

## **Impacts of Climate Change on Plant Productivity World-Wide. Prediction of Phenotype from Genotype. Data Integration for Analysis and Prediction Across Process Scales**

### **Organizers**

**Melanie Correll.** Assistant Professor. Dept. of Agricultural and Biological Engineering. Univ. Florida. Gainesville, FL 32611-0570. Email: correllm@ufl.edu Phone: (352) 392-1864 Ext. 209. Fax: (352) 392-4092 Abiotic stress in plants, root physiology, root functional genomics. Education and outreach to high school students in the Student Science Training Program (SSTP) and 4-H, undergraduate mentor in Science for Life Program at UF (HHMI).

**Ruth Grene.** Professor. Plant Physiology, Virginia Tech. Blacksburg, VA 24061. Phone: (540) 231-6761. Fax: (540) 231-5755. E-Mail: grene@vt.edu Functional genomics of abiotic stress responses in plants. Diversity of stress resistances within crop species. Teach Biological Paradigms for Bioinformatics (grad course for computational students). Faculty Liaison for Virginia Tech Alliance for Minority Participation ( part of VA-NCAMP) in STEM fields.

**T.M. Murali,** Assistant Professor, Computer Science, Virginia Tech, Blacksburg, VA 24061. Phone (540) 231- 8534. E-Mail: tmurali@vt.edu. Computational and systems biology. Cellular response networks and their building blocks. Design and analysis of algorithms. Teach Computational and Systems Biology

**Stephen Welch.** Professor. Theoretical Plant Modeling. Kansas State Univ. Manhattan, KS. Tel: +1-785-532-7236. Fax: +1-785-532-6094. E-mail: welchsm@ksu.edu Crop systems simulation modeling with emphasis on applications, genomics and crop modeling. Electronic information delivery and decision support systems, especially in agriculture.

**Jeffrey W. White.** Plant Physiologist. Arid Land Agriculture Research Center, USDA ARS, Maricopa, AZ. Tel: +1-520-316-6368. FAX: +1-520-316-6330. E-mail: Jeffrey.White@ars.usda.gov Using ecophysiological models and geo-spatial tools, studies plant response to global change factors in agroecosystems.

**Pamela Ronald.** Plant Pathologist. UC Davis Tel 1530 752-1654, email Pcronald@ucdavis.edu UC Davis. Ronald studies the role that genes play in the rice plant's response to its environment, with a focus on disease resistance and flood tolerance. She leads a K-12 program for students about rice genetics and bioenergy and written about genetic engineered for the general public.

### **Summary**

Anticipated global climate change will require directed adaptations of crop species on an unprecedented magnitude in order to sustain agricultural production. Our Grand Challenge seeks to dramatically improve quantitative prediction of phenotypes by facilitating the characterization of "molecular pathways" of ecologically and agriculturally important traits affected by climate change. The project will create computational tools that enable the integration of phenotypes across diverse species. A key focus will be the development of infrastructure tools for the global integration of all available high throughput data for effects of abiotic stress factors, associated with climate change, that shape crop and non-crop plant performance in natural environments. Such integration can greatly accelerate the generation and testing of new hypotheses. The most promising hypotheses can be implemented in process-based simulation tools for further testing and application to climate change research issues ranging from plant breeding strategies to regional impact and mitigation studies.

Within the next few years, enormous amounts of functional genomics, proteomics, metabolomics, and comparative genomics data will be acquired. Multidisciplinary approaches are needed that use current biological knowledge to relate genetic changes to phenotypic outcomes, as are young scientists that are trained in more than one discipline. The existing datasets are often incomplete, dated, difficult to access, lacking in quality control, and error-prone. Nevertheless, they contain valuable data that are under-utilized. Using statistical and computational methods developed within the iPlant context, these data will be integrated to generate more accurate views of gene function and to identify species- or genotype-specific deviations, alterations, and adaptations to abiotic stresses associated with climate change. Such networks can directly inform phenotypic predictions of developmental characteristics in varied tissues and environmental contexts, which can guide plant breeding. This level of modeling can also support how processes are represented in the more integrative crop simulators, which typically provide quantitative predictions of phenotypes at the tissue, organ, plant or community scales and that have been used in climate change studies of both agricultural and natural ecosystems.

This workshop will generate new, and intensify existing, communications concerning form and content of available datasets across different plant communities, and then relating those, where possible and useful, to the model plants *Arabidopsis* and rice, and to major crop species. Through the Grand Challenge proposal development process, we will identify existing data and software resources, computational and mathematical gaps, and relevant models toward the goal of predicting crop phenotypes from genotypes. Three major plant productivity traits affected by climate change will be discussed in detail: (i) ***temperature, atmospheric CO<sub>2</sub>, and ozone induced responses***; (ii) ***water deficit and flooding induced responses***; and (iii) ***phenology***. The proposed Grand Challenge Workshop (GCW) will include participants from the fields of plant biology, genomics, bioinformatics, crop breeding, crop physiology, plant simulation modeling, computer science, mathematics, and education. Expected workshop outcomes include an initial consensus on key limitations in cyberinfrastructure that limit current modeling approaches, a better understanding of how different quantitative models interconnect, prioritization among traits, and identification of working groups needed to finalize the Grand Challenge Proposal.

## Introduction

**Our Grand Challenge seeks to radically improve the characterization of the “molecular pathways” of selected ecologically and agriculturally important traits affected by climate change with the specific, testable goal of improving quantitative prediction of phenotypes.** We are guided by the call from the National Academy of Sciences for “a significant broadening of the National Plant Genome Initiative mission to include the basic biology of economically relevant traits in models and crop species, deeper investigations into plant diversity, and plant adaptation to various ecological niches, and continued expansion of translation to breeders and farmers”<sup>1</sup>.

The focus of this Grand Challenge is on documenting, modeling, and understanding the plant global change response syndrome. This will generate predictive tools to guide crop breeding toward assured future food security and knowledge-based ecosystem preservation<sup>2</sup>. Responses to current and future food shortages require accurate predictions of how individual genotypes within crop species respond to abiotic stresses induced by climate change. This challenge aligns with the “genotype to phenotype problem”<sup>3</sup>, which, in reality, includes the effects of **Genotype, Environment, and Management on Phenotype (GEMP)** and is highlighted as a core research priority for international agriculture (e.g., [www.irri.org/media/press/press.asp?id=126](http://www.irri.org/media/press/press.asp?id=126)).

In the last 40 years, the numerical prediction of plant traits in the field has advanced greatly through models that are important springboards for the GEMP problem. They are used in climate change impact assessment<sup>4</sup>, and efforts have begun to incorporate realistic genetic controls<sup>5, 6</sup>). Genome structure and gene function information in model dicots and monocots (*Arabidopsis*, rice, poplar, maize) provides a reference framework to compare biological processes across species. High-throughput technologies have uncovered a wealth of information on gene expression and regulation in various developmental stages,

conditions, and biological processes. Genome-wide reverse genetics analyses of gene functions are being addressed by mutations and natural variants.

Furthering the quantitative fusion of ecophysiological and “omic” approaches requires a robust cyberinfrastructure, a discovery environment, to integrate and analyze information from gene to phenotype, across species, and from the laboratory to the field. Research tasks to be facilitated by this discovery environment include (1) using genomic and phenotype data to extract approximate network structures responsive to climate change induced stresses, to (2) suggesting critical experiments, to (3) the creation, modification, and testing of models. These activities are not a linear sequence but should fluidly branch and iterate in order to optimally exploit/drive insight creation. Because of the many ecosystem services that plants provide and the ubiquity of climate change, the methods and models to emerge from this discovery environment should find broader use within plant science.

Our Grand Challenge activities emphasize key traits affected by climate change and as expressed in species having large amounts of existing genetic, metabolic, and/or phenotypic data, and already being modeled for environmental responses, albeit with variable levels of skill. This existing knowledgebase will allow for rapid integration into the cyberinfrastructure that we will develop to meet our Grand Challenge. We will focus on three response types affected by climate change, described briefly below.

***Water deficit and flooding induced responses:*** The frequency and severity of droughts and flooding (e.g. Myanmar) are expected to increase in many regions, while decreased runoff impairs reservoir and ground water recharge. This creates a need for drought and submergence tolerant crops, especially those with high water requirements (e.g. drought.irri.org). Inter- and intra-specific gene expression studies have identified multiple pathways that are regulated in response to water stress<sup>7,8,9,10,11</sup> where available sequence data often reveal similar promoter motifs in orthologous, water deficit-responsive genes<sup>12</sup>. In general, osmotic changes, “recorded” at the cell membrane, are transduced in the form of rapid metabolic changes involving generation of reactive oxygen species (ROS), ABA metabolism, phospholipid metabolism, the generation of secondary messengers, calcium sensing, antioxidant signaling and defense pathways, and the synthesis of protective molecules<sup>13,14</sup>. Integration of high throughput datasets across species and stresses will allow us to enhance our understanding of these cellular responses to abiotic stresses across plant species.

Rice and maize have been genetically modified based on modern genetic information for drought tolerance, water use efficiency or flooding tolerance<sup>15,16,17,18</sup>, and some of these genotypes are already dramatically improving yields in farmers’ fields<sup>16</sup>; (Mackill and Ronald, unpublished). Various crop improvement programs targeting drought-prone environments already use simulation models to interpret genotype and environment interactions

***Temperature, atmospheric CO<sub>2</sub>, and ozone induced responses:*** Elevated carbon dioxide [CO<sub>2</sub>] and ozone [O<sub>3</sub>] are increasing worldwide due largely to human activities. Many plant species experience stress in this altered atmosphere, defined as the deviation from normal, evolutionarily shaped homeostatic conditions<sup>19,20,21,22</sup>. Increases of just 20% over the natural concentration of ozone, which are common, reduce yield in a number of crops<sup>23</sup>. The necessity to increase crop production worldwide will require enhanced efforts in breeding both ozone-resistant genotypes, and genotypes adapted to function in a different CO<sub>2</sub> atmosphere from the one in which central metabolism evolved over millions of years. Relevant mechanisms include, in addition to those mentioned above: altered signaling between carbon and nitrogen assimilation pathways, and multiple control points for the expression of photosynthetic genes and proteins, partially through sugar sensing (which is also linked to stress responses<sup>24</sup>). Greenhouse gasses and temperature can interact as stressors. Under high CO<sub>2</sub>, soybean plants in the field were warmer than controls (Bohnert, pers. comm.), and gene expression data from plants under elevated CO<sub>2</sub> showed induction of several heat shock proteins<sup>22</sup>. This temperature sensitivity adversely affects crop yield<sup>25</sup>. Mechanistic models of leaf photosynthetic metabolism will be critical tools for establishing direct impacts of climate change factors on plant productivity<sup>26,27,28,29,30,31</sup>. Extant data and modeling environments will facilitate quantitative lipid modeling<sup>32</sup>, and brassica oil seed crops like rape and Lesquerella<sup>33</sup> present opportunities for leveraging from Arabidopsis. Warming and elevated CO<sub>2</sub> impacts

were estimated by ecophysiological models by the IPCC<sup>4</sup>, but the reliability of model assumptions remains controversial<sup>34</sup> since the models failed to consider stressors including O<sub>3</sub> and heat shock.

**Phenology:** The timing of key events such as anthesis and maturity is often a primary determinant of net productivity and also influences other responses to abiotic and biotic stresses. Warmer temperatures typically accelerate flowering and maturity, and productivity reductions related to global warming partially reflect this response. Over 100 loci influence flowering time in Arabidopsis, and many of their functions are conserved in monocots such as rice and wheat<sup>35</sup>. A large body of molecular information exists for these gene interactions, and responses to environmental factors are being dissected at the genome level. Molecular and genotypic data have been integrated into ecophysiological models. An NSF-funded FIBR project (co-PI Welch) models flowering behavior of diverse Arabidopsis ecotypes in natural environments using simplified genetic networks. Major loci affecting response to photoperiod and vernalization are known for many crop species, but there is less certainty about loci for traits such as earliness per se. Comparisons with Arabidopsis suggest there are cases both of homology and of independent evolution of mechanisms (e.g., for vernalization in wheat). Gene-based prediction of phenology in crop species shows promise, but the models are still rudimentary (e.g., <sup>36</sup>; <sup>37</sup>; <sup>38</sup>).

### **Current state of resources and computational thinking in the field**

One approach to predicting phenotype from genotype is based on quantitative genetics<sup>39</sup>, which involves linear, algebraic, population-level theory that (1) deals with trait means and (co)variances and (2) tracks phenotypic changes across generations. The genetic factors can be resolved into quantitative trait loci, which can sometimes be identified at the gene level, such as tomato fruit size and shape (<sup>40</sup>;<sup>41</sup>), rice heading date<sup>42</sup> and yield<sup>42</sup>. Ecophysiological models, pioneered by de Wit<sup>43</sup>, mimic how plants respond to temperature, solar radiation, water and nutrient levels. Response components include: (i) phenology; (ii) dry matter production and partitioning; and (iii) growth processes. These models are nonlinear differential equations whose solutions are time series of variables such as plant biomass. Morphological realism can be improved via L-systems, sets of rules describing organ differentiation (algorithmicbotany.org/FSPM07) as influenced by time, genetic switches<sup>44</sup> and morphogenetic factors<sup>45</sup>. An L-system was used<sup>46</sup> to link Arabidopsis growth and development with genes.

A third approach for predicting phenotypes has emerged based on genetic, metabolic, and system-level networks<sup>47</sup>; <sup>48</sup>, <sup>49</sup>; <sup>50</sup>; <sup>51</sup>; <sup>52</sup>; <sup>53</sup>; <sup>54</sup>; <sup>55</sup>; <sup>56</sup>. Arabidopsis gene network models exist for stress and for metabolic processes<sup>57,58,19</sup> (Lee and Marcotte, in prep.) and are being extended to other plants (Krishnan & Pereira, in prep.; Ronald and Marcotte, in prep.). Metabolic systems models can provide the nuclei of genomic level hypotheses: their outputs are time series of metabolite concentrations, which can be directly compared to measured data<sup>29</sup>. The inputs to kinetic models are enzyme activities or concentrations obtainable from proteomics data. These network models provide a framework to append functionalities from crop interaction models. However, existing cyberinfrastructure poorly supports comparing and integrating of large datasets across species and genotypes. The VirtualPlant software system is a notable exception (www.virtualplant.org) (Katari, Shasha, Coruzzi and Gutierrez, in prep.). We propose to further develop this capability to facilitate the discovery of stress responsive mechanisms that may, or may not, be specific to given genera, species, or genotypes.

### **Identification and accessibility of available datasets**

Genomic data are well-curated through databases such as TAIR (www.arabidopsis.org), which includes the complete Arabidopsis genome sequence along with data on gene structure, gene products, known metabolic pathways, gene expression (provided by Geneinvestigator), phenology of insertion mutants (in some cases), DNA and seed stocks, genome maps, and genetic and physical markers. However, TAIR has no repository of metabolomic data, nor are there records of physiological data that, in many cases, accompanied the expression and phenological data. Other plant genomic databases include

GrainGenes ([wheat.pw.usda.gov/GG2](http://wheat.pw.usda.gov/GG2)), Gramene ([pathway.gramene.org](http://pathway.gramene.org)), TIGR Rice Genome Annotation ([www.tigr.org/tdb/e2k1/osa1](http://www.tigr.org/tdb/e2k1/osa1)), Soybase ([www.soybase.org](http://www.soybase.org)), PlantGDB ([www.plantgdb.org](http://www.plantgdb.org)), SGN ([sgn.cornell.edu](http://sgn.cornell.edu)) for Solanales including tomato, potato, and peppers, and several Medicago databases. SGN also provides SolCyc, and Gramene provides RiceCyc, which are biochemical pathway databases that have “Omics” Viewers allowing users to upload high throughput data and paint them onto a metabolic map. However, aside from Arabidopsis and rice, there is no public repository of high throughput gene expression data even for genera for whom genomic data are curated, nor is there one for metabolomics data for any species. A plant lipidomics database will come online during the iPlant project, adding another data type (Welti and Roth, pers. comm.). Further examples are the Arabidopsis Lipid Gene Database ([lipids.plantbiology.msu.edu/](http://lipids.plantbiology.msu.edu/)), the rice kinase database ([rkd.ucdavis.edu/](http://rkd.ucdavis.edu/)), the rice glycosyltransferase database (Cao and Ronald, submitted), and KEGG ([www.genome.ad.jp/kegg](http://www.genome.ad.jp/kegg)), which contains rice and Arabidopsis sites but lacks genes/enzymes, data that are needed, especially for rice. The above databases have rudimentary quantitative data on phenotypes, environments, and plant management, having emphasized qualitative or semi-quantitative traits. Yet such data are essential to fully specify the dependent variable, “P”, in the GEMP problem. The International Crop Information System (ICIS) is the largest data repository for crop species and contains pedigree, phenotypic, genetic, management, and environmental data. The rice version, IRIS, contains over 2·10<sup>6</sup> germplasm identifiers, representing landraces, crosses, breeding populations, selections, or lines evaluated in trials or nurseries. SGN contains about 7000 accessions and 15,000 images, links genomes to phenomes, currently mainly tomato, based on open plant and gene ontologies. Data types include images, literature associations, and free text descriptions. SGN will soon support data reanalysis via an in silico QTL functionality. The Generation Challenge Program (GCP; [www.generationcp.org](http://www.generationcp.org)), an international consortium working on drought in major crops, also generates publicly available phenotypic, genotypic, and genomic data relevant to this proposal.

### **Computational models and cyberinfrastructure support for the GEMP problem is limited.**

Current approaches for reverse-engineering molecular interactions and modeling GEMP relationships<sup>59</sup> are limited; they (a) only mine potentially interesting data patterns in data, needing extensive manual review to translate results into hypotheses and experiments; (b) make simplifying assumptions to account for under-determined systems; (c) require detailed and large-scale knowledge about molecular interaction networks, information that is unavailable even for *Arabidopsis*; and (d) struggle with increasingly intractable data fusion issues for analytic software arising from ever newer data types. The cyberinfrastructure for our Grand Challenge must overcome such limitations by being transparent (equations and assumptions not buried in code), breaking down scaling barriers, allowing high throughput computations, integrating diverse data sets, and facilitating model and algorithm specification and testing. The modeling and support systems should explicitly tackle four issues: (a) uncertainty representation at several biological interaction network levels when multiple models are consistent with experimental data; (b) needs analyses to identify levels of model uncertainty compatible with specific applications; (c) experiment suggestion so as to best reduce localized model uncertainties in a cost-effective way; and (d) model refinement by incorporating new experimental results.

### **Participation of both plant and computational researchers**

We seek to address the GEMP problem by interdisciplinary research and modeling using the three outlined quantitative approaches. As the proposed participant list shows, collaboration already exists between plant and computational researchers and will expand in the project. Genomics specialists will work with experts on systems biology, algorithms, distributed computing, and data mining to make sequence, gene expression, and metabolomic data available through (actually or virtually) unified, analytic tool, model, and user interfaces. Bioinformaticists will participate in efforts involving stochastic models, extracting models from data, and assessing model, data, and parameter uncertainty. Genetic

regulatory networks associated with climate change factor responses, both those common across species, and those unique to species, or genotypes within species, identified by these computational tools can link to metabolic models. Joint work with ecologists, plant breeders, agronomists, and crop modelers will bridge to field scales. The latter disciplines will provide use cases, field expertise, and portions of required data, while using data and methods from other disciplines to produce genetically-aware ecophysiological models.

### **Education, Outreach and Training Opportunities**

The simulation models that will be developed in parallel with our Grand Challenge efforts can be built upon to address numerous issues in plant biology, genetics, evolution, and ecology. Therefore, education, outreach and training opportunities on our Grand Challenge topic, effects of global climate change abiotic stress factors on plant productivity and the GEMP problem, will need to reach a broad-range of disciplines and age groups. The common modeling framework to be developed would provide students from all age groups an in silico laboratory, allowing them to conduct virtual experiments in diverse environments, for example improving understanding of how water availability and greenhouse gasses affect productivity, and how genetic changes may affect these relations. As students gain experience and knowledge, they can progress to models of increasing complexity without having to learn new software interfaces or data formats.

Existing programs that can be leveraged for potential outreach activities for our Grand Challenge education and outreach efforts will be identified and selected for proposal development. A NSF-funded art-science K12 outreach program (developed by UC Davis) introduces the concept of protein networks to elementary school students and can be built upon to promote the tools and science of our Grand Challenge. This NSF project reaches approximately 60,000 visitors at a Picnic Day at UC Davis and visitors to the New York Museum of Modern Art. Options directed toward minorities and under represented groups in Science and Engineering will be highlighted during the workshop and included for proposal development. The Scientific Thinking and Educational Partnership (STEP) at the University of Florida partners with faculty and their graduate students to tailor multi-media outreach programs designed to effectively communicate their research to target audiences. These can be utilized for our Grand Challenge.

A primary test-bed for our iPlant effort will be crop improvement. Outputs relating to prediction of responses of phenology, plant architecture, and molecular and metabolic events related to climate change can readily be incorporated into plant breeding. Phenology also has direct application to decision support using crop models, assessment of climate change impacts. We thus anticipate that the project will provide test cases affecting decisions of plant breeders, agronomists or farmers within four years. Since the project involves plant breeding and climate change groups, initial transfer should be direct. We would conduct additional workshops for interested breeders or other members of the plant science community, such as workshops prior to the annual meetings of the Crop Science Society of America and the American Society of Plant Biologists.

### **Goals for the Grand Challenge Workshop.**

The immediate goal of this workshop is to reach an initial consensus on the feasibility of different approaches for data and model integration across organizational levels, with emphasis on lessons from existing tools such as Virtual Plant, the Regulatory Network of Marcotte and Ronald, and ICIS. We expect to identify key limitations in cyberinfrastructure that limit current modeling approaches and to understand better how different quantitative modeling approaches interrelate (Appendix 1). The workshop will also allow further prioritization among traits and refinement of the strategy for outreach. We expect that the interchanges will allow us to produce a draft for the Grand Challenge Proposal by meeting's end.