Data-intensive application environment for DarwinTree

Li Jian-hui, Zhou Yun-chun, Meng Zhen Lin Xiao-guang

Scientific Data Center
Computer Network Information Center
Chinese Academy of Sciences
Scientific databases and data-intensive application environment

IN CAS

Application environment for Phylogenetic analysis of land plants

DarwinTree

———Molecular data analysis and application environment
Background

- Large scientific facilities produce huge data
  - +20 being operation
  - +20 under construction
- Long-Term field observation stations
  - +100 stations including Ecology, Environment, Space, etc.
- Long-Term Research data need to be archived and curation and sharing
  - 100+ institutes
Scientific Data infrastructure

- Application enabled environments and typical applications
- Middle ware (Scientific data grid middleware, internet-based storage service middleware…)
- Software and Toolkits (scientific data collection, curation, and publishing, data analyzing and visualization…)
- Scientific databases
- Massive storage system
- Data-intensive computing facility
- High speed network
Massive Storage System and Data Storage Service

• Internet-based storage service system
  – Data backup and storage for large scientific facilities
  – Archiving and curation of research data and databases in CAS
  – Supporting on-line data accessing, analyzing
  – Supporting data-intensive applications

• Capability (by 2010)
  – 1PB online Disk storage
  – 5PB tape storage
  – ~500 CPU computing capability + 100Tflops
  – 2.5Gbps network connected with CSTNet, CNGI, GLORIAD
Structure of Massive storage system

- CSTNET
- LAN
- SAN
- Data Server
- Database Server
- App Server
- Mgmt Server
- Process Server
- Offline Tape Cabinet
- Nearline Tape Library
- Nearline Diskarray
- Online Diskarray
- HPC Env.
- MDS
Scientific Databases (SDB)

• A Long-term mission started in 1986 which funded by CAS
  – many institutes involved
  – long-term, large-scale collaboration
  – data from research, for research

• Collecting multi-discipline research data and promoting data sharing
  – 337 research databases by 53 institutes
  – 147TB data available to open access and download

http://www.csdb.cn
There were 113 institutions directly under CAS, including 92 research institutes (including three botanical gardens), six universities & supporting organizations (including two universities, one supporting organization, one documentation and information unit, two publication units), 12 management organizations that consist of the headquarters and branches, and three other units. There were 26 legal entities affiliated to CAS institutions. There were 21 CAS invested holding enterprises.
Scientific Databases (cont.)

• SDB Contents
Scientific Databases (cont.)

• Database integration
  – Resource database
  – Reference database
  – Application oriented database
Scientific Databases (cont.)

• 8 Resource databases
  – Geo-Science
  – Biodiversity
  – Chemistry
  – Astronomy
  – Space Science
  – Micro biology and virus
  – Material science
  – Environment

• 2 Reference databases
  – China Species
  – compound

• 4 application-Oriented databases
  – High Energy (ITER)
  – Western Environment Research
  – Ecology research
  – Qinghai Lake Research
Access Scientific Databases

- All databases can be accessed online, most data are free and open (over 70%)
  - No restriction and no charge in China

- From 2003 to now
  - Data download volume: over 90TB
  - Data Access Visitors: over 9 Millions
CAS Scientific Data Grid

• Scientific Data Grid Middleware (SDG)
  – SDGM is built upon the Scientific Database, supporting to find and access large scale, distributed and heterogeneous scientific data uniformly and conveniently in a SECURE and proper way.

• Building scientific data application grid according to domain requirements
  – Integrate distributed data, analysis tools and storage and computing facilities, providing a uniform data service interface.
  – 4 pilot grids
    • bioscience grid
    • geoscience grid
    • Chemistry grid
    • Astronomy and space science grid
CAS Scientific Data Grid

Scientific Data Grid Middleware

Scientific Data Grid Applications

Chemistry Grid

Bioscience Grid

Geosciences Grid

Other Grids

Scientific Databases
Scientific databases and data-intensive application environment

IN CAS

Application environment for Phylogenetic analysis of land plants

DarwinTree
—–Molecular data analysis and application environment
Phylogeny of Origin

– Darwin
Origin of species

– Linnaeus
Modern taxonomy and systematics
Molecular phylogenetic analysis

Data Extraction

Complementary sequences

Phylogeny Programs

analysis toolkits

Reconstruction TOL
Tree of Life / Phylogenetic tree

TOL is to bring all living species (existing and extinct) linked to a phylogenetic tree with the huge amount of information:

- clarify the origin of life
- biological evolution pattern
- categories of biological evolution and phylogenetic relationships
- the survival of biodiversity patterns and dynamics of change.

Challenge facing the life sciences:

- Construction of TOL
- to fully exploit and to take advantage of the information

DarwinTree

- Molecular data analysis and application environment
How to make use of public databases of DNA sequence information, existing?

How to effectively filter the information?

How to quickly auto-generate the tree of life which reflects the evolutionary history of different biological taxa?

How to fully exploit and utilize the tree of life implied in the great information?
DarwinTree

traditional methods & bioinformatics

Future: Google in Life System
Fairylake Botanical Garden (FBG) & Institute of Botany (IB)
Computer Network Information Center (CNIC), CAS
Organizations

• The goals
  • provide automatically work environment
  • data gathering
  • phylogenetic analysis for scientific research
  • share information with other databases
  • to link information from other databases through phylogeny
  • to replace the non-essential error-prone manual labor
  • collection, sharing and preservation environments
  • Process-based data-intensive analysis & high-performance computing environment
Workflow of DarwinTree

- Integrated multi-source data
- Integrated massive analysis tools & models
- Process-oriented analysis
- Visualization, interactive and convenient environment
# Overall framework

## Data Acquisition
- Remote public database interaction, Automatically updated from time to time
- Complementary sequences
- Sequence search model
  - Repeated strategy
  - Timing rules

## Large quantities of data reduction

## Data Cleaning
- Remove the pseudogenes, horizontal transfer genes, organellar copies, recent paralogs and others exceeding of the expected rules etc.
- Cleaning sequence model, multiple sequence alignment model, other expected rules

## Reconstruction of the TOL
- Different branches reconstruction
  - (ML, MP, mrbayers)
  - Assembly of large trees
- Choose & organization of phylogenetic tree models, optimization models and assembly of large trees

## Data Platform Integration
- Workflow organization
- Visualization of the tree of life
- Database of species samples, etc. organization

---

**Database system**

**Operating environment support for model & algorithm**
Data Acquisition
- Remote public database interaction, automatically updated from time to time
- Complementary sequences
- Sequence search model, repeated strategy, timing rules

Large quantities of data reduction

Data Cleaning
- Remove pseudogenes, horizontal transfer genes, organellar copies, recent paralogs, and others exceeding the expected rules etc.
- Cleaning sequence model, multiple sequence alignment model, other expected rules

Reconstruction of the TOL
- Different branches reconstruction (ML, MP, mrbayers)
- Assembly of large trees
- Choose & organization of phylogenetic tree models, optimization models, and assembly of large trees

Data Platform Integration
- Workflow organization
- Visualization of the tree of life
- Database of species samples, etc. organization

Database system
- Operating environment support for model & algorithm
Before

Hand search data from a large number of database
To know the peer's research
### NCBI taxonomy (update 2010-02-10 01:00)

<table>
<thead>
<tr>
<th>NCBI taxonomy</th>
<th>Embryophyta</th>
<th>Tracheophyta</th>
<th>Marchantiophyta</th>
<th>Bryophyta</th>
<th>Anthocerotophyta</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbcL</td>
<td>21231</td>
<td>18988</td>
<td>1206</td>
<td>955</td>
<td>82</td>
</tr>
<tr>
<td>matK</td>
<td>16222</td>
<td>16195</td>
<td>20</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>atpB</td>
<td>4170</td>
<td>3982</td>
<td>135</td>
<td>47</td>
<td>6</td>
</tr>
<tr>
<td>nad5</td>
<td>1662</td>
<td>388</td>
<td>243</td>
<td>984</td>
<td>47</td>
</tr>
<tr>
<td>matR</td>
<td>867</td>
<td>867</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>440</td>
<td>428</td>
<td>1</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>ALL</td>
<td>44772</td>
<td>41024</td>
<td>1605</td>
<td>2007</td>
<td>136</td>
</tr>
</tbody>
</table>
FBG-1 taxonomy (update 2010-02-10 01:00)

<table>
<thead>
<tr>
<th>FBG_1 taxonomy</th>
<th>Embryophyta</th>
<th>Tracheophyta</th>
<th>Marchantiophyta</th>
<th>Bryophyta</th>
<th>Anthocerotophyta</th>
</tr>
</thead>
<tbody>
<tr>
<td>rbcL</td>
<td>6947</td>
<td>5145</td>
<td>1000</td>
<td>736</td>
<td>66</td>
</tr>
<tr>
<td>matK</td>
<td>2889</td>
<td>2862</td>
<td>19</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>atpB</td>
<td>1260</td>
<td>1131</td>
<td>80</td>
<td>43</td>
<td>6</td>
</tr>
<tr>
<td>nad5</td>
<td>914</td>
<td>86</td>
<td>164</td>
<td>629</td>
<td>35</td>
</tr>
<tr>
<td>matR</td>
<td>239</td>
<td>239</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>18S rRNA</td>
<td>108</td>
<td>99</td>
<td>1</td>
<td>8</td>
<td>0</td>
</tr>
<tr>
<td>ALL</td>
<td>12371</td>
<td>9576</td>
<td>1264</td>
<td>1423</td>
<td>108</td>
</tr>
</tbody>
</table>
Data Acquisition

- Remote public database interaction: Automatically updated from time to time
- Complementary sequences
- Sequence search model
- Repeated strategy
- Timing rules

Large quantities of data reduction

Data Cleaning

- Remove the pseudogenes, horizontal transfer genes, organelar copies, recent paralogs and others exceeding the expected rules, etc.
- Cleaning sequence model; multiple sequence alignment model; other expected rules

Reconstruction of the TOL

- Different branches reconstruction (ML, MP, mrbayers)
- Assembly of large trees
- Choose & organization of phylogenetic tree models, optimization models and assembly of large trees

Data Platform Integration

- Workflow organization
- Visualization of the tree of life
- Database of species samples, etc. organization

Database system

Operating environment support for model & algorithm
Before

In the data analysis process a large number of analytical tools each process needs the results of the analysis from one tool to another.
Generation of the phylogenetic trees
**Data Extraction**
- acquiring sequence data from the international public gene databases [Genbank, DDBJ, and EMBL]
- submit researchers own sequences.

**Data Editing & Cleaning**
- manual work of sequences resulting from MSA to eliminate the noise.
- cutting out the blank spaces; deleting the singular sequences, etc.

**Multiple Sequence Alignment (MSA)**
- biological sequences (3+)
  [protein, DNA, or RNA]
- sequence homology can be inferred
- to assess the sequences’ shared evolutionary origins.
Format Converting

- convert the .aln (Clustal) format into some others
- .phy (phylips format) & .nex (nexus format)

Tree Reconstruction

- constructing phylogenetic trees from sequences with specific tools
- MP (maximum parsimony) & ML (maximum likelihood)
- available implementation of these optimal algorithms.

Tree Visualization

- show the resulting phylogenetic tree with specific visualization tools
- some applets to show graphs on web dynamically.
Welcome User: xlf  Group User Panel

View Data
Know the sequences from public databases(GenBank, Etc.) of specific gene(XXX, atpB, Etc.).

Select Gene:

Taxon Name:

Modularization Analysis
- Multiple Alignment
- Clustering & Editing
- Tree Reconstruction
- Phylogeny Evaluation
- Visualization

Account Management
- Modify Password
- Modify Information

Peer-Know
The ones having no sequences of the specific gene(XXX, atpB) print here.

Peer-Know

Excel Output

Evoke New Task
Customize default values of workflow
Evoke a new workflow task

Cooperated by Fairylake Botanical Garden & Institute of Botany & Computer Network Information Center, CAS
CopyRight © 2009 Computer Network Information Center, Chinese Academy of Science. All Rights Reserved
Email: pshp@cnic.cn
1. Search taxon
2. Add Accession
3. Add own sequence in fasta format
4. Add own sequence in fasta file

Any or all to use
next
6 tips
Welcome User: Common User Panel

Batch Extraction Result: 1250664640272 fasta

Please choose the method to alignment multiple:

- ClustalW (Emma, Sequence)
- ClustalW-MPI (local PC)
- ClustalW-MPI (cluster)
- ClustalW-MPI (Dempcomp 7000)

Parameters Setting Panel

Number of Processes: 16

Submit And Next
summary

• Scientific databases and data-intensive application environment
  – Massive Storage System and data Storage Service
  – SDB
  – SDG

• DarwinTree
  – application environment for Phylogenetic analysis of land plants
  – Workflow
  – Use surface
  – Future
Thank you!

DarwinTree@cnic.cn

palpp@cnic.cn

http://phylo.csdb.cn:8080/palpp/

http://darwintree.cn(since May)