Cheminformatics platform for drug discovery application

Hsi-Kai, Wang
Academic Sinica Grid Computing
EGI User Forum, 13, April, 2011
• Introduction to drug discovery
• Computing requirement of high throughput virtual screening
• Cheminfomatics case study
Computational chemistry / Molecular modeling

useful across the pipeline, but
very different techniques

aim for success, but if not:
fail early, fail cheap

## Strategy in drug discovery

<table>
<thead>
<tr>
<th>Receptor (3D structure) unknown</th>
<th>Ligand unknown</th>
<th>Ligand known</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>CombiChem</td>
<td>Pharmacophore</td>
</tr>
<tr>
<td></td>
<td>HTS</td>
<td>Similarity</td>
</tr>
<tr>
<td></td>
<td>Virtual Screening</td>
<td>QSAR</td>
</tr>
<tr>
<td>Receptor (3D structure) known</td>
<td>Receptor-bases searching</td>
<td>Structure-based drug design</td>
</tr>
<tr>
<td></td>
<td>De novo design</td>
<td>Receptor-ligand interaction</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Docking</td>
</tr>
</tbody>
</table>
• What is grid
  • Many definitions exist in the literature
    • Foster and Kesselman, 1998. “A computational grid is a hardware and software infrastructure that provides dependable, consistent, pervasive, and inexpensive access to high-end computational facilities.”
  • Grid can provide
    • Large scale and on-demand resources
      • Computing resources (computing grids)
      • Storage resources (data grids)
Problem
- Millions of compounds and drugs molecules are presently available for screening
- But developing efficient assay in laboratory for such a work is time-consuming and very expensive

Solution
- Grids offer high-speed computing and huge-data managing capability
- Possible variant targets can be studied quickly by present modelling applications.
- This will help medicinal chemists to respond to major instant threats.

chemical compounds

Receptor structures

Molecular docking ....takes years

Data challenge on Grid ....can be done in weeks

Hits sorting and refining

In vitro screening of best ~50 hits
GAP Service Architecture

- **Core Framework**
  - LSA (Local System Agent) is responsible for dealing with the underlying computing environment.
  - VQS (Virtual Queueing System) directly faces the client and communicates with LSA.
  - VQSCClient - The client APIs for communicating with VQS

- **Application Framework**
  - A set of common APIs for handling advanced application logic
  - Also application specific APIs

- **Presentation Framework**
  - Java based UI APIs
• A lightweight framework for parallel scientific applications in master worker model,

• The framework takes care of all synchronization, communication, and workflow management details on behalf of application
The profile of a DIANE job

- Each horizontal line segment = one task = one docking
- Unhealthy workers are removed from the worker list
- Failed tasks are rescheduled to healthy workers

**good load balance**

**the “bad” worker removed**
• 280 DIANE worker agents were submitted as LCG jobs
• 200 jobs (~71%) were healthy
  – ~16% failures related to middleware errors
  – ~12% failures related to application errors

DIANE utilizes ~95% of the healthy resources

stable throughput
GVSS application: dengue virus

Worldwide dengue distribution

Areas infested with *Aedes aegypti*

Areas with *Ae. aegypti* and dengue epidemics

http://en.wikipedia.org/wiki/Aedes
Dengue NS3 protease

## Dengue Fever Data Challenge / resources & 1st result

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total number of completed docking jobs</td>
<td>300,000</td>
</tr>
<tr>
<td>Estimated needed computing power</td>
<td>4,167 CPU*days</td>
</tr>
<tr>
<td>Duration of the experiment</td>
<td>60 days</td>
</tr>
<tr>
<td>Cumulative computing results</td>
<td>42.5 GB</td>
</tr>
<tr>
<td>Total Computing Recourses in EUAsia VO</td>
<td>268 Cores</td>
</tr>
<tr>
<td>Number of used Computing Elements</td>
<td>6</td>
</tr>
</tbody>
</table>
Joint Computing Resources & Users

- Accumulating Computing Resources in EUAsia VO: 268 cpu-cores (100 – ASGC (TW), 2 – TH, 4 – VN, 18 – MIMOS (MY), 80 – UPM (MY), 64 – CESNET (CZ))
  - lcg-infosites --vo euasia ce
- Registered VQS account:
  - 6 users (TW)
  - 17 user (PH, 15 in AdMU, 2 in ASTI)
  - 2 user (TH, 1 in NECTEC, 1 in HAll)
  - 1 user (MY, UPM)
  - 1 user (ID, ITB)
  - 2 user (VN, IAMl)
  - 1 user (FR, HealthGrid)
Integration of SG & DG by EDGES
Scenario 1 – DG to SG via bridge

- User entry point is DG – using SG is completely transparent from user’s point of view
Scenario 2 – SG to DG via bridge

Desktop Grid 1

Desktop Grid n

Edges Services
DG CE + EDGeS
AR

WMS and other EGEE services

EGEE VO

SG (EGEE) user
(using EGEE UI machine or portal)

- User entry point is SG
- using DG is transparent from user’s point of view
Scenario 3 – SG/DG resources but not through EDGGeS bridges
Web UI Service Architecture

[Diagram showing the architecture of a web UI service, including components such as CE, LBS, gLite-based Service Grid, BOINC Server, Desktop Grid, GridFTP, Web UI, Cell-Master, and connections between these components.]
Simulation of drug discovery workflow

Preparing ligand & protein

Docking

Generating conformation

Scoring

Analyzing & ranking data

- Ligand
- Protein
Protein Database

A Resource for Studying Biological Macromolecules

The PDB archive contains information about experimentally-determined structures of proteins, nucleic acids, and complex assemblies. As a member of the worldwide RCSB PDB, we curate and annotate PDB data according to agreed upon standards.

The RCSB PDB also provides a variety of tools and resources. Users can perform simple and advanced searches based on annotations relating to sequence, structure and function. These molecules are visualized, downloaded, and analyzed by users who range from students to specialized scientists.

Welcome to the PDBbind Database!

The PDBbind database is developed by Dr. Shaomeng Wang's group at the University of Michigan. The PDBbind database is designed to provide a collection of experimentally measured binding affinity data (Kd and IC50) exclusively for the protein-ligand complexes available in the Protein Data Bank (PDB). All of the binding affinity data compiled in this database are cited from original references. Click here for a brief introduction to the PDBbind database.

Current release: The current release is version 2007, which means that it takes account for the protein-ligand complexes officially released in PDB before Jan 1st, 2007. In summary, this release provides binding data of 2,524 protein-ligand complexes. For each entry in this set, a table summarizing the basic information is available on our web site. A total of 1,300 protein-ligand complexes are selected to form the "refined set" with concerns on the quality of structures and binding data, which is compiled particularly for docking/scoring studies. For each entry in this set, properly processed structure files are provided so that they can be readily utilized by molecular modeling programs. Our website also supports interactive structure/similarity search within this data set. Another "core set" provides non-redundant sampling of the "refined set". Currently, a total of 70 different proteins and 250 entries are included in this set.

Availability: Academic/governmental users may access the PDBbind database for free under a license agreement. Industrial/commercial users may access this database with a moderate amount of subscription fee. Please follow the instructions on the "enroll" page.

References
Genetic algorithm
- is a search heuristic that mimics the process of natural evolution. It generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.
  - AutoDock, GOLD…

Molecular dynamics
- is used to find poses by force-fields. The generated conformations usually consists of a simulated annealing to locate the global optimum in a large search space.
  - AMBER, CHARMM…

Shape complementarities
- is a description of the molecules, including solvent-accessible surface area, geometric constraints, H-bond, hydrophobic/hydrophilic interaction between all atoms in the complex.
  - DOCK, FRED…
General class of scoring function

- **Force Field**
  - affinities are estimated by intermolecular van der Waals, electrostatic interaction et al. between all atoms of the two molecules in the complex.
  - **AMBER**…

- **Empirical**
  - count the number of interactions and assign a score based on the number of occurrences. Example H-bond, ionic, hydrophobic/hydrophilic interaction.
  - **LUDI, X-Score**…

- **Knowledge-base**
  - observe known protein/ligand structures, and favor interactions and geometries that are seen often.
  - **DrugScore, PMF**…
Tools of docking and scoring

AutoDock

Welcome!
- What is AutoDock?
- What is AutoDock Vina?
- How do I get started with AutoDock?
- What's new?
- What is AutoDockTools?
- Where is AutoDock used?
- Why use AutoDock?
- Run AutoDock on World Community Grid!

What is AutoDock?
AutoDock is a suite of automated docking molecules, such as substrates or drug candidate structures.

X-SCORE

Scoring function for predicting protein-ligand binding affinities

X-Score is basically a “scoring function”, which computes the binding affinities of the given ligand molecules to their target protein. It can be applied to structure-based drug design studies in combination with molecular docking or de novo structure generation programs. X-Score is developed by Dr. Renmiao Wang in Dr. Shengmeng Wang’s group at the Department of Internal Medicine, University of Michigan Medical School. The first paper that reported X-Score was published in Journal of Computer-Aided Molecular Design, 16: 11-26, 2002. Note that X-Score was formerly known as X-Score for a short while. To learn more about the X-Score program, please read the X-Score online manual.

X-Score is released to the public for free. The latest release is X-Score v1.2. You can download the program by clicking the link below. You will go through a license agreement and fill in some necessary registration information. Once we have received your signed license agreement, we will send you instructions of how to log on our server and download the X-Score package. The X-Score v1.2 package includes the program (executable and source codes), user manual, examples, references and the protein-ligand complex data set originally used for developing X-Score.

Click here to get the X-Score v1.2 package now!

Ref: AutoDock, http://autodock.scripps.edu/
Simulated Condition

• Ligand and Protein
  • PDBBind database v2010 (3429 complexes)

• Docking
  • software: AutoDock
  • computing time: 30 ~ 50 min per docking

• ReScoring
  • software: X-Score
  • computing time: 1 ~ 2 min per scoring
Free energy in AutoDock, X-Score

- **AutoDock**
- **X-Score**
Free energy $R^2$ in ligand molecular weight

<table>
<thead>
<tr>
<th>MW</th>
<th>Number</th>
<th>AutoDock</th>
<th>X-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>~300</td>
<td>1071(31.2%)</td>
<td>0.295</td>
<td>0.858</td>
</tr>
<tr>
<td>300~400</td>
<td>801(23.4%)</td>
<td>0.298</td>
<td>0.618</td>
</tr>
<tr>
<td>400~500</td>
<td>732(21.3%)</td>
<td>0.166</td>
<td>0.546</td>
</tr>
<tr>
<td>500~</td>
<td>825(24.1%)</td>
<td>0.216</td>
<td>0.505</td>
</tr>
<tr>
<td>all</td>
<td>3429</td>
<td>0.355</td>
<td>0.711</td>
</tr>
</tbody>
</table>
### Free energy $R^2$ in protein enzyme type

#### Enzyme type

<table>
<thead>
<tr>
<th>Enzyme type</th>
<th>Number</th>
<th>AutoDock</th>
<th>X-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oxidoreductases</td>
<td>74(2.2%)</td>
<td>0.359</td>
<td>0.458</td>
</tr>
<tr>
<td>Transferases</td>
<td>889(25.9%)</td>
<td>0.335</td>
<td>0.756</td>
</tr>
<tr>
<td>Hydrolases</td>
<td>1314(38.3%)</td>
<td>0.477</td>
<td>0.698</td>
</tr>
<tr>
<td>Lyases</td>
<td>203(5.9%)</td>
<td>0.230</td>
<td>0.666</td>
</tr>
<tr>
<td>Isomerases</td>
<td>79(2.3%)</td>
<td>0.441</td>
<td>0.964</td>
</tr>
<tr>
<td>Ligases</td>
<td>65(1.9%)</td>
<td>0.145</td>
<td>0.594</td>
</tr>
<tr>
<td>Other</td>
<td>805(23.5%)</td>
<td>0.217</td>
<td>0.706</td>
</tr>
<tr>
<td>all</td>
<td>3429</td>
<td>0.355</td>
<td>0.711</td>
</tr>
</tbody>
</table>

#### $R^2$ Values

- Oxidoreductases: 0.458
- Transferases: 0.335
- Hydrolases: 0.477
- Lyases: 0.230
- Isomerases: 0.441
- Ligases: 0.145
- Other: 0.217
- All: 0.355
### RMSD in AutoDock, X-Score

<table>
<thead>
<tr>
<th>RMSD</th>
<th>&lt; 1.0</th>
<th>&lt; 1.5</th>
<th>&lt; 2.0</th>
</tr>
</thead>
<tbody>
<tr>
<td>AutoDock</td>
<td>870(25%)</td>
<td>1365(40%)</td>
<td>1777(52%)</td>
</tr>
<tr>
<td>X-Score</td>
<td>869(25%)</td>
<td>1343(39%)</td>
<td>1721(50%)</td>
</tr>
<tr>
<td>Total</td>
<td>1048(31%)</td>
<td>1589(46%)</td>
<td>2044(60%)</td>
</tr>
<tr>
<td>Same</td>
<td>691(20%)[66%]</td>
<td>1119(33%)[70%]</td>
<td>1454(42%)[71%]</td>
</tr>
</tbody>
</table>
Future work

• Finish implement Web-based Virtual Screening Service with EDGeS infrastructure.
• The 691 proteins x 691 ligands docking tasks complete and data analysis.
• Other proteins are classified by enzyme code.
Thank you for your attention