AraGEM, a Genome-Scale Reconstruction of the Primary Metabolic Network in Arabidopsis[1][W]

Cristiana Gomes de Oliveira Dal'Molin, Lake-Ee Quek, Robin William Palfreyman, Stevens Michael Brumbley, and Lars Keld Nielsen*

Australian Institute for Bioengineering and Nanotechnology, University of Queensland, Brisbane, Queensland 4072, Australia

Genome-scale metabolic network models have been successfully used to describe metabolism in a variety of microbial organisms as well as specific mammalian cell types and organelles. This systems-based framework enables the exploration of global phenotypic effects of gene knockouts, gene insertion, and up-regulation of gene expression. We have developed a genome-scale metabolic network model (AraGEM) covering primary metabolism for a compartmentalized plant cell based on the Arabidopsis (*Arabidopsis thaliana*) genome. AraGEM is a comprehensive literature-based, genome-scale metabolic reconstruction that accounts for the functions of 1,419 unique open reading frames, 1,748 metabolites, 5,253 gene-enzyme reaction-association entries, and 1,567 unique reactions compartmentalized into the cytoplasm, mitochondrion, plastid, peroxisome, and vacuole. The curation process identified 75 essential reactions with respective enzyme associations not assigned to any particular gene in the Kyoto Encyclopedia of Genes and Genomes or AraCyc. With the addition of these reactions, AraGEM describes a functional primary metabolism of Arabidopsis. The reconstructed network was transformed into an in silico metabolic flux model of plant metabolism and validated through the simulation of plant metabolic functions inferred from the literature. Using efficient resource utilization as the optimality criterion, AraGEM predicted the classical photorespiratory cycle as well as known key differences between redox metabolism in photosynthetic and nonphotosynthetic plant cells. AraGEM is a viable framework for in silico functional analysis and can be used to derive new, nontrivial hypotheses for exploring plant metabolism.

The past two decades have seen impressive progress in the mapping of plant genes to metabolic function, culminating in the complete sequencing and partial annotation of Arabidopsis (*Arabidopsis thaliana*; Dennis and Surridge, 2000), rice (*Oryza sativa*; Goff, 2005), and sorghum (*Sorghum bicolor*; Paterson et al., 2009), with the maize (*Zea mays*) genome under way (Maize Genome Sequencing Project, 2009). Simultaneously, tools for accurately manipulating gene expression in plants have developed rapidly. However, attempts to use these tools to engineer plant metabolism have met with limited success due to the complexity of plant metabolism. Genetic manipulations rarely cause the predicted effects, and new rate-limiting steps prevent the accumulation of some desired compounds (Sweetlove et al., 2003; Gutierrez et al., 2005). In a bid to improve our understanding of plant metabolism and thereby the success rate of plant metabolic engineering, a systems-based framework to study plant metabolism is needed (DellaPenna, 2001; Gutierrez et al., 2005). Systems biology involves an iterative process of experimentation, data integration, modeling, and generation of hypotheses (Kitano, 2000, 2002a, 2002b). "Omics-based analysis," in the form of transcriptomics (Yonekura-Sakakibara et al., 2008; Minic et al., 2009; van Dijk et al., 2009), proteomics (Thiellement, 1999; Dai et al., 2007; Cui et al., 2008), and metabolomics (Hall, 2006; Morgenthal et al., 2006), is well advanced in plant biology. In contrast, data mining beyond blind statistical analysis, modeling, and hypothesis generation have suffered from a lack of a mathematical modeling framework to handle the complexity. The synergistic use of model and omics data is necessary to understand the complex relationships and interactions among the various parts of the system. Although the ultimate goal is the development of dynamic models for the complete simulation of cellular systems (Tomita et al., 1999; Sugimoto et al., 2005), the success of such approaches has been severely hampered by the current lack of kinetic information on the dynamics and regulation of metabolic reactions.

The genome-scale modeling approach addresses the problem from the opposite direction, building on well-established network features. The foundation is a systemic (two-dimensional) annotation that lists all components (in all their states) of a system in one dimension and all the links connecting the components in the second dimension. Thus, the biochemical network contains all the metabolites in one dimension and all the reactions connecting the metabolites in the

[1] This work was supported by the Cooperative Research Centre for Sugar Industry Innovation through Biotechnology.

* Corresponding author; e-mail lars.nielsen@uq.edu.au.

The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantphysiol.org) is: Lars Keld Nielsen (lars.nielsen@uq.edu.au).

[W] The online version of this article contains Web-only data.

www.plantphysiol.org/cgi/doi/10.1104/pp.109.148817

second dimension. This approach is readily extended to genes and proteins to describe transcription and translation as well as interactions and regulation. At present, however, genome-scale understanding is largely constrained to biochemical networks. Systemic annotations have been termed BIGG (for biochemically, genetically, and genomically structured) databases, and BIGG databases have been reconstructed for several organisms, including *Escherichia coli* (Feist et al., 2007) *Saccharomyces cerevisiae* (Duarte et al., 2004), mouse (*Mus musculus*; Sheikh et al., 2005; Quek and Nielsen, 2008), and human (Duarte et al., 2007).

A BIGG database defines the topology of the biological network. Network topology can be used in mining omics data (Cakir et al., 2006; Oliveira et al., 2008). Topological properties on their own, however, provide limited potential for predicting actual functional states of the network until they are combined with constraints reflecting the physicochemical nature of the system. Generally, systems are studied in metabolic steady state (i.e. $Sv = 0$), where S is the stoichiometric matrix with rows representing metabolites and columns representing individual reactions, and v is the flux for each reaction. The fluxes are subject to maximum capacity constraints ($v_{\mathrm{min}} \leq v \leq v_{\mathrm{max}}$), and some reactions are irreversible (i.e. the corresponding $v_{\mathrm{min}} = 0$). Furthermore, the model may be constrained by regulation (e.g. two reactions may be mutually exclusive) and by experimental data.

The BIGG database together with the constraints constitutes the genome scale model (GEM). This model can be used with a broad range of constraint-based analysis tools to predict, for example, optimal growth rate, yields on different substrates, and the effects of gene deletions (Varma and Palsson, 1994; Bonarius et al., 1997; Fong et al., 2003; Forster et al., 2003; Duarte et al., 2004; Price et al., 2004). In silico models have enabled hypothesis-driven biology, including the study of adaptive evolution in microbial systems (Fong et al., 2003, 2005) and the discovery of new metabolic roles of individual genes (Reed et al., 2006).

GEMs may greatly facilitate the study of the complex metabolism in plants, with its duplication of many pathways in multiple organelles and greatly varying metabolic objectives between different tissues. A partial annotation was recently published for the green alga *Chlamydomonas reinhardtii* (Boyle and Morgan, 2009). This annotation covers central and intermediary metabolism and consists of 484 reactions in cytosol, chloroplast, and mitochondria linked to protein identifier (ID). For plants, an Arabidopsis GEM was recently published (Poolman et al., 2009). Although the reconstructed model is capable of producing biomass components (amino acids, nucleotides, lipid, starch, and cellulose) in the proportion observed experimentally in a heterotrophic suspension culture, it does not consider the localization of reactions in organelles and cannot describe autotrophic metabolism. Moreover, no two-dimensional annotation was provided to link functions to genes.

In this study, a genome-scale model was developed for Arabidopsis. It represents, to our knowledge, the first attempt to characterize the metabolism of a compartmentalized photosynthetic plant cell at the genome scale and represents a structured compilation of cell components and interactions, which enables systematic investigation of metabolic properties and interrogation of available omics data sets for plants. The genome-scale model has been validated against a number of classical physiological scenarios: photosynthesis, photorespiration, and respiration.

RESULTS AND DISCUSSION

Large-Scale Metabolic Reconstruction

A gene-centric organization of metabolic information was adopted, in which each known metabolic gene is mapped to one or several reactions. The core of the Arabidopsis genome-scale model (AraGEM version 1.0) was reconstructed from the Arabidopsis gene

Table I. *Online resources for the reconstruction of the metabolic network of Arabidopsis*

Database	Link
Genome database	
The Arabidopsis Information Resource (TAIR version 9)	http://www.arabidopsis.org/index.jsp
Pathway databases	
Kyoto Encyclopedia of Genes and Genomes (KEGG)	http://www.genome.jp/kegg/pathway.html
AraCyc version 5.0	http://www.arabidopsis.org:1555/ARA/
Biocyc	http://biocyc.org/
ExPASy Biochemical Pathways	http://www.expasy.ch/cgi-bin/search-biochem-index
Reactome (Arabidopsis)	http://www.reactome.org
Enzymes databases	
ExPASy Enzyme Database	www.expasy.org
Brenda	http://www.brenda-enzymes.org
Enzyme/protein localization and others databases	
AraPerox (Arabidopsis Protein from Plant Peroxisomes)	http://www.araperox.uni-goettingen.de/
SUBA (Arabidopsis Subcellular Database)	http://www.plantenergy.uwa.edu.au/applications/suba2/index.php
PPDB (Plant Proteome Database)	http://ppdb.tc.cornell.edu/default.aspx
UniproKB/SwissProt	http://ca.expasy.org/sprot/relnotes/relstat.html

Table II. *Network characteristics of the reconstructed metabolic network of Arabidopsis*

Metabolic Characteristics	Total
Gene reaction-association entries	5,253
ORFs (unique)	1,419
Metabolites	1,748
Unique reactions	1,567
Cytosolic reactions	1,265
Mitochondrial reactions	60
Plastidic reactions	159
Peroxisomal reactions	98
Modified reactions	36
Biomass drains and transporters	148
Biomass drains	47
Transporters (intercellular)	18
Transporters (interorganellar)	83
Gaps (unique reaction IDs) filled by manual curation	75
Singleton metabolites	446

and reaction database publicly available from the Kyoto Encyclopedia of Genes and Genomes (KEGG; release 49.0, January 1, 2009; Kanehisa et al., 2008). The reconstruction process was automated and used the same procedure previously applied to the GEM of mouse (Quek and Nielsen, 2008). The reconstruction retained all reaction attributes from KEGG, including unique reaction and compound IDs and reaction reversibilities.

KEGG does not capture the compartmentalization of metabolism in eukaryotes. The plant cell model was compartmentalized into cytosol, mitochondrion, plastid, peroxisome, and vacuole based on literature information and the available databases: The Arabidopsis Information Resource and AraPerox (a database of putative Arabidopsis proteins from plant peroxisomes; Reumann et al., 2004; Cui et al., 2008; Table I). This process was performed manually based on known organelle functions and localization calls made for the many isozymes based on the best available data.

Characteristics of the Reconstructed Network

The reconstructed plant metabolic network contains 5,253 gene-reaction association entries (Table II). A total of 1,567 unique reactions and 1,748 metabolites are part of the reconstructed network. The active scope of AraGEM includes glycolysis (plastidic and cytosolic), pentose-P pathway (PPP; plastidic and cytosolic), tricarboxylic acid (TCA) cycle, light and dark reactions (Calvin cycle), NADH/ NADPH redox shuttle between the subcellular compartments, fatty acid synthesis, β-oxidation, glyoxylate cycle, and photorespiratory cycle.

Thirty-six unique KEGG reactions were modified to give a proper stoichiometric coefficient and/or consistent nomenclature and/or reversibility constraints. Modified reactions from KEGG have been marked by adding an "N" (new identity) to the KEGG reaction ID (e.g. R01175 became R01175N; Supplemental Table S1).

A total of 148 biomass drains and transporters were introduced in the model. Forty-seven biomass drain equations describe the accumulation of carbohydrates, amino acids, fatty acid, cellulose, and hemicellulose, representing the major biomass drain for a plant cell (Table III). At present, fatty acid biosynthesis is limited to palmitic acid biosynthesis in plastids. Fatty acids derived from cytosolic extension and desaturation reactions are not included, nor is the use or production of lipid alcohols in membrane phospholipids described. Hemicellulose production is currently described as a Xyl drain only, ignoring other sugars used in hemicellulose biosynthesis. The biosynthetic pathways of a limited number of vitamins and cofactors have been curated to date. Eighteen intercellular exchange reactions (cytoplasm-extracellular) have been included to describe the uptake/secretion of light (photons), inorganic compounds (CO_2, H_2O, O_2, NO_3, NH_3, H_2S, SO_4^{2-}, PO_4^{3-}), translocation of sugars (Suc, Glc, Fru, and maltose), and amino acids (Gln, Glu, Asp, Ala, and Ser). Together with biomass drains, the intercellular exchangers define the broad physiological domain of the model (i.e. the curated aspects of primary C_3 plant metabolism captured by AraGEM). A total of 83 interorganelle transporters were introduced in the model to achieve the desired functionality (for a complete list of biomass drains and transporters, see Supplemental Table S2). Most of the transporters have been assigned to putative open reading frames (ORFs).

Flux balance analysis revealed 75 step reactions (unique reaction IDs) with essential metabolic functions that were not assigned to any particular gene, three of which are autocatalytic reactions (Table IV; for a complete list, see Supplemental Table S3). These were involved in important metabolic functions, like pyruvate metabolism, starch and Suc metabolism, lignin biosynthesis, fatty acids biosynthesis, carbon fixation, vitamin biosynthesis, and biosynthesis/metabolism of some amino acids. Apart from the three autocatalytic reactions, we found KEGG reaction ID association for all metabolic reactions added to fill the

Table III. *List of biomass components*

Components	Major Drains
Carbohydrates and sugars	Starch, Suc, Fru, Glc, maltose
Cell wall	Lignin (4-coumaryl alcohol, coniferyl alcohol, sinapyl alcohol), cellulose, hemicellulose (Xyl)
Amino acids	Ala, Arg, Asp, Asn, Cys, Lys, Leu, Ile, Glu, Gln, His, Met, Phe, Pro, Ser, Tyr, Trp, Val
Nucleotides	ATP, GTP, CTP, UTP, dATP, dGTP, dCTP, dTTP
Fatty acids	C16:0 (palmitic acid)
Vitamins and cofactors	Biotin, CoA, riboflavin, folate, chlorophyll, nicotinamide, thiamine, ubiquinone

Table IV. *Metabolic reactions not assigned to any particular gene and with essential metabolic function*

No. of Gaps (75)	Metabolic Function (Core Metabolism)
24	Starch and Suc metabolism, lignin biosynthesis, glycerophospholipid metabolism, fatty acid biosynthesis, carbon fixation and pyruvate metabolism, citrate cycle, lipopolysaccharide biosynthesis; *N*-glycan biosynthesis; vitamins
23	Metabolism of amino acids (Ala and Asp metabolism; Gly, Ser, and Thr metabolism; Met metabolism; Phe, Tyr, and Trp biosynthesis; Lys biosynthesis and His metabolism; Val, Leu, and Ile biosynthesis; metabolism of amino groups)
25	Purine metabolism
3	Autocatalytic reactions

gaps in AraGEM; however, they were not associated with the Arabidopsis genome.

The final model has 446 singleton or dead-end metabolites (i.e. internal metabolites only used in a single reaction; Supplemental Table S4). A total of 512 reactions are linked to the use or production of dead-end metabolites. These reactions will by definition have zero flux in flux balance analysis. There are several causes of singletons. Most are linked to vitamins, cofactors, and secondary metabolites not currently included in the model (i.e. not described by biomass drains or intercellular transporters). Some result from true gaps in the network, where one or more essential reactions have not yet been assigned. Finally, some result from KEGG's automatic annotation, in which every reaction known to be catalyzed by a particular EC enzyme in some organism will be included by default, whether or not other reactions required for functionality of the particular reaction are present. Continued curation efforts will focus on resolving these singletons.

In its current form, the model has 183 degrees of freedom. Of these, 47 are associated with the production of biomass components and reduce to a single degree of freedom (growth rate) when a fixed biomass composition is assumed. Another 18 degrees of freedom are associated with intercellular transport. The remaining 118 degrees of freedom represent the maximum cellular scope for using alternative pathways to achieve identical outcomes in terms of growth rate and net transport (e.g. the use of cytosolic or chloroplastic glycolysis). The real scope is further limited by irreversibility constraints as well as regulatory constraints.

AraGEM version 1.0 is available in System Biology Markup Language (SBML) format (Supplemental Data File S1).

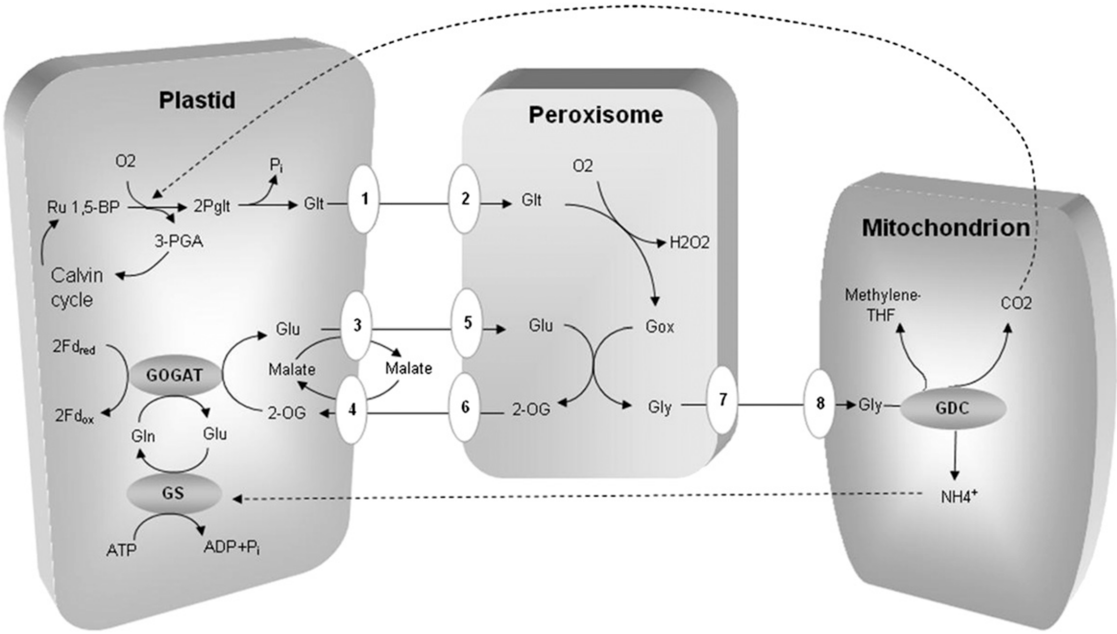

Figure 1. A simplified scheme of the current textbook photorespiratory cycle. Numbers are as follows: 1, plastidic glycolate transporter; 2, peroxisomal glycolate transporter; 3, plastidic Glu-malate translocator; 4, plastidic 2-oxoglutarate-malate translocator; 5, peroxisomal Glu transporter; 6, peroxisomal 2-oxoglutarate transporter; 7, peroxisomal Gly transporter; 8, mitochondrial Gly transporter. GDC, Gly decarboxylase; Glt, glycolate; Gox, glyoxylate; GS, Gln synthetase; methylene-THF, methylene-tetrahydrofolate; 2-OG, 2-oxoglutarate; 3-PGA, 3-phosphoglycerate; 2-Pglt, 2-phosphoglycolate; P_i, inorganic phosphate; Ru 1,5-BP, ribulose 1,5-bisP.

Table V. *Constraints imposed to represent each scenario*

Case 1 is photosynthesis, case 2 is photorespiration, and case 3 is respiration/nitrogen assimilation in nonphotosynthetic cells. −, Flux limited, constraint to zero; +, free flux, estimated through optimization.

Inputs, Outputs, and Constraints	Case 1	Case 2	Case 3
Carbon source: CO_2 uptake (free flux)	+	+	−
Carbon source: Suc uptake (free flux)	−	−	+
Photon uptake (free flux)	+	+	−
Rubisco; EC 4.1.1.39 (carboxylation-oxygenation ratio)	+1:0	+3:1	−
Fd-GOGAT; EC 1.4.7.1 (plastid)	+	+	−
NADH-GOGAT; EC 1.4.1.14 (plastid)	−	−	+
Optimization: minimize uptake of	Photons	Photons	Suc
Biomass rate (estimated and fixed)	Leaf	Leaf	Root

In Silico Fluxomics: Photosynthesis Versus Photorespiration

AraGEM is a generic plant cell model capable of representing both photosynthetic and nonphotosynthetic cell types. In order to evaluate the model, its ability to reproduce classical physiological scenarios of plant cell metabolism was explored. The first scenario explored was photorespiration.

Carbon fixation in photosynthetic cells is mediated by the carboxylation reaction of Rubisco: D-ribulose 1,5-bisP + CO_2 + H_2O = 2 3-phospho-D-glycerate. In the presence of oxygen, Rubisco also catalyzes an oxygenation reaction:

$$\text{D-ribulose 1,5-bisP} + O_2 = \text{3-phospho-D-glycerate} + \text{2-phosphoglycolate}$$

which reduces the energy efficiency of photosynthesis in C_3 plants (Wingler et al., 2000; Eckardt, 2005). The subsequent metabolism of glycolate produced by the oxygenation reaction is known as the photorespiratory cycle. The current "textbook" photorespiratory cycle is displayed in Figure 1.

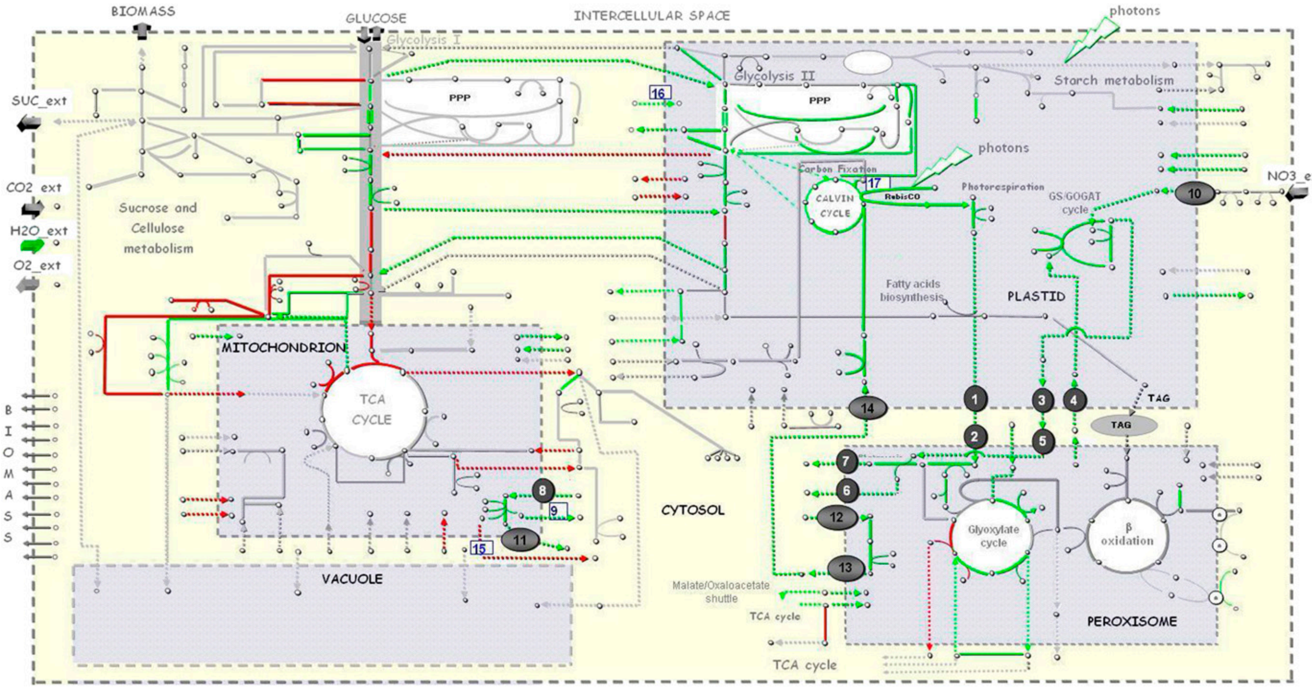

Figure 2. Flux map for the photorespiratory cycle. Comparison of fluxes between photorespiration and photosynthesis (no oxygenation of ribulose 1,5-bisP). In the diagram, solid lines represent fluxes and dashed lines represent the transporters. Green and red lines highlight fluxes that have increased and decreased, respectively, during photorespiration when compared with photosynthesis. Gray lines represent flux values that have not changed significantly. Steps 1 to 17 represent the order of events that complete the photorespiratory cycle. Transporters are described in the Figure 1 legend. Boxes 9, 15, and 16 represent ammonium and CO_2 being transported by diffusion. Box 17 represents CO_2 being fixed in the Calvin cycle.

The effect of photorespiration was explored by comparing the optimum flux distribution predicted with pure photosynthesis (Table V, case 1) with the optimum flux distribution predicted with a flux ratio of 3:1 (Wise and Hoober, 2007) for the carboxylation and oxygenation reactions by Rubisco (Table V, case 2). The optimum flux distribution is here defined as the flux distribution that minimizes photon uptake for a fixed rate of biomass synthesis. The rate of biomass synthesis was chosen based on literature values (Poorter and Bergkotte, 1992; Niemann et al., 1995). Glutamate synthase (Fd-GOGAT, NADH-GOGAT) was constrained based on gene expression and enzyme activity studies (Table V; Coschigano et al., 1998; Trepp et al., 1999; Turano and Muhitch, 1999; Lancien et al., 2002).

The metabolic contrast between case 1 and case 2 is illustrated in Figure 2. The numbers 1 to 8 in Figure 2 refer back to the transporters proposed in the current textbook photorespiratory cycle depicted in Figure 1 (Linka and Weber, 2005). Steps 9 to 17 complete the photorespiratory cycle scheme and were not presented in Figure 1 for the sake of simplicity. As indicated by the green lines, AraGEM predicts that the optimal (i.e. most photon-efficient) way of handling photorespiration is indeed to use the classical cycle. Moreover, the

Table VI. *Up-regulation of key target enzymes during photorespiration and respiration relative to photosynthesis highlighted by AraGEM*

Metabolic Pathway	Enzymes (EC) Up-Regulated during Photorespiration[a]		Enzymes (EC) Up-Regulated during Respiration[a]	
Glycolysis (cytosolic and plastidic)	4.2.1.11	Phosphopyruvate hydratase	2.7.1.40	Pyruvate kinase
	5.3.1.1	Triose-P isomerase	4.2.1.11	Phosphopyruvate hydratase
	1.2.1.13	Glyceraldehyde-3-phosphate dehydrogenase (NADP$^+$)	1.2.1.12	G3PDH (NAD$^+$)
	4.1.2.13	Fru-bisP aldolase	5.4.2.1	Phosphoglycerate mutase
	2.7.2.3	Phosphoglycerate kinase	5.1.3.3	Aldose 1-epimerase
	5.3.1.9	Glc-6-P isomerase	2.7.1.1	Hexokinase
	3.1.3.11	Fru-bisP	2.7.1.11	6-Phosphofructokinase
Suc metabolism			3.2.1.21	β-Glucosidase
			2.4.1.14	Suc-P synthase
			3.2.1.20	α-Glucosidase;
			3.2.1.26	β-fructofuranosidase
			2.4.1.13	Suc synthase
			2.7.1.1	Hexokinase
			3.2.1.4	Cellulase
			2.7.1.4	Fructokinase
			3.2.1.26	β-Fructofuranosidase
			2.4.1.12	Cellulose synthase (UDP forming)
Carbon fixation	4.1.1.39	Rubisco		
	2.7.1.19	Phosphoribulokinase		
	4.1.1.31	PEP carboxylase		
	1.1.1.39	Malate dehydrogenase (decarboxylating)		
PPP (plastidic)	5.3.1.6	Rib-5-P isomerase	2.7.1.15	Ribokinase
	4.1.2.13	Fru-bisP aldolase	5.4.2.2	Phosphoglucomutase
	5.1.3.1	Ribulose-P 3-epimerase	5.1.3.1	Ribulose-P 3-epimerase
	2.2.1.1	Transketolase	5.3.1.9	Glc-6-P isomerase
	2.2.1.2	Transaldolase	1.1.1.49	Glc-6-P dehydrogenase
	5.3.1.9	Glc-6-P isomerase		
	3.1.3.11	Fru-bisP		
TCA cycle			1.1.1.37	Malate dehydrogenase
			2.3.3.1	Citrate (Si)-synthase
			1.2.4.2	Oxoglutarate dehydrogenase
			1.1.1.41	Isocitrate dehydrogenase (NAD$^+$)
			4.2.1.2	Fumarate hydratase
			6.2.1.5	Succinyl-CoA synthetase
			1.3.99.1	Succinate dehydrogenase
Glyoxylate cycle	1.1.1.37	Malate dehydrogenase	1.1.1.37	Malate dehydrogenase
	1.1.3.15	Glycolate oxidase		
	4.2.1.3	Aconitate hydratase		
	3.1.3.18	Phosphoglycolate phosphatase		
	1.1.1.29	Glycerate dehydrogenase		
	2.7.1.31	Glycerate kinase		
GS/GOGAT	1.18.1.2	Ferredoxin-NADP$^+$ reductase	1.4.1.14	Glutamate synthase (NADH)

[a] Relative flux to photosynthesis.

model predicts that photon uptake rate is increased by more than 40% during this process in order to maintain the same rate of leaf biomass biosynthesis. This number is consistent with the experimental estimation that photorespiration drains away 30% to 50% of the carbon fixed by the Calvin cycle (Pessarakli, 1996; Wise and Hoober, 2007).

Figure 2 represents one optimal solution. Flexibility in the network means that there are generally many solutions achieving identical optimal performance. For example, if it is assumed that both plastids and mitochondria have transporters for both Gln and Glu, the Glu-Gln shuttle (Linka and Weber, 2005) is an equally photon-optimal alternative to the diffusive ammonia transport assumed in the textbook model of photorespiration. In contrast, the less efficient Orn-citruline shuttle would only be predicted if ammonia diffusion were deliberately constrained and no mitochondrial Gln transporter were assumed to exist. More generally, the existence of multiple solutions with characteristic usage of particular enzymes can be used to generate testable hypotheses.

Apart from the use of the classical photorespiratory cycle, AraGEM also predicts system-level changes in the optimum flux distribution (Fig. 2; Table VI). For example, increased flux through cytosolic malic enzyme and phospho*enol*pyruvate (PEP) carboxylase is predicted during photorespiration. These enzymes provide plants with metabolic flexibility for the metabolism of PEP and pyruvate. Malic enzyme converts malate to pyruvate and makes it possible for plant mitochondria to oxidize both malate and citrate to CO_2 without involving pyruvate delivery from Glc (Wedding, 1989; Taiz and Zeiger, 2006). While these predictions still need to be experimentally validated, they do highlight the power of genome-scale models to generate new nontrivial hypotheses based on global network properties.

Respiration in Nonphotosynthetic Cells

AraGEM was also used to contrast metabolism in nonphotosynthetic cells (Table V, case 3) to that in photosynthetic cells (Table V, case 1). Under normal conditions, Suc is the main respiratory and growth substrate of higher plants (Dennis and Miernyk, 1982), and the optimal flux distribution for nonphotosynthetic cells was defined as the flux distribution that minimizes Suc uptake for a fixed rate of biomass synthesis. The energy derived from respiration is used for growth and maintenance processes and is higher for roots than for tops (Hansen and Jensen, 1977).

As expected, AraGEM predicted an up-regulation of pathways representing the classical respiratory cycle

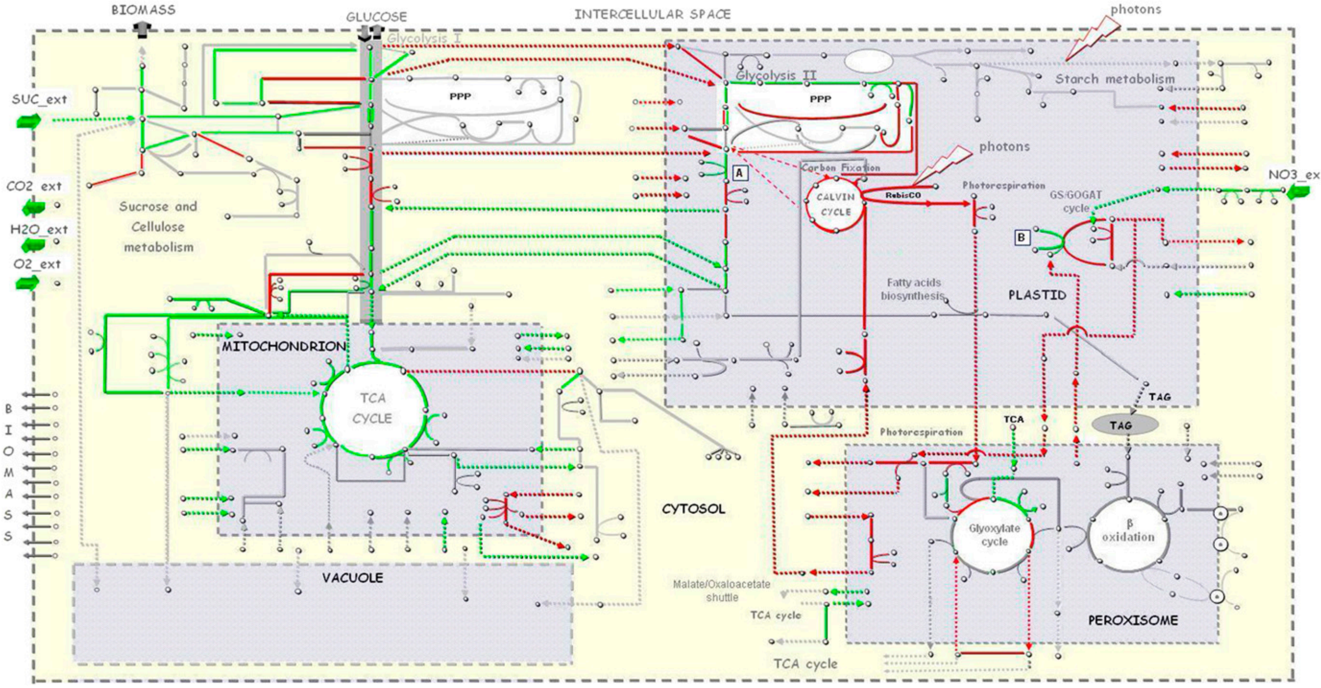

Figure 3. Flux map for respiration of a nonphotosynthetic plant cell. Flux comparison is shown between a nonphotosynthetic cell (respiration) and a photosynthetic cell (photosynthesis). Green and red lines represent increased and decreased flux, respectively, of a nonphotosynthetic cell compared with a photosynthetic cell. Gray lines represent flux values that have not changed significantly compared with photosynthesis. Boxes A and B represent the metabolic activity of some isoenzymes typically activated in nongreen tissues. Box A in plastidic glycolysis (glycolysis II) represents the plastid form of GAPDH [NAD (H)-dependent] activity. Box B represents the metabolic activity of NADH-GOGAT involved in ammonia assimilation in nongreen plastids.

(Suc metabolism, glycolysis, and the TCA cycle; Fig. 3). In addition, system-level changes in the flux distribution were also observed (Fig. 3; Table VI). Predictions in relation to redox metabolism can be used to illustrate how AraGEM facilitates the interrogation of the global properties of the plant metabolic network.

AraGEM predicts an increase in metabolic activity of key isoenzymes responsible for the supply of redox equivalents in plastids during the respiratory cycle. In silico analysis suggests an increase in metabolic activity of the plastid form of NAD-glyceraldehyde-3-phosphate dehydrogenase (GAPDH [GapCp]), an isoenzyme closely related to the cytosolic NAD-GAPDH (Fig. 3, boxes A and B). This prediction is consistent with the observation that this gene in normally expressed in nongreen tissues (Petersen et al., 2003).

AraGEM also predicts that the use of NADP(H) by GAPDH [and not NAD(H)] is energetically favorable for photosynthetic cells during the light period, indicating a preferential coenzyme use under that condition (Fig. 2). Consistent with this, it has been observed that the bispecific chloroplast form of GAPDH (EC 1.2.1.13) uses NADP(H) as a coenzyme under physiological conditions during the photoperiod (leaves; Wolosiuk and Buchanan, 1978; Backhausen et al., 1998).

Another hypothesis derived from the model is that, during light induction (in leaf), the reducing power generated by the photosynthetic transport chain is used in the reduction of nitrite to ammonia and in the Gln synthetase/GOGAT cycle (Fd-GOGAT; ferredoxin-NADP$^+$ reductase dependent). In the dark (no light induction or for nonphotosynthetic cells), on the other hand, the plastidic PPP supplies reducing power to the Gln synthetase/GOGAT cycle. Again, this prediction is in agreement with published observations (Abrol et al., 1983).

Optimality Criterion

The agreement between experimental observations and in silico predictions based on optimization suggests that plant cell metabolism could be regulated to achieve resource efficiency (photons in photosynthetic tissue and Suc in nonphotosynthetic tissue) in individual tissues. This, however, may only be true in a macroscopic sense (i.e. the metabolic flexibility encoded by the genome has evolved to ensure resource efficiency under the highly divergent conditions seen in different tissues [e.g. photosynthetic leaf tissue versus respiring root tissue]).

In order to establish if resource utilization is indeed optimized in individual tissues, it is necessary to compare predicted fluxes with actual fluxes. The current flux estimates are necessarily low estimates, since no value has been assigned to maintenance requirements. Moreover, tissues have been considered in isolation, whereas in reality there is an exchange of material between tissues and photosynthetic tissue must capture carbon for the whole plant.

The modeling framework is readily extended to scenarios where multiple tissues, each represented by a subset of the full model, exchange resources, and we are currently collating the growth parameters necessary to explore whole plant metabolism.

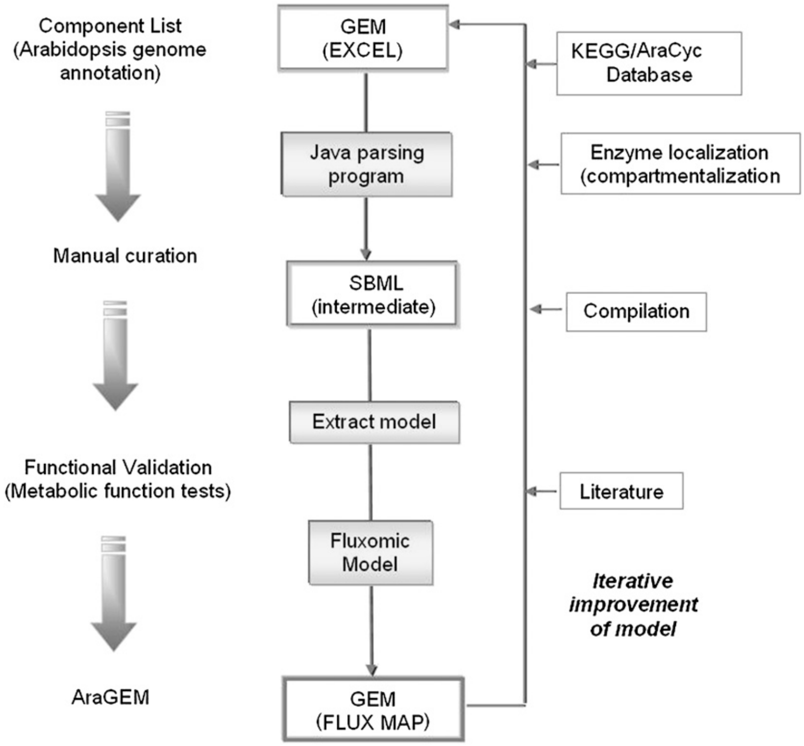

Figure 4. Information flow from a genome-scale model repository (Excel file) to a model document (SBML) to a flux map (Excel file). An initial list of components was compiled from the Arabidopsis genome, and the network was manually curated. The Java parsing program separates reactions into reactants and product stoichiometry and removes repeated reactions. Calculation output from MATLAB was used to update the flux map.

CONCLUSION

AraGEM is an extensively curated, compartmentalized, genome-scale model of plant cell primary metabolism. The reconstruction process led to the identification of 75 reactions essential for primary metabolism for which genes have yet to be identified. Continued curation efforts will focus on refining lipid metabolism, closing gaps in secondary metabolism, and resolving gene product targeting, where this has yet to be established.

The use of AraGEM for in silico flux predictions illustrates the potential of using genome-scale models to explore complex, compartmentalized networks and develop nontrivial hypotheses. The optimality criterion itself is one such hypothesis.

MATERIALS AND METHODS

Flux Balance Analysis

Flux balance analysis was used both in manual curation to confirm that each biomass component could be synthesized and in the subsequent model validation. AraGEM was compiled and curated in Excel (Microsoft) for ease of annotation and commenting (Fig. 4). From this gene-centric database, a two-dimensional reaction-centric SBML (www.sbml.org) database was generated using a Java script (Sun Microsystems). There is currently no specific element in SBML allocated to store the gene-protein reaction associations (e.g. splice variants, isozymes, protein complex). Instead, these were added as notes to the reaction elements.

The stoichiometric matrix, *S*, as well as reversibility constraints (defining v_{min}) were extracted from the SBML database in MATLAB (version 7.3; The MathWorks), and the relevant linear programming problems (see below) were solved using the MOSEK Optimization Toolbox for MATLAB (version 4). Finally, flux simulations were visualized on a metabolic flux map (which represents the central metabolism of a compartmentalized plant cell) drawn in Excel.

Manual Curation

In addition to compartmentalization, the manual curation process consisted of checking the model for reaction consistency, introducing biomass drains and transport reactions, and closing network gaps.

While KEGG is generally logically coherent, a number of inconsistencies make it impossible to use the automatically generated database directly. One is the use of multiple labels to describe the same compound (e.g. the use of nonspecific and specific references to sugar stereoisomers [D-Glc versus α-D-Glc]). Each such multiplicity was resolved as described previously (Quek and Nielsen, 2008). Another is the presence of nonbalanced reactions, typically for (1) the synthesis or breakdown of polymers (e.g. DNA + nucleotide = DNA), (2) the use of generic groups "R", and (3) the consumption or production of H_2O, H^+, and redox equivalents [e.g. NAD(P)H]. In AraGEM, polymers were described in the form of their corresponding monomers, and the use of the generic atom R was avoided.

Biomass drain reactions are incorporated into AraGEM as the accumulation terms of the biomass precursors. It is useful to describe these accumulation terms individually (e.g. "Sucrose = Sucrose_biomass") in order to simplify the task of uncovering the pathway gaps in each of the biosynthetic routes separately. The list of biomass components considered is shown in Table III and includes major structural and storage components as well as trace elements such as vitamins. Once the network gaps were filled (see below), these individual accumulation terms were combined into an overall biomass synthesis equation, with the appropriate coefficients assigned to each precursor to define the composition of biomass. The overall biomass synthesis equation is tissue specific; therefore, the composition of biomass has been estimated to represent photosynthetic or nonphotosynthetic tissues (e.g. leaf, stem, and root; Poorter and Bergkotte, 1992; Niemann et al., 1995). Trace elements were not included in the biomass equation, since their contribution to overall flux is trivial.

Exchange equations are used to describe the intercompartmental exchange of metabolites between cytoplasm and extracellular space and between cytoplasm and the organelles: plastids, peroxisomes, mitochondria, and vacuole. Annotation of transporters is far less developed than annotation of metabolic enzymes, and the annotation was largely manual based on normal metabolic functions described in the literature and transport requirements predicted from network gap filling.

The final step in the manual curation process is the identification of network gaps based on the model's ability to produce biomass components from substrates. For each biomass component in Table III, the following linear programming problem was formulated and solved

$$\text{maximize } v_{\mathrm{i}}$$

$$\text{subject to } Sv = 0$$

$$v_{\mathrm{min}} \leq v \leq v_{\mathrm{max}}$$

where v_{i} is the corresponding biomass drain reaction. The problem was solved for both photosynthetic tissues (photons as energy source, CO_2 as carbon source, and nitrate as nitrogen source) and nonphotosynthetic tissues (Suc as energy and carbon source, nitrate and/or amino acids as nitrogen source). Where the maximum production rate of a biomass component was zero, gap analysis was performed. Some gaps were readily filled based on inspection of the corresponding pathways in KEGG and AraCyc. Others, such as inconsistent irreversibility constraints, stoichiometry errors, compound names, compartmentalization errors, or missing transporters, required sequential tracing through the model to identify break points and careful evaluation of the possible causes.

Model Simulations

The final model was evaluated through the estimation of the flux distributions in three standard physiological scenarios: photosynthesis, photorespiration, and respiration in nonphotosynthetic tissues. The flux distributions were determined using linear programming

$$\text{minimize photon or Suc utilization}$$

$$\text{subject to } Sv = 0$$

$$v_{\mathrm{biomass}} = b$$

$$v_{\mathrm{min}} \leq v \leq v_{\mathrm{max}}$$

(i.e. the distributions that minimize the use of the key energy substrate [photons or Suc] while achieving a specified growth rate).

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Table S1. Modified reactions and gene-enzyme association.

Supplemental Table S2. List of biomass drains and transporters.

Supplemental Table S3. Reactions with no ORF association.

Supplemental Table S4. List of singletons.

Supplemental Data File S1. AraGEM_vs1.0, SBML format.

Supplemental Data File S2. FluxMapC3, Excel file.

Supplemental Data File S3. c3.m, MATLAB file.

ACKNOWLEDGMENTS

We thank Simon Boyes for complementing data curation analysis in AraGEM. We also acknowledge our partners, BSES Ltd. and Metabolix, Inc.

Received October 4, 2009; accepted December 24, 2009; published December 31, 2009.

LITERATURE CITED

Abrol YP, Sawhney SK, Naik MS (1983) Light and dark assimilation of nitrate in plants. Plant Cell Environ **6:** 595–599

Backhausen JE, Vetter S, Baalmann E, Kitzmann C, Scheibe R (1998) NAD-dependent malate dehydrogenase and glyceraldehyde 3-phosphate dehydrogenase isoenzymes play an important role in dark metabolism of various plastid types. Planta **205:** 359–366

Bonarius HPJ, Schmid G, Tramper J (1997) Flux analysis of underdetermined metabolic networks: the quest for the missing constraints. Trends Biotechnol **15:** 308–314

Boyle NR, Morgan J (2009) Flux balance analysis of primary metabolism in Chlamydomonas reinhardtii. BMC Syst Biol **3:** 4

Cakir T, Patil KR, Onsan ZI, Ulgen KO, Kirdar B, Nielsen J (2006) Integration of metabolome data with metabolic networks reveals reporter reactions. Mol Syst Biol **2:** 50

Coschigano KT, Melo-Oliveira R, Lim J, Coruzzi GM (1998) *Arabidopsis* gls mutants and distinct Fd-GOGAT genes: implications for photorespiration and primary nitrogen assimilation. Plant Cell **10:** 741–752

Cui J, Li P, Li G, Xu F, Zhao C, Li YH, Yang ZN, Wang G, Yu QB, Li YX, et al (2008) AtPID: Arabidopsis thaliana protein interactome database. An integrative platform for plant systems biology. Nucleic Acids Res **36:** D999–D1008

Dai SJ, Chen TT, Chong K, Xue YB, Liu SQ, Wang T (2007) Proteomics identification of differentially expressed proteins associated with pollen germination and tube growth reveals characteristics of germinated Oryza sativa pollen. Mol Cell Proteomics **6:** 207–230

DellaPenna D (2001) Plant metabolic engineering. Plant Physiol **125:** 160–163

Dennis C, Surridge C (2000) A. thaliana genome. Nature **408:** 791

Dennis DT, Miernyk JA (1982) Compartmentation of non-photosynthetic carbohydrate metabolism. Annu Rev Plant Physiol Plant Mol Biol **33:** 27–50

Duarte NC, Becker SA, Jamshidi N, Thiele I, Mo ML, Vo TD, Srivas R, Palsson BO (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data. Proc Natl Acad Sci USA **104:** 1777–1782

Duarte NC, Herrgard MJ, Palsson BO (2004) Reconstruction and validation of Saccharomyces cerevisiae iND750, a fully compartmentalized genome-scale metabolic model. Genome Res **14:** 1298–1309

Eckardt NA (2005) Photorespiration revisited. Plant Cell **17:** 2139–2141

Feist AM, Henry CS, Reed JL, Krummenacker M, Joyce AR, Karp PD, Broadbelt LJ, Hatzimanikatis V, Palsson BO (2007) A genome-scale metabolic reconstruction for Escherichia coli K-12 MG1655 that accounts for 1260 ORFs and thermodynamic information. Mol Syst Biol **3:** 121

Fong SS, Burgard AP, Herring CD, Knight EM, Blattner FR, Maranas CD, Palsson BO (2005) In silico design and adaptive evolution of Escherichia coli for production of lactic acid. Biotechnol Bioeng **91:** 643–648

Fong SS, Marciniak JY, Palsson BO (2003) Description and interpretation of adaptive evolution of Escherichia coli K-12 MG1655 by using a genome-scale in silico metabolic model. J Bacteriol **185:** 6400–6408

Forster J, Famili I, Fu P, Palsson BO, Nielsen J (2003) Genome-scale reconstruction of the Saccharomyces cerevisiae metabolic network. Genome Res **13:** 244–253

Goff SA (2005) A draft sequence of the rice genome (Oryza sativa L. ssp. japonica) (April, pg 92, 2002). Science **309:** 879–879

Gutierrez RA, Shasha DE, Coruzzi GM (2005) Systems biology for the virtual plant. Plant Physiol **138:** 550–554

Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol **169:** 453–468

Hansen GK, Jensen CR (1977) Growth and maintenance respiration in whole plants, tops, and roots of Lolium multiflorum. Physiol Plant **39:** 155–164

Kanehisa M, Araki M, Goto S, Hattori M, Hirakawa M, Itoh M, Katayama T, Kawashima S, Okuda S, Tokimatsu T, et al (2008) KEGG for linking genomes to life and the environment. Nucleic Acids Res **36:** D480–D484

Kitano H (2000) Perspectives on systems biology. New Generation Computing **18:** 199–216

Kitano H (2002a) Computational systems biology. Nature **420:** 206–210

Kitano H (2002b) Systems biology: a brief overview. Science **295:** 1662–1664

Lancien M, Martin M, Hsieh MH, Leustek T, Goodman H, Coruzzi GM (2002) Arabidopsis glt1-T mutant defines a role of NADH-GOGAT in the non-photorespiratory ammonium assimilatory pathway. Plant J **29:** 347–358

Linka M, Weber APM (2005) Shuffling ammonia between mitochondria and plastids during photorespiration. Trends Plant Sci **10:** 461–465

Maize Genome Sequencing Project (2009) http://www.maizegenome.org (January 3, 2010)

Minic Z, Jamet E, San-Clemente H, Pelletier S, Renou JP, Rihouey C, Okinyo DPO, Proux C, Lerouge P, Jouanin L (2009) Transcriptomic analysis of Arabidopsis developing stems: a close-up on cell wall genes. BMC Plant Biol **9:** 6

Morgenthal K, Weckwerth W, Steuer R (2006) Metabolomic networks in plants: transitions from pattern recognition to biological interpretation. Biosystems **83:** 108–117

Niemann GJ, Pureveen JBM, Eijkel GB, Poorter H, Boon JJ (1995) Differential chemical allocation and plant adaptation: a Py-Ms study of 24 species differing in relative growth rate. Plant Soil **175:** 275–289

Oliveira AP, Patil KR, Nielsen J (2008) Architecture of transcriptional regulatory circuits is knitted over the topology of bio-molecular interaction networks. BMC Syst Biol **2:** 17

Paterson AH, Bowers JE, Bruggmann R, Dubchak I, Grimwood J, Gundlach H, Haberer G, Hellsten U, Mitros T, Poliakov A, et al (2009) The Sorghum bicolor genome and the diversification of grasses. Nature **457:** 551–556

Pessarakli M (1996) Handbook of Photosynthesis. Marcel Dekker, New York

Petersen J, Brinkmann H, Cerff R (2003) Origin, evolution, and metabolic role of a novel glycolytic GAPDH enzyme recruited by land plant plastids. J Mol Evol **57:** 16–26

Poolman M, Miguet L, Sweetlove LJ, Fell DA (2009) A genome-scale metabolic model of Arabidopsis and some of its properties. Plant Physiol **151:** 1570–1581

Poorter H, Bergkotte M (1992) Chemical composition of 24 wild species differing in relative growth rate. Plant Cell Environ **15:** 221–229

Price ND, Reed JL, Palsson BO (2004) Genome-scale models of microbial cells: evaluating the consequences of constraints. Nat Rev Microbiol **2:** 886–897

Quek LE, Nielsen LK (2008) On the reconstruction of the Mus musculus genome-scale metabolic network model. Genome Inform **21:** 89–100

Reed JL, Patel TR, Chen KH, Joyce AR, Applebee MK, Herring CD, Bui OT, Knight EM, Fong SS, Palsson BO (2006) Systems approach to refining genome annotation. Proc Natl Acad Sci USA **103:** 17480–17484

Reumann S, Ma CL, Lemke S, Babujee L (2004) AraPerox: a database of putative Arabidopsis proteins from plant peroxisomes. Plant Physiol **136:** 2587–2608

Sheikh K, Forster J, Nielsen LK (2005) Modeling hybridoma cell metabolism using a generic genome-scale metabolic model of Mus musculus. Biotechnol Prog **21:** 112–121

Sugimoto M, Takahashi K, Kitayama T, Ito D, Tomita M (2005) Distributed cell biology simulations with E-Cell System. *In* Lecture Notes in Computer Science. Springer-Verlag, Berlin, pp 20–31

Sweetlove LJ, Last RL, Fernie AR (2003) Predictive metabolic engineering: a goal for systems biology. Plant Physiol **132:** 420–425

Taiz L, Zeiger E (2006) Plant Physiology, Ed 4. Sinauer Associates, Sunderland, MA

Thiellement H (1999) Proteomics, an holistic approach for plant biology. Arch Sci **52:** 41–45

Tomita M, Hashimoto K, Takahashi K, Shimizu TS, Matsuzaki Y, Miyoshi F, Saito K, Tanida S, Yugi K, Venter JC, et al (1999) E-CELL: software environment for whole-cell simulation. Bioinformatics **15:** 72–84

Trepp GB, van de Mortel M, Yoshioka H, Miller SS, Samac DA, Gantt JS, Vance CP (1999) NADH-glutamate synthase in alfalfa root nodules: genetic regulation and cellular expression. Plant Physiol **119:** 817–828

Turano FJ, Muhitch MJ (1999) Differential accumulation of ferredoxin- and NADH-dependent glutamate synthase activities, peptides, and transcripts in developing soybean seedlings in response to light, nitrogen, and nodulation. Physiol Plant **107:** 407–418

van Dijk JP, Cankar K, Scheffer SJ, Beenen HG, Shepherd LVT, Stewart D, Davies HV, Wilkockson SJ, Lelfert C, Gruden K, et al (2009) Transcriptome analysis of potato tubers: effects of different agricultural practices. J Agric Food Chem **57:** 1612–1623

Varma A, Palsson BO (1994) Metabolic flux balancing: basic concepts, scientific and practical use. Biotechnology (N Y) **12:** 994–998

Wedding RT (1989) Malic enzymes of higher plants: characteristics, regulation, and physiological function. Plant Physiol **90:** 367–371

Wingler A, Lea PJ, Quick WP, Leegood RC (2000) Photorespiration: metabolic pathways and their role in stress protection. Philos Trans R Soc Lond B Biol Sci **355:** 1517–1529

Wise RR, Hoober JK (2007) Synthesis, export and partitioning of end products of photosynthesis. *In* RR Wise, JK Hoober, eds, Structure and Function of Plastids, Vol 23. Springer, Dordrecht, The Netherlands, pp 274–288

Wolosiuk RA, Buchanan BB (1978) Activation of chloroplast NADP-linked glyceraldehyde-3-phosphate dehydrogenase by ferredoxin-thioredoxin system. Plant Physiol **61:** 669–671

Yonekura-Sakakibara K, Tohge T, Matsuda F, Nakabayashi R, Takayama H, Niida R, Watanabe-Takahashi A, Inoue E, Saito K (2008) Comprehensive flavonol profiling and transcriptome coexpression analysis leading to decoding gene-metabolite correlations in *Arabidopsis*. Plant Cell **20:** 2160–2176