Laboratory of

Neuro Imaging

Visual Informatics and Computational Genomics using the Graphical Pipeline Environment

Ivo D. Dinov

http://www.LONI.ucla.edu
http://Pipeline.loni.ucla.edu
Outline

• The Pipeline Environment
  - Distributed multi-client/server computing
  - Efficient resource integration environment
  - Data I/O Interface for external DB access

• Pipeline Library of Tools
  - Biomedical image processing tools
  - Shape representation, modeling and analysis
  - Statistical analysis tools

• Pipeline Applications & Genomics Demo
  - Brain Mapping
  - Informatics/Genomics
    • Motivation
    • Integrated Protocol for analyzing Genomics Data
    • Interoperable Tools: MAQ, SAMtools, Bowtie, etc.
    cranium.loni.ucla.edu, fgene1.bic.uci.edu, pws.loni.ucla.edu, ...

• Computational Infrastructure
The Pipeline Environment

http://Pipeline.loni.ucla.edu

• Design, validation, execution and dissemination of heterogeneous workflows
• Tool discovery
• Tool interoperability
• Distributed computing
• User-friendly access to data, hardware infrastructure and computational neuroscience expertise

Pipeline Tool Library
Tested Pipeline Genomics and Informatics Tool Library

- Bioinformatics BLAST
- EMBOSS Bioinformatics Workflows
- mrFAST
- GWASS Genomics
- PLINK GWAS
- Mapping and Assembly with Qualities (MAQ)
- Sequence Alignment and Mapping, SAMtools
- Bowtie, GATK, etc.

http://pipeline.loni.ucla.edu/support/pipeline-workflows/
Statistical Analysis Tools

LONI Statistics
- Chi Square Model Fit
- False Discovery Rate Correction
- False Discovery Rate Correction (Text)
- Fisher Exact
- Fligner Killeen
- Friedman Test
- Independent Kruskal Wallis
- Kolmogorov Smirnoff
- Multiple Linear Regression
- One Sample T Test
- One-Way ANOVA
- Simple Linear Regression
- Split Volume
- Two Independent T-Test (Summary Only)
- Two Independent Wilcoxon Test
- Two Paired Signed Rank Test
- Two Paired Signed T-Test (Summary Only)
- Two Paired T Test
- Two-Way ANOVA
- Volume Kruskal Wallis
- Volume One Sample T-Test
- Volume Regression Analysis
- Volume Regression Analysis (Output Directly)
- Volume Two Independent Samples T-Test
- Volume Two Independent Samples Wilcoxon T
- Volume Wilcoxon Two Paired Samples Rank-Sum

MNI
- PCA Registration
Applications & Demo

- **Brain Mapping**
  - Global and Local Shape Analyses
    - These workflows take raw un-skull-stripped brain volumes for multiple subjects (1,000’s) from several groups, or a Study-Design, and generate a scene files containing the models of the ROIs where the groups are different (globally, per ROI, or locally, per vertex on the mean shapes)

- **Informatics/Genomics**
  - Integrated genomics data analysis Protocols
  - Interoperable Tools: MAQ, SAMtools, Bowtie, GATK
  - Multiple Servers
Infrastructure - Databases

- Raw Data (e.g., imaging, genetics, phenotypic, meta-data)
- Derived Data (e.g., Atlases, models, shapes, masks, labels)
Infrastructure - Grid Computing

- Pipeline Grid manager provides an efficient control of back-end hardware computational resources
- Job submission, user management and support
  - SGE
  - Permissions
  - Ticketing
  - Tutorials
  - Batch/Pipeline
  - SVN/CVS
  - Dashboard

www.loni.ucla.edu/Resources/clustervisualization
# Computational Infrastructure

<table>
<thead>
<tr>
<th>Description</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Grid</strong></td>
<td></td>
</tr>
<tr>
<td>Number of Grid Nodes</td>
<td>380 nodes / 1,256 cores</td>
</tr>
<tr>
<td>RAM</td>
<td>8 – 16 Gigabytes / node</td>
</tr>
<tr>
<td>Speed</td>
<td>2.5+ GHZ per core</td>
</tr>
<tr>
<td>Specs</td>
<td>Sun V20z and Sun X2200</td>
</tr>
<tr>
<td>Usage Stats</td>
<td>~16,000 average jobs completed/day (past 3 months)</td>
</tr>
<tr>
<td>Number Users</td>
<td>165 unique users (past 3 months)</td>
</tr>
<tr>
<td><strong>Networking</strong></td>
<td></td>
</tr>
<tr>
<td>Specs</td>
<td>Mixed 1GB production and 10GB HPC networks</td>
</tr>
<tr>
<td>Usage</td>
<td>Average: 20GB/sec. Max: 80GB/sec</td>
</tr>
<tr>
<td>Bandwidth</td>
<td>100Gb+ total throughput to cluster</td>
</tr>
<tr>
<td><strong>Disks</strong></td>
<td></td>
</tr>
<tr>
<td>Capacity (online/offline)</td>
<td>250TB online capacity w/ 4PB+ Offline (tape) virtual storage</td>
</tr>
<tr>
<td>Specs (latency, bandwidth)</td>
<td>Peak max 3 Gigabytes/sec</td>
</tr>
<tr>
<td>Number of Files</td>
<td>10,000,000,000’s</td>
</tr>
<tr>
<td><strong>Web Services</strong></td>
<td></td>
</tr>
<tr>
<td>IDA</td>
<td>1,000’s users per week</td>
</tr>
<tr>
<td>iTools</td>
<td>100’s users per week</td>
</tr>
<tr>
<td>Pipeline - web-server</td>
<td>100’s users per week</td>
</tr>
<tr>
<td><strong>Pipeline</strong></td>
<td></td>
</tr>
<tr>
<td>Queue</td>
<td>pipeline.q</td>
</tr>
<tr>
<td>Usage</td>
<td>~12,000 avg jobs completed/day (past 3 months)</td>
</tr>
<tr>
<td>Node Allocation</td>
<td>Dynamic, approximately 75% of LONI’s HPC Resources</td>
</tr>
<tr>
<td>Users/Accounts</td>
<td>700+ authenticated users</td>
</tr>
<tr>
<td><strong>IDA (database)</strong></td>
<td></td>
</tr>
<tr>
<td>number of projects</td>
<td>55</td>
</tr>
<tr>
<td>number of users</td>
<td>&gt;1,200</td>
</tr>
<tr>
<td>number of volumes</td>
<td>DTI: 2,748; fMRI: 1,569; HISTO: 4; MRA: 1,204; MRI: 56,248; PET: 2,678</td>
</tr>
<tr>
<td>disk-space</td>
<td>1PB</td>
</tr>
<tr>
<td>Average Monthly Uploads (2009)</td>
<td>1,200</td>
</tr>
<tr>
<td>Average Monthly Downloads (2009)</td>
<td>25,000</td>
</tr>
</tbody>
</table>
Integrated MAQ, SAMtools, Bowtie Workflow

Folded Pipeline Workflow
(Abstracting detailed calculations)
Integrated MAQ, SAMtools, Bowtie Workflow

Unfolded Pipeline Workflow
(Illustrating calculation details)
Interactive Hands-on Pipeline Demo - mrFAST

- **Pipeline Web-Start (PWS)**
  [http://pipeline.loni.ucla.edu/PWS](http://pipeline.loni.ucla.edu/PWS)

- **Workflows Location**
  [http://pipeline.loni.ucla.edu/PWS](http://pipeline.loni.ucla.edu/PWS)
  [www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics](www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics)
  [www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMRFAST](www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMRFAST)

- **Load Workflows and run on PWS Server**
  - Open the Workflow
    - mrFAST_Indexing_Mapping.pipe
  - Connect to PWS server (should be auto-connected as guest)
    - pws.loni.ucla.edu
  - Tools ➔ Change Server to PWS Server
  - Click the Run button to execute workflow
  - Inspect results (right-click on Mapping module, View Output Files)
Interactive Hands-on Pipeline Demo - mrFAST
Interactive Hands-on Pipeline Demo - miBLAST

**Pipeline Web-Start (PWS)**
http://pipeline.loni.ucla.edu/PWS

**Workflows Location**
http://pipeline.loni.ucla.edu/PWS
www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics
www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoBLAST

**Load Workflows and run on PWS Server**
- Open the Workflow
  - miBLAST_Workflow.pipe
- Connect to PWS server (should be auto-connected as guest)
  - pws.loni.ucla.edu
- Tools → Change Server to PWS Server
- Click the Run button to execute workflow
- Inspect results (right-click on NCBI-BLAST module, View Output Files)
Interactive Hands-on Pipeline Demo - miBLAST

Documentatwn: http://www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoBLAST

NCBI BLAST
Complete 1 job in 5 seconds

DB Formatting
Complete 1 job in 978 milliseconds

DB Index Key
Complete 1 job in 5 seconds

miBLAST
Complete 1 job in 5 seconds

BLAST Search Word Size
Optional 0-40 seconds

Source Directory
cebelrh5.data.cluster.loni.ucla.edu:8008

Execution Location
cebelrh5.data.cluster.loni.ucla.edu:8008

BLAST FASTA DB File
cebelrh5.data.cluster.loni.ucla.edu:8008

MakeDir
Optional 1 job in 0 seconds

Initialize BLAST
Complete 1 job in 5 seconds

BLAST FASTA Query Instructions
cebelrh5.data.cluster.loni.ucla.edu:8008

BLAST Filtering
Complete 1 job in 1 second

BLAST Results
Complete 1 job in 1 second

NCBI BLAST
Complete 1 job in 5 seconds

Sequences producing significant alignments:
(bits) Value gi|1790777|gb|AE000503.1|AE000503
Escherichia coli K-12 MG1655 s... 26 0.69 gi|2367246|gb|AE000436.1|AE000436 Escherichia coli K-12 MG1655 s... 26 0.69 gi|1788338|gb|AE000294.1|AE000294 Escherichia coli K-12 MG1655 s... 26 0.69 gi|1786819|gb|AE000166.1|AE000166 Escherichia coli K-12 MG1655 s... 26
Interactive Hands-on Pipeline Demo - Genomics Tools Interoperability

- Pipeline Web-Start (PWS)
  http://pipeline.loni.ucla.edu/PWS

- Workflows Location
  www.loni.ucla.edu/twiki/bin/view/CCB/PipelineWorkflows_BioinfoMAQ

- Load Workflows and run on PWS Server
  - Open the Workflow:
    MAQ_SAMtools_Bowtie_Integrated_Cranium.pipe
  - Connect to PWS server (should be auto-connected as guest)
    - pws.loni.ucla.edu
  - Tools ➔ Change Server to PWS Server
  - Click the Run button to execute workflow
  - Inspect results (right-click on NCBIBLAST module, View Output Files)
Additional Interactive Hands-on Pipeline Demos are available Online

- **Workflows Location**
  
  www.loni.ucla.edu/twiki/bin/view/LONI/Pipeline_GenomicsInformatics
  
  www.MyExperiment.org/workflows
Acknowledgments

- **Collaborators**
  - UCLA LONI: Arthur Toga, Alen Zamanyan, Alex Genco, Sam Hobel, LONI Pipeline Team: Petros Petrosyan, Zhizhong Liu, Paul Eggert
  - UCI: Fabio Macciardi, Federica Torri, Harry Mangalam
  - USC: Andrew Clark, Jim Knowles, Ben Berman, Zack Ramjan
  - BIRN: Joseph Ames, Carl Kesselman

- **Funded** by National Institutes of Health
  - U54 RR021813, P41 RR013642, R01 MH71940, U24-RR025736, U24-RR021992, U24-RR021760 and U24-RR026057

- **Other contributions from**
  - Members of the Laboratory of Neuro Imaging (LONI)
  - Biomedical Informatics Research Network (BIRN)
  - National Centers for Biomedical Computing (NCBC)
  - Clinical and Translational Science Award (CTSA) investigators

- **Publications/Citations:** [http://pipeline.loni.ucla.edu/downloads/acknowledgmentscredits](http://pipeline.loni.ucla.edu/downloads/acknowledgmentscredits)
Questions, Comments, Critiques

- **Forum:** [http://Pipeline.loni.ucla.edu/forum](http://Pipeline.loni.ucla.edu/forum)
- **URL:** [http://Pipeline.loni.ucla.edu](http://Pipeline.loni.ucla.edu)
- **Email:** Ivo.Dinov@loni.ucla.edu