POLICY FORUM: GENOMICS

Plant Biology in 2010

Chris Somerville* and Jeff Dangl

The imminent completion of the *Arabidopsis* genome sequence represents the culmination of a 10-year effort by the plant biology community. By the time the community was ready to undertake high-throughput sequencing in 1996, a broad international infrastructure and a history of international cooperation were in place to support a collegial and efficient sequencing project.

As the project draws to a close, a new 10-year project of comparable import is about to begin (*1*) (Table 1). At the heart of the proposed program is the ambitious goal of knowing the function of all plant genes by the year 2010. More boldly, it is envisioned that this knowledge will facilitate the development of a virtual plant—a computer model that will use information about each gene product to simulate the growth and development of a plant under many environmental conditions. In response to a proposal from the community, The U.S. National Science Foundation has recently announced a new funding program that is expected to provide up to $25 million per year for the project (*2*). In conjunction with programs for funding *Arabidopsis* functional genomics in other countries, the NSF 2010 Project will certainly accelerate progress toward understanding basic plant biology within the next decade. However, we believe that the project will also change the way in which academicians study plant biology. Here, we have attempted to outline some of the issues that we believe will drive these changes, and to offer opinions on how to attain desirable outcomes.

There are about 25,900 genes in *Arabidopsis*, of which only about a thousand have been assigned a function by direct experimental evidence (*3*, *4*). Approximately 55% of *Arabidopsis* genes can currently be assigned a putative function, but not a biological role, by sequence comparisons to genes in public databases. The 2010 Project envisions that, 10 years from now, every gene in *Arabidopsis* will have been subjected to one or more experiments in which the gene is inactivated or overexpressed. The resulting phenotypes will be examined by all available criteria, including full-genome mRNA expression profiling and metabolic profiling. Comprehensive information will be available about where and when every gene is expressed; where the protein is localized; how the protein is modified; and what, if any, other proteins the gene product interacts with (*5*). The substrates will be known for the more than 900 protein kinases (*6*), and the promoters controlled by each transcription factor will be defined. The information from these and other assays will be available in powerful databases that will interrelate *Arabidopsis*-specific information with information available for homologs from other species. In those cases where genes have apparent functional duplicates in the genome, the phenotype of the multiple mutant will be established. This information will be used to infer not only the specific function of the gene products, but also the role of every gene in the growth and development of the organism. Finally, this knowledge will be applied to the study of all higher plant genes, not just *Arabidopsis*.

Several recently formed companies are already racing with the academic community to complete the task of studying all *Arabidopsis* genes. One of these, Paradigm Genetics, claims to be able to create and analyze the phenotypes of mutations in 5000 genes per year with only a few hundred scientists (*7*). Thus, there is little doubt that the *Arabidopsis* community of several thousand people can marshal the personnel and technical resources for the task. However, doing so will necessitate some significant changes in the allocation of *Arabidopsis* research funds.

As almost any research with a genetic component can be considered to be "functional genomics," the opening of the public coffers for expanded research on the function of all *Arabidopsis* genes could simply devolve into a situation in which more funding becomes available for hundreds of individual laboratories pursuing their specific interests. This would produce many important discoveries but would almost certainly leave a large proportion of the genome unexplored for decades to come and would be a very inefficient use of resources. For instance, it would be much more efficient to have one

REPRESENTATIVE GOALS OF THE 2010 PROJECT

1- to 3-Year Goals
Develop essential genetic tools, including the following: •comprehensive sets of sequence-indexed mutants, accessible via database search •whole-genome mapping and gene expression DNA chips •facile conditional gene expression systems.
Produce antibodies against, or epitope tags on, all deduced proteins.
Describe global protein profiles at organ, cellular, and subcellular levels under various environmental conditions.
3- to 6-Year Goals
Create a complete library of full-length cDNAs.
Construct defined deletions of linked, duplicated genes.
Develop methods for directed mutations and site-specific recombination.
Describe global mRNA expression profiles at organ, cellular, and subcellular levels under various environmental conditions.
Develop global understanding of posttranslational modification.
Undertake global metabolic profiling at organ, cellular, and subcellular levels under various environmental conditions.
10-Year Goals
Plant artificial chromosomes.
Identify *cis* regulatory sequences of all genes.
Identify regulatory circuits controlled by each transcription factor.
Determine biochemical function for every protein.
Describe three-dimensional structures of members of every plant-specific protein family.
Undertake systems analysis of the uptake, transport, and storage of ions and metabolites.
Describe globally protein-protein, protein-nucleic acid, and protein-other interactions at organ, cellular, and subcellular levels under various environmental conditions.
Survey genomic sequencing, and deep EST sampling from phylogenetic node species.
Define a predictive basis for conservation versus diversification of gene function.
Compare genomic sequences within species.
Develop bioinformatics, visualization, and modeling tools that will facilitate access to all biological information about a representative virtual plant.

Table 1. Goal for investigating *Arabidopsis* functional genomics.

The authors are in the Department of Plant Biology, Carnegie Institution, Stanford, CA 94305, USA, and in the Biology Department, University of North Carolina, Chapel Hill, NC 27599, USA, respectively.

*To whom correspondence should be addressed. E-mail: crs@andrew2.stanford.edu

laboratory determine the DNA sequences flanking a hundred thousand transposon insertion sites than to have hundreds or thousands of scientists isolating and cloning insertions in their favorite genes one at a time. More generally, the 2010 Project envisions that there are many kinds of experiments that can be done most efficiently and most rapidly as a community service by a small number of research groups, or "technology centers," that can implement high-throughput methods. The current U.S. centers that provide community access to large numbers of *Arabidopsis* insertion mutations or DNA microarrays are a first step in this direction (*8*).

Although high-throughput approaches can provide a wealth of useful information, we do not expect the 2010 Project to change the modus operandi of most plant biologists. Rather, the information generated by genomics approaches should empower the average plant biologist to pursue detailed knowledge of specific aspects of plant biology. We consider it likely that in order to accomplish the goals of the 2010 plan, it will be necessary to have relatively large numbers of laboratories using genomic methods to study a subset of the genome (such as all genes containing a particular motif) rather than the whole genome. This raises the problem that if one group is funded to characterize a class of genes, it may be wasteful to fund other groups to study the same genes by the same criteria. Thus, the first group to propose a broad, large-scale investigation of a class of genes could obtain a de facto monopoly on funding.

We believe that grants to study a large number of genes must be explicitly service oriented; all results and materials generated must be made freely and rapidly (before publication) available to the community as they are produced, in much the same way that the sequencing projects released data with minimal delay. To provide transparency to the community, each project will be obliged to post the accession numbers of their target genes in a public database so that anyone can determine the probability of duplication of effort. As with the sequencing projects, such a service-oriented approach will necessitate changes in how the academic system rewards contributions in biology (*9*), or perhaps, some aspects of the work should be directed to companies or technology centers. Whatever the case, principal investigators should not anticipate using these "community-service projects" as a means of funding their traditional research projects.

Perhaps the most radical concept in the 2010 Project is that it specifically endorses the allocation of public research support to studying plant genes that have no sequence similarity to any gene of known function. This necessitates a departure from the traditional model of supporting research, in which scientists seek grant support to study a gene or genes, by arguing that the genes are a key to understanding an important biological phenomenon. Grant review panels typically rank grants on the basis of their agreement with the proposed significance of the likely outcome and the feasibility of achieving that outcome. How will panels evaluate a proposal to study a family of genes that have no similarity to any other gene of known function? How can the qualifications of an applicant be evaluated if there is no hint of what aspect of biology is to be studied? This could resemble "cassette science" in which a generic set of experimental approaches are used to study any set of genes. In the extreme, a standard proposal could be written with empty blanks for the accession numbers of the genes to be analyzed. In addition to the unacceptable effect this could have on creativity, it would probably not be desirable to have individuals randomly allotted to study genes that may bear no relation to their expertise and interests.

One way of dealing with these problems is to focus initially on supporting whole-genome approaches to generate preliminary information about each gene. Proposals to identify all protein-protein interactions or all subcellular localizations might fall in this class. Similarly, proposals to develop and exploit whole-genome microarrays or gene chips to study the expression of all genes under a wide variety of conditions may permit clustering of most genes into known pathways or processes (*10*). For instance, it might be feasible to determine the function of all of the genes of unknown function that are simultaneously induced during cold acclimation or infection by a pathogen. This model seems well suited to single investigators or consortia of scientists with shared goals and complementary expertise.

An unresolved issue associated with the new push toward functional genomics concerns the different standards for release of data and materials around the world. A good example of international collegiality is the Genomic *Arabidopsis* Resource Network (GARNet) in the United Kingdom, which provides unrestricted access to resources for functional genomics (*11*) and two initiatives funded by the Biotechnology and Biological Science Research Council (*12*). One project will produce 30,000 transposon gene trap insertions. Another will sequence the flanking DNA of large numbers of insertions. All information is to be public, and all materials will be freely available through the Nottingham *Arabidopsis* Stock Center.

By contrast, the European Union (EU) and several European countries currently view research funding for plant functional genomics as an investment in regional or national competitiveness and do not appear to be supportive of immediate release of data and materials. For instance, the EU Fifth Framework Program (FP5), which has recently provided major funding for *Arabidopsis* functional genomics (*13*) "has been conceived to help solve problems and to respond to major socio-economic challenges facing the European Union") (*14*). FP5 provides funding for consortia of academic and company researchers under circumstances in which participating companies may obtain preferential access to results, and there are no requirements for early release of data or materials. If two groups in different countries are both funded to study the function of all genes of a certain class, and only one group is obliged to instantly release all data and materials, this group may be significantly disadvantaged (*15*). However, by providing grant support for achievement of technical, community-oriented goals, the sting of releasing data and materials to a potentially uncooperative competitor should be diminished. The groups that participated in the *Arabidopsis* sequencing project, including the EU-funded consortium, released data very rapidly. It is to be hoped that the relevant European funding agencies will recognize the inherent unfairness of encouraging proprietary research in functional genomics projects that are large-scale, international, and public.

It is axiomatic that scientific discoveries cannot be planned, but arise unexpectedly from individuals following the imperatives of personal curiosity. Also, great progress has been made in many fields when scientists working toward specific large goals are supported. Indeed, large-scale projects work best when they are focused around technical goals. In this respect, the 2010 Project should probably be viewed as a technical program that will enable revolutionary discoveries by the plant biology community.

References and Notes

1. J. Chory *et al.*, *Plant Physiol.* **123**, 423 (2000); the full text is available at http://arabidopsis.org/workshop1.html
2. www.nsf.gov/pubs/2001/nsf0113/nsf0113.htm
3. D. Bouchez, H. Höfte, *Plant Physiol.* **118**, 725 (1998).
4. C. R. Somerville, S. C. Somerville, *Science* **285**, 380 (1999).
5. P. Utez *et al.*, *Nature* **403**, 623 (2000).
6. G. MacBeath, S. L. Schreiber, *Science* **289**, 1760 (2000).
7. www.paragen.com
8. http://afgc.stanford.edu
9. L. Rowen *et al.*, *Science* **289**, 1881 (2000).
10. S. Chu *et al.*, *Science* **282**, 699 (1998).
11. www.york.ac.uk/res/garnet/garnet.htm
12. www.bbsrc.ac.uk/science/initiatives/Welcome.html
13. News in Brief, *Nature* **407**, 667 (2000).
14. www.cordis.lu/fp5/management/particip/v-gfp2.htm#Objectives
15. R. S. Eisenberg, *Nature Rev. Genet.* **1**, 70 (2000).
16. The concepts of the 2010 Project were developed in collaboration with R. Beachy, J. Chory, J. Ecker, S. Briggs, M. Caboche, G. Coruzzi, D. Cook, C. Gasser, S. Grant, M. L. Guerinot, S. Henikoff, S. Kay, K. Keegstra, R. Martiensen, S. McCouch, D. Meinke, E. Meyerowitz, K. Okada, N. Raikel, E. Ward, D. Weigel, and S. Wessler.