The Multinational Coordinated *Arabidopsis* 2010 Project

Functional Genomics and the Virtual Plant: A blueprint for understanding how plants are built and how to improve them

Report from an NSF-sponsored workshop held at The Salk Institute for Biological Studies, January 13-14, 2000

TABLE OF CONTENTS

Executive Summary

About 10 years ago, plant scientists adopted the widespread use of an easily manipulated model called Arabidopsis thaliana, and established an international research effort called "The Multinational Coordinated Arabidopsis thaliana Genome Research Project." An outcome of this effort is the entire DNA sequence of this plant will be completed by mid-2000. For the first time, we will know the sequence of 25,000 genes that are necessary for a plant to function as a flowering plant. As a follow-up to the Arabidopsis genome sequencing efforts, an important and revolutionary new initiative - to exploit the revolution in plant genomics by understanding the function of all genes of a reference species within their cellular, organismal and evolutionary context by the year 2010 [the 2010 Project] - has been proposed by the community of plant biologists. Implicit in this mission statement is an endorsement of the allocation of resources to attempts to assign function to genes that have no known phenotype, which represents a

significant departure from the common practice of defining and justifying a scientific goal based on the biological phenomena. The rationale for endorsing this radical change is that for the first time it is feasible for plant biologists to envision a whole system approach to study of plant form and function. This report contains the recommendations of an ad hoc committee representing the community of Arabidopsis and other plant researchers that met at The Salk Institute for Biological Studies, La Jolla, CA, on January 13 and 14, 2000, to discuss the feasibility of commencing a publicly-funded program to determine the function of all Arabidopsis genes during the next decade, using a systems approach. The committee has identified specific project goals, including:

- Development of expanded genetic toolkits as a service to the research community
  - Whole-systems approach to identification of gene function from the molecular to evolutionary levels
  - Expanding the role for bioinformatics
  - Development of human resources
  - International collaboration

The committee discussed the impact that this project would have on the progress of basic plant research as well as on the strategic interests of the United States as they relate to agriculture, energy, the environment and human health. The committee recommended that this program, as outlined in this document and in more detail in an accompanying report from a meeting held earlier, should commence as soon as possible.

INTRODUCTION

Biologists are assimilating a new paradigm. After a generation of characterizing genes one or a few at a time, scientists now have access to the complete genome sequences of many bacterial species and several eukaryotes. Most biologists can now access the complete genome sequence of their favorite organisms, or a proxy thereof, in powerful electronic databases. Access to this information, and new tools that exploit it, will profoundly alter the ways in which we pose and answer questions in biology.

About 10 years ago, plant scientists adopted the widespread use of an easily manipulated model called *Arabidopsis thaliana*, and established an international research effort called "The Multinational Coordinated *Arabidopsis thaliana* Genome Research Project." An outcome of this effort is the entire DNA sequence of this plant will be completed by mid-2000. For the first time, we will know the sequence of 25,000 genes that are necessary for a plant to function as a flowering plant. Indeed, the success of the Multinational *Arabidopsis* Genome Project stands as an example of productive foresight.

The purpose of this document is to outline what is required in the next ten years, to meet the challenges offered by the new paradigm of Functional and Integrative Genomics in plant biology, taking the opportunity opened up by the availability of

complete sequence of the Arabidopsis genome. Exploitation of the genome information will establish Arabidopsis as the reference organism for other plants, as a blueprint for comparative genomics, and for understanding evolution of plants as well as other organisms.

Mission Statement:

To exploit the revolution in plant genomics by understanding the function of all genes of a reference species within their cellular, organismal and evolutionary context.

Long Term Goal:

In order to most efficiently and safely manipulate plants to meet growing societal needs, we must in essence create a wiring diagram of a plant through its entire life cycle: from germinating seed to production of the next generation of seeds in mature flowers. These processes are guided by genes and the proteins they encode. They are directed by both intrinsic developmental cues and environmental signals. The long-term goal for plant biology following complete sequencing of the Arabidopsis genome is to understand every molecular interaction in every cell throughout a plant lifecycle. In essence, to understand the function of every gene, by the year 2010.

The ultimate expression of our goal is nothing short of a virtual plant which one could observe growing on a computer screen, stopping this process at any point in that development, and with the click of a computer mouse, accessing all the genetic information expressed in any organ or cell under a variety of environmental conditions. Completion of the *Arabidopsis* genome sequence provides significant leverage for future plant genome projects. The reference genome is a platform from which useful comparisons are simplified. We will ultimately be able to predict the evolution of new gene function by comparative genomics with other key plant species.

Where will the 2010 project lead us in the future?

Whole systems based knowledge of the entire biology of a reference species confers predictive power that will enable the following:

1) Predictable outcomes to directed experimental genetic changes.

2) Directed genetic changes that accelerate domestication of wild species.

3) Facile genetic manipulation that ensures maintenance of, and expansion of germplasm bases.

4) A description of the underlying mechanisms of heterosis, and the ability to use this phenomenon more effectively.

5) Enhanced understanding of the genetic basis of phenotypic plasticity, which will have a profound impact not just in plants, but also in animals, including humans.

6) Knowledge of "the minimum gene set" required for plant life.

7) Understanding of the genetic basis of plant evolution which will enrich our understanding of the diversity of life on earth.

8) An understanding of interactions between plants and other organisms in their environment, up to the level of ecosystems.

Scientific Objectives:

New experimental tools that investigate gene function at the subcellular, cellular, organ, organismal and ecosystem level need to be developed. New bioinformatics tools to analyze and extract meaning from increasingly systems-based datasets will need to be developed. These will require, in part, creation of entirely new tools. An important and revolutionary aspect of the 2010 Project is that it implicitly endorses the allocation of resources to attempts to assign function to genes that have no known function. This represents a significant departure from the common practice of defining and justifying a scientific goal based on the biological phenomena. The rationale for endorsing this radical change is that for the first time it is feasible to envision a whole-systems approach to gene and protein function. This whole-systems approach promises to be orders of magnitude more efficient than the conventional approach.

Exploitation of the Arabidopsis genome sequence to achieve the goals outlined above will require significant new investment in dedicated, focused, Genome Technology Centers. These Centers should be both dispersed and single site. They will serve the research community at large by providing services and by producing new tools using economies of scale. The Centers will be dedicated to the creation of genome-wide tools, rather than the application of genome-wide tools to solving specific research problems. They will thus be explicitly public service oriented, as was the Arabidopsis genome sequencing project.

In creating genome-wide tools, the Centers must complement and significantly enable investigators throughout the United States, and indeed the world. Individual investigators will be at once the main clientele for the Centers and the dispersed creators of knowledge. The value of this project therefore depends on significant support being available for individual research laboratories throughout the plant biology research community to leverage investment in both the

Arabidopsis genome sequencing project and the proposed Centers to solve a wide range of specific biological problems.

1) Expanded Genetic Toolkit:

A key strength of Arabidopsis as a model is its facile forward genetics. One can isolate mutants disrupted in any process and study the effects of each mutation. Currently, roughly 40% of the genes found in the genomic sequence do not encode a protein of predictable function. In order to identify functions for these genes, we need to develop a more sophisticated genetic toolkit for both forward and reverse genetic screens. This toolkit will include the following:

1-3 Year Goals:

_ Develop the essential genetic toolkits, including: (1) comprehensive sets of sequence indexed mutants, accessible via database search, (2) whole genome mapping chips, and (3) facile conditional expression systems for sensitized and saturation screens for rare alleles

3-6 Year Goals:

_ Establish facile analysis of natural variation via development of many combinations of large families of Recombinant Inbred Lines (RILs).

_ Comprehensive construction of defined deletions of linked, duplicated genes.

_ Develop methods for directed mutations and site specific recombination.

10 year Goal:

_ Plant artificial chromosomes.

2) Whole-Systems Identification of Gene Function

a. Global Analysis of Gene Expression

A first step towards determining the function of all plant genes is to catalog each gene's expression through the life cycle. The integration of all gene regulation events defines the development of an organism, and algorithms currently exist which can predict common gene expression patterns given global expression data. This information will become a platform from which the concerted action of gene sets in the formation of tissues and organs can be elucidated. We need to use, and improve upon, currently available technologies in global analysis of gene expression in order to describe the regulatory logic underlying the wiring circuitry for plant development in the virtual plant.

1 Year Goal:

_ Construction of gene specific DNA probes for expression analysis.

1-3 Year Goal:

_ Definition of full length cDNAs to facilitate annotation of the genome and subsequent analyses of protein expression

3-6 Year Goal:

_ Global mRNA expression profiles at organ, cellular and subcellular levels under a wide variety of environmental conditions.

10 Year Goals:

_ Identification of cis regulatory sequences of all genes.

_ Identification of regulatory circuits controlled by each transcription factor in the genome.

b. Global Analysis of Protein Dynamics

The sum of gene expression changes is translated through development into the proteins from which cellular machines are built. Hence, understanding protein dynamics will enable prediction of what machines exist, and how they work, through a plant life cycle. The technologies required to achieve these goals are nascent, and will be improved upon using Arabidopsis as a reference organism, to the benefit of research efforts in any system.

1-3 Year Goals:

_ Production of antibodies against, or epitope tags on, all deduced proteins.

_ Global protein profiles at organ, cellular and subcellular levels under a wide variety of environmental conditions.

3-6 Year Goal:

_ Global understanding of post-translational modification.

10 Year Goals:

_ Biochemical function determined for every protein.

_ Three-dimensional structures of members of every plant specific protein family.

c. Metabolite Dynamics

Plant growth and development is dictated, to a large degree, by the uptake, trafficking, storage and use of low molecular weight metabolites. Plant cellular factories produce a bewildering array of secondary metabolites, upon which a large amount of drug and product discovery are based. Understanding metabolite dynamics will result in more efficient use of soil and water based nutrients and will allow rationally designed food and pharmaceutical production in plant factories.

3-6 Year Goal:

_ Global metabolic profiling at organ, cellular, subcellular levels under a wide variety of environmental conditions.

10 Year Goal:

_ Systems analysis of the uptake, transport and storage of ions and metabolites.

d. Global Catalogues of Molecular Interactions

The ultimate arbiters of cellular function are the complex protein machines encoded by the mRNA population in each cell at any time during development. The ultimate understanding wiring diagram is a catalogue of molecular interactions which occurs in each cell of the organism through its lifecycle. This ambitious experimental layer incorporates the four above into what will be the final expression of the virtual plant.

10 Year Goal:

_ Global description of protein-protein, protein-nucleic acid, protein-metal, and protein-small molecule interactions at organ, cellular, subcellular levels under a wide variety of environmental conditions.

e. Comparative Genomics

Completion of the Arabidopsis genome sequence provides significant leverage for future plant genome projects. The reference genome is a platform from which useful comparisons are simplified. We will ultimately be able to predict the evolution of new gene function by comparative genomics. We can glimpse the power of comparative genomics as a tool to understand plant evolution and

diversification through the recent strides made in the understanding of plant disease resistance gene structure.

3-Year Goal:

_ Identify species for survey genomic sequencing based on an expanded definition of phylogenetic nodes

10-Year Goals:

_ Survey genomic sequencing, and deep EST sampling from phylogenetic node species.

_ Define a predictive basis for conservation versus diversification of gene function.

_ Within species genomic sequence comparisons.

_ Develop tools for whole genome population biology.

An Expanding Role for Bioinformatics

To achieve the goals above will require significant investment in and development of bioinformatics tools and databases from which the information required to build the virtual plant will be stored and extracted. Bioinformatics will be a key to the success of the 2010 project. A significant effort in this area must be expended in close coordination with the biological aspects of the project.

Ultimately, the database that we envision will provide a common vocabulary, visualization tools, and information retrieval mechanisms that permit integration of all knowledge about an organism into a seamless whole that can be queried from any perspective. Of equal importance for plant biologists, an ideal ATIR will permit a user to use information about one organism to develop hypotheses about less well-studied organisms. This is important because it is unrealistic to expect someone working on a plant such as wheat or poplar to know the intimate details about the latest results with Arabidopsis or maize. Thus, our goal, is to develop facile tools that permit an individual working outside the model species to formulate a query based on the organism of interest, have that query directed to the relevant knowledge for the plant models, and present the information about the models in a way that can be understood by the plant biology community at large.

Ongoing Goals:

_ A wide range of datasets will need to be incorporated into global databases. Database architecture allowing easy integration with other databases will be an essential component of this effort. New and existing analytical tools will need to be developed to integrate divergent types of data (e.g. expression array data and in situ hybridization). The raw data which gives rise to the virtual plant will need to be archived. The ability to generate the datasets described above will easily outstrip the ability to rationally maintain, manage, and extract utility from this data. Hence, there is a critical need to invest in novel approaches to bioinformatics and to also bolster support for current databases. All tools developed should be available through the Internet and designed to take full advantage of the simplicity and ubiquity of platform-independent browsers.

_ We will need to develop new cellular and whole plant visualization tools. These will include both virtual tools to manipulate informatics-based data, and new in vivo imaging systems with which we can observe in real time the changes in genome expression which dictate plant development.

_ A critical bioinformatics need for plant biology is to educate biologists in the use of the tools that are in place and those that will be available soon. It will be essential to attract computer science students and trainees to plant biology where they can participate directly in plant research initiatives and where their talents can be productively applied to bioinfomatics.

Development of Human Resources

The Arabidopsis community has developed into an excellent training ground for plant scientists. The changing paradigm of functional genomics will require new sorts of training to encourage and facilitate lateral, interdisciplinary approaches to problem solving.

Ongoing Goals:

_ An explicit endorsement of Genome Technology Centers providing services and economies of scale for systems-based data generation is not consistent with the traditional training of doctoral and post-doctoral researchers, and the traditional output measurement of publications. Therefore, we need to recruit and incent a core of dedicated technical assistants and research associates as personnel for these facilities. Note that recent undergraduates are often recruited into these positions and they become the next generation of eventual doctoral seekers.

_ We will still need traditionally trained doctoral and post-doctoral researchers with skills in the areas below. We also will need to design and

implement new training programs which recognize the growing importance of computer science, statistics and physical sciences in plant biology. We should encourage interdisciplinary training which specifically seeks a systems based approach for both undergraduate and graduate level students.

_ Short and long term post-doctoral fellowships are a necessary component of this training. Plant biologists who wish to enhance their informatics or systems based knowledge would be well suited to this. Also, individuals trained in peripheral disciplines who wish to enter plant science should be encouraged.

_ Summer courses or other intensive specialized workshops are effective means for retooling and continuing education of established investigators in a rapidly moving field like plant biology.

International Collaboration

The Multinational Arabidopsis Steering Committee and the North American Steering Committee have been, over the last ten years, important liaisons between the community at large and policy markers in north America, Europe and the Pacific-rim. In the 2010 Project, we will continue to work with partners in these areas, many of whom have already begun efforts like those described here. To date, the Arabidopsis community is a model of how these structures can eliminate duplication of effort and enhance research resources worldwide. An expanded role for the Multinational Steering Committee should be a part of the 2010 Project. International Workshops to monitor/advise on Project 2010 will be essential for international coordination of these efforts.

2010 Salk Workshop Participants

Participants

Joanne Chory, Co-Chair, The Salk Institute for Biological Studies

Joseph R. Ecker, Co- Chair, University of Pennsylvania

Steve Briggs, Novartis Agricultural Discovery Institute

Michele Caboche, Unité de Recherche en Génomique Végétale INRA France

Gloria Coruzzi, New York University

Doug Cook, Texas A&M University

Jeff Dangl, University of North Carolina

Sarah Grant, University of North Carolina

Mary Lou Guerinot, Dartmouth College

Steve Henikoff, Fred Hutchinson Cancer Research Center

Rob Martiensen, Cold Spring Harbor Laboratory

Kiyotaka Okada, Kyoto University, Japan

Natasha Raikel, Michigan State University

Chris Somerville, Carnegie Institute of Plant Biology, Stanford University

Detlef Weigel, The Salk Institute for Biological Studies

Observers

Dr. Mary Clutter, National Science Foundation

Dr. Machi Dilworth, National Science Foundation

Dr. Jane Silverthorne, National Science Foundation

Dr. Greg Dilworth, Department of Energy

Dr. Peter Bretting, U S Department of Agriculture