NONLINEAR CONSTRAINT-BASED MODELING OF THE FUNCTION AND EVOLUTION OF C4 PHOTOSYNTHESIS

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Elijah Lane Bogart

August 2015

NONLINEAR CONSTRAINT-BASED MODELING OF THE FUNCTION AND EVOLUTION OF C4 PHOTOSYNTHESIS

Elijah Lane Bogart, Ph.D.

Cornell University 2015

C4 plants, such as maize, concentrate carbon dioxide in a specialized compartment surrounding the veins of their leaves to improve the efficiency of carbon dioxide assimilation. The C4 photosynthetic system is a key target of efforts to improve crop yield through biotechnology, and its independent development in dozens of plant species widely separated geographically and phylogenetically is an intriguing example of convergent evolution.

The availability of extensive high-throughput experimental data from C4 and non-C4 plants, as well as the origin of the biochemical pathways of C4 photosynthesis in the recruitment of enzymatic reactions already present in the ancestral state, makes it natural to study the development, function and evolution of the C4 system in the context of a plant's complete metabolic network, but the essentially nonlinear relationship between rates of photosynthesis, rates of photorespiration, and carbon dioxide and oxygen levels prevents the application of conventional, linear methods for genome-scale metabolic modeling to these questions.

I present an approach which incorporates nonlinear constraints on reaction rates arising from enzyme kinetics and diffusion laws into flux balance analysis problems, and software to enable it. Applying the technique to a new genome-scale model, suitable for describing metabolism in the leaves of either *Zea mays* or generic plants, I show it can reproduce known nonlinear physiological re-

sponses of C3 and C4 plants.

In combination with a novel method for inferring metabolic activity from enzyme expression data, I use the nonlinear model to interpret multiple channels of transcriptomic and biochemical data in the developing maize leaf, showing that the predicted metabolic state reproduces the transition between carbon-importing tissue at the leaf base and carbon-exporting tissue at the leaf tip while making additional testable predictions about metabolic shifts along the developmental axis.

Adapting a method for simulating transition paths in physical and chemical systems, I find the highest-fitness paths connecting C3 and C4 states in the model's high-dimensional parameter space, show that such paths reproduce known aspects of the evolutionary history of the C4 position, and study their response to variation in environmental conditions and C4 biochemistry.

BIOGRAPHICAL SKETCH

Eli Bogart was born in Oregon. At an impressionable age he read James Gleick's *Chaos* and proceeded to spend several days in the back of his fifth-grade classroom ignoring the teacher and implementing simulations of predator and prey population dynamics on the school's aging Commodore 64s. His interest in using mathematics and computation to understand the world continued through his studies in physics at Harvey Mudd College, where he graduated in 2005, and two years in the mathematical biology IGERT program at the University of Utah, before he came to Cornell in 2008. He is a somewhat lapsed violist.

ACKNOWLEDGEMENTS

This thesis has been made possible by the help of many people, directly and indirectly.

It has been a great pleasure to work with Chris Myers, whose wide-ranging interests, diverse and extensive knowledge, open-mindedness, and patience make him an ideal advisor for exploratory, interdisciplinary projects such as this.

Among our excellent collaborators in the world of C4 photosynthesis, Tom Brutnell's support has been critical from the beginning of the project and I regret I have not been able to follow up on all of the interesting ideas that always come up in talking with him. Tim Nelson provided helpful comments on the paper that became chapter 2 below, and he, Lori Tausta, Lin Wang, Qi Sun, and Zehong Ding have all provided data, insight, and fruitful questions to ask, at various times. Financial support for the work presented here was provided by National Science Foundation grant IOS-1127017 and a grant to the International Rice Research Institute from the Bill and Melinda Gates Foundation.

I thank Sol Gruner and Haiyuan Yu for their service on my special committee. The late Roger Spanswick's enthusiasm for science was infectious, and the courses I took from him made a notable impact on me through the sense he conveyed of inviting his students to join a vibrant, decade-spanning, worldwide effort to work out how life functions at the cellular and biochemical level. Encouragement and kindness from Carl Franck before, during, and after my time in his lab is much appreciated. My experiences studying with Aaron Fogelson and James Keener at the University of Utah continue to shape my approach to all problems that combine mathematics and biology.

I was lucky to work in an office full of philosophically-minded friends. I enjoyed many discussions with Sarabjeet Singh, Jason Hindes, and Oleg Kogan. Lei Huang has been a good friend and a major intellectual influence; our long conversations over the years made me think critically about my approaches to metabolic modeling, science, the grad school experience, and life. This was great fun and I am better off for it. Brandon Barker has provided an important perspective on many metabolic modeling ideas as well as the driving force behind our DREAM8 team and other collaborative endeavours.

In the past few years I have learned a lot from my teachers and fellow students in the Cornell Aikido Club, with whom I wish I had spent more time training.

Finally, I am deeply grateful for the encouragement and patience of my friends outside of Cornell and my family. I won't name everyone who deserves it, but I particularly appreciate the friendship, hospitality, and advice of Katy Perdue (whose suggestion that I go talk to John Milton one day in senior year at Mudd has ended up shaping my academic trajectory for nearly a decade) and Cal Pierog. Without the support of my parents, Debra and Barry Bogart, which has been unwavering from their efforts to give me opportunities to pursue my earliest interests in science to their moral support in the late phases of the thesis-writing process and in so many other ways in between, I would not have made it to this point. The love and companionship of Adrian Baur has brightened many days.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Acknowledgements . . . iv
Table of Contents . . . vi
List of Tables . . . ix
List of Figures . . . x

1 Introduction **1**
1.1 C4 photosynthesis . . . 1
1.2 Evolution of the C4 system . . . 3
1.3 Modeling C4 and other plant metabolism . . . 4
1.3.1 Nonlinear physiological models . . . 5
1.3.2 Kinetic models . . . 5
1.3.3 Constraint-based models . . . 6
1.3.4 Review of constraint-based models for higher plants . . . 7
1.4 Incorporating Rubisco kinetics in constraint-based models . . . 23
1.5 Outline . . . 25

2 Multiscale modeling of metabolism in the developing maize leaf **26**
2.1 Introduction . . . 26
2.2 Results . . . 26
2.2.1 Metabolic reconstruction of *Zea mays* . . . 26
2.2.2 Nonlinear flux-balance analysis . . . 29
2.2.3 Flux predictions in the developing leaf based on multiple data channels . . . 31
2.3 Discussion . . . 47
2.3.1 Reconstruction . . . 47
2.3.2 Nonlinear optimization . . . 53
2.3.3 Data fitting . . . 55
2.3.4 The whole-leaf model . . . 58
2.4 Methods . . . 60
2.4.1 Reconstruction process . . . 60
2.4.2 Mesophyll-bundle sheath model . . . 61
2.4.3 Leaf gradient model . . . 62
2.4.4 Optimization calculations . . . 64
2.4.5 Integrating biochemical and RNA-seq data . . . 65

3 Genome-scale modeling of the evolutionary path to C4 photosynthesis **70**
3.1 Introduction . . . 70
3.2 Methods . . . 72
3.2.1 Modeling photosynthetic metabolism . . . 72
3.2.2 Finding optimal evolutionary paths . . . 80

3.2.3 Combining the metabolic and evolutionary pathfinding models . . . 83
3.2.4 Limitations of the approach . . . 85
3.3 Results . . . 85
3.3.1 Fitness increases and path geometry . . . 85
3.3.2 Development of the C4 system . . . 87
3.3.3 Comparison to the model of Heckmann et al. . . . 91
3.3.4 Clustering analysis . . . 94
3.3.5 Varying environmental conditions . . . 96
3.3.6 Varying decarboxylation subtypes . . . 98
3.3.7 Combined environmental and biochemical variation . . . 103
3.4 Assessing the elastic band approximation to the highest-fitness path . . . 104
3.5 Discussion . . . 108

A Development of a flux balance analysis model for maize 114
A.1 Exporting the CornCyc FBA model from Pathway Tools . . . 114
A.2 Discarding reactions . . . 116
A.2.1 Polymerization reactions . . . 116
A.2.2 ATPases . . . 116
A.2.3 Reactions involving generic electron donors and acceptors 117
A.2.4 Duplicates . . . 118
A.2.5 Non-metabolic reactions . . . 118
A.2.6 Glucose-6-phospate . . . 118
A.2.7 UDP-glucose . . . 119
A.3 Minor revisions to achieve basic functionality . . . 119
A.3.1 Mitochondrial electron transport chain . . . 119
A.3.2 Photosynthesis: light reactions . . . 120
A.3.3 Key reactions in biomass component production and nutrient uptake . . . 121
A.3.4 Ascorbate-glutathione cycle . . . 130
A.3.5 Gamma-glutamyl cycle . . . 130
A.3.6 Methionine synthesis from homocysteine . . . 131
A.3.7 Basic import and export . . . 131
A.3.8 Defining the biomass components . . . 132
A.4 Compartmentalization . . . 133
A.4.1 Intracellular transport . . . 134
A.4.2 Photorespiratory pathway . . . 135
A.4.3 Various ferredoxin-consuming pathways . . . 135
A.4.4 Ascorbate production . . . 136
A.4.5 Ascorbate-glutathione cycle . . . 136
A.5 Gene associations for compartmentalized reactions . . . 137
A.5.1 NADH dehydrogenases . . . 138
A.5.2 Pyruvate dehydrogenases . . . 138

A.6 Testing and consistency checking . . . 139
A.7 SBML export . . . 140
A.7.1 Component names . . . 140
A.7.2 Gene annotations . . . 140
A.8 Model refinement . . . 141
A.8.1 Phosphoribulokinase . . . 141
A.9 Biomass equation . . . 142
A.9.1 Fatty acids . . . 142
A.9.2 Hemicellulose . . . 144
A.9.3 Total carbohydrates . . . 144
A.9.4 Organic acids . . . 145
A.9.5 Protein and free amino acids . . . 145
A.9.6 Lignin . . . 146
A.9.7 Nucleic acids . . . 146
A.9.8 Nitrogenous compounds . . . 147
A.9.9 Inorganic materials . . . 147
A.9.10 Total biomass reaction . . . 148
A.9.11 Protonation . . . 148
A.9.12 Oxalate . . . 149
A.10 Plasmodesmatal transport reactions . . . 150

B Supplementary tables **154**
B.1 Overrepresented pathways in the k-means clusters of Fig. 3.6. . . . 154

C An alternative photorespiratory pathway **158**

D Theoretical and practical considerations in solving nonlinear flux balance analysis problems with IPOPT **162**
D.1 The Karush-Kuhn-Tucker conditions . . . 162
D.2 IPOPT . . . 163
D.3 Constraint degeneracy . . . 167

Bibliography **170**

LIST OF TABLES

2.1 Parameters contributing to the effective maximum rate of regeneration of phosphoenolpyruvate in the genome-scale model. . . . 30

3.1 Glossary of symbols and values of parameters used in the nonlinear FBA calculations. 75
3.2 Efficiency parameters used to obtain C4 endpoints of different biochemical subtypes. 100

A.1 Fatty acid proportions in biomass. 143
A.2 Carbohydrate species in biomass. 151
A.3 Nitrogenous biomass breakdown. 152
A.4 Breakdown of total biomass. 153

B.1 Cluster a (140 parameters) 155
B.2 Cluster b (134 parameters) 155
B.3 Cluster c (47 parameters) 156
B.4 Cluster d (38 parameters) 156
B.5 Cluster e (30 parameters) 156
B.6 Cluster f (13 parameters) 157
B.7 Cluster g (11 parameters) 157
B.8 Cluster h (9 parameters) 157

LIST OF FIGURES

1.1 Schematic of the C4 system. 2

2.1 Maize plant and models 27
2.2 CO_2 assimilation rates (A) predicted by the C4 photosynthesis model of von Caemmerer [25], compared to predictions from the present nonlinear genome-scale model, maximizing CO_2 assimilation with equivalent parameters. 31
2.3 Source-sink transition along the leaf as predicted by optimizing the agreement between fluxes in the nonlinear model and RNA-seq data. 35
2.4 Predicted phloem transport of nitrogen and sulfur 36
2.5 Operation of the C4 system in the best-fitting solution. 38
2.6 Photosystem II in mesophyll and bundle sheath. 39
2.7 Bundle sheath PEPC flux in the best-fitting solution. 40
2.8 Agreement between RNA-seq data and predicted fluxes. 42
2.9 Data and predicted fluxes for a linear pathway and a metabolic branch point. 43
2.10 Summary of predictions for the gradient model using the least-squares method without per-reaction scale factors. 45
2.11 Summary of predictions for the gradient model using the E-Flux method. 47
2.12 Summary of predictions for the gradient model using the E-Flux method with fixed biomass composition. 48
2.13 Summary of predictions for the gradient model with fixed biomass composition. 49
2.14 Predicted biomass production rates in mesophyll and bundle sheath cells with fixed biomass composition. 50
2.15 Predicted variable values in an FBA calculation that does not incorporate expression data, compared to the best-fit and E-Flux methods. 51

3.1 Simulated fitness landscape between C3 and C4 states. 86
3.2 Alternate view of shapes and fitness landscapes of straight-line and elastic band paths for various spring constants. 88
3.3 Changes in activity levels of key enzymes and rates of biomass synthesis through the simulated C3-C4 transition 89
3.4 Fraction of biomass produced in the bundle sheath 90
3.5 Comparison of predicted paths to the simulations of Heckmann et al. 92
3.6 Trends in values of the evolving parameters through the simulated C3-C4 transition 94
3.7 Simulated C3-C4 transitions for varying external CO_2 levels. . . . 97

3.8 Predicted shifts in the timing of adaptation of the Rubisco kinetic parameters in response to atmospheric CO_2 levels. 99
3.9 Simulated paths from the C3 state to C4 states using six different combinations of decarboxylating enzymes. 100
3.10 Differences by subtype in light use in bundle sheath and mesophyll along simulated evolutionary paths. 103
3.11 Hierarchical clustering of results for eighteen combinations of intercellular CO_2 levels and decarboxylation subtypes. 105
3.12 Angles between elastic band path tangent vectors and approximate local directions of steepest improvement in fitness. 109

C.1 Hypothetical photorespiratory bypass system without CO_2 loss . 161

CHAPTER 1

INTRODUCTION

1.1 C4 photosynthesis

In the process of photosynthesis, plants use energy from incident light to incorporate carbon from atmospheric carbon dioxide into larger organic molecules. Carbon dioxide and oxygen bind competitively to the same active site of the enzyme Rubisco, which is responsible for photosynthetic CO_2 assimilation [1]. Rubisco-catalyzed carboxylation leads to net CO_2 assimilation, while Rubisco-catalyzed oxygenation leads to photorespiration, in which CO_2 is released [2]. C4 photosynthesis is an anatomical and biochemical system which improves the efficiency of carbon dioxide assimilation in plant leaves by restricting Rubisco to specialized bundle sheath compartments, surrounding the leaf veins, where a high-CO_2 environment is maintained that favors CO_2 over O_2 in their competition for the active sites, thus suppressing photorespiration [3].

In most C4 plants, the CO_2 concentrating system uses phosphoenolpyruvate carboxylase (PEPC) in mesophyll cells to transiently incorporate CO_2 into four-carbon molecules, which cross into the adjacent bundle sheath cells and are decarboxylated again by one of three enzymes (NADP-dependent malic enzyme, NAD-dependent malic enzyme, and phosphoenolpyruvate carboxykinase), with C4 plants conventionally divided into three subtypes based on which decarboxylating enzyme they primarily employ [4]. Figure 1.1(a) illustrates the biochemistry of the system, including key reactions of both the NADP-ME and PEPCK subtypes.

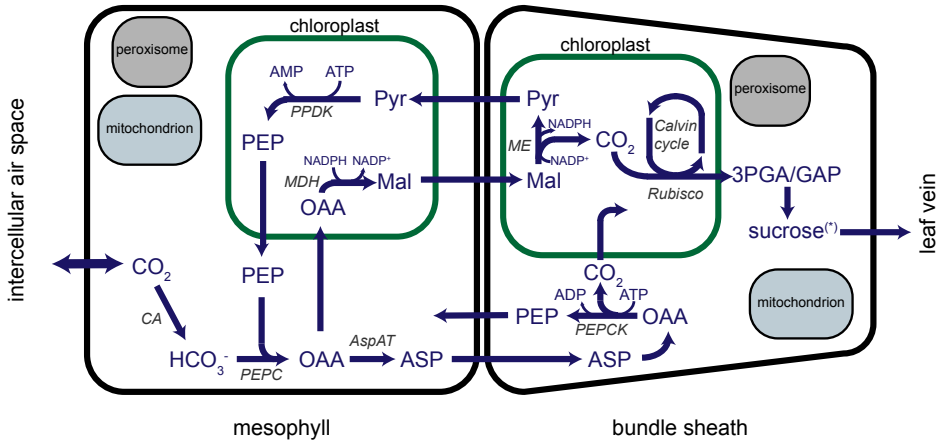

Figure 1.1: **Schematic of the C4 system.** Key reactions of C4 photosynthesis in mesophyll and bundle sheath cells are shown. The system shown is a combination of the NADP-ME and PEPCK subtypes; the NAD-ME subtype is similar to the NADP-ME, though NAD-ME is typically located in the bundle sheath mitochondrion. For simplicity, conversion of triose phosphate to sucrose has been drawn in the bundle sheath, though it is thought to occur in the mesophyll in many C4 species [4]. CA, carbonic anhydrase; PEPC, phosphoenolpyruvate carboxylase; OAA, oxaloacetate; AspAT, aspartate aminotransferase; PEPCK, phosphoenolpyruvate carboxykinase; PEP, phosphoenolpyruvate; Mal, malate; MDH, NADP-malate dehydrogenase; ME, NADP-malic enzyme; Pyr, pyruvate; PPDK, pyruvate, orthophosphate dikinase; 3PGA, 3-phosphoglycerate; GAP, glyceraldehyde 3-phosphate.

The C4 system confers increased tolerance of low atmospheric CO_2 levels as well as improved nitrogen and water use efficiencies, and a number of key crop species, including maize, sugarcane, and sorghum, are C4 plants, as are many weeds [5]. The agricultural and ecological significance of the C4 system, the prospect of increasing yields of C3 crops by artificially introducing C4 functionality to those species [6, 7], and its remarkable evolutionary history, discussed below, have made it the object of intense study. The core biochemical pathways are now generally understood [4] but many areas of uncertainty remain, including the genetic regulation of the C4 system [8], the importance of partic-

ular components of the system to its function (e.g., [9]), and the significance of inter-specific variations in C4 biochemistry [10].

1.2 Evolution of the C4 system

Despite the complexity of the system, which involves extensive coordinated biochemical and anatomical changes from the ancestral (C3) state, C4 photosynthesis has evolved independently in more than 60 distinct plant lineages, widely separated both geographically and phylogenetically [11], at various times in the last 30 million years, following a decline in atmospheric CO_2 levels at the end of the Oligocene epoch [12, 13].

A sequential model for the evolution of the C4 system was proposed by Sage [14], based on evidence from plant species showing intermediate phenotypes with both C3 and C4 characteristics. In summary, the process is believed to begin with preconditioning, changes which do not immediately modify the photosynthetic process but facilitate the later acquisition of C4-like traits. This may include gene duplication (allowing existing genes to be recruited to new functions while maintaining their original functions [15]) as well as anatomical changes (increases in vein density and bundle sheath size [16], in addition to increased numbers of chloroplasts and mitochondria in the bundle sheath cells). After preconditioning, metabolic changes begin with the loss of activity of glycine decarboxylase – the enzyme responsible for photorespiratory CO_2 release – in the mesophyll, and the establishment of a shuttle system in which glycine produced in the mesophyll is decarboxylated in the bundle sheath. As a result, the bundle sheath CO_2 level rises. Next, mesophyll PEPC levels and the

activity of the C4 cycle increase, and the enzymes of the C3 photosynthetic system are redistributed between mesophyll and bundle sheath cells to exploit the increased bundle sheath CO_2 level. Finally, assorted changes occur to optimize enzyme kinetics and metabolic regulation for the new C4 system.

Beyond this general (and flexible) outline, many aspects of C4 evolution remain incompletely understood, and the area is an active topic of research. Among other issues, recent work has explored the relationships among the different C4 decarboxylation subtypes in an evolutionary context, presenting evidence and theoretical arguments that the subtypes are not well-defined – as multiple decarboxylating enzymes often act in combination in the same plant – and that the importance of individual decarboxylases may change in the course of evolution of a particular C4 lineage [10, 17, 18].

Precisely how environmental conditions drove the evolution of the C4 system and the subsequent ecological success of C4 species is also unclear [19], with roles proposed for conversion of woodlands to open habitats by mammalian megaherbivores [20] and fire [21], regional changes in rainfall patterns [22], and the need to limit water loss by transpiration in warm, low-CO_2 conditions [23], in addition to the direct effect of low atmospheric CO_2 on C3 carbon assimilation rates [24].

1.3 Modeling C4 and other plant metabolism

A variety of approaches to computational and mathematical modeling of C4 photosynthesis, or photosynthetic physiology and metabolism more broadly, have been pursued.

1.3.1 Nonlinear physiological models

High-level nonlinear models of photosynthetic physiology [25] start from a small number of well-supported equations relating enzyme activities, light levels, and atmospheric CO_2 levels (including kinetic laws for PEPC and Rubisco, mass balance equations, diffusion laws, empirical relationships between incident light levels and rates of the photosynthetic electron transport reactions, and requirements that photosynthetic ATP production balance its consumption by the Calvin cycle and the C4 cycle) and solve them to find the rate of CO_2 assimilation by leaves under various types of conditions.

These models have been widely applied to infer biochemical properties of C3 and C4 plants from macroscopic experiments (typically, measurements of gas exchange in leaves under varying conditions) and for theoretical explorations of the function of C3, C4, and C3-C4 intermediate photosynthesis. They are widely accepted and quite tractable, but describe many aspects of plant metabolism in only extremely abstract terms.

1.3.2 Kinetic models

Detailed kinetic models describe metabolic systems as systems of ordinary differential equations, with the rate of change of metabolite concentrations depending on the rates of the reactions which produce and consume them, which are determined in turn by kinetic laws. Recently, such models have been used to explore the optimal allocation of resources to enzymes in C3 plants [26] and NADP-ME type C4 plants [27], and probe the relationship between the three C4 decarboxylation types [28].

Such models can offer great insight, particularly into non-steady-state questions such as responses to environmental fluctuations [29]. However, they require the specification of many parameter values – over 200 for [27] (though a few may be left variable) – which may be difficult to compile or unknown, and the effect of uncertainty in the parameters on the predictions of the model is rarely addressed in a comprehensive, rigorous way.

1.3.3 Constraint-based models

Constraint-based models make predictions about metabolic reaction rates which are consistent with the structure of a network representation of the chemical species and metabolic reactions believed to be in a cell or collection of cells, as encapsulated in a stoichiometry matrix, whose entries are the stoichiometric coefficients of each species in each reaction [30]. Generally, few parameters (other than the stoichiometric coefficients) are required. The technique is typically applied to so-called 'genome-scale' models, which attempt to capture all or a large fraction of of a cell's metabolic repertoire as defined by the reactions for which catalyzing enzymes are encoded in its genome.

Such detailed, large-scale metabolic models offer particular advantages for the investigation of connections between the C4 system and a plant's metabolism more broadly (for example, partitioning of nonphotosynthetic functions between mesophyll and bundle sheath, or the evolutionary recruitment of nonphotosynthetic reactions into the C4 cycle) and for interpreting high-throughput experimental data from C4 systems.

However, the development of genome-scale metabolic reconstructions for

C4 plants, and multicellular plants more generally, has lagged behind the development of such models for single-celled organisms for several reasons. Most obviously, genome-scale reconstructions are typically built starting from a genome sequence for the organism of interest, and plant genomes began to be sequenced [31] only some years after bacterial [32] and fungal [33] genomes; similarly, understanding of metabolism, and tools for predicting enzyme function from gene sequence, remain less developed in plants than in key single-celled model organisms.

Below, we review key prior work in the area of plant constraint-based model construction, before turning to a fundamental problem with the application of standard constraint-based modeling techniques to photosynthetic systems (in section 1.4).

1.3.4 Review of constraint-based models for higher plants

The following list includes all large-scale higher plant constraint-based models published through early 2012, and several particularly notable models published since that date. Among the types of data given for the models are:

Method of construction Fundamentally a constraint-based model consists of a list of reactions among a set of chemical species in a standard naming scheme. To build a model of limited scope this list can be assembled by hand. For 'genome-scale' models this list is drawn from one or more databases which predict the reactions that will be catalyzed by enzymes encoded in a plant's genome (on the basis of sequence similarity to genes for enzymes of known function as well as manual curation based on the

literature). In practice such reconstructions must be supplemented with additional reactions to allow the model system to perform its expected metabolic functions (gap-filling).

Compartmentalization To what extent reactions are localized to subcellular organelles or other sub-compartments of the model, and which such sub-compartments are present.

Metabolic scope What pathways, systems, and functions the model can describe. We have summarized descriptions provided by the authors rather than evaluating each model independently.

Consistency checking An umbrella term for efforts to ensure that reactions in a constraint-based model are elementally balanced and, ideally, charge-balanced, and thus to guarantee that the predicted metabolic steady states satisfy conservation of total mass, per-element mass, and charge. Polymerization reactions (often written in the unfortunate form 'polymer + monomer = polymer') and reactions described in the source database with generic reactants ('acceptor', 'an aldehyde', etc.,) are frequent sources of conservation problems, as are inconsistent assumptions about the state of protonation of weak acids. The level of attention to these details varies significantly between models.

Reproducibility Only included for models where we made significant efforts to reproduce the published results.

Grafahrend-Belau barley seed CBM [34]

Description A large (but not genome-scale) model intended for studies of grain yield and composition under perturbations.

Method of construction Manually compiled; a variety of databases including KEGG, MetaCyc, AraCyc, RiceCyc and Reactome were consulted, and specific references to the primary literature are provided for nearly every reaction (mostly to barley-specific work, with gaps filled from literature on wheat, rice, and maize or, in rare cases, dicots).

Metabolic scope Primary metabolism of the developing barley seed (thus, no photosynthesis or photorespiration): glycolysis, pentose phosphate pathway, TCA cycle, amino acid metabolism, starch synthesis.

Size 257 reactions, 234 metabolites.

Compartmentalization Cytosol, mitochondrion, amyloplast.

Consistency checking Not described in detail.

Results and comparison to experiment Authors report that "predicted growth rate and the active metabolic pathway patterns under anoxic, hypoxic and aerobic conditions predicted by the model were in accordance with published experimental results."

Gene-reaction associations No comprehensive table.

Biomass sink reaction Includes all components accounting for more than 1% of dry weight of barley seeds: in practice this meant various carbohydrates and amino acids, in proportions determined from literature seed composition data.

AraMeta [35]

Description A genome-scale model describe heterotrophic (i.e., non-photosynthetic) growth of Arabidopsis cells in suspension culture.

Method of construction From the AraCyc metabolic pathway database, with corrections and additions to allow biomass production.

Metabolic scope Includes glycolysis, pentose phosphate pathway, TCA cycle and mitochondrial respiration; synthesis of amino acids, fatty acids, nucleotides, carbohydrates. No light reactions of photosynthesis.

Size 1253 metabolites and 1406 reactions, before correction for blocked reactions.

Compartmentalization Limited: key reactions of mitochondrial respiration are localized to the mitochondrion, but all other reactions are combined in one main compartment.

Consistency checking Fairly thorough: all reactions are elementally balanced, except for hydrogen; polymerization handled carefully.

Comparison to experiment Extensive. In a later paper [36]predicted fluxes were compared against fluxes measured by 13C-MFA and found to correlate well across several different stress conditions, though there were areas of clear disagreement (model predictions bypassed the pentose phosphate pathway, e.g.).

Gene-reaction association None, but these could be determined from AraCyc.

Biomass sink reaction Includes cell wall components, lipid, starch, nucleic acid, and amino acids, in proportions determined experimentally by the same authors.

Reproducibility Good but imperfect: my determination of the minimum-flux biomass-producing solution uses a slightly smaller set of reactions than the authors'; exact proportions of individual biomass components are not always clear.

AraGEM [37]

Description A comprehensive *Arabidopsis* genome-scale metabolic model, probably the best-known plant CBM.

Method of construction From *Arabidopsis* gene-reaction associations in KEGG as of January 2009 (release 49.0), with additions as necessary to allow biomass production, and additional manual curation.

Metabolic scope Primary metabolism including glycolysis, pentose phosphate pathway, TCA cycle, light and dark reactions of photosynthesis, fatty acid synthesis, β-oxidation, glyoxylate cycle, photorespiration.

Size 1567 reactions and 1748 metabolites (per text of paper).

Compartmentalization Thorough, manually determined from literature and database sources (including TAIR and the peroxisomal proteome database AraPerox). Includes cytoplasm, mitochondrion, peroxisome, plastid, vacuole.

Consistency checking Not described in detail but efforts were made to achieve consistent nomenclature and handling of polymerization reactions.

Results and comparison to experiment Optimizing photon use efficiency when a rubisco carboxylation/oxygenation ratio was imposed reproduced the classical photorespiratory pathway; various comparisons between fluxes under light and dark conditions were consistent with prior literature reports. No direct comparison to experiment.

Gene-reaction association Thorough; 5,253 gene-reaction associations, involving 1419 genes.

Biomass sink reaction Estimated from literature, includes different drains for for photosynthetic and non-photosynthetic tissues; 148 separate biomass

components including "carbohydrates, amino acids, fatty acid, cellulose, and hemicellulose" as well as nucleotides, biotin, CoA, riboflavin, folate, chlorophyll, nicotinamide, thiamine, ubiquinone.

Reproducibility Though the published SBML file was not strictly compliant to the standard, this was correctable and we were able to achieve flux through the system. However, it became clear that many reactions of interest (PEPC, for example) were parts of unrealistic cycles, which were plentiful in the published flux results as well (e.g., no source or sink of H^+ was explicitly included, so two unbalanced reactions acted together to create or destroy it).

Radrich Arabidopsis model [38]

Description A general-purpose genome-scale network model of *Arabidopsis* metabolism, prepared to illustrate a method of creating metabolic network models from independent databases semi-automatically, on the plausible conjecture that reactions supported by both sources are likelier to be accurately represented and biologically realistic. Not a complete model suitable for calculations (no biomass sink, nutrient sources).

Method of construction Compounds in the AraCyc and KEGG databases were matched, using an automated system to identify clear matches and propose others for manual verification. Reactions in AraCyc were then assigned to corresponding reactions in KEGG, where possible, in an iterative process which used reaction identifications to resolve additional compound matches. Three network models were created. The core network included only reactions and compounds confidently matched between

KEGG and AraCyc. An intermediate network added those reactions from either database where all the substrates or all the products could be confidently matched, and all species involved in those reactions. A complete network added all remaining components from both databases.

Metabolic scope Very broad coverage of primary metabolism and some secondary metabolism, though note that, particularly in the core and intermediate models, complete pathways may not be represented and would not be functional in calculations. (No light reactions of photosynthesis.)

Size Core model, 753 reactions, 914 metabolites; intermediate model, 1388 reactions, 1248 metabolites; complete model, 2315 reactions, 2328 metabolites.

Compartmentalization None.

Consistency checking Networks were checked systematically for conservation violations (except of hydrogen and charge, as protonation state was ignored throughout the process); the core network was consistent with conservation, but the intermediate and complete network had multiple conservation issues, which the authors attributed to generic reactants.

Results and comparison to experiment No flux predictions; graph properties (degree distribution, clustering coefficient distribution, betweenness and closeness centrality, etc.,) were calculated and found to be broadly consistent with other analyses of metabolic networks.

Biomass sink reaction None.

Gene-reaction associations An annotated SBML version of the core network includes links from (an unspecified number of) enzymes to gene records in TAIR.

Reproducibility Not assessed.

Though this work did not produce a model immediately usable for FBA, it is notable for its method of construction, network-theoretical analyses, and the inclusion of a brief comparison of the representation of the TCA and glyoxylate cycles in the new model with that in AraMeta and AraGEM, identifying various omissions in the latter two and quite a few discrepancies in cofactor use between the three, which gives some sense of the overall level of agreement.

C4GEM [39]

Description Intended to allow detailed metabolic modeling of all three C4 subtypes, including photosynthesis, considering both mesophyll and bundle sheath cell types.

Method of construction From reactions assigned in KEGG to maize, sorghum, or sugarcane genes with gaps filled by reactions from AraGEM.

Metabolic scope Primary metabolism including glycolysis, pentose phosphate pathway, TCA cycle, light and dark reactions of photosynthesis, fatty acid synthesis, β-oxidation, glyoxylate cycle.

Size 1588 reactions and 1755 metabolites, per text of paper (1243 reactions and 1432 species in published SBML file).

Compartmentalization Inherits AraGEM compartmentalization information including mitochondrial, plastidic, peroxisomal and cytosolic compartments (as well as a plasmodesmata pseudo-compartment for facilitating transport between mesophyll and bundle sheath).

Consistency checking Not described in detail. In practice, carbon is not conserved and polymerization reactions are not handled properly.

Results and comparison to experiment Reproduces the classical C4 cycle; mesophyll/bundle sheath distribution of chloroplast reaction flux under photosynthetic condition correlated well with mesophyll/bundle sheath maize chloroplast proteomics studies.

Gene-reaction association Extensive, the main strength of the paper. Reactions have been associated, where possible, with genes in maize (total 11623), sorghum (3557), and sugarcane (3881).

Biomass sink reaction The text of the paper indicates the following biomass components were considered: carbohydrates, cell wall components, amino acids, nucleotides, one fatty acid (palmitic acid), biotin, CoA, riboflavin, folate, chlorophyll, nicotinamide, thiamine, ubiquinone.

Reproducibility Very poor. The SBML model distributed with the paper is inconsistent with the text in many ways, from the total number of reactions and species onwards, and is missing reactions necessary for the production of some biomass components, preventing the reproduction of any results. Inquiries to the authors yielded an 'updated version' as a collection of Matlab files and no explanation of or comment on the discrepancies.

Pilalis Brassica model [40]

Description A model of central metabolism in the Brassicacae emphasizing oil accumulation in developing seeds.

Method of construction No genome for *B. napus* was available, so the model was based on the related *Arabidopsis* (also a member of the Brassicacae) instead. Reactions were taken from AraCyc 6.0, and those reactions not

active in biomass-producing flux solutions were removed, effectively excluding secondary metabolism from the model. It is not clear from the description whether any gap-filling was necessary.

Metabolic scope Primary metabolism including glycolysis, TCA cycle, mitochondrial respiration, pentose phosphate pathway, light and dark reactions of photosynthesis (no photorespiration), amino acid, lipid and starch synthesis,

Size 313 reactions and 262 metabolites.

Compartmentalization Cytosol, mitochondrion, and chloroplast, with reactions manually assigned based on literature or textbook sources or the BRENDA database.

Consistency checking Not discussed in detail.

Results and comparison to experiment For growth on a medium of sucrose, alanine and glutamine, with realistic limitations on uptake rates, predicted growth rates agreed well with observed growth rates. Flux variability results were qualitatively consistent with the targets of the WRINKLED1 transcription factor having significant control over oil synthesis.

Biomass sink reaction Oil, starch and protein, with amino acid and triglyceride compositions specified from existing data.

iRS1563 (maize) [41]

Description A large-scale maize-specific CBM.

Method of construction Reactions in AraGEM were associated to *Arabidopsis* genes and maize orthologs of those genes were determined; maize genes

not assigned a function in this process were annotated by BLAST searches against an NCBI database, and further manual curation was performed, including the addition of some reactions without gene associations but with direct literature evidence. Some reactions were included to facilitate two-cell simulations but note that these were not actually performed.

Metabolic scope Primary metabolism, photosynthesis, photorespiration, lignin biosynthesis, flavonoid biosnythesis, other secondary metabolite synthesis pathways.

Size 1985 reactions, 1825 metabolites.

Compartmentalization Partial, including cytoplasm, plastid, peroxisome, mitochondrion, vacuole, and extracellular compartments. Compartmentalization was based on plant proteomics database information (PPDB, SUBA), with reactions assumed to be cytoplasmic without specific information otherwise, and transporters added as necessary to allow flux. In some cases this led to weird results, e.g., the Calvin cycle is split between chloroplast and cytoplasm in a non-traditional way.

Consistency checking All reactions elementally and charge balanced.

Results and comparison to experiment Predictions agree with yield changes associated with mutations in lignin biosynthesis genes.

Gene-reaction associations 1563 genes associated to reactions.

Biomass sink reaction Determined from literature data on dry weight composition of maize plants; includes amino acids, carbohydrates, cell wall components, lipids, organic acids, nucleotides, ions.

iRS1597 (Arabidopsis) [41]

Description To allow fair comparison between the iRS1563 maize model and Arabidopsis, the authors prepared a revised version of AraGEM, incorporating information from new gene annotations.

Method of construction Some reactions with conservation issues were removed from AraGEM and 228 reactions and associated metabolites were added; additional gene-protein-reaction associations were established. The source of the new reactions and genetic information is not specified.

Metabolic scope As for AraGEM, with improved coverage of some secondary biosynthesis pathways.

Size 1798 reactions, 1820 metabolites.

Compartmentalization The same compartments as AraGEM were used. Some old compartment assignments were revised based on Uniprot and the subcellular proteomics database SUBA.

Consistency checking Not discussed.

Results and comparison to experiment None.

Biomass sink reaction Not discussed.

Gene-reaction associations To 1597 genes.

bna572 (Brassica) [42]

Description Much like the model of Pilalis et al above, bna572 describes central metabolism and storage compound synthesis in developing seeds of *B. napus*.

Method of construction Developed manually from descriptions of known pathways in the literature and the KEGG and AraCyc databases (a 'bibliomic' reconstruction). *B. napus* and *Arabidopsis* .

Metabolic scope Primary metabolism including glycolysis, pentose phosphate pathway, TCA cycle, glyoxylate cycle, beta-oxidation, mitochondrial respiration, light and dark reactions of photosynthesis, photorespiration.

Size 572 reactions, 376 metabolites (counting metabolites present in multiple compartments only once).

Compartmentalization Fully compartmentalized into nine subcellular compartments (apoplast, cytosol, peroxisome, mitochondrial intermembrane space, mitochondrial inner membrane, mitochondrial matrix, plastid stroma, thylakoid membrane, thylakoid lumen,) based on literature information.

Consistency checking All reactions elementally balanced but conservation of protons was generally not enforced.

Results and comparison to experiment 33 fluxes were determined uniquely by the requirement that the solution use substrates and light as efficiently as possible; (under photoheterotrophic conditions); these predicted fluxes compared well to fluxes measured in previous 13C-MFA experiments [43].

Biomass sink reaction Includes oil, protein, starch, sucrose, glutamine, cell wall components, and nucleic acids. Biomass fractions were determined experimentally for embryos developing under various conditions, with detailed composition of proteins, oils, nucleic acids, etc., taken from prior work.

Gene-reaction associations *Arabidopsis* genes corresponding to many of the reactions are provided.

Mintz-Oron 'compartmentalized' model (Arabidopsis) [44]

Description A comprehensive Arabidopsis reconstruction created as a basis for the generation of tissue-specific submodels.

Method of construction From AraCyc and KEGG, with automatic gap-filling with preference for reactions attested in other plant species.

Metabolic scope 176 metabolic functions tested, including synthesis and degradation of secondary metabolites.

Size 1363 reactions among 1078 species.

Compartmentalization Extensive, including "cytosol, plastid, mitochondrion, endoplasmic reticulum, peroxisome, vacuole, and Golgi apparatus"; reactions assigned using information from SUBA and an automatic method that minimizes the number of transport reactions necessary.

Consistency checking All reactions automatically checked for proton and oxygen balance (and possibly other atom balances as well).

Results and comparison to experiment Subcellular localization predictions compared well to independent subcellular metabolomics data; predictions for response to knockdown of pyruvate kinase in the seed-specific model correlated significantly with independent 13C-MFA results.

Biomass sink reaction Biomass components include amino acids, sugars, cell wall components, nucleic acids, coenzyme A and palmitate; composition apparently based on AraGEM.

Gene-reaction associations 1065 included.

Poolman rice model [45]

Description Genome-scale model of a rice leaf cell, used to study responses to varying light levels.

Method of construction From RiceCyc, supplemented with "modules" providing key reactions in chloroplast and mitochondrion (including Calvin cycle and light reactions, TCA cycle and mitochondrial electron transport chain, etc.;) mitochondrial module adapted from AraMeta.

Metabolic scope Production of key biomass components under photosynthetic conditions.

Size 1736 reactions among 1484 species.

Compartmentalization Partial compartmentalization of key reactions in the chloroplast and mitochondrial modules.

Consistency checking Automated checks for atomic balance for individual reactions, careful handling of polymerization reactions, checks for overall conservation of carbon, nitrogen, phosphorus and sulfur; checks to ensure ATP and reducing equivalents cannot be supplied when no input or output to the system is allowed.

Results and comparison to experiment Many responses of mitochondrial respiration and photorespiration to varying light levels, changes in number of photons required per C assimilated, etc., qualitatively agree with literature results; some experimental observations not predicted by the model are also identified.

Biomass sink reaction Biomass components produced included cell wall components, amino acids, nucleotides, lipids, and starch.

Gene-reaction associations 790 reactions with associated genes.

Cheung 'diel' model (Arabidopsis/generic CAM plant) [46]

Description An integrated model for diurnal variation of metabolism in C3 or CAM (Crassulacean acid metabolism) plants. (Like C4 plants, the CAM plants also suppress photorespiration by restricting Rubisco activity to a high-CO_2 environment, but do so temporally rather than spatially: carbon transiently fixed by PEPC at night is released for fixation by Rubisco during the day.)

Method of construction Two copies of an updated version of AraMeta [47] were used to represent the day and night phases of metabolism, with transport between them corresponding to metabolite accumulation during one phase for use in the next.

Metabolic scope Describes photosynthetic production of sucrose and amino acids for export to the rest of the plant through the leaf veins.

Size 5609 reactions among 5505 species.

Compartmentalization In addition to the notional day and night compartments, chloroplast, mitochondrion, peroxisome and vacuole.

Results and comparison to experiment Qualitatively reproduces experimentally observed aspects of diurnal shifts in C3 and CAM metabolism.

Others The maize genome-scale model of Simons et al. [48], a successor to iRS1563, is discussed below, as is a multi-organ model for barley [49].

1.4 Incorporating Rubisco kinetics in constraint-based models

The standard method for making quantitative predictions in a constraint-based model is flux balance analysis (FBA) [50], which predicts reaction rates $v_1, v_2, \ldots v_N$ in a metabolic network by optimizing a biologically relevant function of the rates subject to the requirement that the system reach an internal steady state,

$$\begin{aligned} \max_{(v_1,v_2,\ldots,v_N)\in\mathbb{R}^N} \quad & f(\mathbf{v}) \\ \text{s.t.} \quad & S\cdot\mathbf{v}=\mathbf{0}, \end{aligned} \tag{1.1}$$

where the stoichiometry matrix S is determined by the network structure as discussed above. Assuming the objective function $f(\mathbf{v})$ is linear, this is a linear programming problem, which may be readily and efficiently solved by any of a number of well-established, user-friendly computational tools.

However, photosynthesis is difficult to describe using this approach because the relationship between the rate v_c of carbon fixation by Rubisco and the rate v_o of the Rubisco oxygenase reaction depends nonlinearly on the ratio of the local oxygen and carbon dioxide concentrations (here expressed as equivalent partial pressures),

$$\frac{v_o}{v_c} = \frac{1}{S_R}\frac{P_{O2}}{P_{CO2}} \tag{1.2}$$

where S_R is the specificity of Rubisco for CO_2 over O_2. In the C4 case, the CO_2 level in the bundle sheath compartment is itself a function of the rates of the reactions of the C4 carbon concentration system and the rate of diffusion of CO_2 back to the mesophyll.

With the addition of (1.2), the problem (1.1) becomes nonlinear and cannot be solved with typical FBA tools; instead (as the problem is also nonconvex [51]), a

general-purpose nonlinear programming algorithm is required to numerically solve it.

Prior constraint-based models of plant metabolism have typically ignored the constraint (1.2) or assumed the oxygen and carbon dioxide levels P_{O2} and P_{CO2} are known and fixed v_o/v_c accordingly [37,41]. While this approach is suitable for mature C4 leaves under many conditions, where v_o/v_c is approximately zero, it may break down in some of the most important targets for simulation: developing tissue, mutants, and C3-C4 intermediate species, where P_{CO2} in the bundle sheath compartment is not necessarily high.

In other recent work, a high-level physiological model was used to determine v_o, v_c, and other key reaction rates given a few parameters, which were then fixed in order to solve eq. (1.1) [52] . This method yields realistic solutions, but its application is limited by the lack of a way to set the necessary phenomenological parameters (e.g., the maximum rate of PEP regeneration in the C4 cycle) based on lower-level, per-gene data (e.g., from transcriptomics or experiments on single-gene mutants).

Here, we introduce a more general solution to the problem: incorporating the nonlinear constraint (1.2) directly into the optimization problem (1.1) and solving the resulting nonlinear program numerically with the IPOPT package [53], using a new computational interface that we have developed, which allows rapid, interactive development of nonlinearly-constrained FBA problems from metabolic models specified in SBML format [54].

Using a new model describing interacting mesophyll and bundle sheath cells in the leaves of either *Zea mays* or generic C3, C4, or intermediate plants, based

on a novel genome-scale reconstruction of the maize metabolic network developed with particular attention to photosynthesis and related processes, we confirm that this approach can reproduce the nonlinear responses of well-validated, high-level physiological models of C4 photosynthesis [25], while also providing detailed predictions of fluxes throughout the network. We then apply these tools to study two key topics in the field of C4 photosynthesis: the shifts in metabolic state along the gradient in from immature to mature tissue in developing leaves, and metabolic adaptation along the evolutionary path from the C3 state to the C4 state.

1.5 Outline

In chapter 2 the new genome-scale metabolic model for the C4 grass *Zea mays*, new software for metabolic modeling with nonlinear constraints, and a novel method for inferring metabolic activity from enzyme expression levels are combined to interpret experimental data from the developing maize leaf. In chapter 3, the model and nonlinear modeling software are combined with techniques from theoretical chemistry to simulate plausible evolutionary paths through the fitness landscape connecting the C3 and C4 phenotypes. Appendix A details the process of reconstruction of the maize metabolic network model. Appendix B contains additional tables of results from chapter 3. Appendix C briefly presents the results of a related study in which constraint-based techniques were used to design a hypothetical metabolic pathway that could bypass the photorespiratory system without releasing CO_2. Appendix D discusses some theoretical and practical issues in nonlinear optimization with implications for the design and solution of nonlinear constraint-based metabolic modeling problems.

CHAPTER 2

MULTISCALE MODELING OF METABOLISM IN THE DEVELOPING MAIZE LEAF

2.1 Introduction[1]

Maize leaves display a developmental gradient along the base-to-tip direction, with young cells in the immature base and fully differentiated cells at the tip [55, 56]. Here, after validating the method for nonlinear metabolic modeling described above and introducing a new maize metabolic reconstruction, we combine the results of enzyme assay measurements and multiple RNA-seq experiments and apply a new method to infer the metabolic state at points along a developing maize leaf (Fig. 2.1a) using a model of mesophyll and bundle sheath tissue in fifteen segments of the leaf, interacting through vascular transport of sucrose, glycine, and glutathione. We compare our results to radiolabeling experiments.

2.2 Results

2.2.1 Metabolic reconstruction of *Zea mays*

A novel genome-scale metabolic model was generated from version 4.0 of the CornCyc metabolic pathway database [57] and is presented in two forms. The

[1]The material in this chapter and Appendix A is adapted from the paper "Multiscale metabolic modeling of C4 plants: connecting nonlinear genome-scale models to leaf-scale metabolism in developing maize leaves", Eli Bogart and Christopher R. Myers, arXiv:1502.07969 [q-bio-MN] (2015), which has been submitted and is currently under review.

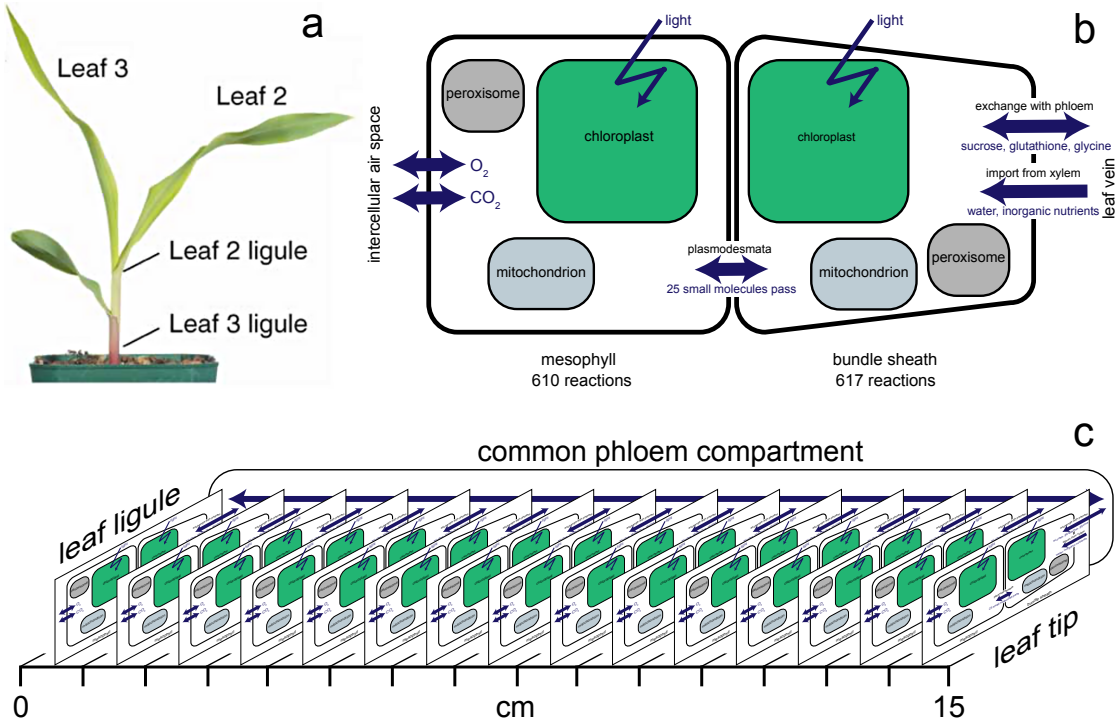

Figure 2.1: **Maize plant and models.** (a) Nine-day-old maize plant (image from [55]). (b) Organization of the two-cell-type metabolic model, showing compartmentalization and exchanges across mesophyll and bundle sheath cell boundaries. (c) Combined 121-compartment model for leaf 3 at the developmental stage shown in (a). Fifteen identical copies of the model shown in (b) represent 1-cm segments from base to tip.

comprehensive reconstruction involves 2720 reactions among 2725 chemical species, and incorporates CornCyc predictions for the function of 5204 maize genes, with 2064 reactions associated with at least one gene. A high-confidence subset of the model, excluding many reactions not associated with manually curated pathways or lacking computationally predicted gene assignments as well as all reactions which could not achieve nonzero flux in FBA calculations, involves 635 reactions among 603 species, with 469 reactions associated with a total of 2140 genes.

Both the comprehensive and high-confidence models can simulate the production of all major maize biomass constituents (including amino acids, nucleic

acids, fatty acids and lipids, cellulose and hemicellulose, starch, other carbohydrates, and lignins, as well as chlorophyll) under either heterotrophic or photoautotrophic conditions and include chloroplast, mitochondrion, and peroxisome compartments, with key reactions of photosynthesis (including a detailed representation of the light reactions), photorespiration, the NADP-ME C4 cycle, and mitochondrial respiration localized appropriately. Gene associations for reactions present in more than one subcellular compartment have been refined based on the results of subcellular proteomics experiments and computational predictions (as collected by the Plant Proteomics Database, [58]) to assign genes to reactions in appropriate compartments.

A model for interacting mesophyll and bundle sheath tissue in the leaf was created by combining two copies of the high-confidence model, with transport reactions to represent oxygen and CO_2 diffusion and metabolite transport through the plasmodesmata, and restricting exchange reactions appropriately (nutrient uptake from the vascular system to the bundle sheath, and gas exchange with the intercellular airspace to the mesophyll). A schematic of the two-cell model is shown in Fig. 2.1b.

Both single-cell versions of the model and the two-cell model, designated iEB5204, iEB2140, and iEB2140x2 respectively (based on the primary author's initials and number of genes included, according to the established naming convention [59]), have been made available in SBML format (e.g., as ancillary files to [60]).

2.2.2 Nonlinear flux-balance analysis

To solve nonlinear optimization problems incorporating the constraints discussed above, we developed a Python package which – given a model in SBML format, arbitrary nonlinear constraints, a (potentially nonlinear) objective function, and all needed parameter values – infers the conventional FBA constraints of eq. (1.1) from the structure of the network, automatically generates Python code to evaluate the objective function, all constraint functions, and their first and second derivatives, and calls IPOPT through the pyipopt interface [61]. Source code for the package is available in the ancillary files of [60] and online (`http://github.com/ebogart/fluxtools`). The software has been used to successfully solve nonlinear FBA problems with over 84000 variables and 62000 constraints.

Figure 2.2 demonstrates that, as expected, optimizing the rate of CO_2 assimilation in the two-cell-type model with nonlinear kinetic constraints [eqs. (2.3), (2.4), (2.5)] produces predictions consistent with the results of the physiological model of [25]. Note that the effective value of one macroscopic physiological parameter may be governed by many microscopic parameters in the genome-scale model. In the figure, the effective maximum PEP regeneration rate V_{pr} is controlled by the maximum rate of three decarboxylase reactions in the bundle sheath compartment, but with an appropriate choice of parameter values any of at least 10 reactions of the C4 system could become the rate-limiting step in PEP regeneration, and in the calculations below, expression levels for any of the 42 genes associated with these reactions (Table 2.1) could influence the net PEP regeneration rate.

reaction	name in model	associated genes
malate dehydrogenase (NADP)	MALATE_DEHYDROGENASE_NADP__RXN_chloroplast	1
alanine aminotransferase	ALANINE_AMINOTRANSFERASE_RXN	10
aspartate aminotransferase	ASPAMINOTRANS_RXN	7
NAD-malic enzyme	EC_1_1_1_39	2
NADP-malic enzyme (cytosol)	MALIC_NADP_RXN	4
NADP-malic enzyme (chloroplast)	MALIC_NADP_RXN_chloroplast	2
PEPCK	PEPCARBOXYKIN_RXN	6
PPDK	PYRUVATEORTHOPHOSPHATE_DIKINASE_RXN_chloroplast	2
adenylate kinase	ADENYL_KIN_RXN_chloroplast	6
pyrophosphatase	INORGPYROPHOSPHAT_RXN_chloroplast	2

Table 2.1: **Detailed parameters contributing to the effective PEP regeneration rate: reactions in the genome-scale model which contribute to the effective maximum PEP regeneration capacity, and the number of genes associated with each.** In addition to the reactions listed, transport capacities of pyruvate, PEP, alanine, aspartate and malate across the plasmodesmata and pyruvate, PEP, malate and oxaloacetate across the chloroplast inner membrane could limit this rate; the model currently associates no genes with these transport reactions.

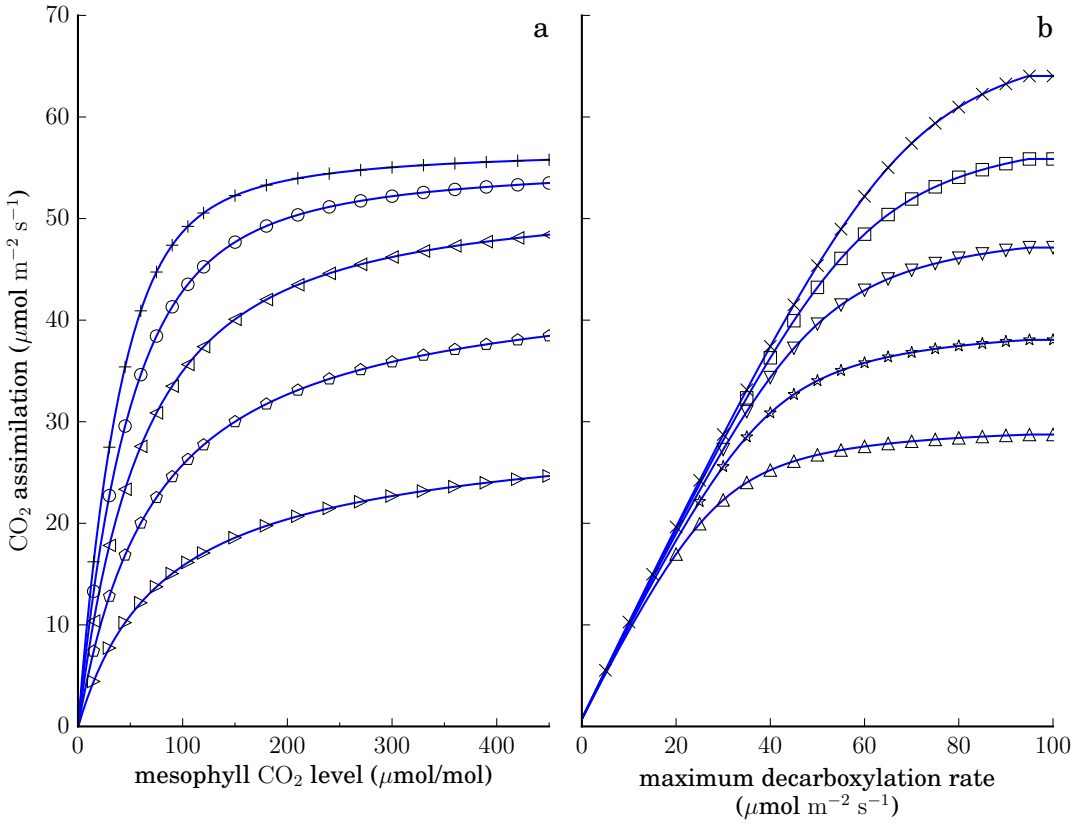

Figure 2.2: **CO_2 assimilation rates (A) predicted by the C4 photosynthesis model of [25], solid lines, and the present nonlinear genome-scale model (markers) maximizing CO_2 assimilation with equivalent parameters.** Left, A vs mesophyll CO_2 levels with varying PEPC levels (top to bottom, $V_{p,\max} = 110$, 90, 70, 50, and 30 µmol m^{-2} s^{-1}). Right, A vs total maximum activity of all bundle sheath decarboxylase enzymes (equivalent to the maximum PEP regeneration rate V_{pr}) at varying Rubisco levels (top to bottom, $V_{c,\max} = 70$, 60, 50, 40, and 30 µmol m^{-2} s^{-1}). Other parameters as in Table 4.1 of [25], except with nonphotorespiratory respiration rates $r_d = r_m = 0$.

2.2.3 Flux predictions in the developing leaf based on multiple data channels

To explore variations in metabolic state along the leaf developmental gradient, we combined the RNA-seq datasets of Wang et al. [62] and Tausta et al. [63] to estimate expression levels (as FPKM) for 39634 genes in the mesophyll and

bundle sheath cells at 15 points, representing 1 cm segments of the third leaf of a 9-day-old maize plant, which includes a full gradient of developmental stages. The combined dataset provides expression information for 920 reactions in the two-cell model (460 each in mesophyll and bundle sheath cells).

A whole-leaf metabolic model, iEB2140x2x15, was created from fifteen copies of the two-cell model, each representing a 1-cm segment, interacting through the exchange of sucrose, glycine, and glutathione through a common compartment representing the phloem. The resulting 121-compartment model, Fig. 2.1c, involves 18780 reactions among 16575 metabolites.

Subject to the requirements that reaction rates in each of the 15 segments obey both the FBA steady-state constraints (eq. 1.1) and the nonlinear constraints governing Rubisco kinetics (eqs. 2.3, 2.5, and 2.4, presented in detail below) we determined the set of rates v_{ij} for each reaction i at each segment j which were most consistent with the base-to-tip variation in the gene expression data, by optimizing the objective function

$$F(v) = \sum_{i=0}^{N_r} \sum_{j=1}^{15} \frac{\left(e^{s_i} \left|v_{ij}\right| - d_{ij}\right)^2}{\delta_{ij}^2} + \alpha \sum_{i=0}^{N_r} s_i^2 \tag{2.1}$$

where $N_r = 920$ is the number of reactions associated with at least one gene present in the expression data, d_{ij} and δ_{ij} are the expression data and associated experimental uncertainty for reaction i at leaf segment j, and s_i is an optimizable scale factor associated with reaction i.

Effectively, this calculation – similar to the method of Lee et al. [64] or FALCON [65] – performs a constrained least-squares fit of the fluxes to the expression data. Allowing the scale factors s_i to vary emphasizes agreement between fluxes and data in their trend along the developmental gradient, rather than in

their absolute value: if the data associated with reaction R_i has average value 100 FPKM, a solution in which R_i has mean flux 10 μmol m^{-2} s^{-1} but correlates well with the data can achieve (with appropriate choice of scale factor) a lower cost than a solution in which R_i has mean flux 100 μmol m^{-2} s^{-1} but is anticorrelated. The penalty term $\alpha \sum s_i^2$ favors solutions in which, generally, reactions with larger associated expression data carry higher fluxes. The parameter α controlling the tradeoff between these criteria was set arbitrarily to 1.0 in the work presented here. We require $s_a = s_b$ if reactions a and b are mesophyll and bundle sheath instances of the same reaction.

To constrain the overall scale of the fluxes and further improve accuracy, we incorporated enzyme activity assay data from [62] for seventeen enzymes (including Rubisco and PEPC) along the 15 leaf segments as additional constraints on the optimization problem, requiring for each enzyme k and segment j

$$E_{jk} \geq |v_{k1}| + \ldots + |v_{kn}| \tag{2.2}$$

where E_{jk} is the measured maximal activity of the enzyme at that segment and the sum on the right hand side includes all the reactions which represent enzyme k in the mesophyll, bundle sheath, and subcompartments of those cells if applicable.

Solving the optimization problem yielded predictions for reaction rates and other variables. Upper and lower bounds on selected variables were determined through flux variability analysis (FVA) [66], allowing the objective function to increase by 0.1% from its optimal value.

Predicted source-sink transition

As shown in Fig. 2.3, in the outer, more photosynthetically developed, portion of the leaf, our optimal fit predicts net CO_2 uptake, with most of the assimilated carbon incorporated into sucrose and exported to the phloem. Near the base of the leaf, sucrose is predicted to be imported from the phloem and used to drive a high rate of biomass production, with some concomitant net release of CO_2 to the atmosphere by respiration.

This transition between a carbon-exporting source region and a carbon-importing sink region is well known, and the predicted transition point between the two, approximately 6 cm above the base of the leaf, can be compared to the ^{14}C-labeling results of Li et al. [55] in the same experimental conditions. Fig. 2.3b shows the location of labeled carbon in leaf 3 after feeding labeled CO_2 to leaf 2 (center image) or leaf 3 (bottom image). Li et al. [55] identified the sink region as the lowest 4 cm of the leaf; the transition is not perfectly sharp and quantitative comparison of exchange fluxes is not possible, but the nonlinear FBA results appear to slightly overestimate the size of the sink region.

Agreement might be improved under a different assumption about net sucrose import or export by leaf 3 (here, we have assumed that the import visible in the center image is exactly balanced by the export suggested by the high density of labeled carbon at the absolute base in the lower image).

The net rate of CO_2 assimilation predicted in the outer, most mature leaf segments, 8-11 $\mu mol\, m^{-2}\, s^{-1}$, is lower than that typically measured in more mature maize plants (e.g., rates of 20-30 $\mu mol\, m^{-2}\, s^{-1}$ in 22-day-old wild-type plants under comparable conditions [9]), but photosynthetic capacity may still be in-

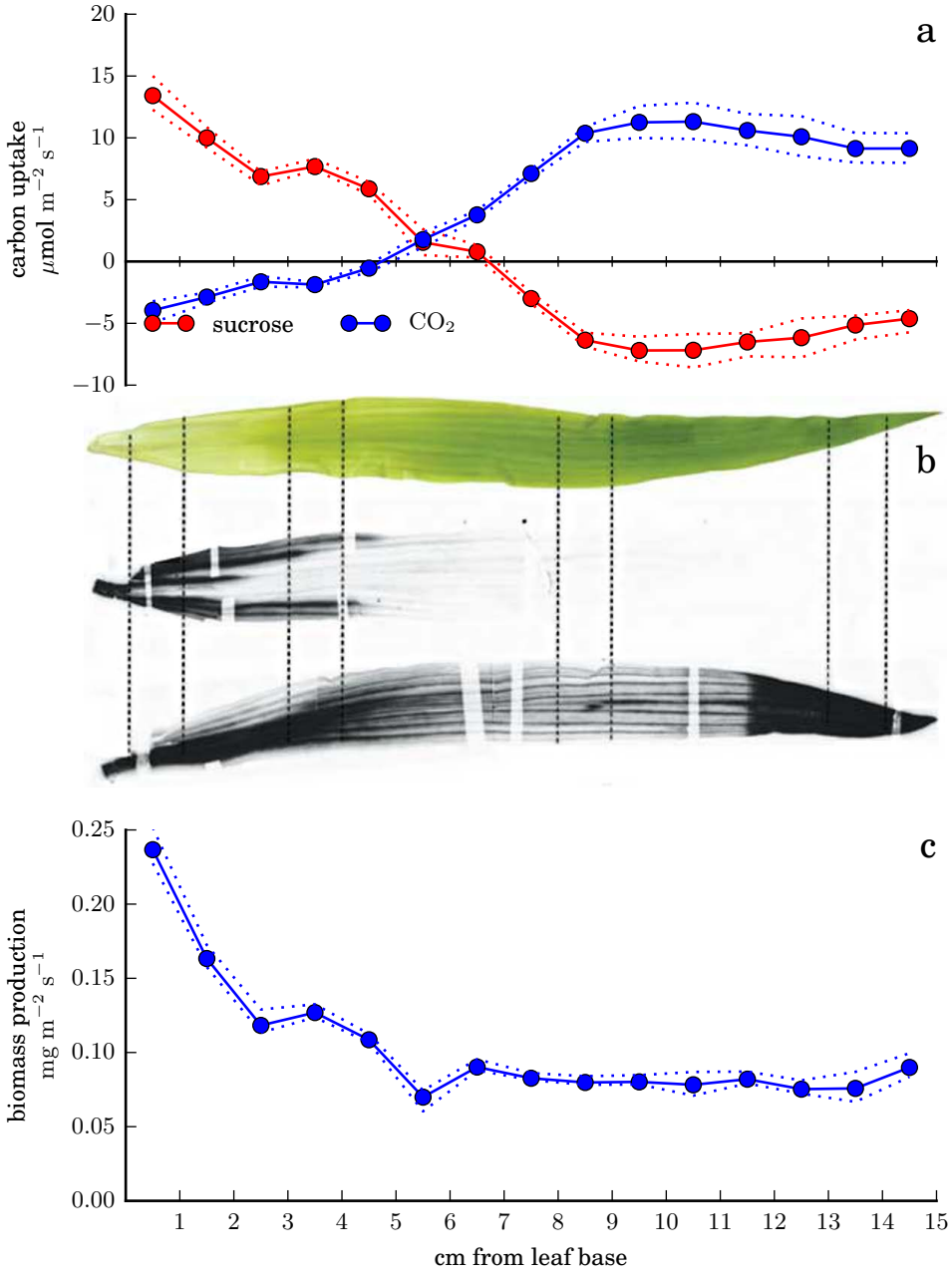

Figure 2.3: **Source-sink transition along the leaf as predicted by optimizing the agreement between fluxes in the nonlinear model and RNA-seq data.** (a) Predicted rates of exchange of carbon with the atmosphere and phloem along the leaf. (b) Experimental observation of the source-sink transition, reproduced from [55]. Upper image, photograph of leaf 3; middle image, autoradiograph of leaf 3 after feeding $^{14}CO_2$ to leaf 2; lower image, autoradiograph of leaf 3 after feeding $^{14}CO_2$ to the tip of leaf 3. (c) Total biomass production in the best-fitting solution. In panels a and c, dotted lines indicate minimum and maximum predicted rates consistent with an objective function value no more than 0.1% worse than the optimum.

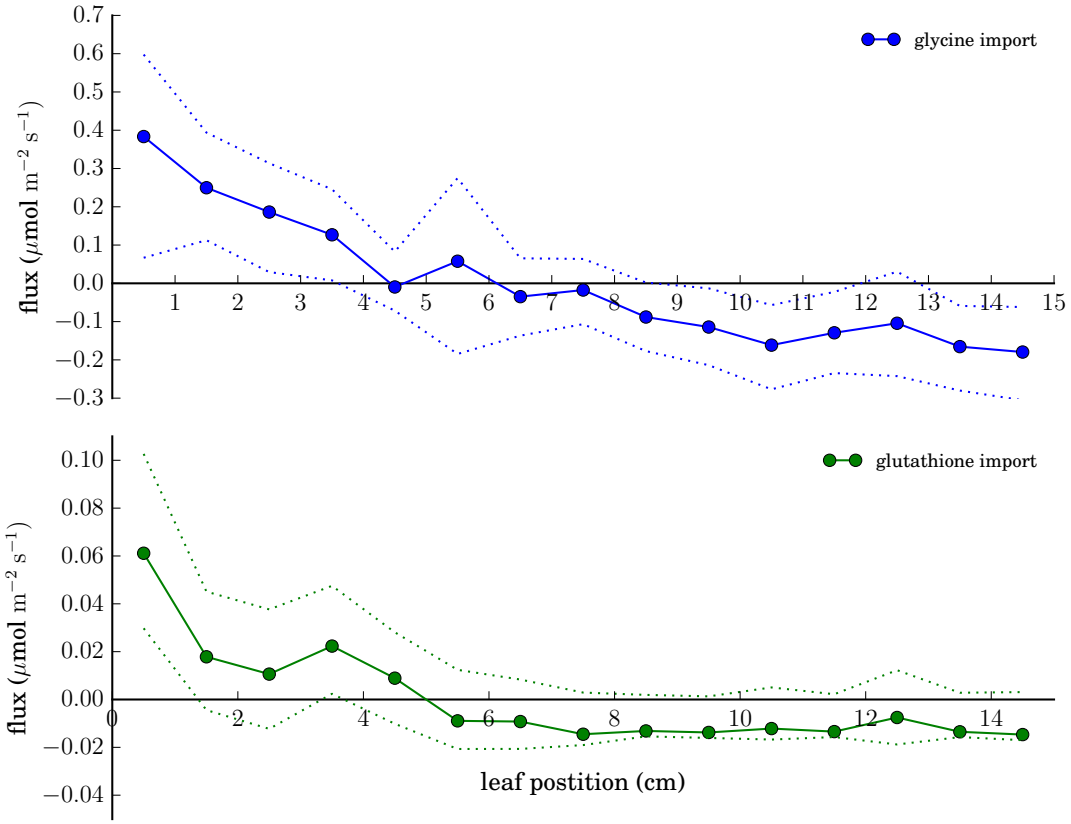

Figure 2.4: **Transport of nitrogen (upper panel) and sulfur (lower panel) through the phloem in the best-fitting solution.** Dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.

creasing even in these segments.

In addition to sucrose, glycine and glutathione are predicted to be exported from the source region through the phloem and reimported by the sink region, consistent with our expectations that nitrogen and sulfur reduction will occur preferentially in the photosynthesizing region (Figure 2.4). Note that this behavior emerges from the data even though there is no explicit requirement in the model that net phloem transport occur in a basipetal direction.

Predicted C4 system function

Figure 2.5 shows predicted rates of key reactions of the C4 system and CO_2 and O_2 levels in the bundle sheath. As expected, the model predicts that a C4 cycle will operate in the source region of the leaf, elevating the CO_2 level in the bundle sheath. The CO_2 level is also elevated in the source region; this is an immediate consequence of respiration in the bundle sheath and eq. (2.5). It may be overestimated here because we have assumed a constant value for the bundle sheath CO_2 conductivity (as measured by Bellasio et al. [67]); in fact, gene expression associated with synthesis of the diffusion-resistant suberin layer between bundle sheath and mesophyll peaks at 4 cm above the leaf base [62], so g_s is presumably higher below that point.

In the Calvin cycle, most reactions are predicted to be bundle-sheath specific, but the reductive phase is active in both cells, with approximately half the 3-phosphoglycerate produced in the bundle sheath transported to the mesophyll and returned as dihydroxyacetone phosphate (Fig. 2.5c); this is a known aspect of NADP-ME C4 metabolism connected to reduced photosystem II activity in the bundle sheath cells [68], which is also predicted here (Figure 2.6). Consistent with conclusions drawn independently from the transcriptomic data, as well as proteomic data from the same system [55, 62, 69], the model does not predict a C3-like metabolic state as a developmental intermediate stage. As expected in maize [70], a significant role for phosphoenolpyruvate carboxykinase (PEPCK) as a decarboxylating enzyme operating in the bundle sheath in parallel with NADP-ME is predicted (Fig. 2.5b).

While the predictions are generally consistent with the standard view of the C4 system in maize, there are minor discrepancies. In the mesophyll, our calcu-

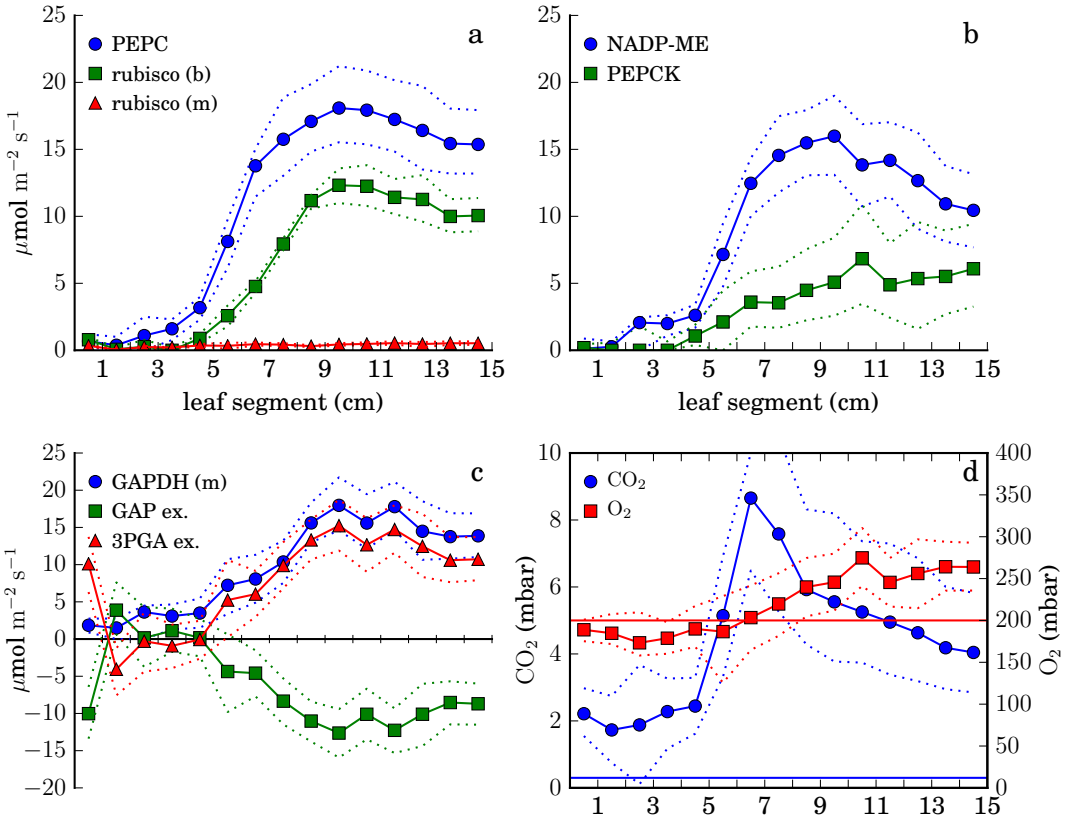

Figure 2.5: **Operation of the C4 system in the best-fitting solution.** (a) Rates of carboxylation by PEPC in the mesophyll and Rubisco in the mesophyll and bundle sheath. (b) Rates of CO_2 release by PEP carboxykinase and chloroplastic NADP-malic enzyme in the bundle sheath. (c) Transport of 3-phosphoglycerate and glyceraldehyde 3-phosphate from bundle sheath to mesophyll (or the reverse, where negative) and glyceraldehyde 3-phosphate dehydrogenation rate in the mesophyll chloroplast, showing the involvement of the mesophyll in the reductive steps of the Calvin cycle throughout the source region. (d) Oxygen and carbon dioxide levels in the bundle sheath. Straight lines show mesophyll levels. Throughout, dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.

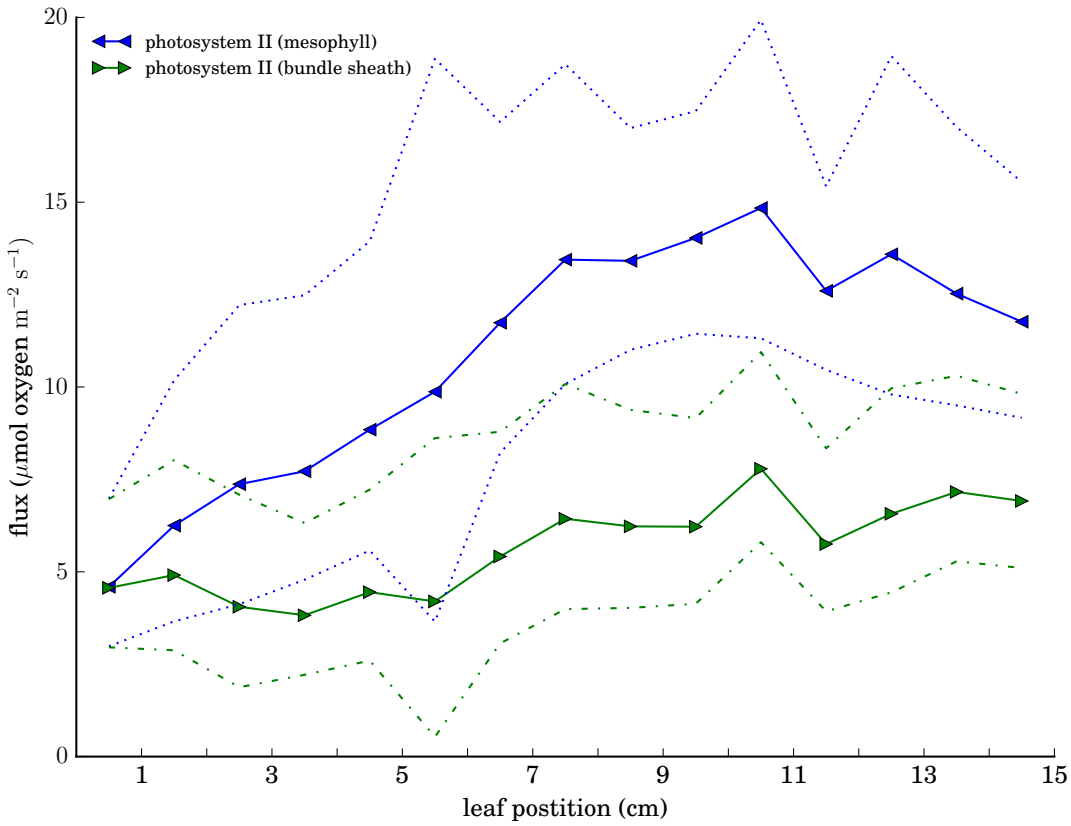

Figure 2.6: **Photosystem II in mesophyll and bundle sheath.** Dashed and dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.

lations predict that malate production occurs in the mitochondrion, rather than the chloroplast. In both mesophyll and bundle sheath, phosphoenolpyruvate is formed by pyruvate-orthophosphate dikinase (PPDK) in the chloroplast at a higher rate than necessary to sustain the C4 cycle; the excess is converted again to pyruvate by pyruvate kinase in the cytoplasm, with the resulting ATP consumed by the model's generic ATPase reaction. Finally, in the bundle sheath, a modest rate of PEPC activity is predicted, recapturing CO_2 only to have it released again by the decarboxylases (Figure 2.7). Further refinement of the associations of genes to reactions in the model might resolve some of these discrepancies.

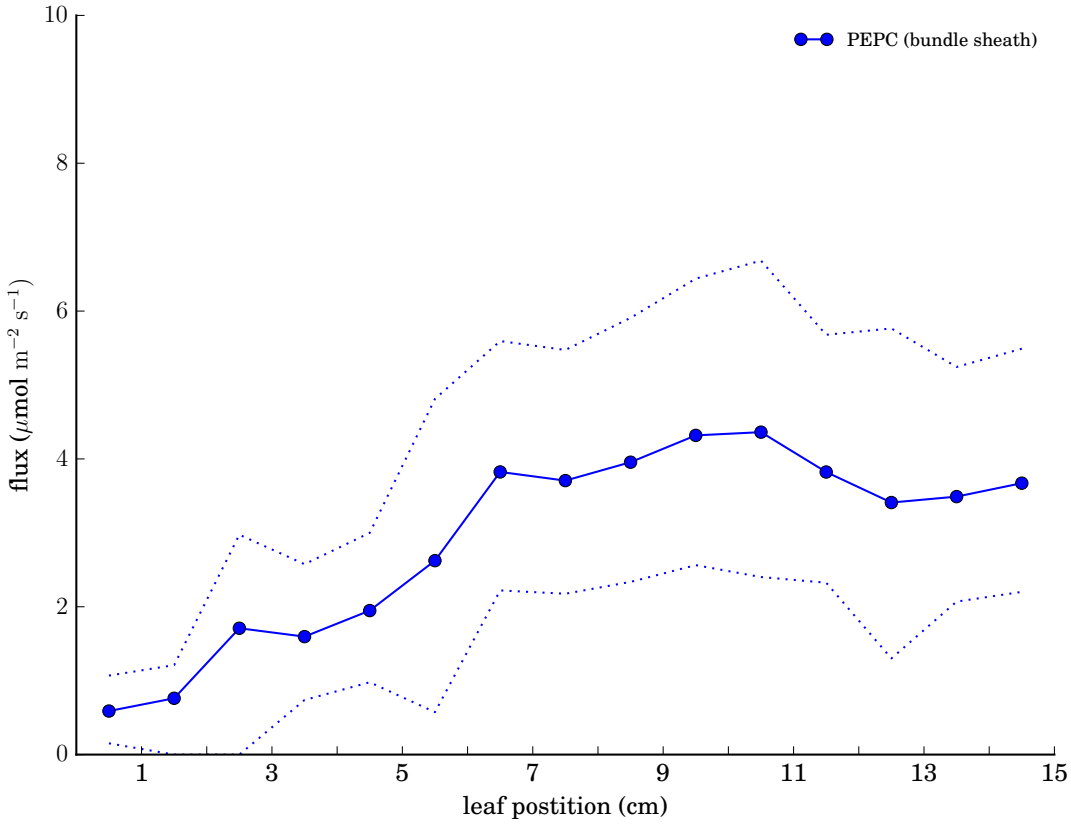

Figure 2.7: **Bundle sheath PEPC flux in the best-fitting solution.** Dotted lines indicate minimum and maximum predicted values consistent with an objective function value no more than 0.1% worse than the optimum.

Global agreement between fluxes and data

Figure 2.8 summarizes overall properties of the predicted fluxes. It is not clear why agreement between data and predicted fluxes is poorer at the base, as shown in Fig. 2.8a. As discussed below, the cell-type-specific RNA-seq data from Tausta et al. [63] does not extend below the fourth segment from the base of the leaf; at the base we have assumed expression levels for all genes are equal in mesophyll and bundle sheath. Though proteomics experiments on the same system [69] generally found limited cell-type specificity at the leaf base, this assumption is likely an oversimplification, and could limit the ability of the al-

gorithm to find a flux prediction consistent with the data there.

For most reactions, the correlation between the base-to-tip expression pattern and the base-to-tip trend in predicted flux is high. The cumulative histogram in Fig. 2.8b shows that the Pearson correlation $r > 0.92$ for more than half of the reactions in the model with associated expression data.

Differences in expression levels between different reactions, however, correlate only weakly with the differences in fluxes between those reactions, as shown for segment 15 in Fig. 2.8c (blue circles). After rescaling fluxes by the optimal per-reaction scale factors, a clear relationship emerges (Fig. 2.8c, red circles), confirming that the scale factors are functioning as intended. Of course we should not expect a perfect correlation between data on transcript levels and predicted fluxes through associated reactions. The limited correlation between fluxes and expression data across different reactions presumably follows, in part, from the imperfect correlation between expression data and protein abundance across different genes, as illustrated in Fig. 2.8d with data from the same experimental system [71], as well as from the different catalytic capabilities of different enzymes, posttranslational regulation, differences in substrate availability, etc.

Reconciling expression data and network structure

Figure 2.9 illustrates the operation of the fitting algorithm in detail, using two regions of the metabolic network with simple structure as examples.

In Fig. 2.9a, expression data for eight reactions of the pathway leading to chlorophyllide a are shown. Expression levels for the different reactions at any

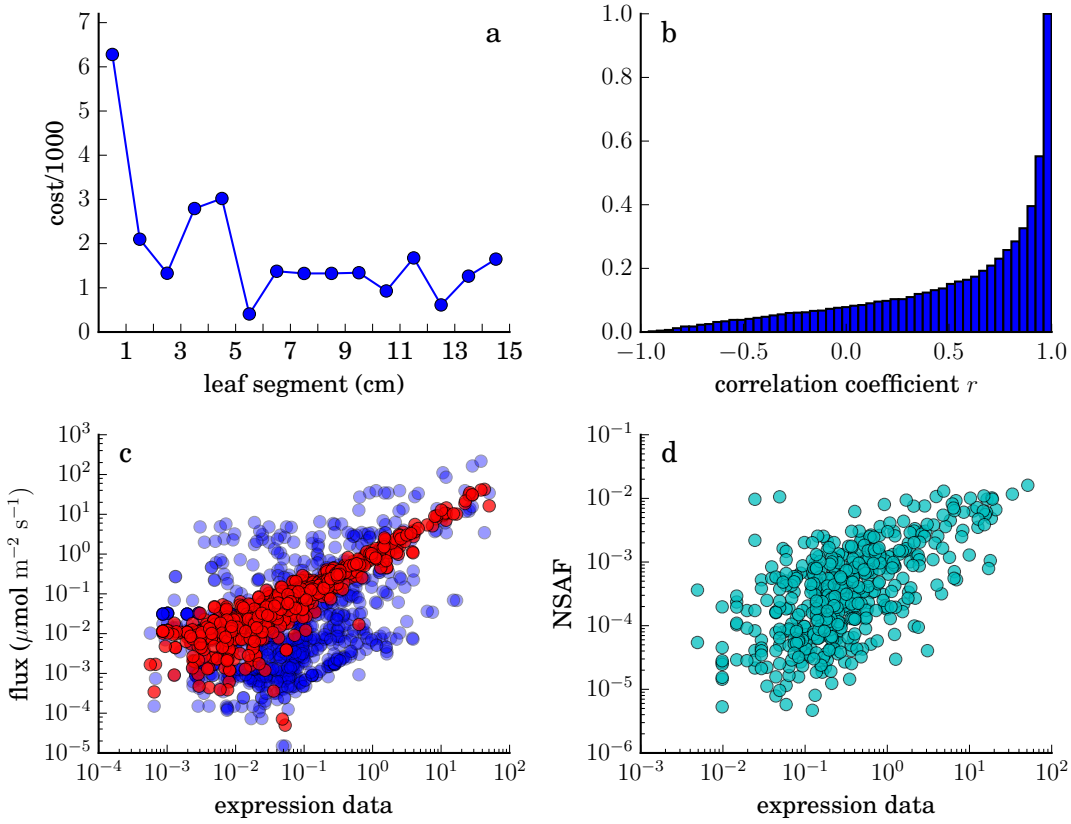

Figure 2.8: **Agreement between RNA-seq data and predicted fluxes.** (a) Contribution of each segment to the objective function (eq. (2.1), excluding costs associated with scale factors). (b) Cumulative histogram of Pearson correlations between data and predicted fluxes for all reactions. (c) Predicted fluxes versus expression data at the tip of the leaf (blue, raw fluxes; red, after rescaling each flux v_i by the optimal factor e^{s_i} of eq. (2.1)). Some outliers with very low predicted flux are not shown. (d) Relationship between RNA-seq and proteomics measurements for 506 proteins in the 14th segment from the base, redrawn from the data of [71]. NSAF, normalized spectral abundance factor.

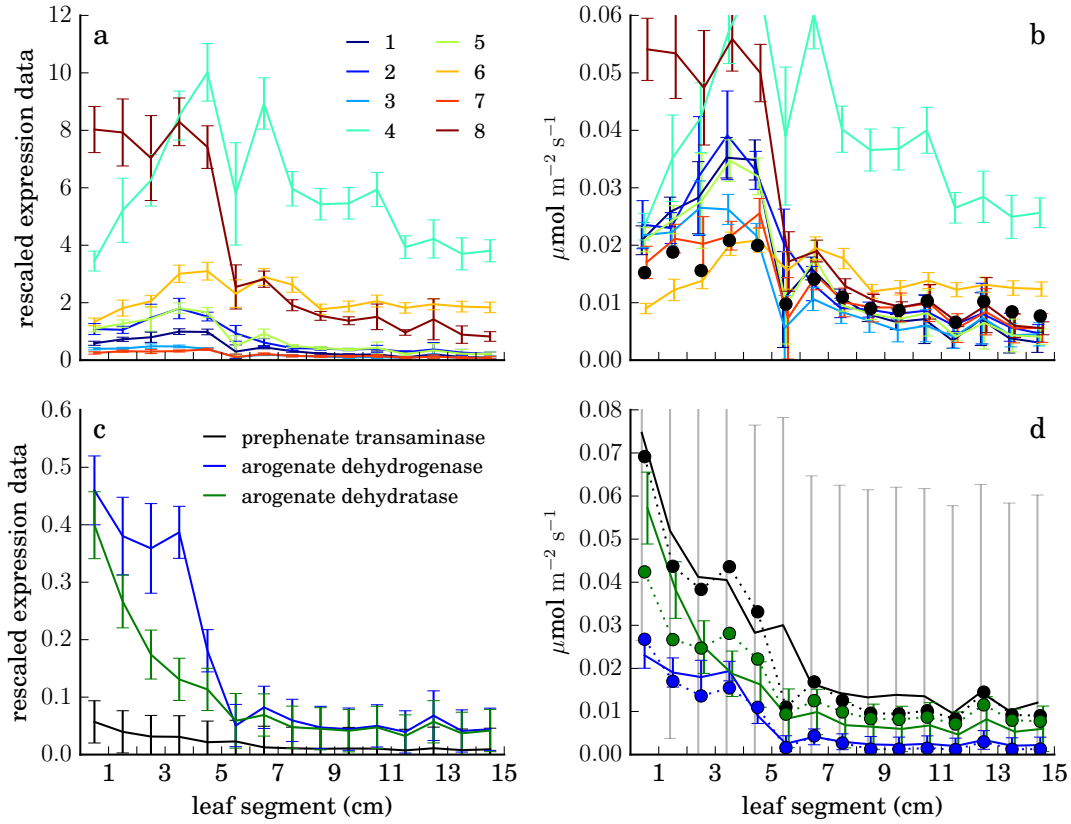

Figure 2.9: **Comparison of RNA-seq data to predicted fluxes for a linear pathway and around a metabolic branch point.** Upper panels, chlorophyllide a synthesis in the mesophyll; lower panels, production of arogenate in the bundle sheath by prephenate transaminase and its consumption by arogenate dehydrogenase and arogenate dehydratase. Left, aggregate RNA-seq data and experimental standard deviations for each reaction rescaled by a uniform factor (see text). Right, same data and errors further rescaled by reaction-specific optimal factors (e^{-s_i}, in the variables of eq. 2.1) to best match data with predicted fluxes (solid circles). Fluxes are equal for all reactions of the linear pathway (1, uroporphyrinogen decarboxylase, 2, coproporphyrinogen oxidase, 3, protoporphyrinogen oxidase, 4, magnesium chelatase, 5, magnesium protoporphyrin IX methyltransferase, 6, magnesium protoporphyrin IX monomethyl ester cyclase, 7, divinyl chlorophyllide a 8-vinyl-reductase, 8, protochlorophyllide reductase). Error bars represent standard deviations of expression measurements across multiple replicates.

point on the leaf may span an order of magnitude or more, but the FBA steady-state assumption requires the rates of all reactions in this unbranched[2] pathway to be equal at each point. Applying the optimal rescaling determined for each reaction's expression data, shown in panel b, allows the flux prediction for the pathway (solid dots) to achieve reasonable agreement with the data. (Note that data for reaction 4 cannot be further scaled down because of the lower limit $\exp(-5)$ on its scale factor $\exp(s_4)$, imposed for technical reasons.)

Figure 2.9c shows data for a three-reaction branch point in aromatic amino acid synthesis. To balance production and consumption of arogenate, the prephenate transaminase flux must equal the sum of the fluxes through arogenate dehydrogenase (to tyrosine) and arogenate dehydratase (to phenylalanine) but expression is consistently lower for the transaminase than the other enzymes. After rescaling (Fig. 2.9d), the data agree well with the stoichiometrically consistent flux predictions (solid dots). The predicted ratio of dehydrogenase to dehydratase flux reflects data for downstream reactions.

Comparison to other methods for integrating RNA-seq data

Figure 2.10 shows predictions that result when the scale factors s_i of eq. (2.1) are fixed to zero. The source-sink transition is apparent but the C4 cycle operates at lower levels, the example pathways of Fig. 2.9 (and a number of others) show little or no activity, and predicted fluxes along the leaf are not as tightly correlated with their associated expression data.

Figure 2.11 shows the metabolic state predicted by applying the expression data for each reaction as an upper bound on the absolute value of the reaction

[2]The branch leading to heme production is not included in the reconstruction.

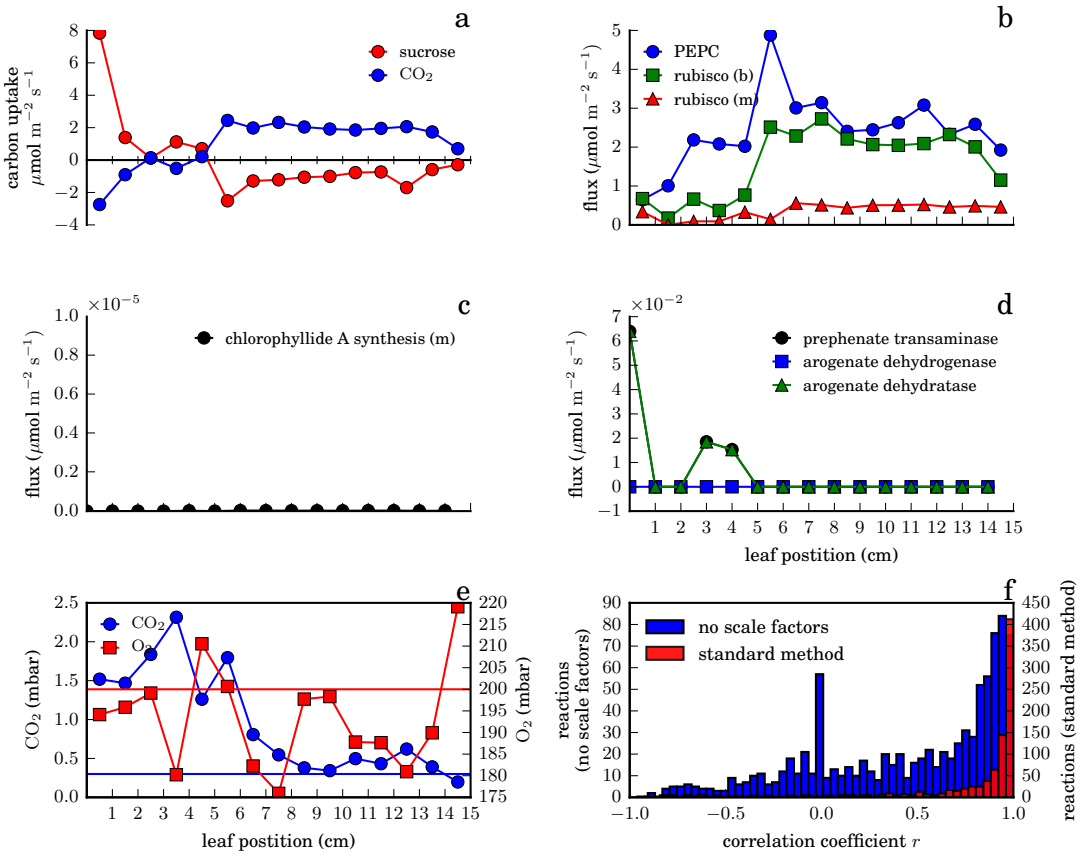

Figure 2.10: **Summary of predictions for the gradient model using the least-squares method without per-reaction scale factors.** In eq. (2.6), $s_i = 0$ for all reactions i. (a) Sucrose and CO_2 uptake rates (compare to figure 2.3a). (b) Rates of carboxylation by PEPC and Rubisco (compare to figure 2.5b). (c) Predicted rate for the reactions of the chlorophyllide A synthesis pathway (compare to figure 2.9b). (d) Predicted rates at the arogenate branch point (compare to figure 2.9d). (e) Predicted oxygen and carbon dioxide levels in the bundle sheath, with straight lines showing mesophyll levels (compare to figure 2.5d). (f) Distribution of correlation coefficients between data and predicted fluxes for each reaction (blue, this method; red, standard method). Correlation coefficients for reactions with zero predicted flux are taken to be zero, resulting in the visible peak in the histogram.

rate as in the E-Flux method [72] to the fifteen-segment model with the same RNA-seq data. The C4 system is predicted to operate, but no source-sink transition is apparent, and typical data-predicted flux correlations are poor. Imposing a realistic biomass composition restores the source-sink transition and somewhat improves correlation between data and fluxes (Figure 2.12; in contrast, as shown in Figure 2.13, fixing the biomass composition has limited effect on the method presented above, except for the suppression of production of some species, such as chlorophyll, which are not included in the composition, and a slightly higher rate of total biomass synthesis at the leaf base, as shown in Figure 2.14). Fluxes predicted by E-Flux are generally smaller than those predicted by the least-squares method, with or without per-reaction scale factors.

Figure 2.15 compares the fluxes predicted at the tip by optimizing agreement with the data through the non-biological objective function (eq. 2.1), fluxes predicted at the tip with an explicit biological objective function (maximizing CO_2 assimilation) constrained by the experimental data in the E-Flux method, and fluxes predicted in an FBA calculation which ignores the data entirely (minimizing total flux while achieving the same CO_2 assimilation rate as predicted at the tip by the least-squares method). Both data-integration methods lead to predictions very different from the unconstrained FBA calculation.

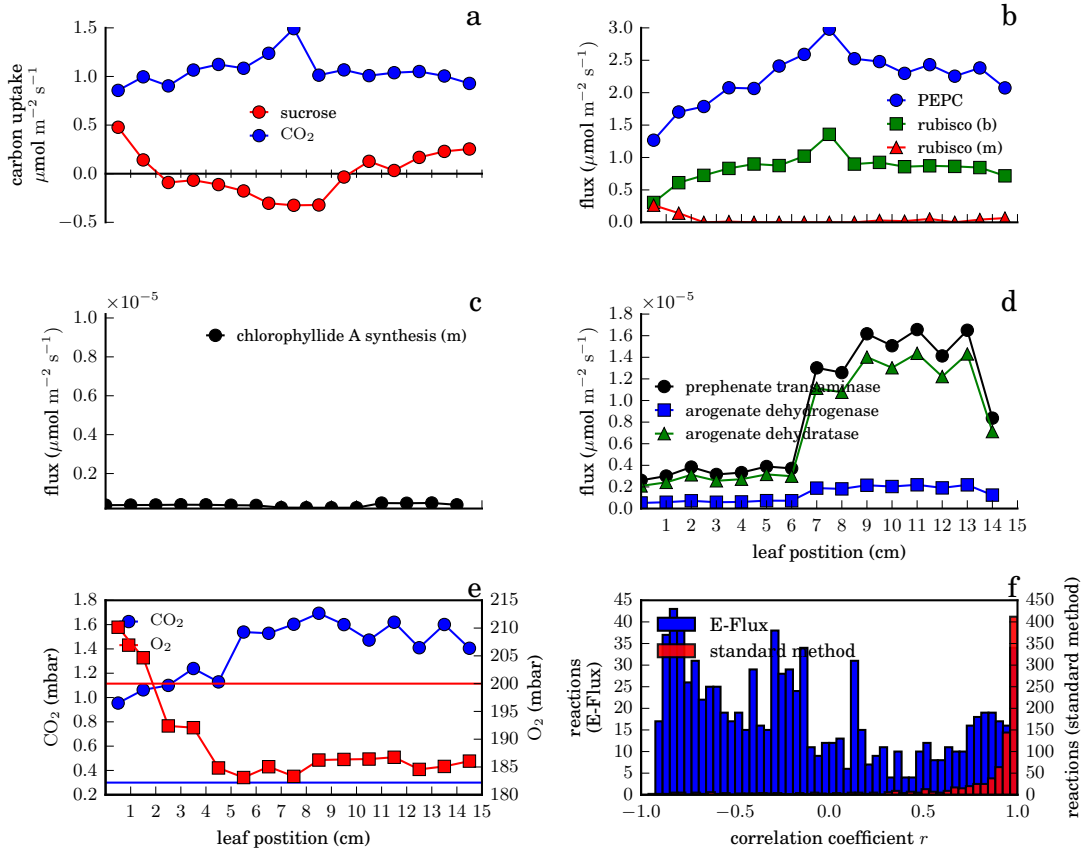

Figure 2.11: **Summary of predictions for the gradient model using the E-Flux method.** For explanation of each panel, see Figure 2.10.

2.3 Discussion

2.3.1 Reconstruction

Our model is the fourth published genome-scale metabolic reconstruction of the major crop plant *Zea mays*, and the first such reconstruction developed solely from maize data sources, rather than as a direct or indirect adaptation of the *Arabidopsis thaliana* model AraGEM [37].

Direct reaction-to-reaction comparison of iEB5204 with C4GEM [39],

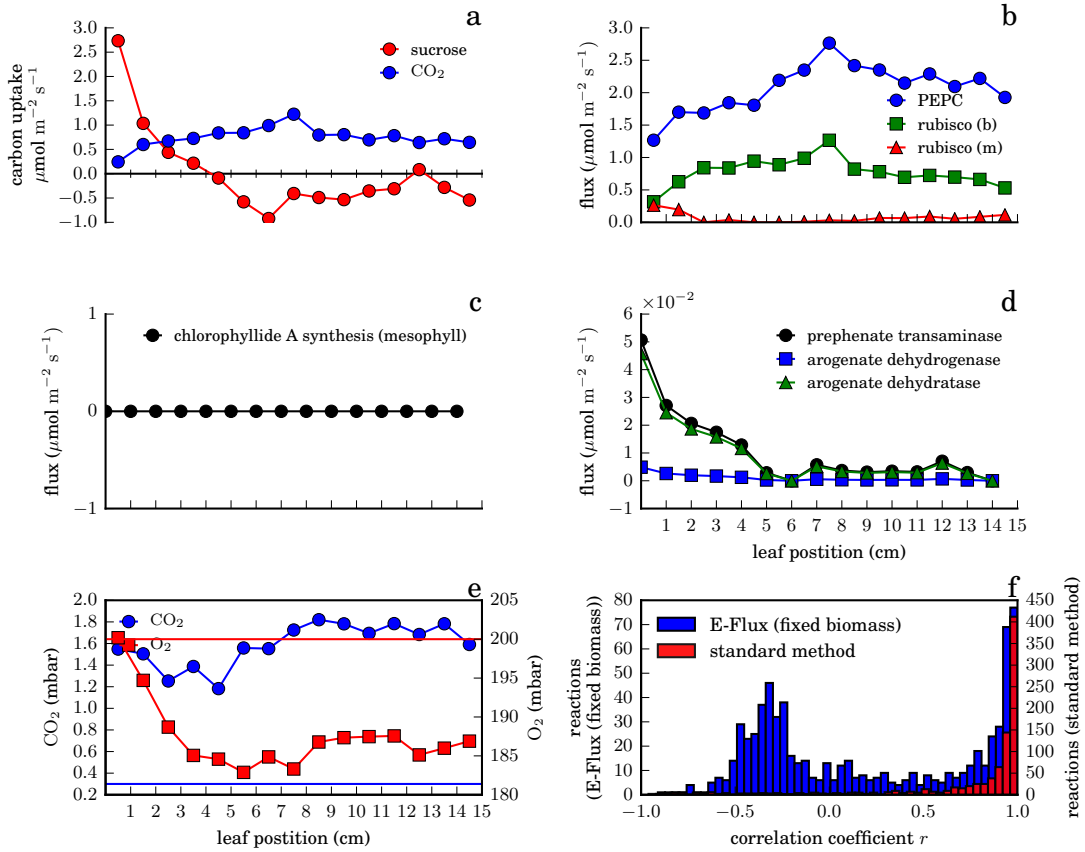

Figure 2.12: **Summary of predictions for the gradient model using the E-Flux method with fixed biomass composition.** The biomass composition is fixed to that used by iRS1563, as adapted (see Appendix A). For explanation of each panel, see Figure 2.10. Note that the chlorophyllide A synthesis pathway is blocked when the fixed biomass composition is used.

iRS1563 [41], and its successor model [48] is difficult because those models use a naming scheme for compounds and reactions ultimately based on KEGG [73,74] while this model, like its parent database, uses the nomenclature of MetaCyc and the BioCyc database collection. The models are broadly similar in size and biological scope. As published, C4GEM included 1588 reactions associated with 11623 maize genes; iRS1563, 1985 reactions associated with 1563 genes; the model of Simons et al. [48], 3892 unique reactions and 5824 genes; and iEB5204, 2720 reactions with 5204 genes. All models can simulate the production of sim-

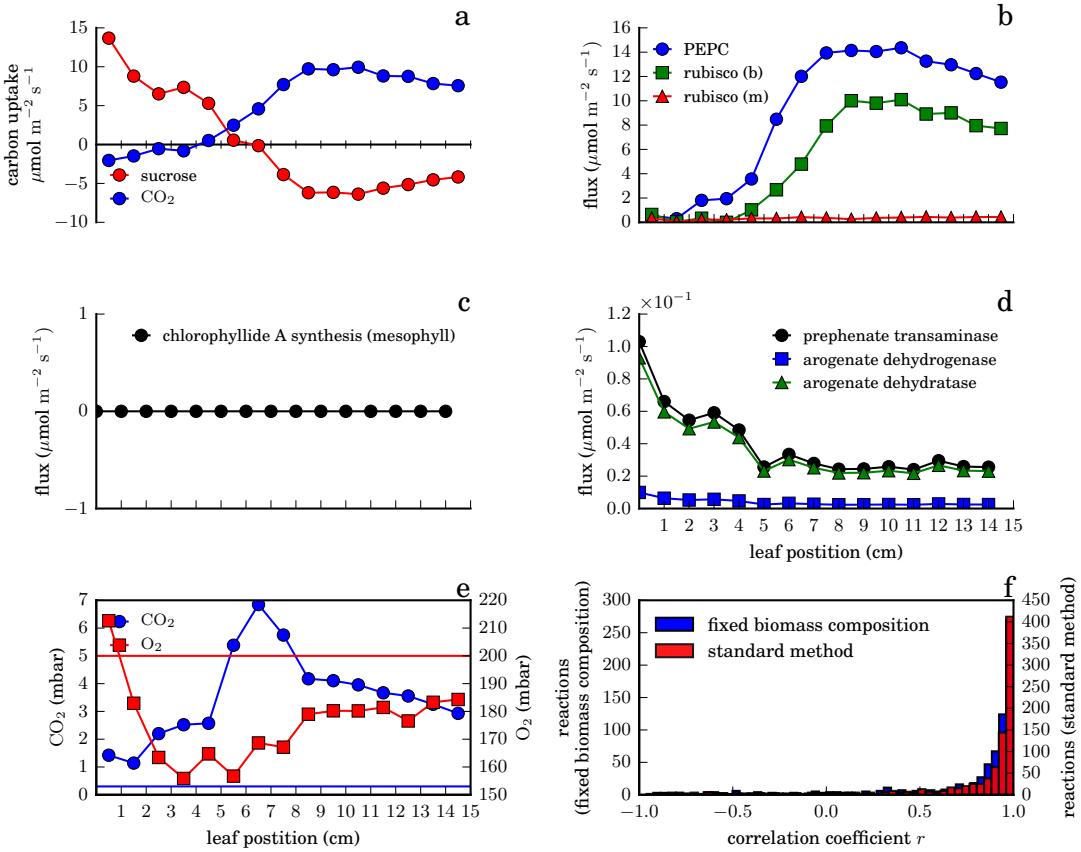

Figure 2.13: **Summary of predictions for the gradient model with fixed biomass composition.** For explanation of each panel, see Figure 2.10. Note that the chlorophyllide A synthesis pathway is blocked when the fixed biomass composition is used.

ilar sets of basic biomass constituents (including amino acids, carbohydrates, nucleic acids, lipids and fatty acids, and cell wall components) under photosynthetic and non-photosynthetic conditions and include key reactions of the C4 cycle. The model of Simons et al. [48] also offers extensive coverage of secondary metabolism.

However, the present model has several advantages which make it particularly suitable for integration with transcriptomics data:

Gene associations The gene associations included in iEB5204 are those pre-

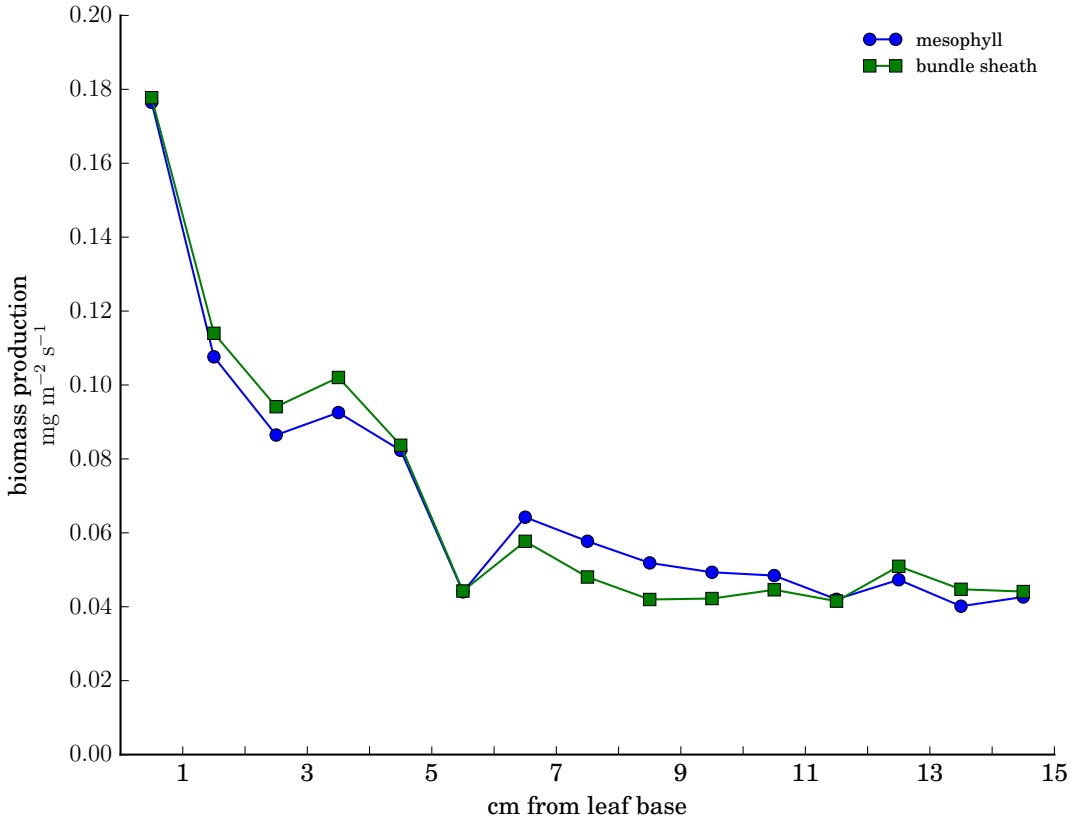

Figure 2.14: **Predicted biomass production rates in mesophyll and bundle sheath cells with fixed biomass composition.**

sented in CornCyc [57], which are generated by the PMN Ensemble Enzyme Prediction Pipeline (E2P2) [75], a homology-based protein sequence annotation algorithm trained on a reference dataset of experimentally validated enzyme sequences. The E2P2 approach is more comprehensive and scalable than the development procedures of the previous maize reconstructions (which involve, for example, obtaining gene associations by transferring annotations from Arabidopsis genes to their best maize BLAST hits and manually selecting annotations for remaining maize genes from among BLAST hits in other species). The entire set of gene associations in the FBA model may be readily updated based on improvements

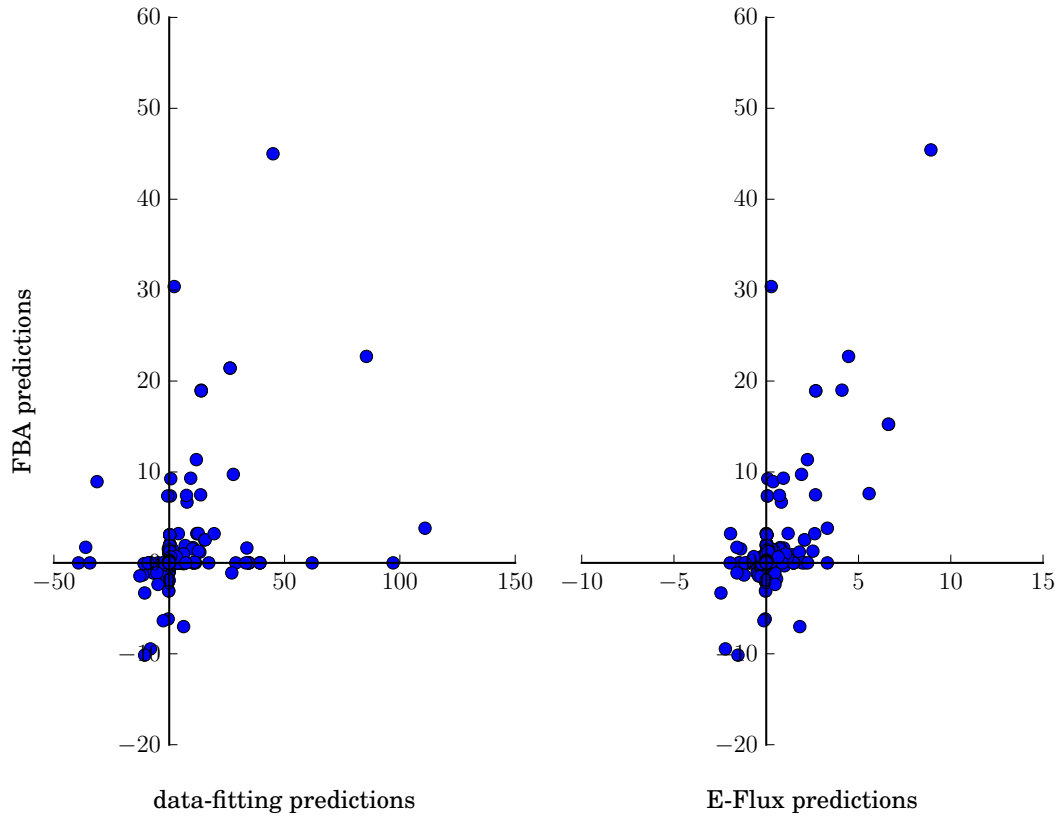

Figure 2.15: **Predicted variable values in an FBA calculation that does not incorporate expression data, compared to the best-fit and E-Flux methods.** The FBA calculation minimizes total flux while achieving the same total rate of CO_2 assimilation as predicted at the tip of the leaf in the fitting results. Left panel, FBA reaction rates vs. reaction rates predicted at the tip of the leaf in the best-fitting solution; right panel, FBA reaction rates vs. reaction rates predicted at the tip of the leaf by the E-Flux method. Axis limits exclude a small number of reactions of particularly large flux. Fluxes in $\mu mol\, m^{-2}\, s^{-1}$.

in the E2P2 prediction algorithm.

High-confidence submodel In developing the fitting algorithm we found that, to obtain plausible metabolic state predictions, a conservative reconstruction was preferable to a comprehensive one. For example, early tests with the comprehensive version of the model suggested that the fitting algorithm often found low-cost solutions involving high fluxes through reactions which, on investigation, we determined were unlikely to be active

in maize. Because of the model's connection to the CornCyc database, it was straightforward to create a reduced, high-confidence version of the model by preferentially excluding reactions not included in any manually curated plant metabolic pathway, even if candidate associated genes had been identified computationally, leading to more realistic results.

Reproducibility In an effort to improve the reusability of the model and encourage its application to other data sets, we have published the full source code for all calculations presented here (as ancillary files of [60] and online at `http://github.com/ebogart/fluxtools`) as has been recommended (see, e.g., [76]).

Previous reconstructions do offer two features absent from this model: gene associations for intracellular transport reactions, and gene associations which take into account the structure of protein complexes. Both should be considered in future work.

In agreement with [77], we found that building the model starting from a metabolic pathway database was considerably more straightforward than the standard process of *de novo* reconstruction [78]. Reasonable effort was still required to bring the model to a functional state by identifying reactions or pathways present in the CornCyc database which could not be handled automatically by the Pathway Tools export facility (for example, because they involved polymerization, or could not be checked automatically for conservation violations) and determining how to represent them appropriately in the FBA model.

The model construction process here could readily be adapted to generate metabolic models describing any of the more than 30 crop and model plant species for which Pathway Tools-based metabolic pathway databases [79] have

been developed by the Plant Metabolic Network [80], Sol Genomics Network [81], Gramene [82], and others (e.g., [83–85]) allowing the present data-fitting method to be applied to RNA-seq data from those organisms. The level of model development effort required and quality of fit results will vary depending on the extent of curation of the pathway database and quality of the gene function annotations.

2.3.2 Nonlinear optimization

In contrast to the linear and convex optimization methods employed in nearly all prior constraint-based modeling work, general constrained nonlinear optimization algorithms typically require more effort from the user (who might be required to supply functions which evaluate the first and second derivatives of all constraints with respect to all variables in the problem). They are slower, are more sensitive to choices of starting point and problem formulation, are not guaranteed to converge to an optimal point even if one exists, and, when they do converge to an optimum, cannot guarantee that it is globally optimal.

The software package we present allows the rapid and effective development of metabolic models with nonlinear constraints despite these complications. All necessary derivatives of constraint functions are taken analytically, and Python code to evaluate them is automatically generated. A model in SBML format may be imported, nonlinear constraints added and removed, and the problem repeatedly solved to test various design choices, solver options, and initial points, all within an interactive session, with a minimum of initial investment of effort in programming.

In the present case, agreement between nonlinear FBA calculations maximizing growth and the predictions of classical physiological models confirmed that the true, globally optimal CO_2 assimilation rate was found successfully. For the data-fitting calculations, where the true optimal cost is not known, we cannot exclude the possibility that there exist other optimal solutions, qualitatively distinct from the flux distributions and quasi-optimal regions presented above, with equivalent or lower costs. In practice, we encountered occasional cases in which reaction or pathway fluxes were initially predicted to be zero even when associated with nonzero data, despite the existence of a superior alternative solution with nonzero predicted fluxes. A step to detect and correct these situations was incorporated into the fitting algorithm.

Many future applications for the software are possible. Our approach to Rubisco kinetics may easily be extended to other models of C4 metabolism or, more generally, to any FBA calculation in a photosynthetic organism where the CO_2 level at the Rubisco active site, and thus the Rubisco oxygenation/carboxylation ratio, is not known *a priori*. A recent genome-scale metabolic reconstruction of the model alga *Chlamydomonas reinhardtii*, for example, was identified by the authors as being deficient in describing algal metabolism under low CO_2 conditions due to the fact that the Rubisco carboxylase and oxygenase fluxes were treated as independent and not competitive, as we have done here [86].

Ensuring that rates of Rubisco oxygenation, Rubisco carboxylation, and PEPC carboxylation are consistent with our knowledge of their kinetics is a special case of the more general problem of integrating kinetic and constraint-based modeling, to which diverse approaches have been proposed (e.g., [87–92]).

To our knowledge, no prior work has simply imposed kinetic laws as ad-

ditional, nonlinear constraints in the ordinary FBA optimization problem. Our results demonstrate the potential of this approach in systems where the kinetics of a few well-understood reactions are crucial. It remains to be seen how many kinetic laws may be incorporated in this way at once, and to what extent their introduction usefully constrains the space of possible steady-state flux distributions even when relevant kinetic parameters are not known (but instead are treated as optimizable variables, an approach with connections to ensemble kinetic modeling [93]).

Nonlinear constraints may also be of use in enforcing thermodynamic realizability of flux distributions, and relaxing requirements of linearity or convexity may stimulate the development of novel objective functions – either for data integration purposes, as here, or as alternatives to growth-rate maximization.

2.3.3 Data fitting

The expression of a gene encoding a metabolic enzyme need not correlate with the rate of the reaction that enzyme catalyzes. The relationship between transcription and degradation of mRNA and control of flux is indirect, mediated by protein translation, folding, and degradation, complex formation, posttranslational modification, allosteric regulation, and substrate availability. Indeed, as reviewed by [94], experimentally observed correlations among RNA-seq or microarray data (each itself an imperfect proxy for mRNA abundance or transcription rate), protein abundance, enzyme activity, and fluxes are variable and often weak.

For example, RNA-seq and quantitative proteomic data obtained from maize

leaves at the same developmental stage studied here, harvested simultaneously from plants grown together, showed Pearson correlation approximately 0.6 across the entire dataset, but some significantly lower values were found when correlations were restricted to genes of particular functional classes, and measured mRNA/protein ratios for individual genes varied up to 10-fold along the gradient [71]. A subset of this data is shown in Fig. 2.8d.

The most comprehensive study of the issue in plants so far [95] found so little agreement between RNA-seq and 13C-MFA data from embryos of two *Brassica napus* accessions that the authors concluded the inference of central metabolic fluxes from transcriptomics is, in general, impossible.

In this light, it is not surprising that methods for integrating transcriptomic data with metabolic models to predict reaction rates have met with limited success. Machado and Herrgård [96] reviewed 18 such methods and assessed the performance of seven of them on three test datasets from *E. coli* and *Saccharomyces cerevisiae* where experimentally measured intracellular and extracellular fluxes were available for comparison. None of the methods consistently outperformed parsimonious FBA simulations which completely ignored transcriptomic data.

In contrast, in the present work the use of transcriptomic data (and a limited number of enzyme activity measurements) allowed the correct prediction of a metabolic transition from the base of the leaf to the tip, which could not have been expected based on FBA calculations alone: without such data, all points along the gradient would be identical, and the biomass-production-maximizing solution would be the same at each. The predicted position of the source-sink transition is not perfectly accurate, and the overall performance of the model

cannot be evaluated until the predicted reaction rates are compared to detailed experimental flux measurements. Nonetheless, the results are encouraging. We offer two explanations for this apparent success.

First, the metabolic transition between the heterotrophic sink region at the base and the photoautotrophic source region at the tip is particularly dramatic, involving a large number of reactions which are effectively absent in one region but carry high fluxes in the other [55]; so long as even a slight correlation between transcript levels and fluxes exists, such a reconfiguration should be apparent from expression data.

Second, although the developing maize leaf is biologically more complex than microbial growth experiments, the relationship between expression levels and fluxes may be actually be closer in the leaf. Leaf development is a stereotyped, frequently repeated, relatively slow, one-way process, in which the precise sequence of events is subject to evolutionary optimization. Coordination of transcription with required fluxes will lead to efficient use of resources. In contrast, the test cases of [96] involve microbial responses to varying environmental conditions and under- and over-expression mutations. Environmental responses must be rapid, flexible and reversible – criteria a complex, scripted transcriptional response may not satisfy – while transcriptional responses to novel mutations, by definition, cannot have been evolutionarily optimized. This hypothesis could be tested by evaluating performance of the present method on RNA-seq data from mutant maize plants, or plants subject to environmental challenges.

We note also that methods that did not constrain or optimize the growth rate predicted zero growth rates in almost all the test cases studied by Machado

and Herrgård [96]. The present method also does not constrain or optimize the growth rate but consistently does predict nonzero growth as reflected in nonzero biomass production (whether with a flexible biomass composition was used, as above, or a fixed biomass composition, as in Figure 2.13 and Figure 2.14).

2.3.4 The whole-leaf model

Large-scale metabolic models of interacting cells of multiple types first appeared in 2010, with C4GEM [39] and a model of human neurons interacting with their surrounding astrocytes [97]. Many more complex multicellular FBA models have since appeared, including studies of the metabolism of interacting communities of microbial species in diverse natural environments or artificial co-cultures [98–104] (also [105] at a smaller scale) and of the metabolic capacities of host animals and their symbionts [106] or parasites [107]. In plants, diurnal variation in C3 and CAM plant metabolism has been simulated with a model which represents different phases of the diurnal cycle with different abstract compartments, with transport reactions representing accumulation of metabolites over time [46].

In the most direct antecedent of the present work, Grafahrend-Belau and coauthors developed a multiscale model of barley metabolism [49] which represented leaf, stem, and seed organs as subcompartments of a whole-plant FBA model, with nutrients exchanged through the phloem. Combining the FBA model with a high-level dynamic model of plant metabolism allowed them to predict changes in metabolism over time, including the transition between

a biomass-producing sink state and a fructan-remobilizing source state in the stem late in the plant's life cycle.

The whole-leaf model presented here occupies an intermediate position between prior C4 models, with single mesophyll and bundle sheath cells, and multi-organ whole-plant models such as [49]. It represents the first attempt to model spatial variations in metabolic state within a single organ, allowing the study of developmental transitions in leaf metabolism by incorporating data from more and less differentiated cells at a single point in time, rather than modeling development dynamically.

Other interacting cell models incorporate *a priori* qualitative differences in the metabolic capabilities of their components (e.g., leaf, stem, and seed, or neurons and astrocytes). In contrast in the work presented here, in order to allow the metabolic differences between any two adjacent points to be purely quantitative, the same metabolic network must be used for all points. This simplifies the process of model creation but implies that meaningful predictions of spatial variation depend entirely on the integration of (spatially resolved) experimental data. The ability of the model to capture the experimentally observed shift from sink to source tissue along the developmental gradient based on RNA-seq and enzyme activity measurements shows that this may be done successfully with high-resolution -omics data and careful model construction.

2.4 Methods

2.4.1 Reconstruction process

A local copy of CornCyc 4.0 [57] was obtained from the Plant Metabolic Network and a draft metabolic model was created using the MetaFlux module of Pathway Tools 17.0 [77]. The resulting model, including reaction reversibility information, was converted to SBML format and iteratively revised, as described in detail in Development of a flux balance analysis model for maize, until all desired biomass components could be produced under both heterotrophic and photosynthetic conditions and realistic mitochondrial respiration and photorespiration could operate.

An overall biomass reaction was adapted from iRS1563 [41] with minor modifications to components and stoichiometry, as detailed in Development of a flux balance analysis model for maize. To allow calculations with flexible biomass composition, individual sink reactions were added for most species participating in the biomass reaction, as well as several relevant species (including chlorophyll) not originally included in the iRS1563 biomass equation.

Core metabolic pathways were assigned appropriately to subcellular compartments (e.g., the TCA cycle and mitochondrial electron transport chain to the mitochondrion; the light reactions of photosynthesis, the Calvin cycle, and some reactions of the C4 cycle to the chloroplast; and some reactions of the photorespiratory pathway to the peroxisome) and the intracellular transport reactions necessary for their operation were added.

The model was thoroughly tested for consistency and conservation viola-

tions, confirming that no species could be created without net mass input or destroyed without net mass output (except species representing light, which can be consumed to drive futile cycles).

In the SBML files, gene association rules for reactions with associated genes in CornCyc are provided following COBRA conventions [108]. Additional annotations give the record in the CornCyc database associated with each reaction and species, where applicable.

To produce the higher-confidence version of the reconstruction, iEB2140, reactions in the base model which were not associated with any identified metabolic pathway in CornCyc, and those for which no genes for a catalyzing enzyme had been identified by computational function prediction, were removed from the model if their removal did not prevent photosynthesis, photorespiration, or the production of any biomass component. Then, all reactions which could not achieve nonzero steady-state rates were removed.

2.4.2 Mesophyll-bundle sheath model

A model for leaf tissue was created by taking two copies of the high-confidence model, representing mesophyll and bundle sheath cells, and adding reactions representing transport through the plasmodesmata which connect the cytoplasmic spaces of adjacent cells. Though in principle most small molecules can cross the plasmodesmata by diffusion [109], unrealistic concentration gradients may be required to drive high diffusive fluxes, and processes other than simple diffusion may play a role in the rapid exchanges which do occur [110]. Given this uncertainty we conservatively restricted such transport to species known or ex-

pected to be exchanged between cell types (under at least some circumstances); a complete list is given in Development of a flux balance analysis model for maize.

Net import or export of metabolites from the system was limited to the mesophyll, for gases exchanged with the intercellular airspace, or the bundle sheath, for soluble metabolites exchanged with the leaf's vascular system. Reactions were not otherwise restricted *a priori* to a particular cell type. To facilitate integration with cell-type-specific RNA data, gene associations in this model are tagged with the relevant cell type, e.g. 'bs_GRMZM2G039273' vs 'ms_GRMZM2G039273'.

2.4.3 Leaf gradient model

The choice of phloem transport metabolites (other than sucrose) is a compromise. Glycine is the most abundant amino acid in maize phloem [111], and glutathione is a putative phloem sulfur transport compound [112], but many other amino acids are present in the phloem sap, and other compounds (e.g., S-methyl-methionine [112]) may play roles in phloem sulfur transport. However, we found that the available data did not adequately constrain rates of phloem transport if multiple transport species of each type were allowed, resulting in high rates of transport from the base towards the tip, against the direction of bulk flow in the phloem.

For simplicity, export of metabolites from the leaf to the rest of the plant through the phloem was neglected and net import of sucrose was not allowed. Each segment was taken to have the same total area, so that a 1 $\mu mol\, m^{-2}\, s^{-1}$

rate of sucrose loading in one segment exactly balanced a 1 µmol m^{-2} s^{-1} rate of sucrose unloading in another segment.

Note that the whole-leaf model is constructed dynamically within the data-fitting code, rather than being loaded from an SBML file.

Physiological constraints

Rubisco carboxylase and oxygenase rates v_c and v_o in mesophyll and bundle sheath chloroplasts were constrained to obey Michaelis-Menten kinetic laws with competitive inhibition,

$$\begin{aligned} v_c &= \frac{v_{c,\text{max}}\,[\text{CO}_2]}{[\text{CO}_2] + k_c\left(1 + \frac{[\text{O}_2]}{k_o}\right)} \\ v_o &= \frac{v_{o,\text{max}}\,[\text{O}_2]}{[\text{O}_2] + k_o\left(1 + \frac{[\text{CO}_2]}{k_c}\right)}, \end{aligned} \tag{2.3}$$

and the relationship $v_{o,\text{max}}/v_{c,\text{max}} = k_C/(k_O \cdot S_R)$ was imposed, from which eq. (1.2) follows [25]. The Michaelis-Menten constants for oxygen and carbon dioxide k_C and k_O and the Rubisco specificity S_R were set to values typical of C4 species: k_C, 650 µmol mol^{-1}; k_O, 450 mmol mol^{-1}; S_R, 2590 [25].

The rate of PEP carboxylation in the mesophyll was bounded above by an appropriate kinetic law,

$$v_p = \frac{v_{p,\text{max}}\,[\text{CO}_2]}{k_{C,p} + [\text{CO}_2]} \tag{2.4}$$

with $0 \leq v_{p,\text{active}} \leq v_{p,\text{max}}$ and an appropriate $k_{C,p}$ (80mmol mol^{-1}, [25]).

The parameters $v_{p\text{max}}$ and $v_{c,\text{max}}$ representing the total amount of Rubisco and PEPC available may be fixed to permit comparison to models parameterized in those terms or allowed to vary.

Rates of oxygen and carbon dioxide diffusion from the bundle sheath to the mesophyll, L and L_O, were constrained to obey the relationship

$$\begin{aligned} L &= g_{BS}\left(\mathrm{CO}_{2,BS} - \mathrm{CO}_{2,ME}\right) \\ L_O &= g_{BS,O}\left(\mathrm{O}_{2,BS} - \mathrm{O}_{2,ME}\right) \end{aligned} \tag{2.5}$$

with $g_{BS,O} = 0.047 g_{BS}$ [25]. All simulations used the bundle sheath CO_2 conductivity measured by [67] for maize plants grown under high light, 1.03 ± 0.18 µmol m^{-2} s^{-1}. While g_{BS} undoubtedly varies along the developmental gradient, its deviation from this value (measured in fully-expanded leaves, 3-4 weeks after planting) is likely greatest below the region of high suberin synthesis identified 4 cm from the leaf base [62]; as the C4 cycle was not predicted to operate at high rates in this region, the impact of this discrepancy should be limited.

Resistance to CO_2 diffusion from the intercellular airspace to the mesophyll cells was neglected; ref. [113] reported $g_m \approx 1$ mmol m^{-2} s^{-1} in maize under a variety of conditions, suggesting the mesophyll and intercellular CO_2 levels would differ only slightly at the rates of CO_2 assimilation and release dealt with here. Similarly, all intracellular compartments were taken to have equal CO_2 concentrations.

2.4.4 Optimization calculations

The nonlinear modeling package uses the libsbml python bindings to read SBML files [114] and an internal representation of SBML models derived from the SloppyCell package [115, 116]. IPOPT calculations used version 3.11.8 with the linear solver ma97 from the HSL Mathematical Software Library [117]. Where not specified, convergence tolerance was 10^{-5}, or 10^{-4} in FVA calcula-

tions. To solve purely linear problems (e.g., to test the production of biomass species during the reconstruction process, where nonlinear constraints were not used) the GNU Linear Programming Kit, version 4.47 [118], was called through a Python interface [119].

2.4.5 Integrating biochemical and RNA-seq data

RNA-seq datasets

To obtain mesophyll- and bundle-sheath-specific expression levels at 15 points, we combined the non-tissue-type-specific data of Wang et al. [62], measured at 1-cm spatial resolution, with the tissue-specific data of Tausta et al. [63] obtained by using laser capture microdissection (LCM) – measured 4 cm, 8 cm and 13 cm from the leaf base (the upper three highlighted positions in Fig. 2.3b). This integration was achieved by determining for each gene at each of those points with LCM data the ratio of the average RPKM in the mesophyll (M) to the sum of the average RPKM values for mesophyll and bundle sheath ($M+B$); furthermore, we assumed that the $M/(M+B)$ ratio at the leaf base was 0.5 (based on the proteomic experiments of Majeran et al. [69], which showed only limited mesophyll-bundle sheath specificity there), and linearly interpolating to estimate $M/(M+B)$ ratios at all 15 points. For very weakly expressed genes, we did not impose cell-type specificity: where the sum of mesophyll and bundle sheath RPKM in the LCM data was less than 0.1, we assumed $M/(M+B) = 0.5$. We then divided the mean whole-leaf FPKM measurement at each point into mesophyll and bundle sheath portions according to these ratios.

To associate expression data with a reaction, data for its associated genes

were summed, dividing the data for a gene associated with multiple reactions in the model equally among them. The uncertainties δ_{ij} in the objective function (eq. (2.1)) were estimated in an ad hoc way by splitting the standard deviations of the FPKM values over multiple experimental replicates according to the $M/(M+B)$ ratios and then summing the uncertainties for all genes associated with a particular reaction, imposing a minimum relative error of 0.05 and a minimum absolute uncertainty corresponding to 7.5 FPKM.

To globally rescale the expression data to be comparable to expected flux values, data for PEPC and Rubisco were compared to the enzyme activity measurements discussed below and a simple linear regression performed, yielding a conversion factor of 204 FPKM $\approx$ 1 µmol m^{-2} s^{-1} for these enzymes. All expression data were divided by this factor before solving the optimization problem.

Enzyme activity measurements

Enzyme activities constrained by measurements in [62] were alanine aminotransferse, aspartate aminotransferase, fructose bisphosphate aldolase, glyceraldehyde 3-phosphate dehydrogenase (NADPH), glyceraldehyde 3-phosphate dehydrogenase (NADH), glutamate dehydrogenase (NADH), malate dehydrogenase (NADH), malate dehydrogenase (NADPH), PEPC, phosphofructokinase, phosphoglucomutase, phosphoglucose isomerase, phosphoglycerokinase, Rubisco, transketolase, triose phosphate isomerase, and UDP-glucose pyrophosphorylase.

For Rubisco and PEPC, enzyme data constrained the sum of the variable kinetic parameters $v_{c,\max}$ and $v_{p,\max}$ in mesophyll and bundle sheath compart-

ments, rather than the sum of the associated fluxes. Enzyme data in nanomole per minute per gram fresh weight was converted to micromole per second per square meter of leaf surface area assuming a fresh weight of 150 g m^{-2}.

Handling reversible reactions

The objective function (eq. (2.1)) optimizes the agreement between the absolute value of the flux through each reaction with its data, but IPOPT requires a twice continuously differentiable objective function. We use a reformulation F' representing each absolute value $|v_{ij}|$ as the product of the flux and a parameter σ_{ij} representing its sign:

$$F'(v) = \sum_{i=0}^{N_r} \sum_{j=1}^{15} \frac{(e^{s_i}\sigma_{ij}v_{ij} - d_{ij})^2}{\delta_{ij}^2} + \alpha \sum_{i=0}^{N_r} s_i^2 \tag{2.6}$$

Similarly, the enzyme activity data constraint, eq. (2.2), was rewritten to replace absolute values in this way. Reaction rates with positive (negative) sign parameter were required to take values greater than a small negative (less than a small positive) tolerance, typically 1.0.

Choosing the σ_{ij} to optimize F' is a very large scale mixed-integer nonlinear programming problem. We arrive at an approximate solution using a heuristic method similar in spirit to that of [64], with three steps.

1. The subproblems representing each segment of the leaf are solved separately, with all scales s_i set to zero and modest upper and lower bounds on the reactions representing nutrient exchange with the phloem. Within each segment, a sign for the reversible reaction r_1 with the highest associated expression data is chosen by first setting its sign σ_1 to $+1$, finding

the minimum-flux best-fitting flux distribution $\mathbf{v}^+$ ignoring the costs associated with all other reversible reactions (but including costs associated with all irreversible reactions), then finding the cost c^+ of the best-fitting flux distribution $\mathbf{v'}^+$ considering the costs of the reversible reactions with nonzero fluxes in $\mathbf{v}^+$ (temporarily setting their signs according to their values in that case). A cost c^- is determined analogously after setting the sign σ_1 to -1 , and if $c^- < c^+$, $\sigma_1 = -1$ is chosen; otherwise, $\sigma_1 = +1$. Then the reversible reaction with the second-highest expression data r_2 is treated in the same way, considering r_1 to be irreversible.

2. When signs for all reversible reactions have been chosen at a segment, a final best-fitting flux distribution given those signs is determined. Then the full optimization problem, combining all fifteen segments, is solved with the chosen sign parameters fixed, using those flux distributions to provide a nearly-feasible initial guess.

3. The sign-choice process in each subproblem is then solved again, fixing the scale factors s_i and rates of metabolite exchange with the phloem to those determined in the full problem. If no signs change, or if the new signs do not decrease the objective function value, fitting stops; otherwise, step 2 is repeated.

4. Finally, for each reaction j with nonzero data and maximum absolute flux less than 0.0001 at any point in the leaf model, a lower bound of $-0.99d_i$ is imposed on the term $(e^{s_i}\sigma_{ij}v_{ij} - d_{ij})$ in the objective function, for $i = 1, \ldots, 15$, and the full fifteen-segment optimization problem is solved again.

The final step addresses the observation that the optimization process occasion-

ally converged to a solution in which a few reactions with associated data were predicted to have zero flux when a better solution with nonzero flux existed. In some cases (e.g. the $s_i = 0$ case shown in Figure 2.10) this step did not lead to an overall reduction in the objective function and was omitted.

Steps 1 and 3 take between one and eight hours per segment using an AMD Opteron 6272 and may be easily parallelized across up to 15 processors. Step 2 may take up to 2 hours in the first iteration but is often faster in later iterations, when the initial guess is closer to the optimum. Typically the procedure stops after 4-5 iterations, requiring about 24 total hours of wall time using 15 processors.

Special cases

The Rubisco oxygenase, Rubisco carboxylase, and mesophyll PEPC fluxes are excluded from the objective function. Instead, terms are added comparing the transcriptomic data for those enzymes to the variables which explicitly represent their activity level: for Rubisco, $v_{c,\mathrm{max}}$ in mesophyll and bundle sheath compartments, and for PEPC, $v_{p\mathrm{max}}$ in the mesophyll. Scale factors for the mesophyll and bundle sheath Rubisco activities are not constrained to be equal.

CHAPTER 3

GENOME-SCALE MODELING OF THE EVOLUTIONARY PATH TO C4 PHOTOSYNTHESIS

3.1 Introduction[1]

In section 1.2, several issues of uncertainty in the current account of the evolution of the C4 system were identified. Mathematical and computational models can assist in the study of these questions by making concrete predictions about how the evolution of the C4 system could play out under various sets of assumptions. Despite progress identifying regulatory mechanisms and amino acid sequence changes associated with the C4 phenotype [8, 120, 121], the complexity of the system and the extent to which its genetic control is still unknown, make it impractical to construct an detailed model which identifies many specific, relevant genetic loci and evaluates the fitness associated with each possible genotype to simulate selection directly. Instead, a more abstract model which describes the photosynthetic phenotype with a manageable number of parameters must be used.

Recently, Heckmann and coauthors [52] adapted a well-validated biochemical and physiological model of C3-C4 intermediate photosynthesis [25] to describe a six-dimensional fitness landscape between C3 and C4 states, parameterized by the fraction of total Rubisco activity in the mesophyll, the fraction of photorespiratory decarboxylation in the mesophyll, mesophyll PEPC level, PEPC Michaelis-Menten constant for bicarbonate, the conductance of the bun-

[1] This chapter is adapted from "Genome-scale modeling of the evolutionary path to C4 photosynthesis", Eli Bogart and Christopher R. Myers (manuscript in preparation).

dle sheath to CO_2 diffusion, and the Rubisco turnover rate (which was used to determine other Rubisco kinetic parameters, following an empirical power-law relationship [122]). Dividing the C3-C4 changes in each parameter into five discrete levels, using CO_2 assimilation rate as a proxy for overall fitness, they analyzed the properties of the landscape, estimated relative mutation probabilities for each parameter, and simulated 5000 realizations of the C3-C4 transition. They found the model described a "Mt. Fuji-like" landscape, in that the C4 state was a unique, global fitness maximum which could be reached from any other point in the parameter space along a path of continuous fitness increase, and showed that the random paths exhibited a modular structure and were distributed narrowly around a mean path that was consistent with biochemical data from C3-C4 intermediate species.

While this work represented a significant advance in quantitative simulation of the C3-C4 transition, it did not examine the influence of environmental factors [123] and its high level of biochemical abstraction precluded the study of differences between possible evolutionary histories of the different decarboxylation subtypes, or of changes in metabolism outside the core photosynthetic pathways in response to the development of the C4 system. It also remains to be seen whether the observed simple structure of the fitness landscape is maintained when the transition is described at the level of changes in expression of individual enzymes in the mesophyll and bundle sheath compartments.

To address these questions, we present here an approach to modeling the C3-C4 transition which is similar in spirit to [52] but replaces the six-dimensional, coarse-grained physiological model with a genome-scale metabolic model that maintains consistency with the nonlinear relationships captured in the physio-

logical model of [25]. The 793 parameters of the larger model include maximum rates for hundreds of reactions in primary metabolism in both mesophyll and bundle sheath compartments as well as Rubisco and PEPC kinetic parameters and the conductance of the bundle sheath compartment to CO_2 diffusion. We explore the resulting simulated transition paths and their responses to changes in atmospheric CO_2 levels and decarboxylation subtypes.

3.2 Methods

3.2.1 Modeling photosynthetic metabolism

Underlying metabolic network model

To obtain a flexible large-scale metabolic model of photosynthesizing leaf tissue, capable of describing C3, C4, or C3-C4 intermediate plants we adapted the model iEB2140x2 described above. Figure 1.1(b) illustrates the structure of the model, which includes mesophyll and bundle sheath compartments, connected through the plasmodesmata, with mitochondrial, peroxisomal and chloroplastidic subcompartments in each cell type. The model includes 1268 reactions and can describe the synthesis of a wide array of biomass components – including carbohydrates, amino acids, cellulose, hemicelluloses, lignins, nucleic acids, fatty acids and lipids, and chlorophyll – under photosynthetic and heterotrophic conditions.

In previous work, the model was applied to infer changes in the metabolic state of maize leaf tissue from RNA-seq data sampled along a developmental

gradient from the base to the tip of a growing leaf, and successfully captured the transition between the heterotrophic, biomass-producing sink region at the base and the more mature photosynthetic, sucrose-exporting upper region.

Though iEB2140x2 was developed specifically for leaves of the C4 grass *Zea mays*, it makes no *a priori* assumptions about localization of any function to mesophyll or bundle sheath; instead, all reactions (except exchanges with the leaf's vasculature or the intercellular air space) are present in a bundle sheath copy and a mesophyll copy. As the bundle sheath compartment and all the reactions of the C4 system are also present, playing different roles, in C3 plants [124,125], the model can also realistically simulate C3 or C3-C4 intermediate leaves, given an appropriate choice of parameters.

Basic nonlinear physiological constraints

The model also incorporates nonlinear relationships between reaction rates, CO_2 levels, and the ratio of Rubisco oxygenase and carboxylase reactions. As described above, we obtain predictions for the CO_2 assimilation rate A, the rates v_i of every reaction in the model, and the bundle sheath O_2 and CO_2 levels through a nonlinear flux-balance analysis approach, maximizing A subject to the usual FBA steady-state constraint [50],

$$S \cdot \mathbf{v} = \mathbf{0} \tag{3.1}$$

(where the stoichiometry matrix S encodes the structure of the metabolic reaction network,) the requirement that

$$v_j \geq 0$$

if reaction j is irreversible, and a collection of kinetic laws, as follows:

- For the Rubisco carboxylation and oxygenation rates v_c and v_o,

$$\begin{aligned} v_c &= \frac{v_{c,\text{active}}\,[\text{CO}_2]}{[\text{CO}_2] + k_c\left(1 + \frac{[\text{O}_2]}{k_o}\right)} \\ v_o &= \frac{v_{o,\text{active}}\,[\text{O}_2]}{[\text{O}_2] + k_o\left(1 + \frac{[\text{CO}_2]}{k_c}\right)}, \end{aligned} \tag{3.2}$$

  Note that these equations apply separately to the mesophyll and bundle sheath compartments, which have different O_2 and CO_2 levels, v_{active} values, etc. In each compartment we further require

$$v_{o,\text{active}}/v_{c,\text{active}} = k_C/(k_O \cdot S_R), \tag{3.3}$$

  where S_R is the specificity of Rubisco for CO_2 over O_2; the simpler relationship

$$\frac{v_o}{v_c} = \frac{1}{S_R}\frac{[\text{O}_2]}{[\text{CO}_2]} \tag{3.4}$$

  follows.

- For the mesophyll PEP carboxylase rate v_P,

$$v_p = \frac{v_{p,\text{active}}\,[\text{CO}_2]}{k_{C,p} + [\text{CO}_2]}. \tag{3.5}$$

- For the diffusive leakage of CO_2 and O_2 from bundle sheath to the mesophyll,

$$\begin{aligned} L &= g_{BS}\left([\text{CO}_{2,BS}] - [\text{CO}_{2,ME}]\right) \\ L_O &= g_{BS,O}\left([\text{O}_{2,BS}] - [\text{O}_{2,ME}]\right) \end{aligned} \tag{3.6}$$

Values of the parameters k_c, k_o, and S_R are determined as discussed below; allowed ranges of values for $k_{C,p}$ and g_{BS} are given in Table 3.1; and $g_{BS,O} = 0.047 g_{BS}$. (While we have written these equations in terms of concentrations, internally all CO_2 and O_2 values are represented as equivalent partial pressures.)

Symbol	Definition	Value	units
$E_{R,\text{relative}}$	relative Rubisco efficiency	1.0-2.588	none
K_p or $k_{C,p}$	PEPC Michaelis-Menten constant for CO_2	80.0-200.0	µbar
g_s	Bundle sheath conductivity to CO_2 diffusion	1.0-15.0	µmol m^{-2} s^{-1} $mbar^{-1}$
A	rate of CO_2 assimilation	variable	µmol m^{-2} s^{-1}
β	fraction of Rubisco in mesophyll	0.0-0.95	none
ξ	fraction of glycine decarboxylase in bundle sheath	0.0-1.0	none

Table 3.1: **Glossary of symbols and values of parameters used in the nonlinear FBA calculations**. Kinetic values are taken from [52].

The resulting nonlinear optimization problem is solved by IPOPT [53] (version 3.11.8, with the ma97 linear solver from the HSL Mathematical Software Library [117]) through the fluxtools Python package (`http://github.com/ebogart/fluxtools`). Effectively, these nonlinear constraints ensure that the flux distributions predicted by the model are also consistent with well-established physiological models of photosynthesis such as [25].

Connections to the model of Heckmann et al.

To facilitate comparison with the results of [52] we adopted the kinetic parameter values (or ranges of values) used there (Table 3.1) and imposed additional nonlinear constraints to reflect the tradeoff between Rubisco efficiency and specificity for CO_2 over O_2 observed by [122]. In the model of [52], this is encapsulated in the power-law relationship:

$$\begin{aligned} k_c &= 16.07 k_{ccat}^{2.36} \\ \frac{k_c}{k_o} &= 3.7 \cdot 10^{-4} k_{ccat}^{1.16} \\ S_R &= 5009.75 k_{ccat}^{-0.6} \end{aligned} \tag{3.7}$$

where k_{ccat} is the Rubisco carboxylase turnover number. Using eq. 3.3, we reformulate this as

$$\begin{aligned} k_C &= 288.6\,\mu\text{bar} \cdot E_{\text{R,relative}}^{2.36} \\ k_O &= 188.62\,\text{mbar} \cdot E_{\text{R,relative}}^{1.2} \\ \frac{v_{o,\max}}{v_{c,\max}} &= 0.272 \cdot E_{\text{R,relative}}^{-0.56} \end{aligned} \tag{3.8}$$

where $E_{\text{R,relative}}$ is the ratio of the efficiency of the Rubisco carboxylase reaction (that is, k_{ccat}) to its C3 value.

We note in passing that [52] discussed a limited approach to integrating non-

linear physiological constraints with an FBA model (to support their use of atmospheric CO_2 assimilation as a proxy for overall fitness). There, the physiological model was first solved to obtain predictions for A, Rubisco and PEPC rates, L, etc., and then an FBA problem was solved with the corresponding fixed to the values thus obtained. That approach would be inadequate here: to explore the integration of the C4 system and the broader metabolic network we wish to predict the optimal level of A that emerges from constraints on the maximum rates of each of the many reactions in the network, but the physiological model alone cannot take these into account. (Instead, in the approach of [52], information flows only the other way: reactions in the larger-scale FBA model are constrained by the small number of parameters supplied to the physiological model.)

Maximum rates for enzymatically catalyzed reactions

Finally, we impose consistency with a set of reaction-specific maximum rate parameters, in two steps. First, we require

$$-v_{\text{max},i} \leq v_i \leq v_{\text{max},i}, i = 0, 1, \ldots N_r,$$

except for the following special cases: mesophyll PEPC, where $0 \leq v_{p,\text{active}} \leq v_{p,\text{max}}$; mesophyll and bundle sheath Rubisco carboxylase reactions, where $0 \leq v_{c,\text{active}} \leq v_{c,\text{max}}$; mesophyll and bundle sheath Rubisco oxygenase reactions, whose maximum rates follow from the maximum carboxylase rates through eq. 3.3; and internal and external transport reactions (as the model does not describe in detail which are active and which passive). Note these constraints supplement, rather than replace, the constraints forcing irreversible reaction rates to be positive.

Second, we decompose each maximum rate

$$v_{\text{max},i} = P_i E_i$$

where P_i notionally represents the total weight of protein (per square meter of leaf surface area) invested in the enzyme that catalyzes reaction i, and E_i represents the efficiency of enzyme i (i.e., turnover number per unit weight of enzyme). Here again Rubisco carboxylase is a special case, with

$$v_{c,\text{max}} = P_R E_R^0 E_{R,\text{relative}}$$

As it is impractical to obtain reliable experimental estimates of *in vivo* turnover numbers for all of the several hundred reactions of the model, we have (except as noted below) taken $E_i = 1$ μmol $m^{-2}s^{-1}$ for all reactions except Rubisco, where $E_R^0 = 0.5$, representing (very conservatively) the fact that Rubisco is highly inefficient compared to many or most other enzymes [1]. (Note this definition leaves the units of the P_i values arbitrary).

Finally, to exclude the possibility that fitness will increase simply through an overall, uniform increase in enzyme levels, we add a maximum value for the total protein level,

$$P_1 + \ldots + P_N \leq P_{max}, \tag{3.9}$$

and ensure that the maximum value is reached at the start and end of the paths simulated below.

Boundary conditions

In the calculations below, we assumed light uptake, nitrate uptake, and sulfate uptake were not limiting. To ensure simulated flux distributions captured a

range of functions carried out in photosynthesizing leaf tissue, we required that biomass production and the export of sucrose, nitrogen (as glycine), and sulfur (as glutathione) to the phloem occur in a set ratio (adapted from the rates inferred from experimental data near the tip of the maize leaf, in [60]): for each mole of carbon assimilated, one gram of total biomass is produced, 0.02 moles of nitrate and 2 millimoles of glutathione are exported, with the balance of the assimilated carbon being exported as sucrose.

Corrections to iEB2104

Early tests of the procedure described below uncovered a small number of problematic behaviors of the model which were not observed when using the very different conditions and objective function of [60], mostly unrealistically high rates through pairs of reactions which could act in concert as a transhydrogenase (oxidizing NADPH to reduce NAD^+ or oxidizing NADH to reduce $NADP^+$). To suppress these, two non-essential reactions were inactivated: an NADPH-dependent glutamate dehydrogenase (EC 1.4.1.4, CornCyc `GLUTDEHYD-RXN`,) which (at least in bacteria and yeast) plays a role primarily in high-ammonia conditions [126], and a proline dehydrogenase (EC 1.5.99.8, CornCyc `RXN-821`) which is an oversimplified representation of a reaction which should donate electrons directly to the mitochondrial electron transport chain [127].

3.2.2 Finding optimal evolutionary paths

With the above constraints in place, the model can predict a CO_2 assimilation rate, given an external CO_2 level, values of g_s, k_p and $E_{R,\text{relative}}$, and enzyme levels for each of 790 reactions. Taking CO_2 assimilation (which is proportional to the plant's overall biomass synthesis rate) as a proxy for fitness, the model describes a 793-dimensional phenotypic fitness landscape, where the phenotype is the overall pattern of expression of metabolic enzymes in mesophyll and bundle sheath, combined with the bundle sheath resistance to CO_2 diffusion and the kinetic properties of Rubisco and PEPC.

Random (or greedy) walks in parameter space

To characterize paths from C3-like points in this landscape to C4-like ones, a natural approach would be to start at a fixed C3 state and perform a biased random walk, perturbing individual parameters serially, with a preference for perturbations which increase fitness. Alternatively, we could proceed as in [52] by specifying both C3 and C4 values for each parameter and choosing which parameter to advance towards its C4 value at each step.

It is reasonable to expect that, in a high-dimensional space, this simulation process will be inefficient, as many parameters will have a limited effect on fitness. The structure of the FBA model makes this inefficiency even more acute: marginal changes in $v_{\text{max},i}$ can have no effect on the solution unless $v_i = v_{\text{max},i}$, which will generally be true for only one reaction unless the v_{max} parameters have been carefully tuned.

Additional complications arise. If we optimize CO_2 assimilation subject

to eq. 3.9, allowing the protein levels P_i to vary, we will generally find that $\|v_i\| = v_{\text{max},i} = P_i E_i$ for all reactions subject to a maximum rate constraint (because otherwise, 'spare' enzymatic capacity could be redistributed to improve the assimilation rate).

But consider for example a linear pathway

$$A \xrightarrow{R_1} B \xrightarrow{R_2} C \xrightarrow{R_3}, \ldots$$

where (under the FBA steady-state constraint) $v_1 = v_2 = v_3$. If it is also true that $v_{\text{max},1} = v_{\text{max},2} = v_{\text{max},3}$, *no single parameter change* can increase the optimal flux through the pathway. Effectively, all the protein levels in the pathway have flux control coefficient zero (a failure of the model, as metabolic control analysis guarantees that the coefficients should sum to 1.0 [128, 129]).

In this situation, it is clear that very few random parameter changes will lead to increases in flux through the pathway, though decreases are readily accomplished. (Note that here any proposed parameter change must increase the levels of some proteins and decrease others, to ensure eq. 3.9 is obeyed).

Of course, we expect that very few random mutations to real C3 plants would lead in the direction of the C4 state, with most being detrimental or neutral, but these technical issues make progress towards the C4 state in a sense more difficult in the model than it was *in vivo* – where a single decrease in expression of mesophyll glycine decarboxylase, for example, would naturally increase the concentration of photorespiratory intermediates, which could diffuse to the bundle sheath, tending to increase the GDC flux there, and thus the CO_2 level. In the model, a large number of other mutations would need to occur coincidentally to achieve the same effect, a larger number still to activate the C4 cycle, and a larger number still to increase flux through Rubisco and the

Calvin cycle in the bundle sheath. Direct simulations of such a process proved prohibitive.

The elastic band method

Instead, we turn to so-called 'chain of states' methods applied in theoretical chemistry and molecular dynamics to find minimum-energy transition paths between different configurations of a simulated system (see [130] and references therein). In such methods, multiple replicas of a system of interest are simulated simultaneously, with one in the initial state, one in the final state, and the others constrained in some way to lie between the endpoints in configuration space. Perhaps the simplest such approach is the elastic band method [131], in which the replicas are connected by imaginary springs. As the replicas move towards lower-energy configurations, the springs prevent them from all falling into the lowest-energy state, so that the chain of replicas converges to a low-average-energy path from the initial state to the final state. Ideally, the result is a good approximation of the minimum-energy path, though paths which achieve only local minimization of the energy may be found if the energy landscape is rough.

More sophisticated alternatives exist, but the elastic band method is simple to describe and implement and has proved adequate for our purposes (but see section 3.4 below). To find an N-replica path between $\mathbf{x}'$ and $\mathbf{x}''$ for a system described each described by an m-dimensional configuration vector $\mathbf{x}$ and an energy function $f(\mathbf{x})$, constrained to obey $\mathbf{g}(\mathbf{x}) = \mathbf{0}$, $\mathbf{h}(\mathbf{x}) \geq \mathbf{0}$, we simply optimize

$$\sum_{i=1}^{N} f(\mathbf{x}_i) + k \sum_{i=1}^{N-1} \|\mathbf{x}_{i+1} - \mathbf{x}_i\|^2 \tag{3.10}$$

subject to

$$\begin{aligned} &g(\mathbf{x}_1) = \mathbf{0}, g(\mathbf{x}_2) = \mathbf{0}, \ldots, g(\mathbf{x}_N) = \mathbf{0} \\ &h(\mathbf{x}_1) \geq \mathbf{0}, h(\mathbf{x}_2) \geq \mathbf{0}, \ldots h(\mathbf{x}_N) \geq \mathbf{0} \\ &\mathbf{x}_1 = \mathbf{x}', \mathbf{x}_{N-1} = \mathbf{x}'' \end{aligned} \tag{3.11}$$

where the spring constant k must be chosen to tune the tradeoff between even spacing of the replicas and minimization of the path energy.

3.2.3 Combining the metabolic and evolutionary pathfinding models

To apply the method to the metabolic model, we first obtain predictions for the initial C3 and final C4 states. For the C3 endpoint, we fix the parameters in table 3.1 to their C3 values where applicable, set $P_{max} = 1000$, and maximize A subject to the usual constraints given above, plus additional constraints requiring that 5% of total Rubisco activity be located in the bundle sheath, 5% of total biomass production take place in the bundle sheath, the bundle sheath CO_2 level be less than that in the mesophyll, and metabolite diffusion through the plasmodesmata occur only outward (from bundle sheath to mesophyll) except for oxygen, CO_2, and sucrose. For the C4 case, we maximize A subject to the usual constraints and $P_{max} = 1000$, allowing the parameters in table 3.1 to take any value in their allowed range (verifying afterwards that their C4 values are chosen and that the solution is otherwise C4-like).

Then we make M copies (typically 25) of the model, each obeying the usual constraints above and $P_{max} = 1000$, and follow the procedure sketched above

(eqs. 3.10-3.11), minimizing a slightly modified objective function

$$-\sum_{i=1}^{M} A_i + \sum_{i=1}^{M-1} \sum_{j \in Z} \left(x_{ij} - x_{(i-1)j}\right)^2 . \tag{3.12}$$

That is, the elastic term in the objective function takes into account only a subset Z of the variables in each replica of the model – specifically the values of the protein levels P_i, the kinetic parameters k_p and $E_{R,\text{relative}}$, and the diffusive conductance g_s, which we take to be the only variables under direct genetic control and call the 'evolving' variables in each replica. (Non-'evolving' variables in the model include the rates of the reactions catalyzed by these enzymes, which we assume follow from the protein levels through post-transcriptional and kinetic regulation; the rates of passive or incompletely described transport reactions; rates of biomass synthesis reactions; variables such as k_C, k_O, S_R, $v_{c,\text{active}}$, $v_{o,\text{active}}$ which are constrained or set by the values of evolving variables in various ways; and CO_2 and O_2 levels, which are set as parameters, or controlled by reaction rates.)

We denote the 793-dimensional vectors of the values of the evolving variables, as $\mathbf{z}_0, \mathbf{z}_1, \ldots, \mathbf{z}_M$. The endpoints $\mathbf{z}_0$ and $\mathbf{z}_M$ fixed to the C3 and C4 values found above. (Most precisely, Z includes the protein levels, g_s, and auxiliary variables constrained to equal, somewhat arbitrarily, $10 \cdot k_p$ and $100 \cdot E_{R,\text{relative}}$, as the small absolute changes allowed in the unrescaled parameters k_p and $E_{R,\text{relative}}$ would otherwise make their contributions to the objective function unreasonably small, given their accepted importance to the transition process.)

3.2.4 Limitations of the approach

Obviously, the results of such calculations require careful interpretation. The procedure is not intended to be a mechanistic model of the real process of evolution. Changes in parameters along the optimal paths through the space cannot be understood as being determined by a process of selection acting on variation in a population, even in a highly abstract sense; moreover, the number and precise positions of the replicas along the path are not particularly meaningful in themselves, and they do not represent individual mutations or other biologically defined stages or steps. Finally, there is no direct notion of time; even a simulated path which accurately reflects the real progress of a plant population from a C3 phenotype to a C4 phenotype will provide no information about the number of generations separating any two phenotypic states of interest. Nonetheless, we hypothesized that the simulated high fitness paths would tend to share qualitative features with real evolutionary histories, if only because both evolution and the simulation process would tend to avoid regions of decreased fitness between the endpoints, so long as alternative paths existed.

3.3 Results

3.3.1 Fitness increases and path geometry

Figure 3.1 shows the surface defined by the simulated transition paths at different values of the spring constant k, using default parameter values with $C_i = 200$. To visualize the paths through the 793-dimensional space of evo-

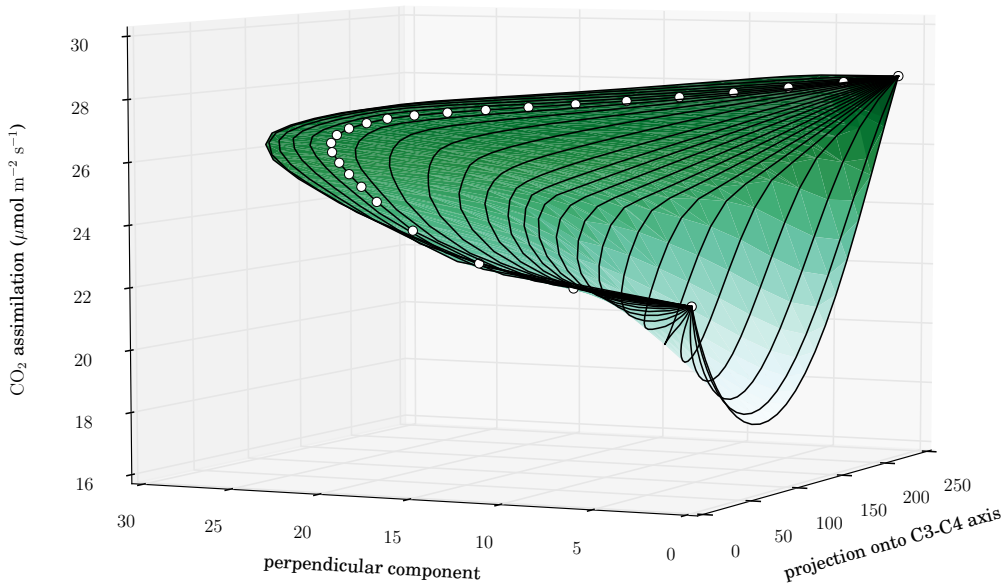

Figure 3.1: **Simulated fitness landscape between C3 and C4 states.** Each dark line represents a different value of the spring constant, which varies (evenly spaced on a logarithmic scale) from 0.01 for the left-most, highest-fitness path to 25.0 for the nearly straight path at right. At low values of the spring constant, the elastic band method successfully finds higher-fitness paths that avoid the fitness barrier separating the endpoints along the more direct path. White circles show the positions of individual replicas along the path for $k = 0.049$.

lutionary parameters, for each step i the parameter-space displacement $\mathbf{z_i} - \mathbf{z_0}$ in the has been decomposed into a component $\mathbf{x}_i$ parallel to the overall C3-C4 parameter shift, $\mathbf{z}_{25} - \mathbf{z}_0$, and a perpendicular component $\mathbf{y}_i$ (not necessarily parallel to $\mathbf{y}_j$, $j \neq i$); the fitness (CO_2 assimilation rate) is plotted versus $\mathbf{x}$ and $\mathbf{y}$ for each path.

For the largest values of the spring constant, the path is a nearly straight line from the C3 state to the C4 state, and a distinct low-fitness valley separates

the two endpoints. As the spring is loosened, the path stretches away from the straight line and the average fitness increases; eventually, the path escapes the valley, following a route of monotonic fitness increase from C3 to C4.

As the spring is loosened, later steps tend to bunch together near the high-fitness C4 state while earlier steps spread out, inefficiently sampling the important early stages of the transition and potentially concealing fitness variations (see Figure 3.3.1, which also shows the fitness along an exact straight line from the C3 to C4 points, corresponding to $k = \infty$). Spring constants below $k = 0.01$ appeared to lead to progressively less regular sampling while maintaining very similar path geometries and fitness results, so this value was used in the remaining calculations to balance regular image spacing and exploration of the landscape.

3.3.2 Development of the C4 system

Figure 3.3 shows the emergence of the C4 phenotype along the simulated transition path in detail. Consistent with the established theory of the C3-C4 transition as discussed above, in the first steps, photorespiratory glycine decarboxylase activity rapidly moves from the mesophyll cells to the bundle sheath cells, increasing the carbon dioxide level there. The rate of bundle sheath photorespiration then falls as the C4 system activity (represented by PEPC in the mesophyll and the decarboxylating enzymes NAD-ME, NADP-ME, and PEPCK in the bundle sheath) increases, further elevating the bundle sheath CO_2 level.

In contrast to the conventional view, the migration of Rubisco from mesophyll to bundle sheath and the increase in C4 system activity begin alongside

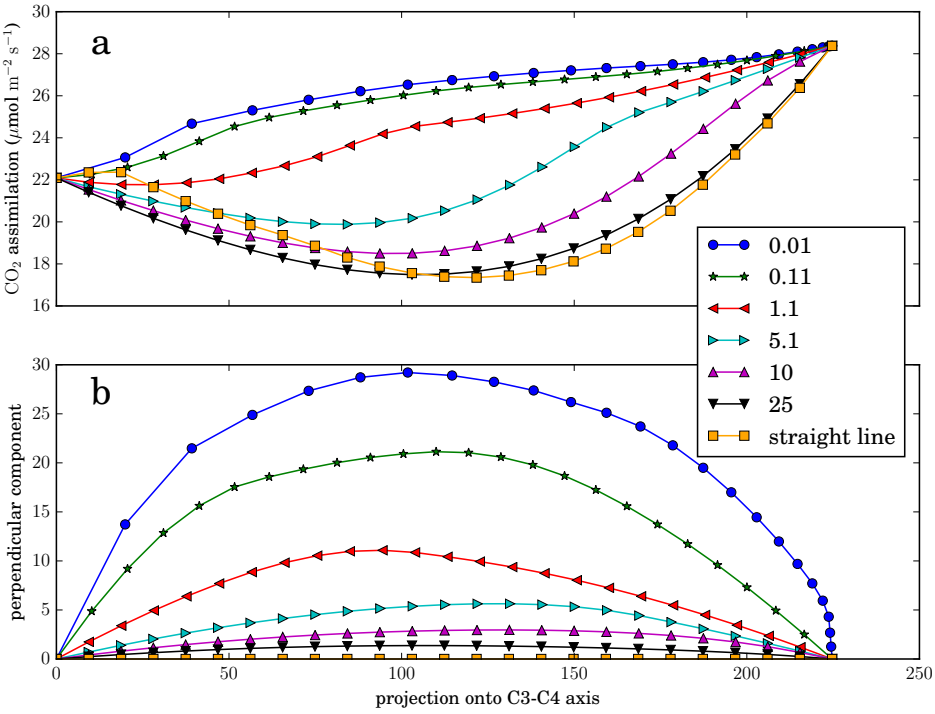

Figure 3.2: **Alternate view of the shapes and fitness values of the straight-line and elastic band paths for various spring constants.** For a subset of the results plotted in figure 3.1, plus the straight-line path from the C3 to C4 points in parameter space, the fitness (i.e., CO_2 assimilation rate; upper panel) and displacement perpendicular to the C3-C4 axis in parameter space (lower panel) are plotted versus the projection of the parameter space position onto the C3-C4 axis.

the shift in photorespiration and continue after it is complete, rather than marking a distinct phase that begins when the photorespiratory shuttle is already well-established.

It is inevitable that the spring term in the elastic band objective will round out sharp corners in parameter space to some degree, so that boundaries between separate phases of the transition will tend to blur; however, the early migration of Rubisco to the bundle sheath results not (only) from this effect but

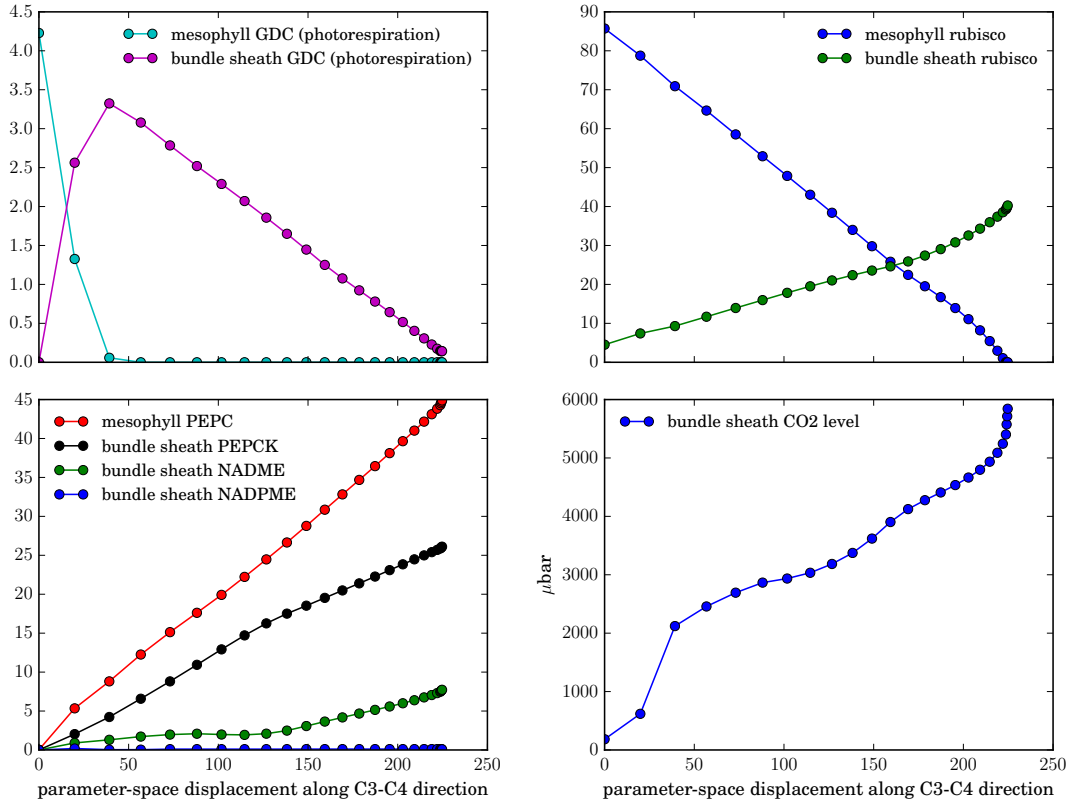

Figure 3.3: **Changes in activity levels of key enzymes and rates of biomass synthesis through the simulated C3-C4 transition.** Enzyme levels are plotted versus the projection of their parameter-space displacement from the C3 starting point onto the C3-C4 axis, normalized by the parameter-space distance between the C3 and C4 states.

from the extremely rapid increase in the bundle sheath CO_2 level, which exceeds the mesophyll level almost immediately after deviating from from the C3 end state as a result of the increased bundle sheath GDC activity and the rapid decrease in the bundle sheath CO_2 conductance g_s (shown below).

Apparent role for anatomical preconditioning

Also notable is the rapid shift in biomass production from the mesophyll to the bundle sheath, with the bundle sheath share of production immediately rising

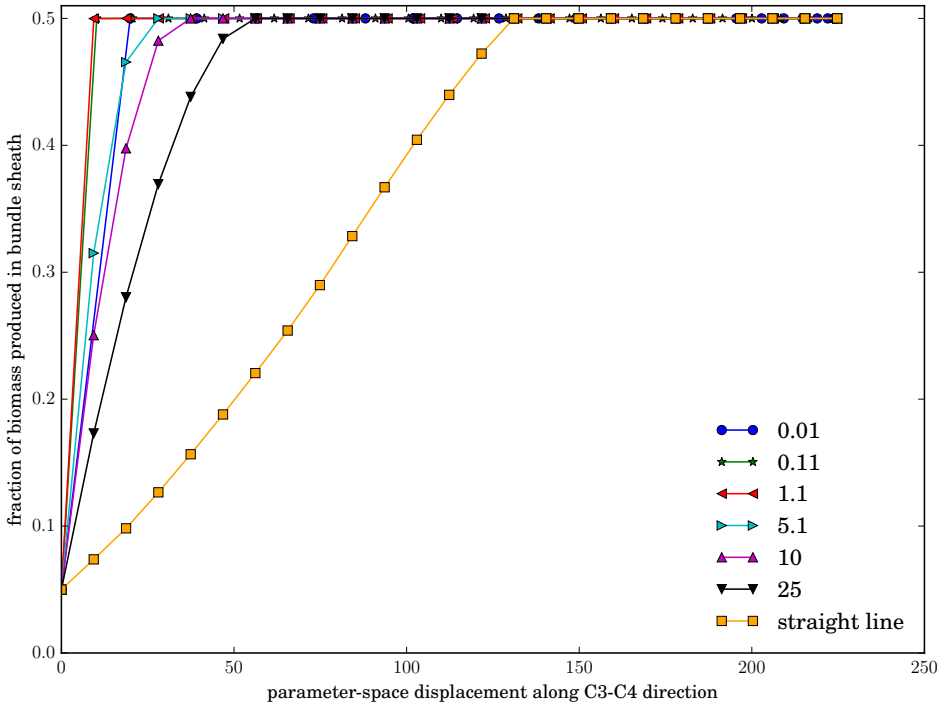

Figure 3.4: **Fraction of biomass produced in bundle sheath.** Default parameter values.

from 5% at the C3 end almost to the maximum allowed value, 50%. As most of the individual biomass components are unlikely to be produced in one cell type and transferred to the other, this behavior, which is not sensitive to the choice of spring constant (section 3.3.2), represents an extremely rapid reconfiguration of leaf anatomy to expand the size of the bundle sheath compartment – through expansion of the bundle sheath cells themselves and/or decreased spacing between veins – well before the C4 system is fully operational. (As the surface area of the mesophyll-bundle sheath interface would increase in the process, this further implies that the concurrent decrease in total conductivity arises from an even more rapid decrease in permeability on a per-area basis.)

These results are consistent with the theory that C3-C4 transitions are likely to be preceded by substantial 'anatomical preconditioning', that is, to occur in C3 species whose leaves have taken on C4-like characteristics for non-photosynthetic reasons– such as improved wind resistance or decreased loss of water to evaporation [14] and maintenance of hydraulic conductivity in dry environments [132]– or through random drift with little impact on fitness [16]. At least in the grasses, this theory is supported by phylogenetic studies: specifically, it appears the C4 system was likelier to evolve in species where the bundle sheath cells made up a larger proportion of the leaf tissue [16].

However, it is not immediately clear how the early increase in the bundle sheath share of biomass increases fitness within the framework of the model, and such a rapid shift in (implied) leaf anatomy may also suggest that the model systematically underestimates the 'evolutionary distance' between C3 and C4-like leaf anatomies, in terms of the number of independent mutations required, the probabilities of their occurrence, and the possibility that they might come at a cost in fitness not reflected in the modeled photosynthetic rate (e.g., the need to invest more resources to produce a leaf of equivalent area).

3.3.3 Comparison to the model of Heckmann et al.

Figure 3.5 shows predicted changes in the six values corresponding to the parameters of the model of Heckmann et al. [52]. There, a highly modular path was predicted, with the migration of photorespiratory decarboxylation to the bundle sheath followed by a phase of increase in PEPC and the bundle sheath Rubisco level, followed by shifts in the PEPC Michaelis-Menten constant, fol-

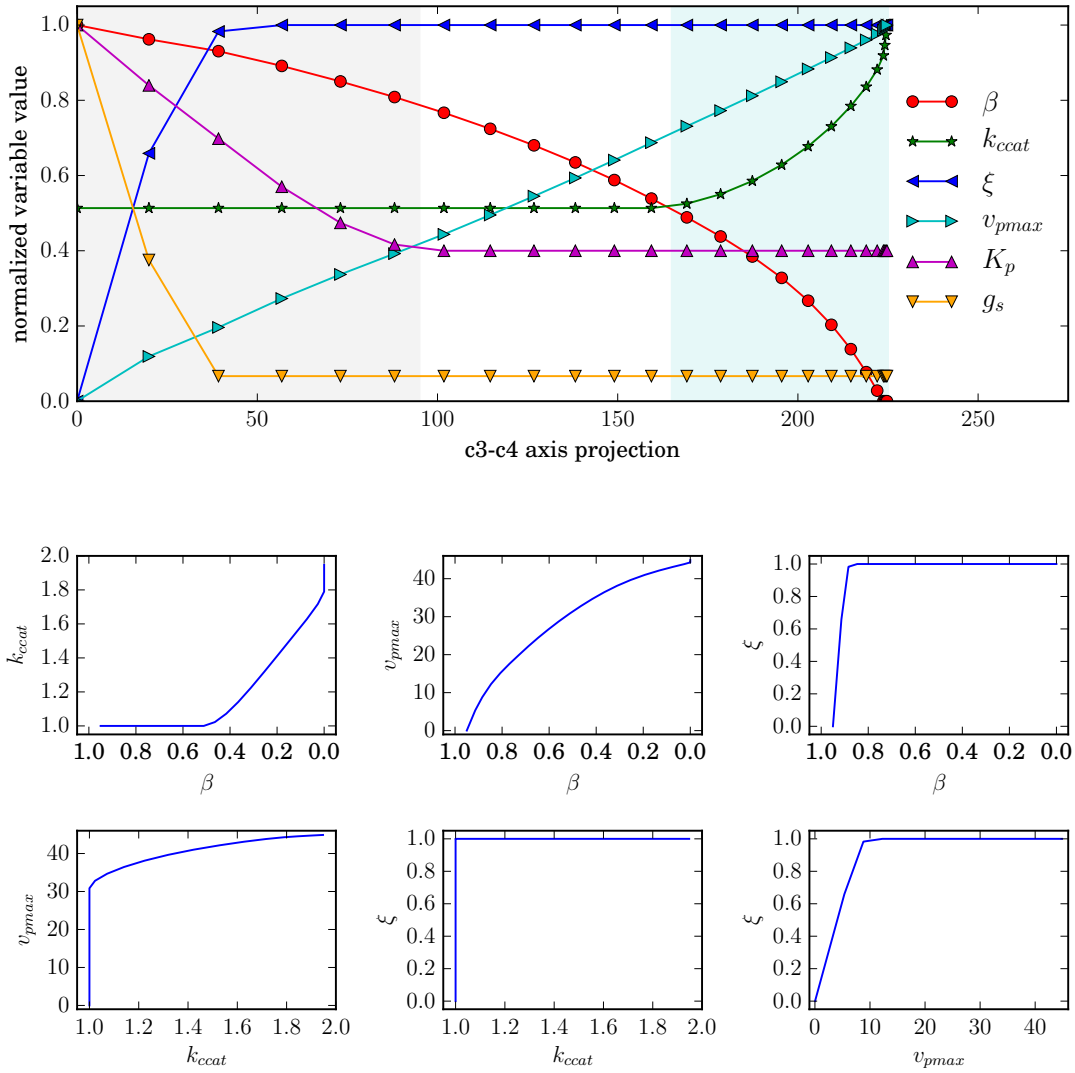

Figure 3.5: **Comparison of predicted paths to the simulations of Heckmann et al.** Upper panel: trajectories of parameters in the present model corresponding to the six variables of the model of [52] (compare to their fig. 4). Each has been normalized to its maximum value along the path. Shaded background highlights the three modules described in the text. Lower panel: selected two-dimensional projections of the simulated path (compare to figure 5 of [52]).

lowed by decreases in the bundle sheath conductivity, followed by final tuning of the Rubisco kinetic parameters, which also varied slightly in earlier stages of the transition.

In contrast, here, the migration of decarboxylation to the bundle sheath is

accompanied by rapid decrease in the conductivity, with both transitions nearly complete after the short fraction of the path represented by the first two replicas in the band. The decrease in the PEPC K_m also begins immediately and is complete within the first half of the path. As noted above, the increase in PEPC levels and migration of Rubisco to the bundle sheath begin immediately and continue throughout the simulated transition, with the latter process accelerating towards the end. The increase in Rubisco efficiency again occurs at the C4 end of the path; here, no changes in this parameter are seen in earlier stages of the transition.

Effectively, where the earlier work found five distinct modules in the evolution from C3 to C4, this model predicts three: an initial phase in which GDC moves completely to the bundle sheath, g_s and K_p fall to their mature C4 values, and migration of Rubisco and the rise of the C4 cycle begin (highlighted in grey in fig. 3.5); an intermediate phase in which Rubisco continues to move to the bundle sheath and the C4 system continues to rise; and a final phase in which the migration of Rubisco accelerates and its kinetic parameters are optimized (highlighted in blue in fig. 3.5).

Despite these differences in the timing of events, the six two-dimensional projections of the path shown in fig. 3.5b are broadly consistent with those predicted by Heckmann et al., and by extension with the biochemical data from C3-C4 intermediate species presented there.

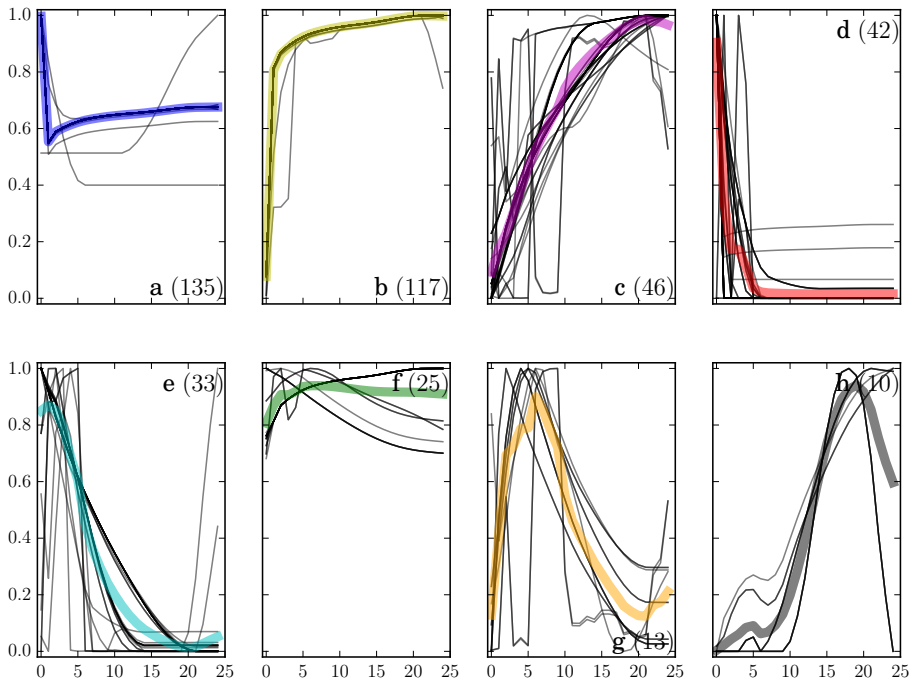

Figure 3.6: **Trends in values of the evolving parameters through the simulated C3-C4 transition.** Colored highlighting indicates the centroid of each cluster. Numbers in parentheses give the total number of parameters assigned to each cluster.

3.3.4 Clustering analysis

Looking beyond the key photosynthetic enzyme levels and kinetic parameters, 421 of the 793 evolving parameters take on values greater than 10^{-3} in at least one point on the simulated transition path. The patterns of changes in these parameters along the path are diverse. Figure 3.6 illustrates some characteristic responses. There, the values of each parameter have been normalized by its maximum value along the path, and the normalized trajectories divided into eight categories by k-means clustering. (Eight clusters were used because cluster shapes became less distinct for larger values of k.)

Overrepresented pathways in the k-means clusters of Fig. 3.6 presents an analysis of metabolic pathway overrepresentation in the clusters. From those calculations, and direct examination of the cluster compositions we conclude that the largest two clusters (**a** and **b**) include primarily protein levels tightly correlated with overall biomass production rates in the mesophyll and bundle sheath respectively. Early in the transition, the parameters in cluster **a** fall sharply and those in cluster **b** rise sharply, as biomass production shifts between cell types; parameters in both clusters then rise as the total rate of biomass synthesis increases along with the assimilation rate. The PEPC and Rubisco kinetic parameters also are grouped with cluster **a**.

Cluster **c**, showing a steady rise from C3 to C4, includes protein levels corresponding to the bundle sheath Calvin cycle, light reactions of photosynthesis, glycolysis, the mitochondrial electron transport chain and TCA cycle, and sucrose synthesis, as well as the C4 cycle enzymes (PEPC and carbonic anhydrase in the mesophyll, PEPCK in the bundle sheath, and aspartate aminotransferase in both cell types) and lower glycolysis in the mesophyll.Many mesophyll counterparts for these reactions are found in cluster **e**, which shows a moderately paced decrease from C3 to C4; in addition to the mesophyll Calvin cycle, it includes the reductive phase of the bundle sheath Calvin cycle.

Cluster **d**, showing a rapid decrease to low levels in the first few steps, includes mesophyll photorespiration, mitochondrial electron transport, and nitrogen assimilation (as well as the bundle sheath copies of several enzymes of the C4 system which are localized to the mesophyll in the C4 state but are active at low levels in the C3 endpoint for unclear reasons, including PEPC and pyruvate, orthophosphate dikinase).

Cluster **f** includes reactions whose rates change only modestly (in relative terms) across the transition, including the reactions from bundle sheath TCA cycle, bundle sheath sulfur reduction, bundle sheath lower glycolysis, and the mesophyll light reactions.

Cluster **g** includes reactions which climb to an early peak and then fall, primarily the reactions of bundle sheath photorespiration, also bundle sheath pyruvate kinase.

Cluster **h**, including reactions which climb rapidly mid-path to a late peak, then fall, includes the Mehler reaction and superoxide radicals detoxification in the bundle sheath chloroplast, and reactions involved in PEP regeneration in the mesophyll (PPDK, pyrophosphatase, and adenylate kinase).

3.3.5 Varying environmental conditions

Varying atmospheric CO_2 levels have historically contributed to the emergence of the C4 phenotype [24]. To examine the influence of this aspect of the environment on the fitness landscape, we simulated transition paths for six values of the intercellular carbon dioxide partial pressure, ranging from 50 to 300 microbar. Figure 3.7a shows the predicted increase in fitness along the paths. As expected, the C3 state is much more sensitive to reductions in the atmospheric CO_2 level than the C4 state, so a greater increase in fitness along the path is observed for lower levels. Although the C3 and C4 ends of the paths have been aligned here for clarity, enzyme allocations in the end states vary significantly with CO_2, as can be seen in fig. 3.7b.

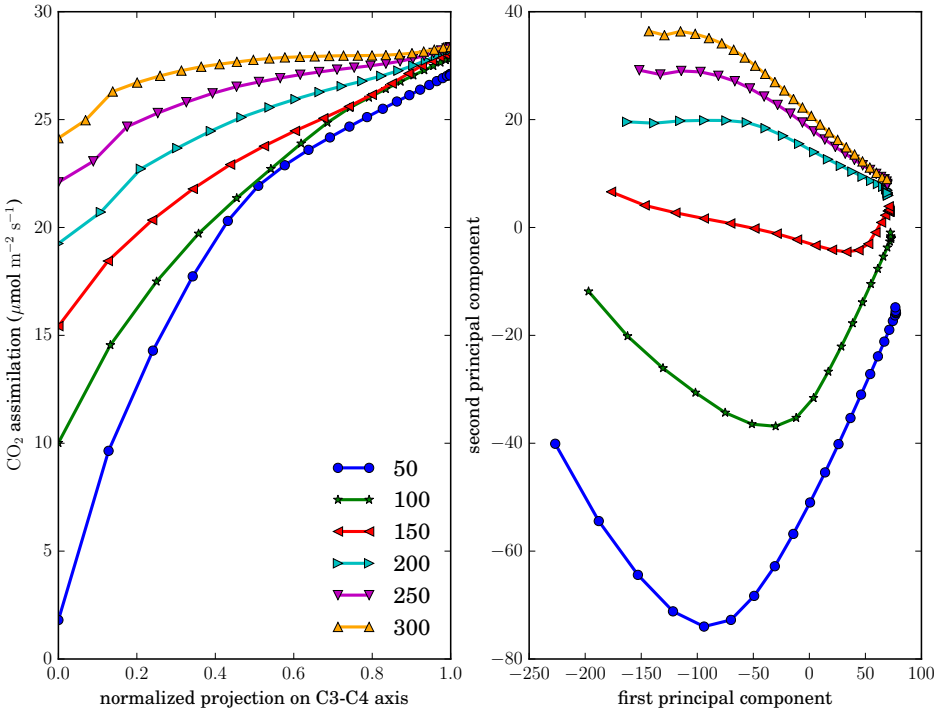

Figure 3.7: **Simulated C3-C4 transitions for varying external CO_2 levels.** Left panel, fitness versus path position, projected in each case onto the straight line in parameter space between the C3 and C4 endpoints and normalized by the length of that line. Right panel, positions in the space of evolutionary parameters, projected onto their first and second principal components; for each path, the C3 endpoint is at left.

There, the positions in the space of evolving parameters have been projected onto their first and second principal components. The first principal component corresponds to (moving right to left in the figure) the migration of Rubisco, the light reactions, and the Calvin cycle from mesophyll to bundle sheath, accompanied by increases in the reactions of the C4 cycle; the second represents (moving from bottom to top) decreases in the levels of mesophyll Rubisco and PEPC, increases in the levels of mesophyll Calvin cycle and light reactions, and decreases in the levels of C4 system enzymes. The pronounced downward swing

in the simulated paths at 50 and 100 microbar thus indicates those paths maintain higher levels of mesophyll PEPC and Rubisco relative to other enzymes while accelerating the establishment of the C4 cycle (though note that some of the Rubisco is inactive).

This is a predictable low-CO_2 response: higher protein levels of PEPC and Rubisco are required per unit flux, the cost associated with the spring stretch prevents the high C3 Rubisco protein level from falling as fast as the Rubisco flux, and the drastic fitness disparity between the C3 and C4 states promotes rapid establishment of the C4 system. These two straightforward components together explain 97% of the observed variance in parameter-space position, suggesting no unexpected qualitative changes in the structure of the transition path occur in response to this environmental perturbation. Examination of individual variables generally confirms this. An exception is the timing of the changes in Rubisco kinetic parameters, which occurs earlier at lower CO_2 levels, relative to the migration of Rubisco to the bundle sheath (Fig. 3.8).

3.3.6 Varying decarboxylation subtypes

By adjusting the efficiency parameters for NAD-malic enzyme, NADP-malic enzyme, PEP carboxykinase, and Rubisco, it is possible to control the combination of decarboxylating enzymes active in the bundle sheath in the most efficient C4 state. Figure 3.9 shows simulated evolutionary paths between a common C3 state and six different C4 states, representing pure and mixed varieties of the NADPME, NADME, and PEPCK subtypes. In each 'pure' case, the primary decarboxylase accounts for over 99% of the C4 decarboxylase activity; in each

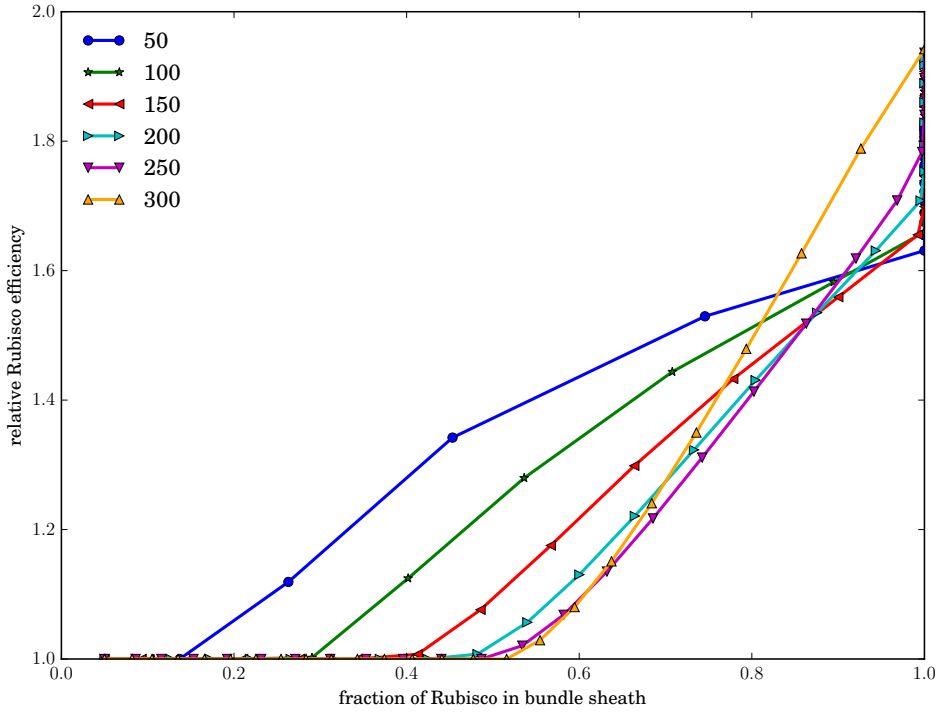

Figure 3.8: **Changes in the timing of adaptation of the Rubisco kinetic parameters, relative to the migration of Rubisco to the bundle sheath, in response to changes in the atmospheric CO_2 level.**

'mixed' case, 70% to 80%. The efficiency parameters for each simulation are given in Figure 3.3.6.

As above, principal components analysis is used to visualize the simulated transition paths and their differences. Here, the first principal component corresponds to migration of Rubisco, the Calvin cycle, the light reactions, and mitochondrial electron transport from mesophyll to bundle sheath, accompanied by increased Rubisco efficiency and increases in components of the C4 system, both those which are not specific to any one subtype (PEPC, carbonic anhydrase, and PPDK, for example) and (at generally lower rates) those which are (the decarboxylating enzymes themselves, aspartate aminotransferases in both

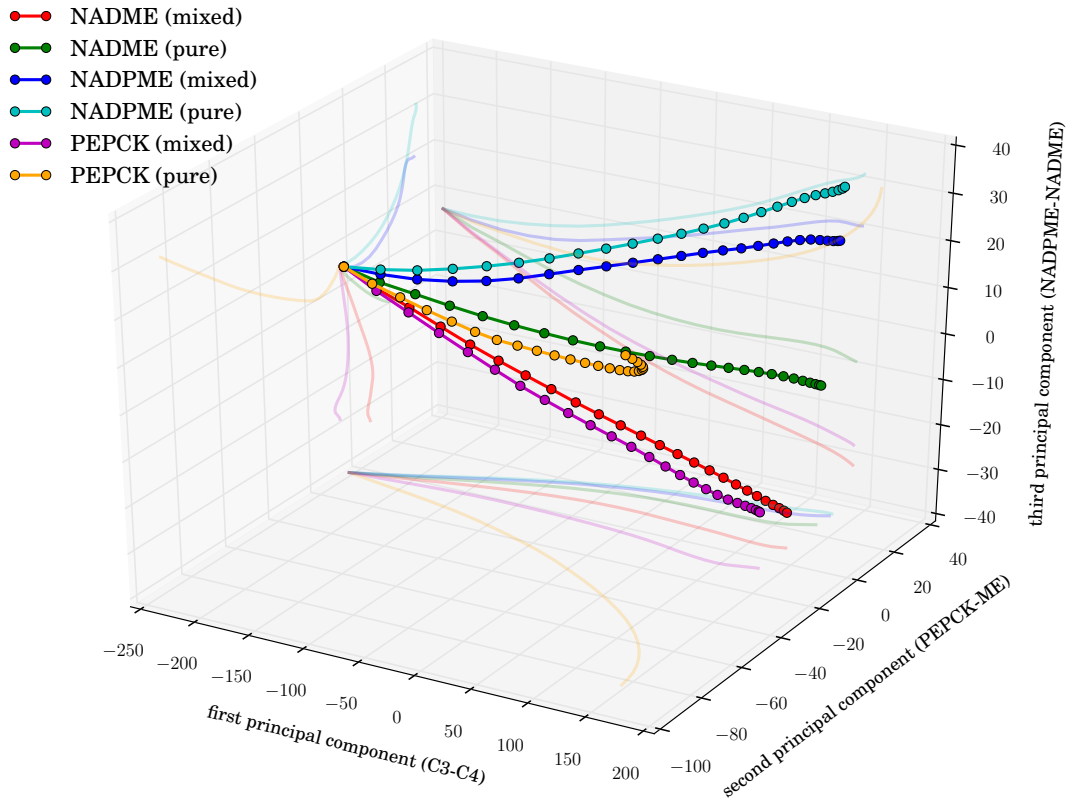

Figure 3.9: **Simulated paths from the C3 state to C4 states using six different combinations of decarboxylating enzymes.** Positions in the high-dimensional space of evolutionary parameters have been projected onto their first three principal components, here labeled according to their interpretations in the text. Faint lines show projections of the paths onto the three two-dimensional subspaces. C3 endpoint at left. Intercellular CO_2 level, 200 µbar.

Subtype	E_{PEPCK}	E_{NADME}	E_{NADPME}
mixed NADPME	0.1	0.5	1.0
mixed NADME	0.25	0.5	0.25
pure NADME	0.25	1.0	0.25
pure NADPME	0.25	0.25	1.0
mixed PEPCK	0.5	0.5	0.5
pure PEPCK	1.0	0.25	0.25

Table 3.2: **Efficiency parameters for the bundle sheath decarboxylating enzymes applied to obtain the differing C4 endpoints of fig. 3.9.**
NADPME efficiencies were applied to both the cytosolic and chloroplastic forms. For these calculations, the efficiency of Rubisco was set to 0.2 (rather than the usual (very conservative) 0.5); this amplified the effect of the changes in decarboxylase efficiencies.

cell types, malate dehydrogenases, etc.); effectively, this component separates C3 states and generic C4 states.

The second principal component encompasses (moving from front to back in the figure) decreases in PEPCK in the bundle sheath, reallocation of photosystem I and (to a lesser extent) photosystem II activity and the reductive Calvin cycle from bundle sheath to mesophyll, increases in pyruvate kinase, NADME and NADPME in the bundle sheath. The third (moving upwards) includes to further decrease in PEPCK, decrease in NADME, increase in NADPME and NADP-MDH, shift in photosystem I from the mesophyll to the bundle sheath. Though the correspondence is not perfect, generally the first component separates paths with PEPCK-type C4 endpoints from those with malic-enzyme-dominated C4 endpoints, while the third separates NADP-ME and mixed NADP-ME/NAD-ME type C4 systems from NAD-ME and mixed NAD-ME/PEPCK C4 systems.

These components collectively explain 98.1% of the observed variation in the points along these six simulated paths. As with the paths simulated at different CO_2 levels, the differences are largely predictable, following from the different levels of the decarboxylases themselves, their accessory enzymes, and their different energy requirements (e.g., in the malic enzyme types the reductive Calvin cycle is localized to the mesophyll and driven by higher light reaction levels in the mesophyll chloroplast). Variation in intermediate points directly reflects variation in the endpoints rather than any apparent flexibility in the optimal sequence of events leading from the C3 state to the C4 state.

Comparison to other models of the biochemical subtypes

Wang et al. [28] analyzed a kinetic model of the C4 system and argued that plants which mixed a malic enzyme pathway with a PEPCK pathway required lower metabolite concentrations to drive necessary rates of diffusion between cell types, had a more flexible distribution of energy use requirements across the mesophyll and bundle sheath, and so could better tolerate fluctuations in incoming light; and further, that pure PEPCK-type systems were unlikely to arise because they required an unrealistically high proportion of incoming light energy to be absorbed in the bundle sheath, which is naturally shaded to some extent by the surrounding mesophyll cells. The present calculations agree that the pure PEPCK type is distinguished by requiring a much larger share of light uptake to occur in the bundle sheath (section 3.3.6, panel a), but we find limited differences in fitness between the subtypes (section 3.3.6, panel b) and those that are seen are sensitive to our assumptions about relative enzyme efficiencies. Thus, while kinetic and leaf-geometry considerations as considered by [28] may promote the evolution of mixed malic enzyme/PEPCK types, all three pure types, as well as mixtures, are accessible at a purely stoichiometric level.

Further supporting the idea that the observed characteristics of the C4 subtypes are not controlled by the metabolic network structure alone, we find that all six simulated subtypes use aspartate rather than malate as the sole carrier of carbon from the mesophyll to the bundle sheath (usually, either solely malate or a mix of malate and aspartate is expected, outside of the pure PEPCK subtype [28];) nitrogen balanced is maintained by exporting glutamate from bundle sheath to mesophyll, with 2-ketoglutarate returned in exchange (rather than, as expected, exchanging alanine for pyruvate).

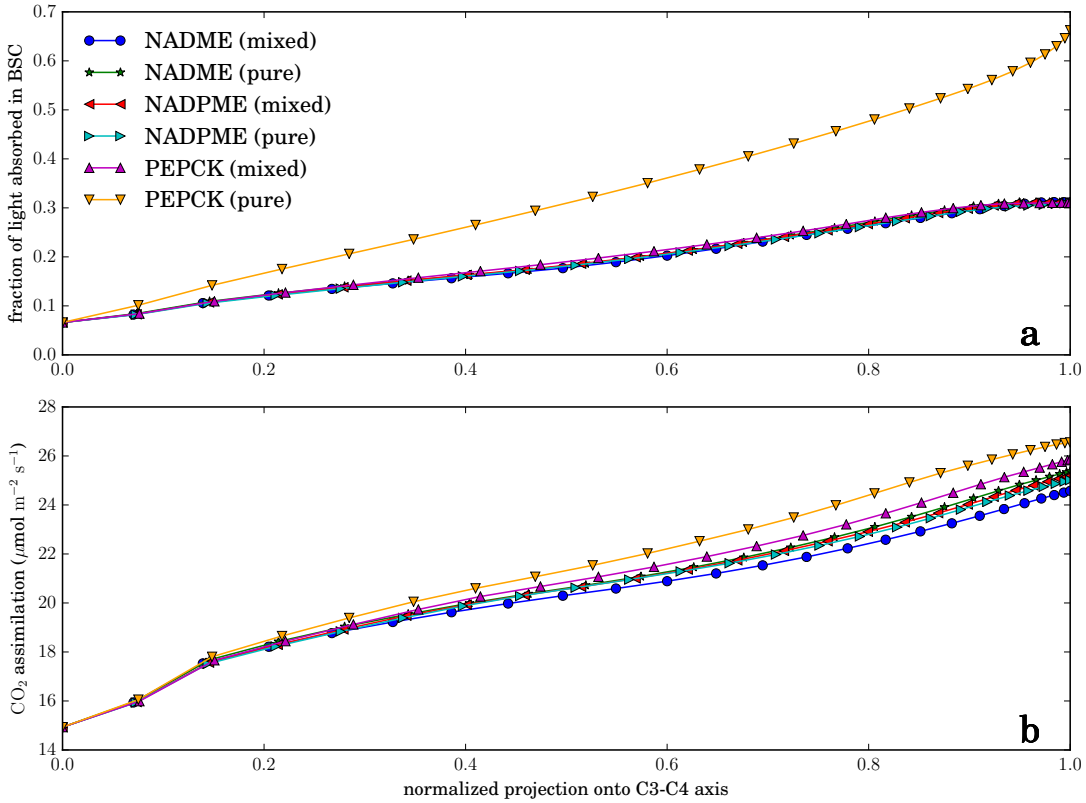

Figure 3.10: **Fraction of total light use taking place in bundle sheath cells along the evolutionary paths for the six simulated subtypes (a), and fitness differences along those paths (b).** Intercellular CO_2 level 200 µbar.

3.3.7 Combined environmental and biochemical variation

Finally, to explore the biochemical and environmental axes of variability in tandem, we repeated the simulations for the six different decarboxylase combinations at intercellular CO_2 levels of 100 and 300 µbar. After normalizing protein level parameters by the overall CO_2 assimilation rate in each image (to minimize apparent differences between the paths solely due to the well-understood reduction in A with decreasing CO_2), vectors of results for each evolutionary parameter and each combination of CO_2 level and decarboxylation subtype were hierarchically clustered using a correlation-coefficient-based metric (after zero-

centering, using scipy.cluster.hierarchy with the UPGMA method [133]).

subsection 3.3.7 shows the result. Even after the normalization by overall CO_2 assimilation rate, transition paths using one decarboxylation subtype are generally more similar to paths using a different decarboxylation subtype at the same CO_2 level than they are to paths of the same subtype at a different CO_2 level, but (as with the results for the reference decarboxylation type, above) changes in the behavior of individual parameters in response to variation in CO_2 generally appeared to be quantitative rather than qualitative (e.g., changes in sequence of events, recruitment of different enzymes, and so on). The hierarchical structure of the evolving parameters generally recapitulates patterns seen in the k-means clustering, above, with biomass-synthesizing reactions in mesophyll and bundle sheath forming two large, fairly tight clusters, while reactions directly involved in photosynthesis (C3 and C4), photorespiration, energy metabolism, and inorganic nutrient assimilation show greater variability.

3.4 Assessing the elastic band approximation to the highest-fitness path

The chain-of-states method described above differs slightly from that typically used in applications. It was presented, as the 'plain' elastic band, by Jónsson et al. [131] primarily as a pedagogical device to motivate the development of the 'nudged' variation – where spring forces act only parallel to the path and forces associated with the potential energy surface act only perpendicular to it – in which form the elastic band method has generally been applied. However, the nudged elastic band forces are non-conservative and cannot easily be in-

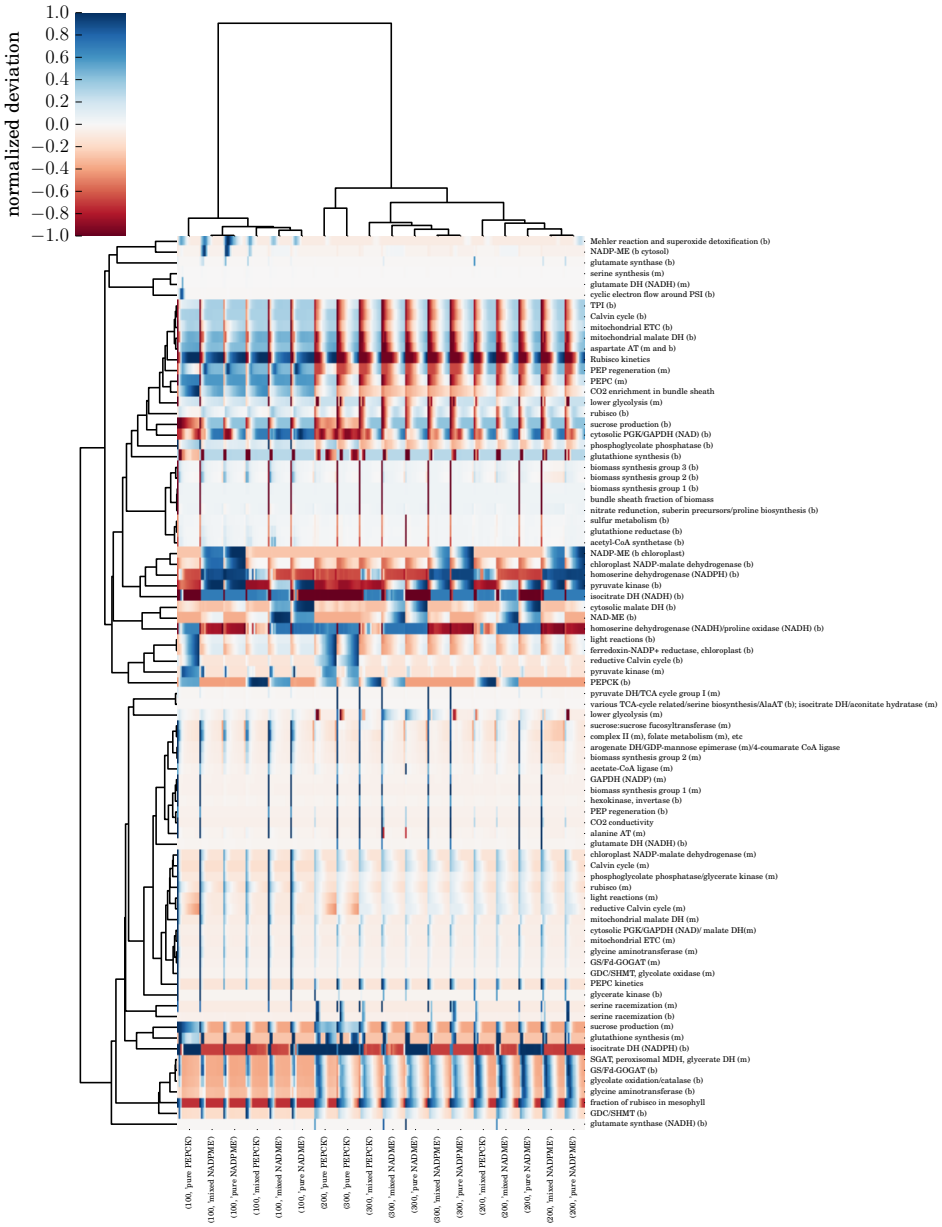

Figure 3.11: **Hierarchical clustering of results for eighteen combinations of intercellular CO_2 levels and decarboxylation subtypes.** Each row of the heat map presents eighteen sequences of 25 values, representing one complete path from C3 to C4 under each of the eighteen conditions. For visualization, each row has been shifted to have *overall* mean zero, and scaled to have maximum absolute value 1.0. Many of the 793 evolutionary parameters are very tightly or perfectly correlated, such as protein levels for reactions which must occur in a set stoichiometric ratio in steady state; here we have presented a summary view in which only one parameter corresponding to each process or group labeled on the right is shown.

corporated in the optimization framework developed to combine eq. 1.1 with eqs. 2.3-3.9. The plain elastic band was thus more practical to implement, but Jónsson indicates that for some problems the resulting approximation to the minimum-energy path may be quite poor for any choice of spring constant k.

We have attempted to assess the performance of the plain elastic band on the present problem by comparing the tangent vector to the elastic band path at each non-endpoint replica with the local direction of steepest ascent in fitness, which is everywhere parallel (or antiparallel) to the true minimum-energy/maximum fitness path. This comparison is complicated by the fact that the tangent vector and the steepest ascent direction can be determined only approximately. In the case of the tangent vector, this results from the discretization of the path. In the case of the direction of steepest ascent, considerations outlined in section 3.2.2 imply that we cannot, e.g., read off the direction of fastest improvement in parameter space from the Lagrange multipliers associated with constraints fixing the evolvable parameters to a given set of values, and although the gradient of the objective function is readily evaluated, to determine its exact projection onto the tangent space of the feasible manifold, respecting all equality and inequality constraints, is highly nontrivial.

It is however straightforward to determine the optimal perturbation δ_z, of norm less than or equal to a given bound c, which could be applied to a point $\mathbf{z}_0$ in the model's parameter space while respecting all constraints (i.e., by adding the constraint $\|\mathbf{z}_o - \mathbf{z}\| \leq c$ to the model and solving normally to obtain the optimal $\mathbf{z}$). For small c, the optimal δ_z will lie along the direction of steepest increase in fitness.

Figure 3.4 shows the results for $c = 0.01$ and $c = 1.0$ (blue and green curves).

The angles range from 10-65 degrees, suggesting local agreement between the elastic band direction and the objective function gradient is generally poor. However, this does not necessarily indicate a large discrepancy between the elastic band path and the true path: not only is the approximation to the tangent vector inexact, but very small deviations from the ideal path could destroy the local alignment of the tangent vector and the objective function gradient if the derivatives of the objective function change rapidly on spatial scales small compared to the overall path length. We may safely conclude that this condition is met in the present problem (for example, consider a reaction playing a key role in synthesis of a very minor biomass component: the associated protein level will be small on an absolute scale, but (if all the available enzymatic capacity is used) decreasing it by 10% (say) will necessarily decrease overall CO_2 assimilation by 10% (as biomass production and carbon assimilation occur in a fixed ratio).

To examine optimality of the path on a coarser scale, the red curve in section 3.4 shows the angle between the vector pointing from each image to the next, $\mathbf{z}_{i+1} - \mathbf{z}_i$, and the optimal step of the same distance $\|\mathbf{z}_{i+1} - \mathbf{z}_i\|$ away from $\mathbf{z}_i$ in any direction (consistent with the constraints). Here the angles are smaller, from 40 degrees to a minimum of 0 degrees (achieved in the step from the penultimate image to the C4 point, which is a global optimum).

As a further test of the sensitivity of the predictions to these apparent deviations between the predicted path and the true maximum fitness path, we considered the modified paths obtained by shifting the images in the reference elastic band path by the optimal perturbations of lengths 0.1 or 1.0 calculated above. The results presented in figs. 3.1-3.3.7 above were essentially unchanged

in the modified paths (data not shown.)

This is somewhat more encouraging but still suggests considerable room for improvement in the approximation to the true maximum-fitness path. Future work should explore more sophisticated chain-of-states methods for transition path simulation, perhaps with particular attention to the finite-temperature string method [134], which is well suited for rough potential energy surfaces. The results from the plain elastic band method, meanwhile, at least provide a lower bound on the extent to which the C3-C4 paths through parameter space may be optimized, and the plain elastic band paths and their features remain interesting as representative of one potential class of paths along which fitness increases monotonically, even if other, somewhat higher fitness paths may also exist.

3.5 Discussion

The approach taken here to the simulation of plausible paths through very high-dimensional fitness landscapes necessarily involves a great deal of abstraction and simplification. Even setting aside the various simplifying assumptions entailed by the underlying flux balance analysis model and limiting our analysis to an objective function that maximizes CO_2 assimilation rate per unit of nitrogen invested in enzymes in photosynthetic tissue, we might still expect that neglecting the (as yet mostly unknown) details of the genetic control of enzyme expression levels in mesophyll and bundle sheath cells – and so forgoing any direct estimation of mutation and fixation rates – would make it impossible to obtain realistic results.

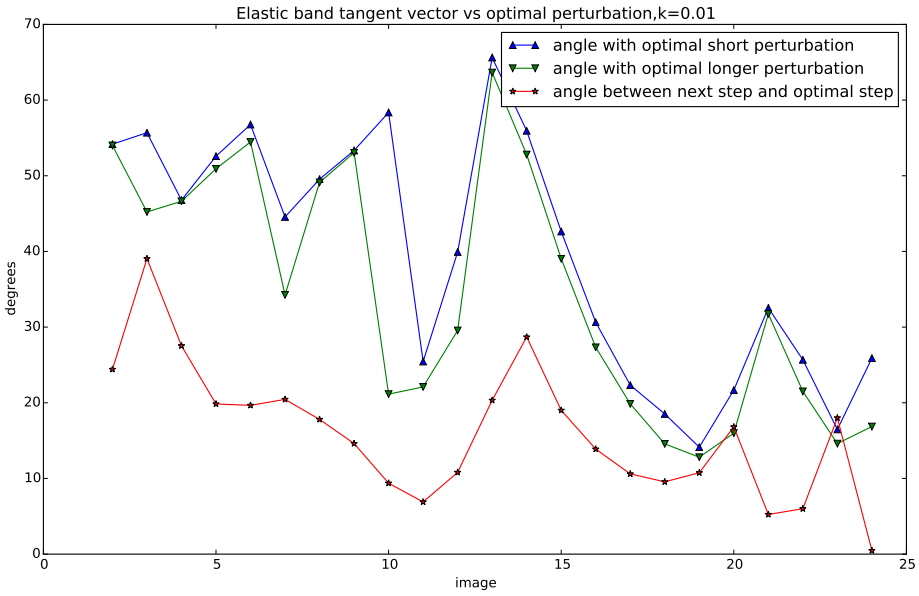

Figure 3.12: **Angles between elastic band path tangent vectors and approximate local directions of steepest improvement in fitness.** At each non-endpoint image in the band with default parameter settings, an approximation to the tangent vector (blue and green curves: centered difference approximation; red curve: direction towards the next image in the band) and the direction of the optimal parameter-space perturbation (blue curve: optimal perturbation of length 0.01; green, length 1.0; red, length equal to the distance to the next image) approximating the direction of steepest increase in fitness, have been computed and the angle between them determined.

Instead, we find that the paths obtained through the purely phenotypic fitness space successfully replicate several widely accepted aspects of historical C3-C4 transitions, most notably the early localization of photorespiratory glycine decarboxylase to the bundle sheath and the late tuning of Rubisco kinetic parameters [14].

As shown in fig. 3.5b, the paths predicted here have a very similar structure to those obtained by Heckmann et al. [52] when projected onto the lower-dimensional parameter space of their model (and thus also generally agree with their compiled data from C3-C4 intermediate species), despite the numerous

methodological differences between our approaches (793 continuous parameters relating to all aspects of central metabolism, biomass production and photosynthesis versus 6 parameters directly related to photosynthesis and photorespiration, discretized into five steps each; multiple parameters allowed to vary between images in arbitrary combinations vs one parameter change per step; all parameters treated symmetrically versus specific per-parameter mutation rate estimates, et cetera).

As well, both approaches predict paths with a distinct modular structure, though each leads to a different number and composition of modules; both indicate that significant 'sign epistasis' can occur, where the same change in one parameter leads to an increase or decrease in fitness depending on the value of other parameters (shown here, for instance, by the difference between the straight-line and optimized paths between identical C3 and C4 states in fig. 3.1;) and both agree that the sequence of events in the transition is quite stable–over many random realizations, in their case, or in response to varying CO_2 levels and biochemical perturbations, here. A notable exception is the prediction here that shifts in Rubisco kinetics will begin earlier along the path at lower CO_2 (fig. 3.8), which we expect would also be obtained from the model of Heckmann et al. if applied over the same range of CO_2 levels.

Heckmann et al. characterized their model's fitness landscape as 'Mt. Fuji'-like, in that no local fitness optima existed other than the C4 state; that is, the C4 state was reachable from every other point in parameter space along a path of nondecreasing fitness. While we have not demonstrated this exhaustively in the current model, all our results suggest the same is true in our higher-dimensional space.

The predicted paths also disagree with established theories and previous simulations in some ways. The increases in PEPC and C4 cycle activity and bundle sheath Rubisco activity begin immediately, rather than in a distinct phase of the transition after the photorespiratory pump has reached a threshold level of activity (and in consequence changes in the PEPC Michaelis-Menten parameter occur early in the transition as well); also, changes in the bundle sheath conductance to CO_2 diffusion occur at the very beginning of the path in parallel with these other changes rather than near the end as in [52].

These aspects of the results are all related to the very rapid predicted increase in the bundle sheath CO_2 level which the fast decrease in g_s allows. As discussed above, this, and the nearly immediate increase in the bundle sheath share of biomass production from 5% to 50%, agree nicely with the apparent importance of anatomical preconditioning in promoting the evolution of the C4 system, but they could also indicate that the model systematically underestimates the significance of the anatomical remodeling of the leaf which such changes would require. Future refinements to the model could address this possibility in two ways.

First, the explicit cost of such changes could be increased. This could be done simply by further rescaling the conductivity contribution to the elastic band cost function, as discussed above, and adding one for the bundle sheath biomass fraction, or through the addition of a more detailed description of leaf anatomy to the model, allowing it to take into account anatomical preconditioning and potentially make predictions about the relative timing of shifts in vein spacing and, e.g., their relationship to light use efficiency [135].

Second, the model's assumptions about the relative costs of shifts in enzyme

expression levels could be improved. Currently, the migration from mesophyll to bundle sheath of 100 enzymes, each expressed at a level corresponding to a maximum rate of 1 µmol per square meter per second, is considered as 'large' a step in parameter space as the migration of a single reaction expressed at a level corresponding to 10 µmol per square meter per second; thus biomass production, which involves many reactions carrying relatively small fluxes, migrates more readily than photosynthesis and photorespiration, which involve a moderate number of reactions of large flux.

However, it would be equally plausible to assume that all mutations which increase expression of an enzyme by a certain percentage are equally 'large', regardless of the enzyme's absolute expression level. This suggests a transformed system, in which the elastic band objective function depends on the logarithms of the current protein level variables, might be more realistic.

Such a refinement could also enhance the model's capacity to examine some of the issues for which a genome-scale model of the C3-C4 transition should be most useful: how reactions outside the core photosynthetic machinery behave in the transition path, and how those responses depend, if at all, on environmental factors and details of the C4 biochemistry. The simulations above offer limited insight into these questions, because the majority of non-photosynthetic processes are tightly correlated with biomass synthesis in one cell type or the other and respond in approximately the same way under all combinations of conditions. While there are likely numerous pathways, outside the core areas of photosynthesis, photorespiration, and energy metabolism, which are not strongly affected in real C4 transitions, such negative results may also be artifacts of an inadequate representation of the evolutionary process, or the conse-

quence of inadequate detail and flexibility in the underlying metabolic model. In future work it may be best to identify and focus on key pathways of potential interest, studying and if necessary expanding the range of behaviors they can display in the model before generating simulated evolutionary paths.

The most important challenge in future extensions of this work will be establishing more direct connections to experimental data. In principle it should be possible to combine this model with a sequence-based enzyme function prediction method to make connections to large-scale genomic and transcriptomic data from either C3-C4 intermediate species or closely related C3-C4 pairs but it is unclear how most of the evolutionary changes simulated here would appear, if at all, in such data, as most cannot distinguish mesophyll and bundle sheath expression and cell-type specificity cannot yet be inferred from regulatory sequences in general.

More broadly, the methods developed here provide a blueprint for the study of other evolutionary transitions between distinct states of large-scale metabolic models. For most existing models, with linear constraints and objective functions, such transitions will not be particularly interesting (with piecewise linear optimal paths), but the development of methods to incorporate nonconvex, nonlinear constraints into FBA models [60] and methods for hybridizing constraint-based and kinetic models [88, 136] will likely give rise to more and more complicated effective fitness landscapes, affording more opportunities to apply these or related techniques.

APPENDIX A

DEVELOPMENT OF A FLUX BALANCE ANALYSIS MODEL FOR MAIZE

This appendix describes the of creation of a metabolic model for maize from the CornCyc database. It covers the creation of an SBML model with exchange and biomass reactions and limited subcellular compartmentalization which can successfully simulate the production of many biomass components and photosynthetic carbon dioxide assimilation, the adaptation of the biomass equation from iRS1563, some considerations in the process of expanding the model to describe interacting mesophyll and bundle sheath compartments, and some modifications made in response to preliminary fitting results.

Sections A.1 through A.7 explain in detail the process of constructing the underlying metabolic model at the one-cell level. Section A.8 discusses in detail changes made to gene associations based on early data fitting results. Section A.9 describes changes to the iRS1563 biomass equation. Section A.10 discusses plasmodesmatal transport in the two-cell model. Filenames referred to are in the `model_development` subdirectory of the project source code.

A.1 Exporting the CornCyc FBA model from Pathway Tools

CornCyc 4.0 [57] was obtained from the Plant Metabolic Network and upgraded from from Pathway Tools 16.5 to 17.0 locally.

The frame `PWY-561` was removed from the database because otherwise some of the reactions of that pathway were excluded from the FBA export, apparently due to a bug.

A simple FBA problem was solved using the Pathway Tools FBA functionality [137], producing an output file which includes all reactions in the FBA model Pathway Tools generates internally, both those which are active in the solution to the FBA problem and those which are not. Note that this list of reactions is distinct from the list of reactions in the database itself; the Pathway Tools software prepares this set of reactions through an extensive process of excluding reactions which are unbalanced or otherwise undesirable while expanding reactions with classes of compounds as products or reactants into sets of possible specific instantiations which respect conservation of mass [77]. Working with the Pathway Tools FBA reaction set (rather than, e.g, an SBML export of the CornCyc database) allows us take advantage of this pre-processing; however, it comes at the cost of needing to reintroduce into the FBA model many reactions which are present in the CornCyc database but are excluded from the FBA export for one reason or another.

Reaction data was extracted from the FBA output file, and reactions were translated to refer to species by their CornCyc frame ID (to allow easy reference to the database and comparison with previous work, and avoid possible ambiguities). Reactions were then added and removed from the model as described below.

A.2 Discarding reactions

A.2.1 Polymerization reactions

Pathway Tools attempts to include an expanded representation of certain polymerization reactions in the exported FBA model, but this function is considered experimental [137]; these reactions were ignored. Note that some reactions representing polymer growth were added manually later in the process.

A.2.2 ATPases

We removed all reactions from CornCyc which have the effective stoichiometry

```
{'ADP': 1.0, 'ATP': -1.0, 'PROTON': 1.0,
'WATER': -1.0, '|Pi|':  1.0}
```

There are nine such reactions:

- RXN-11109,
- 3.6.4.6-RXN,
- RXN-11135,
- RXN0-1061,
- ADENOSINETRIPHOSPHATASE-RXN,
- 3.6.4.4-RXN,
- 3.6.4.9-RXN,
- 3.6.4.5-RXN,
- 3.6.4.3-RXN

all treated as reversible by the Pathway Tools export procedure. Typically these are simplified representations of the metabolic effect of enzymes whose complete function is outside the scope of the database, as, for example, EC 3.6.4.3, the microtubule-severing ATPase.

In their place, we added a single generic ATPase reaction to represent cellular maintenance costs, etc., with no associated genes.

A.2.3 Reactions involving generic electron donors and acceptors

Numerous reactions in the database are written with generic representations of electron carrier species ('a reduced electron acceptor', 'an oxidized electron acceptor'). Most of these reactions are outside the areas of emphasis of the model (e.g., brassinosteroid biosynthesis), have no curated pathway assignment, or also appear in forms which do specify the electron carrier species (e.g., the generic nitrate reductase reaction, `NITRATEREDUCT-RXN`, vs `NITRATE-REDUCTASE-NADH-RXN`,) and so could be safely neglected. A small set of exceptions identified in early drafts included reactions of fatty acid synthesis, handled as discussed below, and proline dehydrogenase, `RXN-821`, catalyzed by a mitochondrial-membrane-bound flavoprotein which donates electrons directly to the mitochondrial electron transport chain [127]. Because we have not thoroughly compartmentalized amino acid metabolism, we implemented this reaction as donating electrons to NAD^+ instead.

A.2.4 Duplicates

A number of other reactions were removed because they appeared to be exact (possibly unintentional) duplicates, down to gene associations, of other reactions in the database; or because they were being replaced by modified forms as discussed below. These are given in `reactions_to_remove.txt`.

A.2.5 Non-metabolic reactions

A number of reactions present in CornCyc were removed because the database indicated, e.g. through the Enzyme Commission summary for the relevant EC number, that they were primarily involved in extrametabolic functions (e.g., cell movement, regulation). These included the GTPases `RXN-5462`, `3.6.5.2-RXN`, and `3.6.5.5-RXN`.

A.2.6 Glucose-6-phospate

In the reduced model (discussed below) only one reaction, myo-inositol-1-phosphate synthase, consumes the generic glucose-6-phosphate species, rather than alpha-G6P or beta-G6P. To ensure that this reaction was appropriately connected to other G6P producing and consuming reactions we manually split it into two instances, one for alpha-G6P and one for beta-G6P.

A.2.7 UDP-glucose

For apparently all reactions in CornCyc involving UDP-glucose, the instantiation procedure produced one version involving generic UDP-D-glucose and one version involving UDP-alpha-D-glucose, the only child of the UDP-D-glucose class. UDP-alpha-D-glucose participated in almost no reactions other than these instantiations (in the reduced model, described below, only one: UDP-sulfoquinovose synthase, EC 3.13.1.1). As such there is little to distinguish the generic and specific versions of the reactions, which add complexity to the model and degeneracy to optimization predictions without providing significant information about the function of the system, so we removed the specific versions and changed the UDP-sulfoquinovose synthase to act on a generic UDP-D-glucose substrate.

A.3 Minor revisions to achieve basic functionality

A.3.1 Mitochondrial electron transport chain

The CornCyc representation of the mitochondrial electron transport pathway (`PWY-3781`, plus the mitochondrial ATPase (`ATPSYN-RXN`, EC 3.6.3.14)) was adjusted. Some reactions excluded from the initial Pathway Tools export because the balance state of reactions involving cytochrome C could not be determined were readded manually; ubiquinones/ubiquinols were uniformly represented as ubiquinone-8/ubiquinol-8, and compartments were assigned to reactants and products to properly represent the transport of protons between the

mitochondrial matrix and the mitochondrial intermembrane space. In CornCyc, as in MetaCyc and other related databases, transport of protons across the membrane is represented explicitly for complex I but not for complex III and complex IV; in agreement with the standard description of mitochondrial electron transport (see, e.g., [138]) proton transport was added to these reactions with a stoichiometry of 2 H+/e- for complex III and 1 H+/e- for complex IV. The stoichiometry of complex IV was further adjusted to include the H+ from the mitochondrial matrix that binds to oxygen to form water.

A.3.2 Photosynthesis: light reactions

Similarly, some modifications were made to the light reactions of photosynthesis (`PWY-101`). Reactions involving plastocyanins were not exported and were added manually; a chloroplastic ATP synthase and a reaction describing cyclic electron transport around PS I were added; and the stoichiometry of proton transport was adjusted in accordance with recent literature, assuming a Q cycle and ratio of 14 H+/3 ATP for the chloroplast ATP synthase [139].

Reduction of oxygen to superoxide at photosystem I (the Mehler reaction) was added to allow flux through the pathways of chloroplastic reactive oxygen species detoxification: superoxide dismutase and the ascorbate-glutathione cycle, including a reaction representing the direct, non-enzymatic reduction of monodehydroascorbate by ferredoxin [140, 141].

A.3.3 Key reactions in biomass component production and nutrient uptake

Several components of biomass required either manual adjustment of reactions from the database or the addition of abstract synthesis reactions summarizing the behavior of pathways which could not easily be represented in more detail.

Starch

Starch synthase (GLYCOGENSYN-RXN) is not exported from CornCyc by default (it is a polymerization reaction, and marked as unbalanced in the PGDB); it was added manually in a form that produces the equivalent of one 1,4-alpha-D-glucan subunit.

The starch branching enzyme EC 2.4.1.18 (RXN-7710) is not exported from CornCyc by default (one reactant, starch, has an unspecified structure); it was added manually as

$$\text{a 1,4-alpha-D-glucan subunit} \rightarrow \text{an amylopectin subunit}$$

Note that this stoichiometry is not intended to suggest that the branching enzyme introduces branches at each subunit.

CornCyc provides a detailed reconstruction of the reactions of starch degradation (PWY-6724) which is by nature difficult to convert to a form suitable for FBA calculations, as many of the stoichiometry coefficients are undefined. To incorporate the effects of the glucan-water and phosophoglucan-water dikinases, for example, we would need to specify how many glucosyl residues must be

phosphorylated (and then dephosphorylated) to produce "an exposed unphosphorylated, unbranched malto-oligosaccharide tail on amylopectin" of a given length; modeling the release of maltose from that tail would require an estimate of the typical unbranched length of such tails, etc. Rather than estimate average values for these parameters, we divide the reactions of the pathway into two types: those which condition starch for depolymerization , and actual depolymerization reactions. The first class (the dikinases above plus isoamylase) share the abstract stoichiometry

$$\text{a starch subunit} \rightarrow \text{an exposed starch subunit}$$

(neglecting any ATP costs), while the second class (beta amylase and disproportionating enzyme) convert exposed starch subunits to sugars appropriately.

The beta-maltose releasing reactions of the starch degradation pathway in CornCyc have no associated genes. We temporarily associated these reactions with the beta amylase record in the database (`RXN-1827`, EC 3.2.1.2) pending further review.

During transient starch degradation, beta-maltose and glucose are exported into the cytosol, where maltose is split, releasing one glucose molecule and donating one glucosyl residue to a cytosolic heteroglycan, from which it may be released in turn as glucose-1-phosphate [142]. In Arabidopsis, specific enzymes (DPE2 and PHS2) are known to be implicated in this process [143]. In simulations with this CornCyc-based FBA model we find the typical mode of breakdown of cytosolic maltose is to alpha-D-glucose and alpha-D-glucose-1-phosphate via `AMYLOMALT-RXN`,

$$\text{maltotriose} + \text{beta-maltose} \rightarrow \text{maltotetraose} + \text{beta-D-glucose}$$

and RXN0-5182,

$$\text{maltotetraose} + \text{phosphate} \rightarrow \text{maltotriose} + \text{alpha-D-glucose-1-phosphate}$$

effectively the standard pathway but with maltotriose/maltotetraose playing the role of the cytosolic heteroglycan pool. This approximation leads to a reasonable effective stoichiometry but it is possible that the genes associated with these reactions do not accurately represent the genes involved in the true underlying process; we have not systematically looked for maize counterparts of the Arabidopsis genes, for example.

Cellulose

The UDP-forming cellulose synthase, EC 2.4.1.12, is not exported from CornCyc by default (it is a polymerization reaction, and marked as unbalanced in the PGDB); it was added manually in a form that produces the equivalent of one subunit.

Hemicellulose

Similarly, the following hemicellulose polymerization reactions were added manually:

- 1,4-beta-D-xylan synthase, EC 2.4.2.24,
- reactions RXN-9093 (EC 2.4.2.-) and RXN-9094 (EC 2.4.1-), representing the addition of arabinose and glucuronate to xylan to form arabinoxylan and glucuronoxylan respectively (note that the corresponding subunits notionally consist of one xylan subunit plus arabionose/glucuronate),

- glucomannan synthase, EC 2.4.1.32,
- `RXN-9461` (EC 2.4.2.39), representing the addition of xylose to a glucan (as implemented, cellulose) to form xyloglucan (again, the corresponding effective subunit corresponds to one glucan subunit plus xylose)– note this representation ignores the previous step in CornCyc's xyloglucan biosynthesis pathway, xyloglycan 4-glucosyltransferase (EC 2.4.1.168).

In addition to these explicit descriptions of hemicellulose formation from CornCyc, we added generic reactions representing the donation of the following sugar residues from activated donor molecules to unspecified generic polysaccharides:

- arabinose (from UDP-L-arabinose)
- galactose (from GDP-L-galactose)
- galacturonate (from UDP-D-galacturonate)
- glucose (from UDP-glucose)
- glucuronate (from UDP-D-glucuronate)
- mannose (from GDP-alpha-D-mannose)
- xylose (from UDP-alpha-D-xylose)

These reactions allow the model to represent flux of these sugars towards hemicelluloses or other polysaccharides without explicit synthesis pathways in CornCyc, or the construction of a hemicellulose term in the biomass equation in terms of the overall composition of hemicellulose without reference to specific synthesis reactions, as in our adaptation of the biomass reaction of iRS1563 (see the biomass reaction discussion, below).

Miscellaneous cell wall components

The following additional cell wall component production reactions from CornCyc were added manually:

- 2.4.1.43-RXN, representing the formation of homogalacturonan from galacturonate
- RXN-9589 (EC 2.4.2.41), representing the addition of xylose to homogalacturonan to form xylogalacturonan (note the resulting xylogalacturonan subunit notionally consists of one galacturonate plus xylose)
- 13-BETA-GLUCAN-SYNTHASE-RXN (EC 2.4.1.12), representing the formation of callose from glucose.

Suberin production is not represented in CornCyc in detail but pathways for the synthesis of three key precursors, N-feruloyltyramine, octadecenedioate, and docosanediotate, are provided. Sinks for N-feruloyltyramine and octadecenedioate were added to the model to represent the flow of material towards suberin production; docosanedioate was neglected because no genes are associated with the reactions of its synthesis pathway. N-feruloyltyramine may be produced from trans-caffeate via either ferulate or caffeoyl-CoA; the branch through ferulate was initially dropped from the reduced version of the model used for data analysis because it relies on trans-feruloyl-CoA synthase, EC 6.2.1.34, which has no associated genes, but it was preserved in subsequent versions of the model because high expression levels for caffeate O-methyltransferase suggest this branch is indeed active.

(In CornCyc, the tyramine N-feruloyltransferase that produces N-feruloyltyramine from feruloyl-CoA could also catalyze the production of other hydroxycinnamic acid tyramine amides (cinnamoyltyramide, sinapoyltyramide, p-coumaroyl-tyramine) but we have neglected these for now.)

Fatty acids and lipids

Plant fatty acid and lipid biosynthesis is rich in complexity (see, e.g., [144]), and attempting to describe it in the FBA model at the level of detail at which it is currently understood would require a daunting number of reactions among the species representing the combinations of lipid head groups and acyl chains. Though CornCyc presents some pathways of lipid metabolism at such a high resolution, we have adopted a simplified approach which aims to include enough detail to allow the model to:

- predict based on RNA-seq data the total flow of biomass into fatty acids and lipids
- coarsely predict differences in the types of lipids and fatty acids produced, based on RNA-seq data
- approximately preserve the iRS1563 biomass equation.

The model describes in detail the sequence of reactions by which fatty acids up to lengths of 16 and 18 are synthesized in the chloroplast (though currently these reactions occur in the cytoplasmic compartment!), and the formation of oleate (as oleoyl-ACP) by the stearoyl-ACP desaturase (`PWY-5156`; [144,145]). In practice, these fatty acids may then enter the 'prokaryotic' pathway of glycerolipid synthesis in the chloroplast or leave the chloroplast and enter the 'eukaryotic' pathway of glycerolipid synthesis in the endoplasmic reticulum, with further desaturation of the acyl chains occurring after their incorporation into lipids.

We simplify this process by effectively decoupling the synthesis of different types of lipids (as distinguished by head groups) from the desaturation

of their associated acyl chains. Reactions from lipid synthesis pathways are implemented as if all lipid species had one 16:0 and one 18:1 acyl chain, by implementing the glycerol-3-phosphate O-acyltransferase and 1-acylglycerol-3-phosphate O-acyltransferase reactions (RXN-10462 and 1-ACYLGLYCEROL-3-P-ACYLTRANSFER-RXN), written in the database with generic acyl-acp substrates, with oleoyl-ACP and palmitoyl-ACP as substrates respectively. (This corresponds to the prokaryotic pathway; in the eukaryotic pathway oleoyl-CoA and palmitoyl-CoA would supply the acyl groups for diacylglycerol formation instead [144]. However the same genes are associated with the reactions of diacylglycerol synthesis in the two pathways (`PWY-5667`; `PWY0-1319`) in CornCyc and so they cannot be distinguished based on expression data alone; we have chosen one arbitrarily.)

This supply of diacylglycerol is sufficient to allow, without further modification to the CornCyc FBA export, the synthesis of a variety of lipids, including:

- phosphatidylcholine, phosphatidylethanolamine, phosphatidylglycerol, phosphatidylinositol;
- sulfoquinovosyldiacylglycerol.

UDP-glucose epimerase is exported from CornCyc in the UDP-glucose-producing direction by default; we allowed it to run in the reverse direction as well, consistent with literature evidence [146, 147], which allowed the production of mono- and digalactosyldiacylglycerol.

In sphingolipid metabolism, dihydrosphingosine, 4-hydroxysphinganine and sphinganine 1-phosphate may be produced, and sink reactions were added for them. Production of the ceramides and their derivatives would require the

choice of a particular fatty acid source for the sphinganine acyltransferase, written by default with the generic substrate 'a long-chain acyl-coA'; per the CornCyc description page for PWY-5129, in leaf sphingolipids C20 to C26 fatty acids are typical. Currently, the FBA model lacks a detailed implementation of production of very long chain fatty acids by elongation (a generic representation is present in CornCyc), so no supply of C20-26 fatty acids is available. We have deferred this issue to future work.

Separately, we model the desaturation of oleate to linoleate and linolenate and palmitate to palmitoleate. These (along with palmitate and stearate) are the fatty acid components of the iRS1563 biomass reaction, which originally incorporated them as triglycerides; our modified biomass equation consumes free fatty acids, rather than attempt to specify the precise ratios in which they are to be found in different lipid species in the leaf.

The CornCyc pathways for linoleate and linolenate produce them as lipid linoleoyl groups and lipid linolenoyl groups respectively, incorporated in generic lipid molecules; to allow these reactions to balance, and to provide linoleate and linolenate for the biomass reaction, we added lipases which release free linoleate/linolenate from the lipid linoleoyl and lipid linolenoyl groups, regenerating the pool of generic 'lipid' species (which participate only in the linoleate pathway, within the FBA model). Note, however, that other reactions within the model but outside the indicated synthesis pathways are capable of producing linoleate and linolenate as well.

CornCyc includes no complete pathway for the production of palmitoleic acid; as there is experimental evidence it is produced in maize leaves (see the discussion of the biomass equation, below) we introduced the acyl-ACP

Δ9-desaturase reaction from the palmitoleate biosynthesis pathway of AraCyc (RXN-8389, 1.14.99.-), producing palmitoleoyl-ACP from palmitoyl-ACP [148], which restores this functionality (in combination with the palmitoleoyl-ACP hydrolase, RXN-9550, which is present in CornCyc). Note that there is some evidence that the stearoyl-ACP desaturase enzyme may also catalyze this reaction [149].

The oleoyl-acyl carrier protein hydrolase (EC 3.1.2.14) from CornCyc is unbalanced with respect to hydrogen; a version with an additional proton on the right hand side was added manually.

The Δ9-desaturase and the desaturases producing linoleate and linolenate (RXN-9667 and RXN-9669) were written originally with generic electron donor and acceptor species. Initial review of the extensive literature on plant fatty acid desaturation suggests that the electron source for desaturases depends on their location within the cell, with chloroplastic desaturases accepting electrons from ferredoxin while desaturases in the endoplasmic reticulum accept electrons from NADH via cytochrome b5 or fused cytochrome domains (see, eg, [150–152]). As discriminating between chloroplastic and extrachloroplastic fatty acid desaturation is not a high priority for the model, NADH was used as the sole electron donor for all three of these reactions.

The ferredoxin-dependent stearoyl-ACP desaturase RXN-7903, not exported from the database by default because it is marked as unbalanced, was added in a form adjusted for hydrogen and charge balance. Ferredoxin-NADP oxidoreductase was made reversible to ensure NADPH can drive this reaction in the dark, as is observed [152].

Nucleic acids polymerization

Reactions representing the pyrophosphate-releasing incorporation of (d)NTPs into RNA and DNA were added and associated with the DNA-directed DNA polymerase and DNA-directed RNA polymerase reactions in the database. (In each case, it is assumed that all nucleotides occur with equal frequency.)

A.3.4 Ascorbate-glutathione cycle

To allow the NADPH-monodehydroascorbate reductase reaction to function in the cycle as curated, we split the L-ascorbate peroxidase reaction (EC 1.11.1.11) into its two subreactions, which by default are not exported in the FBA problem.

A.3.5 Gamma-glutamyl cycle

The gamma-glutamyltransferase was lumped together with `GAMMA-GLUTAMYL-CYCLOTRANSFERASE-RXN`, originally written in terms of the instanceless class '`L-2-AMINO-ACID`' which appeared in no other stoichiometries in the FBA export, and the dipeptidase `RXN-6622`, which is the only reaction that can consume the cysteinylglycine product of the gamma-glutamyltransferase, forming a combined reaction which can carry flux. The combined reaction retained the gene associations of the gamma-glutamyltransferase, as the other two reactions have no associated genes.

A.3.6 Methionine synthesis from homocysteine

The methionine synthase reaction of CornCyc's methionine biosynthesis pathway, HOMOCYSMET-RXN, EC 2.1.1.14, specifically requires 5-methyltetrahydropteryltri-L-glutamate as a cofactor. Polyglutamylation of folates is present in CornCyc in an abstract representation (with tetrahydrofolate synthase catalyzing the addition of a glutamyl group to a 5-methyltetrahydropteryl with n glutamyl groups); we have not converted this into an explicit representation in the FBA model. Instead, HOMOCYSMETB12-RXN, EC 2.1.1.13, acts to produce methionine from homocysteine; the effects of this possible inaccuracy on the behavior of the rest of the network should be limited.

A.3.7 Basic import and export

The following species are given overall import/export reactions:

- WATER
- CARBON-DIOXIDE
- OXYGEN-MOLECULE
- PROTON
- NITRATE
- SULFATE
- |Pi|
- |Light|
- MG+2

These reactions exchange species inside the cell with species in meaningfully labeled compartments where possible (eg, oxygen and CO_2 are exchanged with the intercellular air space, mineral nutrients with the xylem, etc.).

In addition, to facilitate exchange among compartments in the whole-leaf model, a number of exchanges with a phloem compartment were set up: these included sucrose, glycine (as a representative of the amino acids detected in maize phloem sap by Ohshima et al [111]), and the potential phloem sulfur transport compound glutathione [112].

Note that these reactions should be inactive, or restricted to the exporting direction only, when not modeling transport within the leaf (except for sucrose, where a free supply should be allowed in heterotrophic conditions).

A.3.8 Defining the biomass components

Two types of biomass reactions are added to the model:

- Sinks for individual species, for simulations (e.g, fits to RNAseq data) where the relative rates of production of different components are unknown. The species given such sinks are listed in `biomass_components.txt`.
- A set of reactions producing a combined biomass species, made up of assorted components in fixed proportions, for simulations where the maximum rate of production of biomass is of interest, and an approximately realistic biomass composition needs to be enforced directly. These reactions were taken with minor modifications from [41]; their adaptation is described below and they are listed in are listed in `adapted_irs1563_biomass.txt`.

To conceptually and practically separate these types of biomass reactions, which in general should not both be active in any one calculation, the biomass species they produce are located within two separate abstract biomass compartments in the SBML model.

In general, the biomass sink reactions have no gene associations, but an exception was made for the twenty reactions representing incorporation of amino acids into protein, which inherit the gene associations of the corresponding tRNA ligase reactions in CornCyc. (In principle these could be distinguished from sink reactions representing the expansion of free amino acid pools as cells grow and divide, but we have ignored this issue for now.)

Note that, to support the adapted iRS1563 biomass equation, a reaction representing the production of free galactose from GDP-L-galactose was introduced (otherwise, release of galactose from UDP-galactose was catalyzed by two reactions in the pathways of indole-3-acetyl-ester conjugate biosynthesis and indole-3-acetate activation, likely not a major route for carbohydrate production). Free galactose is not included in the individual biomass species used for data fitting.

A.4 Compartmentalization

Approaches differ to the subcellular compartmentalization in FBA models of eukaryotes, ranging from the assignment of compartments to a few key pathways known to function primarily outside the cytosol, as in the mitochondrial and chloroplastic "modules" of AraMeta [35] and RiceMeta [45] to the extremely comprehensive, data-driven approach of [44]. Here, we did not attempt to com-

prehensively assign reactions to their proper compartments; instead, we started with a modular approach similar to [45] in which some core metabolic pathways were compartmentalized (in our case, the TCA cycle and mitochondrial electron transport chain in the mitochondrion, the light reactions of photosynthesis, Calvin cycle, and some reactions of the C4 and photorespiratory pathways in the chlorophyll, and some reactions of the photorespiratory pathway in the peroxisome, with transport reactions added as necessary).

We then refined the compartment assignments of other reactions and pathways as needed to permit key metabolic functions and compartmentalize a limited number of additional reactions whose incorrect assignment to the cytosol we judged particularly likely to lead to misleading results.

More details on individual compartmentalization choices and transport reactions are given below.

A.4.1 Intracellular transport

Sources (beyond those detailed below) informing the addition of intracellular transport reactions in the model included the transport reactions present in AraMeta [35], reviews of photorespiratory metabolism with attention to compartmentalization [153, 154], a review of chloroplast transporters [155], and a review of transport processes in C4 photosynthesis [156].

In most cases we have not tried to reflect the mechanisms of the transport systems, where those are known, in any detail (exceptions include the triose phosphate-phosphate and PEP-phosphate transporters across the chloroplast

envelope), nor have we associated genes with the transporters, even when they are known. Future work should pay greater attention to this aspect of the system.

A.4.2 Photorespiratory pathway

Following [2] we assumed that reducing power was supplied to the peroxisome through an oxaloacetate-malate shuttle and NAD(H)-dependent malate dehydrogenase, and added an oxaloacetate-malate antiporter and a copy of `MALATE-DEH-RXN` to the peroxisome. Reactions of the pathway were localized following [153] and [154]. Note that glycine decarboxylase was assigned exclusively to the mitochondrion, while serine hydroxymethyltransferase was present in both the mitochondrion and the cytoplasm, where it plays a role in one-carbon metabolism [157].

A.4.3 Various ferredoxin-consuming pathways

The model includes several pathways or reactions (e.g., sulfite and nitrite reduction and the chlorophyll cycle) which rely on ferredoxins for reducing power, and are localized to the chloroplast, where, in the light, reduced ferredoxins may be supplied by the photosynthetic electron transport chain.

Rather than assign the reactions of these pathways to compartments appropriately, we added a reaction exchanging reduced ferredoxins and oxidized ferredoxins across the chloroplast boundary to supply ferredoxin-driven pathways in the cytosol. We emphasize that this is a convenient simplification and

is not intended to represent a realistic mechanism.

A.4.4 Ascorbate production

The L-galactonolactone dehydrogenase responsible for the final step of the ascorbate production pathway in CornCyc reduces cytochrome C and has been experimentally localized to the mitochondrial inner membrane, with its catalytic site facing outwards, into the intermembrane space [158]. As the outer membrane is generally permeable to small molecules we have treated this reaction as acting directly on cytoplasmic galactonolactone and ascorbate. A sink for ascorbate as a biomass component was added, as it is found in substantial quantities in leaves (see, e.g., [159,160]).

A.4.5 Ascorbate-glutathione cycle

This cycle is present in multiple cellular compartments; in the model we included only cytosolic and chloroplastic instances (of which only the chloroplastic was ultimately expected to be relevant, as there was no supply of superoxides in the cytosol). Note that none of the genes associated with monodehydroascorbate reductase could be assigned to the chloroplast under the rules described below: two had curated location in the peroxisome while `GRMZM2G320307` had no curated location and TargetP prediction of mitochondrial (`GRMZM2G320307_P01`) and cytoplasmic (`GRMZM2G320307_P02`, `GRMZM2G320307_P03`) locations. Reduction of monodehydroascorbate may also proceed non-enzymatically (see above) so this (enzymatic) reaction was

removed from the chloroplast in favor of direct reduction by ferredoxin.

A.5 Gene associations for compartmentalized reactions

Where a reaction was present in more than one compartment– that is, when two or more reactions in different compartments were associated with the same reaction record in CornCyc– we examined the genes associated with those reactions in CornCyc and assigned them to the instance of the reaction in the most appropriate compartment, as far as possible.

Where the Plant Proteome Database [58] provided manually curated location assignments for genes, those were used; otherwise, we used automatic location predictions by TargetP [161] or in some cases referred to the gene's annotation (both also provided by PPDB). In general we assumed the appropriate location for a gene product was the cytoplasmic compartment absent a specific prediction of localization in the chloroplast, mitochondrion, or peroxisome. Where proteins were predicted to occur in a compartment where an no instance of a particular reaction was present, those gene associations were generally dropped from the model.

When a gene was associated with a reaction in more than one compartment and also a reaction present in only one compartment, in general the association with the reaction in only one compartment was dropped, except for reactions which we believed based on literature evidence (including comments in CornCyc and PPDB) were assigned to the cytoplasmic compartment only because our compartmentalization process was incomplete.

Some details on the judgment calls made in this process are provided in the comments to the file `gra_overrides.txt`; we comment here on a few unusual cases.

A.5.1 NADH dehydrogenases

Cyclic electron transport around Photosystem I may occur through the chloroplast NADH dehydrogenase complex or an alternate pathway which in Arabidopsis involves PGR5 [162, 163]. In C3 plants the PGR5-dependent pathway may play the major role in tuning the photosynthetic ATP/NADPH ratio, while the NADH dehydrogenase pathway is implicated in stress responses [163]. In contrast, in C4 plants the expression of the chloroplast NADH-dehydrogenase appears to correlate with photosynthetic ATP demand, while PGR5 expression does not, suggesting it is the NADH-dehydrogenase CET pathway which allows increased the increased ATP production required by the C4 system [164]. Thus, genes associated in CornCyc with the NADH dehydrogenase reaction for which a chloroplast location was predicted were reassociated with the model's cyclic electron transport reaction (despite the fact that our somewhat abstract cyclic electron transport reaction may not accurately represent the biochemistry of the NADH-dependent pathway).

A.5.2 Pyruvate dehydrogenases

In practice, pyruvate dehydrogenase complexes are found in the mitochondrion and chloroplast, but here we have not fully compartmentalized the chloroplas-

tic pyruvate dehydrogenase and the pathways it supplies, instead leaving it in the cytosol. Thus, genes associated with the reactions of the complex with predicted chloroplast localization were associated instead with the cytosolic version. Genes with no curated or predicted location were left associated with both forms (splitting their expression data between them, in the fitting process).

A.6 Testing and consistency checking

The compartmentalized single-cell model was checked in detail for conservation violations by testing the feasibility of net production or consumption of a unit of each internal species with all external transport and biomass sink reactions suppressed.

Where such production was found feasible, the reactions involved were carefully inspected and stoichiometry coefficients adjusted to restore balance if necessary. In practice, this led only to the correction of erroneous reactions added by hand; as expected, no balance issues were found with reactions exported from CornCyc.

In the final version, no such unrealistic processes are possible in the model under normal conditions. (Note that the species representing light input may be consumed in isolation, but the use of light energy to drive a futile cycle is not unrealistic, though we have not examined the details of the process found by the consistency checker in any detail.) Of course, demonstrating that no such production/consumption is feasible does not guarantee that all reactions in the model are properly balanced.

Testing also verified that all individual biomass sink reactions, and the combined biomass reaction, could proceed at nonzero rates.

A.7 SBML export

A.7.1 Component names

SBML distinguishes a component's name from its ID. Reactions and species in the SBML model were given name attributes according to the by calling the Pathway Tools `get_name_string` function on the frames in the database from which they derive, if any. The IDs of the SBML components were derived from the frame handles, replacing special characters with underscores as necessary to conform to the SBML sID standard.

Note that for some reactions in CornCyc, the result of `get_name_string` is an EC number different from the EC number indicated by the label of the frame (e.g, `2.7.1.133-RXN`, for which 'EC 2.7.1.159' is returned). The frame in CornCyc (if any) from which each reaction in the SBML model is ultimately derived is preserved as a comment in the reaction's Notes element, to resolve any ambiguity.

A.7.2 Gene annotations

Each reaction in the FBA model associated with a particular parent frame in CornCyc was given an association rule that combined all genes associated with

that reaction in CornCyc, as well as all genes associated with all generic reactions of which the parent reaction is a specific form, in a logical 'or' relationship, stored in the reaction's Notes element per the COBRA standard.

A.8 Model refinement

A.8.1 Phosphoribulokinase

In early attempts to fit the model to the leaf gradient data, high costs were associated with the mesophyll phosphoribulokinase reaction in the source tissue when the bundle sheath CO_2 level was high. We noted that in CornCyc 4.0 several genes were associated with both PRK and glyceraldehyde-3-phosphate dehydrogenase. To clarify the role of these genes we referred to annotations in the Plant Proteome Database [58] and best hits in the Conserved Domain Database [165] (accessed through NCBI). Of the eight genes associated with PRK in CornCyc, three (`GRMZM2G039723`, `GRMZM2G337113`, `GRMZM2G162845`) appeared to encode GAPDH enzymes (per PPDB annotations and the presence of `Gp_dh_N` and `Gp_dh_C` domains), three (`GRMZM2G162529`, `GRMZM2G463280`, `GRMZM2G026024`) appeared encode to encode genuine phosphoribulokinases (per PPDB annotations and the presence of PRK domains), and two appeared to encode CP12-type regulatory proteins, with no obvious evidence for any individual protein sharing more than one of these roles. The regulatory role of CP12 does involve forming a complex with PRK and GAPDH, but this reduces, rather than enhancing or enabling, their individual activities [166]. We removed the PRK associations of the GAPDH and CP12 genes from our model. PPDB as-

signed these three GAPDH genes to a plastidic location based on experimental evidence, so we associated them with those reactions exclusively (removing associations with the cytosolic instances of EC 1.2.1.13 and/or EC 1.2.1.12).

A.9 Biomass equation

We developed a biomass equation following that used in [41]. Our calculations are based on supplementary file S4 of that paper[1], in particular sheet 2, 'Biomass_rxn'.

That sheet derives a biomass equation corresponding to the production of one gram of plant dry weight, based on literature data on biomass composition; the description is divided into subreactions forming (e.g.) 'nitrogenous compounds', 'lignin', etc., which then participate in an overall biomass reaction. The units of the stoichiometric coefficients are mmol.

We have adopted most of the biomass composition assumptions of Saha et al wholesale, with gratitude for their efforts in compiling this data from the literature. However, we have made some minor adjustments, resulting in a different overall stoichiometry for biomass production.

A.9.1 Fatty acids

Saha et al represent the total lipid/fatty acid contribution to biomass as a pool of triglycerides in proportions apparently based on a maize oil measurement and

[1]Specifically, `journal.pone.0021784.s004.xls`, as downloaded from the PLoS One web site 20 November 2013

thus probably reflective of seed triglyceride composition.

We substitute measurements of the fatty acid content of mature maize leaf membrane lipids [167] and write a biomass sub-reaction which consumes the relevant free fatty acids (rather than their derivatives in the form of triacylglycerols, membrane lipids, etc.,) as shown in Table A.1.

Fatty acid	CornCyc compound	mol. wt. (g/mol)	mole fraction
palmitic	PALMITATE	255.42	0.104
palmitoleic	CPD-9245	253.4	0.056
stearic	STEARIC_ACID	283.47	0.011
oleic	OLEATE_CPD	281.46	0.044
linoleic	LINOLEIC_ACID	279.44	0.132
linolenic	LINOLENIC_ACID	277.43	0.646

Table A.1: Fatty acid proportions in biomass.

Weighting the molecular weights by the mole fractions, we find one mole of fatty acid in appropriate proportions weighs 272.4 g. Dividing the mole fractions by the overall molar weight and multiplying coefficients by 1000 to convert to millimoles, we arrive at the final equation:

$$0.382\ \texttt{PALMITATE} + 0.206\ \texttt{CPD-9245} + 0.04\ \texttt{STEARIC\_ACID} + 0.162\ \texttt{OLEATE\_CPD} + 0.485\ \texttt{LINOLEIC\_ACID} + 2.372\ \texttt{LINOLENIC\_ACID} = \texttt{fatty\_acids\_biomass}$$

where the left-hand side represents 1 g.

Fractions add to less than 1.0 because we ignore trace (mole fraction ≤ 0.01) amounts of C14:0 and C20:0 fatty acids. Note that the leaf fatty acid composition is known to change along the developmental gradient, so specifying any single composition is an approximation; see [168].

A.9.2 Hemicellulose

We adopted the hemicellulose production reaction as is, using the species added to the model for this purpose, 'polysaccharide_[sugar]_unit'. The resulting equation is:

0.548 polysaccharide_arabinose_unit +
1.248 polysaccharide_xylose_unit +
0.301 polysaccharide_mannose_unit +
0.144 polysaccharide_galactose_unit +
3.254 polysaccharide_glucose_unit +
0.166 polysaccharide_galacturonate_unit +
0.166 polysaccharide_glucuronate_unit
= hemicellulose_biomass.

A.9.3 Total carbohydrates

We recalculated the stoichiometries of the carbohydrate-producing reaction to account for the differing molecular weight of our representation of cellulose ('CELLULOSE_monomer_equivalent', effectively a glucose molecule), account for the fact that one unit of hemicellulose represents one gram, not one (milli)mole, and express pectin in terms of polysaccharide_galacturonate_unit, reflecting a belief that UDP is released in the formation of pectin from UDP-D-galacturonate, rather than retained in the polymer [169].

It is not clear what form the 'mannose' referred to by Penning de Vries et

al should be assumed to take, as free mannose is not found in plants under most circumstances (see, e.g., [170–172]). Here we somewhat arbitrarily choose mannose-6-phopshate.

Table A.2 shows the calculation, resulting in the equation:

0.259 polysaccharide_galacturonate_unit + 0.067 RIBOSE + 0.278 GLC + 0.111 FRU + 0.039 MANNOSE-6P + 0.056 GALACTOSE + 0.146 SUCROSE + 2.220 CELLULOSE_monomer_equivalent + 0.400 hemicellulose_biomass = carbohydrates_biomass.

A.9.4 Organic acids

We adopt this reaction as is. In the terminology of our model, the resulting equation is:

0.556 OXALATE + 0.676 GLYOX + 1.515 OXALACETIC_ACID + 0.746 MAL + 1.562 CIT + 1.724 CIS-ACONITATE = organic_acids_biomass.

A.9.5 Protein and free amino acids

We adopt these reactions as is. In the terminology of our model, the resulting equations are:

1.15 L-ALPHA-ALANINE + 0.0959 ARG + 0.414 L-ASPARTATE + 0.0313 CYS + 1.53 GLT + 0.0445 GLY + 0.0915 HIS + 0.465 ILE +

1.51 LEU + 5.71e-05 LYS + 0.123 MET + 0.314 PHE + 0.762 PRO + 0.612 SER + 0.175 THR + 0.00409 TRP + 0.244 TYR + 0.25 VAL = protein_biomass

and

0.624 L-ALPHA-ALANINE + 0.319 ARG + 0.418 L-ASPARTATE + 0.231 CYS + 0.378 GLT + 0.740 GLY + 0.358 HIS + 0.424 ILE + 0.424 LEU + 0.380 LYS + 0.373 MET + 0.337 PHE + 0.483 PRO + 0.529 SER + 0.467 THR + 0.272 TRP + 0.307 TYR + 0.475 VAL = free_aa_biomass.

A.9.6 Lignin

We adopt this reaction as is. In the terminology of our model, the resulting equation is:

2.221 COUMARYL-ALCOHOL + 1.851 CONIFERYL-ALCOHOL + 1.587 SINAPYL-ALCOHOL = lignin_biomass.

A.9.7 Nucleic acids

We adopt this reaction as is (though note that, as discussed above, nucleotide triphosphates are not necessarily the appropriate best representation for polymerized nucleic acids). In the terminology of our model, the resulting equation is:

0.247 ATP + 0.239 GTP + 0.259 CTP + 0.258 UTP + 0.255 DATP + 0.247 DGTP + 0.268 DCTP + 0.259 TTP = nucleic_acids_biomass.

A.9.8 Nitrogenous compounds

We use the same nitrogenous compound weight fraction breakdown, but recalculate the stoichiometric coefficients accounting for the fact that the protein biomass, free amino acid biomass, and nucleotide biomass species each represent one gram, so that the appropriate stoichiometric coefficients of those species for the production of one total gram of nitrogenous compounds are simply the weight fractions; see Table A.3.

The resulting equation is

0.100 free_aa_biomass + 0.870 protein_biomass + 0.030 nucleic_acids_biomass = nitrogenous_biomass.

A.9.9 Inorganic materials

We ignore these entirely, as they play no other role in the model. (Note that even in iRS1563 the two species involved, potassium and chloride, participate only in source and sink reactions.)

A.9.10 Total biomass reaction

We drop the inorganic materials term (note that weight fractions now add to 0.95) and recalculate the stoichiometric coefficients, accounting for the fact that the component biomass subspecies each represent one gram; see Table A.4.

The final equation is

0.230 nitrogenous_biomass + 0.565 carbohydrates_biomass + 0.025 fatty_acids_biomass + 0.080 lignin_biomass + 0.050 organic_acids_biomass = total_biomass.

Saha et al additionally incorporate an ATP cost in their overall biomass reaction, based on that used in an earlier Arabidopsis model (AraGEM [39]) Combining this ATP hydrolysis with a sink of total biomass, we arrive at the overall equation for biomass production and growth (CombinedBiomassReaction):

1.0 total_biomass + 30.0 ATP + 30.0 WATER = 30.0 ADP + 30.0 Pi + 30.0 PROTON

A.9.11 Protonation

Throughout, note that the molecular weights of species in our model may differ somewhat from those used in the iRS1563 table because of differing assumptions about protonation. The practical consequences of this difference should be limited.

A.9.12 Oxalate

Early drafts of the model could not produce oxalate. CornCyc indicates its production as resulting only from ascorbic acid catabolism with concomitant production of L-threonate. Recent reviews suggest this is the primary pathway of oxalate production in plant species which form calcium oxalate crystals, with the threonate ultimately being oxidized to tartrate [173–175], though the pathways of production of soluble oxalate are less clear [174]. We found little immediate evidence that tartrate (or threonate) is formed in maize leaves at levels comparable to that of oxalate, or of pathways which could further metabolize the tartrate.

Of the three reactions in iRS1563 which could produce oxalate, only one has an associated gene: oxalate carboxylase (oxalate = formate + CO2); KEGG R00522 (EC4.1.1.2). The gene, 'ACG37538', may correspond to `GRMZM2G103512`, whose best Arabidopsis hit is `AT1G09560.1` (germin-like protein 5); it may thus be more likely to be an oxalate-consuming oxidase [176] than an oxalate carboxylase, though no function was computationally predicted for `GRMZM2G103512` in CornCyc.

We decided the available information did not allow us to accurately model oxalate production in maize. However, to retain the iRS1563 biomass equation and ensure that mass and elemental balance was preserved, we allowed production of oxalate from oxaloacetate by oxaloacetase (EC 3.7.1.1; PlantCyc `OXALOACETASE-RXN`, [177]). This simple reaction has been observed in fungi [178] but is considered unlikely to be widespread in plants [174].

A.10 Plasmodesmatal transport reactions

Species allowed to be exchanged between cell types through the plasmodesmata included:

- carbon dioxide and oxygen;
- known C4 cycle metabolites alanine, aspartate, malate, PEP, and pyruvate;
- the Calvin cycle intermediates glyceraldehyde 3-phosphate and 3-phosphoglycerate;
- photorespiratory metabolites glycerate, glycolate, serine, and glycine;
- nutrients sucrose, phosphate, nitrate, ammonia, sulfate and magnesium;
- glutamate and 2-ketoglutarate;
- and cysteine and glutathione [179].

The inclusion of compounds involved in NAD-ME C4 or C3-C4 intermediate photorespiratory carbon concentrating mechanism is not meant to suggest such a system is necessarily active in maize but merely reflects our knowledge that significant transport of those species between mesophyll and bundle sheath can occur under at least some circumstances.

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Ribose	RIBOSE	150.053	0.010	0.067
Glucose	GLC	180.063	0.050	0.278
Fructose	FRU	180.063	0.020	0.111
Mannose	MANNOSE-6P	258.120	0.010	0.039
Galactose	GALACTOSE	180.063	0.010	0.056
Sucrose	SUCROS	342.116	0.050	0.146
Cellulose	CELLULOSE_monomer_equivalent	180.160	0.400	2.220
Hemicellulose	hemicellulose_biomass	1000.000	0.400	0.400
Pectin	polysaccharide_galacturonate_unit	193.130	0.050	0.259

Table A.2: Carbohydrate species in biomass.

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Amino acids	free_aa_biomass	1000.000	0.100	0.100
Proteins	protein_biomass	1000.000	0.870	0.870
Nucleic acids	nucleic_acids_biomass	1000.000	0.030	0.030

Table A.3: Nitrogenous biomass breakdown.

Component	Species in model	unit wt (mg)	wt fraction	units/g product
Nitrogenous compounds	nitrogenous_biomass	1000.000	0.230	0.230
Carbohydrates	carbohydrates_biomass	1000.000	0.565	0.565
Lipids	fatty_acids_biomass	1000.000	0.025	0.025
Lignin	lignin_biomass	1000.000	0.080	0.080
Organic acids	organic_acids_biomass	1000.000	0.050	0.050

Table A.4: Breakdown of total biomass.

APPENDIX B

SUPPLEMENTARY TABLES

B.1 Overrepresented pathways in the k-means clusters of Fig. 3.6.

A list of metabolic pathways from CornCyc [57] represented in the model was compiled and the set of protein level parameters for the reactions associated with each pathway – which may not represent the complete pathway – was tested for overrepresentation in each of the clusters using a one-sided Fisher's exact test. Listed below for each cluster are all those pathways for which the resulting p-value was less than 0.01, and their p-values.

This approach is intended as an informal, heuristic method for identifying the predominant metabolic activities associated with each cluster and not a formal statistical test of the null hypothesis that cluster assignment is independent of pathway assignment, which would be inappropriate here (among other issues, in many cases the FBA steady-state assumption implies that the fluxes of two or more reactions within a pathway will always be equal – or related by a constant factor – thus perfectly correlated, so that their cluster assignments are not independent). We have also ignored the issue of multiple testing.

Note that reactions are frequently assigned to more than one pathway, and a pathway may be detected in a cluster even if its component reactions are actually fulfilling a different metabolic function in the model than that suggested by the name of the pathway.

Table B.1: Cluster a (140 parameters)

8.5993e-21	mesophyll palmitate biosynthesis II (bacteria and plants)
4.3671e-06	mesophyll suberin biosynthesis
2.5938e-05	mesophyll pyruvate decarboxylation to acetyl CoA
0.00015309	mesophyll chorismate biosynthesis from 3-dehydroquinate
0.00015309	mesophyll isoleucine biosynthesis I (from threonine)
0.00023188	mesophyll phenylpropanoid biosynthesis
0.00089798	mesophyll ascorbate biosynthesis I (L-galactose pathway)
0.00089798	mesophyll photosynthesis light reactions
0.00089798	mesophyll valine biosynthesis
0.0038775	mesophyll proline biosynthesis I
0.0038775	mesophyll stearate biosynthesis II (bacteria and plants)
0.0052349	mesophyll fatty acid biosynthesis initiation I
0.0052349	mesophyll linoleate biosynthesis I (plants)
0.0052349	mesophyll phenylalanine biosynthesis II
0.0052349	mesophyll tyrosine biosynthesis II

Table B.2: Cluster b (134 parameters)

3.5818e-21	bundle sheath palmitate biosynthesis II (bacteria and plants)
3.5378e-06	bundle sheath suberin biosynthesis
2.1669e-05	bundle sheath pyruvate decarboxylation to acetyl CoA
0.00013186	bundle sheath chorismate biosynthesis from 3-dehydroquinate
0.00013186	bundle sheath isoleucine biosynthesis I (from threonine)
0.00018228	bundle sheath phenylpropanoid biosynthesis
0.0034588	bundle sheath stearate biosynthesis II (bacteria and plants)
0.0047896	bundle sheath fatty acid biosynthesis initiation I
0.0047896	bundle sheath linoleate biosynthesis I (plants)
0.0047896	bundle sheath phenylalanine biosynthesis II
0.0047896	bundle sheath sulfate reduction II (assimilatory)

0.0092497	bundle sheath pentose phosphate pathway (non-oxidative branch)

Table B.3: Cluster c (47 parameters)

7.9344e-14	bundle sheath Calvin-Benson-Bassham cycle
5.2648e-11	bundle sheath sucrose biosynthesis
6.9097e-09	bundle sheath glycolysis I
3.6089e-08	bundle sheath gluconeogenesis I
8.882e-08	bundle sheath glycolysis II (from fructose-6P)
2.2103e-06	bundle sheath glycolysis IV (plant cytosol)
1.1074e-05	bundle sheath photosynthesis light reactions
0.00019808	bundle sheath starch biosynthesis
0.00033894	bundle sheath C4 photosynthetic carbon assimilation cycle, PEPCK type
0.0006483	bundle sheath sucrose degradation III
0.00075909	bundle sheath aerobic respiration (cytochrome c)
0.001116	bundle sheath pentose phosphate pathway (non-oxidative branch)
0.0034686	bundle sheath GDP-glucose biosynthesis
0.0034686	bundle sheath glutathione biosynthesis
0.0034686	mesophyll CO2 fixation into oxaloacetate (anapleurotic)
0.0089615	bundle sheath xylose degradation IV

Table B.4: Cluster d (38 parameters)

5.3157e-07	mesophyll photorespiration
9.4917e-05	bundle sheath serine biosynthesis
9.4917e-05	mesophyll nitrate reduction II (assimilatory)
0.0002504	bundle sheath sucrose degradation III
0.00036736	mesophyll aerobic respiration (cytochrome c)
0.00088869	bundle sheath sucrose degradation VI (anaerobic)
0.002137	bundle sheath CO2 fixation into oxaloacetate (anapleurotic)
0.002137	mesophyll ammonia assimilation cycle II
0.0049581	mesophyll gluconeogenesis I

Table B.5: Cluster e (30 parameters)

3.9157e-17	mesophyll Calvin-Benson-Bassham cycle
3.2487e-13	mesophyll sucrose biosynthesis
1.5688e-07	mesophyll glycolysis I
2.5527e-06	mesophyll glycolysis II (from fructose-6P)
4.9596e-05	mesophyll starch biosynthesis
5.84e-05	mesophyll glycolysis IV (plant cytosol)
0.00010699	mesophyll sucrose degradation III
0.0001828	mesophyll gluconeogenesis I
0.00018749	mesophyll pentose phosphate pathway (non-oxidative branch)
0.0013958	bundle sheath ammonia assimilation cycle II
0.0013958	mesophyll GDP-glucose biosynthesis
0.003564	bundle sheath glutamine biosynthesis III
0.0040881	mesophyll starch degradation I

Table B.6: Cluster f (13 parameters)

4.2522e-10	bundle sheath photorespiration
0.0084395	bundle sheath folate transformations II

Table B.7: Cluster g (11 parameters)

0.0047199	bundle sheath xylose degradation IV

Table B.8: Cluster h (9 parameters)

0.00011551	bundle sheath serine racemization
0.00011551	mesophyll glutathione biosynthesis
0.00011551	mesophyll serine racemization

APPENDIX C

AN ALTERNATIVE PHOTORESPIRATORY PATHWAY

The constraint-based modeling approach presented in subsection 1.3.3 can be used to describe hypothetical metabolic networks as well as naturally occurring ones. Combining all the enzymatically catalyzed reactions described in many different organisms into an omnibus metabolic network model, then finding flux predictions which satisfy certain requirements while using a relatively tractable total number of reactions, allows modeling-assisted design of synthetic metabolic systems which could then in principal be implemented by transforming genes encoding the necessary enzymes into a single organism. For example, Bar-Even and coauthors [180] used a model comprising all metabolic reactions (≈ 5000) from any organism in the KEGG LIGAND database [74] to search for possible CO_2 pathways that could offer alternatives to the Calvin cycle.

We adapted the approach of Bar-Even et al to study alternatives to the process of photorespiration. In the pioneering work of Kebeish et al [181], five bacterial enzymes (from glycolate catabolism in *E. coli*) were expressed in the *Arabidopsis* chloroplast, allowing glycolate to be converted to glycerate and returned to the Calvin cycle entirely within the chloroplast, bypassing the ordinary process of photorespiration (which spans the chloroplast, peroxisome and mitochondrion). The resulting plants showed increased growth rates and soluble sugar contents. This result suggested that other transgenic pathways for glycolate recycling could also have technological applications.

In particular, although part of the growth advantage seen in the transgenic plants was attributed to an increase in CO_2 concentration in the chloroplast (because the engineered pathway releases CO_2 there, through glyoxylate carboxyli-

gase, rather than glycine decarboxylase in the mesophyll), we speculated that it might be possible to improve growth rates still further by designing a synthetic pathway which recycled glycolate to a Calvin cycle intermediate without any release of CO_2 at all.

To seek the shortest such pathway, we formed a large metabolic model from the reactions in KEGG release 54.1, allowing import of carbon, oxygen, and various cofactors, and export of triose phosphates. We then inactivated all CO_2-releasing reactions, constrained the flux through the Rubisco oxygenase reaction to a nonzero value, and minimized total flux through all reactions outside the Calvin cycle. (Although minimizing total flux is mathematically distinct from minimizing the total number of reactions used, it is much more computationally straightforward and we found it to be a reasonable proxy for simplicity in practice.)

The solution, shown in Fig. C, involved 19 reactions not normally involved in the Calvin cycle or photorespiration. Feasible pathway design, as discussed in [180], involves consideration of thermodynamics and kinetics as well as reaction stoichiometry; we did not investigate those issues for this pathway because we anticipated that the practical challenges associated with transforming the necessary genes for the 19 enzymes from multiple different species into plants, and ensuring that the enzymes were functional, would be prohibitive. Even if this could be done, and the pathway was thermodynamically and kinetically feasible, the nitrogen costs associated with expressing the enzymes at high enough levels to catalyze the reactions at the high rate usually associated with photorespiration might well outweigh any other advantages to the plant.

We conclude that in further efforts to design or optimize photorespiratory bypasses of manageable complexity, it must be accepted that some CO_2 release is inevitable.

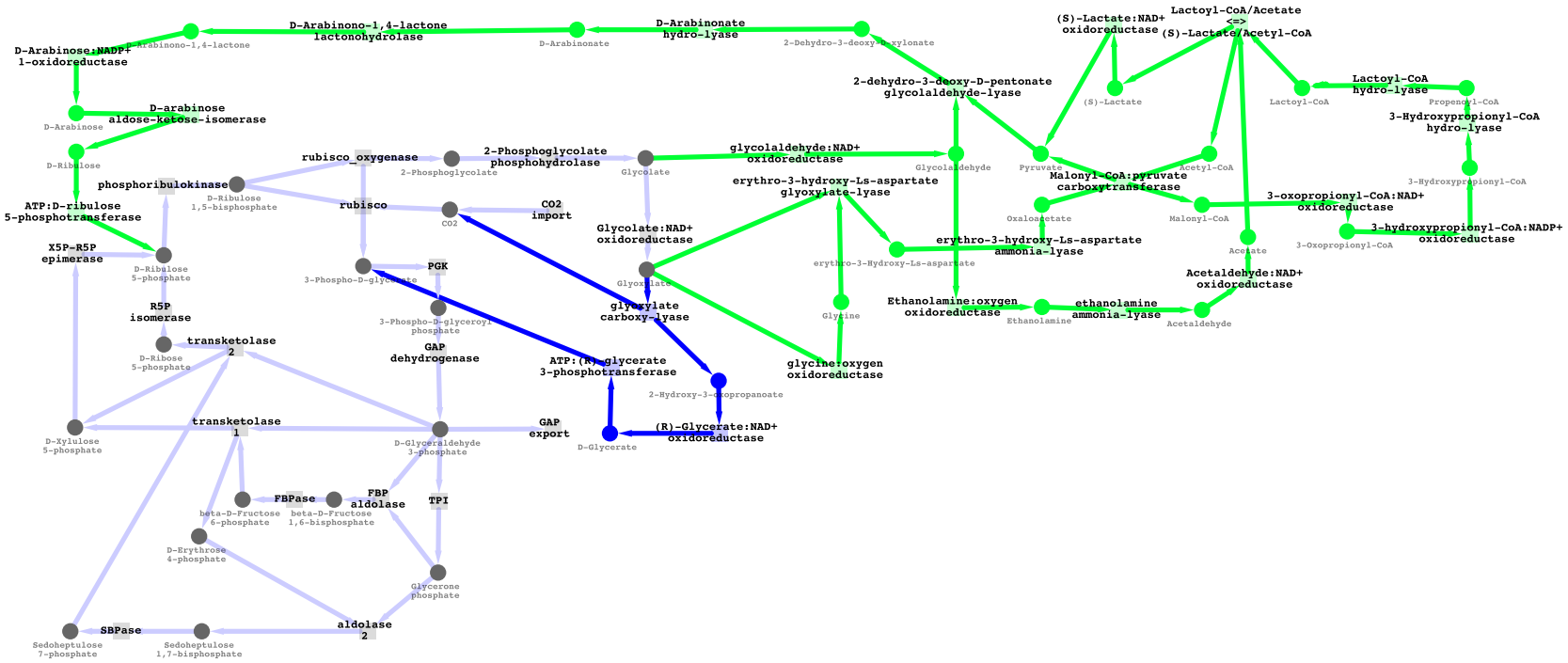

Figure C.1: **Hypothetical photorespiratory bypass system without CO_2 loss.** In the graph, circles represent chemical species and squares represent reactions, with edges between them indicating whether a reaction consumes or produces a metabolite. Grey components are normally present in the Calvin cycle or the usual system of photorespiration. Components in blue belong to the photorespiratory bypass of Kebeish et al. [181]. Components in green belong to the computationally determined minimal system for recycling glycolate to any Calvin cycle intermediate, assembled from all known enzymes in the KEGG database [74] as of June 2010, as discussed in the text.

APPENDIX D

THEORETICAL AND PRACTICAL CONSIDERATIONS IN SOLVING NONLINEAR FLUX BALANCE ANALYSIS PROBLEMS WITH IPOPT

D.1 The Karush-Kuhn-Tucker conditions

Any general nonlinear programming problem with n_g equality and n_h inequality constraints may be placed in the following form:

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^N} f(\mathbf{x}) \\ &\text{s.t.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ &\qquad \mathbf{h}(\mathbf{x}) \geq \mathbf{0}. \end{aligned} \tag{D.1}$$

By introducing a new 'slack' variable s_i for each inequality constraint function h_i and requiring $s_i = h_i$, $i = 1, \ldots, n_h$, we may further reformulate any such problem as

$$\begin{aligned} &\min_{\mathbf{x}\in\mathbb{R}^{N'}} f(\mathbf{x}) \\ &\text{s.t.} \quad \mathbf{g}(\mathbf{x}) = \mathbf{0} \\ &\qquad x_i \geq 0, \qquad i = 1, \ldots, n_h \end{aligned} \tag{D.2}$$

Any solution $\mathbf{x}^*$ to (D.2) (strictly speaking, any solution which meets certain regularity conditions, discussed below) must satisfy the Karush-Kuhn-Tucker

equations [182, 183],

$$\begin{aligned} \nabla \mathbf{g}(\mathbf{x}^*) \cdot \lambda + \mathbf{z} &= \nabla f(\mathbf{x}^*) \\ \mathbf{g}(\mathbf{x}^*) &= \mathbf{0} \\ x_i^* &\geq 0, \qquad i = 1, \ldots, n_h \\ z_i &\geq 0, \qquad i = 1, \ldots, n_h \\ z_i h_i(\mathbf{x}^*) &= 0, \qquad i = 1, \ldots, n_h \end{aligned} \tag{D.3}$$

for some values of the Lagrange multipliers $\lambda = (\lambda_1, \ldots, \lambda_{n_g})$ and $\mathbf{z} \in \mathbb{R}^{N'} = (z_1, \ldots, z_{n_h}, 0, 0, \ldots, 0)$. Effectively these conditions generalize the ordinary Lagrange multiplier approach to optimization with only equality constraints by formalizing the requirement that the Lagrange multipliers associated with inactive inequality constraints (those where $h_i(\mathbf{x}^*)$ is not zero) are zero, and that inequality constraints can 'push' the solution in only one direction.

(We omit here any discussion of additional second-order necessary conditions for $\mathbf{x}^*$ to be an optimal point in the problem (D.1); the reader may refer to [182, 183] or other textbooks on constrained optimization theory.)

D.2 IPOPT

The IPOPT package [53] obtains numerical solutions to the problem (D.2) by solving a sequence of related problems

$$\begin{aligned} \min_{\mathbf{x} \in \mathbb{R}^{N'}} & f(\mathbf{x}) - \mu \sum_{i=1}^{n_h} \log(x_i) \\ \text{s.t.} \quad & \mathbf{g}(\mathbf{x}) = \mathbf{0} \end{aligned} \tag{D.4}$$

with decreasing values the barrier parameter μ (which acts to keep the variables $x_1, \ldots, x_{n_h}$ nonnegative) until the process converges to an adequate solution to an adequate approximation of the original problem. (We omit a detailed justification of this approach; see the references to [53].)

At each iteration, the primal variables $\mathbf{x}$ and dual variables λ and $\mathbf{z}$ are updated by applying Newton's method to the Karush-Kuhn-Tucker equations for D.4 with the current value of μ. (Strictly, Newton's method is used only to find a search direction in the space of primal and dual variables, along which a line search is then performed, but this is irrelevant to the discussion here – for details on this, the control of μ, generalization to upper as well as lower bounds, and the other niceties which contribute to IPOPT's good performance on a wide range of problems, the reader is referred to [53] and the package's online documentation [184].)

To calculate the Newton's method update, a matrix equation must be solved, with the matrix elements being first and second partial derivatives of the constraint functions g and h and the objective function f (or combinations thereof). The performance of the method is largely determined by how quickly, and how precisely, IPOPT (through the linear solver packages it calls) is able to solve this linear system at each step. Several considerations in the design of models intended to be solved by IPOPT follow immediately:

Sparsity The sparsity of the relevant matrix is determined by the number of variables which participate in each constraint. At large scales the problem generally needs to be fairly sparse to be numerically solvable at all, and the linear solver codes (such as ma97) used by IPOPT are typically intended for the solution of sparse systems. In FBA problems with nonlinear con-

straints, as studied here, typically the nonlinear constraints are outnumbered by the linear constraints arising from the conservation laws, and each reaction rate variable typically participates in only a small number of conservation laws (corresponding to its reactants and products) – that is, the overall structure of the problem is sparse because the stoichiometry matrix S is. (Note that the stoichiometry matrix is the transpose of the Jacobian of the linear conservation constraints.)

However, marginal improvements in sparsity can be achieved by formulating additional constraints (or contributions to the objective function) in the problem appropriately, and in some cases this will improve convergence. This may be seen for example in the design of the data-fitting code, where individual auxiliary variables are introduced to represent the contribution to the objective function of each reaction with associated data at each leaf segment, and the total cost associated with the data set is the sum of the auxiliary variables. This is an alternative to a less-sparse design which would directly constrain the variable representing the data set's overall cost to equal a complex expression involving many thousands of data parameters, uncertainty parameters, and reaction rates. As a bonus, such modular design is also often easier to implement and maintain.

Constraint derivative scaling If the derivatives of the constraints with respect to the variables of the problem – taken at the starting point of an optimization calculation, an optimal point, or any intermediate iterate – span many orders of magnitude, so will the entries of the matrix solved in the Newton's-method calculation, and it will tend to have a high condition number, making it more difficult to solve the system precisely (see any numerical linear algebra text for a more precise discussion of this issue.)

IPOPT has the capability to automatically rescale variables (in a way that is transparent to the user) to decrease large derivatives [53,184], and does so by default. However, this does not occur on an iteration-by-iteration basis; instead it is performed at the beginning of the calculation, evaluating the derivatives at the user-provided starting point (or a random sample of points, an optional behavior we never tested because it relies on the MC19 package, which we did not use). Also, it does not correct derivatives which may be too small.

Thus, the user's choice of problem formulation still is key to maintaining good derivative scaling, and we have found choice of units and variable bounds, in particular, can have a great effect on solver performance.

One example occurs in the implementation of the Rubisco kinetics, PEPC kinetics, and CO_2/O_2 diffusion law constraints, which depend on the CO_2 and O_2 concentrations, expressed as equivalent partial pressures. Internally, these are represented in units equivalent to 1 mbar and 10 mbar respectively; this scaling (which may be adjusted through the parameters `co2_scaling_factor` and `o2_scaling_factor`, in `reduced_model.py` for the data-fitting source code and for the elastic band source code in `setup_better_physiology.py`) led to much faster convergence compared to the initial, naive approach of simply expressing both quantities in microbar. (Some experimentation was done to arrive at the current values but further optimization may be possible.)

Another example occurs in the data-fitting code: the scale factors s_i are explicitly required to lie in the range $(-5, 5)$, not just kept to a reasonable size by the penalty term $\alpha \sum_i s_i^2$ in the objective function, because the derivative of the cost with respect to the rate of a reaction (with data) v_i is

proportional to $\exp(s_i)$, and convergence issues arose when one or more such exponential terms became too large or small.

D.3 Constraint degeneracy

In nonlinear problems which incorporate flux balance analysis constraints, the gradients of the constraint functions will typically form a linearly dependent set everywhere. This degeneracy arises for two reasons:

- A typical stoichiometry matrix has a nontrivial left null space (connected to the existence of "conserved moieties": if an element or chemical group is neither produced nor consumed by any reactions in the system, the stoichiometry of one chemical species containing that element or group in a reaction can always be determined from the stoichiometries of all the other species which contain it; thus, the associated row in S is a linear combination of other rows, and represents a redundant constraint [185].) The rows of S are the gradients of the FBA steady-state constraints.
- Many constraints on reaction reversibility are redundant: for example, if one reaction in an unbranched linear pathway is irreversible, it follows that the others cannot run in reverse either. If an explicit lower bound is set for the other reactions (because, for example, they are also believed to be thermodynamically irreversible under biological conditions) it is straightforward to show the gradients of the "constraint functions" enforcing those bounds (which are equal to the unit vectors along the coordinate axes corresponding to the reaction rates) can be written as linear combinations of the gradients of the conservation constraints (rows in the

stoichiometry matrix) and the unit vector along the axis corresponding to the first reaction. The same holds true in more complicated situations where the feasible signs of one reaction rate are in practice determined by the signs allowed for other reaction rates. This is essentially equivalent to the issue discussed in section 3.2.2 above: there, *upper* bounds were redundant.

Two issues arise when the set of gradients of the equality constraints and active inequality constraints is linearly dependent at a (proposed) optimal point.

The first is theoretical: the most common "constraint qualifications", conditions under which the KKT equations are necessary conditions for a point to solve the nonlinear programming problem, may no longer hold. The simplest such condition, the linearly independent constraint qualification, is (as it sounds) precisely the requirement that the set of constraint gradients be linearly independent. One condition of the weaker Mangasarian-Fromovitz constraint qualification (see section 12.6 of [182]) also requires the gradients of the equality constraints to be linearly independent.

We have largely ignored this issue as in practice we have been often able to find acceptable optimal points which do satisfy the KKT equations even when those qualifications do not hold. (It is possible that other more arcane qualifications exist in the literature which do apply to the sorts of problems we have usually solved. Section 12.6 of [182] notes that it is sufficient for all active constraints to be linear functions; we speculate that a similar result could be derived for the special case where the constraint gradients are linearly independent except for degeneracies among a set of purely linear constraints, but have not explored this.)

The second issue is practical. In several particularly large-scale problems, including the data-fitting calculations above, we initially found that IPOPT converged slowly or not at all, but performed better after we removed as many redundant linear constraints from the problem as possible. When the set of active constraint gradients is linearly dependent at an optimal point $\mathbf{x}^*$ for which the KKT conditions do hold, infinitely many choices of multipliers λ and $\mathbf{z}$, and the multipliers may become very large; it is plausible that this will slow the process of convergence to a single choice λ^* and $\mathbf{z}^*$ (with the impact being limited for small-scale problems), though we have not worked out the details.

To facilitate the solution of such large-scale problems, the `simplification` submodule of the `fluxtools` package provides a method to automatically identify and remove redundant linear constraints and variable bounds, which is applied to the basic two-cell model before setting up the data-fitting and elastic band problems solved above. It is important to note that the resulting simplified problem is equivalent to the original problem, but the simplified problem after changing a constraint or variable bound need not be equivalent to the original problem with the same change applied.

BIBLIOGRAPHY

[1] Roy H, Andrews TJ (2000) Rubisco: Assembly and mechanism. In: Leegood R, Sharkey T, von Caemmerer S, editors, Photosynthesis: Physiology and Metabolism, Boston: Kluwer Academic Publishers.

[2] Douce R, Heldt HW (2000) Photorespiration. In: Leegood R, Sharkey T, von Caemmerer S, editors, Photosynthesis: Physiology and Metabolism, Boston: Kluwer Academic Publishers.

[3] von Caemmerer S, Furbank RT (2003) The C(4) pathway: an efficient CO(2) pump. Photosynthesis Research 77: 191–207.

[4] Kanai R, Edwards GE (1999) The biochemistry of C4 photosynthesis. In: Monson RK, Sage, Rowan F, editors, C4 Plant Biology, San Diego: Academic Press, pp. 49–87.

[5] Brown RH (1999) Agronomic implications of C4 photosynthesis. In: Monson RK, Sage RF, editors, C4 Plant Biology, San Diego: Academic Press, pp. 473–507.

[6] Covshoff S, Hibberd JM (2012) Integrating C4 photosynthesis into C3 crops to increase yield potential. Current Opinion in Biotechnology 23: 209–214.

[7] von Caemmerer S, Quick WP, Furbank RT (2012) The development of C4 rice: current progress and future challenges. Science 336: 1671–1672.

[8] Hibberd J, Covshoff S (2010) The regulation of gene expression required for C4 photosynthesis. Annual Review of Plant Biology 61: 181–207.

[9] Studer AJ, Gandin A, Kolbe AR, Wang L, Cousins AB, et al. (2014) A limited role for carbonic anhydrase in C4 photosynthesis as revealed by a ca1ca2 double mutant in maize. Plant Physiology 165: 608–617.

[10] Furbank RT (2011) Evolution of the C4 photosynthetic mechanism: are there really three C4 acid decarboxylation types? Journal of Experimental Botany 62: 3103 –3108.

[11] Sage RF, Christin PA, Edwards EJ (2011) The C4 plant lineages of planet earth. Journal of Experimental Botany 62: 3155–3169.

[12] Pagani M, Zachos JC, Freeman KH, Tipple B, Bohaty S (2005) Marked decline in atmospheric carbon dioxide concentrations during the Paleogene. Science 309: 600–603.

[13] Christin PA, Osborne CP, Sage RF, Arakaki M, Edwards EJ (2011) C4 eudicots are not younger than C4 monocots. Journal of Experimental Botany 62: 3171–3181.

[14] Sage RF (2004) The evolution of C4 photosynthesis. New Phytologist 161: 341–370.

[15] Monson RK (2003) Gene duplication, neofunctionalization, and the evolution of C4 photosynthesis. International Journal of Plant Sciences 164: S43–S54.

[16] Christin PA, Osborne CP, Chatelet DS, Columbus JT, Besnard G, et al. (2013) Anatomical enablers and the evolution of C4 photosynthesis in grasses. Proceedings of the National Academy of Sciences 110: 1381–1386.

[17] Christin PA, Samaritani E, Petitpierre B, Salamin N, Besnard G (2009) Evolutionary insights on C4 photosynthetic subtypes in grasses from genomics and phylogenetics. Genome Biology and Evolution 1: 221–230.

[18] Christin PA, Petitpierre B, Salamin N, Büchi L, Besnard G (2009) Evolution of C(4) phosphoenolpyruvate carboxykinase in grasses, from genotype to phenotype. Molecular Biology and Evolution 26: 357–365.

[19] Roalson EH (2008) C4 photosynthesis: Differentiating causation and coincidence. Current Biology 18: R167–R168.

[20] Sage RF (2001) Environmental and evolutionary preconditionsfor the origin and diversification of the C4 photosynthetic syndrome. Plant Biology 3: 202–213.

[21] Scheiter S, Higgins SI, Osborne CP, Bradshaw C, Lunt D, et al. (2012) Fire and fire-adapted vegetation promoted C4 expansion in the late Miocene. New Phytologist 195: 653–666.

[22] Osborne CP, Beerling DJ (2006) Nature's green revolution: the remarkable evolutionary rise of C4 plants. Philosophical Transactions of the Royal Society of London B: Biological Sciences 361: 173–194.

[23] Osborne CP, Sack L (2012) Evolution of C4 plants: A new hypothesis for an interaction of CO2 and water relations mediated by plant hydraulics. Philosophical Transactions of the Royal Society B: Biological Sciences 367: 583–600.

[24] Christin P, Besnard G, Samaritani E, Duvall M, Hodkinson T, et al. (2008) Oligocene CO2 decline promoted C4 photosynthesis in grasses. Current Biology 18: 37–43.

[25] von Caemmerer S (2000) Biochemical models of leaf photosynthesis. Collingwood: CSIRO Publishing.

[26] Zhu XG, Sturler Ed, Long SP (2007) Optimizing the distribution of resources between enzymes of carbon metabolism can dramatically increase photosynthetic rate: A numerical simulation using an evolutionary algorithm. Plant Physiology 145: 513–526.

[27] Wang Y, Long SP, Zhu XG (2014) Elements required for an efficient NADP-ME type C4 photosynthesis. Plant Physiology 164: 2231–2246.

[28] Wang Y, Bräutigam A, Weber APM, Zhu XG (2014) Three distinct biochemical subtypes of C4 photosynthesis? A modelling analysis. Journal of Experimental Botany 65: 3567–3578.

[29] Stitt M, Zhu XG (2014) The large pools of metabolites involved in intercellular metabolite shuttles in C4 photosynthesis provide enormous flexibility and robustness in a fluctuating light environment. Plant, Cell & Environment 37: 1985–1988.

[30] Bordbar A, Monk JM, King ZA, Palsson BO (2014) Constraint-based models predict metabolic and associated cellular functions. Nature Reviews Genetics 15: 107–120.

[31] Arabidopsis Genome Initiative (2000) Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 408: 796–815.

[32] Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, et al. (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae rd. Science 269: 496–512.

[33] Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, et al. (1996) Life with 6000 genes. Science 274: 546–567.

[34] Grafahrend-Belau E, Schreiber F, Koschützki D, Junker BH (2009) Flux balance analysis of barley seeds: A computational approach to study systemic properties of central metabolism. Plant Physiology 149: 585–598.

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

[36] Williams TC, Poolman MG, Howden AJ, Schwarzlander M, Fell DA, et al. (2010) A genome-scale metabolic model accurately predicts fluxes in central carbon metabolism under stress conditions. Plant Physiol 154: 311–323.

[37] de Oliveira Dal'Molin CG, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) AraGEM, a genome-scale reconstruction of the primary metabolic network in Arabidopsis. Plant Physiology 152: 579–589.

[38] Radrich K, Tsuruoka Y, Dobson P, Gevorgyan A, Swainston N, et al. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks. BMC Systems Biology 4: 114.

[39] Gomes de Oliveira Dal'Molin C, Quek LE, Palfreyman RW, Brumbley SM, Nielsen LK (2010) C4GEM - genome-scale metabolic model to study C4 plant metabolism. Plant Physiology 154: 1871–1885.

[40] Pilalis E, Chatziioannou A, Thomasset B, Kolisis F (2011) An in silico compartmentalized metabolic model of Brassica napus enables the systemic study of regulatory aspects of plant central metabolism. Biotechnology and Bioengineering 108: 1673–1682.

[41] Saha R, Suthers PF, Maranas CD (2011) Zea mays iRS1563: A comprehensive genome-scale metabolic reconstruction of maize metabolism. PLoS ONE 6: e21784.

[42] Hay J, Schwender J (2011) Metabolic network reconstruction and flux variability analysis of storage synthesis in developing oilseed rape (Brassica napus L.) embryos. The Plant Journal 67: 526–541.

[43] Hay J, Schwender J (2011) Computational analysis of storage synthesis in developing Brassica napus L. (oilseed rape) embryos: flux variability analysis in relation to 13C metabolic flux analysis. The Plant Journal 67: 513–525.

[44] Mintz-Oron S, Meir S, Malitsky S, Ruppin E, Aharoni A, et al. (2012) Reconstruction of Arabidopsis metabolic network models accounting for subcellular compartmentalization and tissue-specificity. Proceedings of the National Academy of Sciences 109: 339–344.

[45] Poolman MG, Kundu S, Shaw R, Fell DA (2013) Responses to light intensity in a genome-scale model of rice metabolism. Plant Physiology 162: 1060–1072.

[46] Cheung CYM, Poolman MG, Fell DA, Ratcliffe RG, Sweetlove LJ (2014) A diel flux balance model captures interactions between light and dark metabolism during day-night cycles in C3 and crassulacean acid metabolism leaves. Plant Physiology 165: 917–929.

[47] Cheung CYM, Williams TCR, Poolman MG, Fell DA, Ratcliffe RG, et al. (2013) A method for accounting for maintenance costs in flux balance analysis improves the prediction of plant cell metabolic phenotypes under stress conditions. The Plant Journal 75: 1050–1061.

[48] Simons M, Saha R, Amiour N, Kumar A, Guillard L, et al. (2014) Assessing the metabolic impact of nitrogen availability using a compartmentalized maize leaf genome-scale model. Plant Physiology 166: 1659–1674.

[49] Grafahrend-Belau E, Junker A, Eschenröder A, Müller J, Schreiber F, et al. (2013) Multiscale metabolic modeling: Dynamic flux balance analysis on a whole-plant scale. Plant Physiology 163: 637–647.

[50] Orth JD, Thiele I, Palsson BO (2010) What is flux balance analysis? Nature Biotechnology 28: 245–248.

[51] Boyd S, Vandenberghe L (2004) Convex Optimization. Cambridge: Cambridge University Press.

[52] Heckmann D, Schulze S, Denton A, Gowik U, Westhoff P, et al. (2013) Predicting C4 photosynthesis evolution: Modular, individually adaptive steps on a Mount Fuji fitness landscape. Cell 153: 1579–1588.

[53] Wächter A, Biegler LT (2006) On the implementation of an interior-point filter line-search algorithm for large-scale nonlinear programming. Mathematical Programming 106: 25–57.

[54] Hucka M, Finney A, Sauro HM, Bolouri H, Doyle JC, et al. (2003) The

systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. Bioinformatics 19: 524–531.

[55] Li P, Ponnala L, Gandotra N, Wang L, Si Y, et al. (2010) The developmental dynamics of the maize leaf transcriptome. Nature Genetics 42: 1060–1067.

[56] Nelson T (2011) The grass leaf developmental gradient as a platform for a systems understanding of the anatomical specialization of C4 leaves. Journal of Experimental Botany 62: 3039 –3048.

[57] Plant Metabolic Network (PMN) (2013). CornCyc 4.0. `http://pmn.plantcyc.org/CORN/organism-summary` on `www.plantcyc.org`.

[58] Sun Q, Zybailov B, Majeran W, Friso G, Olinares PDB, et al. (2009) PPDB, the plant proteomics database at Cornell. Nucleic Acids Research 37: D969–974.

[59] Reed J, Vo T, Schilling C, Palsson B (2003) An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). Genome Biology 4: R54.

[60] Bogart E, Myers CR (2015) Multiscale metabolic modeling of C4 plants: connecting nonlinear genome-scale models to leaf-scale metabolism in developing maize leaves. arXiv:1502.07969 [q-bio.MN] .

[61] Xu E (2011). Pyipopt. `http://github.com/xuy/pyipopt`.

[62] Wang L, Czedik-Eysenberg A, Mertz RA, Si Y, Tohge T, et al. (2014) Comparative analyses of C4 and C3 photosynthesis in developing leaves of maize and rice. Nature Biotechnology 32: 1158-1165.

[63] Tausta SL, Li P, Si Y, Gandotra N, Liu P, et al. (2014) Developmental dynamics of Kranz cell transcriptional specificity in maize leaf reveals early onset of C4-related processes. Journal of Experimental Botany 65: 3543–3555.

[64] Lee D, Smallbone K, Dunn WB, Murabito E, Winder CL, et al. (2012) Improving metabolic flux predictions using absolute gene expression data. BMC Systems Biology 6: 73.

[65] Barker B, Sadagopan N, Wang Y, Smallbone K, Myers CR, et al. (2014) A

robust and efficient method for estimating enzyme complex abundance and metabolic flux from expression data. arXiv:1404.4755 [q-bio.MN] .

[66] Mahadevan R, Schilling C (2003) The effects of alternate optimal solutions in constraint-based genome-scale metabolic models. Metabolic Engineering 5: 264–276.

[67] Bellasio C, Griffiths H (2014) Acclimation to low light by C4 maize: implications for bundle sheath leakiness. Plant, Cell & Environment 37: 1046–1058.

[68] Hatch MD (1987) C4 photosynthesis: a unique blend of modified biochemistry, anatomy and ultrastructure. Biochimica et Biophysica Acta (BBA) - Reviews on Bioenergetics 895: 81–106.

[69] Majeran W, Friso G, Ponnala L, Connolly B, Huang M, et al. (2010) Structural and metabolic transitions of C4 leaf development and differentiation defined by microscopy and quantitative proteomics in maize. The Plant Cell 22: 3509–3542.

[70] Wingler A, Walker RP, Chen ZH, Leegood RC (1999) Phosphoenolpyruvate carboxykinase is involved in the decarboxylation of aspartate in the bundle sheath of maize. Plant Physiology 120: 539–546.

[71] Ponnala L, Wang Y, Sun Q, van Wijk KJ (2014) Correlation of mRNA and protein abundance in the developing maize leaf. The Plant Journal 78: 424–440.

[72] Colijn C, Brandes A, Zucker J, Lun DS, Weiner B, et al. (2009) Interpreting expression data with metabolic flux models: Predicting Mycobacterium tuberculosis mycolic acid production. PLoS Comput Biol 5: e1000489.

[73] Kanehisa M, Goto S, Sato Y, Kawashima M, Furumichi M, et al. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. Nucleic Acids Research 42: D199–205.

[74] Kanehisa M, Goto S (2000) KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Research 28: 27–30.

[75] Plant Metabolic Network (PMN) (2014). Enzyme functional annotation method. http://www.plantcyc.org/about/databases_

overview.faces#e2p2 on www.plantcyc.org. Accessed October 16, 2014.

[76] Sandve GK, Nekrutenko A, Taylor J, Hovig E (2013) Ten simple rules for reproducible computational research. PLoS Computational Biology 9: e1003285.

[77] Latendresse M, Krummenacker M, Trupp M, Karp PD (2012) Construction and completion of flux balance models from pathway databases. Bioinformatics 28: 388–396.

[78] Thiele I, Palsson BO (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. Nature Protocols 5: 93–121.

[79] Karp PD, Paley SM, Krummenacker M, Latendresse M, Dale JM, et al. (2010) Pathway tools version 13.0: integrated software for pathway/genome informatics and systems biology. Briefings in Bioinformatics 11: 40–79.

[80] Plant Metabolic Network (PMN) (2014). PMN database content statistics. http://www.plantcyc.org/release_notes/content_statistics.faces on www.plantcyc.org. Accessed January 13, 2015.

[81] Fernandez-Pozo N, Menda N, Edwards JD, Saha S, Tecle IY, et al. (2014) The sol genomics network (SGN)—from genotype to phenotype to breeding. Nucleic Acids Research : D1149-D1155

[82] Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, et al. (2014) Gramene 2013: comparative plant genomics resources. Nucleic Acids Research 42: D1193–D1199.

[83] Urbanczyk-Wochniak E, Sumner LW (2007) MedicCyc: a biochemical pathway database for medicago truncatula. Bioinformatics 23: 1418–1423.

[84] Naithani S, Raja R, Waddell EN, Elser J, Gouthu S, et al. (2014) VitisCyc: a metabolic pathway knowledgebase for grapevine (Vitis vinifera). Frontiers in Plant Science 5: 644.

[85] Jung S, Ficklin SP, Lee T, Cheng CH, Blenda A, et al. (2014) The genome

database for rosaceae (GDR): year 10 update. Nucleic Acids Research 42: D1237–D1244.

[86] Chang RL, Ghamsari L, Manichaikul A, Hom EFY, Balaji S, et al. (2011) Metabolic network reconstruction of chlamydomonas offers insight into light-driven algal metabolism. Molecular Systems Biology 7: 518.

[87] Mahadevan R, Edwards JS, Doyle FJ (2002) Dynamic flux balance analysis of diauxic growth in Escherichia coli. Biophysical Journal 83: 1331–1340.

[88] Smallbone K, Simeonidis E, Broomhead DS, Kell DB (2007) Something from nothing - bridging the gap between constraint-based and kinetic modelling. FEBS Journal 274: 5576–5585.

[89] Jamshidi N, Palsson BØ (2010) Mass action stoichiometric simulation models: Incorporating kinetics and regulation into stoichiometric models. Biophysical Journal 98: 175–185.

[90] Feng X, Xu Y, Chen Y, Tang YJ (2012) Integrating flux balance analysis into kinetic models to decipher the dynamic metabolism of Shewanella oneidensis MR-1. PLoS Computational Biology 8: e1002376.

[91] Cotten C, Reed JL (2013) Mechanistic analysis of multi-omics datasets to generate kinetic parameters for constraint-based metabolic models. BMC Bioinformatics 14: 32.

[92] Chowdhury A, Zomorrodi AR, Maranas CD (2014) k-OptForce: Integrating kinetics with flux balance analysis for strain design. PLoS Computational Biology 10: e1003487.

[93] Tan Y, Lafontaine Rivera JG, Contador CA, Asenjo JA, Liao JC (2011) Reducing the allowable kinetic space by constructing ensemble of dynamic models with the same steady-state flux. Metabolic Engineering 13: 60–75.

[94] Hoppe A (2012) What mRNA abundances can tell us about metabolism. Metabolites 2: 614–631.

[95] Schwender J, König C, Klapperstück M, Heinzel N, Munz E, et al. (2014) Transcript abundance on its own cannot be used to infer fluxes in central metabolism. Frontiers in Plant Science 5:668.

[96] Machado D, Herrgård M (2014) Systematic evaluation of methods

for integration of transcriptomic data into constraint-based models of metabolism. PLoS Computational Biology 10: e1003580.

[97] Lewis NE, Schramm G, Bordbar A, Schellenberger J, Andersen MP, et al. (2010) Large-scale in silico modeling of metabolic interactions between cell types in the human brain. Nature Biotechnology 28: 1279–1285.

[98] Salimi F, Zhuang K, Mahadevan R (2010) Genome-scale metabolic modeling of a clostridial co-culture for consolidated bioprocessing. Biotechnology Journal 5: 726–738.

[99] Zhuang K, Izallalen M, Mouser P, Richter H, Risso C, et al. (2011) Genome-scale dynamic modeling of the competition between Rhodoferax and Geobacter in anoxic subsurface environments. The ISME Journal 5: 305–316.

[100] Zomorrodi AR, Maranas CD (2012) OptCom: A multi-level optimization framework for the metabolic modeling and analysis of microbial communities. PLoS Comput Biol 8: e1002363.

[101] Zengler K, Palsson BO (2012) A road map for the development of community systems (CoSy) biology. Nature Reviews Microbiology 10: 366–372.

[102] Khandelwal RA, Olivier BG, Röling WFM, Teusink B, Bruggeman FJ (2013) Community flux balance analysis for microbial consortia at balanced growth. PLoS ONE 8: e64567.

[103] Chiu HC, Levy R, Borenstein E (2014) Emergent biosynthetic capacity in simple microbial communities. PLoS Computational Biology 10: e1003695.

[104] Zomorrodi AR, Islam MM, Maranas CD (2014) d-OptCom: Dynamic multi-level and multi-objective metabolic modeling of microbial communities. ACS Synthetic Biology 3: 247–257.

[105] Stolyar S, Dien SV, Hillesland KL, Pinel N, Lie TJ, et al. (2007) Metabolic modeling of a mutualistic microbial community. Molecular Systems Biology 3: 92.

[106] Bordbar A, Lewis NE, Schellenberger J, Palsson BØ, Jamshidi N (2010) Insight into human alveolar macrophage and m. tuberculosis interactions via metabolic reconstructions. Molecular Systems Biology 6: 422.

[107] Heinken A, Sahoo S, Fleming RMT, Thiele I (2013) Systems-level characterization of a host-microbe metabolic symbiosis in the mammalian gut. Gut Microbes 4: 28–40.

[108] Becker SA, Feist AM, Mo ML, Hannum G, Palsson BØ, et al. (2007) Quantitative prediction of cellular metabolism with constraint-based models: the COBRA toolbox. Nature Protocols 2: 727–738.

[109] Weiner H, Burnell JN, Woodrow IE, Heldt HW, Hatch MD (1988) Metabolite diffusion into bundle sheath cells from C4 plants relation to C4 photosynthesis and plasmodesmatal function. Plant Physiology 88: 815–822.

[110] Sowiński P, Szczepanik J, Minchin PEH (2008) On the mechanism of C4 photosynthesis intermediate exchange between Kranz mesophyll and bundle sheath cells in grasses. Journal of Experimental Botany 59: 1137–1147.

[111] Ohshima T, Hayashi H, Chino M (1990) Collection and chemical composition of pure phloem sap from Zea mays L. Plant and Cell Physiology 31: 735–737.

[112] Bourgis F, Roje S, Nuccio ML, Fisher DB, Tarczynski MC, et al. (1999) S-methylmethionine plays a major role in phloem sulfur transport and is synthesized by a novel type of methyltransferase. The Plant Cell 11: 1485–1497.

[113] Kromdijk J, Griffiths H, Schepers HE (2010) Can the progressive increase of C4 bundle sheath leakiness at low PFD be explained by incomplete suppression of photorespiration? Plant, Cell & Environment 33: 1935–1948.

[114] Bornstein BJ, Keating SM, Jouraku A, Hucka M (2008) LibSBML: an API library for SBML. Bioinformatics 24: 880–881.

[115] Gutenkunst RN, Atlas JC, Casey FP, Daniels BC, Kuczenski RS, et al. (2007). Sloppycell. `http://sloppycell.sourceforge.net`.

[116] Myers C, Gutenkunst R, Sethna J (2007) Python unleashed on systems biology. Computing in Science and Engineering 9: 34–37.

[117] HSL (2013). A collection of fortran codes for large scale scientific computation. `http://www.hsl.rl.ac.uk`.

[118] GLPK (2011). GNU linear programming kit, version 4.47. http://www.gnu.org/software/glpk/glpk.html.

[119] Finley T (2008). pyglpk. http://tfinley.net/software/pyglpk.

[120] Ludwig M (2013) Evolution of the C4 photosynthetic pathway: events at the cellular and molecular levels. Photosynthesis Research 117: 147–161.

[121] Brown NJ, Newell CA, Stanley S, Chen JE, Perrin AJ, et al. (2011) Independent and parallel recruitment of preexisting mechanisms underlying C4 photosynthesis. Science 331: 1436 –1439.

[122] Savir Y, Noor E, Milo R, Tlusty T (2010) Cross-species analysis traces adaptation of rubisco toward optimality in a low-dimensional landscape. Proceedings of the National Academy of Sciences of the United States of America 107: 3475–3480.

[123] Beer KD, Orellana MV, Baliga NS (2013) Modeling the evolution of C4 photosynthesis. Cell 153: 1427–1429.

[124] Aubry S, Brown NJ, Hibberd JM (2011) The role of proteins in C3 plants prior to their recruitment into the C4 pathway. Journal of Experimental Botany 62: 3049 –3059.

[125] Leegood RC (2008) Roles of the bundle sheath cells in leaves of C3 plants. Journal of Experimental Botany 59: 1663–1673.

[126] Plant Metabolic Network (PMN) (2015). *Zea mays* mays pathway: glutamate biosynthesis III. http://pmn.plantcyc.org/CORN/NEW-IMAGE?type=PATHWAY&object=GLUTSYNIII-PWY on www.plantcyc.org. Accessed March 29, 2015.

[127] Elthon TE, Stewart CR (1982) Proline oxidation in corn mitochondria: involvement of NAD, relationship to ornithine metabolism, and sidedness on the inner membrane. Plant Physiology 70: 567–572.

[128] Kacser H, Burns J (1973) The control of flux. Symposia of the Society for Experimental Biology 27: 65–104.

[129] Heinrich R, Rapoport TA (1974) A linear steady-state treatment of enzymatic chains. European Journal of Biochemistry 42: 89–95.

[130] Sheppard D, Terrell R, Henkelman G (2008) Optimization methods for finding minimum energy paths. The Journal of Chemical Physics 128: 134106.

[131] Jónsson H, Mills G, Jacobsen K (1998) Nudged elastic band method for finding minimum energy paths of transitions. In: Berne B, Ciccoti G, Coker D, editors, Classical and Quantum Dynamics in Condensed Phase Simulations, Singapore: World Scientific. pp. 385–404.

[132] Griffiths H, Weller G, Toy LFM, Dennis RJ (2013) You're so vein: bundle sheath physiology, phylogeny and evolution in C3 and C4 plants. Plant, Cell & Environment 36: 249–261.

[133] Jones E, Oliphant T, Peterson P, et al. (2001–). SciPy: Open source scientific tools for Python. `http://www.scipy.org/`.

[134] E W, Ren W, Vanden-Eijnden E (2005) Finite temperature string method for the study of rare events. J Phys Chem B 109: 6688–6693.

[135] Ogle K (2003) Implications of interveinal distance for quantum yield in C4 grasses: a modeling and meta-analysis. Oecologia 136: 532–542.

[136] Chakrabarti A, Miskovic L, Soh KC, Hatzimanikatis V (2013) Towards kinetic modeling of genome-scale metabolic networks without sacrificing stoichiometric, thermodynamic and physiological constraints. Biotechnology Journal 8: 1043–1057.

[137] SRI International, Menlo Park, CA (2013) Pathway Tools v. 17.0 user manual.

[138] Brownleader M, Harborne J, Dey P (1997) Carbohydrate metabolism: primary metabolism of polysaccharides. In: Dey P, Harborne J, editors, Plant biochemistry, San Diego: Academic Press. pp. 111-142.

[139] Allen JF (2003) Cyclic, pseudocyclic and noncyclic photophosphorylation: new links in the chain. Trends in Plant Science 8: 15–19.

[140] Asada K (1999) The water-water cycle in chloroplasts: Scavenging of active oxygens and dissipation of excess photons. Annual Review of Plant Physiology and Plant Molecular Biology 50: 601–639.

[141] Foyer C, Harbinson J (1997) The photosythetic electron transport system:

efficiency and control. In: Foyer C, Quick W, editors, A Molecular Approach to Primary Metabolism in Higher Plants, London: Taylor and Francis.

[142] Fettke J, Hejazi M, Smirnova J, Höchel E, Stage M, et al. (2009) Eukaryotic starch degradation: integration of plastidial and cytosolic pathways. Journal of Experimental Botany 60: 2907–2922.

[143] Streb S, Zeeman SC (2012) Starch metabolism in Arabidopsis. The Arabidopsis Book : e0160.

[144] Li-Beisson Y, Shorrosh B, Beisson F, Andersson MX, Arondel V, et al. (2013) Acyl-lipid metabolism. The Arabidopsis Book : e0161.

[145] Ohlrogge J, Browse J (1995) Lipid biosynthesis. The Plant Cell 7: 957–970.

[146] Dörmann P, Benning C (1998) The role of UDP-glucose epimerase in carbohydrate metabolism of Arabidopsis. The Plant Journal 13: 641–652.

[147] BRENDA (2013). Information on EC 5.1.3.2 - UDP-glucose 4-epimerase. http://brenda-enzymes.org/enzyme.php?ecno=5.1.3.2. Accessed October 9, 2013.

[148] Plant Metabolic Network (PMN) (2014). *Arabidopsis thaliana* col pathway: palmitoleate biosynthesis II. http://pmn.plantcyc.org/ARA/NEW-IMAGE?type=PATHWAY&object=PWY-5366 on www.plantcyc.org. Accessed October 16, 2014.

[149] Gibson KJ (1993) Palmitoleate formation by soybean stearoyl-acyl carrier protein desaturase. Biochimica Et Biophysica Acta 1169: 231–235.

[150] Sperling P, Schmidt H, Heinz E (1995) A cytochrome-b5-containing fusion protein similar to plant acyl lipid desaturases. European Journal of Biochemistry 232: 798–805.

[151] Harwood JL (1996) Recent advances in the biosynthesis of plant fatty acids. Biochimica et Biophysica Acta (BBA) - Lipids and Lipid Metabolism 1301: 7–56.

[152] Shanklin J, Cahoon EB (1998) Desaturation and related modifications of fatty acids. Annual Review of Plant Physiology and Plant Molecular Biology 49: 611–641.

[153] Reumann S, Weber APM (2006) Plant peroxisomes respire in the light: Some gaps of the photorespiratory C2 cycle have become filled—others remain. Biochimica et Biophysica Acta (BBA) - Molecular Cell Research 1763: 1496–1510.

[154] Foyer CH, Bloom AJ, Queval G, Noctor G (2009) Photorespiratory metabolism: Genes, mutants, energetics, and redox signaling. Annual Review of Plant Biology 60: 455–484.

[155] Weber AP, Linka N (2011) Connecting the plastid: Transporters of the plastid envelope and their role in linking plastidial with cytosolic metabolism. Annual Review of Plant Biology 62: 53–77.

[156] Bräutigam A, Weber APM (2011) Chapter 11 transport processes: Connecting the reactions of C4 photosynthesis. In: Raghavendra AS, Sage RF, editors, C4 Photosynthesis and Related CO2 Concentrating Mechanisms, Dordrecht: Springer, pp. 199–219.

[157] Hanson AD, Roje S (2001) One-carbon metabolism in higher plants. Annual Review of Plant Physiology and Plant Molecular Biology 52: 119–137.

[158] Bartoli CG, Pastori GM, Foyer CH (2000) Ascorbate biosynthesis in mitochondria is linked to the electron transport chain between complexes III and IV. Plant Physiology 123: 335–344.

[159] Foyer C, Rowell J, Walker D (1983) Measurement of the ascorbate content of spinach leaf protoplasts and chloroplasts during illumination. Planta 157: 239–244.

[160] Smirnoff N (1996) The function and metabolism of ascorbic acid in plants. Annals of Botany 78: 661–669.

[161] Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their n-terminal amino acid sequence. Journal of Molecular Biology 300: 1005–1016.

[162] Munekage Y, Hashimoto M, Miyake C, Tomizawa KI, Endo T, et al. (2004) Cyclic electron flow around photosystem I is essential for photosynthesis. Nature 429: 579–582.

[163] Shikanai T (2007) Cyclic electron transport around photosystem I: Genetic approaches. Annual Review of Plant Biology 58: 199–217.

[164] Takabayashi A, Kishine M, Asada K, Endo T, Sato F (2005) Differential use of two cyclic electron flows around photosystem I for driving CO2-concentration mechanism in C4 photosynthesis. Proceedings of the National Academy of Sciences of the United States of America 102: 16898–16903.

[165] Marchler-Bauer A, Bryant SH (2004) CD-search: protein domain annotations on the fly. Nucleic Acids Research 32: W327–W331.

[166] Lopez-Calcagno PE, Howard TP, Raines CA (2014) The CP12 protein family: a thioredoxin-mediated metabolic switch? Frontiers in Plant Science 5:9.

[167] Rizov I, Doulis A (2000) Determination of glycerolipid composition of rice and maize tissues using solid-phase extraction. Biochemical Society Transactions 28: 586–589.

[168] Leech RM, Rumsby MG, Thomson WW (1973) Plastid differentiation, acyl lipid, and fatty acid changes in developing green maize leaves. Plant Physiology 52: 240–245.

[169] Plant Metabolic Network (PMN) (2014). *Zea mays* mays pathway: homogalacturonan biosynthesis. `http://pmn.plantcyc.org/CORN/NEW-IMAGE?type=PATHWAY&object=PWY-1061` on `www.plantcyc.org`. Accessed October 16, 2014.

[170] Herold A, Lewis DH (1977) Mannose and green plants: Occurrence, physiology and metabolism, and use as a tool to study the role of orthophosphate. New Phytologist 79: 1–40.

[171] Schnarrenberger C (1990) Characterization and compartmentation, in green leaves, of hexokinases with different specificities for glucose, fructose, and mannose and for nucleoside triphosphates. Planta 181: 249–255.

[172] Plant Metabolic Network (PMN) (2014). PlantCyc pathway: D-mannose degradation. `http://pmn.plantcyc.org/PLANT/new-image?object=MANNCAT-PWY` on `www.plantcyc.org`. Accessed October 16, 2014.

[173] Franceschi VR, Loewus FA (1995) Oxalate function and biosynthesis in plants and fungi. In: Khan SR, editor, Calcium oxalate in biological systems, CRC Press.

[174] Franceschi VR, Nakata PA (2005) Calcium oxalate in plants: Formation and function. Annual Review of Plant Biology 56: 41–71.

[175] Debolt S, Melino V, Ford CM (2007) Ascorbate as a biosynthetic precursor in plants. Annals of Botany 99: 3–8.

[176] Lane BG, Dunwell JM, Ray JA, Schmitt MR, Cuming AC (1993) Germin, a protein marker of early plant development, is an oxalate oxidase. Journal of Biological Chemistry 268: 12239–12242.

[177] Plant Metabolic Network (PMN) (2014). PlantCyc Reaction: 3.7.1.1. http://pmn.plantcyc.org/PLANT/NEW-IMAGE?type=REACTION-IN-PATHWAY&object=OXALOACETASE-RXN on www.plantcyc.org. Accessed October 16, 2014.

[178] Hayaishi O, Shimazono H, Katagiri M, Saito Y (1956) Enzymatic formation of oxalate and acetate from oxaloacetate. Journal of the American Chemical Society 78: 5126–5127.

[179] Burgener M, Suter M, Jones S, Brunold C (1998) Cyst(e)ine is the transport metabolite of assimilated sulfur from bundle-sheath to mesophyll cells in maize leaves. Plant Physiology 116: 1315–1322.

[180] Bar-Even A, Noor E, Lewis NE, Milo R (2010) Design and analysis of synthetic carbon fixation pathways. Proceedings of the National Academy of Sciences 107: 8889–8894.

[181] Kebeish R, Niessen M, Thiruveedhi K, Bari R, Hirsch HJ, et al. (2007) Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in Arabidopsis thaliana. Nature Biotechnology 25: 593–599.

[182] Nocedal J, Wright SJ (2006) Numerical optimization. New York: Springer.

[183] Biegler LT (2010) Nonlinear programming: concepts, algorithms, and applications to chemical processes. Philadelphia: Society for Industrial and Applied Mathematics.

[184] Kawajir Y, Laird C, Vigerske S, Wächter A. Introduction to IPOPT: A tutorial for downloading, installing, and using IPOPT. http://www.coin-or.org/Ipopt/documentation/.

[185] Sauro HM, Ingalls B (2004) Conservation analysis in biochemical networks: computational issues for software writers. Biophysical Chemistry 109: 1–15.