BioinfoGRID
Bioinformatics Grid Application for life science

In silico docking against malaria: the WISDOM initiative

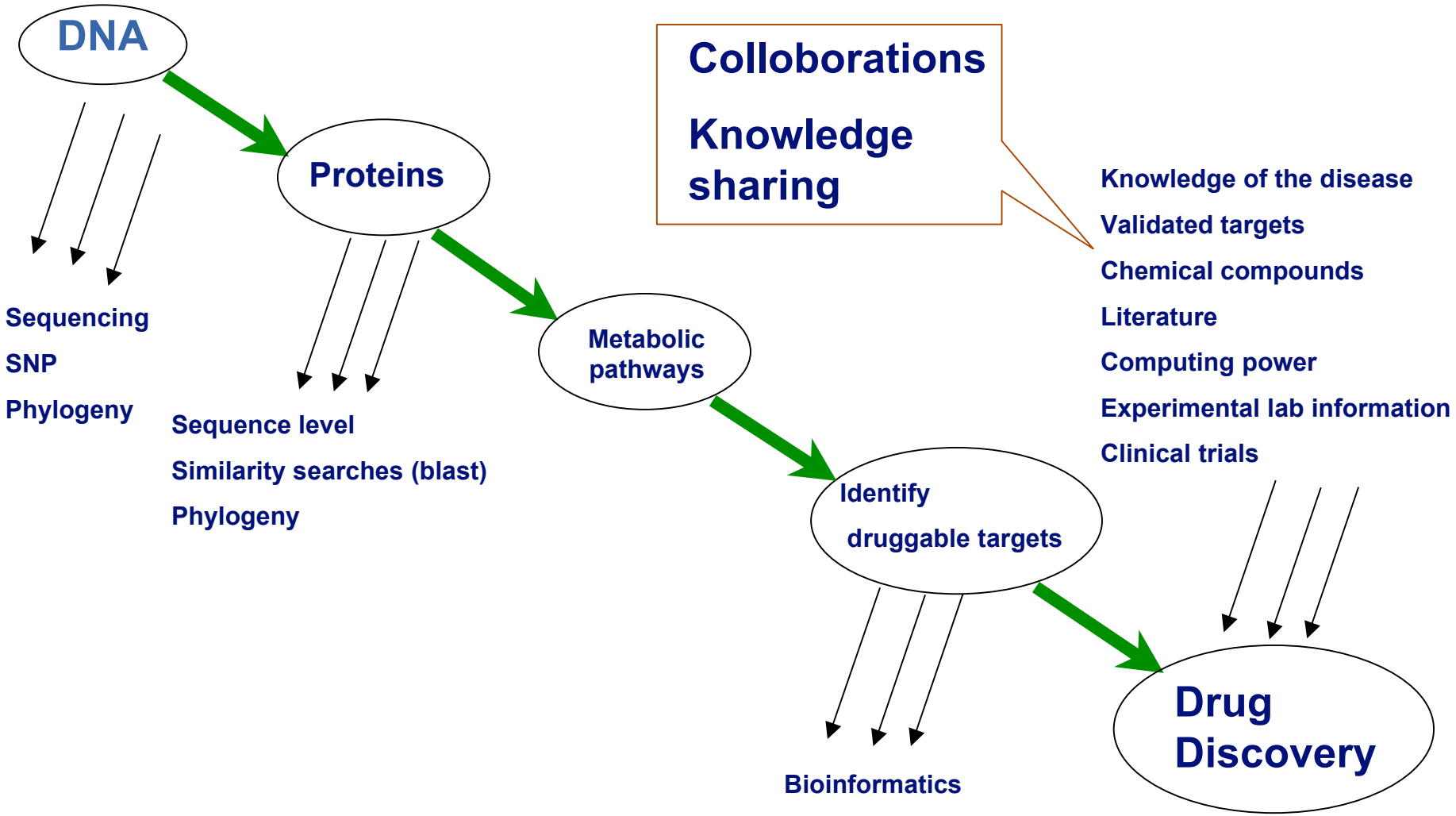
Presented by

Vinod Kasam

Bioinformatics Africa, 31 May 2007, Nairobi

Information Society and Media

Bioinformatics and Drug discovery

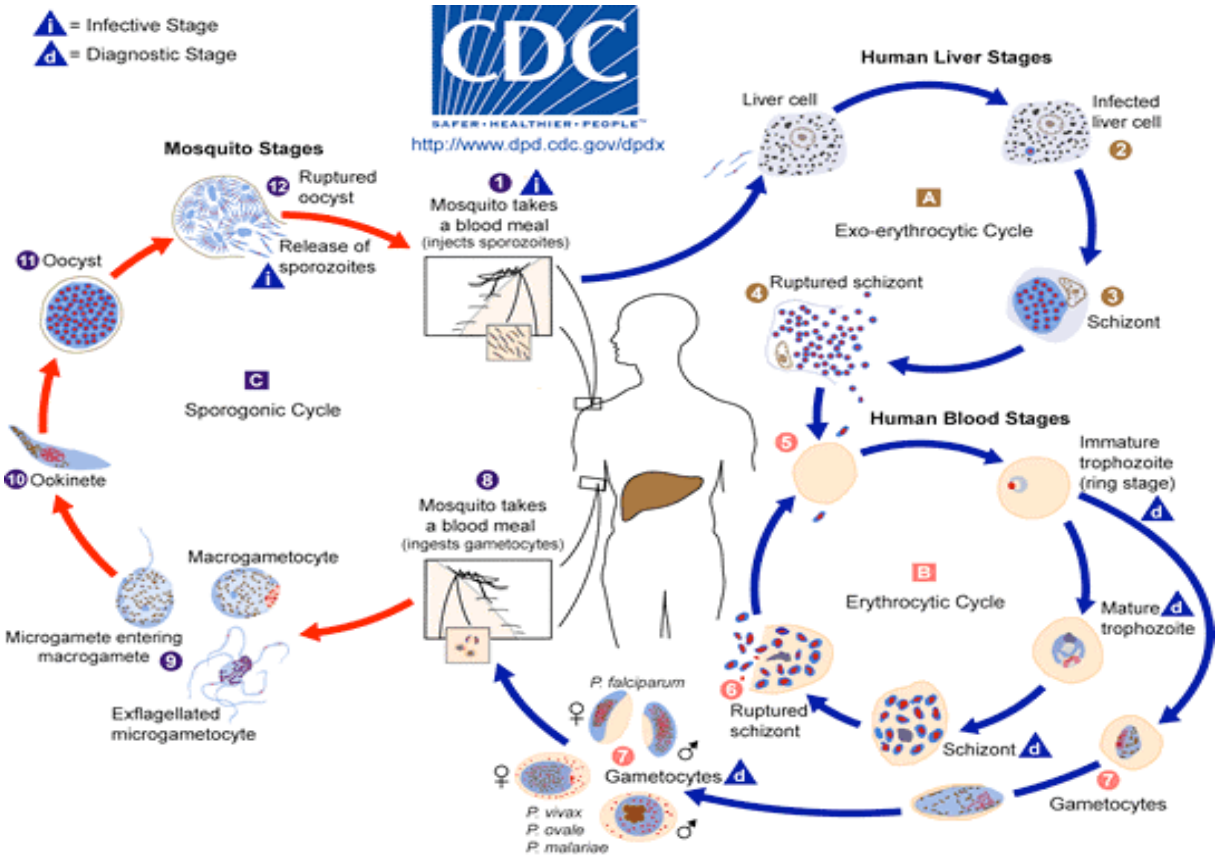


Outline

- Malaria
- Drug discovery and Screening
- Computational Grids
- WISDOM, Wide In silico Docking on Grid
- Resources used in Wisdom project
- Results
- Issues
- Conclusions
- Vision and long term vision

Introduction to the disease : malaria

- ~300 million people worldwide are affected
- 1-1.5 million people die every year
- Widely spread
- Caused by protozoan parasites of the genus *Plasmodium*



Complex life cycle with multiple stages

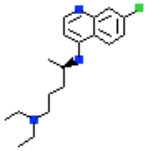
High Throughput Virtual Docking in WISDOM-II

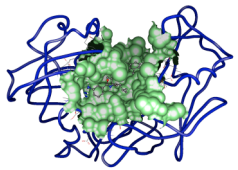
BioinfoGRID

Millions of chemical compounds available

High Throughput Screening
1-10$/compound. Very expensive

Chemical compounds
(ZINC database): 4.3 million
Chembridge ~300,000

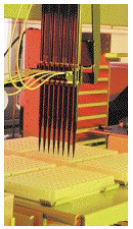




Targets (PDB): PvDHFR
PfDHFR, GST, tubulin

Molecular docking (FlexX)
~413 CPU years, 1.738 TB data
~100,000 dockings per minute

Data challenge on EGEE
~90 days on ~5000 computers



Hits screening using assays performed on living cells

- Leads
- Clinical testing
- Drug

WISDOM overview

◆ WISDOM project aims to build a collaboration platform for drug discovery using the Grid computing technology.

◆ This project intends to solve large-scale computation and data intensive scientific applications in the fields of drug discovery, Bioinformatics and Biology with the help of computational grids

◆ 4.3 million compound database with 3-D structure and physicochemical properties are screened against 4 different targets implicated in malaria to identify potential drug candidates.

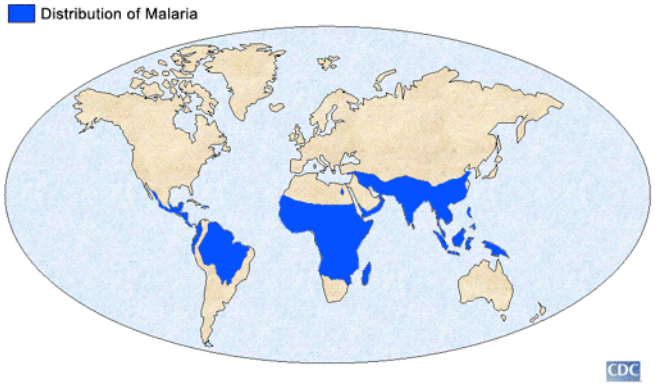
➢In WISDOM-I, on the biological side, three scaffolds have been identified against Plasmepsin and in vitro tests on the best compounds is under process on the best 30 compounds.

WISDOM : Wide In Silico Docking On Malaria

BioinfoGRID

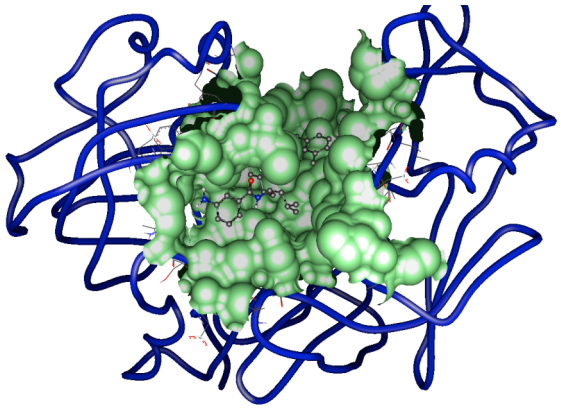
- Biological goal

  Proposition of new inhibitors for a different proteins produced by Plasmodium



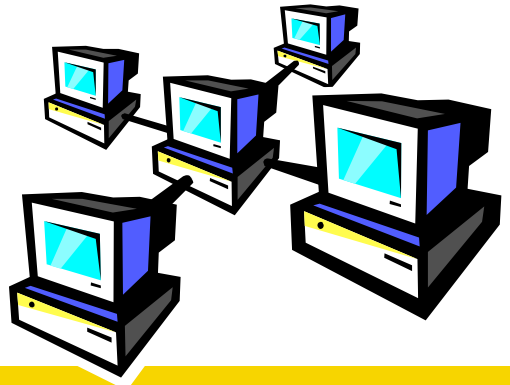
- Biomedical informatics goal

  Deployment of *in silico* virtual docking on the grid



- Grid goal

  Deployment of a CPU consuming application generating large data flows to test the grid operation and services => “data challenge”

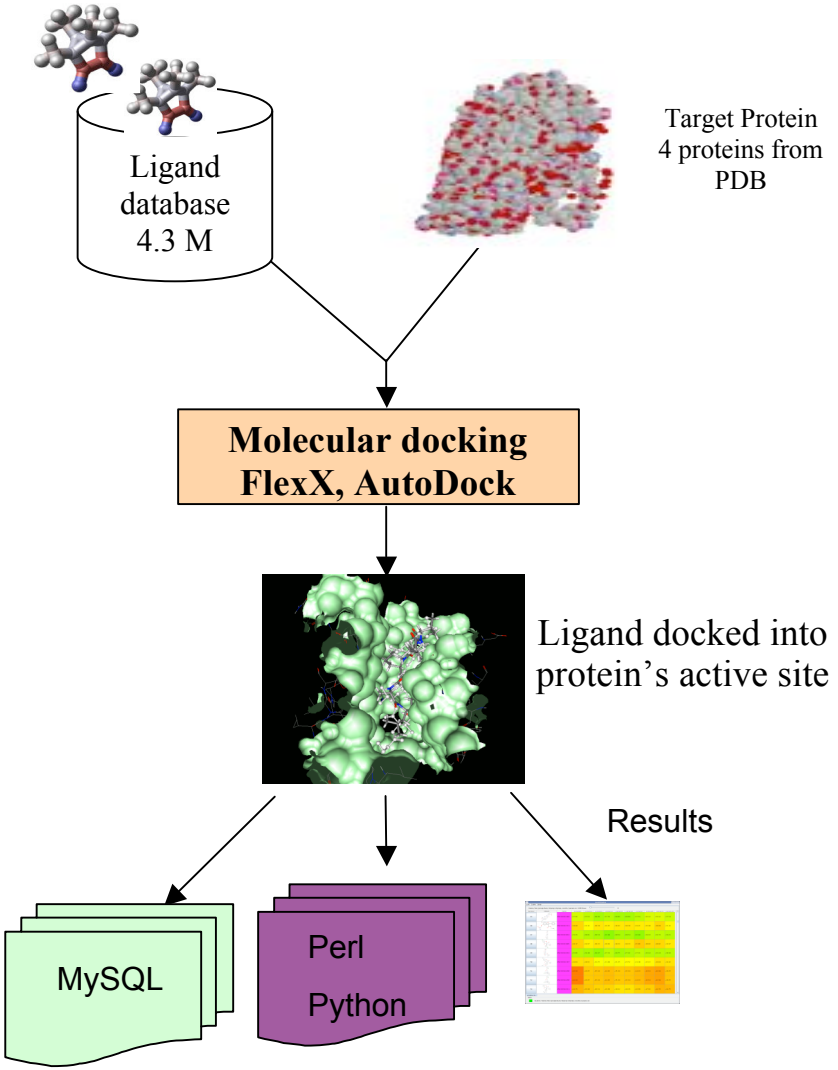


WISDOM-II - second large scale docking deployment

Malaria target	Involved in	Biology partners
GST from *Plasmodium falciparum*	Parasite detoxification	U. of Pretoria, South-Africa
DHFR from *Plasmodium vivax*	Parasite DNA synthesis	U. of Los Andes, Venezuela U. of Modena, Italia
DHFR from Plasmodium falciparum	Parasite DNA synthesis	U. of Modena, Italia
Tubulin from Plasmodium/plant/ mamal	Parasite cell replication	CEA, Acamba project, France

Materials and Procedure in WISDOM

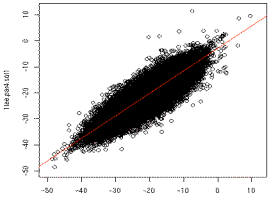
- Targets
  - Different targets from Protein database and homology models
- Chemical Compounds
  - ZINC compounds
- Docking tools
  - FlexX
  - AutoDock
- Grid Infrastructure
  - EGEE, EELA,EUChinaGrid
- Results
  - Python and Perl scripts
  - VS explorer
  - MySQL databases



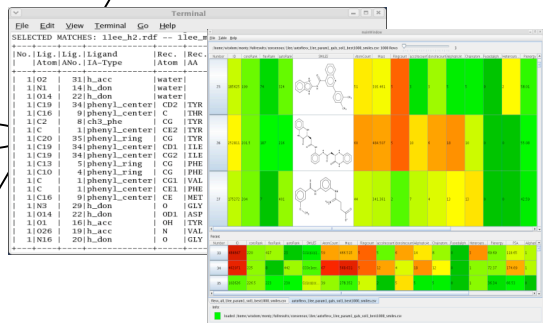
Filtering process employed in WISDOM-I

1,000, 000 chemical compounds

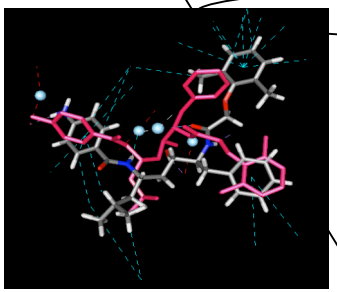
Sorting based on scoring in different parameter sets;
Consensus scoring



10,000 compounds selected

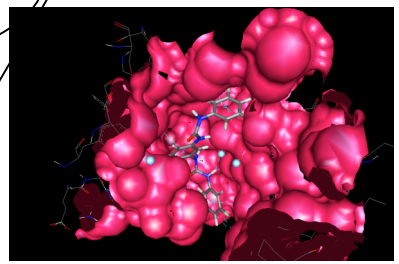


Based on key interactions



1,000 compounds

Key interactions, binding
modes, descriptors,
knowledge of active site



100 compounds

MD

30 compounds to be tested in experimental lab

Grid Computing

- A grid is the combination of networked resources and the corresponding middleware, which provides services for the user

- Grids are unique tools for
  - **Collecting and sharing information**
  - **Networking experts**
  - **Mobilizing resources routinely or in emergency**

The different kind of grids

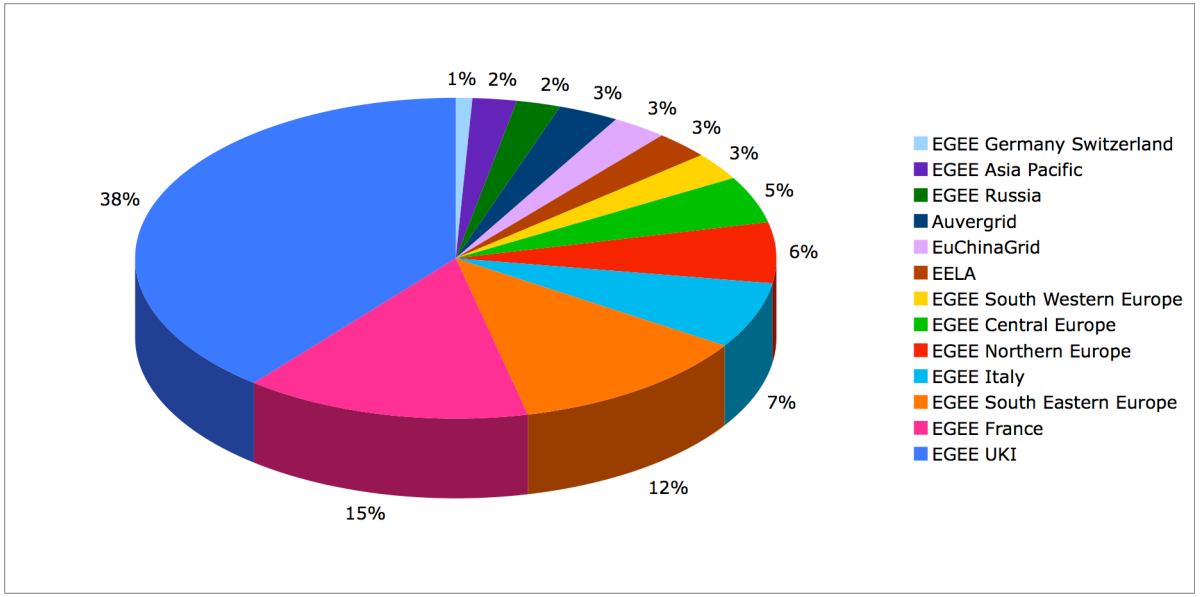
- Computing grid
  - Distributed processors
  - Services to submit jobs, to collect results
  - Impact: in silico search for new drugs or vaccines
- Data grid
  - Distributed data: databases, flat files
  - Services to collect, query, move and analyze the distributed data
  - Impact: collection and sharing of medical data
- Knowledge grid
  - Knowledge space using ontology to manipulate concepts and run complex in silico experiments
  - Impact: integration of “wet” laboratories in a collaboration space

Instances on different infrastructures

BioinfoGRID

Target Structures	Number of instances deployed
GST (A chain)	4 on EGEE
GST (B chain)	4 on EGEE
2BL9 (P. vivax wild type DHFR)	3 on EGEE, 1 on EELA
2BLC (P. vivax double mutant DHFR)	3 on EGEE, 1 on Auvergrid
Dm_vivax (P. vivax DHFR 2BLC minimized)	4 on EGEE
Wt_vivax (P. vivax DHFR 2BL9 minimized)	4 on EGEE
1J3K (P. falciparum Quadruple mutant DHFR)	4 on EGEE
1J3I (P. falciparum Wild type DHFR)	3 on EGEE, 1 on EuChinaGrid

Deployment on different infrastrucures
Distribution of jobs



Statistics of deployment

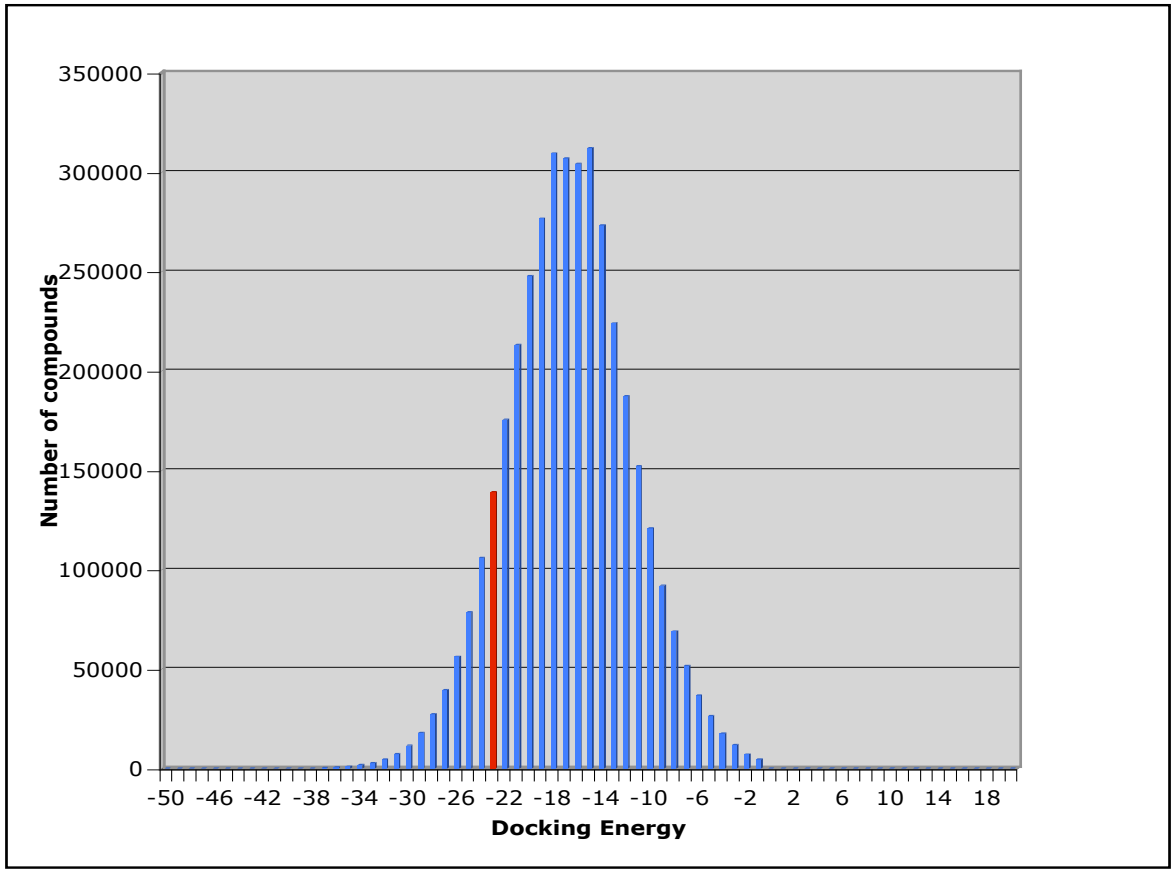
- First DC:
  - 80 CPU years
  - 1 TB
  - 1700 CPUs used in parallel
  - July 1st - August 15th 2005
- 2nd DC
  - 100 CPU years
  - 800 GB
  - 1700 CPUs used used in parallel
  - April 1st - May 15th 2006
- 3rd DC
  - 413 CPU years
  - 1.7 TB
  - Up to 5000 CPUs in parallel
  - 1st October 2006 - 31 January 2007

Number of Jobs	77,504
Total Number of completed dockings	156,407,400
Estimated duration on 1 CPU	413 years
Duration of the experiment	76 days
Average throughput	78,400 dockings/hour
Maximum number of loaded licences (concurrent running jobs)	5,000
Number of used computing elements	98
Average duration of a job	41 hours
Average crunching factor	1,986
Volume of output results	1,738 TB

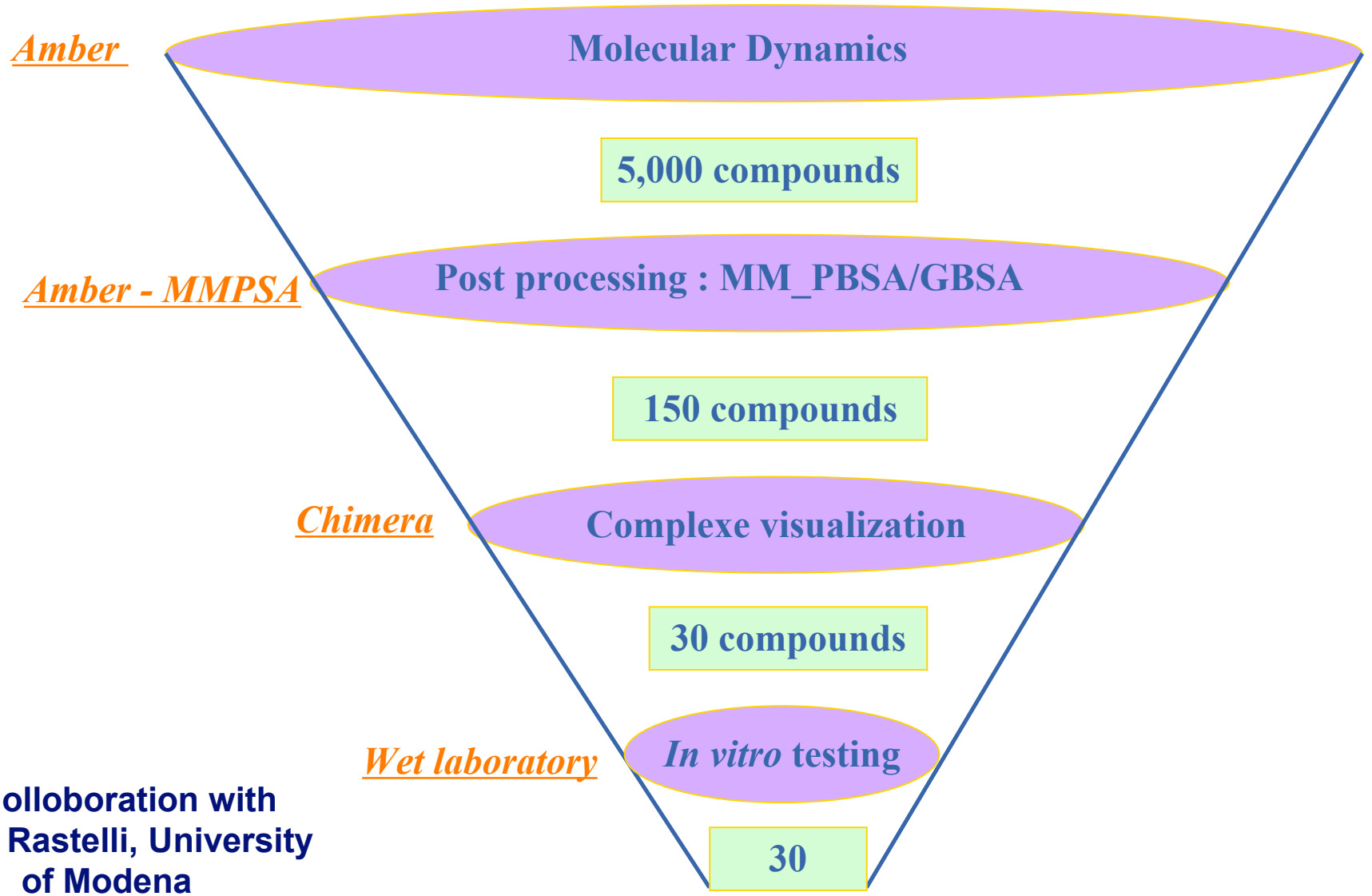
Biological results

The repartition of docking energies of the ZINC database against GST A structure.

(The red column represents a score of -24kj/Mol, the docking score of a co-crystallized ligand (GTX) of GST A chain)



Molecular dynamics Workflow



In Colloboration with Giulio Rastelli, University of Modena

Where grids can help medical development in Africa

- 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 drug delivery and vector control
  - Improve epidemics warning and monitoring system

- Improve the ability of African countries to undertake health innovation
  - Strengthen the integration of African life science research laboratories in the world community
  - Provide access to resources
  - Provide access to bioinformatics services

Grid added value

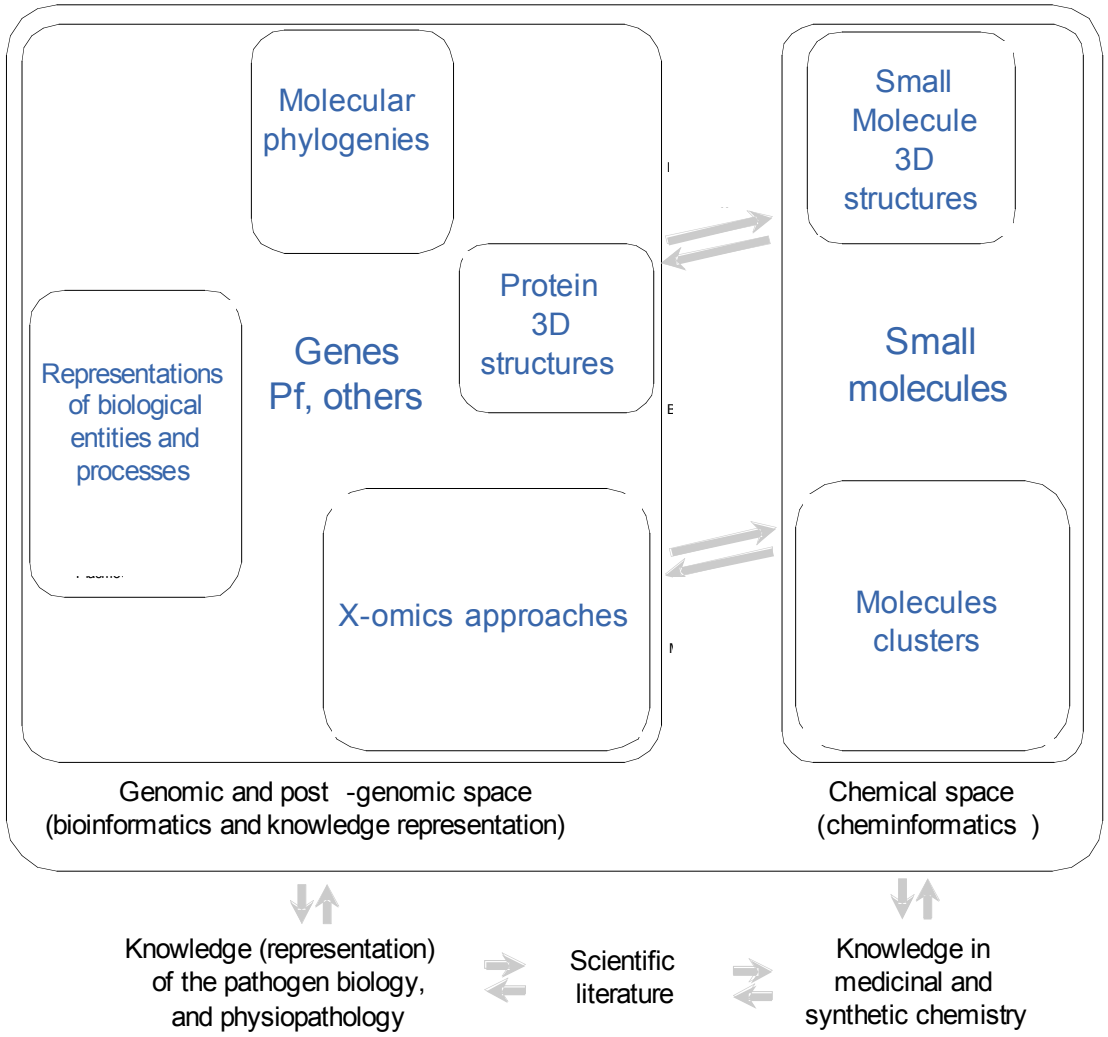
- Grids offer unprecedented opportunities for resource sharing and collaboration

- Grids open exciting perspectives to handle the information flows needed to fight neglected diseases
  - Deployment of services for healthcare and research centers in endemic regions
  - Deployment of infrastructures (federation of databases) to collect biomedical data and improve disease monitoring
  - Cross-organizational collaboration space to share data and resources

- Challenges
  - Infrastructure capacity building in Africa
  - Grid technology must provide the services for data and knowledge management
  - IT expertise and willingness to share information is needed from the participating healthcare centers

Beyond virtual screening, a chemogenomic space for malaria

Chemogenomic knowledge space



Goals:

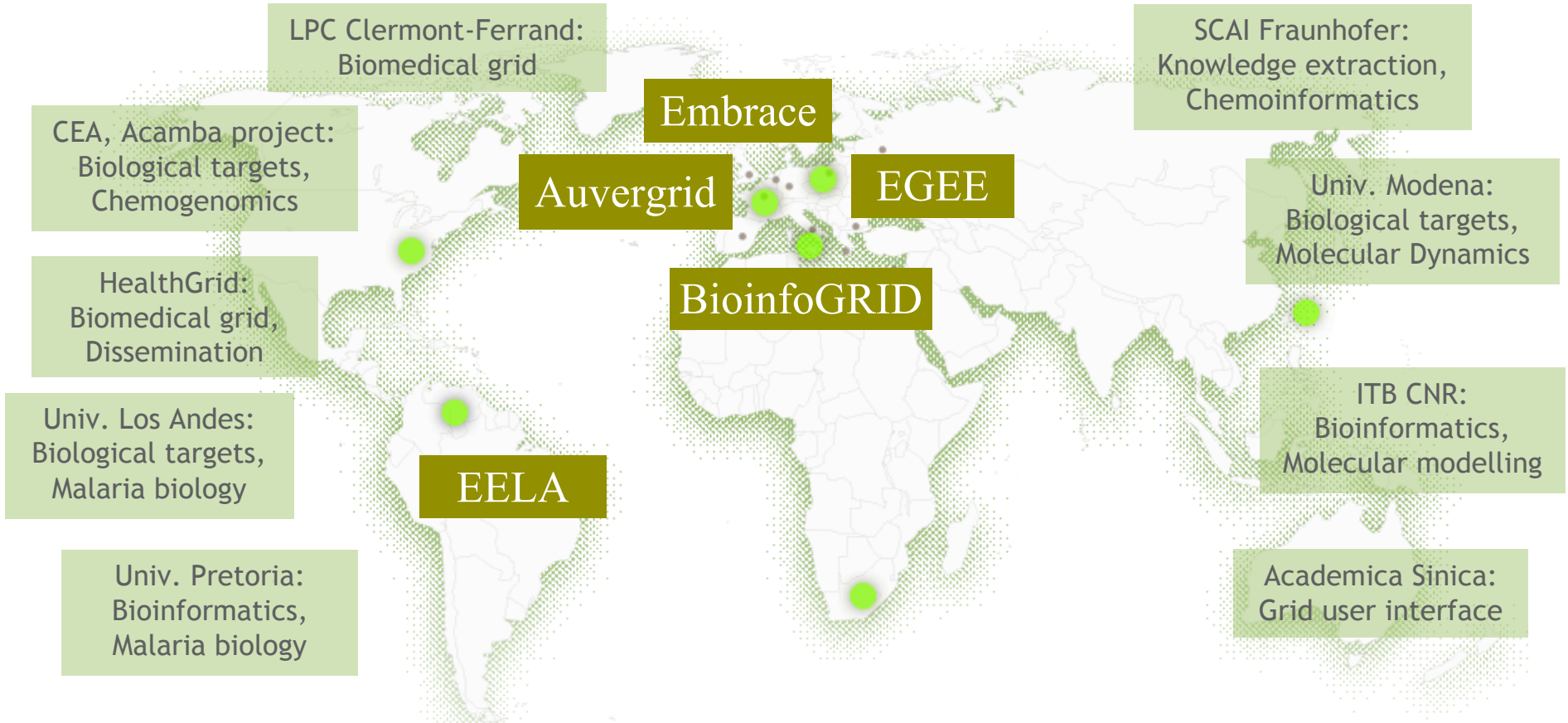
- comparison of protein sequences
- high throughput reconstruction of molecular phylogeny
- representation of biological processes particularly metabolic pathways
- integration of genomic data, biological representations and functional profiling after drug treatments
- determination and prediction of protein structures
- virtual docking with drug candidate structures

Source: Birkholtz L.-M. et al., Malaria Journal, 2006

Conclusion

- WISDOM proposes a new approach to drug discovery thanks to the grid
  - Rapid deployment of very large scale virtual screening
  - Collaborative environment for the sharing of data in the research community
- WISDOM fully exploits EGEE services and resources.
  - AMGA allows to store securely results and statistics immediately
  - Web Service Interface using WS-I profile guarantees interoperability
- First biochemical results demonstrate grid relevance to the drug discovery community
  - Grid is a superior tool to discover new drugs

Long term vision: 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,

Perspectives on Malaria

- EGEE infrastructure open to host other CPU intensive applications relevant to research on Malaria i.e.
  - Search for drugs: virtual screening
  - Search for vaccines: data analysis

- Grids offer also unique opportunities for
  - Early detection
  - Epidemiological watch
  - Prevention
  - International collaboration

- Contact: kasam@clermont.in2p3.fr
  breton@clermont.in2p3.fr

Acknowledments

Academia Sinica
BioSolveIT
CNR-ITB
CNRS
CEA
Healthgrid
IN2P3
LPC
SCAI Fraunhofer
Università di Modena e Reggio Emilia
Université Blaise Pascal
University of Pretoria
University of Los Andes

Auvergrid
Accamba
BioInfoGRID
EGEE
EMBRACE
EUChinaGRID
EUMedGRID
SHARE
TWGrid

Conseil Regional d'Auvergne
European Union

wisdom.healthgrid.org