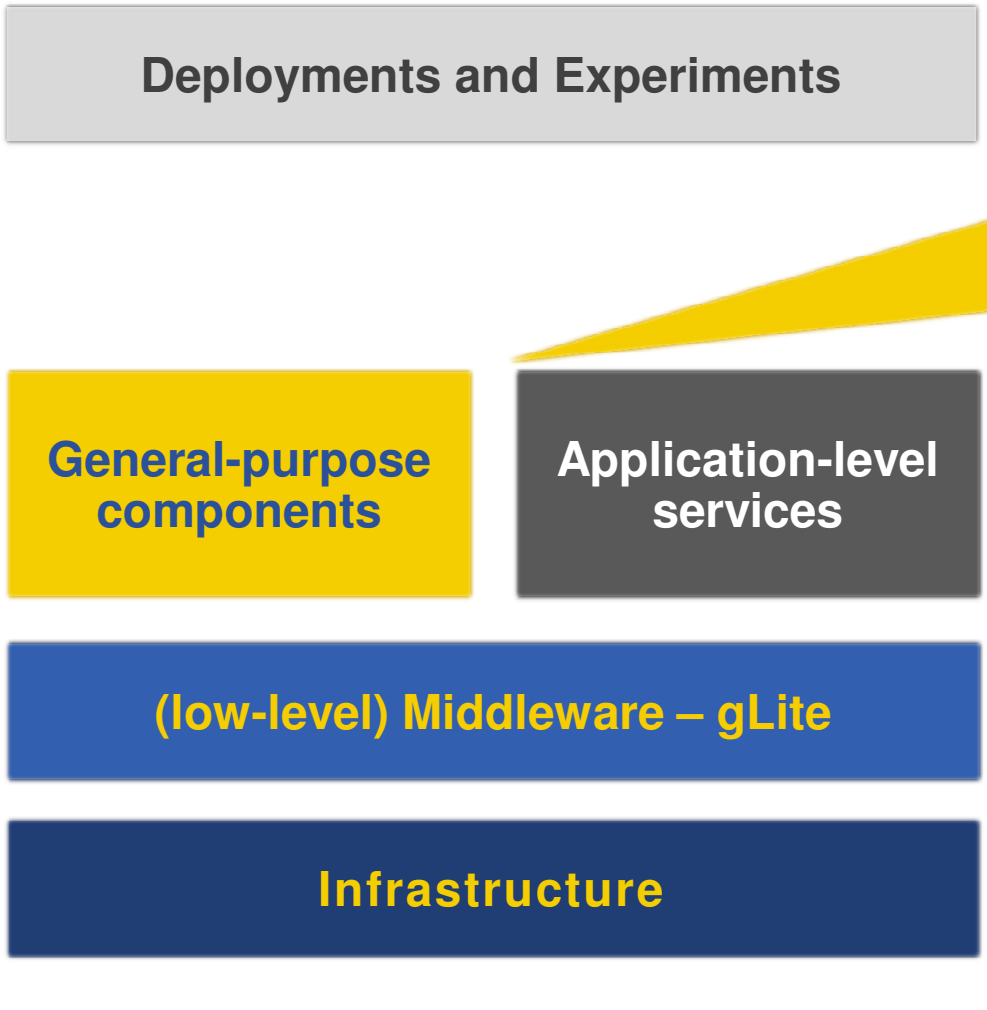
Deployments and Experiments

Middleware – gLite

Infrastructure

- Feedback on requirements for Enhanced security
  - ACLs for SEs and Catalogues.
  - Inclusion of 3rd party components, such as Keystores (Hydra).
- Shorter job management delays
  - Support for specific queues for Bulk Short deadline jobs.
  - Extensive interest on pilot jobs.
- Metadata storage
  - Collaborate with AMGA.

Biomed Activity – Bottom-up Approach



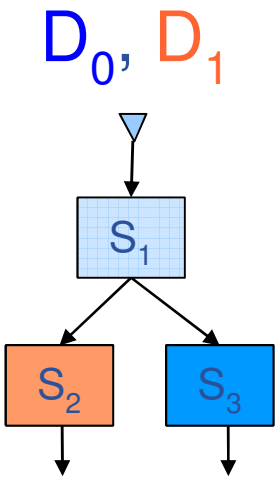
- General components that are not directly supported by the middleware or other communities' components.
- Complex workflow management: Moteur.

- Specific components related to the application area
  - Medical Image Federation (MDM, TRENCADIS).
  - Problem Solving Environments (VI-e-Med).
  - Medical Grid Nodes (Pandora Box).

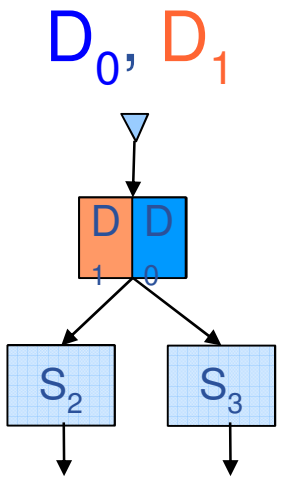
Efficient parallel execution of data flows

- **Enacting services on a batch-oriented grid infrastructure**
  - Submission web service.
- **From workflow manager to grid execution**
  - Execution engine independent from grid middleware.
  - Interfaced to different grid middlewares (gLite/LCG2, DIET, OAR...).
- **Enhanced support to 3 kinds application parallelization.**
- **Jobs grouping strategy in sequential branches in order to reduce grid latency.**

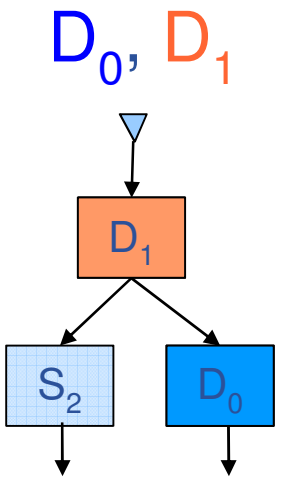
Workflow parallelism



Data parallelism



Service parallelism

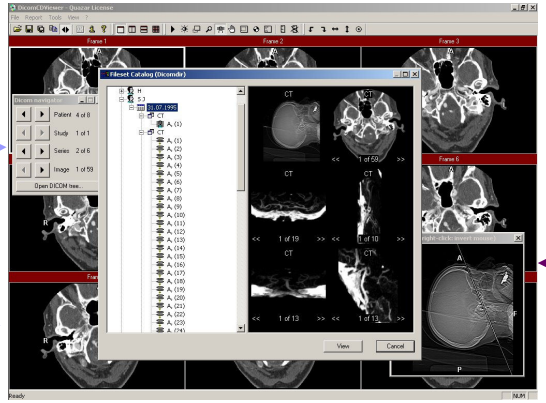


http://modalis.polytech.unice.fr/softwares/moteur/start

Medical Image Federation in EGEE: Medical Data Manager

- **Objectives**
  - Expose a **standard grid interface** (SRM) for **medical image servers** (DICOM).
  - Use native DICOM storage format.
  - Fulfill medical applications security requirements.
- **Enables storing and retrieving Medical Images and metadata on the EGEE Grid.**

DICOM clients



DICOM server

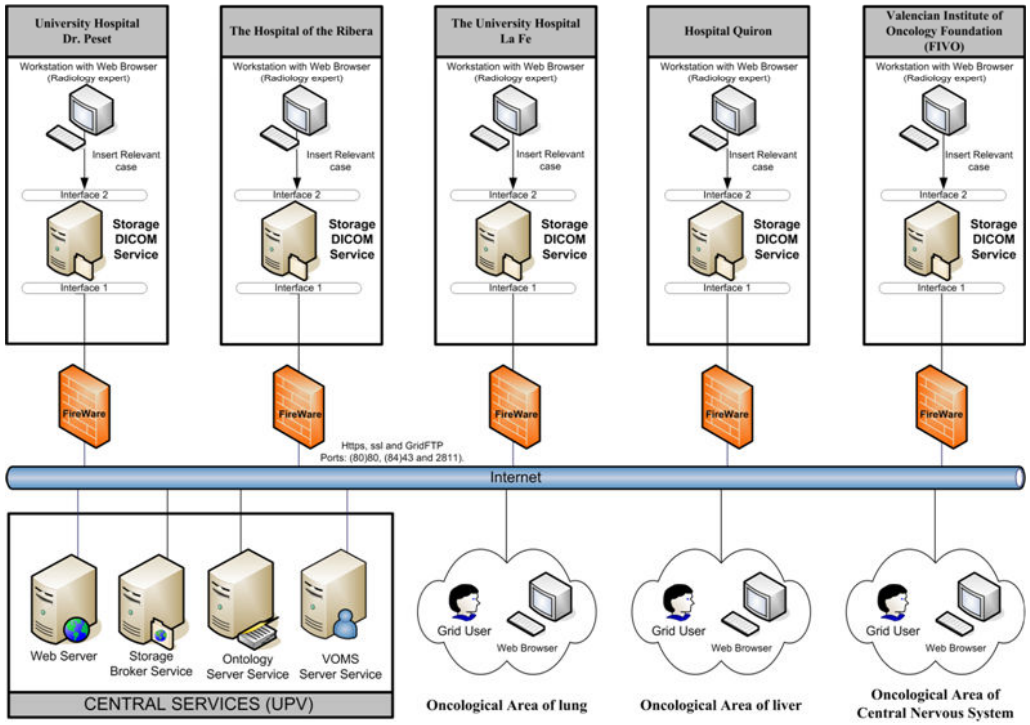
DICOM Interface SRM

Worker Nodes

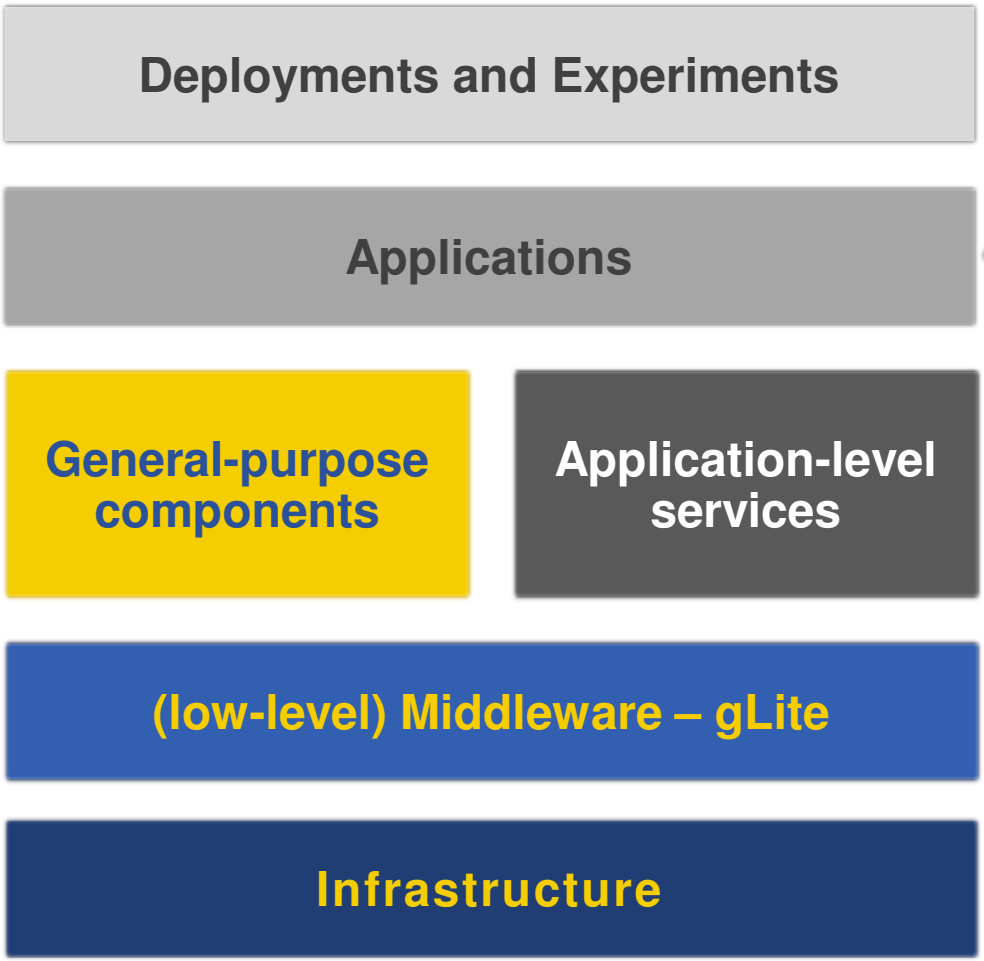
User Interfaces

Radiology Reports and Image Federation in EGEE: TRENCADIS

- **Towards a gRid ENvironment for proCessing and shAring DIcom objectS (TRENCADIS)**
  - Software Architecture, based on the WSRF and gLite.
  - Integrates different local storages of DICOM objects from several centers.
  - Different storage resources are virtualized providing a common interface.
  - It organizes data by communities.
  - Indexing is performed locally services keep references to the storages where information relevant to each community is stored.



Biomed Activity – Bottom-up Approach

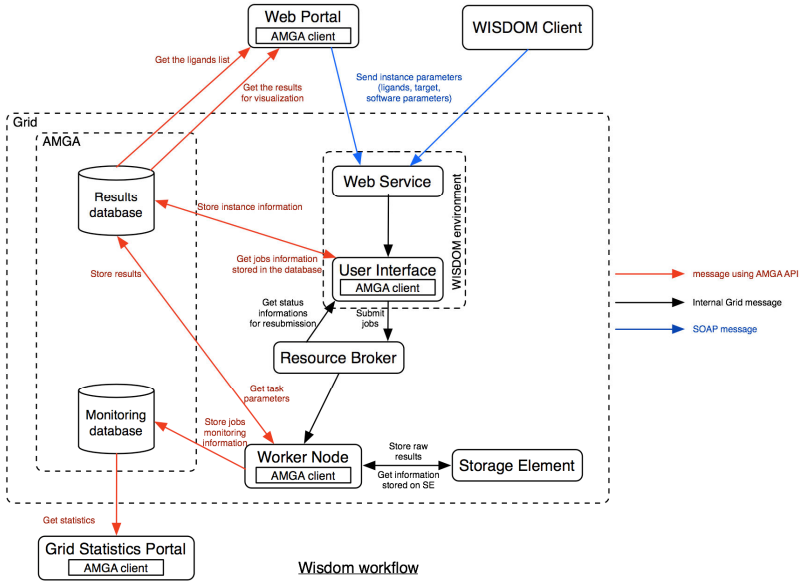


- BLAST in Grids Metagenomics analysis, gPS@, System Biology
- IntegraEPI, Virtual City,
- HOPE, ThIS.
- HeC, Neurogrid, Neurolog, CVIMO.
- WISDOM Environment.
- gPTM3D, Pharmacokinetic modelling.
- SimMRI, gCamaec.

WISDOM Production Environment

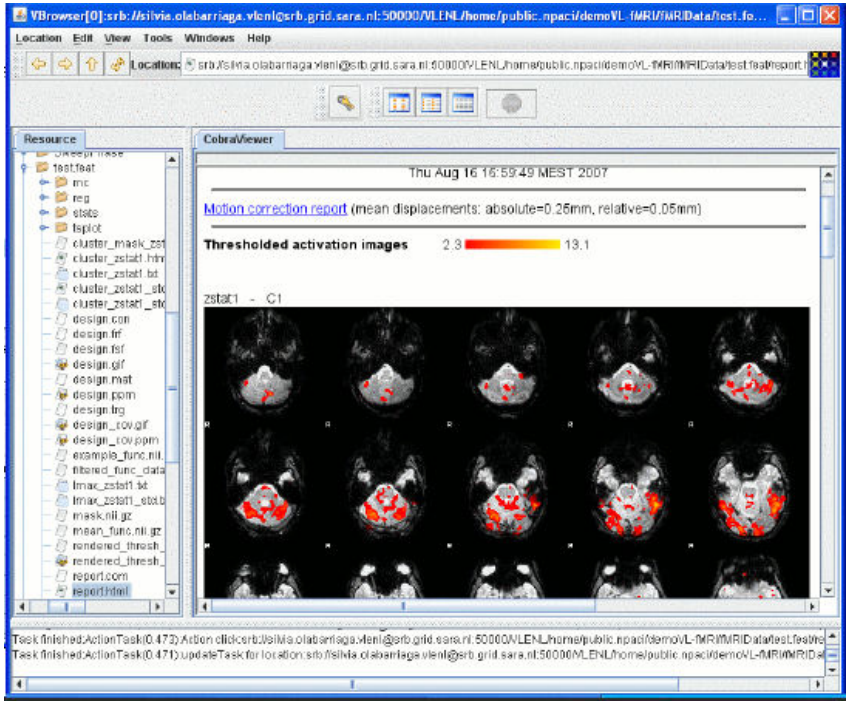
- **The WISDOM Production Environment aims at managing and monitoring the jobs on the EGEE Grid.**
- **It uses most of the EGEE Grid services.**
- **Automatically creates and Submit jobs, the jobs using multithreaded submission.**
  - Check the status of the jobs using multithreaded check.
  - Resubmit jobs if needed.
  - Re-initialize voms proxy if needed.
  - Update instance information in AMGA.
- **Developed in Java.**
- **Uses its own ranking with BDII and own data.**
- **Dynamic storage and query of data using AMGA.**
- **Web Service interface.**

http://wisdom.eu-egee.fr/



VL-e-Med

- **The VL-e Med project has the goal of building grid-enabled problem solving environment for medical imaging. The resulting platform is based on web services activated from a user-friendly graphical interface.**
  - Front-end
    - Virtual Resource Browser.
  - Data
    - SDSC SRB (Storage Resource Broker).
    - gLite LFC.
  - Workflows
    - Scufl (Taverna workbench).
    - MOTEUR.
  - Job submission/monitoring
    - gLiteWMS.



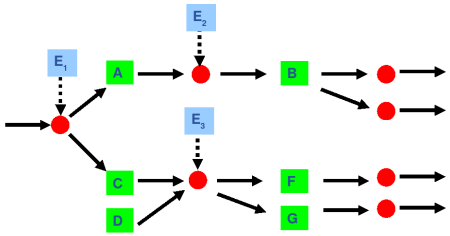
Successful Stories

Success Stories

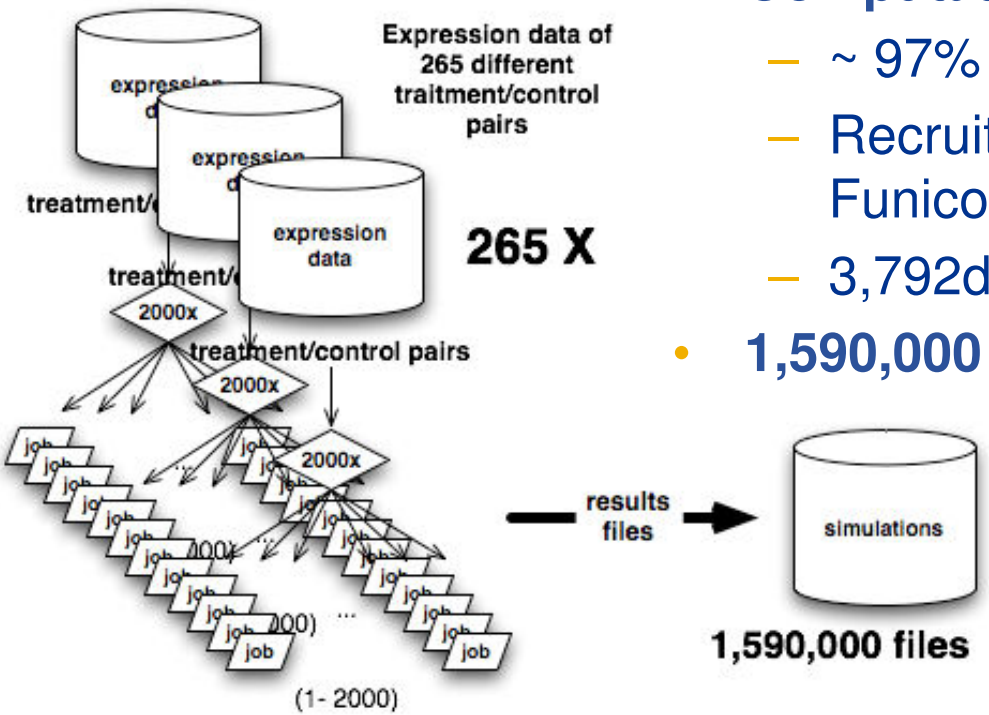
- **Publication in top-level journals in Biology and Health normally requires several years of research.**
- **Biomed community has really started to work using Grids in scientific production since recently**
  - 658 articles in Google Schoolar (EGEE+grid+biomed).
  - 30 articles in the ISI Web of Knowledge (EGEE+grid, only JCR indexed journals in health and life sciences)
    - 2 Publications in Nature and Nature Genetics.
    - 3 in Bioinformatics.
- **Patents on research results.**

Systems Biology on Grid

MBRACE Grid
Network of excellence



- **Systems biology investigates the behaviour of complex systems of interacting components**
  - experimental data on the response of cancer cell lines to 164 different small-molecule perturbagens.
  - Monte Carlo approach with PyBIOS.
- **Computation: 102 years (910,000 h)**
  - ~ 97% after 17 days (acceleration 2200x)
  - Recruiting up to 4,000 CPUs with Funicolare.
  - 3,792different WNs identified.
- **1,590,000 result files => 1,35 TB of data**



Herwig@molgen.mpg.de,
Christophe.Blanchet@ibcp.fr

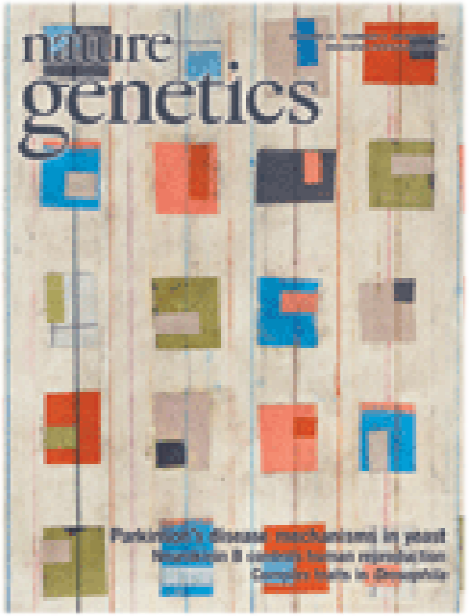
MPIMG

Large scale docking deployments against malaria

Malaria target	Biology partners	Involved in	Status
GST from *Plasmodium falciparum*	U. of Pretoria, South-Africa	Parasite detoxification	In vitro tests
DHFR from *Plasmodium vivax*	U. of Los Andes, Venezuela, U. of Modena, Italy	Parasite DNA synthesis	In vitro tests
DHFR from *Plasmodium falciparum*	U. of Modena, Italy	Parasite DNA synthesis	In vitro tests
Plasmepsin	SCAI Fraunhofer Chonnam Nat. Univ.	Hemoglobine degradation	Molecules patented

Genome-wide haplotype association study on risk for coronary artery disease

- **Goal: study the impact of DNA mutations on human coronary diseases.**
- **Very CPU demanding analysis to study the impact of correlated (double, triple) DNA mutations.**
- **Deployment on EGEE Grid**
  - 1926 CAD (Coronary Artery Diseases) patients & 2938 healthy controls.
  - 378,000 SNPs (Single Nucleon Polymorphisms = local DNA mutations).
  - 8.1 millions of combinations tested in less than 45 days (instead of more than 10 years on a single Pentium 4).
- **Results published in *Nature Genetics March 2009* (D. Tregouet et al)**
  - Major role of mutations on chromosome 6 was confirmed.



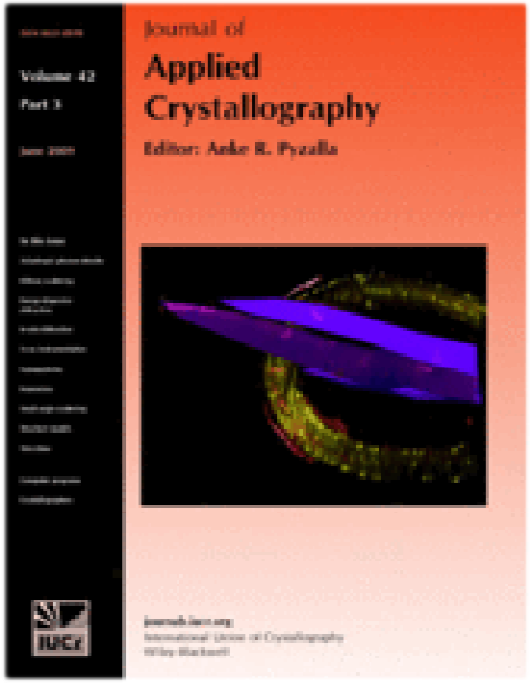
***Nature Genetics* 41, 283 - 285 (2009) ,**

David-Alexandre Trégouët, et. al.

RECálculo de las estructuras 3D de proteínas en PDB

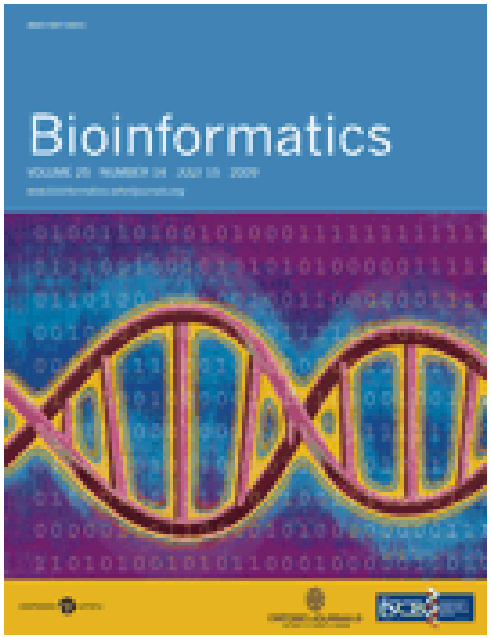
- **The PDB data base gathers publicly available 3D protein structures**
  - Full of bugs.
- **Goal: redo the structures by recalculating the diffraction patterns**
  - PDB-files: 42.752.
  - X-ray structures: 36.124.
  - Successfully recalculated: ~36.000.
  - Improved R-free: 12.500/17000.
  - CPU time estimate 21.7 CPU years.
  - Real time estimate: 1 month on Embrace VO on EGEE.

R.P Joosten et al, Journal of Applied Cristallography, (2009) 42, 1-9



Quality of the phylogenetic annotation of bacteries

- Comparative phylogenetic experiment on a soil sample with respect to different releases of the NR Gene Bank Database.
- Many of the associations of sample fragments to biological families have changed, even recently.
- The changing rate does not decreases as time goes by, being increased in many cases.
- This reveals that the complete diversity of such communities is not sufficiently well described on current data bases.



CSISP Centro Superior de Investigación en Salud Pública

Institut Cavanilles de Biodiversitat i Biologia Evolutiva

BIOINFORMATICS LETTER TO THE EDITOR Vol. 24 no. 18 2008, pages 2124–2125
doi:10.1093/bioinformatics/btn355

Genome analysis

Metagenomics reveals our incomplete knowledge of global diversity

Miguel Pignatelli[1,2], Gabriel Aparicio[3], Ignacio Blanquer[3], Vicente Hernández[3], Andrés Moya[1,2] and Javier Tamames[1,2,*]

[1]Instituto Cavanilles of Biodiversity and Evolutionary Biology, University of Valencia, Apdo 22085, 46071 Valencia, [2]CIBER of Epidemiology and Public Health (CIBERESP) and [3]Grid and High-performance Computing Group, ITACA, Politechnic University of Valencia, Camino de Vera s/n, 46022 Valencia, Spain

Received on May 23, 2008; revised and accepted on July 10, 2008
Advance Access publication July 13, 2008
Associate Editor: Dmitrij Frishman

Conclusions

- **The Biomed community is much more heterogeneous and diverse than other communities.**
- **Biomed Applications require services**
  - Rely on services and components provided by the infrastructure as much as possible.
  - Feedback requirements to middleware developments.
- **Develop application domain specific components with a general view**
  - Rely on standards and components.
  - It was sometimes painful in the past when releases change heavily the components.
- **Developing general components implies a hard dedication and a long-term support**
- **EGEE is a daily tool in bioinformatics, and innovative medicine and a consolidating tool in medical imaging**
  - VPH is used to HPC facilities with great interest on Grids.