Deciphering the Parts List for the Mechanical Plant

Chris Somerville

Abstract: The development of inexpensive DNA sequencing technologies has revolutionized all aspects of biological research. The proliferation of plant genome sequences, in conjunction with the parallel development of robust tools for directed genetic manipulation, has given momentum and credibility to the goal of understanding several model plants as the sum of their parts. A broad inventory of the functions and interrelationships of the parts is currently under way, and the first steps toward computer models of processes have emerged. These approaches also provide a framework for the mechanistic basis of plant diversity. It is hoped that rapid progress in this endeavor will facilitate timely responses to expanding demand for food, feed, fiber, fuel, and ecosystem services in a period of climate change.

CHRIS SOMERVILLE is the Philomathia Professor of Alternative Energy and Director of the Energy Biosciences Institute at the University of California, Berkeley. His current research focuses on plant cell-wall polysaccharide synthesis and conversion to liquid fuels. His work has appeared in *Science*, *Proceedings of the National Academy of Sciences*, and *Current Biology*, among other publications.

As the end of the previous millennium drew near, I accepted an invitation from a leading biomedical journal to summarize the major advances in knowledge of mechanistic plant biology during the preceding one hundred years.[1] By *mechanistic*, I mean an intellectual framework that seeks to understand an organism as the sum of its parts: that is, in terms of the chemical structures and reactions that support life. One of the challenges of summarizing this vast topic was disentangling the many advances that revolutionized our understanding of plants nonspecifically – a rising tide of knowledge that lifted all boats. In spite of their long separation from a common ancestor, plants and animals (and fungi and bacteria) contain many proteins and genes with significant structural similarity. Thus, much of what is known about the function of plant proteins and genes, and the molecular and cellular processes they participate in, has been inferred from knowledge of the function of homologous genes in other types of organisms. In this sense, the study of plant biology is a subset of a broad campaign to understand all life forms. However, plants and animals are thought to have separated about 1.6 billion years ago from a

common unicellular ancestor; so in addition to the fundamental differences in how they obtain energy, there are many interesting differences in how multicellularity and adaptive responses evolved.

After surveying the previous century, I concluded that early biologists were astute observers, and that succeeding generations had not obviously improved on that aspect of inquiry but had better experimental tools, analytical devices, and context. One key factor underlying major advances was technological improvements that facilitated compelling experiments. Thus, for instance, Calvin's elucidation of the path of carbon during photosynthesis was enabled by the availability of the newly discovered ^{14}C isotope from the nearby Berkeley accelerator. Similarly, advances in plant biology over the past ten years were also largely attributable to improvements in one technology – DNA sequencing – and one sociological phenomenon: enthusiasm for model species. In 2000, an international consortium completed the first full genome sequence of a higher plant, *Arabidopsis thaliana*, and placed it in the public domain via a graphical Web interface that allowed users to browse the genome and connect individual genes to the scientific papers describing their functions. Because it was completed before the development of the very high-throughput sequencing technologies that are widely used today, the DNA sequence is estimated to have cost more than $75 million to produce. By comparison, an essentially complete DNA sequence of an Arabidopsis plant can be obtained today for about $5,000 – a dramatic testament to the development of DNA sequencing technology during the last decade.

The second major factor underlying progress in mechanistic plant biology was the widespread adoption of Arabidopsis as a model organism. Interest in Arabidopsis began in the early 1980s, when the first generation of plant biologists to engage in gene cloning and characterization recognized the virtues of an easily cultivated organism with a small diploid genome and a short life cycle. In conjunction with some parallel developments in recombinant DNA technologies, the availability of the Arabidopsis genome sequence changed mechanistic plant biology more profoundly than any plant-specific discovery of the previous century.

To understand why the genome sequence combined with the widespread use of model species is enabling, it is useful first to reflect on how the mechanistic aspects of plants, and most other model organisms, are currently understood. In brief, all organisms can be viewed as adaptive machines that exist to make copies of themselves, or hybrid copies of themselves and their sexual partners. DNA encodes a parts list and some instructions for how many parts to make under various circumstances. Some parts – proteins and RNA molecules – have characteristic lifetimes that may vary by several orders of magnitude and according to information about where to locate themselves within a cell, which other parts to interact with, and how to carry out those interactions. Proteins (and to a lesser extent some RNAs) have the ability to make or modify other parts – usually simple chemicals such as lipids, amino acids, nucleic acids, and sugars that are the building blocks of cells – or to carry excited electrons that are used to power life.

A long-term goal of many plant biologists is to understand what each of the roughly 33,600 genes in Arabidopsis and other plants does and to integrate that information into a predictive mechanistic model. Ideally, this would be a computer model: the cyberplant. Given that at least sixteen thousand scientists worldwide use Arabidopsis as a model species

for research, we could collectively obtain detailed experimental information about the function of every gene in a plant like Arabidopsis within a decade. We could meet this objective by organizing the community to eliminate duplication of effort and maximize information sharing. I believe that a first pass at a complete description of a plant is well within reach. Because all flowering plants (angiosperms) evolved from a common ancestor within the past 125 million years, the mechanisms underlying many aspects of growth and development have been conserved at the molecular level. Thus, a detailed analysis of all genes in several strategically selected angiosperms will provide a broad base of knowledge that is applicable to all higher plants. Although I have emphasized Arabidopsis here because it is the most advanced model plant, I anticipate that similar approaches will be implemented in several other species – for example, rice (*Oryza sativa*) – that represent divergent nodes of angiosperm diversity, and that knowledge of all plants will ultimately involve interpolation from deep knowledge of strategically placed nodes.

The standard method for interrogating gene function is to increase or decrease the activity of the gene and observe how that change affects relevant processes or the organism as a whole. Activity levels can be altered by genetically increasing or decreasing the amount of the mRNA or protein encoded by the gene or, frequently, by altering the catalytic activity of the protein, its ability to bind other proteins, or its location. In one technique, the so-called reverse genetics approach, the investigator replaces an endogenous gene with an altered copy and observes the effect. The challenge of this powerful approach is knowing which aspect of the organism to test for an effect. Indeed, the investigator may find it necessary to become an expert in all aspects of biology in order to design useful tests. Thus, the more broadly useful approach has been to randomly mutagenize the genome, then screen for mutants in which a process of interest is altered. This approach exploits the investigator's deep knowledge of a specific aspect of biology and facilitates the testing of many variants of a gene, revealing not only more or less of the gene product but also more subtle effects, such as loss of regulatory factors. The challenge in this case is to identify the corresponding gene. Additionally, both approaches may be complicated by the presence of duplicated genes that mask the effects of a mutation in only one of the genes. The two approaches are complementary and are frequently used simultaneously.

The availability of a complete genomic DNA sequence makes it relatively easy to identify the mutation corresponding to any genetic difference between two accessions of Arabidopsis. In the simplest cases, a researcher might identify an interesting new mutant and, by genetically mapping the mutation underlying the phenotype to the DNA sequence, can identify the corresponding gene. Similarly, natural variation between accessions can be resolved by genetic crosses to single genetic loci and mapped onto the DNA sequence to identify the basis for differences. Thus, any aspect of plant growth and development that can be marked by a mutation can be linked to a change in one or more specific genes. Because it has now become so inexpensive to obtain the complete genome sequence of an Arabidopsis plant, some researchers have resequenced the entire genome of the mutant – approximately 150 million base pairs – to find the change in DNA sequence corresponding to a mutation.[2] The relatively small size of the Arabidopsis genome compared to most other higher plants makes genome resequencing particularly easy. However, similar approaches are possible for many

species of higher plants, although some species are easier than others because of genome size, ploidy, self-incompatibility, and related features.

The benefits of having a large number of scientists working on one or a few model organisms are manifold. First, the community benefits from the availability of research tools and reagents of broad utility. Indeed, the mere fact that a large user community exists encouraged some scientists to invest large amounts of time and effort in building tools that would benefit the community. An important example is the "Arabidopsis insertion collection": several hundred thousand accessions of Arabidopsis in which a fragment of exogenous DNA is randomly inserted in the genome, frequently within a gene. To create the collection, several groups produced large numbers of transgenic plants with random insertions of exogenous DNA. Then, for each of the hundreds of thousands of DNA insertions, the DNA flanking the insertion was recovered and sequenced so that the location of the insertion in the genome could be determined. Seeds of each of the insertion mutants were made freely available at several stock centers around the world. Thus, when a researcher wishes to investigate the function of a gene, he or she can log into an electronic database and request seeds from one or more lines that specifically lack a functional copy of that gene. Variation in the properties of the gene can be explored by genetically transforming the mutant with natural or synthetic variants of the gene or its relatives.

As another benefit of the community approach, researchers studying different phenomena frequently discover that they are observing different aspects of a common process. Thus, for instance, a researcher who discovers a protein that catalyzes a specific chemical reaction might find that the protein was previously found to be required for some other process, such as disease resistance, by a colleague with no knowledge of the catalytic function of the protein. This kind of second-order knowledge creation, which is essential to the eventual development of a comprehensive understanding, is accelerating through the prevalence of electronic data resources and the expanding ability of the community to link biological information to specific genes.

Another DNA sequence-based approach that has revolutionized plant biology exploits very high-throughput RNA or DNA sequencing technologies or DNA hybridization methods in order to measure simultaneously the abundance of mRNAs for each gene in a tissue sample. At the most basic level, this method catalogs which genes contribute to the state of a tissue sample under the condition in which the sample was taken. Investigators can compare samples taken from different tissues or conditions to observe how the organism reshapes gene expression in order to respond to different developmental states or environmental conditions. Perhaps more important, by using computational methods to compare the data compiled from large numbers of experiments, it is possible to identify genes that are coregulated (that is, expressed at the same time and place). Searching for highly coregulated genes frequently allows investigators to identify previously unknown components of processes. For instance, by searching for genes that were highly coregulated with the known subunits of the enzyme complex that synthesizes cellulose, my colleagues and I recently identified genes that had not previously been implicated in the process.[3]

At present, this gene-centric approach sheds light primarily on isolated mechanisms rather than the operation of the organism as a whole. The field of plant

biology is to a large extent still in an inventory phase, in which the genes that contribute to all aspects of growth and development are being identified and ordered into networks and pathways. Thus, for instance, the genes that contribute to flower development or disease tolerance have been identified by searching for mutations that alter these processes, cloning the corresponding genes, assigning probable function to the genes by comparing the gene sequences to databases of all previously known genes, identifying the genes that are coregulated, analyzing mutations in those genes, and so on. Ultimately, these analyses allow placement of genes into pathways that sequentially carry out specific tasks. Listing the parts and placing them in pathways and networks will presumably be followed by a phase in which the many mechanisms that comprise a whole organism will be conceptually integrated. That phase seems likely to take place on computers that will generate testable hypotheses and identify where experimental measurements are needed to populate models and simulations. I expect that phase to mark the arrival of theoretical biology as a mainstream activity.

Mechanistic research on plant biology can be artificially subdivided into five major, intersecting topics: evolution, development, adaptation, biotic interactions, and molecular and cellular mechanisms. Developmental biologists are systematically describing the networks of genes and the cellular processes that underlie the seemingly miraculous development of a multicellular plant from a single cell. For the time being, much of the work focuses on describing how each tissue or cell type develops. For instance, the surface of plants is usually punctuated by the presence of large numbers (for example, thousands per cm^2) of pairs of cells (stomata) that open and close, like a pair of lips, to regulate the flow of gases and water vapor in and out of the leaf. Mutant analysis appears to have identified all the genes that are involved in this process. Specifically, careful examination of what fails to take place when each gene is altered has enabled biologists to observe the sequence of events and describe causes and effects in molecular detail. This effort has provided insight into how the differentiation of stomata from leaf epidermal cells takes place.[4] The current hope is that such knowledge will provide a road map for how other specialized cell types might develop. However, I think this paradigm is a stopgap measure based on the fact that we are in the midst of a large discovery process. I believe that in the longer term, scientists will describe in detail how every cell type develops in many plant species. In other words, unless we assume that research on plants will conclude at some point, the current use of models and examples will gradually be supplanted by complete descriptions of enough model species to allow predictive manipulation of any plant species.

As I use it here, *adaptation* refers to the ability of an organism to modify its morphology or composition in response to environmental cues. Because plants are sessile, most have a large repertoire of adaptive responses. For instance, in response to attacks by pests and pathogens, plants may activate pathways for the production of toxins. In response to low temperature or drought, some plants undergo a wide variety of changes in chemical composition that facilitate survival under those conditions. If exposed to toxic minerals, plants induce the expression of factors that sequester the toxins. Because light quality varies at different locations in the canopy, plants may alter the composition of the photosynthetic apparatus to make better use of light, or they may

stimulate growth to facilitate better access to light. The ability to measure the expression of all genes simultaneously has greatly facilitated a description of underlying mechanisms involved in these and many other adaptive responses. At the same time, the development of genetic methods has similarly facilitated the identification of the genes that control and participate in such responses. These discoveries have triggered the rational development of genetically modified plants that are better able to withstand stressful conditions. The first generation of transgenic crop plants engineered to better withstand drought conditions recently obtained regulatory approval, and the approval of freezing-tolerant trees is pending.

One of the richest areas of discovery during the past decade has been the elucidation of many of the mechanisms plants use to survive biotic interactions. As any gardener knows, there is a large number of organisms that can devastate plants: viruses, nematodes, insects, fungi, bacteria, slugs, and vertebrates of many kinds. It is remarkable that any plants survive in the natural world. Indeed, approximately 40 percent of agricultural productivity in Africa and Asia is reportedly lost to pests and pathogens.[5] Thus, one of the most promising avenues to increasing the availability of food and fiber is to explore ways to reduce such losses. One of the first transgenic crops grown commercially employed an insecticidal protein to reduce losses to insects that were not controlled adequately by other methods, such as the application of insecticide. Interestingly, most plants are resistant to most pests and pathogens. Crop damage is largely caused by highly specialized pests and pathogens that have become adapted to only a few host species. Understanding the factors that allow pests and pathogens to identify their hosts might facilitate the development of molecular cloaks of invisibility. This is essentially how "mosquito repellents" containing DEET provide protection: the active ingredient blocks one or more of several receptors used by mosquitoes to identify a host.[6] Additionally, most plants have a suite of defensive responses to pests and pathogens. In the most extreme case, infection by a pathogen triggers the death of cells in the vicinity of the infection, thereby starving the pathogen. In this and most other kinds of defensive reactions, there is a cost to the host, so the defensive mechanisms are not activated until necessary. Many aspects of the mechanisms by which plants sense and respond to these events have been discovered during the past ten years, presenting new opportunities to breed or engineer pest and pathogen resistance.

Beyond the defense mechanisms of plants, much recent progress has contributed to a broad understanding of basic mechanisms that operate at the cellular and molecular levels. Some advances entail comparative biology, in which the details of a molecular process are first worked out in an organism that has advantages for basic research, and then the homologous mechanism in several plant species is described. However, many important aspects of plant biology are unique to plants and, therefore, cannot be approached by using convenient model species such as yeast, nematodes, flies or mice. Thus, understanding light-mediated signaling, phytohormone-mediated responses, some aspects of pathology, and aspects of development have been hot topics during the past decade. Some plant-specific subjects, such as photosynthesis and cell wall biosynthesis, attract relatively small followings at present; but trends shift, and the recent interest in lignocellulosic fuels has ignited new interest in plant cell-wall biochemistry.

Essentially all aspects of knowledge about plant biology have surged in the

ten years since the completion of the Arabidopsis genome sequence and the subsequent DNA sequencing of many other plant species. One important outcome of the proliferation of genome sequences has been insight into the mechanistic basis of diversity. For instance, following the completion of the poplar genome, poplar trees and Arabidopsis were found to have remarkably similar gene content.[7] The significant differences in morphology and life cycle are manifestations of differences in the regulation of a very similar suite of genes. Likewise, progress in understanding the molecular basis of flower development has provided fascinating insights into why the Linnaean system of plant classification, which is based on flower morphology, has been so broadly useful. Knowledge of how gene action leads to floral morphology explains how a small number of changes in key genes can lead to very large differences in morphology.[8] The present is an exciting time for evolutionary biologists, who can now trace with precision the DNA rearrangements that accompanied or gave rise to the formation of new species. Plants are characterized by a high degree of polyploidy, and the remnants of ancient genome duplications can be seen in their genome sequences. The dynamic nature of plant genomes has become clear.

Looking forward, I predict that research in plant biology will be shaped by several major trends that seem certain to have broad impacts and by some peculiarities of the field. A central fact of studying plants is that there are many species we care about (roughly 180 are used by humans for nondecorative purposes). In contrast to biomedical research, in which most resources are directed toward one species, much of the effort and resources in plant biology are used to translate knowledge gained from models into other species or to generate insights without the technical benefits of working on a model. Support for work on model species is under pressure both from agricultural commodity groups that favor devoting research funds to species of economic importance, and from a large community of biologists for whom the important questions in biology are related to diversity, ecosystem function, and levels of explanation far removed from molecular processes. Thus, for the foreseeable future, a lack of financial support for research means that we will not discover the function of all genes in the model species or integrate such knowledge into a coherent understanding of how the organisms function. Ideally, this important quest will trickle along to completion later this century. I recognize that it is also important to understand the mechanistic basis of plant diversity and to apply knowledge to the plants of utility; but it is unfortunate that national priorities, particularly the enthusiasm for military force, limit progress toward knowledge that has enormous potential to affect human well-being. Indeed, I think that we are entering an era in which we will recognize a pressing need for deep knowledge about all aspects of plant biology.

We cannot know the future, but we can make some predictions based on the intersection of four key trends: the expanding human population, the growing economies of some less-developed nations, the impact of climate change, and the declining rate of discovery of new petroleum reserves. Put simply, the expansion of the human population to nine or ten billion will require production of more food, feed, and fiber. Effects of climate change on the distribution of rainfall will make it more difficult to produce crops in some regions. Economic expansion is associated with increased demand for animal protein, which, in turn, strongly increases demand for animal feed. Finally, both climate change and

declining petroleum reserves have created interest in production of energy from biomass, creating potential competition for land that might otherwise be used for food and feed production. A large amount of land worldwide can be used to expand agriculture and forestry, including as much as a billion acres that was farmed in the past and later abandoned to agriculture. However, because of market failures and poverty, even if that land is brought back into production, it may not be enough to prevent expansion of agriculture or forestry onto land that has never been cultivated. In tension with this fact is the widespread awareness that natural ecosystems are an important resource that must be preserved in large contiguous blocks in order to maintain the biological diversity contained therein. In the case of the Amazon, it is also possible that a large region of forest must be preserved in order to maintain the climate on which the forest depends for existence – particularly in the face of concurrent climate change. These trends will create incentives to intensify and optimize the production of most types of domesticated plants, including trees, so as to restrain the expansion of agricultural land.

At present, most of the knowledge required to intensify and optimize plant production is not of the mechanistic variety. Indeed, some of the most important opportunities may arise from broad knowledge of which types of plants can be used to produce food, feed, fiber, or fuel on marginal land. However, I believe that when we are able to understand in detail how several plant species operate as machines, we will be able to predict how those machines can be modified through breeding or genetic engineering to optimize production in the many climatic zones, photoperiods, and soil types that are available around the world. We will know how to accelerate breeding from the current eight- or ten-year cycles that may involve tens of thousands of test plots and that are too expensive to facilitate development of improved cultivars of most crops. We will know how to breed or engineer plants to resist key pests and pathogens, or to withstand drought, or to grow on saline soils. We will know how to generate hybrid vigor to achieve the greatest possible yield every year, rather than only occasionally. Plant-derived foods will contain optimal balances of nutrients to maximize feed efficiency and human nutrition. Perhaps we will be able to convert important annual species into perennials so that the energy and environmental costs of annual tilling will be reduced. Ultimately, we will realize the inherent potential of each of the many domesticated plant species, and we will use mechanistic knowledge to accelerate the domestication of many more. Mechanistic knowledge will liberate us from the thousands of years of trial and error that have produced the domesticated species we rely on today.

ENDNOTES

[1] Chris Somerville, "The 20th Century Trajectory of Plant Biology," *Cell* 100 (2000): 13 – 25.

[2] Ryan S. Austin, Danielle Vidaurre, George Stamatiou, Robert Breit, Nicholas J. Provart, Dario Bonetta, Jianfeng Zhang, Pauline Fung, Yunchen Gong, Pauline W. Wang, Peter McCourt, and David S. Guttman, "Next-Generation Mapping of Arabidopsis Genes," *The Plant Journal* 67 (2011): 715 – 725.

[3] Staffan Persson, Hairong Wei, Jennifer Milne, Greer Page, and Chris Somerville, "Large-Scale Coexpression Analysis Reveals Novel Genes Involved in Cellulose Biosynthesis," *Proceedings of the National Academy of Sciences* 102 (2005): 8633 – 8638.

[4] Juan Dong and Dominique Bergmann, "Stomatal Patterning and Development," *Current Topics in Developmental Biology* 91 (2010): 267 – 297.

[5] George N. Agrios, *Plant Pathology*, 5th ed. (New York: Academic Press, 2005).

[6] Mathias Ditzen, Maurizio Pellegrino, and Leslie B. Vosshall, "Insect Odorant Receptors are Molecular Targets of the Insect Repellant DEET," *Science* 319 (2008): 1838 – 1842.

[7] Gerry A. Tuskan et al., "The Genome of Black Cottonwood, *Populus trichocarpa* (Torr. & Gray)," *Science* 313 (2006): 1596 – 1604.

[8] Przemyslaw Prusinkiewicz, Yvette Erasmus, Brendan Lane, Lawrence D. Harder, and Enrico Coen, "Evolution and Development of Inflorescence Architectures," *Science* 316 (2007): 1452.