iPlant:
Cyberinfrastructure for Plant Sciences (and Beyond)

Nirav Merchant
nirav@email.arizona.edu
One Big Problem...
One Big Problem...

2008

2011
One Big Problem...
One Big Problem...

- Transfer
- Storage
- Analysis
- Visualization
- Metadata Mark-up
- Search and Discover
- Share/Collaborate
- Publish
got data?
Flavors of Bio-Information

- Sequences (most popular)
- Structures
- Images
- Video
- Audio
- Pathways (graphs)
- Text (Publications)
- Traces
- Combination (eg Video & Traces)
- And much more …
Growth

• The Human genome project (1990-2003)
  • 13 years • 23 labs • $500 Million.

• A Human genome today (2010)
  • ~3 days • ~1 machine. • ~$10,000.

• Many 1000 and 10,000 genomes projects (human, plant etc)

• Bench-top sequencers are

• “Next-next-next” generation sequencers will produce $500 genome in 5 years
High-throughput Data Acquisition
High-throughput Data Acquisition

In 11 Days
High-throughput Data Acquisition

In 11 Days

• Generates 4TB of raw data
High-throughput Data Acquisition

In 11 Days
• Generates 4TB of raw data
• 600,000,000,000 bases of DNA sequence
High-throughput Data Acquisition

In 11 Days
• Generates 4TB of raw data
• 600,000,000,000 bases of DNA sequence
  • (200 human genomes)
High-throughput Phenotyping
(Watching Grass Grow)

• $70K for ~30 camera sets
• ~200 movies of plants undergoing a dynamic growth process
• “Only” 4GB a day
iPlant Layered Services and Access

Community Facing Resources

- iPlant Discovery Environment
- Educational Interface
- External Access
- User-created Applications

Public APIs and Semantic Web Services

- Job
- I/O
- Data
- Profile
- Auth
- Event
- Apps

Low-Level Services, Security, Access, etc.

- iRODS (Federated Storage)
- Shibboleth (Single Sign-on)
- Globus/Unicore (Grid Resources)
- TeraGrid XSEDE

iPlant Hardware Resources

- Cloud Systems
- High Performance Computers
- Databases
- Storage
iPlant Layered Services and Access

Community Facing Resources

- iPlant Discovery Environment
- Educational Interface
- External Access
- User-created Applications

Public APIs and Semantic Web Services

<table>
<thead>
<tr>
<th>Job</th>
<th>I/O</th>
<th>Data</th>
<th>Profile</th>
<th>Auth</th>
<th>Event</th>
<th>Apps</th>
</tr>
</thead>
</table>

Low-Level Services, Security, Access, etc.

- iRODS (Federated Storage)
- Shibboleth (Single Sign-on)
- Globus/Unicore (Grid Resources)
- TeraGrid XSEDE

iPlant Hardware Resources

- Cloud Systems
- High Performance Computers
- Databases
- Storage
Different Users, Different Access Needs:
One Data Store
Welcome!

You've reached the login page for the Discovery Environment. The Discovery Environment integrates powerful, community-recommended software tools into a system that:

- Can handle terabytes of data
- Utilizes high-performance supercomputing resources as needed
- Hides the complexity needed to do these tasks.

©2011 iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation (#DBI-0735191).
Welcome!

You've reached the login page for the Discovery Environment. The Discovery Environment integrates powerful, community-recommended software tools into a system that:

- Can handle terabytes of data
- Utilizes high-performance supercomputing resources as needed
- Hides the complexity needed to do these tasks.

©2011 iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation (DBI-0735191).
Welcome!

You've reached the login page for the Discovery Environment. The Discovery Environment integrates powerful, community-recommended software tools into a system that:

- Can handle terabytes of data
- Utilizes high-performance supercomputing resources as needed
- Hides the complexity needed to do these tasks.

©2011 iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation (MDI-0735191).
Welcome!
You've reached the login page for the Discovery Environment. The Discovery Environment integrates powerful, community-recommended software tools into a system that:

- Can handle terabytes of data
- Utilizes high-performance supercomputing resources as needed
- Hides the complexity needed to do these tasks.

©2011 iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation (DBI-0735191).

data

Welcome!
You've reached the login page for the Discovery Environment. The Discovery Environment integrates powerful, community-recommended software tools into a system that:

- Can handle terabytes of data
- Utilizes high-performance supercomputing resources as needed
- Hides the complexity needed to do these tasks.

©2011 iPlant Collaborative. The iPlant Collaborative is funded by a grant from the National Science Foundation (DBI-0735191).
iPlant Data Store

Free Your Data

WebDAV

DE

API

i-drop

i-commands

Thursday, March 1, 2012
Atmosphere: Servers and Software on Demand

http://atmosphere.iplantcollaborative.org

Use Your iPlant Credentials
Atmosphere

http://atmosphere.iplantcollaborative.org

**Flapjack**: New software tools for graphical genotyping and haplotype visualization are required that can routinely handle the large data volumes generated by high throughput SNP and comparable genotyping technologies. Flapjack is a new visualization tool to facilitate analysis of these data types. http://bioinf.scri.ac.uk/flapjack/

**Tablet**: Tablet is a lightweight, high-performance graphical viewer for next-generation sequence assemblies and alignments. Supporting a range of input assembly formats, Tablet provides high-quality visualizations showing data in packed or stacked views, allowing instant access and navigation to any region of interest, and whole contig overviews and data summaries. Tablet is both multi-core aware and memory efficient, allowing it to handle assemblies containing millions of reads, even on a 32-bit desktop machine. http://bioinformatics.oxfordjournals.org/content/26/3/401.abstract

**Taverna-workbench**: Taverna is an open source and domain independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation. http://www.taverna.org/

**Kepler**: Kepler is a free-software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows. Kepler’s facilities provide process and data monitoring, provenance information, and high speed data movement solutions. Workflows in general, and scientific workflows in particular, are directed graphs where the nodes represent discrete computational components, and the edges represent paths along which data and results can flow between components. http://en.wikipedia.org/wiki/Kepler_scientific_workflow_system
Atmosphere
http://atmosphere.iplantcollaborative.org

Biotools

Flapjack: New software tools for genome and proteome analysis. Flapjack is a new visualization tool to handle the large data volumes generated from high-throughput sequencing technologies. Flapjack is a lightweight, high-performance tool for displaying data in packed or stacked views, allowing instant access and navigation to any region of interest, and whole contig overviews and data summaries. Flapjack is both multi-core aware and memory efficient, allowing it to handle assemblies containing millions of reads, even on a 32-bit desktop machine. http://bioinformatics.oxfordjournals.org/content/26/3/401.abstract

Tablet: Tablet is a lightweight, high-performance tool for displaying data in packed or stacked views, allowing instant access and navigation to any region of interest, and whole contig overviews and data summaries. Tablet is both multi-core aware and memory efficient, allowing it to handle assemblies containing millions of reads, even on a 32-bit desktop machine. http://bioinformatics.oxfordjournals.org/content/26/3/401.abstract

Taverna-workbench: Taverna is an open source and domain independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation. http://www.taverna.org/

Kepler: Kepler is a free-software system for designing, executing, reusing, evolving, archiving, and sharing scientific workflows. Kepler's facilities provide process and data monitoring, provenance information, and high speed data movement solutions. Workflows in general, and scientific workflows in particular, are directed graphs where the nodes represent discrete computational components, and the edges represent paths along which data and results can flow between components. http://en.wikipedia.org/wiki/Kepler_scientific_workflow_system
Atmosphere

http://atmosphere.iplantcollaborative.org

Flapjack: New software tools for graphical alignment
Tablet: Tablet is a lightweight, high-performance alignment.
Taverna-workbench: Taverna is an open-source environment for designing and executing scientific workflows.

From: atmo@iplantcollaborative.org
Subject: Your Atmosphere Cloud Instance
Date: August 4, 2011 5:26:35 PM MST
To: Eric Lyons

Your Atmosphere cloud instance is ready.
Instance id: i-3C8206D5
IP: 150.135.78.154
SSH Username: elyons
SSH Password:

Atmosphere, iPlant Collaborative
“Powered by iPlant CI”

- Make it easy for community to develop applications based on iPlant CI
- Authentication
- Data store
- Compute
- Applications can be distributed or hosted by iPlant
- Provide developer with tool kit to get started
Training end users!!

Uploading data

There are several ways to upload your data. These include:

- Directly to the Discovery Environment from your desktop (for files under 2 GB)
- Directly to the Discovery Environment from a URL (for all files)
- Using the DAVIS Web Interface (for files under 2 GB) (video tutorial)
- Using the iRODS web client (file size limit unknown, may also be limited by web browser limit of 2 GB)
- Using WebDAV (file size limit unknown)
- Using iDrop (file size limit unknown) (video tutorial)
- Using icommands (ideal for bulk transfers and best for large data and files over 2 GB)
- From Atmosphere using FUSE (useful if you have already been working in a virtual machine (VM) environment)

Colors indicate relative difficulty: green=easier, yellow=intermediate, red=more difficult. For most users, we recommend green, th

Accessing data

There are several ways to access your data. These include:

- Directly from the Discovery Environment: viewing, managing, deleting
- Using the Davis web interface
- Using the iRODS web client
- Using WebDAV
- Using iDrop
- From Atmosphere virtual machines using FUSE
- Using icommands
- Installing the iRODS FUSE client for Linux to create a mountable volume
Training end users !!

iPlant Tools and Services Workshop, UC Davis. March 12th and 13th 2012

Date and Time

<table>
<thead>
<tr>
<th>Date</th>
<th>Time</th>
<th>Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>March 12</td>
<td>8:00am - 2:30pm</td>
<td>SciLab 2060</td>
</tr>
<tr>
<td>March 13</td>
<td>9:30am - 4:00pm</td>
<td>Genome Center 4202</td>
</tr>
</tbody>
</table>

Location

University of California, Davis

Details

Registration for this event is CLOSED.

Biological research depends increasingly on high-throughput data collection methods and complex computational analyses. This free two-day workshop provides a comprehensive look at the tools and services provided by the iPlant Collaborative, a major cyberinfrastructure project of the National Science Foundation (www.iplantcollaborative.org). Workshop topics build progressively to cater to the needs of general and advanced audiences.
Training end users !!

Agenda

Monday, March 12th, 2012

<table>
<thead>
<tr>
<th>Time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>08:00 AM - 08:30 AM</td>
<td>Arrive / Sign-in / Verify iPlant Accounts</td>
</tr>
<tr>
<td>08:30 AM - 09:00 AM</td>
<td>Welcome - Presenter/Participant Self-Introductions</td>
</tr>
<tr>
<td>09:00 AM - 09:20 AM</td>
<td>An Overview of the iPlant Collaborative</td>
</tr>
<tr>
<td>09:20 AM - 09:45 AM</td>
<td>Overview of the iPlant Discovery Environment (DE)</td>
</tr>
<tr>
<td>09:45 AM - 10:00 AM</td>
<td>Break</td>
</tr>
<tr>
<td>10:00 AM - 10:30 AM</td>
<td>iPlant Data Store - Managing &quot;Big Data&quot;</td>
</tr>
<tr>
<td>10:30 AM - 11:30 AM</td>
<td>Using the DE to Examine Differential Expression Within an RNA-seq Dataset</td>
</tr>
<tr>
<td>11:30 AM - 12:00 PM</td>
<td>Atmosphere - Custom Cloud Computing</td>
</tr>
<tr>
<td>12:00 PM - 12:30 PM</td>
<td>Lunch</td>
</tr>
<tr>
<td>12:30 PM - 01:00 PM</td>
<td>Building and Using Workflows Within the DE; Phylogenetics</td>
</tr>
<tr>
<td>01:00 PM - 01:45 PM</td>
<td>Extending the DE for Your Research: Tool Integration and Customization</td>
</tr>
<tr>
<td>01:45 PM - 02:30 PM</td>
<td>Collaborating with iPlant: Future Projects and Workshop Summary</td>
</tr>
</tbody>
</table>

Tuesday, March 13th, 2012

<table>
<thead>
<tr>
<th>Time</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>09:30 AM - 11:30 AM</td>
<td>Advanced ChiP-seq in the iPlant Cyberinfrastructure</td>
</tr>
<tr>
<td>11:30 AM - 12:00 PM</td>
<td>Comparative Phylogenetic Methods in the iPlant Cyberinfrastructure</td>
</tr>
<tr>
<td></td>
<td>Data Clinic Appointments</td>
</tr>
<tr>
<td>01:30 PM - 02:00 PM</td>
<td>Powered by iPlant - Consuming iPlant Services in Your Portals</td>
</tr>
<tr>
<td>02:00 PM - 02:30 PM</td>
<td>Using XSEDE for Bioinformatics</td>
</tr>
</tbody>
</table>
Future directions

• More data driven collaboration interfaces (tickets/tokens*)
• Deeper Metadata handling (extending AVU)
• Content/Repository searching (Solr)
• Data enrichment, decoration (enhanced markup, semantic integration)
• Social computing (Data de-duplication, suggestion engine)
How to get access:
www.iplantcollaborative.org
How to get access:
www.iplantcollaborative.org
iPlant APIs
Resources
iPlant APIs
Resources
Cyberinfrastructure for Life Sciences
Scalable
Capable
Extensible

iPlant APIs
Resources
Where to Get More Information

• Register:
• iPlant DE:
• iPlant Data Store:
• iPlant Atmo:
• iPlant Quick-Start Guides:
• iPlant Tutorials:
• iPlant Forums:
• iPlant Wiki:
iPlant’s Building Blocks

Executive Team:
Steve Goff
Dan Stanzione

Postdocs:
Barbara Banbury
Jamie Estill
Bindu Joseph
Christos Noutsos
Brad Ruhfel
Stephen A. Smith
Chunlao Tang
Lin Wang
Liya Wang
Norman Wickett

Students:
Peter Bailey
Jeremy Beaulieu
Devi Bhattacharya
Storm Briscoe
Ya-Di Chen
John Donoghue
Steven Gregory
Yekatarina Khartianova
Monica Lent
Amgad Madkour
Aniruddha Marathe
Kurt Michaels
Dhanesh Prasad
Andrew Predoehl
Jose Salcedo
Shalini Sasidharan
Gregory Striemer
Jason Vandeventer
Kuan Yang

Metadata
Data
Tools
Workflows
Viz

Faculty Advisors & Collaborators:
Ali Akoglu
Greg Andrews
Kobus Barnard
Sue Brown
Thomas Brutnell
Michael Donoghue
Casey Dunn
Brian Enquist
Damian Gessler
Ruth Grene
John Hartman
Matthew Hudson
Dan Kliebenstein
Jim Leebens-Mack
David Lowenthal
Robert Martienssen
B.S. Manjunath
Nirav Merchant
David Neale
Brian O’Meara
Sudha Ram
David Salt
Mark Schildhauer
Doug Soltis
Pam Soltis
Edgar Spalding
Alexis Stamatakis
Ann Stapleton
Lincoln Stein
Val Tannen
Todd Vision
Doreen Ware
Steve Welch
Mark Westneat

Staff:
Greg Abram
Sonali Aditya
Roger Barthelson
Brad Boyle
Todd Bryan
Gordon Burleigh
John Cazes
Mike Conway
Karen Cranston
Rion Dooley
Andy Edmonds
Dmitry Fedorov
Michael Gatto
Utkarsh Gaur
Cornel Ghiban
Michael Gonzales
Harolf Häfele
Matthew Hanlon
Anthony Heath
Barbara Heath
Matthew Helmke
Natalie Henriques
Uwe Hilgert
Nico Hopkins
Eun-Sook Jeong
Logan Johnson
Chris Jordan
B.D. Kim
Kathleen Kennedy
Mohammed Khalfan
Seung-jin Kim
Lars Koersterk
Sangeeta Kuchimanchi
Kristian Kvillekval
Aruna Lakshmanan
Sue Lauter
Tina Lee
Andrew Lenards
Zhenyuan Lu
Eric Lyons
Naim Matasci
Sheldon McKay
Robert McLaey
Angel Mercer
Dave Micklos
Nathan Miller
Steve Mock
Martha Narro
Praveen Nuthulapati
Shannon Oliver
Shiran Pasternak
William Peil
Titus Purdin
J.A. Raygoza Garay
Dennis Roberts
Jerry Schneider

Andrew Lenards
Zhenyuan Lu
Eric Lyons
Naim Matasci
Sheldon McKay
Robert McLaey
Angel Mercer
Dave Micklos
Nathan Miller
Steve Mock
Martha Narro
Praveen Nuthulapati
Shannon Oliver
Shiran Pasternak
William Peil
Titus Purdin
J.A. Raygoza Garay
Dennis Roberts
Jerry Schneider

Bruce Schumaker
Sriramu Singaram
Edwin Skidmore
Brandon Smith
Mary Margaret Sprinkle
Sriram Srinivasan
Josh Stein
Lisa Stillwell
Kris Urie
Peter Van Buren
Hans Vasquez-Gross
Matthew Vaught
Fusheng Wei
Jason Williams
John Wregglesworth
Weijia Xu
Jill Yarmchuk
Managing and Integrating: Data, Tools, Analysis
Managing and Integrating: Data, Tools, Analysis