The WISDOM initiative
Wide In Silico Docking On Malaria

Yannick Legré, CNRS/IN2P3
on behalf of the WISDOM Consortium

Slides credit: N. Jacq & V. Breton, CNRS-IN2P3
Y-T Wu & H-C Lee, Academia Sinica

www.eu-egee.org
Content

• Presentation of the WISDOM initiative
• Need for new drugs to fight malaria
• Challenges of the High Throughput Docking
• Development of the grid environment for a large-scale deployment
• Achieved deployment on EGEE infrastructure
• **Biological goal**
  Proposition of new inhibitors for a family of proteins produced by Plasmodium *falciparum*

• **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”
• **Partners**
  - Fraunhofer SCAI, Germany (Project PI: Martin Hofmann)
  - LPC Clermont-Ferrand, France (CNRS/IN2P3)
  - CMBA, France (Center for Bio-Active Molecules screening)
  - BioSolveIT
  - HealthGrid

• **Representing different projects:**
  - EGEE (EU FP6)
  - Simdat (EU FP6)
  - AuverGrid (French Regional Grid)
  - Accamba project (French ACI project)
Introduction to the disease: malaria

- ~300 million people worldwide are affected
- 1-1.5 million people die every year ➔ 1 person each 20 seconds !!!
- Widely spread
- Caused by protozoan parasites of the genus *Plasmodium*

Complex life cycle with multiple stages
There is a real need for new drugs to fight malaria (WHO)

• Drug resistance has emerged for all classes of antimalarials except artemisinins.
  – Resistance to chloroquine, the cheapest and the most used drug, is spreading in almost all the endemic countries.
  – Resistance to the combination of sulfadoxine-pyrimethamine which was already present in South America and in South-East Asia is now emerging in East Africa (65% in Western Tanzania)

• All countries experiencing resistance to conventional monotherapies should use ACTs (artemisinin-based combination therapies)

• But there is even the threat of resistance to artemisinin too, as it is already observed in murine Plasmodium yoelii
Identification of new malarial targets

- The available drugs focus on a limited number of biological targets => cross-resistance to antimalarials
- There is a consensus that substantial scientific effort is needed to identify new targets for antimalarials
- With the advent of the plasmodium genome, many targets came into light
- The potential antimalarial drug targets are broadly classified into three categories, and each category has many individual targets.
  - Targets involved in human hemoglobin degradation (proteases)
  - Targets involved in parasite metabolism (Folate, phospholipid… )
  - Targets engaged in parasite membrane transport and signalling (choline carrier etc).
Plasmepsins role in human hemoglobin degradation

- Plasmepsins are involved in the hemoglobin degradation inside the food vacuole during the erythrocytic phase of the life cycle.

- The sequence homology between the plasmepsins is high (65-70%).

- The sequence homology with its nearest human aspartic protease is fortunately low (35%).

- Presence of X-crystallographic data in Protein Data Base.
Molecular Docking: Predict how small molecules, such as substrates or drug candidates, bind to a receptor of known 3D structure.

Target discovery
- Target Identification
- Target Validation

Lead discovery
- Lead Identification
- Lead Optimization

Clinical Phases (I-III)

**in vitro** experimentation

**in silico** experimentation

Duration: 12 – 15 years, Costs: 500 - 800 million US $
High Throughput Virtual Docking

**Chemical compounds (ZINC):**
- Chembridge – 500,000
- Drug like – 500,000

**Targets (PDB):**
- Plasmepsin II (1lee, 1lf2, 1lf3)
- Plasmepsin IV (1ls5)

**High Throughput Screening**
- 1-10$/compound, nearly impossible

**Molecular docking (FlexX, Autodock)**
- ~80 CPU years, 1 TB data

**Data challenge on EGEE**
- ~6 weeks on ~1700 computers

**Hits screening using assays performed on living cells**

**Leads**

**Clinical testing**

**Drug**
Molecular docking and modeling

- **Target scenarios**
  - number of water molecules in the active site

- **Software scenarios**
  - Docking methods (Autodock)
  - Water molecules place and max overlapping volume (FlexX)

- **Compounds preparation**
  - Yet drug like
  - Hydrogens added

- **Target preparation**
  - X-ray crystal structures of 5 plasmepsins (PDB)
  - Active site created from native crystal ligand
The WISDOM application, ISGC 2006 – Taipei – May 1st – 4th, 2006

EGEE, international project of grid infrastructure

- Started in 2004, >70 partners in the world
- Project leader: CERN
- 7 scientific domains with >20 applications deployed
- ~200 grid nodes, ~20,000 CPUs, several PetaBytes of data, 10,000 concurrent jobs

Countries with nodes contributing to the data challenge WISDOM
- **FlexX license server**:  
  - 3000 floating licenses offered by BioSolveIT to SCAI  
  - Maximum number of concurrent used licenses was 1008
Objective of the WISDOM development

• Objective
  – Producing a large amount of data in a limited time with a minimal human cost during the data challenge.

• Need an optimized environment
  – Limited time
  – Performance goal

• Need a fault tolerant environment
  – Grid is heterogeneous and dynamic
  – Stress usage of the grid during the DC

• Need an automatic production environment
  – Execution with the Biomedical Task Force
  – Grid API are not fully adapted for a bulk use at a large scale
Deployment preparation on AuverGrid, a French regional project

- Started in 2005 for 3 years
- Interconnecting the main laboratories of the Auvergne region using EGEE middleware
- Share technologies, competences and resources

<table>
<thead>
<tr>
<th>Metrics</th>
<th>100,000 docking runs in 500 jobs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total CPU time</td>
<td>188 days (6.3 months)</td>
</tr>
<tr>
<td>Duration</td>
<td>40 hours</td>
</tr>
<tr>
<td>Crunching factor</td>
<td>150</td>
</tr>
<tr>
<td>CPU time for 1 job</td>
<td>9 hours</td>
</tr>
<tr>
<td>Grid overhead time for 1 job</td>
<td>30 minutes</td>
</tr>
<tr>
<td>Data transfer time for 1 job</td>
<td>2.5 minutes</td>
</tr>
</tbody>
</table>
1: Intensive submission of FlexX jobs with Chembridge ligands base
2: Resubmission
3: Intensive submission of FlexX jobs with drug like ligands base
4: Resubmission
5: Intensive submission of Autodock jobs with Chembridge ligands base
6: Resubmission
The following institutes contributed computing resources to the data challenge:

IPP-BAS, IMBM-BAS and IPP-ISTF (Bulgaria); CYFRONET (Poland); ICI (Romania); CEA-DAPNIA, CGG, IN2P3-CC, IN2P3-LAL, IN2P3-LAPP and IN2P3-LPC (France); SCAI (Germany); INFN (Italy); NIKHEF, SARA and Virtual Laboratory for e-Science (Netherlands); IMPB RAS (Russia); UCY (Cyprus); AUTH FORTH-ICS and HELLASGRID (Greece); RBI (Croatia); ASCC (Taiwan); TAU (Israel); CESGA, CIEMAT, CNB-UAM, IFCA, INTA, PIC and UPV-GryCAP (Spain); BHAM, University of Bristol, IC, Lancaster University, MANHEP, University of Oxford, RAL and University of Glasgow (United Kingdom).
### Exploitation metrics

<table>
<thead>
<tr>
<th>Metrics</th>
<th>FlexX + Autodock phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total CPU time</td>
<td>80 years</td>
</tr>
<tr>
<td>Number of jobs</td>
<td>72751</td>
</tr>
<tr>
<td>Number of grid nodes</td>
<td>58</td>
</tr>
<tr>
<td>Number of jobs running in parallel on the grid</td>
<td>1643</td>
</tr>
<tr>
<td>Volume of output data</td>
<td>946 GB</td>
</tr>
<tr>
<td>Volume of transferred data (input+output)</td>
<td>6302 GB</td>
</tr>
</tbody>
</table>
## Performance metrics

<table>
<thead>
<tr>
<th>Metrics</th>
<th>FlexX + Autodock phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cumulated millions number of docked ligands</td>
<td>41.27</td>
</tr>
<tr>
<td>Number of docked ligands / h</td>
<td>46,475</td>
</tr>
<tr>
<td>Effective CPU time</td>
<td>67.15 years</td>
</tr>
<tr>
<td>Effective duration</td>
<td>37 days</td>
</tr>
<tr>
<td>Crunching factor</td>
<td>662</td>
</tr>
<tr>
<td>Average transfer rate</td>
<td>0.8 MB/s</td>
</tr>
<tr>
<td>Peak rate</td>
<td>62.1 MB/s</td>
</tr>
</tbody>
</table>
### Efficiency metrics (1/2)

<table>
<thead>
<tr>
<th>Metrics</th>
<th>FlexX + Autodock phases</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success rate</td>
<td>77 %</td>
</tr>
<tr>
<td>Success rate after results checking</td>
<td>46.2%</td>
</tr>
<tr>
<td>Success rate after results checking without WISDOM failures</td>
<td>63 %</td>
</tr>
</tbody>
</table>

- **Efficiency depends on:**
  - Heterogeneous and dynamic nature of the grid
  - Stress usage
  - Automatic jobs (re)submission ("sink-hole" effect)
- Quick overview on very large log-files
- Sorting and merging of files
- Storing and retrieval in databases
• Example: Ligand plot of 1lee (Plasmepsin II) with inhibitors R36 500
Preliminary results of the first data challenge

- **Score of an output is independent of the grid resource where the job runs (conditions controlled)**

- **10% compounds of Chembridge (ZINC) may are hits**
  - Top scoring compounds possess basic chemical groups like thiourea, guanidino, andamino acroleinas core structure.
  - Identified compounds are non peptidic and low molecular weight compounds
  - But the identified compounds look like human thrombin inhibitors
• WISDOM (Wide In-Silico Docking On Malaria) is the first large scale drug discovery initiative on an open grid infrastructure
  – About 80 CPU years to produce TB of data

• Future works on the results
  – Qualitative comparisons of docking tools
  – Ligand similarity based clustering of results

• Future works on the hits
  – simulation on 1000 hits for reranking (EU BioinfoGrid FP6 project)
    ▪ 100 CPU years
    ▪ Docking well fitted for cluster grids, Molecular Dynamics well fitted for supercomputers
  – Finally in vitro testing and structure activity relationships
Data challenge on avian flu: grid facts

• A collaboration of 5 grid projects: Auvergrid, BioinfoGrid, EGEE-II, Embrace, TWGrid
• Data challenge parameters:
  – One docking software: autodock
  – 8 conformations of the target (N1)
  – 300,000 selected compounds
  – 100 year CPU to dock all configurations on all compounds
• Timescale:
  – First contacts: March 1st 2006
  – Kick-off: April 1st 2006
  – Targeted duration: 4 weeks

• See Hurng-Chun Lee's talk

Credit: Y-T Wu
Perspectives

• **Extension of in silico workflow (Embrace)**
  – Virtual docking service at a large scale on gLite (EGEE) with Taverna

• **Second Data Challenge is ongoing (see Hurng-Chun Lee's talk)**

• **Third large scale docking on EGEE in fall 2006**
  – Several new foreseen targets on malaria, dengue and other neglected diseases.
  – Resources needed: up to 80 years CPU per target
  – Supported by EGEE-II and EELA European projects, Swiss BioGrid initiative, AuverGrid, …
  – Any other Asian-Pacific project interested in?

• **We will be pleased to welcome you in the WISDOM initiative!**

• **Grid-enabled *In Silico* Drug Discovery Workshop June 6th 2006 in Valencia (Spain) within the HealthGrid'06 conference**
LPC (CNRS/IN2P3)
- V. Breton
- N. Jacq
- J. Salzemann
- Y. Légré
- M. Reichstadt
- F. Jacq

Fraunhofer SCAI
- M. Hofmann
- M. Zimmermann
- A. Maaß
- M. Sridhar
- K. Vinod-Kusam
- H. Schwichtenberg

EGEE
- Biomed Task Force
- EIS team
- JRA2 team
- SA1 sites
"The only thing necessary for the triumph of Evil is for good men to do nothing!"
Edmund Burke

Questions ?

4th HealthGrid conference
6th – 9th June Valencia (Spain)
http://valencia2006.healthgrid.org