

Grand Challenge Workshop: Mechanistic Basis of Plant Adaptation
Sept 30- Oct 2, 2008 at Biosphere 2, Oracle, Arizona, USA

Notes taken by a workshop attendee

Tuesday Night: 30 Sept 08

Atkins talk: get NSF blue ribbon panel, look at Nano hub
Jorgensen: Manage expectations

Matt Vaughn: 6 programmers at CSHL (most there, more managers at Arizona, too many chiefs and not enough ??)

Wednesday 1 October 2008

8:30 – 9:00 David Salk

Why are we here?

What is expected:

What is long term goal of this workshop?

To convince iPlant to build, deploy and maintain a cyberinfrastructure for our community (those of us here... we can not just put a few sentences to a proposal and go to a URL in a year).

Or – the money is spent and on the table, and be sure we get down what we want to get down

David's thread: The workshop is built around 7 themes that we will need to define for a cyberinfrastructure.

1. Organisms, phenotypes, and other data types (soil, temp, rain, wind, etc) to be represented (both existing and new)
2. Existing and new data sets that need to be incorporated.
3. Formalized data acquisition platforms to support data collection in the laboratory, field, and across sites.
4. Data analysis tools that would be needed, including next generation sequencing data.
5. Modeling tools for hypothesis generation and annotation.
6. Data visualization...
7. Outreach – community science

9:00 – 9:30 John Willis – The Grand Challenge: Mechanistic Basis of Plant Adaptation

The Mimulus guy – huge variation – including self/not self – coast to Sierras, etc

Lila Fishman – ones in toxic thermal hot springs of Yellowstone; myco and bacterial symbionts
Serpentine endemics, etc.

What is the molecular genetic basis of ecological diversity? (e.g. *Mimulus* species only at certain elevations and substrates, what are genetic constraints/adaptations)

How do plants adapt to their environments? Includes: soil nutrients and water content, disease resistance, herbivore resistance, abiotic stress tolerance, flowering time, pollination biology, and many other factors. *We must unstand these at several scales across space-time.*

Why study molecular basis of plant adaptation?

- Basic questions in evolutionary biology and ecology – adaptive evolution leads to new species, increased biodiversity – what are evolutionary and molecular mechanisms?
- Basic questions in plant biology – natural selection has given diversity to cope with extreme environments – how do the developmental, physiological molecular mechanisms work
- societal impacts? How can we translate findings from basic research to improve food/etc..

Now is the time for answering these critically important questions:

- we finally have tons of data to address questions in wide range of plant systems (or will soon)
- we have ability to synthesize/integrate data
- looming global climate change crisis
- urgent societal needs for food/biofuels/etc

Recent talks/paper on critical situation on crops (Dick Flavell, ceres)

World cereal production from 1860 – present – nearly doubled crop production, but what is striking is now the rates of improvement have leveled off if not diminished

.. if plot per capita yield – it peaked in 1985, and we have erased even green revolution gains with population growth (and this is taking livestock feed out of equation)...

...so need to be at higher target with LESS land, LESS inputs, MORE extreme weather conditions... and this is just FOOD, does not even include need for biofuels.

So understanding basis of plant adaptation is critical for basic AND applied reasons...

Until recently we have been limited by data ... connecting adaptive trait variation to molecular mechanisms difficult without genomics..

Now or soon we have more –omic data for many species:

- genomic sequence variation (SnP, indels, CNV, etc)
- gene expression variation (cell specific)
- Variation in epigenetic modifications
- proteomic, metabolomic, ionic variation (at cellular level)
- accurate, dynamic high throughput phenotyping

Chromosome 4 At **genome browser** – imagine this for different genotypes and different environments (Rob/Rebecca slide)

Joe Ecker slide: Sequencing across the Genome-Phenome Divide – year 2000 to 2006 to 2008

Illumina now with 2 postdocs and 7 days at \$7,000 per genome at 50x coverage

Future by 2014 ? Pacific Bioscience single molecule reads?

Have 2-3 minutes for \$70 for an Arabidopsis thaliana genome – easily assembled!

So a standard grant will let you sequence WHOLE populations!

How are we going to handle the data so that we can answer the big questions?

Current approaches in ongoing projects:

Example 1: At variation. David Salt Ionomics facility/high throughput to link genetics with high throughput elemental profiling.

Developed e-laboratory systems for sample management and data visualization

Natural variation in ionome of A. thaliana (assayed 96 Norburg lines to start) Using bulk segregant approaches and other methods that Justin Borevitz and others pioneered.

Example 2: Ed Buckler and maize diversity project. 7600 mutants, 6000 RILS, 8500 NILS, wild and land race accessions (not even including teosinte)...

- Informatics challenges – 36 million common polymorphisms in low copy regions (just SNP, not CNV or rearrangements)
- 15,000 genetic stocks created (RILs, NILs, diverse associations) and being evaluated
- 128 – 10,000 SNPs on these stocks

How do we make a high-throughput platform for QTL dissection?

- Forward Genetic
- Gene-level resolution

- Genome scans
- Community resources

How are they being integrated? Screened germplasm for 25 most diverse lines of maize, and crossed all to B73 for a reference design. Now have Nested Association Mapping (= NAM)

Genotyping – Now at \$5 per line! Keep on cutting edge

Phenotyping – accelerate phenotyping of monitoring millions of plants at daily intervals

- Use barcoding by hand, and remote sensing with balloons and cameras (not quite yet to individual plants, but can see size of plant and disease state)

NAM Activity/Future of NAM -

POINT: These two projects BOTH had to develop their own cyberinfrastructure, and existing tools not sufficient.

Why aren't current tools adequate?

1. Stand alone, not integrated into public sites.
2. Often not designed to handle sheer masses of data.
3. Genomics: most tools developed for Sanger sequencing to reference human genome – plants can have 20-50 times higher polymorphisms
4. Genomes: Need ways to represent diversity without misleading “reference” – within and across species.
5. Phenomics: no robust pipelines to deal with huge sample sizes, complex image, biochemistry..etc

Computational and statistical issues relating genomic variation to adaptive phenotypes:

- Need to account for more realistic breeding designs, population structures
- need to address different sets of genes...heterogeneous traits
- Need to deal with severe statistical issues with infinite parameters/hypotheses with limited sample sizes

How will we answer the grand challenge of understanding plant adaptations?
Work with iPlant.

How we can start to think about cyberinfrastructure that would help us address grand challenge of plant adaptation: (Discussion)

1. What are the main questions we want to answer?
2. What barriers could be overcome with new data?
3. What barriers could be overcome with new methods to effectively use existing new data?

4. What education, outreach, and training opportunities are related to our particularly grand challenges in understanding plant adaptation?
5. Big underlying challenge –share data and collaborate!

Build shared germplasm, like collaborative ...

Q&A/Discussion:

Annie Schmitt – Seems focus is on rapid evolutionary change in changing environments,
Domestication angle/invasive weeds.

Justin Borevitz – Chicago - Climate workshop at satellite to land “scale” and we are on population turnover questions/fine scale environmental sensing – what is driving adaptation within species to speciation – plant communities.

Torbert Rocheford – Purdue - What are ALL the genes that control plant adaptation across environments? And how the genes interact?
(Pires amendment about structural variation, polyploidy, classes of genes, etc)

Brian Dilkes – Purdue – Rapid evolution but how do natural systems deal with change over space (vs. past? Vs. climate change group?) . Human impacts on the environment beyond climate change – ecosystem services (water, air, soil) beyond climate change.

Dan Rokhsar – UC Berkeley – We are DESCRIBING mechanisms but need to then make PREDICTIONS and then REDESIGN (Breed?) as part of the challenge

Jean-Luc Jannink - USDA Cornell – in two examples – there were 7 genes in ionome case and x genes in Buckler case... is that answer to question?
Or What is environment?

John McKay – Colo State – Mol Bio types want genes to know how function, others want to know how variation maintained so start with phenotypes so getting genes is just the start.

David Neale – UC Davis – see “Pattern of Adaptation” and not just mechanisms.

Bill Beavis – data SHOULD be distributed for lots of reasons, no user-database, need to be realistic

Jillian Maloof - UC Davis – Also handle gene networks and complex interactions – not always simple genes but networks/interactions...

Diane Byers - Illinois State – I am more ecologist – want to not think just about genes but integrate phenotypes/gene expression across different environments (I think of genes as alphabet)... I look at roots and not exactly high throughput phenotype

Ed Buckler – Second the point that we don't want just the genes, that is great way to go across species but really want phenotype/physiology

Scott Hodges – UCSB – question the spatial/temporal scale of adaptation – salt tolerance in Arabidopsis but if get across systems... so COMPARATIVE aspect important.

Leonie Moyle – Indiana – Also need comparative framework, if work in tomato it is hard to get to genes to easier to link out from tomato to other systems. If you can't get to gene than difficult to do genotype-phenotype. So if does not work in Arabidopsis can go out to others. Like Scott Hodges said, can

Lance Waller - Emery - Driving questions important but companion idea is what data: remote sensing to plant phenotype? How bounce back and forth genetic outcome to environmental, and back to gene expression data. Development of statistical tools.

Torburt Rocheford – Purdue– what is "environment" and how measure? I think about genes more than environment and not sure how to specify?

Dave Hyten - ARS-Beltsville. From breeding point of view, if I increase protein does that increase/affect other traits? How correlated?

Graham McLaren – CGIAR – Ag perspective – plant adaptation is what makes a better crop – so how translate short term plant adaptation to phenotype you need to measure?
Phenotype can be deceptively complex – like yield – not simple to "get genes". Will require good physiological models and genetic simulations... even with massive genome data, even is space with good breeding value, can we simulate genetic control?
So cyberinfrastructure needs to go from genome to genetic systems to traits to phenotype.

Justin Borevitz – Chicago – yes, we are dissecting multivariate traits and environments – will have temperature/light/soil conditions with various traits for various backgrounds.

Also citizen science and public outreach?

Jean Luc Jannink – USDA – I said gene and I understand different levels of answer are “genes” (big problem) or also networks/interactome. I onome example was “easy” example with salt uptake candidate gene, but yield will be more complex?

Forward Genetics (?) and Reverse Genetics (what are all genes involved in this phenotype) both! In silico forward versus real life forward?

Annie Schmitt – Brown – A candidate trait, phenology (flowering time, bud burst, seasonal responses), a good one where there are networks worked out and common genetic networks over species, and crop model approaches where know inputs well (vernalization, temp, etc).

Could scale to remote sensing – and National Phenology Network (now based on Arizona).

Herbert Sauro – UW-Seattle – Top-down vs. bottom up – collecting genomic data does not always match – will there be effort to support physiological studies?

Echo Phenology need (active network of citizen scientists, project budburst on 100 species of crops, ornamentals, and wild plants, lilac data back to 50 years, but can REALLY bring in education and outreach).

Caution: note that this is one workshop/approach – and QTL/association mapping heavy and other workshops look at other modeling.

David Salt: we want formalized data acquisition platforms.

Rich: iPlant can not collect data – but it will become apparent that there are data needs and we can support SYNERGISTIC projects and want to build with their investment.

Julin Maloof: we are imagining there will be synergies – but how do we get value beyond gee-whiz this all lines up. How can data from my lab or growth chamber on Arabidopsis help maize breeder?

REMOTE Ute Krämer Heidelberg: Plant adaptation research needs to bring together many different aspects in order to realize its potential for generating powerful insights. Therefore, a major barrier is connecting people of different disciplines and helping them to communicate and understand

each other. The cyber infrastructure should enable a scientist in one discipline to begin with the resources and knowledge generated by another scientist in another discipline to complement these results from a different angle. A searchable, widely understandable database with easily understandable requests for research that the lab of origin cannot do, with short accessible and structured results summaries, resources available and other helpful information. Example (maybe not the best one): someone has available a number of genotyped progeny of a cross and knows they vary for drought tolerance under controlled lab conditions in the UK. And then a young ecologist with no knowledge in genetics or molecular biology starts a group in the Mediterranean region and goes to the database and searches for research topics/projects he could do. Then this material should pop up so that he gets the ideal material to start a field competition experiment.

10-10:30 Tea/Coffee Break

10:45 – 11:15 AM Demos of existing cyberinfrastructure

FIRST DEMO: Mike McLennan

Drive by of nanoHUB video "Imagine a technology that allows you to explore a new area of science...that learns from researches on line... not from blurry videos on line and on line discussion... fingertip access to simulation tools...easy to launch...with intuitive controls...that are well documented... results that are readily visualized...and not just one tool, but dozens of tools that simulate many phenomenon... tools that are not just giant applets, but real codes attached to national resources... and if you have questions about results or underlying models you can ask experts from around the world, or add your voice to rest of community... imagine a new way of publishing... where you can upload your own resources to share with others... to build your reputation or outreach...

STOP IMAGINING , THE CYBERINFRASTRUCTURE IS HERE – IT IS HUB.ORG OR NANOHUB.ORG...now that same software Hub0 is now powering several hub communities."

Ok, apologies for goofy music and narrator voice.

We have over 77,000 users not including robots, and all top 50 US Engineering schools,

14% are .edu domains, 333 international ed institutions, etc.

Why? We made easy-to-use software – not much "nano" about it except content so extracted out hub0 to help others.

What is hub0 ?

- Unique middleware for simulations and modeling
- content management system for scientists
- collaborations and social networks

(sort of like YouTube for scientists)

Other networks: using HUBzero:

- NIH grant on health care; - heat transfer applications.
- pharmaceutical developments
- advanced manufacturing, - global engineering hub, cancer care engineering, etc.

Example: You have written code in C+, Matlab, etc... and we have system to register, create, upload, install, test, approve, and publish...and go back and make fixes and publish tool again as version 1.2...1.3... and web publishing takes care of that. Once launch and test you can "approve" or "oops, found mistake, but fixed, now use this version"

So guide through this – we have 200 projects! And engaged hundreds of grad students and PIs, last year we had 110 active projects and 89 active developers... while nanoHUB and HUBzero only have 5-20 people (but only one guy spends half his time on updating the 120 projects, everybody else is working on tool development) ... interaction graph of center and users, blue/green dots for whether ready for use or not. Gray – new development, green = draft, blue = ready to use (rate of development matches semester schedule).

WEB management: NO bottlenecks! Email to Joe does not fall thru cracks.

Users: 77,000 visitors, 6,000 hardcore simulation users, 500 contributors, (He put online application and in one year had 1200 hits and 2 citations)

HUBzero is changing... easier to share information, expectations of experimentalists and educators, increase pace of tool deployment from years to weeks, outreach/intuitive

Questions:

Put in \$75 K for 2 years, and \$20 K for each year after... have pricing structure for hosting service... but then in future make open source – so partner with iPlant – opensource will be out by October 2010 (but hopefully summer 2009 to iPlant and others).

Sasquatch/Source Forge – what is difference with HUBzero?

HUBzero runs on their hardware and terragrid machines – so also simulations.

If developed QT, can move to HUBzero in a matter of hours. Have a cluster that gets rid of window manager garbage.

SECOND DEMO

Damian Gessler NCGR - Semantic Web

The Logical Web

What is the most/biggest thing is the web! Greatest limitation of web is information is semantically dead. If I go to google and type in "gene" how do I know if I have biological gene or Gene Simmons or whatever. How put two things together: the web with a wide variety of questions. So to bring together we are making a "logical web".

The Challenge: Science needs to integrate information – but gap in ability to find and integrate disparate data and services in high-throughput manner – to address this gap, machines require semantics in order to be efficient.

The reality: tomorrow is here today – high throughput sequencing on Illumina/Solexa

2 Gbp every two days/machine. 20 Gbp produced by 2009.

1 TB image data/run \$500 K capital costs

200 machines in the field (just one company and just genomic information) – that is HUGE amount of data – so will open up genome if not transcriptome across populations.

High throughput data available to you via low throughput gateways.

1001 genomes. The data is README files, testq files suitable for use with most aligners and tracking spreadsheets... so problem, get your own programmer or NCBI or ?

For rest of biology, problem is worse: Consider access the gene record of ADH1 from Gramene, maizeGDB, or Soybase... in each case you need a postdoc to get data from these 3 URLs... so both fragility and rigidity a problem.

The Vision/Solution: Put an agent on web to find data for you. A logical web where resources describe themselves in a rich semantic, amenable to reasoning by external agents;

A web where machines can assume much of the burden, etc...

Decision Tree slide: Do I need web services? Do I need semantic web services? Do I have the programmatic commitment to implement a new technology? Usually "yes" to all of these.

SSWAP: Simple Semantic Web Architecture and Protocol

Distributed web-based systems for data and service (algorithm) integration. Uses peer-reviewed, community, accepted ontologies such as GO...etc.

The power of reasoning: "I have data on Medicago"... "I have a service that works on plants"

The three types of questions we can ask:

1. INPUTS: I have this type of data (genomic, proteomic) and/or with x properties, who can operate on my data?
2. OUTPUTS: I want data
3. SERVICES: ...

RDF: Resource Description Framework: Subject – predicate (property/relationship) – object

E.g. bob teaches students. URL of book to classes of books or ISBN numbers

OWL: Web Ontology Language

Reasoned inferences in RDFS & OWL

Morris rdf: type Cat Cat: mammal implies morris is a mammal (other examples besides subclass, such as transitivity and other less obvious examples that are still reasoned).

Virtual Plant Information Network: SoyBase: 13 services, maps, QTLs

Legume Information Services (LIS): 14 services, sequences,

Gramene: 13 services, QTL

NAR web servers and databases : 1500 front end services

Sswap.info front end looks like google interface – type in taxa and comes up with services.

Example: finds gramene and link and why discovered (pedagogy) and 20 lines of code that has common name in it.

Assume community standard ontologies and community smarter than us, so deploy those and have swapmeet we can coordinate ontologies across services as defined by community in a peer-reviewed manner via swap.

Q&A: he used “taxa” example whereas one person would use “taxon” differently.

Agreed upon term between soybase and gramene....turns out three different definitions of taxa or taxon...

We were NOT impressed by “ontologies by committee” and like better the “marketplace of term adoptions”.

Second example: look at ADH1 across legumes? Show how go from datatype to service – how does this all fit together to allow biologists to work together? Key is infrastructure to bring together disparate services.

**DEMO 3: Graham McLaren – IRRI-CYMMYT
Crop Research Informatics Laboratory & Generation Challenge
Program**

Graham has been working on rice with IRRI for 15 years and now expanding/generalizing to wider number of crops for crop improvement programs.

“A Swiss Army Knife for Crop Research” – <http://www.icis.cgiar.org>

ICIS is a database system with tools for integrated management and utilization of genealogy, nomenclature, evaluation, and characterization crop information....etc

ICIS Components and Tools – Genealogy (pedigree) Management system – Genetic Resources (GRC)

Wheat database – idea of wheat germplasm/pedigree and tracing alleles by descent” or

“mapping as you go” (Pioneer also has).

CIMMYT Internation Wheat Trial Sites – shows what data gathered and what needed ?

Another tool:

People do not rush to web but like to collect their own data, plant breeders don't like EXCEL files or web or web 2.0... so wheat database trials and how data annotated – it is stored in central repository,

Norman Borlaug Generations Challenge Program – “I challenge the next generation to use new scientific tools and techniques to address the problems that plague the world's poor.”

<Http://www.generationscp.org>

What is it? CGIAR hosted international research consortium launched in 2003, a wide community of global researchers (Cornell in USA, CYMMIT in Mexico, etc) –

INTEROPERABILITY needed globally

Plan: <http://pantheon.generationcp.org>

A key element is domain model with middleware (Pantheon) to connect views to data sources.

Views are both end user applications ((data entry tools, query applications, data consumers) and web service providers)) and data sources (customs formatted data files, local databases, etc).

Tradeoffs of free-for all and rules.

Enthusiastic about HUBzero application to these databases! (IRRI example shown with ISIS version of rice/IRRI data, but could integrate across crops).

Q&A: How update old datasets of variable quality? In some cases can salvage and in other cases need to reacquire data in a coordinated fashion (e.g., adaptation of wheat in different environments).

Not reinvent wheel: templates for data collections!...but then how get people to follow rules.

NO, need to keep system open but annotate carefully – researchers are independent minded about how to measure water use and so forth but can still integrate phenotyping.

Heidi Appel question: – semantic web seems the way to go, but when read Lincoln Stein's Nature paper he identifies that there are no industry standards – so when will this happen?

Graham: Damien may say when exactly but no using precursor of semantic web that lacks logical connection, hopefully we can retrofit with that connection.

Damien: it is money – the difficulty of adding semantics to lexical data is nobody knows how to do it, but since we can't go back to legacy data in financially reasonable way, that puts us in unique position in biology since we have been creating ontologies over the last decade. So 5-10 years of iPlant is real good time frame to get semantic integration.

LUNCH

1:35 (delayed) Justin Borevitz on ideas on what it might look like for us

Genetics of Adaptation: From Model Organism to Model Ecosystems
Balance big questions with what we can do in 5 years, so lay out a "simple" database

Indiana Dunes National Lakeshore – genotype/environment in controlled environment and landscape, take what learned with Arabidopsis in the field out to other organisms and community.

Genetics of Adaptation:
Transgenetics 1, Mutants 10, Families 100...(see below)..

Outline:
SNP/Tiling Microarrays (SFP/SNP)
Genetic Diversity
Phenotyping in Natural Environments
Next Species: Aquilegia, Switchgrass (Mimulus?)
Ecological Plant Communities: Indiana Dunes, Tall Grass Prairies

Genetic Diversity

SNP/Tiling Microarrays (SFP/SNP)
Universal Whole Genome Array:
RNA: Gene/exon discovery, alternative splicing, antisense transcription, transcriptome atlas, CHIP, Allele specific expression
DNA:

Which arrays should be used? – see Resequencing array – add multiple species and microbial communities? At data sets – SNP array

Browser: Genomic profile of cellular systems responding to their environment – generalize these platforms from Arabidopsis to others – make data available after 12 month embargo or published

Genetic Structure: Family structure: 100-1000 SNP, Deep racial history 10,000 SNP, Genetic association, 100K SNP – can do this cheaply now

At variation throughout environments around Chicago

“Front End Cyber”:

- Sample Tracking,
- Meta Data (Field...GPS, Time/Date, Species stamp, iPhone to barcode for plant pot);
- Phenology/Developmental Stage data
- Environmental Data (light, temp, humidity, wind, soil, etc)
- Data archival and referenced

“Back End Cyber” goal

Trait x Environmental dissection

- Yield...growth, development, harvest index, partitioning
- Environmental partitioning – microclimate mapping
- Genetic variation can specify trait/environment networks

“Back End Cyber” implementation

- Quality control – filtering
- Real time data visualization: Google Earth layers, animation
- Subset summary data
- Statistical analysis – Pop/Quant Gen/QTL-Association mapping
- Community structure
- Env. Filtering/overdispersion ...

Phenotyping in Natural Environments

Arabidopsis varieties in greenhouse from a field – variation of flowering time (Share field sites with Joy Bergelson, now monitoring sites with solar powered instruments)

Distribution of common haplotypes of Arabidopsis across USA, yet Genetic Variation within a Midwest Field (PNA) – haplotypes frequency in USA

...vs. worldwide accessions of Arabidopsis (Geoff Morris)

Selecting mapping population – 5309 lines at 142 SNP, 1841 lines with xxx SNP...final xxx lines chosen as maximum diversity set

Seasons in the Growth Chamber – Program “Solar Calc II” will reproduce environments in your growth chambers.

Phenotyping can be as sophisticated as you want (movies of seedlings/time lapse cameras)

Early results of whole genome association mapping for flowering time ...and disease resistance traits (with Joy Bergelson) ...and ionome with BSA mapping with David Salt (Baxter et al)...

Next species...to ecosystem genomics
From the field to web pages..TOOLS...(list)

Next Species: Aquilegia, Switchgrass (Mimulus?)

Aquilegia/Scott Hodges project – diverse plants and pollinators – hummingbirds, bees, hawkmoths, recent Nature paper shows red evolved twice from blue, so great example of recent adaptive radiation with multiple independent events – same genes/alleles recruited or ?

So bring BAC libraries, EST, JGI sequencing...another model organism like Mimulus

Genetics of Speciation along a hybrid zone – put flowers in 96 well plates and high throughput sequence.

Indiana Dunes Collection Sites – Aquilegia here, have cameras with GPS and on Yahoo Flicker Site...genotype samples on a landscape...but only 10% of SNPs segregating in California are segregating in Indiana so need more markers...lots of ecological variation even between front and back dunes for flowering time/etc.

Lake Michigan Sand Dunes – can date years stable from 250 year old dunes to 12,000 year old dunes

Real Time Ecosystem Monitoring: Fermilab Americflux site (Timothy J. Martin) and HPWren in San Diego

Remote Sensing (micro NEON) – real time sensing, 100s of cameras, temperature, humidity, light spectrum/intensity, wind speed direction, air quality (CO₂, NO_x...) water levels/quality (salts etc), soil temp, moisture

Ecological Plant Communities: Indiana Dunes, Tall Grass Prairies

Ecological Genomics – Deep Green Workshop

1 Target Foundation Prairie Species

-454 EST – 250 K sequences and get 10,000 SNPs as did for *Aquilegia* macro and micro evolutionary importance

2. Landscape genetics – sample 1000s of genotypes across tall grass region, identify pop structure and diversity mix

3. Use diversity mix for restoration – monitor local adaptation and do longitudinal studies

Questions/Discussion for Borevitz

Xx Experimental design for plant adaptations in the environment – GPS/field may MOSTLY be measuring mostly environmental variation..then go to lab/greenhouse...but how bring back to field/common garden? Going from lab to field does not always hold up?

NEED genetic tractability and seeds/rapid cycling plants.

John Willis- I like NAM/maize because the circularity of association mapping is groups caused by traits we study, so need to follow up or lead in with QTL.

So need diversity sets

- haplotypes selected before genes are (Bill Beavis agrees)

Tim Kelly – Indiana. Computational question – from different infrastructures being developed, they have own API and operating procedures from bottom up – do we want to use things that already exist and maintain base code.

Yes – take what you can off the shelf – but need to have data organized in a way for future experiments and citizen science – then can worry about all the genes.

NOW – enough big thoughts – move onto pragmatics for remaining talks and breakouts

2:10 Kirsten Bomblies/David Neale/Bill Beavis – Organisms/phenotypes/other data types

Plant Adaptation is Beautiful and Important – pretty pictures, get excited and get public excited

(A quick non-comprehensive tour)

What Organisms, phenotypes, and data types we prioritize depend on who we are: (some types):

I. Molecular Biologists: controlled environments, screening/mapping, SNPs, array data, sequence information, metabolomics, proteomics, etc (e.g., Arabidopsis studies of her, Borevitz, others)

2. Breeders: also field trials, soils tests, QTL/association mapping (crop plants)
3. Evolutionary biologists: phenotypes in “nature”, reciprocal transplants, pop patterns, landscape genetics, fitness surfaces (see below)... (large diversity of organisms)
4. Ecologists: interactions among species...etc

Our goals and needs may differ but ultimately we all want the same thing...genotype/phenotype

Model systems are crucial for in depth understanding of some systems (can do reverse genetics or candidate gene approaches) knowledge of a pathway/genetic network – apply to non-models.

Reverse genetics: genotype to phenotype (Forward genetics = phenotype to genotype)

Popular model systems:

Arabidopsis (flowering time, salinity, drought tolerance, latitude, herbivore, pathogens...but does not have all adaptations)

Mimulus guttatus (pollinator adaptation, mating system, heavy metal tolerance, elevation, flowering time)

Aquilegia Formosa (pollinator adaptation, soil type, mating systems)

Rice (flooding tolerance, drought tolerance, temp)

Populus and Grape

A model family – Brassicaceae – relatives of *Arabidopsis* – *A. lyrata* (mating system evoln, herbivory, flowering time), *A. halleri* (heavy metal tolerance)...
Boechera howellii (waters), *Thlaspi* (heavy metals), *Leavenworthii* and *Ionopsidium* (adaptive evolution of inflorescence architecture)

Forward Genetics: Start with phenotype and go to genotype: QTL analysis, “classical” genetics, SNP/phenotype association studies, physiological systems

QTL mapping is a useful tool for studying adaptation – an example from *Mimulus guttatus* and *M. lewisii*. Schemske and Bradshaw 1999. Procedure: genotype all recombinant progeny for markers that differ between parents. Test for significant associations between...find QTL...clone

Many adaptive traits have been QTL mapped – here are a few:

Bratteler et al 2006 – serpentine adaptation in *Silene vulgaris*

Doebley – perennialism in teosinte; Rieseberg – *Helianthus* example, Pine tree example of cold tolerance...

Information can be integrated for QTL across accessions or even across species

Getting to the genes can be tedious but there are some cloned gene examples:

- Hoballah et al. Plant Cell 2007. Petunia species – loss of myb transcription factor AN2 gives white/pink color transitions – appears that white flower has been due to multiple independent losses on AN2
- Hanikenne et al. Nature 2008. A. halleri – Zn hyperaccumulation – due to triplication of HMA-4 and cis regulator changes.

Another example of pollinator preference in plants: highlights potential use of fitness landscapes in understanding plant adaptation.

--Whibley et al. Science 2006. Fitness landscape estimated for Antirrhinum flowers – observed (in nature) versus potential (lab) phenotypes.

Phenotypes – flower color and shape

Likely adaptation = pollinator preferences

(Could use fitness landscape in a cross or other systems)

Adaptation from the perspective of a plant breeder (**Bill Beavis**):

Grain yield (for example) is complex quantitative trait of interest to breeders. In breeding for yield, this may select for trait at exclusion of other traits

- but Darwinian fitness in nature is not just high yield (offspring number) but really offspring survival).

-Syngenta dataset: access to experimental databases : 40 genotype-environment experiments across (maize, tomato, Soybean, rice, At) for numerous traits across 60 locations (Beavis slides)

-CYMMIT/IRRI/INGEN datasets since 1975 – 20 nurseries per year, 25 lines per nursery, 20 environments per nursery – a mini-cyber-plant infrastructure in place.

Note: studying plant adaptation not limited to small plants:

Dendrome: (**David Neale**; //dendrome.ucdavis.edu) – doing comparative re-sequencing in Pinaceae. Staggering genetic resources: 15 million progeny being phenotyped!

Traits for trees: bud burst, bud set, growth (ht, diameter, mass), emergence, cold hardiness, etc

Environmental variables on maps for tree sites: temp, rain, aridity, frost dates, elev., and lat/long

Objective 1: Genetic structure of adaptive traits: these traits put on map (Oreg, Wash, BC)

Finally, German distribution maps highlight the specificity of plant adaptation

Can be useful for learning/outreach tool!
So can look up any species and dots for where on for grid...and
GIS based Mapservers may help pinpoint candidate causal environmental differences.

Data? Genotypes/phenotypes/...climate...geological... need to be able to add layers not thought of yet...

2:40 QUESTIONS for Bomblies/Discussion

We need some specs, not "everything in there" but keep flexible for new data.

Justin Borevitz: Want them all, but what exemplary organisms for data sets? (Just save for after TMO talk since he has list of phenotypes)

3:00 Thomas-Michell Olds/ Lila Fishman/John Burke – Existing and new data sets that need to be incorporated

What tools, datasets, and other resources should be incorporated in the proposed cyberinfrastructure? Get a list, so broke into categories:

OMICS data: metabolomics, proteomics, etc...

PHENOMICS: morphometrics, stress resistance, plasticity, etc

Derived data: QTL, env. Data, etc

Environmental context: Ecological, historical

Deal with environmental context – grow an iPlant in an iPod.

How will plant communities respond to climate change?

Using climate models to predict community composition

Ecological communities: herbivores, etc.

Phylogeny of all plants can elucidate ecological processes: why spp diversity high in tropics?

Evol. Data: phylogenies and related spp. comparative genomic info, connections to genes, etc

Distribution data: species geographic and ecological distributions, bioclimatic envelope info, distributions of subspecies and genotypes, distributions of symbionts, herbivores, etc

Maize landrace examples: integration of genetic and environmental data , can see different maize races grow in different environments

Ruiz Corral, 2008, Crop Sci 48: 1502

Genetic resources: accessions, RILs, NILs, and their phenotypes and genotypes

SUGGESTION: Genetic resources

(1) iPlant implements online ICIS database for Mimulus, Aquilegia, A. lyrata, Boechera, etc.

(2) User communities input data! (it is database that already works)

Sequences and genotypes – lots to put online

Lots of reinvention of the wheel, so software needs:

- Automated pipelines to: filter and assemble data, annotate, compare to reference genomes, identify SNPs and indels, prepare derivative data

(divergence, polymorphisms, etc)

- user input: raw reads, name of close relatives

- output: annotated sequences linked into comparative genome browsers (not as easy as it sounds so should be feasible)

Slides from Todd Vision and others:

Biodiversity, Encyclopaedia of Life, etc.

Off topic for 3-4 slides: Modeling complex systems

=Although not a major part of this group it is of interest: what is recent history and why care?

Yin and PC Struik 2008. New Phytologist 170: 629. Model crop systems

Welch et al. 2005. Flowering time (he is talking in climate change session)

Machine learning analyses of protein interaction networks (systems biologist call a hairball, but perhaps closer to a rat's nest)

- Zhu et al. 2007. Plant Physiol. 145: 513-526. Photosynthesis modeling group for Stephen Long at Univ. of Illinois.

This is reminiscent of metabolic control theory in evolutionary context

Jonsson et al. 2005. Bioinformatics 21 (Suppl 1): i232-240. Meristem

modeling from Meyerowitz and collaborators...so not too far from what we do.

Q&A/discussion: Annie Schmitt – historical records of climate and pest outbreaks needed

John McKay – confidence intervals around data sets:

Heidi Appel – plant metabolites? (In metabolomics)

Computational question: citizen science – anybody can upload data – how check quality?

Chris Pires – want herbivores/mutualists – but when does iPlant become iLife? (limits – is that up to board?)

Diane Byers/Illinois – Need for statistical models/approaches (second by Dan Rokhsar)

Annie Schmitt – LOTS of geographical representation of data – so need GIS for dummies

Paul Quick - Annotation basics?—what needed for community NOW

David Bubenheim – whatever list we supply will be inadequate and will have to be modified as time goes on, so need overarching structure with ability to add future data sets for global change group or other groups – need to be able to add modules from other community efforts... to allow us to go 20 years into future.

Bill Beavis - Iowa State – we will talk about statistical analyses later – but we do NOT YET KNOW what they will be yet... so need to be patient.

HUBzero can upload any script...

3:00 – 3:30 Tea/Coffee Break

3:30 – 4:15 Tim McPhillips – UC Davis Genome Center/Mark Schildhauer – NCEAs – Formalized data acquisition platforms to support data collection in the laboratory, field and across sites.

Tim McPhillips: Collaborative instrument control/sample mounting robot – a successful collaboratory from bed at home.

Challenge: Determine 3D structures for all proteins in a single cell organism...

Data acquisition takes place at many steps in addressing a Grand Challenge: experimental work distributed across six institutions and three core teams

Lessons: Don't start from scratch – leverage, integrate, and extend existing cyberinfrastructure

Enable researchers to use their favorite tools. Adoption is critical

Don't view web interface as a panacea.

Be in production from day one. Build the cyberinfrastructure to accelerate and scale up what researchers already do.

There are two very different kinds of information to manage in data-intensive research.

Data that could be useful to researchers in other projects:

Data for one publication.

Grand challenges require data integration:

Many kinds of data collected by many different researchers, using other researcher's data and sharing your data with others. Capturing **metadata** is essential to make useful to others – both raw and derived data. You want scientific workflow system that can automatically record the provenance of data products.

Two kinds of data integration:

One kind is data integration with data of fundamentally the same type:
schema mapping approaches.

Other kind is data integration of multiple types – requires transformation or
scientific workflow automation.

General recommendations for iPlant: Ask iPlant to: focus on data management infrastructure, employ existing standards..etc.

Form collaborations and apply for additional funding to...develop and adapt your own tools for doing more data analysis that takes advantage of iPlant infrastructure.

Adopt a framework for doing analyses that captures provenance automatically.

Mark Schildhauer (NCEAS Director of computing)

Recent issue: EXCEL should not be used (still) for random number generation and so forth...so what is alternative? Whatever you ultimately spin up for iPlant, no matter how easy it is (put Mozilla on desktop and get on web? Not that easy), you will need training/adoption procedures.

Cyberinfrastructure for Holistic Biology:

- Increased need for collaboration and synthesis to solve vital complex questions in biology – from gene to ecosystem.
- Cyberinfrastructures supports synthesis by:
 - providing data access infrastructure;
 - dealing with the integration of heterogenous data
 - basing analysis and modeling on robust, shared code
 - standardizing and exchanging protocols, methods (work hard on this, but not exciting)
- work variety of informatics research projects (KNCB, SEEKI, MaNIS, CIPRes, VegBank, Kepler, GEON...etc) – look at existing activities that are relevant to iPlant puzzle! (ISIS, NanoHUB, SSWAP talked about earlier today, but look at –omics/phylogenetics field).

Data discovery and integration challenges

Ecological data are highly heterogenous:

- variable syntax, structures, and semantics
- highly dispersed holdings (floppies to PC in closet) and few repositories
- derived from many disciplines (genomics, morph, phys...sociology?)

Collaboration and data sharing

Personal data management problems in your own lab from 3-5 years ago – in your file cabinet or 3.5 inch floppy – those are VASTLY compounded in collaboration. So need for better:

- data organization – standardized formats and structures
- data documentation – standardized descriptions of data (metadata), loose coupling but compatible (otherwise just in your head and not metadata, even collaborators don't understand each others files)
- data analysis – documented and executable
- data & analysis preservation – archived, discoverable, retrievable, and interpretable (archives for both)(versus ark in Raiders of the Lost Ark put in box in giant room and can't find...)

Technological solutions:

- confederated data sharing framework (standardized protocols, rich metadata, controlled vocabularies/ontologies, compatible querying mechanisms, distributed management and ownership) (NOT an UBER-dump site)
- analytical software that is scripted, verifiable, re-usable (eg., R, Matlab, SAS, C) and orchestrated with scientific workflows (e.g., Kepler to allow heterogeneous execution environments)
- free, open source, multi-platform software for data management & analysis (whenever possible)
- Virtualized “central collaborative workspace” for organizing communications about data, analyses, protocols, etc)
(NanoHUB is one option, PLONE base is another, WIKI is another...so many options besides HUBzero).

Ecoinformatics Products (NCEAS, UCS, SDSC, KU, LTER and other collaborators)

- Ecological metadata language – structure, semantics, and context of scientific data
- Morpho – desktop metadata and data management software
- Metacat – distributed data server (ESA)
- VegBank – plot, species, and community vegetation data (North America, others in Australia)
- EcoGrid (and EarthGrid) – interfacing distinct data systems and networks
- Kepler – analysis and modeling

Other cyberinfrastructure Efforts (he has been involved with so just tip of iceberg):

SONet: community driven scientific observation networks to achieve semantic interoperability of Environmental and Ecological Data – OCI Interop

VDC:

Others...

Other considerations: Existing and emerging standards: metadata, etc
Technologies – NSF, NMI, W3C
iPlant should not reinvent technology wheels: partner/collaborate with
other efforts
entomologists/zoologists want to work with you!

Data acquisition for iPlant – GCW should identify what are the main CI
needs!
Data contributions/acquisitions (how to collect and organize contributions
from “autonomous” researchers)
Data discovery/sharing
Data storage and delivery
Data mining and pattern detection in massive data
Computational bottlenecks
IP/access and attribution concerns

Recommend: “Use Case/Scenario” method of specific CG to clarify major
issues; then generalize for CI
**- advocate semantic web approach/community annotation and
controlled vocabularies over user-database...**

Q&A: Discussion of balance of standards and so forth

4:30 – 5:15 Breakouts:

Breakout 1 here
Breakout 2 in cassita 1900
Breakout 3 in visitor center room

5:15-5:45 Presentations from breakout groups

Dinner/posters/demonstrations

Thursday 1 October 2008

8:45 Dan Rokhsar and Jean-Luc on computational tools

-How do genotype and environment? Xx

Phenotype = G + G*E (datasets)

Genetic variation (list) phenotype (list) environment (list)

Genetic approaches: classic pedigree methods, genome wide association
methods, and combinations (NAM, Indiana Dunes, etc)

New sequencing technologies – faster, cheaper, better

**Tools for new-seq data? Raw data sets huge – physically move hard-
drive at UCB faster than over network.** Best to transform into sequence

and quality files at core facilities (next to Solexa machine, not iPlant), so submit that to iPlant?

- standard uses are to map to "reference" sequence to identify and report variation (question: enter this data into iPlant or?)

Can use short reads/SNPs

Beyond SNPs – three times non-SNP variation between human and chimp – 36 Mb SNP, but more than 100 Mb of insertions and deletions

Possible value of Platonic idealized genome vs (randomly chosen)

"reference" with its idiosyncratic load of mutants/mutations

Genetic mapping: basic problem: regression of traits vs. alleles/genotypes

Multiple existing software tools with related but distinct methods, strengths, weaknesses

(labs typically develop a comfort level with one package, can iPlant enable testing/application of multiple tools, some software is computationally intensive)

"Traits"/phenotypes : anything that can be reliably measured

most can't be readily standardized across taxa (tower of babel vs Esperanto)

make available templates for phenotyping that could become de facto standards, prevent reinvention (unless a new wheel is needed!)

balancing data privacy with public release (initial sharing within research team to take advantage of iPlant tools, clear path and timeline to public release)

More information about (fill in blank) experiment standards and reporting format) (cf. microarray MIAME standards)

Variation: see Langlade et al. PNAS (Enrico Coen group on leaf shape variation) – like Ed Buckler saying "p=p", you did not need PCA but points used to capture leaf shape amenable to that. 3 PCs of snapdragons describe range, and variation at -15 loci/QTL can describe "path" from one species to another (and this was without a genome as Leonie Moyle/tomato example)

A sample workflow: Dark arrows from genetic association studies/markers Get to gene (marker, synteny, homology) to chromosome segment

(at this point may want to reach out to other genomes if have not model organisms to get candidates)

Get a configurable report on gene(s)/candidate(s) and gene function and other associations

Then get other functional datasets from external databases (iPlant will just have interoperability code to all the databases out there)

Get their phenotypes associated with a locus? (or from candidate gene list)

Get candidate genes and move back to trait dissection, breeding, and/or physical model

www.phytozome.net - advertising his comparative stuff (z left over from metazoan)

A tool for green plant comparative genomics

Arabidopsis – Populus – Vitis; Sorghum-Oryza – Selaginella, Physco, Chlamy

Example: synteny at HMA4 locus – colors tell you which genes are in this family, so not only similar to each other but also syntenic... can do many other things with this.

How share with iPlant/Eric Lyons?

GeneGroups™ - various methods define “groups” of genes (typically within a species)

- coexpression, protein complexes, pathways
- some defined ontologies but others may be ad hoc or based on particular datasets

- Many groups in literature/supplemental tables (bulk load these with attributions, descriptions)

- Allow users to define (and describe) their own groups on the fly

Provide mapping across species via orthology

- Given critical mass, can start to reason with groups across species to drive formulation of hypotheses.

Challenge of using new technologies to address GC of genetic basis of adaptation

(Data will be easy to get, ease of using that data set to answer YFQ towards GC (iPlant bridge that curve)

Jean-Luc

Genetic basis of adaptation

Pieces exist to identify genes within a species

Power of discovery is enhanced by comparative approaches across species, related traits, related environmental stresses

BARRIERS: resources for different comparisons are dispersed and not amenable to assembly into a systematic or efficient workflow

(back to Dan’s talk, how get a “report” that gives list of orthologous or genes, like a literature review, but also connects to phenotypes and may be data not published or not searchable)

David Neale also pointed out that adaptation is rich with patterns/comparisons and no easy way to make generalities.

ALL COMPARATIVE – need hooks to comparative information:

- homologous loci

- physical interaction
- syntenic loci
- correlated traits
- synthetic lethal
- correlated expression
- response to environment

Can link all these by entering into any of those from a visualization tool – from genes or traits or environment (again $p=p$ of Ed's equation, a correlation of phenotypes)

Enter "environmental stress" and get: Homologous traits, known pleiotropic loci, correlated traits, known mechanisms, other species responses, datasets with variation in stress.

Discussion: David Salt lots of discussion of vertical and horizontal integration – need both comparative stuff across species/environments. Vertical = deep in few systems (At, maize) versus horizontal (shallow across systems).

John McKay: Dan's talk good – iPlant may not be place for reference sequence/assembly – but then later said reference is a moving target... what do?

Dan R: we need a "pan-Arabidopsis reference set deleted in Columbia but in other species"

(Dilkes uniqueiome) – Human committee may be first or Arabidopsis? Not worked out yet?

As for where initial mapping gets down, somewhere you will have to farm out genotyping AND assembly...

REMOTE: Lance Waller/Emory University – sequence and phenotypic data – but what about environmental data? Salt intake and coastal data/meteorology or local data are also a data to consider for a hook for environmental stressors. Lance is expert in spatial statistics – what can we adopt? On one hand you have Google Earth to geo-locate and people just figuring out how to do phylogeography of viruses (vs. trees).

Brian Dilkes: Agree with concern of handling Solexa data remotely is problem and need to reduce down, but at same time iPlant has capacity to potentiate research where no reference genome from BACs/Sanger reads. So need platform for "no reference genome" – how move and manipulate data.

Dan R: yes, thousands of genomes being sequenced – others will figure it out. But Buckler and others doing a "reduced representation" and

resequence only at those loci as another type of dataset...not only resequence whole genomes (Sequenome?).

BD: Problem not only moving data around

Justin Borevitz: Comparative hooks of genes and traits, but also geospatial tags – what other studies there? GIS layers? Environmental history? Species/pop/haplotypes distribution maps.

Nick Lauter: another hook is picture of the plant – click on leaf and what studies done there

Graham: another hook is strain/pedigree

Bill Beavis: I like how this conversation is going, but back to spatial statistics –

Ed Buckler email: I think we need to specifically mention 10 major datasets that really exist or is funded now:

Arabidopsis 1001 genome data - Nordborg

Maize NAM - Buckler

Maize Hybrid - Beavis

Wheat CAP - Jannink

Barley CAP- Jannink

Mimulus - Willis

Pine - Neale

Rice - McCouch and IRRI

More ecological datasets?

Does Justin have real a data set that can be incorporated?

I would like to see some remote sensing datasets - we will contribute some of our stuff, but I am sure someone has something better.

Cheers- Ed

9:30 Herbert M. Sauro: UW-Seattle

What is Systems Biology?

Two flavors:

Top down: Systematic Biology - cells(DW – Dave Weston)

Bottom down; “network physiology” – look at small pieces within cell and make predictive models, moves into “synthetic biology” which is an engineering discipline. He (HMS) is between network physiology and synthetic biology.

Philosophy and methodology:

Top Down ' omics
System: whole cells
Model: statistical correlations
Data: high throughput
Visuals: arrays, hierarchies

Bottom up "mechanistic"
- networks, pathways
- mechanistic, biophysical
- quantitative, single cell
Visuals: dynamic curves, networks

Motivation for mechanistic studies:

- understanding the dynamic behavior of cellular networks
- discovering general principles of operation
- engineering new or existing networks to affect the phenotype in specific ways

(now think biophysical model of cytoplasm is correct!)

Infrastructure requirements for mechanistic systems biology

1. Experimental data from hypotheses – driven experiments.
2. Databases/exchange standards
3. Controlled vocabularies and semantic annotation
4. Software provision for computational work

A real network example (E. coli, synthetic biology)

Enter IPTG, output GFP, can alter strength of repression by altering promoter sequence

First set of experiments he did over last 2 years, as engineer need to "tune" the repression level.

(not nearly as complex as ecology, all done in 96 well plates)

(future involve more microscopy)

Huge list of modeling software tools – over 100 on his slide!

So have "model exchange standards: SBML, CellML

SBML is a way to describe the biology of cellular networks

(not all support Cell ML – a math based description from underlying biological)

SBML in a nutshell "Systems Biology Markup Language"

Other proposed standards: SBGN – Systems Biology Graphical Notation (Japan lead)

www.sbgm.org/Main_Page

Now a consensus on how to visually describe mechanistic models

(e.g., ATP attaches to everything so don't show all the wires from that)

Benefits of SBML – unambiguous model exchange – to both

Simulator comparisons and compliance, and

Databases to semantic annotations and journals (which feed to similar also)

Have test suites (180?) and compliance testing suites, models get loaded into 12 simulators and all data gets compared – a sort of competition to improve software at outcome over last two years

Model repositories: Nicolas Le Novère

BioModels.net

185 curated models as of Aug 2008; 74 uncurated models.

Worst way to store models is paper format journals – too many errors, so curate online.

MIRIAM: Minimum Information Requested in the Annotation of Biochemical Models

Not a file format but minimum specifications.

Reference correspondence: encoding a model in a recognized public standardized machine-readable format.

Semantic Annotations:

1. SBO: Systems Biology Ontology (what is K in equation; quantitative terms)
2. MIASE: Minimum Information about a Simulation Experiment (how to graph!)
3. TEDDY: Terminology for the description of dynamics (is it an oscillation or a ?)
4. KISAO: Simulation Algorithm Ontology
5. Missing: an audit trail of a modeling process – fragments that disappear, how record modeling decisions?

SBO: Systems Biology Ontology terms have id numbers “

Software Development Challenges:

Systems biology has history back to 1940s, but lots of reinventing the wheel
Development of large monolithic applications, then die.

Possible solutions:

Split software development into two lines:

1. Development of specialist software libraries.
2. Tie libraries together to form end-user applications.

Software libraries can be used to carry out specialist functions which can be combined by application developers.

Issues to consider: Library stays while underlying science changes, choice of languages important (C++ is very agnostic and can link to Java/Python)...and many C/C++ platforms: libSBML, SOSLib, Antimony, libStructural, ESS, libMFA (Metabolic Flux Analysis)

Finally, where sit with other systems biology community: integrating statistical and mechanistic modeling (traits: flowering time, photosynthesis, etc outputs with abiotic/biotic stress inputs).

Dave Weston comments

DISCUSSION

NEED face to face meetings with collaborators/competitors and learn to trust each other to get over "no"... then have hackathons of 20-25 people, need buy in and not policing).

Smaller group of committed people most important.

Annie Schmitt: fantasy of inputting real time environmental onto phenotype, fitness, and ultimately predict adaptation.

Can we set up network evolution (he does not know)

Dave Weston: need to reduce transcriptome/network to just one/few factors to put into network simulation. So take out hub genes/peripheral genes

10:20 Coffee Break

11:00 – David Bubenheim and Timothy Kelley (CNSC/Indiana University – <http://cns.slis.indiana.edu>)

Data Integration Challenge – Source and Scale – arrays to landscape

Yellow star thistle – amazing strong invasive species

Tim Kelley: in book "information visualization" you can detect one dark pencil in white pixels, can search through 15 million pixels in a minute – can see missing data easily,

Also put multiple people in front of big tv screen and reason together, human ability unchanged for thousands of years while computers doubling...

REFS: Hibbs et al.. Visualization methods for statistica analysis in array clusters. BMC Bioinformatics 6: 115 , 1471-2105 (2005).

- Colizza V et al....

KW Boyack. 2006. Map of Scientific Paradigms. Sci Tech Strategies.

Links of references

Color of dots is speed of consensus in a field – much faster in some than others...

Visualizing Biological Data:

Z Xiang et al. 2007, CRC View: A web server Bioinformatics 23. 14, 1843

Univ. of Calgary – genomes/proteins in a virtual reality environment –

developing technology for this or collaborate with them.

Food Webs software: Visualizing Balance and Imbalance in Complex Networks...

(film looks hokey but excellent tool for visualizing parameters – dynamic simulations –

3-d visualizations along with running graph – can also see if cut out species or add invasive species... (e.g., effect of local extinction of a parasite ultimately leads to extinction of half of species in the web).

David Bubenheim: Multiple Satellite Observations Provide Global Perspectives

Consensus? Or science changes one funeral at a time (Max Planck)

Satellites: TOPES, TRIMM, GRACE, Cloudsat, CALIPSO, Aqua, GIFTS, Landsat, NOAA,

... we have many satellines in US, commercial, international, lots of data that covers the globe.

Other remote sensing: airplanes and at field... are real spatial/temporal issues...and when does the satellite go over? He likes MOTIS that flies over twice a day without great resolution but good temporal component, while LANDSAT has great resolution but only every 7 days and need a clear sky.

Spectral analysis: Usually working with an image broken up into pixels and component spectra.

Local Scale Biophysical Mapping

Local Scale Flux Mapping

Vegetative Index: map boundaries of native/exotic vegetation over time – resolution of a pixel is 500 meters – not great but get every day so nice for looking at “green-ess.”

So iPlant can push forward remote sensing to core sciences.

Determining IS Distribution: different layers can find Tamarisk, star thistle, and cheat grass.

Then link to simulation/growth models with fire (fed by cheat grass) and make risk maps.

Target species: cheatgrass (*Bromus tectorum*), yellow star thistle (*Centaurea solstitialis*), water primrose (*Ludwigia hexapetalae*).

Controlled environment studies in growth chambers – experimental cuvettes and energy/mass balance.... Also model back to space with climate change, and USDA field sites.

Web based viewer to integrate in the west: can see what studies done when

and where, remote sensing built in, was useful as a library (we are or were here) and to develop habitat models and associations but not a real landscape/ecosystem tool.

So, we found boring FORTRAN based 10 year old model that I love:
Soil Water Assessment Tool (SWAT) :

- A water shed/river basin simulation to predict the impact of land management practices on daily water, sediment, and agricultural chemical movements in watersheds with varying soils and land uses.

- Ecosystem Simulation and Visualization: Hydrologic Unit/Watershed System: Map ag land, non-cultivated land, urban land, rivers/floodplains, point sources of pollution, almanac fields.

Dave works on non-ag/non-native/invasive plants.

Regional/national Assessment – and Mississippi river basin – and define land units and water quality – publication will come out soon = problems not always point source, can be farms that are biggest polluters. This was so successful in ag that blue ribbon panel wants nation wide assessment of all range lands.

SWAT accepts many input types – GIS, arcview, weather, hydro, topo, ...many GIS layers...

Ecological forecasting: integrate surface, satellite, and climate data with ecosystem models.

Terrestrial observations and prediction systems.

Daily “nowcasts” for California (meterology, hydrology, vegetation, ecosystem (primary productivity) and can zoom in/out and scale. But hard to use with 1 foot thick manual, so “user view” is eye candy/dirty front end with Google Earth type visual (Santa Rosa is test site/watershed). Can select landscape units...

NASA Advanced Supercomputing (NAS) Division – SGI LTIX and CRAY Opteron cluster
Pitch iPlant to this with USDA?

DISCUSSION: Have user interface that is most comfortable.

Annie Schmitt – thought experiments you can do with technology – what happens if know out predator or knock out regulatory gene?

can you ground truth all models?

A. If have good stream flow data – and can match stream flows, you can

really define where you are (or even go back or forward in time). Now have seven different watersheds across the west, but 10 years away to do all the west.

Dan R: You showed example where you can separate spectra of oak, cheat grass, and star thistle.

LUNCH:

Rich prompted unplanned joint agenda between both workshops for all following sessions and breakouts

12:45 Unplanned discussion with Rich Jorgensen and both workshops:

Project proposal options: – self form 10 best people? Or 80 people here? Nov 7 for Stacy Harmer: Modeling plant responses to the environment over a day or over a year.

But way proposal will be evaluated is how much “community buy in” will there be?

So needs leadership and management plan – with roles specified well – “Our needs” and “How I want to work with iPlant” ... then MOU on both sides... if goals not met then project ends.

Length – these will NOT be 5 years, but maybe 1-2 years and then community takes off.

iPlant prepared to pay for superusers: sabbatical, summer time, postdoctoral program managed jointly by GCW team (not \$100K to everyone in the room).

Apply for NSF RCN ? Like MORPH/Plant Form, other agencies also?

1:15 Ruth Green: 10 questions:

1. What are the genes and pathways responsible for genetic variation of responses of plants to environment?
2. Given large gap of molecular genetics and ecophysiology, what tools/methods can link these for
3. What are best levels of description for capturing adaptations....multiple scales, multiple models...levels include genes, networks...and changing microenvironments.
- 4.
5. What data and tools from individuals should iPlant provide the tools first? How can ?? deep comparison across species can be made?
Reduce costs of sharing data and tools?

How find genes ... reproductive success...crop production...under stress...
Ecosystem services over 5/10/20 years?

What is cumulative effect of microenvironments...to climate change.

Worley: ridiculum – the yeast protein-protein interactions – hairballs – very busy spider...

What kind of IT environment: plug and play, common libraries, comparative ridiculomics, transfer among species, ontological terms, learn from different species...

David Salt – unauthorized summary – we have adaptation questions, synthesis was Brian Dilkes idea... key threads: we need to synthesize 3 datasets:

Genome anchored, landscape anchored, biochemical/physiological/phenotype anchored... allow those three anchors/hooks...interoperability...existing and new datasets that will need to be integrated, platforms needed for data acquisition, data analysis (primary and meta), models and predictions (human friendly), both mechanistic models and predictive models, visualization tools are pointy end for qualitative biologists, those are key points.

**1:20 EOT talk: Kay Havens, Communicating about plants
Director, Division of plant science and conservation, Chicago Botanic Garden**

Reasons why public understanding of plant science is critical:

An informed public (including youth) are more likely to:

- be supportive of plant science research, including its funding
- consider plant science as a career or as a volunteer activity

Citizen science:

- is a venue to engage the public in meaningful research
- improve scientific literacy via participation in the scientific process (youth and adult)

-

Tools that rely on smart phone technology

- common sense – environmental monitoring and ...

Common Sense: mobile environmental sensing platforms – from propaganda on common sense website (used in SF and elsewhere to influence politicians and so forth) –

iTrail: an iPhone application for athletes to see how fast run or ski, but can also measure a transect for time, speed, altitude, max gained with phone (also GPS)

Cybertracker: mostly used in S. Africa and Europe. Used by animal trackers but could not read or write – based on images or words. Can download taxonomic keys, get georeferenced data, ...map global ecosystem in real time”

Citizen Science Projects:

Project Budburst and the National Phenology Network (3 years old, HQ at U. Arizona)

See www.budburst.org (NCAR, Montana, etc)... 2/3 are submission of children 12-18.

Focus on iconic, easy to identify species (Saguaro cactus), Info sheets/photos on all the target species – you can add an “other” but primarily an education vehicle – data sheet is “where observe” (google earth of lat/long) and phenophase data.

(first flower, second flower, first bud, etc pics)...near building? Microclimate? North slope?

- **National Phenology Network** has cloned lilac phenology – observed since 1956, so 52 years of data, also recent 120 year old dataset from New York (very consistent since only 3 people did).

Core protocols and intensive protocols.

Floral Report Card: Project Design – identical gardens at each of 7 botanic gardens (16’ x 16’ raised beds of green material) – 4 clonal replicates of 4 ecotypes of Penstemon, Monarda, Baptisia, Schizarcum, Panicum)

Date collected by volunteers

Aim for network on 50 gardens, with satellite areas with youth centers across country, all connected by internet... kiosks, GIS layers, etc.

Costs \$10,000 per unit to add one... now piloting with 7 gardens to raise awareness and get data

National Institute of Invasive Species Science: Had cyberinfrastructure grant – powered by Global Organism Detection and Monitoring System, Also Eco-? Data from these projects have been contributed to data commons Partnering with NASA for **HoneybeeNet** – beekeepers as citizen scientists and weigh hives, have found nectar weight peaked in 1970s 4 years early.

Cornell Lab of Ornithology – longest citizen science is birdwatching. CLO: Usefulness of citizen-science data ; protocols (if they register, that cuts from drop in false data from 25% to almost 0%.

Methods to overcome limitations of citizen science data, including auto-check on location/range etc. Ways to recruit and retain citizen scientists.

Technology for online reporting.

The Missing Piece: In most cases, CS's submit data, if they are lucky they get an annual report.

For virtually all projects, they don't have the ability to generate questions and analyze data.

For complete and ongoing engagement, we need to enable this piece.

QUESTIONS/DISCUSSION:

FARMERS?

NATIVE PLANT SOCIETY GROUPS? RETIREES?

HOW A PLANT WORKS?

1:45 MARIA GALLO Univ. of Florida – UF-STEP program goals to enhance outreach efforts

- To assist scientists in communicating their research to targeted audiences through innovative media to enhance grant projects.
- To formally train graduate students in communication, critical thinking, problem solving team process skills development, change management (soft skills and not just technical skills)

The STEP Model

- Frame science to make it accessible and dynamic/interesting
- Address target audiences – future/current grads, K-12 teachers, science journalists/journalism students (shocking how little they know), general public
- Create materials that are convenient and effective for an audience (test materials with audiences)

UFGenetics.com

Cutting edge research framed by entertainment approach: 2-5 minutes – YouTube; WUFT public TV, UF Natural History Museum) 50-60 online, and local WUFT uses as “filler”

Features that are specifically developed for career exploration purposes.

Videos on Miami Blue Butterfly and Butterfly House at museums

You can be very specific and focuses – brain on science can be personal or recruitment based.

Features: videos, lesson plans, etc

iPlant Collaborative: Types of offerings:

1. Active learning online environment/network
2. Distance ed. Offerings for graduate/undergrads/etc
3. Active Learning Environment/Network – website and mobile learning
 - videos, case studies, lesson plans, data analyses, simulations, online quizzes

- web2.0 capability

(She is peanut genomics group, so has peanut genomics training module, case study for public,)

Distance Education Offering for Students

Technical and “soft skills” (e.g. GEM NNF – her USDA grant!)

- critical thinking / decision making/problem solving skills
- working effectively in teams
- leadership training
- science communication (capstone summer STEP internship)

Summer internships/workshops

Training of teachers and/or undergraduates (NSF, HHMI training grants)

- work directly with researchers for technical training (academia, industry, etc)
- bring cohort groups together

Assessment and Evaluation

- Critical thinking instruments? Focus groups? Pre-test/post-tests

DISCUSSION/Q&A

My daughters worked with ePod, and just simulations that are semi-realistic, so GAMES to engage kids also!

School kids go to websites – especially home school kids – so be sure Wikipedia is accurate?

Are school kids involved in crosses (pepper plants easy?)

Funding gap? Her department helps, but iPlant won't fund filming for flash drive.

But discovery environments should be designed in way so K-12 can use.

Also professional societies have web portals and availability to work with this (ASBP, BSA, ASPT, etc).

Martha Narro and Lisa Howells

Higher education piece is important – undergrad/grad/postdoc training

SHOULD be different once there! So think creatively about that and not only

We want interdisciplinary ease – computational thinking and plant biology (with math and stat)

Dave Micklos: I take videos and specific budget to adapt materials and put online... already a budget for each GCW for movies, flash animation, and so forth.

See DNAlearning center of Dolan lab online!

Borevitz: Match outreach with inreach –

Kelly: huge desire for people to come in and help with flash video, data collection, science blog folks, etc. One of biggest innovations of YouTube is creative endeavors getting done for free (just social recognition).

2:30 Breakouts?

Data analysis: stay here...

Modeling, Visitors Center breakout room

Data Visualization Tools,

...or fourth group self form for planning organizational levels

3:30 Break

Discussion:

Damien: Data INTEGRATION is VERY hard – if I gave you four taxa and thousands of genomes across those plants across an environmental gradient, and phenotype – that is a huge task and all we do. Look at individuals across space (intraspecific/intragenetic).

Dan Rohksahar: 1001 Arabidopsis genomes in parallel with 1000 human genomes – let them do it and plants will follow, so have iPlant do something else.

Second, discovery environments – we want Google/surf data to some degree... but that is not a scientific way – just rummage through data. Justin Borevitz gave a great example and there are others.... So full court press on a few specific targeted ways (crop and ecological example) and that is it! That will make it happen in 4-5 years...Ed Buckler's and Justin Borevitz have designed experiments that iPlant can answer, than those projects get NY Times worthy thing done but all of us will benefit.... But that we are all on board to say "go with that" since it will be ahead of me but also help rest of us...

Rich: hard thing: must be NEW and EMERGENT and solve new problems (not same-O same-o)

Also must be done in 2 years to keep going – a small scale example.

We need "poster children" of success – some use cases

ANNE – NSF: Must be biological question and URGENT: why now versus

later?

Nick Lauter: Not just a genome browser on steroids.

Rich: Keep computational people engaged throughout project or won't meet evolving expectation...they are part of keeping everybody honest – did you realize that I changed basic rules so need to keep them involved (with Grand Challenge Team/Integrated Solutions Team).

That is how you do collaborations – not just how to use nanohub and BLAST ...how work with Mike and Tim here...listening is important or computational person goes away...

...also...not, build it and they will come, but “they come and we will build it” ...two of five planned workshops... so if you don't like “adaptation” or “gene” focus then go to others... phylogenetic/comparative group is very gene driven and will help with transfer among groups...

also proposals we have not received: ecologists driven by ecologists not here, epigenetics/chromatin, etc.

Annie Schmitt: I think we have consensus here, lots of data coming in but advantage is that plants don't move so you can examine genotypes versus clines/environment (and reciprocal transplants) to make predictions. What else to talk about so just do it!

TMO: So I am confused – what is important enough is to do things we can not do otherwise.

Caveat: Five investigators could cobble together cyberinfrastructure. So what are emergent examples?

- 1 – At genome and TAIR – created a democratization – no big paper and NY Times article, but led to many studies.

- 2 – Human genetics and infrastructure - lots of good, not only susceptibility of type 2 diabetes but many people can do this. More democratization and dozens of papers.

So goal is not just 5 PIs and 5 nice papers, but 100s of papers from 100s of labs, only iPlant can do that.

Adaptation and Climate Change: very similar – all need genotype/phenotype/environment well characterized, all plug-and-play for 100 papers (1000 points of light :) to emerge.

Most pieces on shop floor globally (not just USA) so mostly interoperability.

Rich: so driven by community and grand challenges, this is what congress/public can see results, and democratization moves it forward (for science and citizen scientists)... moved things forward.

Jean-Luc: Now I am confused, democratization is not grand challenge.

Rich: yes, but grand challenge is not pipes... the social stuff is critical and Tom really put his finger on how efficiently it can get done.

Steve Roundsley: We need human diabetes 2 type example from you – so we need that type of example.

Drought can be sold but as plant physiologist of USDA-ARS, after 12 years our weakest link is global warming and phenology – Steve Welch can do this with Arabidopsis and spatial stats, we need to move to crops and disease/CO2/salinity/water/ozone/

PHENOLOGY is tractable over 2-3 years...

Ruth Grene: We are ready to do something? But I don't see it? So grand challenge is to integrate vertically, I work on draught...so that is my interest...

Rich: need to tell us that data are there, and if drought can't be done in 5 years than maybe it is flowering time/phenology.

Yes, flowering time/phenology is good system – and citizen scientist network – so let's use that as our user case

Rich – Loren Rieseberg would also like flowering time, salt tolerance, etc... so guess which is best to unravel, it is flowering time network/MADS box genes, and not salt tolerance or ...

Rich – a biological revolution here may be superceded by social evolution

TMO – back to type 2 diabetes example – possible because SNP density huge and sample size above 12,000 people, so want plant empirical data set out there for this?

China knows this and invested BILLIONS in transgenic agriculture.

David Salt – we are biologists, focus on questions – they are different.

Let iPlant figure out that if 80% similar than we need to have common goals.

FORMAT OF PROPOSAL: Must have execution plan if move forward later

GC Question(s) and subquestions/aims

Concept maps/

USE CASE SCENARIOS (or more abstract storyboard/

tools/AND DESIGN SPECIFICATIONS MUST BE THERE at some point...

GRAND CHALLENGE TEAMS MUST HAVE COMPUTER SCIENTISTS (or Ivan Baxter-like hybrids) THAT WILL WORK WITH IPLANT TO BE SURE WHAT IS BEING BUILT IS SERVING THE PLANT COMMUNITY.

- Feasibility – existing datasets
- ALSO EOT/outreach component...
- synergisms with other GC groups
- Leadership plan/team members/super-users/second and later generation of users

...page limits on this...15 pages or less plus Bios/management plan...

...feasibility...

deadline of Jan 31 (BOD meets March 9, so need external review during February).

...later will lead to full proposal and execution plan...deadline?...then make discovery environment...series of conference calls and then meet March 15 ?
...will lead to MOU...and process to collaborate on a common goal

6:00 (VP debates)

6:30 - 8:00 Dinner