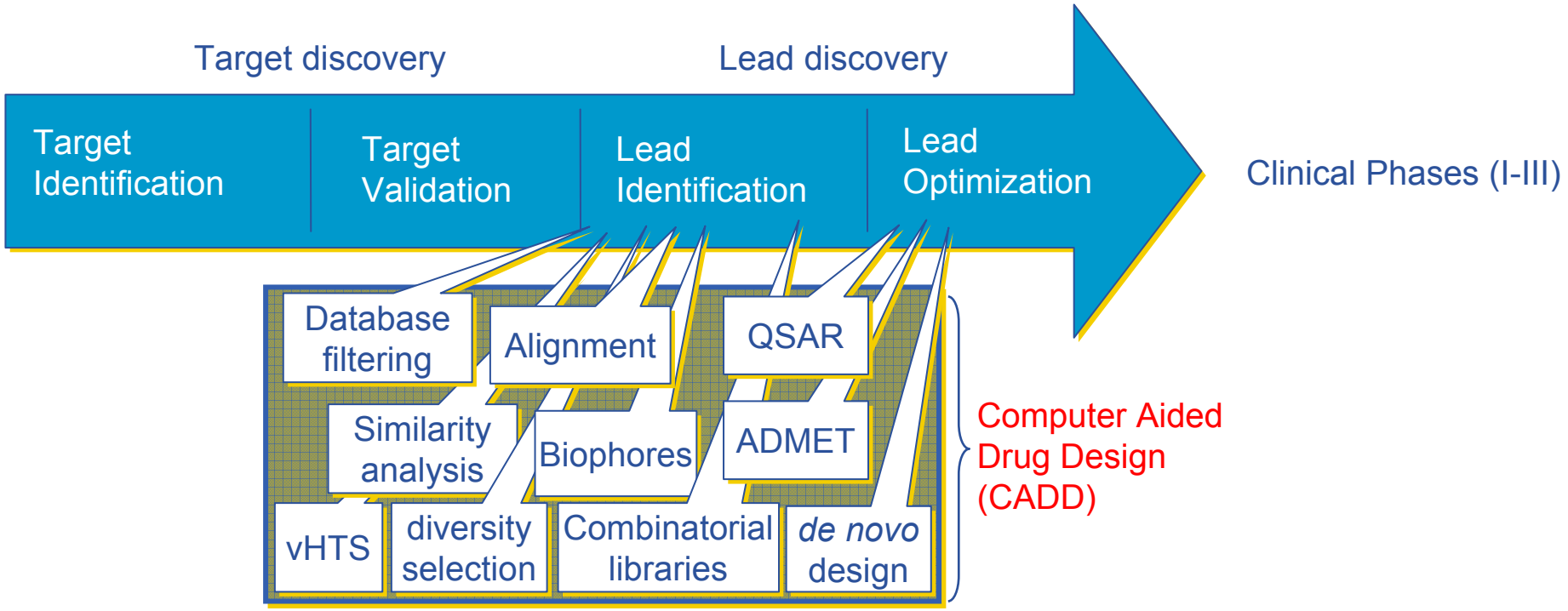
Grid-enabled drug discovery to address neglected diseases

N. Jacq – CNRS-IN2P3

EGAAP meeting - Athens 21 April 2005

Objective

- **Demonstrate the relevance and the impact of the grid approach to address Drug Discovery for neglected diseases.**

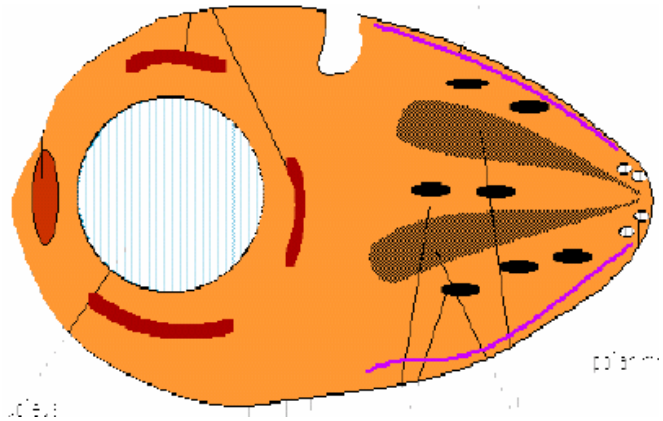


Duration: 12 – 15 years, Costs: 500 - 800 million US $

Use case

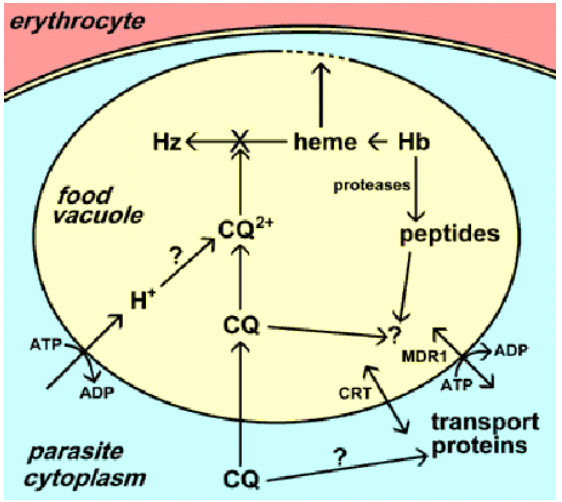
- **Propose new inhibitors for the targets implicated by malaria and dengue by using a docking approach on the GRID.**

Plasmodium structure



- Well known organism
- Multiple crystal structures
- Multiple bound inhibitors
- Structural similarity between multiple species

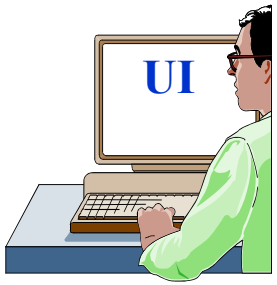
Action mode for chloroquine



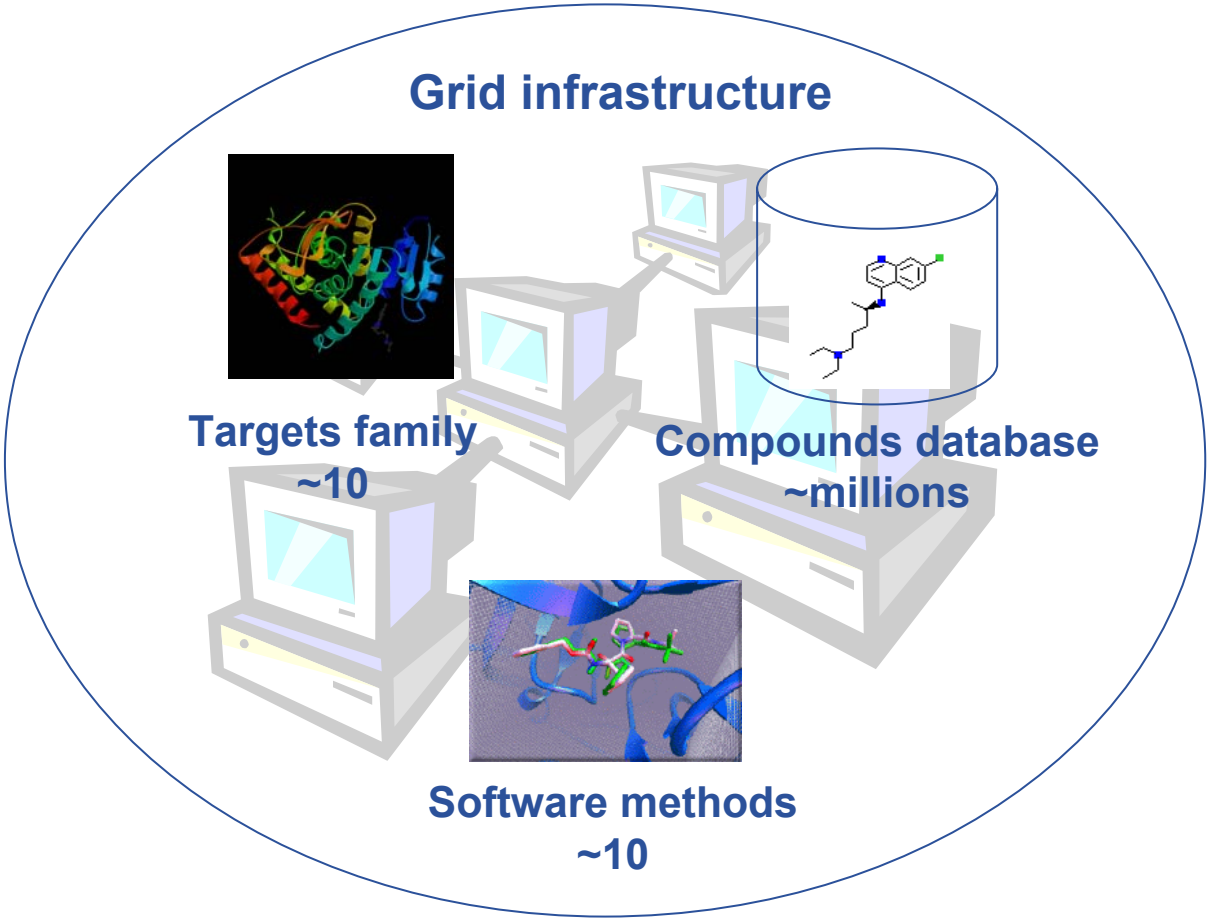
- The one more selective
- Acts on multiple targets
- The one with active in low quantities
- Shows good pharmacokinetics properties
- Good pharmacodynamic properties

Docking platform components

- **Predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure**



Parameter / scoring settings



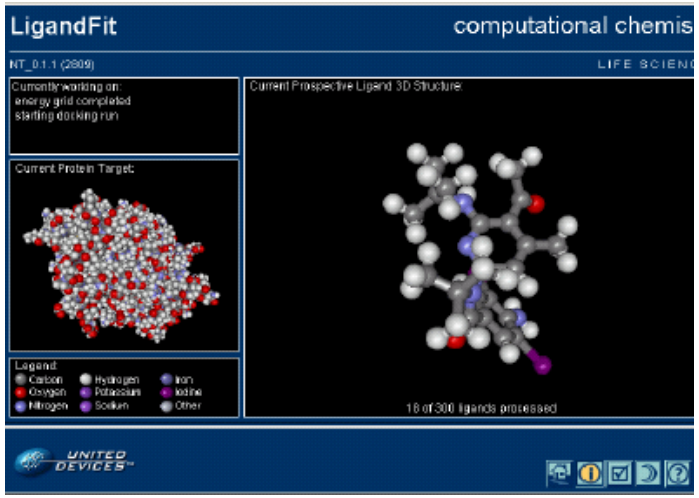
Pervasive grid on docking (1/2)

- **Grid.org**
  - Global grid of United Device
  - World's largest computational grid dedicated to life science research
  - More than 3 million registered computers
    - people's home computers
    - computers from numerous universities
    - a large number of corporations
  - Grid computing projects on docking to screen 35 million of potential drugs (Computational Chemistry of University of Oxford) against several protein targets
  - Reducing the time required to develop a commercial drug

Pervasive grid on docking (2/2)

- **Anthrax Research project (2002/02)**
  - Realised in 24 days instead of years
  - 300,000 ranked hits to be refined and analysed
  - Intel, Microsoft

- **Smallpox Research Grid (2004/11)**
  - For post-infection anti-viral agents to counter smallpox infections resulting from bioterrorism
  - 39000 years/CPU for 8 targets
  - US Department of Defence, Accelrys, IBM



- **Cancer Research Grid (2004/11, phase 1)**
  - 1 target / 400 hits selected for the phase 2
  - 2-4% of hits real activity > 0.1% expected by pharmaceutical industry from in silico screening
  - National Foundation for Cancer Research, Accelrys

Pervasive Grids on life science research

- **World Community Grid**
  - new resource sponsored by IBM for massive-scale research projects of global significance
- **Human proteome folding project**
  - Collaboration between Grid.org and World Community Grid
  - Predicting the protein structures based on known Human Genome sequence data
    - Examining the entire human genome could require up to 1,000,000 years of computational time on an up-to-date PC.
    - Using a commercial 1000 node cluster would require 50 years and, while faster, would still be impractical.
  - Institute of Biology Systems, University of Washington, IBM
- **Decrypthon**
  - AFM (French Muscular Dystrophy Association), CNRS, IBM
  - A pervasive grid, with people's home computers (United Devices)
  - A supercomputers grid, with 3 French universities (not defined technology)
  - Genomics pilot applications

Added value of clusters' grid

- Perennially
- Permanent availability => 7/7, 24/24, user support
- Robustness, reliability => Experiments reproducibility
- Flexibility
- Security
- Confident results

A grid-enabled docking service

- First wide *in silico* docking platform on a production infrastructure
- Deployment of a bioinformatic service for diseases (dengue, rare diseases…)
- Proof of concepts with malaria use case
- Data challenge for the scalability

First deployment

- **Malaria targets sent by the inputSandbox**
  - Lactate dehydrogenase (Energy production, inhibited by chloroquine)
  - Default parameter / scoring settings

- **Compounds database deployed on each SE of biomedical VO**
  - NCI, National Cancer Institute compounds database
    - 2000 compounds
  - Ambinter, subset of ZINC : a free database of commercially-available compounds for virtual screening
    - 416 000 compounds, 3GB

- **Docking software**
  - Autodock : automated docking of flexible ligands to macromolecules
    - ~2,5 mn by target – compound job
    - Sent on each CE of the biomedical VO
  - FlexX : commercial prediction of protein-compound interactions
    - ~1mn by target – compound job
    - Available on SCAI node, soon on LPC node

Submission to EGEE

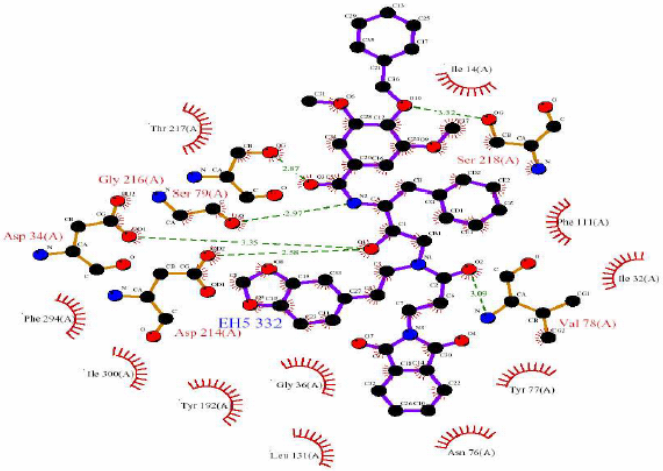
- **Tests**
  - RBs
  - CEs
  - SEs
- **Deployment**
  - software
  - database
- **Submission**
  - Automatic
  - Optimization
  - Fault tolerance
  - Statistics report
  - Results
- **35 submitted tickets to the Global Grid User Support since January**

Performance results

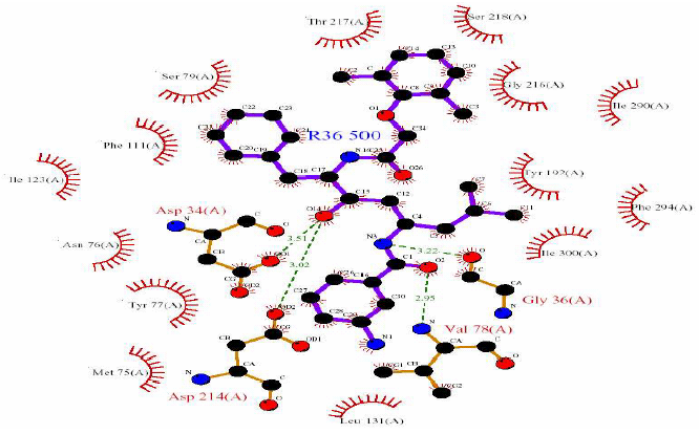
	1 target vs 2000 compounds – 50 jobs	1 target vs 100 000 compounds – 500 jobs (begin of April)
Total CPU time for jobs	2,5 days	188 days
User script time	2,5 h	40 h
Gain of time for the user	25	150
CPU time for 1 job	*1,2 h*	*9h*
Input and output transfer time between SE and CE for 1 job	*< 1mn*	*2,5 mn*
Waiting time for 1 job due to the grid	*7,2 mn*	*30 mn*

Output analysis

- 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

Data challenge during the summer

- **5 different structures of the most promising target**
  - Plasmepsin II, aspartic protease, involved in the hemoglobin degradation of *Plasmodium*
  - Structures under preparation
- **ZINC**
  - 3,3 million compounds, ~25 GB
  - To be deployed on each SE
- **Autodock**
  - ~80 years/CPU
  - ~35 000 jobs of 20h
  - To be deployed on each CE
- **Output Data**
  - 16,5 million results, ~10 TB
  - Will be stored on SEs

Collaboration

- **Fraunhofer SCAI**
  - Martin Hofmann
  - Marc Zimmermann
  - Kai Kumpf
  - Horst Schwichtenberg
  - Astrid Maass
- **CNRS/IN2P3**
  - Vincent Breton
  - Nicolas Jacq
  - Jean Salzemann
- **Biozentrum Basel**
  - Torsten Schwede
  - Michael Podvinec
  - Konstantin Arnold
- **CSCS**
  - Marie-Christine Sawley
  - Patrick Wieghardt
  - Sergio Maffioletti