The iPlant Collaborative: Cyberinfrastructure for the Plant Sciences

Steve Goff
iPlant Overview
Iowa State University
May 13th, 2010
What is iPlant?

- iPlant’s mission is to build the CI to support plant biology’s Grand Challenge solutions
- Grand Challenges were not defined in advance, but identified through engagement with the community
- A virtual organization with Grand Challenge teams relying on national cyberinfrastructure
- Long term focus on sustainable food supply, climate change, biofuels, ecological stability, etc
- Hundreds of participants globally… Working group members at >50 US institutions, USDA, DOE, etc.
iPlant Collaborative Org

- National Science Foundation
  - Community Grand Challenge Teams
  - Executive Team
    - Cyberinfrastructure Development & Engagement Teams
    - Administrative Support Team
    - Science Opportunities Teams
    - Education, Outreach, and Training Team
  - Board of Directors
  - Internal Advisory Board
Why Cyberinfrastructure is Needed

Global Trends

- Population
- Crop Production
- Crop Uses

Drivers in Biology Research:

- Biology research is moving fast, lots of data now
- Not enough time to keep up in multiple fields
- Pressure to be narrowly focused
- Larger problems need a broad perspective
- Collaboration across fields is needed
- Synthesis of information is important
Global Population Estimates

8000 BC - 5 million – Start of Ag
1650 - 500 million
1850 - 1 billion
1930 - 2 billion
1975 - 4 billion
1996 - 5.8 billion
2000 - 6 billion
2025 - 8 billion?
2035 - 10 billion?
2065 - 14 billion?

> 1,000-fold Increase
Worldwide Crop Yields

- 99% of crop production from 24 species
- 60-70% from Rice, Corn, & Wheat
- Doesn’t include sugar

[Bar chart showing yields of various crops, with the top three (Rice, Corn, Wheat) having much higher yields compared to others.]
World Grain Production
(million tonnes)
Yield Gains Are Slowing

(Percent change)
World Grain Area

(million hectares)
Crops for Renewable Energy Increases Demand

**US grain feeding cars**

<table>
<thead>
<tr>
<th>Million tonnes of US grain used to make ethanol for cars</th>
<th>Millions of people the grain could feed</th>
</tr>
</thead>
<tbody>
<tr>
<td>100</td>
<td>350</td>
</tr>
<tr>
<td>20</td>
<td>300</td>
</tr>
<tr>
<td>30</td>
<td>250</td>
</tr>
<tr>
<td>40</td>
<td>200</td>
</tr>
<tr>
<td>50</td>
<td>150</td>
</tr>
<tr>
<td>60</td>
<td>100</td>
</tr>
<tr>
<td>70</td>
<td>50</td>
</tr>
<tr>
<td>80</td>
<td>0</td>
</tr>
</tbody>
</table>

**Source:** Earth Policy Institute, USDA, UN
Changing Lifestyles

- Activity is being engineered out of daily life
- Adults eat equivalent of 4 ½ meals a day
- US Food production @ 3,800 calories /person/day
- Need ~2,000 calories/day at current activity level
Cereal Consumption

Rice in Asia = 0.9-1.1 lb/day/person

Corn in US = ~3.3 lb/day/person *

Wheat in Europe = 0.8-1.2 lb/day/person

* Includes entire food chain

Note: Cereal Production is ~550-600 Million tons/year

Sugar Production is 1 Billion tons/year
Keeping up with Science

Publications in PubMed May 10th, 2010:

- United States 7.6 Million
- England 2.9 Million
- Germany 1.2 Million
- Japan 606k
- France 491k
- Korea 243k
- China 227k
- India 116k
- In 2009 Globally 846K (>2,300 pubs per day)

How many papers can anyone read in a day?
Publications over time: PubMed 1950 to 2009

1950                                                    1990                    2009
<table>
<thead>
<tr>
<th>Organism</th>
<th>ESTs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Human</td>
<td>8,301,471</td>
</tr>
<tr>
<td>Mouse</td>
<td>4,852,146</td>
</tr>
<tr>
<td>Maize</td>
<td>2,018,857</td>
</tr>
<tr>
<td>Pig</td>
<td>1,621,000</td>
</tr>
<tr>
<td>Cattle</td>
<td>1,559,485</td>
</tr>
<tr>
<td>Arabidopsis</td>
<td>1,527,299</td>
</tr>
<tr>
<td>Zebrafish</td>
<td>1,481,930</td>
</tr>
<tr>
<td>Soybean</td>
<td>1,422,983</td>
</tr>
<tr>
<td>Xenopus</td>
<td>1,271,375</td>
</tr>
<tr>
<td>Rice</td>
<td>1,249,110</td>
</tr>
</tbody>
</table>
Why Cyberinfrastructure is Needed

Fundamental Questions:

- Are there particular conditions that select for novelty and high mutation or recombination rates?
- What is the relationship between specific viral genes and the genes of other organisms?
- Are there conditions that select for cooperative behavior?
- What features convey robustness to systems?
- How does robustness trade off against adaptability?
- At what level does evolutionary selection operate most strongly?

Foundation for Functional Annotation

- **E coli**: 62%
- **Yeast**: ~30%
- **Arabidopsis**: ~15%
- **Rice or maize**: ~1%

= total genes
= genes with experimentally-demonstrated function
Arabidopsis 2010 Project

Goals:
- Complete Genome Sequence
- Develop Genomics Tools to Facilitate Research
  - Knockout lines
  - Overexpression lines (genes and RNAi)
  - Markers
  - Arrays
- Understand Function of All Genes

Genome Complete in 2000
- Only 8% of genes with experimental evidence

Most progress made via HT Genomic Approaches Not via Focus on Individual Genes or Families (2005 Report)

Now developing 2020 project!
Known Maize Genes

- Genome draft released in 2008
- Mutants of Maize - 938 mutants described*
  - 3% of genes assuming 30k total
- Maize Alleles ~2350 in maize genetics db**
  - Many alleles in same or duplicate genes
- 1,060 Phenotypes

*Neuffer, Coe, & Wessler 1997 edition
** www.maizegdb.org
Domestication Decreases Genetic Diversity

Wild Species $\rightarrow$ Early Domesticated Species $\rightarrow$ Modern Varieties

How can we accelerate introduction and use of genetic diversity?

Tanksley and McCouch
Science 1997 August 22; 277: 1063-1066
What is a Gene Post-ENCODE?

- Assumed most genes encode proteins
- 1.2% of human genome encodes proteins
- 85-97% of human genome is transcribed
- 3.8% of non-coding regions conserved
- 2/3 of non-rRNA in cytoplasm is non-coding
- ENCODE = 1% of Human Genome
- FANTOM = Similar Mouse Genome Study
New View on Gene Architecture

Mattick Science 309, p1527 (2005)
Examples of Many Gene Products

Mark B. Gerstein et al. Genome Res. 2007; 17: 669-681
Example of Overlapping and Opposite Strand Genes

10 Products
One strand

3 Products
Opposite strand

4 Genes or 13 Genes?

Mark B. Gerstein et al. Genome Res. 2007; 17: 669-681
Some genes are trans-spliced

KIAA0510 - 5,596 NT
FLJ00128 - 5,178 NT

Chromosome 14
Chromosome 1

Claverie Science 309; p1529 (2005)
New Classes of RNAs

• Types of RNAs

We knew about these:

Coding RNA
- mRNA
- Transcription RNAs
  - rRNA
  - tRNA

Noncoding RNAs
- Small RNAs
  - siRNA
  - miRNA
  - snoRNA
  - snRNA
Fairly New Classes of RNAs

- Types of RNAs

Coding RNA
  - mRNA
  - Transcription RNAs
    - rRNA
    - tRNA

Noncoding RNAs
  - Small RNAs
    - siRNA
    - miRNA
    - snoRNA
    - snRNA

Not much about these
Tiling Arrays vs Expression Arrays

- Tens of Thousands of genes
- Hundreds of Thousands of Oligos
- 10-20 Oligos / gene, designed from protein coding genes
- ~10k chips of data
- Fluorescent probes from polyA+ RNA
- Multiple RNAs (sense and nonsense) for each gene?
- Is variation between “Features” telling us something?

Gene 1

Gene 2
Probe Hybridization Signal as a Function of Gene Position

Target Gene

Gene code Ctrl_ZMU29158-M_x_at

Id1, SS1, hybrid
The iPlant Collaborative

Single National Collaborative funded by NSF (Emerging Frontiers, PSCIC)
Partners:
  - University of Arizona
  - Cold Spring Harbor Laboratory
  - Texas Advanced Computing Center
  - University of North Carolina, Wilmington

Community-appointed Board of Directors
  - Chair: Gwen Jacobs, Montana State
NSF Cyberinfrastructure Vision

- High Performance Computing
- Data and Data Analysis
- Virtual Organizations
- Learning and Workforce (*Critical to iPlant, but beyond the scope of this talk*)

Scientific Cyberinfrastructure Examples

GEON – Distributed Geoscience CI (http://www.geongrid.org/)
National Virtual Observatory – Astronomy (http://www.us-vo.org/)
NEES – Network for Earthquake Engineering & Simulation (http://www.nsf.gov/news/special_reports/nees/about.jsp)
Cactus – Astrophysics (http://www.cactuscode.org)
NanoHub – Nanotechnology (http://www.nanohub.org/)
LSST - Large Synoptic Survey Telescope (http://www.lsst.org/lsst)

Biological Cyberinfrastructure

BeeSpace – Integrative Biology (http://www.beespace.uiuc.edu/)
NESCent – National Evolutionary Synthesis Center (http://www.nescent.org/index.php)
NCEAS – National Center for Ecological Analysis & Synthesis (http://www.nceas.ucsb.edu/)
NIMBioS – National Institute for Mathematical & Biological Synthesis (http://www.nimbios.org)
Brief History

- Formally approved by National Science Board – 12/2007
- Funding by NSF – February 1st, 2008
- iPlant Kickoff Conference at CSHL – April 2008
  - ~200 participants
- Grand Challenge Workshops – Sept-Dec 2008
- CI workshop – Jan 2009
- Project Recommendations – March 2009
- Project Kickoffs – May 2009 & August 2009
- First Release of Discovery Environments – April 2010
Grand Challenge & CI Workshops

- Mechanistic Basis of Plant Adaptation (9-30-08)
- Impact of Climate Change on Plant Productivity: Prediction of Phenotype from Genotype (9-30-08)
- Developing common models for molecular mechanisms, crop physiology, and ecology (11-7-08)
- Assembling the Tree of Life to Enable the Plant Sciences (11-19-08)
- Computational Morphodynamics of Plants (12-15-08)
- Botanical Information & Ecology Network
- CI Workshop
GC Projects Recommended by the iPlant Board of Directors
March 2009

Initial Grand Challenge Projects:
Plant Tree of Life – iPToL – May ‘09
  + Taxonomic Intelligence
  + APWeb2
  + Social Networking Website

Genotype to Phenotype – iPG2P – Aug ‘09
  + Image Analysis Platform
  + GLM by GPGPU and FPGA
iPlant Tree of Life Working Groups

- **Trait Evolution**, Brian Omeara
  - Post-tree analysis and mapping of ancestral traits
- **Tree Reconciliation**, Todd Vision
  - Large-scale reconciliation of gene trees, co-evolving parasites, etc., with species trees
- **Big Trees**, Alexandros Stamatakis
  - HPC Phylogenetic inference with 500K taxa
- **Tree Visualization** Michael Sanderson; Karen Cranston
  - Cross cutting group for the viz needs of all
- **Data Integration**, Val Tannen, Bill Piel
  - Cross cutting group for the data integration needs of all
- **Data Assembly**, Doug Soltis, Pam Soltis, Michael Donoghue
  - Community and network building, data assembly
iPlant Genotype to Phenotype Working Groups

- **NextGen Sequencing**
  - Establishing an informatics pipeline that will allow the plant community to process NextGen sequence data

- **Statistical Inference**
  - Developing a platform using advanced computational approaches to statistically link genotype to phenotype

- **Modeling Tools**
  - Developing a framework to support tools for the construction, simulation and analysis of computational models of plant function at various scales of resolution and fidelity

- **Visual Analytics**
  - Generating, adapting, and integrating visualization tools capable of displaying diverse types of data from laboratory, field, in silico analyses and simulations

- **Data Integration**
  - Investigating and applying methods for describing and unifying data sets into virtual systems that support iPG2P activities
First Release Discovery Environment

Beta Release of iPlant Discovery Environment Coming Soon!

iPlant Co-PI Dan Staniziale announced that the first public beta release of the iPlant Collaborative's discovery environment (DE) software will take place in the first quarter of 2010. With this release, the iPlant DE will:

- demonstrate the component model and framework for iPlant DEs;
- demonstrate the new iPlant user interface (UI) and UI standards;
- serve as the basis for all future development and tool integration in the DEs;
- debut the authorization and authentication framework to be used by iPlant;
- demonstrate the ability of the DEs to run jobs on TeraGrid and other large systems;
- demonstrate integration of existing bioinformatics tools into an integrated workflow; and
- debut collaboration facilities of DEs.

Beyond serving as a reference technology platform for all future iPlant DEs, this first release will offer functionality that will be immediately useful to practicing biologists. In particular, this release will address a sample problem identified in the iPlant Tree of Life (IToL) Grand Challenge Project: the application of Phylogenetic Independent Contrasts (PIC) to large phylogenetic trees. Phylogenetic researchers using the first release will be immediately able to:

- upload a tree of the user's choosing in a variety of file formats, or to import a tree from the popular Treebase or Morpheusbank databases;
- edit and manipulate tree and taxa data online via the web interface;
- apply PIC analysis using the “Contrast” tool from the Phyisp package;
- visualize trees via either an integrated web-based viewer, or to select a new prototype scalable viewer now under development at iPlant;
- share results with other collaborators, including the ability of a user to grant various levels of access to other users and groups; and
- upload results to public databases, including Treebase and Morpheusbank.

This initial release will be tested and validated with trees for up to 50,000 taxa. Future releases will be designed to support up to 500,000 taxa.
CI Development & Functionality Introduction

- Developer Toolkit
- Dynamic Workflows
- Tree Reconcil’n
- Big Tree Inference
- Taxon Name Resol’n
- G to P Associations
- Distributed Computing
- Cloud Computing
- Collaboration Tools
- Big Tree Visualization
- UHT Sequencing
- Visualize Complex Data
- Semantic Data Model
- Image Analysis
- Semantic Web
- Ancestral Chara Trait R
- AP Web2

... Foundational CI

Trait Evolution

... DNA Subway
Plant Biology CI Empowerment Strategy

- Evolutionary Biology
- Plant Ecology
- Phenotyping
- Plant Genomics
- GC Solutions
Genomics in Education:
Gene Annotation and Comparison

Goals

- Enable students to learn gene, genome, evolution concepts by working with uncharacterized sequence data
- Provide a learning environment that brings annotation and comparison tools together (DNA Subway)
- Connect educators and students to projects needing analysis
This site ties together key bioinformatics tools and databases used to annotate genes and analyze genome data. Roll over any of the “stations” on the subway map to find out more about the analysis steps. Start a project by selecting one of the three “subway lines” (red, yellow, blue). Register and login if you want to save or share your results.
Tour the DNA Subway

http://dnasubway.iplantcollaborative.org/

Go to this link and log on as a “Guest” for a short tour
First ELIXIR Funding Announced
- 10M £ (~$20M) to EBI to establish core dbs
- DOE FOA for “Knowledgebase” made public
- Four objectives to be addressed by applicants:
  - Methods to integrate multiple data types
  - Methods to infer and curate (meta)genomic functional annotations
  - Methods to couple cellular pathways and processes
  - Methods to model whole cellular processes
- $5M for ~15 Awards in 2010
- Gates Foundation projects (C3/C4 rice, IBP)
Scope: What iPlant *will* do

- Provide storage, computation, hosting, & programmer effort to support grand challenge efforts
- Work with the community to support & develop standards
- Provide forums to discuss the role & design of CI in plant science
- Help organize the community to collect data
- Provide appropriate funding for time spent helping iPlant design and test the CI
Scope: What iPlant won’t do

- iPlant is not a funding agency
  - A large grant shouldn’t become a bunch of small grants
- iPlant does not fund data collection
- iPlant (probably) will not continue funding for `<favorite tool x>` whose funding is ending
- iPlant will not seek to replace online data repositories
- iPlant will not *impose* standards on the community
iPlant 2010 Conference Purpose

- Describe CI tools built to date & the future development plans
- Describe how CI will benefit plant phylogenetics & help facilitate connections between genes & traits
- Present areas of plant science outside scope of iPToL/iPG2P (physiology, adaptation, global diversity, tree biology, plant nutrition)
- Explore CI beneficial to these areas of plant science
- Demonstrate iPlant software developments made to date
- Present EOT developments (DNA Subway)
- Present iPlant activities beyond the GC Projects (APWeb2, Image Analysis Platform, Taxonomic Intelligence, Phylogenetics Social Networking, GLM solutions via GPGPU, Reproducibility of Bioinformatics Experiments, Provenance, etc)
- Solicit community input to develop a roadmap to future CI
Collaborating Institutions

- CSHL iPlant CI
- EMEC External Evaluator
- TACC iPlant CI
- UNCW iPlant CI
- Field Museum Natural History
- MoBot APWeb2
- BIEN Taxonomic Intelligence
- UCSB Image Platform
- UWISC Image Platform
- Boyce Thompson Inst. iP2G2P
- KSU iP2G2P
- UCD iP2G2P
- VA Tech iP2G2P
- Brown iPToL
- UFL iPToL
- UGA iPToL
- UPenn iPToL
- UTK iPToL
- Yale iPToL
Soft Collaborators

- 1kP Consortium
- ARS at USDA
- BRIT: Botanical Research Institute of Texas
- CGIAR and Generation Challenge Program
- Cyberinfrastructure for Phylogenetic Research (CIPRES)
- The Croquet Consortium
- NIMBioS: National Institute for Mathematical and Biological Synthesis
- Pittsburgh Supercomputing Center
- pPOD: processing PhyloData
- Syngenta Foundation
- NanoHub & HubZero
- ELIXIR
- Fluxnet.
- Howard Hughes Medical Institute
- Knowledgebase
- NPN: National Phenology Network
- PEaCE Lab: Pacific Ecoinformatics and Computational Ecology Lab
- MORPH: Research Coordination Network (RCN)
- NCEAS: National Center for Ecological Analysis and Synthesis
- NEON: National Ecological Observation Network
- NESCent: National Evolutionary Synthesis Center
Retrospective

- Two year planning period?
- Alternative Grand Challenge identification process?
- Create Foundational CI first?
- Build cross disciplinary teams first?
Acknowledgments

University of Arizona
Rich Jorgensen
Greg Andrews
Kobus Barnard
Rick Blevins
Sue Brown
Vicki Bryan
Vicki Chandler
John Hartman
Travis Huxman
Tina Lee
Nirav Merchant
Martha Narro
Sudha Ram
Steve Rounsley
Suzanne Westbrook
Ramin Yadegari

Cold Spring Harbor Laboratory, NY
Lincoln Stein
Matt Vaughn
Doreen Ware
Dave Micklos
Sheldon McKay
Jerry Lu
Liya Wang

Texas Advanced Computing Center
Dan Stanzione
Michael Gonzales
Chris Jordan
Greg Abram
Weijia Xu

University of North Carolina-Wilmington
Ann Stapleton

Funded by NSF
Thanks for your attention

Questions or Comments?
If Darwin Had A Web Browser, He Would Never Have Written *Origin of the Species*

One feature of Darwin's work as a scientist was that it proceeded slowly, very, very slowly..... Scientists in Darwin's day had hours to kill on long voyages, took long walks out in the field, and waited while their scientific correspondence leisurely wended its way across oceans or continents.......