Intellectual Merits. Signal transduction pathways hold the key to understanding the early response of higher plants to abiotic stresses, such as drought, flooding, heat, cold, ozone, and salt. Traditionally, signaling pathways have been difficult to synthesize and impossible to manipulate computationally. Current pathway tools and databases provide only limited support for manipulating pathways as networks. In particular, it is not possible for a biologist to take an existing signaling pathway and ask hypothetical questions that are answered computationally. This project builds the Beacon system to provide computational support for biologists’ questions about signaling pathways. The user interface is an editing environment that represents and manipulates a pathway in a standard graphical notation called SBGN Activity Flow language. The project includes the curation of signaling pathways for plant abiotic stress response by recognized authorities and their archiving in a Beacon database. The database representation allows the biologist to impose a semantics, such as boolean semantics, on a pathway. The Beacon simulation engine is then able to computationally implement those semantics and provide results that can be queried and visualized. Adding or removing relationships in a pathway (adding or removing edges in a network) raises the difficult question of causal inference: Does activity of molecule A cause activity of molecule B? Such questions can be assessed computationally and statistically using such tools as Granger causality, Bayesian networks, and structural equation modeling. The Beacon inference engine provides these tools to allow the biologist to develop testable hypotheses about pathway components. The Beacon system is not specific to plant signaling pathways but rather enjoys more general usability.

Broader Impacts. The project will include training and education through well-established outreach programs at Virginia Tech. These include the Multicultural Academic Opportunities Program (MAOP), which provides research opportunities for students from pre-college through the doctorate; S-STEM, which enhances the competitiveness of undergraduates using research experiences; and PREP and PREP-U, which use NSF funding to expose high school and undergraduate students to biological research that contrasts phenotypes of Arabidopsis mutant and wild-type plants.
PROJECT DESCRIPTION

1 Motivation

A signal transduction pathway is a collection of interacting cellular components that perform the response of the cell to an external or developmental signal. A pathway is most conveniently represented by a network with the components as nodes and the interactions as directed edges. This representation assists both computational manipulation and user visualization. Signal transduction pathways are keys to understanding the dynamic logic of, for example, plant responses to stress.

Responses and accommodations to adverse changes in the environment, such as drought, salinity, cold, and oxygen availability, are at the heart of many plant activities. Climate change events are expected to exacerbate the severity and duration of current adverse environmental conditions. Elucidation of the genetic response networks regulating plant dynamic responses to changing environments is daily becoming more of a reality [24, 59, 87, 88, 99, 100, 116, 123]. Bioinformatics approaches are increasingly available to address these questions. However, the full details of even one stress signaling pathway are still not clear. Furthermore, although cross-signaling is clearly an important part of adaptive responses [59, 75, 86, 184, 185], it is unclear to what extent recognition and response pathways for the various abiotic stresses overlap in any one case.

We propose to expedite the consolidation of existing understanding of stress signaling in plants, facilitate its expansion, and broaden participation in knowledge building across the plant biology community. To this end, we propose to facilitate a systems biology depiction of interlocking molecular signaling pathways in plants that respond to abiotic stresses, using extensible curated pathways drawn by authorities in a common graphical language as our starting point.

Deep sequencing of non-model species, of several ecotypes of the model plant Arabidopsis, and of specific crop genotypes, is rapidly increasing our understanding of plant genomic responses to abiotic stress [26, 38, 97, 151, 166, 189, 191]. Information is available about the nature of cell-specific signaling in some cases [9, 34, 119]. Regulatory events at multiple, post-transcriptional, organizational levels are integral components of stress responses [18, 63, 117, 188]. Potentially of equal or greater importance for plants is the recently discovered role of small RNAs in epigenetic events that modify gene expression in response to abiotic stress [104, 155].

Challenges Identified. Existing cyberinfrastructure does not support a systems view of signaling pathways in plants. Current tools for signaling pathways are drawing tools, some of which only result in images [50, 133], while others store true networks [73, 74]. The work of creating a signaling pathway is painstaking, requiring a thorough understanding of the literature and a careful presentation to convey all the concepts present in the literature. The result of such work is typically a figure in an article and perhaps an entry in a signaling pathway database. The result is only suited for limited computational manipulation, because the semantics of the signaling pathway remain in human-readable form only. This limitation has a number of negative consequences. Biologists are not able to take an arbitrary signaling pathway and paint it with omics data from experiments to gain insight into the operation of the pathway under experimental conditions. Generally applicable simulation facilities for signaling pathways are limited, so that “what-if” (e.g., knockout or knockdown) experiments are rarely performed in silico. Inferring new causal links in an existing pathway remains a difficult task; there is limited computational support for suggesting possible new causal links. Finally, there are no widely accepted standards for the syntax and semantics of signaling pathways.
Figure 1: Prototype discovery environment based on iPlant Collaborative Web Services API. A user uploads results of a multi-experiment microarray analysis and invokes VAPrototype, an HTML/Ajax-based web application running at the iPlant Collaborative, which performs GO enrichment analysis for each experiment. Results are returned as a JSON-formatted data structure and rendered by VAPrototype (top panel) in a table color-coded by significance. The user can mouse over a table cell to see the list of genes associated with the term. Selecting a cell submits the cell’s gene list to a web service implementing the GeneMANIA functional network inference tool [113, 171]. Results (lower left) are rendered using Cytoscape Web such that seed genes are yellow circles and inferred genes are magenta circles. The user can mouse over edges and nodes to see a full list of attributes. Selecting a node invokes the eFP browser [179], which shows (lower right) the gross anatomical expression profile for the gene of interest.

Proposed Solution. We propose Beacon, a new systems biology tool, to address these challenges. Within Beacon, an expert curator can construct and edit signaling pathways, allowing the integration of current and future data over multiple scales of cellular organization and across species, with a central focus, in this first phase, on abiotic stress response pathways. The proposed consolidation of expert knowledge across plant stress biology will eliminate wasted effort, unite a section of the plant biology community, and afford the possibility of fresh insights into fundamental biological mechanisms. Equally important, Beacon will serve as a valuable teaching tool available to students of plant biology. Users of Beacon will be able to upload inferred causal relationships among entities, based on their own data. They will be able to store the resulting amended pathway, and share it. Beacon will interact with the cyberinfrastructure of the iPlant Collaborative (see letter from Dr. D. Stanzione, Co-PI of the iPlant Collaborative). Beacon will use an omics data integration workflow instituted in iPlant, in the Data Analysis and Visualization Working Group (of which Grene is the co-leader). A prototype of this workflow exists (see Figure 1). Beacon will use the iPlant web services API, which will provide tools shown in the prototype and also access to data converters, storage, and computational resources, and the ability to invoke other visualization and analysis tools through a consistent web service interface. The iPlant team will provide staff resources to support the Beacon developers in the use of the API, both through training and direct technical support.
As a computational system, Beacon will be able to represent, visualize, and model signaling pathways. While Beacon will be general enough to support a large class of signaling pathways (see Section 4.2), we will focus on abiotic stress signaling pathways in higher plants. A Beacon system diagram is in Figure 2. The central component is the Beacon user interface, which supports editing and visualizing pathways in the Systems Biology Graphical Notation Activity Flow language (see Section 3). The user interface also controls the movement of data, model simulation, and inference. In particular, Beacon will import and export pathways in several standard formats, including BioPax and SBML and will import omics data and inferred pathways from tools available from the iPlant Collaborative and other sources (see Section 4.2). Beacon will allow the user to select semantics for the components in a pathway and to simulate “what if” scenarios implementing those semantics (see Section 4.3). Beacon will support the inference of new pathway components from imported omics data (see Section 4.3). Omics data and simulation results can be visualized and manipulated with the incorporation of the ViVA interactive visual tool into Beacon (see Section 4.2).

Sample Questions to be Addressed Using Beacon. What is understood about specific events in abiotic stress signaling in model plants, crop species, and specific genotypes? What similarities or differences exist among responses to different abiotic stresses? What are the unique features of the responses of specific cell types or tissues? In each case, what inferences can we make about the causal relationships that drive cellular responses to abiotic stress and their impact on phenotype?

The Availability of Interdisciplinary Scientists. Few plant biologists possess the skills to develop a system such as Beacon. However, the combination of expertise in the Beacon team provides the requisite skill set for the project. The Genetics, Bioinformatics and Computational Biology Program at Virginia Tech provides the requisite training for the next generation (see Section 7). Our proposed work will expand this training to other participating institutions through the virtual community created by the authorities and the tools described below.

2 Objectives

To attain the Beacon system and database, we have the following specific objectives.

- **Objective 1**: Provide a curation tool for the SBGN Activity Flow language to a set of authorities to create an initial collection of pathways. Provide curation tool education in a workshop.

- **Objective 2**: Expand the curation tool to support data import and export and to maintain data in the Beacon database. Make the database available on the Web. Incorporate the ViVA visualization tool.

- **Objective 3**: Add the simulation and inference engines to the Beacon system.
<table>
<thead>
<tr>
<th>Authority</th>
<th>Affiliation</th>
<th>Stress</th>
</tr>
</thead>
<tbody>
<tr>
<td>Julia Bailey-Serres</td>
<td>University of California, Riverside</td>
<td>Flooding</td>
</tr>
<tr>
<td>Mikael Brosche, Jaakko Kangasjavri</td>
<td>University of Helsinki</td>
<td>Ozone</td>
</tr>
<tr>
<td>Tzyy-Jen Chiou</td>
<td>Academia Sinica, Nankang, Taipei</td>
<td>Phosphate, small RNAs</td>
</tr>
<tr>
<td>Viswanathan Chinnusamy</td>
<td>Indian Agricultural Research Institute</td>
<td>Cold</td>
</tr>
<tr>
<td>Alejandro A. Covarrubias, Jose Reyes</td>
<td>Instituto de Biotecnologia-UNAM, Mexico</td>
<td>Small RNAs and stress</td>
</tr>
<tr>
<td>John C. Cushman, Mel Oliver</td>
<td>USDA/ARS</td>
<td>Root responses to drought stress</td>
</tr>
<tr>
<td>Sixue Chen, Radhika Desikan</td>
<td>University of Florida, Imperial College, London</td>
<td>Guard cells and stress</td>
</tr>
<tr>
<td>Ruth Finkelstein</td>
<td>University of California, Santa Barbara</td>
<td>Abscisic acid</td>
</tr>
<tr>
<td>Jose M. Pardo, Björn Usadel</td>
<td>Instituto de Recursos Naturales y Agrobiologia, Spain</td>
<td>Salt</td>
</tr>
<tr>
<td>Olena Vatamaniuk</td>
<td>Cornell University</td>
<td>Heavy metals</td>
</tr>
</tbody>
</table>

Table 1: Plant stress authorities who have agreed to provide curated pathways. In addition, Grene and Pereira will provide generalized drought stress pathways.

- **Objective 4**: Validate Beacon simulation predictions in collaboration with PREP.

## 3 Background

### 3.1 Abiotic Stress Signaling

Much data already exist, pointing to genotype-specific variation in the responses of herbaceous plant and tree species to abiotic stress [8, 21, 23, 55, 85, 102, 114, 138, 164, 170, 176, 177, 181]. Valuable insights have been obtained through these studies identifying genes or QTLs that indicate possible molecular pathways explaining intra-specific differential responses leading to stress resistance. To date, it appears that genotype-based variation in stress responses may well offer clues as to bases for differential resistance to stress. “Pre-adapted” states are also important for superior stress tolerance [35, 61, 92, 93, 102]. Nonetheless, such variation, to date, is nearly exclusively based on fitting genes into known, putative, and, in some instances, even conjectured, “resistance” pathways.

We propose to first lay out generalized images of molecular stress response pathways in plants that represent current, conventionally held, views. (See below for some examples.) Expert authorities will provide specific information in visual form during the first phase of the implementation of Beacon. Pathway graphs have been promised (see Table 1), and the responses of the labs contacted have been very positive. As a second phase, the interlocking pathways will be further elaborated to show variations in responses and/or constitutive states that are documented for individual genotypes and/or crop cultivars, together with any phenotypic observations that are available.

**Abscisic Acid (ABA) and Abiotic Stress.** ABA is a major regulator of plant responses to abiotic stresses such as drought, salinity, and cold. These responses are mediated by changes in gene expression as well as rapid effects on ion fluxes, such as those involved in stomatal closure. Extensive genetic and biochemical studies have implicated well over 100 loci in the ABA response and have demonstrated both direct and indirect interactions among their products. These loci comprise numerous families with both redundant and distinct functions at every level of the ABA signaling network: multiple receptor families, protein phosphatases (PP2Cs) that interact with the
PYR/PYL/RCAR receptors, kinases (e.g. SnRKs) regulated directly by the PP2Cs, and transcription factors, ion channels, and numerous other substrates of the SnRK and other kinase families (reviewed in [31, 137]). Although some signaling components participate in multiple ABA responses, others are specific to subsets of responses, probably through a combination of limited expression and preferential interactions. Currently, the best characterized ABA signaling network is that controlling guard cell function in stomatal regulation, which has been developed as an example of a systems biology approach to signaling (reviewed in [9]). While many aspects of this single-cell system may function in other ABA responses, the diversity of cellular responses to this one signal indicates that it is unlikely that there is a universal ABA response module. There is a wealth of information that could be assembled to display interactions among specific members of the families described above, regulation by secondary messengers, and their collective effects on gene expression and cellular metabolism in seeds and seedlings, as well as in guard cells. Over half of ABA-induced changes in gene expression are induced by drought and salinity [144].

The Central Role of Reactive Oxygen Species (ROS) Signaling in Responses to Abiotic Stress. Organisms can sense changes in their environment through the employment of controlled ROS-generating and reporting systems [109, 121]. ROS-mediated response pathways constitute a major component of plant cell responses to biotic and abiotic stresses, allowing adjustment of functional processes to altered conditions [7, 89]. For example, significant flexibility and precise control over photosynthesis-related processes is provided by metabolism driven by redox processes [19, 132]. Mediating environmental awareness are ROS generators, transmembrane facilitators, signal amplifiers, calcium-dependent elements, and genes whose activity is controlled by the cellular redox state, resulting in altered hormone biosynthesis and metabolic transport processes as crucial downstream components [66]. ROS responses, which are condition-(i.e. stress)-specific, are suggested by gene clusters that imply the existence of networks. Distinct signaling cascades target regulatory factors that shuttle between cytosol and nucleus [66]. Redox signaling is initiated through ligands, enzymes, or changes in membrane potential at, for example, the plasma membrane that are perceived and internalized by protein tyrosine phosphatases (PTPs), heteromeric proteins, and transmitted through MAPK cascades [89, 108].

An Example of ROS-Mediated Signaling — Ozone Stress. Plant exposure to elevated atmospheric ozone at acute levels can lead to cell death. Ozone enters the leaves through open stomata and immediately degrades to other ROS in the apoplast, including superoxide and hydrogen peroxide. This ROS production initiates a signaling cascade that involves apoplastic GR1 (GRIME REAPER 1) and an, as yet unidentified, receptor(s) that transfer the signal to the inside of the cell [182]. As a consequence of exposure to ozone, a downstream interaction with plant hormone signaling pathways occurs, which have a crucial role — salicylic acid and ethylene promote cell death, whereas jasmonic acid is proposed to contain cell death [125]. Based on chemical treatments, phosphorylation and Ca2+ are involved in this signaling pathway, although the exact proteins are not identified yet. The rcd1 (RADICAL-INDUCED CELL DEATH) mutant is ozone sensitive, showing higher rates of cell death than wild type Arabidopsis plants, indicating that RCD1 regulates some aspects of cell death, possibly related to its interaction with several transcription factors [65]. ROS scavenging through antioxidant defenses are important to control the level of cell death; low ascorbic acid mutants (vtc) are ozone sensitive [29]. Ozone entry to the apoplast leads to ROS production and activation of as yet unidentified receptor(s) that transfer the signal to the inside of the cell. The Ca2+ transporter DND1 (DEFENSE NO DEATH 1) mediates an essential signaling step, since ozone induced changes in gene expression are absent in dnd1 [183]. Inside the cell, one
of the earliest events detected in response to ozone is activation of MAPK signaling (MPK3 and MPK6) and transport of MPK3 and MPK6 from the cytosol to the nucleus [1]. The plant hormones salicylic acid and ethylene act as positive regulators, and NPR1 acts as a negative regulator.

**ROS, ABA, and Secondary Messengers: Water Deficit and Flooding-Induced Responses.** The frequency and severity of droughts and flooding are expected to increase in many regions of the world, while decreased runoff can be expected to impair reservoir and ground water recharge. Inter- and intra-specific gene expression studies have identified multiple pathways that are regulated in response to water stress [80, 110, 164, 173, 174], and flooding [117, 118]. In the case of severe flooding and submergence, plants are faced with an energy crisis that requires effective management of energy reserves. Submergence tolerant rice adopts a quiescence strategy, limiting gibberellin-activated elongation growth and conserving carbohydrates [15, 42]. Study of low oxygen and submergence responses in Arabidopsis and wild species (e.g., Rumex) indicate that a quiescence strategy can also outweigh the benefit of rapid growth to escape submergence. In general, in the case of drought, osmotic changes, “recorded” at the cell membrane, are transduced in the form of rapid metabolic changes involving generation of ROS, ABA biosynthesis, and long distance transport from its site of synthesis to cellular sites of action via transporter-mediated processes, where it binds to specific receptors, phospholipid metabolism, which includes the generation of secondary messengers such as phosphatidic acid, calcium sensing and the action of calcium on signaling pathway components, antioxidant signaling and defense pathways, and the synthesis of protective molecules [16, 45, 91, 137, 141, 186]. It is increasingly clear that the ubiquitination machinery also plays a crucial role in the modulation of drought signaling pathways [27, 134, 192].

**Role of Small RNAs in Stress Signaling.** The many different types of small non-coding RNAs, their genes, expression characteristics, and influence on the coding part of the transcriptome are already beginning to revolutionize our views of how phenotypes are determined [120, 193]. Small RNAs have also been implicated in the regulation of proline biosynthesis under salt stress [17]. Regulatory effects of specific microRNAs (miRNAs) exerted by modulating transcript amount or turnover of those genes that are their targets have been shown for drought, salt, nutrient (phosphate, nitrogen, and sulfur), and temperature stress in a range of plant species [25, 76, 126, 156, 157]. In addition to transcriptional control, small RNAs may also exert inhibitory effects on translation [20]. At this time, various stress-responsive miRNAs have been identified in various plant species, and putative targets have been assigned for many of them; however, these data remain to be verified (reviewed in [30]). The identity of the validated targets indicate that miRNAs are implicated in the regulation of transcripts encoding a wide variety of proteins, e.g., transcriptional factors (involved in cold, drought, salinity and ABA responses), antioxidant enzymes, ion transporters, F-box proteins, AGO1, hormone metabolism and signaling, among others (reviewed in [30, 67]).

### 3.2 Computational Background

**Signaling Pathway Databases.** PathGuide [14] lists 62 signaling databases, most of which support rendering pathway drawings (visualization) and many of which store pathways as networks, allowing the export of pathways in standard formats such as SBML and BioPax. We review some key signaling pathway databases here. The Arabidopsis Reactome [159] is a general plant systems biology resource that is built on a generalized notion of “reaction”. KEGG [74] is a general resource for pathway information that emphasizes metabolic pathways. NetPath [73] is a database of curated human signaling pathways stored as networks and with reactions annotated by the published literature; it is supported by the PathBuilder system [72]. Science’s Signal Transduction Knowledge
Environment (STKE; [49, 50]) is a small database of curated pathways, supported by the literature and available only as images. Reactome [106] is a general resource for human and other pathways that includes analysis of user-supplied omics data to identify events or pathways that are overrepresented in the data. TRANSPATH [28] is a signaling pathways database augmented with Array-Analyzer software to interpret microarray (up- and down-regulation) data to extract subnetworks from the genes or proteins identified as significant in the microarray. The DMSP database [169] takes protein-protein interaction information as the basis for mathematical modeling and simulation of signaling pathways using the traditional ordinary differential equation (ODE) style of modeling. SPIKE [36] is a signaling pathway database combined with an algorithmic engine that does basic graph-theoretic analysis of paths and connections as well as identifying pathways that are enriched with respect to a set of user-supplied genes. One signal transduction database of popular interest is Wikipathways [77, 133], which follows the Wiki paradigm of collaborative editing and hence includes a visual editing tool PathVisio [162]. The system does not support any standard graphical language and only represents a pathway as a drawing, not as a network. These systems are generally limited by a lack of a means to infer new pathway components and by pathway semantics that are restricted to the ODE view. Beacon will overcome both of these limitations (see Section 4.3).

**Systems Biology Graphical Notation (SBGN).** SBGN [122] is a standard visual notation for systems biology diagrams, consisting of three languages: Process Description (PD), Entity Relationship (ER), and Activity Flow (AF). PD supports temporal changes in the activity of metabolic pathways [64]. AF depicts the information flow in a biological network and is most suited to representing signaling pathways, where intermediary metabolism plays a lesser role. A node in an AF diagram is either an activity node or a container node (for encapsulation; using SBGN, Beacon will be able to support cellular compartments). Most activity nodes represent the activity of some biological entity, with the remainder representing either a perturbation or a phenotype (broadly interpreted, ranging from apoptosis to life span). An arc (edge) typically represents either a positive influence, with a →, or a negative influence, with a ⊣. AF also supports boolean logic (AND, OR, and NOT). See Figure 3 for an example; it is a rendering of the ABA signaling pathway in the nucleus of guard cells in Figure 2(b) of Raghavendra et al. [137]. It was drawn with SBGN-ED (see below). The polygon at the top labeled “ABA” is a perturbation node, representing a change in ABA levels. Rectangles represent the activities of biological entities; for example, rectangle “ABI5” represents the DNA binding of transcription factor ABI5, while rectangle “transcription” with the decoration “ABRE” represents transcription regulated by an ABA responsive element. This AF diagram is a typical rendering of signaling pathways found in the literature, where biological information flow, either positive or negative, is represented. mEPN [40] is another proposed visual representation but does not enjoy the level of community support that SBGN does.

**Pathway Drawing Tools.** Most pathway drawing tools are especially for metabolic pathways, but a few have functionality specific to signaling pathways. Few are based on any graphical standard. We do not survey general network visualization tools, such as Cytoscape; they are not well-suited to drawing signaling pathways and annotating their semantics, because they take in large networks for visualization and rely on automatic layout algorithms. PathText [78] supports construction of pathways by automatic text mining of the literature; it uses the Payao system and CellDesigner to visualize pathways. Payao [105] is a collaborative web service platform for curating SBML models. CellDesigner [43] provides the drawing environment for Payao and uses the SBGN Process Description language, making it most suitable for metabolic pathways; CellDesigner is not open source. Arcadia [168] translates SBML into SBGN Process Description language and is specifically...
Figure 3: SBGN Activity Flow language. ABA is made under drought conditions. This ABA increase relieves inhibition of OST1, which activates ABI5. Increases in ABA also result in an increase in calcium, culminating in activation of transcription factors (TFs), which bind to ABRE with ensuing gene activation. CPK4 is a calcium-dependent protein kinase. Adapted from [137].
4 Proposed Research

4.1 Objective 1: Curation

Provide a curation tool for the SBGN Activity Flow language to a set of authorities to create an initial collection of pathways. Provide curation tool education in a workshop.

We will have authorities on individual signaling pathways (see Table 1) curate pathways in higher plants. To facilitate curation, initial development will concentrate on the editing capabilities of the user interface. The result will be a curation tool that will allow authorities to enter pathways in SBGN Activity Flow language, edit them, annotate them, and save them to an initial Beacon database. The user interface will be built on the network visualization tool VANTED [69] and SBGN-ED, an extension of VANTED. We will extend SBGN-ED to create our curation tool. Usability of the tool will be addressed via a continual feedback loop between users and developers.

A 2-day training workshop for two members of each authority’s group will be held at Virginia Tech early in Year 2. Attendance at the workshop will be capped at 20 visitors. Co-PIs Grene and Heath and their students will serve as instructors. One session will be devoted to interactive discussions and informal presentations by each visitor to establish a common focus within the group. Other sessions will be hands-on, in groups of two, with a final session dedicated to presentations in Beacon. Students and post-doctoral fellows will give the presentations. For groups not attending the workshop, we will have subsequent online training sessions on the curation tool.

4.2 Objective 2: Initial Beacon System

Expand the curation tool to support data import and export and to maintain data in the Beacon database. Make the database available on the Web. Incorporate the VIVA visualization tool.

Beacon Pathway Representation. Pathway representation is central to the power and flexibility of Beacon. SBGN activity nodes, as used in Beacon, can be molecules, perturbations (e.g., “moisture level” or “ABA”), phenotypes (e.g., “apoptosis” or “stomatal closure”), or unknown mechanisms. The last three types of nodes are found in ad hoc signaling pathways from the literature, such as the one in the Assmann review [9], but they are not available in STKE, SBML, or BioPax. Standard edge types, such as “positive influence” and “negative influence” in SBGN Activity Flow language, are typically used when Boolean semantics (i.e., nodes are either switched “on” or switched “off”) are used to interpret (model) the pathway. Beacon will fully support Boolean semantics, including arbitrary logical functions for nodes and propagation delays on edges. Beacon will also support more general variable values at nodes, including more than two discrete levels and continuous values (see Section 4.3). Hence, the Beacon pathway representation will include semantic information, such as variables and transfer functions, for each node. This semantic information will constitute a semantic model for the pathway, which can be simulated using Beacon (see Section 4.3). Annotations for nodes, such as cellular location, will be provided in the database and the visualizations. Note that pathways may overlap (have common components). Beacon will support links between overlapping pathways and merging them into more comprehensive pathways.

Beacon Data Import and Export. Beacon can import and export pathways in the standard formats SBML and BioPax. Many pathway databases support export in one of these formats, so it will be feasible to import pathways into Beacon from those databases. Note that BioPax is an OWL-based ontology for metabolic and signaling pathways. Beacon will import and export BioPax signaling pathways only. Beacon has the capability to import omics data to overlay on the pathway visualization and to use in inferring possible new relations or annotations for the pathway (see Section 4.3).
**Beacon Database.** The signaling pathway database implements the Beacon pathway representation (see above) in a relational database. The database supports versioning, so that multiple versions of a particular pathway can be present and accessible at any time. For simplicity, there will always be a default (most current) pathway. (In the case of controversy, there could be two or more current pathways.) There can, in fact, be multiple Beacon signaling pathway databases, either local or remote; a simple addressing scheme will be used to select among non-default databases. Any Beacon database can host multiple versions of a single pathway; this is done to facilitate hypothetical exploration of different elaborations of one version of the pathway. For example, the same simulations could be run on multiple versions to determine which version best fits a particular data set.

**ViVA.** As ViVA (see Section 3.2) is written in Java and is available as open source, incorporating it into the Beacon system will be straightforward. ViVA has two uses in Beacon. First, it is an alternate way to visualize imported omics data (see above). Second, simulations will generate large quantities of data (thousands of runs) that can be readily visualized and analyzed in ViVA.

### 4.3 Objective 3: Simulation and Inference Engines

*Add the simulation and inference engines to the Beacon system.*

**Simulation Engine.** The simulation engine design will be based on semantic network models that have been used to provide meaning for signaling pathways [2, 12, 79, 152, 158], including Boolean networks [5, 9, 10, 62, 94, 112], multilevel logic [139], and differential equations [44, 180]. The various models have two important subtypes, deterministic and stochastic.

The simulation engine will support Boolean semantics, including arbitrary logical functions for nodes and propagation delays on edges, as the default pathway semantics. The Boolean simulation functionality, which will be implemented first, will be an extension of that in [4, 94], supporting synchronous, asynchronous, and user-supplied timing models for the edges and user-controlled logic and noise models for nodes. We will generalize the Boolean model to allow multiple discrete levels at nodes, which could be useful in a drought stress setting, for example, where there may be five moisture levels (perturbations) of interest. Effects such as a gene knock-out can be simulated by setting a variable at the corresponding node to a fixed constant (such as “off”). In both Boolean and multilevel models, stochastic effects can be introduced by having some variables set randomly; a simulation is then run multiple times to obtain statistically significant results.

While not usually suited to signaling pathways, continuous semantics will be supported by Beacon through its ability to exchange pathways as SBML models. In particular, we will access COPASI [60] through the iPlant API to provide both ODE and stochastic simulation in continuous semantics. Many of the situations (e.g., abiotic stress) in which pathway knowledge will be applied involve time-varying environments. Yet a major shortcoming of many systems biology modeling packages (e.g., SBML) is the inability to link inputs to specific sources of real environmental time-series data. Beacon will build upon and extend features currently existing in COPASI parameter estimation procedures to address this issue (see letter from Dr. Steve Welch). Results of simulations can be output in tabular form or visualized in the user interface as plots in ViVA.

**Inference Engine.** Network inference is the problem of reconstructing a biological network from biological data, also called reverse engineering biological networks in the DREAM (Dialogue on Reverse-Engineering Assessment and Methods) project [154], which addresses network inference in metabolic pathways, gene regulatory networks, molecular interaction networks, and signaling pathways. Numerous methods have been used to infer network connections, including correlation [96, 115, 124], regression [82, 128], Bayesian approaches [41, 84], probabilistic graphical mod-
els [32, 187], information theory [172, 194], algebraic methods [153], logical methods [51, 62, 81], machine learning [71, 111], and evolutionary algorithms [148]. Most of these methods are not applicable to signaling networks due to the directed edges that connote causality.

Computational causal inference [46, 131] narrows the problem considerably, as each causal relationship (directed edge) is taken to mean that the antecedent is the result of the precedent through some, perhaps unknown, mechanism. Such causal relationships can be suggested by various types of omics data. We first discuss time-course data and then contrast data (e.g., wild type versus mutant) in which an experimental intervention (e.g., gene knockout) sets some cellular activity to always on or always off (for boolean semantics). Some prior work specifically utilizes time-course data in network inference [6, 32, 115, 145], with some of it especially for signaling networks [48, 68, 90]. Granger causality is a popular method for causal inference from time-course data [56, 115, 142, 195]. This method uses statistical correlation to infer whether one variable (e.g., expression of gene A) is causally related to another (e.g., expression of gene B). The Beacon inference engine will implement Granger causality, which is generally applicable when there are adequate time-series data [195]. Beacon will borrow from existing code for Granger causality (e.g., [142]).

For experimental data that shows a contrast between treatments, Pearl [130] demonstrates how to use directed graph-based models and suitable algorithms to infer the presence or absence of causal links. He covers Bayesian networks [37, 46, 187], which are based on directed acyclic graphs, and pathway analysis in structural equation modeling [83, 95, 143], which may have directed cycles and hence are more directly applicable in the context of signaling pathways. (Note that MetaReg [161] does causal inference in Bayesian networks and hence does not support directed cycles.) Application of these models to causal inference proceeds as follows: A small number of models (graphs) to test are proposed, and the implications of each model are evaluated statistically against empirical data [83]. In the case of structural equation modeling, the statistical analysis can be provided by OpenMx, open source software based on the R programming language; Beacon will incorporate OpenMx into the inference engine. The Beacon inference engine will return the models in a ranked list in order of probability given the data. In most cases from the literature, each model has to be constructed from scratch by integrating whatever omics and other data is available. In Beacon, we have the advantage of already having curated signaling pathways (see Section 4.1) and only needing to extend or edit a pathway with additional nodes and edges. Given general guidance from the user, Beacon can automatically generate all the models to be tested.

There is also the issue of removing edges that may be redundant, disproven, or represent an indirect rather than a direct causal influence. In the literature, this is addressed with the graph-theoretic algorithm of transitive reduction [3, 4]. In the Beacon inference engine, we will also implement transitive reduction, or use the existing NET-SYNTHESIS software [70], to suggest edges that may be redundant.

### 4.4 Objective 4: Experimental Validation

*Validate Beacon simulation predictions in collaboration with PREP.*

Predictive roles of genes involved in signaling pathways and the direction of signal transduction needs to be validated to be able to base further studies on integrating the pathways. As a proof of concept, phenotypic characterization of Arabidopsis mutations in selected genes predicted in Beacon simulations to play a crucial role in stress signaling will be carried out in Years 3 and 4, in collaboration with Dr. Erin Dolan of PREP and PREP-U (see Section 7). Drought and salt stress pathways will be experimentally dissected for early signaling and cross-talk. A number of useful
standardized phenotypic assays and conditions for assessing the relative tolerance of Arabidopsis have been described [167], and several of these assays (1–3 below) will be used in the proposed project. The effect of mutations in signaling genes will be assessed for perturbations in the abiotic stress pathways by comparing the effects of stress imposition on mutant and wild type genotypes on: (1) growth characteristics: leaf number, leaf area, biomass accumulation, seed yield, root length, and growth; (2) phenology: bolting and flowering time; (3) leaf water content, photosynthetic rate, and stomatal conductance; and (4) the expression of selected genes that are predicted by Beacon to be downstream of the mutant gene in each case, using quantitative RT-PCR.

**Drought Stress Screens.** Arabidopsis plants will be grown in Metro-Mix 350 in a controlled environment chamber (Conviron, 16 hour photoperiod, 22° day, 18° night). Water will be withheld until 50% RWC is reached in each case, after which time the plants will be re-watered and allowed to recover until the wild type genotype has reached pre-stress levels. (The extent of recovery from drought stress at the 50% RWC level in the mutants will be revealed in the course of the experiments.) The time of drought imposition will be ca. 20 days after sowing. Photosynthesis and leaf water content will be measured daily from the time of imposition of stress. One (maximum) stress and one recovery time point will be taken.

**Salinity Stress.** Plants will be grown as described above for drought stress exposure prior to the stress imposition step. Salinity stress will be initiated at the same developmental stage as drought stress. Leaf number will be measured at the time that the stress is initiated. At the start of the experiment, plants will be watered with 0 (control), or 150 mM NaCl in steps of 50mM separated in time to allow for adjustment by the plant [47].

**Phenology.** Bolting and flowering time will be recorded for each genotype and each treatment, as well as leaf number at bolting, and seed yield. (Note that these measurements can be represented as SBGN phenotype nodes.) Gas exchange will be measured with a Li-Cor photosynthesis meter, with Arabidopsis chamber, as is routine in the Grene laboratory [103, 102, 175]. In each case, statistical tests (e.g., ANOVA) will be applied to the data, and differences between genotypes at the 95% confidence level determined.

**Quantitative RT-PCR (qRT-PCR).** qRT-PCR will be carried out on selected samples by the method of Vandesompele et al. [163], which is routinely used in the Grene laboratory [102, 164, 165]. The control transcripts used for standardization will be UQ10 for salinity stress [47] and adenosine kinase for drought stress [164].

### 5 Beacon User Scenario

This scenario is fictional, although relevant to work to which our group contributed [92].

**Experiment.** Two genotypes of Arabidopsis thaliana are subjected to long-term, low level, exposure to the oxidizing air pollutant ozone. The genotypes differ only by a knockout mutation at a single locus G, a candidate regulatory gene, belonging to a large transcription factor (TF) family. The time course experiment has several samples taken over the expected normal life cycle of Arabidopsis for gene expression and phenotypic measurements. Since one phenotypic manifestation of ozone stress, at least at acute levels, is early senescence, “senescence status” is assessed through measurements of photosynthesis, chlorophyll levels, and the expression of well-characterized Senescence Associated Proteins (SAGs) and of the regulatory gene WRKY53 [11, 39, 107, 190].

**Question Posed.** What effect does knocking out gene G have on molecular defenses against long-term ozone stress, such as that experienced by field-grown plants under current and expected future climatic conditions?
<table>
<thead>
<tr>
<th>Objective</th>
<th>Year 1</th>
<th>Year 2</th>
<th>Year 3</th>
<th>Year 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>Objective 1: Curation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective 2: Initial Beacon system</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective 3: Simulation and inference engines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Objective 4: Validation</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2: Activity schedule for Beacon objectives

**Data Mining.** The results of applying a GO enrichment algorithm to the transcriptomic results (see Figure 1) suggest that Pathways A and B did not respond in treated mutant genotype plants, and that putative Regulator X was expressed to a lesser extent, and at a later time point, than in ozone-exposed wild type plants. (Expression of X did not differ between the two genotypes in the absence of ozone stress.) The phenotypic data suggest an early onset of senescence in the absence of a functioning Gene G in the mutant genotype, and a partial, gradual, acclimation to ozone stress in the wild type plants. Using the Beacon simulation and inference engines, GeneMania [113, 171], and eFP Browser [179], a view of the current understanding of signaling pathways in which G and/or X are thought to participate is obtained, in addition to actual data about other experimental conditions under which the expression of these genes is affected. Currently, there is no comprehensive repository of stress signaling pathways in plants and no integration of results from different sources. The implementation of Beacon will fill the first need and the iPlant workflow the second.

**Results.** Results obtained computationally suggest that X, the timing and extent of whose expression is partially controlled by TF G, acts downstream to affect the timing of hormone-regulated processes associated with defense against ozone [98, 178], but not previously associated with the onset of senescence. The data-processing tools employed and the nature of the experimental design allow distinctions to be made between early responses to the stress and later acclimation processes. There was also an association with a ubiquitination pathway that was not observed before for ozone action, although this connection was already reported for drought stress [134].

6 Management Plan

Objectives, task distribution, milestones indicated in Table 2. **Objective 1 Curation:** GRA/Grene (PPWS), PD/Pereira (VBI). **Objective 2 Initial Beacon system:** GRA/Heath (CS). **Objective 3 Simulation and inference engines:** GRA/Heath (CS), GRA/Grene (PPWS). **Objective 4 Experimental validation:** GRA/Grene (PPWS), PD/Pereira (VBI). **Responsibility:** Co-PI Grene coordinates the Curation effort in communication with external collaborators (letters of commitment). Co-PI Heath supervises the training of GRAs/PD for accomplishment of computational tasks. PI Pereira supervises tasks under objectives 1 and 4. **Annual milestones and deliverables:** Year 1 — Curation of biological data and workshop for dissemination of methodology to a global consortium of Collaborators. Year 2 — Beacon databases and tools established for interaction with users. Year 3 — Simulation and inference engines available for user studies. Year 4 — Beacon pipeline based predictions validated by case studies. **Internal communication** is done by weekly project meetings, including bi-weekly video/tele-conferencing with collaborators, to discuss specific signaling pathways. Workshop in the second year will be used for initial training of external collaborators. Beacon will be implemented in a portable language (e.g., Python or Java). The MySQL relational database system will be used for the Beacon database (Heath). **Sustainability Plan.** All code for Beacon will be freely available through an appropriate open source license and distributed perma-
nently through SourceForge. Once completed, tools will be made available on top of the iPlant infrastructure, allowing both tools and data to remain on active servers past the end of funding.

7 Broader Impacts

The proposed Beacon system, to serve as a repository and simulation engine for molecular signaling pathways in plants and their relationships to phenotypic effects, will interest undergraduate students in the computational sciences or computational biology, as well as experimental biologists. Well-established outreach programs will be used to provide training and educational outreach activities.

**Multicultural Academic Opportunities Program (MAOP).** The mission of the MAOP program is to encourage and support the academic achievements of a diverse student body in STEM fields from Virginia Tech. Programs and activities serve students from the pre-college through the doctorate level with a continuum of financial, academic, and social support, as well as research opportunities. VBI (the PI’s affiliation) participates in the MAOP program by sponsoring and/or hosting summer research internships for undergraduates [160]. Co-PI Grene has close ties with the MAOP program and has served on the Advisory Board for an NSF-funded project through MAOP from 2007–2009 (see letter of support from Dr. Karen Sanders, Interim Vice President for Diversity and Inclusion).

**S-STEM.** (See letter of support from Dr. Jill Sible, Associate Dean, College of Science, Virginia Tech). The NSF-funded S-STEM Program trains a cohort of undergraduates in the classroom and laboratory to be competitive for careers in biotechnology. S-STEM scholars take common classes and participate in a weekly seminar dedicated to academic and career preparation, research skills, ethics, and diversity. Beginning in their sophomore year, students are encouraged to participate in laboratory research. These research experiences will be enhanced with funds from the proposed project. S-STEM students can be important early users of the Beacon system, performing complete pathway analyses. The Co-PIs will supervise the S-STEM students. Starting with relevant omics data, they will be invited to visualize those data in the context of a Beacon signaling pathway, use the data to infer new pathway fragments with the Beacon inference engine, conduct “what-if” exploratory simulations with the Beacon simulation engine, and commit any new pathway versions to a Beacon database. Their feedback to the team will be highly valued.

**PREP and PREP-U.** We will collaborate with Dr. Erin Dolan at Virginia Tech, principal investigator of the Partnership for Research and Education in Plants (PREP) and PREP for Undergraduates (PREP-U), and David Lally, PREP coordinator. Through PREP (supported by a Science Education Partnership Award from the National Center for Research Resources; see attached support letter), and PREP-U (initiated with funding from the NSF Course, Curriculum, and Laboratory Improvement Program), high school students and undergraduates challenge Arabidopsis mutant and wild-type plants with abiotic stresses, therefore providing opportunities to conduct authentic scientific research. Beacon results will be utilized to frame investigations appropriate for these students. Students will then design and conduct experiments as tests of Beacon-generated hypotheses and report their findings back to project investigators. The PIs will be directly involved in PREP and PREP-U activities in three ways. First, they will collaborate with Dolan and Lally to develop brief video pieces explaining in non-technical language the types of experiments worthy of investigation. Second, they will provide laboratory materials appropriate for students to design and conduct the investigations. Third, in collaboration with Dolan, Lally, and PREP and PREP-U instructors, the PIs will discuss students’ investigations with them, either in person or using web technologies such as PREP/PREP-U discussion boards, wikis, or video chats. These discussions will encourage students
to develop reasoning and argumentation skills by prompting them to support their experimental decisions with evidence (e.g., why an investigation is worth doing, why a particular experimental design makes sense, and what interpretations are supported given their experimental results).

**Genetics, Bioinformatics, and Computational Biology (GBCB).** The interdisciplinary GBCB Ph.D. program is a new area of study tailored to students’ individual backgrounds. Co-PIs Grene and Heath are founding members of GBCB and serve on its steering committee. The program encompasses the applications of molecular biology, genomics, mathematics, statistics, and computer science to the life sciences. GBCB is designed to provide a combination of discipline-specific and cross-disciplinary course work, as well as a multidisciplinary research environment. GBCB faculty members come from 11 departments and the Virginia Bioinformatics Institute. Four specialty tracks are defined: Life Sciences, Computer Science, Statistics, and Mathematics. Each student specializes in a track to which he or she is suited and is required to gain expertise in at least one other track. All students are required to receive some training in the life sciences, regardless of their chosen track. The Beacon project will hire GBCB students for the computational part of the project and will recruit GBCB students in the life sciences track to assist with Objective 4.

## 8 Results from NSF Support

**Current: Plant Genome Award.** Pereira (PI), Grene (Co-PI), and others obtained the Plant Gen- nome Research Program award DBI-0922747 ($2.4M; 9/15/09 – 8/31/12) “Cereal Drought Stress Response and Resistance Networks”. Drought gene interaction networks are being created in rice and maize to provide a systems view of drought gene functions. Transcriptome data from Arabidopsis and cereals were generated for a comparative analysis across plants, using standardized drought treatments to simulate field drought conditions. Differentially regulated genes were compared using orthologous loci, and significant species-common drought down-regulated genes identified represented by the GO-slim categories. Putative drought responsive rice regulatory genes were identified with available Ds/dSpm transposon inserts (approximately 150). The homozygous mutant lines are being generated to characterize the drought responsive phenotypes. The rice-maize common regulatory genes will be used to identify candidate drought responsive orthologs in maize, to be validated by association analysis for dissection of drought response phenotypes. The gene interaction networks developed for Arabidopsis and rice are being used to study other biological pathways.

**Prior: ITR Award.** Heath (PI) and Grene (Co-PI) had “ITR-0219322: Understanding Stress Resistance Mechanisms in Plants: Multimodal Models Integrating Experimental Data, Databases, and the Literature” ($499k, 09/01/2002 – 12/31/2006, and a DCC supplement for work with the International Potato Center in Peru) and developed the multimodal network concept to integrate biological information from functional genomic data on drought stress in Arabidopsis genotypes, online databases, and the literature to represent and analyze molecular stress-resistance mechanisms in plants. XcisClique, a system developed in the Grene/Heath group, associates annotated genome and gene expression data; models known cis-elements as regular expressions; identifies maximal bicliques in a bipartite gene-motif graph; and ranks bicliques based on their computed statistical significance. XcisClique analysis of the promoter regions of Arabidopsis HSP90s and their putative co-chaperones identified candidate co-chaperones for each HSP90 in abiotic stress-responsive genes at the Nottingham web site. Wet lab projects in the areas of drought stress, ozone stress, and effects of elevated carbon dioxide were completed. The central focus was on the responses of particular accessions to stress. Response diversity reflected differences in constitutive and inducible networks. Publications include [52, 57, 58, 92, 93, 102, 103, 129, 146, 147, 149, 150, 164, 165, 173, 174].
References cited


