

Achieving the in Silico Plant. Systems Biology and the Future of Plant Biological Research

Peter V. Minorsky prepared this summary report with assistance from participants of the DOE workshop on Plant Systems Biology.

Plant biology has the potential of providing partial solutions to several of the most daunting problems that our species and our planet face in the 21st century. Among these problems will be an increasing shortage of available food, the depletion of global oil reserves, and a mounting scarcity of freshwater.

Feeding the growing world population is a major challenge. The world's population currently stands at over 6.3 billion and is increasing by 80 million people per year. Already the global community is failing to supply adequate nutrition and a minimal standard of living to 1 billion of our members, and these problems will only become more acute if, as predicted, the global population swells to 8 to 11 billion people by the year 2050.

Oil, a resource of critical importance to the economic vitality of developed nations, is predicted to become ever more scarce in the 21st century. Available evidence indicates that the world's accessible petroleum reserves may be largely exhausted by 2060 (Masters et al., 1991; Kerr, 1998). Although several countries, including China and the United States, have adequate coal reserves to support current energy use rates for several hundred years, the utilization of coal, because of mining-associated pollution and acid rain, is even more destructive to the environment than is the production and consumption of oil. It is clear that the world in the not-too-distant future will need to meet its energy demands in a very different manner than it is doing today.

Freshwater is another natural resource that will become increasingly scarce in the 21st century. Because of increased population growth and economic development, global withdrawal rates of surface water are projected to at least double in the next 2 decades. Worldwide, agriculture is the primary consumer of freshwater. About 70% of all water withdrawn each year from rivers, lakes, and aquifers is used to irrigate 17% of the world's cropland. This irrigated land produces approximately 40% of the world's food supply. Water shortages may become even more acute in some areas as a result of global climate change.

How can plant biology provide a partial solution to these daunting problems? One of the few natural resources that is increasing globally is carbon dioxide. Because plants convert carbon dioxide into bio-

mass by the process of photosynthesis, higher carbon dioxide levels increase the potential for biomass production. Plant biomass can, of course, be burned directly and inefficiently for energy as our ancestors did before recorded history. Using more modern techniques, however, plant biomass can be converted into renewable fuels such as ethanol that are clean burning and whose widespread utilization would not require a complete overhaul of existing technology. Because of population growth, we cannot divert prime agricultural land from the growing of food to the growing of biomass for fuel production. To meet the demands of the future for plants both as food and fuel, it will be necessary to expand agriculture into marginal lands, achieve higher yields per acre, and increase the water use efficiency of crops. Such attributes in crops, of course, have long been goals of traditional plant breeders, but the progress they have made in achieving them, although commendable, has been slow. Because we are faced with a deadline, it is no longer sufficient merely to experiment with the interactions of a small number of genes and hope for the best: We must gain the ability to engineer the design of crop plants predictably. To do this, it is necessary to gain a comprehensive knowledge of how plants function in a relatively short period of time. We must develop our understanding of how plants function to such an extent that we can precisely predict by modeling how a plant will respond to any given genetic manipulation or environmental perturbation. Fortunately, as we step into the 21st century, the tools for achieving this degree of knowledge are available or within grasp.

Using genomic techniques, we can now identify all the genes in a plant, and have successfully sequenced two plant genomes in their entirety—the model organism *Arabidopsis* and the crop plant rice (*Oryza sativa*). Moreover, using microarray and proteomic techniques, we now have the ability to resolve which genes are activated or inactivated during development or in response to an environmental change. However, identifying all the genes and proteins (system elements) in an organism is comparable with listing all the parts of an airplane. Although such a list provides a catalog of the individual components, by itself it is not sufficient to understand the complexity underlying the engineered object. What we really need to know before we can intelligently en-

gineer plants is how all these system elements interact.

Systems biology is a new branch of biology that attempts to discover and understand biological properties that emerge from the interactions of many system elements (Ideker et al., 2001; Kitano, 2002). The major reason why systems biology is gaining interest today is that progress in molecular biology, particularly in high-throughput genomics and proteomics, is enabling scientists to collect comprehensive data sets on a wide variety of plant responses. The power of these new tools has led to an explosion of information unparalleled in the history of biology. The implications of these sweeping changes in plant biology were the topic of discussion among a diverse group of scientists who met on January 19, 2003, after the 22nd Riverside Symposium in Plant Biology. The goal of the meeting was to identify some of the needs and opportunities that lie ahead in attaining the comprehensive understanding of plants that is critical if a predictive computer model of a generic plant is to become a reality (Chory et al., 2000).

BRAINSTORMING PLANT SYSTEMS BIOLOGY

The specific aim of the workshop was to discuss and identify what types of tools and what sorts of data will be needed by plant biologists in the post-genomic era to integrate our knowledge of plants at the molecular, cellular, and organismal level; that is, to gain a systems level understanding of plants. The sheer immensity of the information explosion in plant biology presents new challenges. How are we to organize and store these massive data sets in standard and easily accessible forms? How can we develop new high-throughput experimental tools to gather and configure such data into interactive models? If we are to study interacting proteins, metabolites, and cellular processes dynamically, what advances in plant cell biology are required? What types of software must be created so that model simulations of key plant processes can be tested? Because plant scientists will need to cross boundaries between diverse scientific disciplines to establish research collaborations with engineers, computer scientists, chemists, and other research specializations, what is the best way to accomplish such interdisciplinary collaborations? A summary of the participants' discussion concerning these questions is provided here with the hope that it will stimulate further dialogue within the broader plant science community concerning the future of plant systems biology.

This workshop, sponsored by the Energy Biosciences Program of the Department of Energy, was cochaired by Chris Somerville (Carnegie Institution, Stanford, CA) and Elliot Meyerowitz (California Institute of Technology, Pasadena). The other attendees included Hamid Bolouri and Timothy Galitski (Institute for Systems Biology, Seattle), Joanne Chory and

Joseph Ecker (Salk Institute, La Jolla, CA), Terry Gaasterlund (Rockefeller University, New York), Ken Keegstra (Michigan State University, East Lansing), Rob Last (Max Planck Institute of Chemical Ecology, Jena, Germany), Shauna Somerville (Carnegie Institution), John Shanklin (Brookhaven National Laboratory, Upton, NY), Eric Mjolsness (University of California, Irvine), Klaas van Wijk (Cornell University, Ithaca, NY), Klaus Palme (University of Freiburg, Germany), Mary Wildermuth (University of California, Berkeley), Natasha Raikhel (University of California, Riverside), Laurence Lejay (New York University), Peter Minorsky (Mercy College, Dobbs Ferry, NY), and Walter Stevens, James Tavares, and Sharlene Weatherwax (U.S. Department of Energy). In addition, Gloria Coruzzi (New York University) and John Quackenbush (The Institute for Genomic Research, Rockville, MD) both made contributions in absentia.

SYSTEMS BIOLOGY. A NEW WAY OF PRACTICING BIOLOGY

At the very core of systems biology is the goal of being able to model a living organism. Systems biology examines the structure and dynamics of cellular and organismal function, rather than the characteristics of isolated parts of a cell or organism. The major reason systems biology is gaining interest today is that progress in molecular biology, particularly in genomics, proteomics, and high-throughput measurements, is enabling scientists to collect comprehensive data sets on the mechanisms underlying plant growth and plant responses to perturbations. The new high-throughput tools of genomics have provided biologists with the potential to systematically perturb and monitor biological systems while they are functioning. With the wealth of information provided by these new approaches, plant biological research is becoming more reliant on informational science. This interaction of plant molecular biology with informational sciences will help pinpoint which types of experimental analyses and measurements need to be made.

Systems biology requires quantitative data that are high quality and comprehensive. Comprehensive-ness in systems biology requires three types of measurements. First, we need to measure the expression levels of a large number of mRNAs, proteins, structural polymers, and metabolites simultaneously. Second, we need to heighten the temporal resolution of such molecular changes to model dynamic changes. Third, we need to spatially resolve where these changes are occurring in the plant at the level of the cell type. To expedite the collection of comprehensive and accurate data, technical innovations in high-throughput experimental measurement including microscopy and robotics need to be fostered. To design these new high-throughput tools, plant biolo-

gists will have to work side by side with engineers who design and operate high-precision and high-throughput measurement systems.

GENOMICS AND PROTEOMICS

More has been learned about how plants work in the last decade than in all of preceding history. In no small part, this rapid rate of progress has been due to advances in genetics and genomics. New methods that build on genomics, such as DNA microarrays (DNA chips) have allowed investigators to measure simultaneously which genes are turned on or turned off in a genome in response to an experimental treatment. Thus, many biologists are calling the beginning of the 21st century the "post-genomic era." By this, they are not implying that we should stop sequencing genomes or analyzing the results of such genomic studies but rather that all the major hurdles to obtaining genomic information have been overcome. The next big challenge is to understand the functions of all gene products.

Microarray studies monitor gene expression—which genes are being turned on or off during development or in response to a perturbation. Techniques that profile changes in gene expression permit the analysis of the expression levels of thousands of genes simultaneously. Proteomic methods reveal the proteins translated from the mRNA molecules that are the direct result of gene expression. Two-dimensional gels were the first method to allow the identity of proteins in a proteome, but newer methods employing tandem mass spectrometry enable identification of even more proteins. Even with these new techniques, proteomics is a much thornier problem than genomics. In contrast to DNA and RNA, there are currently no techniques available to amplify proteins of low abundance (the equivalent of PCR for DNA) or to identify them with very high-dynamic resolution. Moreover, all of the cells that compose a plant have the same genome, but each cell type has a different collection of proteins. In addition to these cell-to-cell variations in proteomic expression, we must also consider proteomic variation over time. As a plant cell matures or responds to an environmental perturbation, its proteome and expressed mRNAs are constantly changing.

In proteomics, it is also of interest to determine not just whether a protein is present but how much of the protein is present and how active it is. The activity of many proteins is regulated by interactions with other proteins that form complexes or catalyze structural modifications. Comprehensive proteomics, therefore, requires not only a list of the protein functions but also a detailed understanding of protein-protein interactions within a cell and protein modifications that regulate function, such as phosphorylation, glycosylation, etc. Methods for the global measurements of various protein modifications, including mass spec-

trometry and the use of modification-specific antibodies, are only now being developed. In contrast to genomics in which standardized tools and standards for data storage and sharing are available, the current methodology for handling large proteomics data sets is less uniform and organized.

Most proteins are enzymes that catalyze specific chemical reactions. Thus, changes in gene expression are typically accompanied by changes in the levels of many cellular chemicals. The simultaneous measurement of these collective chemical changes is the emerging field of metabolomics. Like proteomic changes, metabolomic changes are highly dynamic and require precise high-throughput measurements with enough spatial and temporal resolution to be useful for modeling purposes. As with proteomics, major advances in analytical chemistry are required to allow us to identify and accurately measure the thousands of metabolites in a plant. Parallel advances in proteomics and metabolomics are expected to add additional layers of resolution to our ability to integrate all aspects of a biological process. Attempts to store and organize proteomic and metabolomic data in an easy-to-access format are still in their infancy. Thus, there is a real need for software developers and data archivists to establish the tools for creating proteomics and metabolomics databases for use by plant biologists (see <http://pat.sdsc.edu/perl/browser.pl?tax=Arabidopsis%20thaliana> for recent progress in this respect).

CELL BIOLOGY AND IMAGING

Improvements in high-throughput optical measurements will be necessary before we can efficiently determine the genomic and proteomic changes with sufficient spatial and temporal resolution to allow accurate simulations of plant responses (Girke et al., 2003). Optical methods, in particular, provide the key to understanding how the parts of plants make the whole and to learning how the association of parts—proteins, multiprotein machines, and cells—changes as plants grow and function. Protein interactions can be seen by fluorescence resonance energy transfer, and the movement of proteins within cells can be seen by fluorescent labeling of proteins with extra protein sequences such as green fluorescent protein. The real-time observation of cell division patterns is possible in living plants with a combination of fluorescent reporter genes, fluorescent dyes, and laser scanning confocal microscopy. Higher resolution methods require fixed or frozen tissue or purified proteins and protein complexes but have the advantage of providing critical detailed information that can be combined with dynamic information from optical methods. The highest resolution methods are those of x-ray crystallography, which with modern robotics and bright-synchrotron x-ray sources can be adapted for high-throughput measurement. One of

the methods now being developed for determining the structures of large protein complexes, at lower resolution, is cryoelectron microscopic tomography.

It is not possible to model a eukaryotic cell without knowing where each constituent in the cell is located with respect to the compartments, what the properties of these compartments are, and what defines these properties. Obtaining this information is a major unmet challenge. Conceivably, proteomic approaches combined with innovative methods for subcellular fractionation could provide a catalog of what proteins are located in each compartment. Information about the properties of each protein could be derived from *in vitro* studies of each protein or protein complex. However, a similar profile of the metabolite constituents will remain challenging for the foreseeable future (Sweetlove et al., 2003). The development of novel methods for visualizing the concentration of specific metabolites in subcellular compartments of living cells (Fehr et al., 2002) provides an example of what may be possible. The challenge now is to implement these and related methods and to develop methods that will allow simultaneous imaging of many constituents in such a way that they can be integrated with parallel information about other biological processes. Ultimately, it may be impossible to measure the concentration of every constituent in every compartment of every cell. Therefore, we need to develop *in vitro* or computational models for studying how networks of metabolic pathways perform under the conditions found in subcellular compartments *in vivo*. It seems likely that by having a small number of "sentinel" metabolites for which the absolute concentrations and fluxes are known, we can model the corresponding concentrations of most other metabolic constituents.

MODELING AND QUANTIFICATION

The successful modeling of living organisms is the ultimate goal of systems biology. Clearly, useful modeling will depend on having a large amount of high-quality quantitative information about all aspects of biological processes. Many new types of data have to be systematically determined. For example, it is not currently possible to model flux through a metabolic pathway without knowing the rate constants for the enzymes, how much of the enzymes are present, the subcellular distribution of the enzymes and metabolites, the availability of the inputs, and the disposition of the outputs. Only when we have determined the amounts and locations of each protein and the amounts and locations of each metabolite under a variety of conditions may it prove possible to heuristically derive the necessary quantitative information to support modeling. Modeling should greatly simplify our attempts to understand plants. For example, genes with similar responses over a range of conditions are often clustered together to form functional groups (networks). Often,

these associated genes are under the control of common transcription factors. Recent computer simulations of partial or whole genetic networks have demonstrated the complexity of network behaviors and emergent properties that were not apparent from the examination of a few isolated interactions alone (Jönsson et al., 2003).

Assuming that the challenge of obtaining such information can be met, a substantial challenge remains in developing computer programs that will model biological processes. The ideal program will mimic the performance of a cell or tissue or organism. The modeling of intact higher plants will be especially challenging because of the differential responsiveness of various cell types to a given perturbation. The collection of the comprehensive data needed for modeling might initially be most successful using single-cell microorganisms or higher plant cells grown in defined liquid cultures. The modeling of *Escherichia coli* is already under way (Stelling et al., 2002), and this should facilitate the modeling of cells of more interest to plant biologists. To model plant response accurately, a multitude of software programs of the sorts widely used by engineers (e.g. parameter optimization, flux balance analysis, systems analysis, and computer model simulations, to name a few) need to be adapted for the study of plants. The integration of modeling with experimental work will derive many new insights, with greater complexity, and hopefully greater impact on the global problems we face. A key benefit of modeling is that we can rapidly do many "virtual experiments" and select those that are most interesting for the real world. For example, we could assess the tradeoffs "in silico" of different means for increasing the water use efficiency of a given crop given a set of possible genotypes and a range of possible environmental conditions. The generation of phenotype phase diagrams would enable plant biologists to predict with accuracy which factors are limiting to plant growth under different environmental conditions.

If we are to collect data from single cells, and this may be a necessary first step, then there is an urgent need to improve the accuracy and scope of single-cell measurements and even molecular (nanobiological) measurements. The high quality and comprehensiveness of measurements needed by systems biologists will require further automation and the introduction of inexpensive microfluidic systems for handling and analyzing extremely small sample quantities. The determination of the optimal concentrations of precipitants in protein crystallization experiments provides an example of the use of new technology for an old, difficult, and time-consuming problem (Hansen et al., 2002). Perhaps the unique ability of microfluidic devices to produce concentration gradients can be coupled with advances in fluorescent substrates or miniaturization of analytical instruments to simplify studies of enzyme reactions.

As we accumulate more and more high-quality data, mathematical modeling of biological processes will become ever more successful; that is, the responses of plants to genetic manipulations and environmental perturbations will become increasingly predictable. This will enable us to engineer plants quickly and with foresight as to the cost benefit of the proposed “reprogramming” of the plant genome.

DATA MANAGEMENT

The amount of biological information that the revolution in genomics, proteomics, and metabolomics is generating is staggering. This has created a new problem: how to store and handle this massive quantity of information, often from databases present on computers around the world and in varied format. Many new types of software need to be developed for archiving and accessing this information efficiently and in a standard method. Because systems biology requires the sharing of data among members of the scientific community, facilities (data warehouses) must be established and maintained. Traditionally trained biologists, who typically have had little experience manipulating the large data sets of the post-genomic era, have to acquire a working knowledge of data storage and access. Training in the use of such massive data archives must become *de rigueur* for young biologists. Thus, there is need for training grants to teach biologists—young and old—how to gain competency in data storage, retrieval, and analysis. There is also a need within professional biology for biologists to embrace bioinformatic software developers and data archivists as *bona fide* members of our community and not as mere adjuncts to it.

HUMAN RESOURCES

The explosive growth of molecular biology in the last decade has helped to break down many artificial barriers that had been erected between different branches of biology. Now, it is necessary that the walls between biology and other scientific disciplines (e.g. mathematics, computer science, and engineering) be breached. Biologists of this new era need to become conversant with software developers, data storage archivists, mathematicians, analytical chemists, and engineers. If we are to achieve the “*in silico*” plant, the biologists of this new era will need to understand in some depth the models of mathematicians and computer scientists, and biologists will need to supply the accurate measurements of the many parameters that computational scientists will require to make their models robust. As a consequence, there is a strong need for the development of training programs and the initiation of training grants to foster the intellectual development of both

the current and the next generation of scientists in all of the subdisciplines that constitute systems biology. The sociology of science will also need to change. For example, our students need to be trained to thrive in larger and extended research groups, and our universities must acknowledge and reward multidisciplinary collaboration and community service by their faculty. We can learn from successful examples such as high-energy physics and distributed software engineering.

Unfortunately, each of the technical disciplines that systems biology attempts to merge has developed separately, and each has evolved its own technical jargon that seems impenetrable and arcane to the outsider. This barrier to communication must be overcome, and proximity is the best way to do that. This might be achieved through cross-disciplinary research opportunities for graduate students and postdoctoral research associates. Funding special centers devoted to systems biology research and staffed by biologists, computer scientists, mathematicians, and engineers could be another approach to foster the development of systems biology and to achieve the goal of attaining the “*in silico* plant.”

CONCLUSION

The challenges faced by Earth have never been more acute than they are now at the dawn of the 21st century. “*In silico* plants” may provide a unique albeit partial solution to several of the most daunting problems that we will face in the coming decades (the scarcity of oil and freshwater and the need to produce more food). Although there is no single solution to these emerging problems, systems biology is a vehicle for gaining advanced knowledge of mechanistic plant biology that may provide the scientific basis for such partial solutions. There is already ample evidence that plant productivity can be further enhanced for food, feed, and fiber production and that an increased understanding of biotic and abiotic stress tolerance mechanisms may allow us to produce more food with less water (Somerville and Briscoe, 2001). However, to engineer plants with foresight and within the requisite timeframe, we need to achieve a comprehensive understanding of how plants work. We need to understand plants so thoroughly that we can accurately predict how they will respond to any given genetic manipulation or environmental perturbation. We must progress from simply trying to understand the function of each element in a plant individually to trying to understand how all these elements work together in the living plant. Only in this way can we maximize the usefulness of plants to benefit humanity in a realistic time frame.

What do we need to begin the path from our current understanding of the parts of plants to a synthesis that allows the understanding of their interactions? We require the development of new high-throughput

experimental tools and approaches to gather and configure data, with contributions from the traditional individual investigator and through consortium efforts. We require improved and facile plant cell imaging techniques to dynamically track interacting proteins, metabolites, and cellular processes. We require advances in computational biology to generate software that is widely adaptable for creating and testing model simulations of key plant processes. Finally, plant scientists will need to cross boundaries between diverse scientific disciplines to establish research collaborations with engineers, computer scientists, chemists, and other research specializations.

This year, we celebrate the 50th anniversary of the discovery of the structure of DNA. In just half a century, scientists did what was unthinkable at the outset: They went from studying one gene at a time to revealing entire plant genomes. In the post-genomic era, the next generation of scientists are challenging themselves to do what seems unthinkable at this point in time—to put all the pieces of a plant together into a complete model with interacting parts. We believe that achieving the “in silico plant” will be more than just an intellectual triumph: We foresee its creation as a critical stepping stone to the extensive genetic fine-tuning of crop plants that will be necessary in the not-too-distant future if future generations are to have sufficient food, fuel, and freshwater. We believe that we possess all the resources and talents necessary to reach our goal. What we now require is a national resolve to marshal the

talent and facilities necessary to see the “in silico plant” become a reality.

LITERATURE CITED

- Chory J, Ecker JR, Briggs S, Caboche M, Coruzzi GM, Cook D, Dangl J, Grant S, Guerinot ML, Henikoff S et al. (2000) National Science Foundation-Sponsored Workshop Report: “The 2010 Project,” functional genomics and the virtual plant: a blueprint for understanding how plants are built and how to improve them. *Plant Physiol* **123**: 423–426
- Fehr M, Frommer WB, Lalonde S (2002) Visualization of maltose uptake in living yeast cells by fluorescent nanosensors. *Proc Natl Acad Sci USA* **99**: 9846–9851
- Girke T, Ozkan M, Carter M, Raikhel NV (2003) Towards a modeling infrastructure for studying plant cells. *Plant Physiol* **132**: 410–414
- Hansen CL, Skordalakes E, Berger JM, Quake SR (2002) A robust and scalable microfluidic metering method that allows protein crystal growth by free interface diffusion. *Proc Natl Acad Sci USA* **99**: 16531–16536
- Ideker T, Galitski T, Hood L (2001) A new approach to decoding life: systems biology. *Annu Rev Genomics Hum Genet* **2**: 343–372
- Jönsson H, Shapiro BE, Meyerowitz EM, Mjolsness E (2003) Signaling in multicellular models of plant development. *In* S Kumar, PJ Bentley, eds, *On Growth, Form, and Computers*. Academic Press, London (in press)
- Kerr RA (1998) The next oil crisis looms large—and perhaps close. *Science* **281**: 1128–1131
- Kitano H (2002) Systems biology: a brief overview. *Science* **295**: 1662–1664
- Masters CD, Root DH, Attanasi ED (1991) Resource constraints in petroleum production potential. *Science* **253**: 146–152
- Somerville CR, Briscoe J (2001) Genetic engineering and water. *Science* **292**: 2217
- Stelling J, Klant S, Bettenbrock K, Schuster S, Gilles ED (2002) Metabolic network structure determines key aspects of functionality and regulation. *Nature* **420**: 190–193
- Sweetlove LJ, Last RL, Fernie AR (2003) Predictive metabolic engineering: a goal for systems biology. *Plant Physiol* **132**: 420–425