Grid enabled life science applications: status and perspectives

Vincent Breton
LPC IN2P3/CNRS

Credit for the slides: M. Hofmann, N. Jacq, V. Kasam, J. Montagnat

IN2P3
Institut National de Physique Nucléaire et de Physique des Particules

CNRS

Information Society and Media

- Introduction
- EGEE biomedical applications
- Focus on virtual screening

the Vision

Computing Grid

For data crunching applications

An environment, created through the sharing of resources, in which heterogeneous and dispersed data :

- molecular data (ex. genomics, proteomics)
- cellular data (ex. pathways)
- tissue data (ex. cancer types, wound healing)
- personal data (ex. EHR)
- population (ex. epidemiology)

as well as applications, can be accessed by all users as an tailored information providing system according to their authorisation and without loss of information.

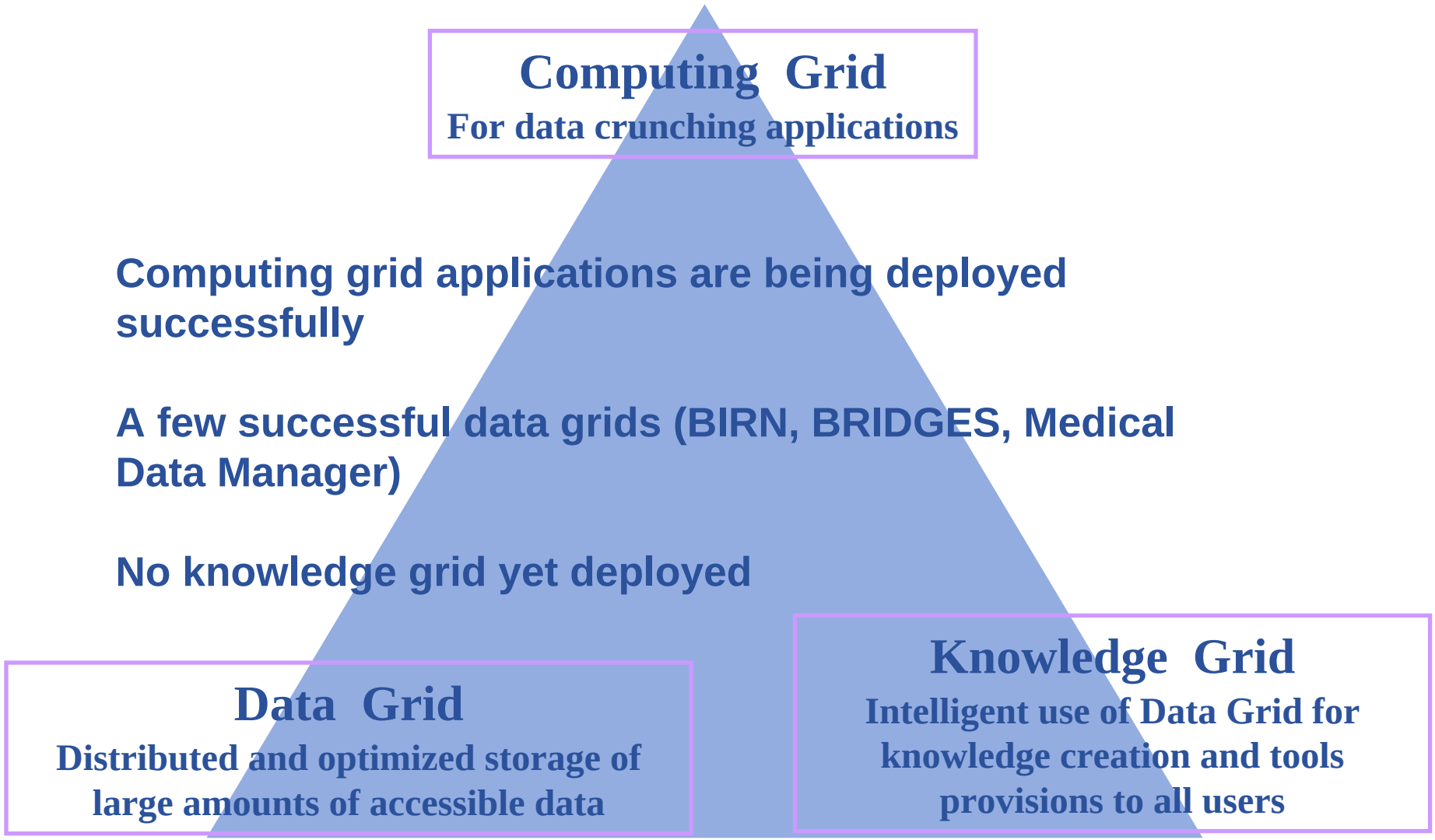
Data Grid

Distributed and optimized storage of large amounts of accessible data

Knowledge Grid

Intelligent use of Data Grid for knowledge creation and tools provisions to all users

Biomedical applications are running today on European grids !



Computing grid applications are being deployed successfully

A few successful data grids (BIRN, BRIDGES, Medical Data Manager)

No knowledge grid yet deployed

EGEE biomedical applications

Biomed Achievements (I)

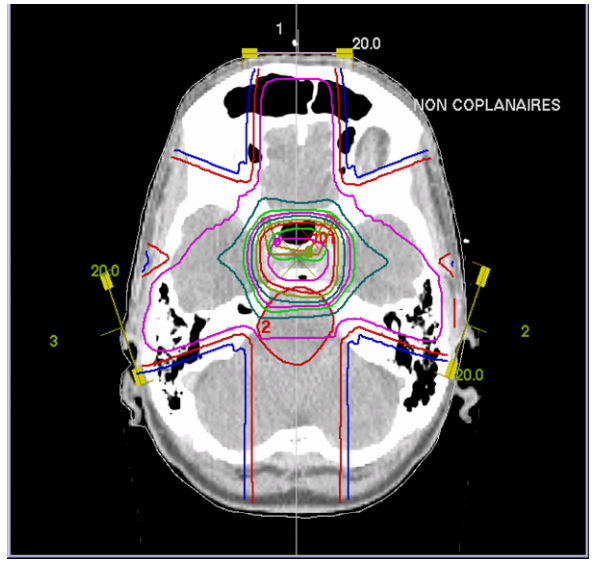
- Goal: demonstrate grid potential for real-scale biomedical applications.
- Start of EGEE in Apr 04: several app. prototypes developed, not yet deployed.
- First year achievements
  - Organization of work
    - Creation of the “Biomed” Virtual Organization
    - Deployment of associated services (guinea pig VO)
      - Definition of application test cases
      - Use of them to test new or updated EGEE components
    - Creation of the “Biomed Task Force”
      - Biomedical user support: GGUS, Data Challenge, tutorials, …
      - Collaboration with middleware and infrastructure activities
  - Application deployment
    - Successful deployment of applications in the field of bioinformatics and medical imaging
    - ~70k jobs from Biomed users reported at EGEE’s first review

Biomed Achievements (II)

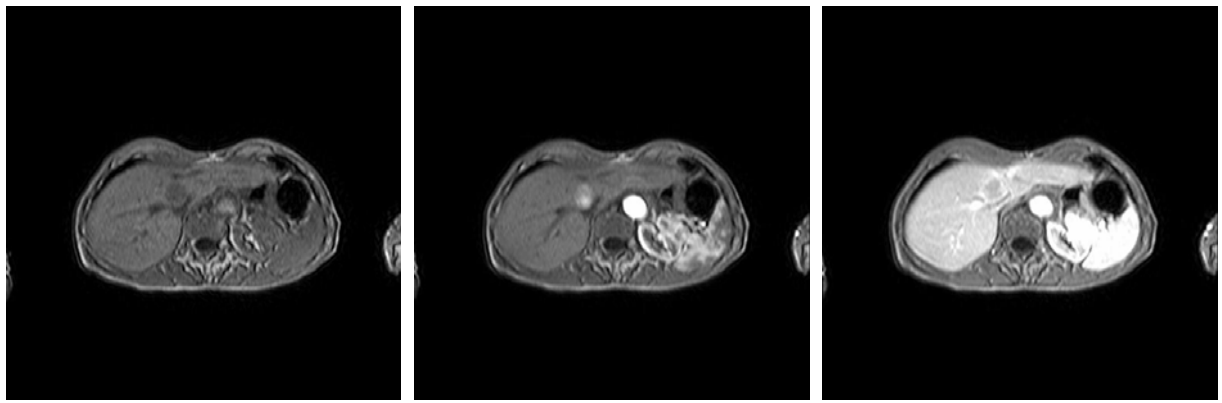
- Development of secured data management and complex data flows on the grid
  - Medical Data Management group has demonstrated complete chain for processing medical images on the grid using these services
- First CPU-intensive grid deployments for bioinformatics in the world
  - *In silico* drug discovery against malaria and bird flu
  - Very large impact in the grid community
  - Biologically-relevant results being processed
- Sustained growth of the “Biomed” VO
  - New apps. interested in joining the VO: 11 in DNA4.4 inventory
  - 3 sub-areas: bioinformatics, medical imaging, drug discovery
  - ~80 users
  - 1000 jobs / day on average

Medical image processing

- GATE: Radiotherapy planning
  - CNRS-IN2P3
  - Monte Carlo simulation
  - Parallel execution on different seeds

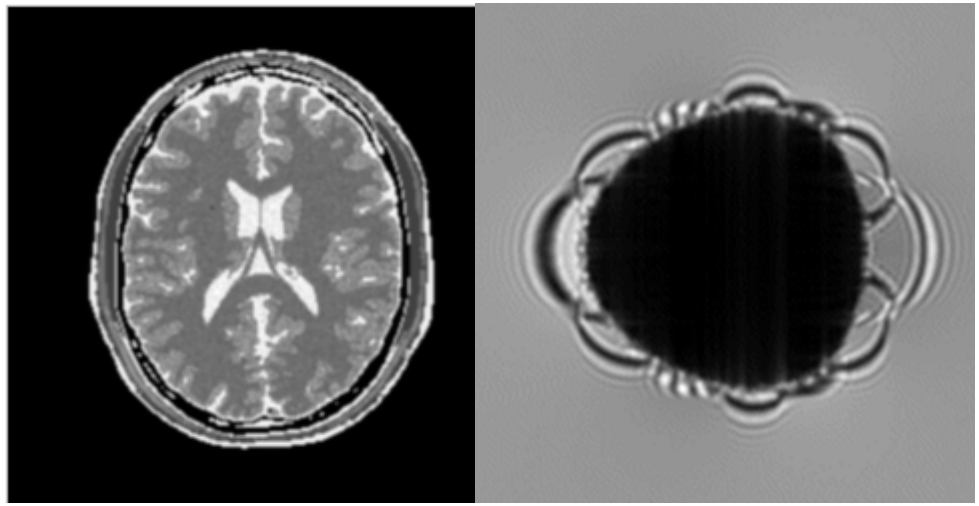


- Pharmacokinetics: contrast agent diffusion study
  - UPV
  - Medical images registration
  - Distribution of registration pairs

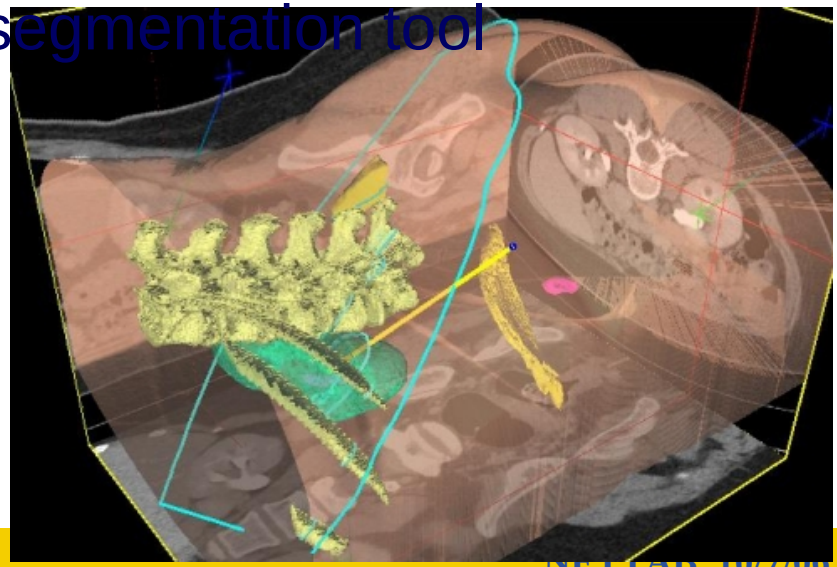


Medical image processing

- SiMRI3D MRI simulation
  - CNRS-CREATIS
  - Magnetic Resonance physics simulation (Bloch's equation)
  - Parallel processing (MPI)

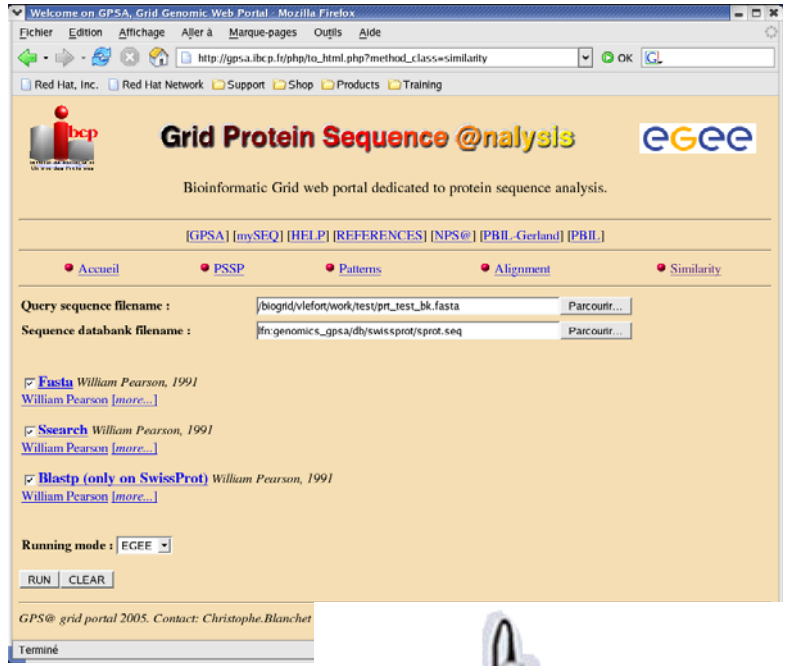


- gPTM3D: Radiological images segmentation tool
  - CNRS-LRI, CNRS-LAL
  - Deformable-contour based segmentation
  - Interactivity through agent-based scheduling



Bioinformatics

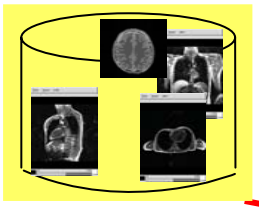
- GPS@: bioinformatics portal
  - CNRS-IBCP
  - http://gpsa.ibcp.fr/ web portal
  - Existing (but overloaded NPSA portal)
  - Tens of bioinformatics legacy code
  - Thousands of potential users



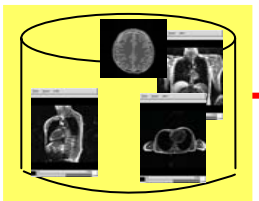
- Electron-microscopic image reconstruction
  - CNB-CSIC
  - Image filtering and noise reduction
  - 3D structure analysis

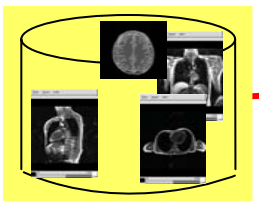
Secure Medical Data Management with EGEE

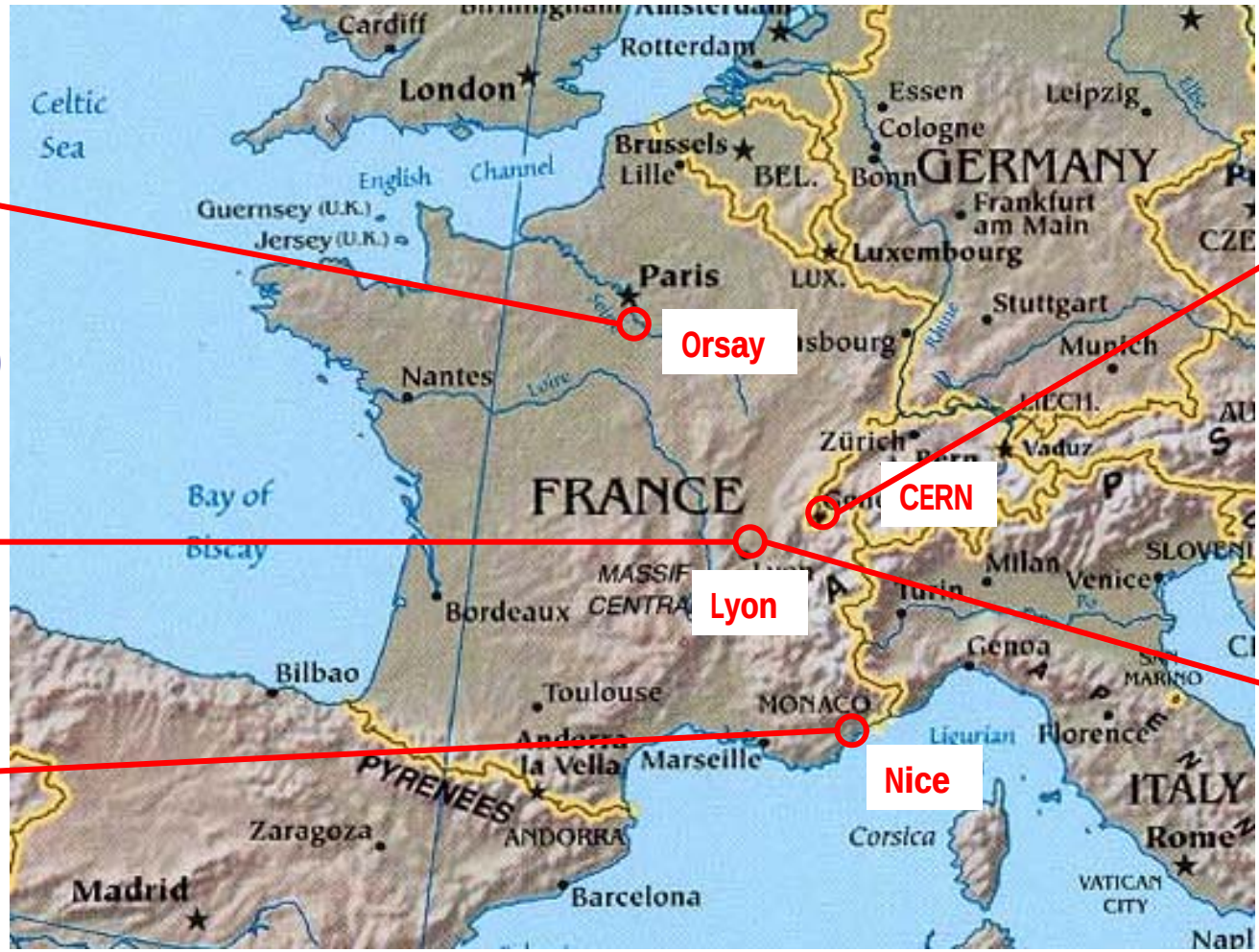
- 3 SRM-DICOM servers — CNRS-I3S; CNRS-CREATIS,CNRS-LAL, CERN
- Distributed secure medical data management services



gLite 3.0







Fireman
File Catalog

Hydra Key store

AMGA Metadata

Potential vs current impact of EGEE applications on scientific community

BioinfoGRID

Application	dev	user	Potentially impacted community	Most limiting factors
GATE	8	12	400	Overhead on jobs execution time Middleware stability Storage space
CDSS	9	9	30 (mental diseases) + 50 (soft tissue tumours) in a short term.	For production: overhead for short jobs. For training the classifiers: computing time (around 1 week)
gPTM3D	0.5	1	Tens (clinical researchers)	Jobs submission response time (in particular queuing delay) Lack of firewall-proof connectivity solution
SiMRI3D	10	10	Several hundreds from the MR physics, medical and image processing communities	Correct handling of MPI jobs (too many errors today). Lack of scheduling time estimation.
Bronze Std	2	4	Tens to hundreds once a proper interface has been set up.	Capacity to handle lot (hundreds) of jobs concurrently in an efficient manner. Currently the speed-up achieved is far from the expected bound. This is still under investigation but probably related to the bottlenecks of centralized RBs / UIs.
Pharmcok.	9	10	Hundreds when the tool will prove stable and accurate enough.	Sufficient computing power dedicated to the application. In production it should represent 3 CPU years per year.
GPS@	3	10	Difficult to estimate as the portal is opened anonymously to the biological community (probably several thousands)	Efficient handling of short jobs. Anonymous users authorization. Automatic replication.
Xmipp_ML	5	10-15	Will be proposed to a NoE: hundreds.	CPU intensive Reliable MPI support High data throughput (data replication)
SPLATCHE	4	5	Limited to a community of specialists in a short term	Sufficient CPUs availability (> 70) to compete with local cluster Multi-data jobs submission capability
WISDOM	8	9	20 in a short term (2006). In the order of 100 later.	Reliability of services, especially WMS Security of data Number of CPUs
GROCK	4	Tens	Thousands	ses in the working nodes.

Reference: EGEE deliverable DNA4.1

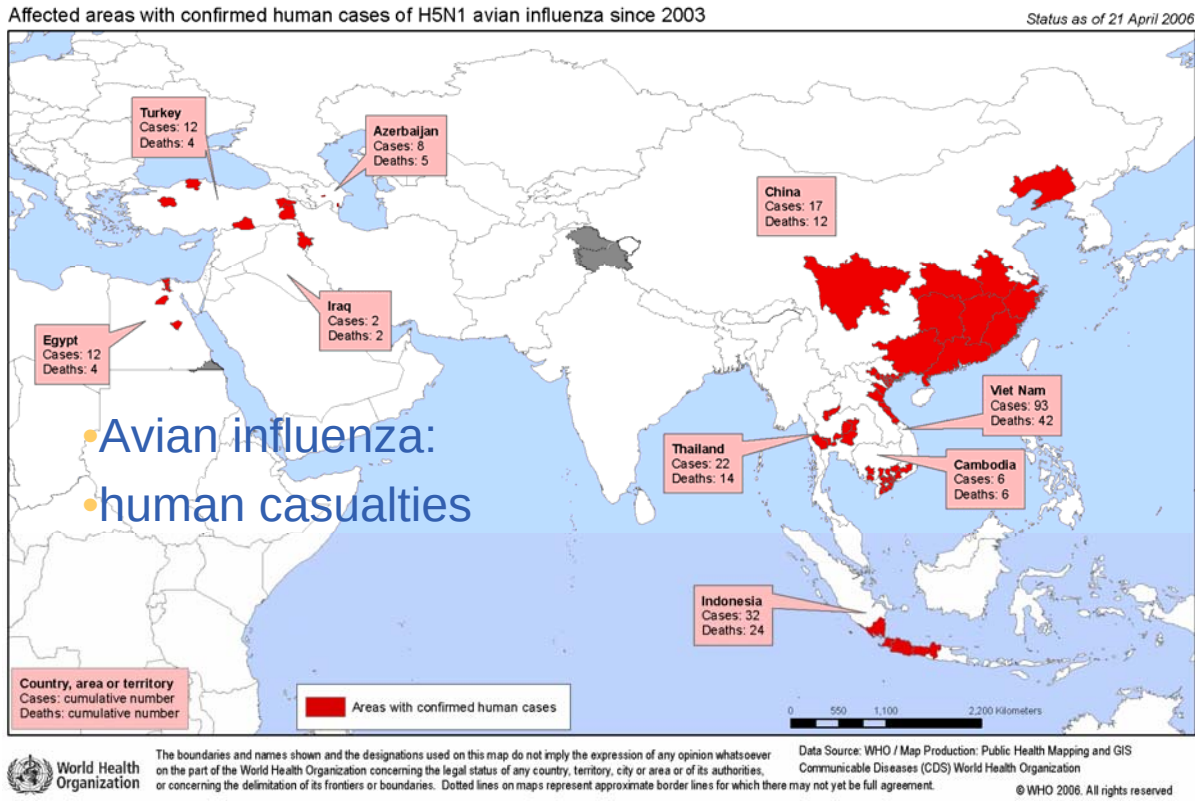
Focus on virtual screening

Addressing neglected and emerging diseases

- Neglected and emerging diseases are major public health concerns in the beginning of the 21st century
  - Neglected diseases keep suffering lack of R&D
  - Emerging diseases are a growing threat to world public health

Both emerging and neglected diseases require:

- **Early detection**
  - **Emergence, resistance**
- **Epidemiological watch**
  - **Emergence, resistance**
- **Prevention**
- **Search for new drugs**
- **Search for vaccines**

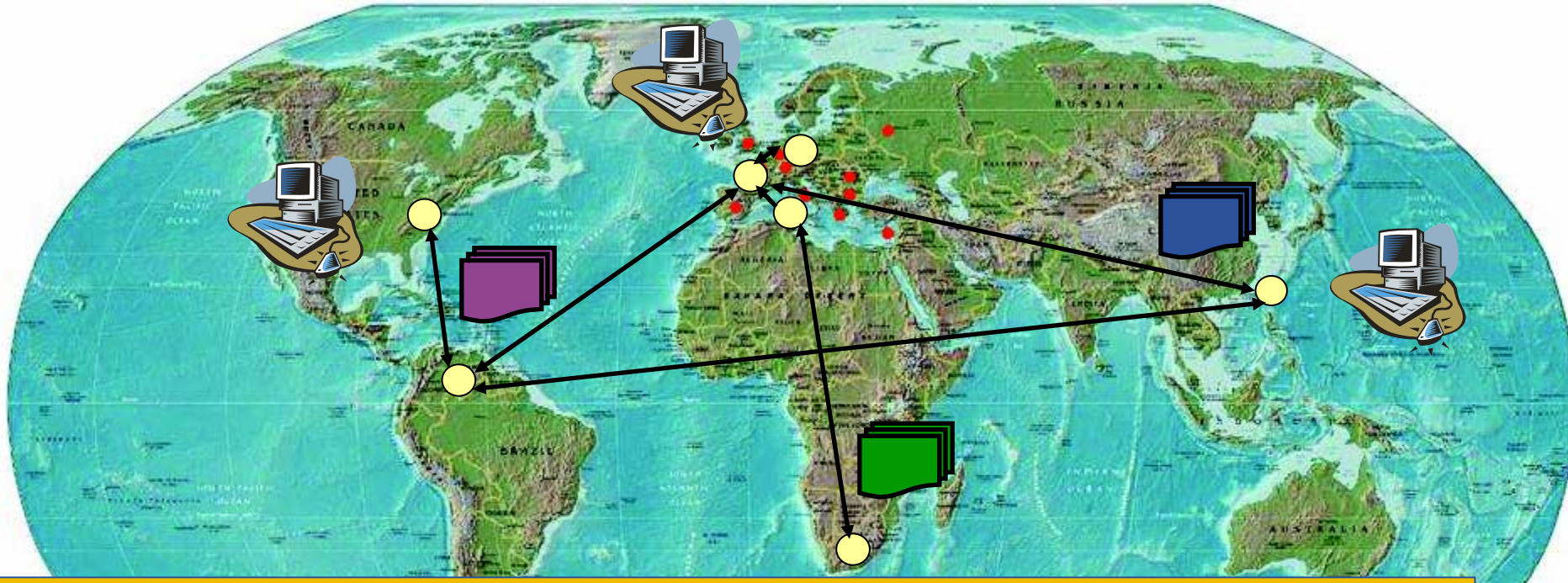


- Avian influenza:
- human casualties

BioinfoGRID

The grid added value for international collaboration on emerging and neglected diseases

- Grids offer unprecedented opportunities for sharing information and resources world-wide

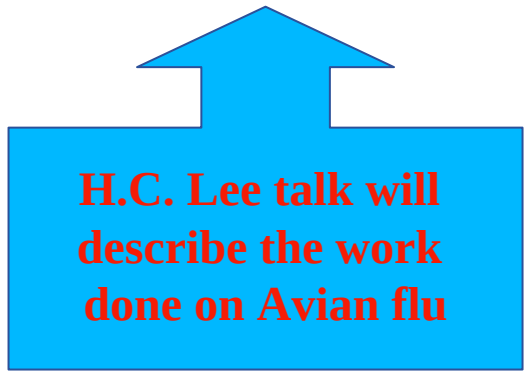


Grids are unique tools for :
-Collecting and sharing information (Epidemiology, Genomics)
-Networking experts
-Mobilizing resources routinely or in emergency (vaccine & drug discovery)

In silico drug discovery against neglected and emerging diseases

- Grids open new perspectives to *in silico* drug discovery
  - Reduced cost for R&D against neglected diseases
  - Accelerating factor for R&D against emerging diseases

- EGEE plays a pioneering role in exploring grid impact
  - Data challenge against malaria in the summer 2005
  - Data challenge against bird flu in April-May 2006



World wide In Silico Docking On Malaria

Burden of Diseases in Developing World

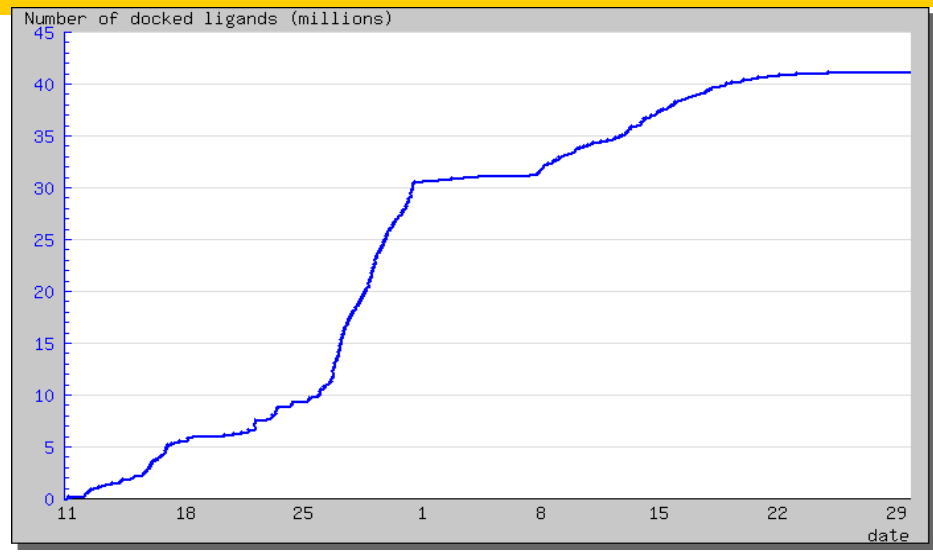
Disease	Endemic Countries	People at Risk (million)	Clinical Incidence/yr (million)	Deaths/yr (million)	Disease Burden (DALYs-million)
HIV/AIDS	180	5.900	40	2.8	86
Malaria	101	2.400	300-500	1.2	44.7
TB	211	1.987	8	1.6	35.4
African trypanosomiasis	36	60	0.3-0.5	0.05	1.5
Chagas Disease	21	100	16-18	0.01	0.7
Leishmaniasis	88	350	12	0.05	2
Filariasis	80	1.000	120	---	5.8
Schistosomiasis	76	500-600	140	0.01	1.7
Onchocerciasis	36	120	18	---	0.5
Leprosy	24	---	0.8	---	0.2

Where grids can help addressing neglected diseases

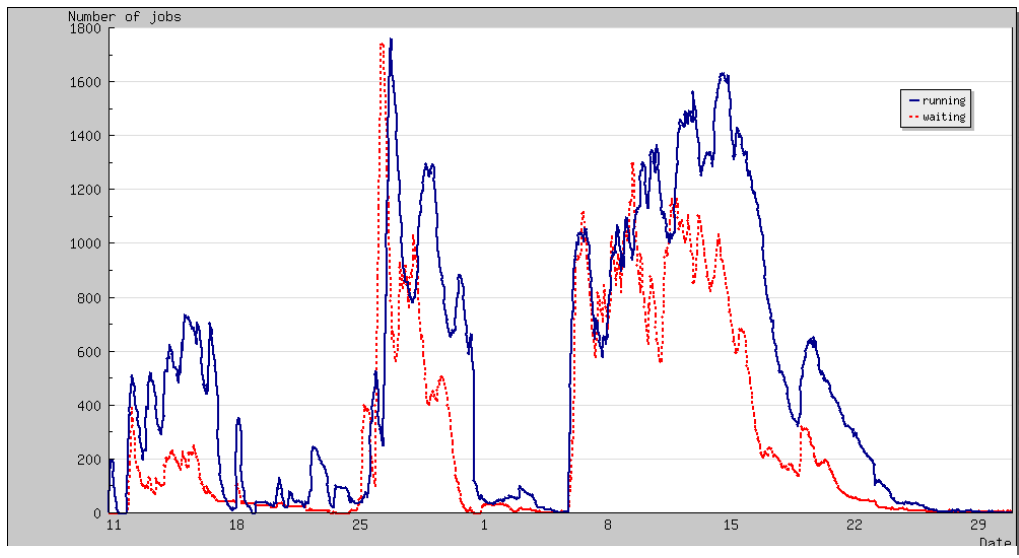
- Contribute to the development and deployment of new drugs and vaccines
  - Improve collection of epidemiological data for research (modeling, molecular biology)
  - Improve the deployment of clinical trials on plagued areas
  - Speed-up drug discovery process (in silico virtual screening)
- Improve disease monitoring
  - Monitor the impact of policies and programs
  - Monitor drug delivery and vector control
  - Improve epidemics warning and monitoring system
- Improve the ability of developing countries to undertake health innovation
  - Strengthen the integration of life science research laboratories in the world community
  - Provide access to resources
  - Provide access to bioinformatics services

First initiative: World-wide In Silico Docking On Malaria (WISDOM)

- Initial partners: Fraunhofer Institute, CNRS – IN2P3
- Significant biological parameters
  - two different molecular docking applications (Autodock and FlexX)
  - about one million virtual ligands selected
  - target proteins from the parasite responsible for malaria
- Significant numbers
  - Total of about 46 million ligands docked in 6 weeks
  - 1TB of data produced
  - Up 1000 computers in 15 countries used simultaneously corresponding to about 80 CPU years
  - Average crunching factor ~600

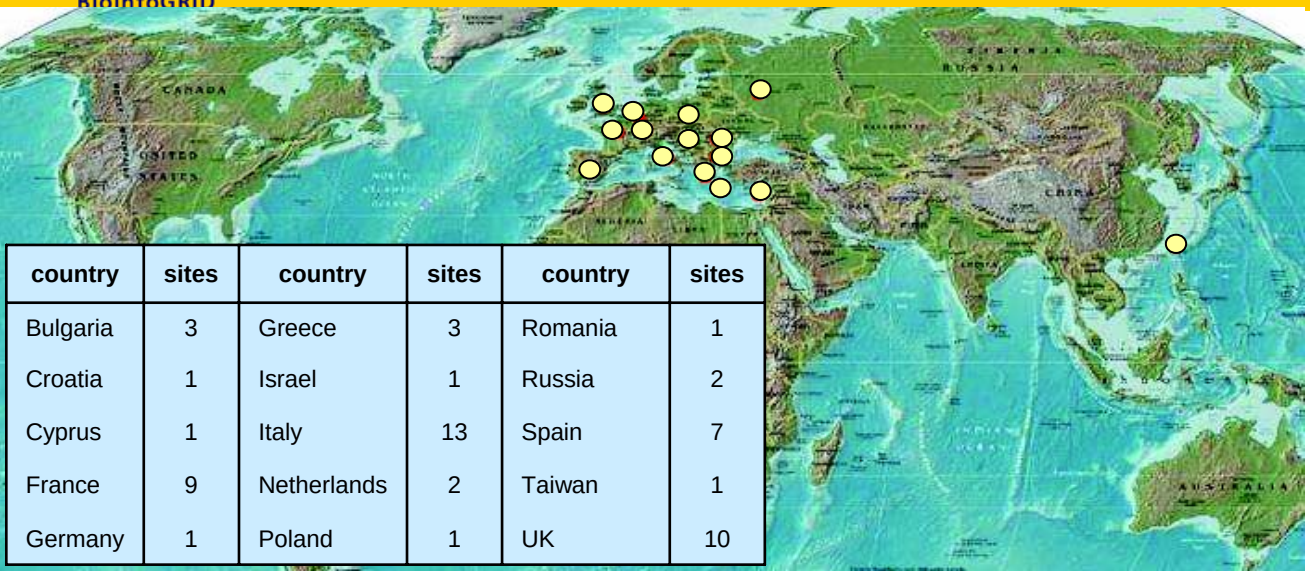


Number of running and waiting jobs vs time



Number of running and waiting jobs vs time

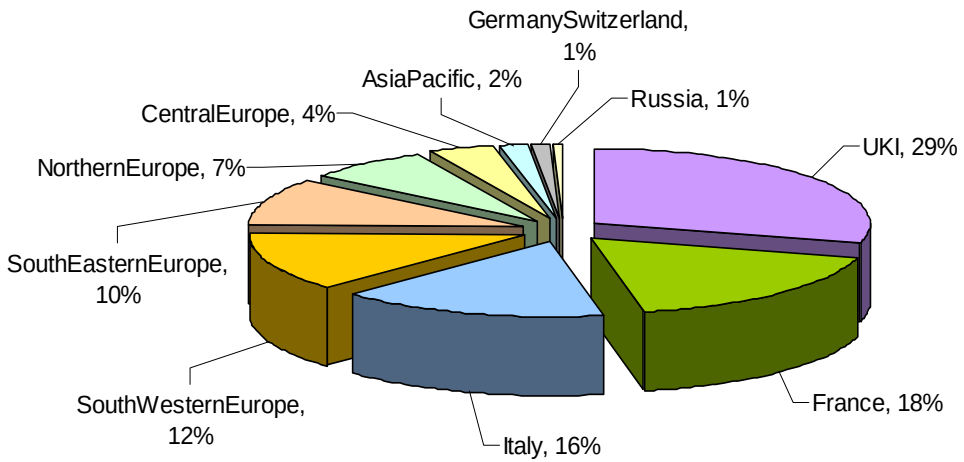
Deployment on EGEE infrastructure, wisdom.eu-egee.fr



country	sites	country	sites	country	sites
Bulgaria	3	Greece	3	Romania	1
Croatia	1	Israel	1	Russia	2
Cyprus	1	Italy	13	Spain	7
France	9	Netherlands	2	Taiwan	1
Germany	1	Poland	1	UK	10

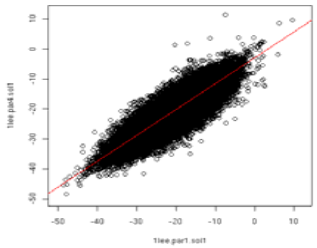
Countries with nodes contributing to the data challenge WISDOM

Total amount of CPU provided by EGEE federation: 80 years



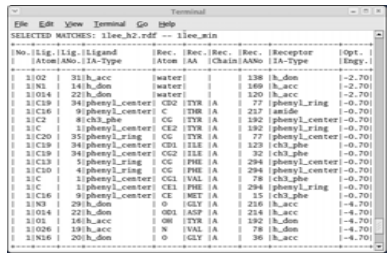
Strategies in result analysis

- Results based on Scoring

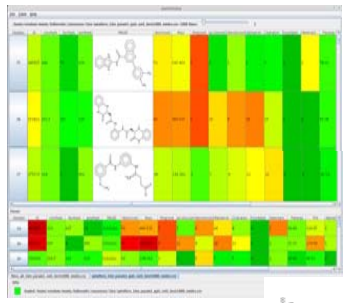


Credit: V. Kasam
Fraunhofer Institute

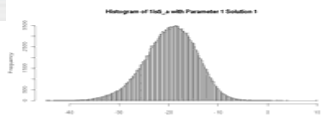
- Results based on match information

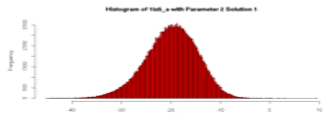


- Results based on consensus scoring

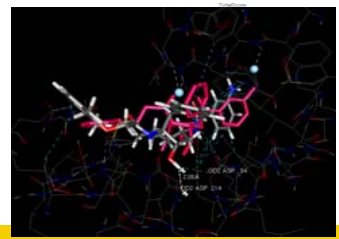


- Results based on different parameter settings

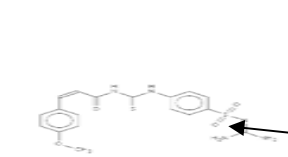




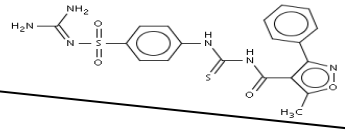
- Results based on knowledge on binding site



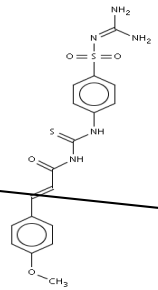
Top 10 compounds by scoring



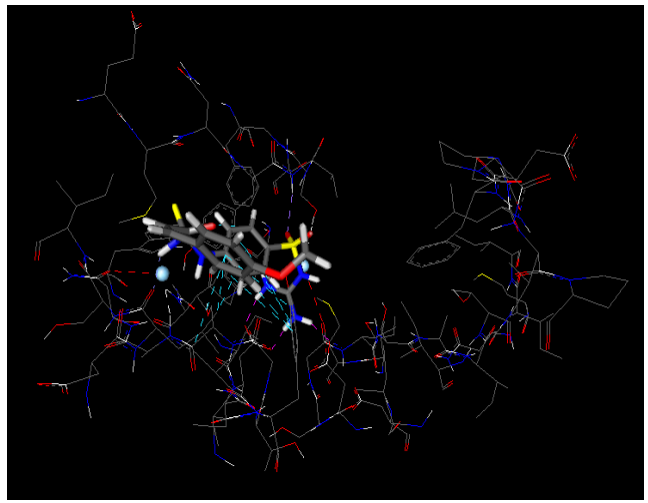
1. WISDOM-490500

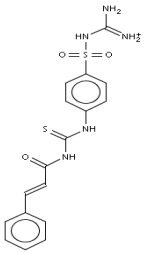


2. WISDOM-491901

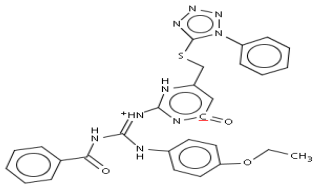


3. WISDOM-490515

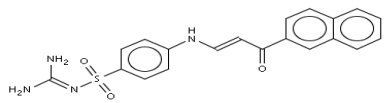




4. WISDOM-490514

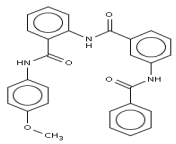


5. WISDOM-462271

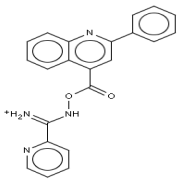


6. WISDOM-235118

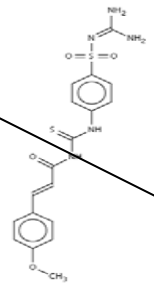
Top scoring but poor binding mode



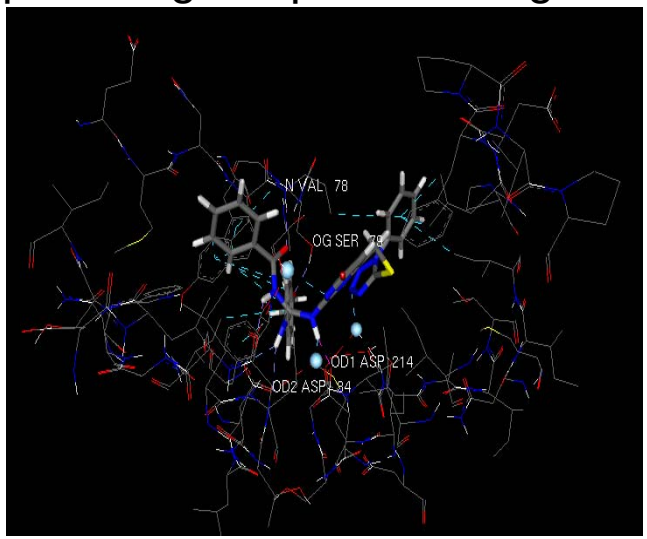
7. WISDOM-278345

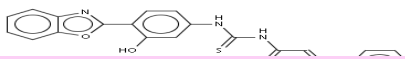


8. WISDOM-360604



9. WISDOM-490502

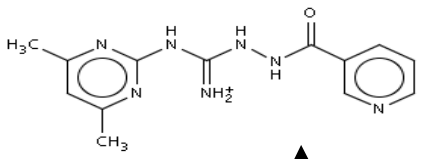


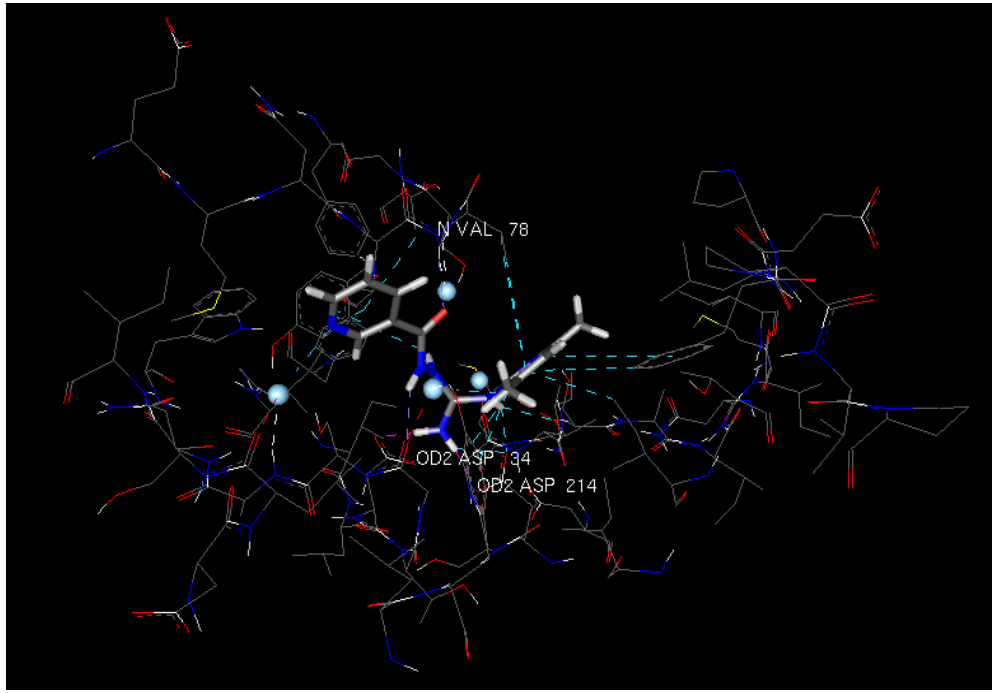


Known inhibitors: thiourea and urea compounds
Potentially new inhibitors: guanidino compounds

Top scoring, good binding mode, interactions to key residues

Compounds for Molecular Dynamics: Guanidino compounds





Terminal

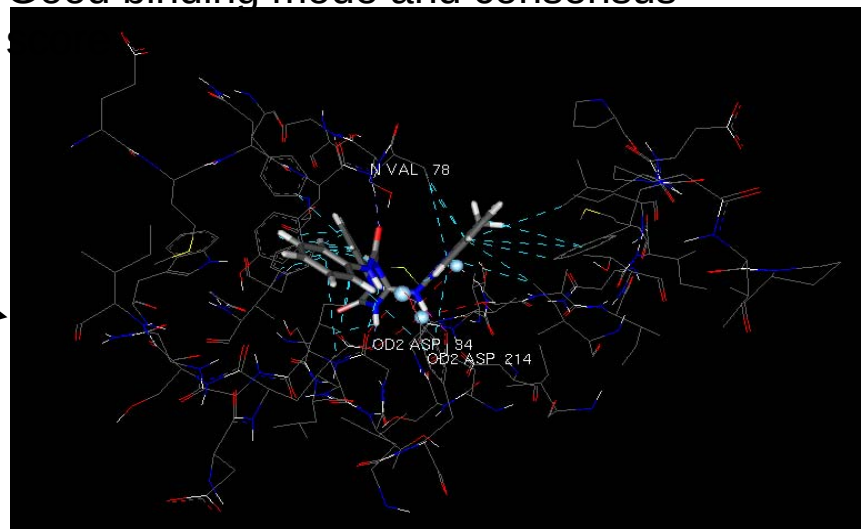
File Edit View Terminal Tabs Help

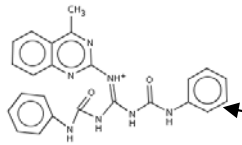
No.	Lig. Atom	Lig. ANo.	Ligand IA-Type	Rec. Atom	Rec. AA	Rec. Chain	Rec. AANo	Receptor IA-Type
1	N1	5	h_acc	water			58	h_don
1	N7	19	h_acc	water			39	h_don
1	N7	19	phenyl_center	C	TYR	A	77	amide
1	C7	13	amide	CG	TYR	A	77	phenyl_center
1	C13	21	ch3_phe	CG	TYR	A	192	phenyl_center
1	N1	5	phenyl_center	CE1	PHE	A	294	phenyl_ring
1	N1	5	phenyl_center	CG2	VAL	A	78	ch3_phe
1	N1	5	phenyl_center	CD1	ILE	A	300	ch3_phe
1	N1	5	phenyl_center	CE2	TYR	A	192	phenyl_ring
1	N1	5	phenyl_center	CG1	VAL	A	78	ch3_phe
1	C3	3	phenyl_ring	CG	PHE	A	294	phenyl_center
1	N3	8	h_don	OG1	THR	A	217	h_acc
1	N3	8	h_don	OD1	ASP	A	214	h_acc
1	N4	10	h_don	OD1	ASP	A	34	h_acc
1	N4	10	h_don	OD2	ASP	A	214	h_acc
1	N4	10	h_don	OD1	ASP	A	214	h_acc
1	C12	20	phenyl_ring	CG	TYR	A	77	phenyl_center
1	N7	19	phenyl_center	CD2	TYR	A	77	phenyl_ring
1	O1	14	h_acc	N	VAL	A	78	h_don
1	N6	12	h_don	O	GLY	A	36	h_acc

Note: Satisfied all criteria, good binding mode, interactions to key residues, good score, appropriate descriptors.

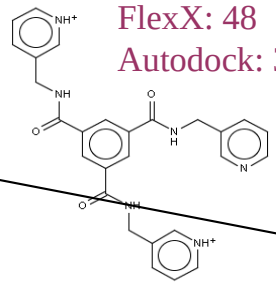
Compounds from consensus scoring

Good binding mode and consensus

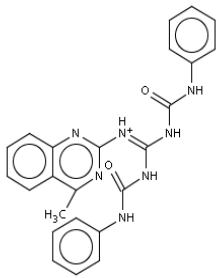




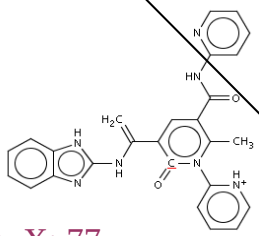
FlexX: 30
Autodock: 24



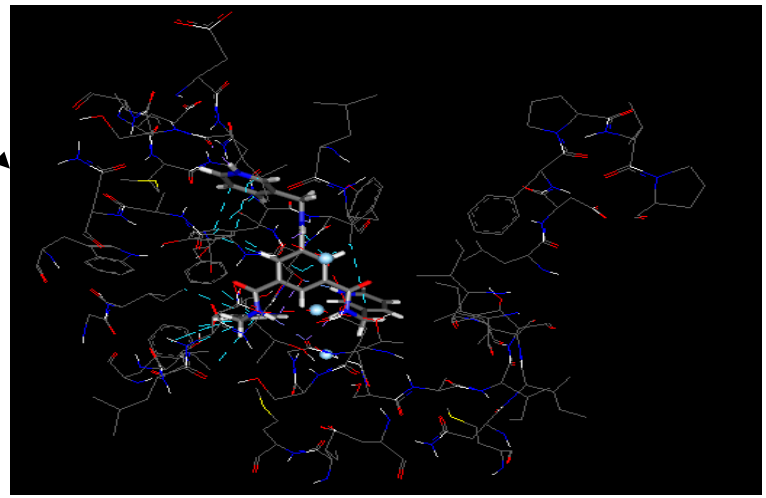
FlexX: 48
Autodock: 33



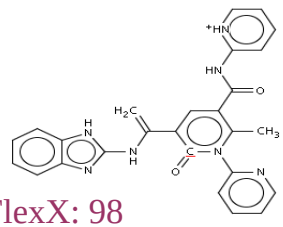
FlexX: 30
Autodock: 97



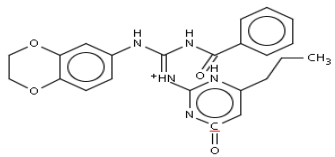
FlexX: 77
Autodock:
130



Good score but bad binding mode



FlexX: 98
Autodock:
110

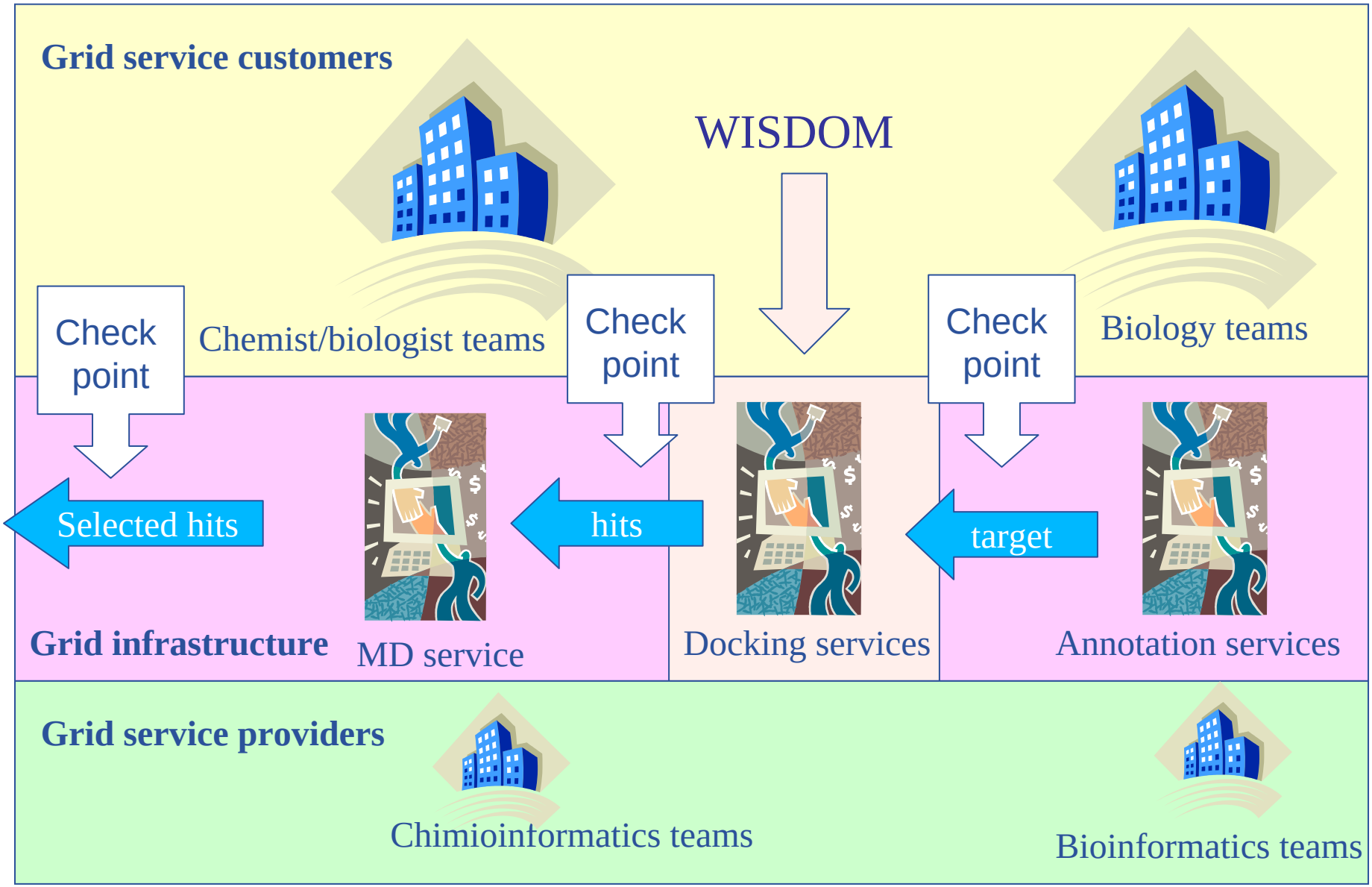


FlexX: 60
Autodock:160

Perspectives

From virtual docking to virtual screening

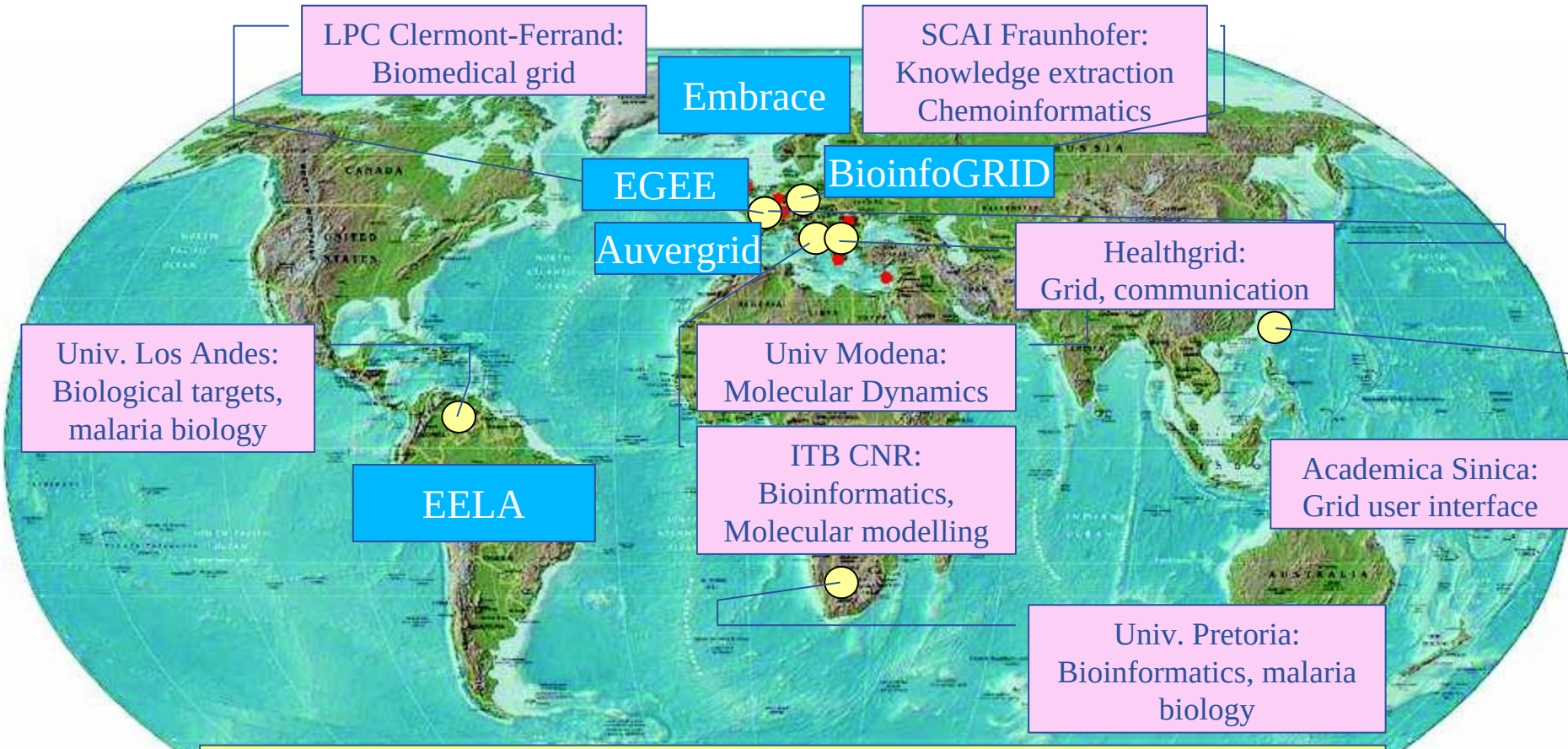
BioinfoGRID



The next steps

- Docking step still requires a lot of manual intervention
  - Goal: reduce as much as possible the time needed for experts to analyze the results
  - Task: improve output data collection and post-docking analysis
  - Contribution from CNR-ITB, within the framework of EGEE-II
- The next step after docking is Molecular Dynamics
  - Goal: grid-enable the reranking of the best hits
  - Task: deploy Molecular Dynamics computations on grid infrastructures
  - Contribution from CNRS-IN2P3, within the framework of BioinfoGRID
- Beyond virtual screening, the long term vision: building a grid for malaria
  - To provide services to research labs working on malaria
  - To collect and analyze epidemiological data

A grid for malaria



Use the grid technology to foster research and development on malaria and other neglected diseases

Contacts also established with WHO, Microsoft, TATRC, Argonne, SDSC, SERONO, NOVARTIS, Sanofi-Aventis, Hospitals in subsaharian Africa,

WISDOM-II

- WISDOM-II is the second large scale docking deployment against neglected diseases
- Biological goals
  - Validation of virtual vs in vitro screening
  - Virtual docking on new malaria targets
    - New targets from Univ. Pretoria, Univ. Los Andes, CEA Grenoble, Univ. Modena
  - New compound libraries
    - Thai library of compounds
  - Possible extension to other neglected diseases
    - Contacts with Univ. Glasgow
- Grid goals:
  - Improve the user interface, the job submission system and the post-processing (BioinfoGRID, EGEE, Embrace)
  - Test the infrastructure at a larger scale (100 -> 500 CPU years)
  - Test the deployment on several infrastructures: Auvergrid, EGEE, EELA

WISDOM timeline

WISDOM	MD reranking			In vitro testing			Further processing											
WISDOM II	Preparation			Deploy ment		Analysis						MD reranking						
Avian Flu	Analysis						MD reranking						Further processing					
Avian Flu II							Preparation				Deploy ment		Analysis					
Month	6 06	7 06	8 06	9 06	10 06	11 06	12 06	1 07	2 07	3 07	4 07	5 07	6 07	7 07	8 07	9 07	10 07	11 07	1 0

Conclusion

- Life science applications are currently running on grid infrastructures
  - Example of EGEE biomedical applications
- Most of these applications are compute intensive
  - Emergence of data grid applications
- Very large scale virtual docking achieved
  - More details from H.C. Lee talk
- Exciting perspectives to develop in silico drug discovery
  - Collaboration of partners and EC projects
  - WISDOM-II, further step towards a malaria grid

Training

- Joint tutorial co-organized by Embrace, EGEE and BioinfoGRID
- Location: Clermont-Ferrand
- Dates: 2 days in October 2006
- Goal: get bioinformaticians to understand the grid to the extent that they can install and use it
- Contact: Florence Jacq, fjacq@clermont.in2p3.fr