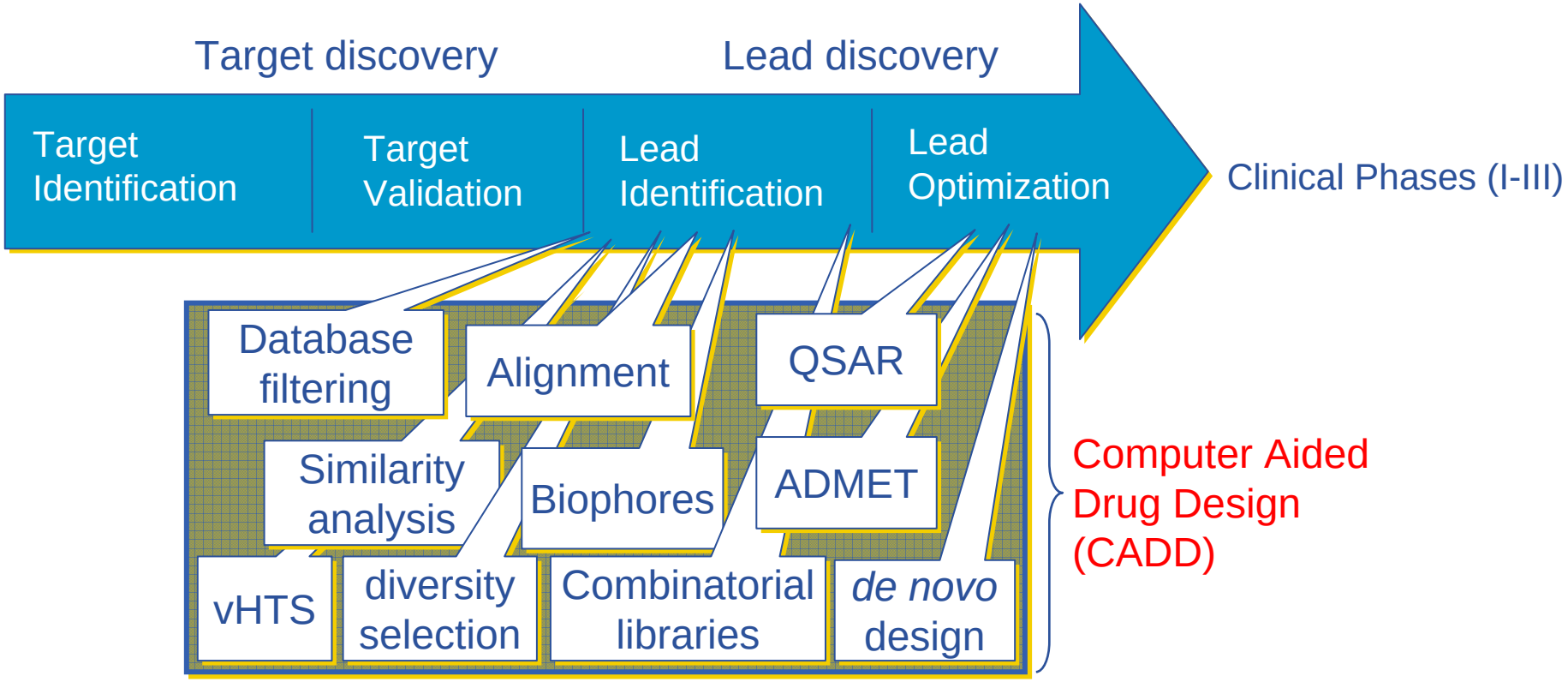
Grid enabled *in silico* drug discovery

Vincent Breton
CNRS/IN2P3
Credit for the slides: N. Jacq

Phases of a pharmaceutical development



Duration: 12 – 15 years, Costs: 500 - 800 million US $

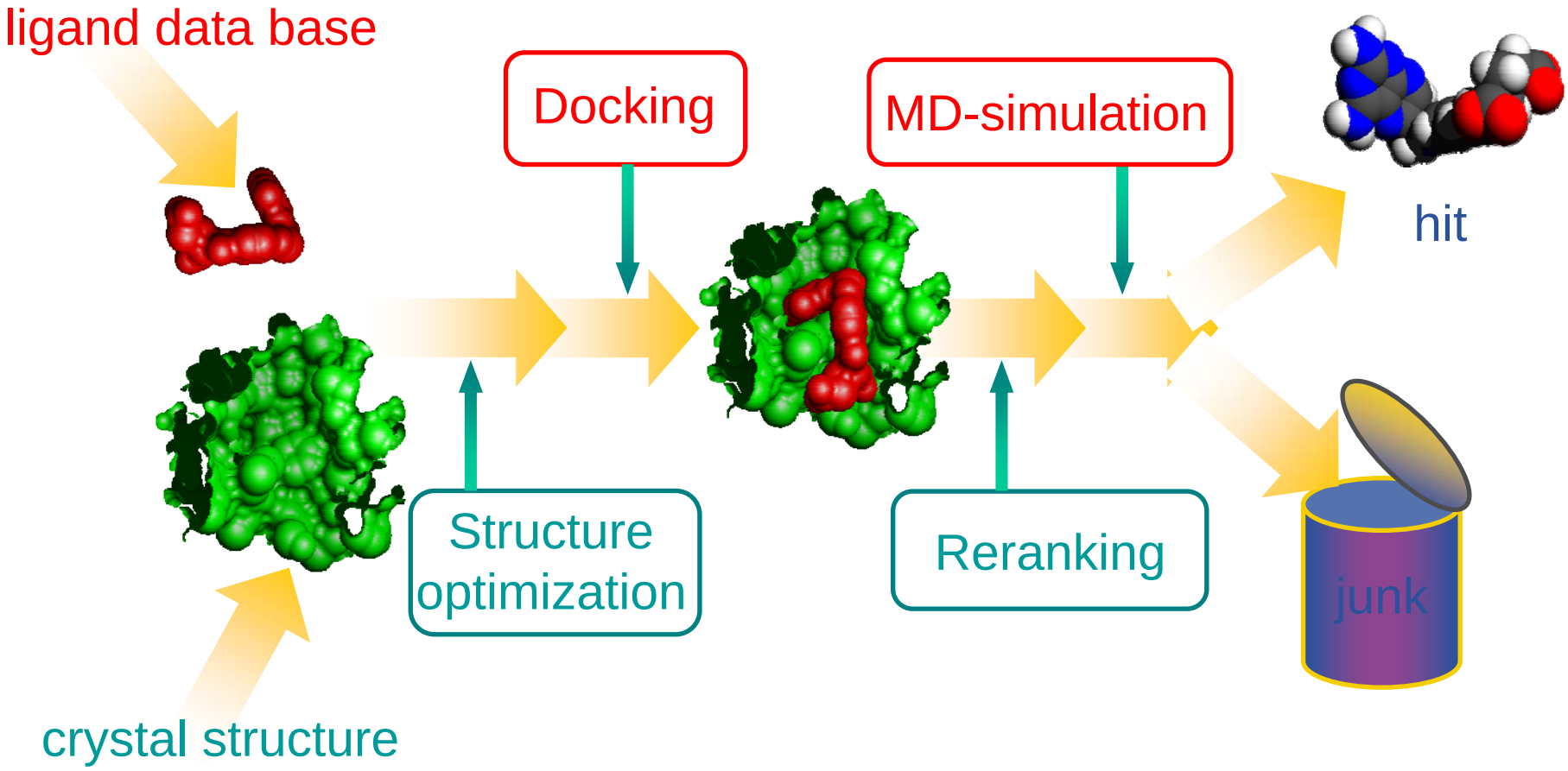
Selection of the potential drugs

- 28 million compounds currently known
- Drug company biologists screen up to 1 million compounds against target using ultra-high throughput technology

- Chemists select 50-100 compounds for follow-up
- Chemists work on these compounds, developing new, more potent compounds

- Pharmacologists test compounds for pharmacokinetic and toxicological profiles
- 1-2 compounds are selected as potential drugs

Dataflow and workflow in a virtual screening



Computational aspects of Drug Discovery virtual screening

- **Enable scientists to quickly and easily find ligands binding to a particular target protein**
  - growth of targets number
  - growth of 3D structures determination (PDB database)
  - growth of computing power
  - growth of prediction quality of protein-compound interactions
- **Experimental screening very expensive : difficult for academic or small companies**
- **Enrichment =** $\frac{\text{Actives molecules}}{\text{Tested molecules}}$ ↗

Grid added value for the first steps of *in silico* drug discovery

- **Target identification and validation**
  - Volume of molecular biology data is exponentially increasing
  - Grid added value: interoperability, sharing of data content and tools
- **Large scale virtual screening to select the most promising compounds**
  - Distributed computing
  - output data management
- **Molecular dynamics to further assess selected compounds**
  - Parallel computing

Grid infrastructures vs pervasive grids

- **A grid infrastructure uses an identified set of resources properly administered behind firewalls**
- **Grid infrastructures vs pervasive grids**
  - Large scale docking on pervasive grid already achieved (Grid.org, Decrypthon, World Community Grid)
    - Centralized job submission and data management
    - Limited security model
    - No output data distribution (web portal)
    - Limited quality of service (no user support)
- **Grid infrastructures vs clusters**
  - Sharing of computing resources
  - Data management: distribution/replication of data
  - Sharing of services (participating groups bring their expertise)

Potential grid services

Grid service customers

Chemist/biologist teams

Biology teams

Grid infrastructure

Selected hits

MD service

hits

Virtual Docking services

target

Annotation services

Grid service providers

Chimioinformatics teams

Bioinformatics teams

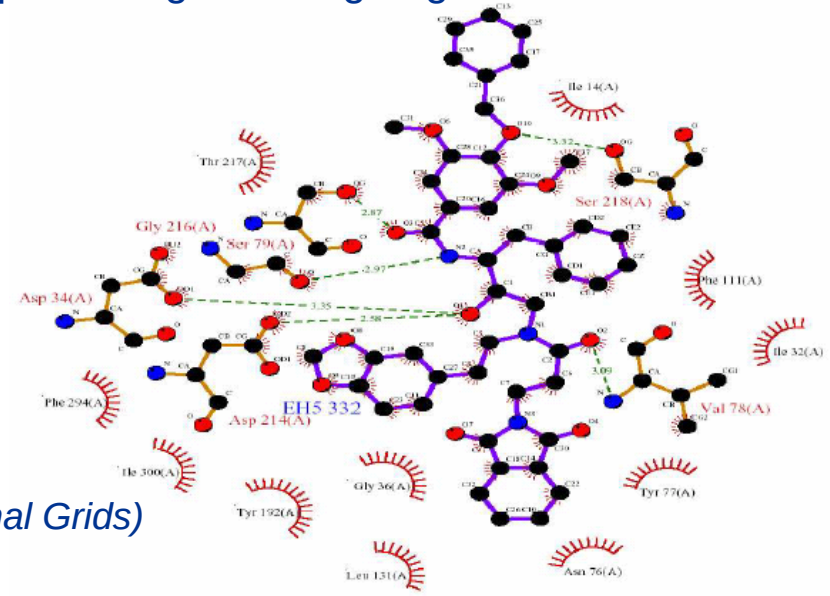
WISDOM : Wide In Silico Docking On Malaria

- **Scientific objectives**
  - start enabling *in silico* drug discovery in a grid environment to address the deadliest infectious disease on earth: malaria
  - Demonstrate to the research communities active in the area of drug discovery the relevance of grid infrastructures

- **Goals of the first "data challenge" (July - September 2005)**
  - Biological goal : Proposition of new inhibitors for a family of proteins produced by *plasmodium falciparum*
  - Biomedical informatics goal : Deployment of *in silico* virtual screening on the grid
  - Grid goal : Deployment of a CPU consuming application generating large data flows to test the grid infrastructure and services.

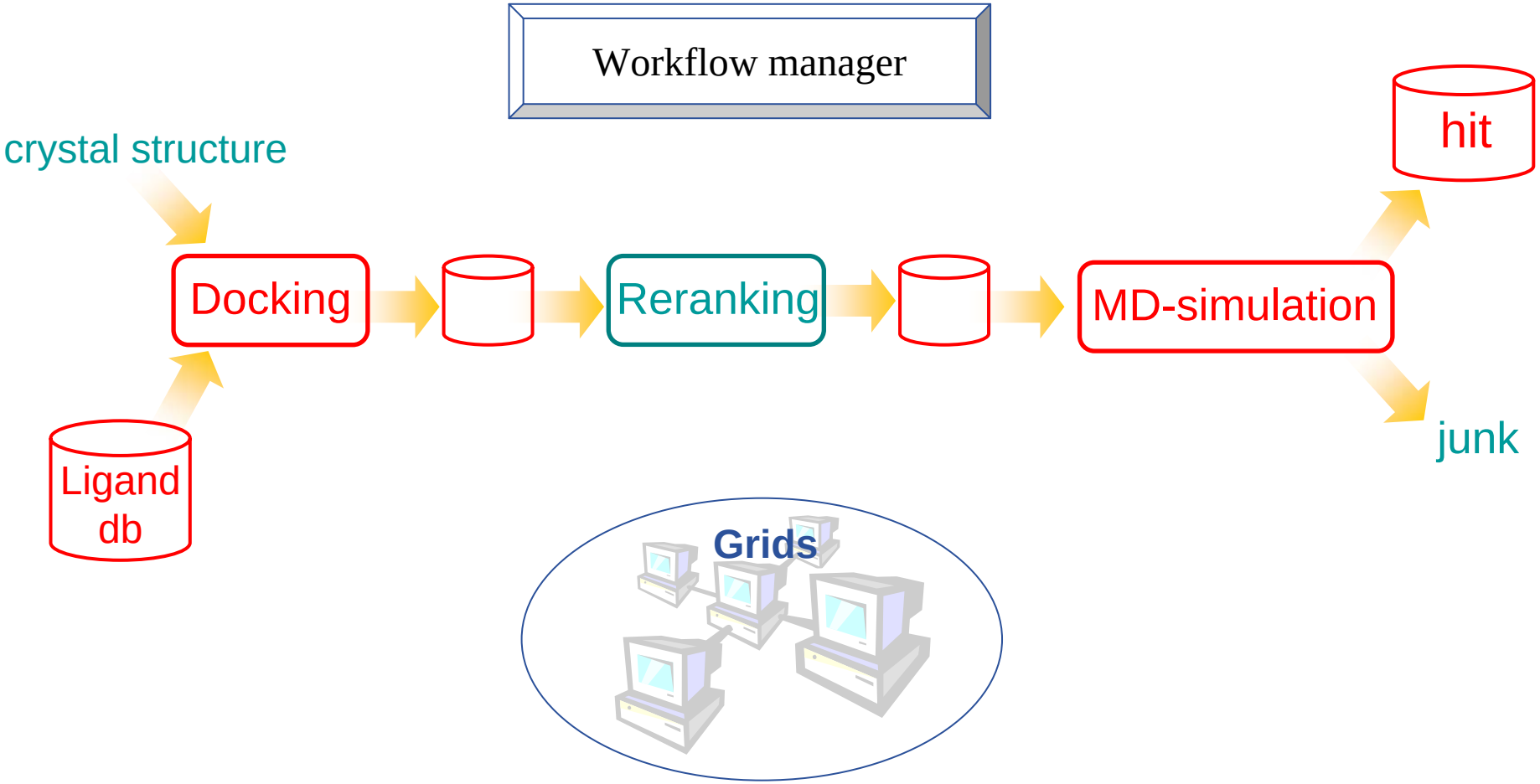
- **Partners**
  - Fraunhofer SCAI
  - CNRS/IN2P3
  - *CMBA (Center for Bio-Active Molecules screening)*

  representing different projects:
  - *EGEE (EU FP6)*
  - *Simdat (EU FP6)*
  - *Instruire and Campus Grid (French and German Regional Grids)*
  - *Accamba project (french ACI project)*



WISDOM workflow

- **Deployment of a virtual screening workflow on grid infrastructures**

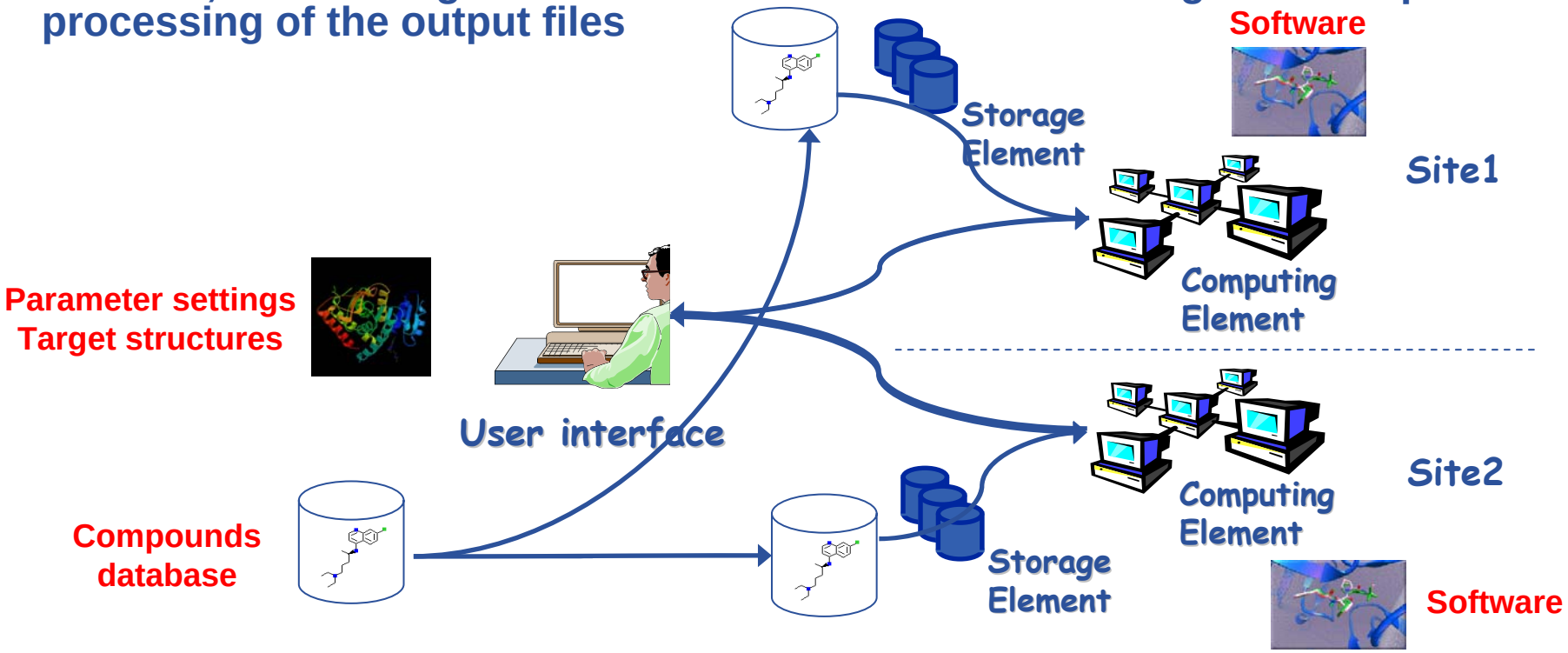


WISDOM elements

- **Biological information**
  - Plasmepsin is a promising aspartic protease target involved in the hemoglobin degradation of P. *falciparum*. 5 different structures are prepared (PDB source)
  - ZINC is an open source library of 3,3 millions selected compounds. They are made available by chemistry companies and are ready to be used
- **Biomedical informatics tools**
  - Autodock is free for academic, with grid based empirical potential and flexible docking via MC search and incremental construction
  - FlexX is licensed required, available for this data challenge during 1 week, with Boehm potential and fragment assembly energy function
- **Grid tools**
  - wisdom_env is an environment for an automatic, optimized and fault tolerance workflow using the grid resources and services
  - The biomedical VO will be the infrastructure with dedicated/no-dedicated resources

WISDOM : Deployment on a grid environment

- Docking is easily distributed once the compound database is available on the grid nodes. Each computing element computes docking probability for a different sample of ligands
- In a first step, docking scores are returned to the user and compared on its local machine.
- Later on, data management services can handle the storage and the post-processing of the output files



Results of the preliminary tests

- Docking application deployed since the summer 2004
- +30,000 jobs since January 2005
- Tests performed with the software Autodock on the biomedical VO

	100,000 compounds 500 jobs
Total CPU time for jobs	6 months CPU
User script time	40 h
Gain of time for the user	150
CPU time for 1 job	*9h*
Input and output transfer time between SE and CE for 1 job	*2.5 mn*
Waiting time for 1 job due to the grid	*30 mn*
Resubmitted Jobs *Aborted jobs %*	*16* *3%*

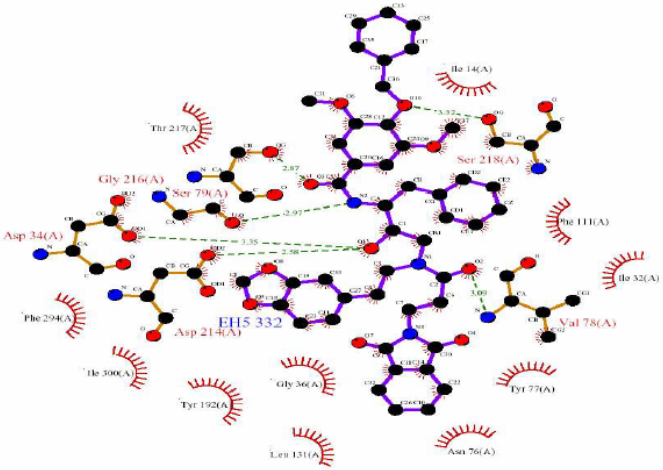
Data challenge scenario

	Scenario 1
Duration	3 weeks
CPU time	80 years CPU
Grid performance	70%
Number of CPU	2,000
Number of grid jobs (20h)	30,000
Storage	2*6 TB
Docking workflow description	
Number of compounds Number of parameters settings	500,000 4
Objective	Selection of the best hits with short analysis

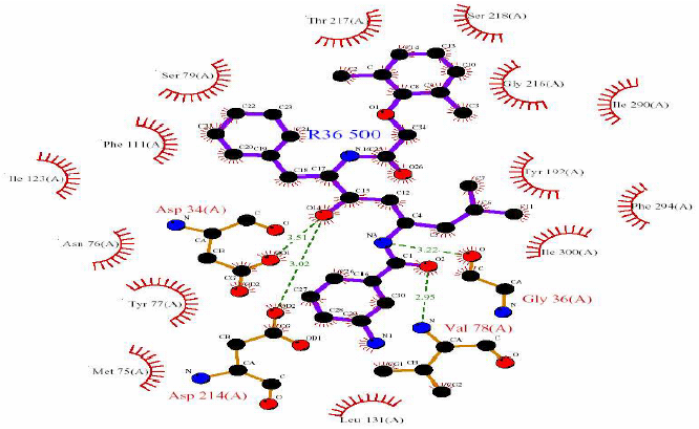
- FlexX running time : 1 mn
- F. output size : 1MB
- F. job output size : 1.2GB
- F. job compressed output size : 250MB
- Autodock running time : 2.5 mn
- A. output size : 1MB
- A. job output size : 0,5GB
- A. job compressed output size : 100MB

Output analysis (Fraunhofer)

- Post filtering
- Clustering of similar conformations
- Checking pharmacophoric points of each conformation
- Doing statistics on the score distribution
- Re-ranking for interesting compounds
- Sorting and assembly of data



Ligand plot of 1LF3 (plasmepsin II) with inhibitor EH5 332



Ligand plot of 1LEE (Plasmepsin II) with inhibitor R36 500

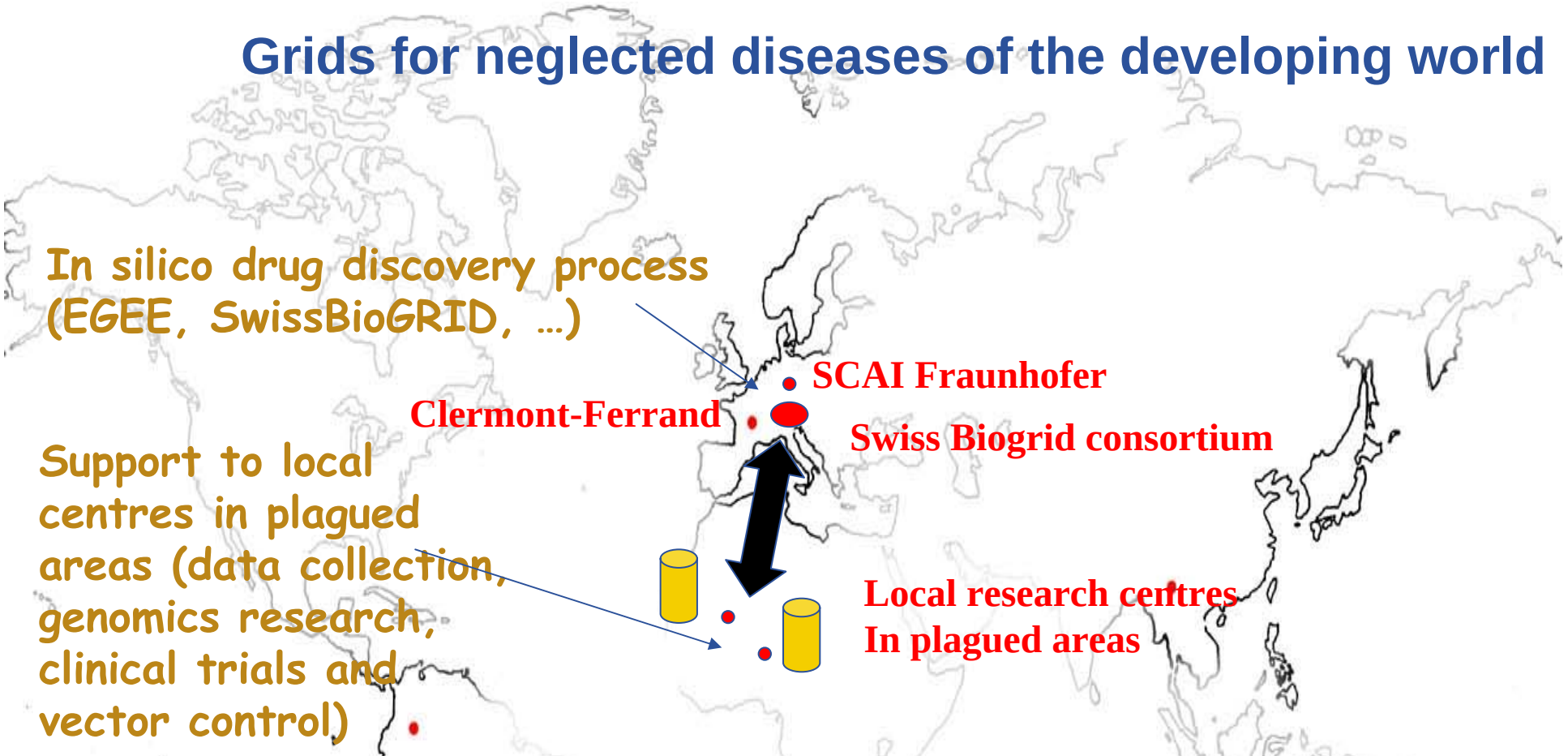
Follow-up of the DC

- **The best hits found by post-treatment will be published and available on a permanent grid storage via a portal**
  - Experimental screening of the most promising hits
- **A knowledge space will be progressively build around these results**
  - to extract and process the most interesting information
  - to enrich the data with the results found later by other *in silico* drug discovery processes
- **The *in silico* drug discovery will be further extend**
  - to include more precise molecular dynamics computations using quantum chemistry software like NAMD

From drug discovery to drug delivery

- **Drug discovery is about finding new drugs**
- **However, the best drugs are useful provided they are made available to the sick**
- **Drug delivery is a huge challenge for developing countries**
  - Lack of healthcare infrastructures
  - Lack of resources to buy drugs
  - Lack of education to deliver them
  - Lack of information on drug efficiency
- **For drug delivery, grids have a real added value**
  - To collect data in endemic areas
  - To provide data and tools to endemic areas (local reseach, training)

Grids for neglected diseases of the developing world



The grid impact :

- Computing and storage resources for genomics research and in silico drug discovery
- cross-organizational collaboration space to progress research work
- Federation of patient databases for clinical trials and epidemiology in developing countries

Grid federation of databases for epidemiology

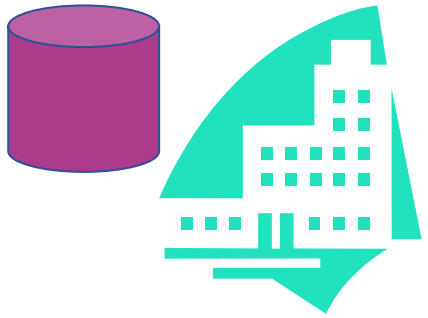
Analysis center
Country A

Epidemiology

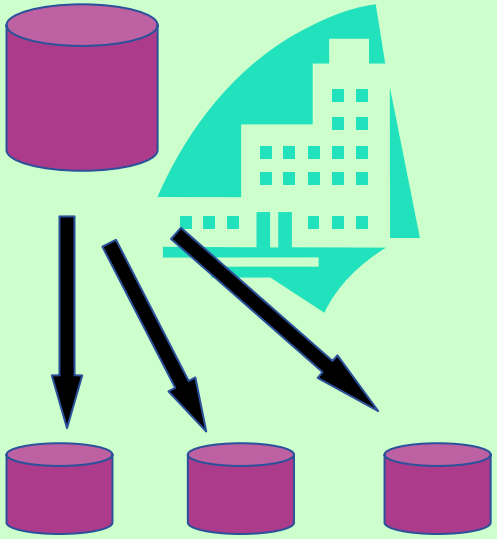
Added value:

- no central repository
- queries on federation of databases
- privacy protected
- telemedecine

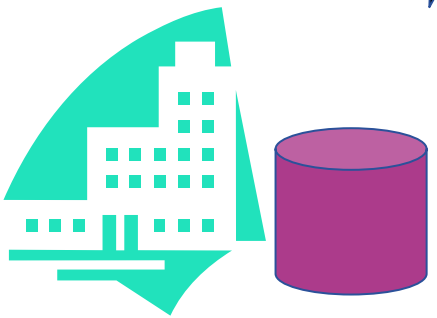
Hospital
Country B



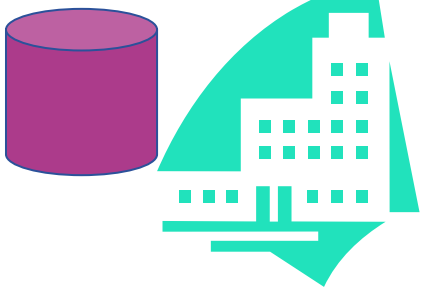
Hospital
Country E



Hospital
Country C



Hospital
Country D



Grid federation of databases for clinical trials

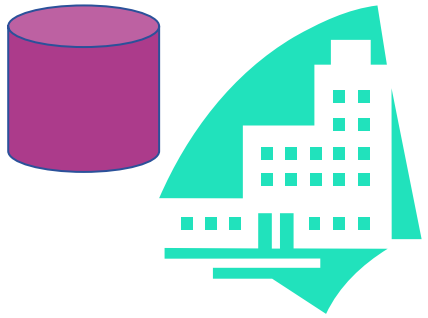
Pharmaceutical laboratory / International organization
Country A

Drug / Vaccine assessment

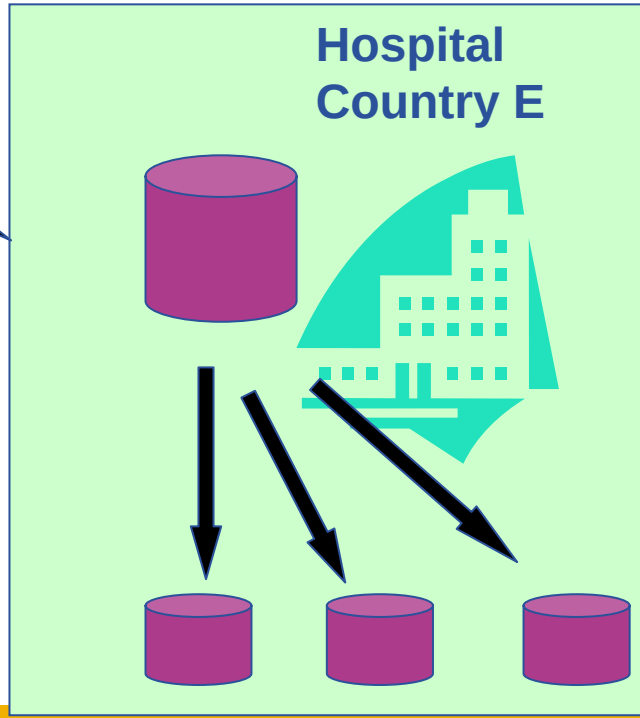
Added value:

- no central repository
- queries on federation of databases
- privacy protected-

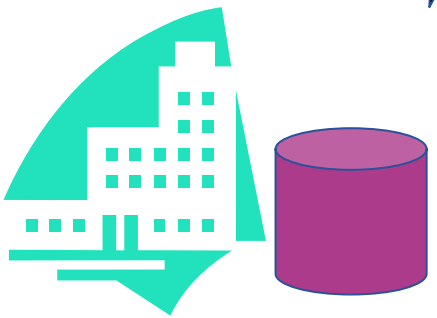
Hospital Country B



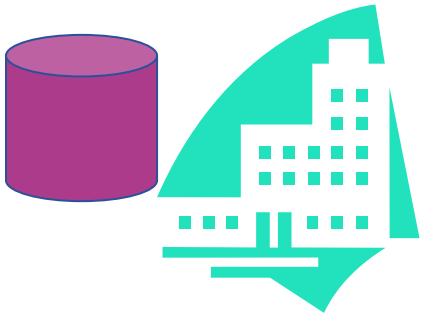
Hospital Country E



Hospital Country C



Hospital Country D



Projects starting on EGEE in relation to drug delivery and telemedecine

- **Grid enabled telemedecine for medical development**
  - Development of neurosurgery in poverty regions of western China
  - Ophthalmology in Burkina-Faso
    - Collaboration with Schiphra dispensary (Ouagadougou, Burkina Faso)

Grid-enabled telemedecine for medical development

Collaboration: NPO Chain of Hope, n°9 Hospital Shanghaï (neurosurgery unit), Chuxiong Hospital (Yunnan), CNRS-IN2P3, Clermont-Ferrand hospitals

Goal: improve patient follow-up by french clinicians

Method: grid-enabled telemedecine web application

Conclusion

- **Grid technologies promise to change the way organizations tackle complex problems by offering unprecedented opportunities for resource sharing and collaboration**
- **Grids should provide the services needed for *in silico* drug discovery**
- **Applied to world health development, grids should also**
  - Help monitor epidemics
  - Strenghthen R&D on neglected diseases
  - Grant easier access to eHealth
  -
- **We are looking for joint pilot projects with a pharmaceutical lab**
  - Develop a grid-enabled drug discovery pipeline for malaria
  - Build a federation of databases to address 1 infectious disease (epidemiology, clinical trials, vector control)
  - Study grid added value for drug delivery