Grids for life sciences: status and perspectives

V. Breton
CNRS-IN2P3
Teratec 2009

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

• **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
• 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
Grid services have made huge progress

- **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?
• 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
• **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

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
• 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
• 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
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.
WISDOM partners

Plateforme de Calcul pour les Sciences du Vivant

LPC Clermont-Ferrand: Biomedical grid

CEA, Acamba project: Biological targets, Chemogenomics

HealthGrid: Biomedical grid, Dissemination

Univ. Los Andes: Biological targets, Malaria biology

Univ. Pretoria: Bioinformatics, Malaria biology

SCAI Fraunhofer: Knowledge extraction, Chemoinformatics

Univ. Modena: Biological targets, Molecular Dynamics

ITB CNR: Bioinformatics, Molecular modelling

KISTI Grid technology

Chonnam Nat. Univ.: In vitro tests

Academica Sinica: Grid user interface
Virtual screening pipeline

Plateforme de Calcul pour les Sciences du Vivant

FLEXX/AUTODOCK

AMBER

CHIMERA

WET LABORATORY

Molecular docking

Molecular dynamics

Complex visualization

in vitro

in vivo
Oseltamivir-Resistant Flu Viruses Increasing
Resistant viruses pose lethal threat to high-risk patients

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 influenca surveillance among more than one million American military personnel during the three flu sessions from 2004 to 2007, and found that subjects in all three sessions 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.
Deployment on EGEE grid...

<table>
<thead>
<tr>
<th>Number of dockings</th>
<th>CPU years</th>
<th>Real Time</th>
<th>CPUs used</th>
<th>Produced Data size</th>
<th>Crunching Factor</th>
<th>Distribution efficiency</th>
</tr>
</thead>
<tbody>
<tr>
<td>4 millions</td>
<td>100</td>
<td>1.5 months</td>
<td>1700</td>
<td>800 GB</td>
<td>900</td>
<td>50 % WISDOM &gt;80% DIANE</td>
</tr>
</tbody>
</table>
Results obtained

Plateforme de Calcul pour les Sciences du Vivant

5 known inhibitors

<table>
<thead>
<tr>
<th></th>
<th>T06</th>
<th>T01</th>
<th>T02</th>
<th>T03</th>
<th>T05</th>
<th>T07</th>
<th>T13</th>
</tr>
</thead>
<tbody>
<tr>
<td>DAN</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>4AM</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>49A</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>GNA</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>G39</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
</tbody>
</table>

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
• 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

Screening Structure

Invitation:
All women > 50

Mammography

Cancer treatment loop

Problem:
Data transfer from labs to screening structure

Anatomical pathology report
Sentinel network

Plateforme de Calcul pour les Sciences du Vivant

Cancer surveillance network on a grid infrastructure

GRID

Screening Structure

Anatomic pathology lab

National / Regional Epidemiology

Crisap database
Architecture

Plateforme de Calcul pour les Sciences du Vivant

Anatomic pathologist 1
Physician

Selective data export
Anatomic pathology server
Grid Server

Hospital Anatomic pathologist service
Physician

Selective data export
Anatomic pathology server
Grid Server

Cancer treatment centre
Physician

Selective data export
Anatomic pathology server
Grid Server

Radiology service
Physician

Image Server (DICOM)
Grid Server

INTERNET

Grid Security authority

Public health

Regional Statistics

Client : National Health
Statistics query

Clients : Associations
Patient data retrieval

Clients : National Health / Region
Statistics query

Client : Patient

National Health
Sanitary watch

Grid Client

Cancer screening structures

user

Business Software

Data import

Grid Server Client & Supplier

Firewall

Patient

Cancer surveillance network on a grid infrastructure
Grid-enabled sentinel network for cancer surveillance, Proceedings of Healthgrid conference 2009
Studies in Health Technology and Informatics
• 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

Molecular Epidemiology

Genome sequencing

Virtual Screening

Homology modelling

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

- 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
Influenza Virus Resource NCBI
ISD Los Alamos National Lab
Influenza Virus Database Beijing Institute of Genomics

New virus

Virtual Screening

New genome

New sequence + informations added to databases

PHYLOGENY PACKAGES
• Freely available
  Phylip
  BioNJ
  PhyML
  Tree Puzzle
  MrBayes
• Commercial
  PAUP
  MEGA

…
Influenza grid pilot architecture

Grid security authority

Influenza Grid portal

Virologist

Security

Interface

Job submission

Public influenza DB

Update & replication

Exportation

Public grid infrastructures

Virologist

Influenza Grid portal

Grid security authority

Job submission

Public influenza DB

Update & replication

Exportation

Public grid infrastructures

Infectious Disorders – Drug Targets
Volume 9, Number 3 (2009) 358-365
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**
• 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
Interoperability =>
Shared protocols
Standard interfaces

Large resources in related disciplines
Specialist biomolecular data resource examples

Medical data resources
Biodiversity data resources
Chemical data resources

Core biomolecular resources

BRENDA
IMGT
Pasteur DBs
SGD
Flybase
MGD
Eumorphia/Phenotypes

Model organism resource examples

Credit: G. Cameron
Plateforme de Calcul pour les Sciences du Vivant

Visionaries
Early Adopters
Early Majority
Late Majority

2007 2009 2011
2014 2016

Infrastructure
Interoperability
Quality of Service
User Friendliness
Knowledge Management Tools and Standards
Data Integration Tools and Standards
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