A Grid Environment for Data Integration of Scientific Databases

Hideo Matsuda¹, Susumu Date¹, Shinji Shimojo¹
(1 Osaka University)
Kentaro Wakatsuki², Takehiro Furudate²
Gen Kawamura¹,³, Yoshiyuki Kido¹,⁴
(2 Hitachi Software, 3 Aztec System, 4 Mitsui Knowledge Industry)
Outline of Talk

- Genome Annotation Pipeline
- Introduction to Japan NAREGI Project
- NAREGI Data Grid
- Scientific DBs (focus: Lifescience DBs)
- Data Integration of Lifescience DBs
Genome Annotation Projects

• Inspired by the Drosophila Genome Annotation Jamboree.
• Several full-length cDNA (FLcDNA = clone of gene transcript) annotation projects have been organized in Japan (mouse, human and rice genomes).
• In the projects, FANTOM (Functional ANnoTation Of Mouse) is the first annotation project for genome FLcDNA sequences.
• Organized by RIKEN GSC.
• 1st FANTOM Meeting (FANTOM1) at RIKEN Tsukuba Institute
  – 21K FLcDNA seqs.
  – >60 researchers (>6 countries)
  – FANTOM DB:
    http://fantom.gsc.riken.jp/
• FANTOM2 (2002) 60K seqs at RIKEN Yokohama Institute
• FANTOM3 (2004) 103K seqs at RIKEN Yokohama Institute
  *Science*, 309:1559-1563, 2005
Annotation Pipeline at FANTOM
(Kasukawa et al., Genome Res. 13:1542, 2003)

1. Clustering of FLcDNA seqs w/ FLAST(DDS)+ CAP3 and ClustalW
2. Masking repetitive elements w/ Repbase (RepeatMasker).
3. ORF (Coding Sequence) Prediction w/ RIKEN Decoder.
4. Sequence homology search (FASTY and BLASTX) against NCBI nr., SwissProt/TrEMBL nr.
5. Motif / Domain search w/ Pfam(estwise), InterPro (InterProScan), etc.
6. Mapping to mouse & human Genome (NCBI build) w/ RIKEN Genomapper.
7. Assignment of Gene Ontology (GO) terms.

blue: tools, red: data or databases
The Ideal World: Ubiquitous VO & user management for international e-Science

Europe: EGEE, UK e-Science, ...

US: TeraGrid, OSG,

Japan: NII CyberScience (w/NAREGI), ...
Other Asian Efforts (APGrid, etc.)...

HEP Grid VO
NEES-ED Grid VO
Astro IVO

Grid Regional Infrastructural Efforts
Collaborative talks on PMA, etc.

Standardization, commonality in software platforms will realize this

By courtesy of Prof. Matsuoka (Tokyo Inst. Tech.)

All different software stacks but users don’t see them
The Reality: Convergence/Divergence of Project Forces

(original slide by Dr. Stephen Pickles, edited by Prof. Matsuoka)
Outline of the NAREGI Project

• NAREGI: National Research Grid Initiative in Japan
• Funded by Japanese Government (MEXT)
• Started from April 2003
• Two main sites:
  – R&D: NII (National Institute for Informatics)
NAREGI Software Stack

Grid-Enabled Nano-Applications

Grid Programming Libraries
- GridRPC
- GridMPI

Grid Visualization
Grid PSE

Grid Workflow
Super Scheduler
Distributed Information Service

Data Grid

Globus, Condor, UNICORE → OGSA: NAREGI implementation & Open Sources

Grid VM

High-Performance & Secure Grid Networking

SuperSINET

NII
IMS
Research Organizations
etc

Computing Resources

National Research Grid Initiative
Project Leader: Ken Miura (NII)

(WP1) Resource Management in the Grid Environment
Satoshi Matsuoka (Tokyo Inst. Tech.)

(WP2) Grid Programming Environment
Satochi Sekiguchi, Yoshio Tanaka (AIST)

(WP3) Grid Application Environment
Hitohide Usami (NII), Shigeo Kawata (Utsunomiya Univ.)

(WP4) Data Grid Environment
Hideo Matsuda (Osaka Univ.)

(WP5) High-performance & Security Grid Networking
Shinji Shimojo (Osaka Univ.), Yuji Oie (Kyushu Inst. Tech.),
Makoto Imase (Osaka Univ.)

(WP6) Grid-Enable Nano Applications
Mutsumi Aoyagi (Kyushu Univ.)
Nano-Science: coupled simulations on the Grid as the sole future for true scalability... between Continuum & Quanta.

Material physics (Infinite system)
- Fluid dynamics
- Statistical physics
- Condensed matter theory

Limit of Idealization

Limit of Computing Capability

Molecular Science
- Quantum chemistry
- Molecular Orbital method
- Molecular Dynamics

Multi-Physics

Old HPC environment:
- decoupled resources,
- limited users,
- special software, ...

Coordinates decoupled resources;
Meta-computing,
High throughput computing,
Multi-Physics simulation w/ components and data from different groups within VO composed in real-time

The only way to achieve true scalability!

National Research Grid Initiative
Simulation Scheme

3D-RISM

- evenly-spaced mesh
- pair correlation functions
  - $g_\gamma(r)$
  - $c_\gamma(r)$
  - $h_\gamma(r)$

Mediator

- find correlations between mesh points
- Data exchange between meshes
- solvent distribution
  - $h_\gamma(r)$
- effective charges on solute sites
  - $q_\alpha$

FMO

- adaptive meshes
- monomer calculations
  - $H^1\Psi^1 = E^1\Psi^1$
  - $H^2\Psi^2 = E^2\Psi^2$
  - ...
- dimer calculation

By courtesy of Prof. Aoyagi (Kyushu Univ.)
Workflow based Grid FMO Simulations of Proteins

By courtesy of Prof. Aoyagi (Kyushu Univ.)
Several Aspects in Data Grid

• Data Transfer
  – GridFTP, RFT, TeraGrid Copy, …
• Data Management
  – SRM, …
• Grid Filesystem
  – SRB, Gfarm, …
• Data Integration
  – OGSA-DAI, myGrid, BRIDGES, …
• Data-intensive Workflow
  – many…. 
NAREGI DataGrid

- GGF-GFS (AIST Gfarm)
- Data Transfer Service between storage and computation nodes (GridFTP)
- OGSA-DAI for database queries
NAREGI Data Grid Environment

Grid Workflow

Data Grid Components

- Data Access Management
- Metadata Construction
- Data Resource Management

Import data into workflow → Place & register data on the Grid → Assign metadata to data → Store data into distributed file nodes

Job 1 → Job 2 → Job n → Grid-wide DB Querying → Grid-wide File System
Implementation of the Data Grid Environment

Data Access Management

Workflow Tool (WFT)

Job 1

Job n

Data 1

Data n

Import data into workflow

Super Scheduler (SS)

Data Staging

Computational Nodes

Job 1

Job 2

Job n

Data 1

Data 2

Data n

Place data on the Grid

Globus Toolkit 4.0.1

Tomcat 5.0.28

Data Resource Management

OGSA-DAI WSRF2.1

PostgreSQL 8.0

MySQL 5.0

Data Resource Information DB

Metadata Construction

Data Specific Metadata DB

National Research Grid Initiative
• Middleware of using virtual huge data resources that gives us the integration of the storages on cluster machines.
• Virtualization of disk resources
• Split allocation of huge file
OGSA-DAI

- OGSA-DAI (Data Access and Integration)
- Middleware to facilitate access to data resources
- Data resources are sources/sinks of data
  - OGSA-DAI currently supports:
    - Relational databases
    - XML data resources
    - Files
- Provides a partial virtualisation of data
  - Hide connection mechanism
  - Move computation to the data
  - Do not hide the underlying data model (relational, XML, file)
Scientific Data and Grid

- High Energy Physics (LHC, D-Grid, HEPGrid, ....)
  - Very huge amount of data
  - High speed & reliable data transfer
- Astronomy (Sloan Digital Sky Survey, AstroGrid, ...)
  - Virtual Observatory: Large amount of image data
  - Heterogeneous observatories and instruments
  - Distributed query processing
Lifescience DBs

- Not so huge amount of data compared to high-energy physics, astronomy, etc.
- Large number of DBs (> 700 DBs)
- Highly heterogeneous and complex in data description
- Need some semantics
• The number of scientific DBs (especially, in life sciences) is very rapidly increasing.
Osaka-U BioGrid Project

• Started from 2002
• Goals
  Technology Development for: Pharmaceutical (drug discovery) and Medical Sciences.
• Leader: Shinji Shimojo (CMC, Osaka Univ.)
• Government Support (MEXT): 5 years (1~4M$/year)
• Web site: http://www.biogrid.jp
BioGrid Project: Application to Drug Discovery

Data Grid Issues:
1. Heterogeneity of data descriptions.
2. Large number of relationships
Databases in BioGrid Project

Disease
Medical Encyclopedia
3K entries
LITerature DB
(Protein Research Foundation)

Pathway
KEGG
18K entries

Protein
SwissProt 163K entries,
60M amino acids
PIR 283K entries,
96M amino acids
PDB 28K protein structures

Genome
Ensembl, DDBJ
Human 24K genes
3.3G nucleotides

Compound
MDDR 150K compounds
PubChem, Ligand Info
1.8M compounds

Interaction
ENZYME
GPCR-DB
NucleaRDB
LGIC DB
Increase of Public Chemical Databases

• **Ligand.Info** ([http://ligand.info/](http://ligand.info/)) : composed of the following data.
  - ChemBank subset 2,344 records
  - ChemPDB subset 4,009 records
  - KEGG subset 10,005 records
  - Anti-HIV NCE subset 42,689 records
  - Drug-likeness NCI subset 192,323 records
  - Not annotated NCI subset 15,237 records
  - AKos GmbH subset 544,391 records
  - Ligand.Info Asinex Ltd. subset 348,276 records
  - Tim Tec subset 7,500 records (in total: 1,159,274 records)

  - 711,361 records

• **Drug Bank** ([http://redpoll.pharmacy.ualberta.ca/drugbank](http://redpoll.pharmacy.ualberta.ca/drugbank))
  - 256 records (company names)
Protein-Compound Interaction Search is one of the most important technologies in drug discovery.
Ontology-based Metadata

- Protein, Gene DBs
- Compound DBs

Interaction Rules:
- X is an agonist of A
- Y is an inhibitor of B

Gene Ontology
- Protein(Gene) A
- Protein(Gene) B

Ligand Ontology

Compound Activity (Proud DDR)
- Compound X
- Compound Y

DB records
Extract relationships between protein functions and compound activities by traversing hierarchical classifications (or ontologies).
Database federation using Grid technology.

Databases for each category are provided as Grid Services.
### Protein Compound Interaction View

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Homology Score</th>
<th>Homology Evalue</th>
<th>Compound</th>
<th>MDDR EXTREG</th>
<th>Molformula</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>212740</td>
<td>212740</td>
<td>C26 H25 N7 O</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>260825</td>
<td>260825</td>
<td>C28 H30 N8 O5</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>317215</td>
<td>317215</td>
<td>C27 H24 F3 N5 O3 S</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>FK-739</td>
<td>193909</td>
<td>C24 H22 N7 Na</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>XR-510</td>
<td>211727</td>
<td>C39 H47 F N5 O6 S K</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>UR-7198</td>
<td>225409</td>
<td>C27 H29 N3 O2</td>
<td>antagonism</td>
</tr>
<tr>
<td>Type-1 angiotensin II receptor (AT1) (AT1AR)</td>
<td>657</td>
<td>0.0</td>
<td>LR-B104</td>
<td>339640</td>
<td>C31 H30 O6 S O4</td>
<td>antagonism</td>
</tr>
</tbody>
</table>

**Diagram:**
- **Protein UK** interacts with **Interaction Japan** leading to the binding of **Drug Japan**.
<table>
<thead>
<tr>
<th><strong>Display Information of Known Compound</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Cas number</strong></td>
</tr>
<tr>
<td><strong>Chemical name</strong></td>
</tr>
<tr>
<td><strong>Company code</strong></td>
</tr>
<tr>
<td><strong>Compound ID</strong></td>
</tr>
<tr>
<td><strong>Molformula</strong></td>
</tr>
<tr>
<td><strong>Molname</strong></td>
</tr>
<tr>
<td><strong>Phase</strong></td>
</tr>
<tr>
<td><strong>2D Regno</strong></td>
</tr>
<tr>
<td><strong>Source1</strong></td>
</tr>
<tr>
<td><strong>Source2</strong></td>
</tr>
<tr>
<td><strong>Trademark</strong></td>
</tr>
<tr>
<td><strong>Molecular Weight</strong></td>
</tr>
<tr>
<td><strong>logp</strong></td>
</tr>
<tr>
<td><strong>SMILES</strong></td>
</tr>
</tbody>
</table>
## Compound Search Result

### Similar Compound

<table>
<thead>
<tr>
<th>Compound Name</th>
<th>Compound ID</th>
<th>Molformula</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ondanetron hydrochloride</td>
<td>30006044</td>
<td>C18H24ClN3O3</td>
<td>0.686</td>
</tr>
<tr>
<td>Ipro 32-0432</td>
<td>10000054</td>
<td>C28H28N4O2</td>
<td>0.676</td>
</tr>
</tbody>
</table>

### Inquiry Compound

<table>
<thead>
<tr>
<th>Compound Name</th>
<th>Compound ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Telmisartan</td>
<td>30005957</td>
</tr>
</tbody>
</table>

### Compounds possibly-interacted to the target protein

- Ondanetron hydrochloride
- Ipro 32-0432
- Telmisartan

Drug Japan → Drug Japan
NAREGI Data Integration w/ Workflow

Access Management System

Data Integration Flow

Metadata Construction System

OGSA-DAI WSRF2.1

MySQL 5.0

Data Integration Results

Storage Storage Storage

Gfarm 1.2 PL4

Globus Toolkit 4.0.1

MySQL 5.0
Data Integration Workflow using Metadata

- Species
- Disease
- Gene
- Protein
- Compound

Metadata
- Species
- Gene2Protein
- Interaction
- Disease2Gene
- Protein
- Compound

DB Category
DB Synonym

Linkage

157 Databases

Organelle DBs
Disease Pathway DBs
Human Gene and Disease DBs
Nucleotide Sequence DBs
Protein Seq. DBs
Structure DBs
Disease to Gene & Gene to Protein Metadata

• Disease to Gene

Extraction of the relationships between diseases and genes from OMIM
Example
Hypertension ↔ PPARG, PPARG1, PPARG2, etc..
Diabetes ↔ IDDM1, IDDM15, IDDM8, IDDM10, etc..

• Gene to Protein

Extraction of translation information from UniGene
Example.
PPARG ↔ swiss-prot P41830
PPARG ↔ swiss-prot P13055
• Two types of metadata
  – Grid file metadata
  – User metadata (domain-specific metadata)
Parallel Search across DBs

- Retrieve in parallel against lifescience DBs by using Globus Toolkit, OGSA-DAI, and Gfarm.
- We extract the information of related records among different lifescience DBs and integrate their results.
Data Integration System Overview

Global File System (~ total 1.5TB)

Globus Toolkit
OGSA-DAI

DB & Metadata

Hit Data

Apache Tomcat

DB & Metadata

Keyword Index

Apache Lucene

Search

Top Page

DBs & IDs

pfg2021 pfg2022 pfg2023 pfg2024
dg2 dg3 dg7
Keyword Search 1

Data Integration Flow

Keyword: hypertension

Protein Sequence Databases
- Blocks
- EMBLCONEXP
- InterPro
- P3MA
- PFAM
- PFAMHMMES
- PFAMSEED
- ProDom
- PIR
- SWISS-PROT
- UniProt
- UNIPRO

Name of Database in Our System

Human Genes And Diseases
- GENOMEREVIEWS
- HGDBASE_HAPLOTYPE
- HUMAN_MITBASE
- MUTRES
- OMIMOFFSET
- P53I
- TAXONOMY

Input Keyword
Keyword Search 2

Data Integration Flow

Click Category Anchor

Count of Retrieval Results

Protein Sequence Databases

<table>
<thead>
<tr>
<th>Database</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Blocks</td>
<td>0</td>
</tr>
<tr>
<td>EMBLCONEXP</td>
<td>0</td>
</tr>
<tr>
<td>InterPro</td>
<td>0</td>
</tr>
<tr>
<td>PfamA</td>
<td>0</td>
</tr>
<tr>
<td>PfamHmmfs</td>
<td>0</td>
</tr>
<tr>
<td>PfamSeeds</td>
<td>0</td>
</tr>
<tr>
<td>ProDom</td>
<td>0</td>
</tr>
<tr>
<td>RemTrembl</td>
<td>0</td>
</tr>
<tr>
<td>Swiss-Protein</td>
<td>43</td>
</tr>
<tr>
<td>Uniprot</td>
<td>0</td>
</tr>
<tr>
<td>Uniref50</td>
<td>0</td>
</tr>
</tbody>
</table>

Human Genes And Diseases

<table>
<thead>
<tr>
<th>Database</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>Genomereviews</td>
<td>0</td>
</tr>
<tr>
<td>Hgbase_Haplotype</td>
<td>0</td>
</tr>
<tr>
<td>Human_Mtbase</td>
<td>0</td>
</tr>
<tr>
<td>Mutres</td>
<td>0</td>
</tr>
<tr>
<td>Omimoffset</td>
<td>1</td>
</tr>
<tr>
<td>Prg</td>
<td>0</td>
</tr>
<tr>
<td>Taxonomy</td>
<td>0</td>
</tr>
</tbody>
</table>
Select dataflow template

Data Integration Flow 1

Data Integration Flow

Protein Sequence Databases >> Human Genes And Diseases >> Structure Databases

Protein Sequence Databases >> RNA Sequence Databases >> Microarray Data And Other Gene Expression Database

Protein Sequence Databases >> Immunological Databases >> Human Genes And Diseases >> Proteome Resources
Data Integration Flow 2

Icon color shows status of process:
- Blue: running
- Green: finished
- Red: failed
Summary and Future Works

• We have developed a system for data integration of >100 lifescience DBs.
• Globus Toolkit and OGSA-DAI integrates distributed DBs and metadata.
• DB keywords and indices are stored in a grid filesystem using Gfarm.

Future Works
• Data are currently downloaded. Need to use web-service (WSDL/SOAP) interface.
• Data-intensive workflow tool is still under development.