Plateforme de Calcul pour les Sciences du Vivant

Grids for life sciences: status and perspectives

V. Breton

CNRS-IN2P3

Teratec 2009

Credit: A. Da Costa, P. De Vlieger, J. Salzemann

http://clrpcsv.in2p3.fr

CENTRE NATIONAL DE LA RECHERCHE SCIENTIFIQUE

IN2P3

INSTITUT NATIONAL DE PHYSIQUE NUCLÉAIRE ET DE PHYSIQUE DES PARTICULES

Introduction

Plateforme de Calcul pour les Sciences du Vivant

- **Grid technology provides services to do science differently, opens new avenues for**
  - Large scale on demand computing
  - Secure data sharing
  - dynamic data analysis

- **Goals of my talk**
  - Share some of our ideas for using grid services in life sciences and healthcare
  - Share my enthusiasm for what is ahead of us

All grid applications described in this talk use gLite as grid middleware

Table of content

Plateforme de Calcul pour les Sciences du Vivant

Grid services have made huge progresses

- **Distributed computing has been available for 5 years for scientific production**
  - ☺ : access to very large number of CPUs (>20.000 for biomed Virtual Organization)
  - ☺ : web service APIs lead to improved interoperability (EGEE, OSG, Digital Ribbon, …)
  - ☹: job efficiency and resource stability are still a problem
  - ☹: MPI is still available on a limited number of clusters (<10% of CPUs on EGEE biomed VO)
- **Distributed data management has recently become available (AMGA)**
  - ☺: secured access
  - ☺: easy installation
  - ☺: good performances
  - ☹: critical mass of developers for software maintenance and evolution

What can I do with these services I could not do before ?

Think bigger !

Plateforme de Calcul pour les Sciences du Vivant

- **Possibility to scale up by one or two orders of magnitude the volume of computations**
  - On demand access to > 20.000 CPU cores instead of cluster
  - Freedom to think big
- **Use cases**
  - Protein structure computations (e-NMR)
  - From docking 1000 drug-like molecules to testing all the compounds currently available on market
  - From updating monthly to updating daily a molecular biology database
  - From studying the impact of single DNA mutations (SNPs) to multiple correlated mutations (Haplotypes) on diseases

Genome Wide Haplotype analyses of human complex diseases with the EGEE grid

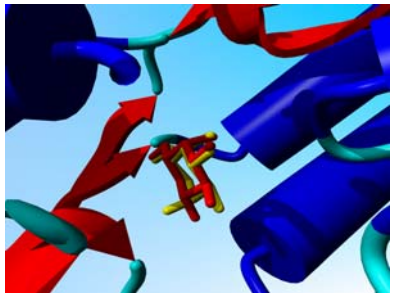
Plateforme de Calcul pour les Sciences du Vivant

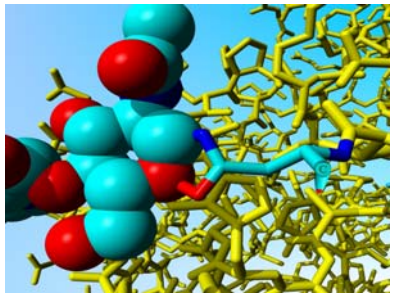
- **Goal: study the impact of DNA mutations on human coronary diseases**
- **Very CPU demanding analysis to study the impact of correlated (double, triple) DNA mutations**
- **Deployment on EGEE Grid**
  - 1926 CAD (Coronary Artery Diseases) patients & 2938 healthy controls
  - 378,000 SNPs (Single Nucleon 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

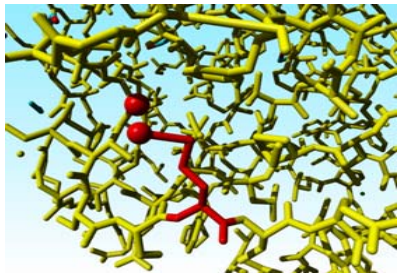
Application: recalculating protein 3D structures in PDB

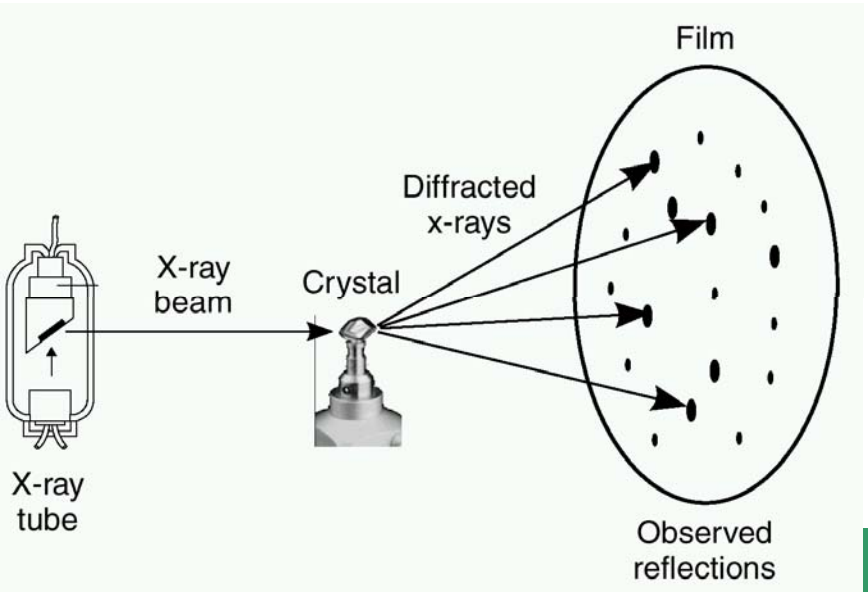
Plateforme de Calcul pour les Sciences du

- **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

Share my data while keeping them !

Plateforme de Calcul pour les Sciences du Vivant

- **Share securely data without having to put them in a central repository**
  - Data are left where they are produced
  - Authorized users have a customized view of a subset of the data
  - Data owners keep a full control of their data
- **Use cases**
  - Federation of mammography databases (MammoGrid) to improve cancer detection
  - Federation of brain medical image databases (BIRN, NeuroLog, NeuGrid) for neurosciences

Monitoring and alert

- **Coupling of grid data management and computing services allows continuous**
  - Data collection
  - Data analysis
  - Updated modeling
  - Towards decision making
- **Use cases**
  - Tsunami alert system
  - Flood alert system
  - Epidemiology

Table of content

Plateforme de Calcul pour les Sciences du Vivant

WISDOM

Plateforme de Calcul pour les Sciences du Vivant

PCSV

WISDOM (World-wide In Silico Docking On Malaria) is an initiative aiming to demonstrate the relevance and the impact of the grid approach to address drug discovery for neglected and emerging diseases.

2005	2006	2007	2008
Wisdom-I Malaria Plasmepsin	DataChallenge Avian Flu Neuraminidase	**Wisdom-II** Malaria 4 targets	DataChallenge Diabetes Alpha-amylase

GRIDS

EGEE, Auvergrid, TwGrid, EELA, EuChina, EuMedGrid

EUROPEAN PROJECTS

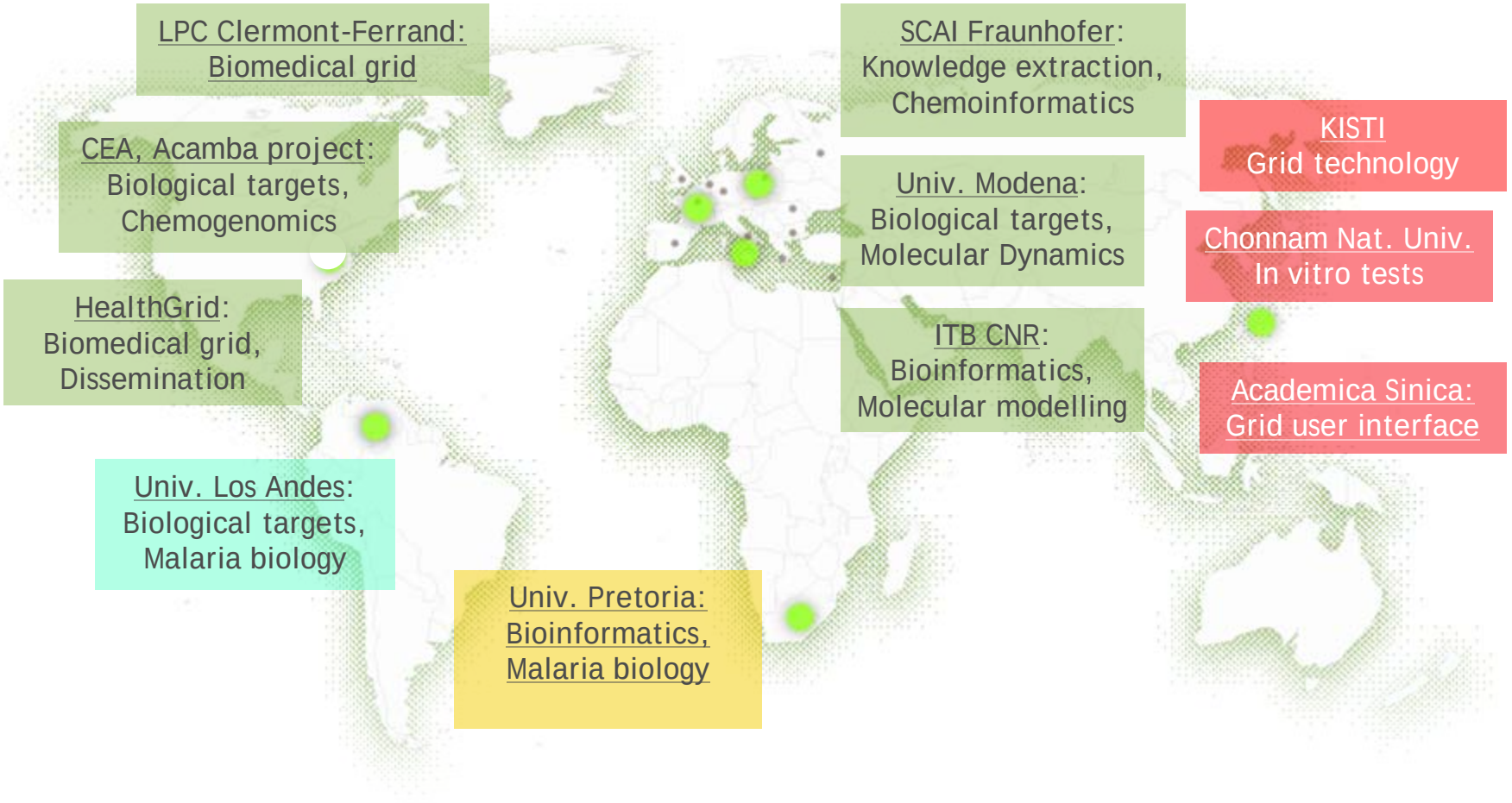
Embrace
EGEE
BioInfoGrid

INSTITUTES

SCAI, CNU
Academica Sinica of Taiwan
ITB, Unimo Univ,, LPC, CMBA
CERN-Arda, Healthgrid, KISTI

WISDOM partners

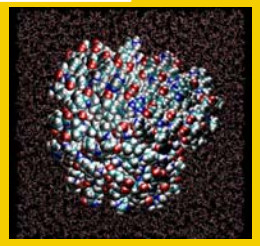
Plateforme de Calcul pour les Sciences du Vivant



Virtual screening pipeline

Plateforme de Calcul pour les Sciences du Vivant

FLEXX/
AUTODOCK



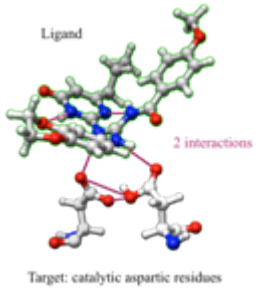
AMBER

Molecular docking

Molecular dynamics

Complex visualization

CHIMERA



in vitro

WET LABORATORY

in vivo

Modern Medicine
healthy patients. healthy practice.
Home Resource Centers CME/CE Medical Economics Careers Communi

Oseltamivir-Resistant Flu Viruses Increasing

Resistant viruses pose lethal threat to high-risk patients

Publish date: Mar 3, 2009

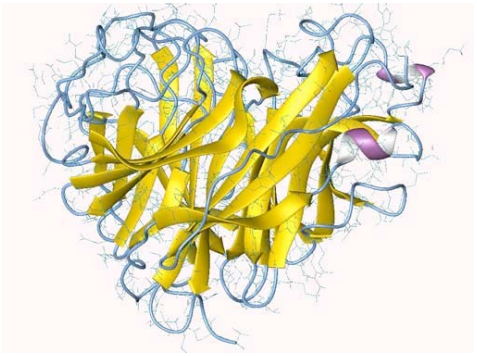
SOURCE: HealthDay

TUESDAY, Mar. 3 (HealthDay News) -- The 2008 to 2009 influenza season will see a higher prevalence of oseltamivir-resistant viruses, and certain strains of the virus are highly pathogenic to high-risk patients, according to two studies published online Mar. 2 in the *Journal of the American Medical Association.* Another study reports that intranasal live attenuated influenza vaccine is associated with more medical encounters than trivalent inactivated vaccine.

Zhong Wang, Ph.D., and colleagues at the Armed Forces Health Surveillance Center in Silver Spring, Md., conducted influenza surveillance among more than one million American military personnel during the three flu seasons from 2004 to 2007, and found that subjects in all three seasons that were vaccinated using trivalent inactivated vaccine had lower incidence of pneumonia and influenza compared to live attenuated influenza vaccination or no vaccination.

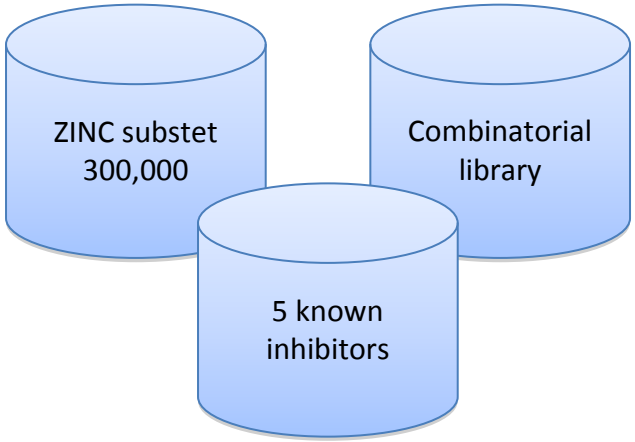
Nila J. Dharan, M.D., of the U.S. Centers for Disease Control and Prevention in Atlanta, and colleagues tested the 2007-2008 season influenza A(H1N1) viruses, and found that of 1,155 viruses tested, 142 (12.3 percent) of them were resistant to oseltamivir. Preliminary data shows that 264 of 268 samples tested from the 2008-2009 season are also resistant. In another report, Jairo Gooskens, M.D., of Leiden University Medical Center in the Netherlands, and colleagues describe the transmission of oseltamivir-resistant influenza A(H1N1) viruses with the H274Y mutation in stem cell transplant and elderly patients.

Oseltamivir is an antiviral drug used in the treatment and prophylaxis of Influenzavirus A



8 variants NA predicted by homology modelling

Autodock Docking Tool



Chemical compound database

Deployment on EGEE grid…

Number of dockings	CPU years	Real Time	CPUs used	Produced Data size	Crunching Factor	Distribution efficiency
4 millions	**100**	**1,5 months**	**1700**	**800 GB**	**900**	**50 % WISDOM >80% DIANE**

WISDOM production environment

Plateforme de Calcul pour les Sciences du Vivant

Client Services

Local Data Repository

FTP

HTTP

Data Manager

Database Service

Transfer Manager

Data Management APIs

WISDOM Information System

AMGA

SE

CE

NGS/EGI

SE

CE

OSG

Task Manager

Tasks Management APIs

Job Manager

Job Submitter

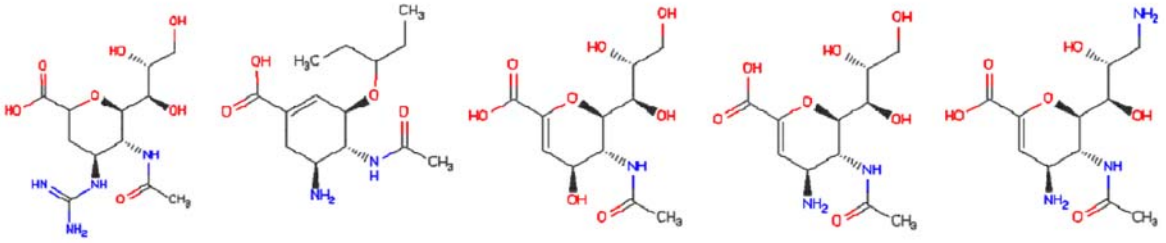
Data

Metadata

Jobs

Results obtained

Plateforme de Calcul pour les Sciences du Vivant



(GNA) (G39) (DAN) (4AM) (49A)

zanamivir oseltamivir

5 known inhibitors

	T06	T01	T02	T03	T05	T07	T13
DAN	-	-	-	-	-	-	-
4AM	-	-	+	+	-	-	-
49A	+	-	+	+	+	-	-
GNA	+	-	+	-	+	-	+
G39	+	-	-	-	+	-	+

Mutation effect on the known inhibitors docking scores.

"+" sign: compound is within top 5%

"-"sign: compound is not within the top 5%.

- Quick evaluation of mutation effect on inhibitor binding

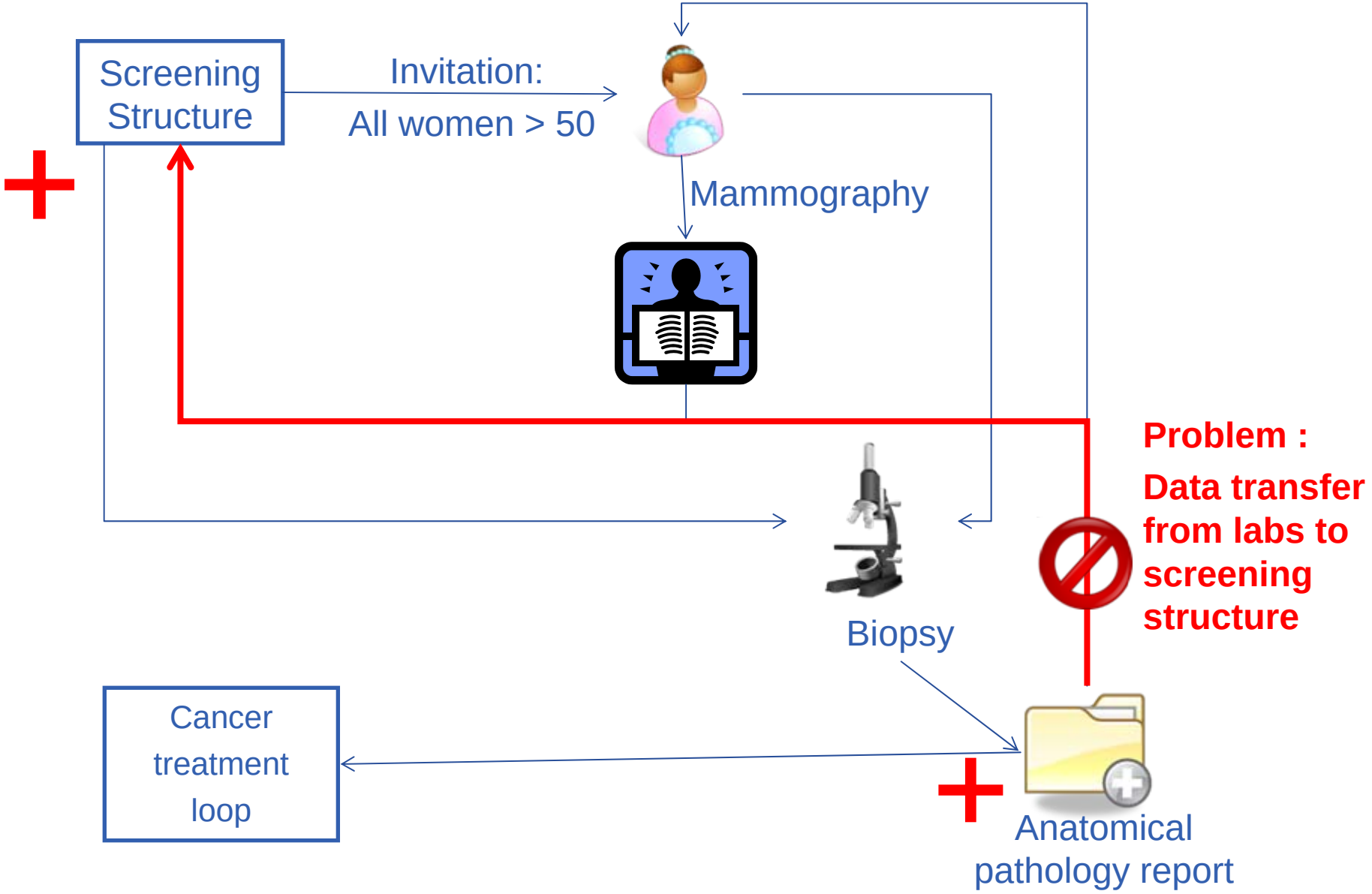
Table of content

Plateforme de Calcul pour les Sciences du Vivant

- **Introduction**
- **Grid added value for**
  - Large scale computing
  - Distributed data management
  - Dynamic data analysis
- **WISDOM, grid-enabled *in silico* drug discovery**
- **Cancer surveillance network**
- **Emerging disease surveillance network**
- **Conclusion**

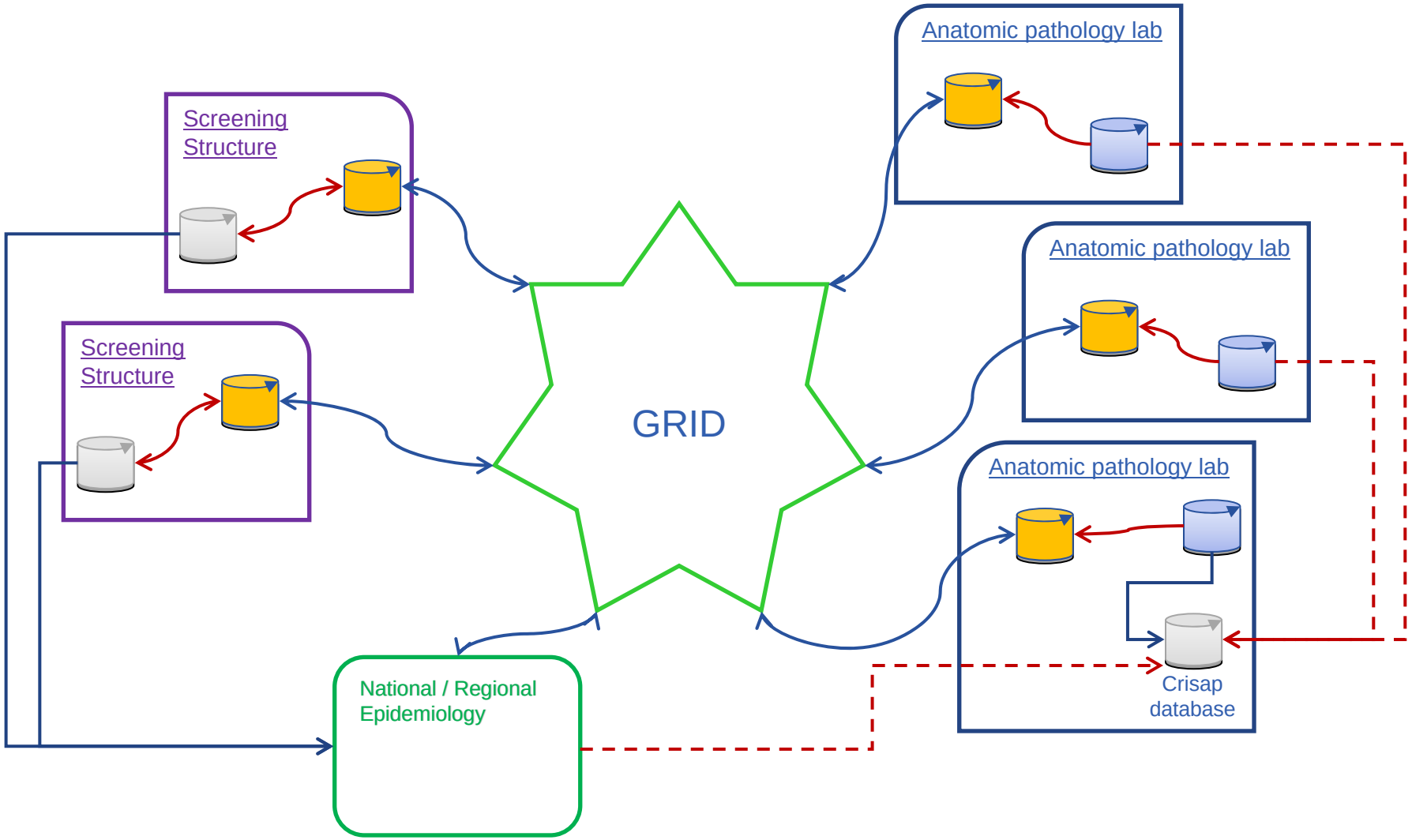
Breast cancer screening

Plateforme de Calcul pour les Sciences du Vivant



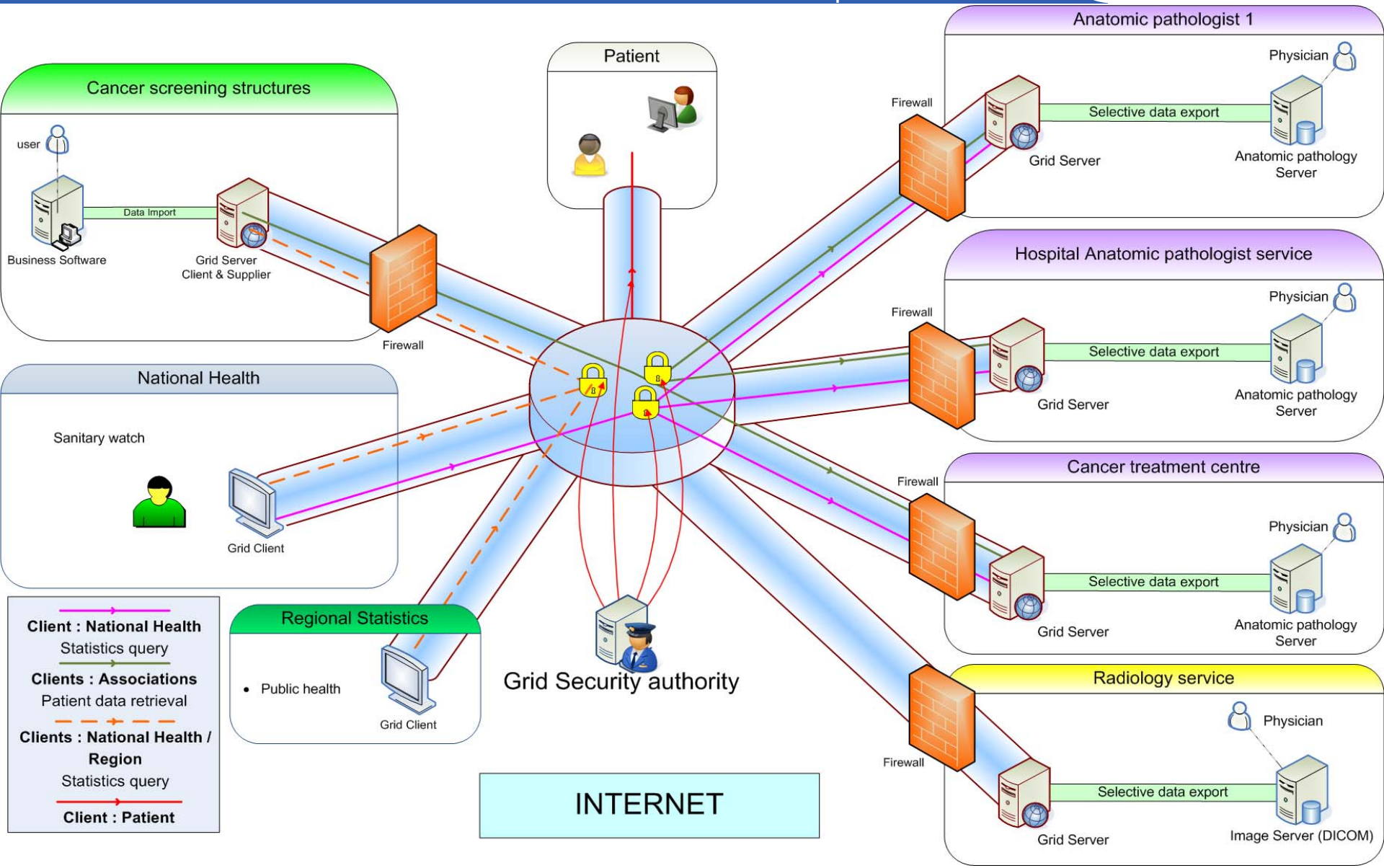
Sentinel network

Plateforme de Calcul pour les Sciences du Vivant



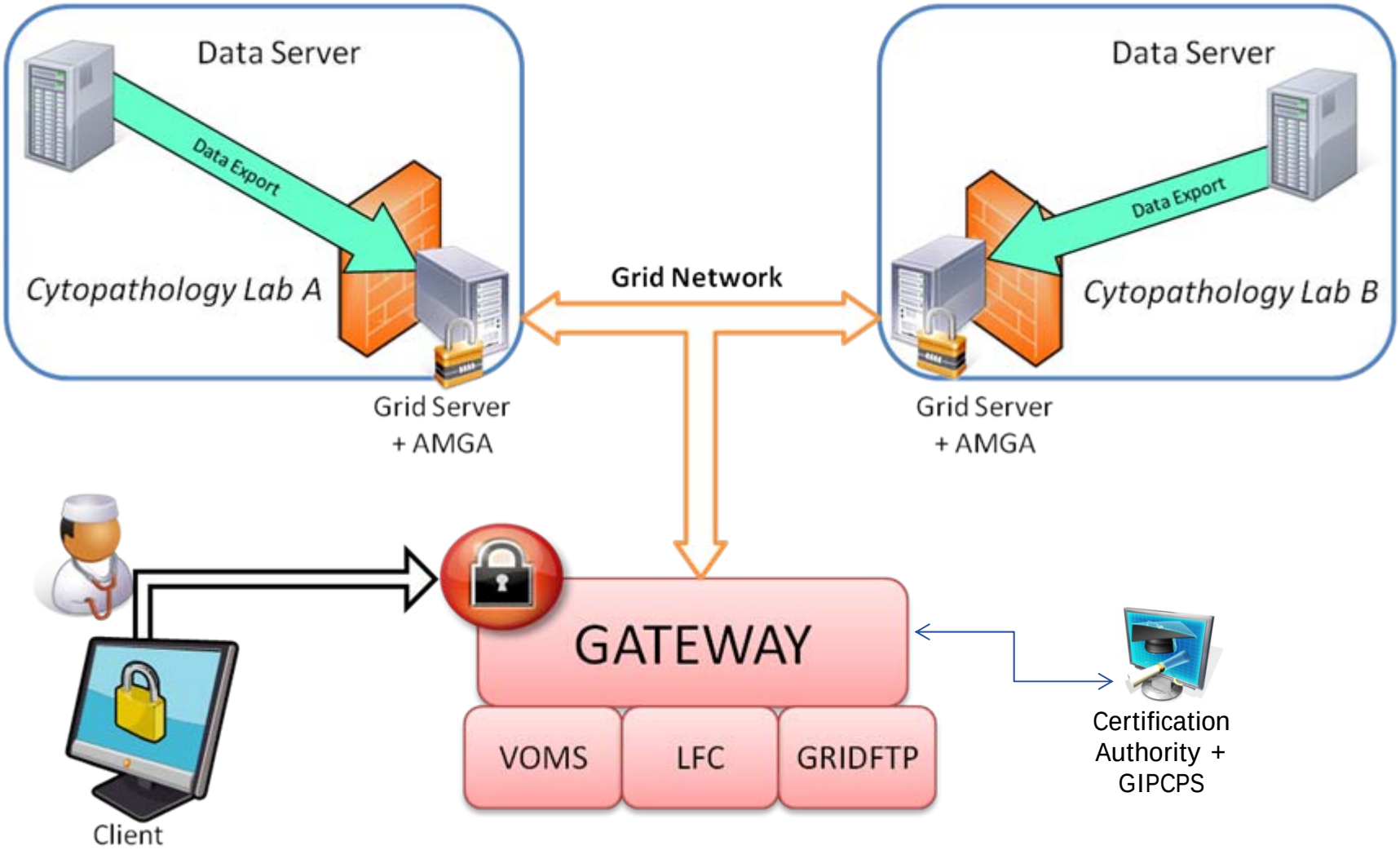
Architecture

Plateforme de Calcul pour les Sciences du Vivant



Technical architecture

Plateforme de Calcul pour les Sciences du Vivant



Grid-enabled sentinel network for cancer surveillance, Proceedings of Healthgrid conference 2009 Studies in Health Technology and Informatics

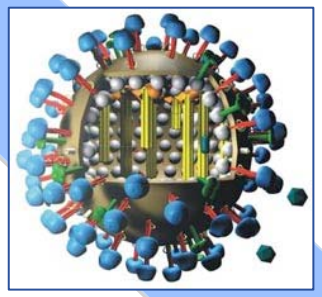
Collaboration: CNRS – MAAT - RSCA

Table of content

Plateforme de Calcul pour les Sciences du Vivant

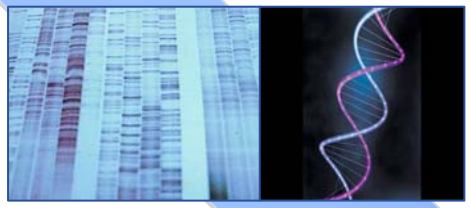
- **Introduction**
- **Grid added value for**
  - Large scale computing
  - Distributed data management
  - Dynamic data analysis
- **WISDOM, grid-enabled *in silico* drug discovery**
- **Cancer surveillance network**
- **Emerging disease surveillance network**
- **Conclusion**

New human or bird case

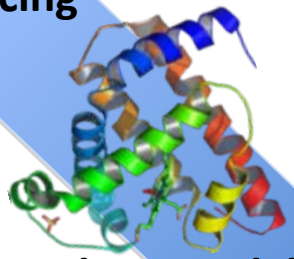


New virus strain

« The only way to track down a virus history Is through its imprint on the viral genome »



Genome sequencing



Homology modelling

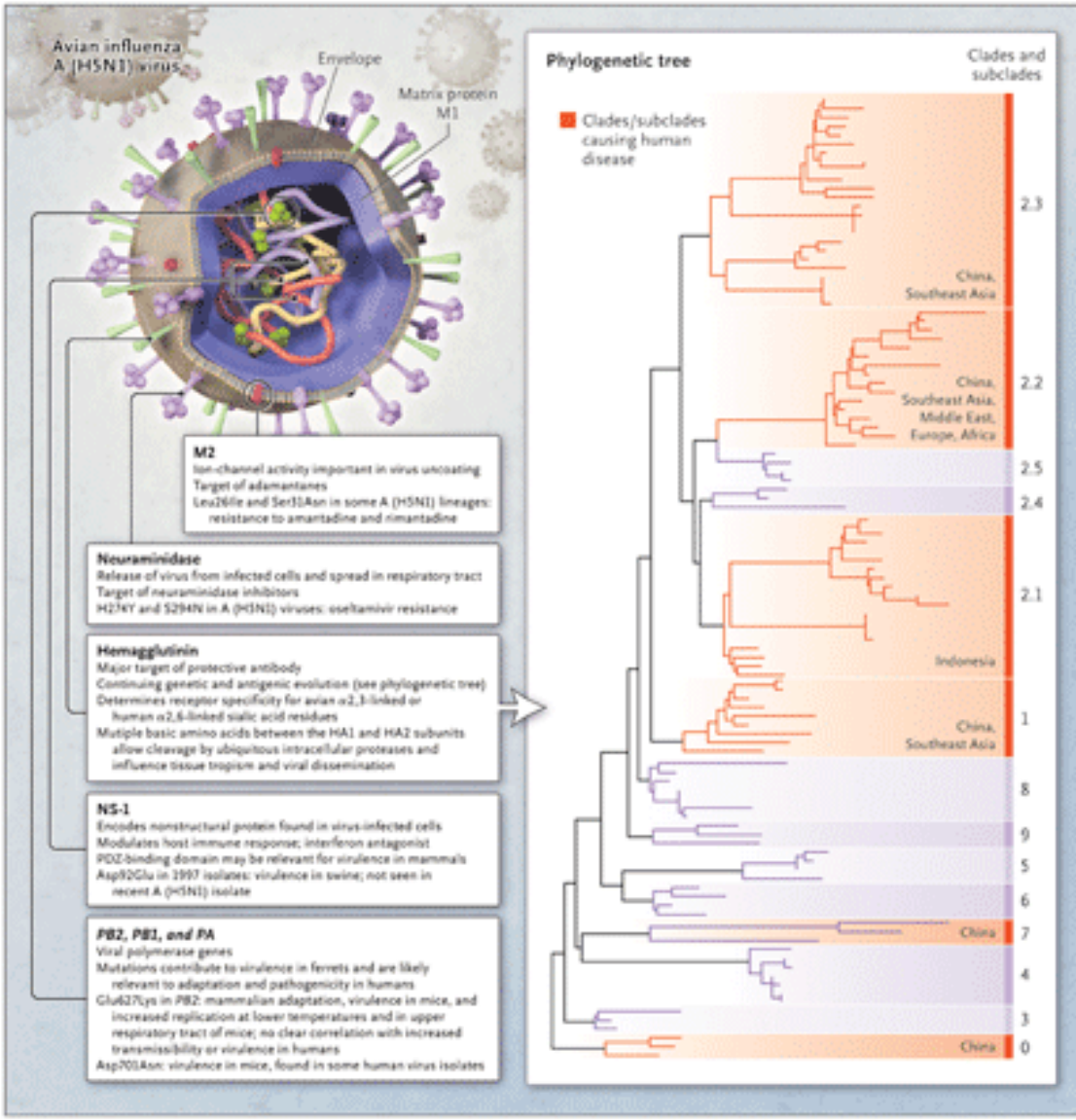
Molecular Epidemiology

Virtual Screening

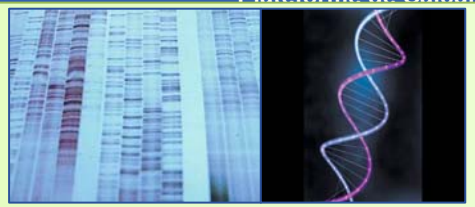
Molecular Epidemiology

Plateforme de Calcul pour les Sciences du Vivant

- Classification of virus strains
- Tracing of transmission of a strain (phylogeography)
- Analyses of outbreaks
  - *Gene rearrangement*
  - *MRCA*
  - *dN/dS*
- Analyses of pathogenesis of virus infection in humans



New virus



New genome

Virtual Screening

Influenza DB

Influenza Virus Resource *NCBI*

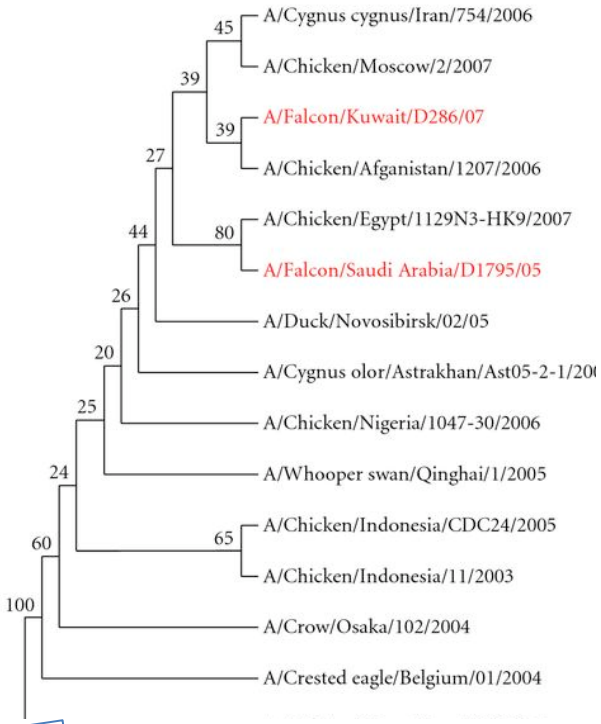
ISD *Los Alamos National Lab*

Influenza Virus Database *Beijing Institute of Genomics*

...

PHYLOGENY PACKAGES

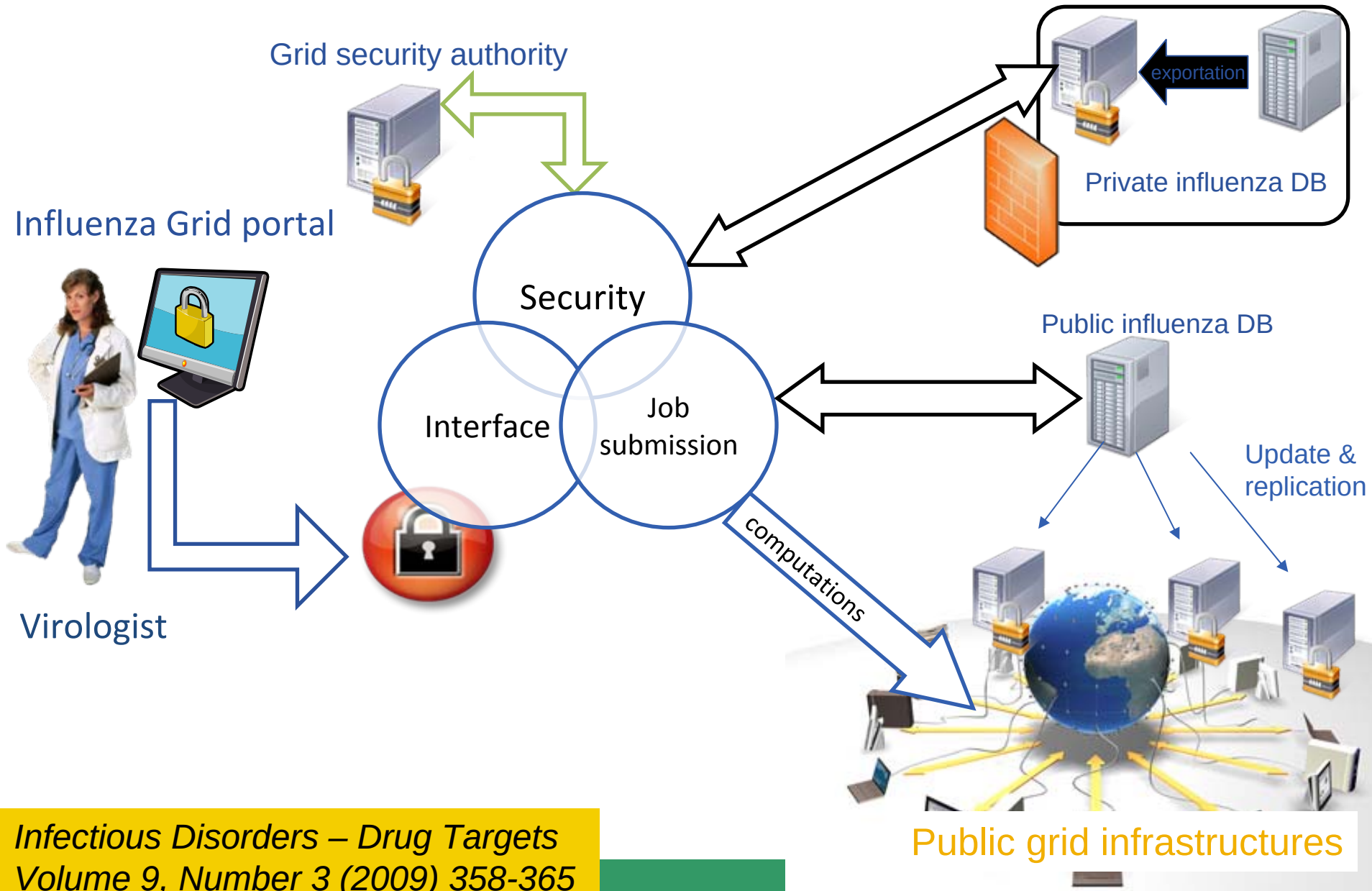
- Freely available
  - Phylip
  - BioNJ
  - PhyML
  - Tree Puzzle
  - MrBayes
- Commercial
  - PAUP
  - MEGA



New sequence + informations added to databases

Influenza grid pilot architecture

Plateforme de Calcul pour les Sciences du Vivant



Infectious Disorders – Drug Targets
Volume 9, Number 3 (2009) 358-365

WISDOM production environment

Plateforme de Calcul pour les Sciences du Vivant

NCBI

FTP

HTTP

Data Manager

Database Service

Transfer Manager

Data Management APIs

WISDOM Information System

AMGA

Client Services

Task Manager

Tasks Management APIs

SE

CE

EGEE

SE

CE

OSG

Job Manager

Job Submitter

Data

Metadata

Jobs

Grid added value for emerging diseases

- **Dynamic monitoring of the outbreak**
  - Virus evolution
  - Primer design for micro array test
  - Epidemic simulation
- ***In silico* drug discovery**
  - Grid-enabled virtual screening
  - Impact of mutations on existing drugs
  - Homology modeling
- **Epidemiology**
  - Epidemic modeling

Show case: Influenza A epidemic

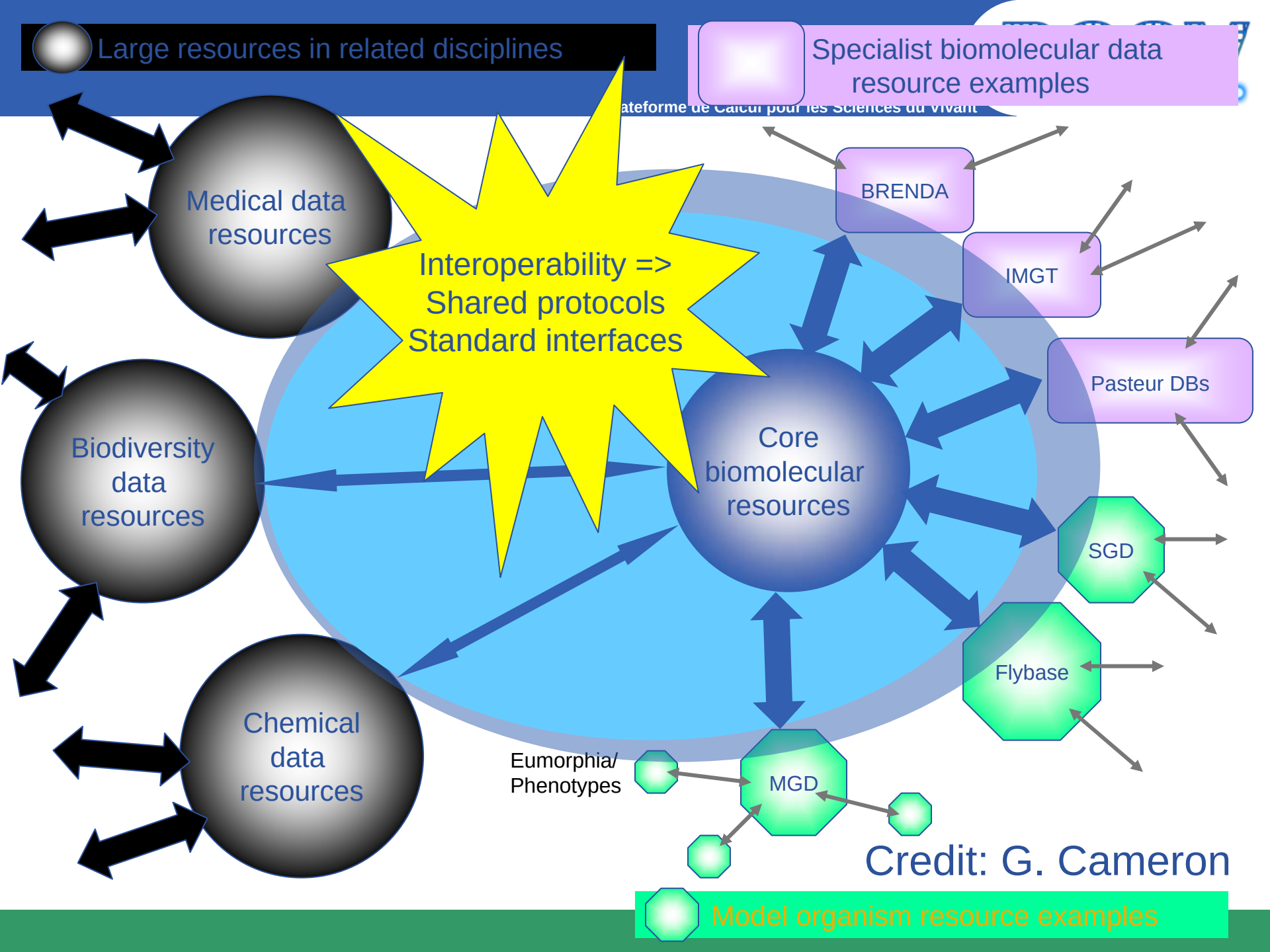
Table of content

Plateforme de Calcul pour les Sciences du Vivant

- **Introduction**
- **Grid added value for**
  - Large scale computing
  - Distributed data management
  - Dynamic data analysis
- **WISDOM, grid-enabled *in silico* drug discovery**
- **Cancer surveillance network**
- **Emerging disease surveillance network**
- **Conclusion and perspectives**

Conclusion

- **Grid services are better than they have ever been**
  - Opportunities to do science differently or at a larger scale
- **Need for more improvements**
  - Improved grid services
    - Installation, operation and maintenance of grid services is still costly
      - *Need for expertise and time*
    - Development of scientific gateways
      - *To allow easy access to grid resources*
    - Interoperability of grid infrastructures
      - *User should be middleware agnostic*
    - Stability is still an issue
  - Towards data integration and knowledge management
    - Data integration is the real challenge for life sciences


Large resources in related disciplines
Specialist biomolecular data resource examples
Medical data resources
Interoperability =>
Shared protocols
Standard interfaces
BRENDA
IMGT
Pasteur DBs
Biodiversity data resources
Core biomolecular resources
SGD
Flybase
Chemical data resources
Eumorphia/
Phenotypes
MGD
Credit: G. Cameron
Model organism resource examples

Quality of Service

Domain Specific Knowledge Management and Ontologies

User Friendliness

Knowledge Management Tools and standards

On Demand Access

Data Integration Tools and Standards

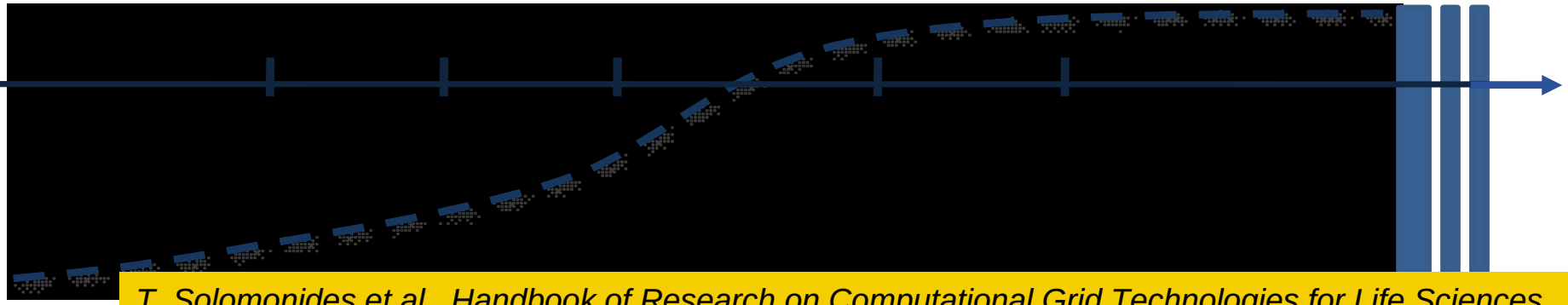
Infrastructure Interoperability

Distributed Data Models

Improved Distributed Data Management

Complexity

Time



T. Solomonides et al , Handbook of Research on Computational Grid Technologies for Life Sciences, Biomedicine, and Healthcare, IGI Global, May 2009