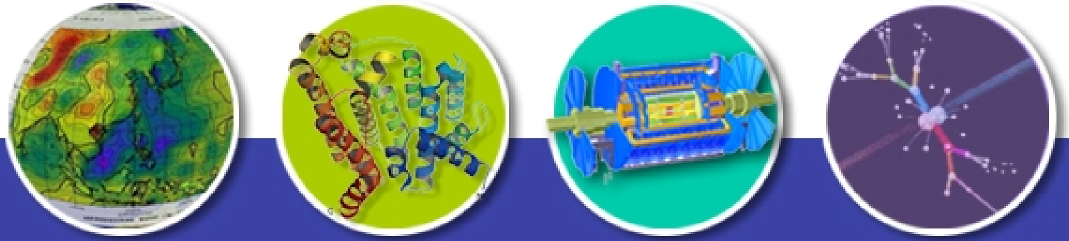
Biomedical activities using Grid Technology

Ludek Matyska, CESNET

on behalf of HealthGrid

National University of Singapore, May 4th, 2010

Credits: A. Da Costa, Y. Legré, V. Breton



Outline

1. **Grid computing to address biomedical challenges**
2. **WISDOM success story**
3. **Biomedical applications in the EUAsiaGrid project**
4. **Conclusion**

Outline

1. **Grid computing to address biomedical challenges**
2. **WISDOM success story**
3. **Biomedical applications in the EUAsiaGrid project**
4. **Conclusion**

Introduction

THE LIFE SCIENCE COMMUNITY NEEDS
BOTH E-SCIENCE AND GRID INFRASTRUCTURES

- **E-science focuses at creating new research environments for biologists**
  - Use of the most recent information technologies (semantics, ontologies)
  - Design of virtual laboratories where the biologist can run experiments and manipulate the knowledge she/he is familiar with
  - Examples: MyGrid (UK) and VLe (Netherlands)
- **Grid infrastructures provide resources needed at different levels**
  - to support bioinformaticians who maintain data bases accessed by e-science environments (update, curate, store/duplicate)
  - To increase resources for e-science environments when needed
  - To enable specific heavy computing or data production projects

Biologists vs Bioinformaticians

1. **Biologists need growing capability to handle all the data relevant to their research topics**
   - Design of complex analysis workflows
   - Knowledge management
1. **Bioinformaticians who are developing the IT services for the biologists need growing resources**
   - To store, update, curate exponentially growing databases
   - To run increasingly complex algorithms on this growing data set
   - To build new databases exploiting the growing body of knowledge
1. **Biologists and bioinformaticians have therefore different needs**
   - Biologists need high level environments and little resources
   - Bioinformaticians need large resources to develop and/or update the services needed by the biologist

Added Value of Grid for Biologists

- **The grid provides the centuries of CPU cycles required on demand**

- The grid provides the reliable and secure data management services to store and replicate the biochemical inputs and outputs

- The grid offers a collaborative environment for the sharing of data in the research community on emerging and neglected diseases

Biomedical challenges

- **Grid computing can help to solve biomedical challenges:**
  - **Data Grid**: storage of vast amounts of complex biological data and federation of distributed data sources
  - **Computational Grid**: mobilize large CPU resources to address growing processing needs to analyse data: from algorithmic and computational modelling
  - **Knowledge Grid**: information management and retrieval to extract knowledge

Life Sciences requirements

- **Identify needs**
  - Access to grids of clusters and to supercomputers
  - Stability and Sustainability
  - Friendly user interfaces
  - Standard access to services (One single API, whatever the middleware)
  - Security of medical data
- **Importance of international standards**
  - Integration of resources into European infrastructures and European initiatives
- ESFRIs
- Virtual Physiological Human
- **Which standards ?**
  - Open Grid Forum
  - Web services

Outline

1. Grid computing to address biomedical challenges
2. WISDOM successful story
3. Biomedical applications in the EUAsiaGrid project
4. Conclusion

In silico Drug Discovery

- **Pharmaceutical development:**
  - Time-consuming: more than 10 years to develop a new medicine
  - Expensive: hundreds millions of dollars
  - Emergent and neglected diseases need fast and cheap answer

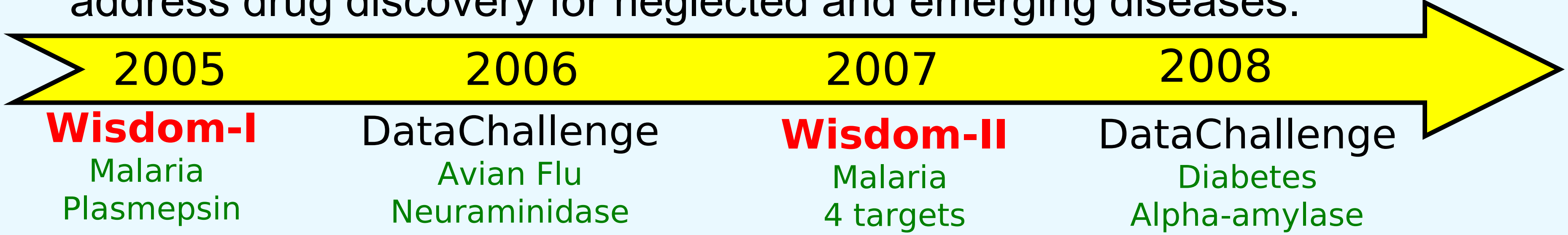
- **Computational tools:**
  - More and more known and registered protein 3D structures
  - More and more libraries of known chemicals
  - More and more computing power available
  - Better quality of prediction for bioinformatics tools, but CPU-consuming

Virtual screening using grid to speed-up the process and minimize the costs

WISDOM

WISDOM (World-wide In Silico Docking On Malaria) is an initiative aims

to demonstrate the relevance and the impact of the grid approach to address drug discovery for neglected and emerging diseases.



2005
Wisdom-I
Malaria
Plasmepsin

2006
DataChallenge
Avian Flu
Neuraminidase

2007
Wisdom-II
Malaria
4 targets

2008
DataChallenge
Diabetes
Alpha-amylase

GRIDS

EGEE, Auvergrid,
TwGrid, EELA,
EuChina,
EuMedGrid

EUROPEAN PROJECTS

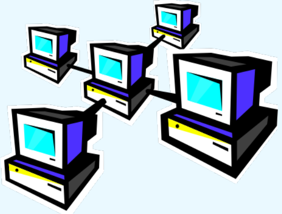
Embrace
EGEE
BioInfoGrid

INSTITUTES

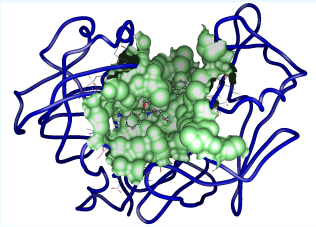
SCAI, **CNU**
Academica Sinica of Taiwan
ITB, Unimo Univ,, **LPC**, CMBA
CERN-Arda, Healthgrid, **KISTI**

Main objectives

- **Computational activities**
  - To show the relevance of computational grids in biomedical applications
  - To develop an environment to monitor the deployments on grid: Wisdom Production environment
  - To provide the grid to non-experts users



- **Biological activities**
  - To establish virtual screening workflow on computational grids
  - To find new inhibitors against neglected diseases



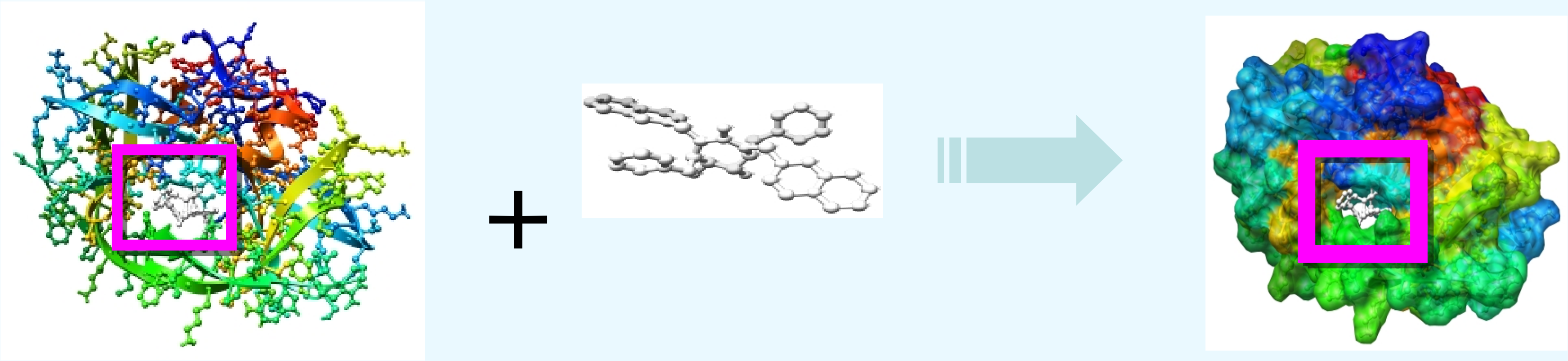
Elements

1. **TARGET: 3D structure for a key protein in a disease**

2. **LIGAND: database of chemical compounds commercially available**

3. **SOFTWARES for virtual screening: docking, molecular dynamics**

- Parallel computations
- Licenses if needed (BioSolveIT and CCDC provided free licenses for specific projects)



Targets and Ligands

Project	Protein	Function	Ligands
Malaria	Plasmepsin PMII	Hemoglobin degradation	ZINC subset 1 million
	Glutathione-S-Transférase GST	Detoxification	ZINC 4, 3 millions
	Dihydrofolate Reductase DHFR	DNA synthesis	ZINC 4, 3 millions
Avian flu	Neuraminidase	Release of new virus	ZINC subset 300, 000 millions + chemical combinatorial library
Diabetes	Amylase/Gluco-amylase	Carbohydrate cleavage	ZINC subset 300, 000 millions
sc-PDB	7000 PDB	all	4000 PDB

Workflow deployed

1. Screening

- High Throughput Docking
  - Autodock/Flexx/Gold
- Scoring function:
  - Binding energy estimation
  - Rank molecules
  - Rank interactions

2. Refinement

- Best conformations from docking
- BEAR (Binding energy after Refinement)
  - Amber package
  - MM-MD-MMPB(G)SA
- Solvation

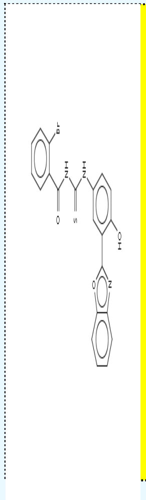
3. Checking

- Best conformations from BEAR
- Complex visualization
  - Chimera UCSF
- Hydrogen bonding
- Hydrophobic interactions

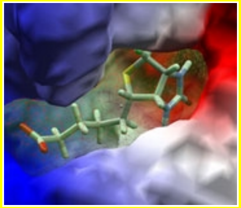
4. Validation

- Few compounds
- Protein synthesis
- Activity assay
  - IC50/Ki

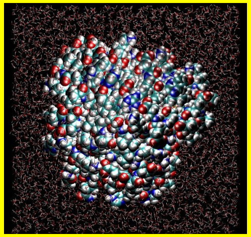
Workflow deployed



FLEXX/
AUTODOCK

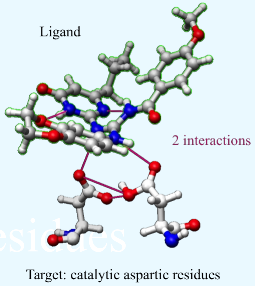


AMBER

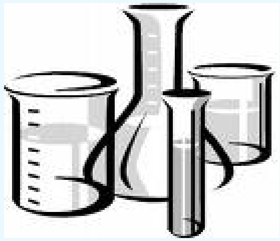


CHIMERA

Catalytic aspartic residues



WET LABORATORY



Molecular docking

Molecular dynamics

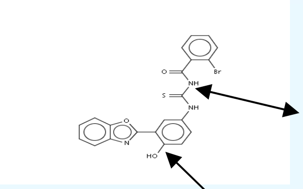
Complex
visualization

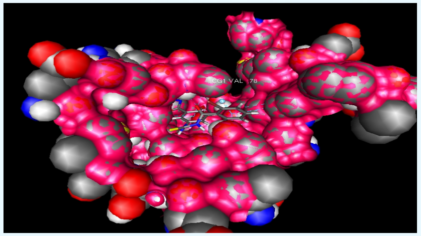
in vitro

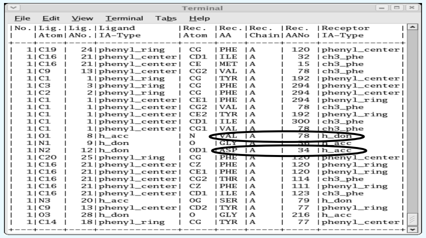
in vivo

2 steps screening on grid

1. **Docking** results based on:
   - Scoring
   - Match information
   - Different parameters settings
   - Knowledge of binding site







```
|No.|Lig.|Lig.|Ligand          |Rec.  |Rec.|Rec.  |Rec.  |Receptor        |
|   |Atom|ANo.|IA-Type         |Atom  |AA  |Chain |AANo  |IA-Type         |
+---+----+----+----------------+------+----+------+------+----------------+
|  1|C19 |  24|phenyl_ring     |  CG  |PHE |A     |  120 |phenyl_center|
|  1|C16 |  21|phenyl_center|  CD1 |ILE |A     |   32 |ch3_phe         |
|  1|C16 |  21|phenyl_center|  CE  |MET |A     |   15 |ch3_phe         |
|  1|C9  |  13|phenyl_center|  CG2 |VAL |A     |   78 |ch3_phe         |
|  1|C1  |   1|phenyl_ring     |  CG  |TYR |A     |  192 |phenyl_center|
|  1|C3  |   3|phenyl_ring     |  CG  |PHE |A     |  294 |phenyl_center|
|  1|C2  |   2|phenyl_ring     |  CG  |PHE |A     |  294 |phenyl_center|
|  1|C1  |   1|phenyl_center|  CE1 |PHE |A     |  294 |phenyl_ring     |
|  1|C1  |   1|phenyl_center|  CG2 |VAL |A     |   78 |ch3_phe         |
|  1|C1  |   1|phenyl_center|  CE2 |TYR |A     |  192 |phenyl_ring     |
|  1|C1  |   1|phenyl_center|  CD1 |ILE |A     |  300 |ch3_phe         |
|  1|C1  |   1|phenyl_center|  CG1 |VAL |A     |   78 |ch3_phe         |
|  1|O1  |   8|h_acc           |  N   |VAL |A     |   78 |h_don           |
|  1|N1  |   9|h_don           |  O   |GLY |A     |   36 |h_acc           |
|  1|N2  |  12|h_don           |  OD1 |ASP |A     |   34 |h_acc           |
|  1|C20 |  25|phenyl_ring     |  CG  |PHE |A     |  120 |phenyl_center|
|  1|C16 |  21|phenyl_center|  CZ  |PHE |A     |  120 |phenyl_ring     |
|  1|C16 |  21|phenyl_center|  CE1 |PHE |A     |  120 |phenyl_ring     |
|  1|C16 |  21|phenyl_center|  CG2 |THR |A     |  114 |ch3_phe         |
|  1|C16 |  21|phenyl_center|  CZ  |PHE |A     |  111 |phenyl_ring     |
|  1|C16 |  21|phenyl_center|  CD1 |ILE |A     |  123 |ch3_phe         |
|  1|N3  |  20|h_acc           |  OG  |SER |A     |   79 |h_don           |
|  1|C9  |  13|phenyl_center|  CD2 |TYR |A     |   77 |phenyl_ring     |
|  1|O3  |  28|h_don           |  O   |GLY |A     |  216 |h_acc           |
|  1|C14 |  18|phenyl_ring     |  CG  |TYR |A     |   77 |phenyl_center|
+---+----+----+----------------+------+----+------+------+----------------+
```

1. **Molecular Dynamics:** from docked poses
   - Distance-dependent dielectric energy minimization (e = 4r)
   - Molecular dynamics on the active site as well as ligand atoms
   - Final re-minimization
   - Re-scoring by MM-PBSA is a program which is used to estimate energies and entropies from the snapshots contained within trajectory files

Grid Performance (Docking)

	Number of dockings	CPU years	Real Time	CPUs used	Produced Data size	Crunching Factor	Distribution efficiency	Model
Malaria I	41 millions	80	6 sem	1700	1TB	400	25%	PUSH
Malaria II	142 millions	400	2,5 mois	Jusqu'à 5000	1,6 TB	2000	40%	
Avian Flu	4 millions	100	1,5 mois	1700	800 GB	900	50 % (>80% DIANE)	
Diabetes	300, 000	40	2,5 jours	7000		6000	85 %	PULL

Wet Lab results

Protein	No compounds tested	Type of analysis	Reference	No Active	IC50	Ki
PM II	30	FRET	Pepstatin A IC50=4.3 nM	26 / 30	4.3 nM-1.8 µM	
Neuraminidase	185	Fluorogenic substrate	Oseltamivir	79/185 (59 > Oseltamivir)		
GST	32 26 (	Colorimétrie	S-hexyl glutathione Ki=35 µM	4 / 32		200-400µM

Patent deposited in South Korea: « Pharmaceutical composition preventing and treating malaria comprising compounds that inhibit Plasmepsin II activity and the method of treating malaria using thereof »

Outline

1. **Grid computing to address biomedical challenges**
2. **WISDOM success story**
3. **Biomedical applications in the EUAsiaGrid project**
4. **Conclusion**

EUAsiaGrid

- Appropriate scientific areas have been identified for porting and deployment of applications to the grid:
  - High Energy Physics
  - Computational Chemistry
  - Social Science Applications
  - Mitigation of Natural Disasters
  - Bioinformatics and Biomedical
  - Digital Culture and Heritage

Fight diseases [1]

Drug Discovery: large-scale deployment of virtual screening software to identify potential inhibitors

1. Avian Flu - DC2 refined (GVSS)

- 8 avian-flu mutant targets from EGEE DC2
- 20,000 highest scored ligands from EGEE DC2 results

1. Dengue fever (GVSS)

- Dengue NS3 target
- 300,000 ligands which are prepared from ZINC.

	Number of dockings	CPU time	Duration	CPU-cores used on EUAsia VO	Produced data size	Status
1	20,000	3 years	25 days	125	12,8 GB	Completed
2	300, 000	12 years	2 months	268	46 GB	Completed

Fight diseases [2]

3 Vietnamese natural products (WISDOM)

- **Goal:** Grid-enabled the study of the potential inhibitory action of chemical compounds extracted from Vietnamese natural products, on important diseases

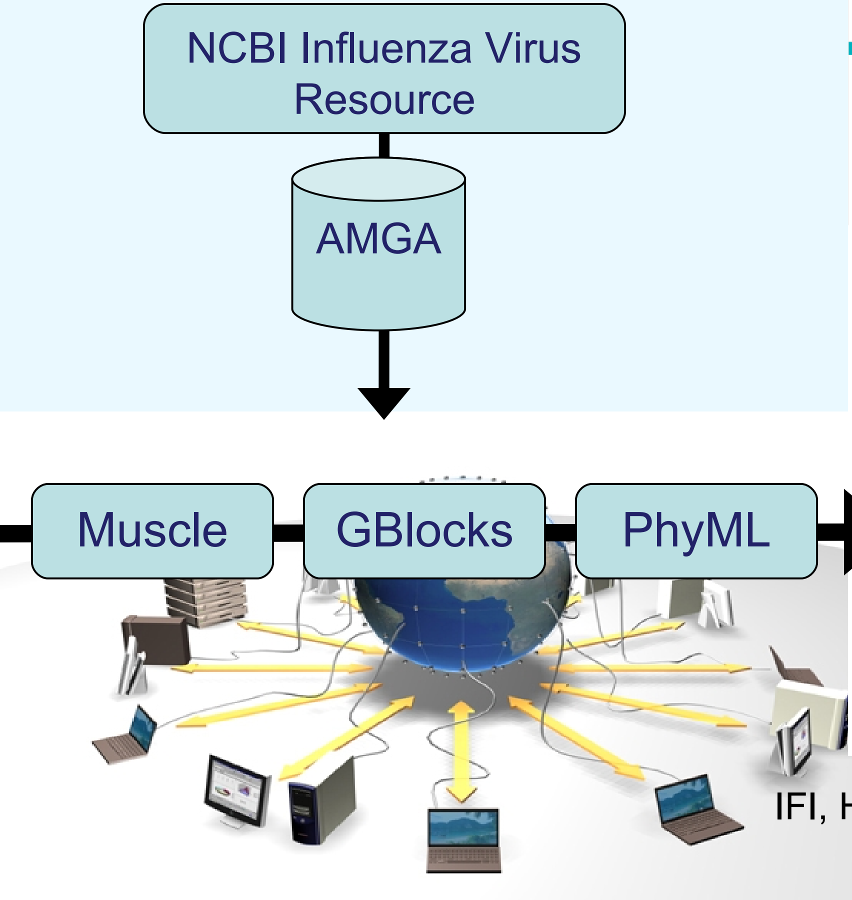
- **Status:**
  - Chemist from INPC spent one month in France (CNRS, HealthGrid)
  - Definition of a project-specific structure for the WISDOM Information System (based on the AMGA metadata catalogue)
  - Workflow definition: 1) First docking 2) Q.S.A.R 3) Second docking
  - “Docking grid service” deployed on the WISDOM Production Environment

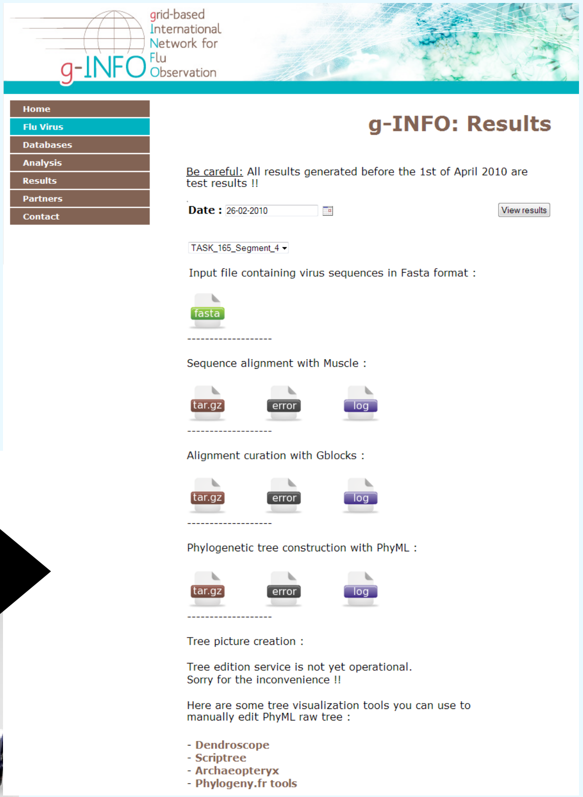
- **Plans**
  - Collect information about **all** chemicals extracted from Vietnamese natural compounds to populate the Information System
  - Select targets of interest
  - Deploy large scale virtual screening on grid

INPC (outside consortium), HealthGrid, CNRSS

Monitor diseases

g-Info: Grid-based International Network for Surveillance to dynamically analyze the influenza molecular biology data, made available on public databases





g-INFO: Results

Be careful: All results generated before the 1st of April 2010 are test results !!

Date : 26-02-2010 View results

TASK_165_Segment_4 ▾

Input file containing virus sequences in Fasta format :

fasta

Sequence alignment with Muscle :

tar.gz error log

Alignment curation with Gblocks :

tar.gz error log

Phylogenetic tree construction with PhyML :

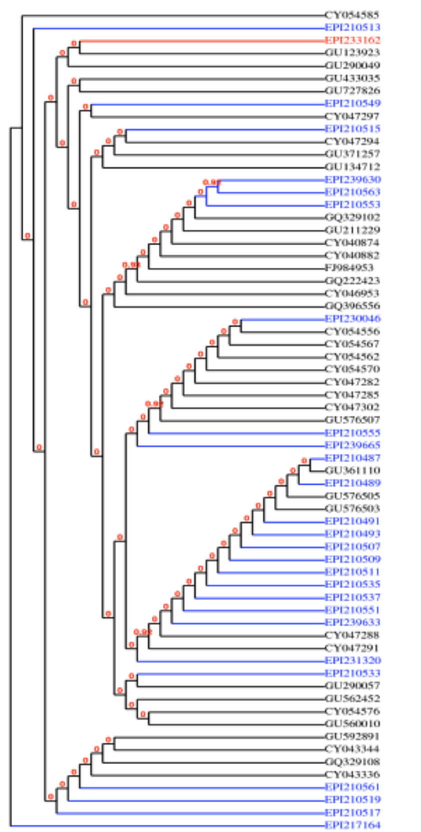
tar.gz error log

Tree picture creation :

Tree edition service is not yet operational.
Sorry for the inconvenience !!

Here are some tree visualization tools you can use to manually edit PhyML raw tree :

- **Dendroscope**
- **Scriptree**
- **Archaeopteryx**
- **Phylogeny.fr tools**



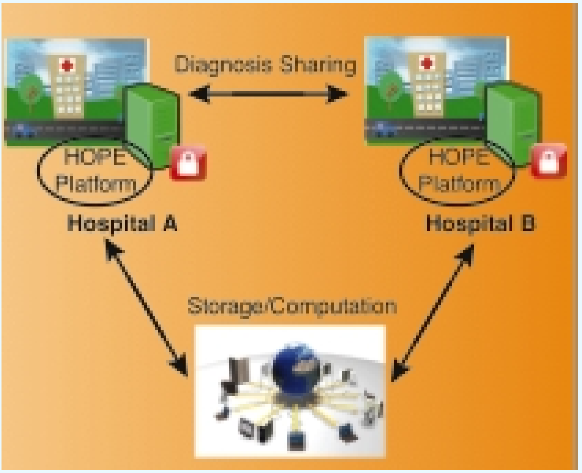
IFI, HealthGrid, CNRS

http://g-info.healthgrid.org/

Share medical data

Help physicians to overcome several medical issues like storing, sharing exchanging, analysing image in a secure way using Grid technology,

- **Telemedicine:** HOPE is a collaborative platform for doctors and physicians (sharing and simulating)
  - Grid site (UI, CE, SE, LFC, WN1, WN2) at IAMI which connected to VinaREN network
  - HOPE installed and configured in the grid node
- **Medical Image Compression** to decrease the size of medical images without any loss of information using JPEG 2000 compression encoder



Integrate Knowledge

- **Development of Dementia Brain (DBRAIN):** integration of gene discovery, protein predictions an drug discovery in a multi-agents systems for diagnosis, therapeutics and treatment of dementia-affected people
- **Grid-enabled research network and infostructure of University of Malaya (GeRaNIUM):** grid-based platform for researchers to access their collective resources, skills, experiences, and results in a secure, reliable and scalable manner
  - Examples:
    - *Neurosciences: altered behaviour in neurodegenerative diseases and addictions*
    - *Grids initiatives for health, biomedical, diseases ecology and biodiversity*

Outline

1. Grid computing to address biomedical challenges
2. WISDOM success story
3. Biomedical applications in the EUAsiaGrid project
4. Conclusion

Conclusion

- **Grid is a well suited to address Life Sciences research and important health concerns of our society**

- **Grid provides to the research community:**
  - Essential data components (genomic sequences, metadata…)
  - Analytical and modelling capabilities
  - Interoperable virtual environment for high performance distributed computation

- **EUAsiaGrid project intends to gather European and Asian scientists to strengthen such research and find solutions using the Grid**

THANKS
FOR YOUR ATTENTION

CONTACTS

Ana Lucia DA COSTA
ana.dacosta@healthgrid.org

Yannick Legré
yannick.legre@healthgrid.org

Vincent BRETON
vincent.breton@clermont.in2p3.fr