

Grand Challenge Workshop: Mechanistic Basis of Plant Adaptation

Sept 30- Oct 2, 2008 at Biosphere 2, Oracle, Arizona, USA

Notes prepared by iPlant designated scribe

October 1st

Goals and deliverables of the workshop, and overview of the organization and agenda of the workshop.

Presenter: David E Salt. 15 min presentation and 15 min discussion/questions.

Why are we here? To define the functionalities we would like to see in a cyber infrastructure that supports investigations into the Mechanisms of Plant Adaptation. Key Questions, agenda, thought processes (for agenda, what are functions?). What are tools and cools stuff as a community that would help promote your research? Do not worry about specifics of deployment, etc. Stay focused on the biology.

What is expected outcome of workshop? To gather enough information and people to develop a Grand Challenge Project to submit to iPLANT by Jan 31st 2009. A proposal that describes the community's cyber infrastructure. How to build and how it works out is another meeting and another game.

What is the long term goal of workshop? Convince iPLANT to build, deploy, and maintain a cyberinfrastructure for our community. Effort needs to be committed to see plan/idea through by everyone in community (some more than others).

THEMES

- 1) Organisms, phenotypes and other data types (soil, temp, rain, wind, etc) new and existing data sets that need to be incorporated.
- 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. (I'm talking about the same thing you are)
- 4) Data analysis tools that would be needed, including next generation sequencing data.
- 5) Modeling tools for hypothesis generation and annotation.
- 6) Data visualization tools for data integration in 4D (spatial and temporal).
- 7) Educational outreach strategies.

Agenda

(overview of agenda for meeting – *see agenda for details*)

Q: David Niel, UC Davis: *Will there be some exchange of the two groups climate change/mechanisms?*

A: David E Salt: *After workshop, we will meet and determine if we should merge or stay separate.*

Q: *hard copy printouts of schedule of climate group in case we want to wander to other meeting?*

A: David E Salt: *Yes, we need to get these hard copy's for everyone.*

The Grand Challenge: Mechanistic Basis of Plant Adaptation. Presenter: John Willis: Duke University

Asked to attend iPLANT kickoff at Cold springs harbor. Gave a big picture talk, was approached by David Salt to think about plant adaptations proposal. Final group writing proposal could be totally new (not necessarily original 5).

Genus Mimulus (100-200 species). From tiny desert plants to large woody plants. Large Phenotypic variation in environment. Large floral diversity, outcrossing, self pollinating. High alpine, salt spray, thermal springs, serpentine soil, toxic copper mine soil (CA), arbuscular mycorrhizal fungi, low nutrient/low water availability, etc. Vary different ecological taxa can be crossed. Can help us answer such questions as; What is the molecular genetic basis of ecological diversity? Why *M. laciniatus* in high granite outcrops in Sierras? How do plants adapt to their environments? (soil, mineral, disease, herbivores, water, abiotic stress, flowering time, pollination biology, etc.), Environmental adaptation at many levels (molecular, cellular, to ecosystem).

Why study molecular basis of plant adaptation

- 1) Adaptive evolution leads to new species, increased biodiversity. What are evolutionary and molecular mechanisms?
- 2) Natural selection has resulted in astonishing diversity of mechanisms that plants use to cope with extreme and often unpredictable environs. How do the developmental, physiological, molecular mechanisms work?

Now is the time to answering these questions

- 1) We finally have tons of data to address questions in wide range of plant systems.
- 2) We have ability to synthesize and integrate data.
- 3) Looming global climate change crisis. How will plant communities will respond and evolve.
- 4) Urgent societal need (need for food, etc) Slide from Richard Flavell, Ceres showing doubling of yields of world cereal production since 1960. Actually a downward trend in rates of improvement in yield production vs time. Per capita level, actually has dropped since 1985 due to massive population increase. How to get to target with given farmland and less water, less fertilizers, more severe weather. This is one place where understanding plant adaptation could have a significant impact.

We have or soon will have more data than we know how to deal with.

Genomic sequence variation (SNPs, indels, CNV, etc.), Gene expression variation (cell specific), Variation in methylation and histone modifications, Proteomic, metabolomic, ionic variation (cell level), Accurate, dynamic, high throughput phenotypes in nature Ecological, geographical, climatological, geological data, "omics" data, histone modification, etc....

Sequencing improvements (dramatic reduction in cost, time, and man hours for Arabidopsis genome). Single molecule based sequencers in the future 2014? 2-3 min to sequence entire Arabidopsis genome for about 70 bucks!!

What about sequencing whole populations?

But how can we handle this data so that we can answer the big questions?

Example: ionomics-linking genetics with high throughput elemental-profiling. Plant growth→sample prep→ICP-MS to analyze ionome. Cyberinfrastructure for this project. Process huge numbers of samples each year (and increasing). Natural variation in ionome of *A. thaliana*. Ts-1 line from coast of Spain had lots more sodium in above ground shoots than others. DNA microarray based BSA mapping of high Na loci in Ts-1. Nailed it down to a single gene which is a known sodium transporter. Yes indeed those plants had greater salt tolerance. Same allele is found in several other worldwide lines (most on coastlines). This is a good example of local adaptation.

Another example: germplasm-maize example. Mutants-7600 stocks, recombinant inbred lines 6000, near isogenic lines, etc. Extensive germplasm resources.

36 million common polymorphisms in low copy regions of maize genome. 15,000 genetic stocks created (RILs, NILs, diverse association) and being evaluated 128-10000 SNPs on these stocks to project 36M polymorphisms. Etc.

How do we make a high throughput platform for QTL dissection??? Forward genetic, gene-level resolution, genome scans, and community resource.

Nested Association Mapping panel (NAM): Screened germplasm for 25 most diverse lines of maize to maximize diversity. Crossed diverse lines to B73 and develop RIL populations from each cross. Project joint efforts of several groups, and should allow QTL mapping to association mapping to eventually level of genes.

Genotyping (maize genotyping with sequencing 5 bucks/line). Phenotyping (accelerated phenotyping of monitoring millions of plants at daily intervals). Remote sensing for daily monitoring of N, water, growth, and disease progression) via balloon with cameras and satellite. Idea to monitor size, disease, etc. Traits being scored by community. Data collection, curating, and storage. Data processing, pipelining, packaging, statistical, graphical.... We should not reinvent the wheel for the cyberinfrastructure!!

Why current tools are not adequate

- 1) Stand alone, not integrated.
- 2) Not designed to handle large data sets.
- 3) New genomics tools to handle 20-50 times higher polymorphism than human genome.
- 4) More realistic breeding designs.
- 5) More realistic population structures.
- 6) Heterogeneous traits (yield or fitness).
- 7) Pink elephant: need statistical grappling of such large data sets.

Conclusion: New data and new infrastructure needed

Discussion: 30 min

1. What are the main biological questions?

Diane Byers Illinois State: how to link the different types of data so that they can make sense with respect to how plants to adapt?

Annie Schmitt Brown University: Evolutionary mechanisms of rapid evolutionary change or response with respect to changing environments?

Justin Borevitz U of Chicago: How to organisms multi-species interactions changing annual or season changes in carbon?

Torbert Rocheford Purdue University: Where all the genes controlling plant adaptation. Can we get at them and identify them?

Brain Dilkes Purdue University: Looking at natural systems, how do they deal with changes in environments over space. Look at evolutionary solutions. Lots of ways that human impacts in environment influence this (ecosystem services: water air, etc).

Dan Rokhsar UC Berkeley: Key step, to predict mechanisms of adaptation and then to design them in plants. This demonstrates we actually got the answer.

USDA at Cornell: All the genes and all the organisms and how they interact and in what environment.

John McKay Colorado State: molecular bio types: id genes for function. Or go from phenotype and find genes. ID genes first and then can understand variations.

David Neale UC Davis: Pattern of adaptation, mechanistic basis and pattern of adaptation (on a population basis). What is actually doable? Large databases of genotypes we can feed phenotypes into. We would like to bring these kinds of data together (can be distributed across several data bases). We need real time statistical analysis of the data.

Julian Maloof UC Davis: Whatever the infrastructure is it should handle gene networks and complex interactions (multidimensional gene space).

Diane Byers Illinois State: Integrate across phenotypes. Links between different levels. Phenotypes to molecule level needs to be integrated. But particularly high throughput phenotype integration and analysis.

Scott Hodges UC Santa Barbara: Spatial and temporal scale in a comparative way so we can identify particular systems that are always involve in a similar way. Some species, it is difficult to get to genes (everything but Arabidopsis!). Adaptation that is single gene or multigene based.

Infrastructure needs to be able to connect genetics with environmental data in a seamless way.

How well do we define environments? What is it and how well do we characterize it?

From a breeding point of view, when changing traits, how does it affect others? Ways around that?

Graham McLaren from CGIAR: what is the phenotype that you need to measure? Yield advantage is a devilishly complicated trait! Problem dissected into many components, trait interactions, gene interactions, gene-gene interactions with environment to produce phenotype. We need good genetic simulations. Build models that simulate genetic control that drive main phenotypic traits of interest. Dissecting complex phenotypes and environments.

Forward and reverse genetics techniques? Need two way streets.

Annie Schmitt Brown University: common genetic networks, crop level, scaling up issues. Bottom up versus top down approach: collecting genomic data does not always support bottom up approach? What about physiological studies? Phenology and project bud burst.

Long term data sets. Very amenable to education and outreach. How do we get value from these combined data sets?

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.

Demo of existing cyberinfrastructures

Michael McLennan – NanoHUB <http://www.nanohub.org/>

Demonstration movie of simulation tools on nanohub. Tools easy to launch well documented, results that are readily visualized.

77,000 unique users from all over the world (top 50 US engineering schools). Usage doubles every month or two. Integrate research tools, researchers, and make it easy to use. Not specific to nano: extracted to package HUBzero. Unique middleware for simulation and modeling, content management system for scientists, collaboration and social networking. Using HUBzero, NIH grant, pharmaceutical, manufacturing, global engineering, cancer care engineering, etc. All Hubs are very similar to Nanohub architecture

Platform like YouTube for tool developers to upload. Register tool, created, upload, install, approve, publish. Web based publishing system. >200 projects on nanohub with 89 active developers.

Movie showing community of developers developing tools. Surge in tool development just as semester begins. Very effective way to update tools. Bottlenecks removed.

Q: *what is the distribution of users.*

A: *77,000 end users, 500 contributors. Over 120 tools online.*

Help facilitate the sharing of information

Change expectations of experimentalists and educators

Increase the pace of tool deployment

Change the face of cyberinfrastructure

Damian Gessler - Simple Semantic Web Architecture and Protocol (SSWAP) and Virtual Plant Information Network (VPIN) <http://sswap.info/> and <http://vpin.ncgr.org/>.

The web has changed the way that we operate. But greatest limitation is that the vast majority of information is lost or dead. The logical web.

Challenge: science needs to integrate info. There is a gap in the ability to find and integrate disparate data on services in a high throughput manner. High throughput sequencing 2 billion bases/2 day per machine. By 2009 20 Gbp/2 day predicted. About 1TB image data/run!! 1000 genome project. Sequence 1000 human genomes. Two different sites in different ftp sites with different files. Cannot get data in a high throughput manner. The vision: a logical web. A web where resources describe themselves in a rich semantic way, amenable to reasoning by external agents. A web where machines can assume much of the burden of data and service discovery, engagement, heterogeneous assimilation, and integration. A way where humans contribute creative decision making increasingly farther downstream. Decision Tree: Do I need web services, do I need semantic web services, do I have the programmatic commitment to implement a new technology. SSWAP: simple semantic web architecture and protocol. Community accepted ontologies. Discover, engage, execute, return. Inputs, outputs, category of service. RDF: resource description framework. Subject:relationship (predicate): object. Reasoning extracts more information out of the underlying data. Eg. Symmetry. Take data and services and describe them in a formal semantic way. Make connections on services. Virtual plant information network. What service works with a given taxa.

Q: example when used the word taxa is not what attendee thinks it should be.

A: Free market place, different groups come up with ontological terms. Market place of terms of ideas.

Q: Example of workflow that allows biologists to actually do something.

A: Bring tech to table that can describe data services. We can all put the definitions in there, but people start using certain definitions over others (the wildfire effect).

Graham McLaren - International Crop Information System.

Bring biology to bear on food security. System for collecting phenotypic data. Advance crop improvement programs. A swiss army knife for crop research www.icis.cgjar.org. ICIS: international crop information system. Community, open source community. Can be implemented for any crop. Software makes new generation (new nursery of lines). Can click through pedigree or make pedigree tree. Guarantees correct annotation and documentation of data collected in field. Problem: People like to collect data with Excel, not weird web interface. Tool allows user to collect, annotate data and feed it into repository. Describe trait, description, condition. Modernize plant breeding to food security. IRRI: international rice research institute. Query IRIS database. International trials: genealogy.

Q: Semantic web architecture seems intuitive. Main disadvantage does not benefit industry wide approval. Where are we in this phase and when will it be an industry standard?

A: Money is bottleneck. Adding semantics is difficult. Legacy data is hard to go back to and add semantics. iPLANT time 5-10 years is a reasonable time frame for this maturation.

Presentation of a draft of the proposed cyberinfrastructure functionalities as a starting point for discussion. Presenter: Justin Borevitz 20 min presentation and 25 min discussion.

Standardization of phenotype and genotype data across multiple species. What is a model ecosystem? Genetics of adaptation. SNP/Tiling microarrays: SFP/SP.

Universal whole genome array: array hybridization: chromatin immunoprecipitation, methylation, polymorphism, comparative genome hybridization, etc.

Tool focus: SNP tiling array. Reference set for community to collect phenotypes. Which genes are controlling which traits? Genomic profile of cellular systems responding to the environment. Population genetics on the landscape: family structure 100-1000 SNPs, genetic association mapping.

Front End Cyber: sample tracking (eg cybertracker prints barcode to transmit into database), meta data eg field (gps, time, date, species). Environmental data (light quality/quantity), temp, humidity, wind soil data, etc.

Trait x environmental dissection

Yield...growth, development, harvest, index, partitioning

Environmental partitioning—microclimate mapping

Genetic variation can specify trait/environment networks.

Back end cyber: real time data visualization in google earth layer with animation. Subset summary data, statistical analysis, data filtering (quality control). Diversity within and between populations (using google earth to map out). Example of how different genotype frequencies change throughout the season. Summary of ecoregional partitioning of Arabidopsis. Solar calc II provides weather from GPS coordinates. Time lapse cameras will be extremely useful for imaging plant growth (time lapse photography). Eg. for whole genome association mapping: flowering time. Results: lots of genetic heterogeneity.

Other species...ecosystem genomics. Genetics of speciation along a hybrid zone. How environmental variation is shaping populations. Ecological genomics: target foundation prairie species, landscape genetics, use diversity mix for restoration.

Q: phenotyping natural population versus laboratory populations...but did you bring it back to the field? Phenotype doesn't always hold up.

A: Explaining environmental versus genotype effects on observed phenotype.

Q: John Willis: alternative approaches?

A: low density fingerprints, put diversity sets back in.

Q: Tim Kelley from Indiana U: functionalities of infrastructure have created their own operating procedures at the bottom up (huge developmental costs). Do we want to reinvent the wheel.

A: Take what you can off the shelf so you can plug in and just use it. Citizen science? Then get genes in there and their interactions with environment.

Session – Cross Education I

Organisms, phenotypes and other data types (soil, temp, rain, wind etc) to be represented (both existing and new). Kirsten Bomblies/David Neale/William

Beavis. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Kirsten Bomblies presenter: Plant adaptation is beautiful and important! What organisms, phenotypes and data types we prioritize depend on who we are.

- 1) Molecular biologists: focus on model systems
- 2) Breeders: focus on crop plants
- 3) Evolutionary biologists: large diversity of organisms
- 4) Ecologists: Large diversity of organisms

However, we all want to know how plant fitness is affected by the environment and how plants deal with it.

Model systems are crucial. An in-depth understanding of some systems can inform about other systems less well studied. Reverse genetics: Genotype→phenotype in different environment or species. *A. thaliana*, *M. guttatus*, *A. Formosa* all have genomes sequences and other resources. A model family- the relatives of *A. thaliana*: used for studying many other model systems (metal tolerance, etc.). If reverse genetics fails, must use forward genetics phenotype→genotype QTL analysis, SNP/phenotype association studies, etc. QTL mapping is a useful tool for studying adaptation-an example from *Mimulus guttatus* x *M. lewisii* (need genetic maps and markers, phenotypes, F2, RIL or other segregating populations). Many adaptive traits have been QTL mapped: cold tolerance, drought and salt tolerance, etc. Information can be integrated across QTL studies. Getting to the genes can be tedious but there are some cloned gene examples. E.g. Petunia: loss of a transcription factor can account for most of the difference in pollinator attraction/flower color. Studying plant adaptation is not limited to small plants. (eg. Dendrome website). Comparative re-sequencing in pinaceae. ~15 million trees. Traits in tree world: growth, emergence, cold hardiness, bud-burst, etc. Environmental variables: temp range, precip, temp fluctuation, aridity, last spring frost, first fall frost, ect. Spatial correlations between traits and genetic structure (Eg. Pacific northwest). GIS based mapservers may help pinpoint candidate causal environmental differences. Layers can show many variables (nickel and other metals in soil, soil acidity, etc.). Then overlay distribution maps of plants.

Existing and new data sets that need to be incorporated. Lila Fishman/John Burke/Thomas Mitchell-Olds. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Presenter Thomas Mitchell-Olds: How will plant communities respond to climate change? Evolutionary data. Distribution of data. Species and climates, bioclimatic envelope info, etc. Genetic resources SUGGESTION: take ICIS database and implement it online. Then establish portals for other model species. Sequences and genotypes. Put an end to reinventing the wheel. RAW DATA IN..RESULTS OUT. Comparison to reference genomes. Users input raw reads and name of close relatives. Biodiversity data (USDA Plants), encyclopedia of life, etc. Modeling complex systems: compartmental models: genes, phenotypes, and environments. Machine learning analysis of protein interaction networks (hairballs and rats nests). Modeling complex systems. Functional models based on known mechanisms.

Q David E Salt: What data types would we like to see in the cyberinfrastructure

A: Annie Schmitt: tracking past genetic changes, or even paleo records (at least for trees).

A: climatic data: global climate model data sets interpolate weather station data so

A: *plant metabolites*

A: *research and citizen science and representation of data: what about quality control? Citing and data credit.*

A: *obvious problems with WIKI. How many authors?*

A: *Li Liao: WIKI like system developed at MIT allows tracking and*

A: *Chris Pires: What about other organisms, nematodes, fungi, bacteria*

A: *Diane Byers: statistical approaches for combining these different sets of data. Variable quality data, large data sets, wasted time on bad data.*

A: *Annie Schmitt: Geographical representation of data: GIS for dummies?*

A: *Paul Quick: series of recommended data annotations from community*

A: *David Bubenheim: acknowledge that data resources list will be inadequate and will need to be modified as time goes on. Identify critical items that need to be improved now.*

A: *Bill Beavis: statistical analysis, we don't know what analysis will be. Need to have capability to incorporate new analysis tools (flexibility).*

Formalized data acquisition platforms to support data collection in the laboratory, field and across sites. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Presenter: Timothy McPhillips: Collaboratory for macromolecular crystallography at SSRL. Blu-ce for data collection: run locally or remotely, multiple copies can run simultaneously, very secure. Challenge: determine structures of all proteins in *T. maritime* built on collaboratory capabilities. Data acquisition takes place at many steps in addressing a grand challenge. Management built cyberinfrastructure around users and it worked!

Lessons

- 1) Don't start from scratch.
- 2) Enable researchers to use their favorite tools.
- 3) Don't view web interfaces as a panacea.
- 4) Be in production from day one.

Two kinds of data

- 1) Data that could be useful to researchers in other projects.
- 2) Information representing the internal state of the project.

Grand challenges require data integration: automation is key, but make sure it is right!

Data integration (same type of data, different types of data)

Look deeply at how data you are using was collected? Need to trace back all the way back to measurements or observations and what their limitations are.

Workflows: not all are linear. But complexity can be encapsulated in subworkflows. Solves problem of recording, reporting, and confirming how computational research is performed.

Ask iplant to:

- 1) Focus on data management infrastructure
- 2) Employ existing and emerging standards for metadata
- 3) Partner with and leverage other research groups efforts developing related technologies

Mark Schildhauer's presentation: Formalized data acquisitions platforms to support data collection in the laboratory, field, and across sites

Don't use excel for statistics!!! But what are the alternatives? We need to think about this. Cyberinfrastructure for Holistic Biology.

Increasing need for collaboration and synthesis to solve vital, complex questions in biology- from gene to ecosystem. Ecological data are highly heterogeneous. Personal data management problems are vastly compounded in collaborations (need standardization: metadata). Data organization, documentation, analysis, preservation and archiving. Technological solutions: Confederated data sharing framework, analytical software that is scripted, verifiable, re-usable. Fee, open-source, multi-platform software for data management and analysis. Virtualized "centralcollaborative workspace" like Nanohub. Broad compatibility with other frameworks. Many other ecoinformatics products: ecological metadata language (EML), morpho, metacat, vegbank, ecogrid, kepler
Other cyberinfrastructure efforts (SoNet, VDC, SEMTOOLS, NSF OCI programs

Conclusion: iplant should partner/collaborate with other CI efforts
Assure broad compatibility with other CI efforts in biological science

Q: data acquisition, central versus distributed data?

A: Semantic web approach is best over generalized global schema. Gives scientist deeper penetration into data they need and interpretation.

Q: what do you do about Arabidopsis annotation which changes every 9 months, keep track?

A: there are approach to tracking deprecation.

Q: geospatial data, climate collections, deparate data that we have to accommodate.

A: Collaborate with OGC which has fast track to isostandardization. Google earth not quite formal and comprehensive enough.

4.0 Discussions of Presentation of summaries from Breakout I

Rich: lot of discussion about data, how does it relate to the grand challenge question. Is the solution of the problem tractable using these data? This needs to be clearly stated

A: Yes!

How to integrate across genotype and phenotypes (technical grand challenge)?

What are the main questions, they are always too narrow? What is the mechanistic basis for plant adaptation? Complexity and diversity: unique feature of adaptation where there is not just one answer. The broad approach is necessary to answer the grand questions (synthetic comparative way). To generalize and find mechanisms? Can we predict patterns? Based on environmental input, can you predict what will happen?

October 2nd

Session – Cross Education II

Data analysis tools that would be needed, including next generation sequencing data. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion. **Dan Rokhsar/Jean-Luc Jannink**

Presenter Dan Rokhsar: Mechanistic basis of plant adaptation. G+GxE→P: Datasets

This meeting focused so far on genetic approaches (relationship between genetics, env, and phenotype). New sequencing technologies, cheaper faster. Tools for new seq data? Raw data sets are huge (actually ship the hard drives!) Transformed into sequence and quality files at core facilities would be best for iPLANT. Sequence mapped to "reference". Does iPLANT want to be involved in this? Maybe not? Beyond SNPs? What is the reference genome? Arabidopsis Col-0 is essentially a random reference. The way in which genetic mapping tools are used. Basic problem: regression of traits vs. alleles genotypes. Multiple existing software tools with related but distinct methods, strengths, weaknesses. Traits: anything that can be readily measured. But most can't be readily standardized across taxa. We should make available templates for phenotyping. Balance data privacy with public release (initial sharing within research team to take advantage of iplant tools, but clear path and timeline to public release. Minimal information about experiment standards and reporting format. Leaf shape in snapdragons example: standardized 30 different coordinates on a leaf, analyzed by principle component analysis. Tools for dimensional analysis would be very useful. For the snapdragons, different species filled out different areas on 3D plots of leaf shapes. Markers associated with trait of interest (want to get down to gene). Use synteny homology to get yourself a list of candidate genes. In one report, get info on these genes. Add other functional datasets from external databases. iPLANT databases does not swallow everything else (other databases should remain intact and supported by iPLANT). iPLANT is not going to be a public repository. iPLANT is however, a place that generates the integrative report. "bridges" are part of the cyberinfrastructure. Example of comparative tool: JGI database www.phytozome.net. Reconstruct the genomes of plants as decedents of ancestral genes (based on inference of function of ancestor gene). Example HMA4 locus: ion transporter. GeneGroups: various natural groups of genes, part of pathway, coexpressed, anyways here is a bag of genes we are interested in (many definitions). Bulk load these gene groups. Could map arbidopsis gene group to tomato for example. Conclusion: data is now easy to get, but to answer your GC, ease of using the data set needs to become easier. This is where iPLANT comes in.

Presenter Jean-Luc Jannink: 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. But there is a barrier: resources for different comparisons are dispersed and not amenable to assembly into a systematic workflow. Generalizing across instances. Adaptation is multiple, many ways of getting at it, and patterns or generalities need to be identified, but currently no resource exists to really do this. Hooks to comparative information: homologous loci, response to environment, correlated expression, synthetic lethal, correlated traits, syntenic loci, physical interation, etc. Correlated traits which you think locus impacts. Eg. Locus effects flowering time and plant height. + other info linked to loci. Also trait centered way of accessing information: known pleiotropic loci, correlated plasticity, homologous traits, correlated traits, multivariate dimension reduction. Or environmental stress: known mechanisms, known loci, same species correlated responses, other species responses, datasets with variation in this stress.

Discussion:

David E Salt: we need both vertical (deep set of info for model species) and horizontal (multiple species) approaches.

John McKay: assembling sequences against a reference may not be done at iplant. What about reference?

Issue of what is the reference genome: reads are projected against pangenome rather than a single reference. Human community may do this first, but Arabidopsis may beat them. This should be done by a genotyping center. In other words, data set should be filtered at the source.

Lance Waller: sequence and phenotype data. However environmental data huge! What do we want to hook this genotype data to? Google earth can geolocate information. Virus evolve very quickly.

Brian Dilkes: Solexa data should be reduced as much as possible before it gets worked on. But iPLANT could catalyze research where a reference genome has not been produced. We must retain the capability to move these large datasets around.

But what do we do with seq with no reference genome.

Justin Borevitz: geospatial tags are also a good way to hook data. Also haplotype maps: where are these types distributed.

Cyberinfrastructure for outreach: pictures of plants georeferenced for students to go find? Also one hook that is missing is the strain!

Bill Beavis: Spatial statistics: engage statistical community to compete with each other over given problems.

Modeling tools for hypothesis generation and annotation. Herbert Sauro/David Weston. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Presenter Herbert Sauro: What is systems biology: systematic biology (top down), network physiology (bottom up), synthetic biology (engineering). Not good overlap between the three areas. Top down "omics" whole cell, statistical correlation models, and high throughput data. Bottom up "mechanistic" models, networks and pathways, mechanistic and biophysical models, quantitative, single-cell data. 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. Our biophysical model of the cytoplasm is more or less correct? Infrastructure requirements for mechanistic systems biology: Experimental data from hypothesis driven experiments. Databases/exchange standards, controlled vocabularies and semantic annotation, software provisions. Example: a real network in E. coli. There are many modeling software tools (huge!!). Tools are not intercompatible. Exchange standard developed **SBML**: system biology model language. SBML focuses on biology, mathematics will be automatically derived. CellML is a math based description from which underlying biological can be inferred (basically matlab scripts). Hackathons to hack out details. Structures: lists of lists: functions, units, compartments, species, etc. Also systems biology graphical notation. Simulator comparison and compliance now available. 185 curated models as of Aug 2008 which are all annotated. www.ebi.ac.uk/biomodels MIRIAM: minimum information requested in the annotation of biochemical models. Semantic annotations (SBO, MIASE, TEDDY, KiSAO, Missing). Terms can be queried programmatically via a web service. Challenges: software continuously gets written, thrown away, then rewritten. The field is littered with partly finished and abandoned applications. Split work into two groups: 1) specialists libraries 2)

enduser applications. Many standard and reusable libraries eg: libMFA metabolic flux analysis.

Integrating statistical and mechanistic modeling:

Q: Is this a very generalizable to any network or pathway?

A: Yes it covers any mechanistic network. You can log in and suggest new branches to tree on website.

Q: Is it flexible enough to annotate every node?

A: yes every tag can be annotated.

Q: Competition for algorithms is great

A: yeah nice!

Q: lots of discussion on standardized vocabulary and so forth, but what is the enforcement mechanism of these standards. It really is difficult to do this, and the only way to do it is via the journals to force the editors to be whipped into shape?

A: Rich: policing is not the answer and as a formal journal editor I can tell you it won't work. The community really has to agree on standards.

Q: compare your approach to other industries like Bowing and Wall street and intel.

A: you do model smaller systems but also at higher levels as well, but it is very modular.

Q: Are your developers a core set or do they come and go.

A: we are very committed and there is a core that does not really change. Some have left some have added.

Q: people can compare the simulators?

A: can the users actually load the model? Time courses compared.

Q: how can we use this to answer questions about mechanistic plant adaptations?

Data visualization tools for data integration in 4D (spatial and temporal). David Bubenheim/Tim Kelley. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Data integration challenge (source and scale). Example invasive species. Plants that know how to adapt like yellowstar thistle. Landscape scale: helps answer questions about how plants are able to adapt to various environments. Influences of grazers and pollinators. Also: Visualization challenge. We want to go beyond what we can analyse and communicate today: (Tim Kelley): Good visualization aides in analysis (data missing? Helps filter data, helps generate hypothesis. The human visual pathway can search through 15 million pixels per minute!!! (looking for dark pixels on white pixels). Good visualization is specific! Problem with plant science is that it is multidisciplinary and links with many other fields. Virtual reality modeling of biomolecules (University of Calgary). Dynamic simulation of ecosystem population (predator prey). Example of deleting one parasite species causes extinction of half the other species!! David: Multiple satellite observations provide global perspectives. Science change is one funeral after funeral. Remote sensing: When does a satellite go over? 2/day? Resolution can be replaced with higher flyover frequency. Vs landsat which flies over 1 to 2/week but gives high resolution (but clouds can influence). Spectral analysis: multispectral sensors and hyperspectral sensors: what kind of process are you trying to get out of it? In some cases you need to generate custom packages to get the data you want out but often established packages can be used. Example local scale biophysical mapping of leaf area index in a plant canopy. Vegetative index: NDVI gives

"green-ness" index. Looking at how things evolve over time is very useful (seasonal variability, annual variability). In one particular pixel (500 m resolution). iPLANT can integrate remote sensing with some of the core based science. Can determine species distribution. Target species: yellow starthistle, waterprimrose, cheatgrass, saltcedar. Controlled environment studies (controlled chamber studies for trace gas exchange with plants). Scale up to ecosystem and find markers at large scale to determine if native plants can compete with invasive species. NASA viewing data. Soil water assessment tool (SWAT) for studying hydrology, erosion, and water runoff with varying soils land use, and management conditions. Ecosystem simulation and visualization hydrologic unit/watershed system. Has been used at large scale: Mississippi river surrounding land to model runoff and water quality. Where are the problems along the river? USDA can focus conservation activities. SWAT accepts many input types (arcVIEW). Daily now casts (icandy:google earth kind of approach). Assessing ecosystem function: can click on map and get any kind of data that was collected. Good structure or template of how to integrate data for iPLANT. Gives you good connection to other data sources like weather and microclimate. Want to make model operable on a supercomputer. Also can provide a nice framework for boundary conditions.

Q: Justin Borevitz: Seems like it quickly gets overwhelming with the amount of data. What is the best way for iPLANT to hook into that?

A: user interface mode and a computer connection. Might have different interfaces at different scales. For a user interface, should base it on something that is comfortable and people are used to doing.

Q: Visualization dynamics: what kind of system perturbation could be done with this sytem for forecasting?

A: yes, that is being expanded right now because it came out of the agricultural section. It's really driven by what plant do you want to work on right now, and it's a good chance to start populating these.

Q: how global is the model?

A: well you can run it..... If we can match the stream flows in new areas, especially if we can go back in time, it gives you a pretty good idea that the system is working. But we feel much more comfortable with some ground truthing. Lots of ground work needs to be done to validate it.

Q: how much resolution can you get with distinct clusters (of various plants).

A: depends on satellites and bands. Aircraft hyperspectral gives you very fine wavelength segregation. Lots of this is data mining, where we don't know exactly what pattern or signal we are looking for for a given plant stress or condition.

Unplanned joint discussion with both workshop groups about GC projects:

Q: how long do you expect projects to go?

A: 0.5-1.5 years. Cyberinfrastructure development will be done by iPLANT development team and/or contracted out.

Q: NSF: consider other agencies for funding sources.

Ruth Grene: 10 questions that arose from climate change group

- 1) What are the genes and pathways responsible for variation in response to environmental stresses
- 2) Gap genetics and ecophys response to climate: what tools need for closing gap for future environments?
- 3) What are best levels of description that capture overall phenomenon (cyberinfrastructure needs)
- 4) To what degree are stress responses coupled?
- 5) What data and tools from individuals should
- 6) Provide env for research where tools can be added from community
- 7) Env. Should enable collaboration
- 8) Understand genes for vegetative and reproduction under stress
- 9) Ecosystem services over 5-20 years
- 10) Best climate models?
- 11) The rediculome (protein protein interactions).
- 12) What kind of IT environment?
- 13) Comparative rediculomics

David E Salt: Over view of Adaptation groups concept

Need to synthesize three data sets (genome, landscape, biochem-physiology anchored data sets) to achieve our biological goal of understanding mechanisms of plant adaptation. This synthesis will require cyberinfrastructure to integrate existing and new data sets and new instruments. Data analysis and visualization tools and models that can integrate these data sets and be predictive and user friendly (mechanistic).

Educational outreach strategies. End with a summary and discussion of topics for the breakout session. 20 min presentation and 25 min discussion.

Presenter: Kay Havens - Reasons why public understanding of plant science is critical. Citizen science programs aims to involve general public in meaningful research. Provides large data sets that are useful. Tools: smart phone technology. Mobile and environmental sensing platforms that can be used to sense pollutants attached to smart phones. iTrail on iPhone but could be used with GPS. Cybertracker developed in Africa. Project budburst: national phrenology network. Best are easy to understand and hard to go wrong on. The national institute of invasive species partnering with honeybee.net.

Q: Justin Borevitz: how many bud burst data points.

A. 6000 this year.

Q: what about farm data?

A. Yeah!

Q: what about other plant organizations that target armies of old people with money.

A: yeah we work with garden clubs.

Q: put at top, we need to get public to understand how a plant works.

A: oh yeah!!

UF-STEP Program Goals

To assist scientists in communicating their research to targeted audiences through innovative media to enhance grant projects. To formally train grad students in

communication, critical thinking, problem solving, team process skills development, change management. Strategies: create adaptable model for science communication, education and outreach. The STEP model: make science accessible and dynamic. Various target audiences. Create materials that are convenient and effective for an audience. Ufgenetics.com. Communicates cutting edge genetics research. Short movie clips now on youtube and other public networks. Videos, short content, career interest, ipods, etc. Lesson Plans middle and high school. For iPLANT collaborative:

- 1) Active learning online environment/network. Website plus mobile learning environments: videos, case studies, lesson plants etc. Web 2.0 based network functions. Eg. Peanut genomics training.
- 2) Distance ed. Offerings for graduate students ect. Technical and soft skills training. Internship.
- 3) Training of underrepresented groups and teachers. Work directly with researchers to get technical training. Bring cohort groups together around a central theme.

Q: simulation of how plants actually work and grow and get used. What about games for kids? What about school assignments and where do they go to get the info. Wikipedia? Make sure that is accurate.

A: target audience of home schoolers really go online heavily for research.

Q: instructional videos fitted to teachers or what?

A: it's a two way street. Minimal standards for teachers, presentation and content.

Q: Business model used. How do you get people involved to actually get it out?

A: that is why iPLANT is potentially so interesting and powerful! Sustaining momentum is not easy.

Q: can iPLANT build cool flash animations and media about my research project?

A: iPLANT can help with the cyberinfrastructure piece and others can run it.

A: Martha: The higher education piece..think about what you can do differently in your teaching with the iPLANT cyberinfrastructure? Think hard about that. The education outreach and training piece needs to be interdisciplinary just as the science using iPLANT cyberinfrastructure. Dave: This is about computation thinking. Dave Michilas: work with grand challenge development to adapt material and create material for those online challenges. So yeah we can make flash animations about your research. Example, check out www.dnalc.org.

Presentation of summaries from joint Breakout II

iPLANT can't sequence, but can build tools to manage the sequences (genome, transcriptome). The plant community doesn't have a genome to annotate. Landmarks on genome. Upload many kinds of data and during so, they become integrated. Therefore give us the upload tools, analysis tools, and interactive tools. Community shared projects, interconnects the big data sets and provides space for interaction between scientists. Provide tools for the whole cycle (laying over many data layers). Need genome picture rather than single gene. Genome to drought, genome to salt, genome to climate, etc. No boundary between cell bio, physiology, biochem, etc. It's all plant bio. "complexity is how does genome interact with environment to produced the observed phenotype". Data integration is hard, not just throwing data into a database and putting a front end on it.

Intraspecific analysis of genetic variation (sequencing individuals). 1000 human genomes already being done. Probably not a good idea to suggest that plants can also be done at same time. Central data repository which we can surf is not really a scientific problem. We need to propose something that we can't do in short term, but can have a large impact when completed or answered. Cyberinfrastructure needs to be focused. What are the genetic and physiological responses to stresses? Projects in 4-5 years. Tools should focus on these projects. Genomic, phenotypic data, environmental data. Integrating data is a great challenge, but not a grand challenge. What about data that already exists? It's not if we build it they will come, it's if they come, we will build it. Must know biological goals first. Is there a supposition, that if it is an iPLANT, then it has to be a massive data set? It is not about genomics, it's about plant science. Integrating across scales is highly desirable (molecular to ecological). The key is what is novel, what is new, etc. High throughput sequencing is just one idea. Top down and bottom up type of approaches. Plants don't move so you have lots of geographical information. Just do it. But, can't the top five investigators cobble together some of these grand challenges without iPLANT? Real goal is not just five good papers from five good labs but hundreds of papers from all over the world. Diabetes II example is what we need from plant science. Drought is a good one, some think it is a bad one. Temperature, CO₂, water, salinity, ozone, nitrogen dep. The grand challenge is to integrate vertically all levels in plants (models and measurements). Tons of phenotypic data out there. Comparative data is not hard to get. Drought is a complex issue (what was timing, what was initial conditions, etc.). Genes duplicated or triplicated in dicots. Flowering time network is under a lot of selection, and can potentially respond quickly to global change faster than others? Peak water, food, oil? It's going to be the plant science community that people will turn to. That is the grand challenge...food, water, and fuel production. Biological and social revolution will occur in the current century. More people the better...What are the poster child examples? Photosynthesis: past 50-60 years from engineering has not really increased very much. Huge amount of data on it. Provides a nice test bed for multiscale modeling. Ironic that people want to throw out drought because it's too hard but are looking for a grand challenge. What makes C₄ photosynthesis work (has involved multiple times separately, high light, dry, high oxygen). More efficient in more stressful conditions?