Predictive Metabolic Engineering: A Goal for Systems Biology[1]

Lee J. Sweetlove, Robert L. Last[2]*, and Alisdair R. Fernie

Department of Plant Sciences, University of Oxford, South Parks Road, Oxford, OX1 3RB, United Kingdom (L.J.S.); Department of Genetics and Evolution, Max-Planck-Institute of Chemical Ecology, Jena, Germany (R.L.L.); Department of Lothar Willmitzer, Max-Planck-Institut für Molekulare Pflanzenphysiologie, Am Mühlenberg 1, Golm 14476, Germany (A.R.F.)

The landmark sequencing of *Escherichia coli*, Brewer's yeast (*Saccharomyces cerevisiae*), Arabidopsis, and human genomes has facilitated the emergence of systems biology—a science that is currently gaining the attention of molecular biologists. In addition to satisfying our intellectual desire to have a more detailed understanding of plant biology, we need to understand how plants work well enough to undertake truly predictive engineering of morphology and chemical composition. This would allow us to efficiently produce plants that are more productive sources of food and fiber as well as novel nutritional and industrial chemicals. What is currently described as genetic engineering of plant metabolism is really tinkering. Very rarely do our genetic manipulations cause the predicted effects—we usually discover new rate-limiting steps that prevent accumulation of the desired compound or induction of diversionary or catabolic pathways, or we observe undesirable pleiotropic effects that limit the usefulness of the modification.

To enable true engineering of plants will require major changes in the way we approach science as well as major technological breakthroughs in analytical chemistry, bioinformatics, and other areas of genomics. We must generate robust mathematical models of metabolic pathways that accurately describe current experimental observations. These models can then be used to generate experimentally testable hypotheses, the results of which can be used to refine the model (Fig. 1). The goal is to go through this process enough times to create a model that begins to predict the behavior of the plant following perturbations by adding transgenes, making mutations altering cell fate, or environmental perturbations.

WHAT IS METABOLIC SYSTEMS BIOLOGY?

At its most basic level, the goal of plant metabolic systems biology is to create a comprehensive multidimensional representation of all of the biosynthetic reactions in a plant. This sounds like filling in plant-specific details on the standard wall chart that hangs in many laboratories or the increasingly powerful tools found on the Internet (for example, the Kyoto Encyclopedia of Genes and Genomes metabolic descriptions; http://www.genome.ad.jp/kegg/metabolism.html). However, this model must also indicate metabolic fluxes: the amounts of intermediates and products that accumulate in a specific tissue or cell type and in each organelle at a given moment in time, and the sizes of the "arrows" or the rates at which the enzymatic and nonenzymatic conversions take place. This futuristic virtual wall chart must be dynamic, with a time dimension showing the responses of each metabolite pool and flux to changes in environment (light, pathogen, cold, etc.) and developmental state. Finally, to truly qualify as systems biology, the descriptions need to exist in a computer model from which we can make specific testable hypotheses; for instance, that overexpression of a newly discovered tree fern cytochrome P-450 in specific apical cells would increase auxin biosynthesis in these cells and alter the morphology of an ornamental plant in a desirable way.

This level of sophistication demands a complete description of all pathways in at least one plant, the most obvious candidate being Arabidopsis. However, to be comprehensive and useful for the basic and applied biology communities, we will need to build upon this virtual metabolic network with information about pathways that only exist in specific species. Examples would include pathways of secondary metabolism, central metabolic processes that differ across the plant kingdom (C4 versus C3 metabolism, for instance), or storage of carbon in specialized organs or in chemical forms distinct from our canonical "virtual plant" (e.g. starch in tubers).

As ambitious as it seems, it will not be enough to have a complete understanding of the biogenesis of small molecules in plants and the changes in these pathways that occur during development and in re-

[1] This work was supported by the Max Planck Gesellschaft (to R.L.L. and A.R.F.) and by a Biotechnology and Biological Sciences Research Council David Phillips research fellowship (to L.J.S.).

[2] Present address: National Science Foundation, Plant Genome Research Program, 4201 Wilson Boulevard, Arlington, VA 22230.

* Corresponding author; e-mail robertllast@hotmail.com; fax 703–292–9062.

www.plantphysiol.org/cgi/doi/10.1104/pp.103.022004.

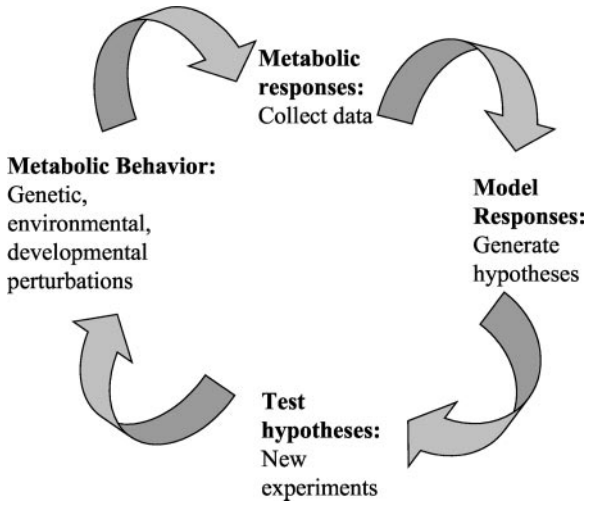

Figure 1. Model building and data accumulation are a cyclical process. A major goal of modeling metabolic responses is to predict which components exert greatest control over regulation of flux through the pathway and to hypothesize the results of altering these constituents. These are then tested experimentally, and the empirical results drive refinement and simplification of the model. Although this process of refinement never ends, the model becomes valuable as soon as it accurately predicts the behavior of the system in response to perturbation.

sponse to the environment. To create a model that is predictive, it will be critical to understand the enzymes of these pathways: when and where they accumulate in the cell, how their activity is modulated by specific covalent modifications, how their turnover is regulated, and how their tertiary structures and higher order interactions modulate their functions. For each of these levels of regulation, we will need comprehensive information about signal transduction pathways, transcriptional regulation, protein trafficking, and more. This means that we cannot do metabolomics in isolation: We need to analyze our data in concert with other large and complex sets of information such as proteomic, global mRNA expression and large-scale structural biological data.

Anaylsis of such complex datasets is a challenging task that requires the adoption of standardized data nomenclatures and formats and the application of sophisticated mathematical and statistical approaches to mine meaningful information from a sea of numbers. Daunting as the task is, a proper integration of transcriptomic, proteomic, and metabolomic datasets will provide the opportunity to analyze metabolic networks at an unprecedented level of detail and will afford a new system-wide level of understanding of the regulation of plant metabolism. Indeed, this approach has already been successful in providing new information about several areas of plant metabolism including specific branches of secondary metabolism (Suzuki et al., 2002), the response to nitrate (Matt et al., 2001), and the response to diurnal changes (Masclaux-Daubresse et al., 2002).

ANALYSIS OF THE METABOLOME: METHODS IN SMALL MOLECULE ANALYSIS

The lofty goal of a predictive model for all of plant metabolism requires accurate and reproducible measurement of all metabolites in the plant. A major impediment to this goal is that there is a tradeoff intrinsic to current analytical technologies: We can either accurately measure a relatively small number of molecules in an assay, or qualitatively analyze larger numbers of compounds. Thus, as a rule, with increasing metabolite coverage, there is a loss of accuracy in detection and quantification, which is schematically represented in Fig. 2.

In the last few years, plant metabolite analysis has shifted from specific enzyme assays and non-coupled chromatographic separations that provide information on single compounds or on mixtures of low complexity toward methods offering both high accuracy and sensitivity in highly complex mixtures of compounds (Fig. 2). The approaches of choice for large-scale metabolome analysis are predominantly reliant on coupled mass spectrometric methodologies. For example, gas chromatography-mass spectroscopy (GC-MS) technologies allow the detection, identification, and robust quantification of a few hundred metabolites within a single extract (Fiehn et al., 2000; Roessner et al., 2001). Furthermore, liquid chromatography-MS (LC-MS)-based methodologies have recently been established that allow reproducible determination of several important classes of

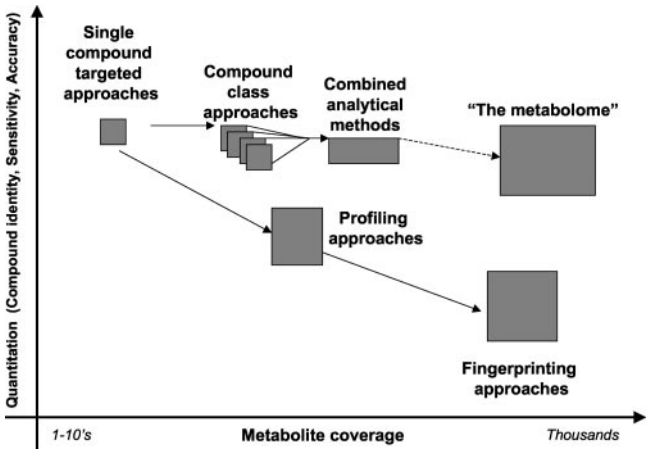

Figure 2. Schematic representation of the different approaches to metabolite measurements. The goal of metabolomics is to accurately identify and quantify every small molecule in the cell (top right sector of the graph). Current methodologies allow very accurate and sometimes high-throughput assays for individual molecules or classes of molecules (top left). Running these assays in parallel ("combined analytical methods") can increase the "metabolome-space" that is covered at the expense of throughput or operating costs. Fingerprinting and profiling approaches such as TOF-MS or FT-MS without prior separation or subsequent fragmentation and analysis allow large numbers of often anonymous compounds to be measured. These methods have relatively poor analytical accuracy, precision, and sensitivity compared with compound class approaches.

secondary metabolites including alkaloids, flavonoids, glucosinolates, isoprenes, oxylipins, phenylpropanoids, pigments, and saponins (for review, see Fiehn, 2002; Fernie, 2003). However, no currently available LC-based method allows the accurate determination of all of these compounds from a single extract, partly due to differences in solubility and stability, and the enormous concentration range of small molecules. LC-MS also suffers from so-called matrix effects, where the ability of the spectrometer to detect and accurately quantify any given ion is influenced by the presence of other ions in the mixture.

Thus, despite the developments described above, the fact remains that the vast majority of metabolites are not measured by approaches currently in use because the sum total of these methodologies covers hundreds rather than the thousands of metabolites estimated to be present in plant cells (De Luca and St Pierre, 2000). Small-molecule analysis therefore currently represents the least comprehensive of the various levels of system analysis, with the current proportional coverage of the profiling technologies decreasing following the order mRNA > protein > metabolite. That said, as is the case in the field of proteomics, progress in analytical chemistry is driving rapid advances in metabolomics.

Further analytical tools that are currently being developed for the multiparallel analysis of metabolites include time-of-flight-MS (TOF-MS), Fourier transformation-MS (FT-MS), FT-infrared spectroscopy, and NMR spectroscopy, with each technique having comparative advantages and disadvantages and selection of a given tool being largely driven by the biological question at hand (for a detailed description, see Sumner et al., 2003). Generally speaking, the choice of method must be determined by evaluating the metabolite coverage it affords with its relative accuracy of compound identification and quantification (summarized in Fig. 2).

Because these have been extensively reviewed elsewhere (Fiehn, 2002; Fernie, 2003; Sumner et al., 2003), we will merely discuss major differences in the technologies. Of these tools, TOF-MS serves a very similar purpose to GC-MS with the exception that TOF mass analyzers are preferred to the quadrupole mass analyzers of standard GC-MS machines, because they can provide a combination of high-mass accuracy and extended range of metabolites detected (Hall et al., 2002). Of course, GC-MS is only useful for volatile compounds or molecules that can be rendered volatile by derivitization, whereas LC-MS can be directly adapted to a wider array of molecules. Nevertheless, GC-MS still has major advantages over LC-MS methods. First, it is easier to achieve complete ionization of molecules in a complex mixture from samples coming off of a gas chromatograph, minimizing interference between molecules in the MS detector. Second, compound identification is greatly facilitated by availability of extensive databases of fragmentation patterns. Finally, tools exist for high-throughput capture of GC-MS data with less time-consuming human intervention than is needed for LC-based methods.

A further MS technology warranting discussion is Fourier Transform Ion Cyclotron Mass Spectrometry (FT-IC-MS), wherein extracts are directly infused into MS instruments using soft ionization techniques to gain fingerprints of the molecular ions presents. This technique is wholly reliant on the presence of a mass analyzer of sufficient accuracy to allow the definitive empirical formulae for several hundred ions. As with all technologies, disadvantages exist in the use of FT-MS; the most problematic of these is the fact that the lack of chromatography renders the technique unable to differentiate between isomers because of their identical molecular masses. Therefore, metabolites with biologically distinct functions such as Glc and Fru cannot be distinguished from one another. Furthermore, given the limited amount of published data (Aharoni et al., 2002) and the lack of documentation of vigorous method validation, it is unclear how useful this method will be when used in isolation. However the coupling of this or any of the other technologies mentioned above with further fragmentation gives far greater accuracy of identification, but it comes with a heavy cost attached in terms of sample throughput.

The use of FT-infrared spectroscopy and NMR has promise but thus far has not been applied to the quantification of large numbers of compounds in a complex mixture. These methods are well known for identification and quantitation of known molecules, especially when pure or in relatively simple mixtures. They are more recently being tested for metabolic fingerprinting in which the presence and intensity of specific spectroscopic signatures are correlated with molecules of interest in the hope of identifying major changes in metabolite levels or in accurate determination of certain chemicals. NMR has the further advantage that it can be used noninvasively and can yield (albeit limited) spatial resolution of solute concentrations.

Even the parallel use of the existing techniques will only manage to capture a proportion of the metabolome. However, the speed of recent advances in separation sciences and analytical instrumentation is unprecedented. This combined with the opportunity to couple various separation and analytical machines together make it likely that quantum leaps will soon be made toward the goal of accurate measurement of thousands of molecules from complex biological samples.

RESOLUTION: CELLULAR AND SUBCELLULAR

Because nearly all plant tissues are composed of many cell types, the development and refinement of

techniques allowing metabolite analysis of the contents of single cells or subcellular compartments are of great importance. Single-cell sampling techniques now combine cell biological methodologies with physical sampling techniques, and the recent development of high-precision laser capture and catapulting methodologies (Eltoum et al., 2002) will most likely produce rapid advances in this area. In addition to these methods, metabolite-linked bioluminescence assays have been established for visualizing metabolite contents in tissue slices, albeit for a limited number of metabolites at low sensitivity (Borisjuk et al., 1998). Thus, at the single-cell level, resolution of metabolite levels is approaching that at the protein and RNA levels.

Obtaining metabolite information at subcellular levels remains particularly problematic. Unlike proteins with their targeting sequences, the location of a metabolite cannot be inferred from their structures. Furthermore, most metabolic intermediates turn over too rapidly to allow measurement after the fractionation procedures regularly used in protein analysis.

Despite these problems, several methods have been developed to obtain subcellular information on metabolite levels in intact plants—two of which appear particularly promising. The first of these, nonaqueous fractionation of lyophilized material, involves the separation of cellular compartments on an organic density gradient and then uses simultaneous equations to estimate metabolite concentrations with respect to marker enzymes of various organelles in the cell (Gerhardt et al., 1983). This approach allows the discrimination of the vacuole, plastid, and cytosol and has been used for a wide variety of plant tissues (see Farre et al., 2001, and refs. therein). However, the inability to distinguish the mitochondrion from the cytosol and the apoplast from the vacuole and the fact that this technique requires in excess of 3 g of tissue limit its utility. The second promising method relies on the generation of specific chimeric proteins created by the fusion of periplasmic binding proteins to green fluorescent proteins that are differentially fluorescent after binding of a given metabolite (Fehr et al., 2002). Although when taken together, these methods offer great potential, it will be challenging to modify these methods for high-throughput analyses.

THE DYNAMICS OF PLANT METABOLISM: MEASURING FLUXES

Although measurements of steady-state levels of metabolites give a useful snapshot of the metabolic network at a given moment in time, the true behavior of plant metabolism can only be gained by direct measurement of metabolic fluxes. Ideally, we need to be able to observe the dynamics of metabolism as it happens. Despite the fundamental importance of metabolic flux, measurement of flux remains a poor cousin to the 'omic relations of transcripts, proteins, and even metabolites. The basic approach of following the distribution of a labeled precursor through metabolism has been used for many years to estimate fluxes through the major pathways of primary metabolism (for a recent example, see Lytovchenko et al., 2002), but relatively crude methods of determining label distribution have severely limited the number of fluxes that can be measured.

The recent emergence of more sophisticated approaches, based on a combination of steady-state stable-isotope labeling and NMR or MS-based detection systems, may revolutionize our ability to measure flux. This approach allows the determination of positional labeling of a range of metabolites. The position of labeling within the carbon skeleton of end products can allow retrospective evaluation of the metabolic route by which they were formed This technique has already been successfully used to investigate the metabolism of lipids and proteins (Schwender and Ohlrogge, 2002) and has the potential to measure a much greater range of specific fluxes, including those for pathways that operate in parallel in different subcellular compartments. Nevertheless, the interpretation of stable-isotope-labeling experiments is not a trivial task, and mathematical frameworks must be established that allow the calculation of pathway fluxes from fractional enrichments of isotopic label in each intermediate. In this respect, the plant community is still some way off from the comprehensive flux frameworks that have been developed for microbial organisms; in part, this can be put down to the increased complexity of the plant metabolic network conferred by multiple cell types and extensive subcellular compartmentation.

FROM NUMBERS TO EQUATIONS: MODELING PLANT METABOLISM

Science in its purest form is the reduction of the behavior of a system to a set of mathematical rules that define it. Ideally, such rules should be sufficiently simple that they themselves provide new insight into system behavior. Constructing models of metabolism has the potential to achieve this insight and is also an excellent way of succinctly representing large data sets. As such, models can be viewed as the mathematical engine of the virtual wall chart. In addition, models also allow the characteristics of the system to be systematically tested in a way that is too time consuming to achieve experimentally and are thus an important source of new hypotheses. Models are also extremely useful in pinpointing those parameters that are the most important in determining a particular function, parameters that need to be the focus of greater experimental precision.

Historically, metabolic models have concentrated on pathway flux due to the availability of theoretical approaches that can be used to derive flux control

structures. Most of the models of plant metabolism to-date are based around the theorems of metabolic control analysis and derive the control-distribution of pathway flux from the kinetic constants of the enzymes involved (Pettersson and Ryde-Pettersson, 1988; Thomas et al., 1997). Although these kinetic models have been of some use in understanding the control of plant metabolic pathways, particularly the photosynthetic Calvin cycle, their reliance on difficult-to-acquire experimental information means that they have not been used more widely in metabolism and remain somewhat limited in scope. Stoichiometric analysis (Schuster et al., 2000) is an interesting alternative approach that has gained much attention recently (Cornish-Bowden and Cardenas, 2002). The basis of this approach is to define elementary flux modes—non-decomposable subnetworks that account for every possible flux within the network. This allows one to mathematically define and describe all metabolic routes that are both stoichiometrically and thermodynamically feasible and is an extremely useful tool for the definition of network structure (Schuster et al., 2000).

BRINGING IT ALL TOGETHER

Although flux models provide a description of metabolic dynamics, for these models to enable predictive metabolic engineering, they must also incorporate experimental information about the changes in gene expression and protein abundances that underlie these fluxes (Fig. 1). Ultimately, one would wish to be able to predict exactly which genes should be altered in expression to generate specific changes in particular fluxes. However, bringing together such disparate data sets presents a considerable challenge. A particular problem is the issue of ontology. Put simply, there are too many different ways of naming, describing, and conceptualizing biological elements, leading to description of the same element in several different ways. Efforts are being made to establish a standard vocabulary that can be applied across all levels of the system hierarchy within an organism and between different organisms (Ashburner et al., 2000). Such ontologies not only furnish standardized descriptors for genes and proteins, but also provide an interactive hierarchical map of the entire system from the level of cellular component, to molecular function, to biological process.

An additional problem when integrating different types of data is that the data set becomes multidimensional. Development of new mathematical approaches that can cope with increased dimensionality will be crucial if data are to be interpreted at a systems level. Machine learning (Kell et al., 2001) provides a promising approach to this problem. This method uses a variety of different algorithms to provide simple rules that map back onto the measured variables and thereby provide explanations about the behavior of the system (Weiss and Kulikowski, 1991; Back et al., 1997). An alternative approach is to establish relevance networks in which mutual information is used to link data points that follow the same pattern of change (Butte and Kohane, 2000).

Although such methods are extremely powerful, they are not readily approachable by those without specialist knowledge. If such techniques are to become a routine part of the analysis of genomic data sets, then software interfaces will be required that allow the user to input and define their data while the algorithms are applied automatically and the results displayed in a readily interpretable format. The display of multidimensional data brings its own challenges, and there are a number of emerging solutions to this problem, examples of which include data touring (http://www.ggobi.org) and the use of glyphs—visual objects onto which many different data attributes are mapped using different visual attributes such as size and color (Pastizzo et al., 2002).

Despite these difficulties in attempting to integrate global data sets, the rewards for successfully doing do so are clear, and several recent papers focusing on subsets of plant metabolic pathways have provided data sets comprising transcripts, enzyme activities, and metabolites. These include studies of various aspects of carbon metabolism including the responses to nitrate (Matt et al., 2001) and to diurnal changes (Masclaux-Daubresse et al., 2002) and studies of individual branches of secondary metabolism such as triterpene saponin biosynthesis (Suzuki et al., 2002). Even though these studies are restricted in their coverage of metabolic space, they nicely demonstrate the power that combining genomic techniques brings to network understanding and pathway elucidation. Expansion of such approaches, even using the tools already at hand, has vast potential in aiding the understanding of complex change underlying diverse biological patterns such as developmental processes or circadian rhythms.

CONCLUSIONS

The ultimate aim of metabolic systems biology is to use the comprehensive experimental data sets describing changes in transcripts, proteins, metabolites, and flux to generate a complete mathematical description of the metabolism of a model plant species such as Arabidopsis. It is envisaged that such a model would allow a truly predictive engineering of plant metabolism. This is an ambitious aim that will require a sustained commitment of resources and unprecedented technological developments to be achieved. Perhaps most important of all, integration of approaches and data sets will be needed at a variety of levels from sociological to computational. As one example of many, we must bring together scientists working in traditionally disparate disciplines such as developmental genetics and biochem-

istry to understand how cell type-specific biochemical pathways are regulated, and use this information to engineer the high-level synthesis of medicinal, food, and industrial products. To take full advantage of the increasing numbers of high-quality metabolite, proteomic, and mRNA data sets will require a completely new way of thinking about and performing biological research. The ability to grow food more efficiently and to put crop plants to brand new uses should be incentive enough for us to rapidly embrace systems biology.

ACKNOWLEDGMENTS

We thank Drs. Joachim Kopka and John Shanklin for constructive comments on the manuscript.

Received February 12, 2003; returned for revision March 7, 2003; accepted March 7, 2003.

LITERATURE CITED

Aharoni A, de Vos CH, Verhoeven HA, Maliepaard CA, Kruppa G, Bino RJ, Goodenowe DB (2002). OMICS **6:** 217–234

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT et al. (2000) Nat Genet **25:** 25–29

Back T, Fogel D, Michalewicz Z (1997) Handbook of Evolutionary Computation. Oxford University Press, Oxford

Borisjuk L, Walenta S, Weber H, Mueller-Klieser W, Wobus U (1998) Plant J **15:** 583–591

Butte A, Kohane I (2000) Pac Symp Biocomput 418–429

Cornish-Bowden A, Cardenas L (2002) Nature **420:** 129–130

De Luca V, St Pierre B (2000) Trends Plant Sci **5:** 168–173

Eltoum IA, Siegal GP, Frost AR (2002) Adv Anat Pathol **9:** 316–322

Farre EM, Tiessen A, Roessner U, Geigenberger P, Trethewey RN, Willmitzer L (2001) Plant Physiol **127:** 685–700

Fehr M, Frommer WB, Lalonde S (2002) Proc Natl Acad Sci USA **99:** 9846–9851

Fernie AR (2003) Funct Plant Biol **30:** 111–120

Fiehn O, Kopka J, Dörmann P, Altmann T, Trethewey RN, Willmitzer L (2000) Nat Biotechnol **18:** 1157–1161

Fiehn O (2002) Plant Mol Biol **48:** 155–171

Gerhardt R, Stitt M, Heldt HW (1983) Physiol Chem **364:** 1130–1141

Hall R, Beale M, Fiehn O, Hardy N, Sumner L, Bino R (2002) Plant Cell **14:** 1437–1440

Kell DB, Darby RM, Draper J (2001) Plant Physiol **126:** 943–951

Lytovchenko A, Sweetlove L, Pauly M, Fernie AR (2002) Planta **215:** 1013–1021

Masclaux-Daubresse C, Valadier MH, Carrayol E, Reisdorf-Cren M, Hirel B (2002) Plant Cell Environ **25:** 1451–1462

Matt P, Geiger M, Walch-Liu P, Engels C, Krapp A, Stitt M (2001) Plant Cell Environ **24:** 177–190

Pastizzo MJ, Erbacher RF, Feldman LB (2002) Behav Res Methods Instrum Comput **34:** 158–162

Pettersson G, Ryde-Pettersson U (1988) Eur J Biochem **175:** 661–672

Roessner U, Luedemann A, Brust D, Fiehn O, Linke T, Willmitzer L, Fernie AR (2001) Plant Cell **13:** 11–29

Schuster S, Fell DA, Dandekar T (2000) Nat Biotechnol **18:** 326–332

Schwender J, Ohlrogge JB (2002) Plant Physiol **130:** 347–361

Sumner L, Mendes P, Dixon RA (2003) Phytochemistry **62:** 817–836

Suzuki H, Achnine L, Xu R, Matsuda SPT, Dixon RA (2002) Plant J **32:** 1033–1048

Thomas S, Mooney PJ, Burrell MM, Fell DA (1997) Biochem J **322:** 119–127

Weiss S, Kulikowski C (1991) Computer Systems That Learn: Classification and Prediction Methods from Statistics, Neural Networks, Machine Learning, and Expert Systems. Morgan Kaufmann Publishers, San Mateo, CA