Applications des grilles aux sciences du vivant

V. Breton

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

Introduction

- **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

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, iRODS)**
  - ☺: 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 !

- **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

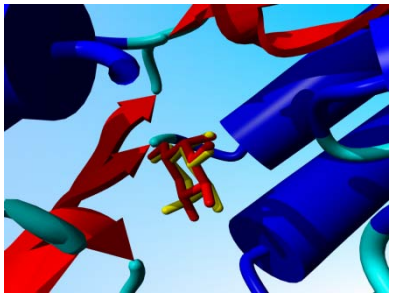
Genome Wide Haplotype analyses of human complex diseases with the EGEE grid

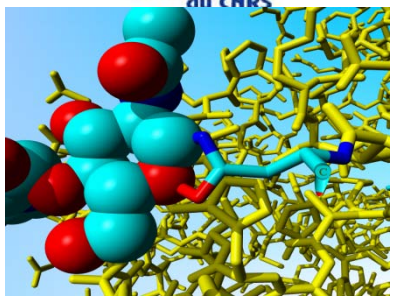
Institut des Grilles du CNRS – Equipe PCSV

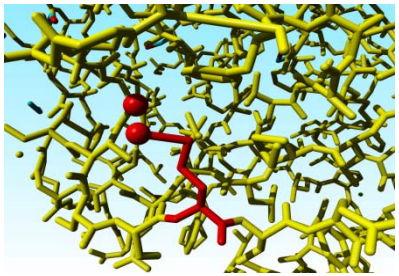
- **Goal: study the impact of DNA mutations on human coronary diseases**
- **Very CPU demanding analysis to study the impact of correlated (double, triple) DNA mutations**
- **Deployment on EGEE Grid**
  - 1926 CAD (Coronary Artery Diseases) patients & 2938 healthy controls
  - 378,000 SNPs (Single Nucleon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

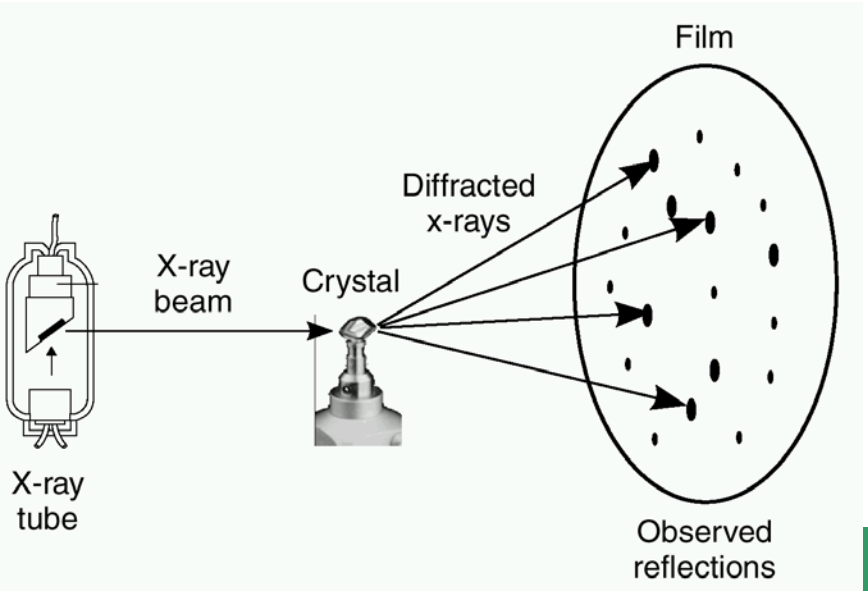
Application: recalculating protein 3D structures in PDB

- **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

WISDOM

Institut des Grilles du CNRS – Equipe 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	2009
Wisdom-I Malaria Plasmepsin	DataChallenge Avian Flu Neuraminidase	**Wisdom-II** Malaria 4 targets	DataChallenge Diabetes Alpha-amylase, maltase	

GRIDS

EGEE, Auvergrid, TwGrid, EELA, EuChina, EuMedGrid

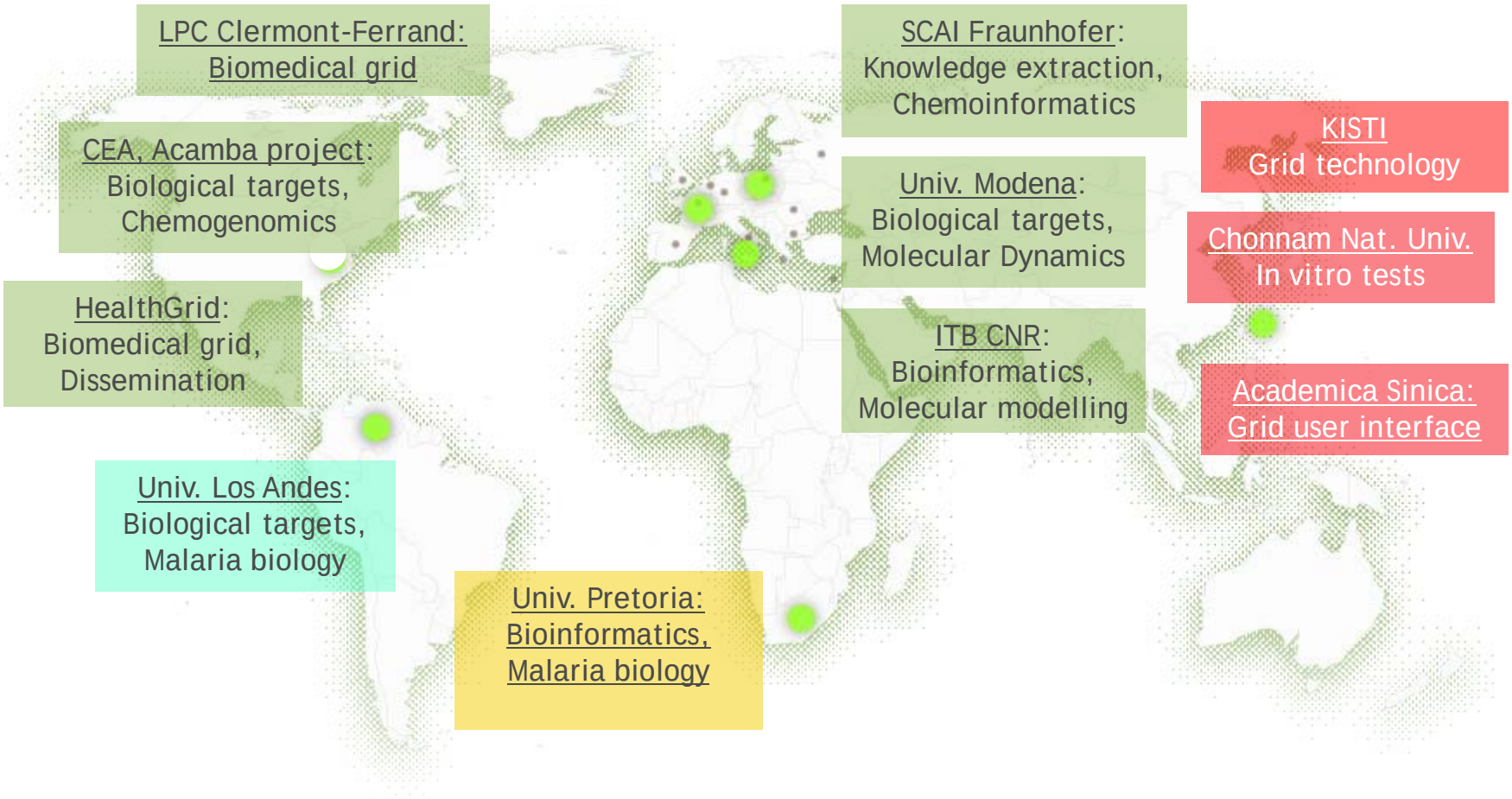
EUROPEAN PROJECTS

Embrace
EGEE
BioInfoGrid

INSTITUTES

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

WISDOM partners

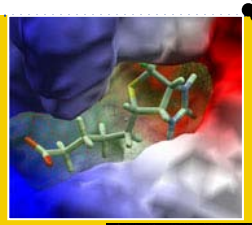


Grid-enabled *in silico* drug discovery

Institut des Grilles du CNRS – Equipe PCSV

1 Million drug-like chemical compounds

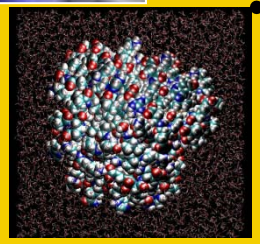
FLEXX/
AUTODOCK



Molecular docking

10.000 drug-like compounds

AMBER

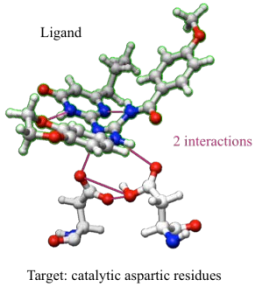


Molecular dynamics

1000 drug-like compounds

Complex visualization

CHIMERA



100 drug-like compounds

in vitro tests

WET LABORATORY

in vivo tests

20% success rate for in vitro tests

Cost for *in silico* experiment:
100 CPU Years
Cost for in vitro tests:
1-10$ per compound

WISDOM production environment

Institut des Grilles du CNRS – Equipe PCSV

Client Services

FTP

HTTP

Local Data Repository

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

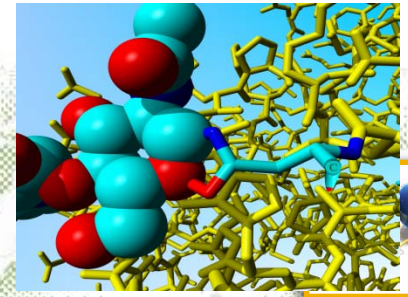
Metadata

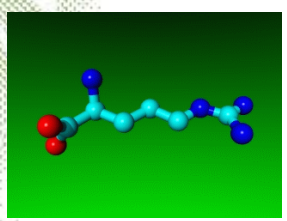
Jobs

Discovering new drugs in Vietnam

Institut des Grilles du CNRS – Equipe PCSV

PDB database
> 50.000 3D structures
including biological targets
for cancer, malaria, AIDS...

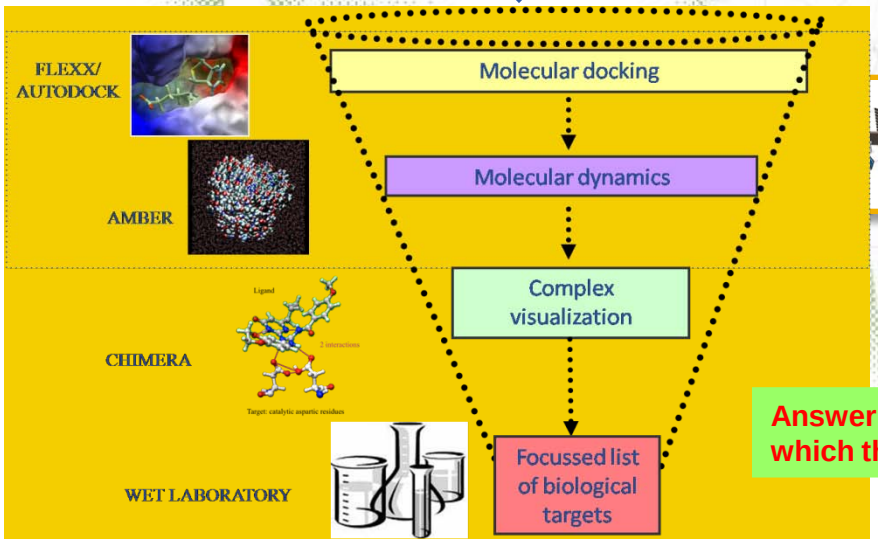




Question: are these products potentially active against cancer, malaria, AIDS ?

Hanoï
INPC

Local DataBase of Natural chemical products extracted from local biodiversity

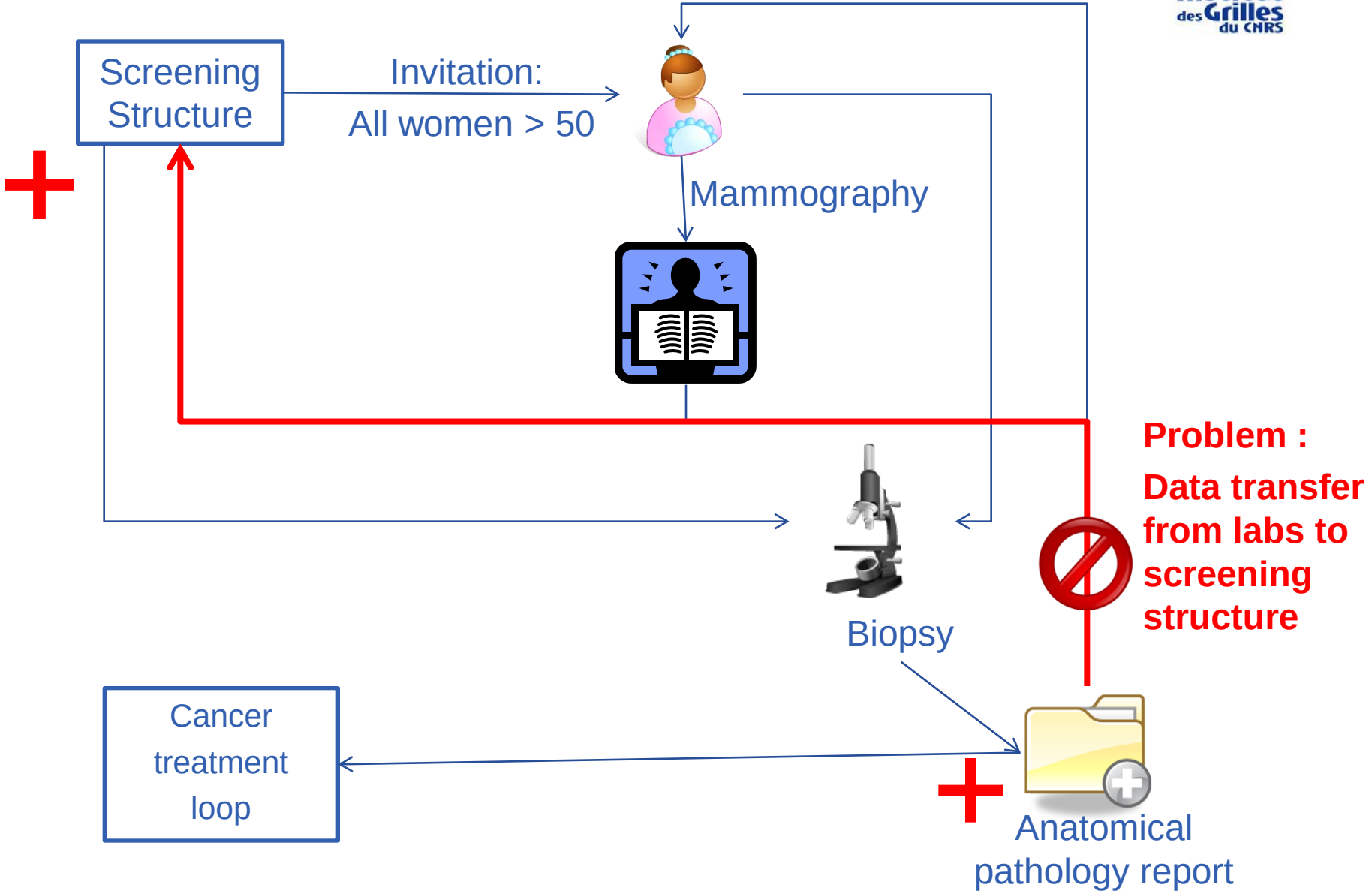


Answer: focussed list of biological targets on which the compound is most active *in silico*

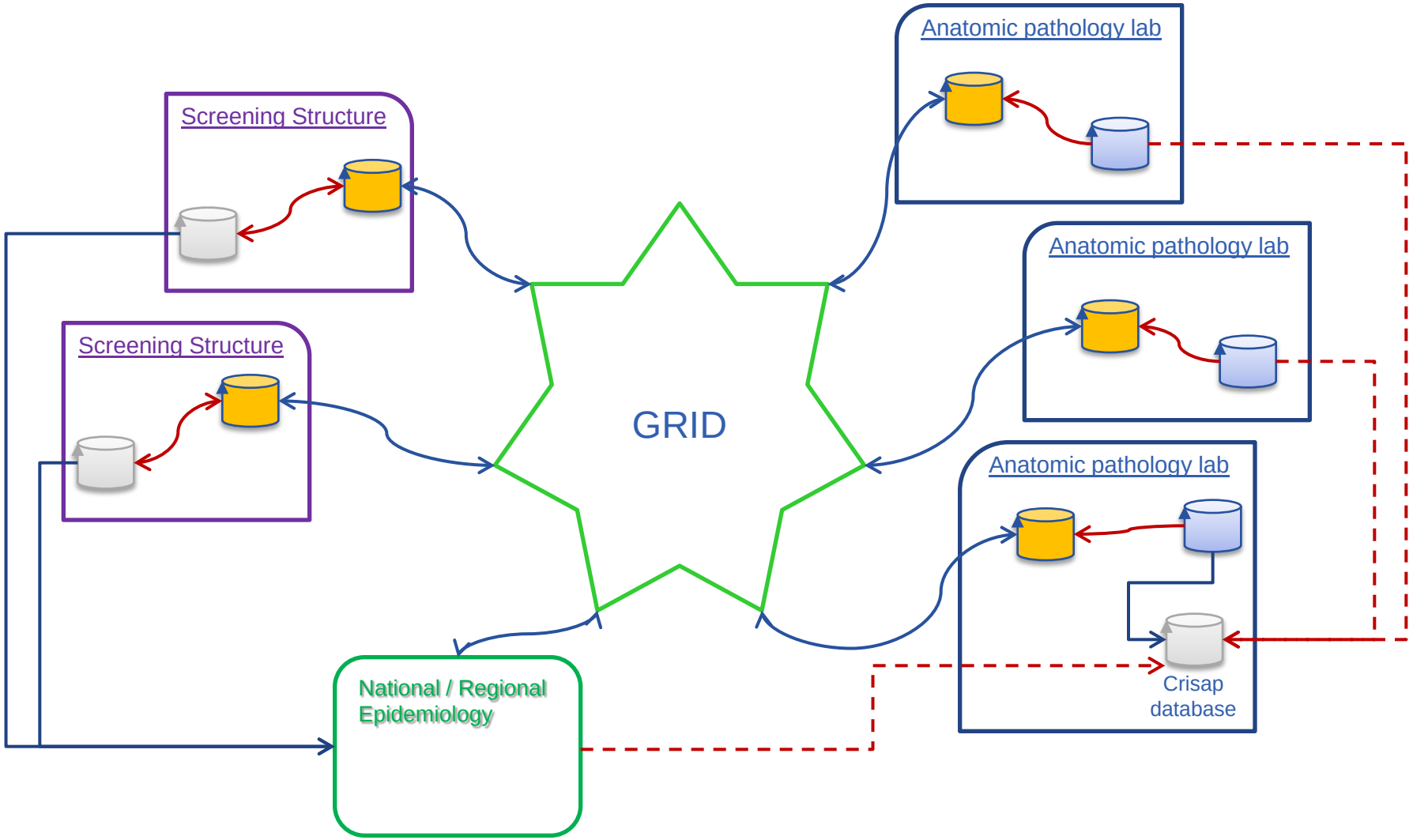
Share my data while keeping them !

- **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

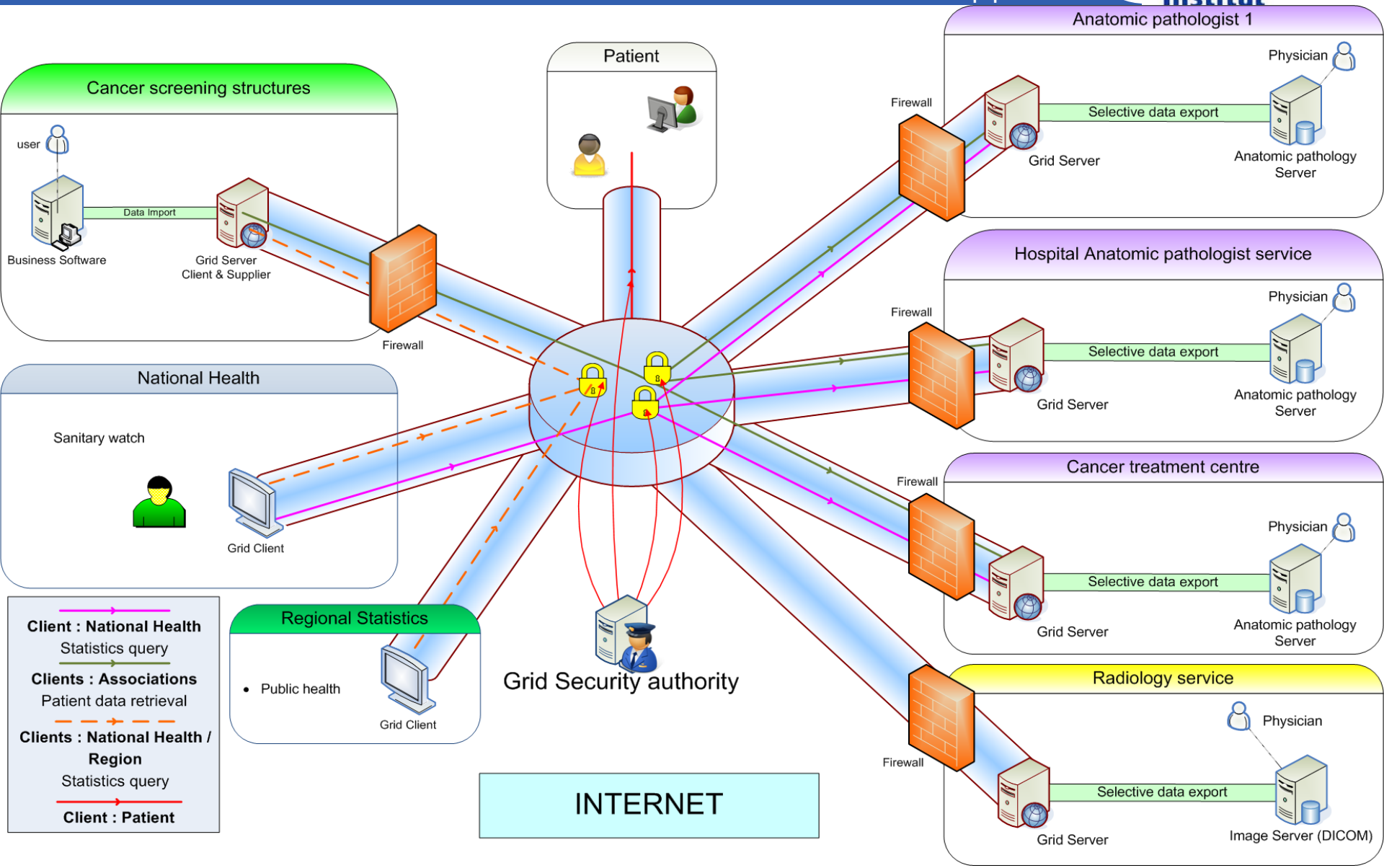
Breast cancer screening



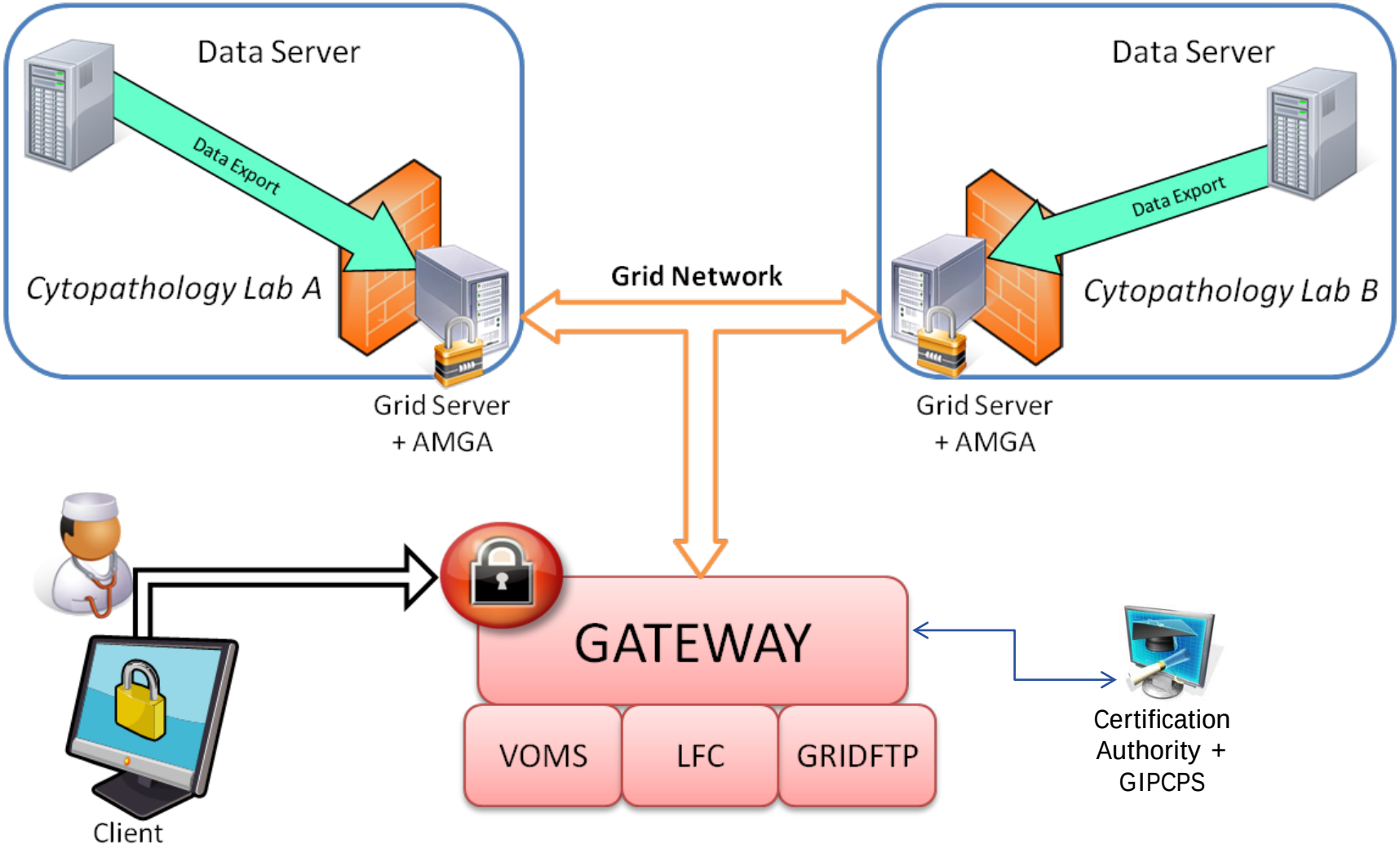
Sentinel network



Architecture



Technical architecture



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

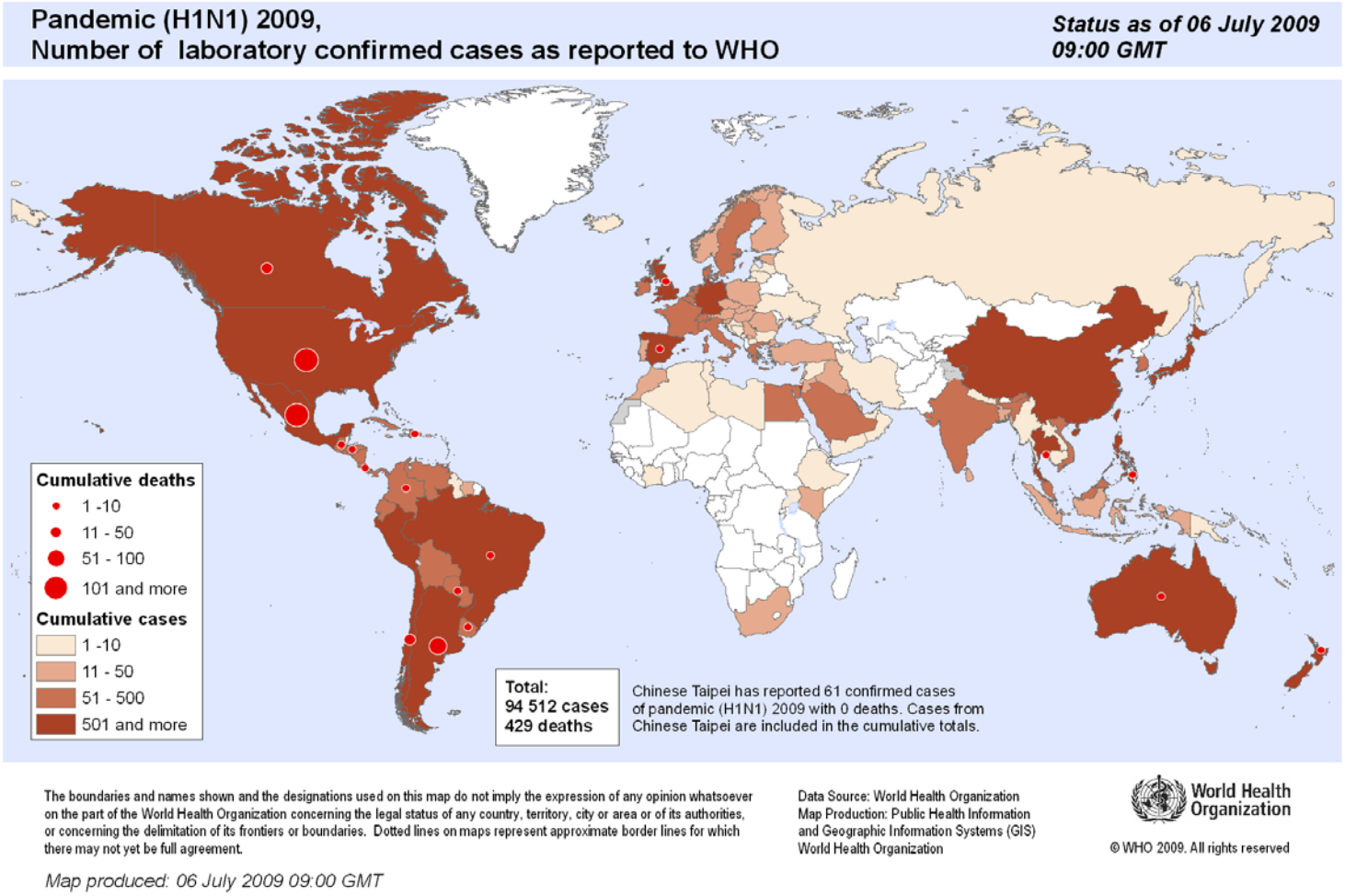
Collaboration: CNRS – MAAT - RSCA

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

Grid-enabled influenza surveillance network

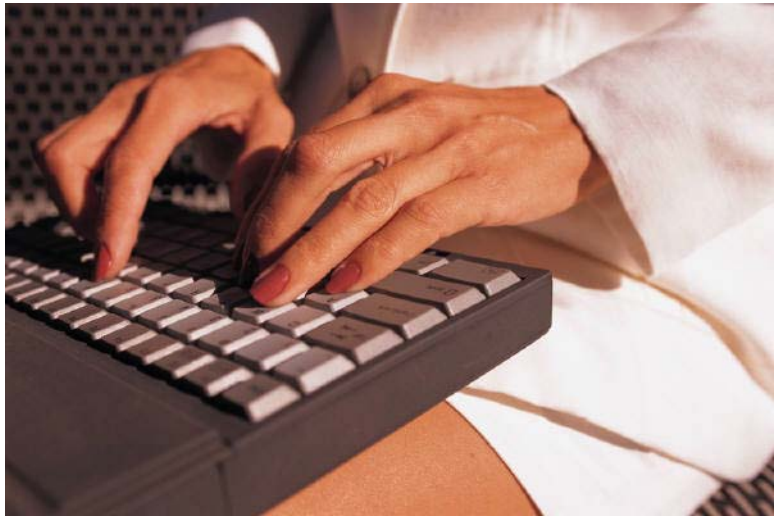
Institut des Grilles du CNRS – Equipe PCSV



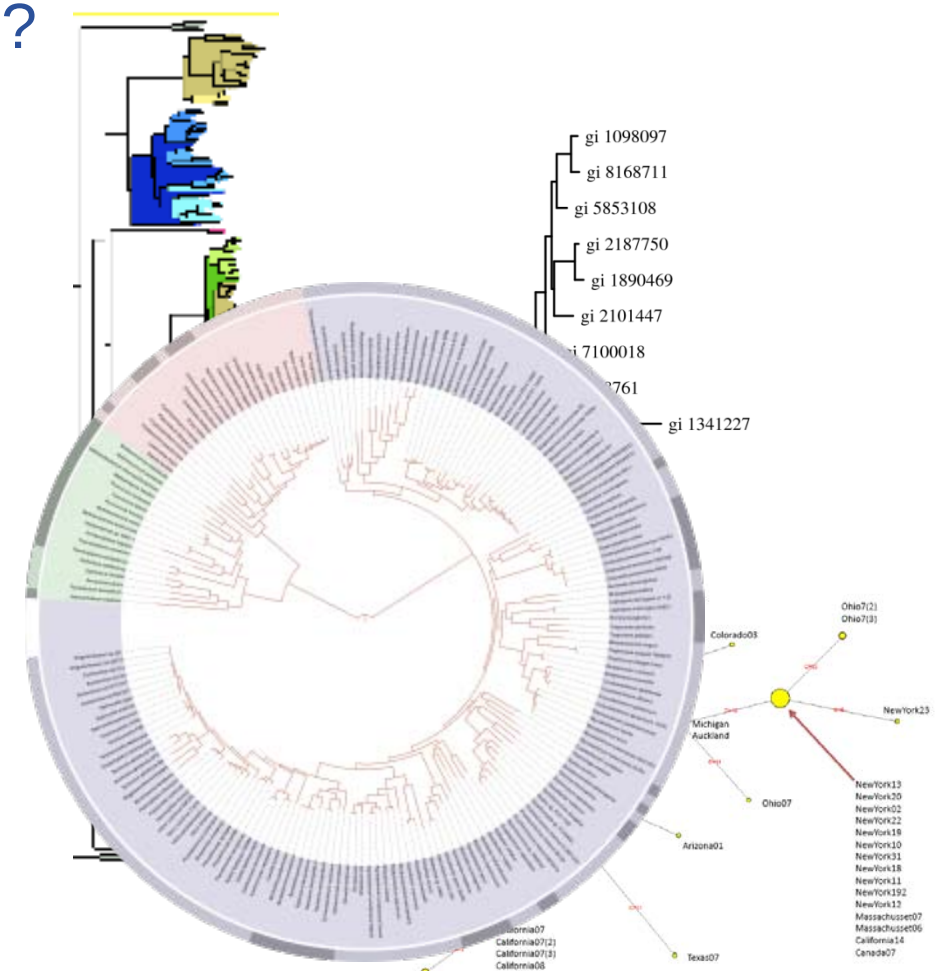
> Is there a way to improve the response to emerging diseases using grids ?

> Elements for Epidemiologists

- Where does the virus come from?
- How does the virus spread?
- How does the virus evolve?



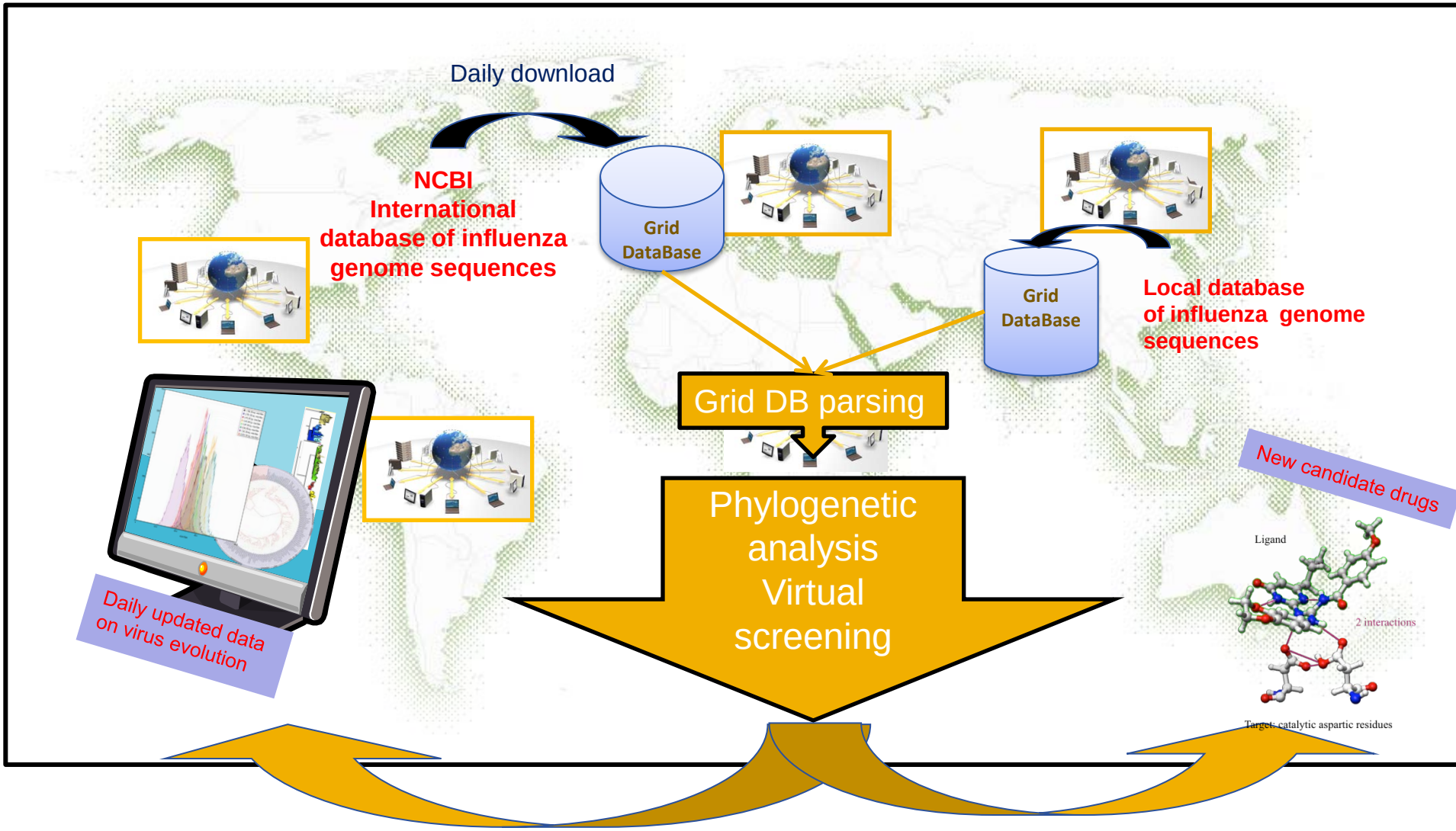
Epidemiologist



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

Monitoring the evolution of influenza viruses

Institut des Grilles du CNRS – Equipe PCSV



Molecular epidemiology pipeline to monitor influenza A virus

AMGA

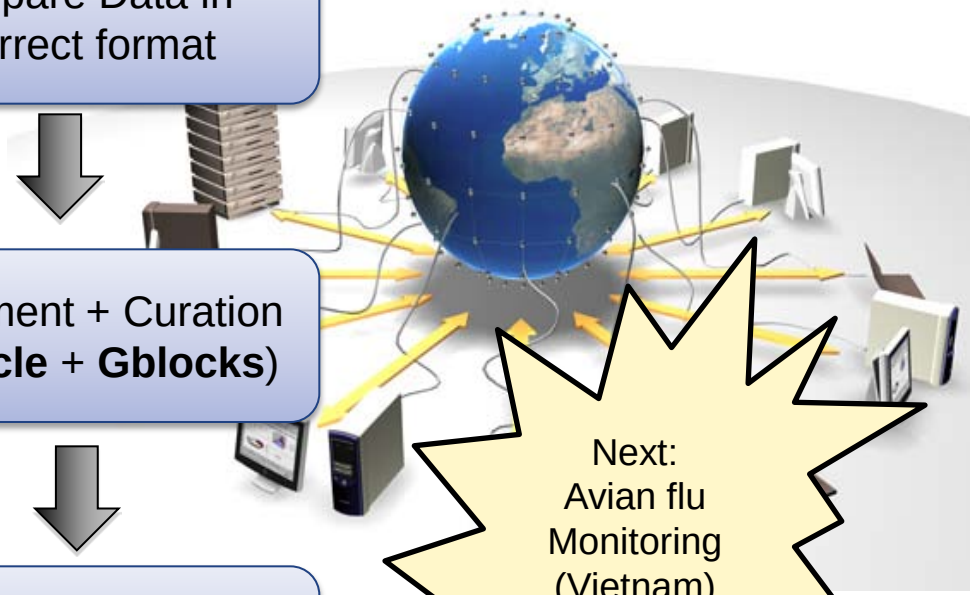
- Metadata
- Sequences
  Protein, Nucleotide, Coding region
- IDs

> Run the phylogenetic workflow on grid

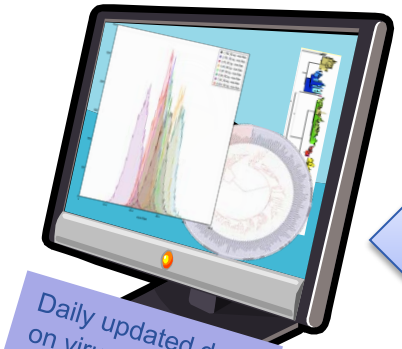
Prepare Data in correct format

Alignment + Curation (**Muscle** + **Gblocks**)

Phylogenetic Analysis (**PhyML**)



Next:
Avian flu
Monitoring
(Vietnam)



Daily updated data on virus evolution

Conclusion

- **Grid services are better than they have ever been**
  - Opportunities to do science differently or at a larger scale
- **Many opportunities for collaboration in the field of life sciences**
  - Virtual Screening
  - Surveillance networks (influenza)
- **HealthGrid conference 2010**
  - Paris, June 28-30 2010
  - Program-registration: http://paris2010.healthgrid.org